Transcriptional Control During Hematopoietic Development

Transcription factor binding and chromatin conformation dynamics

Anita van den Heuvel

ISBN: 978-94-6295-088-7

The work presented in this thesis was performed at the Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands. The department is a member of the *Medisch Genetisch Centrum Zuid-West Nederland*

Cover design: Anita van den Heuvel

Printed by: Uitgeverij BOXPress | | Proefschriftmaken.nl

Transcriptional Control During Hematopoietic Development

Transcription factor binding and chromatin conformation dynamics

Transcriptionele regulatie tijdens hematopoietische ontwikkeling

Transcriptie factor binding en chromatine conformatie dynamiek

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
woensdag 4 februari 2015 om 13.30 uur

Door

Anita van den Heuvel

Geboren te Rotterdam

Promotiecommissie:

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

Contents

Scope of this thesis

A cell's identity is primarily determined by the proteins it produces and therefore by the genes it expresses. During development, correct cell fate specification and determination therefore requires a strictly controlled upregulation or downregulation of lineage-specific gene expression. The experimental work described in this thesis aimed to contribute to unraveling the complex process of such gene expression regulation. The first part of **Chapter 1** gives a general introduction on the mechanism of gene expression regulation, with a main focus on the first step in this process, i.e. gene transcription. This first part is followed by a general introduction on hematopoietic development and differentiation, which is a widely used model system for studying gene expression regulation during cellular differentiation. This part briefly discusses the origin of hematopoietic stem cells (HSCs) and their terminal differentiation towards the different blood cell lineages, with a specific focus on erythropoiesis.

HSCs are known to originate in the Aorta-Gonad-Mesonephros region of the embryo, where they derive from specialized intra-aortic hematopoietic clusters (IAHCs). However, at present, the precise mechanism of HSC specification in these clusters of cells is not known. **Chapter 2** describes the results of our study on the role of the IAHC cells in the development of HSCs.

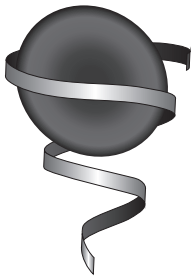
Chromatin structural conformation plays a central role in gene expression regulation, with multiple regulatory elements being located at far distance from their target genes, and often regulating gene expression via the formation of chromatin loops. **Chapter 3** of this thesis discusses the role of distal regulatory elements and chromatin conformation in regulating gene transcription. In addition, it discusses the role of genomic alterations in our noncoding 'regulatory' genome in disease susceptibility and etiology, and discusses several novel therapeutic strategies to target such molecular diseases. **Chapter 4** describes the optimization of the chromatin conformation capture (3C) technology for use in semi-high throughput multiplexed next-generation sequencing analyses. Multiplexed 3C-seq, provides a tool for analyzing the fine-scale chromatin structural conformation and can be used to study the role of chromatin structure on gene-specific expression regulation.

Chapter 5 describes the results of our study on the transcriptional control of murine *Bcl11a*, with a specific focus on the role of the erythroid-specific LDB1 transcription factor complex in this regulation. Data from this study may contribute to unraveling the different mechanisms of transcriptional control used to ensure correct and accurate transcription levels and patterns during hematopoiesis. Moreover, as BCL11a itself is known to be a repressor of *γ-globin* expression, *Bcl11a* repression has become an interesting therapeutic strategy for β-hemoglobinopathies in which elevated γ-globin levels have an ameliorating effect on disease phenotypes. Data from this study may therefore provide new avenues for therapeutic treatment of these globin-related diseases.

Finally, **Chapter 6** presents a general discussion on the experimental data presented in this thesis. It highlights important findings and aims to place this data in a broader perspective by discussing its contribution to the general understanding of the mechanism of gene transcriptional control. It discusses how these results can contribute to the ultimate goal of unraveling the complete mechanisms of hematopoietic development and how these may contribute to the development of new treatments for hematological diseases.

Chapter 1

General Introduction



DNA; the handbook of the cell

The human body contains approximately 3.72 x 10^{13} cells[1], which can be subdivided into more than 400 different cell types, each type having its own function and cellular characteristics.[2] These trillions of cells all originate from a single cell, the fertilized egg, which during the very earliest stages of life proliferates (multiplies) and differentiates (changes) into all the different cell types in the human fetus and adult body. This single diploid cell therefore needs to hold all information for the development and functioning of the complete organism. This information is stored in our DNA[3–6], the genetic material we inherit from our parents in the form of chromosomes. Humans have 46 chromosomes, comprising two copies of 23 chromosomes, one received from each parent (22 autosomes and one of the two different sex chromosomes). The complete set of 46 chromosomes (complemented with some mitochondrial DNA of maternal origin) provides the cell with all information needed for its correct development and functioning. One may therefore think of our genome as the handbook of the cell.

DNA is composed of two long polynucleotide chains, composed of only four types of nucleotides, adenine (A), cytosine (C), guanine (G) and thymine (T), which are wrapped around each other in a left-handed, anti-parallel double helix structure (Figure 1).[7] The two polynucleotide chains are held together by hydrogen bonds between nucleotides of each chain, in which (A) always pairs with (T) and (C) always pairs with (G). A dimer of two nucleotides interacting together is called a base pair (bp). The four nucleotides of this polynucleotide chain form the four-letter code of the DNA. Clues on how such a simple four-letter code can encompass all the information for the development of a multicellular organism, as complex as our human body, came from the discovery that parts of the DNA code for proteins, key structural and functional building blocks of every cell.[8,9] A piece of DNA that codes for the formation of a protein is called a gene. (Of note, later also non-protein-coding genes have been found to exist, which play important roles as well.[10,11] Therefore the definition of a gene has been changed to include not only protein-coding regions, but all DNA sequences coding for any functional molecule, i.e. proteins or RNA molecules.[12]). Since the discovery of DNA as the hereditary material, scientists all over the world have been working on deciphering the complete human genomic sequence, and at present more than 99% of the sequence is known.[13–15] A complete set of nuclear DNA, called our genome[16], is composed of approximately 3 billion base pairs and contains roughly 22,000 protein-coding genes.[14,15]

Cellular divergence; differences in protein production

However, as an exact copy of the genome, encoding for the same 22,000 proteins, is located in almost every cell in the human body, what makes these 400 different cell types so different? This is mainly due to the fact that not all 22,000 proteins are produced in every cell. It is estimated that a typical human cell produces only 30-60% of all 22,000 proteins at a given point in time.[17] Each cell type produces a unique set of proteins to acquire its unique function and cellular characteristics, and even within the same cell type heterogeneity in expression of these genes has be documented.[11,18,19] Without this specification in protein production every cell would produce the same 22,000 proteins and hence all cells would remain identical. It is the dynamics in protein production (and functional molecules produced from non-protein-coding genes) that causes the fertilized egg to suddenly differentiate into different cell types, with different functions and cellular characteristics. Simply said, it is the difference in protein and non-protein-coding RNA levels that distinguishes the skin cells on our arm from the nerve cells in our brain.

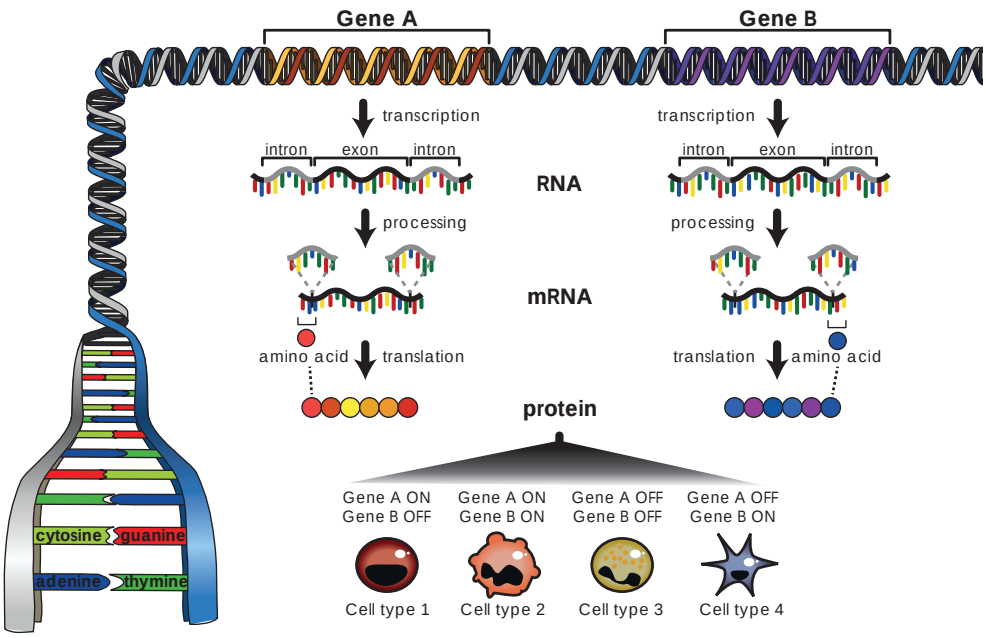


Figure 1. Schematic overview of the mechanism of gene expression. The DNA double helix is composed of 2 polynucleotide strands, composed of 4 different nucleotides, wrapped around eachother in a left-handed double helix. Gene-coding regions are transcribed into RNA by specialized proteins. This RNA is processed by removing all intronic sequences resulting in mRNA, which is translated into proteins by ribosomes. The unique combination of expressed genes, determines each cell's unique cellular identity and function.

Gene expression; protein production

Regulating the production of a specific set of proteins in each cell or cell type is a complex process. In order to produce a protein, the corresponding gene encoding for the protein, needs to be 'expressed', which means that the information encoded in the gene needs to be 'read' and 'translated' into a protein. In short, this gene expression occurs in 3 different steps (Figure 1).[17] First, specialized proteins 'read' the DNA in the gene and transcribe (copy) the code from one of the DNA strands (referred to as the sense strand) into a different kind of nucleic acid, called ribonucleic acid (RNA). This RNA subsequently goes through a maturation process to give rise to a so-called messenger RNA (mRNA) molecule that is mostly comprised of the protein-coding parts of the gene (called exons). The mRNA molecule is then transported out of the nucleus where specialized protein-RNA complexes, called ribosomes, 'read' the mRNA and translate it into a polypeptide which forms the protein. As there are only four different nucleotides in the DNA sequence and 20 different amino acids in a protein, this translation is not a simple one-to-one translation. Instead, ribosomes recognize triplets of nucleotides, called codons. This significantly increases the complexity of the information encoded in the DNA molecule as 64 (4x4x4) different triplet codons can be formed by combining the four nucleotides in different combinations (61 codons encoding for amino acids and 3 encoding a stop signal for translation.[8,9]

The complete process of gene expression, including transcription, RNA processing and translation is a highly complex process which is controlled at various different stages, ranging from gene transcription and RNA processing rates, to RNA and protein stability and degradation.[17] For this introduction I focus on the most upstream part of this process, gene transcription regulation. Gene transcription is regulated by a specialized group of proteins, called transcription factors (TFs). TFs were first introduced by Monod

and Jacob in 1961 (although according to many, Barbara McClintock deserves a considerable share of the credit due to her pioneering work in gene regulation[20]), who were the first to propose that the expression of a gene is regulated by the binding and activity of a regulatory protein.[21,22] As the production of such regulatory proteins themselves also needs to be regulated, this model introduced the idea that gene expression has to be regulated by an intricate regulatory circuit that has to be strictly controlled. Evidenced by the high number of gene expression regulatory proteins that have been identified since then, this hypothesis was irrefutably true. At present more than 1,300 transcription factors have been identified, of which hundreds are being produced per cell.[23]

Gene transcription; RNA polymerase II and general transcription factors

DNA transcription is performed by protein complexes called DNA-dependent RNA polymerases. In humans, three types of DNA-dependent RNA polymerase exist (I, II and III) and the intrinsic details of their transcriptional action differ.[24] The RNA polymerase responsible for transcription of all protein-coding genes (and some non-protein-coding genes) is RNA polymerase II (RNA pol II).

Gene transcription starts from the core promoter of a gene, an approximately 100bp-long DNA sequence at the beginning of the gene (Figure 2A).[25] This core promoter includes several combinations of short DNA sequences, called regulatory elements (e.g. the TATA-box, Inr, DPE, DCE, MTE, BREu, XCPE1)[25] which specify the transcription start site(s) (TSS) of a gene. These elements are recognized by a group of general TFs and cofactors, including TFIIA (Transcription factor for RNA pol II A), TFIIB, TFIID, TFIIE, TFIIF and TFIIH, that bind as complex to the DNA and unwind the double helix to make it available for transcription.[25] The general transcription factors then attract RNA pol II and together form the pre-initiation complex (PIC).[26] Next, TFIIH phosphorylates RNA pol II at the fifth amino acid (a Serine) of a repeated motif located within the C-terminal domain (CTD) of the largest subunit of RNA pol II.[26] This triggers the release of the RNA pol II complex from the promoter and initiates transcription. Serine-5 phosphorylated RNA pol II is therefore often referred to as the 'initiating complex'. This complex is often paused shortly after initiation, by two factors called NELF and DSIF, and prevented from elongating transcription.[27,28] This pause is released by the recruitment of the positive transcription elongation factor complex p-TEFb, composed of CDK9 and CyclinT1, which phosphorylates both DSIF and NELF, and the RNA pol II CTD at the second amino acid of the repeat (also a Serine) (Figure 2B).[29] This releases NELF from the complex and transforms DSIF into a positive regulator, thereby releasing the RNA pol II from the paused state resulting in transcription elongation. The Serine-2 phosphorylated polymerase II is therefore often referred to as the 'elongating complex'.

Transcriptional pausing has been suggested to provide the cell with an additional mechanism to control transcription. The preformation of the PIC prior to functional transcription could ensure a rapid and synchronized response of paused genes to additional regulatory signals.[30–32] In addition, RNA pol II pausing may provide time to attract additional regulatory proteins[33] and to ensure that co-transcriptional processes (e.g. mRNA capping, RNA-splicing or 3'-end maturation) occur correctly.[34–36]. Approximately 30-40% of all genes are found to be paused at some point during transcription, indicating that transcriptional pausing is a generally used mechanism for transcriptional control.[37,38].

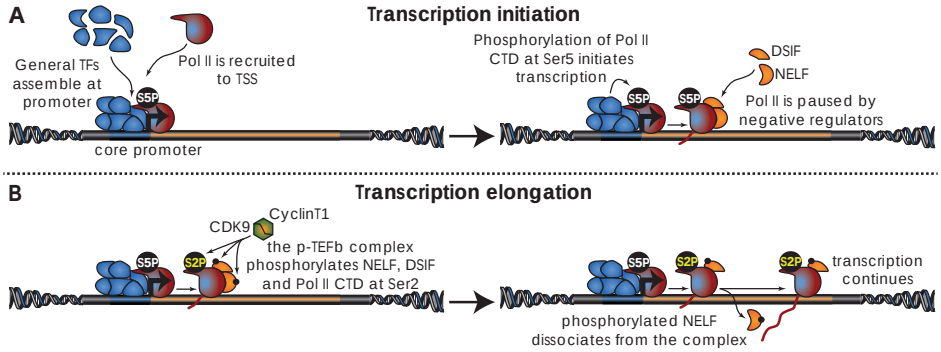


Figure 2. Schematic overview of gene transcription. (A) Transcription is initiated by the assembly of the general TFs at the core promoter. These factors recruit RNA pol II (Pol II) to the TSS, to together form the PIC. Phosphorylation of RNA pol II at Serine 5 (Ser5) of its CTD initiates the release of the transcription machinery from the TSS, to start transcription. Transcription is blocked shortly after, by negative elongation factors like NELF and DSIF. **(B)** The p-TEFb complex, composed of CDK9 and CyclinT1, phosphorylates NELF, DSIF and RNA pol II at at Serine 2 (Ser2) of its CTD. This causes NELF to dissociate from the complex and the trancription machinery to transform into an elongating complex which continues transcription.

Gene transcription regulation; Gene-specific transcription factors and regulatory DNA elements

Even though the formation of the PIC is sufficient to start transcription, this will generally result in short abortive transcripts due to dissociation of the RNA pol II complex shortly after initiation. The general TFs alone therefore are not sufficient for correct gene expression in a cell. Moreover, the formation of the basal PIC cannot account for the high level of variation in expression levels seen between different genes (although studies have shown that there is a significant variation in the components of the PIC assembled at the promoters of different genes, which might account for some level of gene-specific expression variation[39]).

Functional and gene-specific transcription levels require the regulation by additional TFs. Multiple different TFs have been identified and it is their complex interplay that gives a cell its unique characteristics. Multiple key regulators have been identified that are essential for lineage specification in the embryo and the adult animal. Without these factors, the formation of complete cell lineages (e.g. blood, bone, muscle, skin) is affected or even absent. Because of their crucial function they are often referred to as 'master regulators'. Well known examples are SOX2, OCT4 and NANOG, which are essential for the unique characteristics (i.e. pluripotency and self-renewal potential) of mammalian embryonic stem cells (ESCs), the cells that derive from the inner cell mass of pre-implantation stage blastocysts and that later give rise to all cell types of the embryo proper.[40–42] During embryogenesis, the expression of these key regulators decreases and new regulators take over to regulate the development of specific cell lineages (e.g. the erythroid-specific TF GATA1, the myeloid-specific TF PU.1, the muscle-specific TF MyoD).[43] However, although 'master regulators' provide the cell with lineage-specific characteristics, they often provide little information on cell type or differentiation stage specificity.[44,45] This requires additional general and/or cell type-specific factors that interact with these key regulators at different time points during development and cellular differentiation.[45,46]

Like general TFs, these tissue-specific TFs recognize and bind regulatory elements on the DNA and subsequently regulate the transcription of their target gene(s). Generally, TFs are classified into two types of regulators, transcriptional activators or repressors. However, this classification might be too simplistic as it is now known that many TFs can acquire both activating and repressing functions depending on the cofactors they recruit.[46]

Enhancers are DNA regulatory elements containing binding sites for transcription regulators such as TFs. They are characterized by their ability to regulate the transcription of their target gene(s) regardless of their genomic position (though mainly in cis) or orientation relative to the promoter(s) of the gene.[47] The first evidence for the existence of enhancers came in 1981, when Pierre Chambon and George Khoury and their colleagues independently discovered a 72bp repeat sequence in the DNA of the tumor virus Simian Virus 40 (SV40) that was essential for the expression of the early oncogenes of this virus.[48–50] Shortly after this, Walter Schaffner and Pierre Chambon and their coworkers independently revealed the enhancer's capacity to also enhance expression of unrelated genes, even over large distances.[48,51,52] Not long after in 1983, Paul Berg revealed the position and orientation independent nature of enhancers by showing that this SV40 enhancer could regulate expression of a gene even when positioned upstream or downstream of the gene promoter, or even when inserted in reversed orientation.[48,53]

Genes are often regulated by a set of cooperating enhancers. Such sets are referred to as locus control regions (LCRs). The first identified LCR is located roughly 25kb (25000bp) upstream of the *β-globin* gene locus and is comprised of 5 regulatory elements.[54] These elements are essential for the correct expression of the different genes in the *β-globin* gene locus at different time points during development (see below). Like for 'simple' enhancers, their regulatory function is independent of their position relative to the gene locus. However, two important additional characteristic of LCRs are that they can activate gene expression even when positioned in inactive chromatin regions (see section 'Chromatin; compaction and regulation') and that they regulate the expression of their target genes in a copy number dependent manner.[54] After the discovery of the first LCR, many more LCRs have been identified, indicating their general role in transcription regulation.[55]

It was recently found that genes that are key to a cell's identity are often regulated by a group of closely positioned regulatory elements bound by important cell type-specific TFs. These clusters of regulatory elements were referred to as super-enhancers and their activity is shown to be highly cell type-specific.[56] The same is true for the recently identified stretch-enhancers, characterized by their size of more than 3000bp.[57] Analysis of these stretch-enhancers has shown that they act in a highly cell type-specific fashion, with the degree of cell type-specificity correlating with the length of the enhancer. As both super-enhancers and stretch-enhancers share many characteristics with the previously identified LCR (and have been shown to overlap with known LCRs),[57] they are thought to represent specific subgroups of LCRs.

Traditionally, TFs were thought to mainly regulate PIC formation or stability and transcription initiation. However, increasing amounts of evidence reveal an important role for enhancers and interacting TFs in transcription elongation control as well, for example by recruiting the p-TEFb complex to target gene promoters, thereby stimulating the transcription initiation-to-elongation switch.[58–61]

Chromatin; compaction and regulation

At present, many thousands of potential regulatory elements have been identified in the complete human genome.[62–64] Combined with the more than 1,300 TFs that can interact with these sites,[23] they play an important role in the regulation of gene transcription. However, as every cell in the human body contains an identical copy of the DNA, these elements alone cannot explain why a gene is expressed in one cell

type but repressed in another. This difference is at least partially caused by a second layer of regulation and involves the way DNA is organized inside the nucleus of cells, i.e. in the form of chromatin (Figure 3).

As mentioned previously, a complete human genome contains approximately 3 billion bp.[14,15] Furthermore, each individual inherited a complete set of these 3 billion nucleotides from each parent. Completely stretched out, this complete set of DNA has a length of nearly 2 meter.[17] In order to fit this in the nucleus (which on average is only 6 µm in diameter[17]) this DNA needs to be packaged and folded in a compact structure. The compaction starts by wrapping the DNA around octameric protein complexes composed of two copies of four different histone proteins (i.e. H2A, H2B, H3 and H4)[65], which are often held in position by a fifth histone (i.e. H1).[66] This results in a 10nm-thick structure in which the DNA-protein complexes, called nucleosomes, are positioned on the DNA-strand like beads on a string. This histone-folded structure, called chromatin, is further compacted into higher-order structures by an unresolved folding mechanism. It has been generally accepted that the second level of compaction involves short-range nucleosome-nucleosome interactions, resulting in a 30nm fiber, though the exact mechanism of this folding is not clear[66–68] and recently the true *in vivo* existence of such fiber structure has been questioned.[69,70] Additional higher levels of compaction must exist in order for the DNA to fit into the nucleus, however to date the nature of this structure has not been resolved.[66]

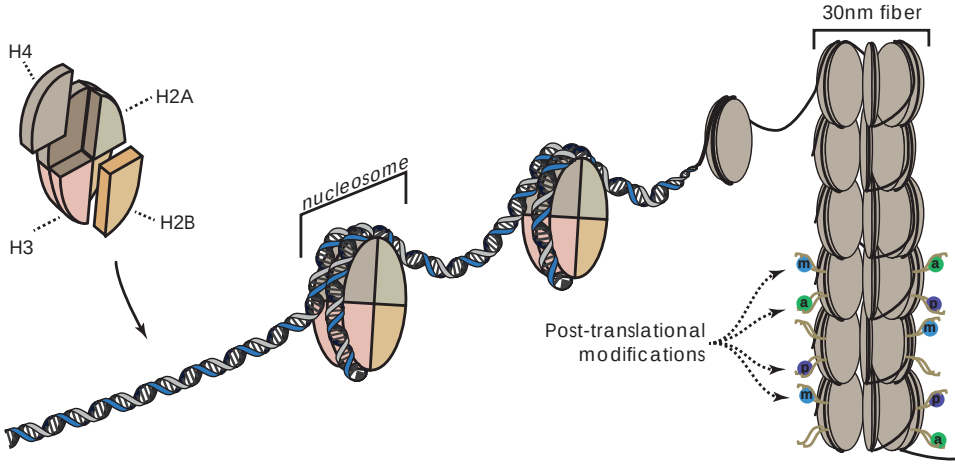


Figure 3. DNA is packaged in the nucleus in the form of chromatin. The DNA double helix is wrapped around an octameric histone complex, composed of two copies of four histone proteins; H2A, H2B, H3 and H4. Together, this complex forms the nucleosome. These nucleosomes are further compacted via the formation of nucleosome-nucleosome interactions to form a 30nm fiber. Ths fiber is schematically depicted, though the exact folding mechanism for the formation of this fiber, or any higher levels of compaction, is unknown. The histones can be post-translationally modified by various modifications. In this figure three examples are given, methylation (blue circle), acetylation (green circle) and phosphorylation (purple circle).

In addition to being essential for DNA compaction, nucleosomes also influence chromatin-contextual processes like gene transcription.[67] When incorporated in a nucleosome, promoters and enhancers can become inaccessible for TFs and the transcription machinery, which causes them to be inactive. In order to be transcribed, the chromatin structure therefore needs to be relaxed and nucleosomes need to be repositioned to open up parts in the DNA that contain regulatory elements. The general proposal is that this is regulated by so-called 'pioneer factors', which are the first to interact with the condensed chromatin region and relax the chromatin structure either passively (i.e. when interaction of this factor alone does not have

any mechanistic consequence for chromatin structure, but merely primes the site for activation by either lowering the number of interactions needed to activate the element by a bigger TF complex or by preventing this site from being inactivated by other factors) or actively (i.e. by actively regulating nucleosome compaction itself or by recruiting additional chromatin remodeling factors).[71,72]

The level of chromatin compaction is regulated by a wide variety of post-translational modifications (PTMs) added to the histones in the nucleosomes (Figure 3).[73–75] These PTMs influence chromatin dynamics and function in two main ways.[73] Firstly, the addition or removal of chemical modifications from the histones can influence the net charge of the nucleosome, thereby influencing the strength of its interaction with the negatively charged DNA or neighboring nucleosomes, resulting in either relaxation or increased compaction of the chromatin fiber. Secondly, PTMs can attract specific effector molecules (so called "readers") that can either influence the general chromatin architecture (e.g. the heterochromatin protein HP1[76]) or can specifically remodel the chromatin by moving, destabilizing, exchanging or restructuring nucleosomes.[77] In both ways, histone PTMs influence the accessibility of enhancers and promoters for TFs and therefore determine which genes can or cannot be transcribed.[78] This information encoded by the PTMs is referred to as the 'histone-code'.[79] Coming back to the metaphor used at the beginning of this Introduction, which described our genome as the handbook of the cell, the histone-code can be seen as all the notes and highlights added in the text or margin. It provides information on what parts of the tremendous amount of information in the DNA is important for the cell at a specific time, and how to interpret this information. Like real notes and highlights in a book, this information does not change the primary information encoded in the DNA, but adds an extra layer of information on top. This is therefore called our epigenome (which means above the genome).

At present 130 different histone PTM signatures have been identified which collectively can be added by more than 150 different histone modifying enzymes (so called "writers").[74,75] These modifications mostly involve the addition or removal of a chemical compound (e.g. an acetyl-, methyl-, phosphate-group) at specific sites of the N-terminal histone tail of every histone in the nucleosome.[73] In addition, nearly 50 additional histone variants have been identified that can replace the four basic histones described above.[75] In the last few years, multiple genome-wide analyses of PTMs have been conducted in an attempt to understand how the respective PTMs affect transcription. As a result various chromatin signatures have been identified. In brief, active gene regions are located in a relatively relaxed chromatin structure, called euchromatin, whereas genes that need to be tightly downregulated/silenced locate in a highly compact structure called heterochromatin.[66] Promoters are marked by a nucleosome-free TSS flanked by H3 trimethylated at Lysine 4 (H3K4me3).[63,80] Enhancers are highly cell-type specific and are characterized by mono-methylation of H3 at Lysine 4 (H3K4me1). In addition, the histone acetyltransferases P300, ATAC and SAGA are often found associated with distinct (though partially overlapping) groups of active regulatory elements, whereas trimethylation of H3 at Lysine 27 (H3K27me3) or Lysine 9 (H3K9me3) marks inactive regions.[63,80,81] Furthermore, transcription elongation correlates with the presence of H3 trimethylated at Lysine 36 (H3K36me3) and H3 dimethylated at Lysine 79 (H3K79me2) with the level of H3K79me2 being a quantitative predictive value for the transcription elongation rate.[82] The fact that so many histone modifications exist suggests that a tremendous amount of information could be encoded in the histone-code. However, to date the true function of all the combinatorial patterns is far from completely understood.

Chromatin higher-order structural conformation; impact on transcription regulation

Inside the nucleus chromatin is further organized into higher-order structural conformations.[83] These are thought to play a central role in transcriptional control. Genome-wide analyses have revealed that the majority of enhancers are located at far distance from their target gene promoter,[83] with enhancer-promoter distances identified as far one 1Mb (1,000,000bp). They can be positioned intragenic as well as intergenic[83] and can even regulate transcription of their target gene over an intermediately positioned gene.[84] In order to regulate transcription over such long distance, these elements need to be brought 'in contact' with their target gene. This often occurs via the formation of long-range chromatin loops, thereby physically bringing the enhancer in close proximity with their target gene (Figure 4A). A well-known example of such long-range interaction is the formation of the active chromatin hub at the *β-globin* gene locus, in which regulatory elements in the LCR located 25kb upstream of the *β-globin* gene-cluster loop towards the gene promoter by looping out all intervening DNA.[85] The formation of these long-range chromatin is essential for the activation of the β-like globin genes,[86] indicating an important role for chromatin conformation in transcriptional control. Aberrations in chromatin structural conformation play important roles in the development of, and susceptibility to, various genetic diseases. The role of chromatin conformation in transcription regulation, together with the impact of its misregulation on disease phenotype and susceptibility, will be discussed in more detail in Chapter 3.

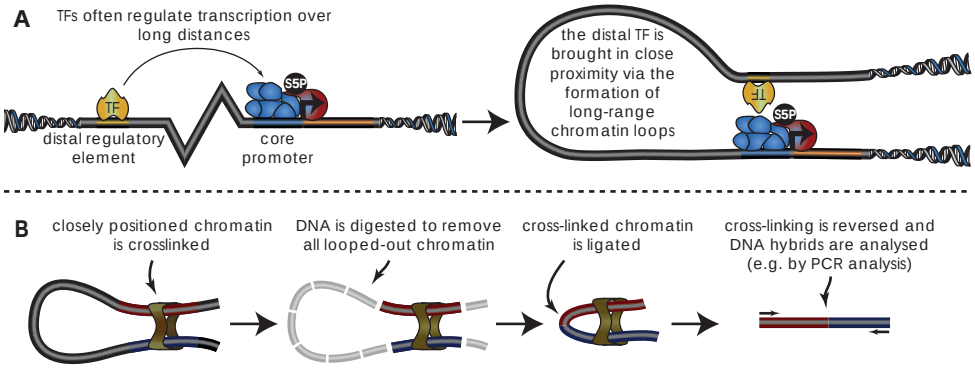


Figure 4. Transcription regulation by distal regulatory elements. (A) Transcription factors bound to distal regulatory elements often regulate their target gene via the formation of long-range chromatin loops. Red shape; RNA pol II, Blue shapes; general TFs, S5P/S2P; RNA pol II CTD phosphorylated at respectively Serine 5 or Serine 2, TF; transcription factor. **(B)** Schematic overview of the chromatin conformation capture (3C) technology.

First evidence for the higher-order structural organization of chromatin came from microscopy studies, revealing nuclear compartmentalization into so-called chromosome territories and (co-)localization of specific active and inactive chromatin regions.[87] However, this could only be studied with limited resolution, due to microscopy resolution limitations. This resolution was significantly increased by the development of the biochemical chromosome conformation capture (3C) technology by Dekker et al. in 2002 (Figure 4B).[88] This approach is based on cross-linking chromatin regions that are in close spatial proximity in the nucleus, followed by DNA digestion to remove all intervening ('looped-out') DNA. The two cross-linked regions are next ligated together to form DNA hybrids, which can be analyzed (e.g. by PCR technologies) to identify couples of closely positioned chromatin regions in the nucleus. Since the development of this 3C technology, numerous 3C variants have been developed to optimize its use for specific functions (e.g. 4C[89–91], 5C[92], 6C[93], ChIA-PET[94,95], GCC[96], Genome-wide 3C[97], TCC[98], Capture-C[99], Hi-C[100], T2C[101]). In Chapter 4

I discuss the development of multiplexed 3C-sequencing, a 3C variant optimized for use in semi-high throughput multiplexed next-generation sequencing.

The complex interplay of TFs and regulatory elements; a highly dynamic process

As mentioned above, more than 1,300 TFs have been identified, hundreds of which are expressed per cell. In addition, more than 43,000 potential enhancers have been identified of which thousands are functionally active per cell type.[62,63] It is the complex interplay between these different factors and regulatory elements that determines the unique transcription patterns per cell. Every gene is on average regulated by 4 to 5 enhancers, and *vice versa* every enhancer regulates about 2 to 3 genes.[62] The high redundancy in enhancers is suggested to ensure both for robust and at the same time accurate expression levels[62,102,103], while the sharing of enhancers would ensure combined regulation of clusters of related genes (as seen for example for the *β-globin* genes[54], the *Hox* genes[104] and the *Igf2-H19* gene locus[105]).

This regulatory process becomes even more complex when considering the fact that a cell's identity is not static and that (especially during development) cells need to be able to rapidly change cellular characteristics upon (external) stimuli and often need to display plasticity in behavior and fate. Regulation by TFs is often a balancing act between various TFs which act in a combinatorial or antagonistic fashion thereby influencing each other's function.[106,107] This cross-antagonism is thought to be important for ensuring complete, robust and correct cell fate decision.[106] In addition, TFs often regulate the expression of additional TFs (or even act auto-regulatory) via feedback or feedforward loops. Auto-regulation may ensure that TF levels can be robustly maintained, while regulatory feedback and feedforward loops may permit cells to rapidly respond to extrinsic cues.[43]

Enhancers are highly cell type-specific and their activity is highly dynamic. During differentiation, cells constantly change their active enhancer repertoire ending with a restricted set of lineage-specific enhancers in the mature cell.[45,108–111] Genome-wide studies during ESC differentiation have shown that lineage-specific enhancers may already acquire a permissive state in early progenitor cells prior to their lineage commitment.[110] Correspondingly, various lineage-specific genes are promiscuously expressed in these early progenitors.[112] This is thought to ensure cellular plasticity and a possibility to rapidly respond to extrinsic stimuli.[113] Upon lineage-commitment this permissive state is only retained at the appropriate lineage-specific regulatory elements, while other regulatory elements are inactivated. This indicates that lineage commitment does not only involve the activation of cell lineage-specific genes, but also the inactivation of genes and regulatory elements from opposing cell lineages. How these enhancers are inactivated is not fully understood, but this can for example involve cross-antagonistic TFs that directly recruit repressive complexes to the regulatory elements of opposing cell lineages, as done by the myeloid-specific factor PU.1 at the binding sites of the erythroid-specific factor GATA1.[107]

Similar early-stage enhancer-priming strategies have also been described in other differentiation processes.[113] However interestingly, two recent studies have identified an additional mechanism for acquiring a lineage-specific enhancer repertoire[108,109]: during hematopoietic cell differentiation a significant number of lineage-specific enhancers acquire *de novo* active enhancer marks upon lineage commitment (rather than prior to lineage commitment).[108,109] They are suggested to be activated by lineage-specific TFs and cofactors, which together attract chromatin remodelers to open up the chromatin structure.[108] However, it has recently been highlighted that

additional non-lineage-specific factors need to play a role as enhancer marks and open chromatin structures are often not restricted to one specific cell-lineage, but are also acquired in different cell-lineages that are derived from the same progenitor cell.[109]

The fact that different tissues use different mechanisms to acquire cell type-specific enhancer repertoires, increases the complexity of transcriptional control and might allow for differences in tissue plasticity and response to (external) stimuli.[109]

Transcription regulation during blood cell development; a widely used model system for cellular differentiation and development

Unraveling the complete mechanism of gene expression, will be essential for understanding the complete development of an organism. In this PhD research I focused on the formation of mammalian blood cells. With a volume of approximately 4-5 liter, the human blood system accounts for roughly 7% of the total body mass. It is involved in various important functions in the body, including immune response, nutrients/waste transport, and O_2/CO_2 transport respectively from the lungs to the tissues and *vice versa*. Blood is therefore essential for the survival of an organism and its correct formation and composition needs to be strictly controlled. Because of this, and its easy accessibility, blood has become the most widely used model system for studying the process of cell differentiation and development. The second part of this introduction will therefore give a brief overview of blood cell development and the role of several TFs and TF complexes in this developmental process.

The blood system; composition and function

Our blood is composed of a complex mixture of various different components which can generally be divided into 4 groups; blood plasma, platelets, white blood cells and red blood cells.[17,114] Roughly 55% of the total blood volume is blood plasma, an aqueous fluid that functions mainly as a solvent for a wide variety of proteins that need to be transported through the body, e.g. nutrients, waste products and electrolytes. In addition, the plasma functions as a suspension liquid for all blood cells allowing efficient transport of these cells through the body's vasculature.

Blood platelets are small cell fragments generated by megakaryocytes. They are responsible for blood clotting and are essential to prevent severe blood loss upon damage of the blood vessels, e.g. due to injury. Upon damage of a blood vessel, the platelets adhere to the endothelial inner cell lining of the vessel and block the breach, thereby preventing any leakage of blood outside the vessel.

The name 'white blood cell' is a collective name used as a counterbalance for the 'red blood cell', and refers to any blood cell type other than the red blood cell. White blood cells include a variety of cells which can be generally divided into 3 groups, the granulocytes, monocytes and lymphocytes.[17,114] Even though they account for roughly only 1% of the total blood volume they are an essential part of the blood system as they play an important role in the organism's immune system. Granulocytes, which can be further subdivided into neutrophils, basophils, mast cells and eosinophils, are involved in respectively phagocytosis (e.g. of microorganisms) or cytokine-driven modulation of inflammatory response (e.g. by histamine secretion). Monocytes can further mature into either macrophages or dendritic cells, which are involved in phagocytosis of microorganisms and of dead or damaged cells. In addition, they can activate other immune system cells via the secretion of cytokines/immunokines. Dendritic cells further stimulate additional immune responses via antigen presentation

to lymphocytes. These lymphocytes consist of B and T lymphocytes and lymphocyte-like Natural Killer cells, and are mainly involved in respectively antibody production and killing of pathogen-infected cells.

Red blood cells, or erythrocytes, comprise the remaining ~45% of the total blood volume and are the most abundant cells in blood. Their main function is O_2 and CO_2 transport respectively from the lungs to the tissues and *vice versa*, and are therefore essential for survival of the organism. In vertebrates, they do this by producing the oxygen-binding protein hemoglobin, which in mature erythrocytes comprises more than 90% of the dry protein weight of erythrocytes.[115] As will be discussed below, this hemoglobin carries iron, which give the erythrocyte its characteristic red color. In mammals, the erythrocyte has a characteristic donut-shaped structure, which it owes to the fact that it no longer carries a nucleus. As will be discussed later in this introduction, this nucleus is, together with the endoplasmic reticulum, mitochondria and ribosomes, extruded from the cell during its final maturation steps. This is essential for giving the erythrocyte its flexible shape, allowing efficient and easy transport through the complete blood vasculature.

Tissue formation; the adult tissue-specific stem cell

During the first stages of embryogenesis, the fertilized egg rapidly proliferates and ultimately differentiates into all basic structures of the embryo. This is followed by a subsequent growth of the fetus, which continues after birth until the organism reaches the adult stage of life. This growth of the fetus, requires the rapid proliferation and generation of the different cell types in the body. This is accomplished by specialized cells, called adult tissue-specific stem cells. Most stem cells are generated during embryogenesis, but remain present in the body throughout life. Tissue-specific stem cells are defined by two main criteria:[116]

1) They are multipotent, meaning that they are, though restricted to one tissue, not confined to one specific cell type and can differentiate towards all different cell lineages in the tissue.
2) They have self-renewal capacity, meaning that they can divide by making at least one identical copy of themselves. It is this capacity that ensures maintenance of the stem cell pool in the body throughout life.

Adult stem cells are important during all stages of life. During life a continuous generation of new cells is required to replenish shortages caused by loss or death of cells. Various organs display cell proliferation and have the capacity to regenerate long after embryonic development, indicating the presence of various adult tissue-specific stem cells in the body.[117] Blood is among the most regenerative tissues. Erythrocytes constantly travel through the body's vasculature, which makes them prone to cell damage. They therefore have a relatively short life span of only 120 days (in humans) and it is estimated that on average 2.5 million new erythrocytes are generated every second in order to replenish the human blood.[118] This number has to even increase in case of severe blood loss, e.g. due to injury. Similarly, throughout life (small) damages are inflicted to vessels, which need to be blocked by blood platelets to prevent bleeding. The average life span of platelets is therefore only 5-9 days and hence on average 2.5 million new platelets need to be generated every second (estimating the total number of platelets being $1.5x10^{12}$)[119]. In addition, when actively fighting an infection, the life span of a white blood cell can be as short as one day and the need and number of white blood cells highly depends on pathogenic stimuli.[114]

Blood cell formation; the hematopoietic stem cell

The adult tissue-specific stem cell responsible for the generation of the different blood cells is the Hematopoietic Stem Cell (HSC)[120,121], which in adults, mainly resides in the bone marrow.[122] The first experimental evidence for the existence of HSCs in the bone marrow was published in 1961, by Till and McColluch.[123] They showed that the transplantation of bone marrow cells from one mouse into recipient mice whose endogenous blood cells were destroyed by irradiation, resulted in the formation of discrete multi-lineage blood cell colonies in the spleen of these irradiated mice. As they found a strong correlation between the number of injected cells and the number of colonies, they concluded that each colony had derived from a small number of cells ("possibly from one cell" [123]) and referred to these cells as colony-forming units. Even though the true stem cell nature of the cells described by Till and McColluch has later been questioned as they do not have long-term self-renewal capacity,[124] these first experiments led to the believe in the existence of a true blood stem cell and started a long search for its true identity.[125]

However, despite more than 50 years of research, the true identity of the HSC still remains unknown. Technological advances now allow for the purification of HSCs to near purity[126], though, a unique identification marker for the HSC has not yet been identified. At present, a true HSC can therefore only be defined retrospectively by its ability to fully reconstitute blood formation-impaired adult recipients (i.e. indicating its pluripotency) both in primary and secondary transplantations (i.e. indicating its self-renewal capacity).[120,121]

The hematopoietic stem cell origin; the AGM region

Although during life HSCs mainly reside in the bone marrow, this is not the site of HSC production. They are generated at a different position during embryonic development and only migrate to the bone marrow just prior to birth.[122] In the following part of the Introduction I give a brief overview of the process of their formation, which has been studied in various species. As all work presented in this thesis is performed in murine hematopoietic cells, this part will focus on HSC formation in mice.

During mouse embryogenesis, the first hematopoietic cells derive in the yolk sac (Figure 5), in which around embryonic day 7.0 (E7.0) large nucleated erythrocytes, referred to as 'primitive' erythrocytes, can be detected[127] together with low numbers of macrophages and megakaryocytes.[128] This 'primitive' hematopoiesis is mainly thought to provide the rapidly growing embryo with enough oxygen and protection during the first embryonic development.[121,127] Starting from E8.5, various different hematopoietic progenitor cells start to appear in the yolk sac[127] and a region named Para-aortic Splanchnopleura (P-Sp).[129] However, transplantation assays revealed that at this stage the cells of neither the yolk sac nor the P-Sp have the ability to long-term reconstitute irradiated adult mice[130] (one of the two main criteria for true HSCs), indicating that at this embryonic stage no true HSCs are present. This primitive hematopoiesis is highly transient and rapidly declines around E9.[127]

The first true (pre-)HSCs can be detected in the embryo a little later, at E10.5, in a region called the Aorta-Gonad-Mesonephros (AGM), where they locate in the dorsal aorta of the embryo (Figure 5).[130,131] Interestingly, HSCs have also been detected in two other major arteries, i.e. the extra-embryonic vitelline artery and the umbilical artery.[132] In the arteries the HSCs reside in so-called intra-aortic hematopoietic clusters (IAHCs).[133–135] These clusters are derived from specialized endothelial cells called hemogenic endothelium, which contain both endothelial and hematopoietic

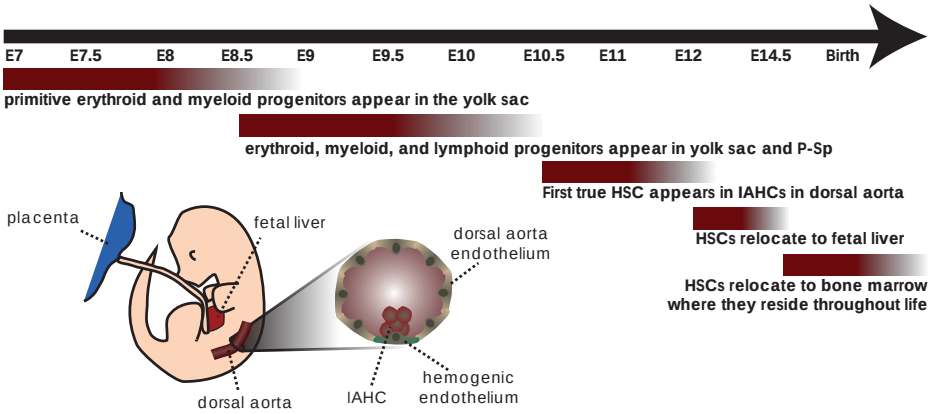


Figure 5. Timeline of hematopoietic development. The first true HSCs appear at E10.5 in the dorsal aorta of the embryo, schematically depicted below the timeline. They originate in intra-aortic hematopoietic clusters, which arise from specialized hemogenic endothelium cells located at the endothelial inner lining on the ventral side of the dorsal aorta.

characteristics[136] and are thought to originate from the aortic endothelium.[133–135] *In vivo* imaging of the dorsal aorta region showed that only a fraction of IAHCs form HSCs.[133] At present, the true identity of the HSCs in these clusters, the role of the IAHCs in the formation of HSCs and the trigger for HSC formation are largely unknown.

Shortly after the first appearance of HSCs in the AGM, HSCs can be detected at other sites in the embryo, e.g. in the yolk sac, placenta and fetal liver.[137–139] Around E12, the HSCs migrate to the fetal liver where the HSC pool expands significantly.[139] The mechanism for this explosive expansion in the fetal liver is not fully understood as it exceeds the expected expansion rates based on average cell division times.[139] This expansion might involve the maturation of an intermediate immature pre-HSC cell into a fully mature HSC.[139] Consistent with this idea is that cells isolated from <E10.5 embryos (prior to the first formation of true HSCs) cannot reconstitute the blood in adult recipients,[130] but can reconstitute the blood of newborns.[140,141] In addition, *ex vivo* cultures of dissected E11.5 AGM regions showed a 150-fold HSC expansion after only 96h of culturing.[142] This rapid expansion could be best explained by maturation of pre-HSCs rather than proliferation of the original HSCs present in the culture. These pre-HSCs have been shown to reside mainly in a VE-cadherin$^+$CD45$^-$ population, which later matures into a VE-cadherin$^+$CD45$^+$ population.[142,143] As the IAHCs also express these markers, it is possible that pre-HSCs exist in the IAHCs.[121] In Chapter 2 of this thesis, I describe the results of our recent study on the HSC formation in the IAHCs. In this study, we focused on the identity of IAHC cells and their role in HSC formation, and we show evidence for the existence of pre-HSCs in IAHCs.

Finally, starting from E14.5-E17 the HSCs relocate to specialized hematopoietic stem cell niches in the bone marrow (described in ref [144]), where they reside throughout life.[122]

Hematopoiesis; blood cell differentiation

The differentiation from HSC towards the different blood cell types is called hematopoietic differentiation or hematopoiesis. Hematopoiesis is classically presented as a branched differentiation process in which the multipotent HSC terminally differentiates towards any of the different blood cell lineages, losing part of its multipotency with each differentiation step (Figure 6) (nicely reviewed by Bryder et al.[125])

In the first step of differentiation, the HSC forms a multipotent progenitor (MPP). This MPP retains its multipotency, but gradually loses most of its self-renewal capacity.[145] This MPP then creates the first branching point in the differentiation tree when it differentiates into either one of two oligopotent progenitors, i.e. (i) the common lymphoid progenitor (CLP)[146], which gives rise to all lymphoid cells, but which is no longer able to differentiate towards myeloid cell lineages and (ii) the common myeloid progenitor (CMP)[147], which gives rise to all myeloid and erythroid cells, but which is no longer able to differentiate towards lymphoid progenitors.

The CLP differentiates into various lineage-specific progenitor cells, which further mature into the different lymphoid cells, i.e. B-lymphocytes, T-lymphocytes and Natural Killer cells and a group of lymphoid-derived dendritic cells. The CMP differentiation, however, involves several intermediate oligopotent progenitor cells. CMPs differentiate further into either a granulocyte-macrophage progenitor (GMP), which subsequently gives rise to monocytes/macrophages, the different types of granulocytes or a specific group of myeloid-derived dendritic cells; or a megakaryocyte-erythrocyte progenitor (MEP), which eventually gives rise to the erythrocytes and the platelet-producing megakaryocytes.[147]

It is important to note that this classical branch-tree representation of the hematopoietic differentiation, though describing important steps in the differentiation pathway, is a simplified and incomplete representation of the full process. Over the years, various alternative and additional intermediate cell stages have been identified and are at present still being identified (such as for example different stages of MPPs[145], heterogeneity in the HSC pool[19] and the presence of lineage-primed progenitors[148,149] respectively). Some examples of additionally identified cell populations and alternative differentiation pathways are depicted in Figure 6. Recent and continuous technological advances (e.g. in cell labeling and sophisticated FACS(-sorting) strategies and the various types of molecular analyses) will continue to improve our understanding of hematopoietic differentiation, and it is likely that more intermediate, transiting cell populations and alternative differentiation pathways will be identified.

Although the different intermediate progenitor cells no longer possess self-renewal capacity, they can proliferate significantly. The multilayer of intermediate progenitor cells therefore provides a means for rapid and cell type-specific expansion of the relevant cells without the need for high rates of HSC differentiation.[125] HSCs therefore mainly reside in a low cell-cycling state. Indeed, a recent study on native hematopoietic differentiation has shown that blood regeneration mainly involves the proliferation and differentiation of such early progenitor cell populations, rather than full hematopoietic differentiation starting from HSCs.[150] Interestingly, as opposed to the general assumption that these progenitors are short-lived and are therefore not suited for long-term reconstitution, this study revealed that early progenitors can reconstitute mice for at least one year. This finding is highly relevant for the development of tissue transplantations therapies, as it suggest a more limited potential for HSCs in tissue reconstitution in contrast with a high reconstitution potential for early progenitors. However, the reliable use for such multipotent early progenitor cells in long-term (life-long) reconstitution still needs to be further investigated.

Erythropoiesis: red blood cell development

In this thesis I mainly focus of the transcription regulation during erythropoiesis, the process of terminal differentiation of hematopoietic progenitors towards fully mature erythrocytes (Figure 7). This process starts with the commitment of an MEP towards

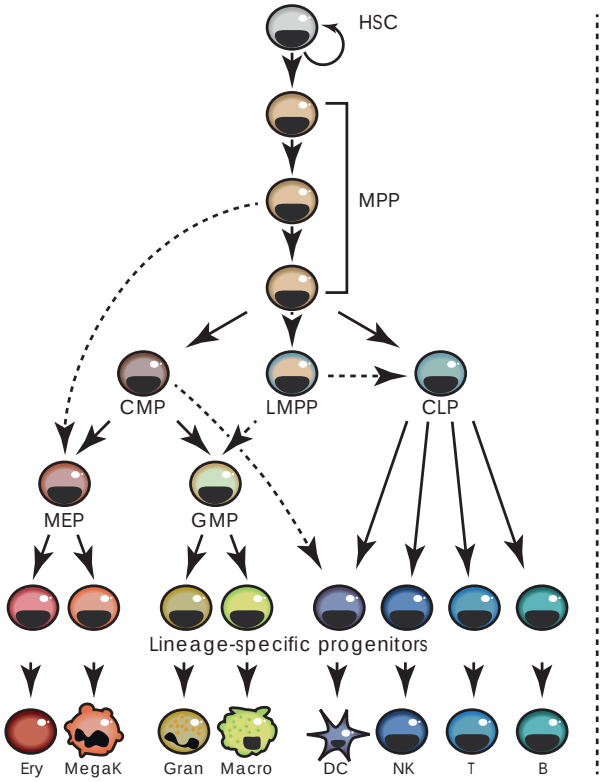


Figure 6. Schematic overview of the hematopoietic differentiation pathway. Solid lines depict the classical branched tree-like differentiation model. Figure adapted from Bryder et al.[125] Dashed lines represent a selection of alternative differentiation pathways identified at later stages. HSC; hematopoietic stem cell, MPP; multipotent progenitor, CMP; common myeloid progenitor, CLP; common lymphoid progenitor, LMPP, lymphoid-primed myeloid progenitor, MEP; megakaryocyte-erythrocyte progenitor, GMP; granulocyte-macrophage progenitor, Ery; erythrocyte, MegaK; megakaryocyte, Gran; granulocyte, Macro; macrophage, DC; dendritic cell, NK; natural killer cell, T; T-lymphocyte, B; B-lymphocyte.

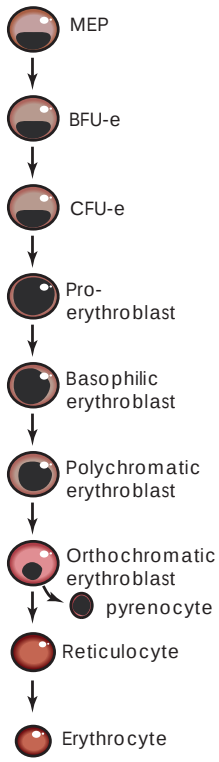


Figure 7. Schematic overview of the erythroid differentiation pathway. MEP; Megakaryocyte-erythrocyte progenitors, BFU-e; erythroid-burst forming units, CFU-e; erythroid-colony forming units. Figure adapted from Dzierzak et al.[151]

the erythroid lineage, thereby losing its oligopotency and becoming fully restricted to the erythroid differentiation pathway.[118,151] This differentiation occurs in erythroblastic islands in the fetal liver and later in the bone marrow. These islands are composed of a central macrophage (also known as the nurse cell) surrounded by erythroid progenitor cells that go through several rounds of differentiation.[152] The first erythroid committed cells are the erythroid-burst forming units (BFU-E) and the slightly more mature erythroid-colony forming units (CFU-E). They were initially defined by their ability to form colonies of mature erythroid cells in semi-solid culture media.[118] Under the control of cytokines (e.g. erythropoietin[153]), these cells differentiate further into morphologically identifiable precursors, including the proerythroblast (ProE), the basophilic erythroblast (BasoE), the polychromatophilic erythroblast (PolyE) and the orthochromatic erythroblast (OrthoE). During this progression, the cells become smaller, start to produce high levels of hemoglobin and their nucleus becomes more condensed. Next, the nucleus is extruded from the cell by an asymmetric cell division yielding the reticulocyte (i.e. the maturing red blood cell) and the pyrenocyte (i.e. “extruded nucleus” [118]). This pyrenocyte is quickly recognized and engulfed by the central macrophage[154] and the reticulocyte is released into the blood, where it further

matures into an erythrocyte by extruding its mitochondria, endoplasmic reticulum and ribosomes, and by adapting its plasma membrane and cytoskeleton to give it its flexible structure needed for its travel through the blood vasculature.[118]

Controlling HSC formation and hematopoiesis; important factors involved

Blood cell development is regulated by a wide variety of TFs. Many of these factors have been functionally tested and show aberrations in hematopoietic development upon their knockdown or knockout.[155] For example, TAL1/SCL and its associated protein partner LMO2 are essential for both primitive and definitive hematopoiesis, as lack of either of these two proteins results in complete lack of blood formation and embryos die around E9 due to severe anemia.[156–160] Both *Lmo2* and *Tal1* expression is found to be under the control of ETV2, indicating an important role for this factor in hematopoietic development.[161] Indeed, *Etv2* deletion also results in a complete loss of blood cell development.[161] RUNX1 and GATA2 are found to be specifically important for definitive hematopoiesis. Deletion of the *Runx1* gene results in a complete lack of IAHCs (the main site of HSC development, as discussed previously), indicating an essential role in HSC formation.[162] GATA2 is thought to be more important for regulating the proliferation status of early hematopoietic progenitors.[163]

Recent technological improvements in transcriptome analyses have now enabled us to study the molecular identity of cells up to single-cell resolution. Using such approach, numerous other TFs were recently identified as potential HSC transcription regulators.[164–166] However, although some hierarchical analyses have been performed,[166] most studies focused on HSC differentiation rather that HSC development and hence the molecular mechanisms in HSC development remain largely unaddressed. In Chapter 2 of this thesis I describe the results of transcriptome analyses of IAHC cells during the first stages of HSC development. We identified 108 TFs being differentially expressed during a short time window surrounding HSC development. This transcriptome analysis may provide important starting information on the role of various TFs in HSC formation and maintenance.

Terminal hematopoietic differentiation towards all different blood cell lineages involves a complex interplay between multiple TFs. As discussed above, lineage specification is often regulated by a subset of key lineage-specific regulators. One such erythroid-specific key regulator is GATA1. Deletion of *Gata1* results in abolishment of erythropoiesis.[167,168] Genome-wide GATA1 binding site analysis has revealed that this factor binds at many (if not all) genes important for erythroid development.[169,170] GATA1 is mostly found associated within multimeric TF complexes. One such complex is the LDB1 TF complex, composed of the five core components LDB1, TAL1/SCL, GATA1, LMO2 and E2A.[171–174] This complex is important for correct erythroid differentiation[173,174] and binds to a large number of erythroid-specific genes and their regulatory elements.[46,175] It binds to DNA via the two DNA-binding proteins GATA1 and TAL1/SCL (which heterodimerizes with E2A)[176], which together recognize the binding motif (C)TGN$_{7-8}$WGATAR.[46] LMO2 functions in this complex as a bridging molecule between TAL1/SCL:E2A and GATA1[172,176], and together with LDB1 acts as a scaffolding protein for additional interacting cofactors. Multiple interacting cofactors have been identified[177] and, depending on the function of the attracted cofactors, this complex is involved in either the activation or repression of its target genes.[46] In order to study the role of the LDB1 complex in transcription regulation of its target genes, in Chapter 5 I studied the role of the LDB1 complex in the downregulation of one of its target genes, *Bcl11a*.

Hemoglobin; the hemoglobin switch

The main function of erythrocytes is transporting oxygen from the lungs to the different tissues in the body, and transporting carbon dioxide from these tissues back to the lungs. Erythrocytes do this by producing the protein hemoglobin. Hemoglobin is composed of two alpha-like globin and two beta-like globin proteins which can bind oxygen via an iron ion-containing heme group incorporated in each of the four folded globins (Figure 8A).[178] In humans, five different beta-like globin genes (i.e. *ε-*, *Gγ-*, *Aγ-*, *β-* and *δ-globin*, respectively) and three α-like globin genes (i.e. *ζ-*, *α1-* and *α2-globin*, respectively) have been identified. These genes are expressed at different stages during development, generating different forms of hemoglobin.[179] In humans, two globin-switches occur. During primitive (and early definitive) erythropoiesis in the yolk sac, the beta-like *ε-globin* and the alpha-like *ζ-globin* are expressed to form the embryonic hemoglobin protein (Figure 8B). The first definitive erythrocytes in the fetal liver mainly express the two beta-like globin genes *Gγ* and *Aγ*, and the two alpha-like globin genes *α1* and *α2*, generating the fetal hemoglobin protein. Shortly after birth, when the main site of erythropoiesis shifts from the fetal liver to the bone marrow, fetal hemoglobin gene expression is reduced to less than 1% of the total hemoglobin level[180] and the majority of the hemoglobin proteins is composed of two beta-like β-globin proteins and two alpha-like α1- or α2-globin proteins (with a minority of hemoglobin containing beta-like δ-globin).

This two-level globin-switching mechanism is relatively unique, as it is only found to occur in human and old-world monkeys.[181] Most species, including mice, have only one switch, in which the primitive erythrocytes express embryonic globin genes, whereas definitive erythrocytes express adult globin genes (Figure 8B). As fetal hemoglobin is found to have a higher oxygen binding capacity than adult hemoglobin, the intermediate fetal hemoglobin switch is thought to be an adaptation to improve the oxygen transfer from mother to fetus.[151,182]

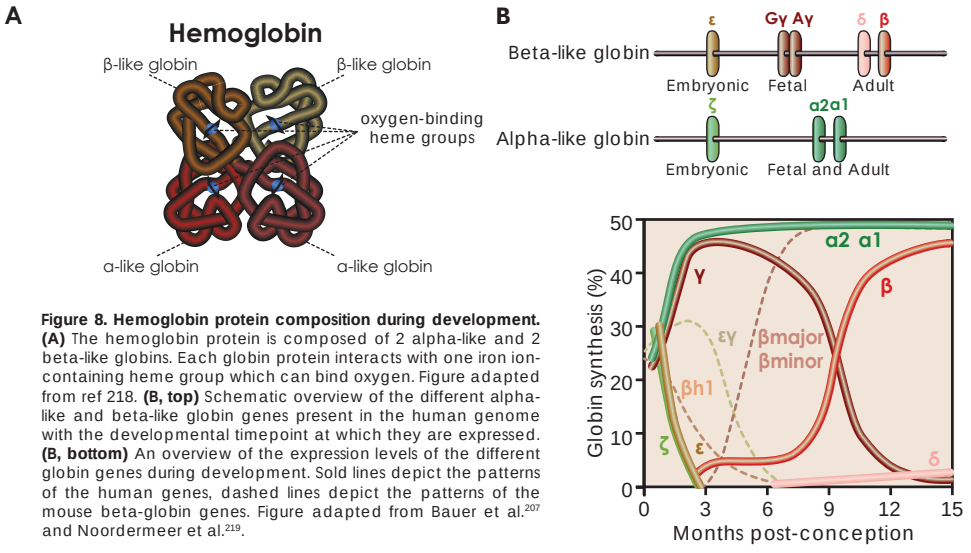


Figure 8. Hemoglobin protein composition during development. **(A)** The hemoglobin protein is composed of 2 alpha-like and 2 beta-like globins. Each globin protein interacts with one iron ion-containing heme group which can bind oxygen. Figure adapted from ref 218. **(B, top)** Schematic overview of the different alpha-like and beta-like globin genes present in the human genome with the developmental timepoint at which they are expressed. **(B, bottom)** An overview of the expression levels of the different globin genes during development. Sold lines depict the patterns of the human genes, dashed lines depict the patterns of the mouse beta-globin genes. Figure adapted from Bauer et al.[207] and Noordermeer et al.[219].

Hemoglobin-related diseases; potential therapeutic strategies

As defects in the production of hemoglobin influence the oxygen transport through the body, such defects often lead to severe anemia and can even lead to death of the patient. Hemoglobin-related diseases, collectively referred to as hemoglobinopathies, are among the most common monogenic diseases on earth, with approximately 6% of the human population carrying a mutation in the globin gene locus.[183,184] Two common hemoglobinopathies are sickle-cell disease (SCD)[185] and β-thalassemia[186]. SCD is caused by a structural mutation of the β-globin protein. In oxygen poor conditions, this mutated protein can aggregate, causing the shape of the cell to change into the typical sickle cell shape. This shape makes the cell less flexibel than normal erythrocytes and these cells are prone to cell damage inflicted during its continuous travel through the body's vasculature. Patients therefore suffer from severe anemia due to increased hemolysis. In addition, the rigid-shaped sickle cells can block small arteries thereby blocking the local blood supply causing local tissue necrosis. β-thalassemia can be caused by a variety of mutations in the *β-globin locus*. These mutations all result in reduced *β-globin* expression levels, leading to a reduced level of functional adult hemoglobin protein in the erythrocytes. Patients therefore develop severe anemia.

SCD and β-thalassemia are often treated with blood transfusions providing the patients with 'healthy' erythrocytes that contain normal levels of functional hemoglobin. However this treatment is only transient due to the normal short life span of erythrocytes of only 120 days. Bone marrow transplantation overcomes this problem by reconstituting the blood system of the recipient with normal stem cells that produce erythrocytes with normal levels of functional β-globin. The first successful bone marrow transplantation in a β-thalassemia patient has already been performed in 1981[187] and currently approximately 85% of the treated patients obtain a hemoglobinopathy-free life after the transplantation.[188] However, the success rate of this treatment is highly dependent on finding a suitable donor and is, also due to the high costs, not widely applicable.

The above mentioned therapies both focus on increasing the level of β-globin in the patient in a non-autonomous manner. As SCD is caused by the production of an abnormal β-globin protein[185] and β-thalassemia can be caused by nearly 200 different mutations affecting the production of adult β-globin[186], drug-treatments to cell-autonomously increase adult β-globin production are not highly effective. However, a positive correlation has been found between the severity of SCD and β-thalassemia and the level of fetal hemoglobin produced during adult life.[189–191] In most people the level of fetal γ-globin is reduced to <1% after birth. However, this level varies significantly among individuals[189] and persistent fetal γ-globin production in SCD and β-thalassemia patients ameliorates disease severity.[190,192] Therefore, reactivation of fetal *γ-globin* has become an intriguing option for therapeutic treatment of SCD and β-thalassemia.

Fetal *γ-globin* reactivation; the BCL11a protein

One protein that plays an important role in the regulation of the switch from fetal to adult hemoglobin is the TF BCL11a. A potential role for BCL11a in globin-switching was first suggested by genetic association studies (e.g. studies in which both DNA sequence and phenotype of multiple individuals are examined in order to try to find associations between small variations in the DNA and specific phenotypes). Various single nucleotide polymorphisms (SNP; commonly occurring variations) in the *BCL11a* gene have been identified, which together account for roughly 15-18% of

the variation in fetal hemoglobin expression levels.[193–198] BCL11a is thereby one of the three main contributors to the naturally occurring variation in fetal hemoglobin levels. Together with SNPs in the *β-globin* locus itself and SNPs in the regulatory elements of the *MYB* gene, *BCL11a* accounts for nearly 50% of the fetal hemoglobin level variation in SCD patients.[193–198] Functional studies showed that *BCL11a* is stage-specifically expressed with functional full-length protein isoforms being expressed only in adult erythroid cells, when fetal hemoglobin is repressed.[199] Reduced levels of *BCL11a* in adult erythroid cells leads to robust fetal globin expression.[199] In addition, a reduction of *BCL11a* expression results in impaired repression of the embryonic globin gene in mice and the fetal globin gene in humans, indicating an important role for BCL11a in hemoglobin switching.[200,201]

The beta-like globin genes are regulated by several regulatory elements surrounding the beta-like globin gene cluster, including the upstream LCR composed of 5 regulatory elements.[54,202] These elements can be bound by various TFs and cofactors,[182] and have been found to interact with the beta-globin gene cluster through the formation of an active chromatin hub structure.[85] In this chromatin hub active genes and regulatory elements are looped towards each other, whereas inactive genes are looped out. These long-range interactions are dynamic as they change during development depending on the globin genes being expressed at that moment. BCL11a has been found to bind at several regulatory elements in the *β-globin* locus, including in the LCR, the *ε-globin* gene and the intergenic region between the *γ-globin* and *δ-globin* gene, where it co-localizes with the erythroid-specific TFs GATA1 and the GATA-partner FOG1 (Figure 9).[203,204] One mechanism by which BCL11a can negatively regulate gene expression is via the attraction of repressive chromatin remodelers (e.g. the NURD complex, LSD1/CoREST, the SWI/SNF complex) to the regulatory elements it binds to, thereby inactivating these elements.[205] Furthermore, BCL11a is involved in modulating the long-range chromatin interactions between the LCR and the different genes in the cluster, at least partially by interacting with SOX6, a repressor of globin expression located at the *γ-globin* gene promoter.[203,206]

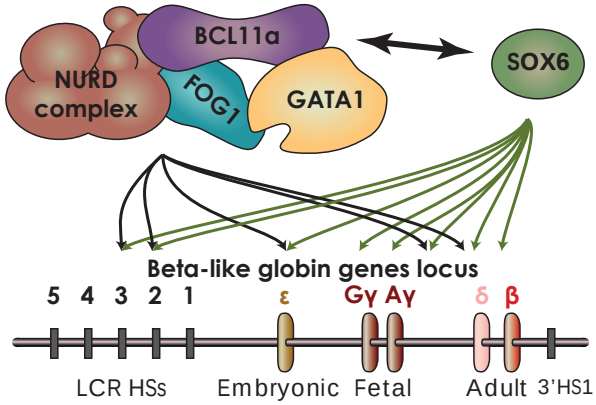


Figure 9. BCL11a regulates beta-like globin genes expression. BCL11a can form a multimeric complex with GATA1, FOG1 and several additional interacting cofactors (including the NURD complex). This complex can bind at various sites in the beta-like globin genes locus (depicted by black arrows). Together with SOX6 it is involved in modulating the long-range chromatin loops between the LCR and the different genes in the locus. The binding sites for SOX6 are depicted by green arrows. LCR; locus control region, HSs; hypersensitive sites. Figure adapted from Xu et al.[203]

These findings highlight *BCL11a* as an interesting potential therapeutic target for SCD and β-thalassemia[207]. Indeed, in a proof-of-principle study, Xu et. al.[208] demonstrated that erythroid-specific deletion of the *BCL11a* gene in an SCD mouse model completely rescues the pathologic defects through elevation of fetal-hemoglobin levels.
TFs have long been thought of as 'undruggable' targets due to their wide function in transcriptional control.[205] However, as will be discussed in Chapter 3, recent technological advances utilizing the cell type specificity of enhancers in enhancer targeting strategies, have now opened doors to the cell type-specific manipulation of gene expression levels. Targeted repression of *BCL11a* expression via such enhancer targeting strategies has therefore become a potentially feasible future therapeutic strategy. However, in order to use *BCL11a* repression for clinical purposes, a full understanding of the transcription regulation of the *BCL11a* gene will be essential.

BCL11a is known to be regulated by the TF KLF1, a key erythroid regulator which plays an important role in hemoglobin switching by both directly increasing *β-globin* transcript levels[209–213] and indirectly decreasing *γ-globin* transcript levels by increasing the expression levels of several *γ-globin* repressor genes, including *BCL11a*.[214–216] Interestingly, expression of both *BCL11a* and *KLF1* in embryonic erythroid cells (expressing embryonic globins) is sufficient to switch the globin expression towards adult *β-globin* genes,[217] whereas *KLF1::BCL11a* compound mutants show reduced β-globin levels and concomitant increased γ-globin levels,[216] indicating an important combinatorial role for these factors in hemoglobin switching. Chromatin immunoprecipitation (ChIP) experiments have revealed that KLF1 directly binds to the promoter of *BCL11a* both in human[214,215] and mice[215] and *KLF1* haploinsufficiency results in significantly reduced *BCL11a* expression levels and increased *γ-globin* expression levels.[214] Indeed, a decrease in *KLF1* levels (e.g. either by knockdown or heterozygous knockout of *KLF1*) significantly decreases *BCL11a* expression levels and increases *γ-globin* expression levels.[214–216] Decreasing KLF1 levels has therefore been suggested as an interesting strategy for *γ-globin* reactivation as therapeutic treatment for β-hemoglobinopathies.[214–216]

In addition to KLF1, recent ChIP-sequencing analyses have revealed that *Bcl11a* is also a target gene of the erythroid specific LDB1 TF complex described earlier in this Introduction.[46] As opposed to KLF1, which interacts with the *Bcl11a* promoter, the LDB1 complex binds at three promoter-distal regions in the *Bcl11a* locus, positioned in intron 2 of *Bcl11a*.[46] The LDB1 complex is known to act as a transcriptional activator for the majority of its target genes,[46] though its precise role in the transcriptional control of *Bcl11a* is unknown. In Chapter 5 of this thesis I discuss the results of our study on the transcription regulation of the highly conserved *Bcl11a* gene in mice, with a specific focus on the role of the LDB1 complex in this transcriptional control. The results from this study may provide important information on the specific TFs and enhancers involved in the transcriptional control of human *BCL11a* as well. These enhancers and TFs may be used as future targets for manipulating *BCL11a* expression as additional therapeutic strategy for hemoglobinopathies like SCD and β-thalassemia.

References

1. Bianconi E, Piovesan A, Facchin F, et al. An estimation of the number of cells in the human body. *Ann. Hum. Biol.* 40(6):463–71.
2. Vickaryous MK, Hall BK. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* 2006;81(3):425–55.
3. O'Connor C. Isolating the Hereditary Material: Frederick Griffith, Oswald Avery, Alfred Hershey and Martha Chase. *Nat. Educ.* 2008;1((1):105):
4. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 1944;79(2):137–58.
5. Griffith F. The Significance of Pneumococcal Types. *J. Hyg. (Lond).* 1928;27(2):113–59.
6. HERSHEY AD, CHASE M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 1952;36(1):39–56.
7. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids. *Nature*. 1953;171(4356):737–738.
8. Crick FHC, Barnett FRS., Brenner S, Watts-Tobin RJ. General Nature of the Genetic Code for Proteins. *Nature*. 1961;192(4809):1227–1232.
9. Nirenberg M. Historical review: Deciphering the genetic code--a personal account. *Trends Biochem. Sci.* 2004;29(1):46–54.
10. Perkins DO, Jeffries C, Sullivan P. Expanding the "central dogma": the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia. *Mol. Psychiatry*. 2005;10(1):69–78.
11. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature*. 2012;489(7414):101–8.
12. Gerstein MB, Bruce C, Rozowsky JS, et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 2007;17(6):669–81.
13. Consortium International Human Genome Sequencing. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931–45.
14. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304–51.
15. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
16. Yadav SP. The wholeness in suffix -omics, -omes, and the word om. *J. Biomol. Tech.* 2007;18(5):277.
17. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of The Cell. Garland Science, Taylor & Francis Group; .
18. Singer ZS, Yong J, Tischler J, et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell*. 2014;55(2):319–31.
19. Copley MR, Beer PA, Eaves CJ. Hematopoietic stem cell heterogeneity takes center stage. *Cell Stem Cell*. 2012;10(6):690–7.
20. COMFORT N. From controlling elements to transposons: Barbara McClintock and the Nobel Prize1. *Trends Genet.* 2001;17(8):475–478.
21. Jacob F, Monod J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* 1961;3:318–356.
22. Yaniv M. The 50th anniversary of the publication of the operon theory in the Journal of Molecular Biology: past, present and future. *J. Mol. Biol.* 2011;409(1):1–6.
23. Vaquerizas JM, Kummerfeld SK, Teichmann S a, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 2009;10(4):252–63.
24. Vannini A, Cramer P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell*. 2012;45(4):439–46.
25. Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* 2010;339(2):225–9.
26. Svejstrup JQ. The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim. Biophys. Acta*. 2004;1677(1-3):64–73.
27. Yamaguchi Y, Takagi T, Wada T, et al. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*. 1999;97(1):41–51.
28. Wada T, Takagi T, Yamaguchi Y, et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* 1998;12(3):343–56.

29. Peterlin BM, Price DH. Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell*. 2006;23(3):297–305.

30. Danko CG, Hah N, Luo X, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell*. 2013;50(2):212–22.

31. Gaertner B, Zeitlinger J. RNA polymerase II pausing during development. *Development*. 2014;141(6):1179–83.

32. Boettiger AN, Levine M. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science*. 2009;325(5939):471–3.

33. Henriques T, Gilchrist DA, Nechaev S, et al. Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Mol. Cell*. 2013;52(4):517–528.

34. Carrillo Oesterreich F, Bieberstein N, Neugebauer KM. Pause locally, splice globally. *Trends Cell Biol*. 2011;21(6):328–35.

35. Kadener S, Fededa JP, Rosbash M, Kornblihtt AR. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl. Acad. Sci. U. S. A.* 2002;99(12):8185–90.

36. Chathoth KT, Barrass JD, Webb S, Beggs JD. A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. *Mol. Cell*. 2014;53(5):779–90.

37. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. 2008;322(5909):1845–8.

38. Min IM, Waterfall JJ, Core LJ, et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev*. 2011;25(7):742–54.

39. D'Alessio J a, Wright KJ, Tjian R. Shifting players and paradigms in cell-specific transcription. *Mol. Cell*. 2009;36(6):924–31.

40. Avilion AA, Nicolis SK, Pevny LH, et al. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev*. 2003;17(1):126–40.

41. Nichols J, Zevnik B, Anastassiadis K, et al. Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. *Cell*. 1998;95(3):379–391.

42. Mitsui K, Tokuzawa Y, Itoh H, et al. The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. *Cell*. 2003;113(5):631–642.

43. Boyer L a, Lee TI, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*. 2005;122(6):947–56.

44. Wu W, Cheng Y, Keller C a, et al. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res.* 2011;21(10):1659–71.

45. Xu J, Shao Z, Glass K, et al. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell*. 2012;23(4):796–811.

46. Soler E, Andrieu-Soler C, de Boer E, et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev*. 2010;24(3):277–89.

47. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell*. 2011;144(3):327–39.

48. Patrusky B. Enhancers: Gene Enhancers. *MOSAIC*. 1986;17(3):37–44.

49. Benoist C, Chambon P. Deletions covering the putative promoter region of early mRNAs of simian virus 40 do not abolish T-antigen expression. *Proc. Natl. Acad. Sci. U. S. A.* 1980;77(7):3865–9.

50. Gruss P, Dhar R, Khoury G. Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proc. Natl. Acad. Sci. U. S. A.* 1981;78(2):943–7.

51. Moreau P, Hen R, Wasylyk B, et al. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* 1981;9(22):6047–68.

52. Banerji J, Rusconi S, Schaffner W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*. 1981;27(2):299–308.

53. Fromm M, Berg P. Simian virus 40 early- and late-region promoter functions are enhanced by the 72-base-pair repeat inserted at distant locations and inverted orientations. *Mol. Cell. Biol*. 1983;3(6):991–9.

54. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell*. 1987;51(6):975–985.

55. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. *Blood*. 2002;100(9):3077–86.

56. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307–19.

57. Parker SCJ, Stitzel ML, Taylor DL, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* 2013;110(44):17921–6.

58. Liu W, Ma Q, Wong K, et al. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell*. 2013;155(7):1581–95.

59. Bai X, Kim J, Yang Z, et al. TIF1gamma controls erythroid cell fate by regulating transcription elongation. *Cell*. 2010;142(1):133–43.

60. Stadhouders R, Thongjuea S, Andrieu-Soler C, et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J*. 2012;31(4):986–99.

61. Sawado T, Halow J, Bender MA, Groudine M. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes Dev*. 2003;17(8):1009–18.

62. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507(7493):455–61.

63. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009;459(7243):108–12.

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

65. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*. 1997;389(6648):251–60.

66. Woodcock CL, Ghosh RP. Chromatin higher-order structure and dynamics. *Cold Spring Harb. Perspect. Biol.* 2010;2(5):a000596.

67. Luger K, Dechassa ML, Tremethick DJ. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* 2012;13(7):436–47.

68. Tremethick DJ. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*. 2007;128(4):651–4.

69. Maeshima K, Hihara S, Eltsov M. Chromatin structure: does the 30-nm fibre exist in vivo? *Curr. Opin. Cell Biol.* 2010;22(3):291–7.

70. Nishino Y, Eltsov M, Joti Y, et al. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *EMBO J*. 2012;31(7):1644–53.

71. Magnani L, Eeckhoute J, Lupien M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet.* 2011;27(11):465–74.

72. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*. 2011;25(21):2227–41.

73. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res.* 2011;21(3):381–95.

74. Tan M, Luo H, Lee S, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*. 2011;146(6):1016–28.

75. Khare SP, Habib F, Sharma R, et al. HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* 2012;40(Database issue):D337–42.

76. Jacobs SA, Khorasanizadeh S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science*. 2002;295(5562):2080–3.

77. Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* 2009;78:273–304.

78. Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* 2011;12(4):283–93.

79. Jenuwein T, Allis CD. Translating the histone code. *Science*. 2001;293(5532):1074–80.

80. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 2007;39(3):311–8.

81. Nagy Z, Riss A, Fujiyama S, et al. The metazoan ATAC and SAGA coactivator HAT complexes regulate different sets of inducible target genes. *Cell. Mol. Life Sci.* 2010;67(4):611–28.

82. Wong P, Hattangadi SM, Cheng AW, et al. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood*. 2011;118(16):e128–38.

83. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489(7414):109–13.

84. Lettice LA. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 2003;12(14):1725–1735.

85. 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(6):1453–65.

86. Deng W, Lee J, Wang H, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*. 2012;149(6):1233–44.

87. Cremer T, Cremer M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* 2010;2(3):a003889.

88. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. 2002;295(5558):1306–11.

89. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* 2006;38(11):1348–54.

90. Zhao Z, Tavoosidana G, Sjölinder M, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* 2006;38(11):1341–7.

91. Van de Werken HJG, de Vree PJP, Splinter E, et al. 4C technology: protocols and data analysis. *Methods Enzymol.* 2012;513:89–112.

92. Dostie J, Richmond TA, Arnaout RA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 2006;16(10):1299–309.

93. 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(7):1171–9.

94. Zhang J, Poh HM, Peh SQ, et al. ChIA-PET analysis of transcriptional chromatin interactions. *Methods*. 2012;58(3):289–99.

95. Goh Y, Fullwood MJ, Poh HM, et al. Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. *J. Vis. Exp.* 2012;(62):

96. Rodley CDM, Bertels F, Jones B, O'Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. *Fungal Genet. Biol.* 2009;46(11):879–86.

97. Duan Z, Andronescu M, Schutz K, et al. A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods*. 2012;58(3):277–88.

98. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* 2012;30(1):90–8.

99. Hughes JR, Roberts N, McGowan S, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* 2014;46(2):205–12.

100. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289–93.

101. Kolovos P, van de Werken HJ, Kepper N, et al. Targeted chromatin capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin*. 2014;7(1):10.

102. Schaffner G, Schirm S, Müller-Baden B, Weber F, Schaffner W. Redundancy of information in enhancers as a principle of mammalian transcription control. *J. Mol. Biol.* 1988;201(1):81–90.

103. Corradin O, Saiakhova A, Akhtar-Zaidi B, 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.

104. Gould A, Morrison A, Sproat G, White RA, Krumlauf R. Positive cross-regulation and enhancer sharing: two mechanisms for specifying overlapping Hox expression patterns. *Genes Dev.* 1997;11(7):900–13.

105. Eun B, Sampley ML, Good AL, Gebert CM, Pfeifer K. Promoter cross-talk via a shared enhancer explains paternally biased expression of Nctc1 at the Igf2/H19/Nctc1 imprinted locus. *Nucleic Acids Res.* 2013;41(2):817–26.

106. Sunadome K, Suzuki T, Usui M, Ashida Y, Nishida E. Antagonism between the master regulators of differentiation ensures the discreteness and robustness of cell fates. *Mol. Cell*. 2014;54(3):526–35.

107. Stopka T, Amanatullah DF, Papetti M, Skoultchi AI. PU.1 inhibits the erythroid program by binding to GATA-1 on DNA and creating a repressive chromatin structure.

108. Lara-Astiaso D, Weiner a., Lorenzo-Vivas E, et al. Chromatin state dynamics during blood formation. *Science (80-.).* 2014;943:

109. Luyten a., Zang C, Liu XS, Shivdasani R a. Active enhancers are delineated de novo during hematopoiesis, with limited lineage fidelity among specified primary blood cells. *Genes Dev.* 2014;28(16):1827–1839.

110. Stergachis AB, Neph S, Reynolds A, et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell.* 2013;154(4):888–903.

111. Nord AS, Blow MJ, Attanasio C, et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell.* 2013;155(7):1521–31.

112. Miyamoto T, Iwasaki H, Reizis B, et al. Myeloid or Lymphoid Promiscuity as a Critical Step in Hematopoietic Lineage Commitment. *Dev. Cell*. 2002;3(1):137–147.

113. Kim T-H, Li F, Ferreiro-Neira I, et al. Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity. *Nature*. 2014;506(7489):511–5.

114. Murphy K, Travers P, M W. Janeway's Immunobiology.

115. Hajjawi OS. Human Red Blood Cells-1. *Am. J. Life Sci.* 2013;1(5):195.

116. Potten CS, Loeffler M. Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt. *Development*. 1990;110(4):1001–20.

117. Spalding KL, Bhardwaj RD, Buchholz B a, Druid H, Frisén J. Retrospective birth dating of cells in humans. *Cell*. 2005;122(1):133–43.

118. Palis J. Primitive and definitive erythropoiesis in mammals. *Front. Physiol.* 2014;5(January):3.

119. Kaufman RM, Airo R, Pollack S, Crosby WH. Circulating megakaryocytes and platelet release in the lung. *Blood*. 1965;26(6):720–31.

120. Domen J, Wagers A, Weissman IL. BONE MARROW (HEMATOPOIETIC) STEM CELLS,. *Regen. Med. Dep. Heal. Hum. Serv. August 2006,*. 2006;Chapter 2.

121. Boisset J-C, Robin C. On the origin of hematopoietic stem cells: progress and controversy. *Stem Cell Res.* 2012;8(1):1–13.

122. Christensen JL, Wright DE, Wagers AJ, Weissman IL. Circulation and chemotaxis of fetal hematopoietic stem cells. *PLoS Biol.* 2004;2(3):E75.

123. Till JE, McCulloch E a. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. 1961. *Radiat. Res.* 2012;178(2):AV3–7.

124. Schofield R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells*. 1978;4(1-2):7–25.

125. Bryder D, Rossi DJ, Weissman IL. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *Am. J. Pathol.* 2006;169(2):338–46.

126. Majeti R, Park CY, Weissman IL. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell*. 2007;1(6):635–45.

127. Palis J, Robertson S, Kennedy M, Wall C, Keller G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development*. 1999;126(22):5073–84.

128. Xu M J, Matsuoka S, Yang FC, et al. Evidence for the presence of murine primitive megakaryocytopoiesis in the early yolk sac. *Blood*. 2001;97(7):2016–22.

129. Godin I, Dieterlen-Lièvre F, Cumano a. Emergence of multipotent hemopoietic cells in the yolk sac and paraaortic splanchnopleura in mouse embryos, beginning at 8.5 days postcoitus. *Proc. Natl. Acad. Sci. U. S. A.* 1995;92(3):773–7.

130. Müller AM, Strouboulis J, Grosveld F, et al. Development of Hematopoietic Stem Cell Activity in the Mouse Embryo. 1994;1:291–301.

131. Medvinsky a, Dzierzak E. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell*. 1996;86(6):897–906.

132. De Bruijn MF, Speck NA, Peeters MC, Dzierzak E. Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo. *EMBO J.* 2000;19(11):2465–74.

133. Boisset J-C, van Cappellen W, Andrieu-Soler C, et al. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature*. 2010;464(7285):116–20.

134. Jaffredo T, Gautier R, Eichmann a, Dieterlen-Lièvre F. Intraaortic hemopoietic cells are derived from endothelial cells during ontogeny. *Development*. 1998;125(22):4575–83.

135. Tavian M, Coulombel L, Luton D, et al. Aorta-associated CD34+ hematopoietic cells in the early human embryo. *Blood*. 1996;87(1):67–72.

136. Garcia-Porrero JA, Manaia A, Jimeno J, et al. Antigenic profiles of endothelial and hemopoietic lineages in murine intraembryonic hemogenic sites. *Dev. Comp. Immunol.* 22(3):303–19.

137. Gekas C, Dieterlen-Lièvre F, Orkin SH, Mikkola HKA. The placenta is a niche for hematopoietic stem cells. *Dev. Cell*. 2005;8(3):365–75.

138. Ottersbach K, Dzierzak E. The murine placenta contains hematopoietic stem cells within the vascular labyrinth region. *Dev. Cell*. 2005;8(3):377–87.

139. Kumaravelu P, Hook L, Morrison AM, et al. yolk sac in colonisation of the mouse embryonic liver 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 colonisat.

140. Yoder MC, Hiatt K, Mukherjee P. In vivo repopulating hematopoietic stem cells are present in the murine yolk sac at day 9.0 postcoitus. *Proc. Natl. Acad. Sci. U. S. A.* 1997;94(13):6776–80.

141. Yoder MC, Hiatt K. Engraftment of embryonic hematopoietic cells in conditioned newborn recipients. *Blood*. 1997;89(6):2176–83.

142. Rybtsov S, Sobiesiak M, Taoudi S, et al. Hierarchical organization and early hematopoietic specification of the developing HSC lineage in the AGM region. *J. Exp. Med.* 2011;208(6):1305–1315.

143. Taoudi S, Gonneau C, Moore K, et al. Extensive hematopoietic stem cell generation in the AGM region via maturation of VE-cadherin+CD45+ pre-definitive HSCs. *Cell Stem Cell*. 2008;3(1):99–108.

144. Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. *Nature*. 2014;505(7483):327–34.

145. Morrison SJ, Wandycz a M, Hemmati HD, Wright DE, Weissman IL. Identification of a lineage of multipotent hematopoietic progenitors. *Development*. 1997;124(10):1929–39.

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

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

148. Luc S, Buza-Vidas N, Jacobsen SEW. Biological and molecular evidence for existence of lymphoid-primed multipotent progenitors. *Ann. N. Y. Acad. Sci.* 2007;1106:89–94.

149. Adolfsson J, Månsson R, Buza-Vidas N, et al. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell*. 2005;121(2):295–306.

150. Sun J, Ramos A, Chapman B, et al. Clonal dynamics of native haematopoiesis. *Nature*. 2014;514(7522):322–7.

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

152. Chasis JA, Mohandas N. Erythroblastic islands: niches for erythropoiesis. *Blood*. 2008;112(3):470–8.

153. Koury MJ, Bondurant MC. The molecular mechanism of erythropoietin action. *Eur. J. Biochem.* 1992;210(3):649–63.

154. Yoshida H, Kawane K, Koike M, et al. Phosphatidylserine-dependent engulfment by macrophages of nuclei from erythroid precursor cells. *Nature*. 2005;437(7059):754–8.

155. Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*. 2008;132(4):631–44.

156. Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W. and Orkin S. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell*. 1996;86:47–57.

157. Shivdasani, R.A., Mayer, E.L. and Orkin SH. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*. 1995;373:432–434.

158. Robb, L., Lyons, I., Li, R., Hartley, L., Kontgen, F., Harvey, R.P., Metcalf, D. and Begley C. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc. Natl. Acad. Sci. U. S. A.* 1995;92:7075–7079.

159. Patterson LJ, Gering M, Eckfeldt CE, et al. The transcription factors Scl and Lmo2 act together during development of the hemangioblast in zebrafish. *Blood*. 2007;109(6):2389–98.

160. Yamada Y, Warren AJ, Dobson C, et al. The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. *Proc. Natl. Acad. Sci.* 1998;95(7):3890–3895.

161. Kataoka H, Hayashi M, Nakagawa R, et al. Etv2/ER71 induces vascular mesoderm from Flk1+PDGFRa+ primitive mesoderm. *Blood*. 2011;118(26):6975–86.

162. Chen MJ, Yokomizo T, Zeigler BM, Dzierzak E, Speck NA. Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature*. 2009;457(7231):887–91.

163. Tsai FY, Keller G, Kuo FC, et al. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*. 1994;371(6494):221–6.

164. Guo G, Luc S, Marco E, et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell*. 2013;13(4):492–505.

165. Gazit R, Garrison BS, Rao TN, et al. Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem cell reports*. 2013;1(3):266–80.

166. Moignard V, Woodhouse S, Fisher J, Göttgens B. Transcriptional hierarchies regulating early blood cell development. *Blood Cells. Mol. Dis.* 2013;51(4):239–47.

167. Pevny, L., Simon, M.C., Robertson, E., Klein, W.H., Tsai, S.F., D'Agati, V., Orkin, S.H. and Costantini F. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*. 1991;349:257–260.

168. Fujiwara Y, Browne CP, Cunniff K, Goff SC, Orkin SH. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc. Natl. Acad. Sci. U. S. A.* 1996;93(22):12355–8.

169. Fujiwara T, O'Geen H, Keles S, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol. Cell*. 2009;36(4):667–81.

170. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*. 2002;3368–3376.

171. Osada H, Grutz G, Axelson H, Forster A, Rabbitts TH. Association of erythroid transcription factors: complexes involving the LIM protein RBTN2 and the zinc-finger protein GATA1. *Proc. Natl. Acad. Sci. U. S. A.* 1995;92(21):9585–9.

172. Wadman IA, Osada H, Grütz GG, 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. *EMBO J*. 1997;16(11):3145–57.

173. Visvader JE, Mao X, Fujiwara Y, Hahm K, Orkin SH. The LIM-domain binding protein Ldb1 and its partner LMO2 act as negative regulators of erythroid differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 1997;94(25):13707–12.

174. Xu Z, Huang S, Chang L-S, Agulnick AD, Brandt SJ. Identification of a TAL1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation. *Mol. Cell. Biol.* 2003;23(21):7585–99.

175. Li L, Freudenberg J, Cui K, et al. Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. 2013;121(22):4575–4585.

176. El Omari K, Hoosdally SJ, Tuladhar K, et al. Structural basis for LMO2-driven recruitment of the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Rep.* 2013;4(1):135–47.

177. Meier N, Krpic S, Rodriguez P, et al. Novel binding partners of Ldb1 are required for haematopoietic development. *Development*. 2006;133(24):4913–23.

178. PERUTZ MF, ROSSMANN MG, CULLIS AF, et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*. 1960;185(4711):416–22.

179. Trimborn T, Gribnau J, Grosveld F, Fraser P. Mechanisms of developmental control of transcription in the murine alpha - and beta -globin loci. *Genes Dev*. 1999;13(1):112–124.

180. Rochette J, Craig JE, Thein SL. Fetal hemoglobin levels in adults. *Blood Rev*. 1994;8(4):213–224.

181. Johnson R, Buck S, Chiu C, et al. Humans and old world monkeys have similar patterns of fetal globin expression. *J. Exp. Zool.* 2000;288(4):318–26.

182. Sankaran VG, Xu J, Orkin SH. Advances in the understanding of haemoglobin switching. *Br. J. Haematol.* 2010;149(2):181–94.

183. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. *Blood*. 2010;115(22):4331–6.

184. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. *Bull. World Health Organ.* 2001;79(8):704–12.

185. Schnog JB, Duits a J, Muskiet F a J, et al. Sickle cell disease; a general overview. *Neth. J. Med.* 2004;62(10):364–74.

186. Cooley T. The beta-Thalessemias. *N. Engl. J. Med.* 1999;341(2):99–109.

187. Thomas ED, Sanders JE, Buckner CD, et al. Marrow transplantation for thalassaemia.

188. Gaziev J, Sodani P, Lucarelli G. Hematopoietic stem cell transplantation in thalassemia. *Bone Marrow Transplant.* 2008;42 Suppl 1(S1):S41.

189. Thein SL, Menzel S, Lathrop M, Garner C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum. Mol. Genet.* 2009;18(R2):R216–23.

190. CONLEY CL, WEATHERALL DJ, RICHARDSON SN, SHEPARD MK, CHARACHE S. Hereditary Persistence of Fetal Hemoglobin: A Study of 79 Affected Persons in 15 Negro Families in Balitmore. *Blood*. 1963;21(3):261–281.

191. EDINGTON GM, LEHMANN H. Expression of the sickle-cell gene in Africa. *Br. Med. J.* 1955;1(4925):1308–11.

192. EDINGTON GM, LEHMANN H. Sickle-cell trait and malaria in Africa. *Bull. World Health Organ.* 1956;15(3-5):837–42.

193. Uda M, Galanello R, Sanna S, 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):1620–5.

194. Lettre G, Sankaran VG, Bezerra MAC, 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):11869–74.

195. Galarneau G, Palmer CD, Sankaran VG, et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* 2010;42(12):1049–51.

196. Menzel S, Garner C, Gut I, 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):1197–9.

197. Sedgewick AE, Timofeev N, Sebastiani P, et al. BCL11A is a major HbF quantitative trait locus in three different populations with beta-hemoglobinopathies. *Blood Cells. Mol. Dis.* 2008;41(3):255–8.

198. Danjou F, Anni F, Perseu L, et al. Genetic modifiers of β-thalassemia and clinical severity as assessed by age at first transfusion. *Haematologica*. 2012;97(7):989–93.

199. Sankaran VG, Menne TF, Xu J, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science*. 2008;322(5909):1839–42.

200. Sankaran VG, Xu J, Ragoczy T, et al. Developmental and species-divergent globin switching are driven by BCL11A. *Nature*. 2009;460(7259):1093–7.

201. Chen Z, Luo H, Steinberg MH, Chui DHK. BCL11A represses HBG transcription in K562 cells. *Blood Cells. Mol. Dis.* 2009;42(2):144–9.

202. Stamatoyannopoulos G. Control of globin gene expression during development and erythroid differentiation. *Exp. Hematol.* 2005;33(3):259–71.

203. Xu J, Sankaran VG, Ni M, et al. Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev*. 2010;24(8):783–98.

204. Jawaid K, Wahlberg K, Thein SL, Best S. Binding patterns of BCL11A in the globin and GATA1 loci and characterization of the BCL11A fetal hemoglobin locus. *Blood Cells. Mol. Dis.* 2010;45(2):140–6.

205. Xu J, Bauer DE, Kerenyi M a, et al. Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc. Natl. Acad. Sci. U. S. A.* 2013;110(16):6518–23.

206. Yi Z, Cohen-Barak O, Hagiwara N, et al. Sox6 directly silences epsilon globin expression in definitive erythropoiesis. *PLoS Genet*. 2006;2(2):e14.

207. Bauer DE, Kamran SC, Orkin SH. Reawakening fetal hemoglobin: prospects for new therapies for the β-globin disorders. *Blood*. 2012;120(15):2945–53.

208. Xu J, Peng C, Sankaran VG, et al. Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. *Science*. 2011;334(6058):993–6.

209. Tallack MR, Whitington T, Yuen WS, et al. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* 2010;20(8):1052–63.

210. Wijgerde M, Gribnau J, Trimborn T, et al. The role of EKLF in human beta-globin gene competition. *Genes Dev*. 1996;10(22):2894–902.

211. Perkins AC, Sharpe AH, Orkin SH. Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature*. 1995;375(6529):318–22.

212. Drissen R, von Lindern M, Kolbus A, et al. The erythroid phenotype of EKLF-null mice: defects in hemoglobin metabolism and membrane stability. *Mol. Cell. Biol.* 2005;25(12):5205–14.

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

214. Borg J, Papadopoulos P, Georgitsi M, et al. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat. Genet.* 2010;42(9):801–5.

215. Zhou D, Liu K, Sun C-W, Pawlik KM, Townes TM. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat. Genet.* 2010;42(9):742–4.

216. Esteghamat F, Gillemans N, Bilic I, et al. Erythropoiesis and globin switching in compound Klf1 :: Bcl11a mutant mice. 2014;121(13):2553–2563.

217. Trakarnsanga K, Wilson MC, Lau W, et al. Induction of adult levels of β-globin in human erythroid cells that intrinsically express embryonic or fetal globin by transduction with KLF1 and BCL11A-XL. *Haematologica*. 2014;99(11):1677–85.

218. King MW. themedicalbiochemistrypage.org. Last modified: January 21, 2014.

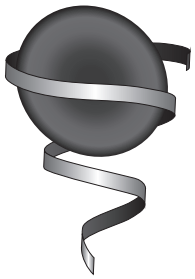
219. Noordermeer D, de Laat W. Joining the loops: beta-globin gene regulation. IUBMB Life. 2008;60(12):824–33.

Chapter 2

Sequential maturation towards hematopoietic stem cells in the mouse embryo aorta

Jean-Charles Boisset[1,2], Thomas Clapes[1,2], Chloé Baron[1], **Anita van den Heuvel**[2#], Supat Thongjuea[3#], Anna Klaus[2], Natalie Papazian[4], Petros Kolovos[2], Jos Onderwater[5], Frank Grosveld[2], Mieke Mommaas-Kienhuis[5], Eric Soler[2,6], Tom Cupedo[4],and Catherine Robin[1,2,7*]

A revised version of this manuscript is accepted for publication in Blood, Oct 9 2014, Under the title *'Progressive maturation towards hematopoietic stem cells in the mouse embryo aorta'*



[1]Hubrecht Institute-KNAW & University Medical Center Utrecht, The Netherlands
[2]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands
[3]MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, United Kingdom
[4]Department of Hematology, Erasmus Medical Center, Rotterdam, The Netherlands
[5]Department of Molecular Cell Biology, Leiden University Medical Center, Leiden, The Netherlands
[6] INSERM UMR967, CEA/DSV/iRCM, Fontenay-aux-Roses, France
[7]Department of Cell Biology, University Medical Center Utrecht, The Netherlands
[#]Equal contribution to the manuscript
*Correspondence: c.robin@hubrecht.eu

Abstract

The nature and role of the Intra-Aortic Hematopoietic Clusters (IAHCs), described a century ago and found in the main arteries of all vertebrate embryos, remains uncertain. IAHCs are derived from the hemogenic endothelium of arteries, and are speculated to contain the first Hematopoietic Stem Cells (HSCs) responsible for blood cell production throughout life. To determine the exact cell composition and function of the clusters, we designed an experimental approach to isolate pure cluster cells. These cells were further tested at the cellular and molecular level. We show that IAHCs contain mainly pre-HSCs, i.e. capable of long-term multilineage reconstitution in newborns, at a time when no HSCs are detected yet. Secondary transplantations and time-lapse imaging demonstrate that IAHC pre-HSCs can mature into HSCs. The molecular analysis of pre-HSCs by RNA-sequencing reveals important successive steps of maturation, occurring *in vivo* at different developmental time points and leading to HSC production. The novel insights in pre-HSC to HSC transition reported here represent an important source of information needed to generate transplantable HSCs *in vitro*, an achievement that would greatly improve current autologous HSC transplantation therapies.

Introduction

Hematopoietic Stem Cells (HSCs) are mainly present in the bone marrow in adults where they maintain blood cell production throughout life. They are identified *in vivo* by their ability to give rise to all hematopoietic lineages at long-term and to self-renew upon transplantation into primary and secondary irradiated adult wild-type (WT) recipients. HSCs are initially generated during embryonic development.[1] They are first detected at embryonic day (E)10.5 of mouse development in the aorta of the Aorta-Gonad-Mesonephros (AGM) region[2-4] and in the vitelline and umbilical arteries.[4,5] After E10.5, HSCs are also detected in the yolk sac, fetal liver and placenta.[2,6,7] After massive expansion in both fetal liver and placenta, HSCs migrate before birth into the bone marrow, the main adult HSC reservoir.[8]

From E9 onward, clusters of cells are observed in the vitelline and umbilical arteries, soon followed by their appearance in the dorsal aorta at E9.5.[9] These clusters named Intra-Aortic Hematopoietic Clusters (IAHCs) are tightly attached to the endothelial layer, facing the lumen of the vessels. They are present in virtually all vertebrate species in the ventral part of the aorta, and also in the dorsal part in the mouse embryo.[9-11] The tight association of IAHCs with the endothelium has, a century ago, led to the hypothesis that IAHC cells might derive from endothelial cells.[12,13] Multiple lines of evidence have now confirmed the so-called hemogenic endothelial origin of IAHCs and HSCs in chick[14,15] and mouse embryos[16-19]. The direct observation of the endothelial to hematopoietic cell-transition (EHT) in the aorta, by time-lapse confocal microscopy *in vivo* in zebrafish embryos[20-22] and *ex vivo* in thick mouse embryo slices[23], definitively confirmed the endothelial origin of IAHCs and HSCs.

Several observations support the hypothesis that HSCs reside in IAHCs.[24,25] Both HSCs and IAHCs (i) express the same surface markers (e.g. c-kit, CD31, CD34, CD41)[26], (ii) are absent in *Runx1* mutant embryos[27], and (iii) are derived from the hemogenic endothelium[16,19].[16,19] However, some discrepancies between HSCs and IAHCs exist, which raise questions on the exact cell composition and role of IAHCs. Indeed, IAHCs appear one day before HSC detection in the aorta (at E9.5 and E10.5, respectively).[2,9] In addition, the number of IAHC cells largely exceeds the number of HSCs estimated per AGM (~700 IAHC cells and 0.2 HSCs at E10.5; ~500 IAHC cells and 1-3 HSCs at E11.5)[2,6,9,28,29].[2,6,9,28,29] Finally, IAHCs are located in both sides of the aorta while HSCs are restricted to the ventral part.[11]

HSCs are not detected before E10.5 (<34 somite pairs, sp) when using the conventional transplantation assay in adult irradiated WT recipients.[2,3] However cells with HSC potential (i.e. self-renewal and multipotency), referred to as HSC precursors or pre-HSCs, have been identified at earlier stages of mouse development by using transplantation assays in other types of recipients. For example, cells from E8 Para-aortic Splanchnopleura (P-Sp) (the prospective AGM area), first cultured as explant and then on S17 stromal cells after tissue dissociation, reconstituted adult irradiated *Rag2*$^{-/-}\gamma c^{-/-}$ recipients after transplantation.[30] In addition, E9.0 yolk sac and P-Sp cells also directly reconstituted busulfan conditioned WT newborns.[31-33] Secondary transplantations performed in WT adult irradiated recipients confirmed the genuine HSC self-renewal potential of the transplanted cells.[32] It is noteworthy that so far no direct connection has been established between pre-HSCs and the first HSCs detected in the AGM.

To determine the exact nature and role of IAHCs, we analyzed the phenotypic evolution, function, and molecular events occurring in IAHCs over successive developmental time points, before and during HSC detection. Our results show that IAHCs are heterogeneous in size, shape and phenotype, but can be isolated to purity solely based on *c-kit* expression by using a new experimental strategy. We found that IAHCs contain no lymphoid progenitors and only very few erythroid-myeloid progenitors. We found that IAHCs essentially contain pre-HSCs, able to sustain *in vivo* long-term hematopoiesis in primary newborn immunodeficient recipients when no HSCs are detected yet. The successful long-term multilineage secondary transplantation of WT irradiated adult recipients (the hallmark assay of adult-type HSCs) definitively proves that IAHC cells can mature into functional HSCs. As shown by RNA-Sequencing analysis, this pre-HSC maturation towards an HSC fate involves significant molecular changes occurring within a short time window between mid-E10 and E11.5.

Results

IAHCs are heterogeneous in shape, but contain seemingly alike cells

To observe in detail the structure of IAHCs and IAHC cells, we imaged by scanning electron microscopy E10 (Figures 1A and S1A) and E11 (Figure S1B) embryo slices. Before tissue fixation, the circulating blood was flushed out of the aorta to ascertain that attached IAHCs, and not circulating cells, were observed. IAHCs, the underlining endothelium and sub-aortic mesenchyme were clearly visible (artificially colored in yellow, pink and blue respectively, in Figure 1B). In addition to single cells (Figure S1C), spheroid (Figure1B), stack (Figure1C) and 'mushroom like' (Figure1D) IAHC shapes were observed independently of the developmental stage. Cells appeared similar in all IAHCs, with a spherical shape and surface microvilli (Figures 1E, S1D and S1E). Although the function of these microvilli is unknown, close-up views suggest a role for cell-cell adhesion and/or communication.

IAHC cells are phenotypically heterogeneous, but can be isolated to purity

To isolate IAHC cells, directly labeled antibodies against c-kit (expressed by all IAHC cells) were injected inside the aorta of non-fixed E10 embryo trunk, prior to AGM dissection and dissociation to individual cells. Our procedure allows flushing out the blood from the aorta, staining all IAHCs[23,34] and isolating pure IAHC cells only (Figure 1F). When the AGM is dissociated before staining, contaminating c-kit^{+} cells from the blood circulation and outside the aorta will also be stained[9,35] (Figure 1G). We verified that the intra-aortic staining procedure did not detach IAHCs, by performing whole-embryo c-kit staining of flushed compared to non-flushed embryos. We found no significant differences in the number of c-kit (IAHC) cells in the aorta or in the shape of IAHCs (data not shown). Therefore, intra-aortic staining allows labeling and isolating IAHC cells to purity, solely based on c-kit expression.

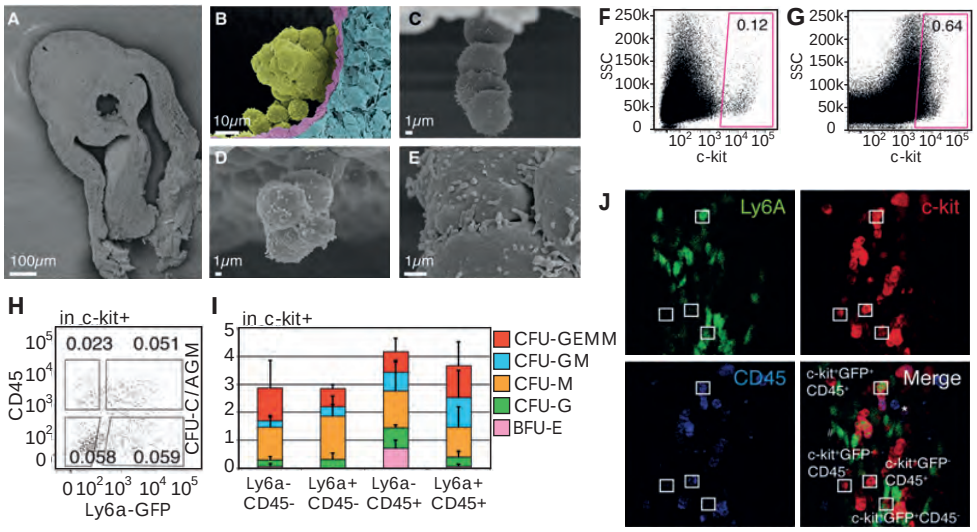


Figure 1. Intra-Aortic Hematopoietic Clusters (IAHCs) are phenotypically heterogeneous and contain very few progenitors at E10. **(A-E)** Scanning electron microscopy of E10 (28-34sp) and E11 thick embryo slices. **(A)** Whole E10 embryo slice where IAHCs are visible inside the aorta. **(B-E)** Close-up views of IAHCs with sphere **(B)**, stack **(C)** and "mushroom like" **(D)** shapes in the aorta of E10 **(B, C)** and E11 **(D)** embryos. IAHC (yellow), endothelium (pink) and sub-aortic mesenchyme (blue) were artificially colored to show the delimitation between the structures. Top, dorsal side; Bottom: ventral side. **(F, G)** Flow cytometry analysis of AGM cells (E10 embryos) stained with c-kit antibody by intra-aortic injection before AGM dissociation **(F)** or on dissociated AGM cells **(G)**. **(H)** CD45 and Ly6a-GFP expression within c-kit+ AGM cells from E10 Ly6a-GFP embryos. Percentages of viable cells are indicated inside the gates. **(I)** CFU-C assays for the four IAHC cell populations isolated from E10 AGMs (28-34sp) and sorted based on **(H)** (n=3). Bars, Means ± Standard Deviation. The number of CFU-C is indicated per flushed AGM. **(J)** Confocal imaging of the aortic endothelium and IAHCs on the floor of a non-fixed E10 Ly6a-GFP (green) embryo. The aorta was stained with antibodies against c-kit (red) and CD45 (blue). White boxes show examples of the four IAHC cell populations expressing c-kit but differentially expressing Ly6a-GFP and CD45 on the merged fluorescent signal panel. The three-dimensional reconstitution of the aortic floor is shown in Movie S1. See also Figures S1 and S2.

All c-kit$^+$ cells in IAHCs display several endothelial and hematopoietic surface markers (Figures S2A-D (E10) and S2E-H (E11.5)). Only Ly6a-GFP and CD45 were differentially expressed in IAHCs among the various markers tested (i.e. Tie2, CD31, VE-Cadherin, Flk1, CD34, CD45, CD41).[9,23,36] The phenotypic heterogeneity within and between IAHCs was independent of IAHC size and shape. Four phenotypically distinct populations (c-kit$^+$Ly6a-GFP$^-$CD45$^-$, c-kit$^+$Ly6a-GFP$^+$CD45$^-$, c-kit$^+$Ly6a-GFP$^-$CD45$^+$ and c-kit$^+$Ly6a-GFP$^+$CD45$^+$, respectively) were therefore detectable by flow cytometry and/or confocal microscopy at E10 (Figures 1H, 1J and Movie S1) and E11.5 (Figure S1F).

IAHCs contain few erythroid-myeloid progenitors and no lymphoid progenitors

To test whether IAHCs contain committed hematopoietic progenitors, IAHC cell populations (sorted based on c-kit and/or Ly6a-GFP and CD45 expression) were tested in three different *in vitro* clonogenic assays. In these assays, CFU (Colony Forming Unit) myeloid progenitors, CFU pre-B lymphoid progenitors and CFU Megakaryocyte progenitors (CFU-Mk) can be identified. We found that IAHCs contained very few myeloid progenitors at E10 and E11.5 (13±3 CFU myeloid progenitors/E10 AGM, n=3, Figure 1I; 19±4 CFU myeloid progenitors/E11.5 AGM, n=3, Figure S1G). At E10, we also found very few CFU-Mk (9±2 CFU-Mk/E10 AGM, n=2) and no CFU pre-B progenitors (also not in c-kit$^+$ yolk sac sorted cells, data not shown). The number and types of myeloid progenitors were similar in the four-sorted IAHC cell populations, indicating that a particular phenotype does not correlate with a specific type of progenitors at E10. Most myeloid progenitors were enriched in the c-kit$^+$Ly6a-GFP$^-$CD45$^+$ and c-kit$^+$Ly6a-GFP$^+$CD45$^+$ fractions at E11.5. Interestingly, most myeloid progenitors at

E10 and E11.5 were in the AGM circulating blood (enriched in the c-kit$^+$Ly6a-GFP$^-$ CD45$^+$ fraction) (Figures S1H-K). This shows that the high number of AGM progenitors described in the literature are in fact circulating cells.

IAHCs contain pre-HSCs able to mature *in vivo* into HSCs

We then tested whether IAHCs contained HSC-like cells, as defined by their ability to long-term multilineage reconstitute newborns or immunodeficient adult recipients but not WT adult recipients (the standard assay to identify HSCs).[30,32] To address this hypothesis, we performed transplantations of c-kit$^+$ IAHC cells isolated from early-E10 (prior to HSC detection) and mid-E10 (onset of HSC detection) AGMs into the liver of *Rag2*$^{-/-}\gamma c^{-/-}$ or WT neonates (Table 1). Early-E10 IAHC cells reconstituted both *Rag2*$^{-/-}\gamma c^{-/-}$ and WT neonate recipients up to 5 months after the minimum injection of 100 and 200 cells, respectively. Mid-E10 IAHC cells reconstituted both types of recipients with as few as 50 cells. In all cases, the reconstitution was multilineage with donor contribution to myeloid and to T and B lymphoid lineages (Figures 2 and S3). Thus, IAHCs contain HSC-like cells and their frequency increases between early and mid-E10. All four c-kit$^+$ IAHC cell populations (sorted based on Ly6a-GFP and CD45 expression) could reconstitute *Rag2*$^{-/-}\gamma c^{-/-}$ neonates (Table 1). Therefore, HSC-like cells in IAHCs are not restricted to a specific phenotype at E10 (based on the presence of Ly6a-GFP and CD45). Our data shows that HSC-like cells not only derive from Ly6a-GFP$^+$ hemogenic endothelial cells (as described previously[23,37]) but also from the Ly6a-GFP$^-$ counterparts.

Table 1. Intra-aortic hematopoietic clusters contain pre-HSCs.

Embryo stage	Sorted cells	Cell number injected/recipient	Recipients# Rag2$^{-/-}\gamma c^{-/-}$	Recipients# WT
E10 (25-32 sp)	c-Kit$^+$ *	10	0/1	0/9
		50	-	0/4
		100	2/2	0/2
		150	-	-
		200	-	3$/5
		750	1¶/1	-
		1350	1/1	-
	c-Kit $^+$GFP$^-$CD45$^-$ **	1000-2500	3/3	-
	c-Kit $^+$GFP$^+$CD45$^-$	250-550	3/4	-
	c-Kit $^+$GFP$^-$CD45$^+$	150-350	1/4	-
	c-Kit $^+$GFP$^+$CD45$^+$	250-650	3/4	-
E10 (33-38 sp)	c-Kit + *	10	0/1	0/6
		50	2/2	2$¶/12
		100	-	1/3
		150	2/2	-
		200	-	1/1

* n=5 independent experiments.
** n=3 independent experiments.
# Number of repopulated recipients per injected neonate recipients 4-5 months post-injection. Mice were considered repopulated when donor derived cells were detected in both bone marrow and spleen, by flow cytometry of the H2kk and Ly6a-GFP donor markers (for Rag2-/-γc-/- recipients) or of the CD45.1 donor marker (for wild-type (WT) recipients).
$Multilineage reconstitution is shown in Figures 2 and S3.
¶Primary recipients used for secondary transplantations (see Table S1).
(-) Not done.

To determine whether IAHC HSC-like cell maturation occurs *in vivo*, secondary transplantations were performed in irradiated adult (*Rag2*$^{-/-}\gamma c^{-/-}$ and/or WT) recipients with bone marrow and spleen cells isolated from primary reconstituted recipients (identified by ¶ in Table 1) (Table S1). In both cases, multilineage reconstitution was observed in bone marrow, spleen and peripheral blood of the recipients up to 4 months post-transplantation. Therefore, IAHC HSC-like cells can mature into fully potent HSCs *in vivo*, and therefore can be referred to as pre-HSCs. To observe IAHC cell maturation, time-lapse confocal microscopy was performed on E10 *Ly6a-GFP* embryo slices (stained with anti-CD31 and anti-c-kit antibodies prior to imaging).[34] All Ly6a-GFP$^-$ IAHC cells started to express GFP during the course of imaging (Figure 3A, two examples shown in Movie S2). To test whether cell proliferation occurred in IAHCs, embryo slices were stained for Phospho-Histone H3.3 (PHH3), CD31 and c-kit (Figure S4A) and proliferating IAHC cells (c-kit$^+$PHH3$^+$) were counted (Figure S4B). Figure 3B shows the percentage of proliferating IAHC cells per total number of IAHC cells (mitotic index). Only few IAHC cells proliferated at early, mid and late-E10 stages. Thus, the majority of IAHC cells mature towards a putative HSC state with low proliferation.

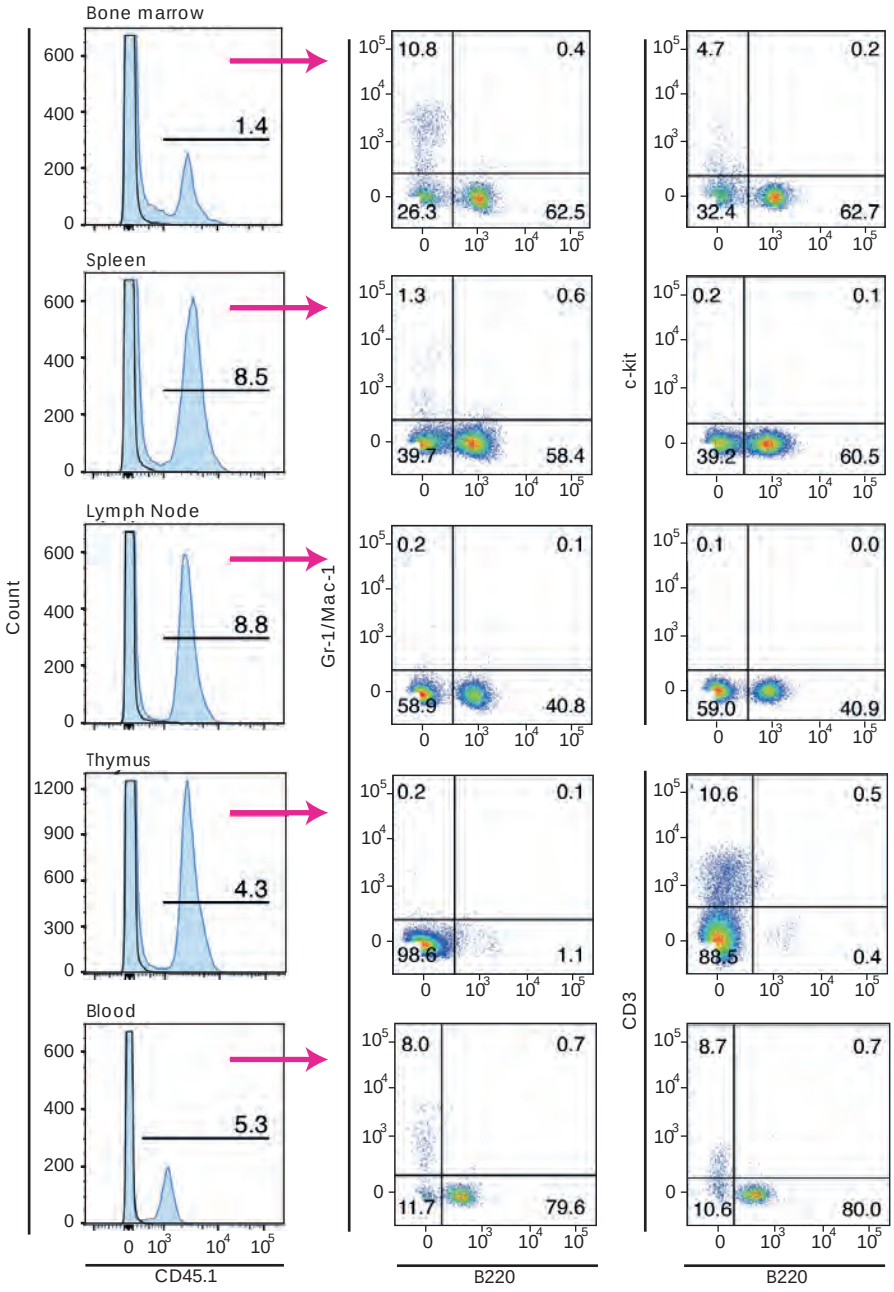


Figure 2. IAHCs from mid-E10 embryos contain pre-HSCs able of long-term multilineage hematopoietic reconstitution after transplantation in WT neonates. Analysis of a WT neonate recipient (CD45.2) transplanted with 50 IAHC c-kit+ cells from mid-E10 (33-38sp) AGM (CD45.1) up to 4 months post-transplantation. FACS analysis show donor cell contribution (CD45.1) in bone marrow, spleen, lymph nodes, thymus and blood, represented in histogram on the left panel (CD45.1: blue, Control: black line). Lines indicate the percentages of donor contribution in the whole tissue. Multilineage donor contribution (dot plots) was analyzed in all organs for myeloid (Gr-1/Mac-1) and B cells (B220), in bone marrow, spleen and lymph node for hematopoietic stem/progenitor cells (c-kit), and in blood and thymus for T cells (CD3). Percentages of each donor population are indicated per quadrant. See also Figure S3 and Table S1.

The progressive decrease of IAHCs (>E10.5) coincides with the beginning of HSC activity in the fetal liver (>E11), and might thus indicate a participation of IAHC cells to the fetal liver HSC production via a maturation process. However, the *in vivo* proof of this hypothesis has been lacking. We detected three of the four IAHC populations in E12 fetal liver by flow cytometry (Figure S4C) and confocal microscopy of thick non-fixed fetal liver slices (Figure S4D). These populations (c-kit$^+$Ly6a-GFP$^-$CD45$^-$, c-kit$^+$Ly6a-GFP$^-$CD45$^+$ and c-kit$^+$Ly6a-GFP$^+$CD45$^+$, respectively) isolated from E11.5 fetal livers were able to reconstitute at long-term *Rag2*$^{-/-}\gamma c^{-/-}$ newborn recipients (data not shown). Repopulation with the Ly6a-GFP$^-$ sub-fractions, which do not contain HSCs (since all HSCs are Ly6a-GFP$^+$ at E11.5[36]) proves that pre-HSCs are present in E11.5 fetal liver, where they might account for HSC production.

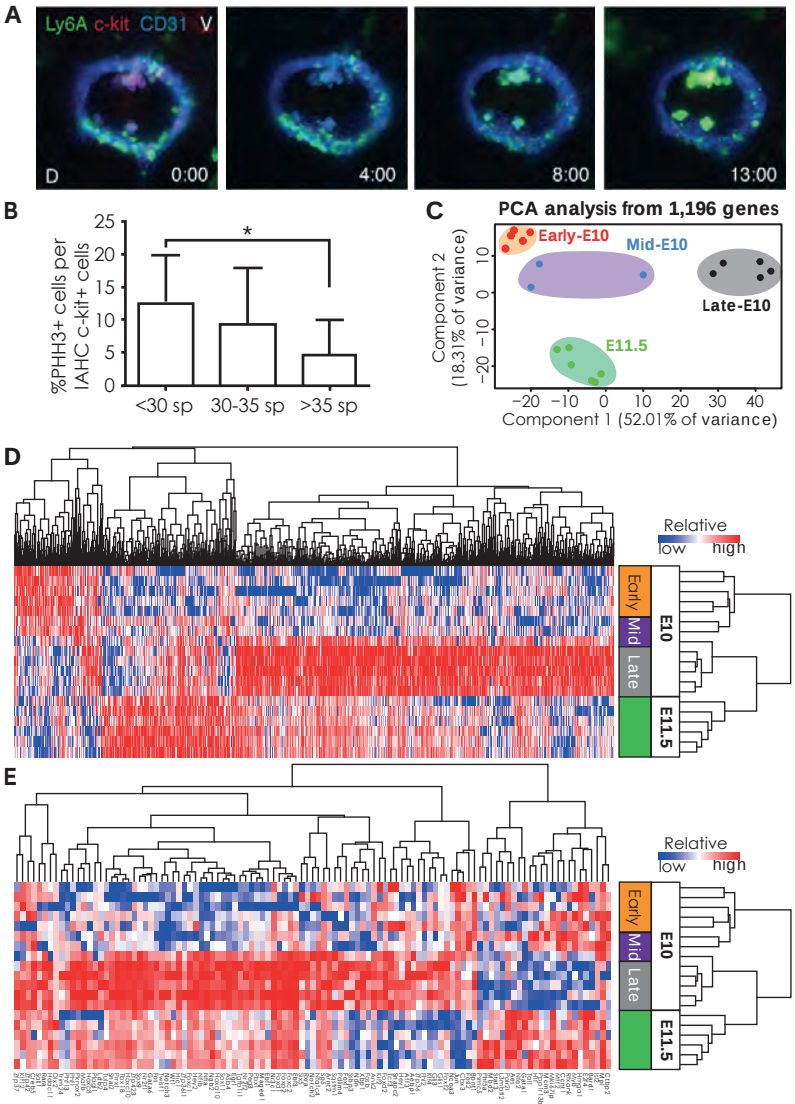


Figure 3. Phenotypic and molecular analysis of pre-HSC maturation towards an HSC fate. (A) Time-lapse series pictures from Movie S2 (Example 1) showing the maturation of IAHC cells. All CD31+c-kit+Ly6a-GFP- in IAHCs progressively express Ly6a-GFP during the time-lapse imaging (13h). V, ventral; D, dorsal; Ly6a-GFP, green; c-kit, red; CD31, blue. **(B)** Percentages of proliferating (c-kit+PHH3+) cells per IAHC (c-kit+) cells in the aorta of early E10 (<30sp), mid-E10 (30-35sp) and late E10 (>35sp) embryos. Bars, Means ± Standard Deviation. *P values <0.05. **(C)** Principal Component Analysis (PCA) from 1,196 differentially expressed genes. Stages of E10 embryos: Early-E10 (26-32sp), mid-E10 (33-35sp) and late-E10 (40-47sp). **(D)** Clustering analysis of 1,196 significantly differentially expressed genes in between early-/mid-E10, late-E10 and E11.5 samples. **(E)** Clustering analysis of 107 differentially expressed transcription factor genes in between early/mid-E10, late-E10 and E11.5 samples. See also Figures S4 and S5.

Punctual molecular changes occur in IAHCs between E10 and E11.5, along with the pre-HSC maturation process

To decipher whether there are fundamental principles underlying the pre-HSC maturation process, we further examined the transcriptional landscape of IAHC cells by RNA-sequencing. We isolated samples of 100 c-kit$^+$ IAHC cells from early, mid and late-E10, and E11.5 embryos. After sample preparation, sequencing and analysis, 1,196 genes were found significantly differentially expressed in between the different samples. A Principal Component Analysis (PCA) was performed to visualize the variance among the different samples (Figure 3C). Samples of the same developmental stage clustered together, and were closer, but not identical, to the successive stages. We then performed unsupervised hierarchical clustering based on the 1,196 genes found. We identified clusters of genes specifically expressed by early/mid-E10, late-E10, and E11.5 IAHC cells (Figure 3D). Mid-E10 IAHC cells seemed to be in a transitional transcriptional state in between early and late-E10 stages. We grouped the differentially expressed genes (up or down-regulated) according to their function by using gene ontology (Figure S5A) and pathway (Figure S5B) analysis. Many genes implicated in cell adhesion, cytoskeleton and extra-cellular matrix remodeling were highly regulated between mid and late-E10. Several known and novel pathways were up-regulated from mid to late-E10 while most of them were down-regulated by E11.5, suggesting that they act only during a short period of time. Restricted clustering analysis identified 107 differentially expressed transcription factor encoding genes that might initiate the different steps of the pre-HSC maturation process (Figure 3E).

Discussion

Although IAHCs have been observed more than a century ago in the main embryonic arteries of all vertebrate species, the role of IAHC cells remains unclear to date. We demonstrate here that IAHCs contain pre-HSCs able to mature *in vivo* into HSCs and therefore could contribute to the increase of the HSC pool at mid-gestation. We also show that the maturation process features several successive molecular changes occurring at precise developmental time points.

The functional characterization of IAHCs has been hindered by the difficulty to isolate IAHC cells to purity. Indeed, contaminating c-kit$^+$ cells are also present in the circulating blood and outside the aorta[9,35].[9,35] By injecting anti-c-kit antibodies directly inside the aorta prior to tissue dissociation, we could specifically label [34] and sort all IAHC cells to purity. All IAHC cells are positive for c-kit and many other hematopoietic and endothelial markers. However, we could further subdivide IAHC cells based on their differential expression of Ly6a-GFP and CD45.[9,23,38] Interestingly, this phenotypic heterogeneity in IAHCs did not correlate to a specific cell function at E10.

The AGM contains very few HSCs at mid-gestation (<3 HSCs at E10/E11)[6,28] though contains large numbers of myeloid progenitors. B cell progenitors were also reported in the AGM after culture on the stromal cell line TSt-4 [39]. We show here that IAHCs, when tested directly after dissection without pre-culture step, contain only very few CFU myeloid progenitors and no CFU pre-B lymphoid progenitors. Most AGM myeloid progenitors were in fact blood-circulating progenitors (as tested in the aortic flushed blood). They most likely derived from the yolk sac, which supports the idea that most hematopoietic progenitors are indeed of yolk sac origin.[40] Thus, HSCs and committed progenitors do not account for the high IAHC cell number. Instead, we demonstrate that IAHCs from early-E10 embryos are a reservoir of pre-HSCs. Indeed, pure IAHC cells sorted from the aorta before HSC production (early-E10), successfully reconstitute primary (immunodeficient newborns) and secondary (WT) recipients. Interestingly, we obtained reconstitution with both CD45$^+$ and CD45$^-$ cell fractions. These E10 CD45$^+$ pre-HSCs were not previously detected by using an AGM re-aggregate culture step

prior to transplantation.[41] We also witnessed by confocal imaging the first maturation steps of IAHC cells into phenotypically defined HSCs *ex vivo* in live E10 embryo slices. A massive HSC expansion rapidly occurs in the placenta and fetal liver at mid-gestation. The fetal liver is most likely colonized by HSCs generated in other organs, but is not capable to autonomously produce them.[42,43] The HSC number goes from an average of 11 to 152 HSCs in the total mouse conceptus in 24h, from E11 to E12.[6,28] It could be due to an increase of HSC self-renewal (HSCs are highly cycling in the fetal liver[44,45]), and/or to a maturation of cells toward HSCs. The second hypothesis was proposed after a massive HSC increase was observed in an *in vitro* AGM re-aggregation culture while most HSCs were slow-cycling.[41,46] However, the *in vivo* relevance of this hypothesis has been lacking so far. We found that the fetal liver contains pre-HSCs at E11.5, potentially coming from the IAHCs. Pre-HSCs were also previously reported in E8/E9 P-Sp and in E9 yolk sac.[30-33] To which extent the pre-HSC populations described so far are connected and contribute after maturation to the massive HSC number increase in the fetal liver still needs to be determined.

We previously observed the emergence of Ly6a-GFP$^+$ IAHC cells from Ly6a-GFP$^+$ hemogenic endothelial cells in non-fixed thick embryo slices.[23] It was suggested that the *Ly6a* transgene exclusively marks hemogenic endothelial cells producing HSCs.[37] Nevertheless, we show here that IAHCs contain pre-HSCs and progenitors in both Ly6a-GFP$^+$ and Ly6a-GFP$^-$ fractions, indicating that Ly6a-GFP is most likely not an exclusive marker of hemogenic endothelial cells.

RNA-sequencing analysis of IAHC cells isolated from early, mid and late-E10 and E11.5 embryos identified clusters of genes specifically expressed by early/mid-E10, late-E10, and E11.5 IAHC cells. Surprisingly, we found only one gene significantly differentially expressed between early and mid-E10, despite that HSCs start to be detected at this time point of development. Mid-E10 IAHC cells seemed to be in a transcriptional transition state in between early and late-E10 stages. Gene ontology and pathway analysis revealed that many genes implicated in cell adhesion, cell remodeling and extra-cellular matrix reorganization were highly regulated between mid and late-E10. Such events are most likely correlated with the progressive detachment of IAHC cells within the aorta.[9] Several important pathways implicating known HSC regulators (e.g. *Tgf*β*, Bmp*, *Wnt*, and *Notch*) were upregulated from mid to late-E10 and downregulated by E11.5. Such pathways may be important in a short time window to trigger pre-HSC maturation that will lead to the onset of HSC activity.

Thus, IAHC cells transitioning from a pre-HSC towards a HSC fate, display different transcriptome signatures during a short developmental window. The sequential steps of maturation, unsuspected so far in the aorta, are most likely needed to reach a fully potent HSC state. The lack of engraftment of pre-HSCs in adult WT recipients, and at low levels in other types of recipients (i.e. immunodeficient recipients, neonates), might thus be the result of a fundamentally immature molecular state. Our data provide an important source of information to identify potent new regulators needed for the maturation of pre-HSC towards a HSC fate *in vivo* during ontogeny. Such information would also be useful to induce hematopoietic specification *in vitro*.

The direct observation of the aorta by time-lapse confocal microscopy, *in vivo* in zebrafish embryos[20,21] and *ex vivo* in non-fixed mouse embryo slices[23], definitely confirmed that EHT occurs in the aorta. We now propose a model where hemogenic endothelial cells do not directly form fully potent HSCs, but rather establish first an intermediate cell population referred to as pre-HSCs (organized in IAHCs in the aorta). These cells will mature towards an HSC fate via successive steps (with phenotypic and molecular changes) that are initiated inside the aorta and possibly completed in the fetal liver and placenta, the two main HSC reservoirs at mid-gestation.[47]

Our study has important clinical significance. One of the goals in the field of regenerative medicine is to reprogram somatic cells into HSCs, notably by combined expression of transcription factors. So far, success has been limited. Our study provides a better understanding of the cellular and molecular events driving pre-HSCs to become fully competent transplantable HSCs *in vivo* during development, which will be essential in defining experimental conditions to mimic this process *in vitro*.

Materials and methods

Ethics Statement

Mice were housed according to institutional guidelines and all animal procedures were carried out in compliance with the Standards for human care and use of laboratory animals.

Embryo generation

Mouse embryos were generated from timed matings. Observation of vaginal plugs was considered as day 0 of embryonic development. WT embryos were generated in the C57BL/6, Ly5.1 or FVB/NJ backgrounds. *Ly6a-GFP* embryos[48] were generated by crossing *Ly6a-GFP*$^{+/-}$ males with (C57Bl/10 x CBA) females, and *Rag2*$^{-/-}\gamma c^{-/-}$ pups by crossing *Rag2*$^{-/-}\gamma c^{-/-}$ males and females.

Scanning electron microscopy

E10 (28-34sp) WT C57BL/6 embryos were separated from placenta, yolk sac and amnion. Head and tail were cut. Blood in the aorta of the remaining caudal half was removed through injection of PBS/FCS (PBS supplemented with 10% fetal calf serum, penicillin (100 U/ml) and streptomycin (100 mg/ml)) inside the dorsal aorta using a pulled glass capillary and a glass syringe. Next, the embryo caudal parts were fixed overnight in osmium tetroxide, washed several times in PBS and subsequently sectioned with a Lancer Vibratome Series 1000 (Technical Products International) into transversal sections of 200 µm. Sections were fixed in 1.5% glutaraldehyde in 0.1M cacodylate buffer, dehydrated in a graded ethanol series, critical point dried over CO_2, sputter-coated with gold-palladium and viewed in a Jeol JSM 7600F Scanning Electron Microscope at 5 kV.

Confocal microscopy of non-fixed embryos (slices or whole)

Tissue preparation was as previously described.[23,34] Briefly, E10 (32-35sp) and E11 *Ly6a-GFP* embryos were freed from placenta, yolk sac, amnion, head and tail. Non-fixed embryo trunks were stained by intra-aortic injection of antibodies or only PBS/FCS (to remove blood), and either cut into 200 µm transversal slices with a tissue chopper (McIlwain) or dissected to remove the dorsal tissues (for whole aorta floor observation).

Tissues (slices or whole) were embedded in agarose gel and observed by confocal microscopy (Leica Microsystem). In the case of *ex vivo* time-lapse imaging, embryos were stained with AlexaFluor647 anti-CD31 and PE anti-c-kit antibodies before slicing. After slicing and embedding, the embryo slices were observed for 13 h. Time-lapse videos were reconstructed using ImageJ. For multicolor z-stack pictures, thick embryo slices were stained after cutting by incubation with Pacific Blue or APC anti-CD34 (RAM34), PE anti-c-kit (2B8), PE anti-flk-1 (Avas12a1), APC anti-CD45 (30-F11), PE anti-CD41 (MWReg30), PE anti-Tie2 (TEK4), AlexaFluor 647 anti-CD31 (MEC13.3), AlexaFluor 647 anti-VE-Cadherin (eBioBV13) antibodies (from BD Pharmingen, eBioscience, Invitrogen, Santa Cruz, BioLegend), and observed by confocal microscopy. Images were edited with the Leica Analysis Software, or with Volocity (Perkin Elmer) for 3D images or video reconstructions. For video 2 and 3, E10 (30-34sp) Ly6a-GFP embryos were used.

Embryonic tissue isolation and cell preparation
Pregnant mice were sacrificed by cervical dislocation to collect E10 and E11 embryonic tissues. *Ly6a-GFP* embryos were checked under the fluorescent microscope for GFP expression. All embryonic tissues were dissected and enzymatically dissociated as previously described.[49] Briefly, after removal of the mother uterus, the embryos were isolated. Intra-embryonic AGM and fetal liver were further carefully dissected away from all other embryonic tissues. The intra-aortic blood was removed by intra-aortic injection of PBS/FCS, or by the injection of antibodies (in the case of intra-aortic staining) with a glass capillary. The flushed blood was collected and filtered on a 40 µm mesh. Flushed AGM were enzymatically digested in a 0.125% collagenase solution in PBS/FCS for 1 hour at 37ºC. Fetal livers were crushed through a 40 µm nylon cell strainer. Single cell suspensions were used for flow cytometry analysis, clonogenic assays and/or transplantations.

Flow cytometry analysis and cell sorting
After tissue isolation and dissociation, cells were counted after Trypan Blue staining to exclude the dead cells in a Bürker Türk counting chamber. Cells were then stained with PE anti-c-kit (2B8) and APC anti-CD45 (30-F11) antibodies for 30 minutes at 4˚C and washed. Alternatively PE anti-c-kit antibodies were injected intra-aortic with a glass capillary prior to AGM dissection and dissociation. Cells were analyzed and sorted with an Aria III flow cytometer (Becton Dickinson). 7-aminoactinomycin D (7-AAD) (Invitrogen) or Hoechst 33258 (Invitrogen, Molecular probes) were added to the cell suspension to discriminate dead from alive cells.

Hematopoietic progenitor assays
For myeloid progenitor clonogenic assay, different doses of cells were plated in methylcellulose (M3434; StemCell Technologies). After 12 days of culture at 37ºC, colonies were identified and counted by microscopic observation. This assay allows identifying five types of clonogenic progenitors (CFU Myeloid): CFU-GEMM (Colony Forming Unit-Granulocytes Erythrocytes Macrophages Megakaryocytes), CFU-GM (CFU-Granulocytes Macrophages), CFU-M (CFU-Macrophages), CFU-G and BFU-E (Burst Forming Unit-Erythroid). For pre-B progenitor clonogenic assay, different doses of cells were plated in methylcellulose (M3630; StemCell Technologies) and the colonies were counted after 7 days of culture. Unfractionated adult bone marrow cells (used as a positive control) provided a similar number of CFU pre-B progenitors as mentioned by the manufacturer (data not shown). For Megakaryocyte progenitor (CFU-MK) assay, cells were transferred into MegaCult®-C Medium (StemCell Technologies Inc.; Vancouver, Canada) supplemented with recombinase human-Thrombopoietin (rh-TPO, 50 ng/ml) and recombinase murine-Interleukin-3 (rm-IL-3, 10 ng/ml). Cultures were prepared according to the manufacturer's instructions and incubated in humidified 5% CO2 incubator at 37°C. After 8 days of culture, slides were fixed in cold acetone (15 min), stained, and CFU-MK colonies were scored. CFU-MKs were identified by the detection of acetylcholinesterase activity of megakaryocytes. A CFU-MK colony was defined as a cluster of three or more MK cells detected by light microscopy. Up to 8 embryo equivalent (ee) of cells were seeded per plate for all CFU assays.

Intra-liver newborn transplantations, analysis, and secondary transplantations
After sort, IAHC cells were washed and suspended in 15-20 µL of PBS before injection in the liver of 1-5 days old *Rag2*$^{-/-}\gamma c^{-/-}$ or WT irradiated newborns (3 and 5 Gy respectively, ^{137}Cs-source). After up to 5 months, donor chimerism was analyzed on bone marrow, spleen and peripheral blood (red blood cells were lysed with IOTest® 3 Lysing Solution, Beckman Coulter) of the grown-up neonates. Thymus and lymph nodes were only analyzed in WT recipients. The presence of donor contribution was determined by flow cytometry (LSR II, Becton Dickinson) after staining with APC anti-H2k^{k} antibody (H100-27.R55) and by Ly6a-GFP expression (both of donor origin) for *Rag2*$^{-/-}\gamma c^{-/-}$

recipients ($H2k^d$), or PE anti-CD45.1 antibody for WT recipients (CD45.2). Dead cells were excluded with DAPI or Hoechst 33258 (Invitrogen, Molecular probes). Expression of $H2k^d$ (recipient) and $H2k^k$ (donor) were tested in BALB/c mice (same background as the recipient *Rag2*$^{-/-}\gamma c^{-/-}$ mice) (data not shown). In addition, bone marrow, spleen and peripheral blood cells were analyzed by semi-quantitative Polymerase Chain Reaction (PCR) for the presence of the *Ly6a-GFP* transgene (data not shown). Multilineage analyses were carried out on bone marrow, spleen and peripheral blood of the transplanted grown-up neonates by using APC-Cy7 or APC efluor780 anti-Mac-1 (M1/70), APC-Cy7 anti-Gr-1 (RB6-8C5), AlexaFluor 700 anti-B220 (RA3-6B2), PE-Cy7 anti-CD45 (30-F11), APC anti-CD3 and Per-CP-Cy5.5 anti-c-kit (2B8) antibodies (from BD Pharmingen, eBioscience, BioLegend).

For secondary transplantations, bone marrow and spleen cells from primary reconstituted recipients were isolated and suspended in PBS for injection in either adult irradiated *Rag2*$^{-/-}\gamma c^{-/-}$ or WT (C57Bl/10 x CBA, or C57Bl/6) recipients (3 Gy and 9 Gy split dose, respectively). $2x10^5$ (C57Bl/10 x CBA, or C57Bl/6) spleen cells were co-injected in the WT recipients. After up to 5 months, secondary recipients were analyzed as described previously for primary recipients. (C57Bl/10 x CBA) recipients were only analyzed for GFP expression by semi-quantitative PCR (as previously described[36]) on bone marrow, spleen and peripheral blood (since they express $H2k^k$ as the donor cells). A mouse was considered repopulated when both bone marrow and spleen contained at least 0.01% of donor contribution.

Confocal microscopy of thick fixed embryo slices for IAHC proliferation assay

FVB/NJ embryos were dissected at early-E10 (<30sp), mid-E10 (30-35sp) and late-E10 (>35sp). The head and tail were removed and the aortic blood was flushed away. The remaining non-fixed embryo trunk was cut into 200 µm slices with a tissue chopper. Slices were fixed 30 minutes in 2% Para-Formaldehyde at 4˚C, washed in PBS and dehydrated in methanol. After rehydration, slices were sequentially incubated with anti-c-kit (2B8)/AlexaFluor 647 anti-rat IgG, biotin anti-CD31 (MEC13.3)/AlexaFluor 594 streptavidin and anti-phospho-Histone H3.3 (PHH3)/AlexaFluor 488 anti-rabbit IgG antibodies as previously described [50]. Slices were cleared with a benzyl alcohol benzyl benzoate (BABB) solution and observed by confocal microscopy (Leica Microsystem). z-stack images of the aorta were obtained with a 20x Epiplan-Neofluar dry lens. Whole mount embryo slices were imaged in 3D (xyz). The number of intra-aortic c-kit$^+$PHH3$^-$ (non-proliferative) and c-kit$^+$PHH3$^{low/+}$ (proliferative) cells was manually counted using the Leica Analysis Software. IAHC cells in focus were counted per each xyz-stack (3 µm in depth) for all z-stack images constituting the embryo slices. 56 c-kit$^+$PHH3$^+$/409 c-kit$^+$ total cells (14 embryo slices), 72 c-kit$^+$PHH3$^+$/769 c-kit$^+$ total cells (27 embryo slices), 18 c-kit$^+$PHH3$^+$/362 c-kit$^+$ total cells (7 embryo slices) were counted at early, mid and late-E10, respectively. Student's *t* test was used to determine statistical significance. **P* values <0.05 were considered significant.

Sample preparation for RNA-Sequencing

c-kit$^+$ IAHC cells (intra-aortic injection) were sorted by flow cytometry in samples of 100 cells from early-E10 (26-32sp, n=20 embryos), mid-E10 (33-35sp, n=3), late-E10/early E11 (40-47sp, n=11), and E11.5 (> 47sp, n=14) embryos. mRNAs were processed and amplified using single-cell procedures as previously described.[51] Briefly, cells were sorted directly into a mild lysis buffer. From the total cell lysate, poly(A)-tailed mRNAs were reversely transcribed using poly(T) primers, with an optimal reverse transcription up to 3 kb from the 3'end of the mRNAs. After poly(A) tailing of the single-stranded cDNAs, double stranded cDNAs were synthesized and amplified by PCR using poly(T) primers. The quality of cDNA synthesis and amplification was assessed by qPCR analysis and electrophoresis and 500 bp-3 kb fragments were size selected from agarose gel. The cDNA samples were sheared by Covaris sonication to 150-250 bp

fragment sizes and purified from agarose gel. This material was used for standard Illumina True-seq DNA sample preparation, starting from the DNA end-repair protocol. The final samples were single-end sequenced on a HS2500 sequencer using HS Rapid flowcells, generating 50 bp reads.

RNA-Sequencing analysis

Reads were aligned using Tophat2[52] against the mouse genome data obtained from the build 37 assembly by NCBI (mm9). Non-uniquely mapped reads were discarded from the analysis. Uniquely mapped reads were counted per exon of each RefSeq gene model and reads per kilobase of a concatenated exon length per million mapped reads (RPKMs) were calculated and assigned to each gene loci using a custom R script. Genes that differentially varied across the entire experiment (adjusted p-value ≤0.05 (using Benjamini-Hochberg method), ANOVA test, and a coefficient of variation (CV) score ≥0.35) were selected for the principal components analysis (PCA) using prcomp function in R. Hierarchical clustering analysis (with distance=1-pearson correlation and method=complete) was done on log2-transformed RPKM values using Gene-E (http://www.broadinstitute.org/cancer/software/GENE-E/). Differentially expressed gene analysis was performed by DESeq[53] using the non-adjusted read counts for each gene loci. Genes, which were significant at an adjusted p-value ≤0.05 (using a Benjamini-Hochberg method) were considered differentially expressed between inter-populations.

Gene ontology and pathway analysis were done using GO-term and pathway analysis tools of MetaCoreTM from Thomson Reuters.

Acknowledgments

We thank the experimental animal center (Experimental Dieren Centrum, Erasmus MC) for mouse care and the Optical Imaging Center of the Erasmus MC for confocal microscope access. We thank Reinier van der Linden and Julien Karrich for help with cell sorting and Elize Haasdijk for initial help with the vibratome. We thank Wilfred van IJcken and the Biomics Department of the Erasmus MC for sequencing. We thank Niels Galjart at the Erasmus MC for access to the Volocity software. We thank Charlotte Andrieu-Soler for initial help with RNA isolation. We thank Jacqueline Deschamps for careful reading of the manuscript.

Competing interest

The authors have no financial, personal, or professional competing interests.

Author contributions

C.R. and J.-C.B. conceived ideas and designed the research. J.-C.B. performed most experiments with the help of C.R. and T. Clapes. C.B. and A. v. d. H. prepared and validated samples for RNA Seq. P.K. initially performed RNA Seq analysis. S.T. performed all RNA Seq analysis. F.G. and E.S. supervised the RNA Seq experiments. J.O. performed scanning electron microscopy, supervised by M.M. N.P. performed the initial neonate transplantations, supervised by T. Cupedo. A.K. performed the experiments in Fig.3a,b, Extended Data Fig.5 and 6. J.-C.B. and C.R. analyzed the data, interpreted the experiments, made the figures and wrote the paper. All authors discussed the results and commented on the manuscript.

Financial disclosure

This work was supported by the ERC grant (project number 220-H75001 EU/ HSCOrigin – 309361) and the Dutch Cancer Genomics Center (project number 93511024). C.R. and J.-C.B. are partly supported by NWO (VIDI) grant (917-76-345), and T. Clapes is supported by Landsteiner Foundation for Blood Transfusion Research (LSBR 1025).

Supplemental Tables and Figures

Table S1. IAHC pre-HSCs are able to reconstitute secondary adult recipients

Primary neonate recipient initially injected with:	Secondary recipients * (Number repopulated/Number injected)**			
	Rag2$^{-/-}\gamma c^{-/-}$ injected with cells from:		wt injected with cells from:	
	Bone marrow	Spleen	Bone marrow	Spleen
750 c-Kit^{+} cells from E10 AGM (25-32 sp)	1/1	1/1	1/1	1/1
50 c-Kit^{+} cells from E10 AGM (33-38 sp)	-	-	3/3	3/3

*Secondary recipients injected with cells from primary recipients (indicated with ¶ in Table 1).
**Number of repopulated recipients per injected recipients 4-5 months post-injection. Mice were considered repopulated when donor derived cells were detected in both bone marrow and spleen, by flow cytometry of the H2kk and Ly6a-GFP donor markers, and semi-quantitative PCR of the Ly6a-GFP transgene.

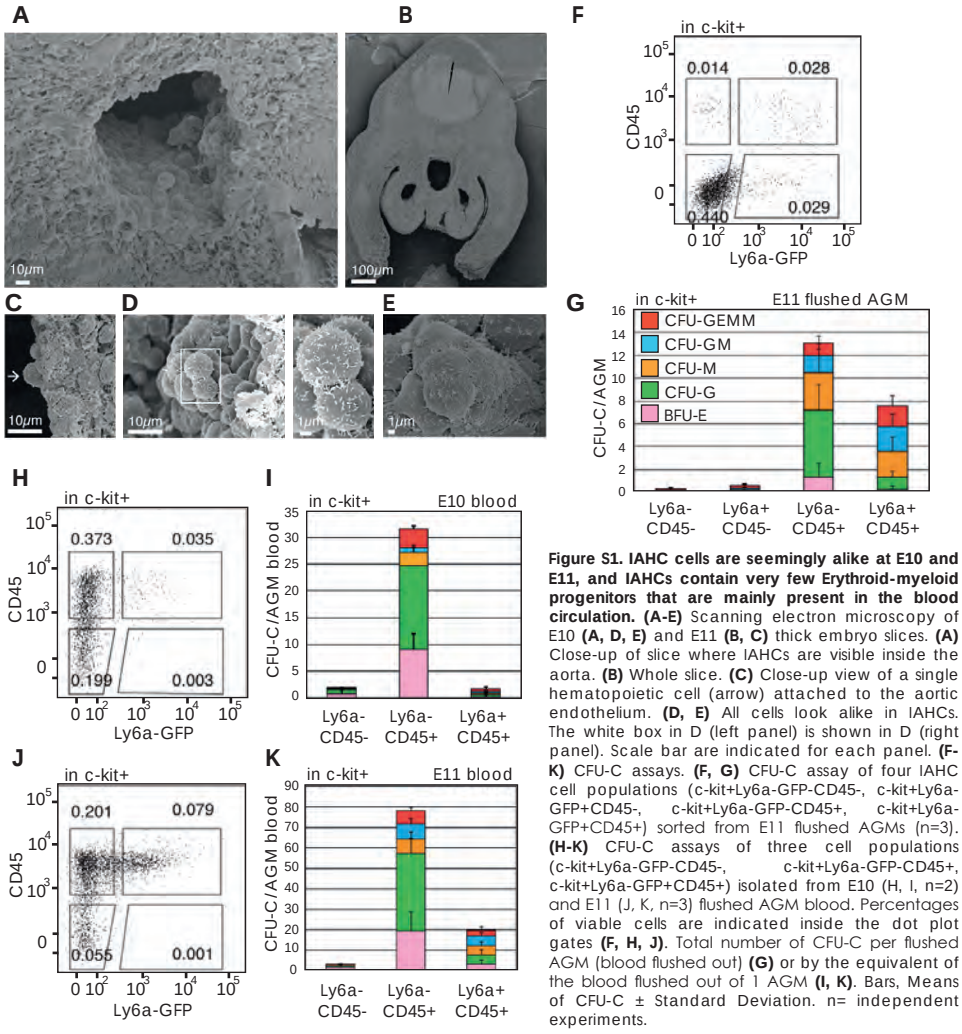


Figure S1. IAHC cells are seemingly alike at E10 and E11, and IAHCs contain very few Erythroid-myeloid progenitors that are mainly present in the blood circulation. (A-E) Scanning electron microscopy of E10 **(A, D, E)** and E11 **(B, C)** thick embryo slices. **(A)** Close-up of slice where IAHCs are visible inside the aorta. **(B)** Whole slice. **(C)** Close-up view of a single hematopoietic cell (arrow) attached to the aortic endothelium. **(D, E)** All cells look alike in IAHCs. The white box in D (left panel) is shown in D (right panel). Scale bar are indicated for each panel. **(F-K)** CFU-C assays. **(F, G)** CFU-C assay of four IAHC cell populations (c-kit+Ly6a-GFP-CD45-, c-kit+Ly6a-GFP+CD45-, c-kit+Ly6a-GFP-CD45+, c-kit+Ly6a-GFP+CD45+) sorted from E11 flushed AGMs (n=3). **(H-K)** CFU-C assays of three cell populations (c-kit+Ly6a-GFP-CD45-, c-kit+Ly6a-GFP-CD45+, c-kit+Ly6a-GFP+CD45+) isolated from E10 (H, I, n=2) and E11 (J, K, n=3) flushed AGM blood. Percentages of viable cells are indicated inside the dot plot gates **(F, H, J)**. Total number of CFU-C per flushed AGM (blood flushed out) **(G)** or by the equivalent of the blood flushed out of 1 AGM **(I, K)**. Bars, Means of CFU-C ± Standard Deviation. n= independent experiments.

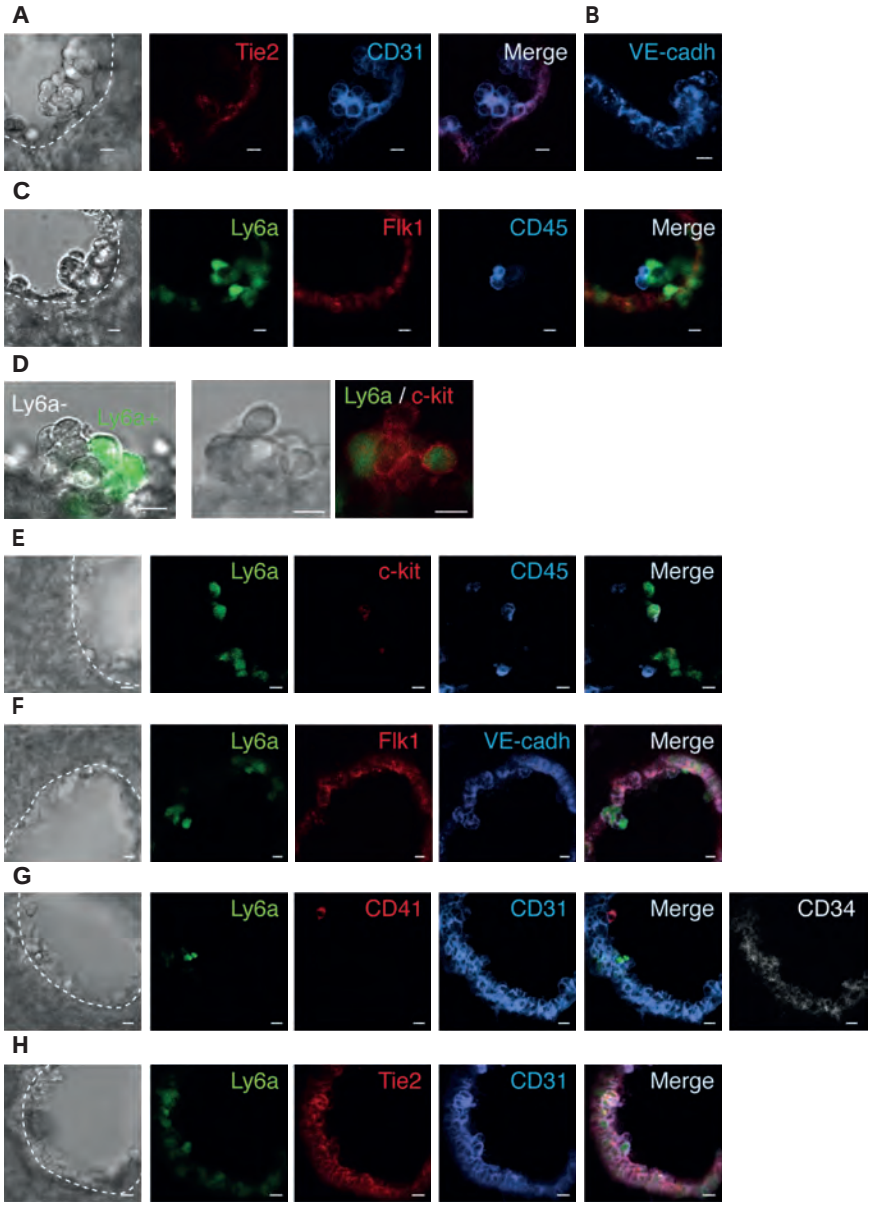


Figure S2. IAHC cells differentially express Ly6a-GFP and CD45. Visualization of IAHC cells by confocal imaging on non-fixed E10 **(A-D)** and E11 **(E-H)** Ly6a-GFP embryo slices (except for **A, B**: WT embryos). Ly6a-GFP (green) embryo slices stained with the indicated antibodies directly labeled with phycoerythrin (red), APC (blue), Alexafluor647 (blue) or pacific blue (grey). Image orientation: ventral side of the embryo downward. Left panels: transmitted light, middle panels: individual fluorescent signals, right panels: merged fluorescent signals. Scale bar, 10 µm. Dashed line: aortic endothelium location. For information, **A** and **B** show two different embryo slices.

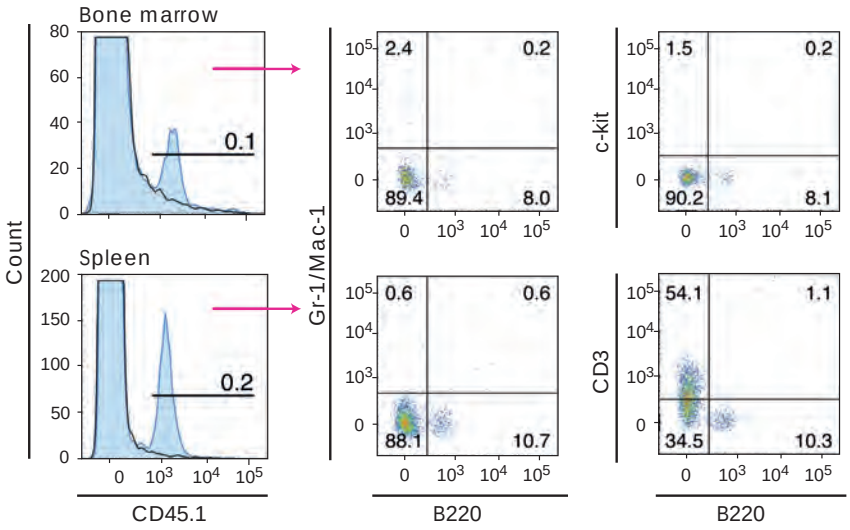


Figure S3. IAHCs from early E10 embryos contain pre-HSCs able of long-term multilineage hematopoietic reconstitution after transplantation in WT neonates. Analysis of a WT neonate recipient (CD45.2) transplanted with 200 IAHC c-kit+ cells from early-E10 (25-32sp) AGM (CD45.1) up to 4 months post-transplantation. FACS analyses show donor cell contribution (CD45.1) in bone marrow and spleen represented in histogram on the left panel (CD45.1: blue, Control: black line). Lines indicate the percentages of donor contribution in the whole tissue. Multilineage donor contribution (dot plots) was analyzed in bone marrow and spleen for myeloid (Gr-1/Mac-1) and B cells (B220), in bone marrow for hematopoietic stem/progenitor cells (c-kit), and in spleen for T cells (CD3). Percentages of each donor population are indicated per quadrant.

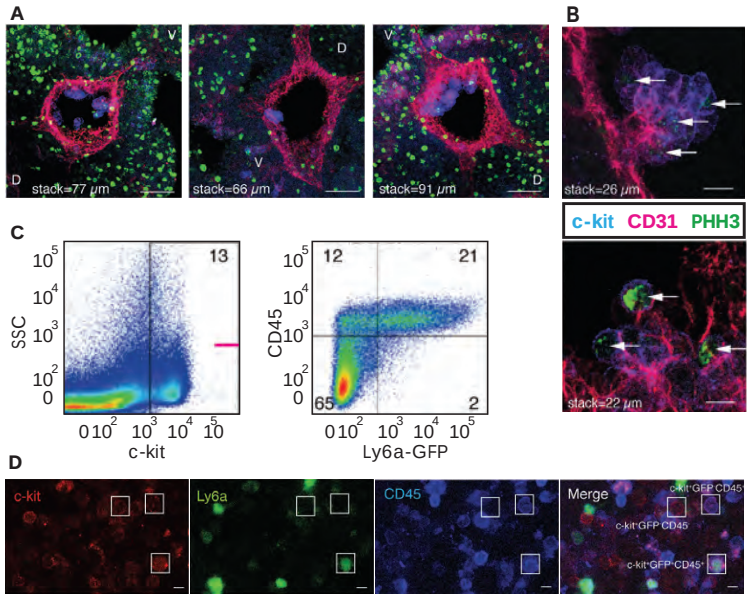


Figure S4. Pre-HSCs are low proliferative in the aorta, and are at later stages present in the fetal liver. (A, B) Visualization of IAHC cells by confocal imaging on fixed thick E10 WT embryo slices. Embryo slices were stained with c-kit (blue), CD31 (red) and PHH3 (green). **(A)** Views of the aorta and surrounding mesenchyme show cell proliferation outside the aorta. Scale bar, 50 µm. **(B)** Close-up of IAHCs showing PHH3+ cells (arrows). Scale bar, 10 µm. V, ventral; D, dorsal. **(C, D)** Visualization of phenotypically defined pre-HSCs by FACS and confocal imaging on fixed thick Ly6a-GFP liver slices. **(C)** FACS analysis of fetal liver cells isolated from E12 Ly6a-GFP embryos, after staining with c-kit and CD45 antibodies. Percentages of viable cells are indicated per gate and quadrant. **(D)** E12 Ly6a-GFP (green) non-fixed fetal liver slices stained with c-kit (red) and CD45 (blue). White boxes show examples of the three c-kit+ cell populations differentially expressing Ly6a-GFP and CD45 on the merged fluorescent signal panel. Scale bar, 10 µm.

A

		GO processes	p-value	Gene counts
Mid-E10 versus Late-E10	Down	Small molecule metabolic process	1.934E-05	18
		Regulation of cell proliferation	5.113E-03	7
		rRNA processing	9.493E-06	6
		Methylation	9.808E-04	5
	Up	Multicellular organismal development	1.364E-18	77
		Cell adhesion	2.754E-27	69
		Extracellular matrix organization	4.833E-39	60
		Positive regulation of transcription, DNA-dependent	2.902E-10	43
		Cell migration	9.601E-09	20
		Cell-cell signaling	3.987E-06	20
		Cellular response to growth factor stimulus	8.859E-05	15
		Regulation of cell migration	8.246E-06	10
		Canonical Wnt receptor signaling pathway	5.010E-05	10
		Positive regulation of Wnt receptor signaling pathway	4.473E-05	6
		Ephrin receptor signaling pathway	6.269E-05	6
		Positive regulation of SMAD protein import into nucleus	1.941E-06	5
Late-E10 versus E11	Down	Multicellular organismal development	5.977E-35	168
		Cell adhesion	2.257E-41	132
		Extracellular matrix organization	2.802E-55	104
		Cell differentiation	4.437E-19	103
		Positive regulation of transcription, DNA-dependent	8.530E-16	88
		Wnt receptor signaling pathway	2.879E-14	43
		Response to cytokine stimulus	6.567E-09	26
		Notch signaling pathway	1.993E-08	26
		Integrin-mediated signaling pathway	8.077E-07	22
	Up	Gene expression	1.916E-32	81
		Translation	6.150E-39	68
		Small molecule metabolic process	2.296E-06	63
		RNA splicing	7.572E-04	17
		Cytokine-mediated signaling pathway	1.270E-03	17
		Leukocyte migration	6.941E-05	13
		Neutrophil chemotaxis	1.243E-06	10
		Hemopoiesis	1.203E-03	9

B

		Pathway maps	p-value	Gene counts
Mid-E10 versus Late-E10	Down	dATP/dITP metabolism	5.248E-03	3
	Up	Cell adhesion_ECM remodeling	5.077E-09	10
		Development_Regulation of epithelial-to-mesenchymal transition (EMT)	5.668E-06	8
		Development_TGF-beta-dependent induction of EMT via SMADs	8.472E-07	7
		Cell adhesion_Chemokines and adhesion	8.767E-04	7
		Development_WNT signaling pathway	1.300E-03	5
		Development_PDGF signaling via STATs and NF-kB	1.406E-03	4
		Cell adhesion_Ephrin signaling	4.992E-03	4
		Cell adhesion_Endothelial cell contacts by junctional mechanisms	7.253E-03	3
		Development_BMP signaling	1.407E-02	3
Late-E10 versus E11	Down	Development_Regulation of epithelial-to-mesenchymal transition (EMT)	2.445E-15	21
		Cell adhesion_Chemokines and adhesion	2.816E-10	20
		Cytoskeleton remodeling_TGF, WNT and cytoskeletal remodeling	1.965E-09	20
		Cytoskeleton remodeling_Cytoskeleton remodeling	1.816E-08	18
		Cell adhesion_ECM remodeling	2.335E-09	14
		Development_TGF-beta-dependent induction of EMT via SMADs	9.524E-11	13
		Development_WNT signaling pathway	3.024E-08	13
		Cell adhesion_Histamine H1 R signaling in interruption of cell barrier integrity	3.761E-08	12
		Cell adhesion_Integrin-mediated cell adhesion and migration	8.268E-08	12
		Development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK	4.731E-07	11
		Development_TGF-beta-dependent induction of EMT via MAPK	5.981E-07	11
		Cell adhesion_Ephrin signaling	2.539E-05	9
		Cell adhesion_Integrin inside-out signaling	1.533E-04	9
		Cytoskeleton remodeling_Integrin outside-in signaling	3.190E-04	8
		Cell adhesion_Endothelial cell contacts by non-junctional mechanisms	1.488E-05	7
		Cell adhesion_Endothelial cell contacts by junctional mechanisms	2.657E-05	7
		Development_Hedgehog signaling	1.166E-03	7
		Development_WNT5A signaling	1.329E-03	7
		Cytoskeleton remodeling_Role of PDGFs in cell migration	1.606E-04	6
		Development_NOTCH-induced EMT	4.150E-03	4
	Up	ATP/ITP metabolism	3.537E-04	9
		Chemotaxis_CCL2-induced chemotaxis	5.531E-05	7
		Development_c-Kit ligand signaling pathway during hemopoiesis	7.167E-04	6
		Chemotaxis_CXCR4 signaling pathway	2.961E-03	4
		Cell adhesion_Integrin inside-out signaling	1.736E-02	4
		Development_Thrombopoetin signaling via JAK-STAT pathway	6.657E-03	3

Figure S5. Variations in gene functions and in biological pathways occurring during pre-HSC maturation. (A) Selected GO categories and **(B)** selected pathways categories derived from up and down regulated genes between mid-E10 and late-E10/ early-E11, and between late-E10/early-E11 and E11.5 samples.

References

1. Boisset JC, Robin C. On the origin of hematopoietic stem cells: progress and controversy. *Stem Cell Res*. 2012;8(1):1-13.
2. Muller AM, Medvinsky A, Strouboulis J, Grosveld F, Dzierzak E. Development of hematopoietic stem cell activity in the mouse embryo. *Immunity*. 1994;1(4):291-301.
3. Medvinsky A, Dzierzak E. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell*. 1996;86(6):897-906.
4. de Bruijn MF, Speck NA, Peeters MC, Dzierzak E. Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo. *Embo J*. 2000;19(11):2465-2474.
5. Gordon-Keylock S, Sobiesiak M, Rybtsov S, Moore K, Medvinsky A. Mouse extraembryonic arterial vessels harbor precursors capable of maturing into definitive HSCs. *Blood*. 2013;122(14):2338-2345.
6. Gekas C, Dieterlen-Lievre F, Orkin SH, Mikkola HK. The placenta is a niche for hematopoietic stem cells. *Dev Cell*. 2005;8(3):365-375.
7. Ottersbach K, Dzierzak E. The murine placenta contains hematopoietic stem cells within the vascular labyrinth region. *Dev Cell*. 2005;8(3):377-387.
8. Christensen JL, Wright DE, Wagers AJ, Weissman IL. Circulation and chemotaxis of fetal hematopoietic stem cells. *PLoS Biol*. 2004;2(3):E75.
9. Yokomizo T, Dzierzak E. Three-dimensional cartography of hematopoietic clusters in the vasculature of whole mouse embryos. *Development*. 2010;137(21):3651-3661.
10. Dieterlen-Lievre F, Pouget C, Bollerot K, Jaffredo T. Are intra-aortic hemopoietic cells derived from endothelial cells during ontogeny? *Trends Cardiovasc Med*. 2006;16(4):128-139.
11. Taoudi S, Medvinsky A. Functional identification of the hematopoietic stem cell niche in the ventral domain of the embryonic dorsal aorta. *Proc Natl Acad Sci U S A*. 2007;104(22):9399-9403.
12. Dantschakoff V. Untersuchungen uber die Entwickelung von Blut und Bindegewebe bei Vogeln. Das lockere Bindegewebe des Huhnchens in Fetalen Leben. *Arch f mikr Anat*. 1909;73:117-181.
13. Jordan HE. Aortic Cell Clusters in Vertebrate Embryos. *Proc Natl Acad Sci U S A*. 1917;3(3):149-156.
14. Jaffredo T, Gautier R, Brajeul V, Dieterlen-Lievre F. Tracing the progeny of the aortic hemangioblast in the avian embryo. *Dev Biol*. 2000;224(2):204-214.
15. Jaffredo T, Gautier R, Eichmann A, Dieterlen-Lievre F. Intraaortic hemopoietic cells are derived from endothelial cells during ontogeny. *Development*. 1998;125(22):4575-4583.
16. Chen MJ, Yokomizo T, Zeigler BM, Dzierzak E, Speck NA. Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature*. 2009;457(7231):887-891.
17. Eilken HM, Nishikawa S, Schroeder T. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature*. 2009;457(7231):896-900.
18. Lancrin C, Sroczynska P, Stephenson C, Allen T, Kouskoff V, Lacaud G. The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature*. 2009;457(7231):892-895.
19. Zovein AC, Hofmann JJ, Lynch M, et al. Fate tracing reveals the endothelial origin of hematopoietic stem cells. *Cell Stem Cell*. 2008;3(6):625-636.
20. Bertrand JY, Chi NC, Santoso B, Teng S, Stainier DY, Traver D. Haematopoietic stem cells derive directly from aortic endothelium during development. *Nature*. 2010;464(7285):108-111.
21. Kissa K, Herbomel P. Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature*. 2010;464(7285):112-115.
22. Lam EY, Hall CJ, Crosier PS, Crosier KE, Flores MV. Live imaging of Runx1 expression in the dorsal aorta tracks the emergence of blood progenitors from endothelial cells. *Blood*. 2010;116(6):909-914.
23. Boisset JC, van Cappellen W, Andrieu-Soler C, Galjart N, Dzierzak E, Robin C. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature*. 2010;464(7285):116-120.
24. Dieterlen-Lievre F, Martin C. Diffuse intraembryonic hemopoiesis in normal and chimeric avian development. *Dev Biol*. 1981;88(1):180-191.
25. Smith RA, Glomski CA. "Hemogenic endothelium" of the embryonic aorta: Does it exist? *Dev Comp Immunol*. 1982;6(2):359-368.
26. Cumano A, Godin I. Ontogeny of the hematopoietic system. *Annu Rev Immunol*. 2007;25:745-785.
27. North T, Gu TL, Stacy T, et al. Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development*. 1999;126(11):2563-2575.
28. Kumaravelu P, Hook L, Morrison AM, 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*. 2002;129(21):4891-4899.

29. Robin C, Ottersbach K, Durand C, et al. An unexpected role for IL-3 in the embryonic development of hematopoietic stem cells. *Dev Cell*. 2006;11(2):171-180.

30. Cumano A, Ferraz JC, Klaine M, Di Santo JP, Godin I. Intraembryonic, but not yolk sac hematopoietic precursors, isolated before circulation, provide long-term multilineage reconstitution. *Immunity*. 2001;15(3):477-485.

31. Yoder MC, Hiatt K. Engraftment of embryonic hematopoietic cells in conditioned newborn recipients. *Blood*. 1997;89(6):2176-2183.

32. Yoder MC, Hiatt K, Dutt P, Mukherjee P, Bodine DM, Orlic D. Characterization of definitive lymphohematopoietic stem cells in the day 9 murine yolk sac. *Immunity*. 1997;7(3):335-344.

33. Yoder MC, Hiatt K, Mukherjee P. In vivo repopulating hematopoietic stem cells are present in the murine yolk sac at day 9.0 postcoitus. *Proc Natl Acad Sci U S A*. 1997;94(13):6776-6780.

34. Boisset JC, Andrieu-Soler C, van Cappellen WA, Clapes T, Robin C. Ex vivo time-lapse confocal imaging of the mouse embryo aorta. *Nat Protoc*. 2011;6(11):1792-1805.

35. Zovein AC, Turlo KA, Ponec RM, et al. Vascular remodeling of the vitelline artery initiates extravascular emergence of hematopoietic clusters. *Blood*. 2010;116(18):3435-3444.

36. de Bruijn MF, Ma X, Robin C, Ottersbach K, Sanchez MJ, Dzierzak E. Hematopoietic stem cells localize to the endothelial cell layer in the midgestation mouse aorta. *Immunity*. 2002;16(5):673-683.

37. Chen MJ, Li Y, De Obaldia ME, et al. Erythroid/myeloid progenitors and hematopoietic stem cells originate from distinct populations of endothelial cells. *Cell Stem Cell*. 2011;9(6):541-552.

38. Mizuochi C, Fraser ST, Biasch K, et al. Intra-aortic clusters undergo endothelial to hematopoietic phenotypic transition during early embryogenesis. *PLoS One*. 2012;7(4):e35763.

39. Ohmura K, Kawamoto H, Fujimoto S, Ozaki S, Nakao K, Katsura Y. Emergence of T, B, and myeloid lineage-committed as well as multipotent hemopoietic progenitors in the aorta-gonad-mesonephros region of day 10 fetuses of the mouse. *J Immunol*. 1999;163(9):4788-4795.

40. Lux CT, Yoshimoto M, McGrath K, Conway SJ, Palis J, Yoder MC. All primitive and definitive hematopoietic progenitor cells emerging before E10 in the mouse embryo are products of the yolk sac. *Blood*. 2008;111(7):3435-3438.

41. Rybtsov S, Sobiesiak M, Taoudi S, et al. Hierarchical organization and early hematopoietic specification of the developing HSC lineage in the AGM region. *J Exp Med*. 2011;208(6):1305-1315.

42. Johnson GR, Moore MA. Role of stem cell migration in initiation of mouse foetal liver haemopoiesis. *Nature*. 1975;258(5537):726-728.

43. Houssaint E. Differentiation of the mouse hepatic primordium. II. Extrinsic origin of the haemopoietic cell line. *Cell Differ*. 1981;10(5):243-252.

44. Bowie MB, McKnight KD, Kent DG, McCaffrey L, Hoodless PA, Eaves CJ. Hematopoietic stem cells proliferate until after birth and show a reversible phase-specific engraftment defect. *J Clin Invest*. 2006;116(10):2808-2816.

45. Morrison SJ, Hemmati HD, Wandycz AM, Weissman IL. The purification and characterization of fetal liver hematopoietic stem cells. *Proc Natl Acad Sci U S A*. 1995;92(22):10302-10306.

46. Taoudi S, Gonneau C, Moore K, et al. Extensive hematopoietic stem cell generation in the AGM region via maturation of VE-cadherin+CD45+ pre-definitive HSCs. *Cell Stem Cell*. 2008;3(1):99-108.

47. Kieusseian A, Brunet de la Grange P, Burlen-Defranoux O, Godin I, Cumano A. Immature hematopoietic stem cells undergo maturation in the fetal liver. *Development*. 2012.

48. Ma X, Robin C, Ottersbach K, Dzierzak E. The Ly-6A (Sca-1) GFP transgene is expressed in all adult mouse hematopoietic stem cells. *Stem Cells*. 2002;20(6):514-521.

49. Robin C, Dzierzak E. Hematopoietic stem cell enrichment from the AGM region of the mouse embryo. *Methods Mol Med*. 2005;105:257-272.

50. Yokomizo T, Yamada-Inagawa T, Yzaguirre AD, Chen MJ, Speck NA, Dzierzak E. Whole-mount three-dimensional imaging of internally localized immunostained cells within mouse embryos. *Nat Protoc*. 2012;7(3):421-431.

51. Tang F, Barbacioru C, Nordman E, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc*. 2010;5(3):516-535.

52. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.

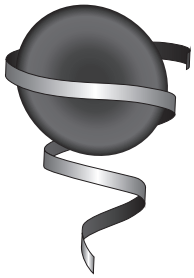
53. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.

Chapter 3

Long-range gene regulation and novel therapeutic applications

Anita van den Heuvel[1], Ralph Stadhouders[1,#], Charlotte Andrieu-Soler[2], Frank Grosveld[1] and Eric Soler[3,*].

Manuscript accepted at Blood as Blood Spotlight manuscript, January 5th 2015



[1] Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands [2]INSERM UMRS 1138, Team 17, Paris, France
[3]INSERM UMR967, CEA/DSV/iRCM, Laboratory of Molecular Hematopoiesis, Fontenay-aux-Roses, France
*correspondence: eric.soler@cea.fr
[#]current address: Centre for Genomic Regulation (CRG), Barcelona, Spain

Abstract

An intimate relationship exists between nuclear architecture and gene activity. Unraveling the fine-scale three-dimensional structure of the genome and its impact on gene regulation is a major goal of current epigenetic research, one with direct implications for understanding the molecular mechanisms underlying human phenotypic variation and disease susceptibility. In this context, the novel revolutionary genome editing technologies and emerging new ways to manipulate genome folding offer new promises for the treatment of human disorders.

Introduction

Fundamental, yet unanswered questions in biology are how genome organization and chromosomal folding influence basic cellular processes such as transcription, how they relate to the development of disease and whether they can be manipulated therapeutically. In mammals, gene regulatory elements are scattered throughout the genome, collectively occupying a significant fraction of the genomic non-coding DNA content. Initially, non-coding DNA (comprising ~98% of the human genome) was considered to be largely 'junk'-DNA, lacking function. However, the fast growing collection of genome-wide datasets describing chromatin features across increasing numbers of cell types has dramatically changed this view. These studies have started to reveal the organizational complexity of mammalian genomes, and it is at present speculated that approximately 40-80% of the genome shows a biochemical signature that could imply functional relevance.[1–3] Transcriptional enhancers represent a critical component of this non-coding 'regulatory genome' as they bestow a unique identity upon cells by establishing cell type-specific spatio-temporal gene expression patterns.[4,5] In line with their essential roles in transcriptional regulation, numerous recent studies have causally linked aberrant enhancer function to human disorders and phenotypic variation, further demonstrating the important roles played by transcriptional enhancers in human biology.[6–19] Through this review article, we aim to provide a concise update on new insights obtained in the past few years concerning the molecular mechanisms by which regulatory elements regulate gene expression, often over large genomic distances, and how disruption of these processes can contribute to the development of human disease. We will also discuss emerging therapeutic strategies aimed at manipulating the function of enhancers for the treatment of human genetic disorders.

Transcriptional regulation by enhancers is often a long-distance event

Many thousands of potential enhancers have been identified in the human genome,[1] of which thousands are active in a given cell type.[20–22] Enhancers are often localized at large distances from the genes they regulate, with an estimated median enhancer-target gene distance of 120kb,[23] although extreme cases of >1Mb have been documented.[8,12,24] They can be positioned both intragenic and intergenic, or even in non-related genes, and do not necessarily regulate transcription of the nearest gene.[25] Enhancers regulate genes over large genomic distances via chromatin looping, bringing distal enhancers and the regulatory protein complexes that bind them in close nuclear proximity to their target genes. Chromatin loop formation has therefore been shown to be a better predictor of enhancer target genes than enhancer-gene linear proximity[25] (although it is important to note that chromatin looping does not functionally connect enhancers and promoters per se).

Well-studied examples of such long-range gene regulation in the hematopoietic system are the erythroid globin[26–28], *Bcl11a*[13], *Myb*[29] and *Kit*[30] gene loci. Gene regulatory

chromatin-looping events are thought to be dynamic and actively modulated during differentiation, to accommodate for the changes in target gene expression necessary during development and cellular differentiation. Although a recent study in Drosophila indicates that enhancer-promoter loops may be remarkably stable during development.[31]

Because of the high degree of complexity and specificity required for enhancer-gene communication, chromosome conformation needs to be highly organized. A substantial body of evidence suggests that the determinants of promoter-enhancer specificity can be very diverse,[32–34] ranging from transcription factors (TFs)[30,35–38], chromatin modifying proteins[39] or so-called 'architectural' proteins (i.e. CTCF, Cohesin and Mediator)[40–44] to non-coding RNAs (e.g. eRNAs, lncRNAs)[45,46]. Enhancer-promoter interactions are promoted by the confinement of such interactions to chromosome structural domains called 'topological-associated-domains' (or TADs), which partition chromosomes into discrete sub-megabase to megabase-sized domains.[47–49] This observation suggests a 'loops within loops' model, where TADs provide a structural environment that prevents enhancer promiscuity.[50] Other studies suggest that active genes regulated by similar TF complexes tend to cluster in the nuclear space or even co-localize in specific nuclear foci referred to as transcription factories.[50–54] Although the occurrence of (active) gene movement towards relatively static transcription factories is presently still under debate[54,55], it is clear that different scales of genome folding are intimately linked to transcription regulation by allowing proper enhancer-gene contacts, placing enhancers and genome spatial organization at the heart of transcription.

Spatial genome organization and human disease

The majority of identified disease-associated genomic mutations are located in non-coding DNA regions, often co-localizing with potential regulatory sequences.[1] For several examples, single nucleotide polymorphisms (SNPs) have been shown to significantly influence long-range chromatin folding (Figure 1).[14,15] In addition, cancer cells typically show massive structural and spatial chromosomal rearrangements that potentially displace regulatory elements from their native into an ectopic environment.[6–11] As a consequence, novel long-range interactions can be established between normally unrelated enhancer-promoter pairs, leading to improper gene regulation and cellular transformation (Figure 1).[7,56–58] An elegant study by the Delwel laboratory recently reported a prime example of how aberrant enhancer 'rewiring' can cause disease.[7] They investigated acute myeloid leukemia (AML) cells bearing inv(3)/t(3;3) chromosomal rearrangements, which are characterized by the transpositioning of a *GATA2* enhancer into the *EVI1* stem cell regulator locus. This chromosomal abnormality causes leukemia through the inappropriate long-range activation of *EVI1* expression by the ectopic *GATA2* enhancer, possibly re-enforced by the accompanying reduction of *GATA2* expression. Studies by the Yamamoto group using a transgenic mouse model of inv(3) AML further confirmed these observations.[58] Another recent study investigated the molecular basis of pediatric medulloblastoma, in which complex chromosomal rearrangements activate the *GFI1* and *GFI1B* oncogenes by placing them under the transcriptional control of unrelated enhancer elements,[56] an event the authors referred to as 'enhancer hijacking'. These studies highlight the potential pathological impact of regulatory element displacement in human disease, underscoring the value of investigating spatial genomic organization when dissecting the molecular events associated with cancer.

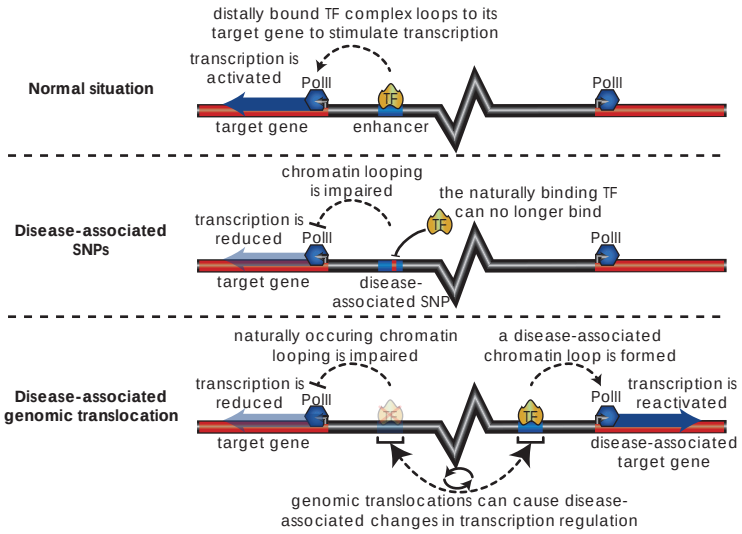


Figure 1. Genomic alterations can affect gene regulation via chromosome conformation. (Top panel) Chromatin folding plays an essential role in transcriptional control by distally located regulatory elements. **(Middle panel)** Disease-associated SNPs located in distal regulatory elements can influence long-range chromatin interactions through their detrimental effect on the recruitment of TF complexes (e.g. by destroying a TF binding motif), resulting in reduced expression of the target gene. **(Bottom panel)** Chromosomal aberrations (e.g. translocations) can relocate distal enhancers near a disease-associated gene, leading to the formation of pathological long-range chromatin interactions that ectopically activate the expression of this gene.

Therapeutical targeting of enhancers

Because of their key role in human disease, targeting these newly established disease-associated long-range interactions has become an attractive therapeutic strategy. Recent advances in genome editing technologies, such as using Zinc-Finger nucleases (ZnFs), Transcription Activator-Like Effector Nucleases (TALENs) or the CRISPR/Cas9 system (See Gaj et al.[59] and Gupta and Musunuru[60] for comprehensive reviews), make targeted enhancer-modifying strategies feasible and open-up exciting new avenues for the therapeutic manipulation of genome topology.

In the above mentioned study by Gröschel and colleagues[7], excision of the oncogenic GATA2 enhancer from the AML genome, using either TALEN or CRISPR/Cas9 mediated genome editing, *in vitro* induced growth-arrest and differentiation of the 'edited' leukemic cells, demonstrating the promising potential of hijacked oncogenic enhancers as therapeutic targets.

Enhancer activity is often highly cell type-specific and even widely expressed genes seem to possess tissue-specific enhancers driving their expression.[61] This cell type-specific nature of enhancers makes them suitable targets for tissue-specific modulation of gene expression. Targeting enhancers (rather than promoters or the gene products themselves) offers the advantage of allowing cell-type specific silencing of gene expression (Figure 2). Such a strategy would even allow reducing the levels of currently undruggable proteins, including TFs.[62] This principle was first explored by Bauer et al., using genome-editing technology to delete an erythroid-specific enhancer of the *BCL11A* TF.[13] BCL11A is widely recognized as an important therapeutic target for the treatment of β-thalassemias and sickle cell anemia (the β-hemoglobinopathies), two common erythroid genetic disorders caused by respectively a quantitative or qualitative defect in adult hemoglobin production.[63,64] In erythroid cells, BCL11A plays an important role in β-globin gene regulation and its depletion in adult erythroid cells leads to a strong reactivation of fetal (β-like) γ-globin gene expression, which can efficiently compensate for the low abundant or defective adult hemoglobin.[65–67] However, as *BCL11A* is a widely expressed transcription factor

and has been implicated in lymphomagenesis,[68,69] targeting the protein itself remains problematic. In erythroid cells, *BCL11A* expression is controlled by intronic enhancers located 55-62kb downstream of the transcription start site.[13] Targeted deletion of these enhancers dramatically reduces *BCL11A* expression specifically in erythroid cells, showing that targeting enhancers allows for the tissue-specific silencing of broadly expressed genes.

Other genome-editing based strategies to manipulate enhancer biology can be envisioned, such as the specific targeting of repressor (domain) fusion-proteins to disease-associated enhancers. Mendenhall and colleagues pioneered this approach by fusing TAL effector repeat domains to the LSD1 histone demethylase. Targeting this fusion-protein to the genome efficiently reduced enhancer activity, resulting in the downregulation of nearby genes.[70]

Manipulating the loop: artificial enhancer-hijacking for therapeutic purposes?

As stated before, TFs are involved in establishing and stabilizing long-range chromatin interactions. The non-DNA binding adaptor protein LDB1 is required for the development of multiple tissues, including the hematopoietic system.[71,72] LDB1 assembles a multi-protein TF complex in erythroid cells, containing two essential DNA-binding TFs GATA1 and TAL1, and LDB1 is required for enhancer-promoter looping at the *β-globin* and *Myb* loci.[29,36,37,73] A direct demonstration that LDB1 is the critical complex component mediating chromatin looping came from an elegant study by the Blobel laboratory.[73] During erythropoiesis, the LDB1 TF complex binds the *β-globin* gene promoter and upstream locus control region (LCR) to achieve LCR-promoter looping and high-level globin gene expression.[36,37] In *Gata1*-deficient erythroid progenitors, LDB1 is only targeted to the LCR (via its interaction with TAL1) but not to the β-globin promoter. In these cells, long-range LCR-promoter interactions are absent and the β-globin genes are therefore not expressed. Artificial ZnF-mediated tethering of LDB1, or the LDB1 dimerization domain only, to the *β-globin* promoter in these *Gata1*-deficient cells was shown to be sufficient for establishing the LCR-promoter loop, resulting in a (partial) activation of *β-globin* gene expression.[73]

This 'forced-looping' strategy was then used in an attempt to reactivate fetal *γ-globin* gene expression in adult red blood cells as a potential new strategy for the treatment of β-hemoglobinopathies.[74] Deng et al. showed that forced chromatin looping of the *β-globin* LCR to the developmentally silenced fetal *γ-globin* gene reactivates its expression in cultured adult erythroid cells. This time, they manipulated local long-

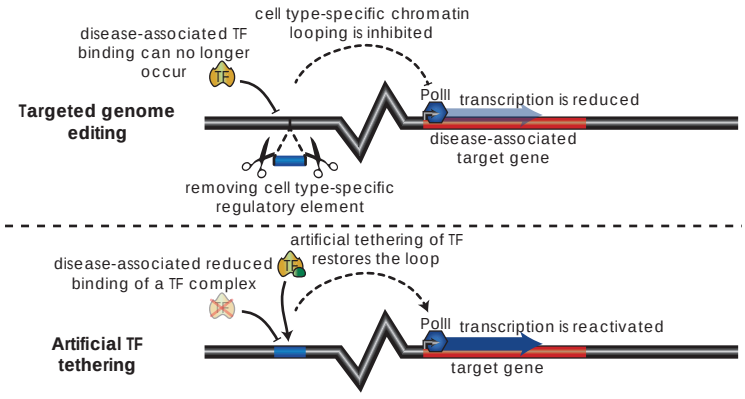


Figure 2. Enhancer targeting strategies with proven therapeutical potential. **(Top panel)** Deletion of enhancers can cell type-specifically modulate disease-associated gene expression. **(Bottom panel)** Artificial tethering of TFs to regulatory elements can restore disease-impaired chromatin looping, leading to reactivation of gene expression. Similar tethering strategies can also be used to artificially reactivate naturally silenced genes for therapeutic use,[74] or for enhancer silencing by tethering inhibitory proteins to the enhancer.[70]

range chromatin interactions by targeting the ZnF-LDB1 dimerization domain fusion-protein to the human γ-globin gene promoter. Importantly, γ-globin expression reached levels that would be sufficient to significantly ameliorate the clinical disease course of β-hemoglobinopathy patients.[75] This study provides the first proof of principle that long-range chromatin interactions can be artificially controlled - potentially even for therapeutic purposes -, providing an invaluable addition to the ever-growing genomics toolbox at our disposal (Figure 2).

Clinical enhancer targeting: remaining challenges

Despite the attractiveness of enhancer targeting as a potential broadly applicable therapeutic approach, several important challenges need to be faced before such strategies become clinically applicable. Whereas ZnF-LDB1 based reactivation of γ-globin did not appear to have a major impact on the majority of a selected set of erythroid genes controlled by the endogenous LDB1 complex, the influence of overexpressed engineered fusion proteins on the activity of endogenous regulatory proteins complexes and the genes they regulate needs to be rigorously tested. Similarly, the off-target effects of genomic engineering technologies are still under intense investigation and several studies have highlighted the need for more optimized targeting strategies.[76,77] In addition, because of the important role of regulatory elements in organizing spatial genome topology, potential side-effects of (even small) genomic alterations on local 3D chromosomal organization and gene regulation will have to be thoroughly investigated. Finally, there is the considerable challenge of efficient and specific delivery of the genome editing constructs to the target cell. With a few exceptions,[78,79] this is still only feasible using *ex vivo* cultured (stem) cells that can be transplanted back into the recipient.[80] Major technological developments in this area are required if the current genome editing methods are to be used for systemic therapy *in vivo*.

Conclusion

Enhancers that regulate gene expression over large distances by chromatin looping processes are critical for proper development and tissue homeostasis. Recent progress has clearly shown that the genetic disruption of enhancer function plays a widespread and important role in human phenotypic variation, disease susceptibility and even disease etiology. As new technologies that allow the targeted manipulation of regulatory elements develop at an astonishing rate, we expect therapeutic strategies aimed at intervening with disease-associated enhancer-gene communication or at establishing therapeutically beneficial enhancer-gene interactions to become feasible in the future.

Authorship

Contribution: All authors contributed to the writing of the paper and approved the final version.

Conflict-of-interest disclosure: All authors declare no competing financial interests

Correspondence: E. Soler INSERM UMR967, CEA/DSV/iRCM, Laboratory of Molecular Hematopoiesis, Fontenay-aux-Roses, France; email: eric.soler@cea.fr

Acknowledgements

We thank members of the Grosveld and Soler laboratories for helpful discussions. We apologize to our colleagues whose work could not be cited due to space limitations.

References

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74.
2. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 2014;111(17):6131–8.
3. Stamatoyannopoulos JA. What does our genome encode? *Genome Res.* 2012;22(9):1602–11.
4. Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* 2011;12(4):283–93.
5. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell.* 2011;144(3):327–39.
6. Farh KK-H, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2014;
7. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell.* 2014;157(2):369–81.
8. Herranz D, Ambesi-Impiombato A, Palomero T, et al. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat. Med.* 2014;20(10):1130–7.
9. Corcoran LM, Cory S, Adams JM. Transposition of the immunoglobulin heavy chain enhancer to the myc oncogene in a murine plasmacytoma. *Cell.* 1985;40(1):71–9.
10. Busslinger M, Klix N, Pfeffer P, Graninger PG, Kozmik Z. Deregulation of PAX-5 by translocation of the Emu enhancer of the IgH locus adjacent to two alternative PAX-5 promoters in a diffuse large-cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* 1996;93(12):6129–34.
11. Fahrlander PD, Sümegi J, Yang JQ, et al. Activation of the c-myc oncogene by the immunoglobulin heavy-chain gene enhancer after multiple switch region-mediated chromosome rearrangements in a murine plasmacytoma. *Proc. Natl. Acad. Sci. U. S. A.* 1985;82(11):3746–50.
12. Lettice LA. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 2003;12(14):1725–1735.
13. Bauer DE, Kamran SC, Lessard S, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science*. 2013;342(6155):253–7.
14. Stadhouders R, Aktuna S, Thongjuea S, et al. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* 2014;124(4):1699–710.
15. 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.* 2012;22(3):446–55.
16. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 2009;106(23):9362–7.
17. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 2014;46(11):1160–5.
18. Kioussis D, Vanin E, DeLange T, Flavell RA, Grosveld FG. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*. 1983;306(5944):662–6.
19. Soudon J, Bernard O, Mathieu-Mahul D, Larsen CJ. c-myc gene expression in a leukemic T-cell line bearing a t(8;14) (q24;q11) translocation. *Leukemia*. 1991;5(1):60–5.
20. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507(7493):455–61.
21. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009;459(7243):108–12.
22. Ho JWK, Jung YL, Liu T, et al. Comparative analysis of metazoan chromatin organization. *Nature*. 2014;512(7515):449–452.
23. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489(7414):109–13.
24. Velagaleti GVN, Bien-Willner GA, Northup JK, et al. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *Am. J. Hum. Genet.* 2005;76(4):652–62.
25. Smemo S, Tena JJ, Kim K-H, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*. 2014;507(7492):371–5.
26. 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(6):1453–65.
27. Hughes JR, Lower KM, Dunham I, et al. High-resolution analysis of cis-acting regulatory networks at the a-globin locus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2013;368(1620):20120361.

28. Vernimmen D. Uncovering Enhancer Functions Using the α-Globin Locus. *PLoS Genet*. 2014;10(10):e1004668.

29. Stadhouders R, Thongjuea S, Andrieu-Soler C, et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J*. 2012;31(4):986–99.

30. Jing H, Vakoc CR, Ying L, et al. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Mol. Cell*. 2008;29(2):232–42.

31. Ghavi-Helm Y, Klein F a., Pakozdi T, et al. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*. 2014;512(7512):96–100.

32. Maksimenko O, Georgiev P. Mechanisms and proteins involved in long-distance interactions. *Front. Genet*. 2014;5:28.

33. Gorkin DU, Leung D, Ren B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell*. 2014;14(6):762–775.

34. Van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol*. 2014;24(11):695–702.

35. Drissen R, Palstra R-J, Gillemans N, et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev*. 2004;18(20):2485–90.

36. Song S-H, Hou C, Dean A. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. *Mol. Cell*. 2007;28(5):810–22.

37. Krivega I, Dale RK, Dean A. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes Dev*. 2014;1278–1290.

38. Stadhouders R, de Bruijn MJW, Rother MB, et al. Pre-B cell receptor signaling induces immunoglobulin κ locus accessibility by functional redistribution of enhancer-mediated chromatin interactions. *PLoS Biol*. 2014;12(2):e1001791.

39. Kim S-I, Bultman SJ, Kiefer CM, Dean A, Bresnick EH. BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc. Natl. Acad. Sci. U. S. A*. 2009;106(7):2259–64.

40. Splinter E, Heath H, Kooren J, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*. 2006;20(17):2349–54.

41. Ribeiro de Almeida C, Stadhouders R, Thongjuea S, Soler E, Hendriks RW. DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation. *Blood*. 2012;119(26):6209–18.

42. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet*. 2014;15(4):234–46.

43. Kagey MH, Newman JJ, Bilodeau S, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*. 2010;467(7314):430–5.

44. Zuin J, Dixon JR, van der Reijden MIJA, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A*. 2014;111(3):996–1001.

45. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends Cell Biol*. 2014;24(11):651–663.

46. Ørom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell*. 2013;154(6):1190–3.

47. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):376–80.

48. Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381–5.

49. Jin F, Li Y, Dixon JR, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. 2013;503(7475):290–4.

50. Soler E, Grosveld F. Transcription regulation in Stem Cells. *Stem Cell Biol. Regen. Med. Chapter 2*, In Press.

51. De Wit E, Bouwman B a M, Zhu Y, et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*. 2013;501(7466):227–31.

52. Schoenfelder S, Sexton T, Chakalova L, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet*. 2010;42(1):53–61.

53. Ghamari A, van de Corput MPC, Thongjuea S, et al. In vivo live imaging of RNA polymerase II transcription factories in primary cells. *Genes Dev*. 2013;27(7):767–77.

54. Papantonis A, Cook PR. Transcription factories: genome organization and gene regulation. *Chem. Rev*. 2013;113(11):8683–705.

55. Cisse II, Izeddin I, Causse SZ, et al. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*. 2013;341(6146):664–7.

56. Northcott PA, Lee C, Zichner T, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*. 2014;511(7510):428–34.

57. Kovalchuk AL, Ansarah-Sobrinho C, Hakim O, et al. Mouse model of endemic Burkitt translocations reveals the long-range boundaries of Ig-mediated oncogene deregulation. *Proc. Natl. Acad. Sci. U. S. A.* 2012;109(27):10972–7.

58. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415–27.

59. Gaj T, Gersbach CA, Barbas CF. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol*. 2013;31(7):397–405.

60. Gupta RM, Musunuru K. Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. *J. Clin. Invest*. 2014;124(10):4154–4161.

61. Kieffer-Kwon K-R, Tang Z, Mathe E, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*. 2013;155(7):1507–20.

62. Koehler AN. A complex task? Direct modulation of transcription factors with small molecules. *Curr. Opin. Chem. Biol*. 2010;14(3):331–40.

63. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. *Blood*. 2010;115(22):4331–6.

64. Bauer DE, Kamran SC, Orkin SH. Reawakening fetal hemoglobin: prospects for new therapies for the β-globin disorders. *Blood*. 2012;120(15):2945–53.

65. Sankaran VG, Menne TF, Xu J, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science*. 2008;322(5909):1839–42.

66. Sankaran VG, Xu J, Ragoczy T, et al. Developmental and species-divergent globin switching are driven by BCL11A. *Nature*. 2009;460(7259):1093–7.

67. Chen Z, Luo H, Steinberg MH, Chui DHK. BCL11A represses HBG transcription in K562 cells. *Blood Cells. Mol. Dis.* 2009;42(2):144–9.

68. Kuo T-Y, Hsueh Y-P. Expression of zinc finger transcription factor Bcl11A/Evi9/CTIP1 in rat brain. *J. Neurosci. Res.* 2007;85(8):1628–36.

69. Satterwhite E. The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. *Blood*. 2001;98(12):3413–3420.

70. Mendenhall EM, Williamson KE, Reyon D, et al. Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotechnol*. 2013;31(12):1133–6.

71. Mukhopadhyay M, Teufel A, Yamashita T, et al. Functional ablation of the mouse Ldb1 gene results in severe patterning defects during gastrulation. *Development*. 2003;130:495–505.

72. Love PE, Warzecha C, Li L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends Genet*. 2014;30(1):1–9.

73. Deng W, Lee J, Wang H, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*. 2012;149(6):1233–44.

74. Deng W, Rupon JW, Krivega I, et al. Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell*. 2014;158(4):849–860.

75. Platt OS. Hydroxyurea for the treatment of sickle cell anemia. *N. Engl. J. Med*. 2008;358(13):1362–9.

76. Fu Y, Foden JA, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol*. 2013;31(9):822–6.

77. Cho SW, Kim S, Kim Y, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*. 2014;24(1):132–41.

78. Komáromy AM, Alexander JJ, Rowlan JS, et al. Gene therapy rescues cone function in congenital achromatopsia. *Hum. Mol. Genet*. 2010;19(13):2581–93.

79. Yin H, Xue W, Chen S, et al. Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nat. Biotechnol*. 2014;32(6):551–3.

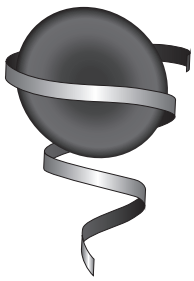
80. Mandal PK, Ferreira LMR, Collins R, et al. Efficient Ablation of Genes in Human Hematopoietic Stem and Effector Cells using CRISPR/Cas9. *Cell Stem Cell*. 2014;15(5):643–652.

Chapter 4

Multiplexed Chromosome Conformation Capture Sequencing (3C-Seq) for rapid genome-scale high-resolution detection of long-range chromatin interactions

Ralph Stadhouders[1]*, Petros Kolovos[1]*, Rutger Brouwer[2,4]*, Jessica Zuin[1], **Anita van den Heuvel**[1], Christel Kockx[2], Robert-Jan Palstra[1], Kerstin Wendt[1], Frank Grosveld[1,3], Wilfred van IJcken[2,6] & Eric Soler[1,3,5,7]

Published in *Nature Protocols*, March 2013; 8(3):509-24.



[1]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands
[2]Center for Biomics, Erasmus Medical Center, Rotterdam, The Netherlands
[3]Cancer Genomics Center, Erasmus Medical Center, Rotterdam, The Netherlands [4]Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands
[5]Laboratory of Hematopoiesis and Leukemic Stem Cells (LSHL), CEA / INSERM U967 Fontenay-aux-Roses, France
[6]Corresponding author for sequencing and bioinformatics: Dr. Ir. Wilfred van IJcken w.vanijcken@erasmusmc.nl
[7]Corresponding author: Dr. Eric Soler, eric.soler@cea.fr
*These authors contributed equally

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 (Figure 1). The

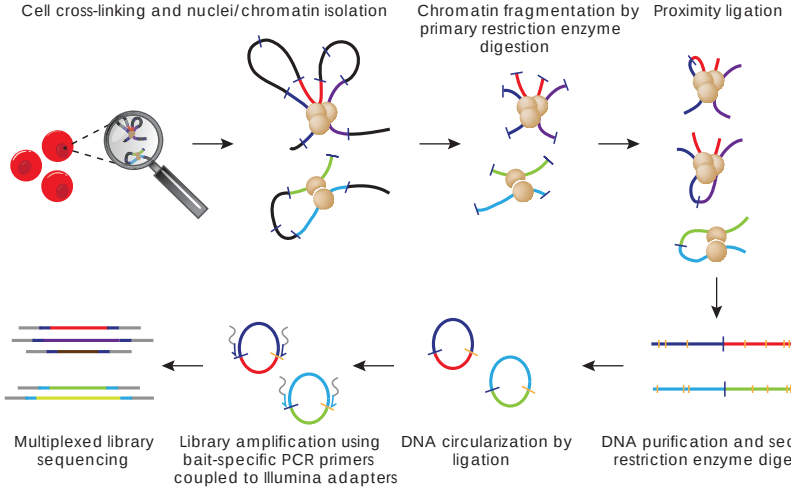


Figure 1. Overview of the multiplexed 3C-seq procedure. Nuclei from cross-linked cells are digested (primary restriction enzyme) and ligated under dilute conditions to physically link in vivo interacting DNA fragments. After a secondary digestion (secondary restriction enzyme) and ligation, inverse PCR 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.

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 (Figure 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 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

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[12,16] (3C-seq or 4C-seq)	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-wideinteractions between all individual fragments simultaneously	Obtaining high resolution requires a massive sequencing effort; expensive, complicated analysis

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 co-association 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 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 (Figure 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 (Figures 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

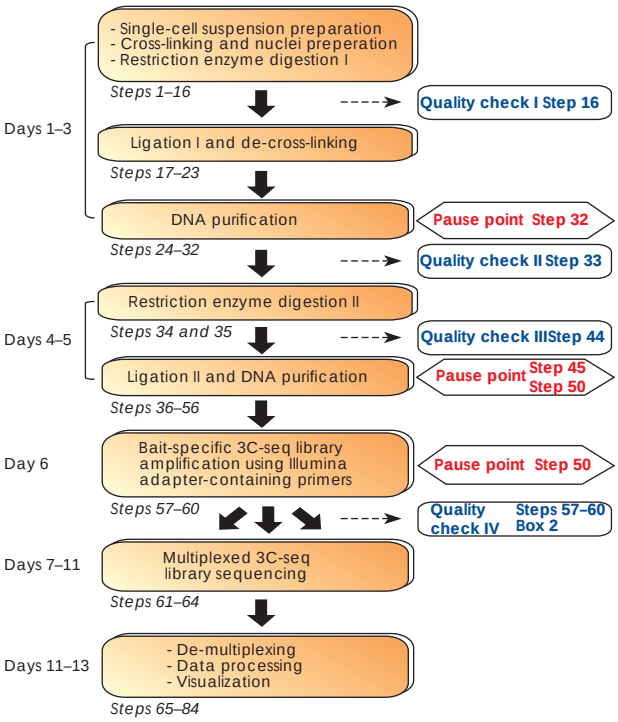


Figure 2. Flowchart of multiplexed 3C-seq data generation and processing. Steps involved in the multiplexed 3C-seq procedure are shown in blue rectangles. Time needed to complete these steps is depicted on the left. Pause points are indicated together with the timing of the different quality checkpoints: 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).

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.

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

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, l11)	Excellent	None
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		
Other mouse tissues (Mouse fetal brain E12.5-15.5)	Good	Use a collagenase treatment (PROCEDURE Step 1) to obtain a single-cell suspension for efficient cross-linking
Rat tissues (liver, heart and lung)		
Human primary melanocytes [33]	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]
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)		

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 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 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.

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 Figure 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

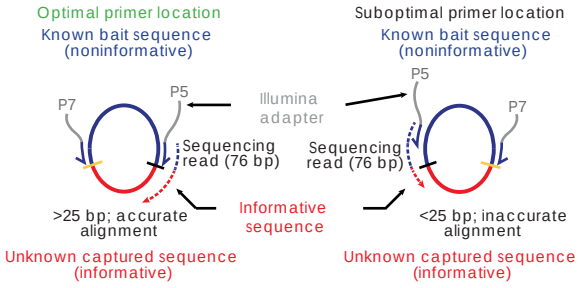


Figure 3. 3C-seq primer design and positioning. Schematic drawing of the location of the inverse PCR primers used to amplify a 3C-seq library. The ring represents a circular DNA molecule composed of the bait fragment (blue) ligated to an unknown captured fragment (red). The two PCR primers are located 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.

BOX 1: 3C-SEQ PRIMER DESIGN

Two primers, a P5-primer and 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 (Figure 3).This primer contains the P5 Illumina adapter sequence (5'-AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' to be placed upstream of the annealing sequence, see Figure 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', Figure 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 100bp of the restriction site).

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

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 co-associations (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.

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 (Roche) to minimize any biases resulting from these differences.[11] Bait-specific primers (without adapters) are first tested for proper linearity and efficiency.

1. Test 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 2 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 due 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 (see 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)

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).

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 (Figure 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 (Figure 4b).

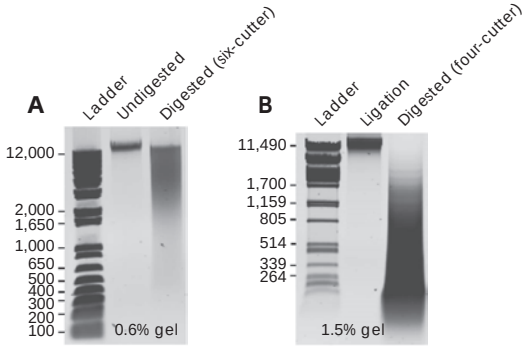


Figure 4. Examples of successful digestion and ligation efficiencies. (A) Agarose gel (0.6%, wt/vol) on which an aliquot of undigested (left lane) and digested (right lane) sample (primary restriction digestion, Step 16) was run. A six-cutter was used, showing a typical smear of DNA fragments (a majority 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).

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 (Figure 4b).

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.

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
Forward primer (25 pmol/µl)	1 µl	25 pmol
Reverse primer (25 pmol/µl)	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.

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 (Figure 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
    ```

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

Multiplexed 3C-seq success primarily depends on digestion efficiencies, 3C-seq PCR setup (Boxes 1 and 2) and Illumina sequencing. Table 3 contains 3C-seq troubleshooting advice, mainly concerning these steps. Digestion efficiencies are also highly dependent on the cell or tissue type used. Table 2 provides additional cell type–specific troubleshooting information. Other published protocols have also provided detailed troubleshooting for the 3C procedure[19,30].

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 Tm is used. Further optimizing the Tm 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

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]

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: 5–6h
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 Figure 5). After uploading the BED output file (Step 84) in a genome browser, interactions with the chosen bait fragments can be observed. Signals are represented as bars (Figure 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 (Figure 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 picture18. 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].

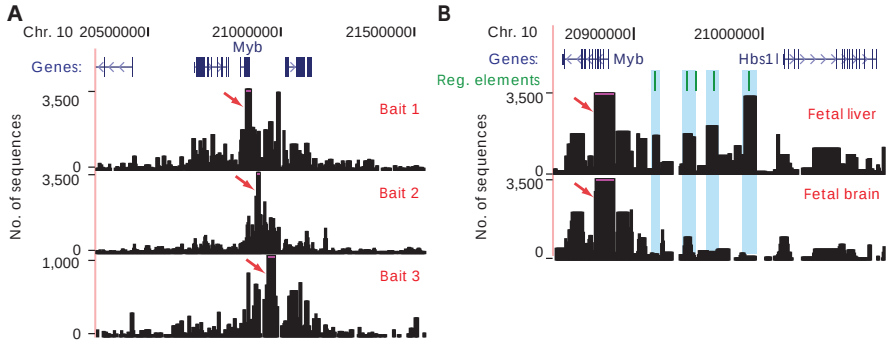


Figure 5. Typical interaction profiles obtained from a multiplexed 3C-seq experiment. (A) 3C-seq interaction profiles in mouse fetal liver cells shown for three bait fragments in the Myb locus[17] (1.2-Mb region shown). Bait signals are depicted by an arrow. **(B)** 3C-seq interaction profiles generated from both mouse fetal liver and brain using the Myb promoter as bait (shown is an ~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.

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 non-multiplexed sample, interaction patterns remain almost identical (Figure 6). Thus, multiplexing 3C samples seems to have little effect on the resulting interaction profiles (Figure 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 (Figure 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.

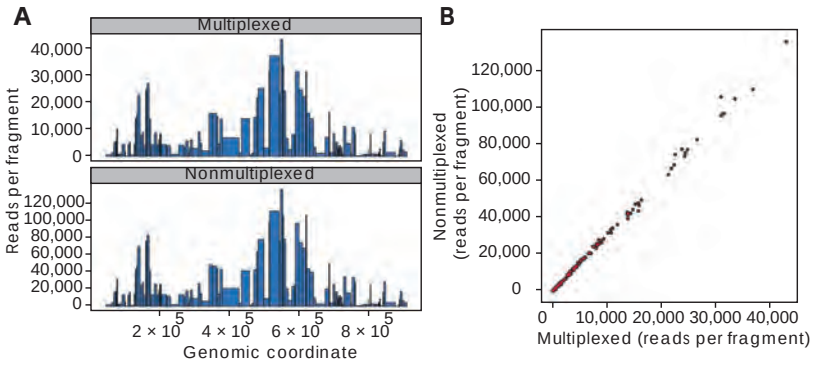


Figure 6. Comparison of interactions detected for the same 3C-seq sample after single or multiplexed library sequencing. (A) Interaction profiles around the bait fragment for a 3C-seq sample after multiplexed (top) or nonmultiplexed (bottom) library sequencing, showing highly similar profiles. **(B)**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

Acknowledgements

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-a-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).

34. Lin, Y.C. et al. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat. Immunol. 13, 1196–1204 (2012).

Summary

Samenvatting

List of Abbreviations

Summary

The development and survival of multicellular organisms relies on the continuous cellular proliferation and differentiation of various tissue-specific stem cells. As a cell's unique identity is primarily determined by its unique gene expression profiles, correct cell fate decision requires the strict regulation of gene expression dynamics. In this thesis I study the transcriptional control during different steps of hematopoietic development. We humans rely on hematopoietic tissue for various important functions in the body (e.g. oxygen/carbon dioxide transport, immune response) and hence correct hematopoiesis is essential for survival. Various defects in hematopoietic development are associated with the development of disease. The first step in the development of a treatment for such diseases will be to unravel the processes of normal hematopoiesis, as only then the effect of defects in these processes can be fully appreciated.

In **Chapter 2**, we study the development of the hematopoietic stem cells (HSCs). HSCs are known to originate during embryonic development in the dorsal aorta of the embryo, where they derive from so-called intra-aortic hematopoietic clusters (IAHCs). However, the cellular identity of the IAHCs and the mechanisms involved in HSC specification in these clusters are largely unknown. In **Chapter 2** we study the (molecular) identity of the IAHC cells in mice. We find that IAHCs are mainly composed of premature HSCs, which, when transplanted into newborn recipient mice, can mature into true HSC with long-term multilineage reconstitution potential. Using transcriptome analyses, we identify 1,239 genes, of which 108 genes encoding for transcription factors, being differentially expressed during a one-day time window covering the time point of the first appearance of (pre-)HSCs in the IAHCs. These transcriptome analyses provide the first step in the identification of potential novel regulators of HSC development.

The structural conformation of chromatin plays a central role in gene transcriptional control. Gene transcription is usually regulated by promoter-distal regulatory elements which are brought in close spatial proximity to their target genes via the formation of long-range chromatin loops. In **Chapter 3** we discuss the role of distal regulatory elements and chromatin structure in the regulation of transcription. We also describe the potential effect of genomic alterations on chromatin conformation, how this could affect transcription regulation, and how this contributes to phenotypic variation and diseases susceptibility and etiology. In addition we describe several novel potential therapeutic strategies to target such molecular diseases.

In order to fully unravel the role of distal regulatory elements in the transcriptional control of their target genes, it will be essential to incorporate information on chromatin conformation and its dynamics. In **Chapter 4** we describe the development of the multiplexed 3C-sequencing technology, a variant of the chromatin conformation capture (3C) technology optimized for use in semi-high throughput multiplexed Illumina next-generation sequencing. Multiplexed 3C-sequencing, provides a tool for analyzing fine-scale chromatin structural conformation and can be used to study the role of chromatin structure in gene-specific transcription regulation.

In **Chapter 5** we study the chromatin conformation dynamics in the murine *Bcl11a* locus during erythroid differentiation using this technology. BCL11a is an important transcriptional repressor of *γ-globin,* the β-like globin gene being expressed during fetal stages of life, but repressed after birth. As elevated *γ-globin* levels ameliorate the disease phenotypes of β-hemoglobinopathies like sickle cell disease and

β-thalassemia, repression of *Bcl11a* has become an interesting therapeutic strategy for these diseases.

In the study presented in **Chapter 5**, we identify an erythroid stage-specific enhancer in the second intron of *Bcl11a*, which is essential for early-erythroid-specific *Bcl11a* expression levels. We show that the activity of this enhancer correlates with the binding of the erythroid-specific LDB1 transcription factor complex and a change in chromatin conformation in the *Bcl11a* locus. Chromatin conformation data show that the enhancer and promoter are not involved in the classically described enhancer-promoter interaction, but rather share a primary interaction site ~5kb downstream of the transcription start site, inside the BCL11a gene. We provide evidence to suggest that this interaction site co-localizes with an RNA-polymerase II pause site and that the *Bcl11a* enhancer is involved in the regulation of transcription elongation.

Data from this study provides important insights in the transcriptional control of *Bcl11a* by the LDB1 transcription factor complex and provides important information on potential novel targets for *γ-globin* reactivation as therapeutic strategy for β-hemoglobinopathies.

Chapter 6 presents a general discussion on the work presented in this thesis and highlights novel insights in the diverse mechanisms involved in the transcriptional control of hematopoiesis, which may contribute to the development of new therapeutic strategies for the β-globin related hemoglobinopathies.

Samenvatting

De ontwikkeling en het voortbestaan van multicellulaire organismen is afhankelijk van de continue proliferatie en differentiatie van een groot aantal verschillende weefselspecifieke stamcellen. Aangezien de unieke identiteit van iedere cel hoofdzakelijk wordt bepaald door zijn unieke gen expressie profiel, is een streng gecontroleerde regulatie van gen expressie noodzakelijk. In dit proefschrift, heb ik de regulatie van de eerste stap van gen expressie, transcriptie, bestudeerd tijdens verschillende stadia van de hematopoietische ontwikkeling. Wij mensen zijn afhankelijk van hematopoiese voor belangrijke processen in het lichaam (waaronder zuurstof/koolstofdioxide transport en immuun respons). Correcte hematopietische ontwikkeling is daarom essentieel voor het leven van zoogdieren en de mens. Een aantal defecten in hematopoiese zijn gecorreleerd aan de ontwikkeling van ziekten. De eerste stap in het ontwikkelen van een behandelstrategie voor deze ziekten is het ontrafelen van de verschillende processen betrokken bij de regulatie van normale hematopoiese. Alleen dan kan de ontwikkeling van deze ziektes volledig worden begrepen.

In **Hoofdstuk 2** bestuderen wij de ontwikkeling van hematopoietische stam cellen (HSCs). Het is bekend dat HSCs ontstaan tijdens de embryonale ontwikkeling in de dorsale aorta van de embryo als intra-aorta hematopoietische cel clusters (IAHCs). Echter, de cellulaire identiteit van de IAHCs cellen, of het mechanisme betrokken bij de ontwikkeling van HSCs in deze IAHCs is nog steeds grotendeels onduidelijk. In **Hoofdstuk 2** bestuderen we de (moleculaire) identiteit van deze IAHCs cellen in muizen. Wij tonen aan dat IAHCs vooral bestaan uit 'ongerijpte' HSCs, welke kunnen uitgroeien tot volwaardige HSCs met multi-differentiatie en reconstitutie potentieel wanneer ze getransplanteerd worden in pasgeboren muizen. Met behulp van transcriptoom analyse identificeren we vervolgens 1.239 genen, waaronder 108 genen coderend voor transcriptie factoren, die significant variërend worden geëxpresseerd tijdens een 24-uurs tijdsinterval rond het tijd punt van de eerste verschijning van (pre-) HSCs in the IAHCs. Deze transcriptoom analyse levert de eerste initiële informatie voor de identificatie van potentiële nieuwe regulatoren van de ontwikkeling van HSCs.

De structurele conformatie van chromatine speelt een belangrijke rol in gen transcriptie regulatie. Gen transcriptie wordt vaak gereguleerd door promoter-distale regulatoire elementen, welke dicht bij hun doelwitgenen worden gebracht doormiddel van het vormen van een chromatine lus. In **Hoofdstuk 3** bespreken we de rol van deze distale regulatoire elementen en de structuur van chromatine in transcriptie regulatie. Daarnaast bespreken we de potentiële effecten van veranderingen in het genoom op de structurele conformatie van chromatine, de mogelijke invloed hiervan op transcriptie regulatie, en hoe dit kan bijdragen aan de ontwikkeling van fenotypische variatie en de ontwikkeling van en vatbaarheid voor ziekte. Als laatste, bespreken we verschillende nieuwe potentiële therapeutische strategieën voor de behandeling van deze moleculaire ziekten.

Om de rol van distale regulatoire elementen in de transcriptie regulatie van hun doelwitgenen compleet te kunnen bepalen, is het essentieel om informatie over de chromatine conformatie (en de dynamiek hierin) mee te nemen in de analyse. In **Hoofdstuk 4** beschrijven we de ontwikkeling we de *multiplexed 3C-sequencing* technologie. Dit is een variant van de *chromatin conformation capture* (3C) technologie, waarbij deze is geoptimaliseerd voor gemultiplexeerde Illumina *next-generation sequencing*. *Multiplexed 3C-sequencing* biedt een methode voor de fijnschalige analyse van chromatine conformatie en kan worden gebruikt om de rol van chromatine structuur in een gen-specificieke expressie regulatie te bestuderen.

Met behulp van deze methode, bestuderen we in **Hoofdstuk 5** de dynamiek van de chromatine structuur van het *Bcl11a* gen tijdens de differentiatie van rode bloedcellen in muizen. BCL11a is een repressor van het γ-*globine* gen, het β-type globine gen dat wordt geëxpresseerd in de foetus, maar na de geboorte wordt geïnactiveerd. Aangezien een verhoogde γ-*globine* eiwit concentratie een positief effect heeft op het ziektebeeld van β-hemoglobinopathieën zoals sikkelcelziekte en β-thalassemie, is het inactiveren van het *Bcl11a* gen een interessante therapeutische strategie geworden voor deze ziekten.

In deze studie identificeren we een erytroïde stadium-specifieke enhancer in het tweede intron van *Bcl11a*, welke essentieel is voor het specifieke expressie patroon van vroege erytroïde voorloper cellen. We laten zien dat de activiteit van deze enhancer correleert met de binding van het erytroïde-specifieke LDB1 transcriptie factor complex en een verandering in de chromatine structuur van de *Bcl11a* locus. Daarnaast toont onze data aan dat de *Bcl11a* enhancer en promoter geen klassieke enhancer-promoter interactie aangaan, maar samen een primaire interactie plaats delen op ~5kb na de transcriptie start positie in het *Bcl11a* gen. Verder suggereren de data dat deze primaire interactie plaats co-lokaliseert met een pauze plaats van RNA polymerase II en dat de *Bcl11a* enhancer betrokken is bij de regulatie van transcriptie elongatie.

Data van deze studie levert belangrijke inzichten in de transcriptionele regulatie van *Bcl11a* en de rol van het LDB1 transcriptie factor complex in deze regulatie. Daarbij, levert deze studie belangrijke informatie voor de ontwikkeling van nieuwe potentiële therapeutische strategieën for β-hemoglobinopathieën.

Samengevat biedt het werk beschreven in dit proefschrift nieuwe informatie over de diverse mechanismen betrokken bij de transcriptionele regulatie van hematopoiese. Verdere ontwikkeling van onze kennis over de transcriptionele regulatie van cellulaire differentiatie is belangrijk voor klinische doeleinden en kunnen mogelijk bijdragen aan de ontwikkeling van nieuwe therapeutische strategieën voor verschillende hematologische ziekten.

List of abbreviations

+58kb enhancer	The enhancer in the Bcl11a gene locus, positioned 58kb downstream of the *Bcl11a* TSS
(pre-)HSC	(Premature) Hematopoietic stem cell
(q)PCR	(Quantitative) Polymerase chain reaction
3C(-seq)	Chromosome conformation capture (coupled to next-generation sequencing)
5'-interacting site ('docking site')	The region at the 5'end of Bcl11a that forms a long-range chromatin interaction with both the +58kb enhancer and the promoter
5'UTR	5' Untranslated region
Ab	Antibody
AGM	Aorta-Gonad-Mesonephros region in the embryo
Bcl11a	B cell CLL/lymphoma 11A
BFU	Burst forming unit
bp	Base pair
CDK9	Cyclin-dependent kinase 9
CFU	colony forming unit
ChIP	Chromatin immunoprecipitation
DNA	Deoxyribonucleic acid
DNAseHS	DNAse I hypersensitivity
DSIF	DRB sensitivity inducing factor complex
E10-E11.5	Embryonic day 10-11.5
ESC	Embryonic stem cell
FACS	Fluorescence-activated cell sorting
GATA1	GATA binding protein 1
GO-term analysis	Gene ontology analysis
H3K27Ac	Histone H3 Lysine 27 acetylation
H3K36me3	Histone H3 Lysine 36 trimethylation
H3K4me1	Histone H3 Lysine 4 monomethylation
IAHC	Intra-aortic hematopoietic cluster
kb	Kilobase (1000bp)
KLF1	Kruppel-like factor 1 (erythroid)
LCR	Locus control region
LDB1	LIM-domain binding protein 1
LMO2	LIM domain only 2
MEL (cells)	Mouse erythroid leukemia (cells)
NELF	Negative elongation factor complex
PIC	Pre-initiation complex
Pol II	RNA polymerase II
PolIISer2P ('elongation complex')	RNA polymerase II phosphorylated at Serine 2 of the repeat motif in the C-terminal domain of its largest subunit
PolIISer5P ('initiation complex')	RNA polymerase II phosphorylated at Serine 5 of the repeat motif in the C-terminal domain of its largest subunit
p-TEFb	Positive transcription elongation factor complex b
PTM	Post-translational modification
RLIM	LIM-domain binding ring finger protein
RNA	Ribonucleic acid
RNA pol II CTD	RNA polymerase II C-terminal domain
SCD	Sickle cell disease
SSBPs	Single-stranded DNA binding proteins
TAL1	T-cell acute lymphocytic leukemia 1
TF	Transcription factor
TIF1γ	Transcriptional intermediary factor 1 gamma
TSS	Transcription start site
WT	wild type
ZEB	Zinc finger E-box binding homeobox 1

Curriculum Vitae

PhD Portfolio

Curriculum Vitae

Personal details

Name : Anita van den Heuvel
Date of Birth : 12 July 1987
Place of Birth : Rotterdam, The Netherlands

Education

2010 – 2015 : **PhD student**, Department of Cell Biology, Promotor Prof. dr. Frank Grosveld, Erasmus Medical Center, Rotterdam, The Netherlands

2008 – 2010 : **Master**, Biomolecular Sciences, Specialization Molecular Cell Biology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

2005 – 2008 : **Bachelor of Applied Science**, Biology and Medical Laboratory Research, Hogeschool Rotterdam, Rotterdam

1999 – 2005 : **High School**, *VWO* level, Comenius College, Capelle aan den IJssel, The Netherlands

Research experience

2010 – 2015 : **PhD project**, Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands
Promotor: Prof. dr. F. Grosveld, Copromotor: Dr. E. Soler
Thesis title: *Transcriptional Control During Hematopoietic Development: Transcription factor binding and chromatin conformation dynamics*

2010 : **Master internship**, Department of Cell Biology, Erasmus Medical Center, Head of Department Prof. dr. Frank Grosveld,
Supervisors: Dr. E. Soler and Prof. dr. Frank Grosveld
Topic: *Transcription regulation of the Bcl11a gene*

2009 : **Master internship**, Department of Biomedical Genomics, Hubrecht Institute, Head of Department Prof. dr. Wouter de Laat,
Supervisors: Ing. E. Splinter and Prof. dr. W. de Laat
Topic: *The role SmcHD1 and Histone MacroH2A in the structural organization of the inactive X-chromosome*

2008 : **Bachelor internship**, Department of Endocrinology and Metabolism, Utrecht University, Head of Department Prof. dr. Dick van der Horst
Supervisors: Dr. H. de Jonge and Dr. J Bogerd
Topic: *Structure determination of the human FSHR protein*
(Received the HAS thesis prize for the best thesis of the study Biology and Medical Laboratory Research 2008 of the Hogeschool Rotterdam, awarded by the Laboratory education foundation of The Hague)

2007 : **Bachelor internship**, Department of Internal Medicine, Erasmus Medical Center, Head of Department Prof. dr. Andre Uitterlinden
Supervisors Drs. S. Stolk and Dr. J. Meurs
Topic: *Genetic variation in the estrogen pathway*

List of Publications

Anita van den Heuvel, Petros Kolovos, Ralph Stadhouders, Supat Thongjuea, Rutger Brouwer, Wilfred van IJcken, Frank Grosveld, & Eric Soler. Dynamic long-range chromatin interactions control *Bcl11a* transcription during erythroid differentiation in mice. Manuscript in preparation.

Anita van den Heuvel, Ralph Stadhouders, Charlotte Andrieu-Soler, Frank Grosveld & Eric Soler. Long-range gene regulation and novel therapeutic applications. Accepted at *Blood* as *Blood Spotlight Review*.

Aissa BenYoucef, Julien Calvo, Laurent Renou, Marie-Laure Arcangeli, **Anita van den Heuvel**, Sophie Amsellem, Maryam Mehrpour, Jérome Larghero, Eric Soler, Irina Naguibneva, Francoise Pflumio. The SCL/TAL1 transcription factor represses the stress protein DDiT4/REDD1 in human hematopoietic stem/progenitor cells. *Submitted to Stem Cells,* under revision.

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. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protoc*. 2013;8(3):509-24.

Stadhouders R, **van den Heuvel A**, Kolovos P, Jorna R, Leslie K, Grosveld F, Soler E. Transcription regulation by distal enhancers: who's in the loop? *Transcription*. 2012;3(4):181-6. Review.

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. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J*. 2012;31(4):986-99.

PhD portfolio

Name PhD student	: Anita van den Heuvel
Erasmus Medical Center Department	: Cell Biology
Research School	: Medisch Genetisch Centrum
PhD Period	: October 2010 – February 2015
Promotor	: Prof. dr. Frank Grosveld
Copromotor	: Dr. Eric Soler

PhD Training	Year
General Courses	
Photoshop and Illustrator CS6 Workshop	2014
Indesign CS6 Workshop	2014
English Biomedical Writing and Communication	2012
Safely working in a laboratory	2011
In-depth Courses	
Biostatistical Methods I: Basic Principles Part A	2013
Literature Course	2012
Cell and Developmental Biology	2012
Next Generation Sequencing (NGS) Data Analysis Course	2012
Genetics Course	2011
Biochemistry and Biophysics Course	2010
International Conferences	
EUTRACC Transcription Networks Symposium, Berlin, Germany	2011
The Operon Model Conference, Institut Pasteur, Paris, France	2011
EUTRACC 2nd Young Scientist Meeting, Dubrovnik, Croatia	2010
Meetings, Workshops and Symposia	
Weekly Monday Morning Meetings, Cell Biology Department, Erasmus Medical Center (Oral Presentations)	2010-2014
24th Annual MGC Symposium, Rotterdam, The Netherlands	2014
21st Annual MGC PhD student Workshop, Münster, Germany (Oral Presentation)	2014
20th Annual MGC PhD student Workshop, Luxembourg city, Luxembourg (Oral Presentation)	2013
23rd Annual MGC Symposium, Rotterdam, The Netherlands	2013
22nd Annual MGC Symposium, Leiden, The Netherlands	2012
21st Annual MGC Symposium, Leiden, The Netherlands	2011
18th Annual MGC PhD student Workshop, Maastricht, The Netherlands	2011
Additional activities	
HLO (Bachelor of Applied Sciences) student supervision	2013-2014
Organizer of the 20th MGC PhD student Workshop, Luxembourg city	2013

Dankwoord

Acknowledgements

Hier ligt hij dan, mijn proefschrift, het eindproduct van vier jaar werk als 'onderzoeker-in-opleiding'. Het feit dat u dit proefschrift kunt lezen betekent dat ik deze opleiding inmiddels met goed gevolg heb afgerond of in zeer korte tijd hoop af te ronden. Hoewel het de keus is van één persoon om een PhD-traject te starten, is het doorlopen van een PhD-traject zeker geen eenmanswerk, maar een gezamenlijk project van vele verschillende mensen. Zonder de hulp en support van vele collega's, vrienden en familieleden had dit proefschrift nu niet voor u gelegen. Ik wil deze mensen dan ook heel erg bedanken voor alle hulp.

Allereerst wil ik uiteraard mijn promotor Frank Grosveld en mijn copromotor Eric Soler bedanken. Frank, hoewel ik al een aantal jaren zeker wist dat dit was wat ik wilde doen, kan ik me nog goed herinneren dat ik bij het starten van mijn PhD-traject lang niet zeker wist óf en zo ja hóé ik dit ooit tot een goed eind zou kunnen brengen. Bedankt dat je na mijn masterstage genoeg vertrouwen in me hebt gehad om mij te laten starten aan een PhD-traject op de afdeling Celbiologie. Tijdens de loop van dit PhD-traject hebben je ontspannen manier van werken en je zeer open-minded aanpak van onderzoeksproblemen me enorm geïnspireerd en er voor gezorgd dat mijn stress niveau (welke bij mij vrij makkelijk op loopt) niet de pan uit rees. Ik ben dan ook enorm blij dat ik de mogelijkheid heb gehad om mijn PhD-project binnen jouw groep te kunnen uitvoeren.

En wel onder de begeleiding van Eric Soler!! Eric, I want to thank you for agreeing to supervise me during my PhD-project, even though you knew you were moving back to France within a few years. I know, and everyone else who ever received one of my chaotic and endless emails knows, that supervising someone over such long distance mostly via email and skype contact is not an easy task. I started this PhD project as a very naive student who did not know much about the big wide world of science. Thanks for all the time you spent on introducing me into this world and in showing me how to become a true scientist. Also many thanks for the reassuring words whenever I had one of my little 'panic attacks', questioning whether I would ever be able to become such a true scientist. I wish you, Charlotte and your three kids all the best in France both in science and outside.

Ik wil daarnaast de leden van de kleine commissie (a.k.a de leescommissie) bedanken. Danny, Sjaak en Raymond, ik weet dat jullie mijn proefschrift hebben ontvangen in misschien wel de drukste periode van het jaar (niet alleen buiten de wetenschap waren er een aantal drukke dagen, maar ik heb begrepen dat er ook binnen de wetenschap een aantal deadlines rond deze periode lagen). Ik wil jullie dan ook bedanken voor alle tijd die jullie hebben gestoken in het lezen en corrigeren van mijn proefschrift en voor alle nuttige commentaren die ik van jullie heb ontvangen.

Ook wil ik Catherine, Marieke en Niels bedanken dat ze deel wilden uitmaken van mijn grote commissie tijdens mijn verdediging. Dat waardeer ik zeer.

Dan komen nu twee andere belangrijke mensen, mijn paranimfen. Guillaume and Andrea, many thanks for accepting to be my paranymphs, I really appreciate it. You two have helped me a lot during the years that we worked together. Guillaume, I enjoyed the many nice discussions we have had about one of our projects or just about some interesting new paper that just came out. I especially enjoyed the many brainstorm sessions we had at the end of a day, when all experiments were finished and we could just relax and discuss about the many questions that came up during our research. This has led to some very interesting brainstorm sessions. Andrea, apart from being a very good scientist who helped me with a lot of questions even though you were not into the field of long-range chromatin interactions during hematopoietic development at all, I think you are perfect when it comes to creating a perfect work

atmosphere. I enjoyed your dinners and your many ideas to organize some group activities (you even arranged a restaurant to open just for us to have lunch! I am impressed!!). I wish you both all the luck with the 702 group in the Erasmus MC. With the impressive number of students that have been working under your supervision in the last year, and the even more impressive speed at which this group seems to expand, I believe that you guys are building a very nice lab in 702.

Natuurlijk zijn er nog vele anderen die ik wil bedanken. Allereerst, de overige mensen van het Grosveld lab. Petros, you are the last of the PhD-students that started in Frank's group during my time in this department. Thanks for all your help with the bioinformatics in my projects. Good luck with finishing your many projects and let me know when it is time to congratulate you as Dr. Kolovos. Ralph, jij loopt juist net iets op me voor. Een aantal van onze projecten hadden veel gemeen en Ik heb veel gehad aan jou ervaring met een aantal dingen. Bedankt voor al je hulp en heel veel succes in Barcelona. Laat me weten wanneer ik jou moet feliciteren met de Nobel prijs. Robert-Jan, voor een lange tijd mijn 'bureau-buurman'(;-)). Niet alleen als 3C-seq expert heb je me heel veel geholpen, ook daarbuiten was je vaak mijn vraagbaak als ik tegen een probleem in mijn projecten aan liep. Dank daarvoor. Hoewel ik het na drie jaar erg fijn vond om met mijn computer te verhuizen naar 'het kantoortje', vond ik het jammer om jou als mijn bureau-buurman te moeten missen. Inmiddels ben je verhuisd naar de 6e verdieping. Heel veel succes in de HIV wereld en we houden contact.

Charlotte, Andrea M., Xiao, Mary, Ruud, Ali en Farzin, you all no longer work at the department, but I want to thank you for the nice time we had when working together and for all the help you gave me. I always enjoyed working with you and I wish you all good luck with your new job. You all now work at diverse places both in the Netherlands and Europe, so who knows, I might meet one of you again when I find a new position. De vele studenten die inmiddels in ons lab hebben gelopen, bedankt voor al jullie gezelligheid en succes met jullie studie of inmiddels jullie PhD-project of nieuwe baan. Ilknur (D.), bedankt dat je het aandurfde om mijn eerste student te zijn en te werken aan het inmiddels beruchte 'lowChIP-project'. Succes met je afstudeerstage. Komt helemaal goed, ik weet het zeker.

Dan de vele collega's buiten de Grosveld groep…om te beginnen de mensen met wie ik aan de verschillende projecten heb gewerkt. Catherine, Chloé, Jean-Charles and Thomas, 'The French gang' who are now at the Hubrecht Institute. Catherine, I am very happy that you asked me to collaborate with you in the research on the HSC development in the IAHCs in the embryonic aorta. Chloé, I enjoyed working with you on the molecular analysis of the IAHC cells. I was pleased to see that unlike the infamous lowChIP-seq, the lowRNA-seq protocol worked like a charm and we got some really nice data out. I know there are still various questions remaining and I hope we can somehow continue to work on these things together.

Supat and all other people from the group that was generally known to us as 'the Norway group'. Thanks for your help with bioinformatics and the analysis of a large number of the ChIP-seq, 3C-seq and RNA-seq experiments I did.

Kerstin and Jessica, I think we can be very happy with the Nature Protocols paper we published together.

Alle mensen van Biomics, met extra dank voor Zeliha, Christel, Rutger, Mirjam en Wilfred. Vooral ook bedankt voor het (soms vanwege tijdsdruk 'verplicht' snelle) optimaliseren van de sequence procedure als ik weer eens met 'vreemde samples' of een nieuw protocol kwam aanzetten. Ik heb het jullie niet altijd makkelijk gemaakt

met mijn 'low-cell-number' protocollen (en het uitspreken van de namen lowChIP en LinDA zorgde soms ook bij jullie voor een licht wanhopige glimlach op jullie gezicht).

Na al deze mensen te hebben bedankt ben ik er nog lang niet. Er nog zoveel meer mensen waarmee ik niet zozeer heb samen gewerkt aan een project, maar die zeker een belangrijke bijdrage hebben geleverd aan het slagen van mijn PhD-project. Ernie, je hebt me meerdere malen geholpen door een experiment aan de praat te krijgen als het bij ons in het lab om mysterieuze wijze niet lukte. Mike (en vroeger Erik) de enzym-mannen, jullie zorgden altijd zonder enig (merkbaar) klagen, waar mogelijk nog dezelfde dag voor het gevraagde enzym. Michael (V.), je hebt me regelmatig voorzien van een suiker boost voor een potentiële lange avond en je hebt regelmatig mijn frustraties aangehoord en me voorzien van bemoedigende woorden na het krijgen van negatieve resultaten voor het 'lowChIP-project'. Alle collega's van 'mijn' gang; de mensen van 706 (Ernie, Mike, Johan, Ali, Maaike, Marti), de 'de high-throughput productiegroep' van 710 (Alex, Dubi, Michael (vd R.), Rick en Rien) en alle mensen van 716, (die later zijn verhuisd naar 1002) (Maria, Ileana, Silvia, Nynke en Thamar) bedankt voor de fijne werksfeer op onze afdeling. En natuurlijk wil ik ook alle mensen met een meer administratieve organisatorische rol bedanken; de mensen van de inkoop en bestellingen, Melle, Leo, Koos en Annet; de mensen van de ICT, beter bekend als de 'computer jongens' en zeker niet te vergeten de secretaresses, Marike, Bep en Jasperina.

Er zijn nog veel meer mensen die ik in mijn tijd op de afdeling Celbiologie van het Erasmus MC heb leren kennen. Helaas kan ik deze mensen niet allemaal één voor één opnoemen en bedanken, aangezien dit dankwoord dan oneindig lang lijkt te worden, maar iedereen van die groep mensen heeft op een of andere manier bijgedragen aan ik leuke tijd die ik heb gehad bij deze afdeling. De vier jaren van mijn PhD-project waren vier super leuke jaren waarin ik enorm veel heb geleerd. Iedereen die mij langer kent dan vier jaar zal kunnen bevestigen dat dit PhD-traject mij enorm heeft veranderd (en hopelijk in de goede zin). Bedankt daarvoor!

Dan is er nog een hele belangrijke groep mensen die ik moet bedanken, mijn familie. Iedereen die een promotie traject doorloopt weet dat dit niet kan zonder de support van familie. Pa, Ma, Diana en Linda, bedankt voor al jullie support tijdens deze vier jaar. Diana, wij hebben ons hele leven alles samen gedaan en dat is met dit PhD-project niet anders geweest. Hoewel jij niets weet over transcriptie regulatie en chromatine vouwing en ik niets over de CD4 positieve-CD8 positieve B cellen (of nee wacht waren het nou CD4 positieve CD8 negatieve T cellen) zit ik dagelijks met je op de gmail-chat en heb ik menig uur met je aan de telefoon gezeten om resultaten van experimenten met je door te spreken. Hoewel we hebben besloten om de verdediging 'eens niet samen' te doen, ben je eigenlijk gewoon mijn derde paranimf. Bedankt voor al je hulp. Voor jou komt het einde van je PhD-traject ook in zicht en ik hoop dat we de rollen nu kunnen omdraaien en ik jou nu kan helpen met afronden.

Pa, Ma en Linda, jullie zijn vaak genoeg het slachtoffer geworden van een van de discussies / brainstormsessies van Diana en mij, waarbij we 'maar niet ophielden over werk', sorry daarvoor ;-D. Bedankt ook voor jullie support. Vooral het laatste jaar heb ik regelmatig dingen afgezegd omdat ik weer eens wat wilde afmaken op mijn werk. Dit was niet altijd leuk, maar jullie hebben me altijd gesteund tijdens dit PhD-traject. Hoewel jullie me misschien niet zozeer met de inhoud van het PhD-project hebben kunnen helpen, hebben jullie me enorm geholpen met menig andere zaken tijdens alle jaren van mijn promotie-traject. Linda, ik vond het geweldig toen ik in mijn werk met een techniek genaamd LinDA begon te werken. Deze techniek leek voor mij te zijn gemaakt aangezien nu het plaatje compleet was en alle drie de zussen een

soort van betrokken waren in mijn werk. Hoewel, LinDA (de techniek) me lichtelijk in de steek heeft gelaten ;-), heb jij, Linda, dat zeker niet gedaan. Bedankt voor je altijd volle vertrouwen dat het me zou lukken om dit traject met goed resultaat af te ronden, ook als ik er zelf niet zo zeker van was.

Pa en Ma, ik heb meer dan eens een beroep gedaan op jullie tijd, voor dingen variërend van boodschappen halen tot in mijn huis aanwezig zijn als er bijvoorbeeld een loodgieter langs zou komen die dag en ik "eigenlijk echt een experiment gepland had". En ook zonder dat ik het vroeg hebben jullie me menig keer geholpen. Zo werd er vaak onverwachts, soms laat in de avond, nog wat te eten voor me klaar gemaakt als ik vanuit mijn werk nog even bij jullie langs kwam en jullie er achter kwamen dat ik nog niet had gegeten. Of ik kwam thuis en er bleken ineens dingen te zijn geregeld, opgeruimd, afgemaakt of langsgebracht (jullie hebben zelfs mijn huis in kerst sferen gebracht deze kerst!!) Ik kan jullie niet genoeg bedanken voor alles wat jullie voor me doen! Zonder jullie hulp was dit allemaal niet mogelijk geweest.

Bedankt iedereen, Thank you all!

Anita