The three-dimensional regulatory landscapes of the globin genes

A thesis submitted for the degree of Doctor of Philosophy of the University of Oxford



A. Marieke Oudelaar
Green Templeton College

Wellcome Trust Doctoral Training Programme in Genomic Medicine and Statistics
Weatherall Institute of Molecular Medicine
University of Oxford

Trinity Term 2018

Abstract

One of the most important outstanding questions in biology involves the precise spatial and temporal regulation of gene activity, which enables different cell types to express the specific set of genes required for their function and is therefore a cornerstone for complex biological life. *Cis*-regulatory elements, such as gene promoters and enhancers, play a key role in controlling gene activity. These elements physically interact with the genes they regulate within structural chromatin domains. The organisation of chromosomes into these domains is critical for specific regulation of gene expression and disruption of these structures underlies common human disease. However, it is not understood how chromatin domains form, how interactions between the *cis*-regulatory elements contained within them are established, or how such interactions influence gene expression.

The major hurdles in addressing these questions are to determine chromatin structures with high resolution and sensitivity and to examine their dynamic nature within single cells. To overcome these, I have developed Tri-C, a new chromosome conformation capture assay that can analyse concurrent chromatin interactions at single alleles at high resolution. By combining Tri-C with conventional chromosome conformation capture techniques, I have analysed the three-dimensional regulatory landscapes of the well-characterised murine globin loci at unprecedented depth. Additionally, to examine the roles of *cis*-regulatory elements in establishing chromatin architecture, I have analysed how engineered deletions in enhancers and CTCF-binding elements in the murine α-globin locus disrupt its chromatin landscape.

These analyses reveal that the chromatin domains containing the globin genes represent compartmentalised structures, which are delimited by CTCF boundaries. The heterogeneity of interactions in these domains between individual cells is indicative for a dynamic process of loop extrusion underlying their formation. Within chromatin domains, preferential structures are formed in which multiple enhancers and promoters interact simultaneously. These complexes provide a structural basis for understanding how multiple *cis*-regulatory elements cooperate to establish robust regulation of gene expression. Importantly, these complex, tissue-specific structures, cannot be explained by loop extrusion alone and indicate other, independent mechanisms contributing to chromosome architecture, likely involving interactions mediated by multi-protein complexes. Together, these analyses show that the current view of genome organisation, in which chromosomes are organised by stable loops and domains that self-assemble into hierarchical structures, is not correct. Rather, chromatin architecture reflects a complex interplay between distinct molecular mechanisms contributing to the formation of regulatory landscapes that facilitate precise, robust control of gene expression.

Acknowledgements

First and foremost, I would like to express my gratitude to Jim Hughes and Doug Higgs, for giving me the opportunity to work in their groups and for all their support, advice and encouragement. It has been a privilege to work under their supervision.

I would also like to thank all the members of the Hughes and Higgs groups, for their help and expertise and the many valuable discussions of our projects. I feel very fortunate to have worked in such a stimulating research environment. In particular, I would like to thank James Davies, for teaching me the basics of Capture-C, Perl, and erythropoiesis, and Lars Hanssen, for our joint efforts to characterise the α-globin locus by Capture-C. I have also valued the discussions with Jill Brown and Veronica Buckle about their work to characterise the structure of the α-globin locus by complementary microscopy-based approaches. Damien Downes and Jelena Telenius have been very helpful and provided support in the lab and with bioinformatic analyses, respectively. Finally, I would like to thank Martin Larke, for sharing our DPhil ups and downs over a good cup of coffee.

I would like to acknowledge Rob Klose and Tom Milne for their advice on my work during our thesis committee meetings. I am also very grateful to Green Templeton College, for providing a friendly and supportive environment for both academic and social life. In particular, I would like to thank Marella De Bruijn, my college advisor, for our valuable conservations and enjoyable dinners. Lastly, I would like to express my gratitude to my family and friends, especially to Bastian Betthäuser, for their support and encouragement.

This work was made possible by generous funding from the Wellcome Trust and the Medical Research Council UK.

Table of Contents

ABSTRACT 2

ACKNOWLEDGEMENTS 3

TABLE OF CONTENTS 4

LIST OF FIGURES 6

LIST OF ABBREVIATIONS 8

CHAPTER 1 – INTRODUCTION 10

SPATIO-TEMPORAL CONTROL OF GENE EXPRESSION 10
GENOME ORGANISATION 12
CHROMOSOME CONFORMATION CAPTURE 19
THE MURINE GLOBIN LOCI 22
THESIS OUTLINE 29
NOTES 30
AVAILABILITY 30

CHAPTER 2 – THE STRUCTURAL ORGANISATION OF THE MURINE GLOBIN LOCI 31

INTRODUCTION 31
RESULTS 31
Characterisation of the murine globin loci by Hi-C *31*
High-resolution analysis of the chromatin domains containing the murine globin loci by Capture-C *37*
DISCUSSION 44
DATA AVAILABILITY 47
ACKNOWLEDGEMENTS 47

CHAPTER 3 – THE ROLES OF *CIS*-REGULATORY ELEMENTS IN ESTABLISHING CHROMATIN ARCHITECTURE 48

INTRODUCTION 48
RESULTS 52
CTCF-binding elements in the α-globin locus mediate the formation of a chromatin domain that contributes to enhancer specificity *52*
Enhancers are not required for the formation and maintenance of the α-globin chromatin domain *56*
The structure of the α-globin domain is not dependent on active transcription *58*
DISCUSSION 59
DATA AVAILABILITY 61
ACKNOWLEDGEMENTS 61

CHAPTER 4 – THE THREE-DIMENSIONAL CONFORMATION OF THE MURINE GLOBIN LOCI IN SINGLE CELLS 62

INTRODUCTION 62
RESULTS 63
New 3C techniques to analyse multi-way chromatin interactions in single cells *63*
Multiple enhancers and promoters form higher-order chromatin structures in which they interact simultaneously *68*
CTCF-binding elements form dynamic interactions supportive of a loop extrusion mechanism underlying chromatin domain formation *73*
DISCUSSION 76
DATA AVAILABILITY 78
ACKNOWLEDGEMENTS 78

CHAPTER 5 – CAPTURING CHROMOSOME CONFORMATION IN SMALL NUMBERS OF CELLS ____ 79
INTRODUCTION ____ 79
RESULTS ____ 79
Approach for the development of a low-input 3C technique ____ 79
Optimisation of 3C library preparation in small numbers of cells ____ 80
Development of low-input Capture-C library preparation approaches ____ 81
Integration of optimised procedures into a low-input Capture-C protocol ____ 90
Bioinformatic analysis of low-input Capture-C data ____ 92
DISCUSSION ____ 93
DATA AVAILABILITY ____ 94
ACKNOWLEDGEMENTS ____ 95

CHAPTER 6 – DISCUSSION ____ 96
A REFINED VIEW OF GENOME ORGANISATION ____ 96
MOLECULAR MECHANISMS UNDERLYING CHROMATIN ARCHITECTURE ____ 98
COOPERATION BETWEEN *CIS*-REGULATORY ELEMENTS ____ 103

CHAPTER 7 – METHODS ____ 105
CELLS ____ 105
Erythroid cells derived from adult mouse spleens ____ 105
Embryonic stem cells ____ 105
Erythroid cells derived from foetal livers ____ 106
HI-C ____ 106
CAPTURE-C ____ 106
3C library preparation ____ 107
Capture-C library preparation ____ 108
Enrichment for fragments of interest by oligonucleotide capture ____ 110
Viewpoints ____ 111
TRI-C ____ 115
C-TRAP ____ 117
LOW-INPUT CAPTURE-C ____ 117
3C library preparation ____ 118
Capture-C library preparation – Low-Input Capture-C ____ 120
Capture-C library preparation – Tag-Capture-C ____ 121
Enrichment for fragments of interest by oligonucleotide capture ____ 123

CHAPTER 8 – DATA ANALYSIS ____ 126
SOFTWARE PACKAGES ____ 126
STATISTICAL ANALYSES ____ 126
HI-C ____ 127
CAPTURE-C ____ 128
TRI-C ____ 129
C-TRAP ____ 131
LOW-INPUT CAPTURE-C ____ 132
AVAILABILITY ____ 132
Custom analysis scripts ____ 132
Sequencing data ____ 133
UCSC Genome Browser Track Hubs ____ 133

REFERENCES ____ 134

List of Figures

Chapter 1 – Introduction

Figure 1.1: Physical aspects of chromosomal interactions. 13
Figure 1.2: Layers of genome organisation. 16
Figure 1.3: Formation of self-interacting domains by stable chromatin loops. 17
Figure 1.4: Formation of self-interacting domains by a dynamic loop extrusion mechanism. 18
Figure 1.5: Chromosome conformation capture techniques. 20
Figure 1.6: Expression of the globin genes during erythropoiesis. 23
Figure 1.7: The murine α-globin locus. 24
Figure 1.8: The murine β-globin locus. 25
Figure 1.9: Transcription factor and RNA polymerase II occupancy at the active murine globin loci. 27
Figure 1.10: The active chromatin hub model. 28

Chapter 2 – The structural organisation of the murine globin loci

Figure 2.1: Hi-C analysis of the extended α-globin locus. 32
Figure 2.2: Hi-C analysis of the extended β-globin locus. 33
Figure 2.3. The Strings & Binders Switch Model. 35
Figure 2.4: SBS models of the extended α-globin locus. 36
Figure 2.5. SBS models of the extended β-globin locus. 37
Figure 2.6: Characterisation of the interaction landscapes of the *cis*-regulatory elements of the α-globin locus. 38
Figure 2.7: Structural conformations of the active and inactive α-globin domain. 39
Figure 2.8: Characterisation of the interaction landscapes of the *cis*-regulatory elements of the β-globin locus. 41
Figure 2.9: Structural conformations of the active and inactive β-globin domain. 42
Figure 2.10: SBS models of the α-globin domain in its active and inactive conformation. 43
Figure 2.11: SBS models of the β-globin domain in its active and inactive conformation. 44

Chapter 3 – The roles of *cis*-regulatory elements in establishing chromatin architecture

Figure 3.1: Schematic overview of the D3839 mouse model. 49
Figure 3.2: Schematic overview of the enhancer knockout mouse models. 50
Figure 3.3: The effects of deletions of enhancer elements on α-globin expression. 51
Figure 3.4: The effects of a CTCF boundary deletion on the interaction landscape of the α-globin locus. 53
Figure 3.5: The effects of a CTCF boundary deletion in the α-globin locus on local gene expression. 55
Figure 3.6: The effects of deletions of enhancer elements on the interaction landscape of the α-globin locus. 57
Figure 3.7: Transcription is not required for maintenance of the α-globin domain structure. 59

Chapter 4 – The three-dimensional conformation of the murine globin loci in single cells

Figure 4.1: Overview of the experimental procedure and data output of C-Trap. 64
Figure 4.2: Overview of the experimental procedure and data output of Tri-C. 65
Figure 4.3: Validation of pair-wise interactions detected by Tri-C. 66
Figure 4.4: Validation of multi-way interactions detected by Tri-C. 67
Figure 4.5: Interactions detected by Tri-C are allele-specific. 68

Figure 4.6: Analysis of multi-way interactions between enhancers and promoters in the α-globin locus. 70
Figure 4.7: Tri-C data are highly reproducible. 71
Figure 4.8: Analysis of multi-way interactions between enhancers and promoters in the β-globin locus. 72
Figure 4.9: Analysis of multi-way interactions between CTCF-binding elements in the α-globin locus. 74
Figure 4.10: Analysis of multi-way interactions between CTCF-binding elements in the β-globin locus. 75

Chapter 5 – Capturing chromosome conformation in small numbers of cells

Figure 5.1: Overview of the Capture-C procedure. 80
Figure 5.2: 3C library preparation efficiency. 81
Figure 5.3:Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the α-globin locus. 83
Figure 5.4: Reproducibility of LI-Capture-C and Tag-Capture-C interaction profiles in biological replicates. 84
Figure 5.5: Correlation between interaction profiles of the α-globin locus generated by different Capture-C protocols. 86
Figure 5.6: Windowed interaction profiles of individual LI-Capture-C and Tag-Capture-C replicates. 87
Figure 5.7: Comparison of characteristics of LI-Capture-C and Tag-Capture-C libraries. 89
Figure 5.8: Interaction profiles generated with LI Capture-C using 10,000-100,000 cells. 90
Figure 5.9:Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the β-globin locus. 91
Figure 5.10: Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the Mitoferrin-1 locus. 92
Figure 5.11: Overview of Capture-C data analysis. 93

Chapter 6 – Discussion

Figure 6.1: Proposed model for the formation of chromatin structures. 100

Chapter 7 – Methods

Figure 7.1. Quality control of 3C libraries. 108
Figure 7.2. Quality control of Capture-C libraries. 109
Figure 7.3. Quality control of Capture-C DNA products. 111
Figure 7.4. Tri-C quality control. 116
Figure 7.5. Quality control of low-input 3C libraries. 120
Figure 7.6. Quality control of LI-Capture-C libraries. 121
Figure 7.7. Quality control of Tag-Capture-C libraries. 123
Figure 7.8. Quality control of low-input Capture-C DNA products. 125

List of Abbreviations

3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
AID	Auxin-inducible degron
ATAC	Assay for transposase-accessible chromatin
BFU-e	Burst-forming unit-erythroid
bp	base pair
CD	Cluster of differentiation
CDK9	Cyclin-dependent kinase 9
CFU-e	Colony-forming unit-erythroid
ChIP	Chromatin immunoprecipitation
chr	chromosome
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
DHS	DNaseI hypersensitive site
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DNaseI	Deoxyribonuclease I
dNTP	deoxyribonucleotide triphosphate
DRB	5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole
ENCODE	Encyclopedia of DNA elements project
ES	Embryonic stem (cell)
F1	Filial 1 (hybrid)
FACS	Fluorescence-activated cell sorting
FBS	Foetal bovine serum
FISH	Fluorescence in situ hybridization
GAM	Genome architecture mapping
Gata1	GATA-binding factor 1
GWAS	Genome-wide association study
H3K4/27	Histone H3 lysine position 4/27
HS	Hypersensitive site (nomenclature for *cis*-regulatory elements of β-globin)
kb	kilo base pair
Klf1	Krüppel-like factor 1
LAD	Lamina-associated domain
LCR	Locus control region

LI-Capture-C	Low-input Capture-C
MACS	Magnetic activated cell sorting
Mb	Mega base pair
MCS	Multispecies conserved sequence
Med1	Mediator complex subunit 1
mRNA	messenger RNA
ns	non-significant
Nfe2	Nuclear factor erythroid 2
Nipbl	Nipped B-like protein
NG-Capture-C	Next-generation Capture-C
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PolII	RNA polymerase II
P-TEFb	Positive transcription elongation factor b
qPCR	quantitative polymerase chain reaction
R	Regulatory element (nomenclature for *cis*-regulatory elements of α-globin)
RNA	Ribonucleic acid
RNase	Ribonuclease
RPKM	Reads per kb per million mapped
rpm	revolutions per minute
Scl	Stem cell leukaemia factor
SDS	Sodium dodecyl sulfate
Smc1	Structural maintenance of chromosomes 1
SNP	Single nucleotide polymorphism
TAD	Topologically associating domain
TALEN	Transcription activator-like effector nucleases
Tris	Tris(hydroxymethyl)aminomethane
U	Unit
UCSC	University of California Santa Cruz
UMI	Unique molecular identifier
WT	Wild type

Chapter 1 – Introduction

Spatio-temporal control of gene expression

Human bodies contain approximately 37 trillion cells (Bianconi et al., 2013) of hundreds of different types and subtypes (Macosko et al., 2015). The information required for the specification and maintenance of these cells is stored in the human genome, which contains ~22,000 protein-coding genes (Pertea and Salzberg, 2010). As this gene number is very similar to that of much simpler organisms such as *C. elegans* (Chen, 2004), the complexity of our species is likely not encoded by these genes alone. Despite a similar number of genes, the human genome is more than an order of magnitude larger than that of *C. elegans*. It contains approximately three billion base pairs, of which only 1.2% constitute protein-coding genes (Lander et al., 2001). Characterisation of the remaining non-coding DNA sequences has revealed a myriad of genomic elements with predicted regulatory functions (ENCODE Project Consortium, 2012). This implies that the complexity of multi-cellular organisms depends on the pleiotropic use of their genes in different cell types, tissues and developmental stages. Precise spatio-temporal regulation of gene expression is therefore a cornerstone for complex biological life.

The regions of non-coding DNA implicated in the control of gene activity are called *cis*-regulatory elements, with the prefix *cis* (‘on this side of’) referring to their position on the same chromosome as the regulated gene. The exact number of these elements in the human genome is not clear. Chromatin profiling has assigned a biochemical function for ~80% of the genome (ENCODE Project Consortium, 2012), though sequence conservation has revealed that only ~5.5% is under positive selection (Lindblad-Toh et al., 2011). As biochemical modification does not necessarily imply biological function, and maintenance of functional *cis*-regulatory elements might not require strict sequence conservation (Ludwig et al., 2000), the answer might be somewhere in between these estimates. Despite the fact that their exact abundance remains unknown, it is clear that *cis*-regulatory elements vastly outnumber protein-coding genes and perform a wide range of critical functions. Different classes of *cis*-regulatory elements have been described, which can essentially be divided into promoters, enhancers and boundary elements.

Promoters are DNA sequences that are typically located directly upstream of the 5’ transcription start sites of protein-coding genes (Maxam et al., 1977). These elements contain binding motifs for the general transcription factors, which recruit RNA Polymerase II to form

the pre-initiation complex and start transcription (Sainsbury et al., 2015). In contrast to promoters, enhancers are scattered around the genes they regulate (Sanyal et al., 2012), and can be located more than a million base pairs up- or downstream (Lettice, 2003). They are defined as short regions of DNA that can activate transcription of the gene they regulate, independent of their relative position, distance or orientation relative to the gene promoter (Banerji et al., 1981). They often contain multiple binding sites for tissue-specific transcription factors. These factors can recruit co-activator proteins including the Mediator complex, which are thought to promote the formation of the pre-initiation complex to initiate transcription and/or overcome RNA polymerase II pausing (Heinz et al., 2015; Long et al., 2016). Many important developmental genes are regulated by clusters of enhancer elements which bind high levels of tissue-specific transcription factors and Mediator. It has been proposed that such clusters form a special class of enhancers called 'super-enhancers' with a distinct role in driving the high levels of expression of critical genes that define cell identity (Whyte et al., 2013). However, at the moment there is no conclusive evidence that such super-enhancers have a functional role that is different from the combined effects of several strong regular enhancer elements (Hay et al., 2016). Effective communication between enhancers and the promoters of the genes they control is regulated by boundary or insulator elements (Kellum and Schedl, 1991), which bind architectural proteins that contribute to the formation of specific structural enhancer-promoter complexes (Bickmore and van Steensel, 2013).

These *cis*-regulatory elements are associated with specific chromatin signatures. Promoters and enhancers are typically located in regions with increased chromatin accessibility in the cells in which they are active. Active promoters are associated with H3K27 acetylation and H3K4 di- and tri-methylation. Active enhancers are also marked with H3K27 acetylation, but are usually correlated with H3K4 mono-methylation and low levels of H3K4 di- and tri-methylation (Bannister and Kouzarides, 2011). Boundary elements are occupied by architectural proteins, such as the highly conserved zinc-finger protein CCCTC-binding factor (CTCF) (Chung et al., 1997), which plays a crucial structural role in mammalian gene regulation (Merkenschlager and Nora, 2016; Ong and Corces, 2014).

Accurate regulation of gene expression depends on robust and specific communication between these *cis*-regulatory elements. The structural organisation of the genome within the cell nucleus therefore plays a critical role in establishing precise, tissue-specific patterns of gene expression.

Genome Organisation

The need for a mechanism to structurally organise DNA in the cell nucleus is evident from the fact that the human genome spans about 2 meters, yet is contained within a nucleus that is only a few microns in diameter. At the smallest scale, DNA helices are wrapped around histone proteins and form nucleosomes every ~200 bp, creating ~11 nm beads-on-a-string structures (Kornberg and Lorch, 1999). Organisation at the scale of individual chromosomes has also been well characterised. Individual chromosomes occupy distinct chromosome territories in which they are folded in such a way that they more frequently interact with themselves in *cis* rather than with other chromosomes in *trans* (Cremer and Cremer, 2001). However, the intermediate layers of genome organisation that guide communication between genes and their *cis*-regulatory elements are less well understood.

The structural organisation of DNA within the cell nucleus is in part determined by the polymeric nature of chromosomes. Their physical properties influence which genomic elements have an opportunity to interact, the kinetics of these interactions and the number of cells in the population in which these interactions occur (Dekker and Mirny, 2016). Physical aspects central for understanding the spatial organisation of the genome are summarised in Figure 1.1.

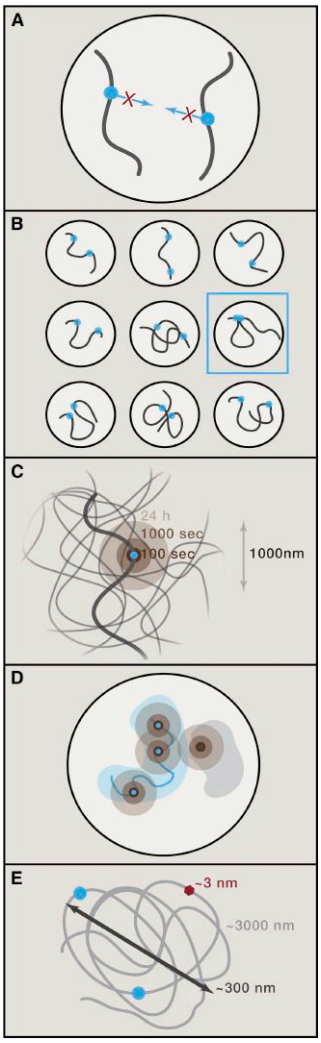


Figure 1.1: Physical aspects of chromosomal interactions. ***(A)*** *Interactions between genomic loci depend on the molecular affinity of protein-protein and protein-DNA interactions, which is limited to 1–10 nm. Such interactions can only take place when genomic loci are already very close in space. Formation of meaningful genomic interactions therefore initially relies on stochastic contacts between genomic loci.* ***(B)*** *Most interactions are extremely infrequent due to the large number or possibilities.* ***(C)*** *The polymeric nature of chromosomes limits mobility of individual loci. Therefore, loci close along the linear genome interact much more frequently than more distant loci.* ***(D)*** *Chromosomal territories increase the probability of interactions within the same chromosome and make interactions between loci located on different chromosomes highly variable between cells.* ***(E)*** *Proteins (~3 nm) are much smaller than chromatin regions (a modest region spanning 100–150 kb is estimated to be ~300 nm in diameter and ~1–3 mm in length), implying a more complex function for architectural proteins than direct physical insulation of interactions between chromosomal loci. Figure adapted from: Dekker and Mirny, 2016.*

These physical properties of chromosomes explain some features of genome organisation, such as relatively frequent interactions between proximal genomic regions and relative spatial insulation of distal regions. However, for large genomes, chromosomal organisation based on these physical properties alone would lead to highly stochastic structures and thus enormous cell-to-cell variation in gene expression (Dekker and Mirny, 2016). Therefore, additional layers

of genome organisation are required. For precise control over regulatory interactions, some proximal regions might need to be insulated, while specific distal interactions might need to be facilitated. These interactions need to be regulated in a cell-type specific manner to allow for precise and robust control of gene expression.

To enable such regulation, chromosomes are organised in different types of sub-chromosomal domains. At the megabase level, chromosomes form large functionally distinct domains, referred to as A/B compartments. Blocks of predominantly active chromatin associate with other active chromatin regions to form A compartments in the interior of the nucleus, whereas the transcriptionally inactive B compartments are located closer to the periphery (Lieberman-Aiden et al., 2009; Rao et al., 2014). These B compartments largely correspond to lamina-associated domains (LADs), which have also been associated with gene repression (Gibcus and Dekker, 2013; Peric-Hupkes et al., 2010).

Within such compartments, smaller domains at the scale of hundreds of kilobases are formed, which are usually referred to as topologically associating domains (TADs). These TADs are defined by the preferential interactions of chromatin regions located within the same domain and the relative depletion of interactions between regions located in adjacent domains. TADs correlate with chromatin signatures, gene activity and replication timing, and are thought to be relatively stable across different cell types and conserved across species (Dixon et al., 2015; 2012; Nora et al., 2012; Rao et al., 2014; Schmitt et al., 2016).

More recent analyses of TADs at higher resolution have revealed previously unrecognised, smaller, self-interacting domains that have been referred to as sub-TADs (Phillips-Cremins et al., 2013), contact domains (Rao et al., 2014), insulated neighbourhoods (Dowen et al., 2014) and frequently interacting regions (Schmitt et al., 2016). Tissue-specific interactions between genes and/or their *cis*-regulatory elements are thought to occur within these domains and are often described as chromatin loops (Bonev and Cavalli, 2016).

A common feature of all self-interacting domains, is that their borders are enriched for boundary elements that among many other proteins commonly bind CTCF. It has therefore been suggested that CTCF plays an important role in the formation of chromatin domains, which often depends on the presence of pairs of CTCF-binding motifs in convergent orientations at the domain borders (de Wit et al., 2015; Gómez-Marín et al., 2015; Guo et al., 2015; Rao et al., 2014; Sanborn et al., 2015; Vietri Rudan et al., 2015).

The definitions of the domains discussed above are rather vague and their identification is somewhat arbitrary and highly dependent on the resolution of the underlying data and the algorithms used to identify these structures (Forcato et al., 2017). The described domains are basically all defined by their self-interacting features, boundary elements, and sizes. To an extent, self-interacting domains appear to be fractal structures in the genome and it is unclear if they represent fundamentally different structures, or if they are variants of a similar organisational process. Based on their conservation across tissues, it has been proposed that TADs are the invariant building blocks of chromosomes, and that in a given cell type TADs of similar chromatin signatures form the larger cell type-specific compartments through a process of self-assembly. The different layers of chromosome organisation are therefore often described as an hierarchy (Gibcus and Dekker, 2013). An overview of the proposed hierarchical organisation of the genome is shown in Figure 1.2.

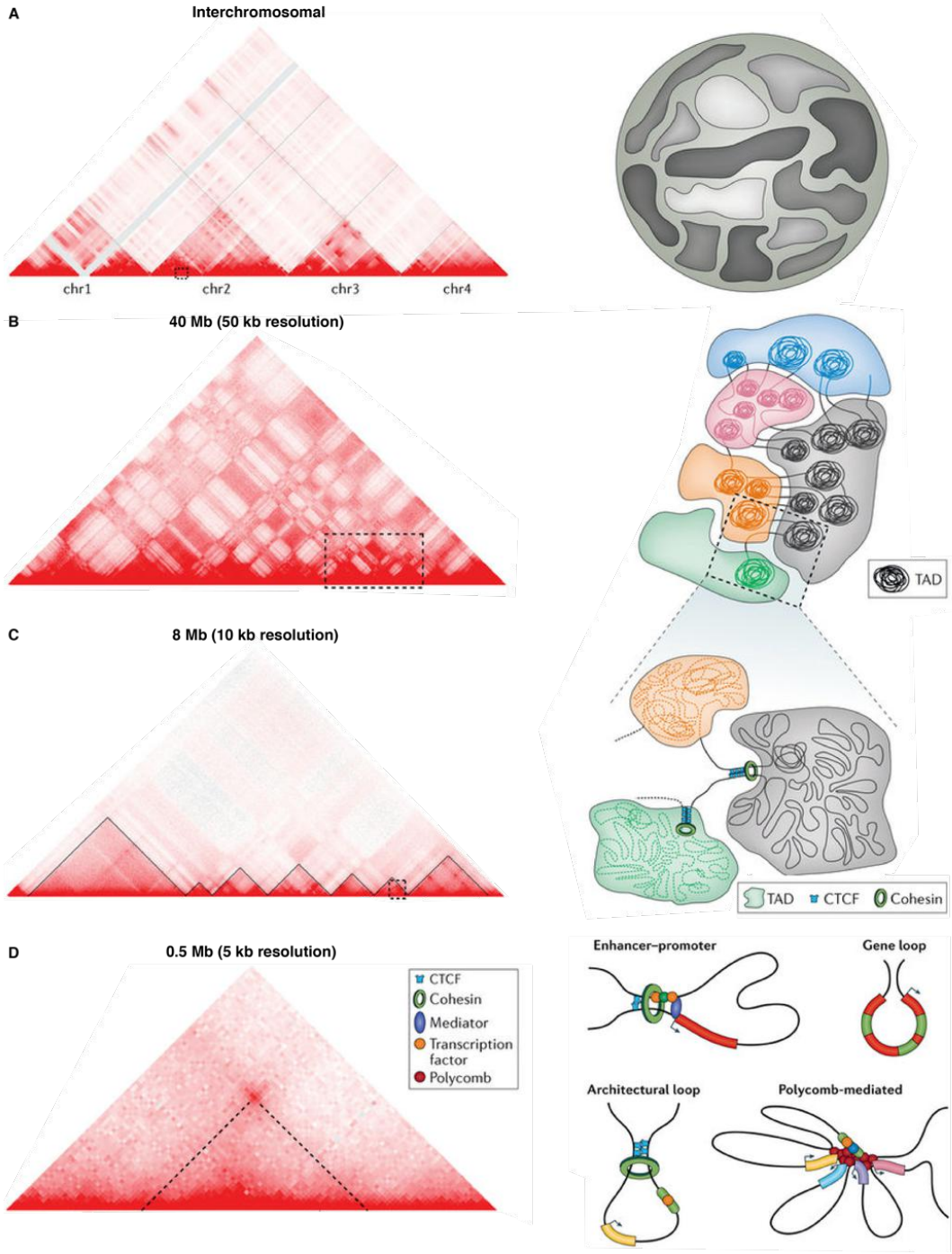


Figure 1.2: Layers of genome organisation. *Overview of the proposed hierarchical organisation of the genome. Hi-C contact matrices (explained below) from GM12878 cells are shown on the left and interpretations of corresponding chromatin structures are shown on the right. Dotted rectangles indicate the regions that are shown at higher magnification and increased resolution in the panel below.* ***(A)*** *Chromosomes occupy distinct territories within the cell nucleus.* ***(B)*** *Within individual chromosomes, compartments are formed by increased inter-domain interactions between domains with similar transcriptional activity and epigenetic signatures (visible in the Hi-C contact matrix as the checkerboard pattern).* ***(C)*** *Chromatin regions form domains of preferential interactions, often referred to as TADs (annotated with solid lines).* ***(D)*** *Within self-interacting domains, chromatin loops are formed between genes and/or* cis*-regulatory elements. Figure adapted from: Bonev and Cavalli, 2016.*

There is increasing evidence that the organisation of the genome in these structures plays an important role in the regulation of gene expression. Various studies have shown that perturbation of CTCF-binding domain borders can lead to changes in chromatin structure that allow enhancers to interact with genes in neighbouring domains, resulting in mis-regulated gene expression (Dowen et al., 2014; Franke et al., 2016; Gómez-Marín et al., 2015; Hanssen et al., 2017; Lee et al., 2017; Lupiáñez et al., 2015; Narendra et al., 2016; 2015; Nora et al., 2012; Symmons et al., 2016). It is therefore believed that self-interacting chromatin domains demarcate coherent *cis* neighbourhoods in which gene regulatory mechanisms take place. What remains unclear however, is how these domains are formed and how they contribute to specific interactions between *cis*-regulatory elements.

Based on the observation that boundaries of self-interacting domains often form punctate interactions in Hi-C contact matrices (Figure 1.2D), these structures have been described as loop domains. Such domains have been interpreted as relatively stable structures, formed by interactions between CTCF proteins located at the loop anchors (Figure 1.3) (Rao et al., 2014).

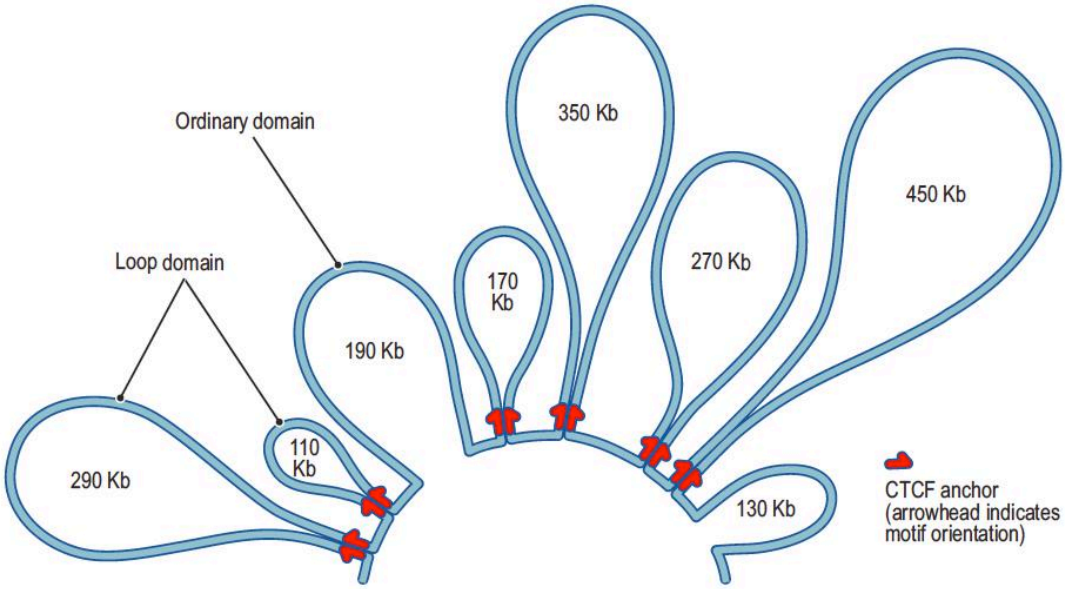


Figure 1.3: Formation of self-interacting domains by stable chromatin loops. *Interpretation of the structure of a 2.1 Mb region on chromosome 20 (48.78–50.88 Mb) in GM12878 cells. The structure is represented by eight domains, ranging in size from 110 kb to 450 kb, in which 95% of the chromatin in the region is contained. Six of the eight domains are demarcated by loops between convergent CTCF-binding sites located at the domain boundaries. Figure adapted from: Rao et al., 2014.*

The formation of stable chromatin loops can explain Hi-C data derived from bulk populations of cells. However, single-cell Hi-C studies (Nagano et al., 2013; 2017; Stevens et al., 2017) and

simulations based on polymer models (Giorgetti et al., 2014) have suggested that chromatin structures in single cells are highly diverse.

In order to explain these observations, a more dynamic mechanism of loop extrusion, borrowed from mitotic chromosome organisation (Alipour and Marko, 2012; Nasmyth, 2001), has been proposed to contribute to the formation of self-interacting domains. Loop extrusion is thought to be an active process in which a loop extruding factor, containing two DNA-binding units, associates with chromatin and travels along the fibre in opposite directions, creating a progressively larger intervening loop, until the factor is stalled at boundary elements, such as convergent CTCF-binding sites. Hetero-dimerization between CTCF proteins or other boundary proteins then tethers the extrusion factor, defining the limits of the self-interacting domain (Fudenberg et al., 2016; Sanborn et al., 2015). Based on its co-localisation with CTCF at domain boundaries, Cohesin, a ring-shaped protein complex that contains ATPase activity and can interact with two double-stranded DNA helices, has been proposed as a main loop extruding factor in the mammalian genome (Figure 1.4) (Hadjur et al., 2009; Parelho et al., 2008; Rubio et al., 2008; Stedman et al., 2008; Wendt et al., 2008; Xiao et al., 2011). It has been suggested that actively transcribing RNA polymerase II could also contribute to chromatin architecture, either by direct polymerase-mediated extrusion (Lee et al., 2015), or by influencing the distribution of Cohesin along chromosomes (Busslinger et al., 2017).

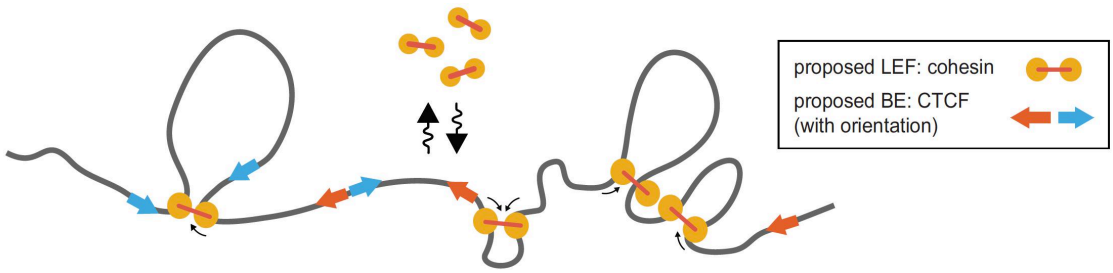


Figure 1.4: Formation of self-interacting domains by a dynamic loop extrusion mechanism. *Schematic of the hypothetical process of loop extrusion, in which self-interacting domains result from* cis*-acting loop extruding factors (LEFs; likely Cohesin), which form progressively larger loops, until stalled at boundary elements (BEs), such as CTCF-binding elements. Figure adapted from: Fudenberg et al., 2016.*

A mechanism of loop extrusion limited by barriers recapitulates many features of interphase chromosome organisation. These include the formation of non-overlapping dynamic domains within which chromatin preferentially self-interacts, insulation between *cis*-regulatory elements located in adjacent domains, enrichment of architectural proteins at the boundaries of chromatin domains, and preferential convergent orientation of CTCF motifs at these boundaries. The model also explains why these domains are formed within rather than between

chromosomes. Furthermore, loop extrusion has advantageous topological properties that would prevent the formation of intra-chromosomal knots and inter-chromosomal entanglements in the genome (Fudenberg et al., 2016; Sanborn et al., 2015).

Though it has been demonstrated that loop extrusion can occur *in vitro* by the yeast Condensin complex (Ganji et al., 2018), it is not clear if and to what extent mammalian genomes are organised by loop extrusion. It also remains unclear if and how loop extrusion could contribute to the formation of tissue-specific chromatin structures, and how and where loop extruding factors would be recruited and distributed along the genome. Therefore, despite extensive investigation of chromosome architecture across different scales, many aspects regarding genome organisation and the underlying mechanisms remain unknown.

Chromosome conformation capture

Though many fundamental features of genome organisation have been described by microscopic studies of the nucleus, in particular using fluorescence *in situ* hybridisation (FISH) (Bickmore, 2013), the majority of the recent discoveries describing chromatin structures at higher resolution have been made using Chromosome Conformation Capture (3C) technology (Dekker et al., 2002). These techniques are based on the principle that DNA sequences that are in close spatial proximity in the cell nucleus can be crosslinked together. Subsequent fragmentation and *in situ* proximity ligation can generate long DNA concatemers, in which fragments that were in close three-dimensional space within the cell nucleus are ligated proximal to each other. Quantification of these newly formed ligation junctions in 3C libraries provides information about the frequencies of interactions between DNA elements in the cell and can therefore be used to infer nuclear chromatin structures (Figure 1.5A) (Dekker et al., 2002).

In the initial 3C protocol, ligation junctions are detected using PCR followed by gel electrophoresis (Dekker et al., 2002). Though early studies using this technique have greatly contributed to our fundamental understanding of interactions between *cis*-regulatory elements (de Wit and de Laat, 2012), this approach is very laborious and can only be used to analyse a small number of suspected interactions. The conventional 3C method has therefore been superseded by more recently developed techniques based on the same principle that allow for unbiased identification of chromosomal interactions genome-wide. A multitude of such 3C based approaches have been developed, which vary in their resolution, throughput and biases. They can be divided in one vs all techniques, in which all interactions with a specific viewpoint

of interest in the genome are analysed, and all vs all techniques, which characterise interactions throughout the entire genome (Figure 1.5B) (Davies et al., 2017).

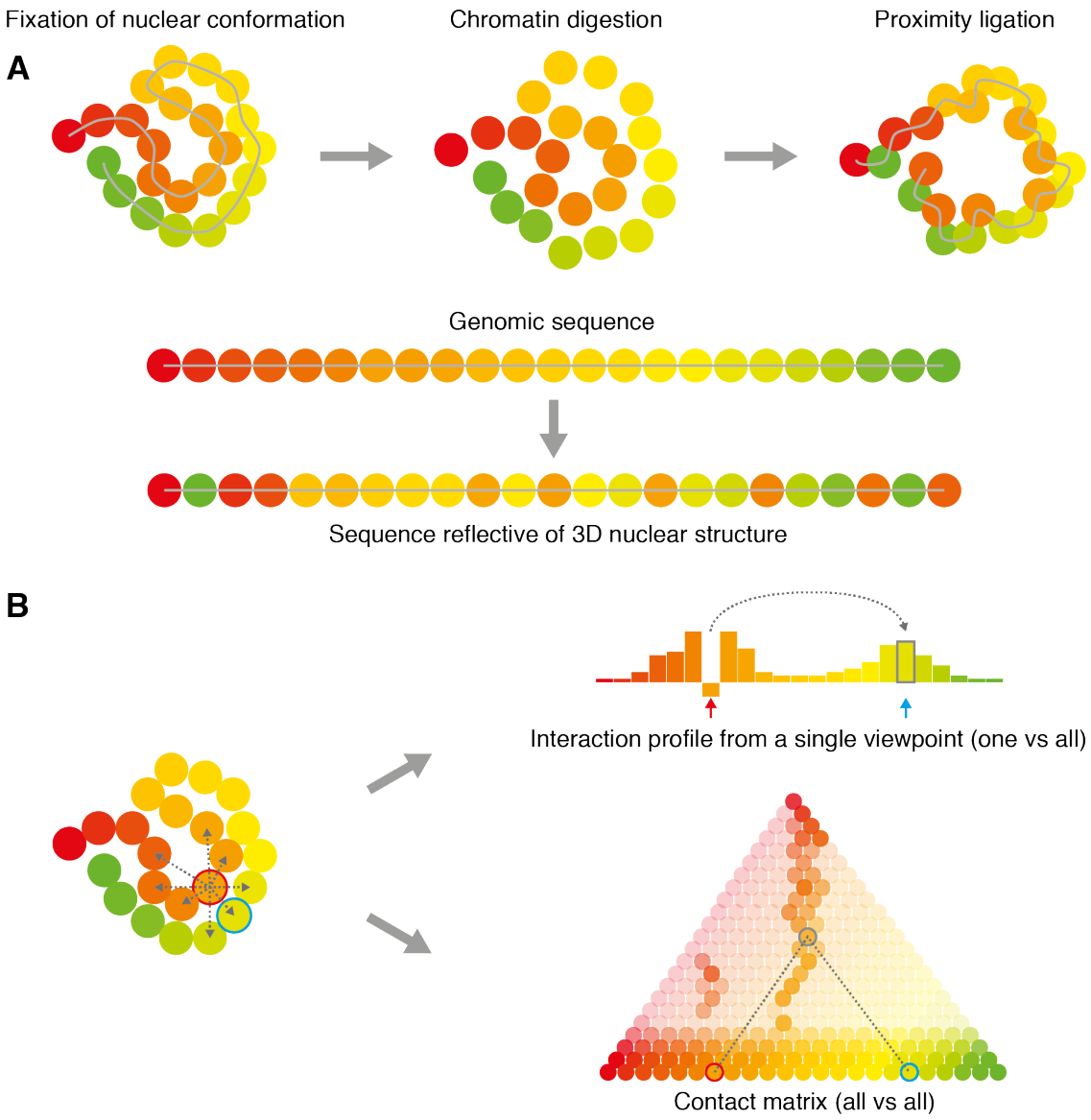


Figure 1.5: Chromosome conformation capture techniques. (A) *3C techniques rely on crosslinking of DNA sequences that are in close nuclear proximity, followed by fragmentation with restriction enzymes and* in situ *proximity ligation. This procedure results in large DNA concatemers in which the order of the ligated fragments reflects the three-dimensional structure of the chromatin in the cell nucleus. Quantification of the newly formed ligation junctions in 3C libraries provides information about the frequencies of interactions between DNA elements.* ***(B)*** *Several techniques based on the 3C principle are used to study different aspects of chromatin architecture. One vs all techniques can identify all interactions with a DNA fragment of interest at high resolution, which are usually represented in genome-wide interaction profiles from the viewpoint of interest. All vs all techniques can analyse structural interactions across the entire genome at moderate resolution, which are typically shown in contact matrices in which the interaction frequency is represented by the colour intensity at the intersection of the interacting loci. Figure adapted from: Davies et al., 2017.*

Though all vs all 3C techniques have the obvious advantage of being capable of identifying interactions across the entire genome, the enormous complexity of mammalian genomes means that 3C libraries need to be sequenced at enormous depth to detect all possible interactions. This limits the sensitivity of these techniques and requires the data to be pooled in bins, which markedly reduces the resolution of these approaches. The most widely used all vs all 3C technique is Hi-C. The Hi-C protocol uses a biotinylated nucleotide fill-in after restriction enzyme digestion, which allows for enrichment of informative ligation junctions in the 3C library after the material has been extracted and sonicated, and therefore increases sensitivity about two-fold (Lieberman-Aiden et al., 2009). Typical Hi-C contact matrices have a resolution of 10-50 kb, though contact matrices at 1-5 kb resolution can be produced using extremely deep sequencing of the libraries (Rao et al., 2014).

Analysis at higher resolution requires focussing on an area of the genome of interest with one vs all techniques. The maximal resolution of these techniques is defined by the choice of the restriction enzyme used to fragment the chromatin. Commonly used enzymes have recognition sequences of six or four bp, resulting in average restriction fragment sizes of 4096 and 256 bp, respectively. The most frequently used one vs all 3C technique is circular chromosome conformation capture (4C). This technique uses conventional 3C library preparation, followed by an inverse PCR with primers targeting a viewpoint restriction fragment of interest to amplify all its interacting partners (Simonis et al., 2006; Zhao et al., 2006). These were initially read out using a microarray, but this has been replaced by high-throughput sequencing (van de Werken et al., 2012). Though 4C has good sensitivity and resolution, it suffers from biases related to the efficiency of the inverse PCR. In addition, interaction frequencies are not accurately quantified because unique interactions cannot be distinguished from PCR duplicates. Furthermore, 4C is relatively laborious, as only one viewpoint can be examined per experiment.

These caveats have been addressed in more recently developed methods, in which hundreds of viewpoints can be analysed in a single experiment. The most sensitive technique of these many vs all approaches is Capture-C, which uses oligonucleotide capture technology to specifically enrich for viewpoints of interest and their interacting partners up to a million-fold. Importantly, the 3C products are sonicated prior to enrichment, which generates a 3C library with uniform small fragments with random ends, which can be used as unique molecular identifiers to collapse PCR duplicates. In addition, several independent 3C libraries (for example, from different cell types) can be analysed simultaneously in a single experiment, as the sonicated libraries are ligated with indexed sequencing adapters prior to enrichment and sequencing. These advantages make Capture-C the technique with the highest throughput and sensitivity,

allowing the data to be expressed at maximal resolution (Davies et al., 2015; Hughes et al., 2014).

The murine globin loci

The 3C techniques described above have been used extensively in the last two decades to study the nuclear organisation of the genome and the role of its structural features in gene regulatory mechanisms. Though genome-wide studies have been instrumental in elucidating the general principles of genome organisation (Bonev and Cavalli, 2016), many discoveries regarding the relationship between genome structure and the regulation of gene expression have been made by examining individual gene loci in detail (de Laat and Duboule, 2013).

The α- and β-globin genes, which code for the subunits of haemoglobin, have a long history as model loci for studying gene regulatory mechanisms. Their *cis*-regulatory elements were initially identified based on naturally occurring mutations which caused rare cases of thalassaemia, characterised by severely reduced expression of haemoglobin despite intact sequences of the globin genes themselves (Curtin et al., 1985; Hatton et al., 1990; Kioussis et al., 1983; Nicholls et al., 1987). These cases highlighted the importance of distal regulatory elements for controlling gene activity and allowed investigation of mutations in such elements in their native chromatin environment. As a result, the *cis*-regulatory elements of the globin genes have been very well characterised. In addition, the specific expression pattern of haemoglobin make these gene loci convenient experimental model systems. Both α- and β-like globin genes are expressed in red blood cells, which are easily obtained from patients and model organisms. Moreover, their activity is gradually upregulated throughout erythroid differentiation in a predictable manner (Figure 1.6). This process can be recapitulated in primary cells *ex vivo*, allowing for analysis of gene regulation in purified cell populations with distinct levels of globin transcription. Importantly, the products of the globin genes are only functionally important at terminal differentiation, unlike other model loci such as the developmental *Hox* genes. Therefore, changes in globin expression do not affect cell fate decisions or differentiation processes, which can drastically confound the interpretation of experimental manipulations.

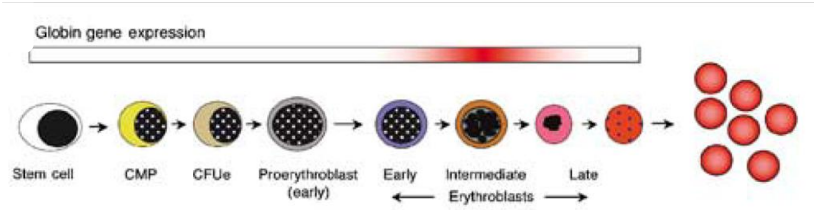


Figure 1.6: Expression of the globin genes during erythropoiesis. *Globin gene expression (indicated in red) is gradually upregulated during erythroid differentiation. Figure adapted from: Vernimmen et al., 2007)*

Consequently, the globin genes have served as important model loci to study gene regulatory mechanisms and their investigation has contributed greatly to our understanding of the general principles of mammalian genome organisation and gene regulation (Higgs et al., 2008; Noordermeer and de Laat, 2008). Though α- and β-globin have undergone different evolutionary paths, many aspects of their regulation are conserved (Philipsen and Hardison, 2017).

The α-globin locus comprises the duplicated adult α-globin genes *Hba-a1* and *Hba-a2*, as well as the embryonic ζ-globin gene *Hba-x*. This gene cluster is located in a region of synteny of ~135 kb, which has been defined by sequence comparisons across 22 species, covering 500 million years of evolution. This region contains four highly conserved non-coding sequences with several transcription factor binding sites. These putative enhancer elements have been defined as Multispecies Conserved Sequences (MCS) R1–4 (Hughes et al., 2005). Three of these elements are located in the introns of the neighbouring housekeeping gene *Nprl3*. In addition to these conserved elements, the murine α-globin locus contains the mouse-specific element Rm and two θ-pseudogenes *Hbq1a* and *Hbq1b*. An overview of the murine α-globin locus is shown in Figure 1.7. This figure also shows the chromatin signature of the locus in erythroid cells, in which the genes are highly expressed, and in mouse ES cells, in which the genes are silent.

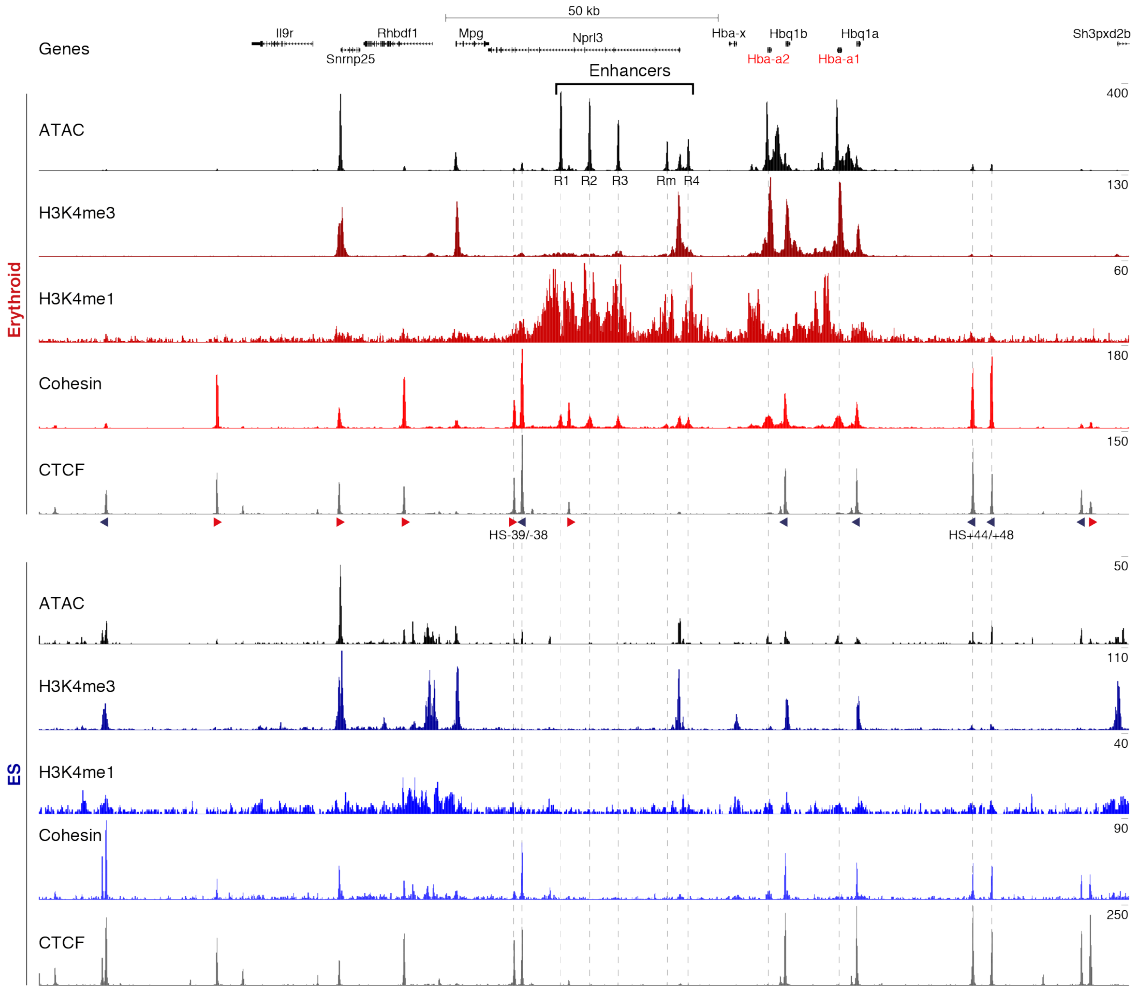


Figure 1.7: The murine α-globin locus. *Overview of the murine α-globin locus, including characteristic histone modifications and occupancy of transcription factors in primary erythroid (top; red) and ES (bottom; blue) cells. From top to bottom, tracks represent gene annotation (α-globin genes highlighted in red), chromatin accessibility, H3K4 tri- and mono-methylation, and occupancy of Cohesin (Rad21 subunit in erythroid cells; Smc1 subunit in ES cells) and CTCF, with arrows at the bottom indicating the orientation of the directional CTCF motifs. Profiles represent normalised (RPKM) ATAC-seq and ChIP-seq data (Hanssen et al., 2017). Promoters, enhancers and boundary elements are highlighted with dashed grey lines. Coordinates: chr11:32,050,000-32,250,000.*

The β-globin locus in mouse contains two adult β-globin genes *Hbb-b1* and *Hbb-b2*, an embryonic β-globin gene *Hbb-bh1* and an embryonic ε-globin gene *Hbb-y*, which are all transcribed from the negative strand of the genome. Expression of these genes is regulated by the conserved upstream Locus Control Region (LCR), which contains six elements. Figure 1.8 shows an overview of the murine β-globin locus.

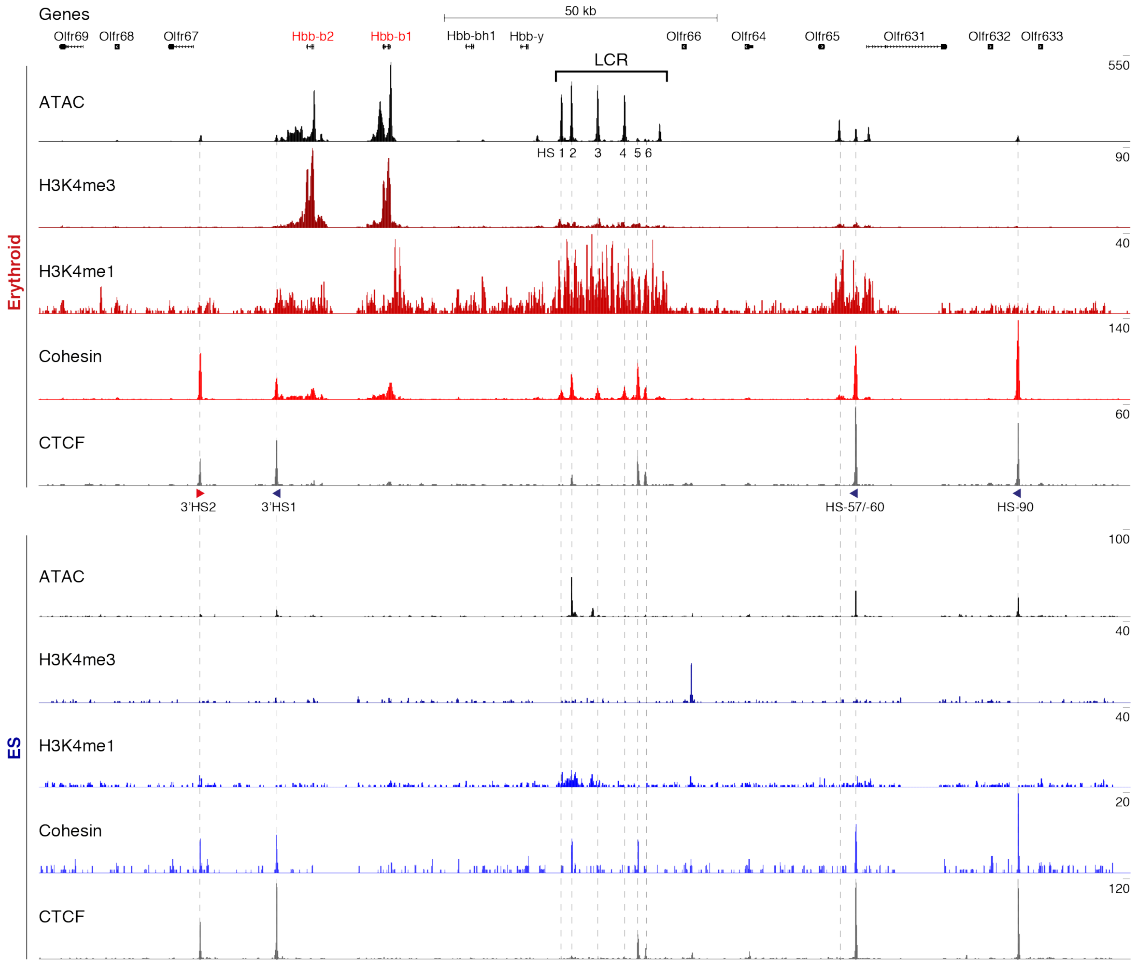


Figure 1.8: The murine β-globin locus. *Overview of the murine β-globin locus, including characteristic histone modifications and occupancy of transcription factors in primary erythroid (top; red) and ES (bottom; blue) cells. From top to bottom, tracks represent gene annotation (β-globin genes highlighted in red), chromatin accessibility, H3K4 tri- and mono-methylation, and occupancy of Cohesin (Rad21 subunit in erythroid cells; Smc1 subunit in ES cells) and CTCF, with arrows at the bottom indicating the orientation of the directional CTCF motifs. Profiles represent normalised (RPKM) ATAC-seq and ChIP-seq data (Hanssen et al., 2017). Promoters, enhancers and boundary elements are highlighted with dashed grey lines. Coordinates: chr7:110,912,000-111,112,000.*

As shown in Figures 1.7 and 1.8, the *cis*-regulatory elements of the globin genes have characteristic chromatin signatures. Both loci are flanked by multiple CTCF- and Cohesin-binding sites. These putative boundary elements demarcate a region containing active promoters and predicted enhancers, marked by high levels of H3K4 tri-methylation and H3K4 mono-methylation, respectively. The enhancers have been further characterised *in vivo* by studying the effects of engineered deletions of these elements in the α-globin (Hay et al., 2016) and β-globin (Bender et al., 2012) locus. This has revealed that the globin enhancers cooperate in an additive manner to upregulate gene expression, but differ markedly in their individual contribution to transcriptional output.

The activity of the *cis*-regulatory elements of the globin loci is controlled by several erythroid transcription factors. These include GATA-binding factor 1 (Gata1), Stem cell leukaemia factor (Scl, also known as Tal1), Nuclear factor erythroid 2 (Nfe2), and Krüppel-like factor 1 (Klf1) (De Gobbi et al., 2007). It has been shown that binding of these transcription factors to the α-globin enhancers is accompanied by binding of the general transcriptional co-factor Mediator (Hay et al., 2016) and recruitment of RNA polymerase II (Vernimmen et al., 2007). Occupancy of these proteins at the active α- and β-globin loci is shown in Figure 1.9.

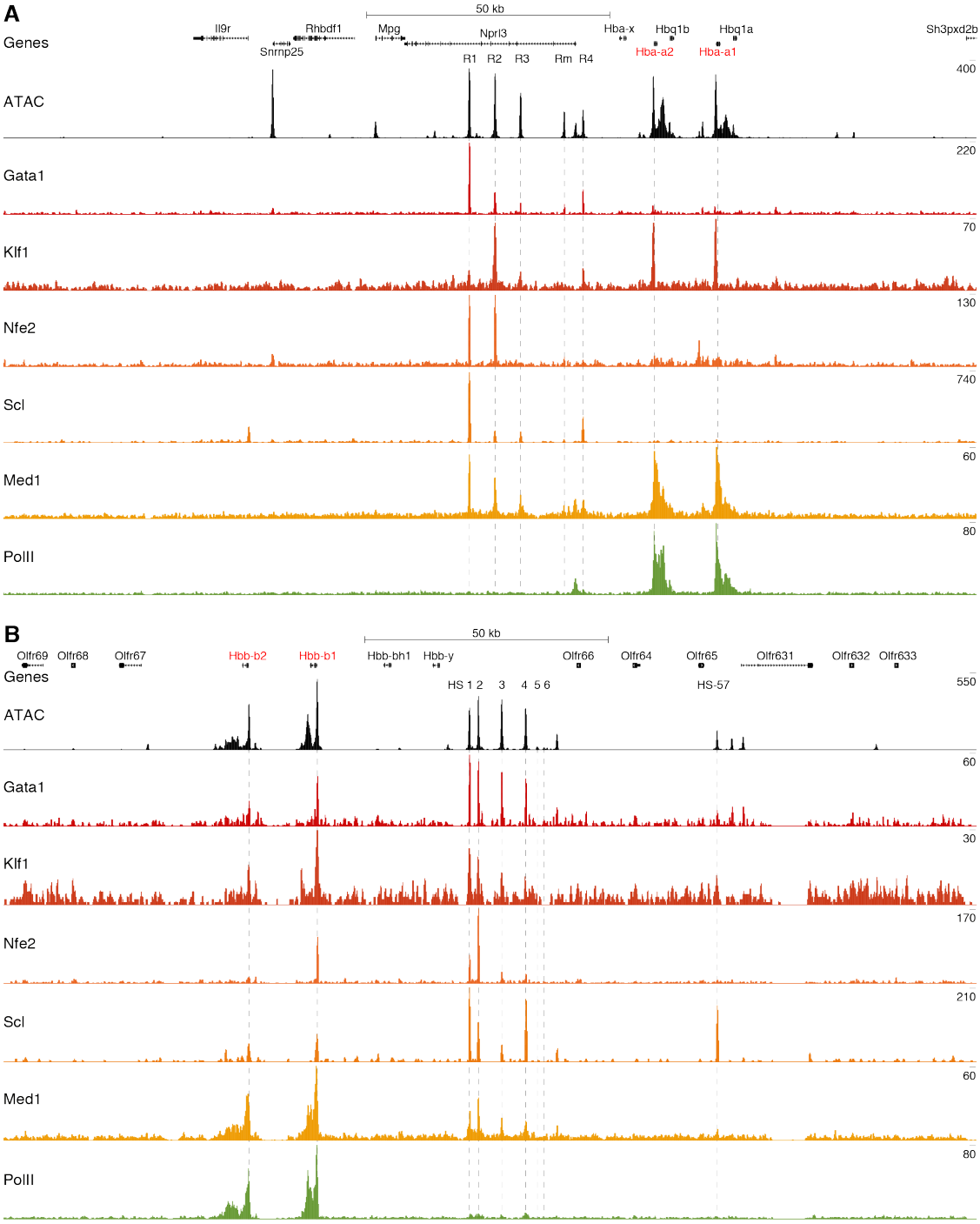


Figure 1.9: Transcription factor and RNA polymerase II occupancy at the active murine globin loci. *From top to bottom, tracks represent gene annotation (globin genes highlighted in red); chromatin accessibility (black); occupancy of erythroid transcription factors Gata1, Klf1, Nfe2 and Scl (red/orange); transcriptional co-factor Mediator (Med1; yellow) and RNA Polymerase II (PolII; green) at the active* ***(A)*** *α-globin and* ***(B)*** *β-globin loci. Profiles represent normalised (RPKM) ATAC-seq and ChIP-seq data in primary erythroid cells (Hay et al., 2016). Promoters and enhancers are highlighted with dashed grey lines. Coordinates: chr11:32,050,000-32,250,000 (α-globin) and chr7:110,912,000-111,112,000 (β-globin).*

In addition to the extensive functional annotation of their *cis*-regulatory elements, the structural organisation of the globin loci has also been well characterised. The murine β-globin locus was one of the first loci in which long-range gene regulation by physical interactions between *cis*-regulatory elements was demonstrated *in vivo*. Based on conventional 3C experiments in the locus, it was hypothesised that these elements form an 'active chromatin hub', in which active genes and *cis*-regulatory elements interact, with intervening inactive chromatin looping out (Figure 1.10) (Tolhuis et al., 2002).

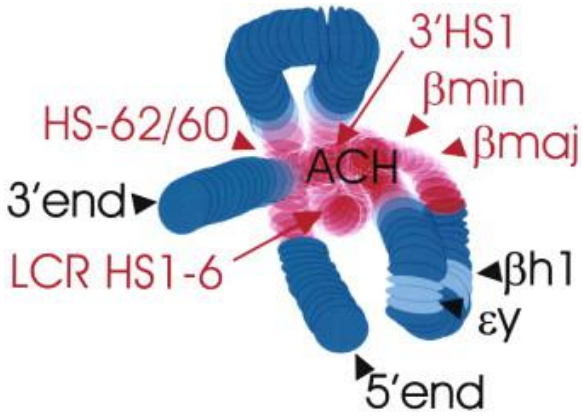


Figure 1.10: The active chromatin hub model. *Proposed three-dimensional structure of the murine β-globin locus, in which active regions (active genes and hypersensitive sites; red) form a hub of hyper-accessible chromatin, with the intervening inactive, more compacted chromatin (grey) looping out. Figure adapted from: Tolhuis et al., 2002.*

Since these initial experiments, further 3C analyses of the β-globin (Davies et al., 2015; Drissen et al., 2004; van de Werken et al., 2012) and α-globin (Baù et al., 2011; Davies et al., 2017; van de Werken et al., 2012; Vernimmen et al., 2007; 2009; Zhou et al., 2006) loci have characterised the interactions between their *cis*-regulatory elements in more detail and at higher resolution. Based on these experiments and investigations of many other loci, interactions between enhancers and promoters are often described as 'enhancer loops' (de Laat and Duboule, 2013).

This model implies that discrete, stable interactions are formed exclusively between enhancers and promoters. Such structures are difficult to reconcile with the dynamic nature of chromatin interactions predicted by polymer models (Giorgetti et al., 2014) and observed in single-cell Hi-C studies (Nagano et al., 2013; 2017; Stevens et al., 2017). However, due to the limited resolution of single-cell Hi-C (100 kb), these studies are not directly informative at the level of individual *cis*-regulatory elements. The dynamic nature of enhancer-promoter interactions in single cells therefore remains unclear.

As a result, it also remains unclear how multiple *cis*-regulatory elements located within a chromatin domain interact together. Like most important developmental genes, the murine globin genes are regulated by multiple enhancer elements, which cooperate in an additive manner to upregulate gene expression (Bender et al., 2012; Hay et al., 2016). Such cooperativity between multiple *cis*-regulatory elements has been observed at many other gene loci (Osterwalder et al., 2018). However, the underlying mechanisms and associated chromatin structures remain elusive, as current techniques are not able to identify concurrent chromatin interactions in single cells with sufficient sensitivity and resolution.

Thesis outline

Though the development of 3C has greatly improved our understanding of the structure of the genome and its contribution to the accurate regulation of gene expression, many unanswered questions remain. More specifically, it is not clear how self-interacting chromatin domains are formed, what the roles of these domains are in facilitating and/or insulating chromatin interactions, and how the multiple *cis*-regulatory elements within these domains communicate and cooperate to establish precise and robust gene expression patterns. This thesis aims to address these questions by building on previous studies at the murine globin loci to characterise their regulatory landscapes in unprecedented detail.

To this aim, I will first characterise the different layers of nuclear organisation of the murine globin loci, using complementary currently available 3C techniques. Next, I will examine the role of specific *cis*-regulatory elements in the α-globin locus in establishing its structural conformation, by analysing the effects of genetic perturbations of these elements. To further investigate how these *cis*-regulatory elements interact and cooperate to upregulate globin transcription, I will then present analyses of the three-dimensional organisation of the globin loci in individual cells using a new 3C technique, which I developed for this purpose. Finally, to increase the applicability of 3C techniques, I will present optimised 3C approaches, which can analyse chromosomal structures in very small numbers of cells.

Though I have analysed the native chromatin structures of both murine globin loci in depth, I have only examined the effects of genetic perturbations of *cis*-regulatory elements in the α-globin locus. The α-globin locus will therefore be the focus of my discussion throughout the thesis, though I will compare my findings to β-globin where possible. I will discuss the presented results briefly at the end of each chapter, and more extensively in a separate chapter,

in which I summarise my findings and discuss how they contribute to our current understanding of genome organisation and gene regulation.

Notes

Unless explicitly stated otherwise, the data and their interpretation presented in this thesis reflect my own work. However, to reflect the collaborative nature of the projects in the Hughes and Higgs groups, I prefer to use the pronoun ‘we’, rather than ‘I’, to describe them. Work from other members of the groups is acknowledged at the end of each chapter.

Availability

Part of the literature review presented in this chapter is published in the following articles:

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nature Methods *14*, 125–134.

Oudelaar, A.M., Hanssen, L.L.P., Hardison, R.C., Kassouf, M.T., Hughes, J.R., and Higgs, D.R. (2017). Between form and function: the complexity of genome folding. Human Molecular Genetics *26*, R208–R215.

Chapter 2 – The structural organisation of the murine globin loci

Introduction

This chapter describes the structural characterisation of the murine globin loci by complementary 3C techniques. Hi-C is used to study the large-scale organisation and structural context of the globin domains at moderate resolution, and Capture-C to analyse the interaction landscape of the *cis*-regulatory elements within these domains at high resolution. To characterise both the active and inactive structural conformations of the globin loci, we performed all analyses in primary erythroid cells and ES cells, respectively.

Results

Characterisation of the murine globin loci by Hi-C

To characterise the large-scale organisation of the active and inactive murine globin loci, we performed Hi-C in primary erythroid cells, and compared these data to an equivalent publicly available dataset from ES cells (Giorgetti et al., 2016). Both Hi-C datasets were analysed at 40 kb resolution. TADs were identified based on the insulation indices derived from the Hi-C matrices. The Hi-C contact matrices and derived structural features are shown in Figures 2.1 and 2.2.

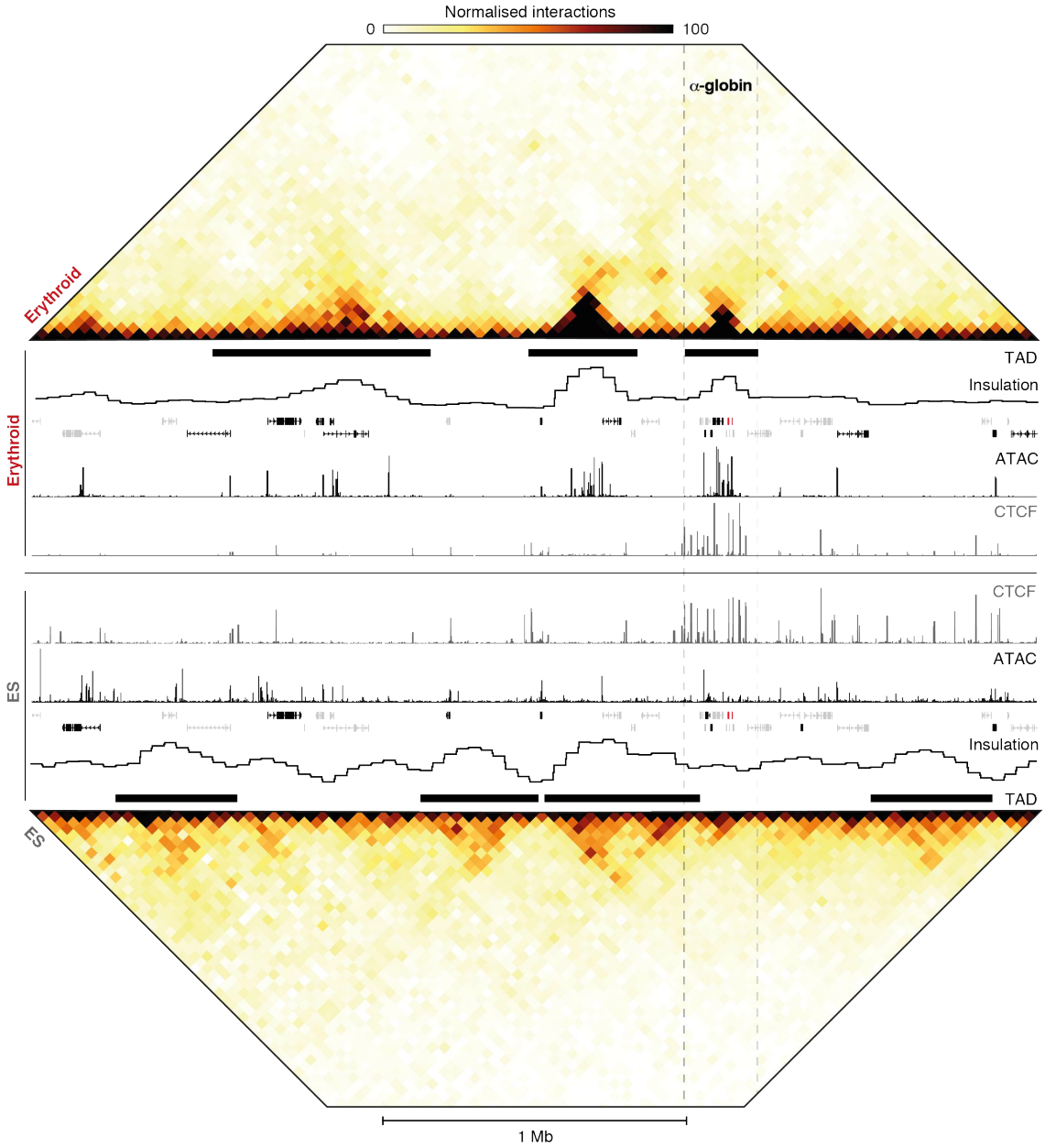


Figure 2.1: Hi-C analysis of the extended α-globin locus. *Hi-C contact matrices of 3.3 Mb regions surrounding the murine α-globin genes in primary erythroid (top) and ES (bottom) cells at 40 kb resolution. Corresponding TADs, insulation indices, gene annotation (active genes in black, inactive genes in grey, α-globin genes in red), chromatin accessibility (defined by ATAC; black) and CTCF occupancy (grey) are shown in the middle and the TAD containing the active α-globin locus is highlighted in a dashed grey box. Coordinates: chr11:29,900,000-32,300,000.*

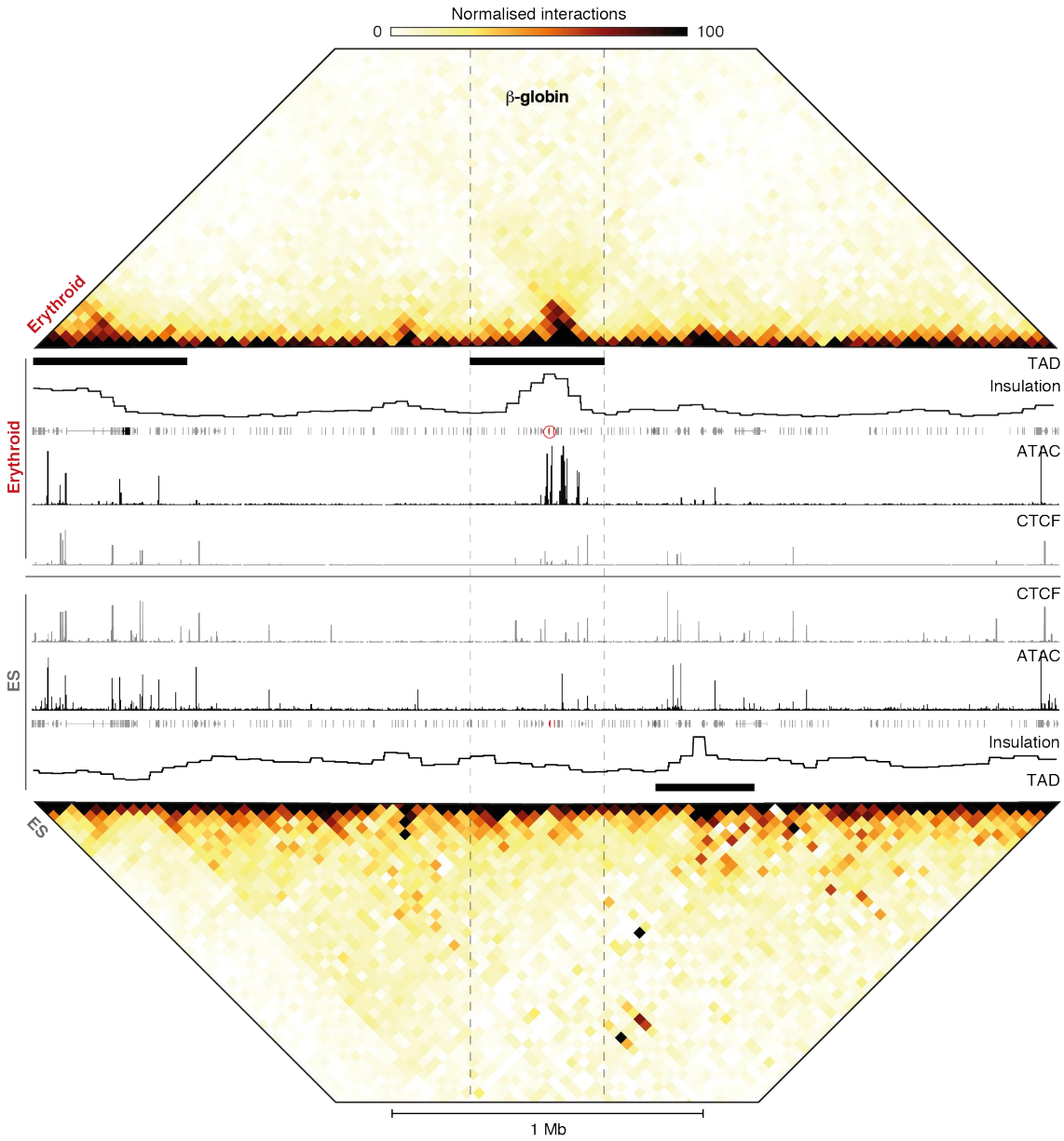


Figure 2.2: Hi-C analysis of the extended β-globin locus. *Hi-C contact matrices of 3.3 Mb regions surrounding the murine β-globin genes in primary erythroid (top) and ES (bottom) cells at 40 kb resolution. Corresponding TADs, insulation indices, gene annotation (active genes in black, inactive genes in grey, β-globin genes in red), chromatin accessibility (defined by ATAC; black) and CTCF occupancy (grey) are shown in the middle and the TAD containing the active β-globin locus is highlighted in a dashed grey box. Note that the Hi-C data in the contact matrix in ES cells are rather sparse due to suboptimal alignment of this region in ES cells (Cast/129 strain) to the reference genome (as described in Chapter 8). Though this complicates direct quantitative comparison of interactions between erythroid and ES cells, differences in domain structure are apparent. Coordinates: chr7:109,300,000-112,600,000.*

These results clearly show that the organisation of TADs around the globin genes varies between erythroid and ES cells. Strong TADs containing the active globin clusters form specifically in erythroid cells. This is in contrast with previous studies, which suggest that TADs are conserved across cell types (Dixon et al., 2012; 2015; Rao et al., 2014; Schmitt et al., 2016). Moreover, these studies describe the organisation of chromosomes into TADs as a pervasive feature of genome organisation. However, our analyses of the globin loci show that only certain regions in the genome form such self-interacting structures and that large chromosomal segments in the genome are less well-defined. This is especially evident in the context of the β-globin locus. The differences in the extent of TAD organisation between the α- and β-globin loci might reflect their different genomic contexts. The α-globin cluster is located in a relatively active region of the genome, which contains many constitutively expressed housekeeping genes. The β-globin cluster is surrounded by olfactory receptor genes, which are exclusively expressed in olfactory sensory neurons and completely silent in both erythroid and ES cells.

As described in Chapter 1, the physical properties associated with the polymer nature of chromosomes constrain their spatial structures within the nucleus. Models based on polymer physics can therefore provide insight into the principles of genome organisation and can be used to visualise three-dimensional conformations of chromosomal structures derived from experimental data which take their physical properties into account (Nicodemi and Pombo, 2014).

The framework of the Strings & Binders Switch (SBS) model recapitulates key features of chromosome organisation in space and time (Barbieri et al., 2012). Within this framework, chromosomes are modelled as polymers which interact with diffusing binders that can crosslink the polymer at specific binding sites. The model can be expanded to consider many polymers, different binder species, and interactions with nuclear landmarks (Figure 2.3).

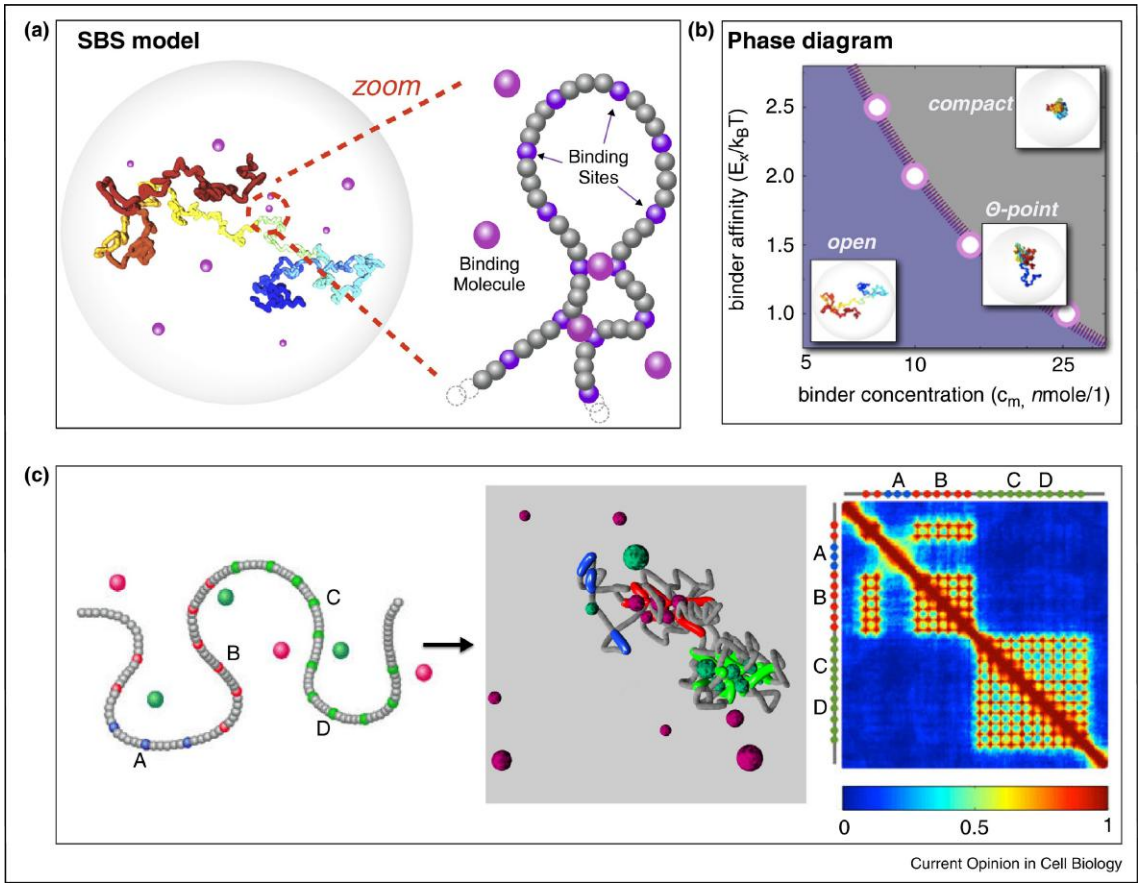


Figure 2.3. The Strings & Binders Switch Model. (A) *In the SBS model, chromatin fibres are modelled as self-avoiding-walk polymer chains with binding sites for diffusing binding molecules. Interactions between the polymer chain and the molecular binders result in specific three-dimensional structures.* ***(B)*** *The basic model parameters are the binder concentration (c_m) and the binder affinity for the polymer binding sites (E_X). The thermodynamically stable phases of the SBS model correspond to different conformational classes, between which chromatin fibres can switch in response to changes in binder concentration or affinity.* ***(C)*** *By specifying polymer binding sites and their binding molecules, the SBS framework can model the formation of chromatin domains. The corresponding contact matrices correlate well with Hi-C contact matrices. Figure adapted from: Nicodemi and Pombo, 2014.*

We used the SBS model to derive three-dimensional conformations of the α- and β-globin loci in their active and inactive conformations in erythroid and ES cells, respectively (Figures 2.4 and 2.5). Simulations were based on normalised Hi-C interactions at 20 kb resolution.

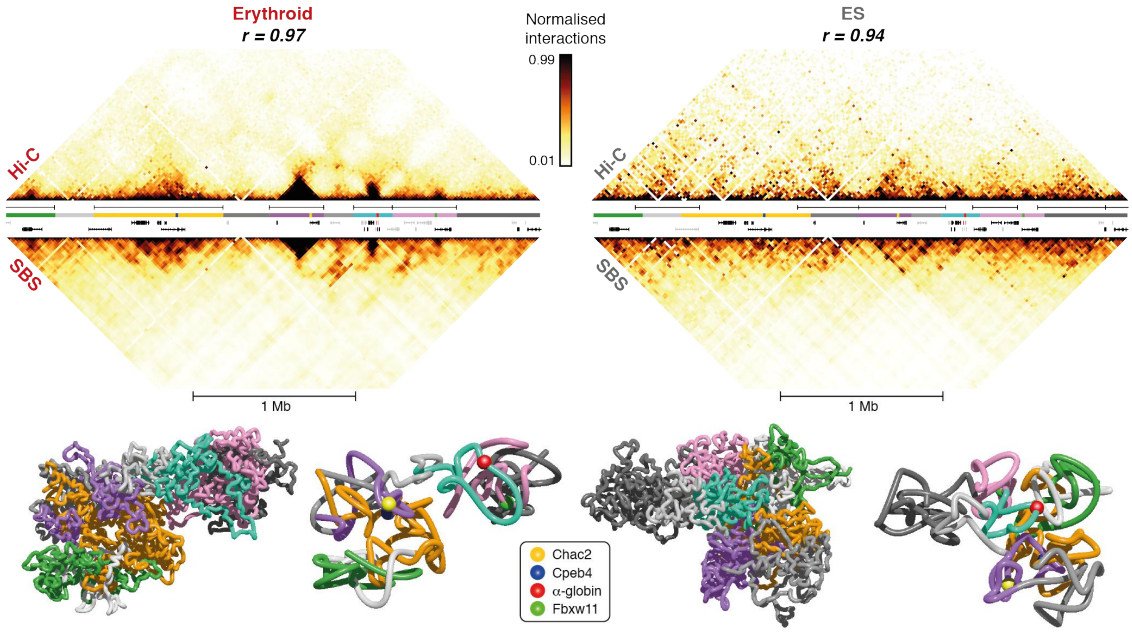


Figure 2.4: SBS models of the extended α-globin locus. *Contact matrices (20 kb resolution) of the α-globin locus derived from Hi-C data (top) and SBS simulations (bottom) in erythroid (left) and ES (right) cells. TADs, colour legends and gene annotation (active genes in black, inactive genes in grey, α-globin genes in red) are shown in the middle of the matrices. The correlation between the experimental and modelled matrices are indicated by the Pearson correlation coefficients (r) at the top. The three-dimensional conformations show full (left) and coarse-grained (right) representations of the α-globin locus at 20 kb resolution derived from the SBS models in erythroid (left) and ES (right) cells, with colours corresponding to the annotation in the contact matrices. Coordinates: chr11:29,900,000-32,300,000.*

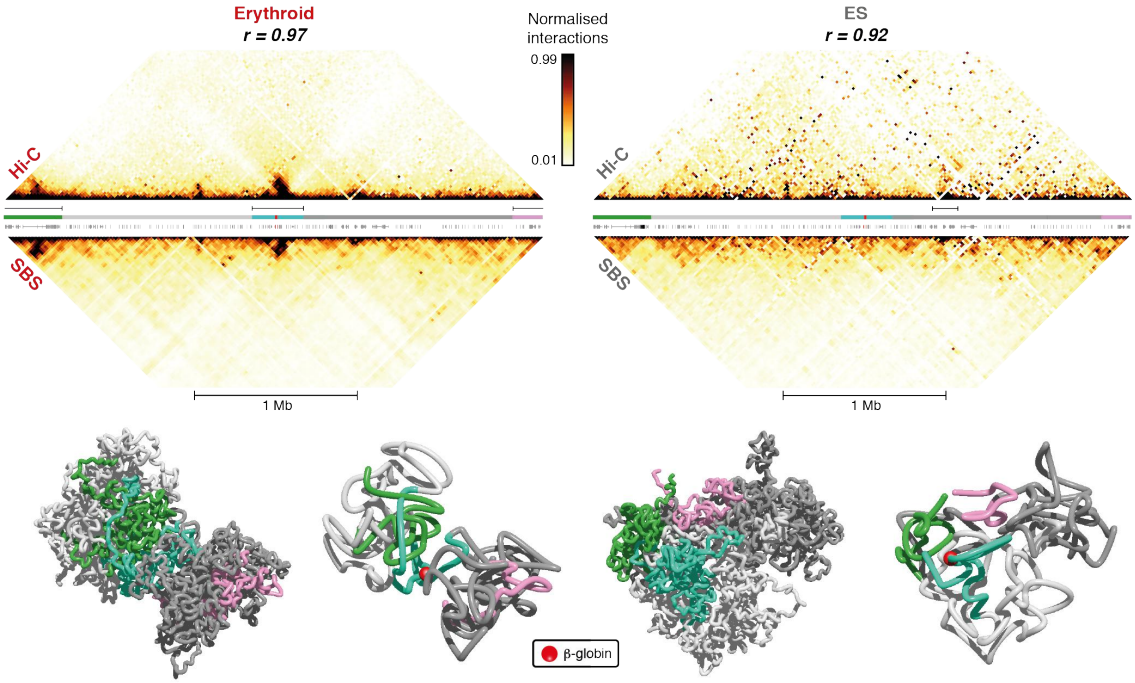


Figure 2.5. SBS models of the extended β-globin locus. *Contact matrices (20 kb resolution) of the β-globin locus derived from Hi-C data (top) and SBS simulations (bottom) in erythroid (left) and ES (right) cells. TADs, colour legends and gene annotation (active genes in black, inactive genes in grey, β-globin genes in red) are shown in the middle of the matrices. The correlation between the experimental and modelled matrices are indicated by the Pearson correlation coefficients (r) at the top. The three-dimensional conformations show full (left) and coarse-grained (right) representations of the β-globin locus at 20 kb resolution derived from the SBS models in erythroid (left) and ES (right) cells, with colours corresponding to the annotation in the contact matrices. Coordinates: chr7:109,300,000-112,600,000.*

The colours in the three-dimensional conformations in Figures 2.4 and 2.5 correspond to TADs identified in erythroid contact matrices at 20 kb resolution. These representations highlight that the organisation of genomic regions into self-interacting domains is tissue-specific.

High-resolution analysis of the chromatin domains containing the murine globin loci by Capture-C

To study the organisation of the chromatin within the globin TADs at higher resolution and characterise the interaction landscapes of the *cis*-regulatory elements that control expression of the globin genes, we performed multiplexed high-resolution Capture-C experiments from ~45 viewpoints (including all known *cis*-regulatory elements) in primary erythroid and ES cells. The resulting data are summarised in interaction profiles and contact matrices (Figures 2.6–2.9).

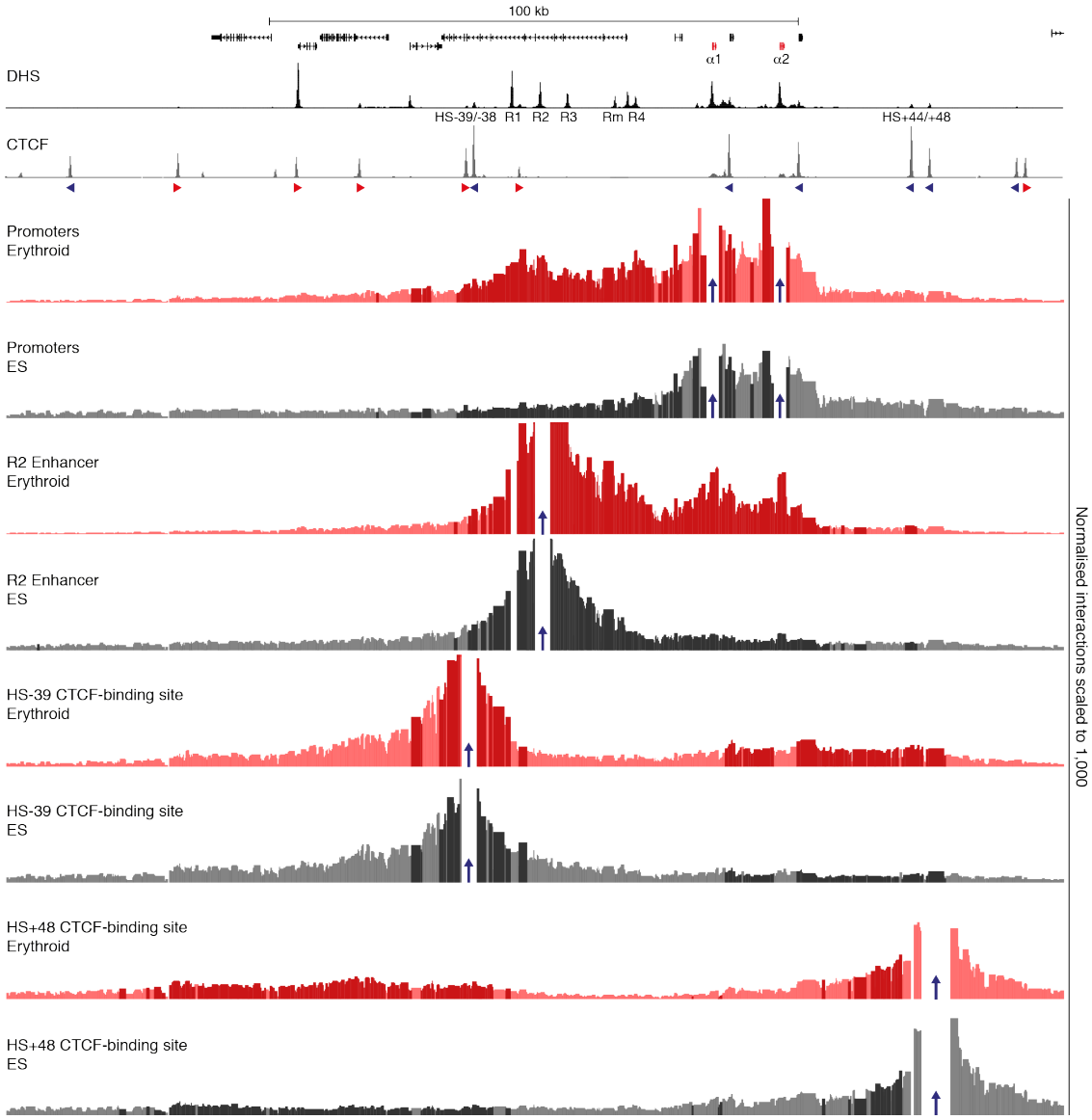


Figure 2.6: Characterisation of the interaction landscapes of the cis-regulatory elements of the α-globin locus. *High-resolution Capture-C interaction profiles of the α-globin locus from the viewpoints (indicated by blue arrows) of the α-globin promoters, the R2 enhancer, and CTCF-binding sites HS-39 and HS+48 in erythroid (red) and ES (grey) cells. Profiles represent the mean number of normalised unique interactions per restriction fragment from three replicates. Significantly different interactions between erythroid and ES cells are highlighted in bold colours. Gene annotation (α-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS) and CTCF occupancy are shown at the top, with arrows indicating the orientation of the CTCF-binding motifs. Coordinates: chr11:32,050,000-32,250,000.*

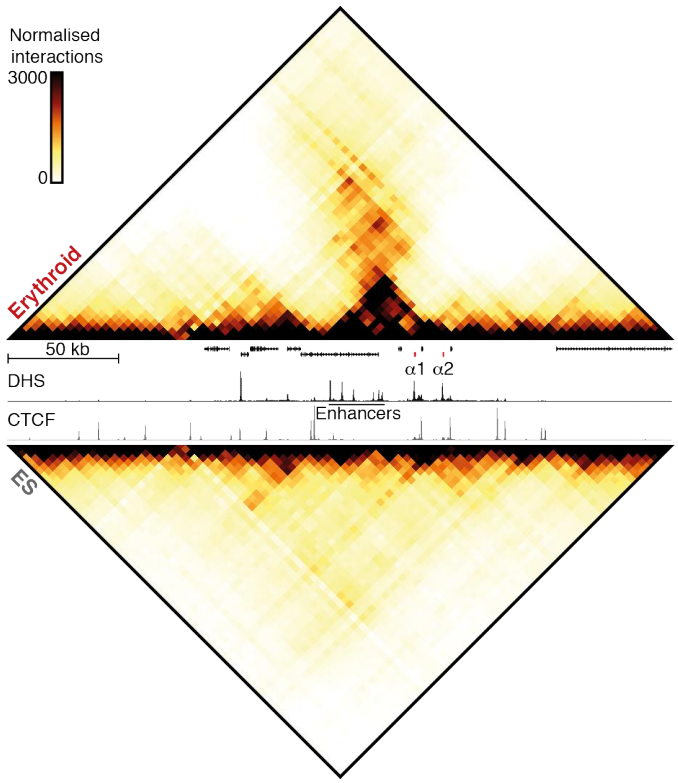


Figure 2.7: Structural conformations of the active and inactive α-globin domain. *Contact matrices (4 kb resolution) of the α-globin locus derived from Capture-C experiments with viewpoints closely spaced across the domain in erythroid (top) and ES (bottom) cells. Matrices represent the mean number of normalised unique interactions from three replicates. Gene annotation (α-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS) and CTCF occupancy are shown in the middle. Coordinates: chr11:32,000,000-32,300,000.*

Figure 2.6 shows interaction profiles in the α-globin domain from the viewpoint of the promoters, the strongest enhancer element R2 (Hay et al., 2016), and CTCF boundary elements on either side of the domain. These profiles show strong reciprocal interactions between the α-globin promoters and all its enhancer elements in erythroid cells, but not in ES cells. The flanking CTCF-binding elements do not participate in these enhancer-promoter interactions, but form erythroid-specific diffuse interactions with the chromatin bound by CTCF on either side of the domain, spanning contiguous regions of ~50 kb. The distinct nature of these interactions patterns is highlighted in the contact matrices derived from the viewpoints tiled across the locus (Figure 2.7).

Although the contacts between the promoters and enhancers are more punctate compared to those between the CTCF-binding sites, interaction frequencies are significantly enriched throughout the domain in erythroid cells. Importantly, this is not consistent with a model in which stable, distinct loops are formed exclusively between *cis*-regulatory elements (Tolhuis et

al., 2002). Rather, all regions of chromatin within the domain contact each other, with increased interactions between enhancers and promoters, and the regions bound by CTCF.

Capture-C analysis of the β-globin domain shows very similar patterns of interactions (Figures 2.8 and 2.9). This suggests that such chromatin domains represent compartmentalised structures, in which every region of chromatin interacts together, and preferred, stabilised interactions are formed between *cis*-regulatory elements.

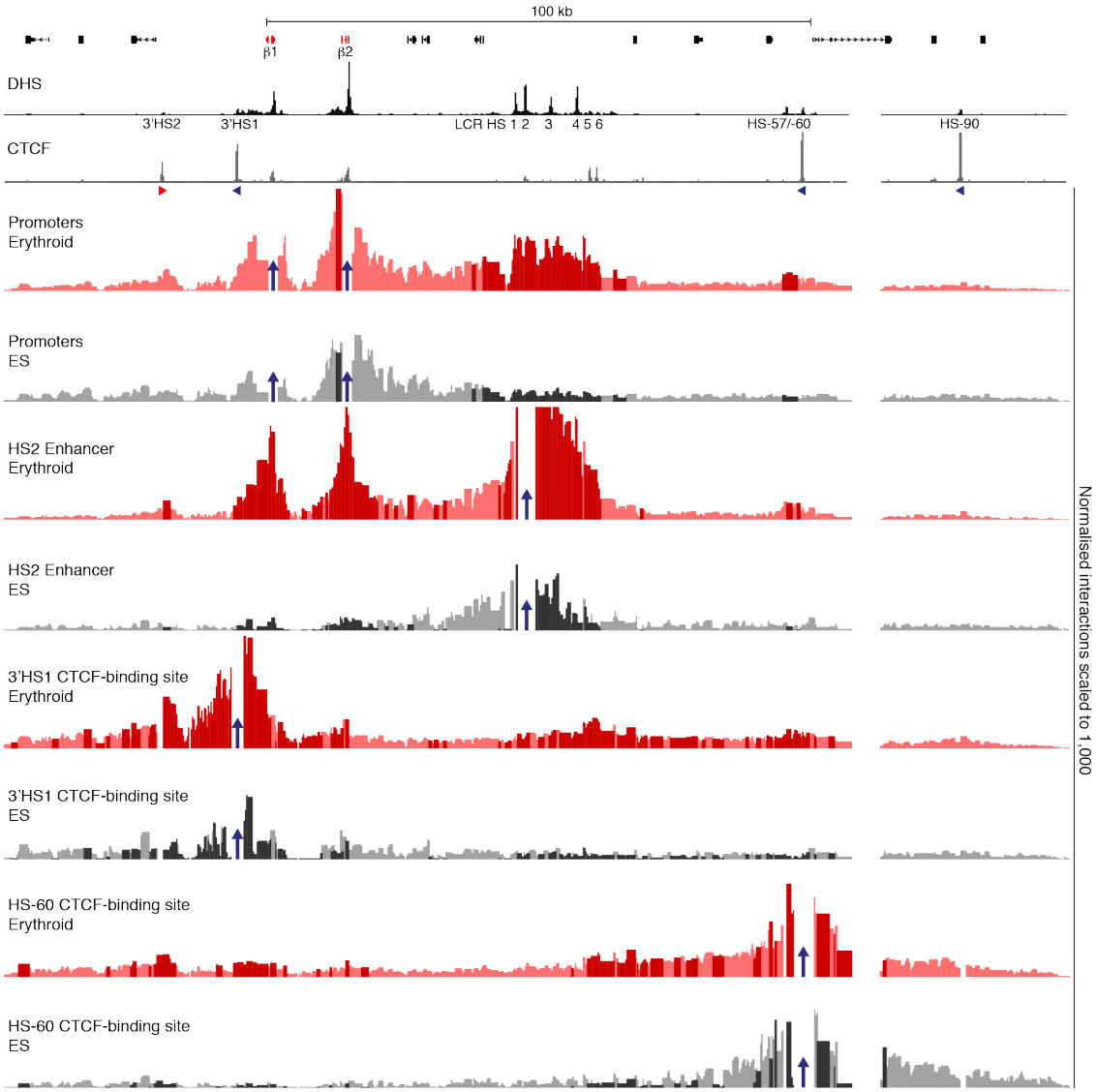


Figure 2.8: Characterisation of the interaction landscapes of the cis-regulatory elements of the β-globin locus*. High-resolution Capture-C interaction profiles of the β-globin locus from the viewpoints (indicated by blue arrows) of the β-globin promoters, the HS2 enhancer, and CTCF-binding sites 3'HS1 and 5'HS-60 in erythroid (red) and ES (grey) cells. Profiles represent the mean number of normalised unique interactions per restriction fragment from three replicates. Significantly different interactions between erythroid and ES cells are highlighted in bold colours. Gene annotation (β-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS) and CTCF occupancy are shown at the top, with arrows indicating the orientation of the CTCF-binding motifs. Coordinates: chr7:110,912,000-111,112,000.*

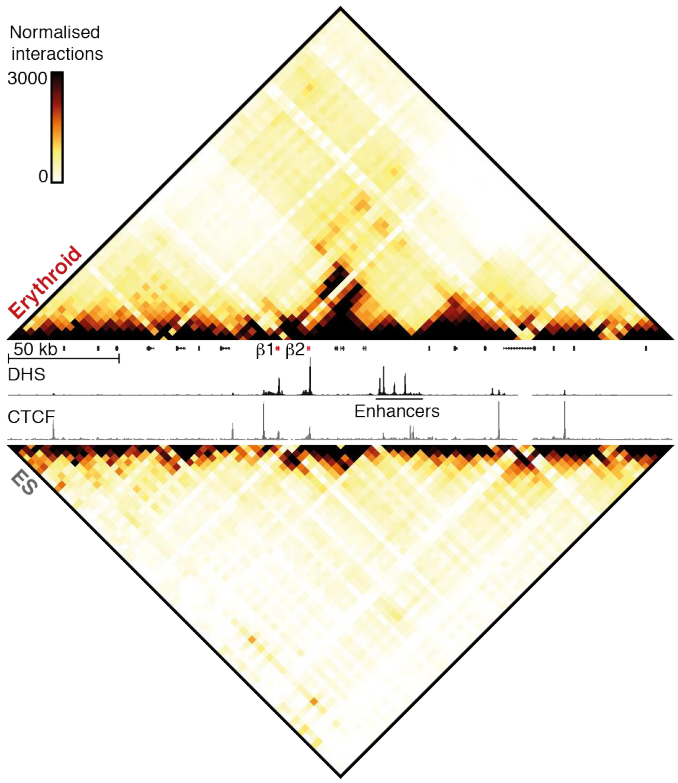


Figure 2.9: Structural conformations of the active and inactive β-globin domain*. Contact matrices (4 kb resolution) of the β-globin locus derived from Capture-C experiments with viewpoints closely spaced across the domain in erythroid (top) and ES (bottom) cells. Matrices represent the mean number of normalised unique interactions from three replicates. Gene annotation (β-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS) and CTCF occupancy are shown in the middle. Coordinates: chr7:110,840,000-111,140,000.*

To further characterise the organisation of the globin domains, we used SBS polymer models to visualise their three-dimensional structures at high resolution in their active and inactive conformations (Figures 2.10 and 2.11). These three-dimensional models clearly show that the globin loci form a single, large domain, characterised by extensive intermingling of chromatin, in ES cells, and more discrete structures in erythroid cells. This is particularly evident for the α-globin locus, where an extruded domain, encompassing the active α-globin genes and their enhancers, is formed in erythroid cells.

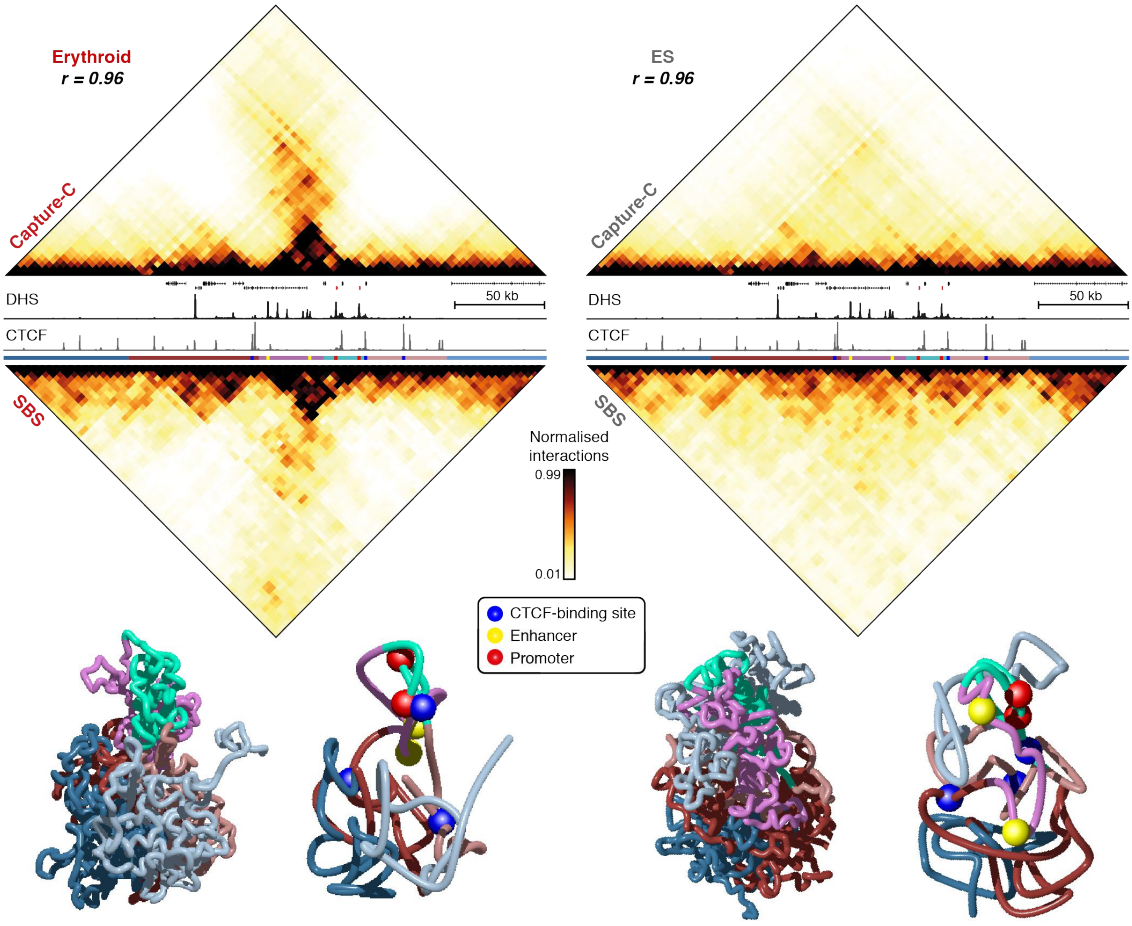


Figure 2.10: SBS models of the α-globin domain in its active and inactive conformation. *Contact matrices (4 kb resolution) of the α-globin locus derived from Capture-C data (top) and SBS simulations (bottom) in erythroid (left) and ES (right) cells. Gene annotation (α-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS), CTCF occupancy and colour legends are shown in the middle of the matrices. The correlation between the experimental and modelled matrices are indicated by the Pearson correlation coefficients (r) at the top. The three-dimensional conformations show full (left) and coarse-grained (right) representations of the α-globin locus at 4 kb resolution derived from the SBS models in erythroid (left) and ES (right) cells, with colours corresponding to the annotation in the contact matrices. Coordinates: chr11:32,000,000-32,300,000.*

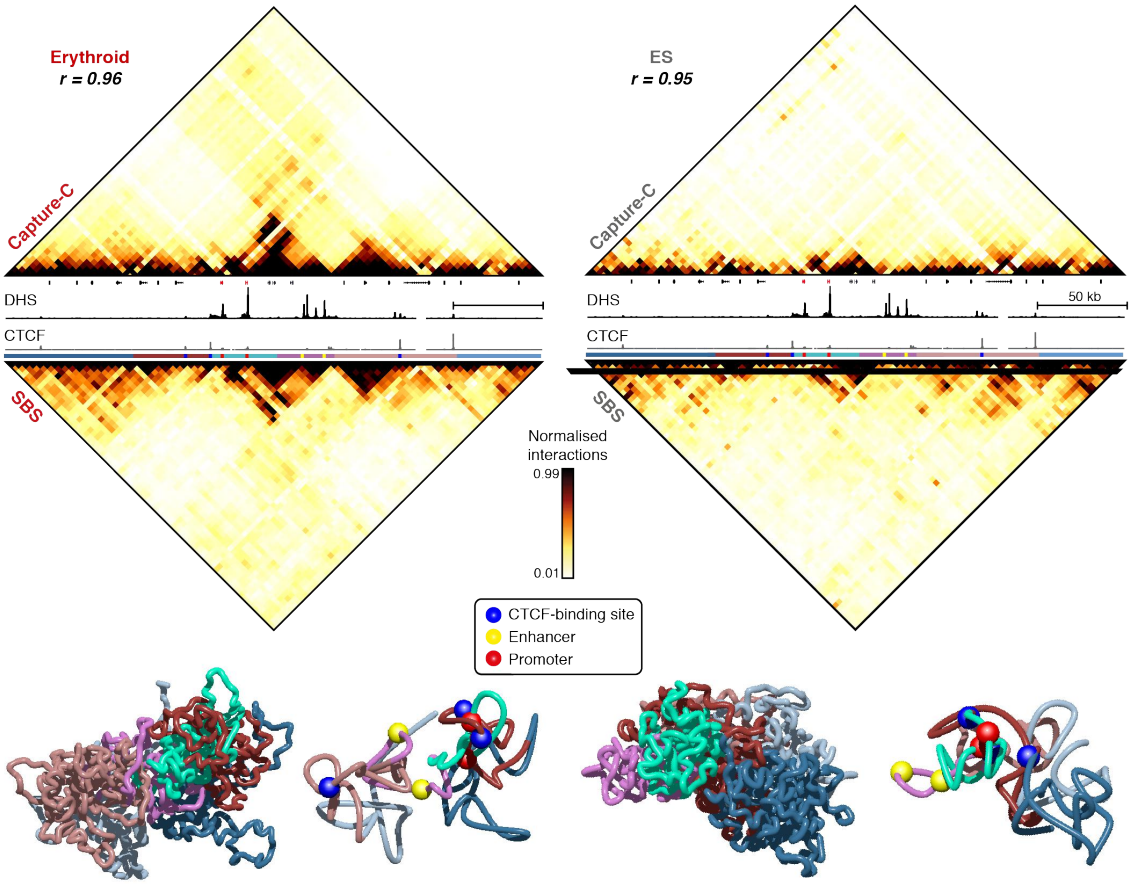


Figure 2.11: SBS models of the β-globin domain in its active and inactive conformation. *Contact matrices (4 kb resolution) of the β-globin locus derived from Capture-C data (top) and SBS simulations (bottom) in erythroid (left) and ES (right) cells. Gene annotation (β-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS), CTCF occupancy and colour legends are shown in the middle of the matrices. The correlation between the experimental and modelled matrices are indicated by the Pearson correlation coefficients (r) at the top. The three-dimensional conformations show full (left) and coarse-grained (right) representations of the β-globin locus at 4 kb resolution derived from the SBS models in erythroid (left) and ES (right) cells, with colours corresponding to the annotation in the contact matrices. Coordinates: chr7:110,840,000-111,140,000.*

Together, these results show that activation of the murine globin loci is associated with the formation of a strongly compartmentalised domain, in which enhancers and promoters interact. The flanking CTCF-binding elements do not participate in these enhancer-promoter interactions, but form preferential interactions with the chromatin bound by CTCF on either side of the domain.

Discussion

Spatial organisation into TADs has been described as a general property of chromosomes, which is pervasive throughout the genome and stable across different cell types (Dixon et al.,

2012; 2015; Rao et al., 2014; Schmitt et al., 2016). However, careful comparison of the nuclear organisation of primary erythroid and ES cells using Hi-C data at 20 kb resolution and SBS models shows that these conclusions do not hold for the extended murine globin loci.

First of all, our Hi-C data and SBS models show that the extent to which the genome is organised into TADs depends on genomic context and that large chromatin regions do not have well-defined structures. Such regions have not been emphasised, as 'typical examples' shown in the literature (Figure 1.2) generally show regions that are organised in adjacent self-interacting domains. However, in the context of investigating changes in TAD organisation upon depletion of Cohesin in HeLa cells, it has recently been described that TAD coverage (as defined by directionality indices) across the wild type HeLa genome is 56% (Wutz et al., 2017), which is in agreement with our observations at the murine globin loci.

Moreover, comparison of the Hi-C data and SBS models in primary erythroid and ES cells clearly shows that the organisation of the extended globin loci into self-interacting domains is tissue-specific. This contrasts with conclusions from the seminal Hi-C papers, which state that the structure of TADs is highly conserved across cell types (Dixon et al., 2012; 2015; Rao et al., 2014; Schmitt et al., 2016). However, close inspection of the data underlying these findings shows that the conclusions of these papers are rather overstated. Dixon et al. analyse two cell types in both mouse and human and show that 16-46% of TADs identified in Hi-C contact matrices at 20 kb resolution are tissue-specific, depending on the cell type analysed (Dixon et al., 2012). Rao et al. compare Hi-C contact matrices at (sub) 5 kb resolution between eight different human cell lines. Their detailed analysis of GM12878 cells shows that 54% of domains called in one of the other cell types were present in GM12878 cells (Rao et al., 2014). Schmitt et al. analyse Hi-C data at 40 kb resolution in 21 human cell lines and primary tissues and show that nearly 10% of all TAD boundaries are present in one cell type only and that less than 40% are conserved between all examined cell types (Schmitt et al., 2016). Though these data show that about half of the identified TADs are conserved, the papers conclude that *"[TADs] are stable across different cell types"* (Dixon et al., 2012);*"Domains tend to be conserved across cell types"* (Rao et al., 2014); *"TAD boundaries are indeed highly conserved across different cell lines and tissues"* (Schmitt et al., 2016). These conclusions might be justified when comparisons between different cell types are contrasted with completely randomised contact matrices. Indeed, Schmitt et al. perform statistical analyses to show that their results are highly significant considering that by chance only 1.7% of TAD boundaries are expected to be shared by all examined cell types (Schmitt et al., 2016). However, such comparisons are not very meaningful, as most cell types share many genomic features and are expected to contain large overlapping regions that are constitutively expressed or silenced. It

would therefore be possible that conserved domain structures reflect areas of the genome with comparable activity in the cell types considered, but that regions in which gene activity is altered are structurally reorganised. In support of this hypothesis, it has been shown that genomic regions of which the domain structures differ between cell types are highly enriched for differentially expressed genes (Dixon et al., 2012). The observed drastic reorganisation of the globin loci in erythroid cells compared to ES cells further supports these results.

In light of these findings, it is not correct to consider TADs as the invariant building blocks of the genome. Rather, these domains often represent tissue-specific structures that are closely related to genome function.

The reorganisation of chromatin structure in erythroid cells allows for the formation of tissue-specific regulatory interactions. Our high-resolution Capture-C analyses show that activation of the globin loci in erythroid cells is associated with the formation of compartmentalised chromatin domains, in which enhancers and promoters interact. These domains are delimited by regions bound by CTCF, which do not participate in the enhancer-promoter structures, but form broad interactions flanking the central compartmentalised chromatin domain.

Our structural characterisation of the active and inactive globin domains highlight the different interaction patterns formed by distinct classes of *cis*-regulatory elements. Based on analyses at lower resolution, interactions between both enhancer and promoters (Tolhuis et al., 2002) and CTCF-binding sites (Rao et al., 2014) have been described as ‘loops’, which are formed exclusively between these elements, without contacting the intervening chromatin. Our analyses at high resolution show that this model is not correct and that enhancer-promoter interactions represent preferred interactions within compartmentalised domains, in which all regions of chromatin form contacts, and which are flanked by diffusely interacting CTCF-binding sites.

However, it remains unclear how these interactions contribute to both the structure of chromatin domains and the activity of the genes contained within them. To get more insight into the function of *cis*-regulatory elements and the relationship between genome structure and function, we analysed mouse models with engineered deletions of enhancer elements and CTCF-binding sites in the α-globin locus. The effects of genetic perturbation of these elements on chromatin architecture and gene expression profiles are described in the next chapter.

Data availability

The data presented in this chapter are published in the following pre-print:

Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P., Telenius, J.M., Schwessinger, R., Liu, Y., Brown, J.M., Downes, D.J., Chiariello, A.M., Bianco, S., Nicodemi, M., Buckle, V.J., Dekker, J., Higgs, D.R., and Hughes, J.R. (2018). Single-cell chromatin interactions reveal regulatory hubs in dynamic compartmentalized domains. BioRxiv*, doi: 10.1101/307405.
**Under revision for Nature Genetics.*

The Capture-C interaction profiles from all the viewpoints in the α- and β-globin loci are available in UCSC Genome Browser Track Hubs and can be accessed with the following links:

http://sara.molbiol.ox.ac.uk/public/hugheslab/CaptureC_alpha_globin/hub.txt
http://sara.molbiol.ox.ac.uk/public/hugheslab/CaptureC_beta_globin/hub.txt

Acknowledgements

Hi-C experiments in primary erythroid cells were performed in collaboration with Professor Job Dekker. SBS polymer models were generated in collaboration with Professor Mario Nicodemi.

Chapter 3 – The roles of *cis*-regulatory elements in establishing chromatin architecture

Introduction

Our analysis of the chromatin structures of the murine globin loci described in the previous chapter has shown that these loci form compartmentalised domains in their active conformations in erythroid cells. These domains are characterised by strong interactions between enhancers and promoters and diffuse interactions between the flanking chromatin regions bound by CTCF. Though we show that the formation of these structures is associated with activation of the globin loci, it remains unclear how these tissue-specific structures are formed and what their contribution is to the regulation of gene expression. More specifically, it is not clear how the functions of *cis*-regulatory elements within such domains depend on their structure, and if and how they contribute to their formation. This chapter addresses these questions by examining established mouse models with deletions in CTCF-binding sites and enhancer elements in the α-globin locus and analysing the effects of these deletions on chromatin architecture and gene expression.

As described in Chapter 1, the zinc-finger transcription factor CTCF is thought to play an important role in the organisation of the genome across different scales. It has been shown that TAD boundaries are enriched for CTCF-binding motifs (Dixon et al., 2012), predominantly (>90%) in a convergent orientation (Gómez-Marín et al., 2015; Rao et al., 2014; Sanborn et al., 2015; Vietri Rudan et al., 2015), and that similar patterns of convergent CTCF-binding sites contribute to specific chromatin structures within TADs (de Wit et al., 2015; Guo et al., 2015). However, the precise roles of CTCF in establishing chromatin structures and/or regulating gene activity remain unclear, and recent studies have described a variety of functions for CTCF-binding elements in different contexts. These include acting as boundaries to chromatin modifications (Cuddapah et al., 2009; Handoko et al., 2011; Weth et al., 2014), facilitating interactions between regulatory elements (Hirayama et al., 2012; Liu et al., 2011) and insulating genes from tissue-specific enhancers (Dowen et al., 2014; Flavahan et al., 2016; Hnisz et al., 2016; Narendra et al., 2015; 2016).

Our analysis in Chapter 2 shows that the predominantly convergent CTCF-binding sites flanking the α-globin locus appear to define the boundaries of the chromatin domain in which the active α-globin enhancers and promoters interact. Based on this observation, these CTCF-

binding sites could have a role in facilitating the interactions between the α-globin enhancers and promoters and shielding the promoters of nearby genes from the activity of the strong α-globin enhancers. Consistent with this hypothesis, the *Nprl3* gene, which is located within the boundaries of the α-globin domain, is six-fold upregulated in erythroid cells compared with ES cells (Hanssen et al., 2017; Kowalczyk et al., 2012). However, expression of the *Mpg*, *Snrnp25* and *Rhbdf1* genes, which are further upstream and therefore outside the domain, is unchanged or repressed (Hanssen et al., 2017). This suggests that the CTCF-binding sites HS-38 and HS-39, located between these genes and the α-globin enhancers, could act as an insulator, shielding the upstream genes from the strong activity of the α-globin enhancers in erythroid cells. To test this hypothesis *in vivo*, we analysed mice carrying a double deletion in the CTCF-binding sites HS-38 and HS-39 (D3839; Figure 3.1), and examined the effects of these deletions on the structure of the α-globin domain and the activity of the genes in and nearby this domain.

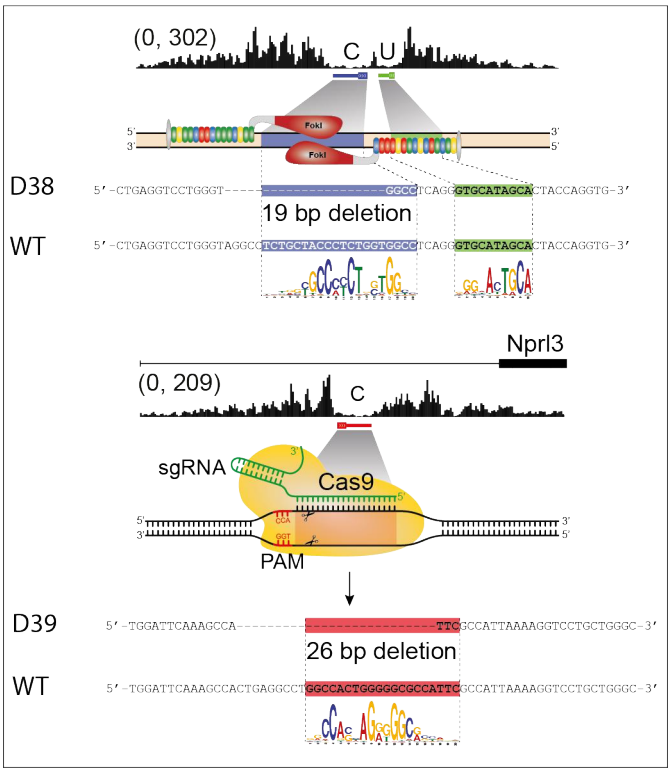


Figure 3.1: Schematic overview of the D3839 mouse model. *TALEN and CRISPR-Cas9 were used to create small deletions in the core (C) CTCF-binding motifs in HS-38 and HS-39 in the α-globin locus. Figure adapted from: Hanssen et al., 2017.*

Our analyses in Chapters 1 and 2 show that the interactions between the CTCF-binding sites surrounding the α-globin locus occur specifically in erythroid cells, despite ubiquitous binding

of CTCF to these sites. The mechanisms that drive these tissue-specific interactions between CTCF-binding sites are unknown. As they are not related to differences in the levels of CTCF bound at these sites, the interactions might be mediated by other tissue-specific processes, such as transcriptional activity, or interactions between active enhancers and promoters by bridging transcription factors. This suggests that such transcriptional processes might not only reflect chromatin architecture, but could also actively contribute to its establishment.

To get more insight into this interplay between genome structure and function, we studied the roles of enhancers in establishing chromatin architecture and regulating gene activity, by analysing mouse models in which individual or multiple enhancer elements in the α-globin locus are deleted (Anguita et al., 2002; Hay et al., 2016; Kowalczyk et al., 2012) (Figure 3.2).

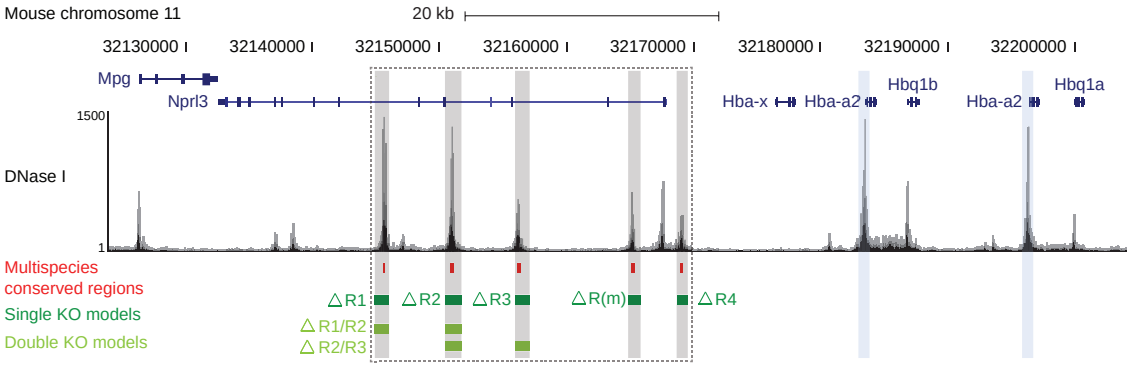


Figure 3.2: Schematic overview of the enhancer knockout mouse models. *Homologous recombination was used to delete enhancer elements (~1-2 kb), both individually and in informative combinations, in the α-globin locus. Figure adapted from: Hay et al., 2016.*

The effects of these deletions on expression of the α-globin genes have been characterised previously and are summarised in Figure 3.3 (Hay et al., 2016).

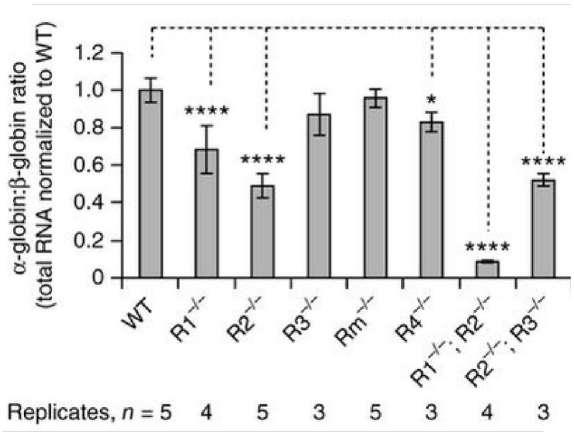


Figure 3.3: The effects of deletions of enhancer elements on α-globin expression. *Quantification of the ratio of α- and β-globin expression by NanoString in primary erythroid cells derived from mouse foetal liver cells of mice at E12.5. The bars represent mean expression levels from a minimum of three independent biological replicates, with the standard deviations indicated by error bars. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001. Figure adapted from: Hay et al., 2017.*

The expression analysis in mice carrying individual enhancer deletions shows that, although all five elements are defined as erythroid enhancers based on their chromatin signatures (Figures 1.7 and 1.9) and transient enhancer assays (Hay et al., 2016), only two of them (R1 and R2) behave as strong enhancers *in vivo*. These strong enhancers are bound by the greatest amount of Mediator and numbers of erythroid transcription factors. Examination of the mouse models in which both active elements ($R1^{-/-}$,$R2^{-/-}$), or an active and an inactive element ($R2^{-/-}$,$R3^{-/-}$), were deleted, shows that these enhancer elements work independently and in an additive fashion with respect to transcriptional output of the α-globin genes.

However, it remains unclear what the contribution of these elements is to forming and/or maintaining the chromatin domain in which they interact. To address these questions, we analysed the effects of the enhancer deletions on chromatin architecture by performing Capture-C experiments in the enhancer knockout models.

To differentiate between effects directly mediated by the enhancer elements themselves and the reduction of transcription following their deletion, we also examined the effects of inhibiting transcription by drug treatments. For these experiments, we used DRB, which blocks the kinase CDK9 in the transcription elongation factor P-TEFb and therefore inhibits transcriptional elongation (Bensaude, 2011).

Results

CTCF-binding elements in the α-globin locus mediate the formation of a chromatin domain that contributes to enhancer specificity

Mutation of the core CTCF-binding motifs in HS-38 and HS-39 resulted in a complete loss of CTCF binding at these two sites, but did not affect binding of CTCF to the other CTCF-binding elements in the domain. To analyse the effects of removal of these CTCF boundaries on the architecture of the α-globin domain, we performed Capture-C experiments from various viewpoints in and around the α-globin locus in primary erythroid cells derived from D3839 mice, and compared these to the wild type Capture-C data presented in Chapter 2. The results of these experiments are summarised in Figure 3.4.

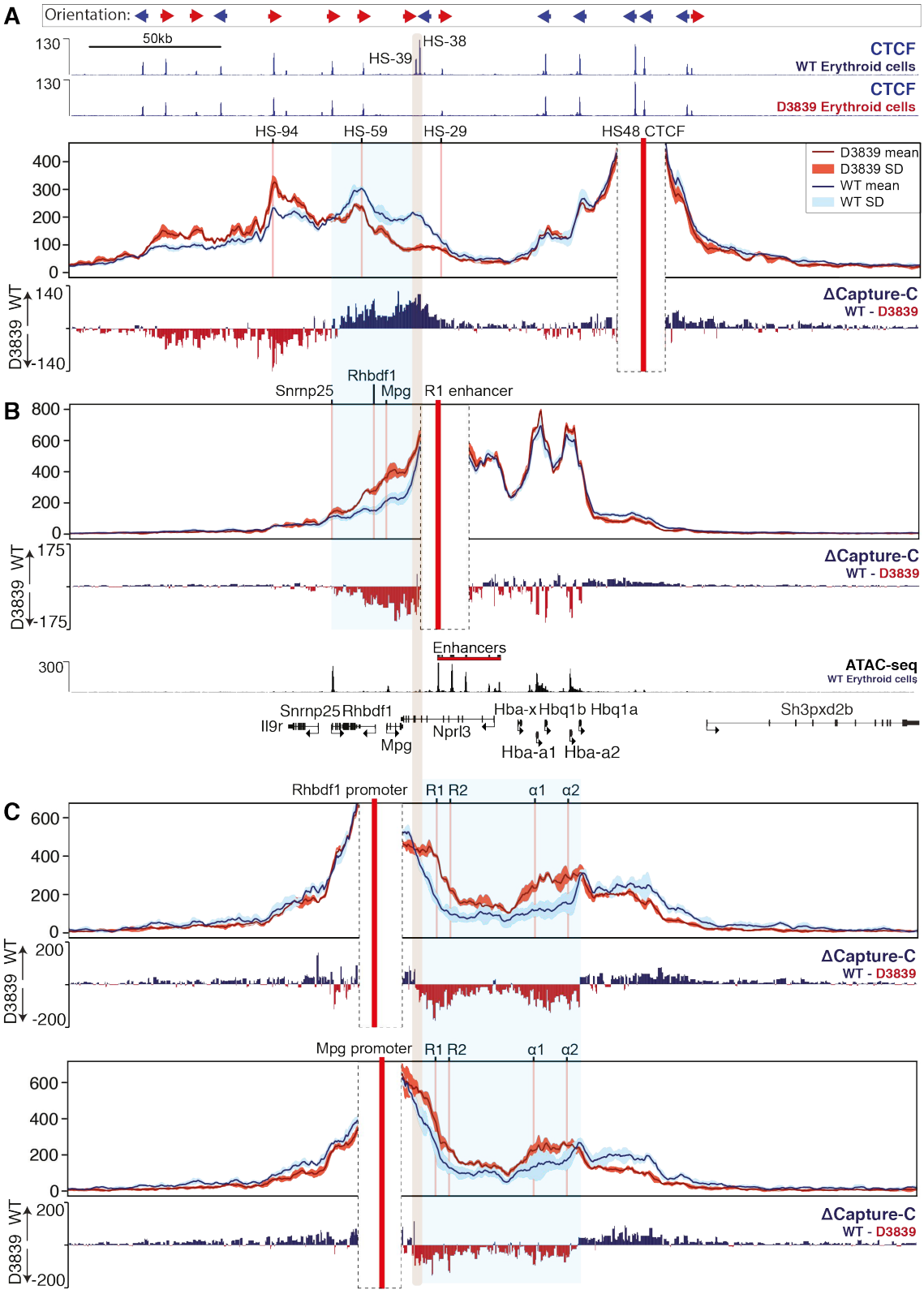


Figure 3.4: The effects of a CTCF boundary deletion on the interaction landscape of the α-globin locus. *Caption on next page.*

Figure 3.4: The effects of a CTCF boundary deletion on the interaction landscape of the α-globin locus. *Capture-C interaction profiles of the α-globin locus from the viewpoints (indicated by red vertical bars) of the* ***(A)*** *HS+48 CTCF-binding site,* ***(B)*** *R1 enhancer, and* ***(C)*** *Rhbdf1 and Mpg promoters in wild type (WT; blue) and D3839 mutant primary erythroid cells (D3839; red). The profiles represent the mean number of unique interactions from three biological replicates, with the standard deviation indicated by the line shadow, and altered interactions highlighted in the differential tracks (ΔCapture-C). The orientation of CTCF-binding motifs and occupancy of CTCF at these sites in wild type and D3839 mutant erythroid cells are shown at the top and chromatin accessibility (defined by ATAC) and gene annotation are shown in the middle. The mutated CTCF-binding sites are indicated with a shaded grey vertical bar and chromatin regions with altered interactions are highlighted with shaded blue boxes. Coordinates: chr11:32,000,000-32,300,000. Figure adapted from: Hanssen et al., 2017.*

To investigate whether mutation of the HS-38/-39 CTCF-binding sites alters the interactions between the chromatin regions flanking the α-globin locus, we analysed Capture-C profiles from the viewpoint of CTCF-binding site HS+48, located downstream of the α-globin genes. Though we still observe diffuse interactions between the regions flanking the α-globin genes in the D3839 mutant, the enriched interactions over the upstream HS-38/-39 sites in wild type erythroid cells are shifted to the next adjacent upstream CTCF-binding site (HS-59) in the D3839 mutant. Analysis from the viewpoint of the R1 enhancer shows that this shift in accompanied by increased interactions between R1 and the region of chromatin upstream of the deleted HS-38/-39 sites, which contains the *Mpg*, *Rhbdf1* and *Snrnp25* genes. However, the interactions between R1 and the α-globin promoters appear unchanged. Capture-C profiles from the viewpoints of the *Rhbdf1* and *Mpg* promoters confirm the increased interactions between these genes and the R1 enhancer, and also show increased interactions with the other α-globin enhancers and promoters.

To examine whether these changes in chromatin architecture are associated with changes in gene expression, we performed RNA-seq and real-time qPCR in primary erythroid cells derived from wild type and D3839 mice (Figure 3.5).

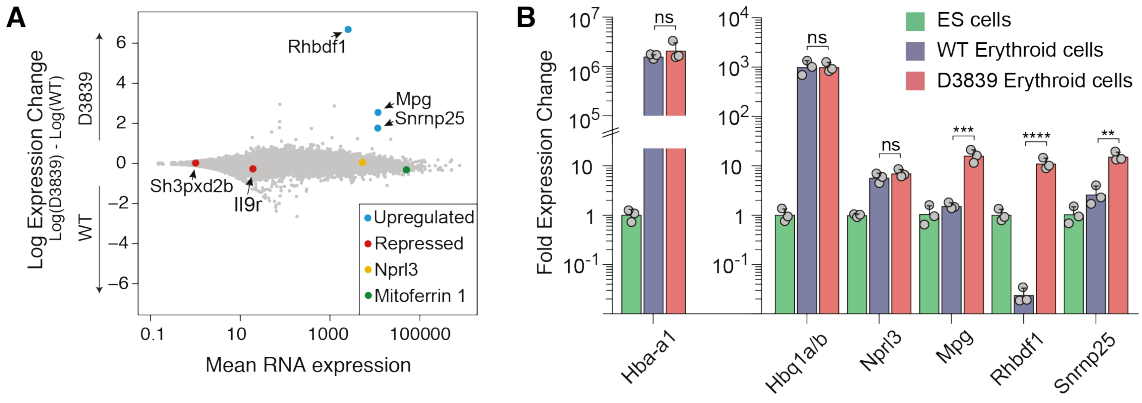


Figure 3.5: The effects of a CTCF boundary deletion in the α-globin locus on local gene expression. ***(A)*** *MA plot of RNA-seq data derived from three biological replicates of wild type (WT) and D3839 mutant primary erythroid cells (D3839), showing the log ratio of changes in gene expression versus the mean expression of these genes. Rhbdf1, Mpg and Snrnp25 (blue) are significantly upregulated in the D3839 mutants. Control genes which are unaffected by the deletions are indicated with different colours and include Mitoferrin-1 (a highly expressed erythroid gene; green), Sh3pxd2b and Il9r (repressed genes in erythroid cells; red), and Nprl3 (a housekeeping gene directly downstream of the α-globin gene cluster; yellow).* ***(B)*** *Real-time qPCR data derived from three biological replicates of wild type (WT) and D3839 mutant primary erythroid cells (D3839), and wild type ES cells. The bars represent the mean expression levels, with the standard deviations indicated by error bars. Individual data points are represented with grey dots. NS, $P > 0.05$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. Figure adapted from: Hanssen et al., 2017.*

Consistent with the unchanged interaction frequencies between the α-globin promoters and its enhancers, we observed no differences in α-globin gene activity between wild type and D3839 mutant erythroid cells. However, the increased interactions between the α-globin enhancers and the *Mpg*, *Rhbdf1* and *Snrnp25* genes are associated with a strong induction of their RNA levels in D3839 erythroid cells. *Mpg* and *Snrnp25* are housekeeping genes, which are active in wild type erythroid cells. Deletion of the HS-38/-39 CTCF boundary increases their expression twelve- and six-fold, respectively. The *Rhbdf1* gene is silenced in wild type erythroid cells by Polycomb group complexes, but is ~600-fold upregulated in the D3839 erythroid cells, well above its normal activity levels in ES cells. By contrast, the *Il9r* gene, which is located upstream of the intact HS-59 CTCF-binding sites, remained inactive in D3839 erythroid cells.

Together, these results suggest that the HS-38/-39 CTCF-binding sites function as a boundary of an erythroid-specific chromatin domain that delimits the region with which the strong α-globin enhancers interact and therefore contributes to enhancer specificity.

Enhancers are not required for the formation and maintenance of the α-globin chromatin domain

To study the roles of the enhancer elements in the establishment of erythroid-specific chromatin structures, we performed Capture-C experiments from the viewpoint of the α-globin promoters in each of the enhancer knockout models (Figure 3.6).

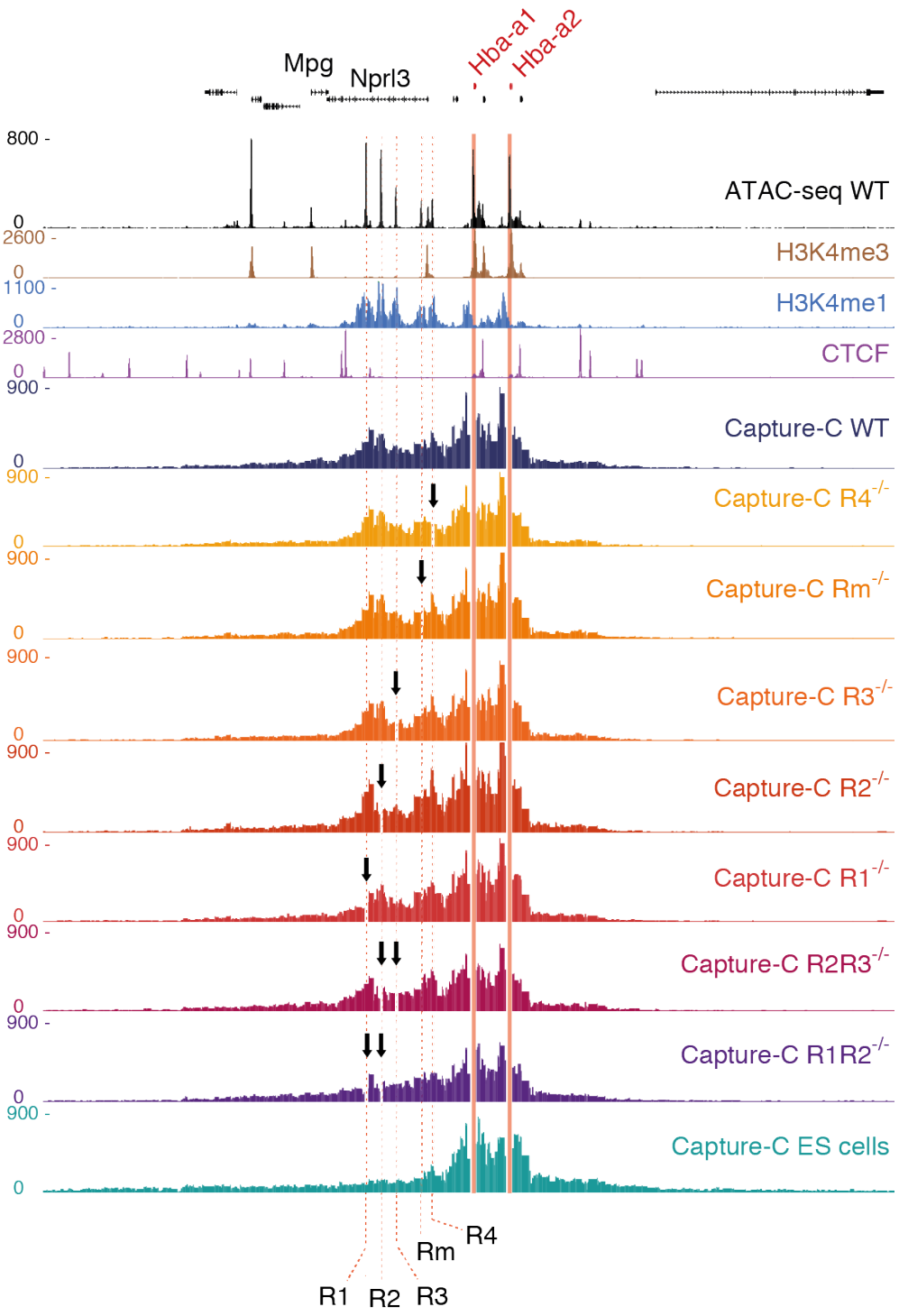


Figure 3.6: The effects of deletions of enhancer elements on the interaction landscape of the α-globin locus. *Capture-C interaction profiles of the α-globin locus from the viewpoints (indicated by red vertical bars) of the α-globin gene promoters in wild type (WT) and enhancer mutant primary erythroid cells, and wild type ES cells. The profiles represent normalised unique interactions. Gene annotation, chromatin accessibility (defined by ATAC), histone modifications (H3K4me3 and H3K4me1) and CTCF occupancy are shown at the top and the enhancer elements are highlighted by dashed red lines. Coordinates: chr11: 32,032,000–32,332,000. Figure adapted from: Hay et al., 2016.*

Comparisons of the Capture-C interaction profiles in enhancer knockout and wild type cells show very little effect of the individual deletions and modest effects of the double deletions on the structure of the α-globin locus in erythroid cells. In the R1/R2 knockout model, interactions between the α-globin promoters and the regions around the R1 and R2 enhancers are weaker compared to the wild type profiles, though still enriched compared to ES cells. The interactions between the promoters and the R4 enhancer are also slightly reduced. The R2/R3 knockout model shows diminished interactions between the α-globin promoters and the R1 enhancer compared to wild type, but no change in interactions with the Rm and R4 enhancers.

Despite these modest effects on interactions between individual *cis*-regulatory elements, these data show that the presence of all α-globin enhancers is not required for the overall formation and maintenance of an erythroid-specific chromatin domain. Even when both strong enhancers are deleted and α-globin transcription is reduced by 90%, the α-globin domain is still formed.

The structure of the α-globin domain is not dependent on active transcription

The subtle changes in the structure of the α-globin domain after deletion of both the R1 and the R2 enhancer suggest that these enhancers have a modest contribution to the formation of chromatin interactions. To investigate whether these changes are directly mediated by the enhancer deletions or reflect the resulting reduced transcription in the locus, we analysed chromatin structures in erythroid cells after inhibiting transcriptional elongation by treatment with DRB for three hours (Figure 3.7).

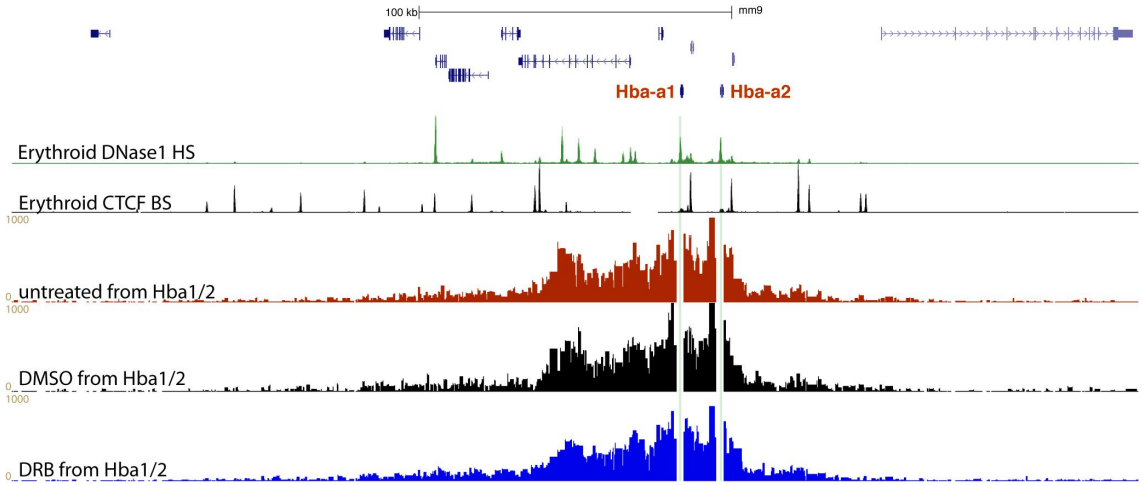


Figure 3.7: Transcription is not required for maintenance of the α-globin domain structure. *Capture-C interaction profiles of the α-globin locus from the viewpoints (indicated by green vertical bars) of the α-globin gene promoters in wild type primary erythroid cells that were untreated (red) or treated with DMSO (black) or DRB (blue) for three hours. The profiles represent normalised unique interactions. Gene annotation (α-globin genes highlighted in red), erythroid DNase1 Hypersensitive Sites (HS) and CTCF-binding sites (BS) are shown at the top, Coordinates: chr11:32,000,000-32,350,000. Figure adapted from: Brown et al., 2017.*

If the altered chromatin structures resulted from reduced transcriptional activity, the modest changes in the interaction profiles of the R1/R2 knockouts would be augmented after inhibiting active transcription throughout the domain. However, our results indicate that inhibition of transcriptional elongation has no effect on the structure of the α-globin locus, showing that active transcription is not required for the maintenance of the α-globin domain.

Discussion

Our investigation of the function of the CTCF-binding sites HS-38 and HS-39 in the α-globin domain shows that these sites act as boundary elements of erythroid-specific chromatin structures. These structures delimit enhancer interactions and therefore contribute to tissue-specific regulation of gene expression. This suggests that enhancers might be promiscuous in nature and that, rather than having inherent specificity for their cognate promoters, their functional interactions with the genes they regulate are constrained by CTCF-mediated chromatin architecture. The conclusions of our detailed *in vivo* examination of the function of CTCF-binding sites in their natural chromatin environment are in agreement with previous studies that have shown that genetic perturbations of CTCF binding in cell lines result in misregulation of gene expression (Dowen et al., 2014; Flavahan et al., 2016; Hnisz et al., 2016; Narendra et al., 2015; 2016).

Though our data provide evidence for a boundary function of CTCF-binding sites, which constrains enhancer activity and contributes to the specificity of their interactions, it remains unclear whether CTCF-mediated structures are also required to facilitate such interactions. Our data do not directly support such a facilitating role for CTCF, as the interactions between the α-globin enhancers and promoters and the expression of α-globin do not change after removal of the HS-38/-39 CTCF-binding sites. However, as the region contains many CTCF-binding sites, it is possible that there is functional redundancy and that the facilitating role is now mediated by the next CTCF-binding site (HS-59), which is only 20 kb further upstream. In support of this hypothesis, we observe increased interactions between the HS-59 CTCF-binding site and the HS+48 CTCF-binding site downstream of the α-globin locus after removal of the HS-38/-39 CTCF-binding sites.

In contrast to the CTCF-binding elements, the major α-globin enhancers appear to have a very modest role in establishing the structural α-globin domain, which is not dependent on active transcription in the locus. Even after combined deletion of the strong R1 and R2 enhancers, the erythroid-specific domain structure is formed, though with slightly reduced interactions between the remaining enhancers and promoters. This suggests that these strong enhancers are not necessary for the formation and maintenance of the domain, but are able to potentiate the interactions between the other *cis*-regulatory elements. However, we have only examined the effects of removing enhancer elements individually or in pairs. It is therefore possible that *some* but not *all* enhancers are required for the formation of tissue-specific chromatin domains, either because this is a specialist function of some enhancer elements or because these elements function redundantly in establishing chromatin architecture.

The complex interactions patterns between the five α-globin enhancer elements, their additive contribution to globin expression, and their dependency on CTCF-mediated chromatin architecture, indicate an intricate interplay between *cis*-regulatory elements to ensure robust regulation of gene expression. However, the precise structural mechanisms underlying these processes remain unclear, as conformational data are derived from populations of cells and therefore reflect multiple dynamic structures that obscure specific higher-order complexes associated with active gene regulation in individual cells. Therefore, to understand how *cis*-regulatory elements cooperate and regulate gene expression, analysis of single-cell chromatin conformations is required. In the next chapter, we present a new 3C technique that enables analysis of concurrent chromatin interactions at individual alleles and provides new insights into structural cooperation between *cis*-regulatory elements and the dynamic formation of chromatin domains.

Data availability

The data presented in this chapter are published in the following articles and pre-prints:

Hanssen L.L.P., Kassouf M.T.*, Oudelaar A.M.*, Biggs D., Preece C., Downes D.J., Gosden M., Sharpe J.A., Sloane-Stanley J.A., Hughes J.R., Davies B., and Higgs D.R. (2017). Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nature Cell Biology 19, 952–961.
**Co-second authorship reflected in alphabetical order on publication.*

Hay D., Hughes J.R., Babbs C.*, Davies J.O.J.*, Graham B.J.*, Hanssen L.L.P.*, Kassouf M.T.*, Oudelaar A.M.*, Sharpe J.A.*, Suciu M.C.*, Telenius J.*, Williams R.*, Rode C., Li P.S., Pennacchio L.A., Sloane-Stanley J.A., Ayyub H., Butler S., Sauka-Spengler T., Gibbons R.J., Smith A.J.H., Wood W.G., and Higgs D.R. (2016) Genetic dissection of the α-globin super-enhancer in vivo. Nature Genetics 48, 895–903.
**Co-third authorship reflected in alphabetical order on publication.*

Brown J.M., Roberts N.A., Graham B.J., Waithe D., Lagerholm C., Telenius J.M., De Ornellas S., Oudelaar A.M., Szczerba I., Babbs C., Kassouf M.T., Hughes J.R., Higgs D.R., and Buckle V.J. (2017). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. BioRxiv*, doi: 10.1101/234427.
**Under revision for Nature Communications.*

Acknowledgements

The data presented in this chapter are part of collaborative projects in the Higgs and Hughes groups. Lars L.P. Hanssen generated the D3839 mouse model and performed the expression analyses; we worked together to produce the Capture-C data to characterise the effects of these deletions on chromatin architecture. Deborah Hay cultured the cells that I used for the Capture-C experiments in the enhancer knockout models. I worked together with Jill M. Brown to culture and treat the cells that I used for the Capture-C experiments after transcription inhibition.

Chapter 4 – The three-dimensional conformation of the murine globin loci in single cells

Introduction

The data presented in the previous two chapters show that the murine globin loci are organised in compartmentalised chromatin domains, in which multiple enhancers cooperate and interact with gene promoters in structures mediated by interactions between CTCF-binding elements. These tissue-specific structures reflect a complex interplay between the *cis*-regulatory elements in the globin domains. However, insights into the precise structural relationships between these elements remain limited, as Capture-C and Hi-C provide average pair-wise interaction frequencies derived from millions of cells. These data therefore inform about the probability that two elements interact in specific cell populations, but do not provide insights into higher-order three-dimensional structures of *cis*-regulatory elements in individual cells.

Analysis of single-cell chromosome conformations can address many important outstanding questions regarding structural gene regulatory mechanisms. For example, dissecting the higher-order configurations and heterogeneity of enhancer-promoter structures in individual cells can provide insights into how multiple enhancers and promoters work together to regulate gene activity. Furthermore, unravelling the dynamic structures of boundary elements can clarify their roles in the formation of chromatin domains and functional enhancer-promoter interactions.

Single-cell Hi-C analyses (Nagano et al., 2013) have provided some insights into chromosomal structures in individual cells at large scales, but despite recent advances in sensitivity (Stevens et al., 2017) and throughput (Nagano et al., 2017), genome coverage is random and incomplete and resolution is low (100 kb or lower). A more recent approach, Genome Architecture Mapping (GAM), uses sequencing of thin nuclear sections to analyse chromatin contacts in single cells (Beagrie et al., 2017). Though the resolution of GAM is significantly higher (30 kb) and enables detection of structures formed by interactions between multiple transcriptionally active loci in single nuclei, it is still not informative at the level of individual *cis*-regulatory elements within a single gene locus.

Because 3C libraries contain large concatemers of re-ligated fragments that were in close proximity in individual nuclei, information about three-dimensional chromosomal conformations in single cells could also be accessed by analysing several neighbouring fragments in the 3C concatemer simultaneously. However, the read-out of multiple ligation

junctions in existing techniques is extremely inefficient (2% or lower). Therefore, analyses of data derived from 4C experiments (Fabre et al., 2017; Jiang et al., 2016), a modified in situ Hi-C approach (Darrow et al., 2016) and in-gel 3C assays (Ay et al., 2015; Gavrilov et al., 2014) have been difficult to interpret due to the low sensitivity, and resulted in conflicting conclusions. The recently developed 3way-4C approach (Olivares-Chauvet et al., 2016) is capable of detecting several thousands of triple ligation events. Though this significant improvement in efficiency enabled the identification of spatial hubs containing multi-way contacts between the Polycomb-repressed Hox clusters in mouse ES cells, the sensitivity of the technique is not sufficient for robust detection of higher-order structures formed by individual *cis*-regulatory elements.

To overcome these limitations, we developed a new 3C technique that can identify multi-way interactions with viewpoints of interest at high resolution with unprecedented sensitivity, and used this technique to characterise the structural conformations of the murine globin loci in single cells.

Results

New 3C techniques to analyse multi-way chromatin interactions in single cells

A 3C technique to analyse multi-way chromatin interactions of interest in individual cells with high efficiency, sensitivity and resolution must be able to identify multiple interacting fragments within the long ligation products in 3C libraries with high accuracy and throughput. Our initial approach to this challenge was C-Trap, which is based on the use of long-read Nanopore sequencing to identify interactions occurring simultaneously with a known interaction between two fragments of interest. It uses long-range PCR amplification with primers targeting these two fragments to enrich for ligation products in a 3C library that contain the interaction of interest and ‘trap’ the intervening fragments that were interacting simultaneously. After size selection to deplete short PCR products (< 700bp), the Nanopore MinION sequencing platform is used to identify all trapped fragments (Figure 4.1A).

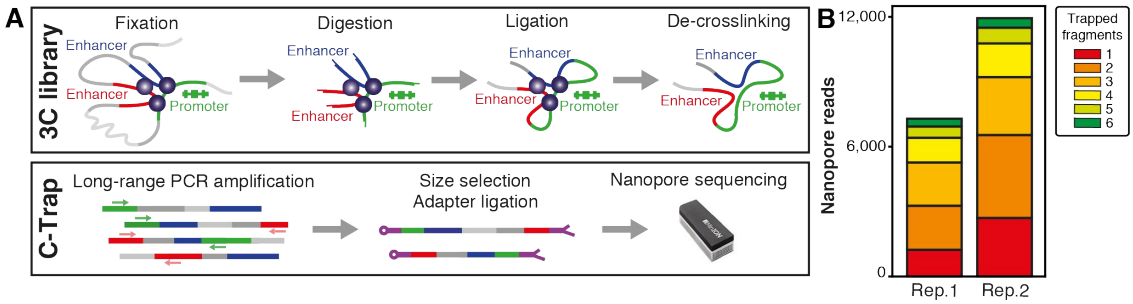


Figure 4.1: Overview of the experimental procedure and data output of C-Trap. (A) *Overview of C-Trap.* ***(B)*** *Number of reads and trapped fragments per read detected in a C-Trap experiment in which the interaction between the R2 enhancer and the α-globin promoters is targeted in erythroid cells (two biological replicates). Read numbers represent specifically amplified reads in which both primer sequences could be detected at the ends.*

The use of the Nanopore sequencing platform, which is capable of producing long reads up to hundreds of kb, has the advantage that it has the potential to detect many fragments in 3C ligation products in individual reads. However, due to several technical limitations, the sensitivity of C-Trap is not very high and the data produced by a C-Trap experiment is relatively sparse (Figure 4.1B). This is mainly due to the low throughput and read quality of the Nanopore MinION. A typical run (MAP-005 chemistry) produces ~60,000 reads with ~85% accuracy (Jain et al., 2015). The high error rate in the Nanopore reads also complicates bioinformatic analysis of the data. PCR duplicates cannot be identified reliably, which hampers accurate quantification of the detected interactions. Though the use of long-read sequencing should enable detection of many simultaneously interacting fragments in ligation products, the majority of the reads contain only a few trapped fragments (Figure 4.1B). This could be due to PCR amplification skew towards preferential enrichment of shorter fragments and/or lower quality of longer Nanopore reads. A final limitation is that C-Trap can only be performed in cell types in which the interaction of interest is strong, as the PCR enrichment will only be efficient if the targeted interacting fragments are in relatively close proximity in most ligation products. It is therefore not possible to produce interaction profiles in cells in which the targeted fragments do not interact, which limits its use for comparative analyses in cell types with different levels of activity in the gene loci of interest.

The technical difficulties that limit the sensitivity of C-Trap could be overcome by using Illumina sequencing, which is superior in both throughput and accuracy compared to all other sequencing platforms. However, the length of Illumina reads is limited. The MiSeq system supports 600 cycles of sequencing, producing up to 50 million 300 bp paired-end reads per run. The NextSeq system can only sequence for 300 cycles, but has a much higher throughput and

can generate up to 800 million 150 bp paired-end reads. The shorter read length means that a method based on Illumina sequencing can only read out a relatively small part of the ligated 3C concatemers. However, identification of only three interacting fragments is sufficient to reconstruct higher-order three-dimensional relationships between *cis*-regulatory elements, if a 3C library, which contains ligation products derived from millions of cells, is sampled in great depth. We therefore built upon the sensitivity of the Capture-C assay (Davies et al., 2015; Hughes et al., 2014), which uses oligonucleotide capture to enrich for interactions of interest up to a million-fold. We modified and optimised this approach to develop the Tri-C technique, which can identify multi-way interactions with viewpoints of interest with high sensitivity (Figure 4.2A).

For efficient detection of multiple ligation junctions between interacting fragments in the 3C concatemer, Tri-C libraries are generated using an enzyme selected to create relatively small DNA fragments (~200 bp) for the viewpoints of interest. We chose the restriction enzyme NlaIII, which has a recognition sequence of four bp and therefore cuts on average every 256 bp. Sonication of these libraries to ~450 bp generates DNA fragments of which about ~50% contain three or more interacting restriction fragments. Highly optimised enrichment using capture oligonucleotides targeting the middle of the small viewpoint fragments allows for efficient detection of both fragments interacting with these viewpoints with conventional short read Illumina sequencing. The combination of efficient enrichment for multi-way interactions with high library complexity and high-throughput sequencing allows for the detection of millions of multi-way interactions, generating 3C profiles at unprecedented depth (Figure 4.2B). Importantly, Tri-C allows for multiplexing both viewpoints and samples, enabling analyses of multiple genomic regions and cell types of interest in a single experiment. Because Illumina sequencing provides accurate identification of the random sonication ends of the reads, these can be used as unique molecular identifiers to filter out PCR duplicates, allowing for quantitative analysis of the detected multi-way interactions.

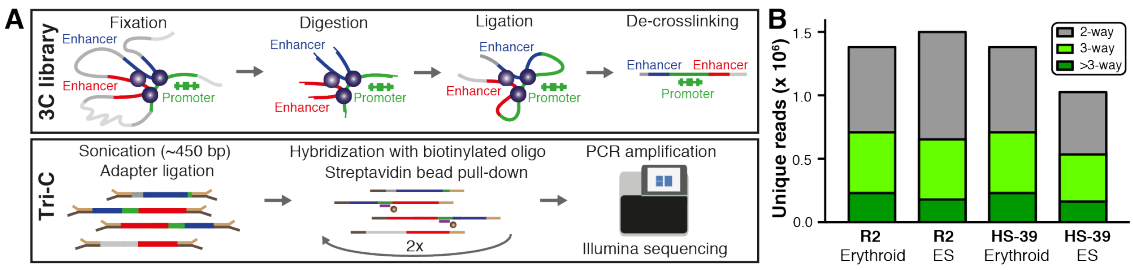


Figure 4.2: Overview of the experimental procedure and data output of Tri-C. (A) *Overview of Tri-C.* ***(B)*** *Number of unique reads containing pair-wise and multi-way interactions detected by Tri-C for viewpoints in the α-globin locus.*

To validate that Tri-C detects reliable multi-way interactions in individual nuclei, we performed several additional experiments and analyses.

To confirm that capturing multiple ligation junctions simultaneously does not introduce a bias in the interactions that can be detected resulting from differences in restriction fragment length or other technical artefacts, we compared the pair-wise interactions derived from Tri-C to Capture-C interaction profiles from the same viewpoint (Figure 4.3). These profiles are very similar, which demonstrates that detection of multiple ligation events simultaneously does not skew the data. The subtle differences between the Capture-C and Tri-C profiles are likely related to the use of different restriction enzymes during 3C library preparation (DpnII in Capture-C; NlaIII in Tri-C).

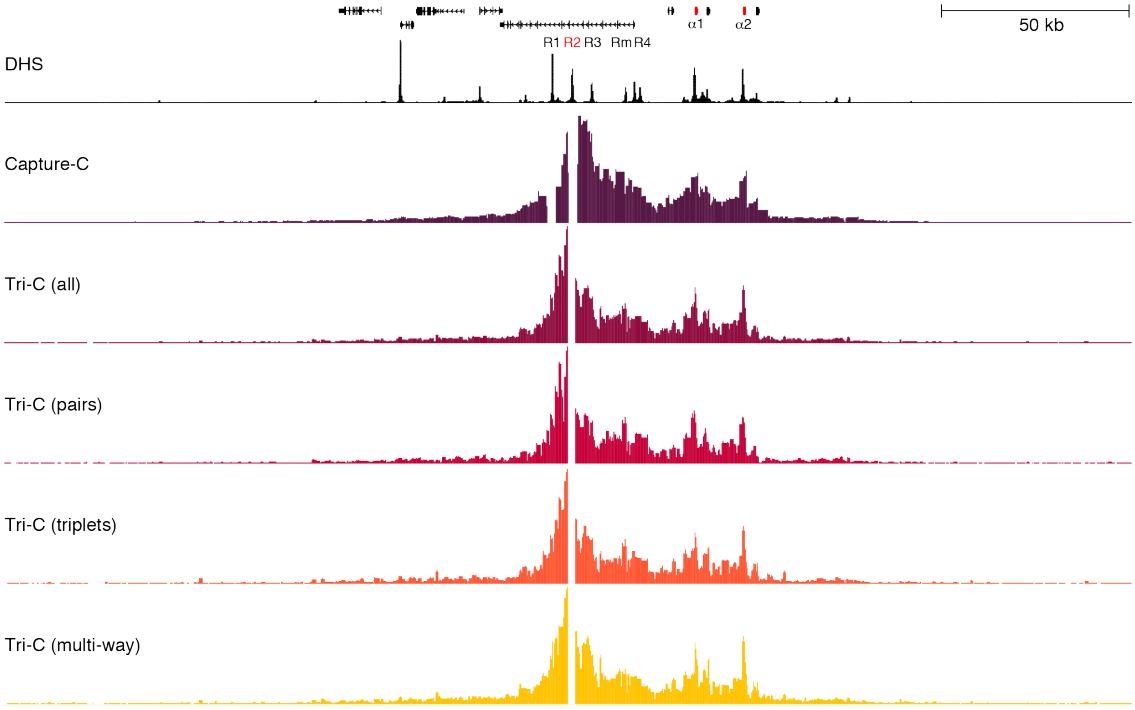


Figure 4.3: Validation of pair-wise interactions detected by Tri-C. *Comparison of the Capture-C interaction profile (purple) from the viewpoint of the R2 enhancer in the α-globin locus to pair-wise interactions derived from Tri-C data. The first Tri-C profile (maroon) shows pair-wise interactions derived from all unique Tri-C reads; the tracks below show pair-wise interactions derived from unique Tri-C reads containing two (red), three (orange) or multiple (yellow) interacting fragments. Gene annotation (α-globin genes highlighted in red) and DNaseI Hypersensitive Sites (DHS) are shown at the top. Coordinates: chr11:32,000,000-32,300,000.*

Though C-Trap produces relatively sparse data, it does provide an independent dataset to further validate the multi-way interactions detected by Tri-C. To this end, we performed C-Trap experiments targeting the interaction between the R2 enhancer and the α-globin

promoters in erythroid cells and compared these data to interactions detected in Tri-C reads containing both the R2 viewpoint and the α-globin promoters (Figure 4.4). These interaction profiles look very similar: interactions are confined to the α-globin domain and enriched over the enhancers. This confirms the validity of the multi-way interactions detected by Tri-C.

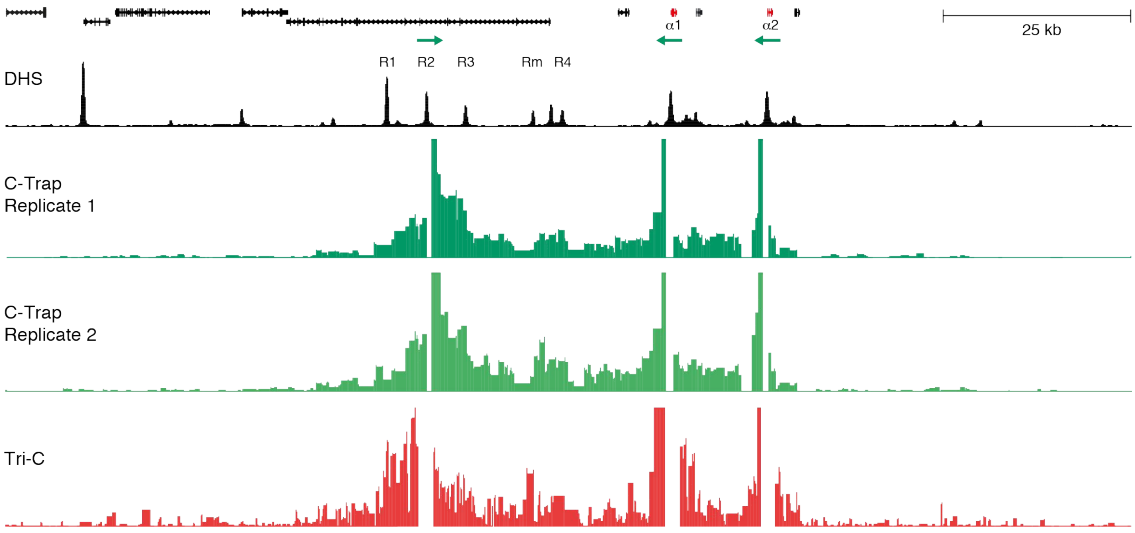


Figure 4.4: Validation of multi-way interactions detected by Tri-C. *Comparison of multi-way interaction profiles generated from C-Trap fragments interacting with the R2 enhancer and the α-globin promoters, and from Tri-C reads containing both the R2 viewpoint and the α-globin promoters, generated from primary erythroid cells. Gene annotation (α-globin genes highlighted in red) and DNaseI Hypersensitive Sites (DHS) are shown at the top. The location of the C-Trap primers is indicated by green arrows. Coordinates: chr11:32,095,000-32,245,000.*

We have previously shown that interactions detected by Capture-C predominantly (>95%) originate from the same allele, and thus represent ligation events in individual nuclei, by analysing the distribution of heterozygous SNPs in interaction profiles generated in F1 hybrids (Davies et al., 2015). To confirm that Tri-C interactions represent single-cell chromosome conformations, we assessed the allele-specificity of the detected multi-way interactions by analysing the distribution of SNPs in the β-globin locus in ES cells (Figure 4.5). Since 97% of the examined reads are allele-specific, this analysis confirms that the multi-way interactions detected by Tri-C are predominantly derived from individual nuclei.

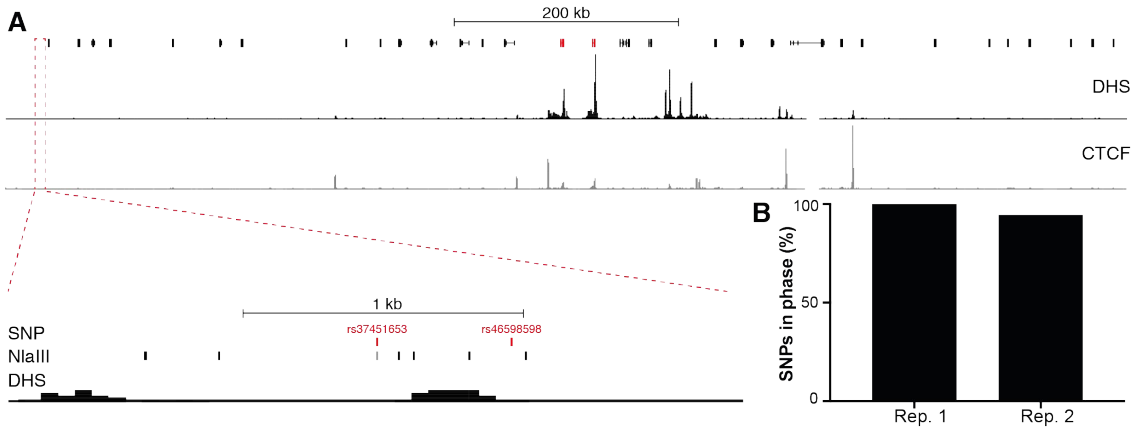


Figure 4.5: Interactions detected by Tri-C are allele-specific. (A) *E14 ES cells are heterozygous for SNPs rs37451653 (C/T) and rs46598598 (G/C), which are located downstream of the β-globin genes. The minor variant of rs37451653 creates a new NlaIII cut site (CACG/CATG). Gene annotation (β-globin genes highlighted in red), erythroid DNaseI Hypersensitive Sites (DHS) and CTCF occupancy are shown at the top. Coordinates: chr7:110,713,500-111,213,500.* ***(B)*** *The small NlaIII fragment created by the minor rs37451653 variant (T) is predominantly (97%) captured with the minor rs46598598 variant (C) in unique reads containing both SNPs in two independent replicates of ES cells.*

Multiple enhancers and promoters form higher-order chromatin structures in which they interact simultaneously

Our Capture-C analyses in Chapter 2 suggest that all enhancers and promoters in the globin loci can form interactions. However, as these data are derived from pair-wise interactions in populations of cells, it remains unclear how these elements interact together in the three-dimensional cell nucleus. It has been proposed that multiple *cis*-regulatory elements can form complex 'hub' structures, in which they all interact simultaneously (Figure 1.10) (Tolhuis et al., 2002). However, as this model is based on two-dimensional data derived from bulk populations, it remains speculative. Current 3C data cannot distinguish the formation of higher-order structures in which multiple enhancers and promoters interact simultaneously within individual nuclei from a model in which elements compete for the formation of exclusive interactions. To resolve this question, we used Tri-C to analyse the three-dimensional structures around strong enhancers in the murine globin loci in erythroid and ES cells.

To visualise these three-dimensional structures, we represent the Tri-C multi-way interactions in contact matrices in which we exclude the invariant viewpoint of interest and plot the frequencies with which two elements were captured simultaneously with this viewpoint. Note that although these matrices contain data derived from populations of cells, they only show interactions occurring simultaneously at single alleles, and therefore provide insight into the

patterns of interactions and three-dimensional relationships between multiple *cis*-regulatory elements in individual nuclei. Mutually exclusive contacts between elements appear as depletions at the intersections between these elements in the contact matrix, whereas preferential simultaneous interactions in higher-order structures are visible as enrichments at these foci.

Analysis from the R2 viewpoint, the strongest α-globin enhancer (Hay et al., 2016), shows that specific multi-way interactions with the other enhancers (R1, R3, Rm and R4) and promoters are formed in erythroid cells (Figure 4.6A). These interactions are absent in ES cells (Figure 4.6B) and direct comparison shows strong, significant enrichments for these multi-way enhancer-promoter interactions in erythroid cells (Figure 4.6C). For example, multi-way interactions between R2, R1 (the second strongest α-globin enhancer (Hay et al., 2016)), and either of the α-globin promoters ('R2-R1-Hba triplets') are over ~4-fold enriched (Figure 4.6D). Importantly, these interactions are still highly significantly enriched in erythroid cells compared to ES cells after correcting for the increased pair-wise interactions between the individual enhancers and the α-globin promoters (~3.5-fold; $p = 0.0005$; Figure 4.6E). To further investigate how the multi-way R2 interactions detected in erythroid cells relate to the underlying pair-wise contact distribution, we compared the R2 Tri-C matrices to the contact matrices derived from the multiplexed Capture-C experiments at 4 kb resolution (Figure 4.6F). This shows that multi-way interactions between R2, the other α-globin enhancers and promoters are significantly enriched, beyond what would be expected based on the pair-wise interactions between these elements in erythroid cells. To examine these enrichments at higher resolution, we analysed 'double-anchored' interaction profiles generated from reads containing both the R2 and R1 enhancer. To investigate whether reads containing both major α-globin enhancers were more likely to interact with the other α-globin enhancers and promoters, we compared these profiles to conventional interaction profiles containing all interactions with the R2 viewpoint. Though the double-anchored profiles have a relatively high signal-to-noise ratio due to the many possible triplets that can be formed, interactions with all the other enhancers and the *Hba-a1* promoter are significantly enriched. Together, these analyses show that the multi-way enhancer-promoter interactions in the α-globin locus represent erythroid-specific higher-order structures, and do not simply reflect increased pair-wise interactions between the participating elements in erythroid cells.

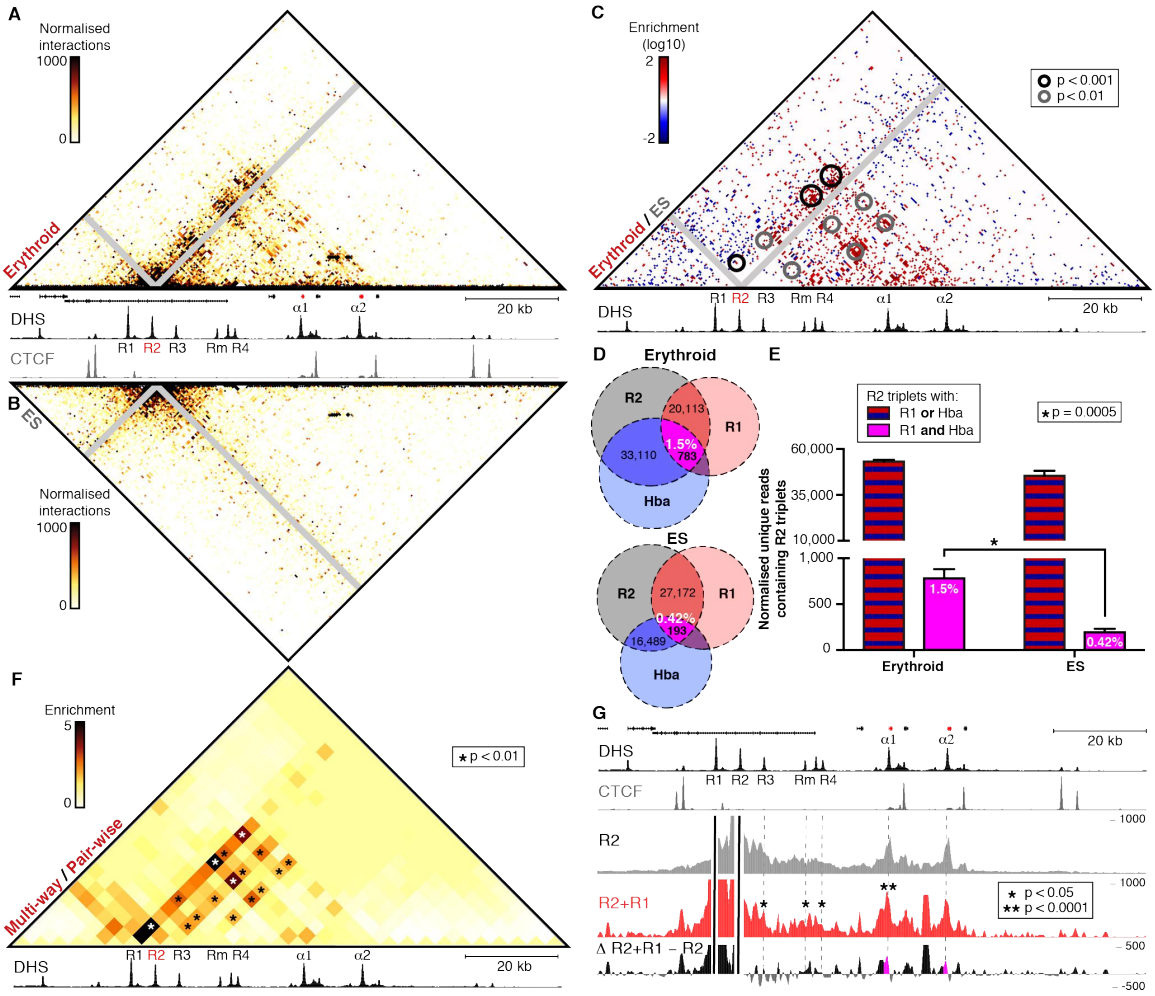


Figure 4.6: Analysis of multi-way interactions between enhancers and promoters in the α-globin locus. *Tri-C matrices represent the mean number of normalised, unique interaction counts in ten erythroid or seven ES replicates, displayed at 500 bp resolution, with proximity contacts around the R2 viewpoint excluded (grey diagonal). The strong signal at the intersection of the two α-globin genes is a bioinformatic artefact resulting from homology between the duplicated genes. Matrices are annotated with gene positions (α-globin genes highlighted in red), erythroid DNaseI hypersensitivity (DHS) and/or CTCF occupancy. Coordinates: chr11:32,120,000-32,240,000.* ***(A)*** *Contact matrix showing multi-way interactions with R2 in erythroid cells.* ***(B)*** *Contact matrix showing multi-way interactions with R2 in ES cells.* ***(C)*** *Contact matrix showing differential multi-way interactions with R2 that are enriched in erythroid (red) or ES (blue) cells. Statistically significant enriched interactions in erythroid cells are highlighted.* ***(D)*** *Venn diagram illustrating the proportion of R2-R1-Hba triplets relative to all triplets containing pair-wise R2-R1 and R2-Hba interactions in erythroid and ES cells. Numbers represent normalised counts of unique reads containing the specified three-way or pair-wise interactions.* ***(E)*** *Comparison of the absolute numbers of R2-R1-Hba triplets (shown on the y-axis) and the proportion of these relative to all triplets containing pair-wise R2-R1 and R2-Hba interactions (represented as percentages in the bar graphs) in erythroid and ES cells. Bar graphs represent average normalised unique read counts in ten erythroid and seven ES replicates, with the standard error indicated by the error bars. Both absolute and relative R2-R1-Hba triplet counts are significantly enriched in erythroid cells (~4-fold, p = 0.0002 and ~3.5-fold, p = 0.0005, respectively).* ***(F)*** *Contact matrix (4 kb resolution) showing enrichment of multi-way interactions with R2 in erythroid cells over pair-wise contact frequencies derived from multiplexed Capture-C data in three biological replicates. Statistically significant enriched multi-way interactions are highlighted.* ***(G)*** *Comparison of a double-anchored*

interaction profile containing fragments interacting with both the R2 and R1 enhancers (red) to the conventional R2 interaction profile (grey) in erythroid cells. Enrichment of multi-way R2-R1 interactions over pair-wise R2 interactions are shown in the differential profile at the bottom (black), with the R2-R1-Hba interactions highlighted (magenta). Profiles represent windowed, normalised, unique interaction counts in ten replicates. Statistically significant enriched multi-way interactions with cis-regulatory elements of α-globin are indicated. Note that the double-anchored interaction profile has a relatively high signal-to-noise ratio and therefore contains some spurious peaks, including a strong, yet insignificant, enrichment of an unusually large (> 1.5 kb) NlaIII fragment containing many repeats in between the two α-globin promoters.

As a further validation of the robustness of the Tri-C technology, the detected multi-way interactions in the α-globin locus are highly reproducible between replicates (Figure 4.7).

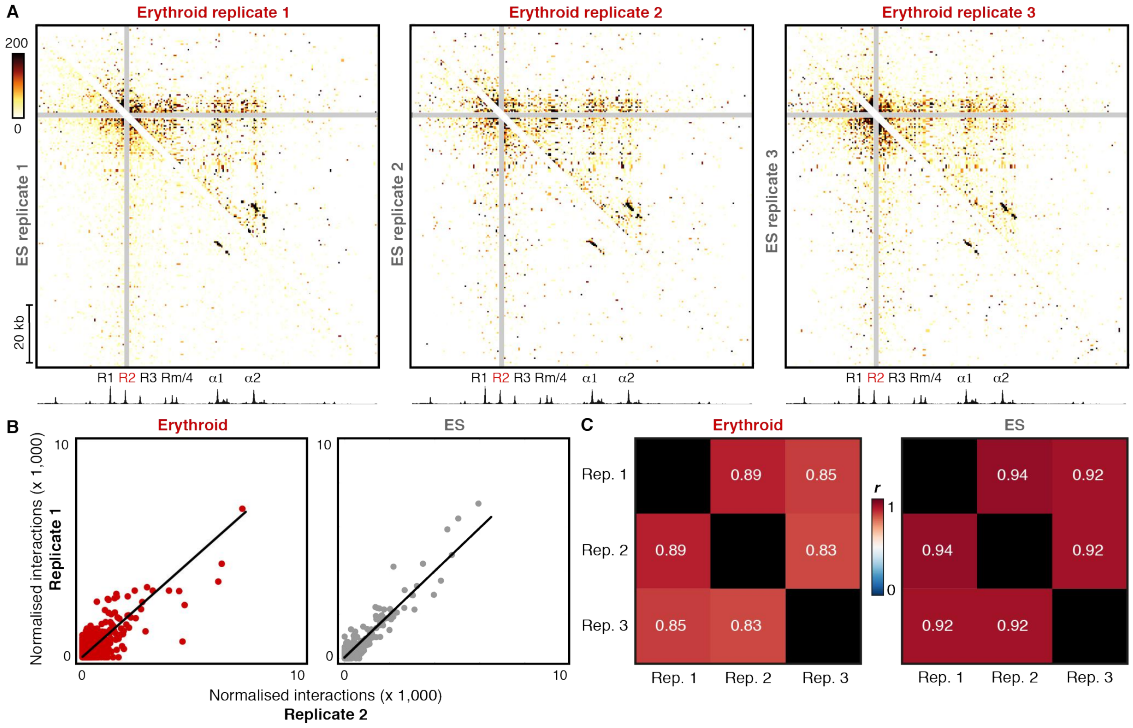


Figure 4.7: Tri-C data are highly reproducible. ***(A)*** *Comparison of Tri-C contact matrices (500 bp resolution) from the viewpoint of the R2 enhancer in the α-globin domain in three biological replicates. Erythroid samples are shown in the top right half of the matrix and ES samples in the bottom left half. Proximity contacts around R2 are excluded (grey diagonals). The strong signal at the intersection of the two α-globin genes is a bioinformatic artefact resulting from homology between the duplicated genes. Erythroid DNaseI Hypersensitive Sites are shown at the bottom of the matrices. Coordinates: chr11:32,120,000-32,240,000.* ***(B)*** *Correlations of multi-way interaction frequencies detected from the R2 viewpoint between individual erythroid (left) and ES (right) replicates.* ***(C)*** *Correlation matrices showing Pearson correlation coefficients (r) of multi-way interaction frequencies detected from the R2 viewpoint between replicates of erythroid (left) and ES (right) cells.*

Tri-C analysis of HS2, one of the strongest β-globin enhancers (Bender et al., 2012), shows very similar structures, in which multiple enhancers and promoters interact specifically in erythroid cells (Figure 4.8).

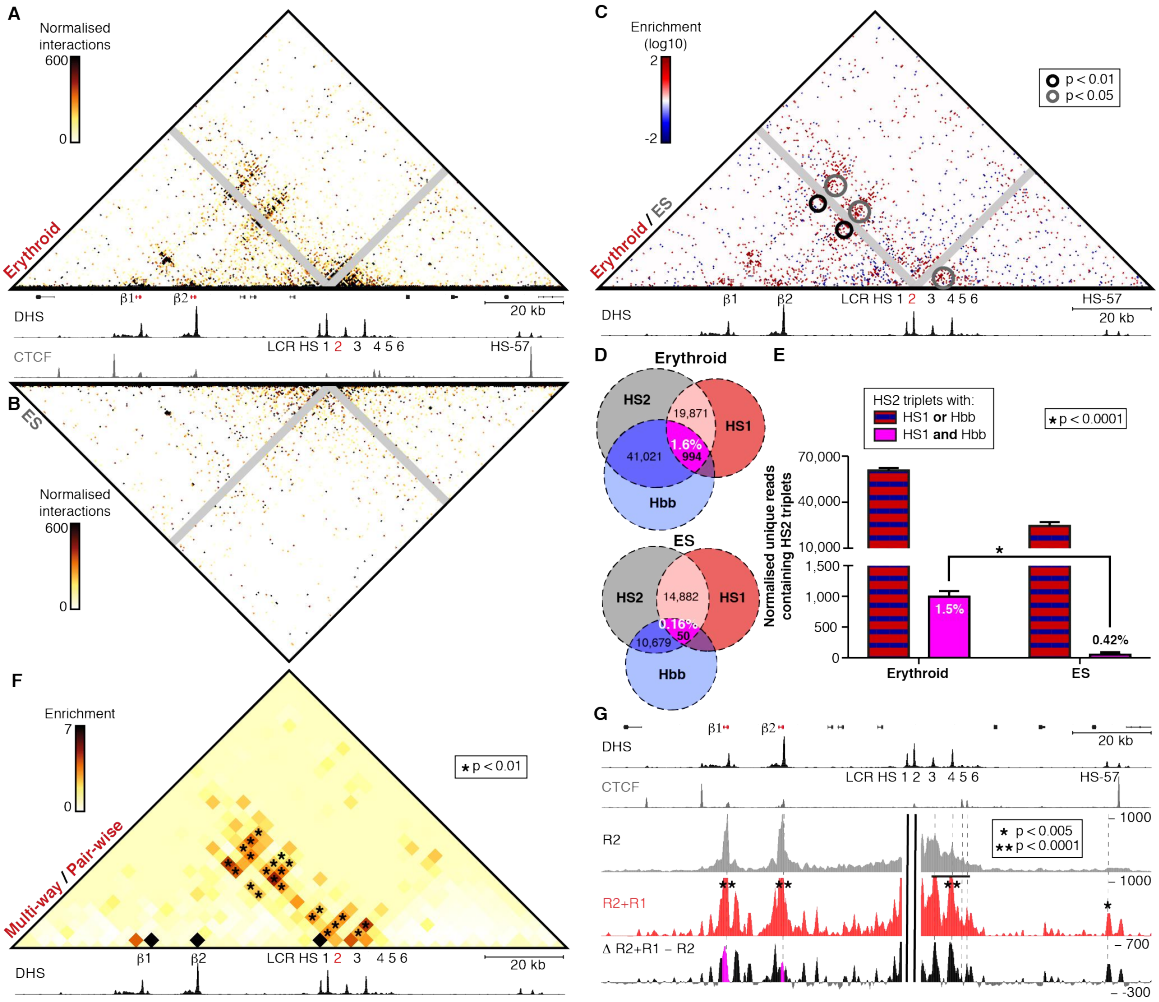


Figure 4.8: Analysis of multi-way interactions between enhancers and promoters in the β-globin locus. *Tri-C matrices represent the mean number of normalised, unique interaction counts in ten erythroid or seven ES replicates, displayed at 500 bp resolution, with proximity contacts around the HS2 viewpoint excluded (grey diagonal). The strong signal at the intersection of the two β-globin genes is a bioinformatic artefact resulting from homology between the duplicated genes. Matrices are annotated with gene positions (β-globin genes highlighted in red), erythroid DNaseI hypersensitivity (DHS) and/or CTCF occupancy. Coordinates: chr7:110,930,000-111,070,000.* ***(A)*** *Contact matrix showing multi-way interactions with HS2 in erythroid cells.* ***(B)*** *Contact matrix showing multi-way interactions with HS2 in ES cells.* ***(C)*** *Contact matrix showing differential multi-way interactions with HS2 that are enriched in erythroid (red) or ES (blue) cells. Statistically significant enriched interactions in erythroid cells are highlighted.* ***(D)*** *Venn diagram illustrating the proportion of HS2-HS1-Hbb triplets relative to all triplets containing pair-wise HS2-HS1 and HS2-Hbb interactions in erythroid and ES cells. Numbers represent normalised counts of unique reads containing the specified three-way or pair-wise interactions.* ***(E)*** *Comparison of the absolute numbers of HS2-HS1-Hbb triplets (shown on the y-axis) and the proportion of these relative to all triplets containing pair-wise HS2-HS1 and HS2-Hbb interactions (represented as percentages in the bar graphs) in erythroid and ES cells. Bar graphs represent average normalised*

unique read counts in ten erythroid and seven ES replicates, with the standard error indicated by the error bars. Both absolute and relative HS2-HS1-Hbb triplet counts are significantly enriched in erythroid cells (~20-fold, $p < 0.0001$ and ~10-fold, $p < 0.0001$, respectively). ***(F)*** *Contact matrix (4 kb resolution) showing enrichment of multi-way interactions with HS2 in erythroid cells over pair-wise contact frequencies derived from multiplexed Capture-C data in three biological replicates. Statistically significant enriched multi-way interactions are highlighted.* ***(G)*** *Comparison of a double-anchored interaction profile containing fragments interacting with both the HS2 and HS1 enhancers (red) to the conventional HS2 interaction profile (grey) in erythroid cells. Enrichment of multi-way HS2-HS1 interactions over pair-wise HS2 interactions are shown in the differential profile at the bottom (black), with the HS2-HS1-Hbb interactions highlighted (magenta). Profiles represent windowed, normalised, unique interaction counts. Statistically significant enriched multi-way interactions with cis-regulatory elements of β-globin are indicated.*

Together, these analyses demonstrate that the globin enhancers do not compete for the formation of exclusive interactions, but form tissue-specific higher-order structures in which they simultaneously interact together to switch genes on.

CTCF-binding elements form dynamic interactions supportive of a loop extrusion mechanism underlying chromatin domain formation

Our analyses in Chapter 2 and 3 have shown that CTCF-binding sites flanking the α-globin locus contribute to the formation of a domain that delimits the region of chromatin within which the observed complexes between enhancers and promoters can be formed. However, the processes underlying the formation of such chromatin domains and their contribution to specific enhancer-promoter interactions remain unclear. As described in Chapter 1, based on Hi-C data it has been suggested that CTCF-binding sites located at domain boundaries form stable, specific loops, which might form more complex multi-anchored and/or nested structures involving multiple CTCF-binding elements. Based on polymer models, an alternative model has been proposed, in which chromatin domains are more dynamic and formed by a process of loop extrusion. Our Capture-C data show that the CTCF-binding sites in the regions flanking the globin loci interact together (Figures 2.6–2.9). However, the pattern of interactions is very broad compared to the interactions between enhancers and promoters and it is unclear whether these interactions represent stable loops between multiple CTCF-binding sites or a more dynamic mechanism by which domain boundaries are formed.

To resolve these structures in individual cells, we performed Tri-C analysis from the viewpoint of the functionally validated HS-39 CTCF-binding site upstream of the α-globin genes (Figure 4.9). Consistent with the pair-wise interaction data (Figures 2.6 and 2.7), the multi-way interactions in erythroid cells are preferentially formed with the region located downstream of

the α-globin locus containing many CTCF-binding sites. A model in which CTCF boundaries are formed by stable, multi-anchored loops would result in specific enrichments of multi-way contacts between these CTCF-binding sites. However, the HS-39 Tri-C contact matrix shows a diffuse enrichment with the entire region on the other side of the α-globin domain in erythroid cells (Figure 4.9A). Comparison of these interactions to ES cells (Figure 4.9B) shows a trend of general enrichment of this diffuse interaction pattern in erythroid cells, though none of these interactions are significantly enriched (Figure 4.9C).

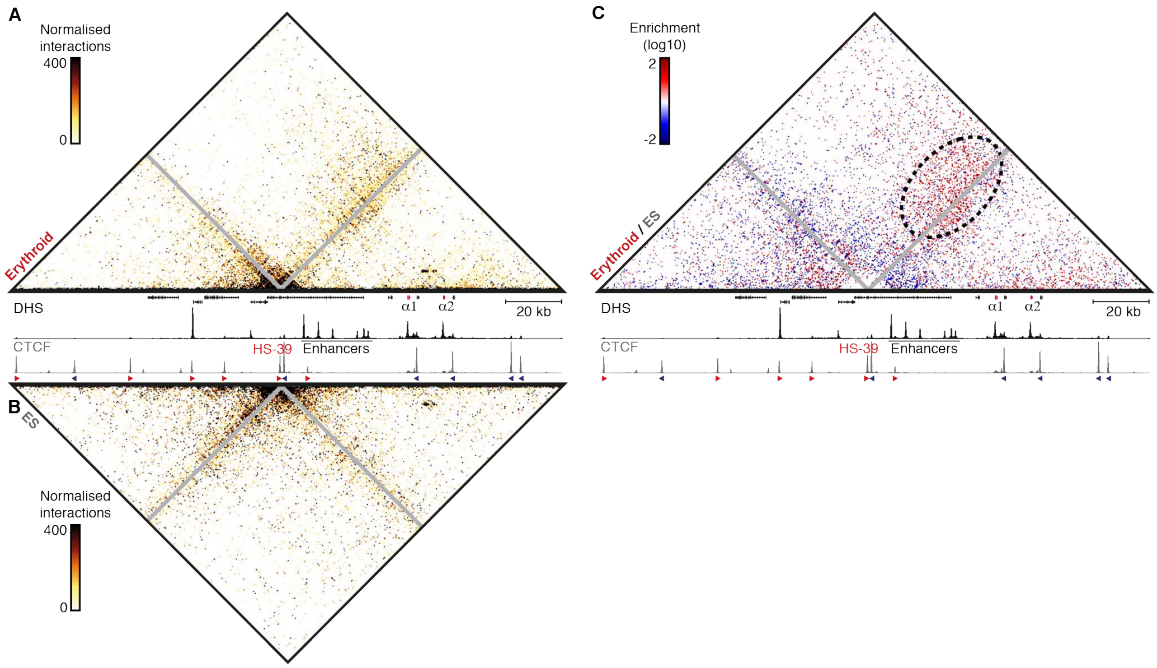


Figure 4.9: Analysis of multi-way interactions between CTCF-binding elements in the α-globin locus. *Tri-C matrices represent normalised, unique interaction counts, displayed at 500 bp resolution, with proximity contacts around the HS-39 viewpoint excluded (grey diagonal). Matrices are annotated with gene positions (α-globin genes highlighted in red), erythroid DNaseI hypersensitivity (DHS) and CTCF occupancy, with arrows indicating the orientation of CTCF-binding motifs. Coordinates: chr11:32,040,000-32,240,000.* ***(A)*** *Contact matrix showing multi-way interactions with HS-39 in erythroid cells.* ***(B)*** *Contact matrix showing multi-way interactions with HS-39 in ES cells.* ***(C)*** *Contact matrix showing differential multi-way interactions with HS-39 that are enriched in erythroid (red) or ES (blue) cells. Although the diffuse interactions with the region spanning the other side of the α-globin domain are enriched in erythroid cells (highlighted), there are no significant enrichments of specific interactions between multiple CTCF-binding sites ($p > 0.5$).*

Tri-C analysis from the viewpoint of the 3'HS1 CTCF-binding site downstream of the β-globin locus shows a similar pattern of diffuse multi-way interactions (Figure 4.10).

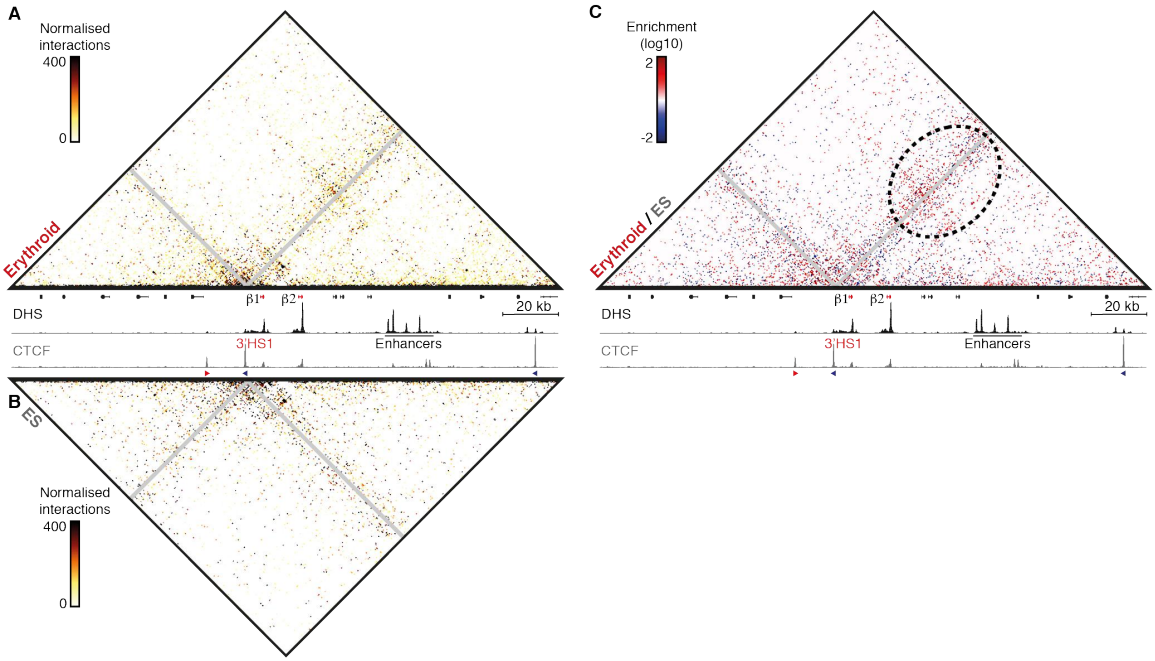


Figure 4.10: Analysis of multi-way interactions between CTCF-binding elements in the β-globin locus. *Tri-C matrices represent normalised, unique interaction counts, displayed at 500 bp resolution, with proximity contacts around the 3'HS1 viewpoint excluded (grey diagonal). Matrices are annotated with gene positions (β-globin genes highlighted in red), erythroid DNaseI hypersensitivity (DHS) and CTCF occupancy, with arrows indicating the orientation of CTCF-binding motifs. Coordinates: chr7:110,870,000-111,070,000.* ***(A)*** *Contact matrix showing multi-way interactions with 3'HS1 in erythroid cells.* ***(B)*** *Contact matrix showing multi-way interactions with 3'HS1 in ES cells.* ***(C)*** *Contact matrix showing differential multi-way interactions with HS-39 that are enriched in erythroid (red) or ES (blue) cells. Although the diffuse interactions with the region spanning the other side of the β-globin domain are enriched in erythroid cells (highlighted), there are no significant enrichments of specific interactions between multiple CTCF-binding sites ($p > 0.5$).*

Because the Tri-C patterns in the murine globin loci do not reveal enriched simultaneous interactions between CTCF-binding elements, our data are not consistent with a model in which chromatin domains are formed by multiple stable loops between CTCF-binding sites at domain boundaries. Rather, we observe erythroid-specific enrichments over a broad region of chromatin containing many CTCF-binding sites. The diffuse interaction patterns can be interpreted as dynamic interactions with the entire region flanking the opposite sides of the globin domains, with these heterogeneous interactions representing single-cell snapshots of the dynamic interactions throughout this chromatin region. These patterns could therefore be explained by continuous translocation of loop extruding factors across chromatin regions and transient stalling at CTCF boundary elements, and provide support for a loop extrusion mechanism underlying chromatin domain formation.

Discussion

As explained in Chapter 1, the proposed loop extrusion mechanism could explain many features of chromosome organisation (Fudenberg et al., 2016; Nasmyth, 2001; Sanborn et al., 2015). However, the process of loop extrusion and the resulting dynamic chromatin structures in individual cells have not been observed *in vivo*. Current evidence is derived from polymer model predictions (Fudenberg et al., 2016; Sanborn et al., 2015) and perturbations of specific components of the loop extrusion machinery (Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). Our analyses of multi-way interactions between CTCF-binding sites in the murine globin loci therefore provide an important new line of support for a dynamic mechanism such as loop extrusion underlying the formation of chromatin domains. Importantly, the diffuse enrichment patterns we observe are indicative of transient CTCF-binding site interactions, which is in agreement with the kinetics of CTCF and Cohesin binding to the genome (Hansen et al., 2017).

The observed diffuse interactions between CTCF-binding sites at opposite ends of the globin loci are formed specifically in erythroid cells and explain the formation of tissue-specific domains (Figures 4.9 and 4.10). It remains unclear what underlies the preferential formation of distinct chromatin domains in specific cell types. As CTCF occupancy in the globin loci is similar in both erythroid and ES cells (Figures 1.7 and 1.8), the erythroid-specific interactions between the CTCF-binding elements are not related to abundance of CTCF at these loci. It is possible that the interactions between active *cis*-regulatory elements could stabilise the interactions between regions bound by CTCF. It is also possible that the different structures reflect differences in the processivity of loop extruding factors such as Cohesin, which could result from tissue-specific recruitment locations and/or extrusion rates. It has been shown that the transcriptional co-activator Mediator forms complexes with Cohesin and the Cohesin-loading factor Nipbl at active enhancers and promoters, (Kagey et al., 2010). As both Cohesin and Mediator bind the globin enhancers in erythroid cells (Figures 1.7–1.9), the erythroid-specific formation of the globin domains could be explained by a loop extrusion model in which Cohesin is recruited to the genome at active enhancers and/or promoters. Recent findings have suggested that Cohesin translocation is stimulated by Nipbl, suggesting that Nipbl abundance determines extrusion rates (Rhodes et al., 2017). Differences in Nipbl distribution could therefore also contribute to tissue-specific interaction patterns, though this remains to be further explored.

Within the dynamic chromatin domains containing the globin gene clusters, our Tri-C analyses reveal higher-order structures in which multiple enhancers and promoters interact

simultaneously in individual cells. These structures appear similar to the proposed 'active chromatin hub' (Figure 1.10) (Tolhuis et al., 2002), with the fundamental difference that they are not formed by specific, stable chromatin loops, but represent preferred interactions in dynamic chromatin domains. The formation of these higher-order structures indicates structural cooperation between multiple *cis*-regulatory elements and shows that enhancers and promoters do not compete for the formation of exclusive interactions. This is highlighted by the fact that the promoter of the housekeeping gene *Nprl3*, which is located between the α-globin genes and its enhancers and is six-fold upregulated in erythroid cells (Hanssen et al., 2017; Kowalczyk et al., 2012), is included in the complexes formed with the α-globin enhancers and promoters (Figure 4.6).

Although a model in which a loop extruding factor such as Cohesin travels along chromosomes in a one-dimensional manner and is transiently halted at CTCF-binding sites can explain the formation of compartmentalised chromatin domains, it does not provide enough complexity and specificity to explain the higher-order three-dimensional structures that are formed by multiple enhancers and promoters. These structures therefore indicate other, independent mechanisms contributing to chromosome architecture. This is consistent with polymer models (Barbieri et al., 2017) and recent studies in which Cohesin binding was perturbed, but interactions between enhancers and promoters still occurred, though more promiscuous (Rao et al., 2017; Schwarzer et al., 2017). Our data therefore suggest that the formation of functional chromatin structures depends on a loop extrusion mechanism that brings *cis*-regulatory elements into close proximity and subsequent formation of more complex structures reflecting further (transient) stabilisation of interactions between these elements.

As described in Chapter 1, genetic variation and mutations that disrupt these complex structures can result in misregulation of gene expression and have been implicated in disease. This highlights that 3C techniques not only contribute to our fundamental understanding of the relationship between genome structure and function, but are also a valuable tool to interpret how genetic variations can perturb regulatory landscapes and cause disease. Since these regulatory structures are highly tissue-specific, it is crucial that 3C experiments are performed in purified, primary cells that are relevant for the disease of interest. However, current 3C techniques require millions of cells for high-resolution analysis of chromatin structures. Many rare cell and tissue types therefore remain inaccessible to 3C analyses. To enable investigation of chromatin architecture in these cells, we developed a new low-input 3C approach, which is described in the next chapter.

Data availability

The data presented in this chapter are published in the following pre-print:

Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P., Telenius, J.M., Schwessinger, R., Liu, Y., Brown, J.M., Downes, D.J., Chiariello, A.M., Bianco, S., Nicodemi, M., Buckle, V.J., Dekker, J., Higgs, D.R., and Hughes, J.R. (2018). Single-cell chromatin interactions reveal regulatory hubs in dynamic compartmentalized domains. BioRxiv*, doi: 10.1101/307405.
**Under revision for Nature Genetics.*

Acknowledgements

C-Trap experiments were performed in collaboration with James O.J. Davies and Helena Ayyub.

Chapter 5 – Capturing chromosome conformation in small numbers of cells

Introduction

As demonstrated in the previous chapters, 3C techniques play a key role in investigating how structural interactions between *cis*-regulatory elements relate to gene expression. Because these interactions are highly tissue-specific, it is crucial that 3C experiments are performed in well defined, purified cell populations. However, current 3C techniques require a large number of cells to generate reproducible interaction profiles at high resolution, which hampers the investigation of chromatin architecture in rare cell populations.

To generate high-resolution contact matrices using Hi-C, two to five million cells are required (Lieberman-Aiden et al., 2009; Rao et al., 2014). The input for 4C experiments is usually ten million cells (van de Werken et al., 2012), though it has recently been shown that UMI-4C can generate robust interaction profiles (containing 5,000–10,000 unique interactions) using 1 µg of 3C library (Schwartzman et al., 2016), which corresponds to ~340,000 cells (assuming 50% loss during 3C library preparation; as explained below). Capture-C is currently the most efficient 3C technique and can generate high-quality interaction profiles (containing ~19,000 unique interactions) from 100,000 cells (Davies et al., 2015).

However, many primary tissues and rare cell populations are not available in these numbers. Therefore, we developed a low-input Capture-C approach, to enable high-resolution analysis of chromatin architecture in small numbers of cells.

Results

Approach for the development of a low-input 3C technique

The Capture-C procedure contains three stages:

(I) 3C library preparation (Figure 5.1; steps 1–4);
(II) Capture-C library preparation (Figure 5.1; steps 5–6);
(III) Enrichment for fragments of interest targeted by capture oligonucleotides (Figure 5.1; steps 7–10).

To optimise the Capture-C procedure for small cell numbers, all three stages need to be performed with maximum efficiency.

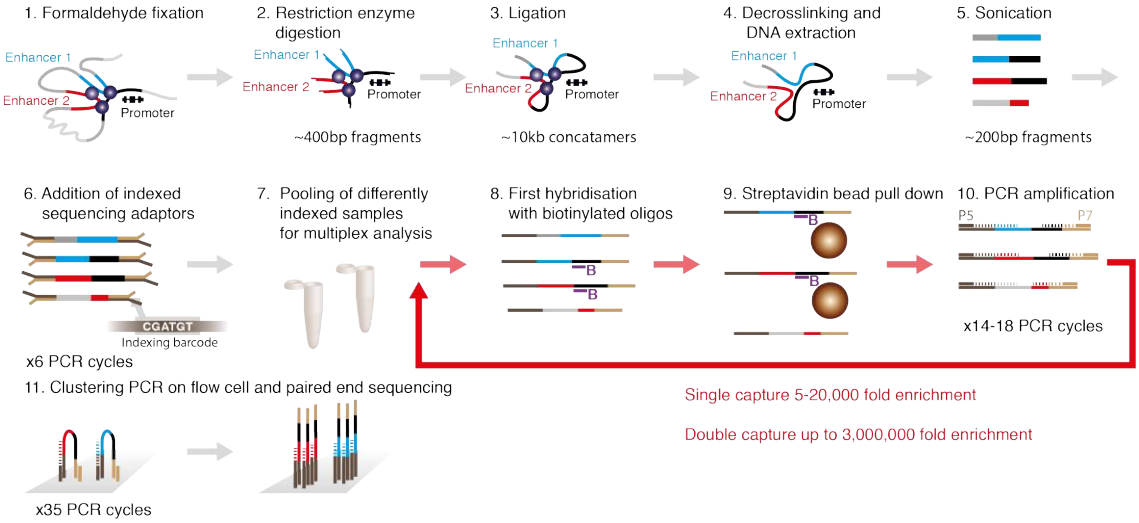


Figure 5.1: Overview of the Capture-C procedure. *Figure adapted from: Davies et al., 2015.*

At the end of Capture-C library preparation, the material is PCR amplified. Subsequent enrichment for viewpoint fragments of interest has been optimised previously and is extremely efficient, with enrichment up to a million-fold (Davies et al., 2015). We therefore focused on improving the efficiency of the first two stages of the Capture-C protocol, as these determine the minimum number of cells required. We optimised the preparation of the 3C library and the Capture-C library independently, to enable careful comparison of different approaches and accurate quantification of gains and losses at every step. The optimised procedures were subsequently combined into a single protocol, which we performed directly on low-input samples to demonstrate the effectiveness of our new approach. We compared the low-input Capture-C profiles to previously published interaction profiles generated using the Next-Generation (NG) Capture-C protocol (Davies et al., 2015), which is the most efficient current procedure for the generation of high-quality 3C data (Davies et al., 2017).

Optimisation of 3C library preparation in small numbers of cells

To optimise 3C library preparation for small cell numbers, we adapted the protocol to a single-tube procedure, in order to minimise losses during digestion and ligation. We also improved the efficiency of the DNA extraction procedure by implementing Phase Lock Gel technology for DNA recovery. The technical details of our optimised procedure are described in Chapter 7.

Using this optimised protocol, we generated 3C libraries from 10,000, 20,000, 50,000 and 100,000 cells. Careful quantification of these libraries by qPCR shows that we were able to recover ~50% of the DNA mass, irrespective of the cell number used (Figure 5.2).

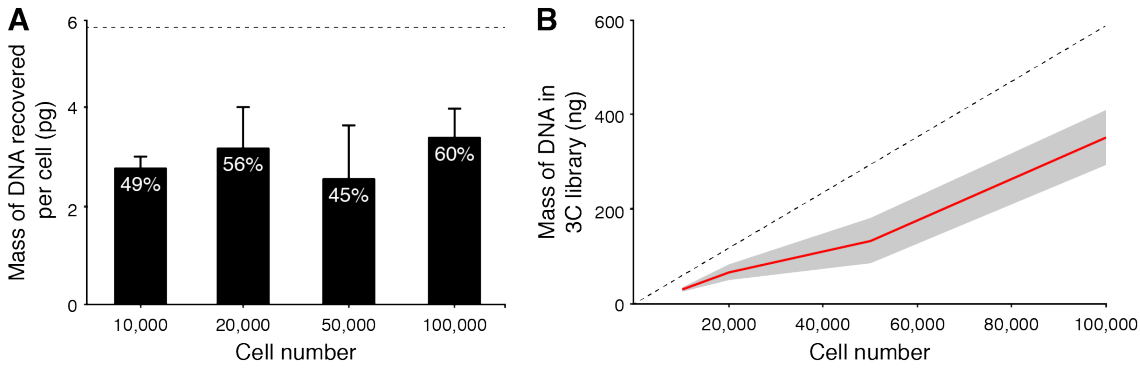


Figure 5.2: 3C library preparation efficiency. (A) *3C library preparation efficiency expressed as mass of DNA recovered per cell. The bars represent the average and standard error of five technical replicates of 10,000–100,000 primary erythroid cells. The dashed line indicates the theoretical maximum of 5.86 pg, which is the mass of the diploid mouse genome. The percentages indicate the mass recovered relative to this maximum.* ***(B)*** *Relationship between the number of cells used as input and the mass of DNA in the generated 3C library. The red line shows the average 3C library mass generated from five technical replicates of 10,000–100,000 primary erythroid cells, with the standard error indicated by the grey line shadow, and the theoretical maximum indicated by the dashed line.*

Development of low-input Capture-C library preparation approaches

During the subsequent Capture-C library preparation, the 3C library is processed into a sequencing library by fragmenting the ligated DNA concatemers and adding indexed sequencing adapters. To increase the efficiency of these experimental steps, we developed two new approaches – Low-Input (LI) Capture-C and Tag-Capture-C – which use sequencing library preparation protocols that are more suitable for small cell numbers. The LI-Capture-C procedure uses a conventional sequencing library preparation strategy, involving sonication and enzymatic end-repair, dA-tailing and adaptor ligation, based on the NEBNext Ultra DNA Library Prep strategy. Tag-Capture-C uses a one-step transposase-catalysed adaptor insertion ('tagmentation'), which has been presented as a more efficient method of sequencing library preparation (Adey et al., 2010).

We optimised both protocols using triplicate aliquots of 50, 100 and 200 ng 3C library, corresponding to ~17,000, ~34,000 and ~68,000 cells, respectively (Figure 5.2). We tested the protocols by analysing the well-characterised α-globin (viewpoints of the *Hba-a1* and *Hba-a2*

promoters), β-globin (viewpoints of the *Hbb-b1* and *Hbb-b2* promoters) and Mitoferrin-1 (viewpoint of the *Slc35a37* promoter*)* loci in primary erythroid cells.

Analysis of the α-globin locus shows that LI-Capture-C interaction profiles are highly reproducible at maximum resolution (individual restriction fragments) when 100-200 ng 3C library is used (Figures 5.3–5.5; Table 5.1). Although interaction profiles generated from 50 ng samples are noisier and less reproducible, the profiles are still highly informative and interpretable, especially when the data are represented in windowed interaction profiles typically used in low-resolution 4C analyses (Figure 5.6).

Tag-Capture-C interaction profiles generated from 100–200 ng 3C libraries are less reproducible than LI-Capture-C profiles. However, when only 50 ng of 3C library was used, Tag-Capture-C outperformed LI-Capture-C, and produced more robust interaction profiles, with less noise and higher reproducibility (Figures 5.3–5.6; Table 5.1).

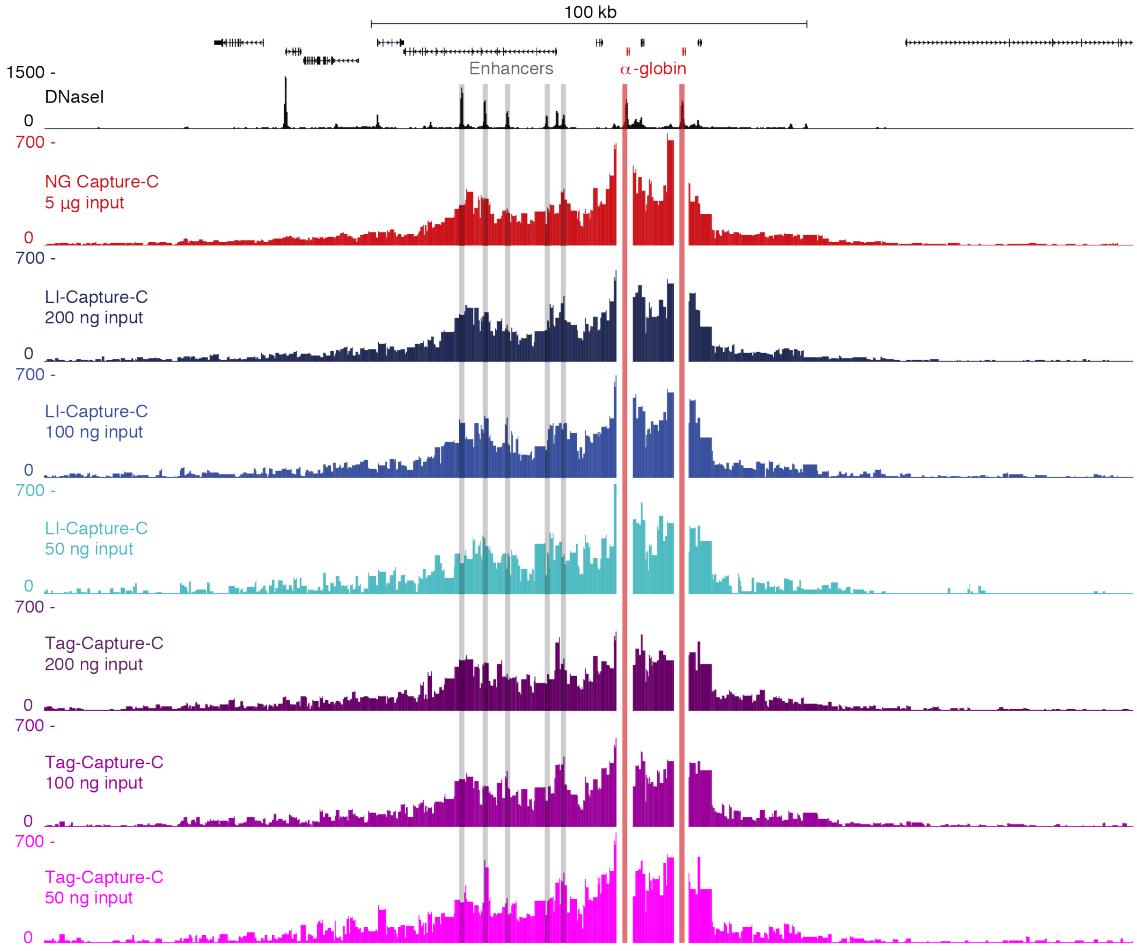


Figure 5.3:Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the α-globin locus. *Interaction profiles from the viewpoint of the α-globin promoters (highlighted in red) generated from decreasing amounts of 3C libraries (prepared from primary erythroid cells) with NG-Capture-C (red), LI-Capture-C (blue) and Tag-Capture-C (purple). Profiles show the mean number of unique interactions per restriction fragment from three replicates, normalised for a total of 100,000 interactions genome-wide. The α-globin enhancers are highlighted in grey. Coordinates: chr11:32,050,000-32,300,000.*

	Hba-a1	***Hba-a2***	***Hbb-b1***	***Hbb-b2***	***Slc25a37***
NG-Capture-C 5 µg input	169,744	166,114	122,794	157,370	119,985
LI-Capture-C 200 ng input	14,012	14,131	12,548	11,730	9,035
LI-Capture-C 100 ng input	6,824	6,749	5,713	5,350	4,251
LI-Capture-C 50 ng input	2,929	2,903	2,549	2,280	1,978
Tag-Capture-C 200 ng input	7,414	7,436	8,059	8,354	6,968
Tag-Capture-C 100 ng input	4,769	4,720	5,011	5,443	4,417
Tag-Capture-C 50 ng input	2,645	2,633	2,908	3,170	2,606

Table 5.1: Unique interactions detected per viewpoint using NG-Capture-C, LI-Capture-C and Tag-Capture-C. *Table shows the average of three biological replicates of erythroid cells.*

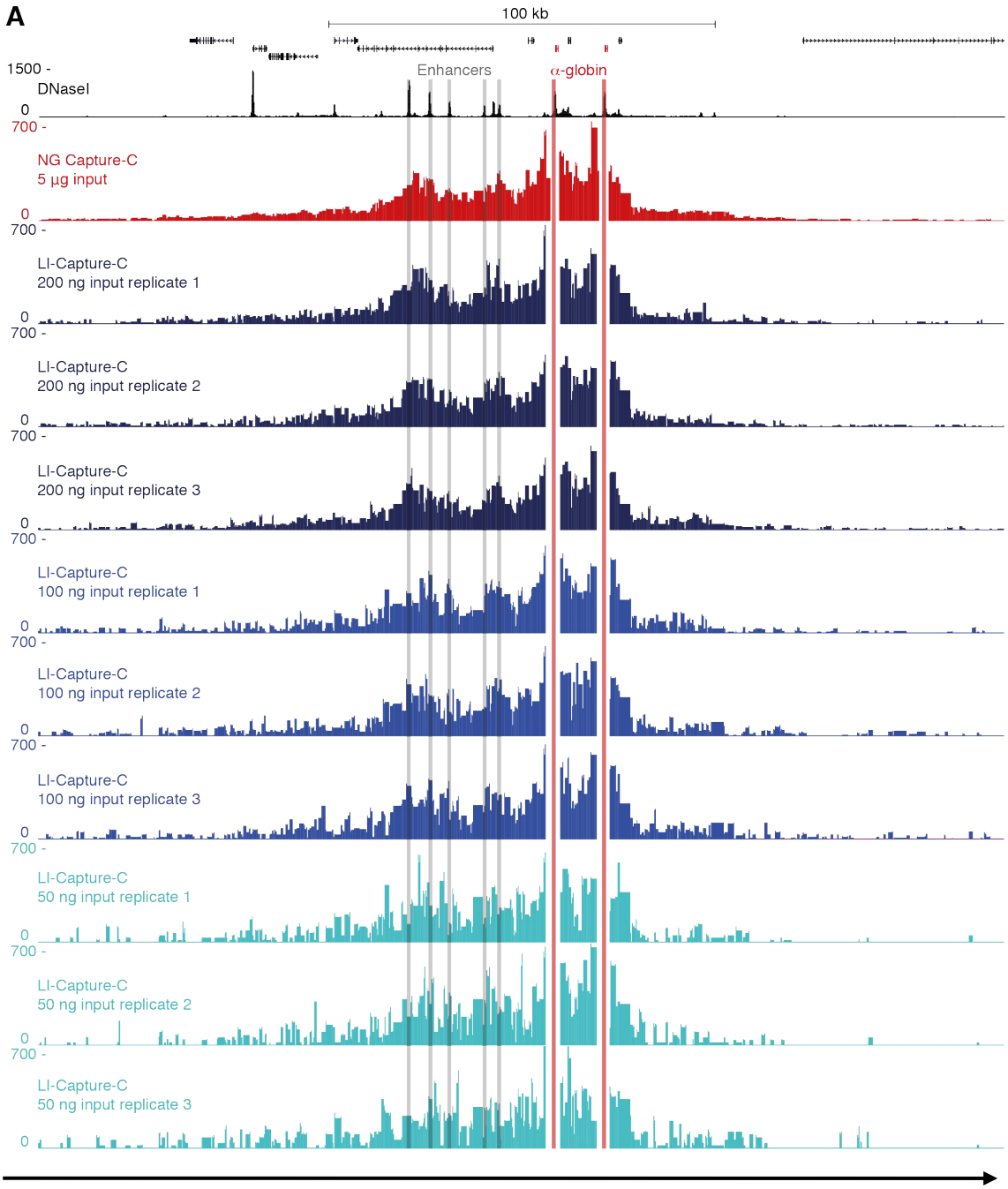


Figure 5.4: Reproducibility of LI-Capture-C and Tag-Capture-C interaction profiles in biological replicates. *Caption on next page.*

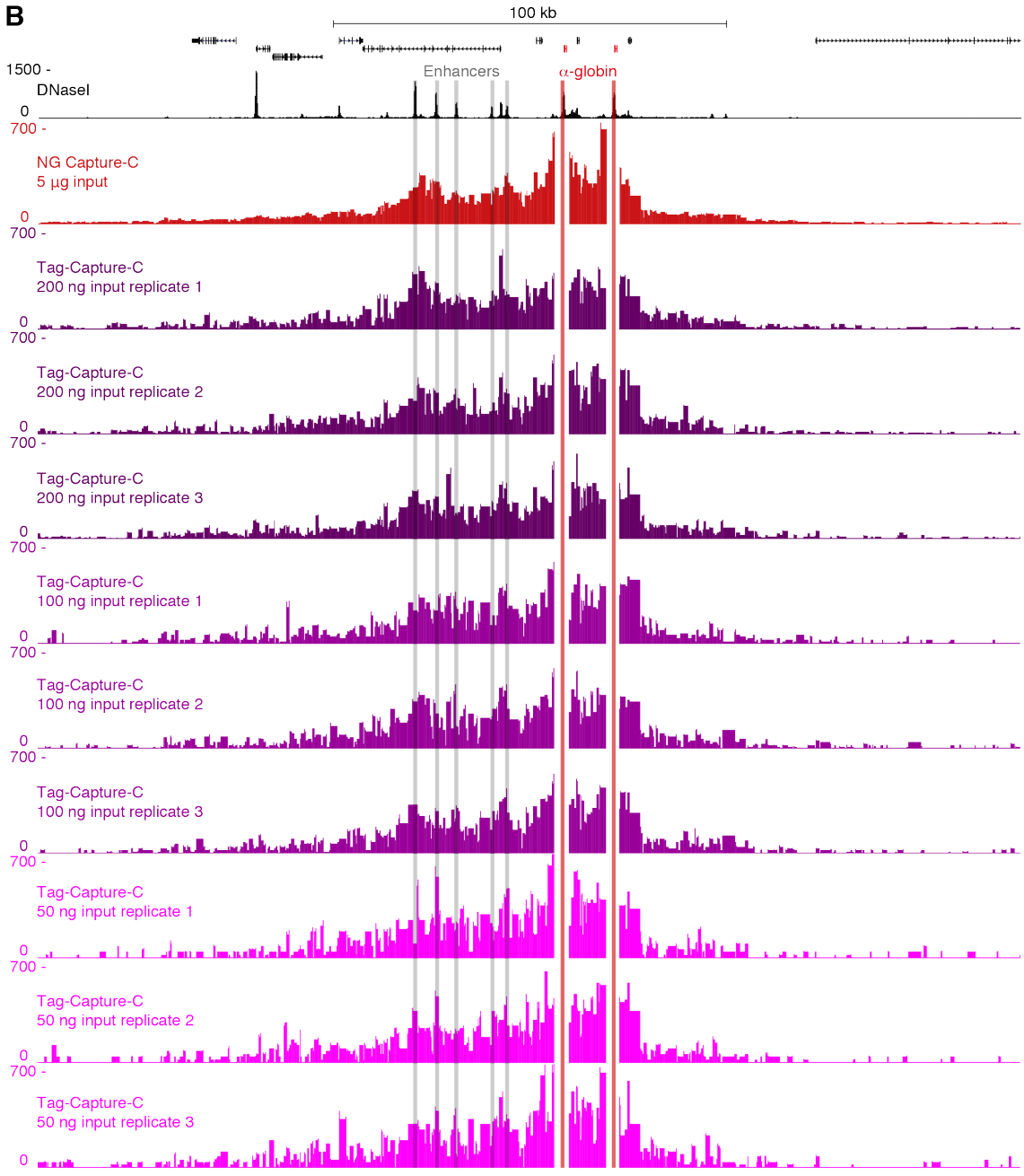


Figure 5.4. Reproducibility of LI-Capture-C and Tag-Capture-C interaction profiles in biological replicates. *Interaction profiles from the viewpoint of the α-globin promoters (highlighted in red) from individual replicates, generated from decreasing amounts of 3C libraries (prepared from primary erythroid cells) with* ***(A)*** *LI-Capture-C and* ***(B)*** *Tag-Capture-C. Profiles show the number of unique interactions per restriction fragment, normalised for a total of 100,000 interactions genome-wide. The α-globin enhancers are highlighted in grey. Coordinates: chr11:32,050,000-32,300,000.*

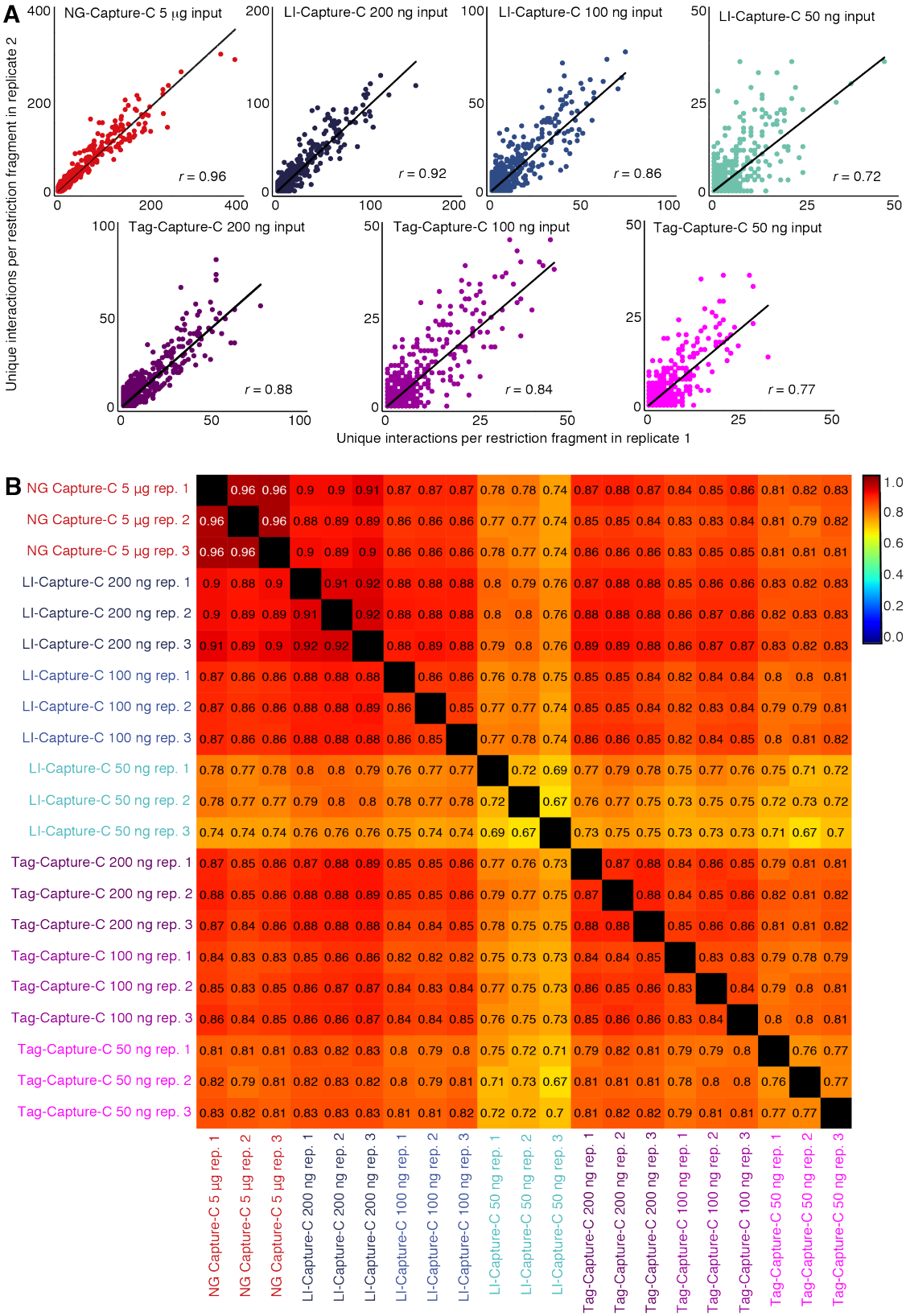


Figure 5.5: Correlation between interaction profiles of the α-globin locus generated by different Capture-C protocols. ***(A)*** *Correlation of chromosomal interactions detected by NG-Capture-C, LI-Capture-C and Tag-Capture-C in individual replicates. The r-values represent the Pearson correlation coefficients.* ***(B)*** *Heatmap showing the Pearson correlation between the α-globin interaction profiles generated by NG-Capture-C, LI-Capture-C and Tag-Capture-C.*

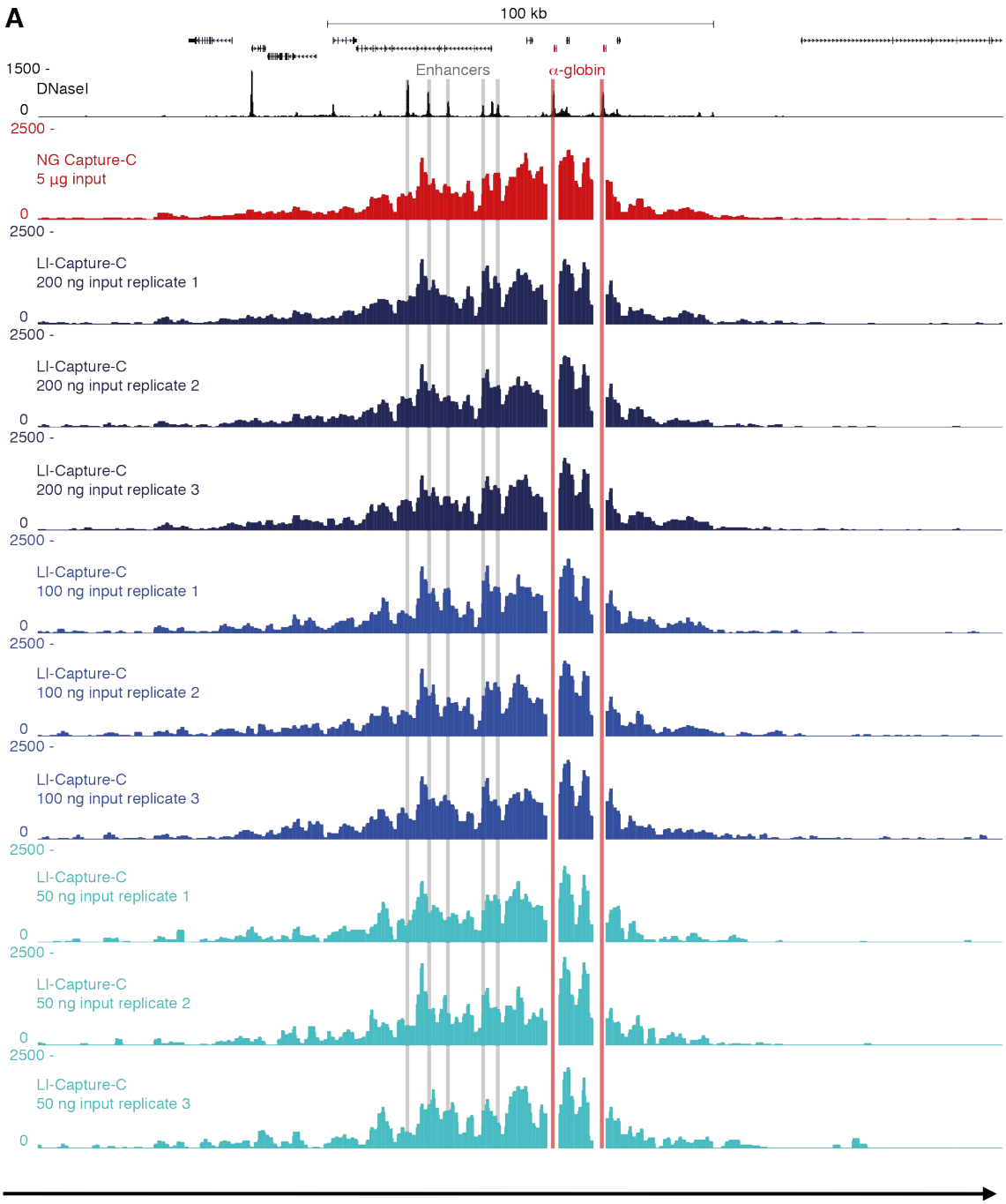


Figure 5.6: Windowed interaction profiles of individual LI-Capture-C and Tag-Capture-C replicates. *Caption on next page.*

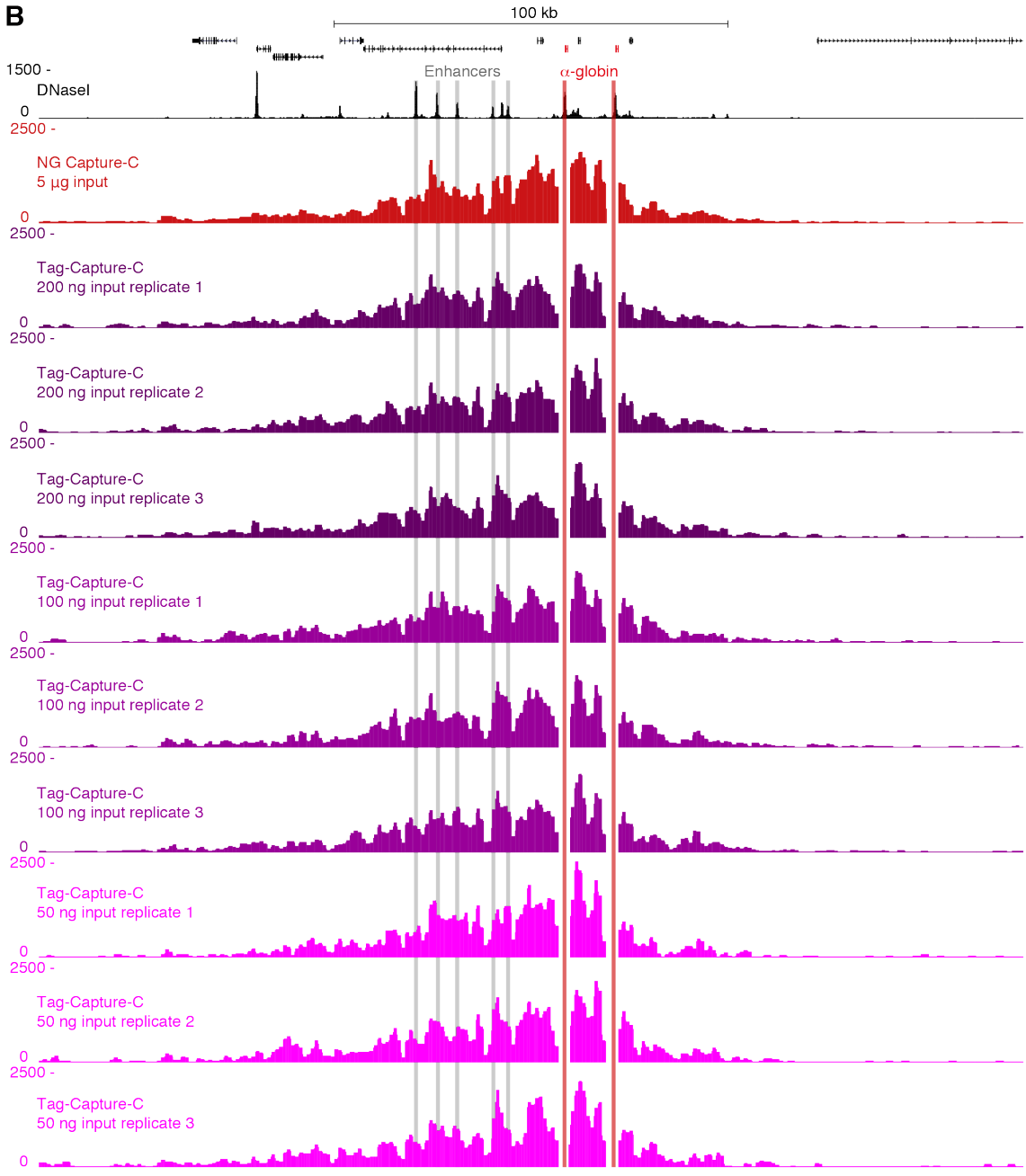


Figure 5.6. Windowed interaction profiles of individual LI-Capture-C and Tag-Capture-C replicates. *Interaction profiles from the viewpoint of the α-globin promoters (highlighted in red) from individual replicates, generated from decreasing amounts of 3C libraries (prepared from primary erythroid cells) with* ***(A)*** *LI-Capture-C and* ***(B)*** *Tag-Capture-C. Profiles show normalised unique interactions with the α-globin promoters (highlighted in red) in a moving window (window size = 2 kb; window increment = 1 kb). The α-globin enhancers are highlighted in grey. Coordinates: chr11:32,050,000-32,300,000.*

The lower performance of Tag-Capture-C compared to LI-Capture-C in the samples with 100–200 ng input is likely related to the different fragmentation mechanisms in the protocols. LI-Capture-C uses sonication, which results in random fragmentation of the 3C library. However, the tagmentation events in the Tag-Capture-C procedure are not completely randomly distributed over the genome (Figure 5.7A). The reported preference for C and G bases (Green et al., 2012; Madrigal, 2015) makes it more likely that independent tagmentation events in different cells take place at exactly the same genomic locations in and around targeted restriction sites. This results in fragments with identical genomic coordinates that are indistinguishable from PCR duplicates and are removed from analysis. The probability of independent tagmentation events occurring at identical genomic locations decreases when fewer cells are sampled, which could explain why the efficiency of Tag-Capture-C improves when the amount of input DNA is reduced. The differences in performance between LI-Capture-C and Tag-Capture-C are unlikely to be related to the differences in fragment size distribution of the libraries, as the number of detected restriction fragments per read is very similar in LI-Capture-C and Tag-Capture-C libraries (Figure 5.7B).

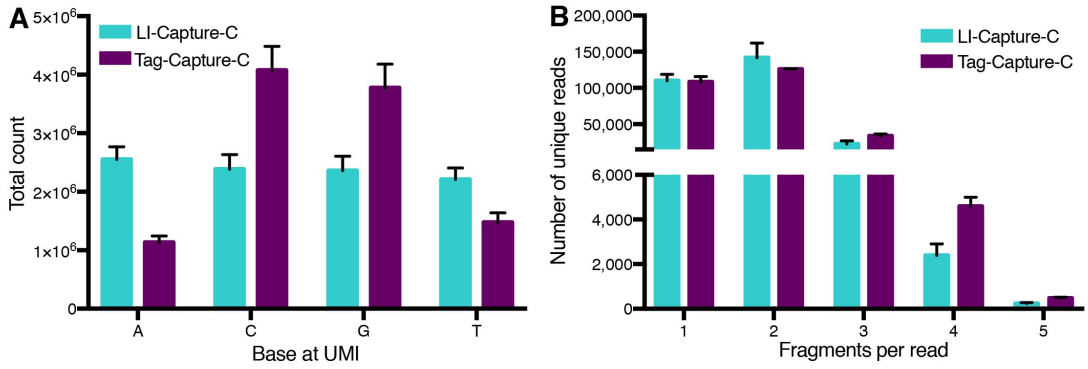


Figure 5.7: Comparison of characteristics of LI-Capture-C and Tag-Capture-C libraries. ***(A)*** *Distribution of nucleobases at the sonication or tagmentation sites that are used as UMIs for PCR duplicate removal in LI-Capture-C and Tag-Capture-C, respectively. Bars represent average counts and standard errors of three replicate experiments using 200 ng 3C library.* ***(B)*** *Comparison of the number of restriction fragments detected per unique read in LI-Capture-C and Tag-Capture-C experiments. Bars represent average read numbers and standard errors of three replicate experiments using 200 ng 3C library.*

Integration of optimised procedures into a low-input Capture-C protocol

To demonstrate the effectiveness of the optimised protocols, we performed the entire LI Capture-C procedure directly on 10,000, 20,000, 50,000, and 100,000 cells (Figure 5.8). These experiments confirm that high-quality interactions profiles can be generated from small numbers of cells.

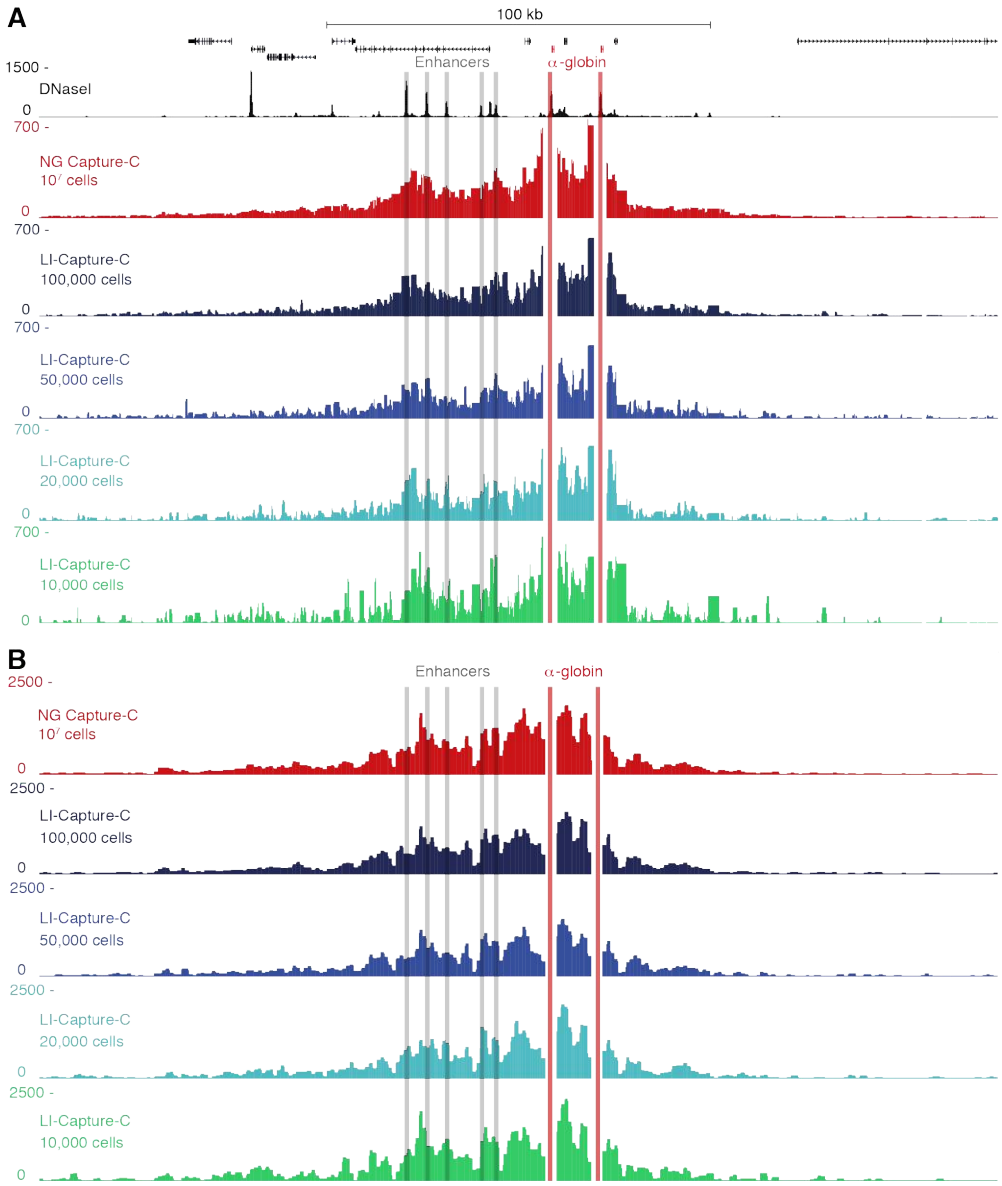


Figure 5.8: Interaction profiles generated with LI Capture-C using 10,000-100,000 cells. *LI Capture-C interaction profiles from the viewpoint of the α-globin promoters (highlighted in red) generated from decreasing numbers of primary erythroid cells, displayed at* ***(A)*** *maximum resolution or* ***(B)*** *using a moving window (window size = 2 kb; window increment = 1 kb). Profiles represent the mean number of unique interactions per restriction fragment from three replicates, normalised for a total of 100,000 interactions genome-wide. The α-globin enhancers are highlighted in grey. Coordinates: chr11:32,050,000-32,300,000.*

Importantly, lowering input does not influence the throughput of Capture-C, as the quality of interaction profiles is related to the complexity of the Capture-C libraries and does not depend on the number of genomic regions sampled (Davies et al., 2015). The low-input Capture-C approaches therefore allow for multiplexing thousands of viewpoints and analysing dozens of samples in a single experiment. In the low-input Capture-C experiments presented in this chapter, the α-globin, β-globin and Mitoferrin-1 loci were analysed simultaneously, and the quality of the corresponding interaction profiles was very similar (Figures 9 and 10; Table 5.1).

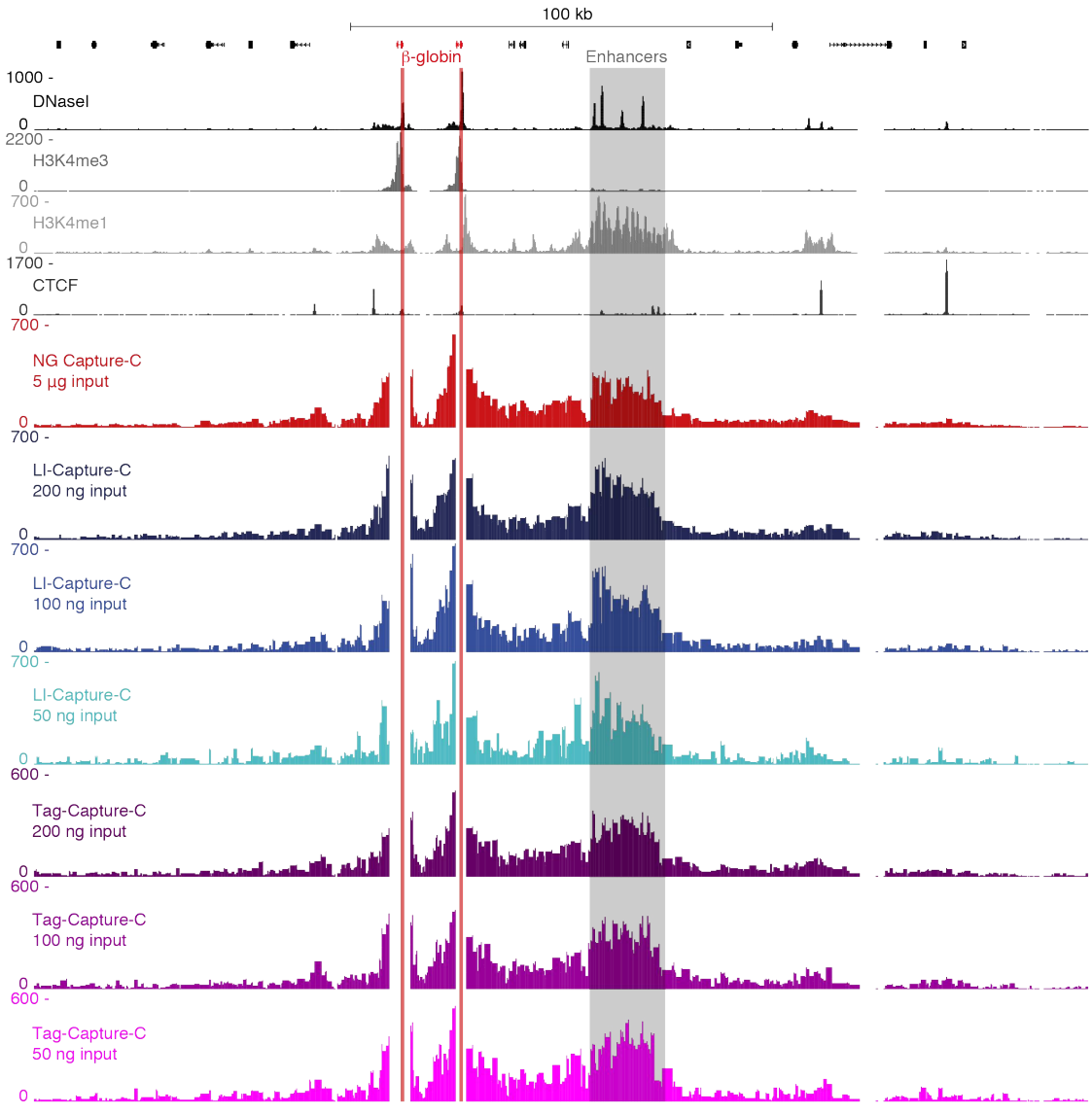


Figure 5.9:Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the β-globin locus. *Interaction profiles from the viewpoint of the β-globin promoters (highlighted in red) generated from decreasing amounts of 3C libraries (prepared from primary erythroid cells) with NG-Capture-C (red), LI-Capture-C (blue) and Tag-Capture-C (purple). Profiles show the mean number of unique interactions per restriction fragment from three replicates, normalised for a total of 100,000 interactions genome-wide. The β-globin enhancers are highlighted in grey. Coordinates: chr7:110,875,000-111,125,000.*

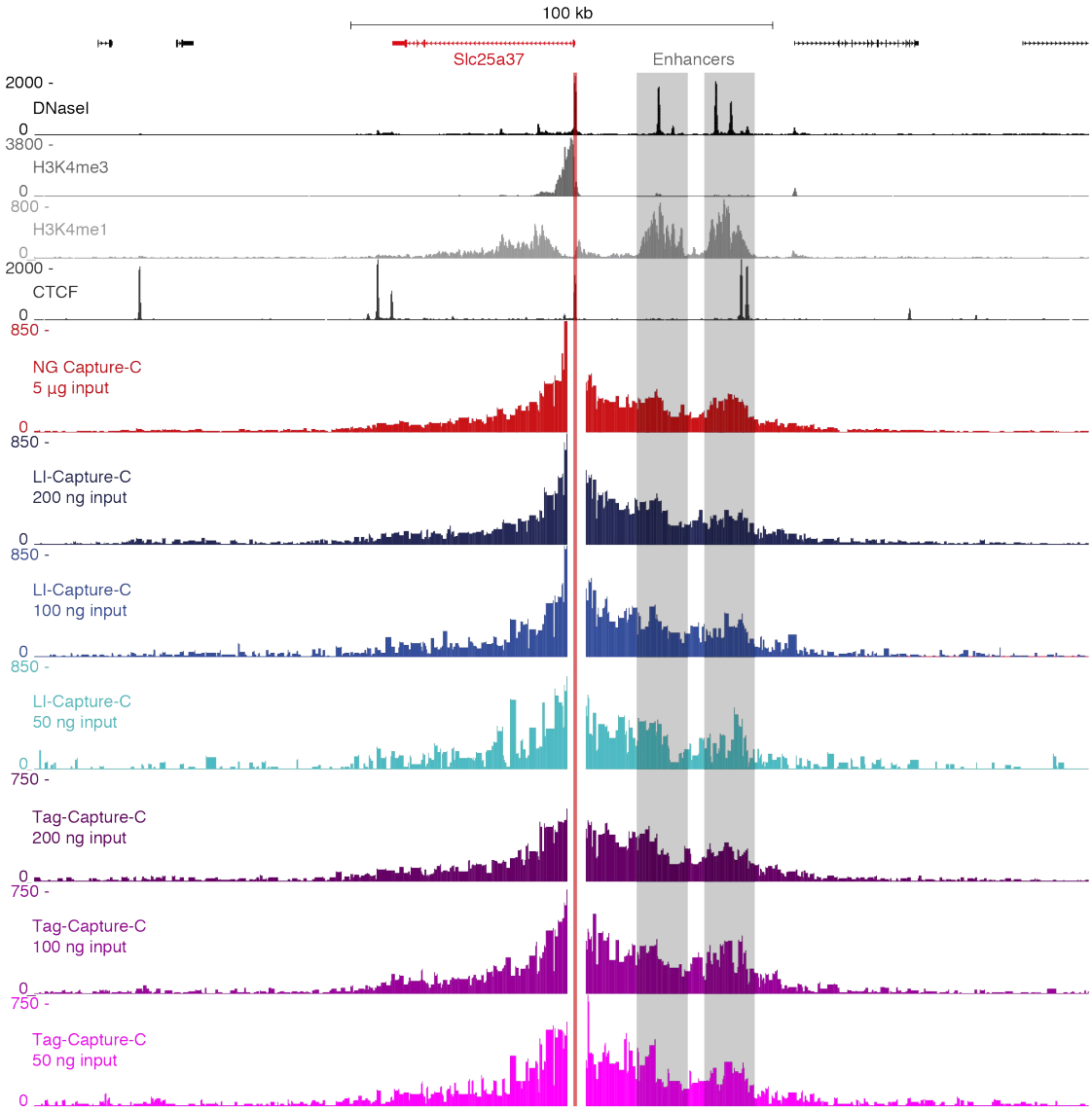


Figure 5.10: Comparison of NG-Capture-C, LI-Capture-C and Tag-Capture-C interaction profiles of the Mitoferrin-1 locus. *Interaction profiles from the viewpoint of the Mitoferrin-1 promoter (highlighted in red) generated from decreasing amounts of 3C libraries (prepared from primary erythroid cells) with NG-Capture-C (red), LI-Capture-C (blue) and Tag-Capture-C (purple). Profiles show the mean number of unique interactions per restriction fragment from three replicates, normalised for a total of 100,000 interactions genome-wide. The Mitoferrin-1 enhancers are highlighted in grey. Coordinates: chr14:69,775,001-70,025,000*

Bioinformatic analysis of low-input Capture-C data

The approach for data analysis is based on the NG-Capture-C analysis pipeline (Figure 5.11) (Davies et al., 2015).

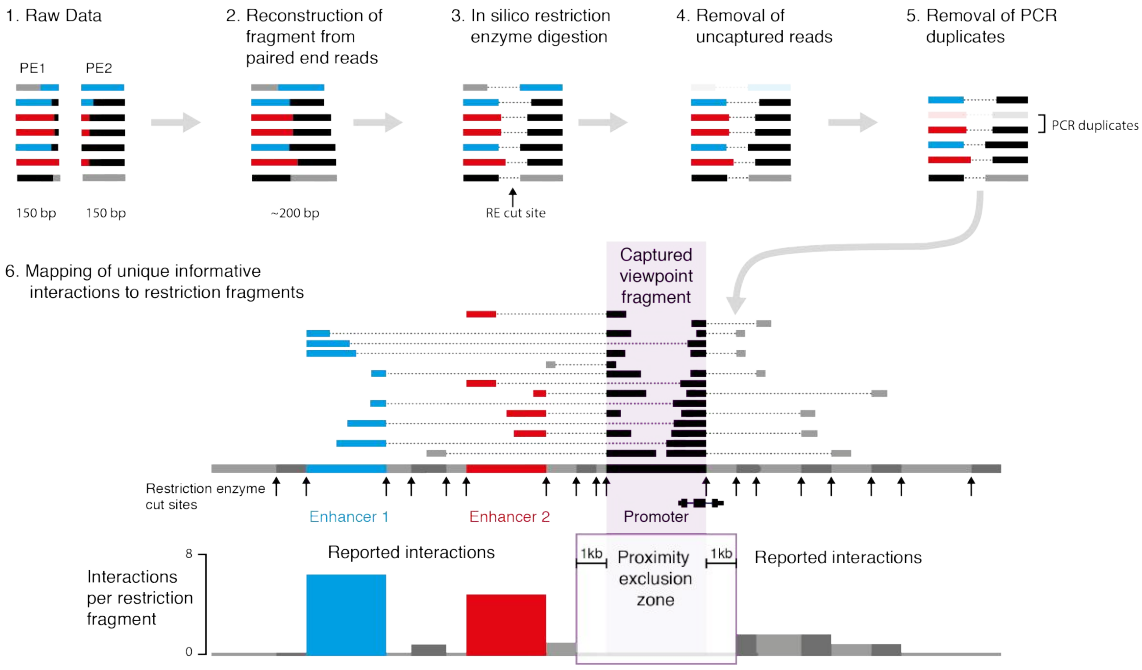


Figure 5.11: Overview of Capture-C data analysis. *Figure adapted from: Davies et al., 2015.*

Removal of PCR duplicates in this analysis strategy is based on the unique ends of the reads resulting from the (near) random sonication or tagmentation step, which are used as two independent UMIs. Reads with the exact same start and end coordinates of each individual restriction fragment are excluded from analysis. However, occasional sequencing and mapping errors can cause 'wobbly' ends, which make PCR duplicates appear as unique reads. In low-input libraries with limited complexity, this can cause spurious peaks in interaction profiles. We therefore created a more stringent duplicate filter, in which all reads with one or two differences in their UMIs are also excluded from analysis. While this approach is very conservative and will also remove genuine unique fragments, therefore giving an underrepresentation of the number of unique interactions detected, it ensures detection of robust interactions only.

Discussion

The data presented in this chapter demonstrate that LI-Capture-C and Tag-Capture-C are able to generate high-quality interaction profiles from ~20,000 cells at maximum resolution, and from ~10,000 cells using windowing based analysis. These new 3C approaches therefore enable analysis of chromatin structures in rare tissues and cell types which are currently intractable.

This is an important technical development, as it allows investigation of gene regulatory landscapes in primary cells that are biologically and clinically relevant. This is particularly relevant with respect to interpreting the wealth of data generated by genome-wide association studies (GWAS) that have been produced over the last decades. As the majority of the detected variants are located in the non-coding part of the genome (Maurano et al., 2012), these likely disrupt gene regulatory mechanisms. However, without knowing the structural conformation of these GWAS loci, it is difficult to predict which genes are affected (Sanyal et al., 2012), which has hampered insights into new disease mechanisms and drug targets. The low-input Capture-C approach presented in this chapter allows for the investigation of thousands of GWAS loci in one multiplexed experiment using only 10,000-20,000 (primary) cells relevant for the disease of interest. This enables linking variants in these loci to the genes they regulate and interpreting the biological processes that are affected by these variants.

It would be interesting to get more insight into how such variants disrupt three-dimensional gene regulatory structures, for which the development of a low-input Tri-C or similar approach would be a useful tool. The optimisation strategy presented in this chapter provides a framework for further development and optimisation of such techniques to make them more widely applicable.

Data availability

The data and procedures presented in this chapter are published in the following article and protocol:

Oudelaar, A.M., Davies, J.O.J., Downes, D.J., Higgs, D.R., and Hughes, J.R. (2017). Robust detection of chromosomal interactions from small numbers of cells using low-input Capture-C. Nucleic Acids Research, doi:10.1093/nar/gkx1194.

Oudelaar, A.M., Downes, D.J., Davies, J.O.J., and Hughes, J.R. (2017).Low-input Capture-C: A Chromosome Conformation Capture Assay to Analyze Chromatin Architecture in Small Numbers of Cells. Bio-Protocol 7, doi:10.21769/BioProtoc.2645.

The LI-Capture-C and Tag-Capture-C data in the α-globin locus are available in a UCSC Genome Browser Track Hub and can be accessed with the following link:

http://sara.molbiol.ox.ac.uk/public/hugheslab/LowInputCaptureCHub/hub.txt

Acknowledgements

The optimisation strategy for the 3C library preparation protocol for small cell numbers was designed with Damien J. Downes and James O.J. Davies, and the low-input 3C libraries were generated by Damien J. Downes.

Chapter 6 – Discussion

A refined view of genome organisation

In the current model of genome organisation (Figure 1.2), chromosomes are believed to be hierarchically structured by central, invariant building blocks, formed by self-interacting domains usually referred to as TADs. Within these domains, genes and their *cis*-regulatory elements are thought to form tissue-specific chromatin loops. On a larger scale, TADs with similar transcriptional activity and chromatin signatures are believed to self-assemble to form larger, functionally distinct compartments (Bonev and Cavalli, 2016). However, careful investigation of the chromatin organisation of the murine globin loci shows that this model is an oversimplification.

First of all, Hi-C analyses of the extended globin loci (~3.3 Mb) show that the extent to which the genome is structured into TADs is highly variable. Though some chromatin regions are organised in consecutive TADs, large segments of chromosomes do not form such well-defined structures (Figures 2.1 and 2.2). While this is in agreement with a recent study in which TAD coverage in the HeLa genome was found to be only 56% (Wutz et al., 2017), it has not been emphasised in the literature, in which 'typical' examples of chromosomal regions usually contain highly structured consecutive domains (Figure 1.2).

Moreover, comparison of Hi-C data between ES cells, in which the globin genes are inactive, and erythroid cells, in which they are highly expressed, clearly shows that TADs containing and surrounding these loci are not conserved between cell types (Figures 2.1 and 2.2). Though this seems to be in contrast with the current literature, careful inspection of the data underlying claims that TADs are stable across different cell types reveals that these statements are not well supported (Dixon et al., 2012; 2015; Rao et al., 2014; Schmitt et al., 2016). As described in Chapter 2, only about half of the TADs identified in these studies are invariant. It is possible that these conserved domains reflect large genomic areas with similar transcriptional activity and/or chromatin signatures in the considered cell types. However, this does not mean that TADs represent stable structures. Our data clearly show a dramatic reorganisation of TAD structures surrounding the globin genes associated with their activation, which is further confirmed by genome-wide analyses showing that tissue-specific TADs are highly enriched for differentially expressed genes (Dixon et al., 2012).

Furthermore, high-resolution analysis of the TADs containing the globin loci using Capture-C reveals that the chromatin interactions within these domains are more complex than simple chromatin loops. Though we identify strong interactions between enhancers and promoters, these interactions are not exclusively formed between these elements, but enriched throughout a region of chromatin delineated by CTCF-binding elements. The nature of these structures is therefore not indicative of the discrete, stable loops between *cis*-regulatory elements that have been described (Figures 1.2 and 1.10) (Tolhuis et al., 2002). Rather, these interactions indicate the formation of a dynamic compartmentalised chromatin domain, in which, at some point, every region of chromatin interacts with every other, and preferred, transiently stabilised structures are formed between *cis*-regulatory elements. The CTCF-binding elements that demarcate these compartmentalised domains do not participate in the enhancer-promoter interactions within the domains, but form distinct structures. These structures represent diffuse interactions between ~50 kb regions of chromatin bound by CTCF on either side of the domains, rather than the described punctate loops between individual CTCF-binding elements (Figures 1.2 and 1.3) (Rao et al., 2014).

The discrepancy between our findings and the chromatin structures described in the literature likely results from differences in sensitivity and resolution of the techniques used to identify these structures. For example, the diffuse interactions between CTCF-binding elements that we detect using high-resolution Capture-C, would appear as a single pixel in a Hi-C contact matrix at 50 kb resolution, and could therefore be interpreted as a specific chromatin loop formed between two individual CTCF-binding elements. This highlights the importance of using technology with sufficient sensitivity to detect chromatin structures with the resolution required to interpret them correctly.

Importantly, our high-resolution analyses of the globin loci can reconcile the disagreement between the stable structures identified in 3C analyses of bulk populations of cells (Bonev and Cavalli, 2016) and the high variability in chromatin architecture observed in single-cell Hi-C studies (Nagano et al., 2013; 2017; Stevens et al., 2017) and polymer models (Giorgetti et al., 2014), as our data show that these structures are much more dynamic than initially assumed.

Together, this suggests that there is no such thing as an invariant building block of the genome. Rather, chromosomes appear to be arranged in tissue-specific structures across all layers of their organisation, which are likely closely related to the function of the *cis*-regulatory elements contained within them.

Molecular mechanisms underlying chromatin architecture

The diffuse patterns of interactions between *cis*-regulatory elements detected by Capture-C suggest that such interactions are variable between cells. However, Capture-C data represent average chromatin structures in bulk populations of cells, and therefore do not provide insight into the multiple dynamic conformations that are formed within individual cells. Therefore, to unravel the dynamic processes underlying domain formation and productive enhancer-promoter interactions, analysis of single-cell chromatin structures is required. As current single-cell 3C techniques do not have sufficient sensitivity to detect interactions at the resolution of individual *cis*-regulatory elements, we developed Tri-C, a new 3C approach that can analyse concurrent chromatin interactions with a viewpoint of interest at individual alleles at maximal resolution.

To get more insight into the formation of chromatin domains, we used Tri-C to study multi-way interactions of CTCF-binding elements flanking the globin loci in single cells (Figures 4.9 and 4.10). Our analyses clearly show that these boundary elements do not form stable loops. Rather, the CTCF-binding elements interact throughout the entire chromatin domains with diffuse enrichments over the CTCF-binding regions flanking the other sides of the domains. These interaction patterns therefore support the recently proposed mechanism of loop extrusion underlying the formation of chromatin domains. In this model, loop extruding factors translocate along chromatin and bring together progressively more distant loci along a chromosome, thereby producing dynamically expanding chromatin loops. Translocation of extruding factors is probabilistically halted at genomic loci that contain barriers to extrusion, such as elements bound by CTCF, which form transient domain boundaries (Figure 1.4) (Fudenberg et al., 2016; Sanborn et al., 2015). The Tri-C interaction patterns of the CTCF-binding elements flanking the globin loci can be explained by this model, as the observed diffuse interactions throughout the globin domains, with broad enrichments over the CTCF-binding chromatin regions on the other sides of the domains, are indicative of a continuous translocation process and transient stalling at CTCF boundary elements.

As described in Chapter 1, loop extrusion could explain many observed aspects of chromosome organisation and provides a mechanism that could explain how molecular interactions between proteins (for example, CTCF) that are a few nanometers in size can organise micron-long chromosomal segments (Figure 1.1) (Dekker and Mirny, 2016; Fudenberg et al., 2018). However, loop extrusion is difficult to assay directly, and remains a hypothetical process. Our analyses of the single-cell dynamics of boundary elements in the globin loci therefore provide an important new line of support for a mechanism of loop extrusion underlying chromatin domain formation.

We also used Tri-C to further investigate the nature of interactions between enhancers and promoters in individual cells. Our high-resolution analysis by Capture-C has shown that these elements do not form discrete chromatin loops, as initially proposed (Tolhuis et al., 2002), but stabilised interactions in the context of compartmentalised chromatin domains. However, what remains unclear is whether these interactions form structures in which multiple enhancers and promoters interact simultaneously, or if these *cis*-regulatory elements compete for the formation of mutually exclusive interactions. This important question cannot be addressed by current 3C techniques, as they are not able to detect concurrent interactions between multiple *cis*-regulatory elements in individual cells with sufficient sensitivity and resolution. Our Tri-C analyses of the major globin enhancers therefore demonstrate for the first time that enhancers and promoters form higher-order structures in which they interact simultaneously. These structures are similar to the proposed 'active chromatin hub' model (Figure 1.10) (Tolhuis et al., 2002), with the important distinction that the interactions do not represent stable, exclusive enhancer-promoter loops, but preferred interactions within dynamic compartmentalised domains.

Such structures, formed by preferred interactions rather than stable loops, are in agreement with a dynamic model of loop extrusion contributing to chromatin architecture. However, both the strength and the complexity of the interactions between enhancers and promoters cannot be explained by a loop extrusion mechanism alone. Chromatin domains that are believed to be formed by loop extrusion, such as TADs, are characterised by an approximately two-fold increase in interactions within the domain compared to neighbouring domains (Dekker and Mirny, 2016). This might explain the compartmentalised structures we observe in our Capture-C data, but cannot account for the strong enriched interactions between enhancers and promoters within these domains. Therefore, a model of loop extrusion alone is not sufficient for reliable formation of strong enhancer-promoter interactions. Furthermore, as loop extrusion is essentially a linear, one-dimensional scanning process, it is not clear how it could contribute to formation of the observed specific, complex, three-dimensional structures between enhancers and promoters.

It is therefore likely that the two-fold difference in interaction frequencies resulting from loop extrusion is amplified by other mechanisms that stabilise specific interactions between *cis*-regulatory elements (Mirny, 2010). In this model (Figure 6.1), loop extrusion brings these elements in close proximity and facilitates further stabilisation of more complex structures. The formation of such stabilised structures could be mediated by protein-protein interactions between transcription factors bound at enhancers and the pre-initiation complex at the promoter. Our observations in the mouse model in which the strongest α-globin enhancers are

deleted provide support for this model. Removal of these strong enhancers weakens the interactions between the remaining α-globin enhancers and promoters, but does not affect the formation or boundaries of the compartmentalised chromatin domain (Figure 3.6). Since the strongest α-globin enhancers are occupied by the greatest amount of transcription factors (Figures 1.7 and 1.9), it is likely that interactions mediated by multi-protein complexes are involved in the formation of enhancer-promoter structures. Further support for this model is provided by recent predictions and simulations suggesting that interactions mediated by proteins bound at *cis*-regulatory elements could be involved in phase separation of multi-molecular assemblies implicated in transcriptional regulation (Hnisz et al., 2017).

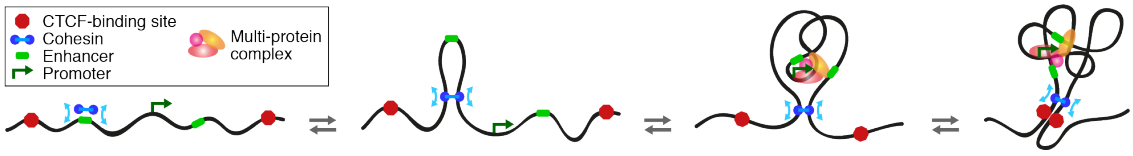


Figure 6.1: Proposed model for the formation of chromatin structures. *Our data are supportive of a loop extrusion mechanism contributing to the formation of compartmentalised chromatin domains. Complex higher-order structures, in which multiple enhancers and promoters interact, are formed within these domains by distinct tissue-specific mechanisms, likely involving the formation of phase-separated protein assemblies of the transcriptional machinery.*

Our analyses in the mouse model with engineered deletions in the CTCF-binding elements upstream of the α-globin genes provide additional support for the model presented in Figure 6.1. Removal of these boundary elements results in increased interactions between the strong α-globin enhancers and the genes further upstream, which are normally shielded from these enhancer elements by the CTCF-binding elements (Figure 3.4). The increased interactions between the α-globin enhancers and these genes are associated with a strong upregulation of their transcription (Figure 3.5). These observations support a model in which loop extrusion halted at CTCF-binding elements contributes to the formation of chromatin domains in which specific interactions between enhancers and promoters take place. However, removal of the CTCF boundaries does not affect the interactions between the enhancers and the α-globin genes, indicating that these interactions are not exclusively and/or directly mediated by the CTCF boundaries of the chromatin domain in which they occur. Our observations also highlight that there is no competition between genes for activation by *cis*-regulatory elements, as the increased interactions between the α-globin enhancers and the genes upstream of the deleted CTCF boundary did not reduce the levels of transcription of α-globin itself. It should be noted though that the α-globin genes are expressed at very high levels and that a small reduction in their activity might be difficult to observe. It would be interesting to further

analyse the conformational changes resulting from the boundary deletion in the α-globin domain by performing Tri-C to investigate whether the upstream genes are incorporated in the higher-order structures formed by the α-globin enhancers and promoters.

Our model of chromatin domain formation is further supported by analyses of the α-globin domain in single cells using FISH, which provides a complementary approach to study different aspects of chromosome conformation. Increased 3C interactions in the α-globin domain associated with activation of the α-globin genes could reflect a more compacted structure, or alternatively, increased dynamic interactions within a de-compacted region of chromatin. Measurements of the volume occupied by the α-globin domain can distinguish between these two possibilities. In a recent study by our groups, large FISH probes covering the entire α-globin domain were used to measure differences in the volume of the α-globin domain throughout erythroid differentiation. These analyses show that the α-globin locus is structured in a de-compacted self-interacting domain in erythroid cells, indicative of the formation of more dynamic contacts, as would be predicted by loop extrusion. Of interest, these structures are absent in ES cells, but form early in erythropoiesis, prior to full activation of the globin genes. Though the expression of the globin genes in these early erythroid cells is absent or very low, their *cis*-regulatory elements are bound by erythroid transcription factors. This indicates that tissue-specific chromatin structures depend of the presence of activated lineage-specific *cis*-regulatory elements for their formation, and subsequently contribute to a permissive environment for gene activation (Brown et al., 2017).

Importantly, our observations in the globin loci, indicating distinct, complementary mechanisms contributing to genome organisation, are supported by genome-wide analyses of chromatin architecture upon perturbations of the loop extrusion machinery. Due to their critical cellular functions, the effects of global loss of CTCF and Cohesin on chromatin architecture genome-wide have been difficult to investigate. Initial depletion studies suggested a modest role in genome organisation for both factors, but these conclusions were likely confounded by incomplete loss of the proteins (Seitan et al., 2013; Sofueva et al., 2013; Zuin et al., 2014). More recent approaches, including use of the auxin-inducible degron (AID) system (Nishimura et al., 2009), have enabled functional analyses of near complete depletion of these proteins, which revealed much more dramatic effects. One of these studies used the AID system to deplete CTCF and showed that this results in a complete loss of TADs across the genome. However, the removal of CTCF does not alter the genome-wide relationship between contact frequency and linear genomic distance between loci, indicating a similar level of chromatin packaging upon loss of CTCF (Nora et al., 2017). Depletion of chromatin-associated Cohesin,

either by deleting its loading factor Nipbl (Schwarzer et al., 2017) or by direct AID-mediated depletion (Rao et al., 2017), results in a similar loss of TAD structure, but also affects the rate of contact frequency decay with genomic distance. These results can be explained by a model of loop extrusion by Cohesin halted at CTCF-binding boundaries. In this model, both factors are required for the formation of TADs. However, as CTCF only plays an instructive role for extrusion activity, loss of CTCF results in perturbed domain structures, but does not affect contact frequency decay. Without Cohesin though, the process of extrusion no longer occurs, resulting in not only the loss of TADs, but also in an altered relationship between contact frequency and genomic distance. Of interest, segregation of chromosomes into active and inactive compartments is unaffected by the CTCF perturbations (Nora et al., 2017), and even reinforced upon Cohesin displacement, after which the disappearance of TADs unmasks a finer compartment structure that more accurately reflects transcriptional activity and chromatin state (Rao et al., 2017; Schwarzer et al., 2017). This is a very important observation, as its demonstrates that chromosome architecture results from the interplay of (at least) two independent mechanisms: CTCF/Cohesin-mediated formation of domains, likely by a mechanism of loop extrusion, and CTCF/Cohesin-independent segregation of the genome into compartments defined by chromatin state. Additional evidence for this model is provided by the observation that increased processivity of Cohesin, by depleting the Cohesin release factor WAPL, increases the strength of TADs and reduces compartmentalisation (Haarhuis et al., 2017; Wutz et al., 2017). Moreover, the independence of the two mechanisms is further supported by the absence of compartments, despite the existence of TADs, in maternal zygotic pronuclei in mice (Flyamer et al., 2017). Though evidence for a mechanism of loop extrusion underlying the CTCF/Cohesin-dependent mode of organisation is accumulating (Fudenberg et al., 2018), the molecular mechanisms contributing to the second mode remain more speculative. It has been suggested that the segregation into active and inactive compartments of similar epigenetic signatures could be established by a phase separation process according to local chromatin states (Hnisz et al., 2017). Recent polymer simulations have demonstrated that a combination of loop extrusion and compartmental segregation can faithfully reproduce Hi-C data, providing support for this hypothesis (Nuebler et al., 2018). Though compartments were originally described as megabase-scale chromatin structures (Lieberman-Aiden et al., 2009), examination of these structures without the overlapping TADs have revealed that compartmental separation takes place at much smaller scales than initially appreciated (Schwarzer et al., 2017). It is therefore likely that this segregation mechanism, bringing together genomic regions of similar epigenetic signatures, could contribute to the formation of large-scale structures defined by their chromatin state, as well as stabilised enhancer-promoter structures within chromatin domains at a much smaller scale. This is further confirmed by the observation that Cohesin-independent interactions are highly enriched for ‘super-enhancers’

(Rao et al., 2017) and thus take place at the level of individual *cis*-regulatory elements. Compartmentalisation across different scales could therefore contribute to the formation and/or maintenance of interactions between *cis*-regulatory elements based on their specific chromatin signatures, which could explain why loss of CTCF and chromatin-associated Cohesin do not lead to widespread misregulation of gene activity (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017). However, removal of these proteins does lead to changes in expression of a subset of genes. Of interest, the majority of the downregulated genes are predicted to be regulated by distal *cis*-regulatory elements. Without the structure imposed by TADs, these distal elements no longer interact with their cognate genes, and instead act on neighbouring genes and/or cryptic promoters located in their near vicinity (Schwarzer et al., 2017). This shows that CTCF/Cohesin-mediated loop extrusion plays an important role in guiding communication between *cis*-regulatory elements and genes, particularly over long distances.

Together, our in-depth analyses of the globin loci and the genome-wide investigations of mechanisms involved in chromosome architecture described in recent studies, indicate that the classic picture of genome organisation (Figure 1.2), in which chromosome structures at different scales represent hierarchically folded entities, is not correct. Our results and the current literature clearly show that chromosomal organisation represents a much more complex interplay between at least two independent mechanisms, with distinct, complementary roles in mediating specific communication between *cis*-regulatory elements required for precise regulation of gene activity (Figure 6.1).

Cooperation between *cis*-regulatory elements

The formation of complex chromatin structures in which multiple enhancers and promoters interact simultaneously provides a structural basis for functional cooperativity between *cis*-regulatory elements. Such cooperativity has been described for the globin (Bender et al., 2012; Hay et al., 2016) and many other genes (Osterwalder et al., 2018), and is characterised by additive effects of individual enhancer elements on transcriptional output.

Transcriptional regulation by multiple *cis*-regulatory elements might serve to ensure robustness of gene expression patterns, as the presence of several enhancer elements that can recruit transcription factors and create a permissive chromatin environment for gene activation could buffer potential deleterious effects of mutations in such elements. Our analyses in mouse models with deletions in individual α-globin enhancers provide support for this hypothesis, since we show that no single element in the α-globin locus is critical for globin expression and

that the chromatin structures associated with active α-globin transcription do not collapse upon deletion of a single enhancer.

Moreover, the observed structural cooperativity suggests a function for apparent redundant enhancer elements (Frankel et al., 2010; Perry et al., 2010). Such 'shadow' enhancers could have a structural role in forming and maintaining effective platforms for assembly of the transcriptional machinery and ensure the formation of robust chromatin complexes, even in the context of mutations in other enhancer elements.

Redundancy between *cis*-regulatory elements could explain the rarity of Mendelian disorders caused by single mutations in *cis*-regulatory elements (Kleinjan and Coutinho, 2009). However, it might also permit regulatory variants with minor effects in isolation to become prevalent in populations. It is possible that these variants have a more significant impact when they are inherited in combination with similar variants. Such combinatorial effects likely account for at least part of the 'missing heritability' of common disease that cannot be explained by the contribution of individual genetic variants discovered in GWAS.

This highlights the importance of careful analyses of the chromatin architecture of GWAS loci. The Tri-C method provides a new tool to analyse how genetic variations might disrupt complex chromatin structures and cause misregulation of gene expression and disease. Furthermore, the low-input Capture-C methods make it possible to study chromatin interactions in small cell numbers, thereby enabling analyses of rare cell and tissue types relevant for disease that were previously intractable.

We hope that the development of these techniques, together with our in-depth analyses of the globin loci as an illustrative example, contribute to our fundamental understanding of genome organisation and its role in the regulation of gene expression in health and disease.

Chapter 7 – Methods

This chapter provides an overview of the experimental methods we used to generate the data presented in this thesis. It contains a detailed description of the newly developed techniques and a more concise description of methods that have been published in detail previously. Experiments that were performed in collaboration with others are explained briefly.

Cells

All 3C based experiments in wild type cells were performed in primary murine Ter119+ erythroid cells derived from adult mouse spleens and/or mouse embryonic stem cells. Capture-C experiments to analyse the effects of the combined deletion of the CTCF-binding sites HS-38 and HS-39 in the α-globin locus were performed in primary murine Ter119+ erythroid cells derived from spleens of adult homozygous mutant mice and wild type litter mates. Since homozygous deletion of both the R1 and R2 α-globin enhancers in mice is lethal and causes embryonic death at ~13.5 days, the effects of the individual and combined enhancer deletions were assessed by analysing Ter119+ erythroid cells derived from foetal livers of homozygous mutant embryos and wild type litter mates at 12.5 days. The effects of inhibiting transcription were also assessed in Ter119+ erythroid cells derived from foetal livers.

Erythroid cells derived from adult mouse spleens

Primary Ter119+ erythroid cells were obtained from spleens of mice treated with phenylhydrazine, which causes haemolytic anaemia resulting in expansion of erythroid cells in the spleen (> 80% of cells are CD71+/Ter119+) (Spivak et al., 1973). Female C57BL/6 mice were treated with three doses of 40 mg/g body weight, given 12 hours apart, and sacrificed after five days by Jacqueline A. Sharpe. Spleens were dissociated into single cell suspensions, which were washed, stained with Ter119 antibodies conjugated with phycoerythrin, and positively selected using anti-phycoerythrin MACS beads, as previously described (Davies et al., 2015).

Embryonic stem cells

Wild type mouse ES cells (129/Ola) were derived from mice at embryonic day 14. Cells were cultured and harvested by Lars L.P. Hanssen and Mira T. Kassouf, as previously described (Davies et al., 2015).

Erythroid cells derived from foetal livers

Primary Ter119+ erythroid cells were obtained from foetal livers from embryos at 12.5 days. Single cell suspensions of liver cells were expanded, after which erythroid precursors were selected and differentiated into mature erythroid cells. This was performed by Deborah Hay and Jill M. Brown, as previously described (Hay et al., 2016). To assess the effects of transcription inhibition by DRB, the differentiated foetal liver cells were treated with 100 μM DRB dissolved in DMSO or DMSO alone (control) for three hours in standard tissue culture conditions.

Hi-C

Hi-C data were generated in collaboration with Job Dekker's laboratory (University of Massachusetts). Aliquots of ~10^7 primary Ter119+ erythroid cells were fixed with 1% formaldehyde by James O.J. Davies, as previously described (Davies et al., 2015). Hi-C experiments were performed by Yu Liu, as previously described (Lieberman-Aiden et al., 2009).

Capture-C

Capture-C experiments in wild type, D38/39 and ΔR1/R2 erythroid cells, and in ES cells were performed in three biological replicates. Experiments in erythroid cells with individual α-globin enhancer deletions were performed in single replicates.

All experiments were performed according to the Next-Generation Capture-C protocol, which has been described in detail previously (Davies et al., 2017) and is explained briefly below. The procedure contains three stages:

(I) 3C library preparation;
(II) Capture-C library preparation;
(III) Enrichment for fragments of interest targeted by capture oligonucleotides.

3C library preparation

Aliquots of ~10^7 cells were fixed with 2% formaldehyde for 10 minutes. The reaction was quenched with glycine. Fixed cells were washed in cold PBS, resuspended in cold lysis buffer (10 mM Tris pH 8, 10 mM NaCl, 0.2% Igepal CA-630, 1x cOmplete Protease Inhibitor Cocktail, water), incubated on ice for 20 minutes, and centrifuged. Pelleted cells were snap frozen, resuspended in restriction buffer and Dounce homogenised on ice. Cells were treated with SDS and Triton X-100 before digestion with DpnII. Three aliquots of 1500 U DpnII restriction enzyme were added several hours apart over a total incubation time of 16-24 hours on a shaking thermomixer at 37 °C. The digestion reaction was heat-inactivated at 65 °C for 20 minutes and cooled on ice. Digested chromatin was ligated with 720 U of T4 DNA ligase overnight on a shaking thermomixer at 16 °C. Ligated DNA was reverse crosslinked and proteinase K treated overnight at 65 °C. Samples were treated with RNase and DNA was purified using phenol-chloroform extraction and precipitation with ethanol and sodium acetate at −80 °C. After washing, the resulting 3C libraries were resuspended in water and stored at −20 °C.

To assess the quality of 3C libraries, controls of non-digested chromatin and non-ligated digested chromatin were prepared in parallel. Assessments of unspecific degradation, digestion efficiency and ligation efficiency were performed by gel electrophoresis of these controls and the 3C library, respectively (Figure 7.1).

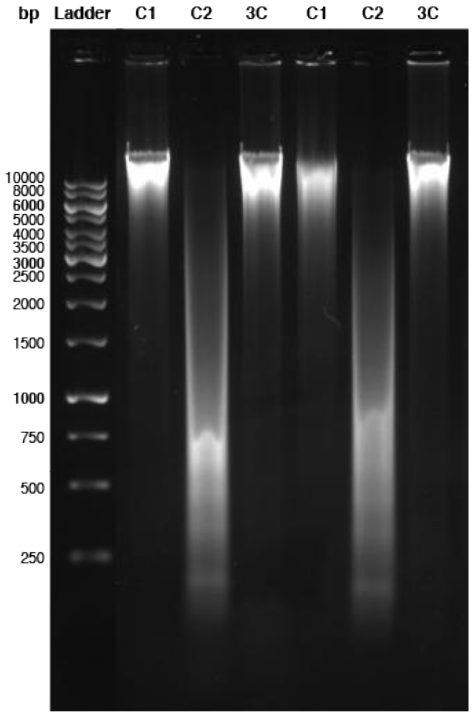


Figure 7.1. Quality control of 3C libraries. *DNA gel electrophoresis (1% agarose gel) of two representative non-digested chromatin controls (C1), digested non-ligated chromatin controls (C2), and 3C libraries (3C).*

Capture-C library preparation

Aliquots of 5 µg of 3C library were sonicated to ~200 bp fragments using a Covaris S220 Focused Ultrasonicator (settings: six cycles of 60s; duty cycle, 10%; intensity, 5; cycles per burst, 200). Illumina TruSeq adaptors were added using NEBNext DNA Library Prep reagents, according to the manufacturer's protocol. Adapter-ligated libraries were amplified and indexed using Herculase II reagents in a 7-cycle PCR reaction, according to the manufacturer's protocol. DNA clean-up steps were performed using AMPure XP beads. Two aliquots of each library were processed in parallel and ligated with the same index to generate maximum library complexity.

Sonication, adapter ligation and amplification efficiencies were assessed using the Agilent Tapestation (D1000) (Figure 7.2).

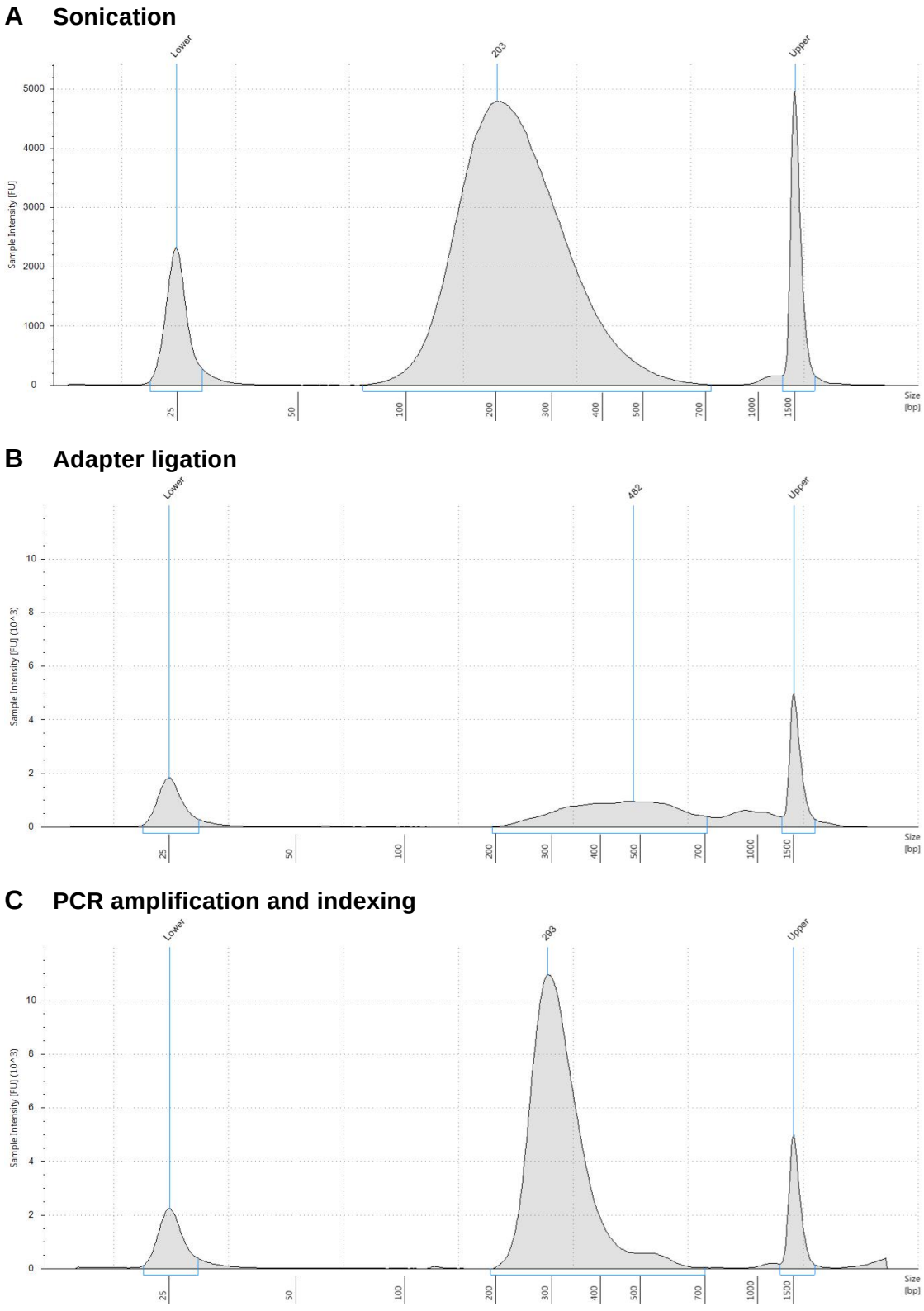


Figure 7.2. Quality control of Capture-C libraries. *Assessment of processed 3C libraries after* ***(A)*** *sonication,* ***(B)*** *adapter ligation, and* ***(C)*** *PCR amplification and indexing, using the Agilent Tapestation (D1000).*

Enrichment for fragments of interest by oligonucleotide capture

Enrichment by oligonucleotide capture was performed using Nimblegen SeqCap EZ reagents and custom biotinylated DNA oligonucleotides targeted against viewpoint restriction fragments of interest. Differently indexed libraries were pooled and hybridised to multiple capture oligonucleotides simultaneously, allowing for multiplexing of both samples and viewpoints. The procedure described below was scaled up to the number of multiplexed libraries and performed in a single tube.

Aliquots of 1-2 µg of differently indexed libraries were mixed with 5 µg mouse Cot-1 DNA and 1 nmol SeqCap EZ HE blocking oligonucleotides (both universal and index-matched). The mixture was dried completely in a vacuum centrifuge and resuspended in SeqCap EZ hybridisation reagents, according to the manufacturer's protocol. After denaturation at 95 °C for 10 minutes, a 13 fmol pool of multiplexed biotinylated capture oligonucleotides was added and hybridisation was performed at 47 °C in the thermocycler for 64-72 hours. Enrichment for DNA fragments hybridised to the biotinylated oligonucleotides was performed by adding Streptavidin beads and incubating at 47 °C for 45 minutes on a shaking thermomixer, followed by several thorough washing steps with SeqCap EZ washing buffers. Enriched DNA was resuspended in water, PCR amplified directly off the beads using KAPA/SeqCap EZ reagents, and purified using AMPure XP beads. All these steps were performed according to the manufacturer's protocol. To increase enrichment, a second round of oligonucleotide capture was performed following the same procedure, using 500-1000 ng of enriched material in a hybridisation reaction of 20-24 hours.

The enriched Capture-C libraries were assessed using the Agilent Tapestation (D1000) (Figure 7.3) and quantified using Qubit dsDNA BR and KAPA Library Quantification reagents, before sequencing using the Illumina MiSeq and NextSeq platforms. At least 500,000 150 bp paired-end reads were generated per viewpoint, per sample in multiplexed libraries to produce sufficient sequencing depth.

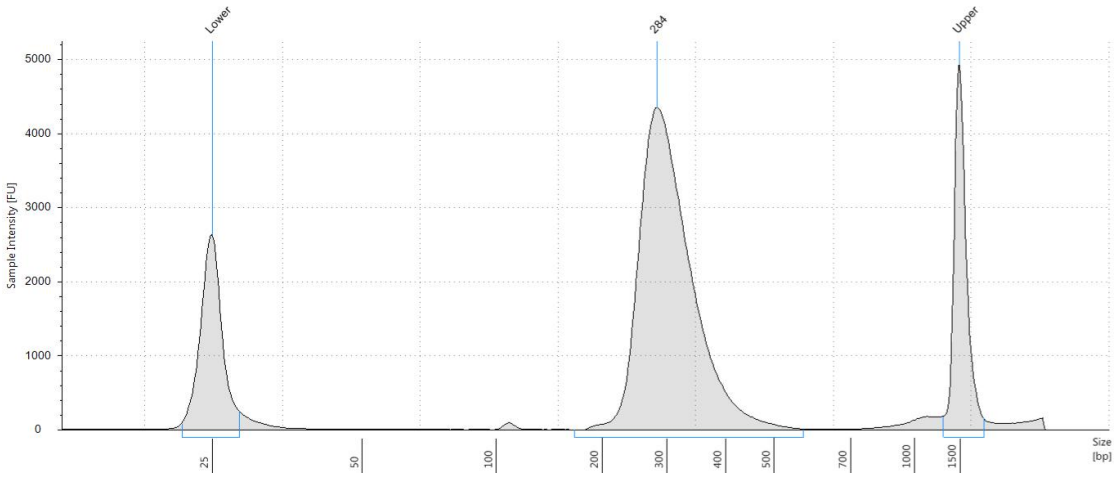


Figure 7.3. Quality control of Capture-C DNA products. *Assessment of Capture-C DNA products using the Agilent Tapestation (D1000).*

Viewpoints

Because exclusion zones around all viewpoints analysed in a multiplexed experiment are removed from analysis, several Capture-C experiments were performed to characterise the complete interaction landscapes of all *cis*-regulatory elements of interest in the globin loci. The following combinations of viewpoints were analysed in six independent experiments:

I. α-globin locus: R1 and R2 (enhancers); β-globin locus: HS1 and HS2 (enhancers);
II. α-globin locus: R3, Rm and R4 (enhancers); β-globin locus: HS3 and HS4 (enhancers);
III. α-globin locus: *Hbq1a* and *Hbq1b* (θ-globin promoters/CTCF-binding sites); β-globin locus: HS5 and HS6 (enhancers/CTCF-binding sites);
IV. α-globin locus: HS-38 and HS-39 (upstream CTCF-binding sites); β-globin locus: HS-57 (enhancer), HS-60 and HS-90 (upstream CTCF-binding sites);
V. α-globin locus: HS+44 and HS+48 (downstream CTCF-binding sites); β-globin locus: 3'HS1 and 3'HS2 (downstream CTCF-binding sites);
VI. α-globin locus: *Il9r*, *Snrnp25*, *Rhbdf1*, *Mpg* and *Nprl3* (promoters);

To cover the remaining regions of the α- and β-globin loci for the generation of a high-resolution all vs all contact matrix, an additional multiplexed experiment with viewpoints tiled across a 300 kb window around both globin clusters was performed.

Capture oligonucleotides were designed using CapSequm (Davies et al., 2015). Overviews of the viewpoint DpnII fragments in the α- and β-globin loci are shown in Tables 7.1 and 7.2, respectively.

Viewpoint	Annotation	DpnII fragment
Hba-a1	Promoter	chr11:32182969-32183821
Hba-a2	Promoter	chr11:32195804-32196638
R1	Enhancer	chr11:32145277-32146568
R2	Enhancer	chr11:32151060-32151883
R3	Enhancer	chr11:32155865-32156383
Rm	Enhancer	chr11:32164966-32165085
R4	Enhancer	chr11:32168680-32169413
Hbq1a	Promoter/CTCF-binding site	chr11:32199491-32200139
Hbq1b	Promoter/CTCF-binding site	chr11:32186467-32187034
HS-38	CTCF-binding site	chr11:32138079-32139000
HS-39	CTCF-binding site	chr11:32137176-32137426
HS+44	CTCF-binding site	chr11:32220918-32221720
HS+48	CTCF-binding site	chr11:32224323-32227298
Il9r	Promoter	chr11:32099819-32100299
Snrnp25	Promoter	chr11:32104921-32105535
Rhbdf1	Promoter	chr11:32122089-32122693
Mpg	Promoter	chr11:32126341-32126674
Nprl3	Promoter	chr11:32167508-32168372
32003k	Model_viewpoint	chr11:32002284-32002785
32012k	Model_viewpoint	chr11:32011827-32012022
32019k	Model_viewpoint	chr11:32019083-32019233
32027k	Model_viewpoint	chr11:32027129-32027804
32037k	Model_viewpoint	chr11:32036007-32037157
32042k	Model_viewpoint	chr11:32042139-32042440
32051k	Model_viewpoint	chr11:32050692-32051800
32059k	Model_viewpoint	chr11:32059255-32059583
32068k	Model_viewpoint	chr11:32067632-32068398
32077k	Model_viewpoint	chr11:32076914-32077034
32084k	Model_viewpoint	chr11:32083717-32083891
32091k	Model_viewpoint	chr11:32090436-32092064
32111k	Model_viewpoint	chr11:32110960-32111206
32119k	Model_viewpoint	chr11:32118902-32119562
32175k	Model_viewpoint	chr11:32175103-32175364
32207k	Model_viewpoint	chr11:32207138-32207281
32216k	Model_viewpoint	chr11:32215331-32215799
32236k	Model_viewpoint	chr11:32235766-32237204
32243k	Model_viewpoint	chr11:32242417-32243041
32253k	Model_viewpoint	chr11:32252649-32252858
32259k	Model_viewpoint	chr11:32258095-32259644
32267k	Model_viewpoint	chr11:32267139-32267638
32275k	Model_viewpoint	chr11:32274955-32275418
32284k	Model_viewpoint	chr11:32283727-32283847
32291k	Model_viewpoint	chr11:32291086-32291835
32298k	Model_viewpoint	chr11:32297594-32298445

Table 7.1. Overview of the viewpoints analysed in Capture-C experiments in the α-globin locus. *Experiments were performed with two 120 bp capture oligonucleotides targeting both ends of the viewpoint DpnII fragments. When one of the ends contained a repetitive sequence, only one oligonucleotide targeting the unique end was used. The table shows that 120 bp end of the restriction fragment instead of the entire fragment in that case.*

Viewpoint	Annotation	DpnII fragment
Hbb-b1	Promoter	chr7:110961967-110962817
Hbb-b2	Promoter	chr7:110975976-110976456
HS1	Enhancer (LCR)	chr7:111007506-111007930
HS2	Enhancer (LCR)	chr7:111009550-111009749
HS3	Enhancer (LCR)	chr7:111014164-111014602
HS4	Enhancer (LCR)	chr7:111018972-111019398
HS5	Enhancer (LCR)	chr7:111021350-111022005
HS6	Enhancer (LCR)	chr7:111023278-111023655
3'HS1	CTCF-binding site	chr7:110955508-110955666
3'HS2	CTCF-binding site	chr7:110940624-110941597
HS-57	Enhancer	chr7:111058110-111058705
HS-60	CTCF-binding site	chr7:111061643-111062050
HS-90	CTCF-binding site	chr7:111091378-111092626
110846k	Model_viewpoint	chr7:110845602-110845722
110852k	Model_viewpoint	chr7:110851723-110852505
110863k	Model_viewpoint	chr7:110862597-110862717
110870k	Model_viewpoint	chr7:110869104-110870170
110879k	Model_viewpoint	chr7:110879128-110879377
110887k	Model_viewpoint	chr7:110885985-110887299
110894k	Model_viewpoint	chr7:110894213-110894515
110902k	Model_viewpoint	chr7:110901673-110903198
110909k	Model_viewpoint	chr7:110908816-110908936
110919k	Model_viewpoint	chr7:110918844-110919286
110928k	Model_viewpoint	chr7:110926852-110928205
110934k	Model_viewpoint	chr7:110934018-110934138
110971k	Model_viewpoint	chr7:110971166-110971561
110983k	Model_viewpoint	chr7:110983084-110983538
110991k	Model_viewpoint	chr7:110990420-110991222
111000k	Model_viewpoint	chr7:110999928-111000287
111027k	Model_viewpoint	chr7:111026798-111027050
111036k	Model_viewpoint	chr7:111036248-111036368
111043k	Model_viewpoint	chr7:111043069-111043411
111051k	Model_viewpoint	chr7:111051223-111051670
111068k	Model_viewpoint	chr7:111068238-111068358
111078k	Model_viewpoint	chr7:111078043-111078649
111087k	Model_viewpoint	chr7:111086801-111087189
111100k	Model_viewpoint	chr7:111100104-111100247
111106k	Model_viewpoint	chr7:111106311-111106563
111122k	Model_viewpoint	chr7:111121964-111122084
111129k	Model_viewpoint	chr7:111129322-111129442
111139k	Model_viewpoint	chr7:111138889-111139009

Table 7.2. Overview of the viewpoints analysed in Capture-C experiments in the β-globin locus. *Experiments were performed with two 120 bp capture oligonucleotides targeting both ends of the viewpoint DpnII fragments. When one of the ends contained a repetitive sequence, only one oligonucleotide targeting the unique end was used. The table shows that 120 bp end of the restriction fragment instead of the entire fragment in that case.*

Tri-C

The R2 enhancer and HS-39 CTCF-binding site in the α-globin locus, and the HS2 enhancer and 3'HS1 CTCF-binding site in the β-globin locus are located on small NlaIII restriction fragments (Table 7.3). The NlaIII restriction enzyme was therefore used for digestion of 3C libraries, which were prepared as described above, using three aliquots of 600 U NlaIII. To be able to capture multiple restriction fragments in individual reads, the subsequent sonication step was optimised to generate DNA fragments of ~450 bp, using a Covaris S220 Focussed Ultra-Sonicator with the following settings: one cycle of 55 s; duty cycle: 10%; intensity: 4; cycles per burst: 200. DNA clean-up and size selection after sonication were performed with AMPure XP beads in a 1:0.7 sample:bead ratio. Subsequent ligation of sequencing adapters and indexing and multiplexing of libraries were performed as described above. To enrich for reads containing the viewpoint fragments of interest, a double oligonucleotide capture step was performed as described above, using a 13 fmol pool of capture oligonucleotides (Table 7.3).

Viewpoint	Annotation	Capture oligonucleotide	NlaIII fragment size (bp)
R2	Enhancer (α-globin)	chr11:32150965-32151084	176
HS-39	CTCF-binding site (α-globin)	chr11:32137199-32137318	136
HS2	Enhancer (β-globin)	chr7:111009445-111009564	216
3'HS1	CTCF-binding site (β-globin)	chr7:110955472-110955591	143

Table 7.3. Overview of the viewpoints analysed with Tri-C. *Experiments were performed with 120 bp capture oligonucleotides targeting the middle of the viewpoint NlaIII fragments.*

Sonication, adapter ligation, amplification and capture enrichment efficiencies were assessed using the Agilent Tapestation (DNA 1000) (Figure 7.4).

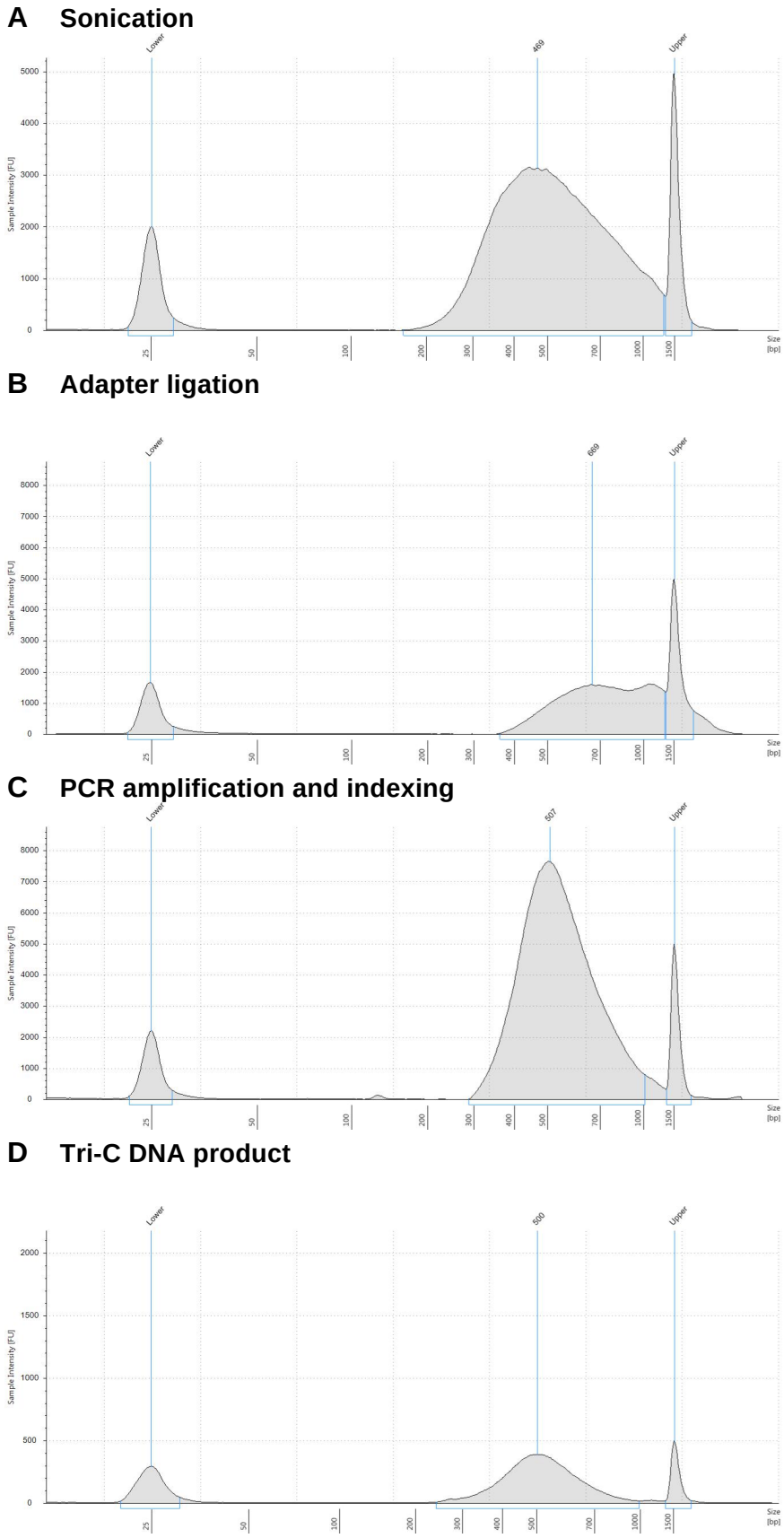


Figure 7.4. Tri-C quality control. *Assessment of processed 3C libraries after* ***(A)*** *sonication,* ***(B)*** *adapter ligation,* ***(C)*** *PCR amplification and indexing, and* ***(D)*** *capture enrichment using the Agilent Tapestation (D1000).*

A pilot experiment with three biological replicates of primary erythroid cells was initially performed, which was sequenced using the Illumina MiSeq platform (V2 chemistry; 250 bp paired-end reads). *In silico* trimming of the reads showed that there was little benefit of using 500 cycles of sequencing compared to 300 cycles for capturing multiple reporters in the reads. Therefore, to generate data at sufficient depth, a second multiplexed experiment was performed with seven additional technical replicates (derived from three biological replicates) of erythroid and ES cells, which was sequenced using the Illumina Nextseq (V2 chemistry; 150 bp paired-end reads).

C-Trap

3C libraries were prepared as described above, using the DpnII restriction enzyme for digestion. Primers targeting the restriction fragments of interest were designed >100 bp away from the restriction sites (Table 7.4), to allow for selection of reads that were selectively amplified and of sufficient quality. PCR amplification was performed using the Takara Prime Star GXL 2-step amplification program for 10-30 kb amplicons, according to the manufacturer's protocol. The AMPure XP system was used to purify the PCR product and select fragments >600 bp. The purified amplicons were adapter-ligated and sequenced using the Oxford Nanopore MinION platform (MAP-005 chemistry), according to the manufacturer's protocol.

Name	Sequence
R2 (forward primer)	AATGAGGCAGAGTTTAATGGGAATG
α-globin promoter (reverse primer)	GGCTCAAGTGAATCTGCTAGC

Table 7.4. Overview of the primers used for PCR enrichment in the C-Trap experiments targeting the interaction between the R2 enhancer and the α-globin promoter.

Low-input Capture-C

Capture-C procedures optimised for small numbers of cells are described in detail below. For steps that have been optimised previously for the Next-Generation Capture-C procedure and that have not been further adapted, we refer to the more concise descriptions in the *Capture-C* section above. The full procedure is available in Bio-protocol (Oudelaar et al., 2017).

3C library preparation

Aliquots of 10,000, 20,000, 50,000 and 100,000 cells were obtained by FACS sorting Ter119+ erythroid cells in 100 µl RPMI media supplemented with 10% FBS. Volumes were made up to 1 ml and fixed with 2% formaldehyde for 10 minutes at room temperature on a tumbling incubator. The reaction was quenched in 125 mM cold glycine, after which cells were centrifuged (15 minutes / 500 g / 4 °C). The supernatant was carefully removed, leaving ~5% behind, to not disturb the pellet. Cells were washed with 1 ml cold PBS, carefully resuspended in 500 µl cold lysis buffer (10 mM Tris pH 8, 10 mM NaCl, 0.2% Igepal CA-630, 1x cOmplete Protease Inhibitor Cocktail, water) and incubated on ice for 20 minutes. Cells were subsequently snap frozen with ethanol and dry ice and stored at -80 °C.

Aliquots of formaldehyde-fixed cells were defrosted and centrifuged (10 minutes / 15,000 g / 4 °C). Lysis buffer was carefully removed, using a P20 pipette to remove residual buffer without disturbing the pellet. Cells were resuspended in 175 µl DpnII buffer supplemented with 0.25% SDS and incubated for 1 hour at 37 °C on a ThermoMixer (1,400 rpm). After adding Triton X-100 to a concentration of 1.67%, cells were incubated for another hour at 37 °C on a ThermoMixer (1,400 rpm). Chromatin was subsequently digested with the DpnII restriction enzyme in a 200 µl reaction, which was incubated overnight at 37 °C on a ThermoMixer (1,400 rpm) and to which three doses of 150 U DpnII were added several hours apart. After heat-inactivation of the restriction enzyme at 65 °C for 20 minutes, the ligation reaction was performed in the same tube, by adding 120 U T4 ligase in ligation buffer and incubating overnight at 16 °C on a Thermomixer (300 rpm). Ligated DNA was reverse crosslinked and treated with proteinase K in an overnight incubation at 65 °C. To extract the DNA, Phenol:Chloroform:Isoamyl (25:24:1, v/v) was added and the mixture was transferred to a Phase Lock Gel Light tube for separation by centrifugation (10 minutes / 15,000 g / 4 °C). DNA was precipitated overnight at -20 °C in 70% ethanol and 80 mM NaOAc, supplemented with GlycoBlue for pellet visualisation. Precipitated DNA was centrifuged (45 minutes / 21,000 g / 4 ºC), washed with 1 ml 70% cold ethanol and resuspended in water. The resulting 3C libraries were stored at -20 °C.

To minimise losses during 3C library preparation, all cells were used for 3C library production and the standard controls (see *Capture-C – 3C library preparation*) were omitted. A novel approach, requiring minimal 3C library input, was employed for quality control. Digestion efficiencies were determined using qPCR analyses of the 3C libraries (1:10 dilution), by comparing C_t values of a primer pair spanning a ligation junction and a primer pair targeting a genomic control (Table 7.5). Libraries with a digestion efficiency >75% were used for further

processing. Ligation efficiencies of the 3C libraries were assessed using the Agilent Tapestation (Genomic) (Figure 7.5). The mass of DNA in the 3C libraries was quantified by qPCR using the primer pair targeting the genomic control and calculated relative to a standard curve generated from genomic DNA.

Table 7.5. Quality control of low-input 3C libraries by qPCR.

Primer	Sequence
Ligation junction; forward	GGAGAAAGAAGGCTGGTGTTAT
Ligation junction; reverse	TATCTGAGTTGGACAGCATTGG
Genomic control; forward	TTATCTTGCATTTGCCAACTCG
Genomic control; reverse	TGGGTTTCCCTGATTCTGAAA

Table 7.5A. qPCR primers.

Reagent	Volume
2x KAPA SYBR (+ROX)	10.4 µl
Primer mix (10µM)	0.6 µl
Water	7 µl
3C library (1:10 dilution)	2 µl

Table 7.5B: qPCR reaction.

Step	Temperature	Time
Step 1	95 °C	20 sec
Step 2	95 °C	3 sec
Step 4	60 °C	30 sec
Step 4	Go to Step 2	40 cycles
Step 5	Melt curve	–

Table 7.5C. qPCR conditions.

	Primers	C_t (average)	C_t (SD)	ΔC_t (DpnII-Genomic)	$\Delta\Delta C_t$ (Genomic-3C)	Digestion efficiency (100 * (1-$2^{\Delta CT}$))
Genomic DNA	Genomic	24.863	0.0897	0.397	-2.475	82.013 %
	DpnII	25.260	0.0245			
3C Library	Genomic	22.384	0.0399	2.872		
	DpnII	25.256	0.0075			

Table 7.5D. Digestion efficiency calculation.

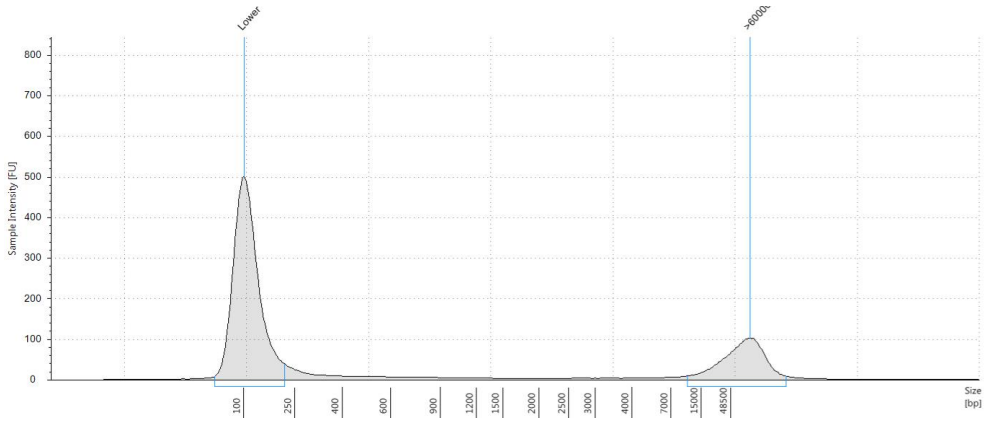


Figure 7.5. Quality control of low-input 3C libraries. *Assessment of ligation efficiency in low-input 3C libraries using the Agilent Tapestation (Genomic).*

Capture-C library preparation – Low-Input Capture-C

3C libraries (up to 1 µg in 120 µl) were sonicated to 200 bp fragments using a Covaris S220 Focused Ultrasonicator (settings: six cycles of 60s; duty cycle, 10%; intensity, 5; cycles per burst, 200). Illumina TruSeq adapters were subsequently added by performing an End Repair and Adapter Ligation reaction using NEBNext Ultra DNA Library Prep reagents, according to the manufacturer's protocol. The libraries were amplified and indexed in 9 rounds of PCR using Herculase II reagents (Table 7.6). All DNA clean-up steps were performed using AMPure XP beads (1:1.8 sample:bead ratio). The material was assessed using the Agilent Tapestation (D1000 High Sensitivity and D1000) (Figure 7.6).

Table 7.6: PCR amplification and indexing of LI-Capture-C libraries.

Reagent	**Volume**
Adapter-ligated 3C library	28.5 µl
NEB Universal primer	5 µl
NEB Index primer	5 µl
Herculase II Reaction Buffer (5x)	10 µl
dNTP Mix (25 mM)	0.5 µl
Herculase II Fusion DNA Polymerase	1 µl

Table 7.6A. PCR reaction.

Step	**Temperature**	**Time**
Step 1	98 °C	30 sec
Step 2	98 °C	10 sec
Step 3	65 °C	30 sec
Step 4	72 °C	30 sec
Step 5	Go to Step 2	9 cycles
Step 6	72 °C	5 min
Step 7	4 °C	Hold

Table 7.6B. PCR conditions.

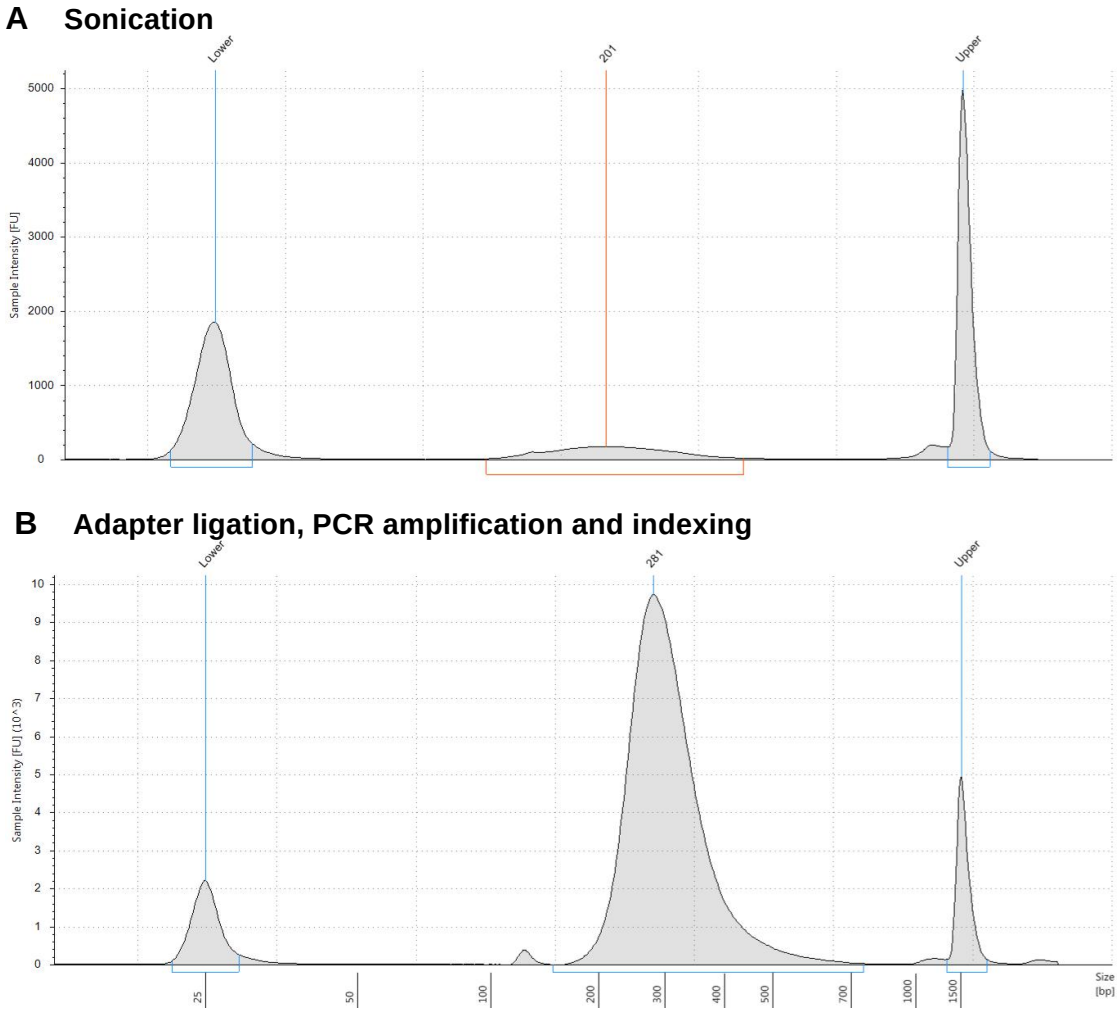


Figure 7.6. Quality control of LI-Capture-C libraries. *Assessment of processed 3C libraries after* ***(A)*** *sonication, and* ***(B)*** *adapter ligation, PCR amplification and indexing, using the Agilent Tapestation (D1000 High Sensitivity and D1000, respectively).*

Capture-C library preparation – Tag-Capture-C

3C libraries were tagmented using Nextera DNA Library Prep reagents. The manufacturer's protocol was optimised to generate fragments within our desired size range: a maximum input of 50 ng 3C library was used per reaction and reaction time was extended to overnight incubation at 55 ºC. Tagmented DNA was purified using Zymo DNA Clean & Concentrator reagents, according to the manufacturer's protocol. When input was greater than 50 ng, several tagmentation reactions were performed in parallel and pooled after purification. Libraries were amplified in 9 rounds of PCR using KAPA Hi-Fi PCR reagents (Table 7.7), as this generated a higher yield per cycle compared to the PCR strategy used in the Nextera DNA Library Prep

protocol. The material was purified using AMPure XP beads (1:1.8 sample:bead ratio) and assessed using the Agilent Tapestation (D1000 High Sensitivity and D1000) (Figure 7.7).

Table 7.7: PCR amplification and indexing of Tag-Capture-C libraries.

Primer	Sequence
Universal	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Index 1	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Index 2	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Index 3	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Index 4	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Index 5	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Index 6	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Index 7	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Index 8	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Index 9	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Index 10	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Index 11	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Index 12	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT

Table 7.7A. PCR primers (Buenrostro et al., 2013).

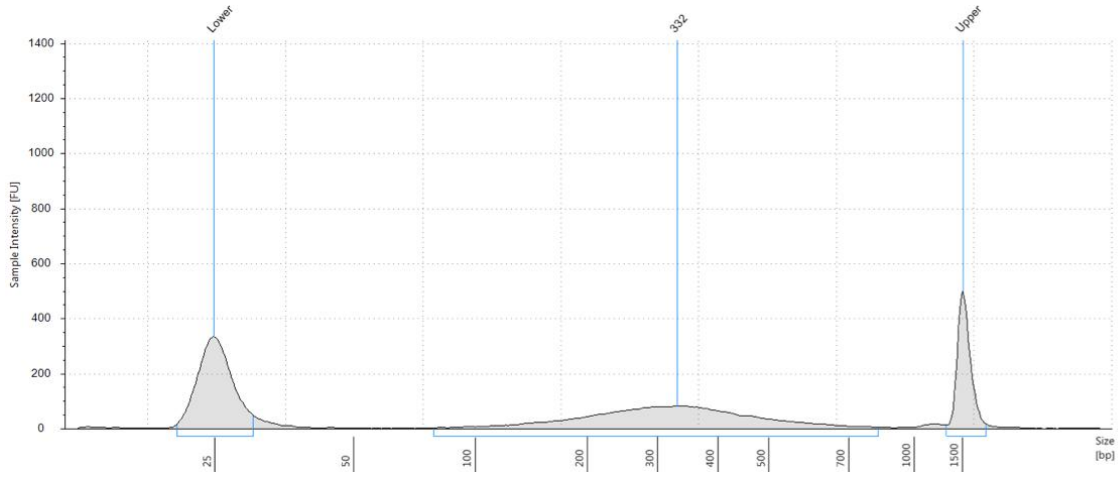
Reagent	Volume
Tagmented 3C library	33.5 µl
Universal primer (25 µM)	2 µl
Index primer (25 µM)	2 µl
KAPA HiFi Fidelity Buffer (5x)	10 µl
KAPA dNTP Mix (10 mM)	1.5 µl
KAPA HiFi DNA Polymerase (1 U/µl)	1 µl

Table 7.7B. PCR reaction.

Step	Temperature	Time
Step 1	72 °C	5 min
Step 2	98 °C	30 sec
Step 3	98 °C	10 sec
Step 4	63 °C	30 sec
Step 5	72 °C	1 min
Step 6	Go to Step 3	9 cycles
Step 7	72 °C	3 min
Step 8	4 °C	Hold

Table 7.7C. PCR conditions.

A Tagmentation



B PCR amplification and indexing

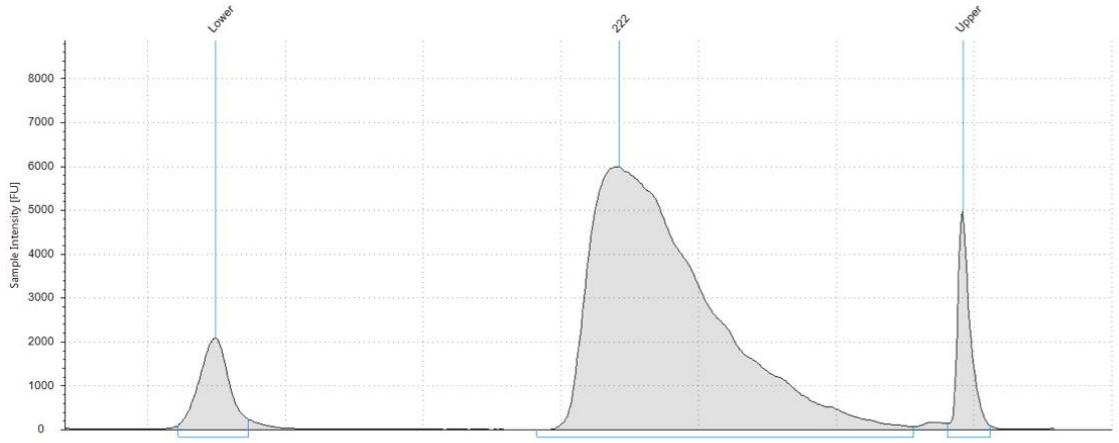


Figure 7.7. Quality control of Tag-Capture-C libraries. *Assessment of processed 3C libraries after* ***(A)*** *tagmentation, and* ***(B)*** *PCR amplification and indexing, using the Agilent Tapestation (D1000 High Sensitivity and D1000, respectively).*

Enrichment for fragments of interest by oligonucleotide capture

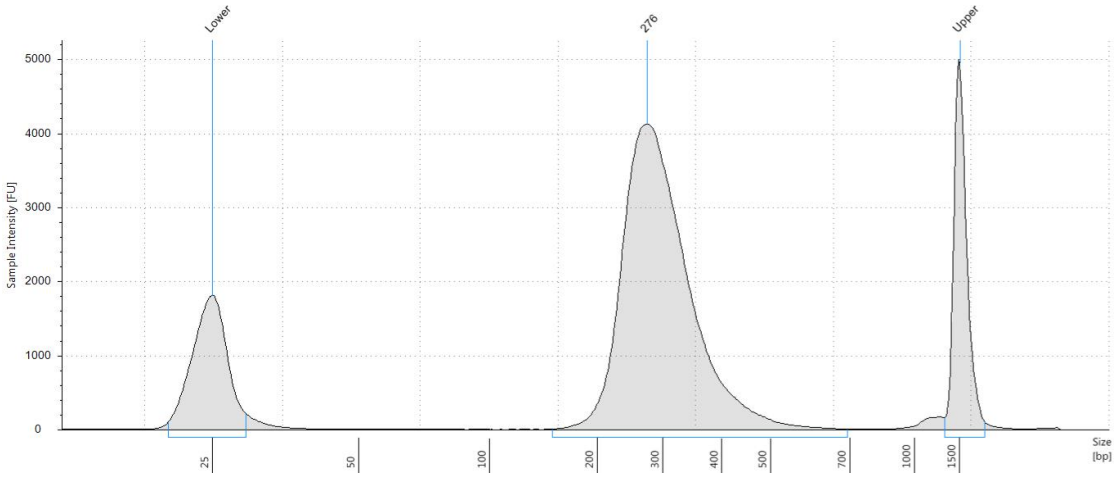
Enrichment by oligonucleotide capture has been optimised previously for the Next-Generation Capture-C protocol and is extremely efficient, with up to three million-fold enrichment for fragments of interest (Davies et al., 2017). Oligonucleotide capture enrichment of LI- and Tag-Capture-C libraries was therefore performed as described in *Capture-C – Enrichment for fragments of interest targeted by capture oligonucleotides*, using 700 ng–1.5 µg of LI-/Tag-Capture-C libraries.

Tag-Capture-C libraries contain sequencing adapters that are not compatible with commercially available blocking oligonucleotides. These oligonucleotides are used to prevent

concatemerisation through annealing of the adaptor sequences (‘daisy chaining’), which markedly reduces hybridization efficiency. The Tag-Capture-C oligonucleotide capture procedure was therefore performed using custom blocking oligonucleotides designed against the tagmentation adapters (Table 7.7) (xGen Custom Blocking Oligos, IDT).

The enriched LI-/Tag-Capture-C libraries were assessed using the Agilent Tapestation (D1000) (Figure 7.8) and quantified using Qubit dsDNA BR and KAPA Library Quantification reagents. Capture-C sequencing libraries generated from small numbers of cells require lower sequencing depth, as the complexity of these libraries is lower. However, similar to Next-Generation Capture-C libraries, at least 500,000 150 bp paired-end reads were generated per viewpoint, per sample in multiplexed libraries, to maximise data generation from low input samples. The libraries were sequenced using the Illumina MiSeq and NextSeq platforms.

A LI-Capture-C DNA product



B Tag-Capture-C DNA product

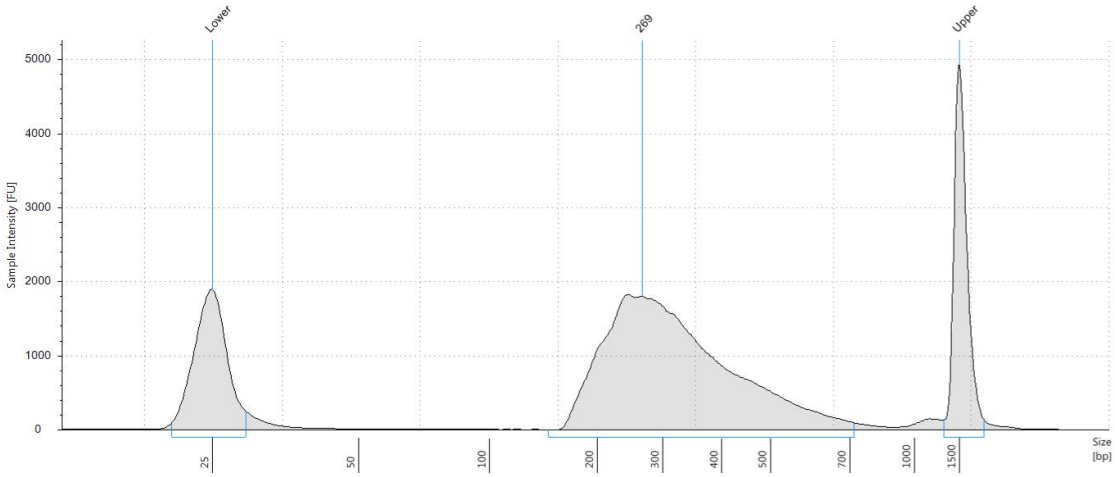


***Figure 7.8. Quality control of low-input Capture-C DNA products**. Assessment of Capture-C DNA products generated with **(A)** LI-Capture-C and **(B)** Tag-Capture-C, using the Agilent Tapestation (D1000).*

Chapter 8 – Data analysis

This chapter provides an overview of the bioinformatic approaches that we used to analyse the data presented in this thesis.

Software packages

The following software packages were used and are referred to throughout the methods section:

- bedtools v2.25.0 (Quinlan and Hall, 2010);
- BLAST v2.3.0 (Altschul et al., 1990; Camacho et al., 2009; Zhang et al., 2000);
- BLAT v3.2.1 (Kent, 2002);
- bowtie v1.1.2 (Langmead et al., 2009);
- bowtie2 v2.2.5 (Langmead and Salzberg, 2012);
- FLASH v1.2.8 (Magoč and Salzberg, 2011);
- poretools v0.5.1 (Loman and Quinlan, 2014);
- samtools v1.3 (Li et al., 2009);
- TrimGalore 0.3.3 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore);
- UCSC browser (Kent et al., 2002).

Custom scripts were written in Perl, Python and R.

Statistical analyses

Statistical significance of differential pair-wise (Capture-C) and multi-way (Tri-C) interactions between erythroid and ES cells was analysed using the Bioconductor package DESeq2 (Love et al., 2014). This method tests for differential signals by use of negative binomial generalised linear models, which is a common approach for the analysis of count-based data generated by next-generation sequencing. As DESeq2 normalises the data internally, we ran the model using raw interaction counts per restriction fragment (Capture-C) or bin (Tri-C). The remaining statistical analyses were performed with Student's two-tailed *t* tests using GraphPad Prism software.

Hi-C

The Hi-C data in primary murine Ter119+ erythroid cells were analysed in parallel with an equivalent dataset in mouse ES cells (Cast/129) (Giorgetti et al., 2016), which was generated using exactly the same experimental procedure, to allow for comparisons between the two cell types. Both datasets were analysed with the HiC-Pro pipeline (Servant et al., 2015), using the default parameters.

Reads were aligned to the mm9 reference genome using Bowtie2, with minor modifications to the recommended options (erythroid: -k 3 --score-min L,-0.6,-0.2; ES: -k 3 --score-min L,-0.6,-0.6) to allow for multi-mapping in the duplicated regions in the globin loci and better alignment of the ES data in the β-globin region, which contains many SNPs in the Cast/129 strain compared to the mm9 reference. Despite these adjustments, alignment of the ES Hi-C data in this region remained somewhat compromised.

The erythroid Hi-C library was sequenced to a depth of 1,256,434,953 read pairs, which resulted in 714,421,216 valid interactions, with 63.3% of the interactions occurring in *cis*.These data are therefore suitable to display at 10-20 kb resolution. However, for comparative analysis with the ES Hi-C dataset (445,453,769 reads; 161,276,943 valid interactions; 51.3% of interactions in *cis*), we down-sampled the erythroid data to 156,921,018 valid interactions. These interactions were normalised using the Iterative Correction and Eigenvector decomposition (ICE) procedure implemented in the HiC-Pro pipeline. TADs were identified based on insulation indices using TADtool (Kruse et al., 2016). For robust TAD calling and visual comparison of structural features, Hi-C contact matrices were analysed at 40 kb resolution.

The SBS models were based on normalised contact matrices at 20 kb resolution, as these provided sufficient depth for the SBS simulations. Note that the TADs called in the corresponding contact matrices (Figures 2.4 and 2.5) are slightly different from TADs identified at 40 kb resolution (Figures 2.1 and 2.2).

Due to compromised mapping of the ES data at the β-globin locus, noise levels in the ES contact matrix of this region are relatively high. Though this complicates direct quantitative comparison of interactions between erythroid and ES cells, differences in domain structure are apparent.

Capture-C

Capture-C data were analysed using a combination of publicly available tools and customised scripts. TrimGalore was used to remove adapter sequences in the reads. Where possible, paired-end reads were reconstructed into single reads using FLASH with interleaved output settings. A custom script was used to perform an *in silico* restriction enzyme digestion, after which the reads were aligned to the mm9 reference genome using Bowtie1 with the following options: -p 1 -m 2 -v 3 --best --strata. The -m 2 option was used to allow for multi-mapping in the duplicated regions in the globin loci. The -v 3 option was used to allow up to three mismatches to improve alignment of the ES data in the β-globin region, which contains many SNPs in the 129/Ola strain compared to the mm9 reference. Because some SNPs were located in close proximity to or even in DpnII cut sites, alignment of reads with viewpoints in these DpnII fragments remained sub-optimal. The quality of some Capture-C profiles in the β-globin region in ES cells is therefore somewhat compromised, though still highly interpretable.

Aligned reads were filtered for PCR duplicates using the random unique ends of the reads resulting from the sonication step as two independent UMIs. Reads with the exact same start and end coordinates of each individual restriction fragment were excluded from analysis. Uniquely aligned reads were analysed with custom scripts to identify captured reads containing targeted viewpoints of interest. Restriction fragments in captured reads were defined as interacting 'reporter' fragments if they were located outside a ~1 kb exclusion zone around the captured viewpoint restriction fragment. All other viewpoint restriction fragments in multiplexed experiments were also excluded from analysis. Unique interactions were therefore reported only when the read was unique and a single capture and one or more reporter restriction fragments were identified in a single read. These unique interactions were filtered to remove regions of spurious alignment due to copy-number differences or potential misalignment of reads from the proximity-exclusion region, which were identified using BLAT. The filtered unique interactions were normalised for the total number of interactions on the chromosome analysed.

Because interactions with the duplicated globin promoters (*Hba-a1* and *Hba-a2*, and *Hbb-b1* and *Hbb-b2*) cannot be distinguished bioinformatically, interaction profiles of the α- and β-globin promoters represent composite profiles in which the data from the duplicated promoter viewpoints are pooled.

Statistical analyses were performed using the DESeq2 package. Statistically significant interactions indicated in Figures 2.6 and 2.8 represent restriction fragments with significantly different interaction frequencies in erythroid or ES cells, with a p-value < 0.01, adjusted for multiple testing.

Tri-C

Tri-C data were analysed using a combination of publicly available tools and customised scripts. Adapter trimming, read reconstruction, *in silico* restriction enzyme digestion, read alignment and PCR duplicate removal were performed as described in the *Capture-C* section. Uniquely aligned reads were analysed with custom scripts to identify captured reads containing the targeted viewpoints of interest and interacting 'reporter' fragments located outside a ~1 kb exclusion zone around the viewpoint restriction fragments. An overview of the numbers of detected reporters is shown in Table 8.1. Reads with two or more reporters were used to calculate multi-way interaction counts between reporter fragments for each viewpoint. These interaction counts were normalised for a total of 1,000,000 interactions on the chromosome analysed. To allow the data to be represented in symmetrical contact matrices, interaction counts were binned in 500 bp bins and corrected for the number of restriction fragments present in each bin. The contact matrices were scaled to 400–1,000 normalised interactions per bin, depending on the viewpoint analysed. To allow for direct comparisons between erythroid and ES cells, equivalent scales were used to display the multi-way interactions with each viewpoint in both cell types.

	Erythroid				ES			
	α: R2	α: HS-39	β: HS2	β: 3'HS1	α: R2	α: HS-39	β: HS2	β: 3'HS1
Reads (total)	1,543,888	1,382,231	1,637,730	3,313,613	1,500,991	1,026,015	949,036	1,398,212
- 1 reporter	806,832	673,174	859,769	1,555,070	847,963	492,204	531,019	652,602
- 2 reporters	514,710	479,530	550,410	1,149,968	473,430	369,812	309,108	505,782
- >2 reporters	222,345	229,527	227,551	608,574	179,598	163,999	108,908	239,828
Reporters (total)	2,55,5425	2,379,973	2,694,822	5,847,628	2,367,219	1,757,001	1,495,350	2,437,140

Table 8.1: Tri-C read statistics. *Overview of the number of reads detected in erythroid and ES samples in the multiplexed Tri-C experiments. Numbers represent unique reads after PCR duplicate removal. The first row shows the total number of reads containing both the viewpoint and at least one reporter fragment. These are subdivided in reads containing one, two or more reporters in the rows below. The bottom row shows the total number of reporters detected.*

Because both globin genes are duplicated, interactions with these genes and their promoters cannot be distinguished bioinformatically. Bowtie maps these duplicated sequences to either of the genes randomly, which explains the high interaction signal between the globin genes in

both erythroid and ES cells. As these signals do not represent reliable interactions, they were not considered in the statistical analyses described below.

Statistical analysis of differential multi-way interactions in ten erythroid and seven ES replicates was performed using the Bioconductor package DESeq2. For these analyses, we considered interactions between all bins on the chromosome containing the viewpoint fragment of interest that contained interaction counts in at least two replicates. However, we would not expect to find chromosome-wide changes in multi-way interactions with *cis*-regulatory elements that normally only interact within a ~100 kb domain. The displayed p-values therefore represent p-values prior to multiple testing adjustment. However, even after adjusting for multiple testing, which is very stringent considering all possible triplets that can be formed between 500 bp bins on an entire chromosome, most reported significant interactions with R2 in the α-globin locus were significant at a level of $p < 0.05$, with the exception of the interactions between R3 and the promoters, and R4 and the promoters, which had an adjusted p-value < 0.1 (Figure 4.6C). All reported significant differential interactions with HS2 in the β-globin locus were also significant at a level of $p < 0.1$ after multiple testing adjustment (Figure 4.8C). In contrast, all enriched interactions in Tri-C matrices from the viewpoint of the CTCF-binding site in both the α- and β-globin locus had adjusted p-values > 0.5 (Figures 4.9C and 4.10C).

To calculate the enrichment of triplets containing two major globin enhancers and one promoter relative to the pair-wise interactions between these elements in both erythroid and ES cells, we counted the number of unique triplets containing all three elements in each replicate. We normalised these counts for a total of 1,000,000 reads containing triplets with the viewpoint of interest and calculated their frequencies relative to the normalised number of triplets containing pair-wise interactions with either of these elements. For example, in the α-globin locus we calculated the normalised R2-R1-Hba triplet counts and expressed these as a proportion of the normalised counts of triplets containing R2 and either R1 or Hba (Figures 4.6D and 4.8D). Using an unpaired, two-tailed t-test, we show that the numbers of R2-R1-Hba triplets in the α-globin locus and HS2-HS1-Hbb triplets in the β-globin locus are significantly enriched in erythroid cells ($p = 0.0002$ and $p < 0.0001$), even after accounting for increased pair-wise interactions between these elements ($p = 0.0005$ and $p < 0.0001$) (Figures 4.6E and 4.8E).

To investigate if the multi-way interactions with enhancers detected by Tri-C in erythroid cells are enriched over the underlying pair-wise contact distribution, we compared these Tri-C

interactions to the pair-wise interactions detected by multiplexed Capture-C experiments. As the resolution of the Capture-C contact matrices is 4 kb, we generated Tri-C matrices at comparable resolution. We normalised these matrices by extracting the regions of interest (α-globin: chr11:32,120,000-32,240,000; β-globin: chr7:110,930,000-111,070,000) and expressing the interactions counts relative to a total of 1,000 interactions in these regions. We calculated the average enrichment of multi-way contacts over the pair-wise distribution and plotted these values in contact matrices. To examine the statistical significance of these enrichments, we compared three biological erythroid replicates and performed multiple unpaired, two-tailed t-tests, with the Holm-Sidak method to correct for multiple comparisons. As the Tri-C data are derived from a specific viewpoint of interest (one vs. all) and the multiplexed Capture-C data represent general interaction frequencies (all vs. all), we excluded the bins containing the R2 or HS2 viewpoint from analysis. Due to strong proximity signals in one vs. all 3C experiments, interactions between elements close to the viewpoint are expected to be enriched in Tri-C. For the bins neighbouring the viewpoint bins, we therefore only considered bins with enrichments above the average enrichment of the entire row/column neighbouring the viewpoint. We also excluded proximity contacts (< 4 kb, i.e. 1 bin form the matrix diagonal) from statistical analysis. We displayed statistically significant enriched interactions with a p-value < 0.01 after adjustment for multiple testing (Figures 4.6F and 4.8F).

To examine whether multi-way interactions in triplets containing two major globin enhancers are more likely to occur with other *cis*-regulatory elements compared to all interactions with a single globin enhancer, we generated double-anchored interaction profiles, which display all fragments interacting in triplets containing two fixed viewpoints. We compared double-anchored R2-R1 interaction profiles in the α-globin locus and HS2-HS1 profiles in the β-globin locus to conventional interaction profiles containing fragments interacting with the R2 or HS2 viewpoint in all triplets. DESeq2 analysis based on raw counts of interacting restriction fragments in each of the ten erythroid samples was used to analyse statistical significance of enriched multi-way interactions. Displayed p-values represent adjusted p-values after adjustment for multiple testing (Figures 4.6G and 4.8G).

C-Trap

The MinION reads were basecalled using Metrichor and converted to FASTA format using Poretools. To select reads that were specifically amplified and of sufficient quality, a BLAST search against a database containing the sequences between the primers and the restriction sites of the targeted fragments was performed. Reads that matched >60% of both primer sequences

were selected using custom scripts. To map the trapped restriction fragments in the reads, a BLAST search against a database containing all the DpnII fragments in the genome was used. Custom scripts were used to iterate through the BLAST results and select the non-overlapping BLAST matches in the reads with a match >60% of fragment length or >100 bp (to avoid a skew towards smaller fragments) in order of significance of the BLAST scores. A summary of the read statistics is shown in Table 8.2.

Rep.	Reads (total)	Reads (QC pass)		Reads (QC fail)		Specific reads	Read length (bp)	Trapped fragments
		Total	Match >60%	Total	Match >60%			
1	59,350	9,902	40%	49,448	7%	7,565	1,445	3.06
2	57,126	12,442	58%	44,684	11%	12,277	1,279	2.77

Table 8.2: C-Trap read statistics. *Overview of the number of reads detected in C-Trap experiments in two replicates (rep.) of primary erythroid cells. The first column shows the number of basecalled reads in each replicate. Approximately half of the reads that passed the Metrichor quality control (QC) matched both primer sequences; these were defined as specific reads and used for analysis. Of the reads that failed QC, only a small percentage could be used for further analysis. The last two columns show the average read length and number of trapped fragments in the specific reads detected.*

Low-input Capture-C

Analysis of low-input Capture-C data was based on the Capture-C data analysis approach described above and adapted for lower cell numbers by implementing an extra stringent PCR duplicate filter, in which all reads with one or two differences in the UMIs were also excluded from analysis. The rationale behind this approach is explained in Chapter 5.

Availability

Custom analysis scripts

Custom analysis scripts have been published on GitHub. The scripts for analysis of (low-input) Capture-C and Tri-C experiments are available in the following repositories.

- https://github.com/oudelaar/CaptureC/
- https://github.com/oudelaar/TriC/

Sequencing data

Sequencing data have been deposited in the Gene Expression Omnibus. Hi-C, wild type Capture-C and Tri-C data are available in SuperSeries GSE107940. Capture-C data in the enhancer knockout models, the D3839 model and in cells treated with DRB are available in SuperSeries GSE78835, Series GSE97871 and Series GSE107675 respectively. Low-Input Capture-C data are available in Series GSE99257.

UCSC Genome Browser Track Hubs

Analysed Capture-C data from all the viewpoints in the α- and β-globin loci and low-input Capture-C data in the α-globin locus are available in UCSC Genome Browser Track Hubs and can be accessed with the following links.

- http://sara.molbiol.ox.ac.uk/public/hugheslab/CaptureC_alpha_globin/hub.txt
- http://sara.molbiol.ox.ac.uk/public/hugheslab/CaptureC_beta_globin/hub.txt
- http://sara.molbiol.ox.ac.uk/public/hugheslab/LowInputCaptureCHub/hub.txt

References

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Alipour, E., and Marko, J.F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Research *40*, 11202–11212.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410.

Anguita, E., Sharpe, J.A., Sloane-Stanley, J.A., Tufarelli, C., Higgs, D.R., and Wood, W.G. (2002). Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. Blood *100*, 3450–3456.

Ay, F., Vu, T.H., Zeitz, M.J., Varoquaux, N., Carette, J.E., Vert, J.-P., Hoffman, A.R., and Noble, W.S. (2015). Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C. BMC Genomics *16*, 121.

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

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Dostie, and Pombo, A. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. Proc Natl Acad Sci USA *109*, 16173–16178.

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., de Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *109*, 1427.

Baù, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol. *18*, 107–114.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Bender, M.A., Ragoczy, T., Lee, J., Byron, R., Telling, A., Dean, A., and Groudine, M. (2012). The hypersensitive sites of the murine β-globin locus control region act independently to affect nuclear localization and transcriptional elongation. Blood *119*, 3820–3827.

Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? Transcription *2*, 103–108.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., et al. (2013). An estimation of the number of cells in the human body. Ann. Hum. Biol. *40*, 463–471.

Bickmore, W.A. (2013). The spatial organization of the human genome. Annu Rev Genomics Hum Genet *14*, 67–84.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics *17*, 772–772.

Brown, J.M., Roberts, N.A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J.M., De Ornellas, S., Oudelaar, M., Szczerbal, I., Babbs, C., et al. (2017). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. BioRxiv, doi: 10.1101/234427.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth *10*, 1213–1218.

Busslinger, G.A., Stocsits, R.R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.-M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature *544*, 503–507.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Chen, N. (2004). WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Research *33*, D383–D389.

Chung, J.H., Bell, A.C., and Felsenfeld, G. (1997). Characterization of the chicken beta-globin insulator. Proc Natl Acad Sci USA *94*, 575–580.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Reviews Genetics *2*, 292–301.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research *19*, 24–32.

Curtin, P., Pirastu, M., Kan, Y.W., Gobert-Jones, J.A., Stephens, A.D., and Lehmann, H. (1985). A distant gene deletion affects beta-globin gene function in an atypical gamma delta beta-thalassemia. J. Clin. Invest. *76*, 1554–1558.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., Shamim, M., et al. (2016). Deletion of DXZ4on the human inactive X chromosome alters higher-order genome architecture. Proc Natl Acad Sci USA *113*, E4504–E4512.

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nat Meth *14*, 125–134.

Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2015). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat Meth *86*, 1202–1210.

De Gobbi, M., Anguita, E., Hughes, J., Sloane-Stanley, J.A., Sharpe, J.A., Koch, C.M., Dunham, I., Gibbons, R.J., Wood, W.G., and Higgs, D.R. (2007). Tissue-specific histone modification and transcription factor binding in globin gene expression. Blood *110*, 4503–4510.

de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes & Development *26*, 11–24.

de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Molecular Cell *60*, 676–684.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell *159*, 374–387.

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

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

Fabre, P.J., Leleu, M., Mormann, B.H., Lopez-Delisle, L., Noordermeer, D., Beccari, L., and Duboule, D. (2017). Large scale genomic reorganization of topological domains at the HoxD locus. Genome Biol. *18*, 149.

Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S.V., Abdennur, N., Razin, S.V., Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature *544*, 110–114.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat Meth *14*, 679–685.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., and Stern, D.L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature *466*, 490–493.

Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., and Mirny, L.A. (2018). Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harbor Symposia on Quantitative Biology 034710.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. CellReports *15*, 2038–2049.

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. Science *360*, 102–105.

Gavrilov, A.A., Chetverina, H.V., Chermnykh, E.S., Razin, S.V., and Chetverin, A.B. (2014). Quantitative analysis of genomic element interactions by molecular colony technique. Nucleic Acids Research *42*, e36–e36.

Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Molecular Cell *49*, 773–782.

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950–963.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575–579.

Gómez-Marín, C., Tena, J.J., Acemel, R.D., López-Mayorga, M., Naranjo, S., la Calle-Mustienes, de, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., et al. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc. Natl. Acad. Sci. U.S.a. *112*, 7542–7547.

Green, B., Bouchier, C., Fairhead, C., Craig, N.L., and Cormack, B.P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA *3*, 3.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *74*, 595–413.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet *43*, 630–638.

Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Elife *6*, e25776.

Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M., Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R., et al. (2017). Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nature Cell Biology *19*, 952–961.

Hatton, C.S., Wilkie, A.O., Drysdale, H.C., Wood, W.G., Vickers, M.A., Sharpe, J., Ayyub, H., Pretorius, I.M., Buckle, V.J., and Higgs, D.R. (1990). Alpha-thalassemia caused by a large (62 kb) deletion upstream of the human alpha globin gene cluster. Blood *76*, 221–227.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat Genet *48*, 895–903.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

Higgs, D.R., Vernimmen, D., and Wood, B. (2008). Long-range regulation of alpha-globin gene expression. Adv. Genet. *61*, 143–173.

Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., and Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. CellReports *2*, 345–357.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Hughes, J.R., Cheng, J.-F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E., and Higgs, D.R. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc Natl Acad Sci USA *102*, 9830–9835.

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet *46*, 205–212.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat Meth *12*, 351–356.

Jiang, T., Raviram, R., Snetkova, V., Rocha, P.P., Proudhon, C., Badri, S., Bonneau, R., Skok, J.A., and Kluger, Y. (2016). Identification of multi-loci hubs from 4C-seq demonstrates the functional importance of simultaneous interactions. Nucleic Acids Research *44*, 8714–8725.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kellum, R., and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell *64*, 941–950.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, A.D. (2002). The Human Genome Browser at UCSC. Genome Research *12*, 996–1006.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Research *12*, 656–664.

Kioussis, D., Vanin, E., deLange, T., Flavell, R.A., and Grosveld, F.G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature *306*, 662–666.

Kleinjan, D.-J., and Coutinho, P. (2009). Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic *8*, 317–332.

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

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. Molecular Cell *45*, 447–458.

Kruse, K., Hug, C.B., Hernández-Rodríguez, B., and Vaquerizas, J.M. (2016). TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics *32*, 3190–3192.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.

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

Lee, H.K., Willi, M., Wang, C., Yang, C.M., Smith, H.E., Liu, C., and Hennighausen, L. (2017). Functional assessment of CTCF sites at cytokine-sensing mammary enhancers using CRISPR/Cas9 gene editing in mice. Nucleic Acids Research *45*, 4606–4618.

Lee, K., Hsiung, C.C.S., Huang, P., Raj, A., and Blobel, G.A. (2015). Dynamic enhancer–gene body contacts during transcription elongation. Genes & Development *29*, 1992–1997.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Liu, Z., Scannell, D.R., Eisen, M.B., and Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell *146*, 720–731.

Loman, N.J., and Quinlan, A.R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics *30*, 3399–3401.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564–567.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell *161*, 1012–1025.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Madrigal, P. (2015). On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. Front Bioeng Biotechnol *3*, 144.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Maxam, A.M., Tizard, R., Skryabin, K.G., and Gilbert, W. (1977). Promoter region for yeast 5S ribosomal RNA. Nature *267*, 643–645.

Merkenschlager, M., and Nora, E.P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Http://Dx.Doi.org/10.1146/Annurev-Genom-083115-022339 *17*, 17–43.

Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci USA *107*, 22534–22539.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature *547*, 61–67.

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E.O., and Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes & Development *30*, 2657–2662.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017–1021.

Nasmyth, K. (2001). Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. Annual Review of Genetics *35*, 673–745.

Nicholls, R.D., Fischel-Ghodsian, N., and Higgs, D.R. (1987). Recombination at the human alpha-globin gene cluster: sequence features and topological constraints. Cell *49*, 369–378.

Nicodemi, M., and Pombo, A. (2014). Models of chromosome structure. Current Opinion in Cell Biology *28*, 90–95.

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., and Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. Nat Meth *6*, 917–922.

Noordermeer, D., and de Laat, W. (2008). Joining the loops: beta-globin gene regulation. IUBMB Life *60*, 824–833.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L. (2018). Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation. Biophys. J. *114*, 30a.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296–300.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics *15*, 234–246.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239–243.

Oudelaar, A.M., Downes, D., Davies, J., and Hughes, J. (2017). Low-input Capture-C: A Chromosome Conformation Capture Assay to Analyze Chromatin Architecture in Small Numbers of Cells. Bio-Protocol *7*.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. Cell *132*, 422–433.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular Cell *38*, 603–613.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. *20*, 1562–1567.

Pertea, M., and Salzberg, S.L. (2010). Between a chicken and a grape: estimating the number of human genes. Genome Biol. *11*, 206.

Philipsen, S., and Hardison, R.C. (2017). Evolution of hemoglobin loci and their regulatory elements. Blood Cells, Molecules and Diseases 1–0.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. Cell *153*, 1281–1295.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305–320.e324.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Rhodes, J., Mazza, D., Nasmyth, K., and Uphoff, S. (2017). Scc2/Nipbl hops between chromosomal cohesin rings after loading. Elife *6*, 11202.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc Natl Acad Sci USA *105*, 8309–8314.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 129–143.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U.S.a. *112*, E6456–E6465.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. CellReports *17*, 2042–2059.

Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., Izraeli, S., and Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. Nat Meth *13*, 685–691.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., H Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Research *23*, 2066–2077.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet *38*, 1348–1354.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. Embo J. *32*, 3119–3129.

Spivak, J.L., Toretti, D., and Dickerman, H.W. (1973). Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. Blood *42*, 257–266.

Stedman, W., Kang, H., Lin, S., Kissil, J.L., Bartolomei, M.S., and Lieberman, P.M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. Embo J. *27*, 654–666.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., and Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev. Cell *39*, 529–543.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Molecular Cell *10*, 1453–1465.

van de Werken, H.J.G., de Vree, P.J.P., Splinter, E., Holwerda, S.J.B., Klous, P., de Wit, E., and de Laat, W. (2012). 4C technology: protocols and data analysis. Meth. Enzymol. *513*, 89–112.

Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. Embo J. *26*, 2041–2051.

Vernimmen, D., Marques-Kranc, F., Sharpe, J.A., Sloane-Stanley, J.A., Wood, W.G., Wallace, H.A.C., Smith, A.J.H., and Higgs, D.R. (2009). Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). Blood *114*, 4253–4260.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports *10*, 1297–1309.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

Weth, O., Paprotka, C., Günther, K., Schulte, A., Baierl, M., Leers, J., Galjart, N., and Renkawitz, R. (2014). CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. Nucleic Acids Research *42*, 11941–11951.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J., et al. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. Embo J. *36*, 3573–3599.

Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity. Molecular and Cellular Biology *31*, 2174–2183.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. *7*, 203–214.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet *38*, 1341–1347.

Zhou, G.-L., Xin, L., Song, W., Di, L.-J., Liu, G., Wu, X.-S., Liu, D.-P., and Liang, C.-C. (2006). Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes. Molecular and Cellular Biology *26*, 5096–5105.

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci USA *111*, 996–1001.