Developmental Dynamics of Transcription and Genome Architecture

Petros Kolovos

ISBN/EAN: 978-94-6299-172-9

Cover illustration: Petros Kolovos

Cover illustation represents the skyline of Rotterdam; adopted from http://vervoersplanoloog.blogspot.nl/

Printed by: Ridderprint BV

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

Developmental Dynamics of Transcription and Genome Architecture

De dynamiek van transcriptie en genoom architectuur tijdens de ontwikkeling

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board

The public defence shall be held on
tuesday 22 September 2015 at 15.30 hrs

by

Petros Kolovos

born in Thessaloniki, Greece

Erasmus University Rotterdam

Doctoral Committee

Promotor: Prof.dr. F.G. Grosveld

Overige leden: Prof.dr. D.F.E. Huylebroeck

Prof.dr. J. Gribnau

Dr. A. Papantonis

Table of Contents

	Scope of this thesis	7
Chapter 1	Introduction	9
Chapter 2	Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions	43
Chapter 3	Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements.	65
Chapter 4	TNFα signalling primes chromatin for NF-κB binding and induces rapid and widespread nucleosome repositioning	85
Chapter 5	The bimodal function of NFκB and the effect of TNFα in the spatiotemporal landscape of responsive and non-responsive genes	101
Chapter 6	Dynamics of the LBD1 complex and the activation of hematopoietic development	121
Chapter 7	Dynamics of essential transcription factors during erythroid differentiation	145
Chapter 8	General Discussion	175
	Summary	188
	Samenvatting	190
	Curriculum Vitae	192
	PhD Portfolio	194
	Acknowledgements	195

Scope of this thesis

Regulation of gene expression is necessary for the control of complex developmental processes. In order to unravel gene regulation, it is necessary to understand the chromatin structure and organization. Furthermore, developmental procedures are controlled by complex combinatorial transcription factor (TF) networks. Hence, unveiling those networks will provide a better insight towards understanding those developmental procedures. The work described in this thesis aims to contribute to decode the genome structure and organization and understand the complex mechanisms controlling developmental pathways.

Chapter 1 provides an introduction to the basic concepts of the DNA, gene regulation and the role of TFs. Subsequently, it describes the aspects of chromatin compaction and its implications towards tethering together genes with their regulatory elements and presents the methods to unravel the 3D genome structure. Finally, it presents insights into the hematopoietic development and its association with combinatorial TF networks.

Chapters 2 to **7** contain the experimental work performed during this PhD study. **Chapter 2** contains the development of 3C-seq; a method to depict the close proximity of DNA segments in the 3D nuclear organization. **Chapter 3** describes the development of T2C; a method to unveil the chromatin interactome and the spatial organization of genome in high resolution with low sequencing costs. **Chapter 4** provides evidence about the rearrangement of nucleosomes along the DNA fiber upon TNFα signaling, to allow establishment of new transcriptional programs. In **Chapter 5** we extend this study to assess the effect of developmental signaling cues such as the TNFα singalling, in the dynamics of NFκB TF, the spatiotemporal chromatin architecture and the interactome of the genome. **Chapter 6** describes the role of the LDB1 complex in hematopoietic development and the dynamics of the "GATA" switch. **Chapter 7** describes a combinatorial TF network, its dynamics and properties in erythropoiesis.

In **Chapter 8** I summarize the results of the experimental research conducted in **Chapter 2** to **7**. Additionally, I highlight important findings and consider the effect of these results in our understanding of genome structure and organization and the complex mechanisms controlling developmental procedures. Finally I propose future perspectives for the continuation and evolvement of the current findings.

Chapter 1

Introduction

Parts of this chapter were published in
Epigenetics & Chromatin
2012; 9;5(1):1

Cell, the "brick" of life

DNA, the "computer" of life

It is estimated there are approximately 8.7 million eukaryotic species on Earth today, with everyone to be different from the other[1]. However, all of them share a common purpose in life; reproduction and inheritance of their genetic information into their offspring. All such species, from the simplest to the most complex ones, originate from single cells which are the vehicles of their hereditary information. The latter is stored in the DeoxyriboNucleic Acid, or most commonly known as DNA[2, 3].

DNA has been evolving the last 3.5 billion years and enabled to store all the necessary information to build an organism in a format consisting of long strings of nucleotides. Each nucleotide consists of sugar (deoxyribose) with a phosphate group attached to it and a base which may be adenine (A), cytosine (C), guanine (G) and thymine (T)[4], with A and G to be complementary with T and C, respectively[3] (Figure 1). DNA was first discovered by Friedrich Miescher in 1869[5] and the three dimensional (3D) structure of the helix was solved in 1953 by James Watson and Francis Crick[3] (a left-handed, anti-parallel double-helix structure) after the pioneering work by Rosalind Franklin[6] and Maurice Wilkins[7]. The bonds between the two DNA strands are weaker than the covalent bonds between the sugar-phosphates which allows the DNA strands to be pulled apart and serve as a template for DNA replication[8] and transcription[9, 10]. Like the notes of music which in a specific order produce a specific melody, nucleotides are arranged in a specific order making up the genes, the functional units of heredity[11]. Genes are long stretches of (usually short) DNA segments (exons) which carry the encoding information, interrupted in most genes by non-coding DNA (introns) (Figure 1). Genes encode a usually single-strand functional transcript containing ribose instead of deoxyribose[12] the Ribonucleic Acid (RNA)[13], where T is replaced by uracil (U) with A, G and C to remain the same.

One of the most important stepping stones of biology, proposed by Francis Crick in 1958[14] and reformulated in 1970[15], is *The Central Dogma* (Figure 1) which postulates that DNA encodes RNA which codes for proteins, and not the other way around. The dogma (Francis Crick later admitted the catch name of Central Dogma would better have been called Central Hypothesis) defines also that DNA can replicate itself to be maintained. When a cell needs a particular protein, the stretch of DNA that encodes that protein, the gene, is used as a template to synthesize RNA (a process called transcription). After removing the introns (splicing) the mature mRNA is used as a template for protein synthesis (a process outside of the cell's nucleus known as translation)[12, 14, 15]. However, there are many genes which

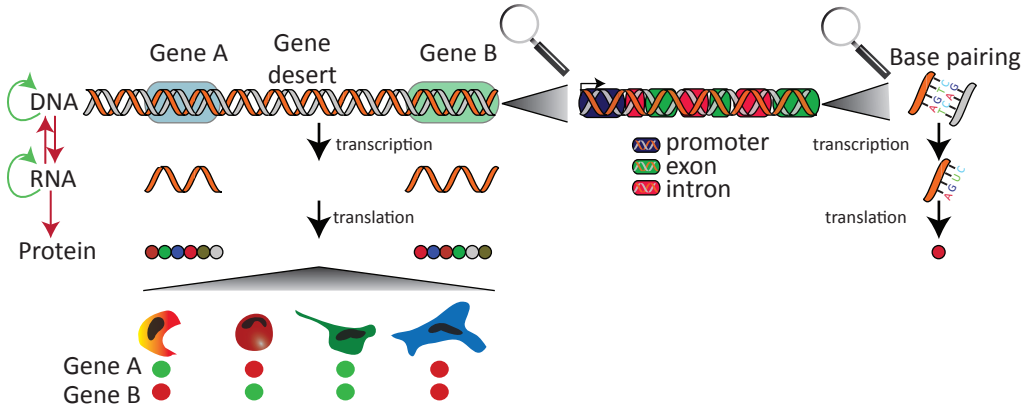

Figure 1: **The Central Dogma of Biology** as it was postulated by Francis Crick and updated by Howard Martin Temin and David Baltimore, illustrating the typical process of DNA (A and G to be complementary to T or C) to proteins. Regions of the genome encoding proteins termed genes (Gene A and B) contain promoter, exons and introns and transcribed into RNA which is subsequently translated into a protein. Activity (green circle) or silencing (red circle) of Genes A and B and their respective proteins, typify specific cell types, giving the transcriptome of each cell type its unique properties.

	Human	Mouse
Assembly	GRCh38.p2	GRCm38.p3
Base pairs (bp)	3.221×10^9	2.8×10^9
Coding genes	20300	22543
Non-coding genes	25159	12420

Table 1: General information about the size and number of coding vs non coding genes of the human and mouse genome.

encode non-coding RNA (ncRNAs), which are not translated into a protein[16]. Howard Martin Temin and David Baltimore determined in 1970 that DNA is transcribed into RNA and not *vice versa*. RNA can be converted to DNA via a process defined as reverse transcription[10, 17]. Furthermore, RNA can replicate itself, a phenomenon often seen in RNA viruses[18].

All the genes of an organism, with the exception of e.g. yeast and of mitochondrial and chloroplast DNA, form its genome which is divided into several long DNA molecules, the chromosomes. Joe Hin Tjio at 1955[19] correctly determined that humans (*Homo sapiens*) have 46 chromosomes (22 autosomes and 2 sex chromosomes, XX in females and XY in males) with two copies of each autosome (1 inherited from the father, 1 from the mother) and sex chromosomes, with the haploid genome totalling 3.2×10^9 base pairs (bp). Mouse (*Mus musculus*) has 40 chromosomes (19 autosomes, and XX in females, XY in males)[20] totalling 2.8×10^9 bp (Table 1). In 2004, the human genome was (almost) fully sequenced. It is still not completely finished as some regions are very repetitive in nature and extremely difficult to sequence properly. Nevertheless it revealed somewhat disappointingly that we have more than 20000 coding genes[21-23] (and not the previously anticipated 100000) with median size of approximately 3.5Kb, making up only 2.2% of the genome. The remainder (initially called "junk" DNA) contains regulatory regions responsible to e.g. activate or silence a gene as well as "gene deserts"; genomic regions with as yet no apparent functional role.

Comparing the human genome to the invertebrate *C.elegans* (a nematode worm, 9.7×10^7 bp containing about 20000 genes), an apparent conundrum concerning the smaller size of this genome and the number of protein-coding genes is revealed[24]. Apparently, the human genome is quite parsimonious (Table 1). A potential explanation may be that those "gene deserts" contain ncRNAs important for gene regulation[25], and it has also become clear that the spacing of regulatory sequences is important. Complex organisms such as human would then have more sophisticated and complicated ways of gene regulation[26], which would require more "DNA space". Interestingly, 94% of the human exons undergo alternative splicing (AS), thereby limiting the genome size, while increasing genome complexity. The relationship between protein-coding genes and genome size follows a logarithmic pattern for eukaryotes compared to non-eukaryotes, which follow a linear pattern[27]. Hence, we observe a positive correlation between the number of ncRNAs, AS events and complex regulatory networks with eukaryotic complexity[28].

Cellular heterogeneity: gene expression and protein production

Although the DNA that is present in all the somatic cells of an organism is identical, specific genes are expressed in specific cells, in addition to many other genes that are expressed in every cell and fulfill more general functions required in each cell (Figure 1). We estimate humans have 411 different cell types[29], each encoding a subset of the >100000 possible proteins coded for by 20000 genes. The 8000 protein-coding genes (including many genes encoding transcription factors (TFs)) ubiquitously expressed in all different cell types are called housekeeping genes[30]. Thus the pattern of gene expression (the transcriptome) of each cell ultimately provides the cell's unique identity. Cells are under a tight gene regulation control program[31-34], with activation or repression of genes to modulate the transcriptome and hence the cell's identity and function. Fairly subtle differences in gene expression, the so-called "spatiotemporal expression of the genes"[35], suffice for this. The gene regulation program is

controlled by cell extrinsic (e.g. environmental stimulations) and intrinsic signals. Cell-to-cell communications via extracellular signal molecules and downstream signal transduction pathways (like those of the BMP/TGFβ family) are types of stimulation which can install and/or maintain the control of specific transcriptional programs[36, 37]. Another example is the NFκB pathway, which is strongly active in stressful, inflammatory and innate immune responses[38-40]. *SAMD4A* is among the first genes to respond to TNFα stimulation[41-43], a cytokine that signals through NFκB to orchestrate the inflammatory response[38] (discussed in **Chapter 4** and **5**).

Key features of transcription

Gene transcription: stepping stones of a complex procedure

Transcription is a complex mechanism, which precedes translation[13,15,16] and requires many essential elements for its completion. Many groups have joined Roger Kornberg's pioneering studies in the unravelling of the molecular mechanisms of eukaryotic transcription[10, 44, 45] (Figure 2). This process is highly mediated by DNA-dependent RNA polymerase whose enzymatic activity was identified by Weis and Gladstone at 1959 in rat liver nuclei[46, 47]. Eukaryotes have in fact three DNA dependent RNA polymerases: RNA Polymerase I (RNAPI), RNA Polymerase II (RNAPII) and RNA Polymerase III (RNAPIII)[48] and they were first described by Pierre Chambon[49] and Jam Tata[50]. Their chromatographic separation was achieved around 1969 by Roeder and Rutter[51, 52]. The enzyme responsible for eukaryotic DNA transcription of most studied genes is the RNAPII, also known as "polII" or "RNA polII"[53].
RNAPII binds to a sequence of DNA at the start of a gene to initiate transcription. This region is functionally known as the promoter and contains any combination of a number of sequence elements (see below) with a starting sequence that is often also containing ATG[39]. They contain specific nucleotide sequence elements and provide the necessary "space" required to harbour not only RNAPII, but also the Mediator (a large complex of proteins) and TFs necessary for the initiation of transcription[54-56] (Figure 2). Transcription initiation studies have unveiled two types of promoter, i.e. the focused and dispersed promoter, respectively[54, 57]. In focused initiation, transcription starts either at single nucleotide or over a narrow segment of nucleotides. Dispersed promoters have multiple start sites over 50 to 100 nucleotides most commonly found in CpG islands on the one hand and in constitutively expressed genes on the other hand[54, 57]. It should be clarified that dispersed promoters should not be confused with alternative promoters which are distinct and are often located far apart from each other. There are promoters which exhibit the properties of both the two aforementioned types of promoter; multiple dispersed start sites, but with one particularly strong and hence predominant start site. Focused promoters could be characterized as "ancient" as they appear to be predominant in simple organisms. However, 70% of the vertebrate genes have dispersed promoters. Biological interest in specific cell types was the main reason why most of studies for RNAPII have been conducted on focused promoters even though they are a minority in vertebrates when compared to dispersed promoters. Those studies led to the discovery of recognition motifs that are enriched in core promoters (Figure 2) such as the TATA box, Inr (initiator), BREu (upstream TFIIB recognition element), DPE (downstream promoter element), MTE (motif ten element), DCE (downstream core element) and XCPE1 (X core promoter element 1). Dispersed promoters generally lack TATA, BRE, DPE and MTE motifs[56, 58] with TATA-lacking promoters also being deficient in ATG triplets[59].
These aforementioned elements are recognized by some additional general TFs (GTFs, Figure 2)[60], such as TFIIA (TF for RNAPII), TFIIB, TFIID, TFIIE, TFIIF and TFIIH to promote the unwinding of DNA at the early steps of transcription[54, 57]. TFIID consists of the TBP subunit, which recognizes the TATA box, and the TAF subunits which recognize the Inr (TAF1, TAF2), the DCE (TAF1) and the DPE (TAF6, TAF9). In summary, RNAPII and GTFs are recruited to the promoter to form the pre-initiation complex (PIC)[39, 61, 62]. Furthermore, the multi-protein complex Mediator facilitates the localization of the PIC to the pro-

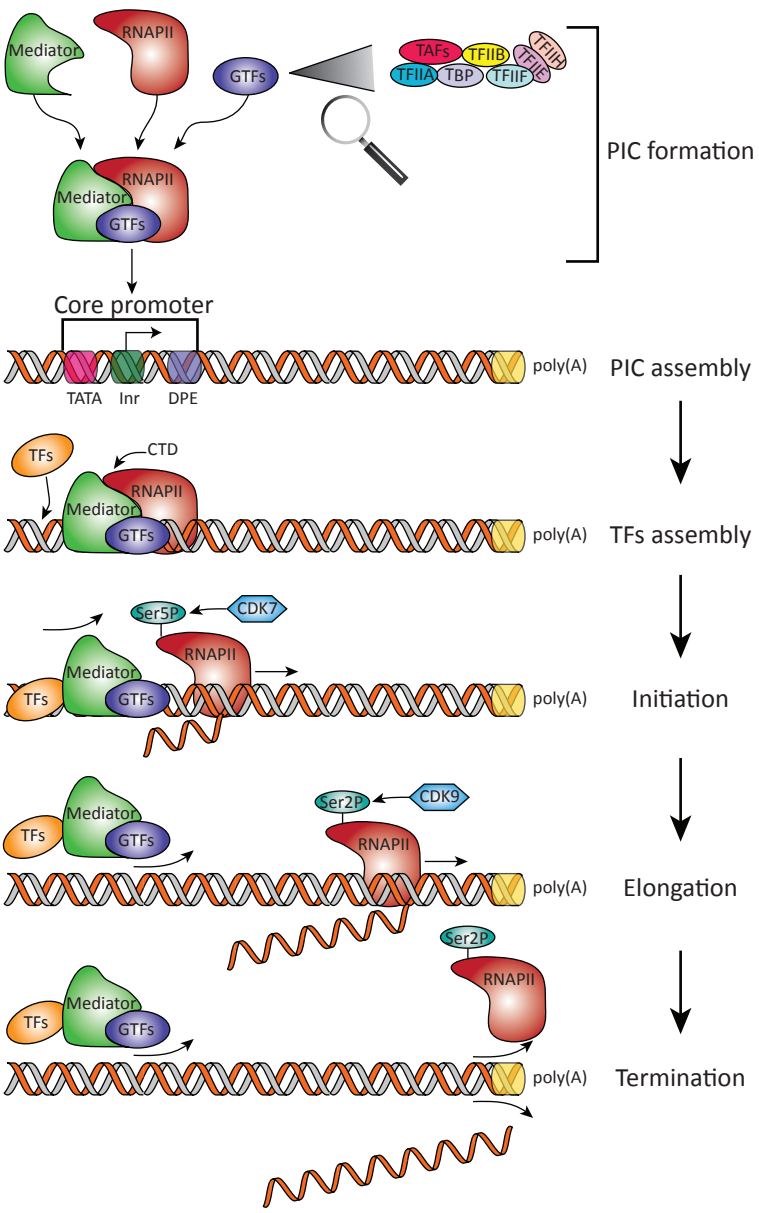

Figure 2: An example of transcription process by RNAPII. PIC (pre-initiation complex) is assembled by RNAPII, the Mediator and GTFs and recruited to the promoter (TSS) of the gene, which is going to be transcribed. TFs are recruited to the vicinity of PIC, CDK7 phosphorylates Ser5 of RNAPII to escape the promoter with the subsequent release of Mediator and GTFs resulting in the initiation of transcription with the formation the initiation complex. Subsequently, CDK9 phosphorylates Ser2 of RNAPII resulting in the elongation of the transcription. When RNAPII transcribes through a poly(A) site, termination proteins pause RNAPII, the RNA transcript is released and transcription is terminated.

moters[63, 64]. Briefly, TFIIA promotes the binding of TBP to the TATA box[65] and subsequently TFIID binds to the TATA box. TFIIB (interacts with TBP and promotes the recruitment of RNAPII to the promoter[65]) followed by TFIIF (stabilizes TBP/TFIIB interaction[66], attracts TFIIE and TFIIH[67]) and RNAPII are recruited to the promoter[68]. Then, TFIIE joins the PIC and facilitates the enzymatic functions of TFIIH[69] and they both promote the transition from transcription initiation to elongation[60, 69]. TATA-less promoters have the TFIID recruited at the 30bp upstream region from the promoter through direct or indirect interactions with TBP associated factors[70].

Often RNAPII will fail to initiate transcription after already a few nucleotides (also referred as “abortive initiation”)[71]. Successful initiation is mediated by TFIIH[72]. CDK7, which is part of TFIIH, phosphorylates (P) the fifth serine (Ser5) of the RNAPII carboxy-terminal domain (CTD)[73] (also known as RNA PolII Ser5P) and triggers the disengagement of RNAPII from the promoter to initiate transcription[72, 74]. That

process is also known as promoter escape or promoter clearance and is characterized by the release of the Mediator and GTFs[75]. At that stage, RNA PolII Ser5P is also known as "initiating complex". The initiation complex can be paused after promoter escape[72, 76], especially by two factors, DSIF[77] and NELF[78]. Forty percent of the genes are estimated to undergo transcriptional pausing[79, 80] which may provide in these cases the time necessary to recruit additional regulatory proteins to the PIC[81]. CDK9 and CyclinT1 phosphorylate NELF, DSIF and RNAPII at Ser2 (RNA PolII Ser2P) and the pause is released[82], resulting in transcription elongation. The RNA PolII Ser2 complex is also known as the "elongating complex". CDK7 and CDK9 are obviously two important kinases in transcription by RNAPII[73, 74,]. Genome-wide analysis of RNAPII Ser5P and Ser2P occupancy shows that RNAPII Ser5P diminishes from the promoter whereas RNAPII Ser2P is present throughout the transcribing gene. Transcription is complete near the end of the gene when the RNAPII transcribes through a poly(A) addition site. Termination proteins recognize that site, pause RNAPII and cleave the RNA transcript as part of 3' maturation process[83] exerted by the so-called CPSF complexes[84, 85]. In fact, CPSF is also able to couple to TFIID at transcription initiation and subsequently associates with the elongating polymerase[86].

Gene transcription and TFs: a relationship of love and hate

Apart from the GTFs, the PIC and the Mediator, gene transcription requires another specialized class or TFs, which direct the transcription machinery to the promoters[87]. Every TF binds a specific sequence motif or is part of a complex binding a specific motif that is distant from the TSS. They can bind at the promoter or at sequences distant from the promoter to activate/inhibit or increase/decrease transcription from the promoter[87]. The operon model of Francois Jacob and Jacques Monod in 1961, defined in the bacterial *E. coli* disaccharide lactose metabolism negative control system, was the stepping stone to the era of transcription regulation by introducing the idea that gene transcription can be repressed by a protein upon its binding to its target gene[88].

The specific DNA sequence motif, which the TFs recognize in order to bind to the DNA, is often between 6-12 bp. Such specific short motifs are distributed throughout the genome and offer a huge variety of possibilities to the TFs to occupy different positions. As a consequence, they can be located in exons, introns, promoters or intergenic/"gene desert" areas. Usually 200-300 bp contain multiple different motifs for different TFs, suggesting that TFs can cooperate in a combinatorial fashion[89] with other TFs (and their co-factors) recruited to the same vicinity, or alternatively that they compete/antagonize each other[87]. The spacing between motifs is often very important but can vary; the hematopoietic TFs, GATA1 and TAL1 have a spacing between them of about 8-11 nucleotides[90, 91] when they are present in the same complex. As mentioned above such motifs can be located in the vicinity of promoters[26] or in regulatory elements distant from promoters such as enhancers, silencers or even insulators[92].

One of the most challenging conundrums is the relationship between TF motifs; the underlying mechanism of recruitment of a specific TF (complex) to a specific motif/position while it is not recruited to other identical or similar motifs/positions throughout the genome. A proposed mechanism is that "pioneering TFs" promote the recruitment of specific TFs to specific genomic locations (discussed in **Chapter 6**). For example, it has been proposed that GATA2 can function as a "pioneer" TF to promote the binding of GATA1 to a subset of specific putative genomic locations[93, 94].

According to a recent study, humans have 1391 TF encoding genes[95], however this number will likely change in the foreseeable future through new studies. As explained above, the spatiotemporal expression of the genes[35] requires the activation of genes in specific tissues or at specific developmental stages, which is driven by tissue or stage specific expression of TFs and their localization to their target genes[87] (Figure 3A). That different expression of different TFs in different cells separates the cells into different categories or lineages and endows them with their unique characteristics[96] (Figure 3A).

In summary, TFs have many different properties (Figure 3B); they can interact with the Mediator or GTFs[97, 98], promote the recruitment of other TFs[95], synergize or compete and/or antagonize other TFs[99],

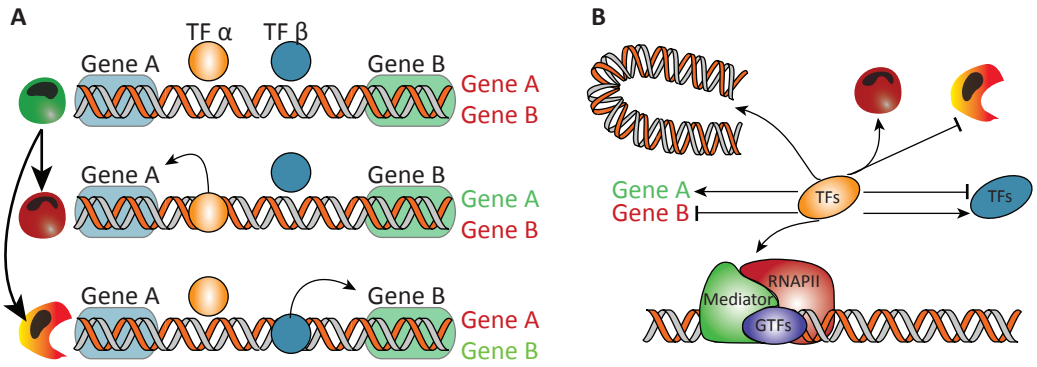

Figure 3: Properties of TFs. (A) TFs provide different gene expression in specific tissues or developmental stages. Gene A is active in one cell type and silenced in another where Gene B is active due to the effect of binding of different TF. **(B)** TFs can have different properties and generate specific cell lineages; they can promote chromatin looping, recruit and/or compete with other TFs, interact with the PIC and activate or repress their target gene(s).

promote specific cell lineages[96], interact with the transcription machinery[87], activate or repress their target gene(s)[92, 100] and also promote chromatin looping (discussed in **Chapter 3** and **5**)[92].

An example of those TFs, whose expression might contribute to different cell types, are the GATA factors which recognize the WGATAR nucleotide sequence (wherein W=A or T, R= A or G). *Gata1* is expressed in testis, megakaryocytes, eosinophils and erythroid precursors[101-108] whereas *Gata2* is expressed in hematopoietic stem cells (HSCs) and hematopoietic progenitors[99, 101, 106-108]. An interesting phenomenon is the "GATA switch"[99, 109, 110] (discussed in **Chapter 7**). *Gata2* expression activates *Gata1*, with GATA1 to subsequently replace GATA2 and repress *Gata2*[99]. Another TF, Friend-of-GATA (FOG1), can sometimes facilitate GATA1 occupancy and the GATA switch[111, 112] via its N-terminal sequence, which binds the NuRD chromatin remodelling complex[113]. This implies that the GATA switch requires chromatin remodelers such as the NuRD complex.

An important TF complex is the "LDB1 complex". Absence of LDB1 results in embryonic death of mouse embryos between E9.5 and E10.5 with hematopoietic and other defects[114-116]. The hematopoietic LDB1 complex consists of LDB1, GATA1/2, TAL1, E2A and LMO2[116, 117] plus a number of other proteins, including CDK9[97]. The LDB1 complex typically recognizes the E-box/GATA motifs whereby GATA1/2 dock at the GATA motifs. TAL1 recognizes the E-box motif[100] with E2A, which heterodimerizes with TAL1[118]. LMO2 forms a bridge between GATA1/2 and TAL1/E2A[90, 118] with LDB1 acting as a scaffold for attracting more TFs (like ETO2, MTGR1, RUNX1, LMO4)[119] to either suppress or activate their target genes[100]. The LDB1 complex is mainly localized in distal regulatory elements. The most prominent and well-studied example is the Locus Control Region (LCR) of the *β-globin* locus[120], which binds the LDB1 complex; this promotes the close proximity of the LCR to the *β-globin* gene[121-123].

As mentioned above, NFκB is a protein dimer consisting of subunits of a TF family that play a critical role in cell survival, inflammation, cell proliferation and differentiation[124]. They were first identified in 1986 as TFs whose specific DNA binding activity[121, 122] is promoted by an extracellular signal to provide an immediate early reaction of the genome to the signal[125]. NFκ-type TFs are important for activation and regulation of important genes upon response to developmental cues or extracellular stimuli[40, 125]. NFκB consists of a combination of five proteins: p65, cREL, RELB, p50 and p52, with the interaction between p65 and Mediator being important for activation of NFκB-dependent genes[126]. A structurally conserved REL homology region is shared between all the subunits; it is responsible for dimerization, nuclear entry, interaction with inhibitory IκB proteins and binding to their DNA target sequences (also known as κB sites)[126]. Prior to its activation, inhibitory proteins such as p100 or IκB members restrict NFκB to the cytoplasm, but upon activation, NFκB is quickly released from these and enters the nucleus to either repress[127] or activate[42] particular genes.

Chromatin compaction: more than meets the eye

Completely unfolded, linear DNA has a length of approximately 2 meters and is packed in a nucleus of 2-10μm diameter[12]. Hence, the DNA must be compacted in higher order structures, with electron microscopy (EM) analysis revealing several potential layers of packaging[128] (Figure 4A). The basic structural element of the packed DNA is the nucleosome; it consists of 146bp of DNA wrapped around an octameric protein complex composed of two copies of the H2A, H2B, H3 and H4 histones [129-131]. H1, the linker histone, is present in most nuclei and provides partial nuclease protection for approximately 20bp of DNA[129]. The compacted DNA with its histones is called "chromatin". When observed in EM, it resembles as "beads on a string" (beads are the nucleosomes and the string is the relaxed DNA). The nucleosome fibre has a diameter of 10nm.

The nucleosomes are connected by short DNA segments termed "linker DNA"[132]. A higher order of structure/compaction is formed by nucleosome-to-nucleosome interactions resulting in the 30nm fiber[132], a model proposed by Finch and Klug in 1976[129]. The nucleosomes are arranged in a zig-zag manner in order for a nucleosome to be close to the second neighboring nucleosome and not the immediate first neighbour[128]. Despite this often being regarded as a very regular structure mainly arising from a popular artistic representation, this is not the actual representation in the original publication[129]. A perpetual question in the field of chromatin biology is how the chromatin is folded into structures beyond the 30nm fiber. In the late 1970s, Laemmli and colleagues showed that chromosomes consist of 90Kb-sized loops; they also postulated that these are in contact with the nuclear matrix during mitosis, forming rosettes composed of approximately 18 loops, with 100 rosettes per mitotic chromosome[131]. In the 1990s, a new chromatin state named the "chromonema" fiber was postulated, which is composed by a 60-130 nm fiber interspersed by loosely folded 30nm segments[133].

The compactness of the genome, the respective ability of TFs and the assembled RNAPII machinery to access the chromatin are the main contributing factors for gene transcription and its regulation (Figure 4B). This can potentially be achieved in a *direct* or *indirect* way; nucleosome repositioning results in genome accessibility, which is orchestrated by chromatin organisation. It also harbours specific histone modifying proteins (histone readers) that target specific histones, yielding subsequently regulation of transcription[134-136]. Nucleosome remodelers and histone-modifying proteins promote the change of the chromatin from a highly compact (with no transcription state) to a relaxed state, enabling transcription, and *vice versa*.

Nucleosome repositioning can be achieved by specific chromatin remodelers via ATP hydrolysis, resulting in a nucleosome-free or dense genomic region with an impact to transcription regulation[136]. DNA can be wrapped around the nucleosomes and the histones in a relaxed or a tight manner, with or without gene expression, respectively. The first state is often termed as "open chromatin" or "euchromatin" and the latter as "closed chromatin" or "heterochromatin"[12, 134] (Figure 4B). In the first *direct* mechanism, TFs cannot access the heterochromatic DNA to promote their functions. However, in the euchromatin, pioneering TFs can attract and/or stabilize those TFs to the nucleosome-free positions and they subsequently can promote their functions (Figure 3B, 4B).

The "euchromatin" state can be easily determined by four different methods. The enzyme DNaseI (when applied in partial digestion conditions) can recognize "open chromatin" genomic positions, which are also called DNaseI-hypersensitivity sites. Digestion with Micrococcal nuclease (MNase) preferentially recognizes linker DNA and nucleosome-free genomic regions, resulting in nucleosomal "ladders" visible after gel electrophoretic separation. DNA methylation footprinting exploits the ability of DNA methyltransferases to add methyl-groups to specific nucleosome-free nucleotide sequences (i.e. to CG) more efficiently than to DNA wrapped around nucleosomes. With formaldehyde-assisted isolation of regulatory elements (FAIRE), chromatin is fixed by crosslinking, and the nucleosome-free DNA released from the interface in a phenol-based extraction selects for DNaseI-hypersensitivity sites and active promoters. All these methods can be coupled to high-throughput next generation sequencing

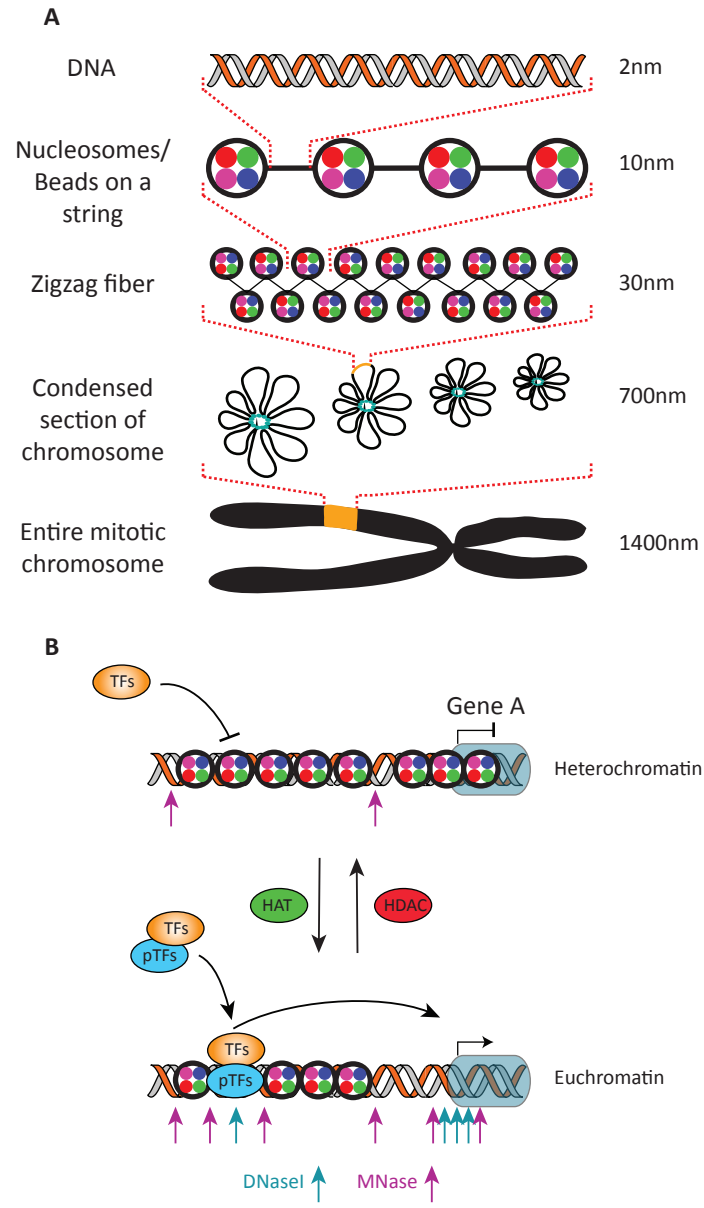

Figure 4: Chromatin packing. (A) Different stages of compactness of chromatin from the double strand DNA, to the 10nm fiber with "beads (nucleosomes) on string" where the naked DNA is the linker between the nucleosomes. That is followed by the 30nm fiber. These fibers form rosettes which condense further to result in mitotic chromosomes[12]. **(B)** Chromatin remodelling, from the heterochromatin "closed" chromatin to the euchromatin "open" chromatin, is achieved by nucleosome repositioning in a reversible manner. HATs (histone acetyltransferases) endorse euchromatin by reducing the positive charge of histones as a result of histone acetylation, whereas HDACs (histone deacetylases) promote heterochromatin due to low levels of histone acetylation. Upon nucleosome repositioning, the pioneering TFs (pTFs) are thought to attract and stabilize the TFs to the "open" chromatin regions which in turn can activate Gene A. "Open" chromatin regions can be identified in a genome wide manner by DNAseI signature (cyan arrows; also termed DNAseI hypersensitivity sites) and/or MNAseI (purple arrows; recognize nucleosome free regions).

(DNaseI-seq, MNase-seq, DNAme-seq, FAIRE-seq) in order to access the "open chromatin" areas in a genome-wide manner[137] (Figure 4B). Prediction of the "open chromatin" genomic regions can be an efficient tool to predict the positions where TFs and/or the transcription machinery are bound to the genome thereby providing information as to which genes are transcribed or poised.

Histone-modifying proteins bind to a specific segment of histones, usually the "histone tail", resulting in specific post-translational modifications (PTMs) of these histones[138]. PTMs were first described by Vincent Allfrey in the early 1960[139] and meanwhile have yielded a great variety of PTMs: acetylation, phosphorylation, methylation, ubiquitylation, sumoylation, neddylation, and others have been described[138]. Such modifications can have either a *direct* or an *indirect* effect. The *direct* effect influences the structure of the genome by changing the charge of histones, resulting in relaxed *versus* condensed chromatin. The first state can be achieved by acetylation and/or phosphorylation which can reduce the

positive charge of histones thereby diminishing the interactions with the negatively charged DNA[138]. Enzymes such as the histone acetyltransferases (HATs) which add certain modifications to the "histone tails" are often called "histone writers", while the "histone erasers" (HDACs; histone deacetylases) remove the histone modifications[140] (Figure 4B). The *indirect* mechanisms underlying gene expression modulation involve the recruitment of specific effector molecules, the "readers", which are separated into three categories, i.e. chromatin architectural proteins, chromatin remodelers, recruiters of other TFs, respectively, depending on their function. These "readers" determine the functional outcome of specific PTMs[135].

The full repertoire of the PTMs make up the epigenome (from the Greek *επί-γονιδίωμα*, on top of the genome). Histone modification signatures can be used to predict the position, the role and function of regulatory elements such as enhancers (discussed below) and promoters (Figure 5B). Promoters of actively transcribed genes are usually marked by H3K4me3, whereas silent promoters are enriched for H3K9me3 and H3K27me3[141]. Interestingly, the intermediate state between "active" and "silenced" promoters referred to as the "poised" promoters, are marked by H3K4me3 and H3K27me3[141]. Those promoters are also called "bivalent" since they have both activating and repressing histone modification signatures[141]. The "silenced" and "bivalent" promoters often contain the Polycomb repressive complexes (PRC1, PRC2) with PCR2 containing EZH2, which catalyzes H3K27 tri-methylation[142]. A large proportion of the "bivalent" PRC2 targeted promoters (inactive at pluripotent cell stage and rapidly induced or inactivated depending the developmental decision) corresponds to important developmental genes, which encode important TFs for successive developmental stages[143]. In addition, H3K36me3 and H3K79me2 correlate with transcription elongation[144]. Large regions of "heterochromatin" are characterized by H3K9me2 and H3K27me3 and are in contact with nuclear lamina associated domains (LADs)[141, 145, 146].

Gene transcription and regulatory elements

The complex linear organisation of metazoan genomes encodes regulatory sequences that can be categorised into two major groups: enhancers and silencers[92] (Figure 5A). TFs can bind regulatory elements and regulate their target gene(s) by either activating or repressing them. Even though those elements have been studied[125, 147, 148] and reviewed extensively[35, 87, 92, 149-152], they still pose a challenge for gene regulation studies.

Enhancers are short motifs (200-300 bp) that contain binding sites for TFs; they activate their target genes independent of enhancer orientation and often over great distance in *cis* (up to 1Mb) or can achieve this also in *trans*[150]. Silencers suppress/downregulate gene expression[152] and/or confine it within specific chromatin boundaries (and thus are sometimes also called 'insulators')[151] (Figure 5A). The interplay between these two types of contrasting regulatory element, their target promoters and epigenetic modifications at all levels of 3D organisation (that is, nucleosomes, chromatin fibres, loops, rosettes, chromosomes and chromosome location)[153, 154] leads to fine-tuning of expression during development and differentiation. Although enhancers and silencers have apparently opposite effects, accumulating evidence suggests they share more properties than intuition would have suggested[155]. The latest studies estimated that 40% of the genome has some regulatory potential[156].

Enhancers

Enhancers were characterised almost 35 years ago[157, 158], especially in DNA tumor viruses at that time, but their current functional definitions vary because of their flexibility of action (whether in *cis* or in *trans*)[159], position (relative orientation and/or distance)[61] and genomic location (in "gene deserts", introns and/or untranslated regions)[150] (Figure 5A). Although sequence conservation between species can be an efficient predictor of enhancer identity, there are examples where genes with identical ex-

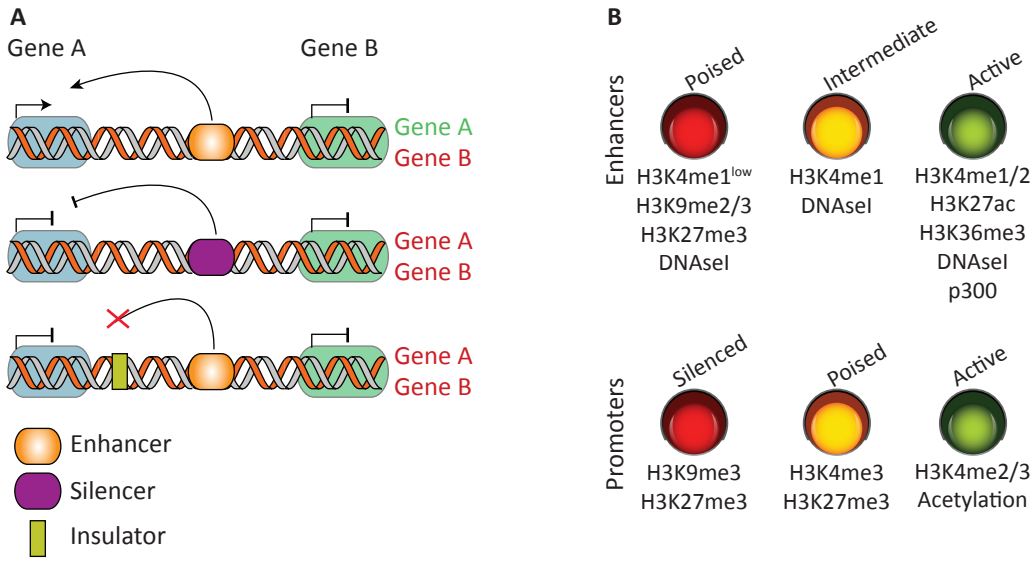

Figure 5: Regulatory elements and their signatures. (A) Enhancers and silencers can either activate or repress their target genes, whereas insulators can block the function of an enhancer or silencer. **(B)** Enhancers can be categorized into poised, intermediate and active depending different histone modification signatures. Similarly, promoters can be categorized into silenced, poised and active.

pression patterns in different species rely on enhancers that bear no similarities whatsoever[160]. Within a single genome, sensitivity to DNaseI and characteristic modifications of histone tails provide a more reliable means of identification of enhancers. “Active” enhancers typically occupy approximately 200bp of “open” chromatin (making them DNaseI-sensitive)[161], are flanked by regions rich in mono- and/or di-methylated lysine 4 of histone H3 (H3K4me1/H3K4me2), acetylated lysine 27 of histone H3 (H3K27ac) and, generally, bind p300[162]. Furthermore, the enhancers display low levels of trimethylated lysine 4 of histone H3 (H3K4me3) and low levels of nucleosome occupancy. Decreasing levels of H3K27ac and H3K36me3 can distinguish active enhancers in different subclasses with decreased activity[162]. Thus, those signatures can be used to profile enhancers in a genome-wide manner[141] (Figure 5B).

Genome-wide studies of genomic regions harboring histone modification markers have revealed that the regulatory elements of the genome have specific signatures, with every class of them having a distinct histone modification profile. Attempts have been made to classify enhancers into subclasses that are differentially used during development. Comparison between mouse embryonic stem (ES) cells, their differentiated derivatives and terminally differentiated murine cells allows distinctions between “active”, “intermediate” and “poised” enhancers (here, additional marks are used, e.g. H3K27me3 and/or H3K36me3)[147, 162]. “Intermediate” enhancers, are characterized by the presence of H3K4me1, DNaseI-hypersensitivity and the absence of H3K27Ac, with their target genes expressed at intermediate level, i.e. between “active” and “poised” enhancer target genes[147, 162]. However, “poised” enhancers have typically replaced H3K27Ac with H3K27me3, H3K9me2 and H3K9me3[147, 162] (Figure 5B). These accessible DNA stretches are often bound (and can thus be identified) by p300, Mediator subunits, chromodomain helicase DNA-binding protein 7 (Chd7), cohesin and/or CCCTC-binding factor (CTCF)[162, 163]. Most importantly, canonical enhancers are characterised by the presence of bound RNAPII[164].

A collection of enhancers can give rise to a Locus Control Region (recently also re-termed super-enhancers), firstly described by the Grosveld laboratory for the *β-globin* locus[120]; this LCR is located 15 to 60 kb upstream from the promoter of the different *globin* genes it regulates. The *β-globin* LCR is a collection of five regulatory elements marked by DNaseI-hypersensitivity (referred to as HS1-5)[120, 165-167]. HS1-4 are formed only in erythroid cells, whereas HS5 can be found in many different lineages[166,

[168]. Interestingly, HS2-4 but not HS1 and HS5 have an enhancing activity[165]. The hallmark of LCRs is that they consist of a group of regulatory elements, regulating a specific target gene and have the ability to maintain a chromatin in an "open" (nucleosome-free) state even when integrated into ectopic genomic sites. Since the identification of the *β-globin* LCR example, many other LCRs have been subsequently determined in other loci[166].

Lately, the Young lab and the Collins lab have characterized what they consider a new category of enhancers termed "super or stretched enhancers"[169, 170]. These consist of a cluster of enhancers spanning 12.5kb or above 3kb, respectively. They contain TF binding sites and control the transcription of important developmental/differentiation genes[170-172]. Genome-wide identification studies of "super or stretched enhancers" revealed that they have a significant overlap with LCRs, including the *β-globin* LCR[170]. Hence, it is debatable whether "super or stretched enhancers" are a new distinct category of enhancers or they simply represent the previously identified LCRs.

Enhancers are transcribed into RNAs (eRNAs) that do not encode proteins, run the length of the enhancer sequence, and appear to stabilise enhancer-promoter interactions[155, 173-175]. eRNAs derived from elements upstream of the Arc promoter depend on the activity of that promoter, as removing the promoter abolishes eRNA production[174]. *β-globin* associated ncRNAs are still produced in the absence of the *β-globin* promoter[174, 176, 177]. However, the rate at which eRNAs are turned over, the exact mechanism by which they function and their abundance (relative to the mRNAs they regulate) all remain to be determined.

The non-coding subset of the genome, the "gene desert", is gradually become regarded as important for the precise regulation by enhancers, silencers and insulators[61]. An additional class of ncRNAs longer than 200 nucleotides (long intergenic ncRNAs (lincRNAs)) were found in a survey of human transcripts, and some exhibited enhancer function[173]. More than 3000 lincRNAs have now been determined[178]. Some seem essential for the activation of particular promoters e.g. the promoter of the thymidine kinase gene, as well as for the expression of neighbouring protein-coding genes (although not all act as *bona fide* enhancers)[35]. For example, HOTTIP (a lincRNA transcribed from the 5' end of the HOXA locus) coordinates the activation of several HOXA genes; chromatin looping brings HOTTIP close to its targets, and this drives H3K4 trimethylation and transcription[179].

Silencers

At the opposite end lie silencers (Figure 5A). They prevent gene expression during differentiation and progression through the cell cycle[180]. This again correlates with RNA production (in some cases, through the generation of RNA duplexes that underlie the methylation of DNA at the promoter)[181, 182]. Accumulating documentation supports a broad and general role of both long and short RNA molecules in transcriptional inhibition. Antigene RNAs (agRNAs) are small RNAs that target promoters and downstream regions[181]. The expression of genes encoding progesterone, low-density lipoprotein, the androgen receptor, cyclooxygenase-2, the major vault protein and huntingtin is inhibited by agRNAs[181, 183]. Similarly, microRNAs (miRNAs), which are 20 to 22 nucleotides long, regulate gene expression post-transcriptionally[184], and they may also act at the level of transcriptional initiation or elongation. This is now supported by deep sequencing of nuclear and cytoplasmic small RNA libraries, where the majority of mature miRNAs localise in the nucleus (and not only in the cytoplasm)[184]. For instance, introduction of miRNA mimics that target the progesterone gene promoter decreases RNAPII occupancy. It also increases H3K9me2 levels in an Argonaute 2 (Ago2) dependent manner and leads to gene silencing[185]. Of note is that mature miRNAs in the nucleus can also act as "enhancers"[186].

PRC1 and PRC2 complexes rely on non-coding transcripts from silencing elements for recruitment to target sites. A range of examples are available: for instance, repression in *cis* in CD4+ T-cells and ES cells, where PRC2-catalysed H3K27me3 recruits PRC1 to prevent chromatin remodelling of targeted loci[187] and the PRC2-HOTAIR interaction, where transcripts produced from the XOXC locus establish

repression of XOXD[178]. In human breast cancer cells, overexpression of HOTAIR results in the promiscuous association of PRC2 with more than 850 targets, which are in turn silenced[188]. Furthermore, in the well-studied cascade of X chromosome inactivation, the ncRNA Xist binds PRC2, which in turn drives H3K27me3[189, 190] and propagation of PRC1's binding to multiple sites along the silenced allele[191]. Here the 3D conformation is also critical for efficient silencing and results in chromatin compaction and/or rearrangement[192]. Such equilibrium may, however, be shifted by the eviction of Polycomb proteins to restore an active state[190].

Insulators

Functionally autonomous domains are strung along the chromatin fibre, and these need to be insulated from their neighbours to prevent the action of irrelevant enhancers and silencers (Figure 5A). Insulator or boundary elements perform this task. These can be further categorised as enhancer blockers (when the insulator is located between a promoter and a cognate enhancer) and barriers (when located between a promoter and a silencer)[193]. Mutating or deleting insulators alters the pattern of gene expression and leads to developmental defects[194].
It has been suggested that insulators evolved from a class of promoters binding a specific subset of TFs that drive chromatin remodelling and long range interactions[155]. Many are marked by DNaseI-hypersensitivity[195] and/or the presence of bound RNAPII. Specifically, in the *Drosophila* Hox gene cluster, stalled polymerases, in conjunction with elongation factors DISF and NELF, insulate four of eight promoters from Hox enhancers, and this correlates with the rearrangement and/or de novo formation of chromatin loops[196]. It has also been known for a long time that the insertion of a gene (i.e. promoter) between an enhancer and its target gene can silence its target gene due to competition for the enhancer[197].
Perhaps the most abundant protein associated with insulator activity is CTCF. In the well-studied example of the IGF2-H19 imprinted locus, CTCF prevents activation of the maternal *Igf2* allele by a distal enhancer. When its cognate binding site is lost, the gene is reactivated[198]. Nonetheless, in this locus, CTCF is a positive regulator of the H19 gene[188]. Moreover, CTCF mediates enhancer-promoter, insulator-insulator and insulator-promoter interactions[155]. The insulator function of CTCF is regulated by cohesins[199-201] with their respective binding sites to coincide in various cell types and loci.
However, the CTCF-cohesin duplet is characteristic of only one type of insulator or boundary. In a comprehensive mapping of such *Drosophila* elements, additional factors, such as boundary element associated factor, GAGA and CP190, were used to pinpoint and classify domain boundaries[202]. Again, DNaseI-hypersensitivity characterises many of these elements, and examples exist where their function is Ago2 dependent (and so transcription-dependent, but RNAi-independent)[203].

Unveiling the 3D structure of the genome; cutting the Gordian knot

Tethering together the regulatory elements; the core of a complex structure

In order for enhancers, silencers or insulators to have their function in gene expression regulation, they have to be in fairly close proximity with their target gene(s). The last decade, growing evidence supports the theory[112, 123, 159, 200, 204-207] that TFs, even ncRNA[208], and more recently also the Mediator and Cohesin complexes, stabilize that proximity[201, 209].
Four models were proposed to describe how those elements may regulate a gene; the *tracking*, *linking*, *relocation* and *looping* model[92] (Figure 6) with the *looping* model being the only one that suggests that proximity is important.
According to the *tracking* model (Figure 6A), a protein loads onto the enhancer and tracks along the chromatin fibre towards the promoter, where it stimulates transcription[210]. This model is based on

investigation of the activator gp45, which is loaded on the DNA and tracks its target gene promoter by sliding on the DNA[210]. The *linking* model is similar (Figure 6B), but here the loaded protein drives polymerisation of proteins in the direction of the promoter[211]. One example is the *Drosophila* Chip protein which interacts with other TFs to promote enhancer-promoter proximity[212]. In the *relocation* model (Figure 6C), a given gene relocates to compartments in the nucleus where enhancer-promoter interactions (and so transcription) are favoured[213, 214]. An evidence for that model is that active RNAPII can be identified in specific focal sites[215] inside the nucleus and that transcriptionally active genes are associated with those sites[216] (the "transcription factories")[146, 217, 218]. The *looping* model (which shares features with the relocation model, Figure 6D) predicts a direct contact between an enhancer and a relevant promoter that loops out the intervening DNA[121] and thus is closely linked to the 3D architecture of the genome[214, 219].

Next, activators that are bound to the enhancer interact with the Mediator complex, which recruits RNAPII and GTFs to the promoter[35, 63]. Recently, TFs[123, 207], Cohesin[201] and the Mediator[209] complex have been implicated in tethering the enhancer to its target promoter in close proximity in the 3D nuclear space thereby also promoting transcription regulation. This last model is now favoured, as it readily explains enhancer-promoter interactions in *trans*[220]. Furthermore, it is supported by a wealth of experimental data derived from chromatin conformation capture technology[112, 121-123, 159, 204-207, 221, 222] and modelling[146, 154, 219].

Similarly, among the three major models proposed for insulator function (roadblock associated with the *tracking* model, sink/decoy associated with the *looping* model, and the topological loop models, which are a combination of the *tracking*/*looping* models), the topological loop model is best supported by experimental data. For example, rearrangement and/or de novo formation of appropriately oriented loops efficiently insulates promoters from enhancer elements[223]. Note also that recent data show how gene repression dependent on gypsy insulators in *Drosophila* propagates between distant loci to be repressed via the organisation of local loops[224].

Gene regulation from distal regulatory elements via local looping or via broader rearrangements in 3D organisation is now a widely accepted concept. The best studied example is the regulation of the *β-globin* gene by its LCR. The two key elements, the LCR and the *β-globin* promoter, are located in close proximity in the 3D nuclear space forming new chromatin loops[112, 121-123, 159, 204-206, 221, 222]. This *β-globin*-LCR relationship is also a well-studied example of tissue and developmentally specific regulation[165-167]. All *cis*-regulatory elements in this locus are in close proximity, where they form an "active chromatin hub". This hub is present in mouse primary erythroblasts (mouse fetal liver; mFL) where the *β-globin* gene is actively transcribed, but not in mouse fetal brain (mFB; where the *β-globin* gene is not expressed)[123, 159, 205, 221, 222]. An active chromatin hub, as defined in the *β-globin* locus paradigm, arises from the 3D clustering of DNA-hypersensitive sites, depends on specific DNA-protein interactions and tethers together all essential components for transcriptional activation[159, 205, 222]. Other examples are GATA1, which represses *Kit* via specific loop formation, while its exchange with GATA2 reforms the enhancer-promoter loop and reactivates *Kit* expression[225]. The *IgH* locus is yet another example; a 2.7Mb region is reorganised spatially during *IgH* locus activation[154]. Similarly, various TFs have been implicated in the formation of regulatory chromatin loops, including EKLF[123]; CTCF[200, 201, 206]; cohesin[201, 226] and LDB1[100, 207]. Knocking them out or down individually results in loss of looping and changes in transcriptional state[112, 123].

Transcription Factories; a plausible explanation for genome organization and gene regulation

On a broader scale, the genome is organised non-randomly in 3D space[146, 154, 219] as a result of a variety of chromatin loops and rosettes[146, 213, 227] (unpublished data), and knowledge is emerging that transcription is also architecturally organised[146, 228]. The traditional model of transcription requires that the polymerase tracks along the DNA template, somewhat like a locomotive, resulting in the synthesis of

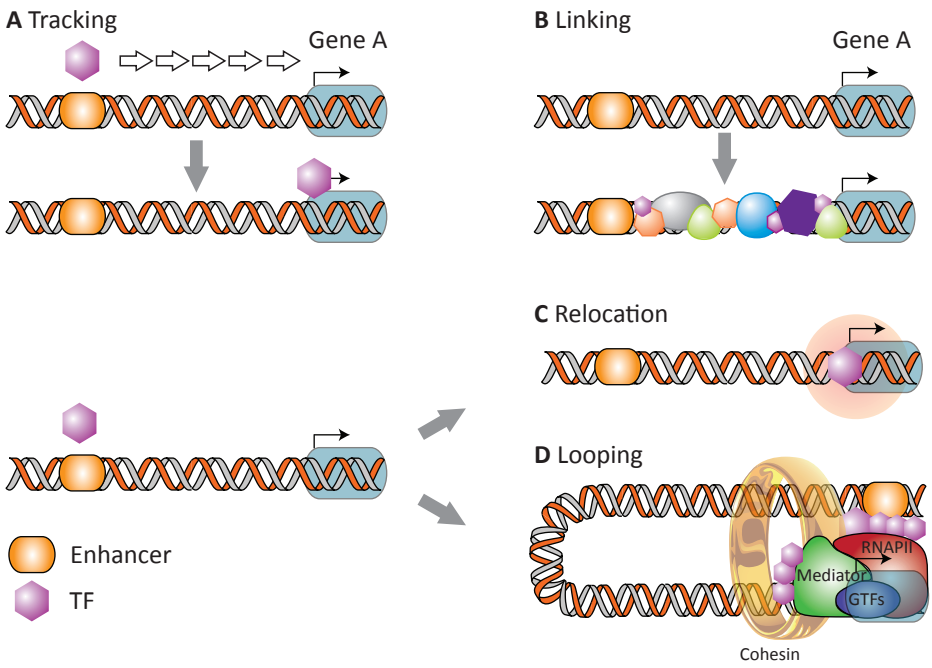

Figure 6: Existing models for the function of enhancers. The four existing models describing gene regulation by enhancers are depicted. **(A)** The *tracking* model, where a TF (purple hexagon) loads onto the enhancer and tracks along the chromatin fibre towards the promoter, where it stimulates transcription by association with the polymerase **(B)** The *linking* model, where the loaded TF drives polymerization of proteins in the direction of the promoter. **(C)** The *relocation* model, where a gene relocates to nuclear subcompartments (pink halo) favouring enhancer-promoter interactions, and so transcription. **(D)** The *looping* model, where the enhancer comes into proximity with the relevant promoter due to protein-protein interactions and stabilized by cohesin. This loops out the intervening chromatin and triggers transcriptional activation.

a transcript[12]. However, lately and increasing evidence suggests an alternative model. It proposes that transcription occurs in nucleoplasmic hot spots (called "transcription factories", see above) where a high local concentration of the required molecular machinery renders the whole process more efficient[146]. This term was first applied in 1968[229] by Cook and colleagues, who have contributed significantly to this concept[41, 42, 146, 214, 215, 217, 230-234]. In the 1990s the "transcription factory" term was used for the nuclear foci where DNA repair, DNA replication and gene transcription take place[215, 235, 236]. HeLa nuclei contain approximately 1µM of RNAPII but the concentration is estimated 1000-fold higher in such factories[237]. The underlying principle is that the polymerase is fixed to the "transcription factories", while the promoter of the gene diffuses to the factories and the DNA template moves through it by using the energy emanating from nucleotide triphosphate hydrolysis[146]. Possibly "transcription factories" are a collection of several "active chromatin hubs".

By definition, "transcription factories" harbour at least two RNA polymerases, each transcribing a different template. RNAPII genes themselves are also transcribed in separate factories from RNAPIII-dependent genes[234]. Actively transcribed genes tend to co-localise in the nucleus[216, 238], but different types of genes seem to cluster in "specialised" transcription factories, where they are co-regulated and co-expressed. The *β-globin* "transcription factory" contains at least two polymerases; one transcribes the enhancer and another transcribes a protein-coding gene. Interestingly, *Eraf* is located 25Mb away from the *β-globin* locus, appears to be in close proximity with the LCR and the *β-globin* locus in a foci rich with the required TFs and RNAPII[100, 216, 239]. Furthermore there is evidence that TNFα responsive genes, such as *SAMD4A* and *TNFAIP2* which are 50Mb apart, are located at the same "NFκB transcription factory"[42, 232]. Although factories can now be isolated and their proteins characterised using mass spectrometry[240], the mechanism by which factories are "marked" by specific TFs, and the relative rep-

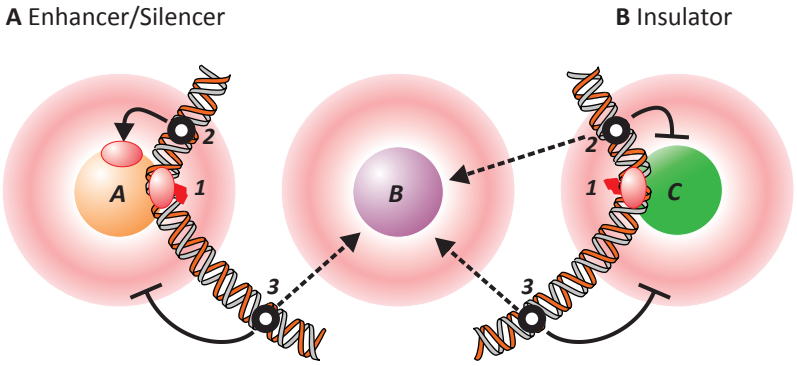

Figure 7: A simple model for the function of regulatory elements into transcription factories. Spheres A, B and C represent factories rich in different sets of TFs and associated halos indicate the probability that promoter 1, 2 or 3 will collide with a factory (red indicates high probability). The low-probability zone immediately around the factory arises because the intrinsic stiffness of the chromatin fibre restricts the formation of very small loops. Curved black arrow indicates collision between promoter and factory that yields a productive initiation. Dashed black arrows indicate the preferred site of initiation (as factory B is rich in the relevant TFs). Blocked red arrows indicate unproductive collisions (as the factory contains few of the relevant factors). **(A)** Enhancers and silencers. Transcription unit 1 is being transcribed by a polymerase in factory A. This tethers unit 2 in a 'hot zone', where it has a high probability of colliding with a polymerase in factory A (which contains high local concentrations of factors necessary for initiation by promoters 1 and 2). As a result, unit 1 acts as an enhancer for unit 2. At the same time, unit 3 is tethered far from factory B (which is rich in the factors required for its initiation). Here unit 1 acts as a silencer of unit 3. **(B)** Insulator. At a different stage in development, a different constellation of TFs are expressed. Chromatin domains containing units 2 and 3 (now be transcribed efficiently only in factory B which is rich in the necessary factors) are separated by unit 1 (now transcribed in factory C, which contains low concentrations of the factors required by units 2 and 3), so they rarely bind to factory B and interact. Here unit 1 acts as an insulator or barrier.

resentation of different subtypes of factories, remain undetermined.

How can these ideas be extended to explain the function of enhancers and silencers and/or insulators? All share common features; for example, DNaseI- hypersensitivity, active chromatin marks and interaction with TFs and RNAPII. Therefore, I propose that canonical regulatory elements are primarily transcription units and that, in order for them to be functional, they need to be transcribed (and so associated with a transcription factory). This hypothesis defines two key aspects of chromatin structure: (i) proximity between distant DNA sequences due to looping, and (ii) tethering of active genes to a factory.

Does the number of factories in a given cell suffice to accommodate all transcription units, including enhancers and/or silencers? To date, the lower estimate of 200 factories concerns murine primary cells and comes from RNAPII immunostaining *ex vivo*[216]. This suggests that about 80 transcription units would share a factory (assuming 16000 active transcription units, the estimated number in HeLa cells)[231] and that a number of them are transcribed outside a factory as well. Other approaches in HeLa cells return a number that is an order of magnitude higher: they propose approximately 2000 factories, each hosting an average of 8 transcription units[233]. Furthermore, the density and diameter of these transcriptional hot spots appear to be constant between cell types, suggesting an underlying topology accessible to transcription units in different nuclear neighbourhoods[231, 241]. The difference between these numbers may be explained by a difference in the sensitivity of the detection protocols used[231]. Nevertheless, the question whether most transcription occurs in factories has not been answered. It seems that most indeed does take place in factories as some estimates indicate that more than 95% of nascent nucleoplasmic RNA is found in factories[228, 233, 241]. Nonetheless, these issues will probably be resolved only by direct, live-imaging of such factories in cultured cells.

Now consider that an enhancer (transcription unit 1) (Figure 7A) tethers its target promoter (located in unit 2) close to factory or hub A that contains the necessary machinery. As a result, the target promoter 2 will diffuse through the nucleoplasm and frequently collide with a polymerase in factory A to initiate transcription. Although another promoter, i.e. unit 3, is also tethered close to the same factory,

it will initiate rarely (because factory A lacks the necessary TFs required by this promoter 3). However, promoter 3 can initiate in factory B (which does contain high concentrations of the relevant factors), but it will do so rarely, simply because it is tethered close to factory A and is kept far from factory B. Next, the transcription unit 1 would then act as an enhancer of unit 2 and as a silencer of unit 3. Then, the addition of histone modifications that mark the various units as active or inactive will enforce the status quo. After that, once unit 1 has been transcribed, these marks will make it more likely that unit 1 or unit 2 will reinitiate in factory A, thereby enabling to create a virtual cycle.
Similarly, at another developmental stage, when a different set of TFs are expressed (Figure 7B), unit 1 might be transcribed in factory C. It is again flanked by units 2 and 3, but these can now be transcribed efficiently only in factory B (which is rich in the necessary factors). As units 2 and 3 cannot stably interact with each other by binding to factory C, unit 1 now acts as an insulator or barrier. As before, histone marks will reinforce this different (virtual) cycle.
The model proposed here (Figure 7) illustrates a case where non-genic transcription unit 1, in its normal genomic location, acts as an enhancer, silencer or insulator or barrier, depending on the target and developmental stage. I imagine that most regulatory motifs normally act in only one way; however, when taken out from their normal context and moved elsewhere (usually the case in most of the assays used for testing the action of these motifs), they will act differently, depending on the new context (which includes proximity to an appropriate factory). This model encapsulates notions of transcriptional activity, epigenetic marks and 3D architecture, which, in combination, provide the context that determines promoter activity.

3D genome organization: important players for a perpetual question

Although the linear composition of the genome is clear, its 3D organization is relatively unknown, despite the increasing endeavours to unveil this uncharted territory. Chromosomes occupy distinct sub-nuclear volumes, the chromosomal territories (CTs)[242] (Figure 8A). A number of recent studies have shown that the genome, along the linear chromosomal axis within a CT, is organized in self-associating domains that are separated by linker regions[243]. These so-called “topological domains” or “topologically associated domains” (TADs) generally range from 300Kb to 1Mb and consist of a series of different types of chromatin loops, still in agreement with earlier models for the genome[153] (Figure 8B). Chromatin loops can be categorized into two groups; functional and structural loops[221]. Structural loops enable the folding of the genome creating TADs[243] with the base of the loop to define the domain boundaries. Functional loops are found within TADs and serve expression of genes[112, 123, 159, 204-207, 221, 222]; less frequently there are interactions between TADs[192, 221, 243]. Thus the “loop-within-loop” model proposes one “structural” TAD to contain a lot of “functional” loops (Figure 8B). A conundrum in the field of chromatin biology is the common misunderstanding that chromatin looping is conceived and perceived wrongly as chromatin interaction. Indeed, a chromatin loop, as detected by any of the chromosome conformation capture (3C) based technologies, only shows that two elements of the genome are in close proximity in the 3D nuclear space rather than documenting true interaction. Hence, whether those elements are actually interacting must be proven by functional experiments.
Within a TAD, the regulatory elements and their target genes are in general located in close proximity in the 3D nuclear space to enable the control and regulation of their target genes (Figure 8B). The observation that “interactions” within TADs are much more frequent than between TADs suggests a modular organization of the genome[221, 243]. Interestingly, depletion of CTCF results in increased “interactions” between TADs and reduced “interactions” within the same TAD, whereas the depletion of cohesin leads only to an overall reduction of “interactions”. Nonetheless, neither CTCF nor cohesin depletion alters the boundaries of TADs[201]. It appears that structure and function of genomes have co-evolved to maximize the expression of genetic information in a physically limited storage space. Even though this relationship seems obvious, the in-depth 3D architecture of genomes, their spatial and

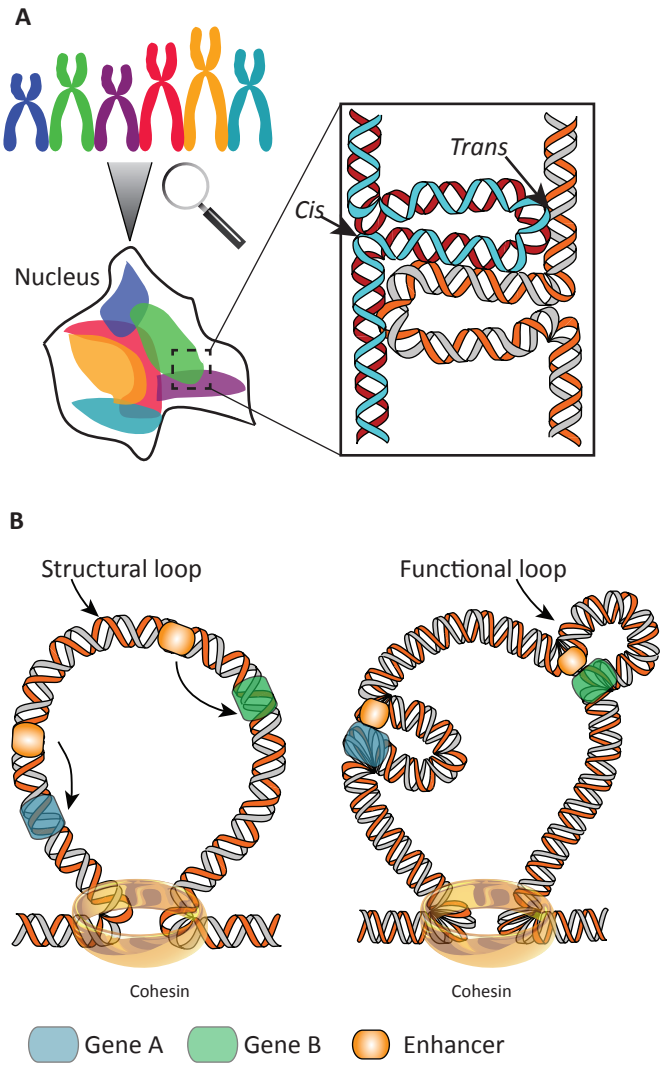

Figure 8: 3D genome organization. (A) Different chromosomes (depicted in different colors) are organized into CTs inside the nucleus occupying specific space and volume. Within and/or between those chromosomes, DNA segments are in close proximity in the 3D nuclear space either in *cis* (i.e. DNA segments of the same chromosome being in close proximity) or in *trans* (i.e. DNA segments of different chromosomes located in close proximity. **(B)** A "loop within loops" hypothetical model of how a TAD is formed; a structural loop of 1-2Mb forms a TAD with its base to define the domain boundary. Inside the structural loop, other smaller loops take place, either structural or functional tethering into close proximity regulatory elements with their target genes.

temporal dynamics of interactions and their relationship to transcription, replication and cell division, are still far from understood.

Towards the effort to unravel the chromatin loops and the subsequent 3D chromatin organization, the development of 3C technology has contributed significantly. The basic principle of various 3C-based methods is to detect which DNA segments are in close proximity (interactome) in the nuclear space, unravelling of the 3D structure of the genome.

Over the last decade a number of different 3C-based technologies have been developed. These are 3C-qPCR[244], 3C on a chip (therefore named 4C)[238, 245], 3C combined with high-throughput sequencing (4C-seq)[246], multiplexed 3C sequencing (3C-seq)[247], Targeted Locus Amplification (TLA)[248], Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)[249], 3C carbon copy (5C)[250], Hi-C[227], Tethered Conformation Capture (TCC)[251], Capture-C[252], capture Hi-C (cHi-C)[253] and Targeted Chromatin Capture (T2C)[221]. Each of these methods invariably involves the crosslinking of the chromatin (to link neighbouring segments of DNA together) and fragmentation by enzyme digestion followed by ligation of the DNA fragments (often termed as "3C template" at this step) at very low concentration. The latter promotes the ligation of linked fragments over ligation of non-linked fragments, hence promoting mono-molecular reactions. This results in the ligation of fragments that are close together in space but which do not need to be close to each other in the linear genome. These new ligation joints are subsequently "quantified" with different approaches and basically determine which genome fragments are in close proximity in the 3D nuclear space. As stated already above, all the aforementioned methods provide only an estimation of the proximity of the ligated products and not hard evidence of a functional interaction between them. Hence, none of them offers "absolute quantifiable" measurement of the true interactome, due to a number of parameters. For example, all the 3C-derived methods involve PCR-based am-

plification at either higher or lower level, which can be challenging due to limiting amounts of available biological material. Two DNA fragments (i.e. an enhancer and a promoter) can be in close proximity in the 3D nuclear space, but if there occurs a functional interaction between them (i.e. the enhancer promoting the transcription of its target gene), this has to be established by functional experiments.

All these techniques offer different advantages and limitations (also discussed in **Chapter 2** and **3**) and have provided valuable information on chromosomal interactions and gene transcription mechanisms[159, 221, 227, 238, 250]. 3C-qPCR, 4C, 3C-seq/4C-seq and TLA are based on fragmentation of the 3C template and measure the number of ligation events with qPCR, microarray or sequence analysis. They require knowledge of the genomic region of interest e.g. the location of at least one enhancer/silencer/promoter or of another genomic region of interest. With the exemption of 3C-qPCR, they all allow a "one to all" approach i.e. the interaction of one specific region (the viewpoint) with all other candidate regions. However, 5C is a "many to many" approach and requires the design of oligonucleotides for the DNA segments of interest for which the interaction network is retrieved. Hi-C and TCC are genome-wide "many to many" methods which provide the proximity network between many individual DNA fragments and the compartmentalization/TADs of the genome. Although these offer a genome-wide proximity map (interactome), they lack resolution as they usually offer a 40Kb resolution on average, which the latest algorithms have improved to 10Kb. However, the latter resolution requires a large sequencing effort (e.g. 3.4 billion mapped paired-end reads from six biological replicates)[254]. The latest improved Hi-C with 1Kb resolution, required an extreme deep-sequencing effort (approximately 6.6 billion paired-end reads) resulting in very high costs[255]. Capture-C[252] describes the interactome of specific DNA fragments of the genome (e.g. 455 promoters) selected with a dedicated oligonucleotide design for the mechanically sheared library and results in an output similar to 3C-seq/4C-seq. cHi-C applies a selection process like Capture-C for 22225 promoter regions of mouse ES cells, providing the "promoter only" interactome[256]. However, it also required massive sequencing (approximately 800 million mapped paired reads for six biological replicates) and binning of the reads covering the digested fragments[253, 256].

3C-seq[247] employs a restriction enzyme fragmentation and ligation at the 3C template that allows a "one to all" approach which can be changed to a "many to all" when multiplexed (multiplexed 3C-seq). Multiplexing offers the possibility of using different baits/viewpoints in the same sequencing lane; they can subsequently be separated based on their different indexes. This provides the simultaneous interactome of different regions of interest.

T2C[221] provides the genome-wide interactome for a selected region of (usually) up to 5Mb, which will comprise several TADs. T2C requires low sequencing efforts (up to 1/10 of a sequencing lane; less than 4 million mapped reads per sample) when compared to Hi-C (or cHi-C). Due to the high coverage and enrichment of the selected regions with mapped reads, T2C yields a much higher signal-to-noise ratio and does not require the binning of the reads (like in Hi-C, cHi-C or Capture-C), resulting in absolute restriction fragment resolution. Furthermore, every DNA fragment in the selected region of interest can be used as a single 3C-seq viewpoint. At the same time T2C offers the possibility to define TADs, their interactions and their boundaries, at a higher resolution when compared to Hi-C. Importantly T2C can easily be multiplexed as well. Hence, T2C is an affordable, cost-effective "two in one" method with single restriction fragment resolution that enables the exploration of the local spatial organization of the genome and chromatin interactions, without requiring massive sequencing efforts. Furthermore, due to the low number of PCR cycles used when compared to all the other methods, it offers more "quantifiable" conclusions, even though it requires a hybridisation step, which itself will introduce some bias. In this thesis I discuss the development and application of two of the 3C derived techniques; 3C-seq (discussed in **Chapter 2, 4** and **5**) and T2C (discussed in **Chapter 3** and **5**).

The next step in the chromatin biology field is to apply 3C technology without crosslinking, which was successful recently. This offers the advantage of a native, unbiased view of the compartmentalization of the genome and of the interaction network, without the interference of chemicals such as formal-

dehyde, which primarily targets lysines for crosslinking and hence introduces a bias in the "observed" interactions; this means that a short DNA fragment will have fewer bound proteins and hence fewer lysines than a larger fragment.

3D chromatin architecture

The chromatin fiber is in continuous motion and goes through subsequent rounds of condensation and de-condensation. Increasing efforts have been dedicated to unveil the shape and the structure of the variety of conformations that the chromatin fiber can adopt. The fiber is often perceived as a polymer. Thus, its physical properties can be explained in terms of so-called "random walk models"[130] (Figure 9). The basic principle of each of these models is that the chromatin consists of a series of rigid and non-flexible segments connected by flexible hinges. Several random walk models have been proposed to describe the dynamic shape and structure of the polymer chain[130]. These models are the *freely jointed chain, self-avoiding chain, worm-like chain, random walk/giant loop (RW/GL)* model*, multiloop subcompartment (MLS)* model*, fractal globule* and *random loop (RL)* model, each one having its own unique properties (Figure 9).
In the *freely jointed* model (Figure 9A), flexible and free to rotate hinges connect a series of rigid segments, also known as Kuhn segments, which can overlap and intersect another chain. The basic difference between the *freely jointed* model and the *self-avoiding* model is that the Kuhn segments cannot cross themselves or other segments[257] (Figure 9B). The *worm like chain* model (Figure 9C), also known as the Kratky and Porod chain, proposes that the polymer chain is continuously flexible (rather than flexible only within the hinges that separate the Kuhn segments like in the *freely jointed chain*). The chromatin structure of yeast was proposed to resemble to the *worm like chain* model[130, 258].
The *RW/GL* model (Figure 9D) proposes that the chromatin fiber is in random motion and consists of large loops of 2-5Mb tethered to loop attachment points with the DNA within the loops to follow a random walk[130, 259]. However measurements of genomic regions within those 2-5Mb loops did not agree with the *RW/GL* model[260]. High resolution microscopy and experimental procedures like FISH, revealed the *MLS* model (Figure 9E) in the late 1990s to describe long range chromatin folding[260-262]. The basic principle of the rosette like MLS model is that the chromatin is folded into 1-2Mb large loops and contain 60-120Kb sized loops attached to a common loop base and connected with a linker[130, 260-262].
In 2009, Lieberman-Aiden, Mirny and Dekker using Hi-C postulated that the genome folds into a shape called "*fractal globule*"[227, 263] (Figure 9F). The basic principle is that due to polymer condensation, a compact polymer emerges as a result of polymer condensation which prevents one region of the chain to pass across another one. This model has been first proposed theoretically by Grosberg[264] stating that interphase DNA can self-organize into a long-lived, non-equilibrium conformation. The "non-equilibrium globule" resembles to a pack of noodles at the moment they are starting to get boiled; dense but un-entangled. However, the "equilibrium globule" resembles to a pack of noodles after cooking where a single noodle is difficult to extract in contrast to the non-equilibrium model.
Evidence in the last years like the measurement of nucleosome distribution[265] as well as *in silico* simulations and Monte Carlo simulations (unpublished data) based on real experimental data make the rosette like *MLS* model most likely. As a result of the high resolution and coverage, T2C data from *in vitro* and *in vivo* biological material suggest that the loops of the rosettes are in the range of 30-100Kb with ~5±1 nucleosomes/11nm. Thus, it is rather tempting to speculate that the 1-2Mb loops are of a "structural" type and form the TADs, with the small 30-100 Kb sized loops to be "functional" to promote gene regulation (Figure 8B). Another more "flexible" model has also been proposed, the *RL* model (Figure 9G), where the loops have a dynamic size at random chromosomal intervals[266]. However, such a model contradicts with current data indicating that TADs remain the same between cell types and species[221, 243]. In order to resolve these different models, new computational methods as well as experimental techniques needed to be developed, which deliver genomic interaction output with high

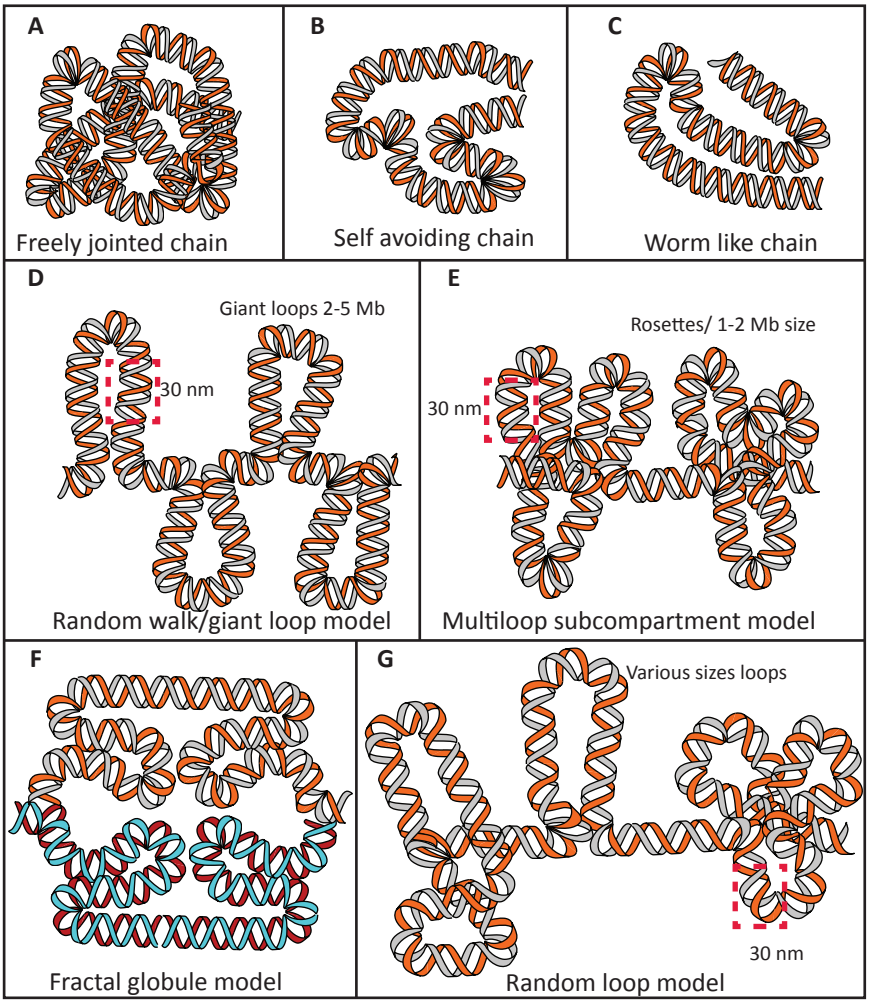

Figure 9: Polymer chain models of the chromatin fiber can be explained in terms of random walk models[130]. **(A)** The *freely joint* model contain flexible and rotate-free hinges connecting rigid segments (Kuhn Segments) which can overlap themselves on the contrary to the **(B)** *self avoiding* model. **(C)** The *worm like chain* model postulates that the polymer chain is continuously flexible comparing to the *freely joint* model. **(D)** The *RW/GL* model describes a randomly in motion chromatin fiber with large 2-5Mb loops whereas the **(E)** *rosette like MLS* model contain 1-2Mb giant loops with 60-120Kb loops inside them (Figure 8B) connected to the next rosette with a linker. The *fractal globule* **(F)** resembles a cauliflower when cut through the middle, everything is symmetrical at increasing levels of architectural organisation. The *RL* model **(G)** hypothesizes that the loops have a dynamic size at random chromosomal intervals.

resolution and coverage (like T2C) in order to elucidate the structure and the shape of the genome.

Unravelling a complex developmental procedure

Hematopoiesis

Hematopoietic development, or hematopoiesis (from the Greek *αίμα* "blood" and *ποιείν* "to make") is the process for the formation of the blood lineages from the hematopoietic stem cells (HSCs)[267]. HSCs were first identified in 1961 by Till and McCulloch; a population of stem cells in the bone marrow of adult mice responsible for the generation of all mature blood cells[268]. The first example of the proper-

ties of HSCs was deduced by HSC transplantation into an irradiated recipient with a depleted endogenous hematopoietic system and thereby rescuing the recipient's blood system[268]. Even though HCSs have been extensively studied with important breakthroughs for the field[267, 269], their true properties and characteristics still remain unknown. Today, HSCs can be quite highly purified based on specific cell surface markers[270] and can be truly characterized by their retrospective property of production of mature blood cells *in vivo* for prolonged periods[271].
The hematopoietic development is characterized by two "waves"; the primitive and the definitive wave[267]. The primitive wave in the mouse takes place in the yolk sac (YS) at approximately E7.5 (days of gestation) and produces sufficient primitive erythrocytes (red blood cells; RBCs), megakaryocytes and macrophages to sustain the growing embryo with the necessary oxygen[267, 272]. A second hematopoietic induction takes place in the dorsal aorta, at approximately E10.5 with the emergence of the first cells that will become definitive HSC[273]. Robin and colleagues were the first to "capture" this birth live[274]. However, HSCs can be also identified in placenta and the yolk sack at E11[267]. Subsequently, these definitive pre-HSCs migrate to the FL (E12) where they mature to HSC and expand significantly (more than 100 fold from E11 to E14) with a mechanism that is poorly understood since the identified numbers of HSCs exceed the expected rates based on the average cell division[275, 276]. This conundrum could be explained by the maturation of an intermediate pre-HSC into a mature HSC rather than the proliferation of the original HSC[276]. Finally, HSCs migrate to spleen, thymus and the bone marrow (E14.5-E17), the latter becoming the main hematopoietic source for the entire lifespan[277-279].
In vitro differentiation models (described below, Figure 10A) indicated that both waves are dependent on cell signaling stimuli. High Activin/Nodal and low Wnt-β-catenin favors the primitive wave, whereas the opposite support the development of the definitive wave[280].

Hemangioblast and hemogenic endothelium

Mesodermal cells from the primitive streak migrate to the YS around E7, aggregate and form blood islands. The central part of the blood islands generates primitive blood cells whereas the peripheral cells differentiate into endothelial cells. This parallel development of the hematopoietic and endothelial cells in these blood islands is in agreement with the hypothesis that they share a common ancestor/progenitor; the hemangioblast[281], which already was proposed at the early 1900s[282, 283].
In vitro differentiation of mouse embryonic stem (mES) cells provides relatively easy access to early developmentally stages which are difficult to study *in vivo* (Figure 10A). This property has been exploited to study the early stages of hematopoietic development. ES cell differentiation can recapitulate early embryonic events and generate three-dimensional, differentiated cell masses called embryonic bodies (EBs)[284, 285]. EBs are subjected to stimuli which direct embryonic development and can differentiate into all three germ layers; endoderm, ectoderm and mesoderm[284, 286]. The mES differentiation system to EBs produces a progenitor cell with the properties of the hemangioblast[287, 288]. Carefully timed EB differentiation (approximately 3.65-4 days) assays led to the characterization of a progenitor termed "blast colony-forming cells" (BL-CFCs). These can give rise to both endothelial and hematopoietic cells[287, 288] leading to the conclusion that BL-CFCs represents the *in vitro* equivalent of hemangioblasts. BL-CFCs express both the mesodermal marker Brachyury (Bry) and the receptor tyrosine-kinase FLK1 providing corroboration that these are mesodermal cells, which undergo specification towards the hematopoietic and vascular lineages[289]. In gastrulating mice and zebrafish embryos, BL-CFCs were identified providing evidence for the *in vivo* existence of hemangioblasts and that BL-CFC is not an artifact of the ES cell differentiation process[285, 290, 291]. This *in vivo* progenitor arises in the posterior primitive streak of the embryo, expresses *Bry* and *Flk1* and displays the same properties like the BL-CFC derived from the EBs[285].
It has recently been demonstrated that *in vivo* hematopoietic cells are derived from the hemangioblast through an intermediate state, i.e. a phenotypically differentiated endothelial cell with haematopoietic

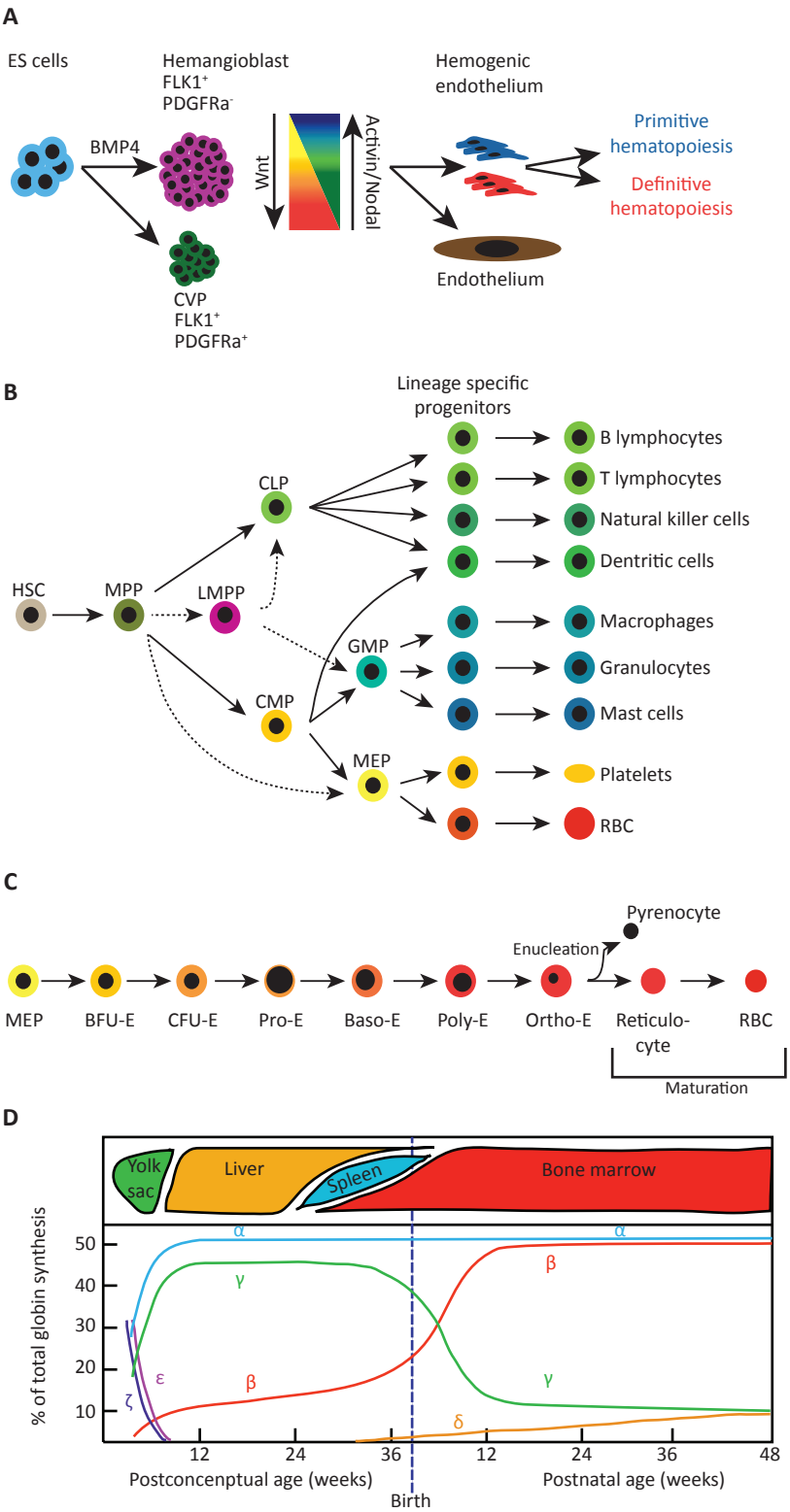

Figure 10: Different aspects of hematopoiesis. (A) *In vitro* model of mES cells differentiation through well-defined stages. Four days of mES differentiation produce EBs containing the *in vitro* equivalent hemangioblast ($FLK1^+$, $PDFRa^-$) and cardiovascular progenitors (CVP; $FLK1^+$, $PDFRa^+$). Subsequently the hemangioblast via the hemogenic endothelium gives rise to primitive or definitive hematopoiesis. This process is under the control of cell signalling stimulus such as the Bmp4, Wnt and the Activin/Nodal pathway. **(B)** Schematic model of hematopoietic development from HSCs towards different lineages and committed cells via a tree like branching process[278]. Each branching results in the cells losing their potency, from initial multipotent progenitors to more committed and mature cells for each lineage. The model is proposed by Irving Weissman[269] and updated by Sten Jacobsen[305] (dashed lines). Part of the hematopoietic development is erythropoiesis **(C)**, which is a well characterized process from MEPs towards enucleated RBCs via different well defined cell types[278]. **(D)** Hemoglobin levels in different developmental stages and tissues[278]. Two switches take place; from ε to γ in embryonic development and from γ to β/δ after birth.

potential; this hemogenic endothelium is transiently generated during BL-CFC development[292]. Interestingly, HSCs apart from the AGM, can be detected in the extra-embyonic vitelline artery and in the umbilical artery[273, 293]. In those arteries, HSCs resides into distinct intra-aortic hematopoietic clusters (IAHCs)[274]. Multiple studies in chicken[294] and mouse embryos[292, 295] have confirmed that IAHCs derive from the aortic hemogenic endothelium. However, recent evidence speculate that hemogenic endothelial cells do not form fully potent HSCs but rather establish first an intermediate cell population referred to as pre-HSCs (organized in IAHCs in the aorta) which will progressively develop to mature HSC in E12 FL[296]. It is important to note that although the hematopoietic and endothelial cells in YS appear simultaneously, the aorta and associated endothelium is formed one day before the emergence of IAHCs[297].

Interestingly, TFs such as *Tal1*[298] and *Runx1*[299] are important for the establishment of the hemogenic endothelium and the generation of the definitive hematopoietic cells from the hemogenic endothelium respectively[292]. *Runx1* expression in the dorsal aorta, umbilical and vitelline arteries, proceeds the emergence of hematopoietic cells suggesting that RUNX1 is a candidate TF to mark the hemogenic endothelium[300]. Its absence still leads to formation of the hemogenic endothelium but no definitive hematopoietic cells are formed[299-301]. The *in vitro* model contains the hemangioblast, which produces two types of endothelial cells; fully differentiated endothelial cells and hemogenic endothelial cells which can engender blood cells through the endothelial to hematopoietic transition[292, 302].

The hemangioblast ($FLK1^+/BRY^+$) via the hemogenic endothelium generates the primitive hematopoietic cells[290, 292]. In contrast, hemogenic endothelium located at different sites throughout the embryonic vasculature later gives rise to definitive hematopoietic cells[303, 304]. Lately, the Keller group has postulated the existence of “two types” of hemogenic endothelium (although still not possibly to be separated with the current cell surface markers) which can generate either primitive or definitive cells respectively. They are derived from two different types of mesodermal cells; primitive or definitive (depending the stimulus from either the Activin/Nodal or the Wnt signalling pathway), with the first one to have higher hemangioblast potential than the latter[280]. Hence, it is rather tempting to speculate that the hemogenic endothelium is a mix of primitive and definitive hematopoietic cells. It may imply, that the hemogenic endothelium is located at both YS (E7.5) and AGM (E10) and as a result of developmental cues or extracellular stimuli can give rise to either primitive or definitive hematopoietic cells.

Hematopoietic differentiation

Definitive hematopoiesis takes place in FL and adult bone marrow where HSCs give rise to three different lineages; the myeloid, the lymphoid and the erythroid[269, 278] (Figure 10B). That complex branched differentiation process includes several distinct stages; from HSCs to distinct multipotent progenitors for every lineage, which progressively differentiate through tightly controlled steps and by losing their multipotency to fully mature cells for each lineage. The laboratories of Irving Weissman[269] and Sten Jacobsen[305] have proposed the currently most accepted model of hematopoietic differentiation. Briefly, the HSCs form the multipotent progenitors (MPPs), which gradually loose their self-renewal ability and create the first branching point by differentiating into three distinct progenitors[269, 305]; the common lymphoid progenitor (CLP)[306] giving rise to all lymphoid cells, the common myeloid progenitor (CMP)[307] which engender the myeloid and erythroid lineage and the lymphoid-primed multipotent progenitor (LMPP) which develops to both CLP and granulocyte-macrophage progenitors (GMP)[305].

CLPs differentiate into different lymphoid cells such as B/T-lymphocytes, natural killer cells and the lymphoid derived dendritic cells[306]. The latter can be also derived from CMPs[308]. Furthermore, CMPs differentiate into either the megakaryocyte-erythrocyte progenitors (MEPs; can also derived directly from MPPs)[305] which give rise to the erythroid lineage or to the GMPs (also produced from LMPPs)[305] which generate the myeloid lineage (macrophage, granulocytes and mast cells)[307]. MEPs can further differentiate and branch into two lineages; megakaryopoiesis which generates megakaryocytes and

platelets and erythropoiesis which engenders red blood cells[278].

Erythropoiesis

During my thesis studies, the erythropoietic differentiation was mainly used to study chromatin dynamics and combinatorial TF networks. Erythropoiesis takes mainly place in erythroblastic islands in fetal liver and adult bone marrow and refers to the differentiation of MEPs towards the erythroid lineage with the subsequent production of mature RBCs[278, 309] (Figure 10C). RBCs can be distinguished into two categories; primitive and definitive. Briefly, YS produces and releases primitive RBCs in the circulation at E7.5[278, 310] whereas, FL produces the first definitive RBCs[267, 271, 278, 310, 311]. Finally, HSCs from the FL migrate to bone marrow to sustain the production of definitive RBCs for the entire lifetime[277-279].
The erythropoietic differentiation process is well defined and characterized with distinct cell types easily characterized based on cell surface markers and expression of specific genes. Erythropoiesis is discriminated into three defined and separated compartments; the progenitor, the precursor and the definitive RBC compartment. The first two are located in extravascular spaces and the third within the vascular network[309].
The progenitor compartment contains the first lineage-committed, definitive erythroid progenitor derived from MEPs, which is the burst-forming unit erythroid cells (BFU-E) followed by the colony-forming unit erythroid cells (CFU-E)[309]. Both BFU-E and CFU-E have the ability to form colonies of mature erythroid cells in semisolid media, with the first to require 7 and 14 days to grow and the latter 2 and 7 days for mouse and human respectively[309].
The second precursor compartment contains four different nucleated precursors, which are produced one after the other in a well-defined order: the proerythroblasts (ProE), the basophilic cells (BasoE), the polychromatophilic cells (PolyE) and the orthochromatic cells (OrthoE). This differentiation process has some well described properties: (i) erythroblast expansion through a limited set of symmetric cell divisions, (ii) accumulation of hemoglobin, (iii) decrease in cell size, (iv) nuclear pyknosis and (v) decrease in RNA content[309]. These different cell types are well characterized and discriminated both morphologically and with distinct cell surface markers providing the necessary tools to isolate and study them[309, 312].
The final step of maturation is enucleation resulting in two populations; the reticulocytes primarily containing hemoglobin[313] and the pyrenocytes consisting of the extruded nuclei[314] which are quickly ingested by the macrophages[315]. Reticulocytes maturation towards RBC is associated with an approximately 20% shrinkage, reduced cell volume and loss of organelles such as mitochondria and ribosomes[309, 316].
The third compartment, consists of the circulation of reticulocytes and RBCs in the vascular network and their perpetual production and release in the bloodstream to compensate the engulfment of senescent RBCs by macrophages in the spleen[317]. Adult human contain approximately 5×10^6 RBCs per microliter of blood with an average life span of 120 days. That results for a 70kg man with 5 litres of blood into 2.5×10^{13} total RBCs. Due to the continuous loss of RBCs we replace on average $1/115^{th}$ of RBCs per day resulting in the incredible number of 2.5×10^6 RBCs that have to be produced per second[278, 309, 318].

Hemoglobin

The main objective of erythropoiesis is the generation of erythrocytes to transfer oxygen from the lungs to the whole body and carbon dioxide in the opposite direction. Hemoglobin, the main protein produced in erythrocytes, binds oxygen and carbon dioxide. It is consists of two α-like and two β-like globin proteins forming a hetero-tetramer, which bind the oxygen via iron ion containing heme groups located in each one of the globin αβ dimers[319]. The α-like globins can be the α and ζ for either human

or mice. The β-like globin proteins for mice are εy, βh1, β^{major} and β^{minor} and for humans ε, γ^{G}, γ^{A}, δ and β[12, 278]. The genes are expressed in different developmental stages forming different types of hemoglobin during primitive or definitive hematopoiesis[320] (Figure 10D).
The first is characterized by the human ε, γ and ζ and the mouse εy, βh1 and ζ forming the following tetramers $\zeta_2\varepsilon_2$ $\zeta_2\gamma_2$ $\zeta_2\beta_2$ and $\alpha_2\varepsilon_2$. During definitive hematopoiesis in fetal liver, $\alpha_2\gamma_2$ (also known HbF) is expressed which allows the embryo to extract oxygen efficiently from the maternal blood. Interestingly, around birth the production of γ is replaced by β which is also expressed in fetal liver. In adults, HbF is usually around 1% of the total hemoglobin (although it varies in different humans) with the main hemoglobin to be $\alpha_2\beta_2$ (HbA1; 97% of the total hemoglobin) and $\alpha_2\delta_2$ (HbA2; 2% of the total hemoglobin)[278]. Intriguingly, some adults contain higher than expected HbF, a condition also known as hereditary persistence of fetal hemoglobin (HPFH)[321]. That provides an advantage when in the case of either sickle cell disease or β-thalassemia in adult patients. The high HbF ameliorates the severity of the symptoms of the diseases[278]. Hence, reactivation of the fetal γ globin in the adults has been one of the "holy grails" for the therapeutic approaches to treat those diseases.

The "labyrinth" of the genome; following Ariadne's Thread

In this biology era, we have been able to understand a good deal of different biological processes. However, it seems that we clearly still have a rather steep hill to climb before we will really be able to understand all the functions and properties of the genome. The studies in **Chapters 2-7** aim to contribute and place another brick in the wall towards deciphering the complicated questions of genome structure and organization in addition to complex mechanisms controlling developmental procedures.

References

1. Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G. & Worm, B. How many species are there on Earth and in the ocean? *PLoS Biol* **9**, e1001127 (2011).
2. O'Connor. C. Isolating the Hereditary Material: Frederick Griffith, Oswald Avery, Alfred Hershey and Martha Chase. **1((1):105)** (2008).
3. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
4. Levene, P. The structure of yeast nucleic acid. *The Journal of Biological Chemistry* **40, 415-424.** (1919).
5. Miescher, F. "Ueber die chemische Zusammensetzung der Eiterzellen" (On the chemical composition of pus cells). *Medicinisch-chemische Untersuchungen* **4 : 441–460** (1871).
6. Franklin, R.E. & Gosling, R.G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740-741 (1953).
7. Wilkins, M.H., Stokes, A.R. & Wilson, H.R. Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738-740 (1953).
8. Meselson, M. & Stahl, F.W. The Replication of DNA in Escherichia Coli. *Proc Natl Acad Sci U S A* **44**, 671-682 (1958).
9. Miller, O.L., Jr., Hamkalo, B.A. & Thomas, C.A., Jr. Visualization of bacterial genes in action. *Science* **169**, 392-395 (1970).
10. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209-1211 (1970).
11. Lenay, C. Hugo De Vries: from the theory of intracellular pangenesis to the rediscovery of Mendel. *C R Acad Sci III* **323**, 1053-1060 (2000).
12. Alberts, B. Molecular biology of the cell. 5th edition. *Garland Science* (2008).
13. Gerstein, M.B. et al. What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17**, 669-681 (2007).
14. Crick, F.H. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138-163 (1958).
15. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).
16. Cech, T.R. & Steitz, J.A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94 (2014).
17. Temin, H.M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**, 1211-1213 (1970).
18. Ahlquist, P. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* **296**, 1270-1273 (2002).
19. Tjio, J.H. & Puck, T.T. The Somatic Chomosomes of Man. *Proceedings of the National Academy of Sciences of the United States of America* **44**, 1229-1237 (1958).
20. Painter, T.S. A Comparison of the Chromosomes of the Rat and Mouse with Reference to the Question of Chromosome Homology in Mammals. *Genetics* **13**, 180-189 (1928).
21. Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. & Conso, I.H.G.S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).
22. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
23. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
24. Eddy, S.R. The C-value paradox, junk DNA and ENCODE. *Current biology : CB* **22**, R898-899 (2012).
25. Prasanth, K.V. & Spector, D.L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes & development* **21**, 11-42 (2007).
26. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147-151 (2003).
27. Hou, Y. & Lin, S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PloS one* **4**, e6978 (2009).
28. Taft, R.J., Pheasant, M. & Mattick, J.S. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**,

288-299 (2007).
29. Vickaryous, M.K. & Hall, B.K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev* **81**, 425-455 (2006).
30. Ramskold, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5**, e1000598 (2009).
31. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
32. Singer, Z.S. et al. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Molecular Cell* **55**, 319-331 (2014).
33. Copley, M.R., Beer, P.A. & Eaves, C.J. Hematopoietic Stem Cell Heterogeneity Takes Center Stage. *Cell Stem Cell* **10**, 690-697 (2012).
34. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226 (2008).
35. 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).
36. Conidi, A. et al. Few Smad proteins and many Smad-interacting proteins yield multiple functions and action modes in TGF beta/BMP signaling in vivo. *Cytokine Growth F R* **22**, 287-300 (2011).
37. Zwijsen, A., Verschueren, K. & Huylebroeck, D. New intracellular components of bone morphogenetic protein/Smad signaling cascades. *Febs Letters* **546**, 133-139 (2003).
38. Smale, S.T. Selective transcription in response to an inflammatory stimulus. *Cell* **140**, 833-844 (2010).
39. Alberts, B. Molecular biology of the cell. 5th edition. *Garland Science,* (2008).
40. Hoffmann, A. & Baltimore, D. Circuitry of nuclear factor kappaB signaling. *Immunological reviews* **210**, 171-186 (2006).
41. Papantonis, A. et al. TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *The EMBO journal* **31**, 4404-4414 (2012).
42. Papantonis, A. et al. Active RNA polymerases: mobile or immobile molecular machines? *PLoS biology* **8**, e1000419 (2010).
43. Diermeier, S. et al. TNFalpha signalling primes chromatin for NF-kappaB binding and induces rapid and widespread nucleosome repositioning. *Genome biology* **15**, 536 (2014).
44. Kelleher, R.J., 3rd, Flanagan, P.M. & Kornberg, R.D. A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell* **61**, 1209-1215 (1990).
45. Kornberg, R.D. An autobiographic conversation with Roger D. Kornberg on his work on transcription regulation. *Cell death and differentiation* **14**, 1977-1980 (2007).
46. Thomas, M.C. & Chiang , C.M. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41(3):105-78.** (2006).
47. Weiss, S. & Gladstone, L. A mammalian system for the incorporation of cytidine triphosphate into ribonucleic acid. *J Am Chem Soc Rev* **81**, 4118–4119.
48. Hsin, J.P. & Manley, J.L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development* **26**, 2119-2137 (2012).
49. Chambon, P., Ramuz, M. & Doly, J. Relation between soluble DNA-dependent RNA polymerase and “aggregate” RNA polymerase. *Biochemical and biophysical research communications* **21**, 156-161 (1965).
50. Widnell, C.C. & Tata, J.R. Evidence for Two DNA-Dependent Rna Polymerase Activities in Isolated Rat-Liver Nuclei. *Biochimica et biophysica acta* **87**, 531-533 (1964).
51. Roeder, R.G. & Rutter, W.J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224**, 234-237 (1969).
52. Roeder, R.G. & Rutter, W.J. Specific nucleolar and nucleoplasmic RNA polymerases. *Proceedings of the National Academy of Sciences of the United States of America* **65**, 675-682 (1970).
53. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular cell* **45**, 439-446 (2012).
54. Juven-Gershon, T. & Kadonaga, J.T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**, 225-229 (2010).
55. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature reviews. Genetics* **13**, 233-245 (2012).
56. Sandelin, A. et al. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews. Genetics* **8**, 424-436 (2007).
57. Juven-Gershon, T., Hsu, J.Y., Theisen, J.W. & Kadonaga, J.T. The RNA polymerase II core promoter - the gateway to transcription. *Current opinion in cell biology* **20**, 253-259 (2008).
58. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics* **38**, 626-635 (2006).
59. Lee, M.P. et al. ATG deserts define a novel core promoter subclass. *Genome research* **15**, 1189-1197 (2005).
60. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes & development* **10**, 2657-2683 (1996).
61. Maston, G.A., Evans, S.K. & Green, M.R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59 (2006).
62. Bieniossek, C. et al. The architecture of human general transcription factor TFIID core complex. *Nature* **493**, 699-702 (2013).
63. Malik, S. & Roeder, R.G. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet* **11**, 761-772 (2010).
64. Carlsten, J.O., Zhu, X. & Gustafsson, C.M. The multitalented Mediator complex. *Trends in biochemical sciences* **38**, 531-537 (2013).
65. Deng, W. & Roberts, S.G. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* **116**, 417-429 (2007).
66. Eichner, J., Chen, H.T., Warfield, L. & Hahn, S. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *The EMBO journal* **29**, 706-716 (2010).
67. Chen, H.T., Warfield, L. & Hahn, S. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nature structural & molecular biology* **14**, 696-703 (2007).
68. Okuda, M. et al. Structural insight into the TFIIE-TFIIH interaction: TFIIE and p53 share the binding region on TFIIH. *The EMBO journal* **27**, 1161-1171 (2008).
69. Okamoto, T. et al. Analysis of the role of TFIIE in transcriptional regulation through structure-function studies of the TFIIEbeta subunit. *The Journal of biological chemistry* **273**, 19866-19876 (1998).
70. Pugh, B.F. & Tjian, R. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes & development* **5**, 1935-1945 (1991).

71. Cheung, A.C. & Cramer, P. A movie of RNA polymerase II transcription. *Cell* **149**, 1431-1437 (2012).
72. Svejstrup, J.Q. The RNA polymerase II transcription cycle: cycling through chromatin. *Biochimica et biophysica acta* **1677**, 64-73 (2004).
73. Glover-Cutter, K. et al. TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. *Molecular and cellular biology* **29**, 5455-5464 (2009).
74. Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Molecular cell* **36**, 541-546 (2009).
75. Luse, D.S. Promoter clearance by RNA polymerase II. *Biochimica et biophysica acta* **1829**, 63-68 (2013).
76. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics* **13**, 720-731 (2012).
77. Wada, T. et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & development* **12**, 343-356 (1998).
78. Yamaguchi, Y. et al. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41-51 (1999).
79. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).
80. Min, I.M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes & development* **25**, 742-754 (2011).
81. Henriques, T. et al. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Molecular cell* **52**, 517-528 (2013).
82. Peterlin, B.M. & Price, D.H. Controlling the elongation phase of transcription with P-TEFb. *Molecular cell* **23**, 297-305 (2006).
83. Kuehner, J.N., Pearson, E.L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature reviews. Molecular cell biology* **12**, 283-294 (2011).
84. Kyburz, A., Friedlein, A., Langen, H. & Keller, W. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Molecular cell* **23**, 195-205 (2006).
85. Preker, P.J., Ohnacker, M., Minvielle-Sebastia, L. & Keller, W. A multisubunit 3' end processing factor from yeast containing poly(A) polymerase and homologues of the subunits of mammalian cleavage and polyadenylation specificity factor. *The EMBO journal* **16**, 4727-4737 (1997).
86. Dantonel, J.C., Murthy, K.G., Manley, J.L. & Tora, L. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. *Nature* **389**, 399-402 (1997).
87. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* **13**, 613-626 (2012).
88. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology* **3**, 318-356 (1961).
89. Wilson, N.K. et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-544 (2010).
90. Wadman, I.A. et al. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *The EMBO journal* **16**, 3145-3157 (1997).
91. Cohen-Kaminsky, S. et al. Chromatin immunoselection defines a TAL-1 target gene. *The EMBO journal* **17**, 5151-5160 (1998).
92. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & chromatin* **5**, 1 (2012).
93. Dore, L.C. & Crispino, J.D. Transcription factor networks in erythroid cell and megakaryocyte development. *Blood* **118**, 231-239 (2011).
94. May, G. et al. Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell stem cell* **13**, 754-768 (2013).
95. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263 (2009).
96. Yamamizu, K. et al. Identification of transcription factors for lineage-specific ESC differentiation. *Stem cell reports* **1**, 545-559 (2013).
97. Borggrefe, T. & Yue, X. Interactions between subunits of the Mediator complex with gene-specific transcription factors. *Seminars in cell & developmental biology* **22**, 759-768 (2011).
98. Allen, B.L. & Taatjes, D.J. The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology* **16**, 155-166 (2015).
99. Bresnick, E.H., Lee, H.Y., Fujiwara, T., Johnson, K.D. & Keles, S. GATA switches as developmental drivers. *The Journal of biological chemistry* **285**, 31087-31093 (2010).
100. Soler, E. et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**, 277-289 (2010).
101. Evans, T. & Felsenfeld, G. The erythroid-specific transcription factor Eryf1: a new finger protein. *Cell* **58**, 877-885 (1989).
102. Tsai, S.F. et al. Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446-451 (1989).
103. Leonard, M., Brice, M., Engel, J.D. & Papayannopoulou, T. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood* **82**, 1071-1079 (1993).
104. Martin, D.I., Zon, L.I., Mutter, G. & Orkin, S.H. Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. *Nature* **344**, 444-447 (1990).
105. Ito, E. et al. Erythroid transcription factor GATA-1 is abundantly transcribed in mouse testis. *Nature* **362**, 466-468 (1993).
106. Tsai, F.Y. et al. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221-226 (1994).
107. Tsai, F.Y. & Orkin, S.H. Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood* **89**, 3636-3643 (1997).
108. Liu, F. et al. Enhanced hemangioblast generation and improved vascular repair and regeneration from embryonic stem cells by defined transcription factors. *Stem cell reports* **1**, 166-182 (2013).
109. Van Handel, B. et al. Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell* **150**, 590-605 (2012).
110. Soler, E. et al. A systems approach to analyze transcription factors in mammalian cells. *Methods* **53**, 151-162 (2011).
111. Pal, S. et al. Coregulator-dependent facilitation of chromatin occupancy by GATA-1. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 980-985 (2004).

112. Vakoc, C.R. et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular cell* **17**, 453-462 (2005).
113. Hong, W. et al. FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1. *The EMBO journal* **24**, 2367-2378 (2005).
114. Mukhopadhyay, M. et al. Functional ablation of the mouse Ldb1 gene results in severe patterning defects during gastrulation. *Development* **130**, 495-505 (2003).
115. Mylona, A. et al. Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis. *Blood* **121**, 2902-2913 (2013).
116. Cantor, A.B. & Orkin, S.H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**, 3368-3376 (2002).
117. Tsiftsoglou, A.S., Vizirianakis, I.S. & Strouboulis, J. Erythropoiesis: model systems, molecular regulators, and developmental programs. *IUBMB life* **61**, 800-830 (2009).
118. El Omari, K. et al. Structural basis for LMO2-driven recruitment of the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell reports* **4**, 135-147 (2013).
119. Meier, N. et al. Novel binding partners of Ldb1 are required for haematopoietic development. *Development* **133**, 4913-4923 (2006).
120. Grosveld, F., van Assendelft, G.B., Greaves, D.R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975-985 (1987).
121. Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. & Grosveld, F. The effect of distance on long-range chromatin interactions. *Molecular cell* **1**, 131-139 (1997).
122. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).
123. Drissen, R. et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* **18**, 2485-2490 (2004).
124. Oeckinghaus, A. & Ghosh, S. The NF-kappaB family of transcription factors and its regulation. *Cold Spring Harbor perspectives in biology* **1**, a000034 (2009).
125. Sen, R. & Baltimore, D. Inducibility of kappa immunoglobulin enhancer-binding protein Nf-kappa B by a posttranslational mechanism. *Cell* **47**, 921-928 (1986).
126. van Essen, D., Engist, B., Natoli, G. & Saccani, S. Two modes of transcriptional activation at native promoters by NF-kappaB p65. *PLoS biology* **7**, e73 (2009).
127. Dong, J., Jimi, E., Zhong, H., Hayden, M.S. & Ghosh, S. Repression of gene expression by unphosphorylated NF-kappaB p65 through epigenetic mechanisms. *Genes & development* **22**, 1159-1173 (2008).
128. Woodcock, C.L. & Ghosh, R.P. Chromatin higher-order structure and dynamics. *Cold Spring Harb Perspect Biol* **2**, a000596 (2010).
129. Finch, J.T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A* **73**, 1897-1901 (1976).
130. Jhunjhunwala, S., van Zelm, M.C., Peak, M.M. & Murre, C. Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**, 435-448 (2009).
131. Paulson, J.R. & Laemmli, U.K. The structure of histone-depleted metaphase chromosomes. *Cell* **12**, 817-828 (1977).
132. Luger, K., Dechassa, M.L. & Tremethick, D.J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature reviews. Molecular cell biology* **13**, 436-447 (2012).
133. Belmont, A.S. & Bruce, K. Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *J Cell Biol* **127**, 287-302 (1994).
134. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).
135. Yun, M., Wu, J., Workman, J.L. & Li, B. Readers of histone modifications. *Cell research* **21**, 564-578 (2011).
136. Saha, A., Wittmeyer, J. & Cairns, B.R. Chromatin remodelling: the industrial revolution of DNA around histones. *Nature reviews. Molecular cell biology* **7**, 437-447 (2006).
137. Bell, O., Tiwari, V.K., Thoma, N.H. & Schubeler, D. Determinants and dynamics of genome accessibility. *Nature reviews. Genetics* **12**, 554-564 (2011).
138. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell research* **21**, 381-395 (2011).
139. Allfrey, V.G., Faulkner, R. & Mirsky, A.E. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **51**, 786-794 (1964).
140. Chi, P., Allis, C.D. & Wang, G.G. Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nature reviews. Cancer* **10**, 457-469 (2010).
141. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics* **12**, 7-18 (2011).
142. Sarma, K., Margueron, R., Ivanov, A., Pirrotta, V. & Reinberg, D. Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. *Molecular and cellular biology* **28**, 2718-2731 (2008).
143. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).
144. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318 (2007).
145. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).
146. Cook, P.R. A model for all genomes: the role of transcription factories. *J Mol Biol* **395**, 1-10 (2010).
147. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).
148. Rada-Iglesias, A. et al. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell stem cell* **11**, 633-648 (2012).
149. Rada-Iglesias, A., Prescott, S.L. & Wysocka, J. Human genetic variation within neural crest enhancers: molecular and phenotypic implications. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120360 (2013).
150. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).
151. Yang, J. & Corces, V.G. Chromatin insulators: a role in nuclear organization and gene expression. *Adv Cancer Res* **110**, 43-76 (2011).
152. Maeda, R.K. & Karch, F. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Current opinion in genetics & development* **21**, 187-193 (2011).

153. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews. Genetics* **2**, 292-301 (2001).
154. Jhunjhunwala, S. et al. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**, 265-279 (2008).
155. Raab, J.R. & Kamakaka, R.T. Insulators and promoters: closer than we think. *Nat Rev Genet* **11**, 439-446 (2010).
156. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506 (2013).
157. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).
158. Khoury, G. & Gruss, P. Enhancer elements. *Cell* **33**, 313-314 (1983).
159. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* **10**, 1453-1465 (2002).
160. Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R. & Eisen, M.B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS genetics* **4**, e1000106 (2008).
161. He, H.H. et al. Nucleosome dynamics define transcriptional enhancers. *Nature genetics* **42**, 343-347 (2010).
162. Zentner, G.E., Tesar, P.J. & Scacheri, P.C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* **21**, 1273-1283 (2011).
163. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).
164. De Santa, F. et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8**, e1000384 (2010).
165. Fraser, P. & Grosveld, F. Locus control regions, chromatin activation and transcription. *Current opinion in cell biology* **10**, 361-365 (1998).
166. Li, Q., Peterson, K.R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077-3086 (2002).
167. Grosveld, F. Activation by locus control regions? *Current opinion in genetics & development* **9**, 152-157 (1999).
168. Li, Q., Zhang, M., Duan, Z. & Stamatoyannopoulos, G. Structural analysis and mapping of DNase I hypersensitivity of HS5 of the beta-globin locus control region. *Genomics* **61**, 183-193 (1999).
169. Whyte, W.A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).
170. Parker, S.C. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17921-17926 (2013).
171. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).
172. Wang, H. et al. NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 705-710 (2014).
173. Orom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).
174. Orom, U.A. & Shiekhattar, R. Long non-coding RNAs and enhancers. *Curr Opin Genet Dev* **21**, 194-198 (2011).
175. Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394 (2011).
176. Ling, J., Baibakov, B., Pi, W., Emerson, B.M. & Tuan, D. The HS2 enhancer of the beta-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a cis-linked globin promoter. *Journal of molecular biology* **350**, 883-896 (2005).
177. Ling, J. et al. HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *The Journal of biological chemistry* **279**, 51704-51713 (2004).
178. Khalil, A.M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).
179. Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).
180. Li, L.M. & Arnosti, D.N. Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. *Current biology : CB* **21**, 406-412 (2011).
181. Janowski, B.A. & Corey, D.R. Minireview: Switching on progesterone receptor expression with duplex RNA. *Molecular endocrinology* **24**, 2243-2252 (2010).
182. Morris, K.V., Chan, S.W., Jacobsen, S.E. & Looney, D.J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305**, 1289-1292 (2004).
183. Janowski, B.A. et al. Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. *Nature chemical biology* **1**, 216-222 (2005).
184. Liao, J.Y. et al. Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers. *PloS one* **5**, e10563 (2010).
185. Younger, S.T. & Corey, D.R. Transcriptional gene silencing in mammalian cells by miRNA mimics that target gene promoters. *Nucleic acids research* **39**, 5682-5691 (2011).
186. Pawlicki, J.M. & Steitz, J.A. Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. *The Journal of cell biology* **182**, 61-76 (2008).
187. Kanhere, A. et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Molecular cell* **38**, 675-688 (2010).
188. Schoenherr, C.J., Levorse, J.M. & Tilghman, S.M. CTCF maintains differential methylation at the Igf2/H19 locus. *Nat Genet* **33**, 66-69 (2003).
189. Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. & Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).
190. Vernimmen, D. et al. Polycomb eviction as a new distant enhancer function. *Genes & development* **25**, 1583-1588 (2011).
191. Bernstein, E. et al. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and cellular biology* **26**, 2560-2569 (2006).
192. Splinter, E. et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development* **25**, 1371-1383 (2011).
193. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**, 703-713 (2006).

194. Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433-440 (2007).
195. Valenzuela, L. & Kamakaka, R.T. Chromatin insulators. *Annu Rev Genet* **40**, 107-138 (2006).
196. Chopra, V.S., Cande, J., Hong, J.W. & Levine, M. Stalled Hox promoters as chromosomal boundaries. *Genes & development* **23**, 1505-1509 (2009).
197. Hanscombe, O. et al. High-level, erythroid-specific expression of the human alpha-globin gene in transgenic mice and the production of human hemoglobin in murine erythrocytes. *Genes & development* **3**, 1572-1581 (1989).
198. Hark, A.T. et al. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486-489 (2000).
199. Parelho, V. et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433 (2008).
200. Wendt, K.S. et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801 (2008).
201. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 996-1001 (2014).
202. Negre, N. et al. A comprehensive map of insulator elements for the Drosophila genome. *PLoS genetics* **6**, e1000814 (2010).
203. Moshkovich, N. et al. RNAi-independent role for Argonaute2 in CTCF/CP190 chromatin insulator function. *Genes Dev* **25**, 1686-1701 (2011).
204. Palstra, R.J., de Laat, W. & Grosveld, F. Beta-globin regulation and long-range interactions. *Advances in genetics* **61**, 107-142 (2008).
205. Palstra, R.J. et al. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* **35**, 190-194 (2003).
206. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-2354 (2006).
207. Stadhouders, R. et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *The EMBO journal* **31**, 986-999 (2012).
208. Orom, U.A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154**, 1190-1193 (2013).
209. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435 (2010).
210. Kolesky, S.E., Ouhammouch, M. & Geiduschek, E.P. The mechanism of transcriptional activation by the topologically DNA-linked sliding clamp of bacteriophage T4. *Journal of molecular biology* **321**, 767-784 (2002).
211. Bulger, M. & Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes & development* **13**, 2465-2477 (1999).
212. Morcillo, P., Rosen, C., Baylies, M.K. & Dorsett, D. Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in Drosophila. *Genes & development* **11**, 2729-2740 (1997).
213. Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature reviews. Genetics* **8**, 104-115 (2007).
214. Cook, P.R. Nongenic transcription, gene regulation and action at a distance. *J Cell Sci* **116**, 4483-4491 (2003).
215. Jackson, D.A., Hassan, A.B., Errington, R.J. & Cook, P.R. Visualization of focal sites of transcription within human nuclei. *The EMBO journal* **12**, 1059-1065 (1993).
216. Osborne, C.S. et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-1071 (2004).
217. Papantonis, A. & Cook, P.R. Transcription factories: genome organization and gene regulation. *Chem Rev* **113**, 8683-8705 (2013).
218. Razin, S.V. et al. Transcription factories in the context of the nuclear and genome organization. *Nucleic acids research* **39**, 9085-9092 (2011).
219. Knoch, T.A., Goker, M., Lohner, R., Abuseiris, A. & Grosveld, F.G. Fine-structured multi-scaling long-range correlations in completely sequenced genomes--features, origin, and classification. *European biophysics journal : EBJ* **38**, 757-779 (2009).
220. Lomvardas, S. et al. Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403-413 (2006).
221. Kolovos, P. et al. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics & chromatin* **7**, 10 (2014).
222. de Laat, W. & Grosveld, F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* **11**, 447-459 (2003).
223. Gohl, D. et al. Mechanism of chromosomal boundary action: roadblock, sink, or loop? *Genetics* **187**, 731-748 (2011).
224. 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. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2294-2299 (2011).
225. 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).
226. Nativio, R. et al. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS genetics* **5**, e1000739 (2009).
227. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
228. Sutherland, H. & Bickmore, W.A. Transcription factories: gene expression in unions? *Nat Rev Genet* **10**, 457-466 (2009).
229. Jungwirth, C. & Launer, J. Effect of poxvirus infection on host cell deoxyribonucleic acid synthesis. *Journal of virology* **2**, 401-408 (1968).
230. Papantonis, A. & Cook, P.R. Genome architecture and the role of transcription. *Current opinion in cell biology* **22**, 271-276 (2010).
231. Pombo, A. et al. Regional and temporal specialization in the nucleus: a transcriptionally-active nuclear domain rich in PTF, Oct1 and PIKA antigens associates with specific chromosomes early in the cell cycle. *The EMBO journal* **17**, 1768-1778 (1998).
232. Papantonis, A. & Cook, P.R. Fixing the model for transcription: the DNA moves, not the polymerase. *Transcription* **2**, 41-44 (2011).
233. Iborra, F.J., Pombo, A., Jackson, D.A. & Cook, P.R. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *Journal of cell science* **109 (Pt 6)**, 1427-1436 (1996).
234. Pombo, A. et al. Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J* **18**, 2241-2253 (1999).
235. Hozak, P., Hassan, A.B., Jackson, D.A. & Cook, P.R. Visualization of replication factories attached to nucleoskeleton. *Cell* **73**, 361-373 (1993).
236. Jackson, D.A., Balajee, A.S., Mullenders, L. & Cook, P.R. Sites in human nuclei where DNA damaged by ultraviolet light is repaired: visualization and localization relative to the nucleoskeleton. *Journal of cell science* **107 (Pt 7)**, 1745-1752 (1994).
237. Cook, P.R. Predicting three-dimensional genome structure from transcriptional activity. *Nature genetics* **32**, 347-352 (2002).
238. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354 (2006).

239. Ragoczy, T., Bender, M.A., Telling, A., Byron, R. & Groudine, M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes & development* **20**, 1447-1457 (2006).
240. Melnik, S. et al. The proteomes of transcription factories containing RNA polymerases I, II or III. *Nat Methods* **8**, 963-968 (2011).
241. Faro-Trindade, I. & Cook, P.R. A conserved organization of transcription during embryonic stem cell differentiation and in cells with high C value. *Molecular biology of the cell* **17**, 2910-2920 (2006).
242. Bolzer, A. et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology* **3**, e157 (2005).
243. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
244. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).
245. Gondor, A., Rougier, C. & Ohlsson, R. High-resolution circular chromosome conformation capture assay. *Nature protocols* **3**, 303-313 (2008).
246. van de Werken, H.J. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods* **9**, 969-972 (2012).
247. Stadhouders, R. et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols* **8**, 509-524 (2013).
248. de Vree, P.J. et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature biotechnology* **32**, 1019-1025 (2014).
249. Fullwood, M.J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).
250. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309 (2006).
251. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* **30**, 90-98 (2012).
252. Hughes, J.R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics* **46**, 205-212 (2014).
253. Jager, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications* **6**, 6178 (2015).
254. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).
255. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
256. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research* **25**, 582-597 (2015).
257. Gennes, P.G.d. Scaling concepts in polymer physics. (Cornell University Press, Ithaca, N.Y.; 1979).
258. Bystricky, K., Heun, P., Gehlen, L., Langowski, J. & Gasser, S.M. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16495-16500 (2004).
259. Sachs, R.K., van den Engh, G., Trask, B., Yokota, H. & Hearst, J.E. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 2710-2714 (1995).
260. Munkel, C. & Langowski, J. Chromosome structure described by a polymer model. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **57 (5B), 5888–5896** (1998).
261. Knoch, T.A. Towards a holistic understanding of the human genome by determination and integration of its sequential and three-dimensional organization. *High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York* (2003).
262. Rauch, J. et al. Light optical precision measurements of the active and inactive Prader-Willi syndrome imprinted regions in human cell nuclei. *Differentiation; research in biological diversity* **76**, 66-82 (2008).
263. Mirny, L.A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **19**, 37-51 (2011).
264. Grosberg, A.Y., Nechaev, S.K. & Shakhnovich, E.I. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. France* **49**, 2095-2100 (1988).
265. Weidemann, T. et al. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. *Journal of molecular biology* **334**, 229-240 (2003).
266. Bohn, M., Heermann, D.W. & van Driel, R. Random loop model for long polymers. *Physical review. E, Statistical, nonlinear, and soft matter physics* **76**, 051805 (2007).
267. Orkin, S.H. & Zon, L.I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631-644 (2008).
268. Till, J.E. & Mc, C.E. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation research* **14**, 213-222 (1961).
269. Bryder, D., Rossi, D.J. & Weissman, I.L. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *The American journal of pathology* **169**, 338-346 (2006).
270. Majeti, R., Park, C.Y. & Weissman, I.L. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell stem cell* **1**, 635-645 (2007).
271. Copley, M.R. & Eaves, C.J. Developmental changes in hematopoietic stem cell properties. *Experimental & molecular medicine* **45**, e55 (2013).
272. 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* **126**, 5073-5084 (1999).
273. Muller, A.M., Medvinsky, A., Strouboulis, J., Grosveld, F. & Dzierzak, E. Development of hematopoietic stem cell activity in the mouse embryo. *Immunity* **1**, 291-301 (1994).
274. Boisset, J.C. et al. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464**, 116-120 (2010).
275. Gekas, C., Dieterlen-Lievre, F., Orkin, S.H. & Mikkola, H.K. The placenta is a niche for hematopoietic stem cells. *Developmental cell* **8**, 365-375 (2005).
276. Kumaravelu, P. et al. Quantitative developmental anatomy of definitive haematopoietic stem cells/long-term repopulating units (HSC/RUs): role of the aorta-gonad-mesonephros (AGM) region and the yolk sac in colonisation of the mouse embryonic liver. *Development* **129**, 4891-4899 (2002).
277. Morrison, S.J. & Scadden, D.T. The bone marrow niche for haematopoietic stem cells. *Nature* **505**, 327-334 (2014).

278. Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harbor perspectives in medicine* **3**, a011601 (2013).
279. Christensen, J.L., Wright, D.E., Wagers, A.J. & Weissman, I.L. Circulation and chemotaxis of fetal hematopoietic stem cells. *PLoS biology* **2**, E75 (2004).
280. Sturgeon, C.M., Ditadi, A., Awong, G., Kennedy, M. & Keller, G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. *Nature biotechnology* **32**, 554-561 (2014).
281. Xiong, J.W. Molecular and developmental biology of the hemangioblast. *Developmental dynamics : an official publication of the American Association of Anatomists* **237**, 1218-1231 (2008).
282. Murray, P. The development in vitro of the blood of the early chick embryo. *Proc R Soc Lond B Biol Sci* **11**, 497–521 (1932).
283. Sabin, F.R. Studies on the Origin of Blood-vessels and of Red Blood-corpuscles as Seen in the Living Blastoderm of Chicks During the Second Day of Incubation. *Contrib. Embryol* **9**, 213–262 (1920).
284. Doetschman, T.C., Eistetter, H., Katz, M., Schmidt, W. & Kemler, R. The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *Journal of embryology and experimental morphology* **87**, 27-45 (1985).
285. Keller, G. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & development* **19**, 1129-1155 (2005).
286. Itskovitz-Eldor, J. et al. Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Molecular medicine* **6**, 88-95 (2000).
287. Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J.C. & Keller, G. A common precursor for hematopoietic and endothelial cells. *Development* **125**, 725-732 (1998).
288. Nishikawa, S.I., Nishikawa, S., Hirashima, M., Matsuyoshi, N. & Kodama, H. Progressive lineage analysis by cell sorting and culture identifies FLK1+VE-cadherin+ cells at a diverging point of endothelial and hemopoietic lineages. *Development* **125**, 1747-1757 (1998).
289. Fehling, H.J. et al. Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. *Development* **130**, 4217-4227 (2003).
290. Huber, T.L., Kouskoff, V., Fehling, H.J., Palis, J. & Keller, G. Haemangioblast commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432**, 625-630 (2004).
291. Vogeli, K.M., Jin, S.W., Martin, G.R. & Stainier, D.Y. A common progenitor for haematopoietic and endothelial lineages in the zebrafish gastrula. *Nature* **443**, 337-339 (2006).
292. Lancrin, C. et al. The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature* **457**, 892-895 (2009).
293. de Bruijn, M.F., Speck, N.A., Peeters, M.C. & Dzierzak, E. Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo. *The EMBO journal* **19**, 2465-2474 (2000).
294. Jaffredo, T., Gautier, R., Brajeul, V. & Dieterlen-Lievre, F. Tracing the progeny of the aortic hemangioblast in the avian embryo. *Dev Biol* **224**, 204-214 (2000).
295. Chen, M.J., Yokomizo, T., Zeigler, B.M., Dzierzak, E. & Speck, N.A. Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature* **457**, 887-891 (2009).
296. Boisset, J.C. et al. Progressive maturation toward hematopoietic stem cells in the mouse embryo aorta. *Blood* **125**, 465-469 (2015).
297. Kissa, K. & Herbomel, P. Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature* **464**, 112-115 (2010).
298. Porcher, C. et al. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47-57 (1996).
299. Okuda, T., van Deursen, J., Hiebert, S.W., Grosveld, G. & Downing, J.R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84**, 321-330 (1996).
300. Swiers, G., Rode, C., Azzoni, E. & de Bruijn, M.F. A short history of hemogenic endothelium. *Blood cells, molecules & diseases* **51**, 206-212 (2013).
301. Yokomizo, T. et al. Runx1 is involved in primitive erythropoiesis in the mouse. *Blood* **111**, 4075-4080 (2008).
302. Eilken, H.M., Nishikawa, S. & Schroeder, T. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature* **457**, 896-900 (2009).
303. Antas, V.I., Al-Drees, M.A., Prudence, A.J., Sugiyama, D. & Fraser, S.T. Hemogenic endothelium: a vessel for blood production. *The international journal of biochemistry & cell biology* **45**, 692-695 (2013).
304. Medvinsky, A., Rybtsov, S. & Taoudi, S. Embryonic origin of the adult hematopoietic system: advances and questions. *Development* **138**, 1017-1031 (2011).
305. Adolfsson, J. et al. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295-306 (2005).
306. Kondo, M., Weissman, I.L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661-672 (1997).
307. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I.L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193-197 (2000).
308. Moore, A.J. & Anderson, M.K. Dendritic cell development: a choose-your-own-adventure story. *Advances in hematology* **2013**, 949513 (2013).
309. Palis, J. Primitive and definitive erythropoiesis in mammals. *Frontiers in physiology* **5**, 3 (2014).
310. McGrath, K.E. et al. A transient definitive erythroid lineage with unique regulation of the beta-globin locus in the mammalian embryo. *Blood* **117**, 4600-4608 (2011).
311. McGrath, K.E. & Palis, J. Hematopoiesis in the yolk sac: more than meets the eye. *Experimental hematology* **33**, 1021-1028 (2005).
312. Wojda, U., Noel, P. & Miller, J.L. Fetal and adult hemoglobin production during adult erythropoiesis: coordinate expression correlates with cell proliferation. *Blood* **99**, 3005-3013 (2002).
313. Liu, J. et al. Quantitative analysis of murine terminal erythroid differentiation in vivo: novel method to study normal and disordered erythropoiesis. *Blood* **121**, e43-49 (2013).
314. McGrath, K.E. et al. Enucleation of primitive erythroid cells generates a transient population of “pyrenocytes” in the mammalian fetus. *Blood* **111**, 2409-2417 (2008).
315. Yoshida, H. et al. Phosphatidylserine-dependent engulfment by macrophages of nuclei from erythroid precursor cells. *Nature* **437**,

754-758 (2005).
316. Johnstone, R.M., Bianchini, A. & Teng, K. Reticulocyte maturation and exosome release: transferrin receptor containing exosomes shows multiple plasma membrane functions. *Blood* **74**, 1844-1851 (1989).
317. Bennett, G.D. & Kay, M.M. Homeostatic removal of senescent murine erythrocytes by splenic macrophages. *Experimental hematology* **9**, 297-307 (1981).
318. Franco, R.S. Measurement of red cell lifespan and aging. *Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie* **39**, 302-307 (2012).
319. Perutz, M.F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **185**, 416-422 (1960).
320. Trimborn, T., Gribnau, J., Grosveld, F. & Fraser, P. Mechanisms of developmental control of transcription in the murine alpha- and beta-globin loci. *Genes & development* **13**, 112-124 (1999).
321. Stamatoyannopoulos, G. Control of globin gene expression during development and erythroid differentiation. *Experimental hematology* **33**, 259-271 (2005).

Chapter 2

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

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

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

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

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

Abstract

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

Introduction

In recent years, it has become evident that the 3D organization of genomes is not random. Numerous studies have implicated long-range chromosomal interactions in several crucial cellular processes, including the regulation of gene expression[1-4]. Indeed, chromatin coassociations mediated by chromatin looping provide a means by which distal enhancers communicate with their target genes and stimulate transcription[5-7]. 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[9-12], 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 IgK loci (Igk), revealing crucial enhancer-gene communications[16-18].

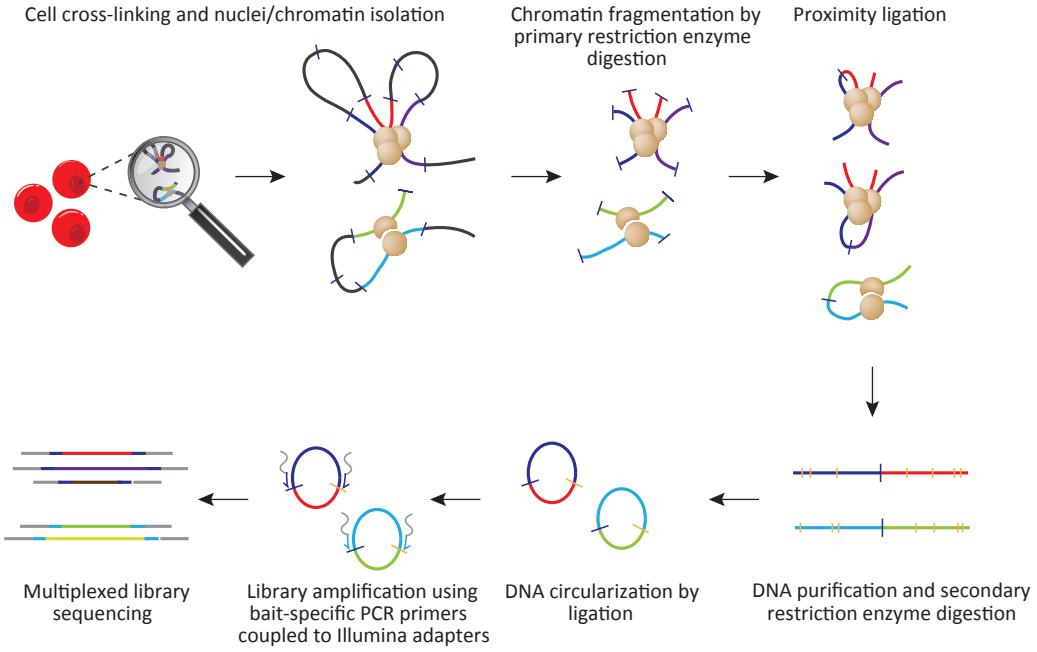

Figure 1: Overview of the multiplexed 3C-seq procedure. Nuclei from cross-linked cells are digested (primary restriction enzyme) and ligated under dilute conditions to physically link *in vivo* interacting DNA fragments. After a secondary digestion (secondary restriction enzyme) and ligation, inverse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

TABLE 1 | Comparison between different 3C variants.

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

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 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[9-12, 16-18]) 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 detection of unknown bait-specific interactions, but is limited by the number of

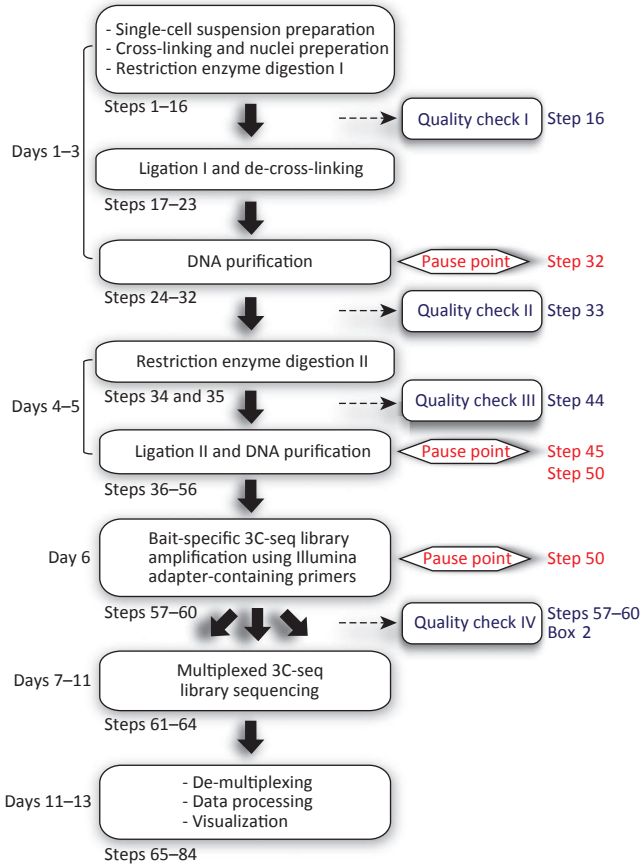

Figure 2: Flowchart of multiplexed 3C-seq data generation and processing. Steps involved in the multiplexed 3C-seq procedure are shown in blue rectangles. Time needed to complete these steps is depicted on the left. Pause points are indicated together with the timing of the different quality checkpoints: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

arrays needed to achieve genome-wide coverage and by the saturation of signals around the bait sequence, preventing the detection of medium- to close-range interactions (up to 200 kb away). The 5C variant[13] overcomes this limitation and offers the possibility of exploring every potential chromatin coassociation in large subchromosomal domains by using primer sets covering all possible interactions. It is, however, difficult to reach genome-wide coverage using 5C, as it requires extremely large numbers of primers for all possible intrachromosomal and interchromosomal interactions. HiC, in contrast, provides a global genome-wide analysis of all possible chromatin associations by coupling a modified 3C procedure to high-throughput sequencing[15]. Although it is extremely powerful, Hi-C requires substantial computational resources, and the number of sequence reads needed to obtain 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 (Figure 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 nonrepetitive 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

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

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

applied to detect interactions between promoter elements and the surrounding regions, or to connect distal enhancers to their target gene(s). With the recent developments in high-throughput chromatin occupancy profiling[21], large numbers of transcription factor binding and chromatin modification data sets are becoming available. Combined with this knowledge, 3C-seq can be used to analyze the functional relationships existing between regulatory elements, sites of active transcription, gene deserts or boundary elements where transitions in chromatin structure or transcription are observed (e.g., insulator elements or initiation sites for productive transcription elongation).

Limitations of 3C-seq

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

Experimental design

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

Box 1 | 3C-seq primer design

Two primers, a P5 primer and a P7 primer, need to be designed for each bait fragment of interest:
The P5 primer must be located as close as possible to the primary restriction enzyme site (usually the six-cutter). As only the sequence located after the restriction site is informative for identifying interacting fragments, the distance between the primary restriction enzyme primer and the restriction site itself should be minimized to ensure unambiguous alignment and identification of the interacting fragments (Fig. 3). This primer contains the P5 Illumina adapter sequence (5′-AATGATACGGCGACCACCGAACACTCTTTCCCTACAC GACGCTCTTCCGATCT-3′ to be placed upstream of the annealing sequence; Fig. 3) from which library sequencing will be initiated. The sequencing reaction starts from the bait fragment, reads through the annealing primer sequence and extends into the unknown captured fragment. To allow more flexibility for primer design and to ensure optimal alignment of the sequences, we use a 76-bp sequencing read length (Step 64).
The second primer, located near the secondary restriction enzyme site (the four-cutter), contains the P7 Illumina adapter sequence (5′-CAAGCAGAAGACGGCATACGA-3′, Fig. 3), and although it is required for the inverse PCR and the Illumina sequencing chemistry it is not sequenced (in contrast to paired-end sequencing, for which a different adapter is required). Therefore, the location of the P7 primer with regard to the secondary restriction site is more flexible (within 100 bp of the restriction site).
Actual primer requirements are similar to those used in standard PCR reactions. Oligo length is kept between 17 and 24 nt to facilitate efficient amplification and annealing temperatures are generally chosen between 54 and 59 °C. We regularly use primer design software (DNAMAN 5.0) to check these parameters and to ensure that primers are not prone to form dimers.

Box 2 | 3C-seq PCR setup and optimization

As 3C-seq library fragments differ in length and abundance, we use the Expand long template system to minimize any biases resulting from these differences[11]. Bait-specific primers (without adapters) are first tested for proper linearity and efficiency.
1. Test the increasing amounts of 3C-seq library DNA (up to 200 ng) using a 50-µl PCR. Reaction components and conditions are described in PROCEDURE Step 57.
2. Analyze PCR products on a 1.5% (wt/vol) agarose gel, where they should appear as a reproducible smear of DNA fragments, usually showing two prominent bands11. These prominent bands are the result of recircularization of the bait fragment in the first ligation step, and of detection of the neighboring fragment owing to incomplete digestion of the primary restriction site on the bait fragment[11].
3. Assess the linear range of the individual primer pairs by quantifying prominent bands in each reaction of the dilution range.
4. Order versions of the primer pairs that perform well, including the P5 and P7 Illumina adapter sequences (Box 1). Test these new primers as described in steps 1–3 of Box 2.
5. Use successful P5 and P7 primers to prepare 3C-seq samples for sequencing (PROCEDURE Steps 57–60).

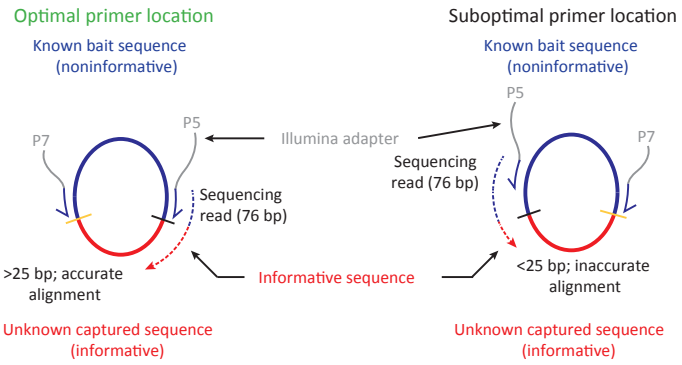

Figure 3: 3C-seq primer design and positioning. Schematic drawing of the location of the inverse PCR primers used to amplify a 3C-seq library. The ring represents a circular DNA molecule composed of the bait fragment (blue) ligated to an unknown captured fragment (red). The two PCR primers are located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

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 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

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).

in parallel (e.g., the same promoter for different biological conditions), small bar codes (2–6 nt) may be added to the primers (PROCEDURE Step 62; Box 3).

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

Materials

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

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

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

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

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

Caution: Formaldehyde is toxic.

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

Caution: Phenol/chloroform is toxic.

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

EQUIPMENT

- Cell strainer, 40 µm (BD Falcon, cat. no. 352340)
- Polypropylene centrifugation tubes (Greiner bio-one, cat. no. 188271)
- Safe-Lock 1.5-ml centrifugation tubes (Eppendorf, cat. no. 0030120.086)
- Thermomixer (Eppendorf, cat. no. EF4283)
- Water bath
- Microcentrifuge (Eppendorf, cat. no. 5417R)
- PCR thermocycler (MJ Research, cat. no. PTC-200)
- Spectrophotometer (NanoDrop 2000c, Thermo Scientific)
- Agilent 2100 Bioanalyzer (Aligent Technologies, cat. no. G2938C) with the 7500 DNA chip (cat. no. 5067-1506)
- Illumina HiSeq2000 high-throughput sequencing machine (Illumina)
- Excel spreadsheet software (Microsoft)
- Computer with a minimum of 8 Gb RAM and 1.5 Tb attached storage running a Linux distribution and the software listed above

REAGENT SETUP

- Complete protease inhibitor, EDTA free

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

- Lysis buffer

Prepare the following solution in Milli-Q H_2O: 10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% (vol/vol) NP-40 and 1× protease inhibitor solution.

Critical: Because protease inhibitors degrade quickly in solution, use freshly prepared lysis buffer for each new experiment.

PROCEDURE

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

1. Obtain single-cell preparations from fresh tissue, FACS-sorted cells or cell lines in 10% (vol/vol) FCS/PBS (see Table 2 for cell types successfully used by us in 3C-seq experiments). Tissues rich in extracellular matrix (e.g., brain) can be treated with collagenase (0.125% (wt/vol) in PBS; incubate the tissues for 30–60 min at 37 °C) first. Filter tissue-harvested cell preparations through a 40-µM cell strainer to obtain single-cell suspensions (see ref. 19). Determine cell concentrations and dilute 0.3×10^6 to 10×10^6 cells (10×10^6 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).

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).

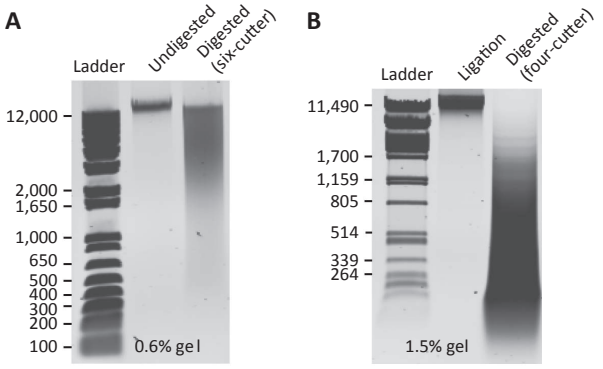

Figure 4: (a) Agarose gel (0.6%, wt/vol) on which an aliquot of undigested (left lane) and digested (right lane) sample (primary restriction digestion, Step 16) was run. A six-cutter was used, showing a typical smear of DNA fragments (a majority of DNA fragments residing between the 12 kb and 4 kb marker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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- *Troubleshooting*

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

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

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

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

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

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

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

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

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].

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

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 trouble shooting 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
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]

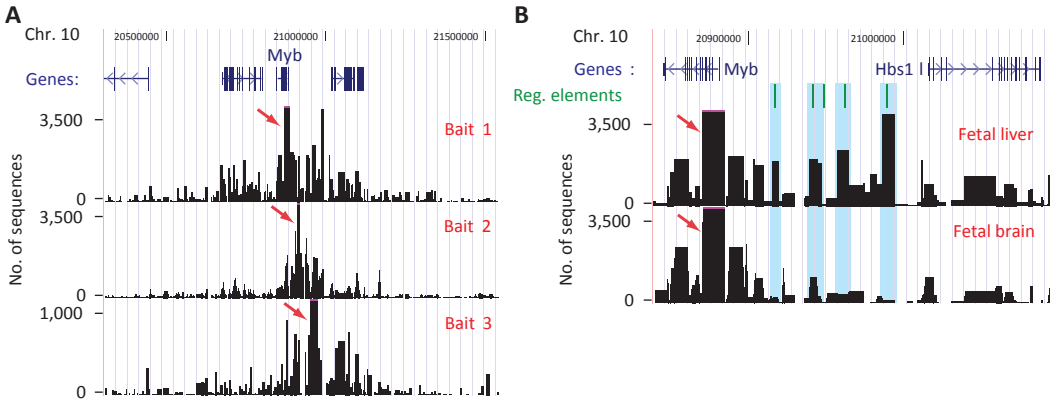

Figure 5: Typical interaction profiles obtained from a multiplexed 3C-seq experiment. (a) 3C-seq interaction profiles in mouse fetal liver cells shown for three bait fragments in the Myb locus17 (1.2-Mb region shown). Bait signals are depicted by an arrow. (b) 3C-seq interaction profiles generated from 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.

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

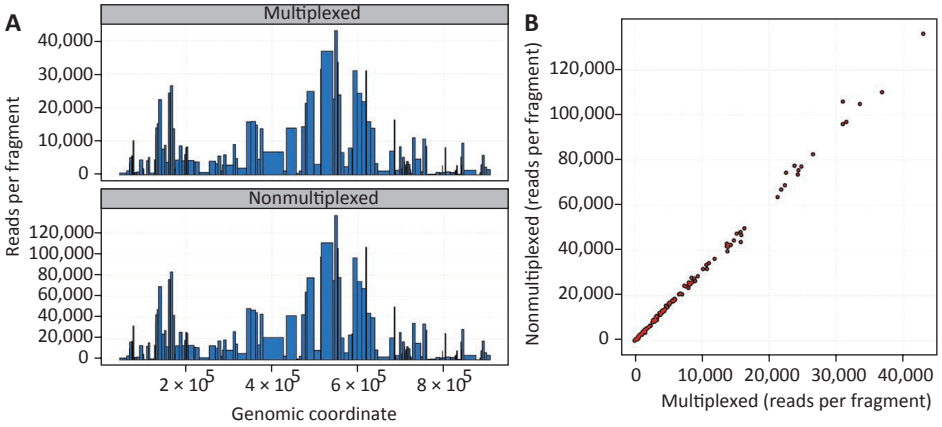

Figure 6: Comparison of interactions detected for the same 3C-seq sample after single or multiplexed library sequencing. (a) Interaction profiles around the bait fragment for a 3C-seq sample after multiplexed (top) or nonmultiplexed (bottom) library sequencing, showing highly similar profiles. (b) Scatter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

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 reads30) 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 picture[18]. Interactions found in trans (generally about 40–50% of the reads) often show low interaction frequencies and appear to be randomly scattered around the genome. Trans-interaction signals therefore need to be interpreted with caution, as their reproducibility may appear questionable in a number of cases. However, several studies have begun to probe their functional relevance in specific cases, in particular in light of chromosomal translocations, and showed correlation between physical proximity and sites of recombination, indicating that physical proximity in trans may be relevant[31, 32].

Multiplexing 3C-seq samples greatly increases the technique's throughput and results in a substantial cost reduction. Even though the total number of reads is lower in a multiplexed sample compared with a nonmultiplexed sample, interaction patterns remain almost identical (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.

Acknowledgments

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

Supplementary information

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

Contributions

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

References

1. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376-380 (2012).
2. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381-385 (2012).
3. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* 489, 109-113 (2012).
4. Splinter, E. & de Laat, W. The complex transcription regulatory landscape of our genome: control in three dimensions. *EMBO J* 30, 4345-4355 (2011).
5. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144, 327-339 (2011).
6. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* 12, 283-293 (2011).
7. Stadhouders, R. et al. Transcription regulation by distal enhancers: who's in the loop? *Transcription* 3, 181-186 (2012).
8. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1306-1311 (2002).
9. Gondor, A., Rougier, C. & Ohlsson, R. High-resolution circular chromosome conformation capture assay. *Nat Protoc* 3, 303-313 (2008).
10. Sexton, T. et al. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nat Protoc* 7, 1335-1350 (2012).
11. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38, 1348-1354 (2006).
12. van de Werken, H.J. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* 9, 969-972 (2012).
13. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* 2, 988-1002 (2007).
14. Fullwood, M.J. et al. An oestrogen-receptor-alpha-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 Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa 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* 58, 192-203 (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 U S A* 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).

Chapter 3

Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements.

Petros Kolovos[1], Harmen J. G. van de Werken[1], Nick Kepper[2], Jessica Zuin[1], Rutger W.W. Brouwer[3], Christel E.M. Kockx[3], Kerstin S. Wendt[1], Wilfred F.J. van IJcken[3], Frank Grosveld[1]†, Tobias A. Knoch[1]†

[1]Department of Cell Biology, Erasmus MC, Dr. Molewaterplein 50, 3015GE Rotterdam, the Netherlands
[2]Deutsches Krebsforschungszentrum (DKFZ) & BioQuant, Im Neuenheimer Feld 280, Heidelberg 69120, Germany
[3]Center for Biomics, Erasmus MC, Dr. Molewaterplein 50, 3015GE Rotterdam, the Netherlands

†Corresponding authors.

Published in:
Epigenetics & Chromatin
2014; 16;7:10

Abstract

Background

Significant efforts have recently been put into the investigation of the spatial organization and the chromatin-interaction networks of genomes. Chromosome conformation capture (3C) technology and its derivatives are important tools used in this effort. However, many of these have limitations, such as being limited to one viewpoint, expensive with moderate to low resolution, and/or requiring a large sequencing effort. Techniques like Hi-C provide a genome- wide analysis. However, it requires massive sequencing effort with considerable costs. Here we describe a new technique termed Targeted Chromatin Capture (T2C), to interrogate large selected regions of the genome. T2C provides an unbiased view of the spatial organization of selected loci at superior resolution (single restriction fragment resolution, from 2 to 6 kbp) at much lower costs than Hi-C due to the lower sequencing effort.

Results

We applied T2C on well-known model regions, the mouse *β-globin* locus and the human *H19/IGF2* locus. In both cases we identified all known chromatin interactions. Furthermore, we compared the human *H19/IGF2* locus data obtained from different chromatin conformation capturing methods with T2C data. We observed the same compartmentalization of the locus, but at a much higher resolution (single restriction fragments *vs.* the common 40 kbp bins) and higher coverage. Moreover, we compared the *β-globin* locus in two different biological samples (mouse primary erythroid cells and mouse fetal brain), where it is either actively transcribed or not, to identify possible transcriptional dependent interactions. We identified the known interactions in the *β-globin* locus and the same topological domains in both mouse primary erythroid cells and in mouse fetal brain with the latter having fewer interactions probably due to the inactivity of the locus. Furthermore, we show that interactions due to the important chromatin proteins, LDB1 and CTCF, in both tissues can be analyzed easily to reveal their role on transcriptional interactions and genome folding

Conclusions

T2C is an efficient, easy, and affordable with high (restriction fragment) resolution tool to address both genome compartmentalization and chromatin-interaction networks for specific genomic regions at high resolution for both clinical and non-clinical research.

Background

A number of recent studies have shown that the genome is organized in self-associating domains[1] that are separated by linker regions. These so-called "topological domains" or "topological associated domains" generally range from 300 kilobasepairs (kbp) to 1 megabasepairs (1Mb) and consist of a series of different types of chromatin loops in agreement with earlier models of the genome ([2] and references therein).

One loop is defined as two distant chromatin regions coming, spatially, into close proximity (interact with each other), thereby creating DNA loops. Such "long-range interactions" have been first observed between promoters and distant enhancers ([3, 4] and references therein) and can bring DNA-elements together that are separated by a large distance on the linear DNA strand ([5, 6] and references therein). These regulatory elements (enhancers or silencers) are short sequences containing several binding sites for transcription factors, which regulate the activation (reviewed in [7]) repression (reviewed in [8]) genes and their subsequent transcription (reviewed in [9]). In the linear genome the distance between enhancer(s) and gene can be quite large, for example, the sonic hedgehog (shh) enhancer is located about 1 Mb away from its target gene *Shh*[10]. Changes or differences within these elements and their interaction with genes can be responsible for changes in gene expression[11], causing intrinsic differences between individuals, disease susceptibility and disease progression.

A number of chromatin loops are thought to be purely structural, that is to enable the folding of the genome creating distinct topological domains, while other loops have a function in the expression of genes. Loops of the latter type are frequently found within topological domains, but are less frequently observed between different topological domains[1, 12]. These regulatory chromatin loops change and depend on a large number of proteins including CTCF[13], cohesin[14] and a series of transcription factors[15-18], which are mostly involved in the transcriptional regulation of genes within the domain.

The recent refinements of the genome structure were largely due to the Chromosome Conformation Capture (3C) technique which allowed the rapid identification of chromatin regions residing in close proximity[19, 20]. The basic principle of the 3C technique is that segments, which are spatially in close proximity within the cell nucleus, can be tethered together by cross-linking. After cross-linking and restriction enzyme digestion of the genome, the proximal segments remain covalently linked and segment-ends can be, subsequently, ligated in dilute conditions. The ligation products can be analyzed using PCR-based methods[19]. A number of different 3C-type techniques have been developed to answer different biological questions including: 3C/3C-qPCR[19, 21, 22], 3C-seq/4C-seq[23, 24], 4C (3C-on-a chip)[25-27], Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)[28], 5C (3C carbon copy)[29] and

Table 1 Comparison between different chromatin conformation capturing techniques (adopted and modified from[23])

Method	Applications	Advantages	Limitations
3C-qPCR	One-to-one	Simple analysis	Laborious, requires knowledge of the locus and proper controls
3C-seq/4C-seq	One-to-all	Good resolution, good signal-to-noise ratio	Restricted to single viewpoint per experiment when multiplexing several viewpoints, analysis requires extra bioinformatics expertise, not an all-to-all genome-wide method
3C-on-chip (4C)	One-to-all	Relatively simple data analysis	Poor signal-to-noise ratio, difficult to obtain genome-wide coverage
5C	Many-to-many	Identifies interactions between many individual fragments	Very laborious, no genome-wide coverage, primer design can be challenging. Analysis requires advanced bioinformatics expertise
Hi-C	All-to-all	Explores the genome-wide interactions between all individual fragments	Very expensive, requires a large sequence effort to obtain sufficient coverage, approximately 10 to 40 kbp resolution, requires advanced bioinformatics expertise
T2C	Many-to-all	Explores the interactome of a selected region in cis but also in trans, high (restriction fragment) resolution, cheaper than Hi-C and 5C, requiring only half a lane of Illumina HiSeq2000	Is restricted to the selected regions ofthe genome, requires advanced bioinformatics expertise

Hi-C[30]. All these techniques have their own advantages and limitations (Table 1) and have provided very valuable information on chromosomal interactions and gene transcription mechanisms[20, 25, 30, 31]. 3C and 4C are quite work and cost intensive, given that they are only one-to-one fragment and one-to-all fragment techniques respectively. Prior knowledge of the locus is necessary to define the region of interest.

The analysis of the interactions of several viewpoints with the aforementioned techniques in 3C and 4C is possible, but the choice for several viewpoints will increase the costs and work effort linearly. However, the number of viewpoints can also be limited due to the (often) limiting amount of available cell material. 5C is demanding in primer design and allows the analysis of interactions only among the primer designed fragments. Furthermore, genome-wide coverage is not possible. Hi-C is very expensive as it requires extremely deep sequencing in order to cover the whole genome, even at a relatively low resolution of 40 kbp. The most recent Hi-C data analysis has used a new algorithm and provided a genome-wide interaction map of 10 kbp resolution. However, an enormous amount of sequencing is required (3.4 billion mapped paired-end reads from six biological replicates)[32]. Such effort is not affordable for most research groups and, in addition, the scientific interest is most of the time focused on a specific question involving a limited set of specific loci or domains. Hence, there is a need for a technique which eliminates most of the aforementioned limitations.

Here we present Targeted Chromatin Capture (T2C), a new 3C method, which does not involve a massive sequencing effort, but which results in a high resolution map of interactions for particular loci of interest. We used the well-studied human *H19/IGF2* locus and compared the results of our new method with data from other chromatin conformation capturing techniques. Using the mouse *β-globin* locus we demonstrated that the method can reliably identify chromatin structural changes between different tissues and also allows the study of the role of individual transcription factors in the chromatin architecture.

Overview of the procedure

To overcome the aforementioned problems of the 5C and Hi-C techniques we have developed the novel method T2C. The method has the advantage that it allows the analysis of the structure of the genome and all the interactions of selected regions of the genome at high resolution (single restriction fragments) without a massive sequencing effort and associated costs.

T2C employs a selective enrichment of the 3C ligation products in preselected regions of interest in order to identify their interactions within a domain as well as the compartmentalization of one or several specific regions of the genome. These regions can be continuous Mb sized genomic regions, but could also be a collection of smaller regions (a few kbp each). Every captured restriction fragment can be used as a single “4C-seq viewpoint” and analyzed accordingly. The results of T2C provide a local interaction map at a restriction fragment-level resolution accompanied with a lower sequencing effort and less intricate bioinformatics analysis than Hi-C. T2C also overcomes the limits of 5C since it identifies not only interactions within the targeted region(s), but also interactions between the targeted region(s) and with regions outside of them.

In brief, we have designed sets of unique oligonucleotide probes (ranging from 62 to 90 nucleotides) specific for all the restriction fragments and as close as possible to the end of the first restriction site (*Mm* – HindIII + NlaIII digest, *Hs* – BglII + NlaIII digest) in our regions of interest, the mouse *β-globin* locus and the human *H19/IGF2* locus (see Methods). Alternative to continuous regions, separate genomic regions within one (or more) chromosomes could be analyzed simultaneously. The oligonucleotides are spotted on an array or can alternatively be captured on beads. Some fragment ends cannot be captured by a designed oligonucleotide due to the presence of repeat elements or the insufficient size of the restriction fragment end. Repetitive sequences are a general problem in all 3C based methods, including Hi-C. The size limitation of the fragment end can be circumvented if necessary by a backup

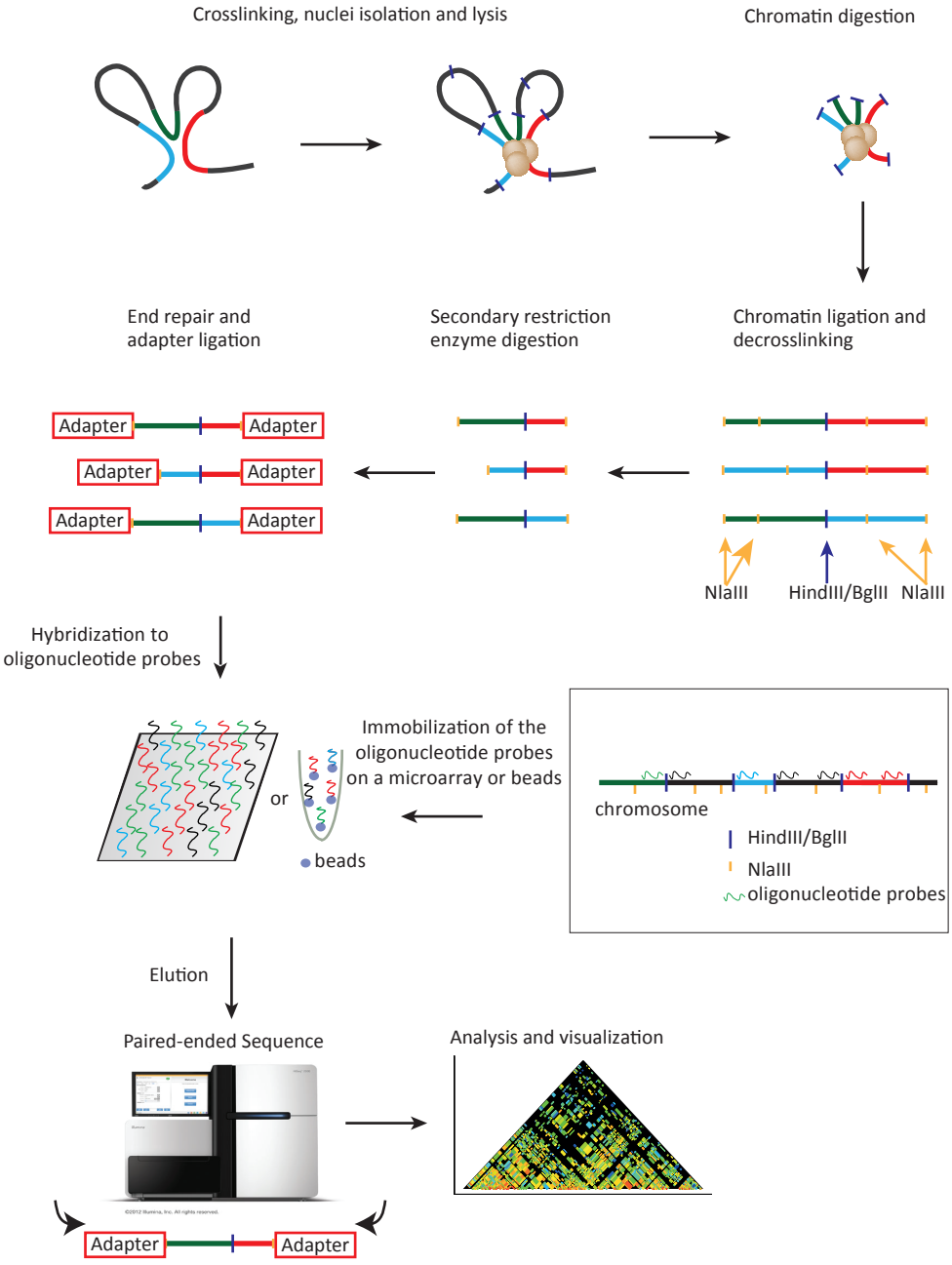

Figure 1. Overview of the Targeted Chromosome Capture (T2C) procedure.
Isolated cross-linked chromatin is digested with a restriction enzyme (dark blue lines) and ligated under diluted conditions to favour ligations between restriction fragments that are spatially in proximity. After de-cross-linking and a secondary digestion (orange lines), the overhangs are repaired followed by adapter ligation. Different address sequences can be used in the adapters for different samples to allow multiplexing of different samples (hybridisation of different samples to the same set of oligonucleotides). The resulting library is hybridized to a set of unique oligonucleotides on an array or oligonucleotides in solution that are captured on beads. The unique oligonucleotides (green, red, black and blue lines) are located as close as possible to the first restriction site. The hybridized DNA, which contains the library of all interactions from the selected area of the genome, is eluted and is pair-end sequenced on an Illumina HiSeq2000 followed by bioinformatic analysis

Table 2 Summary ofinformation about the different experiments

Type	Genome assembly version	Coordinates oligo-nucleotide positions	Size of area of interest (Mb)	Median resolution (kbp)	Raw paired reads (n)	Paired reads that could be mapped to the whole genome (n)	Mapped paired reads between the region ofinterest and the whole genome (n)	Uniquely mapped paired-reads in the whole genome without self-ligation and and non-digestion (n)	Uniquely mapped paired-reads between the region ofinterest and the whole genome without self-ligation and non-digestion (n)	Uniquely mapped paired-reads inside the region of interest without self-ligation and non-digestion (n)	'Interactions' inside the region of interest (n)	Average number of reads/interaction in the region of interest (n)
Mouse fetal liver	mm9	chr7: 109876329-111966581	2.1	2	65,165,916	9,300,108	5,716,401	4,559,952	2,723,515	557,763	4,057	137
Mouse fetal brain	mm9	chr7: 109876329-111966581	2.1	2	84,977,143	6,380,256	3,191,360	3,018,169	1,414,128	271,177	2,369	114
HB2	hg18	chr11: 1100646 - 3173091	2.1	4.1	51,952,969	13,813,662	12,127,051	5,503,770	4,745,779	1,929,245	8,989	215

Table 2: Summary of information about the different experiments. Columns from left to right: Tissue type or cells; genome assembly version; summary of the positions of oligonucleotides (region of interest); the size and the median resolution of the area under investigation; the number of the raw paired-reads (before alignment, *i.e.* all reads from the sequenator); the number of mapped paired reads that could be mapped back to the whole genome; the number of paired reads between the region of interest (fragments with oligonucleotides) and the whole genome; the number of uniquely mapped paired-reads in the whole genome after removal of the self-ligation and non-digestion events (See Methods); the number of uniquely mapped paired-reads between the region of interest (fragments with oligonucleotides) and the whole genome after removal of the self-ligation and non-digestion events; the number of uniquely mapped paired-reads inside the region of interest after removal of the self-ligation and non-digestion events; the number of "interactions" between fragments in the region of interest; average number reads per interaction. The capture efficiency and purification (enrichment) by hybridization is high (*i.e.* how many reads from the region of interest ("specific" reads) are found when compared to total reads *i.e.* the reads that are from other areas of the genome and not containing a sequence from the area of interest ("non-specific" reads)). We find that the "specific" reads represent 61%, 50% and 88% of total reads ("specific" plus "non-specific") for mouse primary erythroid cells, mouse fetal brain cells and HB2 respectively, including the self-ligation and non-digestion events. By removing those events those numbers change to 60%, 47% and 86% respectively. This means for example that 60% of the fetal liver reads (2723515) represent $2.1x10^6$ bases (the region of interest) while the remaining 40% of reads represents 3.10^9 bases (the whole genome), numbers that indicate a high level of enrichment by the hybridization step.

procedure with different enzymes (changing either the first or the second restriction enzyme or both), which generates a new set of end fragments or by mechanically shearing of the chromatin (instead of the second restriction enzyme digestion) which can result in fragment sizes of different length (see discussion).
The first steps of the preparation of the chromatin conformation capturing library are carried out as in 3C-seq[23]. Basically, chromatin is cross-linked, followed by digestion with a 6 bp recognition restriction endonuclease, ligation in diluted conditions and de-cross-linking of the DNA. The library is subsequently digested with a frequently-cutting 4 bp recognition restriction endonuclease or mechanically sheared to obtain small fragments containing the ligation site, followed by end-repair and ligation of an adapter. Within the adapter, different barcodes can be included that would allow multiplexing of different samples. The resulting library is hybridized to the specific oligonucleotide probe set representing the area(s) of interest (either on an array or in a bead capturing procedure) to enrich specifically for the interacting fragments of the region of interest (all fragments positioned at the ends of the original 6 bp-cut fragments after the second 4bp-cut and eliminate all fragments internal to the 6 bp generated fragments). After extensive washing all ligation products including regions covered by the targeting-array are eluted and their sequence determined by Illumina-sequencing (Figure 1). The capture efficiency (the proportion of paired reads of total reads when at least one read of the paired end reads is located on a fragment represented by an oligonucleotide) is between 47% and 86% depending the cell type and the region (see Table 2).

Results

T2C identifies known long-range interactions

We first have chosen the *H19/IGF2* region on human chromosome 11 to test and compare the method to other 3C-methods. Previously, we analyzed the 3D-structure of the locus by 3C to study the role of cohesin and CTCF for chromosomal long-range interactions[33] and also generated 4C-seq data[14] (Figure 2). Hi-C interaction maps were retrieved for IMR90 cells[1].
We selected unique oligonucleotides mapping near the ends of 344 BglII generated fragments spanning 2.1 Mb around the *H19/IGF2* locus (Table 2). This set of 525 oligonucleotides was spotted on a capture array. A ligation fragment library was generated from the breast endothelial cell line 1-7HB2 (abbreviated HB2) after digestion with BglII and NlaIII according to the 3C-seq protocol [23] (see also Figure 1). The library was subsequently hybridized to the capture array. After elution from the capture array the captured DNA fragments were amplified by a PCR with low cycle number (12 cycles) and sequenced by paired-end Illumina sequencing (see Methods).
To demonstrate that T2C reveals a similar overall interaction pattern and compartmentalization of the locus as observed by Hi-C in IMR90 cells[1] we first binned the paired-reads into 40 kbp bins (Figure 2A, B). The interaction patterns at this level of resolution show that the topological domain is maintained between different cell types, HB2[14] versus IRM90[1] with a Spearman's rank correlation coefficient r_s =0.64 ($P < 2.2x10^{-16}$).
However, with T2C we obtained a chromatin interaction map at restriction fragment resolution (Figure 2C, each block represents one restriction fragment), revealing significantly more detail with respect to the general chromatin organization of the region when visualized by a logarithmic and rainbow-like coloured interaction frequency. To first validate T2C in comparison to 3C and 4C-seq we compared the interactions of a single restriction fragment (CTCF AD viewpoint[33]) to interactions detected for this fragment by 3C[33] and 4C-seq[14] (Figure 2D, E, F). Although there are some variations in the read coverage of the individual interactions, similar interactions can be observed by both 4C-seq and T2C. Moreover, both methods detect interactions which we previously observed with 3C[33]. It should be noted that an important difference between 4C-seq and T2C is the number of PCR amplification cycles.

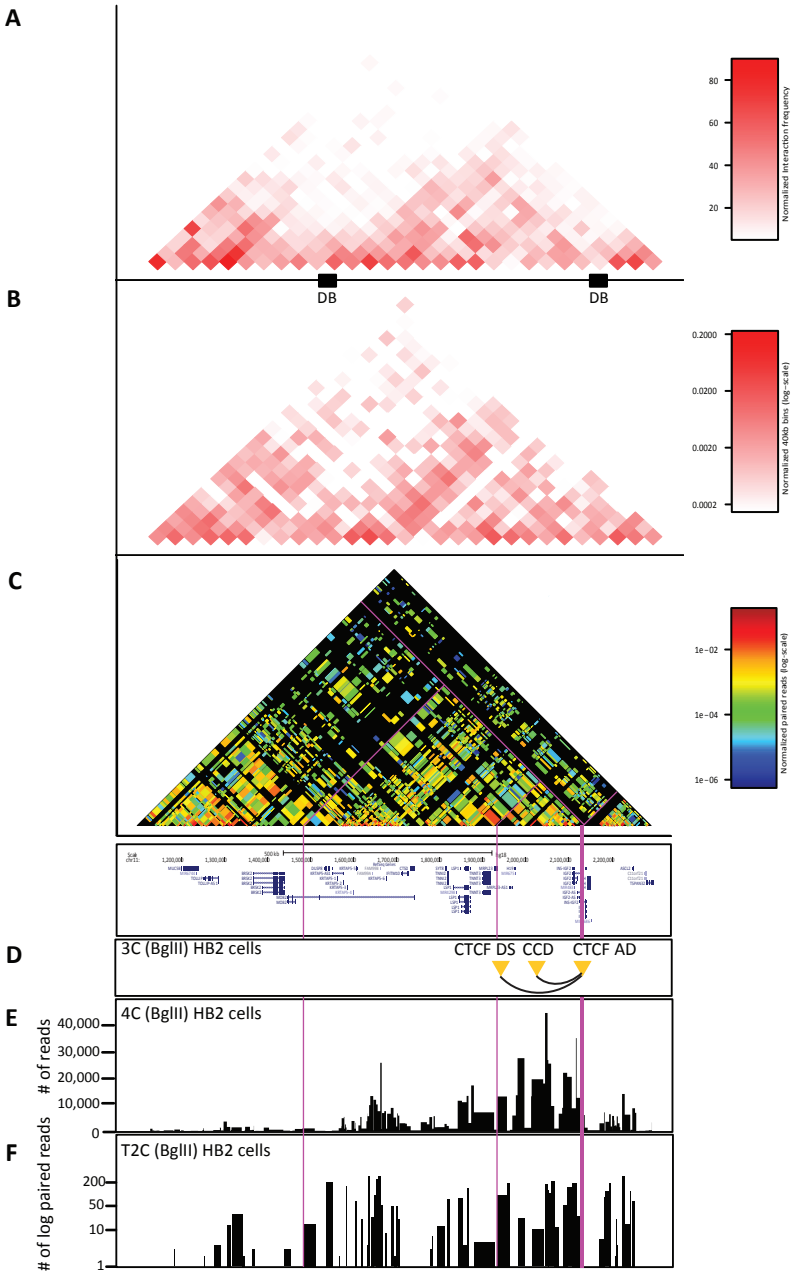

Figure 2. Comparison of interactions detected by T2C for the human chr11p15.5 region with Hi-C and 4C-seq.
(**A**) Hi-C data generated by Dixon *et al.*, for IMR90 cells covering the *H19/IGF2* region of interest, presented at a resolution 40 kbp with their respective domain boundaries (DB) depicted as black boxes[1]. (**B**) T2C interactions in HB2 cells at a 40 kbp resolution. The overall topological domain pattern observed by the two methods is similar (r_s =0.65, $P < 2.2\times10^{-16}$). (**C**) T2C interaction with their actual resolution at restriction fragment level. (**D**) Interactions detected by 3C[33]. The restriction fragments are indicated with yellow triangles. (**E**) 4C-seq interaction data[14], for a viewpoint close to the *IGF2* gene. (**F**) Interactions observed for a particular viewpoint by T2C plotted with logarithmic y-axis. The position of the viewpoint is indicated as bold pink line to allow a direct comparison between the methods. The thin pink lines indicate a couple of interaction fragments for ease of comparison.

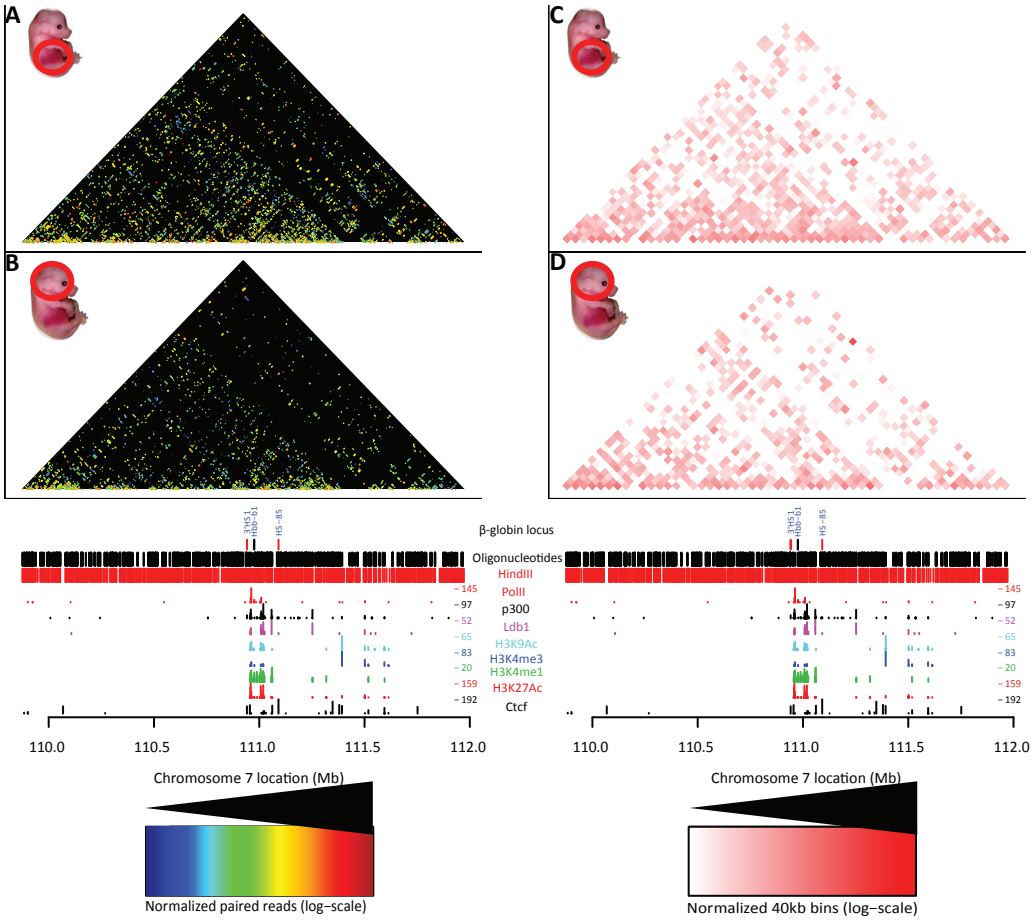

Figure 3. Comparison of the compartmentalization and interactions for the *β-globin* locus.
T2C performed in a 2.1 Mb region around the *β-globin* locus for mouse primary erythroid cells (**A**) and mouse fetal brain cells (**B**) from E12.5 mice. The topological domain patterns between different biological materials are identical and are independent of the number of interactions. Analysis of the interactions obtained with T2C obtained from mouse primary erythroid cells (**C**) and mouse fetal brain cells (**D**) were plotted at 40 kbp resolution to compare T2C to the regular Hi-C binning. The overall topological domain pattern is similar in the two tissues. All the T2C interactions are normalized to the same color code (see color inset). The bottom tracks show a linear representation of the *β-globin* locus, the oligonucleotides probes positions (black lines), *Hin*dIII recognition sites (red lines) and the ChIP-seq derived binding sites of PolII (red lines), LDB1 (purple lines)[38], CTCF (black lines), p300 (black lines) and various histone modification markers (light blue, dark blue, green and red)[37] in mouse erythroleukemia cells.

For T2C this is on average 12 cycles (only after capture) whereas for 4C-seq it is 30 cycles. The lower number of cycles will give less PCR bias of the different fragments relative to each other, because fragments have different PCR efficiencies.
We conclude that the T2C method yields interaction data at a resolution identical to 4C-seq for the individual restriction fragments (median approximately 4 kbp resolution) and that when T2C is performed for a continuous region over 2 Mb it can reproduce the overall topological domain structure that was observed by Hi-C.

T2C identifies different interaction networks based on different biological materials

Next we used the extensively characterized mouse *β-globin* locus as a model system to show that the T2C method can detect reliably conformational changes due to activation of the genes *in vivo* at high resolution (Figures 3 and 4). We further showed, with an intersection between ChIP-seq derived chromatin protein data and T2C, that chromatin proteins may be involved in forming or maintaining the 3D structure of the genome (Figure 5).
The mouse *β-globin* locus undergoes structural changes upon activation in erythroid tissue[20, 34, 35], but is surrounded by silent olfactory receptor genes, which are only expressed in the olfactory epithelium. The major difference between the *H19/IGF2* locus and the *β-globin* locus is that the *β-globin* locus is embedded in a large area of inactive genes. Thus two patterns of interactions may be expected in erythroid cells, those important for the globin locus and those present in inactive chromatin. We selected a region of 2.1 Mb around the locus (Table 2) containing 719 restriction fragments of the restriction enzyme HindIII (6 bp recognition site). About 800 oligonucleotide probes were designed close to the ends of the fragments. To analyze the locus in its active state we used primary erythroid cells from fetal liver which were compared to fetal brain cells as a model of inactive loci. Based on results from previous 3C studies of the locus[20, 35] we expected in primary erythroid cells a higher number of interactions around the *β-globin* gene and between the *β-globin* gene and its regulatory elements.
The analysis of the hybridised fragments shows that almost the entire 2.1Mb appears to be part of one topological domain (with two possible sub-domains, one of which contains the *β-globin* locus) with the next domain starting near the end of the selected sequences (due to the repetitive sequences and the borders of the region of interest, that topological domain cannot be depicted clearly, in agreement with Dixon et al.,[1]) both in mouse primary erythroid cells (Figure 3A, right hand side) and mouse fetal brain cells (Figure 3B) with many interactions within the topological domain (Figure 3C and 3D). Although the topological domain structure between the different biological materials is similar, there appear to be less interactions in mouse fetal brain cells relative to mouse primary erythroid cells due to the inactivity of the locus in the brain (Figure 3). Focusing on the *β-globin* region, all the well-known interactions in the *β-globin* locus are detected in the primary erythroid cells. The known interactions, such as between the *β-globin* promoter and Locus Control Region (LCR) (Figure 4B, adapted and modified from Drissen *et al.*[16], with blue line depicting the interactions for primary erythroid cells and with grey the interactions for mouse fetal brain cells) and between the LCR-3'HS1 are clearly visualized[16, 20, 35] (Figure 4A). These interactions are absent from the fetal brain sample (Figure 4C). Furthermore, the main regulatory region (HS1-6) shows the well-known interaction with the *β-globin* genes and HS1 at the 3' end of the locus in fetal liver cells but not in brain[16, 20]. In addition, for the *β-globin* promoter we identify a few additional interactions further away than the ones previously reported. These are located even approximately 1Mb far from the *β-globin* promoter (Figure 3A). It is unknown whether these interactions are related to the functioning of the *β-globin* genes or whether these DNA elements are in close proximity due to the folding of the domain, although their absence in the fetal brain suggests they have a role in the regulation of the globin *β-globin*. In addition to the interactions *in cis*, the *β-globin* (*Hbb-b1*) gene and the LCR also contact a number of positions on other chromosomes.

T2C in combination with ChIP-seq identifies factor specific interactions

We also compared the interactions of the binding sites of an important regulatory transcription factor in mouse primary erythroid cells, the LDB1 complex, and the insulator binding protein CTCF (Figure 5A-D). LDB1 is highly enriched on the *β-globin* locus and its LCR in mouse primary erythroid cells compared to fetal brain cells[36]. By visualizing only the restriction fragments containing the LDB1 or CTCF binding sites as determined by ChIP-seq in fetal liver derived mouse erythroleukemia cells (MEL)[37, 38],

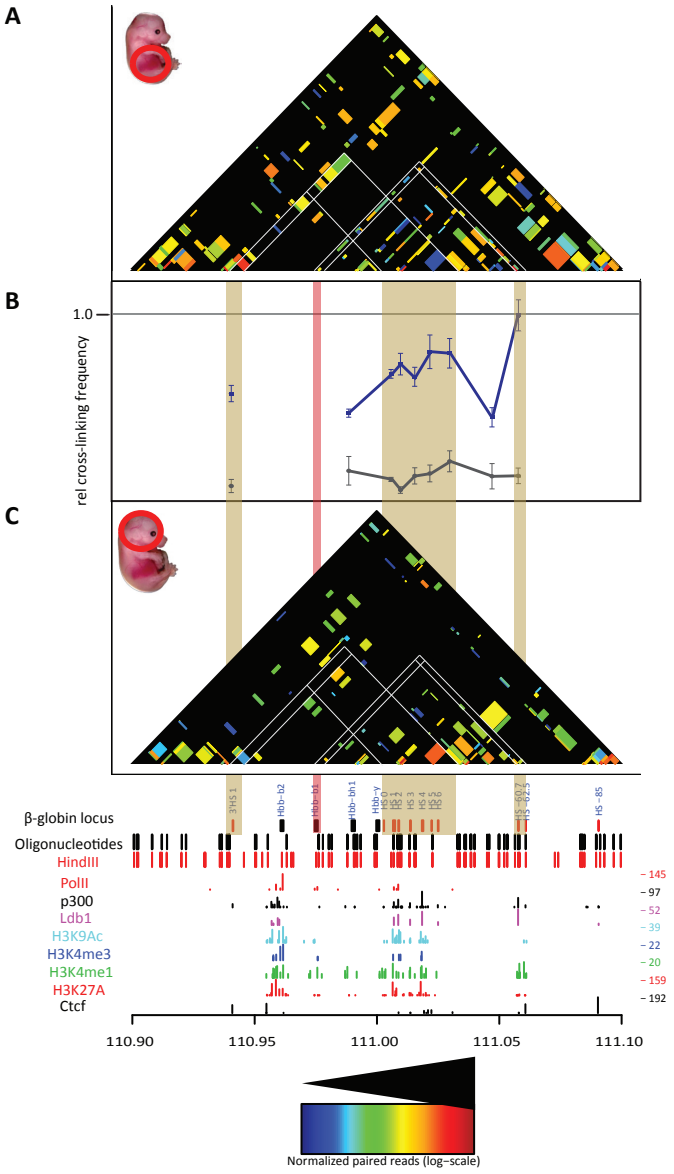

Figure 4. Comparison of T2C with 3C-qPCR for the *β-globin* promoter

T2C for mouse primary erythroid cells (**A**) and mouse fetal brain cells (**C**) from E12.5 mice, revealed the same interactions from the *β-globin* promoter when comparing them to 3C-qPCR (**B**). The 3C-qPCR was adapted and modified from Drissen *et al.*[16] with blue line depicting the interactions for primary erythroid cells and with grey the interactions for mouse fetal brain cells from E12.5 mice. White lines indicate the areas of particular interest (such as 3'HS1, *β-globin* promoter, Locus Control Region (LCR) and 5' HS-60/-62) in the *β-globin* locus. Interactions between LCR, the *β-globin* promoter and the 3'HS1 are lost in mouse brain cells. The shaded vertical bars indicate the comparison between the different panels. The red vertical bar indicates the *β-globin* promoter. All the T2C interactions are normalized to the same color code (see color inset).The bottom tracks show a linear representation of the *β-globin* locus, the oligonucleotides probes positions (black lines), *Hin*dIII recognition sites (red lines) and the ChIP-seq derived binding sites of PolII (red lines), LDB1 (purple lines)[38], CTCF (black lines), p300 (black lines) and various histone modification markers (light blue, dark blue, green and red)[37] in mouse erythroleukemia cells.

we can immediately deduce in which interactions the LDB1 complex (Figure 5E, F) or CTCF (Figure 5G, H) are involved. In addition, we can identify the restriction fragments that represent gene promoter fragments (by Histone 3 Lysine 4 trimethylation (H3K4me3)) or enhancer fragments (marked by H3K4me1, that is in the LCR, HS-60 and −62.5) or neither of these, by plotting the histone modifications ChIP-seq profiles[37]. Interestingly the 3'HS1 and HS-85 belong to the latter class and have robust CTCF but not LDB1 binding sites. This suggests that they are "structural" elements which would fit with the observation that the deletion of the 3'HS1 results in a loss of looping but not in a decrease of *β-globin* mRNA levels[13]. In contrast the enhancer immediately 3' of the *β-globin* enhancer is apparent, but it does not appear to interact with any distal elements. It is also clear that in mouse primary erythroid cells LDB1 (Figure 6A) and CTCF (Figure 6B) occupy restriction fragments that have more interactions with other positions in the locus when compared to mouse brain cells. In addition the median distance on the linear chromosome between two fragments in spatial proximity is larger in primary erythroid

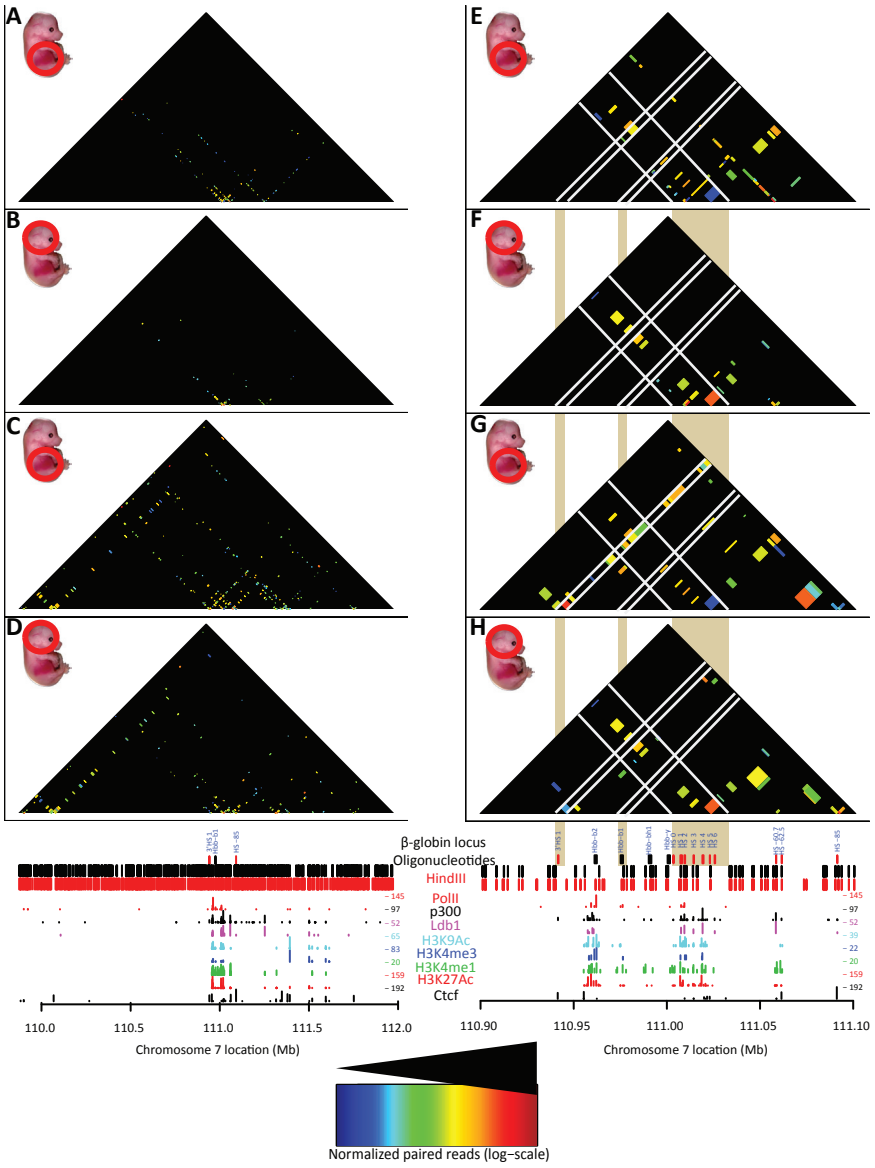

Figure 5: T2C/ChIP-seq intersection plot.
A comparison of the interactions containing one or two fragments with a LDB1 or CTCF binding site. Interactions are plotted, at restriction fragment resolution, over a 2.1 Mb region around the *β-globin* locus for LDB1 (**A**, **B**) or CTCF (**C**), (**D**) for mouse primary erythroid cells (**A**, **C**) and mouse fetal brain cells (**B**, **D**) from E12.5 mice. The topological sub-domain around the *β-globin* locus is clearly depicted in the mouse primary erythroid cells when compared to mouse brain cells. Focusing on the *β-globin* locus, T2C-intersection plots, at restriction fragment resolution, of interactions that contain a LDB1 bound fragment (**E**, **F**) or a CTCF bound fragment (**G**, **H**), for mouse primary erythroid cells (**E**, **G**) and mouse brain cells (**F**, **H**). White lines indicate particular areas of interest (like 3'HS1, the *β-globin* promoter and the Locus Control Region (LCR)) in the *β-globin* locus. The mouse primary erythroid cells interactions between LCR, *β-globin* promoter and 3'HS1 are lost in mouse brain cells. The shaded vertical bars, indicate the comparison between the different panels. All the interactions are normalized to the same color code (see color inset). The bottom tracks show a linear representation of the *β-globin* locus, the oligonucleotides probes positions (black lines), *Hin*dIII recognition sites (red lines) and the ChIP-seq derived binding sites of PolII (red lines), LDB1 (purple lines)[38], CTCF (black lines), p300 (black lines) and various histone modification markers (light blue, dark blue, green and red)[37] in mouse erythroleukemia cells.

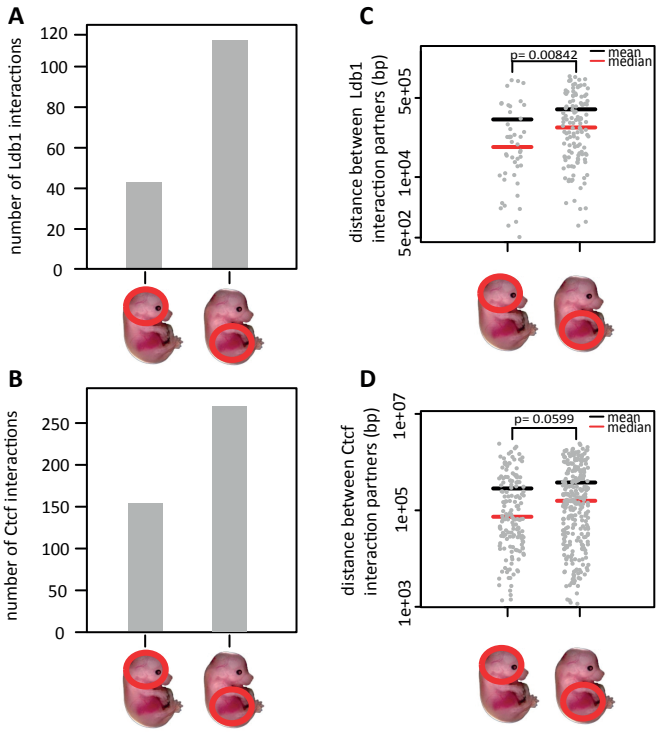

Figure 6: The mean, median and the number of T2C interactions for the LDB1 or CTCF containing fragments.
The number of LDB1 (**A**) and CTCF (**B**) interactions is lower in mouse fetal brain when compared to primary erythroid cells. Furthermore, the mean and the median of the distance between either LDB1 (**C**) or CTCF (**D**) interaction partners is lower in mouse fetal brain cells when compared to mouse primary erythroid cells. *P* -values were calculated using the Mann-Whitney *U* test.

cells for both LDB1 (Figure 6C) and CTCF (Figure 6D) binding sites. This suggests that this area of the genome is less condensed. We conclude from these experiments that T2C indeed detects topological domains and the different interactions between and within domains. These interactions depend on the expression status of the genes such as the active *β-globin* locus in primary erythroid cells versus the same silent locus in fetal brain. In addition, the high level of resolution of the interactions allows novel observations such as shown for the *β-globin* locus LDB1 and CTCF binding sites and immediately shows which of these binding sites interact with each other and where they are positioned on the linear genome.

Discussion

The importance of the role of chromatin interactions in the regulation of the gene transcription is well established[9, 39-42]. However, there is still an increasing need for a quick, easy and affordable technique to provide the information on chromatin interactions and the compartmentalization of the genome. T2C is affordable to most scientific groups and will meet in a satisfactory manner their needs for detecting high resolution chromatin organization of selected loci. Every restriction fragment can serve as a 'viewpoint' and all their interactions, either short or long or to other chromosomes (data not shown), can be identified. Thus multiple 3C-seq, 4C-seq or 5C experiments do not have to be performed. Moreover, with T2C the compartmentalization of the genome can be identified in the regions of interest without requiring the large sequencing effort of Hi-C, which would increase the costs tremendously. Furthermore, due to the T2C design, a better coverage and resolution of the locus is obtained when compared to other genome wide techniques (like Hi-C and 3C with its derivatives) using a 6 bp cutter as first restriction enzyme. Here we multiplexed two samples, but by multiplexing more than two samples the costs are likely to be reduced significantly without sacrificing the quality of the output. We have recently successfully used 13 samples per sequencing lane, including the *β*-globin locus which showed the same interactions (data not shown).

The resolution of T2C is based on the restriction enzyme used. Digesting cross-linked chromatin from primary erythroid cells and HB2 cells with HindIII or BglII, resulted in a median resolution of 2 kbp and 4.1 kbp respectively (Table 2). That provides a significantly better resolution than the usual 40 kbp bins obtained with Hi-C. Furthermore, comparing T2C with 4C-seq[14] and Hi-C[1] for the *H19/IGF2* locus (Figure 2) and with already published 3C-qPCR data for the *β-globin* locus[16, 20, 35], the same topological domains and chromatin interaction networks were identified. Taken together, all these results, reveal

the strengths of the T2C as a tool to identify all the interactions and the compartmentalization of specific regions of the genome.

In addition, the T2C interactions are easily connected to the factors that play a role in these interactions or the type of elements (promoters/enhancers) involved in the interactions. LDB1 and CTCF are important proteins which mediate chromatin interactions. LDB1 is an important transcription factor necessary for primitive mouse hematopoiesis and for the development of megakaryocytes[43, 44] and controls essential hematopoietic pathways in mouse early development[45]. Depletion of *Ldb1* is lethal for mouse embryos after E9.5 with severe effects such as impairment of hematopoietic and vascular development[46]. It is well established that the LCR has higher interaction frequencies with the *β-globin* locus in mouse primary erythroid cells comparing to mouse brain cells[16, 20, 35] and that LDB1 is significantly enriched in the LCR region in mouse primary erythroid cells relative to mouse fetal brain cells[36] (Figure 5E *vs*. Figure 5F). Furthermore, CTCF is an insulator binding protein known to be involved in chromatin conformation[33] and is enriched at the boundaries of topological domains[1]. CTCF mediates long range interactions in the *β-globin* locus[13] (Figure 5C *vs*. Figure 5D and Figure 5G *vs*. Figure 5H). Hence, it is no surprise that for LDB1 and CTCF occupied restriction fragments we observe a higher number of interacting fragments at larger linear distances of fragments that interact in mouse primary erythroid cells than in mouse brain cells (Figure 6). This effect can be explained by the fact that the *β-globin* locus is active in mouse primary erythroid cells. Furthermore, we observe that the boundaries of the topological domain, which contains the *β-globin* locus, are easily observed in mouse erythroid cells (Figure 3A). That is prominent when depicting only the CTCF interacting fragments (Figure 5C *vs*. Figure 5D). Furthermore, the number of interactions within that topological domain, appear higher in the erythroid cells comparing to fetal brain cells (Figure 3A *vs*. Figure 3B, Figure 6A, B). We hypothesize that this is due to the fact that the *β* -globin locus is active with open chromatin in mouse primary erythroid cells. Hence, the chromatin has a different conformation by enabling the interaction between many different elements necessary for the regulation of the gene[34]. However, in mouse fetal brain cells, where *β-globin* locus is not active, that is not necessary and there are no important elements that need to spatially be in close proximity.

The method may be improved by bringing the cost further down. For example each of the *β-globin* locus experiments was carried out by using one sequencing lane on an Illumina HiSeq machine for each different biological sample (mouse primary erythoid cells and mouse fetal brain cells). That yielded after comprehensive data analysis and 271.177 and 557.763 paired-reads within the limits of the region of interest excluding self-ligations and uncut fragments for both fetal brain and liver (see Methods). These reads represented 2.369 and 4.057 distinct interactions with 114 and 137 reads per interaction on average for fetal brain and liver, respectively (Table 2). The read frequency of the highest 20% of the interactions is from 11858 to 202 in fetal liver and from 29637 to 188 (the top 30% is from 11858 to 123 and 29637 to 120 for fetal liver and fetal brain, respectively). The bottom 20% account for 4 reads in both tissues (while 30% account for 9 and 13 for fetal liver and fetal brain, respectively). The question then becomes whether one could do more samples per lane (that is a reduction in cost per sample) which would result in fewer reads per interaction point. The decision on this depends to some extent on the research question asked. Analysis of functional interactions and/or the “rough” overall structure of a locus, can be achieved by using a range between 1/2 and 1/13 of a sequencing lane which will dramatically lower the costs without losing much information.

We also considered using mechanical shearing instead of a secondary restriction enzyme. The advantage of the secondary restriction enzyme over mechanically shearing is that it is very reproducible and provides a better repair step of the ends and hence ligation of the adapters. The possible disadvantage of the second cleavage would seem to be a loss of fragment, because a number of fragments would be represented by one or no oligonucleotide. However when the oligonucleotides are used in excess, as in T2C, there is virtually no statistically significant difference in detecting the reads of fragments represented by two, one, or no oligonucleotides (Figure 7). Mechanically shearing would have the advantage that

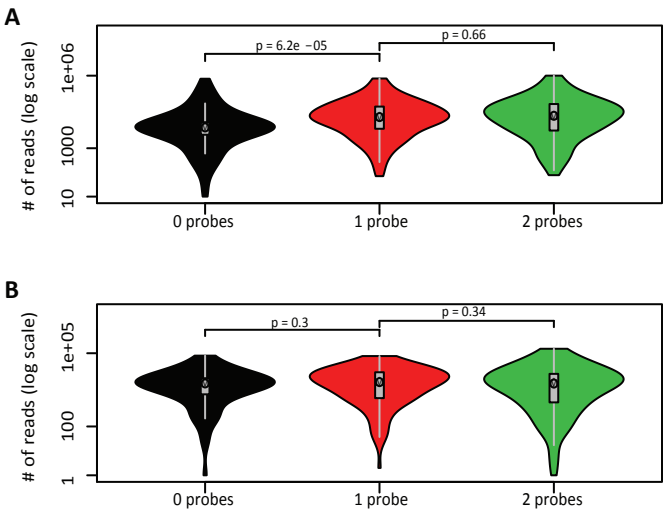

Figure 7: Comparison of capture efficiencies
The efficiency with which each fragment of the selected area is captured was derived from counting all of the reads for any particular fragment, that is all its interactions, its self-ligation and non-cleaved material and plotting these against the presence of two, one or no oligonucleotides (probes) in the fragment (**A**). This shows that the presence of one or two oligonucleotides does not make a difference in the capture as would be expected under conditions where the oligonucleotides are in saturation. When no oligonucleotides are present for a particular fragment, the number of reads will be lower, because the reads due to self-ligation cannot be captured. When the reads are corrected for the self-ligation and non-cleaved fragments this difference largely disappears (**B**). *P*-values were calculated using the Mann-Whitney *U* test

the chance of capturing a fragment is improved, because some of the secondary restriction sites are too close to the primary restriction sites. However the disadvantages are that mechanically shearing is random, which will have the same possible loss addressed above, but more importantly mechanical shearing is difficult to standardize between different laboratories. Using two different sets of oligonucleotides in combination with two different restriction enzymes for the first or second cleavage would give the most advantage because fewer fragments would be lost and the overall resolution and coverage would be further improved.
The "quantification" could be further improved by spiking the samples with control cells preferably from another species, to allow easy recognition of the spike when mapping the sequences back to the genome during the analysis of the ligated fragments. This would also require the addition of a spike specific set of capturing oligonucleotides. Spiking the sample with a DNA sample with a different address sequence at the amplification and sequencing stage of the procedure would also be an improvement, although it would be less quantitative than the spiking with cells at the start of the procedure. The normalization of the signals using the capture efficiency of each of the fragments (Figure 7) also increases the "quantification", although it should be noted these are all relative numbers rather than a real quantification because a number of parameters cannot be controlled or assessed properly.
Because T2C is focused on particular regions of interest, it would be easy to design a set of oligonucleotides for a number of loci that are known to be associated with a particular disease and design a diagnostic kit on that basis that could handle many samples at the same time. Since SNPs are often linked to diseases, dedicated oligonucleotides for them can be designed in order to assess their effect in long range interactions and the regulation of the gene transcription. For non-clinical research purposes the size of the region used in our experiments is sufficient (more than 2 Mb) to extract safe conclusions about the local chromatin interactome and the compartmentalization of the genome.

Conclusions

We conclude that T2C can be used as an affordable, cost-effective diagnostic tool with single restriction fragment resolution to explore the local spatial organization of the genome and chromatin interactions without requiring laborious procedures or massive sequencing efforts.

Methods

Oligonucleotide design

A microarray for the *β-globin* locus was designed containing unique oligonucleotides and physically as close as possible to the HindIII restriction sites spanning 2.1 Mb around the gene (chr7: 109876329-111966581, mm9). For the *H19/IGF2* locus unique oligonucleotides were designed close to BglII restriction sites (chr11:1100646-3173091, hg18) spanning an area of 2.1 Mb (**Table 2**).
The oligonucleotides were designed with the following criteria, they should be: (1) as close as possible to the first restriction site; (2) a unique DNA sequence within the area of interest and preferably in the entire genome; (3) similar melting temperatures, but with different base composition and the length; (4) oligonucleotides which exceed the second restriction site due to very small end fragments, were trimmed keeping in mind to stay close to the same melting temperature.
A custom-made NimbleGen Sequence Capture 2.1M capture array is produced separately for the H19/IGF2 locus and for the *β-globin* locus containing for each one the oligonucleotides which satisfy the aforementioned criteria. The oligonucleotides, 525 for the *H19/IGF2* locus and 800 for the *β-globin* locus, were replicated proportionally and equally up to 2.1M in total for each design, *i.e.* for the *β-globin* locus each of the 800 oligonucleotides was spotted in 2625 spots.

Chromatin isolation and library preparation

Nuclei from approximately 10^7 mouse primary erythroid cells from mouse fetal liver E12.5, mouse fetal brain cells E12.5 and a human breast endothelial cell line (HB2) were isolated, cross-linked (in 2% formaldehyde at room temperature) quenched with 1M glycine and were re-suspended in lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% (vol/vol) NP-40 and 1× protease inhibitor solution). The chromatin was digested with a 6-cutter (400 units of *Hin*dIII for mouse cells and *Bgl*II for the HB2 cells) and ligated using 100 units of T4 DNA ligase (Promega) under conditions favouring intramolecular ligation events. After reversing the crosslink at 65°C overnight, 50 µg of the resulting DNA chromatin library were digested with a frequent 4-cutter (*Dpn*II or *Nla*III for the mouse cells, *Nla*III for the HB2 cells, at a DNA concentration of 100 ng/µL, using 1 unit of enzyme per µg of DNA). All these steps were performed according to the initial steps of 3C-seq protocol, as described previously[23].
The final library is prepared for analysis on the Illumina Cluster Station and HiSeq 2000 Sequencer according to the Illumina TruSeq DNA protocol with modifications (www.illumina.com). In short, the digested library is purified using AMPure XP beads (Beckman Coulter), end-repaired and cleaned using AMPure XP beads. The now blunt-ended fragments were A-tailed using the Klenow exo enzyme in the presence of ATP and purified again using AMPure XP beads. Then indexed adapters provided by Illumina were ligated to the A-tailed DNA fragments with subsequent purification using AMPure XP beads.

Array capturing

The resulting adapter-modified DNA library (300-500 ng) was hybridized in 35 µL for 64 hours at 42°C on a custom-made NimbleGen Sequence Capture 2.1M capture array according to NimbleGen Sequence Capture array protocol (www.nimblegen.com/seqcapez) on the NimbleGen Hybridization System. The captured DNA fragments are eluted from the capture array and purified using MinElute columns (Qiagen). The yield for a positive region (a fragment inside the region of interest) and a negative region (a fragment outside the region of interest) differ by >30 fold on average. The captured DNA fragments are amplified by 12 PCR cycles. PCR products are purified using AMPure XP beads and eluted in 30 µL of re-suspension buffer. One microliter is loaded on an Agilent Technologies 2100 Bioanalyzer using a

DNA 1000 assay to determine the library concentration and to check for quality.

Cluster generation and high throughput sequencing

Cluster generation is performed according to the Illumina Cluster Reagents preparation protocol (www.illumina.com). Briefly, 1 µL of a 10 nM TruSeq DNA library stock DNA is denatured with NaOH, diluted to 9-10 pM and hybridized onto the flowcell. The hybridized fragments are sequentially amplified, linearized and end-blocked according to the Illumina Paired-end Sequencing user guide protocol. After hybridization of the sequencing primer, sequencing by synthesis is performed using the HiSeq 2000 sequencer with a 101 cycle protocol according to manufacturer's instructions. The sequenced fragments were denatured with NaOH using the HiSeq 2000 and the index-primer was hybridized onto the fragments. The index was sequenced with a seven-cycle protocol. The fragments are denatured with NaOH, sequentially amplified, linearized and end-blocked. After hybridization of the sequencing primer, sequencing-by-synthesis of the third read is performed using the HiSeq 2000 sequencer with a 101-cycle protocol.

Targeted Chromatin Capture data analysis.

The generated HiSeq 2000 sequencing reads were trimmed if the reads contained the first enzyme restriction recognition site (*Hin*dIII for the mouse derived reads and *Bgl*II for the human derived reads) For each read with one or more enzyme recognition sites, the DNA sequence after the 3' end of the first site was removed, that is after the trimming procedure the trimmed reads contained and ended with a single restriction recognition site. Subsequently, consecutive bases with a quality score lower than 10 were cut off from the ends of all the reads and the reads that contained less than 12 bases were omitted using Trimmomatic[47]. We used the Burrows-Wheeler Alignment tool (BWA, version 0.6.1) to the whole genome NCBI36/hg18 assembly for the human derived reads and to NCBI37/mm9 assembly for the mouse derived reads, using default settings[48]. Aligned reads that localized between two second enzyme recognition sites that did not contain a first enzyme recognition site, that is all *Nla*III-*Nla*III restriction fragments were removed using BEDtools [51].
In the alignment, paired reads were removed if one of the reads was not uniquely mapped. Furthermore, paired reads that were a result of a self-ligation event, non-digestion/re-ligation event or a ligation of identical ends were removed from the analysis, since these paired reads introduce a common bias in chromosome conformation capture techniques[49, 50]. The alignments were further processed with SAMtools[48] to generate paired-end Binary Alignment/Map (BAM) files. BEDtools[51] was used to remove reads that overlapped more than one restriction fragment. Interaction matrices were generated from the alignments at a resolution of the restriction fragments and at 40 kb resolution (using BEDtools on a 40 kbp binned genome). In addition, the human T2C 40 kb binned data were compared to IMR90 40 kb Hi-C data of the combined replicates[1]. The T2C interaction plots were normalized for capture efficiency of the fragments. For each interaction the number reads of each interaction was normalized through dividing it by the sum of the reads of both fragments involved in the interaction. Similarly, the T2C plots of the 40 kb bins were normalized after all the fragments were divided into 40 kb bins along each chromosome. ChIP-seq and T2C interaction-intersection plots were generated from normalized T2C interaction plots and intersected with fragments that contained a ChIP-seq peak signal of the protein of interest. The statistical software package R (version 3.1.0) was used to generate the interaction plots and to conduct the statistical calculations[52].

ChIP-seq analysis
Published ChIP-seq datasets[37, 38] were obtained and analyzed. MACS[53] was used to identify peaks (fdr≤0.01, peak height≥20 overlapping reads) to intersect their positions with the interacting fragments

obtained from T2C.

Authors' contributions: PK, KSW, TAK, designed the experiments. PK, JZ, and KSW carried out the experiments. HJGvdW performed the bioinformatics analysis. NK and RWWB conducted the initial steps of bioinformatics analysis. CEMK and WFJvI carried out the Illumina sequencing. P.K., HJGvdW and F.G. wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments: We thank the members of the FG Laboratory, Argyris Papantonis, Robert-Jan Palstra and Danny Huylebroeck for discussions and reading the manuscript. PK was supported by grants from EpiGenSys/ERASysBio + /FP7 (NL: NWO, UK: BSRC, D: BMBF). JZ was supported by an NWO ALW grant and KSW by E-RARE/TARGET-CdLS (NL: ZonMW). NK and the grid infrastructure were supported by the BMBF (grant #01AK803A (German MediGRID), and #01IG07015G (Services@MediGRID). HJGvdW was supported by Zenith (93511036) grant from the Netherlands Genomics Initiative (NGI). The work was supported by EpiGenSys/ERASysBio + /FP7 (NL: NWO, UK: BSRC, D: BMBF), the Bluescript EU Integrated Project, the Netherlands Institute for Regenerative Medicine (NIRM), the MEC Booster grant by the Netherlands Genomics Institute (MEC Booster grant).

Accession number: The accession number is SRP042002.

References

1. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
2. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301 (2001).
3. Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. & Grosveld, F. The effect of distance on long-range chromatin interactions. *Mol Cell* **1**, 131-139 (1997).
4. Mueller-Storm, H.P., Sogo, J.M. & Schaffner, W. An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell* **58**, 767-777 (1989).
5. Jhunjhunwala, S. et al. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**, 265-279 (2008).
6. Medvedovic, J. et al. Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. *Immunity* **39**, 229-244 (2013).
7. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).
8. Maeda, R.K. & Karch, F. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Curr Opin Genet Dev* **21**, 187-193 (2011).
9. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5**, 1 (2012).
10. Maas, S.A. & Fallon, J.F. Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. *Dev Dyn* **232**, 345-348 (2005).
11. Lin, Y.C. & Murre, C. Nuclear location and the control of developmental progression. *Curr Opin Genet Dev* **23**, 104-108 (2013).
12. Splinter, E. et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25**, 1371-1383 (2011).
13. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-2354 (2006).
14. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001 (2014).
15. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).
16. Drissen, R. et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* **18**, 2485-2490 (2004).
17. Hou, C., Dale, R. & Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* **107**, 3651-3656 (2010).
18. 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).
19. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).
20. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453-1465 (2002).
21. Hagege, H. et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* **2**, 1722-1733 (2007).
22. Naumova, N., Smith, E.M., Zhan, Y. & Dekker, J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods* **58**, 192-203 (2012).
23. Stadhouders, R. et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection

of long-range chromatin interactions. *Nat Protoc* **8**, 509-524 (2013).
24. van de Werken, H.J. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* **9**, 969-972 (2012).
25. 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).
26. Sexton, T. et al. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nat Protoc* **7**, 1335-1350 (2012).
27. Gondor, A., Rougier, C. & Ohlsson, R. High-resolution circular chromosome conformation capture assay. *Nat Protoc* **3**, 303-313 (2008).
28. Fullwood, M.J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).
29. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* **2**, 988-1002 (2007).
30. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
31. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309 (2006).
32. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* (2013).
33. Nativio, R. et al. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet* **5**, e1000739 (2009).
34. van de Corput, M.P. et al. Super-resolution imaging reveals three-dimensional folding dynamics of the beta-globin locus upon gene activation. *J Cell Sci* **125**, 4630-4639 (2012).
35. Palstra, R.J. et al. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* **35**, 190-194 (2003).
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* **28**, 810-822 (2007).
37. Consortium, E.P. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
38. Soler, E. et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**, 277-289 (2010).
39. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413-417 (2007).
40. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800 (2007).
41. Simonis, M. et al. High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nat Methods* **6**, 837-842 (2009).
42. Stadhouders, R. et al. Transcription regulation by distal enhancers: who's in the loop? *Transcription* **3**, 181-186 (2012).
43. Li, L. et al. A requirement for Lim domain binding protein 1 in erythropoiesis. *J Exp Med* **207**, 2543-2550 (2010).
44. Li, L. et al. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat Immunol* **12**, 129-136 (2011).
45. Mylona, A. et al. Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis. *Blood* **121**, 2902-2913 (2013).
46. Mukhopadhyay, M. et al. Functional ablation of the mouse Ldb1 gene results in severe patterning defects during gastrulation. *Development* **130**, 495-505 (2003).
47. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* (2014).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
49. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
50. van de Werken, H.J. et al. 4C technology: protocols and data analysis. *Methods Enzymol* **513**, 89-112 (2012).
51. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
52. R-Core-Team R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/* (2013).
53. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

Chapter 4

TNFα signalling primes chromatin for NF-κB binding and induces rapid and widespread nucleosome repositioning

Sarah Diermeier[1,9*], Petros Kolovos[2,6*], Leonhard Heizinger[3], Uwe Schwartz[1], Theodore Georgomanolis[4], Anne Zirkel[4], Gero Wedemann[5], Frank Grosveld[2], Tobias A Knoch[6,7], Rainer Merkl[3], Peter R Cook[8], Gernot Längst[1†] and Argyris Papantonis[4,8†]

[1]Department of Biochemistry III, University of Regensburg, Regensburg, Germany.
[2]Cell Biology and Genetics, Erasmus Medical Center, , The Netherlands.
[3]Institute of Biophysics and Physical BiochemistryRegensburg, Germany.
[4]Centre for Molecular Medicine, University of Cologne, Cologne, Germany.
[5]Institute for Applied Computer Science, University of Applied Sciences Stralsund, Germany.
[6]Biophysical Genomics, Erasmus Medical Center, The Netherlands.
[7]BioQuant & German Cancer Research Center Heidelberg, Germany.
[8]Sir William Dunn School of Pathology, University of Oxford,, United Kingdom.
[9]Present address: Cold Spring Harbor Laboratory, USA.

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

Published in:
Genome Biology
2014; 15(12):536

Abstract

Background

The rearrangement of nucleosomes along the DNA fiber profoundly affects gene expression, but little is known about how signalling reshapes the chromatin landscape, in three-dimensional space and over time, to allow establishment of new transcriptional programs.

Results

Using micrococcal nuclease treatment and high-throughput sequencing, we map genome-wide changes in nucleosome positioning in primary human endothelial cells stimulated with tumour necrosis factor alpha (TNFα) - a proinflammatory cytokine that signals through nuclear factor kappa-B (NF-κB). Within 10 min, nucleosomes reposition at regions both proximal and distal to NF-κB binding sites, before the transcription factor quantitatively binds thereon. Similarly, in long TNFα-responsive genes, repositioning precedes transcription by pioneering elongating polymerases and appears to nucleate from intragenic enhancer clusters resembling super-enhancers. By 30 min, widespread repositioning throughout megabase pair-long chromosomal segments, with consequential effects on three-dimensional structure (detected using chromosome conformation capture), is seen.

Conclusions

Whilst nucleosome repositioning is viewed as a local phenomenon, our results point to effects occurring over multiple scales. Here, we present data in support of a TNFα-induced priming mechanism, mostly independent of NF-κB binding and/or elongating RNA polymerases, leading to a plastic network of interactions that affects DNA accessibility over large domains.

Background

The arrangement of nucleosomes along the chromatin fibre profoundly affects genome function[1, 2]. For example, silenced genomic segments and constitutive heterochromatin contain nucleosomes positioned in high-density arrays[1, 3, 4], whereas active and regulatory regions appear more disorganized and 'open'[1, 5, 6]. Although some data exist on the reorganization of the nucleosomal landscape following extra-cellular signalling[7, 8] and differentiation[9, 10], the temporally resolved dynamics of chromatin architecture remain poorly characterized.

Nucleosome positioning can be mapped genome-wide at single-nucleosome resolution using micrococcal nuclease digestion followed by sequencing (MNase-seq)[11, 12]. We applied this technique to primary human umbilical vein endothelial cells (HUVECs) stimulated with tumour necrosis factor alpha (TNFα). This potent cytokine drives the inflammatory response by signalling through the transcription factor nuclear factor kappa-B (NF-κB)[13, 14]; on phosphorylation, NF-κB translocates into nuclei, where it regulates hundreds of genes[15, 16]. Therefore, we correlated nucleosomal repositioning with genome-wide NF-κB binding (assessed by chromatin immunoprecipitation coupled to high-throughput sequencing; ChIP-seq) and gene expression (assessed by sequencing of total RNA; RNA-seq).

We focused on spatial and temporal changes in chromatin architecture during the critical window when 'immediately-early' proinflammatory genes become active: 0, 10 and 30 min post-stimulation. In agreement with the idea that nucleosomes reposition in coincidence with (and/or as a result of) transcription factor binding at cognate sites[1-6], we did not expect to observe widespread repositioning before NF-κB binding was quantitatively detected (that is, 15 min post-stimulation[17, 18]). However, we observed widespread nucleosome repositioning already by 10 min, coinciding with marginal, if any, stable binding of the factor (Figure 1A). Similarly, we expected elongation by pioneering RNA polymerases along TNFα-responsive genes to initiate a 'wave' of repositioning; however, examination of long (>100 kilobase pairs (kbp)) genes that are synchronously activated by TNFα showed that nucleosomes were already repositioned all the way from 5' to 3' ends, despite polymerases having transcribed <50% of their length after 30 min[19, 20]. We attribute this to changes in positioning that nucleate from few selected NF-κB binding clusters embedded in the bodies of such responsive genes. We show that these effects are accompanied by changes in the three-dimensional conformation of the chromatin fibre detected using chromosome conformation capture coupled to deep sequencing (3C-seq[21]).

Results

TNFα induces immediate widespread changes in nucleosome positioning

HUVECs grown to confluence were serum-starved (to promote synchrony), stimulated with TNFα for 0, 10 or 30 min, and treated with MNase to release mononucleosomes. The purified DNA (Additional file 1A) was deep-sequenced to obtain approximately 180 million read-pairs per time point (Figure 1A). When mapped to the reference genome (hg19), reads from two 0- and 30-min biological replicates gave comparable profiles (Additional file 1B).

First, we identified peaks in the MNase-seq read profiles that marked single-nucleosome positions (using findPeaks[22]) and selected those differentially unmasked at 10 or 30 min post-stimulation (that is, those where nucleosomes are repositioned by >10 bp when compared to 0 min). By 10 min, unmasked regions were enriched for binding motifs of proinflammatory transcription factors (for example, NF-κB, AP-1; Additional file 1C), and characterized by Gene Ontology terms associated with cell regulation and cytokine signalling (Additional file 1D). Notably, short interspersed nuclear elements[23], especially *AluY*, *AluSx* and *AluSg*, which all contain NF-κB binding sites[24] and confer enhancer-like characteristics[25], were amongst the most significantly unmasked regions (Table 1). These findings are perhaps surprising, because levels of nuclear NF-κB do not peak before 15 to 17.5 min (Additional file 1E)[18,26,27].

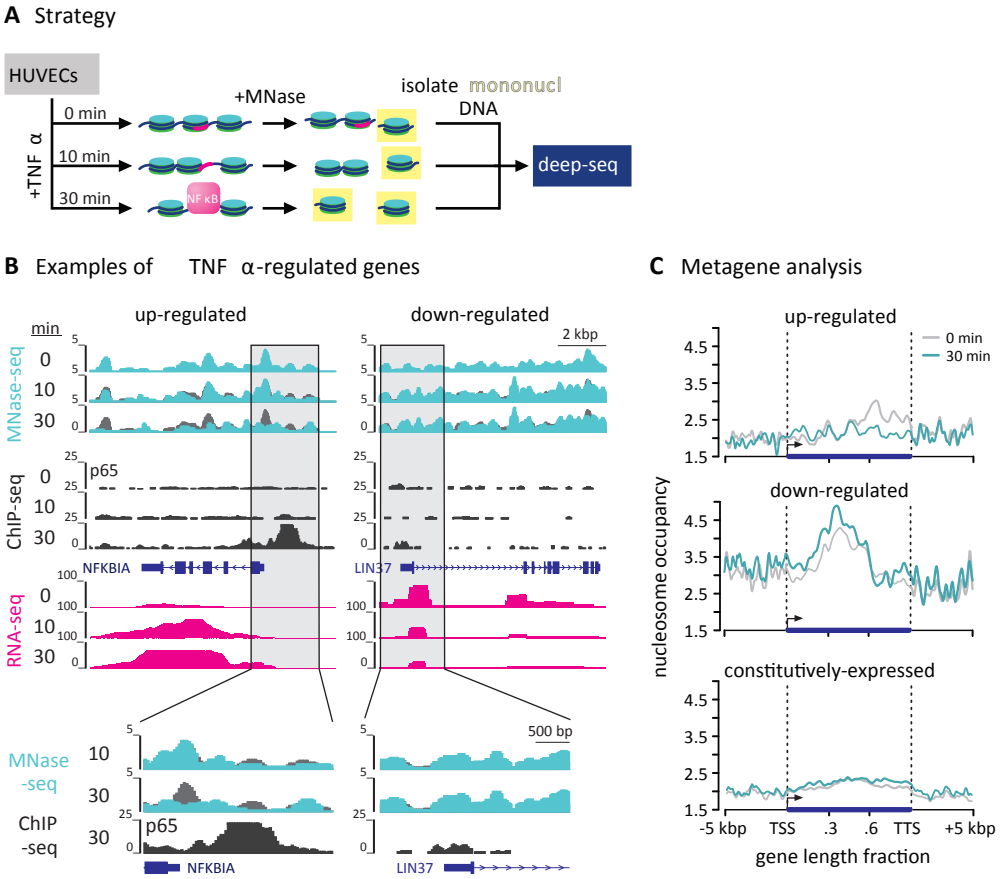

Figure 1. Nucleosome repositioning in TNFα-responsive genes.
(A) Strategy: HUVECs were serum-starved and stimulated with TNFα (0, 10, 30 min), treated with MNase, and DNA associated with mononucleosomes (highlighted yellow) deep-sequenced. Nucleosomes reposition within 10 min to unmask NF-κB binding sites (magenta), before NF-κB enters the nucleus. (B) Browser tracks (vertical axes - reads/million; magnifications of transcription start sites shown below) for typical up- or down-regulated genes obtained by MNase-seq (green; reflects nucleosomal profiles; 0-min levels in grey underlie 10- and 30-min ones to facilitate comparison), p65 ChIP-seq (black; reflects NF-κB binding), and total RNA-seq (magenta; reflects RNAPII activity). (C) Nucleosome occupancy (reads/million; MNase-seq) at 0 (grey) or 30 min post-stimulation (green) along metagenes derived from 109 up-regulated (>0.6 log2 fold-change at 30 compared to 0 min, plus >100 reads mapping to each), 69 down-regulated (<-0.6 log2 fold-change, plus >100 reads mapping to each), and 509 constitutively expressed genes (±0.01 log2 fold-change, plus >100 reads mapping to each). Genes were aligned at transcription start/termination sites (dotted lines), gene bodies divided into 50-bp windows, lengths scaled proportionately, and MNase-seq reads in each window summed; profiles from 5 kbp up- and downstream are also displayed. ChIP-seq, chromatin immunoprecipitation coupled to high-throughput sequencing; kbp, kilobase pair; MNase-seq, micrococcal nuclease digestion followed by sequencing; NF-κB, nuclear factor kappa-B; RNA-seq, sequencing of total RNA; TNFα, tumour necrosis factor alpha; TSS, transcription start site; TTS, transcription termination site.

By 30 min, regulatory regions (for example, CpG islands, promoters, 5′ untranslated regions) and genes (for example, coding regions, exons) were all statistically significantly unmasked (Table 1). These data point to a progressive transition from the 0- to the 10-min, and finally to the 30-min, state.

TNFα induces repositioning in differentially regulated gene subsets

We next examined genes differentially regulated following a 30-min TNFα pulse. They were selected using data obtained after deep sequencing total rRNA-depleted RNA (RNA-seq; approximately 120

million read pairs per time point) and were required to change by at least ±0.6 log_2-fold (that is, ±1.5-fold at 30 min relative to 0 min); constitutively expressed genes (±0.01 log_2-fold) provided controls (Additional file 2A and Additional file 3). We also monitored NF-κB binding using ChIP-seq data (by targeting its p65 subunit) at 10 and 30 min post-stimulation. At 10 min, marginal binding was observed, in agreement with data showing that NF-κB translocation into the nucleus and binding to cognate sites is not quantitatively detected before 15 or 30 min, respectively (examples in Figure 1B and Additional file 1E). At 30 min, more than 80% of up-regulated genes were associated with at least one p65 peak, compared to just 10% of down-regulated ones (compared to 6% and 7% for the 10-min data; Additional file 2B).

Comparison of MNase-seq (raw) read profiles along a typical immediate-early up-regulated gene, *NFKBIA*, showed nucleosomes already repositioned by 10 min, and changes in nucleosome occupancy became more pronounced at 30 min, when density decreased throughout the locus as NF-κB binding increased (Figure 1B, *left*). By contrast, profiles on a typical down-regulated gene, *LIN37*, became heightened and more defined (Figure 1B, *right*). This held true for other up- or down-regulated genes, whilst those of constitutively expressed loci varied little (Additional file 2C).

Global changes in genic nucleosome occupancy were assessed using 'metagene' analyses, by aggregating profiles from all up- or down-regulated genes. In up-regulated genes, the first few nucleosomes downstream of the promoter became more precisely positioned (most likely as transcription start site (TSS)-proximal nucleosomes form well-positioned arrays [1]), and occupancy decreased incremental-

Table 1 Genome Ontology analysis of nucleosome-unmasked regions

10 versus 0 min TNFα stimulation			30 versus 0 min TNFα stimulation		
Annotation	GO group	log P -value	Annotation	GO group	log P -value
rRNA	Basic	-132.6	CpG island	Basic	-18,894.5
-	-	-	coding	Basic	-2,508.6
-	-	-	protein -coding	Basic	-2,202.6
-	-	-	exons	Basic	-2,175.4
-	-	-	promoters	Basic	-1,637.2
-	-	-	5′ UTR	Basic	-1,573.7
-	-	-	rRNA	Basic	-90.7
-	-	-	miscRNA	Basic	-57.6
-	-	-	TTS	Basic	-30.3
-	-	-	miRNA	Basic	-2.3
Alu	SINE	-1,010,851.5	Alu	SINE	-127,523.7
Satellite	Satellite	-138,649.4	AluY	SINE	-254,77.1
AluSx	SINE	-130,806.8	AluJb	SINE	-13,157.4
Simple	Repeat	-109,102.9	AluSx	SINE	-10,162.6
AluSz	SINE	-95,618.8	AluSx1	SINE	-9,723.1
Satellite	Satellite	-84,407.6	AluSz	SINE	-7,150.9
TGn	Repeat	-82,907.0	AluJr	SINE	-6,259.1
Can	Repeat	-82,146.3	AluJo	SINE	-5,837.8
AluSx1	SINE	-81,015.6	AluSz6	SINE	-3,989.9
CATTCn	Satellite	-77,388.9	AluSg	SINE	-3,556.9

A list of the top regions unmasked at 10 and 30 min post-stimulation (looking at nucleosomes identified using findPeaks [23] that were repositioned by >10 bp at 10 or 30 compared to 0 min). Top half: regions associated with 'basic' genome annotation. Bottom half: repeat elements. For each entry, the annotation category, genome ontology group and identification confidence levels (log P-value) are shown; Alu repeats known to bind NF-κB [24] are in bold. GO, Gene Ontology; SINE, short interspersed nuclear elements; TNFα, tumour necrosis factor alpha.

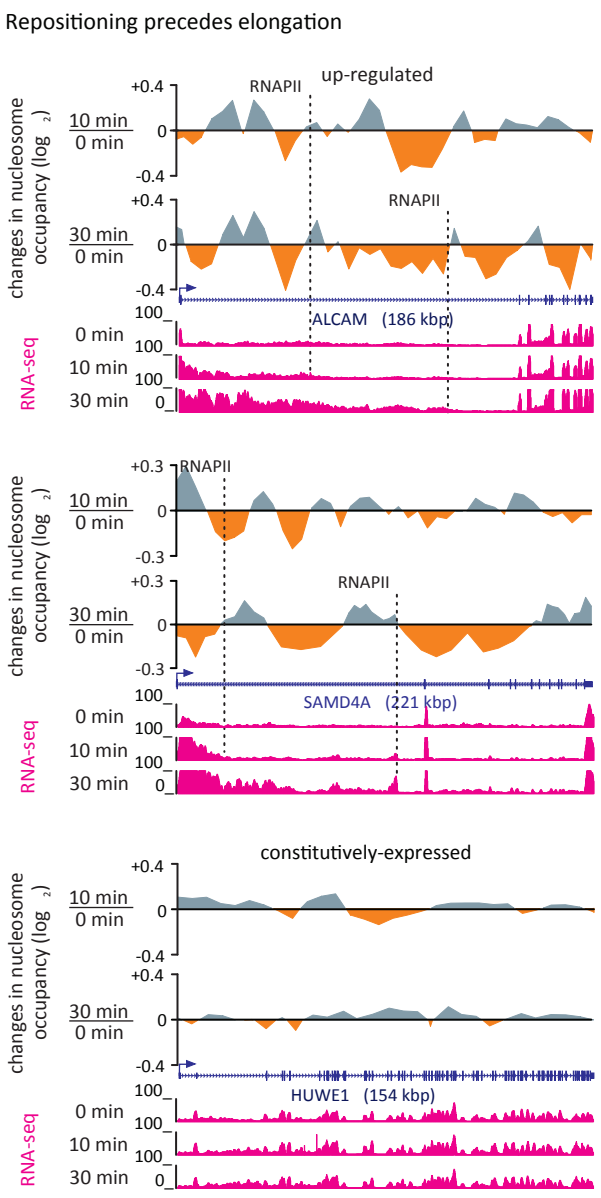

Figure 2. Nucleosome repositioning at 3' ends of long genes precedes transcription by pioneering (elongating) polymerases.
Browser views show (log2 fold) changes in nucleosome occupancy 10 or 30 min post-stimulation calculated using 5-kbp non-overlapping windows and a running-means average along up-regulated long genes ALCAM and SAMD4A. Changes (read enrichment - grey; read depletion - orange) are shown normalized to those in transcriptionally inert genomic regions. Total RNA-seq tracks (magenta) show elongating polymerases generating intronic signal close to the 5' ends of genes after 10 and 30 min, as they have not yet reached termini (dotted lines - positions of pioneering RNAPs after 10 and 30 min). The long, constitutively expressed HUWE1 locus (bottom) serves as a control. Kbp, kilobase pair; RNAP, RNA polymerase; RNA-seq, sequencing of total RNA.

ly towards the 3' end (as nucleosome-rich exons tend to be found more 3'[28, 29]). In down-regulated genes, occupancy increased throughout; again, little change was observed in constitutively expressed loci (Figure 1C).

Nucleosome repositioning precedes transcriptional elongation in long genes

The transcriptional activation of five long genes of >100 kbp has been studied in detail in this experimental model[17-20]. Following treatment with TNFα, pioneering RNA polymerases (RNAPs) initiate synchronously at the TSSs within 15 min, and then elongate at approximately 3 kbp/min. Thus, elongating RNAPs have transcribed less than the first half of these long genes after 30 min (see RNA-seq profiles in Figure 2 and ChIP-quantitative PCR (qPCR) in Additional file 4A). Therefore, one would expect nucleosomes only in the first half of these genes to have been repositioned. To simplify analysis, we initially applied the PeakPredictor algorithm[30] to our MNase-seq data and 'called' single-nucleosome positions along three such long genes. As expected, TSS-proximal regions appeared progressively more depleted of nucleosome peaks (for example, in the first 10 kbp downstream of the TSS of 318-kbp *EXT1*, 41, 38 and 24 peaks were called at 0, 10 and 30 min, respectively; Additional file 4B). Unexpectedly, peak depletion of the same scale spread over hundreds of kilobase pairs from TSS to transcription termination site (TTS) (for example, the number of peaks throughout *EXT1* fell by 12% after 30 min; Additional file 4B), and 'MNase-on-ChIP' verified this effect (Additional file 4C).

Of course the above effect does not accurately describe the phenomenon, as there exist no such long nucleosome-devoid stretches of DNA. Thus, we analysed MNase-seq data throughout each long gene via a custom bioinformatics pipeline to examine whether nucleosome reposi-

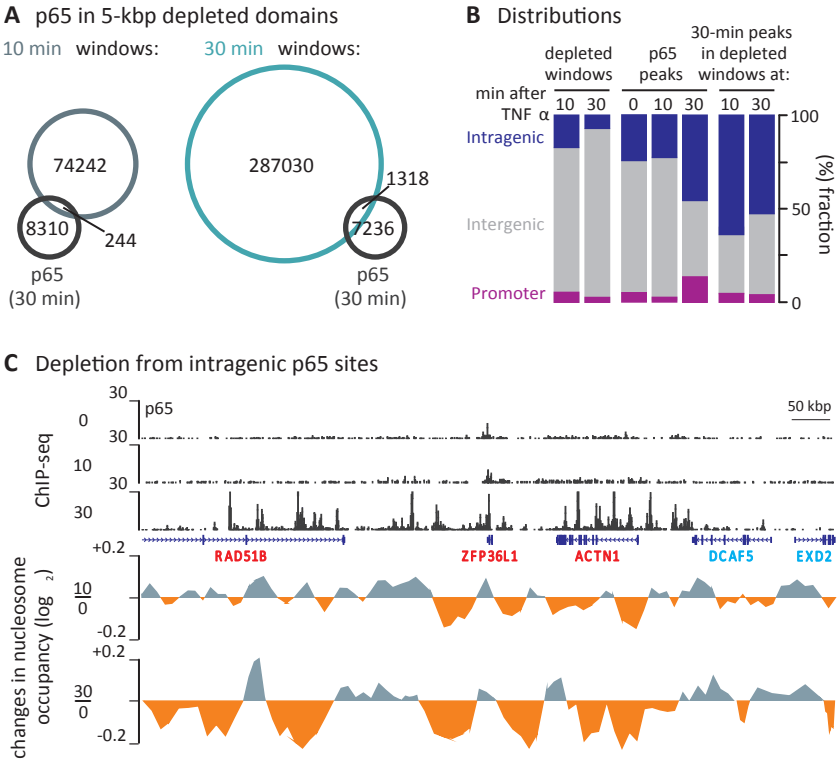

Figure 3. NF-κB binding in TNFα-regulated nucleosomal domains.
(A) A minority of p65 peaks are found in depleted domains. The genome was partitioned into 5-kbp non-overlapping windows, and those depleted of nucleosomes selected (determined as in Figure 2) and compared to the location of p65 binding sites (determined using ChIP-seq data obtained 30 min post-stimulation). By 10 min, 74,486 nucleosome-depleted windows appear, after 30 min 288,377 such windows develop (21,788 of which are also seen at 10 min). By 30 min, 8,554 p65 peaks are seen, but only 244 and 1,318 overlap (≥25% of sequence) with the 10- and 30-min nucleosome-depleted windows, respectively. (B) Nucleosome-depleted p65-containing windows are predominantly intragenic. Bar graphs give the fraction of nucleosome-depleted windows or p65 peaks (0, 10, 30 min) coinciding with regions lying within or outside annotated genes (blue - intragenic; grey - intergenic, or ±1 kbp from the transcription start site (purple - promoter) (C) Browser tracks illustrating changes in nucleosome occupancy (log2 fold-changes determined using 5-kbp non-overlapping windows as in Figure 2) in a 1-Mbp locus on chromosome 14 (TNFα-responsive genes - red, non-responsive - blue); p65 ChIP-seq tracks (0, 10 and 30 min post-TNFα; vertical axes - reads/million) are also shown. ChIP-seq, chromatin immunoprecipitation coupled to high-throughput sequencing; kbp, kilobase pair; TNFα, tumour necrosis factor alpha.

tioning follows RNAP elongation (as might be expected). Genes were divided into 5-kbp non-overlapping windows, and changes in each window scored relative to (background) levels of nucleosome repositioning occurring in transcriptionally inert genomic segments (see Methods). This revealed a decrease in nucleosome occupancy (hereafter termed depletion), which was evident throughout 186-kbp *ALCAM* and 221-kbp *SAMD4A* (Figure 2), as well as in 116-kbp *NFKB1* and 458-kbp *ZFPM2* (Additional file 5A), at both 10 and 30 min, when pioneer RNAPs had advanced for <30 and <100 kbp, respectively. This effect was reproducible between biological replicates (Additional file 5B), and profiles of down-regulated and constitutively expressed genes served as controls (Figure 2 and Additional File 5A).

NF-κB binding is associated with repositioning over great distances

We next examined whether NF-κB binding was enriched in kilobase pair-long genomic segments dis-

playing reduced MNase-seq signal. ChIP-seq collected 10 min post-stimulation showed sparse binding of p65 (approximately 200 peaks genome-wide, most at repeat elements; Additional file 6A), but by 30 min around 8,600 peaks were detected, most found at sites bearing histone marks characteristic of enhancers (high H3K4me1 and H3K27ac, low H3K4me3[31]; Additional file 6A). At the same time, >280,000 5-kbp windows appeared depleted of nucleosomes (defined as above). Remarkably, <20% of p65 peaks (1,318) were embedded in such depleted windows, and the overlap was even smaller when compared to 10-min windows (244 peaks; Figure 3A). This is inconsistent with a simple model where NF-κB binding drives genome-wide nucleosome depletion, especially as little NF-κB has quantitatively bound in HUVEC chromatin by 10 min (Figure 1B and Additional file 1E). Intriguingly, p65-bearing windows significantly associated with gene bodies (Figure 3B).

As p65 binds both close to and in the body of many up-regulated genes (Additional file 2B), we speculated that the TNFα-driven repositioning seen throughout such genes (Figure 1C) might be nucleated from p65 bound at intragenic sites (Figure 3C illustrates one locus). Thus, of all up-regulated genes examined, 72% encompassed ≥1 p65 peak; by contrast, <10% of down-regulated genes contained a p65 peak (Additional file 7A). The physical separation between such intragenic peaks in up-regulated genes is an order of magnitude greater than those between intergenic ones (despite the small fraction of the genome occupied by protein-coding genes); thus, this group of peaks covers a substantial portion of the respective gene bodies (Additional file 7A). These results point to a focused binding of NF-κB, in clusters of 'primed' sites, within genes (even though the transcription factor might be bound at low titres), followed by nucleosome repositioning over several tens of kilobase pairs (Additional file 6B and Additional file 7B).

Multi-scale nucleosome repositioning impacts on higher-order structure

We next used the long arm of chromosome 14 as a model to study how changes in nucleosome density might affect structure at increasingly larger scales (as loci on this chromosome have been extensively studied before[17-20]). The chromosome was divided into non-overlapping windows of 25, 50 and 100 kbp, and nucleosome occupancy examined. By 10 min, alternating enriched and depleted domains were seen at all window sizes; by 30 min most of these further evolved (Additional file 8A) and depleted profiles predominated (also reproducible between replicates; Additional file 8B). In other words, a gradual spreading of nucleosome-depleted domains was observed, and this appeared to be nucleated by the hotspots seen at 10 min (many also engulfing DNase-hypersensitive sites, especially by 30 min post-stimulation; Additional file 8C).

To relate changes in nucleosome occupancy to those in DNA conformation, we performed 3C-seq at 0 and 30 min post-stimulation[21] using the TSSs of TNFα-responsive *SAMD4A* and constitutively expressed *EDN1* as viewpoints. For the *SAMD4A* TSS, we showed previously that stimulation induces development of new contacts throughout the genome[18]; here we focus only on the more abundant intra-chromosomal contacts. At 0 min, *SAMD4A* contacts were scattered throughout the chromosome arm, and after 30 min new ones developed (Figure 4A, *top*). Of the 167 most frequently seen 30-min contacts, 131 formed *de novo* upon TNFα treatment. When correlated with changes in nucleosome occupancy (in 5-kbp windows, as in Figure 2), we found essentially all 30-min and 'shared' contacts embedded in nucleosome-depleted windows (significantly more than 0-min contacts; Figure 4A).

By contrast, the *EDN1* TSS formed fewer new contacts upon stimulation (of the 496 most frequent 30-min contacts 42% were also see at 0 min; Figure 4B, *top*). Moreover, significantly more shared contacts correlated with nucleosome-depleted windows (compared to 0- or 30-min specific ones; Figure 4B). Closer inspection of the two loci shows that contacts (in accord with obtained chromatin interaction analysis by paired-end tag sequencing data[18]) do not form randomly between 'nucleosome-free' regions, but rather share particular features (that is, NF-κB binding, H3K4me1 enrichment and transcriptional activity; Additional file 9).

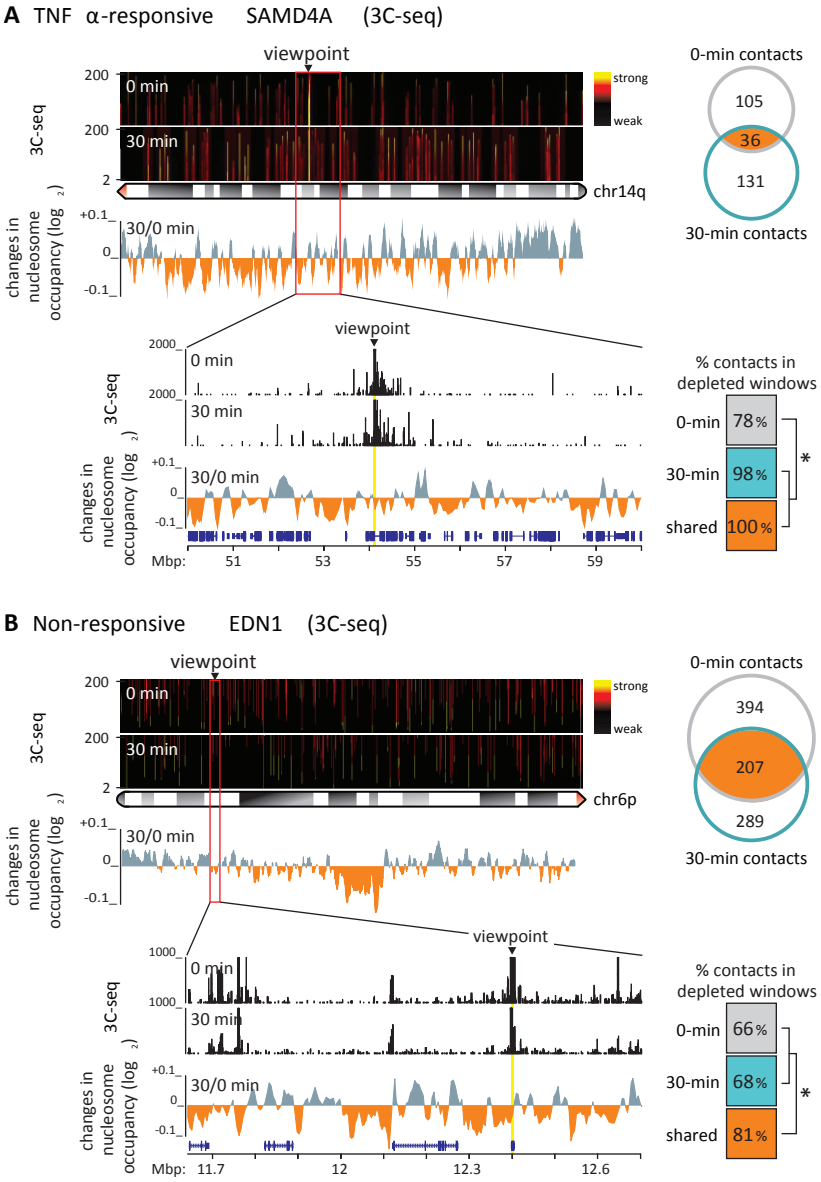

Figure 4. Changes in nucleosome positioning affect higher-order structure.
(A) High-confidence contacts (P <0.05; determined using 3C-seq 0 or 30 min post-stimulation) made by the transcription start site (TSS; arrowhead) of TNFα-responsive SAMD4A with parts of the long arm of chromosome 14 (ideogram) are depicted as a domainogram (y-axis - contacts visualized in 2- to 200-kbp sliding windows). Most contacts are unique for each time point (Venn diagram). The magnified region (red rectangle) compares 3C-seq contacts (y-axis - reads per million) to changes in nucleosome occupancy (determined as in Figure 2). The table (bottom right) gives the fraction of 3C contacts embedded in nucleosome-depleted windows at 0 or 30 min, or shared at both times; a significant increase is seen for 30-min and shared contacts (*P <0.05; Fisher's exact test). (B) Details as in panel (A), for the non-responsive EDN1 TSS (arrowhead) on the long arm of chromosome 6 (ideogram). Almost 40% of high-confidence contacts persist from 0 to 30 min (Venn diagram), and are significantly associated with nucleosome-depleted 5-kbp windows (*P <0.05; Fisher's exact test). 3C-seq, chromosome conformation capture coupled to deep sequencing; TNFα, tumour necrosis factor alpha.

Discussion

We addressed the question: how does TNFα stimulation reshape the chromatin landscape as it establishes the immediate-early proinflammatory transcriptional programme? The cytokine signals through NF-κB[13], and one might envisage that the factor first binds in the vicinity of regulatory elements to induce repositioning of nucleosomes locally. This would then facilitate transcriptional initiation by RNA polymerase, and would in turn open up the bodies of TNFα-responsive genes as polymerases elongate through them[32, 33]. However, changes observed here cannot be reconciled with this scenario.

First, we saw hotspots of nucleosome depletion 10 min post-stimulation (Additional file 8A), before detectable NF-κB binding to cognate sites (Additional file 6A). Although there were approximately 1,300 NF-κB binding peaks in nucleosome-depleted windows after 30 min, most bound NF-κB was not embedded in kilobase pair-long depleted regions (Figure 3A). This also fits with the distribution of typical NF-κB motifs (5'-GGRRNNYYCC-3'): out of >550,000 sites found genome-wide, only 60,000 and 250,000 were embedded in windows depleted of nucleosomes after 10 and 30 min, respectively (with 28,000 being shared and very few being occupied; Figure 3A). It follows that NF-κB binding is highly selective; the first transcription factor complexes to enter nuclei (between 10 and 15 min) must preferentially bind to a small subset of primed domains depleted of nucleosomes, harbouring the highest affinity sites - probably within the critical enhancers that regulate the ensuing cascade and/or on particular *Alu* repeats[24]. This is reminiscent of a subset of NF-κB dimers in macrophages selectively binding to already-accessible chromatin segments where partner regulators constitutively bind[34] - which raises the question of what the endothelial-specific NF-κB partners might be.

Second, results cannot be reconciled with the idea that transcription through nucleosomes by pioneering elongating RNAPs is solely responsible for changes in chromatin structure. Nucleosomes in long TNFα-responsive genes are repositioned throughout, well before elongating polymerases have transversed their full length (Figure 2). Then, what molecular mechanism might drive repositioning at sites many kilobase pairs away from a bound NF-κB or a pioneering polymerase? We can suggest some possibilities that might act singly, or in concert. For example, an effector other than NF-κB might be responsible for priming; then, NF-κB (and/or another effector) could induce chromatin remodelling enzymes to act throughout the surrounding locale - perhaps a chromatin loop or cluster of loops in a topological domain attached to a transcriptional hot spot[35]. Alternatively, transcription could generate supercoiling that remodels one such loop (or cluster of loops) within a topological domain[36]. Lastly, polymerases other than pioneers on responsive genes could drive repositioning - perhaps ones generating enhancer RNAs (like in Additional file 6B)[37]. This is supported by the presence of NF-κB clusters bound within gene bodies at sites marked by histone marks and transcripts characteristic of enhancers; these overlap 'super-enhancers' previously mapped in HUVECs[38] that also show decreased nucleosome density post-stimulation (see examples in Figure 3C and Additional File 6B).

Third, nucleosome repositioning has traditionally been viewed as a local phenomenon, but we detect occupancy changes throughout megabase pair-long segments (see chromosomes 4 and 14 in Additional file 8). (Note that, using semi-quantitative Western blotting with antibodies targeting histones H3 and H4, we verified TNFα stimulation does not affect global histone levels; data *not shown*.) Using 3C-seq, we confirmed the intuition that changes in nucleosome positioning around two megabase pair-long chromosomal loci go hand-in-hand with the development of contacts in three-dimensional nuclear space. Interestingly, a subset of recorded 3C contacts - which predominantly form between regulatory *cis*-modules[39, 40] marked by NF-κB and characteristic histone modifications (Additional file 9) - persist throughout the transition from the unstimulated to the TNFα-stimulated state (Figure 4). This is consistent with pre-looped chromatin facilitating responses to extra-cellular cues[41], and can now be explained also at the level of nucleosomal organization.

Conclusions

Collectively, our data point to TNFα triggering chromatin priming so that most nucleosomes are repositioned independently of NF-κB binding and/or polymerases elongating through responsive genes. This effect is a prelude to the ensuing proinflammatory programme, and it occurs both locally (at the gene level) as well as at considerable distances from, what have hitherto been considered, the major nucleating sites to affect large chromosomal segments. Finally, although 'topological domains' may constitute invariant building blocks within chromatin[41-43], an underlying and plastic network of interactions within a domain must affect DNA accessibility to polymerases, ultimately allowing the rapid transitions that occur as different sets of genes become active and inactive and the inflammatory cascade unfolds[15, 16]. Of course, the molecular machines responsible for priming, their interplay with NF-κB, and the potential role of other factors (like histone H1 eviction or activity of topoisomerases) need be addressed in light of these findings.

Methods

Cell culture

HUVECs from pooled donors (Lonza, Cologne, Germany were grown to 80% to 90% confluence in endothelial basal medium 2-MV with supplements (EBM; Lonza) and 5% foetal bovine serum (FBS); starved for 16 to 18 h in EBM + 0.5% FBS; treated with TNFα (10 ng/ml; Peprotech, Hamburg, Germany); and harvested 0, 10 or 30 min post-stimulation.

Isolation of mononucleosomes, sequencing and mapping

Approximately 5×10^6 HUVECs stimulated with TNFα for 0, 10 or 30 min were digested (3 min at 37°C) with 750 units of micrococcal nuclease (MNase; Sigma-Aldrich, Seelze, Germany). Mononucleosomal DNA was isolated following separation on 1.3% agarose gels using glass beads (Qiagen, Hilden, Germany), and average fragment lengths determined using a 2100 Bioanalyzer (Agilent). Libraries were generated using the NEBNext DNA Library Prep Master Mix Kit (New England Biolabs, Ipswich, USA) and paired-end (2×50-bp) sequenced on a HiSeq2000 platform (Illumina) to comparable depths (that is, 181, 185 and 187 million reads for 0, 10 and 30 min samples, respectively). Obtained reads were processed using the toolkits FastQC[44] and FASTX[45], mapped to hg19 using Bowtie[46].

MNase-seq analysis

Different peak-calling algorithms were applied depending on the downstream application. For Additional file 4 the Peak Predictor/GeneTrack package[30] was used. For motif analyses, as well as Gene and Genome Ontology profiling (Additional file 1 and Table 1), the HOMER software package[47] and findPeaks 3.1[22] were applied (adjusting fragment size to that determined using the Bioanalyzer with the following settings: *–style factor –size 147 –minDist 1 –F 0 –L 0 –C 0*). When comparing two or more datasets, the *getDifferentialPeaks* or *mergePeaks* scripts were used. For visualization, tag directories of mapped reads were generated and .bedGraph files produced using the *makeUCSCfile* (for raw reads) or *pos2bed.pl* (for peaks and other BED-formatted files) scripts; tracks were then visualized with the UCSC Genome browser[48]. Both known and *de novo* motif analyses were performed with *findMotifsGenome.pl* using standard settings and the repeat-masked hg19 genome build. All peak annotations, including histograms, were generated with *annotatePeaks.pl*, and graphs plotted in R[49] with a smoothing spline of 0.2.

Differences in nucleosome positioning between any two time-points (0- compared to 10- or 30-min

datasets) were elucidated statistically using a novel Neyman-Pearson 'normalized log-likelihood-ratio' analysis. Chromosomes 1-X were divided in n non-overlapping windows $w_1, w_2, \ldots w_n$ of a constant size $|w_i|$. In a pre-processing step, MNase-seq data files containing read positions at t_1 and t_2 were used to compile datasets $R = (r_1, r_2, \ldots r_n)$ and $S = (s_1, s_2, \ldots s_n)$; r_i and s_i are the read counts in each w_i observed under treatments t_1 and t_2, respectively. Then hypotheses H_1 and H_2 were tested by computing a log-likelihood-ratio Q according to:

$$Q = \log\frac{R}{S} = (q_1, q_2 \ldots . q_n); \quad q_i = \frac{r_i}{s_i}.$$

This set of log-likelihood-ratio values has a mean of $Q_{mean} = \frac{1}{n}\sum_{i=1}^{n} q_i$ and a normalized distribution $||Q|| = Q - Q_{mean}$. It follows that $||q_i||$ values are centred on zero. The null hypothesis is then that all observed q_i-values from regions that were transcriptionally inert (assessed using RNA-seq data) were due to random fluctuations and not caused by treatments t_1 and t_2. The normalized cumulative distribution N_{cum} was used to determine a p-value $p(||q_i||)$ for $||q_i|| \geq 0$ according to:

$$p(\| q_i \|) = 1 - N_{cum}(\| q_i \|)$$

Thus, the smaller $p(||q_i||)$ is, the lower the probability that the ratio $||q_i||$ is merely due to a stochastic fluctuation of read counts.

Chromosome conformation capture

Nuclei were harvested after 0 or 30 min of TNFα stimulation, cross-linked in 1% paraformaldehyde (PFA; Electron Microscopy Science, Munich, Germany), and processed as described [21] using *Apo*I as the primary restriction endonuclease. Following sequencing on a HiSeq2000 platform (Illumina; approximately 2×10^7 reads), data were analysed using the r3Cseq pipeline[50]. The domainogram in Figure 4 was generated using the top 167 *cis*-contacts on chromosome 14 (on which the viewpoint lies) using publicly available software[51]. In brief, 3C-seq reads are made binary and relative enrichments calculated using sliding windows compared to a randomized background made up of 3,000 fragment ends. Data permutation is then used to determine a threshold of <0.01 false discovery rate (FDR); windows exceeding this threshold are scored as interacting.

Chromatin immunoprecipitation and ChIP-seq analysis

Approximately 10^7 HUVECs were cross-linked (using 1% PFA for 10 min, preceded by 25 min in 10 mM ethyl-glycol-*bis*-succinimidylsuccinate at room temperature, as described previously[18]) 0, 10 or 30 min after TNFα stimulation; chromatin was fragmented by sonication (Bioruptor; Diagenode, Liège, Belgium); then immunoprecipitation was carried out using a rat monoclonal against phospho-Ser2 in the C-terminal domain of the largest subunit of RNA polymerase II (3E10[52]; a gift from Dirk Eick, Helmholtz Institute, Munich, Germany) or a rabbit polyclonal against the full-length p65 subunit of NF-κB (39369, Active motif) on aliquots of approximately 25 µg chromatin. Immunoprecipitated complexes were washed and eluted using the ChIP-It-Express kit (Active motif, Rixensart, Belgium).
For qPCR analysis, a Rotor-Gene 3000 cycler (Qiagen) and Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen, Darmstadt, Germany) were used. Following incubation at 50°C for 2 min to activate the qPCR mix, and 95°C for 5 min to denature templates, reactions were carried out for 40 cycles at 95°C for 15 s, and 60°C for 50 s. PCR primers were designed via Primer3Plus[53] using *qPCR* settings with an optimal length of 20 to 22 nucleotides, a Tm of 62°C, targeting 100 to 200 bp. The presence of single amplimers was confirmed by melting-curve analysis, and data were analysed to obtain enrichments relative to input. *P* values (two-tailed) from unpaired Student's *t*-tests[54] were considered significant

when <0.05.
For deep sequencing, previous (0- and 30-min[18]) and newly generated (10-min) p65 ChIP-seq data were aligned to hg18 and signal peaks detected using MACS[55]. This allowed 68, 214 and 8,583 high-confidence p65-binding events to be detected for 0, 10 and 30 min respectively (FDR ≤0.01, peak height ≥20 reads/million). Peaks were correlated to publicly available ENCODE Hidden Markov chromatin models and HUVEC ChIP-seq data (H3K27ac: GSM733691; H3K4me1: GSM733690[31, 56]) and annotated against RefSeq genomic features (TSS, exon, intron, intergenic region).

Total RNA sequencing and analysis

Total RNA was isolated from 0.5×10^6 HUVECs stimulated with TNFα for 0, 10 or 30 min using TRIzol (Invitrogen), treated with RQ1 DNase (1 unit/µg RNA, 37°C, 45 min; Promega, Leiden, Netherlands), depleted of rRNA (RiboMinus; Epicentre, Madison, USA), chemically fragmented to approximately 350 nucleotides, and cDNA generated using random hexamers as primers (according to the True-seq protocol; Illumina). Adapters were then ligated to cDNA molecules, and libraries sequenced (Illumina HiSeq2000 platform; 100-bp paired-end reads; around 120×10^6 read-pairs per sample). Raw reads were then mapped to hg18 using TopHat[57] and reads aligning to RefSeq gene models were counted using the HTseq package[58]. Statistical analysis of differentially expressed genes was performed with the DESeq Bioconductor package[59] (asking for >100 reads per gene, and for a >0.6, <-0.6, or ±0.01 $\log_2$ fold-change for up-regulated, down-regulated or constitutively expressed genes, respectively; Additional file 3).

Immunofluorescence

HUVECs grown on coverslips etched with hydrofluoric acid were fixed with 4% PFA (Electron Microscopy Science) in phosphate-buffered saline (PBS; 20 min, 20°C), washed once in PBS (5 min, 20°C), permeabilized using 0.5% Triton X-100 in PBS (5 min, 20°C) and blocked with 1% bovine serum albumin (BSA) in PBS (Sigma-Aldrich; 45 min, 20°C). Phosphorylated (at Ser536) p65 was detected using a rabbit monoclonal antibody (1:500 dilution, 0.5% BSA in PBS; #04-1000, Millipore, Nottingham, UK) and Alexa488-conjugated donkey anti-rabbit AffinityPure F(ab')2 Fragment (1.5 µg/ml; Jackson ImmunoResearch, Maine, USA). After DAPI counter-staining, images were collected on a Leica DMI6000 B widefield microscope and analysed using ImageJ[60]; nuclei were encircled, the mean intensity calculated per area, and nuclear fluorescence (arbitrary units) calculated by subtracting the background (measured as the minimum intensity in the image).

Data availability

MNase-seq raw data are available at the GEO database under accession number [GEO: GSE53343], while 3C-seq, p65 ChIP-seq and total (ribo-depleted) RNA-seq data generated here can be accessed at the SRA archive under accession number [SRA: SRP044729].

Abbreviations
3C-seq; chromosome conformation capture coupled to deep sequencing; bp, base pair; BSA, bovine serum albumin; ChIP-seq, chromatin immunoprecipitation coupled to high-throughput sequencing; EBM, endothelial basal medium; FBS, foetal bovine serum; FDR, false discovery rate; kbp, kilobase pair; MNase-seq, micrococcal nuclease digestion followed by sequencing; NF-κB, nuclear factor kappa-B; PFA, paraformaldehyde; PBS, phosphate-buffered saline; RNAP, RNA polymerase; RNA-seq, sequencing of total RNA; TNFα, tumour necrosis factor alpha; TSS, transcription start site; TTS, transcription

termination site.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SD, PK, PRC, GL and AP designed experiments. SD, PK, TG, AZ and AP performed experiments. SD, PK, TG, US, GW and AP performed bioinformatics analyses. LH and RM produced the MNase-seq analysis algorithm. All authors interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Karsten Rippe and Alvaro Rada-Iglesias for discussions; Vladimir Benes (EMBL, Heidelberg, Germany), Wilfred van Ijcken (Erasmus MC, Rotterdam, Netherlands) and Chris Greenman (TGAC, Norwich, UK) for sequencing the MNase-, ChIP-, 3C- and RNA-seq libraries, respectively; and Dirk Eick for the 3E10 antibody. This work was supported by the Epigensys consortium funded by the ERASysBio+/FP7 initiative via the BBSRC (PRC), the BMBF (SD, GL, GW) and the NWO (PK, FG, TAK); by a M.E.C. Booster grant from the Netherlands Genomics Institute (PK); by an SBF960 collaborative grant (GL); by CMMC intramural funding (TG, AP); and by Köln Fortune (AZ).

Supplementary information
Supplementary information is available at the Genome Biology website: Supplementary File 1-9

References

1. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**, 267-273 (2013).
2. Sadeh, R. & Allis, C.D. Genome-wide "re"-modeling of nucleosome positions. *Cell* **147**, 263-266 (2011).
3. Gaffney, D.J. et al. Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**, e1003036 (2012).
4. Li, G. & Zhou, L. Genome-wide identification of chromatin transitional regions reveals diverse mechanisms defining the boundary of facultative heterochromatin. *PLoS One* **8**, e67156 (2013).
5. Cairns, B.R. The logic of chromatin architecture and remodelling at promoters. *Nature* **461**, 193-198 (2009).
6. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **10**, 443-456 (2009).
7. He, H.H. et al. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res* **22**, 1015-1025 (2012).
8. Grontved, L. & Hager, G.L. Impact of chromatin structure on PR signaling: transition from local to global analysis. *Mol Cell Endocrinol* **357**, 30-36 (2012).
9. Teif, V.B. et al. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* **19**, 1185-1192 (2012).
10. Weicksel, S.E., Xu, J. & Sagerstrom, C.G. Dynamic nucleosome organization at hox promoters during zebrafish embryogenesis. *PLoS One* **8**, e63175 (2013).
11. Valouev, A. et al. Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520 (2011).
12. Zhang, Z. & Pugh, B.F. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* **144**, 175-186 (2011).
13. Smale, S.T. Selective transcription in response to an inflammatory stimulus. *Cell* **140**, 833-844 (2010).
14. Natoli, G. NF-kappaB and chromatin: ten years on the path from basic mechanisms to candidate drugs. *Immunol Rev* **246**, 183-192 (2012).
15. Bhatt, D.M. et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279-290 (2012).
16. Hao, S. & Baltimore, D. RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc Natl Acad Sci U S A* **110**, 11934-11939 (2013).
17. Papantonis, A. et al. Active RNA polymerases: mobile or immobile molecular machines? *PLoS Biol* **8**, e1000419 (2010).
18. Papantonis, A. et al. TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *EMBO J* **31**, 4404-4414 (2012).
19. Wada, Y. et al. A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A* **106**, 18357-18361 (2009).
20. Larkin, J.D., Cook, P.R. & Papantonis, A. Dynamic reconfiguration of long human genes during one transcription cycle. *Mol Cell Biol* **32**, 2738-2747 (2012).
21. Stadhouders, R. et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protoc* **8**, 509-524 (2013).
22. Fejes, A.P. et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729-1730 (2008).
23. Deininger, P. Alu elements: know the SINEs. *Genome Biol* **12**, 236 (2011).
24. Antonaki, A. et al. Genomic analysis reveals a novel nuclear factor-kappaB (NF-kappaB)-binding site in Alu-repetitive elements. *J Biol Chem* **286**, 38768-38782 (2011).

25. Su, M., Han, D., Boyd-Kirkup, J., Yu, X. & Han, J.D. Evolution of Alu elements toward enhancers. *Cell Rep* **7**, 376-385 (2014).
26. Ashall, L. et al. Pulsatile stimulation determines timing and specificity of NF-kappaB-dependent transcription. *Science* **324**, 242-246 (2009).
27. Tay, S. et al. Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267-271 (2010).
28. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19**, 1732-1741 (2009).
29. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001 (2009).
30. Albert, I., Wachi, S., Jiang, C. & Pugh, B.F. GeneTrack--a genomic data processing and visualization framework. *Bioinformatics* **24**, 1305-1306 (2008).
31. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
32. Limpert, A.S. et al. NF-kappaB forms a complex with the chromatin remodeler BRG1 to regulate Schwann cell differentiation. *J Neurosci* **33**, 2388-2397 (2013).
33. Xiong, Y. et al. Brg1 governs a positive feedback circuit in the hair follicle for tissue regeneration and repair. *Dev Cell* **25**, 169-181 (2013).
34. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157-171 (2013).
35. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5**, 1 (2012).
36. Naughton, C. et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* **20**, 387-395 (2013).
37. Natoli, G. & Andrau, J.C. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* **46**, 1-19 (2012).
38. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).
39. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).
40. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113 (2012).
41. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).
42. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-782 (2013).
43. Kolovos, P. et al. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin* **7**, 10 (2014).
44. A quality control tool for high throughput sequence data. *www.bioinformatics.babraham.ac.uk/projects/fastqc.*
45. Short read pre-processing tools. *www.hannonlab.cshl.edu/fastx_toolkit*.
46. 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).
47. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
48. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
49. The R project for statistical computing. *www.r-project.org/.*
50. Thongjuea, S., Stadhouders, R., Grosveld, F.G., Soler, E. & Lenhard, B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res* **41**, e132 (2013).
51. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221-230 (2012).
52. Chapman, R.D. et al. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science* **318**, 1780-1782 (2007).
53. Primer 3.0 Plus oligonucleotide designing software. *www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi.*
54. GraphPad statistical software. *www.graphpad.com.*
55. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
56. Hoffman, M.M. et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-841 (2013).
57. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
58. HTSeq, analyzing high-throughput sequencing data with Python. *http://www-huber.embl.de/users/anders/HTSeq/.*
59. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
60. Abramoff MD, Magelhaes PJ & SJ, R. Image processing with ImageJ. *Biophot Int* **11**, 36–42 (2004).

Chapter 5

The bimodal function of NFκB and the effect of TNFα in the spatiotemporal landscape of responsive and non-responsive genes

Petros Kolovos[1]†, Milos Nikolic[2], Anna Koeferle[3], Joshua D. Larkin[3], Wilfred F. van Ijcken[4], Eduardo G. Gusmao[5], Ivan G. Costa[5], Peter R. Cook[3], Frank G. Grosveld[1], Argyris Papantonis[2]†

[1]Department of Cell Biology, Erasmus MC, Dr. Molewaterplein 50, 3015GE Rotterdam, the Netherlands
[2]Center for Molecular Medicine, University of Cologne, 50931 Cologne, Germany
[3]Sir William Dunn School of Pathology, University of Oxford, OX1 3RE Oxford, UK
[4]Center for Biomics, Erasmus Medical Centre, 3015GE Rotterdam, The Netherlands;
[5]IZKF Computational Biology Research Group, RWTH Aachen University Medical School, 52062 Aachen, Germany.

†Corresponding authors.

Manuscript in preparation

Chapter 6

Dynamics of the LBD1 complex and the activation of hematopoietic development

Petros Kolovos[1*], Andrea Martella[1*], Mary Stevens[1], Guillaume Giraud[1], Niels Galjart[1], Wilfred van IJcken[2] , Charlotte Andrieu-Soler[1,3]†, Frank Grosveld[1]†

[1]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands.
[2]Center for Biomics, Erasmus Medical Center, Rotterdam, The Netherlands.
[3]Inserm UMRS872, Institut de Recherche des Cordeliers, Fontenay-aux-Roses, Paris, France

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

Manuscript in preparation

Chapter 7

Dynamics of essential transcription factors during erythroid differentiation

Petros Kolovos[1], Jan Christian Bryne[2], Guillaume Giraud[1], Ernie de Boer[1], Mary Stevens[1], Boris Lenhard[2,4], Wilfred van IJcken[3], Charlotte Andrieu-Soler[1,5], Eric Soler[1,5]† and Frank Grosveld[1]†

[1]Department of Cell Biology, Erasmus Medical Center, Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands.
[2]Computational Biology Unit - Bergen Center for Computational Science and Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway
[3]Biomics Department, Erasmus Medical Center, Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands
[4]Current address: Institute of Clinical Sciences, MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK
[5]Current address: INSERM UMR967 and Laboratory of Excellence GR-Ex, CEA/DSV/iRCM, Fontenay-aux-Roses, France

†Corresponding authors.

Manuscript in preparation

Chapter 8

General Discussion

Summary

Samenvatting

Curriculum Vitae

PhD Portfolio

Acknowledgements

Summary

The genome contains all the necessary information for all developmental processes required for the proper survival of every living species. Genes, which are located on the chromatin fiber, have an important role in those processes. They are regulated by regulatory elements and other genes and the genome has to shape in specific conformations to fit inside the nucleus and to tether specific regulatory elements to their target genes. Although the linear composition of many genomes is largely known, their three dimensional (3D) organization and dynamics are largely unknown. Although it was known that genome conformation was important for the control of all the complex developmental processes, our knowledge is far from complete. Hence the main objectives of this thesis were: 1) to study the genome conformation/interactome and their effect on gene regulation and 2) to unveil the role of transcription factor proteins (TFs) in complex developmental processes.

In order to unveil the interactome of some genomic regions with the rest of the genome in a few hundred Kb radius, a 3C variant was developed, called 3C-seq. It provides information about the interactome of either one (3C-seq) or many (multiplexed 3C-seq) viewpoints with their regulatory elements and the rest of the genome. In **Chapter 2** we described the development of 3C-seq and in **Chapter 4** and **Chapter 5** we provide examples of its application.

In order to overcome the limitation of the different 3C variant methods, Targeted Conformation Capture (T2C) was next developed. T2C provides the genome wide interactome network for a selected region of (usually) up to 5Mb, which will comprise several Topological Associated Domains (TADs) as an alternative to Hi-C. It requires low sequencing depth (up to 1/10 of a sequencing lane) compared to other methods and can be multiplexed. Due to the high coverage that T2C offers, it yields high signal to noise ratios resulting to absolute restriction fragment without the requirement of binning the reads of the fragments. As a result, T2C provides high resolution, high coverage mapping of TADs and their interactions and boundaries. Due to the low number of PCR cycles it also provides more “quantifiable” conclusions. Hence, T2C is a cost effective tool to study the local spatial organization of the genome and its interactome with high resolution. **Chapter 3** describes the development of T2C and **Chapter 5** provides examples of its application. The high resolution and coverage of T2C allows an analysis of how the genome is shaped (unpublished data). *In vivo* and *in vitro* T2C data, suggest that the rosette like *MLS* model is the most likely architecture, with ~5±1 nucleosomes/11nm. Comprehensive analysis postulates that the loops of the rosettes of the *MLS* model are in the range of 30-100Kb and the loops of a rosette together form TADs of 1-2Mb.

Nucleosomes, the basic architectural blocks of the genome, rearrange along chromatin fiber and profoundly affect gene expression. **Chapter 4** shows how TNFα signaling reshapes the conformation of the genome to allow the establishment of new transcriptional programs. Intriguingly, TNFα triggers nucleosome repositioning prior to and independently of NFκB binding. However, approximately half of the nucleosome depleted regions contain typical NFκB motifs. Interestingly, we confirmed that nucleosome repositioning is accompanied by changes in the interactome of responsive genes, with new contacts appearing in nucleosome depleted regions enriched for NFκB and H3K4me1 binding. **Chapter 5** assesses the effect of TNFα signaling on the spatial conformation of the genome and the interactome for both TNFα responsive and non-responsive genes as well as the bimodal function of NFκB based on its consensus DNA motif.

Chapter 6, describes the role of the LDB1 complex and its dynamics in hematopoietic development (starting from the hemangioblast to pro-erythroblast and erythroblasts) and the dynamics of the switch from GATA2 to GATA1. Interestingly, during development the binding of LDB1 complex is redistributed to new genomic regions subsequently affecting its target genes. The first genes bound by the LDB1 complex for both the hemangioblast and pro-erythroblast stage are important for hematopoietic differentiation. Upon establishment of the hematopoietic lineage, the LDB1 complex is also located near genes necessary for erythroid/red blood cell differentiation and maintenance as well as heme

synthesis. The redistribution of the LDB1 complex during development is facilitated by “pioneering” TFs. We postulate the existence of two types of “pioneering” TFs based on the newly acquired data; ones which could attract the LDB1 complex to new genomic regions and others which stabilize it.
At later developmental stages (between the pro-erythroblasts and erythroblasts), GATA1 forms complexes with different TFs such as FOG1 and GFI1B in addition to the LDB1 complex (**Chapter 7**). The GATA1/FOG1 complex bound to regulatory elements at the pro-erythroblast stage, has repressive properties. In contrast GATA1/GFI1B bound regulatory elements at the erythroblast stage regulatory have activating properties. Those different properties of GATA1 when complexed with either FOG1 or GFI1B, demonstrate the dynamic equilibrium of a combinatorial TF network and we propose candidate TFs which may determine the equilibrium.
Finally **Chapter 8** has a general discussion of the findings presented in this thesis and highlights the important aspects of this work. Furthermore, I discuss the importance of understanding the spatial conformation of the genome and the importance of the dynamics of TFs during development and I propose future perspectives for the continuation and evolvement of the current findings.

Samenvatting

Het genoom bevat alle noodzakelijke informatie voor alle ontwikkelingsprocessen die nodig zijn voor in alle levende wezens. Genen, die onderdeel zijn van de chromatine vezel, spelen een belangrijke rol bij deze processen. Genen worden gereguleerd door regulerende elementen en andere genen en het genoom dient specifieke conformaties aan te nemen om in de kern te passen, en er zo voor te zorgen dat specifieke regulerende elementen naar hun doelgenen kunnen vinden. Ook al is de lineaire samenstelling van vele genomen grotendeels bekend, hun driedimensionale (3D) organisatie en dynamiek is grotendeels onbekend. Het was al enige jaren bekend was dat genoom conformatie belangrijk is voor de controle van alle complexe ontwikkelingsprocessen, maar onze kennis hierover is verre van volledig. De hoofddoelstellingen van deze scriptie waren daarom: 1) het bestuderen van de genoom conformatie/interactoom en het effect daarvan op genregulatie en 2) de rol van bepaalde transcriptie factor eiwitten (TF) in complexe ontwikkelingsprocessen op te helderen.

Om het interactoom van enkele gebieden van het genoom met de rest van het genoom binnen een afstand van enkele honderden kb's op te helderen, is een 3C variant ontwikkeld; 3C-seq. Het geeft informatie over het interactoom van één (3C-seq) of een groter aantal (multiplex 3C-seq) posities met hun regulerende elementen en de rest van het genoom. De ontwikkeling van 3C-seq is beschreven In **hoofdstuk 2** we en **hoofdstuk 4** en **hoofdstuk 5** beschrijven voorbeelden van de toepassing hiervan.

Teneinde de beperking van de verschillende 3C varianten te verwijderen, werd hierna Targeted Conformation Capture (T2C) ontwikkeld. T2C biedt het genoomwijde interactoom netwerk voor een geselecteerde regio van (meestal) tot 5Mb, die verschillende Topological Associated Domains (TADs) zal omvatten, als een alternatief voor hi-C. Het vereist veel minder sequencing diepte (tot 1/10 van een sequencing laan) vergeleken met andere methoden en kan worden gemultiplexed. Door het hoog aantal sequenties per fragment in T2C levert het een hoge signaal-ruisverhouding per restrictiefragment zonder de noodzaak van binning de "reads". T2C heeft daarom een hoge resolutie en een hoge dekkingsgraad bij het in kaart brengen van TADs, en hun interacties en grenzen. Door het lage aantal PCR-cycli dat nodig is, geeft het bovendien meer "kwantificeerbare" conclusies. T2C is daarom een kosteneffectief middel om de lokale ruimtelijke organisatie van het genoom en zijn interactoom met hoge resolutie te bestuderen. **Hoofdstuk 3** beschrijft de ontwikkeling van T2C en **hoofdstuk 5** geeft voorbeelden van de toepassing hiervan.

De hoge resolutie en dekking van T2C maakt een analyse van hoe het genoom mogelijk is gevouwen. *In vivo* en *in vitro* T2C data suggereren dat het rozet-achtige *MLS* model de meest waarschijnlijke architectuur is, met ~ 5 ± 1 nucleosomen / 11nm. Uitgebreide analyse postuleert dat de lussen van een rozet in het *MLS* model ergens tussen de 30-100Kb zijn en de lussen van een rozet samen een TAD vormen van tussen 1-2 MB.

Nucleosomen, de fundamentele bouwstenen van het genoom, herschikken zich langs chromatine vezel en hebben een sterke invloed op genexpressie. **Hoofdstuk 4** laat zien hoe TNFα signalering de conformatie van het genoom verandert om zo nieuwe transcriptie programma's mogelijk te maken. Fascinerend is dat TNFa de nucleosoom herpositionering teweeg brengt voorafgaande aan en onafhankelijk van NFκB binding. Echter, slechts ongeveer de helft van de nucleosoom vrije gebieden bevat een typisch NFκB motief. Interessant genoeg kon bevestigd worden dat nucleosoom herpositionering wordt vergezeld door veranderingen in het interactoom van responsieve genen, met nieuwe contacten die verschijnen in nucleosoom verarmde gebieden, die verrijkt zijn met NF-κB en H3K4me1 binding. **Hoofdstuk 5** stelt het effect van TNFα signalering vast op de ruimtelijke ordening van het genoom en het interactoom voor zowel op TNFα reagerende genen als niet op TNFα reagerende genen als ook de bimodale functie van NFκB gebaseerd op zijn DNA motief.

Hoofdstuk 6 beschrijft de rol van het LDB1 complex en de dynamiek van de hematopoietische ontwikkeling; beginnend vanaf de hemangioblast tot pro-erythroblast en erythroblasten – en de dynamiek van de switch van "GATA2 naar GATA1"; een omschakeling van GATA2 naar GATA1. Interessant is dat

tijdens de ontwikkeling, de binding van het LDB1 complex herverdeeld wordt over nieuwe (en oude) genomische regio's, met een effect op de expressie van die nieuwe target genen. De eerste genen die het LDB1 complex bindt tijdens de hemangioblast / pro-erythroblast fase zijn belangrijk voor hematopoietische differentiatie. Tijdens de geboorte van de hematopoietische lichaamscellen, is het LDB1 complex ook gelokaliseerd vlakbij genen die nodig zijn voor erythroïde/rode bloedcel differentiatie en in stand houding alsook de synthese van haem. De herverdeling van het LDB1 complex tijdens de ontwikkeling wordt gefaciliteerd door "pioneer" TFs. Ik postuleer het bestaan van twee soorten "pioneer" TFs op basis van de nieuw verkregen kennis; een groep die het LDB1 complex naar nieuwe genomische regio kan rekruteren en een groep die het LDB1 complex stabiliseert.

In latere ontwikkelingsstadia van pro-erythroblasten naar erythroblasten, vormt GATA1 complexen met verschillende TFs zoals FOG1 en GFI1B naast het LDB1 complex (**hoofdstuk 7**). Wanneer Het GATA1 / FOG1 complex, gebonden is aan regulerende elementen in het pro-erythroblast stadium, heeft het repressieve eigenschappen . Dit in tegenstelling tot GATA1/GFI1B gebonden regulerende elementen tijdens het erythroblast stadium, dat juist activerende eigenschappen heeft. Deze verschillende eigenschappen van GATA1 als het een complex vormt met ofwel FOG1 ofwel GFI1B, laat het dynamisch evenwicht zien van een combinatorisch TF netwerk en we stellen kandidaat TF's voor die dit evenwicht mogelijk bepalen.

Tenslotte heeft **hoofdstuk 8** een algemene discussie van de bevindingen gepresenteerd in dit proefschrift en highlights de belangrijke aspecten van dit werk. Verder bespreek ik het belang van het begrijpen van de ruimtelijke conformatie van het genoom en het belang van de dynamiek van de TF's tijdens de ontwikkeling. Ten slotte beschrijf ik een aantal toekomstperspectieven voor de continuering en de ontwikkeling van de huidige bevindingen.

Curriculum Vitae

Name: Petros Kolovos
Date of birth: 28 July 1984
Place of birth: Thessaloniki, Greece
Nationality: Greek

EDUCATION

2010-2015 **PhD Student**
Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands

2009-2010 **Master by Research in Reproductive Biology**
University of Edinburgh, Scotland, UK

2002-2009 **BSc in Biology**
Aristotle University of Thessaloniki, Greece

TRAINING

2010-2015 **PhD Research**
Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands
Promotor: Professor Dr Frank Grosveld
Subject: *3D organization of the genome and dynamics of TFs in development*

2009-2010 **MSc research**
MRC Centre of Regenerative Medicine, University of Edinburgh, Scotland, UK
Supervisor: Dr Paul de Sousa
Subject: *Isolation of Oct-4 promoter active cell populations from adult mouse tissues by selective ablation*

2007-2009 **BSc research**
Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, Greece
Supervisor: Professor George Thomopoulos
Subject: *Development of Genetic Database of Decapoda, Aves and Reptilia of Greece*

2006-2007 **BSc research/Practical Exercise**
National Agricultural Research Foundation, Thessaloniki, Greece
Supervisor: Dr Emmanuel Vainas
Subject: *Evaluation of fertilizing ability of mammalian spermatozoa*

PUBLICATIONS

1. **Kolovos P**, Nikolic M, Koeferle A, Larkin J, van IJcken WF, Gusma E, Costa I, Knoch T, Cook PR, Grosveld F, Papantonis A. The bimodal function of NFκB and the effect of TNFα in the spatio-temporal landscape of the genome
(Manuscript in preparation)
2. **Kolovos P**, Bryne JC, Giraud G, de Boer E, Stevens E, van IJcken WF, Soler E, Andrieu-Soler E, Grosveld F. Dynamics of essential transcription factors during erythroid differentiation.
(Manuscript in preparation)

3. **Kolovos P***, Martella A*, Stevens M, Giraud G, Galjart N, van IJcken WF, Andrieu-Soler C, Grosveld F. Dynamics of the LBD1 complex and the activation of hematopoietic development. (Manuscript in preparation)
***Equal contribution**

4. Knoch T, Wachsmuth M, Kepper N, Lesnussa M, Abuseiris A, Imam A, **Kolovos P**, Zuin J, Kockx C, Brouwer RWW, Harmen J. G. van de Werken HJ, van IJcken WF, Wendt KS, Grosveld F. The Detailed 3D Multi-Loop Aggregate/Rosette Chromatin Architecture and Functional Dynamic Organization of the Human and Mouse Genomes.
(Manuscript in preparation)

5. Stadhouders R, Cico A, Stephen T, Thongjuea S, **Kolovos P**, Baymaz I, Yu X, Demmers J, Bezstarosti K, Maas A, Barroca V, Kockx C, Ozgur Z, van IJcken WF, Arcangeli ML, Andrieu-Soler C, Lenhard B, Grosveld F, Soler E. (2014) Control of developmentally poised erythroid genes by combinatorial corepressor actions. *Nature Communications*. (In revision)

6. Caputo L, Witzel HR, **Kolovos P**, Cheedipudi S, Looso M, Mylona A, van IJcken WF, Laugwitz KL, Evans SM, Braun T, Soler E, Grosveld F, Dobreva G. The Isl1/Ldb1 complex orchestrates heart-specific chromatin organization and transcriptional regulation. *Cell stem Cell.* 2015 Sept 3; 17, 1-13 .

7. Tresini M, Warmerdam DO, **Kolovos P**, Snijder L, Vrouwe MG, Demmers JA, van IJcken WF, Grosveld FG, Medema RH, Hoeijmakers JH, Mullenders LH, Vermeulen W, Marteijn JA. The core spliceosome as target and effector of non-canonical ATM signalling. *Nature*. 2015 Jul 2;523(7558):53-8.

8. Diermeier S*, **Kolovos P***, Heizinger L, Schwartz U, Georgomanolis T, Zirkel A, Wedemann G, Knoch TA, Grosveld F, Merkl R, Cook PR, Längst G, Papantonis A. TNFalpha signalling primes chromatin for NF-kappaB binding and induces rapid and widespread nucleosome repositioning. *Genome Biology*. 2014 Dec 3;15(12):536.
***Equal contribution**

9. **Kolovos P**, van de Werken HJ, Kepper N, Zuin J, Brouwer RW, Kockx CE, Wendt KS, van IJcken WF, Grosveld F, Knoch TA. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin.* 2014 Jun 16;7:10.

10. Stadhouders R, de Bruijn MJ, Rother MB, Yuvaraj S, Ribeiro de Almeida C, **Kolovos P**, Van Zelm MC, van Ijcken W, Grosveld F, Soler E, Hendriks RW. Pre-B cell receptor signaling induces immunoglobulin κ locus accessibility by functional redistribution of enhancer-mediated chromatin interactions. *PloS Biol.* 2014 Feb 18;12(2).

11. Zuin J, Dixon JR, van der Reijden MI, Ye Z, **Kolovos P**, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IJcken WF, Grosveld FG, Ren B, Wendt KS. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *PNAS.* 2014 Jan 21;111(3):996-1001.

12. 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 Mar;8(3):509-24.
***Equal contribution**

13. 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 Jul-Aug;3(4):181-6.

14. **Kolovos P,** Knoch TA, Grosveld FG, Cook PR, Papantonis A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin*. 2012 Jan 9;5(1):1.

15. van den Driesche S, **Kolovos P**, Platts S, Drake AJ, Sharpe RM. Inter-relationship between testicular dysgenesis and Leydig cell function in the masculinization programming window in the rat. *PLoS One*. 2012;7(1).

PhD Portfolio

Name PhD student:	Petros Kolovos
Department:	Cell Biology
Period:	Sept 2010 - Sept 2015
Promoter:	Prof.dr. F.G. Grosveld

PhD Training

Courses

2011	Biobase course: Functional annotation of experimental data using TRANSFAC
2011	Safety working in the laboratory
2011	Development, Stem Cells and Disease
2011	Technology facility
2010-2011	Molecular and Cell Biology (PhD teaching programme) Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands

(Inter)national conferences and workshops

2015	25th MGC Symposium, Leiden, The Netherlands (*oral presentation*)
2014	IUAP DevRepair meeting, Rotterdam, The Netherlands (*oral presentation*)
2014	Chromatin and Epigenetics: From Omics to Single Cells, Strasbourg, France (*poster presentation*)
2013	RUNX1: Transcription factors in Disease & Development, Wilsede, Germany
2012	ICSB : The 13th International Conference on Systems Biology, Toronto, Canada (o*ral presentation/workshop*)
2012-2014	EpiGenSys/ERASynBio meetings (*3x oral presentation*)
2011	18th MGC PhD Workshop, Maastricht, The Netherlands

Acknowledgements

Here it is, the (final) product of the past few years. Although to some people it would seem that it is a task completed by one person, this is not the case. Completing a PhD thesis is not a one man's job but a team effort. During the past years, I had the extreme luck to, not only collaborate, but work closely together with many excellent scientists. Their collective contributions were of great importance to the fulfilment of this thesis. However, none of these would have been possible without Frank Grosveld, my supervisor and promotor.

Dear Frank, I would like to express my deepest gratitude for providing me the opportunity to perform my PhD research in your laboratory; an inspiring and challenging environment to explore new concepts. It has been a real honour and a privilege. You have been a great mentor, with your door always open for discussions, career tips, brainstorming, exploring great ideas and addressing novel concepts. Your comments and suggestions were always to the point and offered great guidance in addressing the scientific questions. I would like to thank you for all your valuable advice in the writing of my thesis, scientific papers and grant applications. I look forward to more exciting research together.

I would also like to thank all the members of the small/reading committee. Dear Danny, I am really grateful for your time and all the great comments and suggestions that you had. They were truly helpful. Also, I would like to thank you for the help, advice, discussions, career tips and invitation to the IUAP meeting. Your door was always open and that is something that I am grateful. I look forward helping you deciphering your favourite TF! Of course, thank you for bringing some of your amazing Belgian beers, it was really nice comparing them to the Dutch ones!

Dear Joost, I would like to thank you for agreeing to be part of my thesis committee, for taking time to read my thesis and your great advice for grant applications. I anticipate some really exciting results from your favourite Xist.

Dear Aki, I would like to thank you for accepting to be a member of my committee. Our collaboration over the past four years was really exciting with great outcomes and hopefully more cool things to come. Your advice was always accurate and you were always there to offer your insight. Thank you for the great discussions both in person and via skype. I wish all the best to your family and you and success in your new position in Cologne.

I also would like to express my gratitude to the members of my defence committee. Dear Sjaak, thank you for the interesting discussions over the years. They were really entertaining! Dear Niels, thank you for becoming a member of my defence committee and for the great PhD program you organized. Dear Kerstin, thank you both, for being part of my defence committee and for a great collaboration that yielded some really interesting results. I look forward to its continuation. Last but not least, dear Marieke thank you for accepting to participate in my committee and traveling to Rotterdam for my defence.

I would also like to acknowledge our foreign collaborators. Dear Prof. Peter Cook, thank you for the interesting meetings, discussions, insight into our work and valuable comments. I particularly enjoyed our discussions on "transcription factories". I am really grateful for all the help and look forward to some great science to come! Dear Gergana, Luca and Alvaro, thank you for great collaborations and scientific results; I look forward to their continuation!

My paranimfen: Maria, we have been friends for approximately thirteen years. It has always been fun, starting from our bachelor years when we used to have coffee and go for trips and throughout our master and PhD studies. You always have something nice to say and can make people laugh. I wish you good luck with your thesis; it is your turn now and all the best in the future. Guillaume, "monsieur", working together in the same lab was really fun! We got along well from the beginning and became not only colleagues, but friends. I am really grateful for that! I really enjoyed our discussions both scientific and not, as well as our "fights" about how the meat should be "properly" cooked! I wish you all the best!

This work could not have been completed without all the colleagues of the Cell Biology and other departments. I would like to thank all the people of the department and particularly, the members of the Grosveld lab during the last years! Anita (ChIP guru; I wish you all the best in Leiden; see you soon), Ralph (thanks for the advice on the different ways to eat pindakaas; good luck in Barcelona), RJ (the 3C guru; thank you for the great advice and discussions in the lab, good luck at 6th floor), Andrea M. (EB expert), Xiao (FACS support), Andrea A. (the Zeb expert; always fun to discuss in the lab), Mary (the Chuck Norris of the lab), Ernie (if you have a problem, she knows how to fix it), Ali and Farzin (they made HEP differentiation look easy), Ruud (the recombineering expert), Rick (thanks for all the fun chats), Dubi (made the tram travels more interesting) and Michael/Rin (thanks for all the help), thank you all! Also Eric and Charlotte, I am grateful for all the great discussions and collaboration over the years. You taught me the "secrets" of many techniques; thank you for that! I wish you all the best in Paris, with more great science to come! Tobias, thank you for the nice biophysical insight that you have provided. Ali, I have really enjoyed our philosophical discussions, thank you for that!

The Raymond lab; Raymond thank you for the interesting discussions and great to the point comments that you provided. Marti (Barcelona for ever!), Maaike (I wish you all the best with your thesis), Johan (thanks for the nice bioinf discussions) and Mike (the enzyme supplier), thank you all. The Philipsen lab; Thamar, Ileana, Nynke and Silvia thank you for all the great time! Harmen, I am really grateful for all the bionformatical support; you provided me some great insight into that world! Luca the fellow admin, I wish you all the best with your PhD! Jessica, the CTCF/cohesin expert, thanks for all the great chats and thesis advice. Emma, Michaela, Christina and the "ten Berge lab", thank you all! Agnese and Aristea, I wish you all the best with your thesis and your next steps! I would also like to aknowledge the assistance from the secretaries (Marike, Bep, Jasperina and Sonja) as well as the IT guys; was really important during the last years. Thank you all!

From the Genetics department, Wim and Jurgen thank you for the great collaboration and I look forward to some new exciting results. Maria, our discussions were always stimulating and fun and extended to a really cool collaboration. I have really enjoyed working with you; thank you for everything!

This PhD thesis, would have not been completed without the great collaboration with the Biomics department. Wilfred, I really thank you for all the support and willingness to help us with all our crazy ideas and suggestions. Rutger, I am deeply grateful for all the bionformatical help that you provided and the really great discussions about which side of the moon is actually dark! Mirjam and Frank, thank you for the great assistance! Christel, Zeliha Xander and Edwin, thank you for the excellent technical support and never ending willingness to help us complete our projects.

I would like to thank Prof. George Thomopoulos; your suggestions and advice to broaden my horizons were really important and I am grateful for that.

I would like to thank my uncle for his never ending support and long skype discussions about all different topics, which made a lot of things easier for me. I am also deeply grateful to my family, my father, my mother and my sister; your constant support and education was instrumental and shaped me in every step. You have helped me broaden my horizons. Without those things, I could not have been here.

Finally, I would like to thank my wife. Lena, all these years you have been a silent force, a solid rock next to me. You dared to follow me in this adventure and you were always there for me. Your never ending support was instrumental and nothing would have been possible without your help. I thank you for everything…

Petros