New approaches to the investigation of the interactions between enhancers and promoters during haemopoiesis

Thesis submitted for the Degree of Doctor of Philosophy (DPhil)



Dr James Davies

BM BCh, BA, MRCP, FRCPath, EDIC

Wellcome Trust Clinical Research Training Fellow

MRC Molecular Haematology Unit

Weatherall Institute of Molecular Medicine

Trinity College

University of Oxford

Declaration

I, James Davies confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to thank Prof D. Higgs and Prof J. Hughes for excellent supervision during this project.

I would also like to thank the Wellcome Trust for funding this project through the Wellcome Trust Clinical Research Training Fellowship scheme (098931/Z/12/Z).

Dedication

This thesis is dedicated to my father Hugh Davies, who passed away whilst I was writing this manuscript and is greatly missed.

Abstract

Next generation sequencing based methods allow us to identify active genes and potential regulatory elements rapidly across the genome, but an outstanding challenge is to unravel which regulatory elements control which genes. This is problematic because regulatory elements control genes over huge distances (over 1 million base pairs) and they have an unpredictable distribution around the genes they influence. In order to enhance transcription the protein complexes at distal regulatory elements make physical contact with the promoter. These interactions can be detected using Chromosome Conformation Capture (3C) technology and the position of regulatory elements can be deduced from these data. However, these methods are either low throughput or low resolution and they are prone to bias. In this 3C techniques were specifically developed to determine the interactions between regulatory elements and gene promoters promoter. The focus has been the development of Next Generation Capture-C, which allows many genetic loci and samples to be analysed simultaneously, with greater sensitivity and accuracy than has previously been possible. High resolution data can be produced from as few as 100,000 cells, and single-nucleotide polymorphisms can be used to generate allele-specific tracks. Nanostring technology has also been developed for analysis of 3C libraries, as this allows interaction profiles to be determined without PCR or sequencing bias. The Nanostring data show remarkable correlation with the interaction profiles generated by NG Capture-C.

Table of Contents

Declaration 2
Acknowledgements 2
Dedication 2
Abstract 3
Table of Contents 4
Figures 9
Publications 12
Publically Submitted Datasets 13
Abbreviations 14
1 Introduction 17
1.1 Elements controlling gene expression 17
1.2 Nuclear organization 21
1.3 Chromosome conformation capture technology 22
1.3.1 3C library preparation 25
1.3.2 Different methods for analysis of 3C libraries 26
1.3.2.1 Chromosome conformation capture (3C) 30
1.3.2.2 Circular chromosome conformation capture (4C) 30
1.3.2.3 Chromosome Conformation Capture Carbon Copy (5C) 32
1.3.2.4 Hi-C 33
1.3.2.5 Cavalli Hi-C 34
1.3.2.6 Single Cell Hi-C 34
1.3.2.7 Capture-C 35
1.3.2.8 Micro-C 35
1.3.2.9 Capture Hi-C 36
1.3.2.10 T2C 36
1.3.2.11 ChIA-PET 37
1.4 Thesis overview 37
2 Development of the Next Generation Capture-C assay 40
2.1 Rationale 40
2.2 Overview of the method 41
2.3 Improvements to the method 43

2.3.1 Optimisation of oligonucleotide design....43
2.4 Introduction of double capture step....44
2.4.1 Comparison of double capture with single capture....44
2.4.2 Reproducibility of data....46
2.5 Comparison of NG Capture-C with other methods....48
2.6 Detailed methods....53
2.6.1 3C library preparation....53
2.6.2 Quality assurance of 3C libraries....53
2.6.2.1 Agaraose gel electrophoresis of 3C libraries....53
2.6.2.2 Real time PCR assessment of digestion efficiency....54
2.6.3 Discrepancies between digestion efficiencies....55
2.7 Addition of Illumina sequencing adaptors....55
2.7.1 Sonication....57
2.7.2 Ligation of adaptors....57
2.7.3 LM-PCR for addition of Truseq adaptors....58
2.7.4 Reconstitution of capture oligonucleotides....58
2.7.5 Hybridisation with capture oligonucleotides....58
2.7.6 Extraction of captured DNA....59
3 Data analysis....60
3.1 Outline of data analysis.....60
3.1.1 Reconstruction of reads....61
3.1.2 In silico restriction enzyme digestion....62
3.1.3 Alignment to the genome....63
3.1.4 Main analysis script....63
3.1.5 Downstream processing....66
3.1.6 Normalisation of data....66
3.1.7 DESeq2 analysis....67
3.1.8 Circos....67
4 Multiplexed Capture-C....68
4.1 Identification of tissue-specific regulatory elements by direct comparison of chromosome conformation in different cell types....68
4.2 Identification of erythroid specific regulatory elements using ES cells as a control....69

4.2.1 Interaction with the gene body on activation of transcription 71
4.3 Identification of ES cell specific regulatory elements 72
4.3.1 Identification of active elements in both cell types simultaneously 81
4.4 Conclusions 85
5 Analysis of long range cis and trans interactions 86
5.1 NG Capture-C defines previously unidentified interactions at the alpha globin locus 86
5.2 Hi-C data at the alpha globin locus 88
5.2.1 Possible functional effects of long range interactions in *cis* 89
5.3 Long range interactions are present in *cis* 89
5.4 *Trans* interactions 93
5.5 Conclusions 99
6 Capture-C for reduced cell numbers 101
6.1 Aims 101
6.2 Adaptation of the NG Capture-C protocol for reduced cell numbers 101
6.2.1 3C library preparation 101
6.2.2 Library quality assurance 104
6.2.3 Ligation of adaptors and addition of indices 105
6.2.4 Oligonucleotide capture and sequencing 106
6.3 Results 106
6.3.1 Limitations from using reduced cell numbers 108
6.4 Determination of the interaction profile during cellular differentiation. 109
6.4.1 Sorting of cells at different stages of erythroid differentiation. 109
6.4.2 Capture-C on cells from different stages of differentiation 110
6.5 Conclusions 112
7 Allele Specific Capture-C 114
7.1 Production of allele specific tracks 114
7.1.1 Analysis of the phase of SNPs in allele specific tracks 116
7.1.2 Generation of separate tracks from Hba-a1 and 2 118
7.2 Conclusions 120
8 Analysis of a single locus from multiple viewpoints by Capture-C 121
8.1 Aims 121

8.2 Experimental methods 121
8.3 Capture-C from the alpha globin enhancers 122
8.4 Capture-C from the CTCF sites within the alpha globin locus 123
8.5 Capture-C on alpha globin enhancer knockout mice 126
8.5.1 Methods 126
8.5.2 Results 126
8.5.3 Conclusions 132
9 Development of the use of NanoString technology to analyse 3C libraries 135
9.1 Aims 135
9.2 Overview of NanoString technology 135
9.3 Adaption of NanoString technology for analysis of 3C libraries 136
9.3.1 Experimental design 136
9.3.2 Sensitivity of the Nanostring ncounter system 136
9.3.3 Preparation of 3C libraries 137
9.3.4 Hybridisation with NanoString probes 137
9.3.5 Effects of impurities in 3C libraries 137
9.4 Results 139
9.4.1 NanoString can determine an accurate interaction profile 139
9.4.2 Comparison of C-NanoString to Capture-C 141
9.4.3 Conclusions 145
10 C-trap and C-minion 147
10.1 Aims 147
10.2 C PCR-trap 147
10.2.1 Methods 147
10.2.2 Results 152
10.3 C-MinION 155
10.3.1 Aim 155
10.3.2 Methods 155
10.3.3 Data analysis 156
10.3.4 Results 157
10.4 Conclusions 161
11 Discussion 163
11.1 Technical discussion 163

11.1.1 Comparison of 3C based methods 163
11.1.1.1 Resolution 163
11.1.1.2 Sensitivity 166
11.1.1.3 Biases 167
11.1.1.4 Multiplexed detection of interactions 168
11.1.1.5 Multiplexing Samples 170
11.1.2 Limitations of Capture-C 172
11.1.3 Limitations of C-nanostring 173
11.2 General Discussion 173
11.2.1 Insights gained from Next Generation Capture-C 173
11.2.2 NG Capture-C suggests that there is cooperation between enhancer elements 174
11.2.3 Interaction profiles from different viewpoints within the alpha globin locus revels different compartments of interaction 175
11.2.4 Locus control regions (or super-enhancers) may evolve due to enhancer driven compartments of interaction with gene promoters 176
11.2.5 Distal enhancer positioning may occur to allow separate evolution of gene regulation from gene sequence 176
12 References 180

Figures

Figure 1.1 3C strategy 24
Figure 1.2 Probability distribution of ligation events 25
Figure 1.3 Comparison of different 3C based methods 28
Figure 2.1 Overview of the experimental workflow 43
Figure 2.2 Capture oligonucleotide design 43
Figure 2.3 Comparison of single and double oligonucleotide capture 45
Figure 2.4 Reproducibility of NG-Capture-C 47
Figure 2.5 Comparison with 4C-seq 50
Figure 2.6 Comparison of Capture-C data mapped to normal and maximum resolution 52
Figure 2.7 Gel electrophoresis to assess digestion efficiency 54
Figure 2.8 Schema of primers used to assess digestion efficiency 54
Figure 3.1 Outline of initial data analysis 61
Figure 3.2 Script flow diagram (CCanalyser2.pl) 65
Figure 4.1 Capture-C at the beta globin locus 69
Figure 4.2 Capture-C from the mitoferrin 1 (Slc25A37) gene promoter 70
Figure 4.3 Capture-C data from the Cd47 promoter 71
Figure 4.4 Capture-C at the Myc locus. 74
Figure 4.5 Capture-C at the Klf4 locus in ES and erythroid cells. 76
Figure 4.6 Capture-C from the Sox2 promoter in ES and erythroid cells. 77
Figure 4.7 Capture-C from the Pou5f1 (Oct4) promoter. 79
Figure 4.8 Capture-C from the promoter of Nanog 81
Figure 4.9 Capture-C at the Pnpo (pyridoxine 5'-phosphate oxidase) locus. 82
Figure 4.10 Capture-C from the Tal1 gene in ES and erythroid cells. 84
Figure 5.1 Interactions between alpha globin and surrounding genes 87
Figure 5.2 Hi-C data from erythroid cells at the alpha globin locus 89
Figure 5.3 Comparison of local with long range cis and trans interactions 90
Figure 5.4 Capture-C from the alpha globin promoters - whole chromosome in cis 92
Figure 5.5 Cis and trans interactions from alpha globin, beta globin and mitoferrin 94
Figure 5.6 Cis and trans interactions from 30 gene promoters 96

Figure 5.7 Analysis of trans interactions from the alpha globin (Hba-a1&2) promoters....98
Figure 6.1 Mass of DNA in the 3C library from different cell numbers....104
Figure 6.2 Assessment of digestion efficiency with reduced cell numbers....105
Figure 6.3 Robust interaction profiles generated from small numbers of cells.107
Figure 6.4 Comparison of NG Capture-C with different numbers of cells with other methods....108
Figure 6.5 Gating strategy for sorting early and intermediate erythroblasts.....110
Figure 6.6 Changes in the Capture-C profile during erythroid differentiation...111
Figure 7.1 SNP specific interaction profiles....115
Figure 7.2 Allele specific tracks at the Tal-1 locus.....116
Figure 7.3 Phase of SNPs with the viewpoint fragment.....117
Figure 7.4 Specific interaction profiles for *Hba-a1* and *Hba-a2* parologous genes generated using SNP specific Capture-C.....119
Figure 8.1 Interaction profile from the alpha globin enhancers....123
Figure 8.2 Capture-C from multiple viewpoints within the alpha globin locus..124
Figure 8.3 Effects of enhancer knock-out on the Capture-C interaction profile 128
Figure 8.4 Statistical analysis of the Enhancer KO datasets....131
Figure 8.5 Active chromatin hub model....132
Figure 8.6 Domain model of interactions around the alpha globin locus....133
Figure 9.1 Effects of contaminants in the 3C library on NanoString counts....138
Figure 9.2 Replicates of C-nanostring from erythroid cells....139
Figure 9.3 Replicates of C-nanostring from ES cells....139
Figure 9.4 Combined strip chart and barograph of the relationship between nCounter count and amount of input 3C library....140
Figure 9.5 Scatter plot comparing when 2μg and 5μg input 3C library are used 141
Figure 9.6 Statistical analysis of Erythroid and ES cell replicates....142
Figure 9.7 Comparison of C-NanoString with Capture-C and 4-C seq....143
Figure 9.8 Comparison between Capture-C and C-NanoString....144
Figure 9.9 Comparison between Capture-C and C-NanoString....145
Figure 10.1 C-PCR trap experimental design....148

Figure 10.2 Agarose gel electrophoresis of PCR products from C-PCRtrap experiments....151
Figure 10.3 C-PCR trap results....153
Figure 10.4 Sequencing of 2D reads by MinION....156
Figure 10.5 UCSC genome browser MinION profile at the alpha globin locus....158
Figure 10.6 Looping Glyph of MinION data (Washington University Epigenome Browser)....159
Figure 10.7 Map of reconstructed MinION reads....160
Figure 11.1 Comparison of sensitivity and resolution between techniques....171

Publications

Davies J.O.J., Oudelaar A.M., Higgs D.R., Hughes J.R. 'How best to identify chromosomal interactions: a comparison of approaches', Nature Methods 2017, 14 (2), 125-134

Davies J.O.J., Telenius J.M., McGowan S., Roberts N.A., Taylor S., Higgs D.R. and Hughes J.R. 'Multiplexed analysis of chromosome conformation at vastly improved sensitivity', Nature Methods 2016; 13, 74-80

Hay D., Hughes J.R., Babbs C., Davies J.O.J., Graham B.J., Hanssen L., T. Kassouf M.T, Oudelaar A.M., Sharpe J.A., Suciu M.C., Telenius J., Williams R., Rode C., Li P., Pennacchio L.A., Sauka-Spengler T., Sloane-Stanley J.A., Ayyub H., Butler S., Gibbons R.J., Smith A.J.H., Wood W.G., Higgs D.R. 'Testing the Super-Enhancer Concept by *In-Vivo* Dissection', Nature Genetics 2016, 48 (8), 895-903

Paralkar V.R., Taborda C.C., Huang P., Yao Y., Kossenkov A.V., Prasad R., Luan J., Davies J.O.J., Hughes J.R., Hardison R.C., Blobel G.A., Weiss M.J. 'Unlinking an lncRNA from Its Associated *cis* Element' Molecular Cell. 2016 Apr 7;62(1):104-10

Brackley C.A., Brown J.M., Waithe D., Babbs C., Davies J.O.J., Hughes J.R., Buckle V.J., Marenduzzo D. 'Predicting the three-dimensional folding of *cis*-regulatory regions in mammalian genomes using bioinformatic data and polymer models', Genome Biology 2016 Mar 31;17(1):59

Platt J.L., Salama R., Smythies J., Choudhry H., Davies J.O.J., Hughes J.R., Ratcliffe P.J., Mole D.R. 'Capture-C reveals preformed chromatin interactions between HIF-binding sites and distant promoters', EMBO reports (2016) e201642198

Ooi W.F., Xing M., Xu C., Lin S.J., Yao X., Qamra A., Irwanto A., Nandi T., Lim K., Lee-Lim A.P., Chan Y.S., Tay S.T., Lee M., Davies J., Wong W.K., Soo K., Chan W.H., Ong H.S., Chow P., Wong C.Y., Rha S.Y., Liu J.P., Hillmer A., Hughes J., Rozen S., Teh B.T., Ramlee M., Lim M.C., Cao F., Babu D., Qu Z.J., Fullwood M., Li S., Poon L.

'Epigenomic Profiling of Primary Gastric Adenocarcinoma Reveals Super Enhancer Heterogeneity', Nature Communications (2016), 7, 12983

Publically Submitted Datasets

Multiplexed analysis of chromosome conformation at vastly improved sensitivity NCBI Gene Expression Omnibus dataset GSE67959. (GSM1659659-1659674)

Prizes

Radcliffe Department of Medicine Graduate Prize (Oxford University, 2016)
Medical Sciences Division DPhil prize (2nd place) (Oxford University, 2015)

Presentations

Davies J.O.J. 'Novel methods for determining physical interactions in chromatin' Radcliffe Department of Medicine Day Presentation (for RDM Graduate Prize), 2016

Davies J.O.J. 'Next Generation Capture-C: a highly sensitive and scalable method for multiplexed definition of chromatin structure' Radcliffe Department of Medicine DPhil Day – 2nd Prize, 2015

Davies J.O.J. 'Nextgen Capture-C: a highly sensitive technique for determining chromatin interactions' NGS special interest group, 2015

Davies J.O.J., Hughes J.R., 'Chromosome conformation capture breakout session' Epigenesys meeting, 2014

Davies J.O.J., 'High resolution analysis of regulatory chromosomal landscapes' Weatherall Institute of Molecular Medicine Day, 2014

Davies J.O.J, Hughes J.R., Roberts N.A., Hay D., Lynch M., McGown S., Taylor S. and Higgs D.R. 'Linking promoters with their regulatory elements during erythropoiesis'. EMBO nuclear structure and dynamics meeting, 2013.

Abbreviations

2D reads	2 dimensional reads - reads in which both the forward and reverse strands were sequenced sequentially from the same molecule
3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
5C	Carbon copy chromosome conformation capture
ATAC-seq	Assay for transposase-accessible chromatin with high throughput sequencing
ChIP-seq	Chromatin immunoprecipitation analysed by DNA sequencing
Cot DNA	Repetitive DNA used to block non-specific hybridisation from repetitive sequences
CTCF	CCCTC-binding factor
DNA	Deoxyribonucleic acid
DNAse-seq	DNAseI assay analysed by sequencing (for mapping DNAseI hypersensitive sites)
DNAseI	Deoxyribonuclease I (an endonuclease that non specifically cleaves DNA)
DPE	Downstream promoter element
EDTA	Ethylenediaminetetraacetic acid
ENCODE project	Encyclopedia of DNA elements project
ES	Embryonic stem (cell)
F1 (cross)	Filial 1 hybrid
FACS	Flow cytometry assisted cell sorting
FCS	Fetal calf serum
FISH	Fluorescence in situ hybridization
FLC	Fetal liver culture

GTF	General transcription factor
GWAS	Genome wide association study
H3K4/K27	Histone H3 lysine position 4 / 27
Hi-C	Chromosome conformation capture based technique that uses a biotin fill in to selectively sequence ligation junctions genome wide
HS-12	Hypersensitive site -12kb - 5th mouse specific enhancer of the mouse alpha globin locus
KO	Knock out
LCR	Locus control region
MACS	Magnetic activated cell sorting
MCSR	Multispecies conserved sequence regulatory element
MTE	Motif ten element
ncRNA	non coding RNA
NG Capture-C	Next generation Capture-C
NF-Y	Nuclear transcription factor Y
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PERL	Practical extraction and reporting language - programming language developed by Larry Wall in 1987
Pol II	RNA polymerase II
R	Programming language for statistical computing
R1	(Multispecies conserved sequence) Regulatory element 1 at the alpha globin locus
R2	(Multispecies conserved sequence) Regulatory element 2 at the alpha globin locus
R3	(Multispecies conserved sequence) Regulatory element 3 at the alpha globin locus
R4	(Multispecies conserved sequence) Regulatory element 4 at the alpha globin locus

RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
RPMI	Roswell Park Memorial Institute (cell culture media)
SAM	Sequence alignment map (a standard output file format for genome alignment)
SD	Standard deviation
SNP	Single nucleotide polymorphism
Sp1	Specificity 1 transcription factor
SPRI	Solid phase reversible immobilisation
SV40	Simian vacuolating virus 40
TAD	Topologically associated domain
TAF	TATA box binding protein associated factors
TALE	Transcription activator-like effector
TBE	Tris/Borate/EDTA
TBP	TATA box binding protein
Ter119	Antibody useful for distinguishing murine erythrocytes
Th	T-helper
TSS	Transcription start site
WT	Wild type

1 Introduction

Progress in our ability to annotate regulatory elements in the genome and determine their potential function has been driven by technological advances, such as RNA-seq (Wang et al., 2009), ChIP-seq (Mikkelsen et al., 2007; Robertson et al., 2007), DNase-seq (Hesselberth et al., 2009) and ATAC-seq (Buenrostro et al., 2013). Using these techniques it has become relatively straightforward to identify regulatory elements genome wide. However, an outstanding challenge is to understand the mechanisms by which regulatory elements control specific gene promoters at a distance (10s to 1000s kb)(Stamatoyannopoulos, 2016). It is thought that physical interactions are required between the macromolecular complexes at distal regulatory elements and gene promoters in order to for gene expression to be modulated. These physical contacts between different sections of chromatin can be detected using chromosome conformation capture (3C) based assays. Using these techniques it is possible to analyse in detail the interactions between enhancers, silencers, boundary elements and promoters at individual loci at high resolution (de Laat and Duboule, 2013; Dekker et al., 2002; Noordermeer et al., 2011b; Sanyal et al., 2012; Tolhuis et al., 2002; van de Werken et al., 2012b). Recently, our group have shown, using a high-throughput approach (Capture-C), that it is possible to interrogate *cis*-interactions, at hundreds of loci at high resolution in a single experiment (Hughes et al., 2014). Such approaches are of immediate value in defining interactions within the chromosomal landscapes of many loci, and identifying the genes and the functional effects of SNPs that are associated with complex diseases, the majority of which lie in intergenic *cis*-acting regulatory elements (Maurano et al., 2012; Parker et al., 2013; Pasquali et al., 2014).

1.1 Elements controlling gene expression

The complex patterns of tissue specific gene expression seen in higher eukaryotes are in part attributable to the mechanisms by which genes are regulated during differentiation and development. The key elements controlling gene expression are promoters, which tend to be bound by widely expressed transcription factors and lie close to the structural gene and distal regulatory

elements, such as enhancers. In contrast to promoters, enhancers can lie more than 1 Mb up or downstream from the gene body (Lettice et al., 2003) and they tend to be bound by tissue specific transcription factors (Kleinjan and van Heyningen, 2005). A third class of elements, typified by CTCF binding sites, that control genome architecture are also likely to be of fundamental importance in controlling gene expression.

Promoters are typically located directly upstream of the 5' transcription start site and are an important site of RNA polymerase and transcription factor binding. Active promoters are associated with specific chromatin signatures including nucleosome depletion and increased chromatin accessibility as determined by DNAseI hypersensitivity and ATAC-seq. Specific histone modifications are also associated with active promoters including H3K4 di- and tri- methylation and acetylation as well as H3K27 acetylation.

Although the structure of promoters is complex and the classification of promoters can be unclear, two main types of core promoter can be defined in metazoans, based on the distribution of TSSs (Danino et al., 2015). Focused promoters tend to be located between -40 and +40 nucleotides from the TSS and often contain elements the have a positional bias such as the TATA box, Initiator, Motif Ten Element (MTE) and Downstream Promoter Element (DPE) (Rach et al., 2009). By contrast, broad promoters contain multiple weak transcription start sites spread over 50 to 100 nucleotides (Juven-Gershon and Kadonaga, 2010). In vertebrates, these promoters are usually associated with CpG islands, Sp1 and NF-Y sites. Although focused initiation of transcription is found in all organisms, it is only the predominant mechanism in simpler organisms. In vertebrates over 70% of the promoters are broad and these CpG island associated promoters appear to evolve more rapidly than TATA box associated promoters (Carninci et al., 2006).

CpG island promoters often display bidirectional promoter activity (Trinklein et al., 2004). Bidirectional promoters were originally described where two protein coding genes, generate two closely spaced transcripts in both the sense and antisense orientations; allowing coordinated, cell-type specific transcription of the pair of genes. Subsequently, a large number of ncRNA transcripts have been described at mammalian promoters. If these are taken into account the

prevalence of bidirectional promoters rises from an initial estimate of around 10% to nearly 80% (Core et al., 2014; Core et al., 2008). However, the function of these rapidly decaying short transcripts remains elusive (Duttke et al., 2015).

Promoters are composed of short DNA sequences often referred to as core promoter motifs. There are no core promoter elements that are used ubiquitously, but generally these serve as binding sites for the basal transcription machinery, which recruits Pol II is recruited to the promoter. There is great diversity in the make up of the basal transcription machinery at different promoters. However, the general transcription factors (GTFs) TFIIB and TFIID are important components. The TFIID complex has considerable flexibility in it's makeup but this is classically nucleated by TATA box binding protein (TBP) and other modules are usually involved, such as the TBP-associated factors (TAFs), which bind specific core promoter elements (TAF1/2 bind the Inr, for example). It is likely that the core composition of the promoter is important in determining how enhancers can influence gene expression (Zabidi et al., 2015).

'Enhancers' were first described in the early 1980s. The term was initially used to describe the marked increase in expression of a cloned beta globin gene in HeLa cells induced by a repetitive sequence from the SV40 virus (Banerji et al., 1981). For the effect to be seen the SV40 elements had to be in *cis*, although they could be distant from the promoter, and it was independent of the relative orientation of the sequences. Subsequent studies showed that enhancers usually contain multiple transcription factor binding sites and that enhancer activation is associated with DNAseI hypersensitivity and modification of neighboring histones. In particular, H3K27 acetylation and H3K4Me1 in combination correlate with enhancers near active genes. When H3K4Me1 is found without H3K27ac it is thought to be associated with poised enhancers (Creyghton et al., 2010) and in when this occurs H3K27 is often trimethylated to a low degree (Bulger and Groudine, 2011). Active enhancers also tend to be bound by P300, which can acetylate H3K27, and other, tissue specific, transcription factors.

There are several diseases associated with deletion of non-coding regions, which contain enhancers including: alpha thalassaemia (HBA)(Higgs and Wood, 2008); beta thalassemia (HBB) (Kioussis et al., 1983); campomelic dysplasia (SOX9)

(Bagheri-Fam et al., 2001); preaxial polydactyly (SHH) (Lettice et al., 2003) and X-linked deafness (POU3F4) (de Kok et al., 1996). This is likely to be a significant underestimate of the disease burden resulting from polymorphisms in enhancers because over 90% of disease associated single nucleotide polymorphisms (SNPs) in genome wide association studies (GWAS) (Maurano et al., 2012) are in non-coding DNA and are thought likely to effect enhancer sequences. In addition the overwhelming majority of mutations in cancer genomes are also in non-coding DNA (Ley et al., 2008) although the contribution that these make to malignant transformation is undetermined at present.

The most well characterized architectural protein in vertebrates is CTCF, which is a highly conserved zinc finger DNA binding protein found in most bilalaterian animals. It was initially described as a transcription factor that could elicit both augmentation and repression of transcription in reporter assays (Lobanenkov et al., 1990) but is now thought to have a much wider role in mediating long-range interactions in chromatin. CTCF normally binds to the linker regions between nucleosomes (Cuddapah et al., 2009) and about half of CTCF sites are independent of cell type. In terms of location, about 15% of CTCF sites are found at gene promoters and a further 40% are found within the gene bodies (Ong and Corces, 2014). The remainder are found in intragenic regions and surprisingly for a protein which was classically ascribed insulator functions, here they are enriched at enhancer elements. The cell type specificity of CTCF binding is often associated with differential methylation of the binding sites, with the presence of 5-methylcytosine (5mC) correlating with reduced CTCF occupancy (Wang et al., 2012).

There is considerable evidence to support the idea that CTCF has an important role in mediating long-range chromatin interactions (Handoko et al., 2011; Sanyal et al., 2012; Splinter et al., 2006). CTCF binds to the cohesin complex, physically linking it to chromatin and cohesin is nearly always essential for CTCF function (Rubio et al., 2008; Wendt et al., 2008; Xiao et al., 2011). The cohesin complex forms a ring that is able to encircle one or more strands of DNA holding them together and although it has a fundamental role in connecting the sister chromatids during cell division it is becoming clear that it is also vital for in regulating transcription.

CTCF has appears to have a number different functions that probably arise from its regulation of genome architecture. Classically, it was described as an insulator (i.e. it is able to repress enhancer induced transcription and buffer transgenes from position effects) but more recently it is thought likely that more often it facilitates interactions between enhancers and promoters (Ong and Corces, 2014). CTCF is also implicated in a number of other processes, such as V(D)J recombination at the IgH locus, that are likely to be dependent on specific long-range interactions. Around 15% of CTCF sites are found at the boundaries of topologically associated domains (TADs), which are the large architectural features in chromatin (Dixon et al., 2012; Nora et al., 2012).

It seems likely that CTCF has an important role in determining gene expression that is at least in part mediated though its ability to modulate chromatin structure.

1.2 Nuclear organization

There is a growing body of evidence that genomic organization within the nucleus is a major determinant of gene expression (Bulger and Groudine, 2011). Microscopy has revealed several broad principles, which determine nuclear architecture. The chromosomes tend to occupy distinct territories inside the nucleus although there are some areas in which intrachomosomal interactions take place. Generally repressed, heterochromatic genes and gene poor regions tend to reside in the nuclear periphery in association with the nuclear lamina whilst active genes tend to be located in the interior of the nucleus. In keeping with this the gene dense chromosomes have a preferred position in the nuclear interior in many cell types. This is not universally the case; heterochromatin localizes to the center of the nucleus in some specialist cell types, for example, the rods of the retinal epithelium of nocturnal mammals, in which it acts as a lens to deflect light towards the light sensing perimembrane regions of the cells (Solovei et al., 2009).

Some studies have also found that specific gene loci can migrate out of their chromosome territory and that this coincides with gene expression (Noordermeer et al., 2008). There is evidence that artificial tethering of reporter genes to the nuclear periphery results in repression of gene expression although

this is not always the case (Andrulis et al., 1998; Finlan et al., 2008; Reddy et al., 2008). In addition it has been shown using a construct containing a TALE fused to an acidic peptide that decondensation of chromatin is sufficient to induce gene positioning without transcription (Therizols et al., 2014).

There are a number of substructures within the nucleus that can be resolved microscopically, which are likely to be involved in gene expression or repression. For example nucleoli and splicing speckles are likely to be coincident with gene expression whilst polycomb bodies are likely to be associated with repressed chromatin.

FISH shows that inactive loci adopt a compact structure. Upon gene activation there appears to be decompaction and the loci often appear to migrate out of the chromosomal territory (Dostie and Bickmore, 2012).

At present microscopy is limited in spatial resolution; loci usually need to be separated by over 100kb to be resolved by FISH. The throughput is also low; only a small number of interactions can be determined in a few hundred cells. With the development of super-resolution microscopy there are also concerns that the disturbance of chromatin architecture by DNA denaturation may become a limitation on the quality of the data (Dostie and Bickmore, 2012).

This technology is evolving rapidly, particularly with the advent of super resolution microscopy. However, the majority of discoveries regarding sequence specific genome structure, to date, have been made using 3C technology. These include the description of large scale structures such as topologically associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and much smaller scale interactions such as those with regulatory elements (Tolhuis et al., 2002) and within the gene body itself (Tan-Wong et al., 2012).

1.3 Chromosome conformation capture technology

The first assay that used a restriction enzyme digest followed by ligation to detect DNA-DNA interactions was performed in the late 1980s to show that DNA-DNA interactions occur in plasmids (Mukherjee et al., 1988). Subsequently a derivative of this was used technique to demonstrate enhancer promoter interactions at the rat prolactin gene using viral minichromosomes in a cell line

(Cullen et al., 1993). However, it wasn't until the seminal paper by Dekker et al., (Dekker et al., 2002) that it became possible to define interactions in the chromatin of eukaryotic cells and studies demonstrating interactions between regulatory elements in mammalian genomes quickly followed (Tolhuis et al., 2002). The assays are all based on the principle that chromatin can be cut and then rejoined so that sequences in close physical proximity ligate to one another (Figure 1.1a). The ligation junctions that result reflect the three-dimensional organization of the genome and can therefore be used to inform about chromatin structure.

It is important to realize that when the 3C library is being manufactured every restriction fragment in every individual cell could ligate to many partners. If one assumes that the restriction cut site has the highest probability of ligating back to its original partner in the fixed nucleus then there are at least 20 potential ligation partners for every cut site in each cell because the probability of ligating back to the original partner is only around 5% (1 in 20) (Davies et al., 2015) (Figure 1.1b). In addition live cell imaging studies show that chromatin is in a state of constant flux within the nucleus (Lucus et al., 2014) and single cell 3C assays, therefore, show huge variability in the ligation junctions between cells (Nagano et al., 2013).

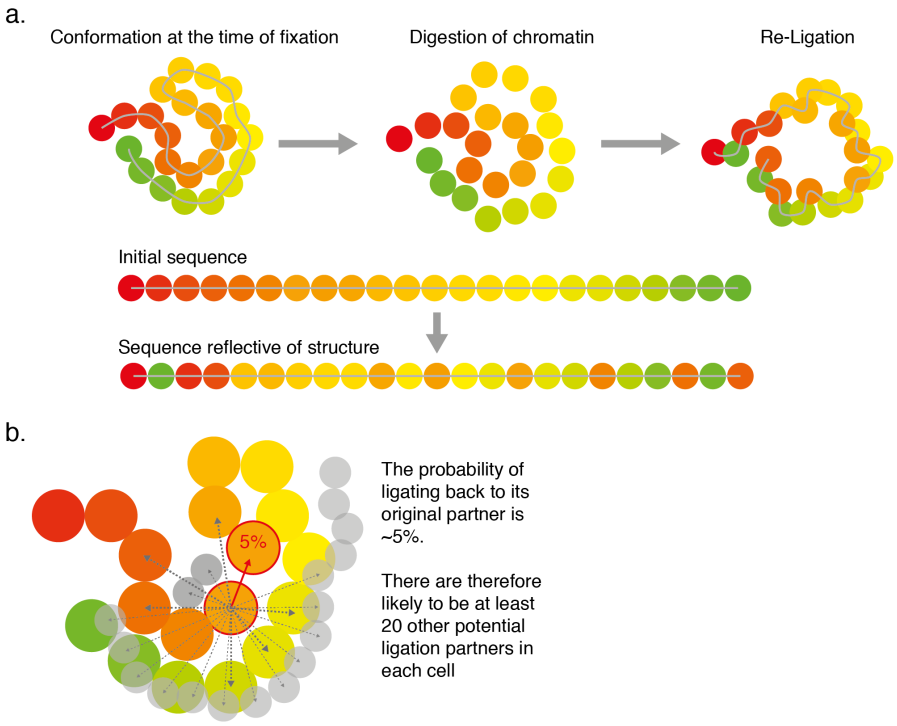


Figure 1.1 3C strategy

a. The chromatin fiber is initially digested into short fragments and then a ligation reaction is performed so that the order of the components in the fragment reflects the position of the fragments relative to one another.
b. Only around 5% of fragments ligate back to their original partner in a DpnII digested library. Assuming that this is the most likely outcome from the ligation reaction there are at least 20 other potential partners for each ligation junction in each fixed nucleus.

The ligation junctions present within the 3C library, therefore, represent a probability distribution of the possible ligation junctions from a group of cells in a particular state. It is therefore vital to quantify the ligation junctions present in the 3C library accurately and with significant depth in order to build up a profile of the probability distribution underlying the ligation junctions (Figure 1.2). Simply stating that a junction can or cannot be found is not particularly meaningful because if one samples the library with sufficient depth then it is likely that a ligation junction could be found with most other fragments in the genome.

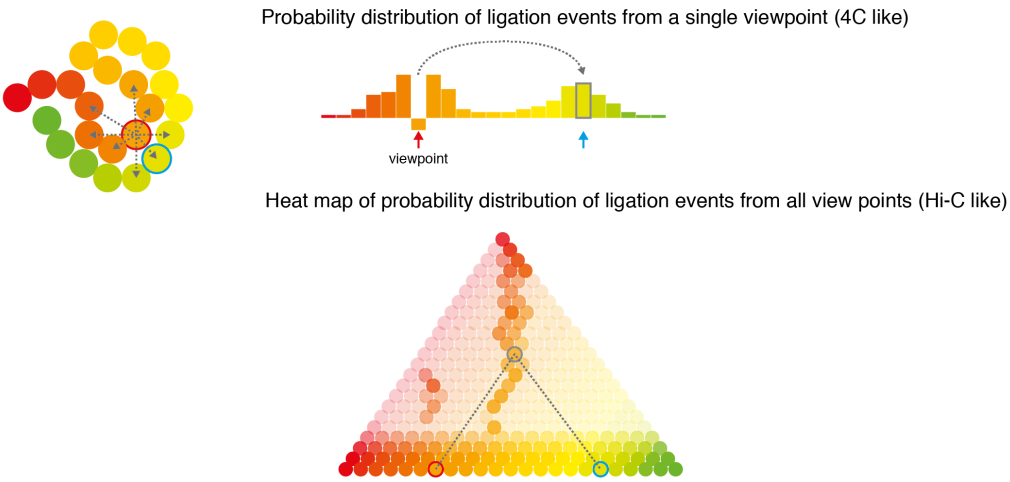


Figure 1.2 Probability distribution of ligation events

This shows the potential interaction partners from a single fragment (highlighted with the red border). A theoretical interacting partner, a regulatory element for example, is highlighted in blue. This can be represented either as an interaction profile from a single viewpoint as is reported by 4C and Capture-C methods. Alternatively heat maps showing the interaction frequencies allow all vs all data types to be viewed, commonly used to report Hi-C and 5C data.

1.3.1 3C library preparation

Generally cells are initially fixed using formaldehyde crosslinking (Figure 1.3). It is possible to manufacture 3C libraries without this fixation step by embedding the cells in agarose plugs, in order to maintain the nuclear shape and structure. This results in similar Hi-C profiles to those generated from formaldehyde fixed material although there is greater background noise (Rao et al., 2014).

In most techniques the chromatin is cut using a restriction enzyme. The choice of restriction enzyme determines the maximum resolution of the data because interactions can only be detected at restriction enzyme cut sites. Generally 3C and 5C are limited to being performed with six-cutter restriction enzymes such as HindIII, EcoR1 and NcoI and this considerably limits the resolution of these methods because these enzymes only cut every 4096bp on average. 4C and

Capture-C generally tend to use four-cutter restriction enzymes, which generate data that has 16 times the resolution (in a random DNA sequence a four-cutter will cut every 256bp on average). Both four (Rao et al., 2014; Sexton et al., 2012) and six (Lieberman-Aiden et al., 2009) cutter enzymes have been used for generating Hi-C libraries although here the resolution is often limited by the sequencing depth rather than the choice of restriction enzyme.

Digestion of chromatin by DNAseI (Weintraub and Groudine, 1976) and restriction enzymes (Pfeiffer and Zachau, 1980) has been extensively used to determine chromatin accessibility and determine active regions of the genome. For 3C library preparation the length of time used for restriction enzyme digestion and the quantities of restriction enzyme used are orders of magnitude greater than those used for chromatin accessibility assays. However, considerable effort is required to optimize digestion efficiency and this can vary considerably between different cell types (Stadhouders et al., 2013). Despite this it is likely that the digestion efficiency will still be dependent on the state of the chromatin to some extent. Although far more widespread digestion occurs it is likely that highly compacted chromatin cuts less well and this leads to potential bias in these assays.

In standard 3C library preparation DNA ligase is added to the chromatin following the restriction enzyme digestion step. The restriction enzyme step leads to over hanging ends to the DNA fragments and the sequences that are in close proximity in the nucleus have an increased probability of ligating to one another. Large concatamers of DNA are generated by this reaction and the sequence of the restriction fragments is dictated by the structure of the chromatin at the time of fixation.

1.3.2 Different methods for analysis of 3C libraries

The challenge with analyzing 3C libraries is that they have enormous complexity because each ligation junction can potentially ligate to every other ligation junction in the genome. In addition, particularly with 6 cutter manufactured libraries, only a small proportion of the DNA contains ligation junctions. Therefore, for large mammalian genomes, simply sequencing the library does not

yield useful information. The key to analysis of these libraries is to enrich the libraries for the ligation junctions of interest. There are a number of different ways this can be achieved, which vary markedly in resolution, throughput and biases (Figure 1.3).

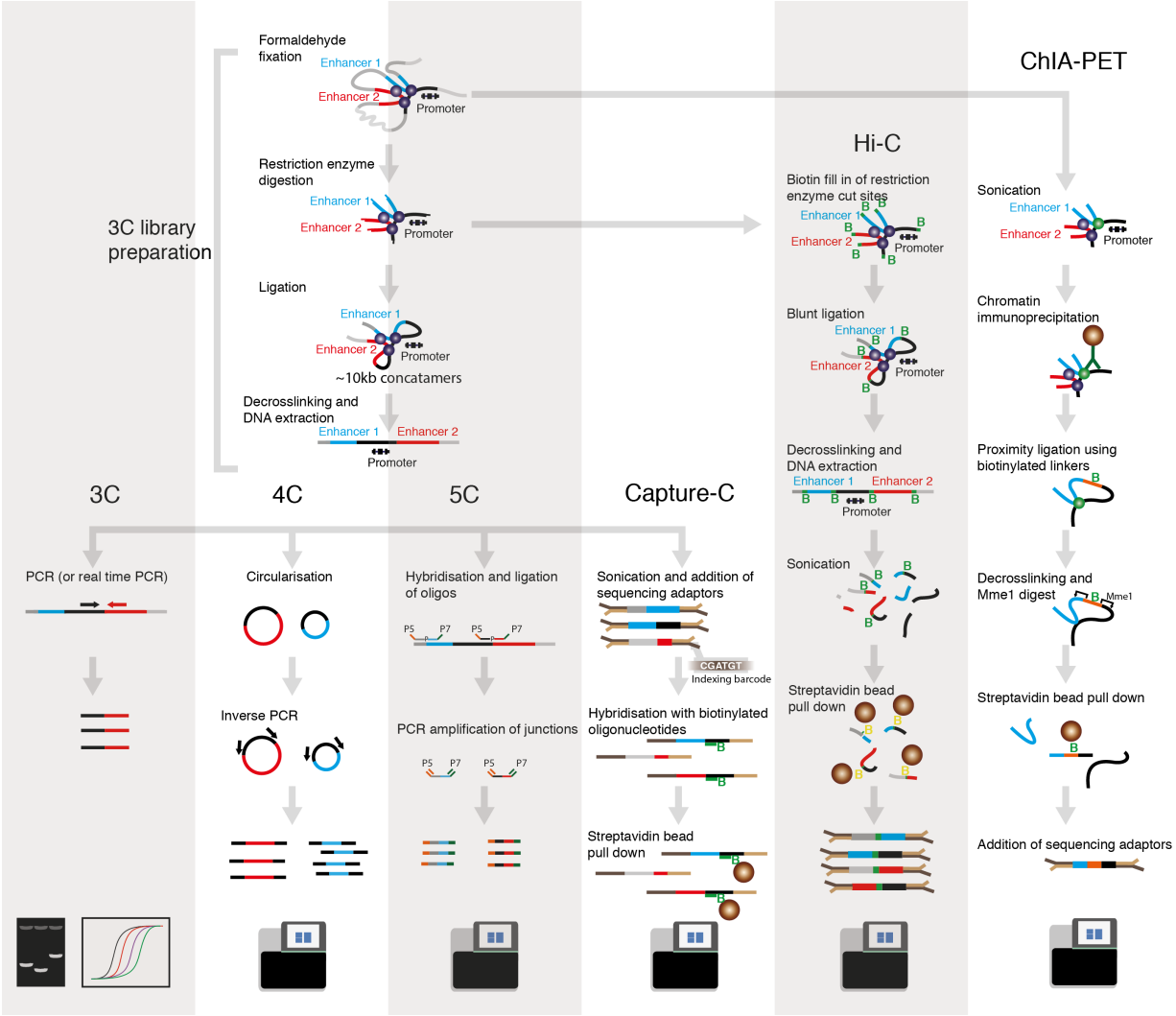


Figure 1.3 Comparison of different 3C based methods

Initially 3C libraries are manufactured by initially crosslinking the chromatin using formaldehyde. Following this a restriction enzyme digest is performed followed by a ligation reaction. The DNA is then extracted following a decrosslinking of the chromatin.

3C libraries can be interrogated by several different techniques. Initially ligation junctions were detected by PCR amplification (3C) and the product of this reaction can be assayed either by gel electrophoresis or real time PCR.

In 4C, 3C libraries need to be circularized by either optimizing the ligation reaction or a second restriction enzyme step followed by a dilute ligation can be performed on the purified 3C library. Inverse PCR is then performed and the product of this reaction is then sequenced.

5C uses a massively multiplexed hybridization with forward and reverse oligonucleotides. A ligation reaction is then performed so that forward and reverse oligonucleotides ligate to each other on either side of ligation junctions. P5 and P7 sequences are then used to amplify these junction products, which are then measured by sequencing.

Capture-C involves sonication of the 3C library followed by ligation of sequencing adaptors. A hybridization reaction and biotin pull down are subsequent performed (this is repeated in NG Capture-C to markedly improve the efficiency of the oligonucleotide capture). This material can then be sequenced.

In Hi-C the 3C library manufacturing process includes a biotin fill in of the overhanging restriction cut sites and a blunt end ligation reaction is performed, which is less efficient than the normal ligation reaction. Following this the material is sonicated and a biotin pull down is performed to concentrate ligation junctions.

ChIA-PET initially involves a chromatin immunoprecipitation step. Biotinylated linkers are ligated to the immunoprecipitated chromatin and following this the linkers are ligated. A MmeI digest is then performed and the material containing the linkers is pulled down using the biotin moiety in the linkers.

1.3.2.1 Chromosome conformation capture (3C)

Initially ligation junctions were detected in chromosome conformation capture libraries by polymerase chain reaction (PCR) followed by gel electrophoresis. Subsequently real time PCR has been used extensively for 3C analysis.

Many important discoveries have been made using 3C. The technique was initially used to define the spatial organization of chromosomes in yeast (Dekker et al., 2002). Interactions at the murine β-globin locus were determined soon after (Tolhuis et al., 2002). This seminal paper showed that the enhancers at the beta globin LCR physically interact with the promoter; a finding that led to the active chromatin hub model, which predicts that multiple enhancers interact with the promoter to induce transcription. Subsequently 3C has been used to analyse many other loci including the T-helper type 2 cytokine locus (Spilianakis and Flavell, 2004), the immunoglobulin light chain locus (Liu and Garrard, 2005) and the alpha globin locus (Vernimmen et al., 2007; Vernimmen et al., 2009; Zhou et al., 2006).

However, conventional 3C is limited because it is only able to detect interactions between a small number of restriction fragments making it difficult to detect interactions with regulatory elements that have not been defined by other techniques. In addition it is limited in resolution because it is usually only able to analyse libraries generated by 6-cutter restriction enzymes such as HindIII. This is because the complexity of 3C libraries increases with the square of the number of fragments so a 4-cutter library will have 256-fold more complexity compared to a 6-cutter library (Lieberman-Aiden et al., 2009). Thus 3C libraries generated by 4-cutter enzymes have a 256 fold reduction in the concentration of individual ligation junctions, making it difficult to detect these reproducibly with PCR. In addition it can be difficult to get reproducible results and correction for biases due to differences in amplification with different primer pairs requires great care. Conventional 3C has, therefore, largely been superseded by other methods.

1.3.2.2 Circular chromosome conformation capture (4C)

The next major advance in the field was the development of 4C, which allows all of the potential interacting partners to be identified from any specific point of

interest in the genome (Simonis et al., 2006; Zhao et al., 2006). 4C uses standard 3C library preparation, often with a 4-cutter restriction enzyme, after which inverse PCR is used from the 'viewpoint' fragment to amplify any interacting partners (Figure 1.3). Initially the interacting partners were read out using a microarray, but this has now been replaced by high throughput sequencing.

For inverse PCR to work efficiently, small circularized fragments must be generated either by optimizing the initial ligation reaction (Zhao et al., 2006) or by the more widely used method of performing a second restriction enzyme digest and ligation of the extracted DNA with a different restriction enzyme (Simonis et al., 2006). The inverse PCR in 4C data is affected by the GC content and size of the interacting fragments to the extent that some large GC-rich fragments simply fail to amplify. This can result in quite serious problems with bias in the interaction profile (Stadhouders et al., 2013) and make the interactions appear more discrete than they really are (Figure 2.5).

In addition one of the main recommended methods uses primers that directly abut the restriction enzyme cut point, making it impossible to distinguish genuine amplification of ligation products from the view point fragment from mis-priming elsewhere in the genome. In addition many groups (van de Werken et al., 2012a) have used 50bp single end reads to analyse the libraries, which makes mapping of the interactions difficult since nearly half of the read is used up sequencing the primer sequence, which makes it challenging to map the other half due to the shortness of the remaining sequence. However, with the improvements in sequencing technology this is much less of a problem than it has been in the past. In the most recently developed adaptation of 4C, UMI-4C (Schwartzman et al., 2016), the bias associated with differences in fragment length has been reduced by the introduction of a sonication step before a nested ligation-mediated amplification.

4C has good resolution and sensitivity, when carefully performed, allowing very long-range interactions to be detected. In order to generate high-quality 4C profiles, it is necessary to perform multiple PCRs to generate material from the DNA of a million or more cells (Simonis et al., 2007). However, it is difficult to determine the sensitivity of conventional 4C accurately, as it is not possible to differentiate between PCR duplicates and unique ligation junctions. In UMI-4C

interactions can be carefully quantified by using the unique fragment ends introduced by the sonication step and this shows that about 5000-10,000 unique read pairs can be detected per μg of input material. UMI-4C also allows for multiplexing up to 20-50 baits, though the individual nested PCR reactions would still need to be optimized for each viewpoint (Schwartzman et al., 2016).

A large number of seminal observations have been made using 4C. It has been widely used to describe interactions between promoters and enhancers and to show how these change during differentiation and development (Andrey et al., 2013). Very long-range interactions between active genes have been demonstrated in *cis* using 4C (de Wit et al., 2013). 4C has also been used to link potential regulatory single nucleotide polymorphisms (SNPs) with genes (Pasquali et al., 2014). It can be used to determine structural changes such as chromosomal rearrangements(Simonis et al., 2009) although this is probably more easily done using conventional karyotype analysis and FISH. It has also been used to define disease mechanisms, including the demonstration that chromosomal translocations can result in distal enhancers becoming juxtaposed to an oncogene leading to malignancy (Groschel et al., 2014).

4C has also been used to link potential regulatory SNPs with genes (Pasquali et al., 2014). It can be used to determine structural changes such as chromosomal rearrangements (Simonis et al., 2009) although this is probably more easily done using conventional karyotype analysis and FISH. It has also been used to demonstrate novel disease mechanisms, for example 4C was used to show that chromosomal translocations can result in a distal enhancer becoming juxtaposed to an oncogene leading to malignancy (Groschel et al., 2014).

1.3.2.3 Chromosome Conformation Capture Carbon Copy (5C)

This technique is based on detecting ligation junctions using massively multiplexed ligation mediated amplification of primers designed adjacent to the restriction enzyme cut site. It uses forward primers that contain a T7 sequence at the 5' end and reverse primers that are phosphorylated on the 5' end and contain a T3c sequence at the 3' (Dostie et al., 2006). Following hybridization,

ligation and PCR amplification using T7 and T3c primers the frequency of ligation junctions can be assayed using microarray or sequencing.

Two possible designs are used for this type of assay. The forward and reverse primers can be designed to the ends of alternating fragments to try and provide a global all v all interaction profile. Alternatively the smaller numbers of reverse primers can be designed to particular viewpoint fragments of interest and forward fragments can be tiled across the ends of the other intervening fragments.

This technique was used to determine interaction profiles at the alpha (Bau et al., 2011) and beta globin loci as well as the pilot regions of the ENCODE project (Sanyal et al., 2012). In addition analysis of the X-chromosome by 5C provided some of the first evidence for the existence of topologically associating domains (TADs) (Nora et al., 2012). Although massively multiplexed 5C can be performed, it requires significant resources for somewhat limited resolution and scale. For example, the ENCODE project had to use 5826 oligos to cover 1% of the genome and they only performed two replicates of three different cell lines.

1.3.2.4 Hi-C

This technique uses a modified method for 3C library preparation. Here a biotinylated nucleotide fill in is performed after the restriction enzyme digestion and following this blunt end ligation is done. This is less efficient than the ligation step in conventional 3C library preparation, in which there is an overhang from the restriction enzyme digest. After DNA extraction and sonication of the material a streptavidin bead pull down is performed to concentrate ligation junctions, which can then be analysed by high throughput sequencing.

Since it was initially reported (Lieberman-Aiden et al., 2009) Hi-C experiments have produced data at increasing resolution predominantly through massive increases in sequencing depth. Initially 40kb bins were used (Lieberman-Aiden et al., 2009) but this was improved to 5-10kb (Jin et al., 2013) and more recently 1kb bins have been reported (Rao et al., 2014). Producing Hi-C data of this quality for large mammalian genomes requires a gargantuan effort; with the

analysis of several billion read pairs. This is beyond the capabilities of most laboratories at present and even with billions of read pairs the number of informative interactions from any individual viewpoint are probably around 100 fold lower than the techniques such as 4C-seq and Capture-C that use a one/many vs all approach.

Hi-C has been used extensively to define topologically associated domains genome wide (Dixon et al., 2012) and to determine the structure of the chromosome during mitosis (Naumova et al., 2013). Hi-C has also been used to link trans interactions with sites associated with chromosomal translocations (Zhang et al., 2012). At present this is the best available technique for determining genome wide all vs all maps of interaction.

1.3.2.5 Cavalli Hi-C

For small genomes such as *Drosophila Melanogaster* it is possible to generate 3C libraries using a four cutter restriction enzyme such as DpnII and sequence the library without any enrichment steps (Sexton et al., 2012). In order to reduce sequencing of uninformative fragments that do not contain ligation junctions, the material is sonicated to ~800bp prior to the addition of sequencing adaptors. The potential problem with this approach is that the sonicated fragments are over twice the size of an average DpnII fragment, so there is a reasonable chance that restriction fragments could lie in between the two restriction fragments at the end of the sonicated fragment, which are the only part of that are sequenced paired end sequencing. Therefore interactions could be identified between regions, when there is no ligation junction.

1.3.2.6 Single Cell Hi-C

A variant of 3C, called single-cell Hi-C has been developed, which impressively allows the Hi-C interactions to be determined from single cells (Nagano et al., 2013). In order to achieve this, single nuclei were picked and individually sequenced following standard 3C library preparation. This study showed that topological domains can be determined in individual cells but there was considerable variability between different cells, particularly with regards to interchromosomal contacts.

It is difficult to know how much of this variability reflects the chromatin state of the individual cells at the time of fixation or the variability inherent in the 3C library process. Data from methods such as Capture-C suggests that the probability of the restriction fragment ligating back to it's original partner is less than 5%. If interacting fragments have a similar or lower probability of ligating to that of the original partner indicating then at least 20 potential ligation partners will exist for each restriction fragment in each cell but only one of these can be reported. Therefore it is likely that considerable variability would be seen if single cell Hi-C could be performed on absolutely identical cells.

1.3.2.7 Capture-C

This approach combines 3C library preparation with a four-cutter restriction enzyme and oligonucleotide capture technology (Hughes et al., 2014). The Agilent SureSelect system, which uses RNA oligonucleotides was used; allowing genome wide interaction profiles to be determined for 450 gene promoters simultaneously. This allowed regulatory elements to be defined at hundreds of loci, which is was a significant advance in the field. It identified all of the known regulatory elements at the very well characterized alpha and beta globin loci. In addition to this it was possible to determine the interaction profiles from all the genes in a single locus, allowing an intricate picture of the regulatory landscape to be defined.

In it's initial format, Capture-C was somewhat inflexible since the Agilent SureSelect system has a minimum design of 40,000 different oligonucleotides. Whilst this is cost effective for analyzing several hundred of genes, it is prohibitively expensive for researchers wishing to analyse a smaller number of genes in multiple samples. In addition, the capture efficiency limits the sensitivity and efficiency of the technique.

1.3.2.8 Micro-C

A further increase in resolution has recently been achieved in *S. Cerevisiae* by substituting micrococcal nuclease (MNase) in place of the restriction enzyme using a Hi-C based protocol (Hsieh et al., 2015). This shows similar but not identical maps of interaction compared to Hi-C. In particular long-range

interactions are poorly captured by Micro-C; it shows that ~90% of all interactions are within 1kb, which is a blind spot of current approaches.

1.3.2.9 Capture Hi-C

This approach combines Hi-C library preparation with oligonucleotide capture technology. It has been used to target every gene promoter in the genome on libraries prepared using a six-cutter restriction enzyme (Schoenfelder et al., 2015). The data from these publications appears to be of relatively low quality compared to 4C-seq. From the authors own analysis of the data there are only 88 significant interactions with the promoter of the alpha globin gene for example and they have only achieved a 10-fold enrichment of the targets they are interested in.

The potential problem with this approach is that the library complexity may be lost during the Hi-C library preparation, which is probably an unnecessary step if the capture oligonucleotides are designed to the ends of the restriction fragments.

1.3.2.10 T2C

In this technique 3C libraries are manufactured using a six-cutter restriction enzyme (Kolovos et al., 2014). A second restriction enzyme digestion of the 3C library is then performed prior to the addition of sequencing adaptors. This material is then captured using oligonucleotides that are designed to capture the ends of all of the restriction fragments in the locus of interest.

The pitfall with this approach is that ligation junctions between restriction fragments inside locus of interest (that is being captured) can be hybridized to two different oligonucleotides. This means that the interaction profile becomes dependent on the combined capture efficiency of the two different oligonucleotides. Although the authors did not find a significant difference between the capture efficiencies between fragments captured by one or two oligonucleotides it seems unlikely that this was genuinely the case for all of the oligonucleotides, which makes interpretation of these data difficult.

1.3.2.11 ChIA-PET

In this technique the material is initially sonicated and chromatin immunoprecipitation (ChIP) is performed to enrich DNA-protein complexes. DNA linkers are then ligated used to ligate sequences held in close proximity. The linkers are biotinylated and contain MmeI restriction sites allowing short fragments of material to be extracted with a streptavidin bead pull down. Paired end tags (PETs) can then be identified by paired end sequencing.

ChIA-PET studies have been performed using ChIP for the oestrogen-receptor-alpha (Fullwood et al., 2009) and PolII (Kieffer-Kwon et al., 2013).

The technique has limitations because the initial ChIP results in only a tiny amount of material being pulled down and so the number of reads used to identify interactions is usually very small. In ChIAPET three read pairs are often used to define significant interactions (it is thought that this corresponds to a false discovery rate of <0.05) (Fullwood et al., 2009). This makes the technique insensitive and vulnerable to artifact. In particular the initial ChIP can introduce bias towards detecting interactions between sequences that bind the protein targeted by the antibody used for the immunoprecipitation. Thus studies that show that CTCF and PolII tend to interact with other sequences bound by these proteins need to be interpreted with caution.

1.4 Thesis overview

Although many variants of 3C have been developed to improve the resolution (Noordermeer et al., 2011b) and number of interactions that can be investigated simultaneously (Lieberman-Aiden et al., 2009; Sanyal et al., 2012). At the start of the project only a small number of genes had been analysed at high resolution in a small number of different cell types because high resolution identification of genome wide interactions can only be performed one gene at a time (Andrey et al., 2013). To circumvent this, our laboratory developed a technique called Capture C (Hughes et al., 2014), which is able to determine the interaction profile of over 450 genes in a single assay, with resolution and sensitivity comparable to 4C seq. However, the original Capture-C protocol (Hughes et al., 2014) used was very inflexible and costly to perform on multiple samples. We wanted to be able

to analyse smaller numbers of regions in many different samples. Furthermore, the sensitivity possible with this design did not allow very long-range *cis*-interactions or *trans*-interactions to be detected robustly and so it was difficult to be certain that important long range interactions were being missed.

The project therefore aimed to develop 3C based techniques and the following attributes were felt to be important aims:

1. Improved flexibility so that the interaction profiles in one to several hundred genes could be determined simultaneously and cost effectively in multiple samples. This would allow comparison if between different genes in multiple samples in replicates in different cell types, which is currently not feasible by any of the currently available techniques without huge effort and expenditure.
2. Improved sensitivity to allow very long rang interactions to be quantified accurately.
3. Improved resolution to give a better picture of the interactions between genes and regulatory elements, which we hope should inform on the underlying mechanisms involved.
4. Improved reproducibility – currently not well defined in any of the techniques.
5. Reduced biases – most of the available techniques have significant PCR associated biases.
6. Ability to use small numbers of cells – the current standard in the field is to use input material of between 10^7 and 10^8 cells. This is particularly important for analysis of primary cells and if cell numbers could be reduced to allow the use of FACS sorted cells to be analysed this would be hugely important and allow many biologically important questions to be answered. In particular it should aid the investigation of SNPs in regulatory elements, in primary human cells.
7. Ability to analyse multiple samples simultaneously and increase the throughput of the techniques.

8. A standardized protocol should be developed so that it should allow any non-specialist molecular biology laboratory to perform these experiments.
9. Improved methods of data analysis including the release of open source software for data analysis.
10. Improved statistical methods for identifying significant interactions.

With these aims in mind the Capture-C protocol was modified to use biotinylated DNA oligos in solution so that each set of capture oligos can be designed specifically to capture from one to many hundreds of regions, in a single experiment and designs can be easily expanded by addition of new oligos to existing pools. Importantly, independent 3C libraries (e.g. from different cell types or different stages of development) can now be captured and processed in a single tube making separately indexed samples directly comparable. This greatly increases throughput and allows meaningful subtractive analysis of chromosome conformation in different cell types. Using this approach we have defined the smallest number of cells required to identify robust interactions and shown how this approach can be used to analyse interactions with specific SNP-containing regions of interest.

In addition two further techniques were developed. C-nanostring, uses nanostring technology to analyse 3C libraries. This potentially provides more reliable quantification of interaction profiles as it does not involve PCR amplification of the library prior to analysis. In addition C-nanostring analyses 12 samples simultaneously, allowing multiple replicates to be performed in different tissues.

Finally a technique was developed to try and define the coincident interactions between promoters and enhancers and nanopore sequencing was used to try and define the simultaneous interactions inside the compartment.

2 Development of the Next Generation Capture-C assay

2.1 Rationale

Using conventional chromosome conformation capture, it is now possible to analyze in detail the interactions between enhancers, silencers, boundary elements and promoters at individual loci at high resolution. More recently, we have shown that using a high-throughput approach (Capture-C) it is now possible to interrogate *cis*-interactions, at hundreds of loci at high resolution (at least equivalent to that of 4C) in a single experiment. Such approaches are of immediate value in defining interactions within the chromosomal landscapes of many loci, and identifying the genes and the functional effects of SNPs that are associated with complex diseases, the majority of which lie in intergenic *cis*-acting regulatory elements.

Although a significant advance on other chromosome conformation methods, the original Capture-C protocol uses a relatively expensive and inflexible platform for oligonucleotide capture. This includes a minimum design of 40,000 oligos (140nt in length) because the oligonucleotides are synthesized on a microarray. Although experiments on this scale are extremely useful, laboratories more often want to analyse a much smaller subset of interactions in multiple samples. Furthermore, the sensitivity of the current design does not readily allow for the analysis of very long-range *cis*-interactions or *trans*-interactions. Finally, because of this limited sensitivity and efficiency, the original Capture-C protocol, although more quantitative than other chromosome conformation methods, does not provide entirely quantitative estimates of chromosomal interactions.

To address these limitations and further develop Capture-C for the analysis of any number of loci, cheaply, and easily, we have made some important and transformational modifications to the protocol and this new protocol was termed next generation Capture-C (or NG Capture-C). First, rather than using a fixed oligo array, we have shown that the Capture-C protocol works efficiently using from 2 to 265 oligos in solution, however, there is no theoretical limit to the number of oligos that could be included. This means that each set of capture

oligos can be designed specifically to capture from one to many hundreds of regions, cheaply, in a single experiment. Importantly, independent chromosome conformation libraries (e.g. from different cell types or different stages of development) can now be captured and processed in a single tube making separately indexed 3C libraries directly comparable and allowing meaningful subtractive analysis of chromosome conformation in different cell types.

The original Capture-C protocol enriches junction fragments in the library by a factor of 5-20,000. Using the double capture described here, provides a further 100 fold enrichment and approximately one million fold overall. This makes it possible to reproducibly detect rare interactions present in only one in 5-10,000 cells. This increased sensitivity combined with the inherent facility for removing PCR duplicates using Capture C, means that almost all unique interactions can be scored and counted in a given library resulting in a truly quantitative analysis of such interaction data.

Finally this streamlined protocol, allows for simultaneous capture of differently labeled samples and easily repeated experiments, which can be analysed using bench top sequencing (MiSeq rather than HiSeq) platform. Using this approach we have defined the smallest number of cells required to identify robust interactions; the effect of different cross linking conditions, and shown how this approach can be used to analyse interactions with specific SNP-containing regions of interest.

2.2 Overview of the method

This technique initially involves the standard method of manufacture of 3C libraries: namely, formaldehyde crosslinking of live cells; restriction enzyme digestion of chromatin (with a four cutter restriction enzyme); ligation under dilute conditions; de-crosslinking and DNA extraction (Figure 1). This generates a library containing concatamers of restriction enzyme fragments, which have been shuffled so that sequences that were in close physical proximity in the nucleus are juxtaposed. These libraries have a high degree of complexity, with

more than 10^{12} ligation junctions being theoretically possible, so some form of enrichment is needed prior to sequencing.

For enrichment by oligonucleotide capture, biotinylated DNA oligonucleotides are designed to capture the restriction enzyme fragments of the desired viewpoints (the fragment containing a gene promoter, for example). To maximize the capture of fragments containing interacting partners, single 120mer oligonucleotides are designed to the ends of the restriction fragment so that they contain the restriction enzyme cut sequence (obviating the need for a biotin fill in to enrich for ligation junctions). These are screened for repetitive sequences to minimize off target capture and we developed an online tool to automate design for multiple targets.

Prior to oligonucleotide capture, 3C libraries are sonicated to 200bp followed by the addition of Illumina paired-end sequencing adapters. Maintaining library complexity here is crucial because ideally over 100,000 fold coverage is required over the captured viewpoints. The amount of input DNA was maximized by performing the sequencing library preparation in duplicate, allowing a total of 10ug of 3C library to be used. In addition losses were minimised prior to PCR amplification of the library with the index specific primers by optimising the SPRI bead clean up steps.

The hybridisation reaction is performed with blocking oligonucleotides to the sequencing adaptors and Cot DNA to prevent secondary off target capture from repetitive sequences. Streptavidin beads are then used to pull down the biotinylated oligonucleotides and capture the hybridized fragments of DNA containing the desired view-point sequences. Following multiple bead washes to remove uncaptured DNA; primers designed to the ends of the Illumina adaptors are used to PCR amplify the captured DNA. The first oligonucleotide capture step enriches the targets around 5000 fold but the captured DNA often only represents around 1% of the sequences present. Therefore we have markedly improved the assay by repeating the oligonucleotide capture, which results in enrichment of the targets over 50,000 fold, so that captured material makes up the majority of the sequenced material (Figure 2.3). This hugely increases the efficiency of the sequencing such that we are able to generate up to 24 tracks (3

genes from 8 samples) from a single run from the Illumina MiSeq (20 million reads). Paired end sequencing has been used with 150 bp reads (300bp total).

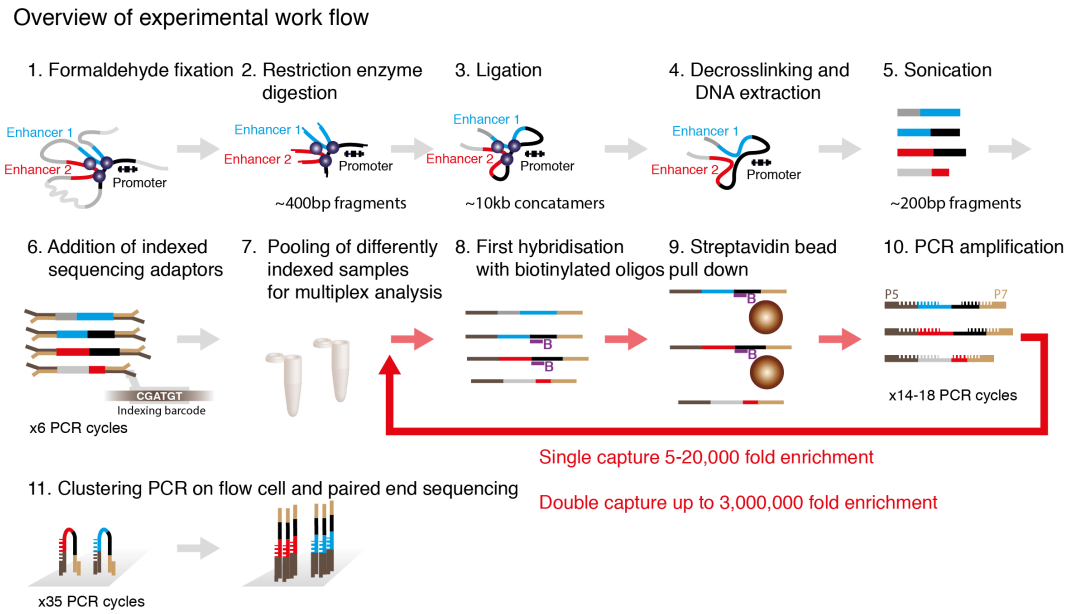


Figure 2.1 Overview of the experimental workflow

2.3 Improvements to the method

2.3.1 Optimisation of oligonucleotide design

The library design has been simplified so that only single biotinylated 120bp DNA oligonucleotides (rather than 3-5 overlapping oligonucleotides) are used to capture each end of the restriction fragment. This not only saves on reagent costs but it should also reduce capture of reads over the capture fragments that do not contain a ligation junction. Because the oligonucleotides contain the restriction enzyme cut sequence (GATC in the case of DpnII) there is no need for biotin fill in to enrich for ligation junctions (a procedure that risks substantial loss of library complexity).

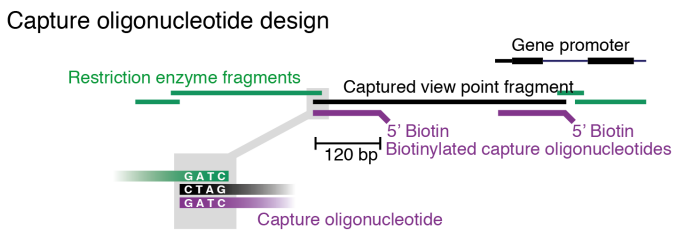


Figure 2.2 Capture oligonucleotide design

120bp oligonucleotides are designed to the ends of the desired viewpoint restriction enzyme fragments. The promoters of genes were used exclusively in this study but any non-repetitive restriction fragment in the genome could be used. The oligonucleotides include the restriction enzyme cut sequence (GATC in the case of DpnII) obviating the need for a biotin fill in. The oligonucleotides include a 5' biotin moiety to allow streptavidin bead capture of the target sequence and any interacting partners. The oligonucleotide sequences are screened using the online Capsequm tool for repetitive sequences and short repeats.

2.4 Introduction of double capture step

We found that when capture is performed using an individual capture site (such as the promoter of the Hba gene) a single oligonucleotide capture step enriches the targets ~ 5-20,000 fold over the capture site. Despite this the captured DNA from this single site only make up less than 1% of the sequenced reads, with the remainder being uncaptured background (Figure 2.3a). In NG Capture-C the use of two sequential oligonucleotide capture steps results in up to 3,000,000 fold enrichment over baseline so that captured material makes up approximately half of the sequenced material (Figure 2.3a).

2.4.1 Comparison of double capture with single capture

This second capture step increases the number of PCR cycles and the number of PCR duplicates sequenced because the library complexity limits the number of unique interactions determined rather than the sequencing depth. However, the PCR duplicates are automatically excluded by removing non-unique reads and no differences are seen in the interaction profiles, GC content or read length compared to a single capture step (Figure 2.3c&d). Paradoxically PCR duplication of the material during the addition of sequencing adaptors improves the quality of the data produced by the assay. This is because PCR generates multiple copies of each ligation junction to be present in the library and this protects against the loss of library complexity, particularly when small cell numbers are used. In addition the PCR steps following the first capture step

increase the PCR duplication of the material, meaning that the second capture step has very little impact on loss of library complexity.

The alpha and beta globin genes are some of the most extensively characterized genes and their regulatory interactions have been interrogated by almost every 3C based method to date (Bau et al., 2011; Hughes et al., 2013; Hughes et al., 2014; Sanyal et al., 2012; Tolhuis et al., 2002; van de Werken et al., 2012b; Vernimmen et al., 2007) so they provide important controls for the validation of any new methodology. The identical profiles generated by both the single and double capture are in agreement with previous 3C analyses and accurately identify every known interaction of the Hba-a1 and Hba-a2 promoters (Figure 2.3b). However, importantly, for the same depth of sequencing the double capture increases the sensitivity of the determined profile (30 fold).

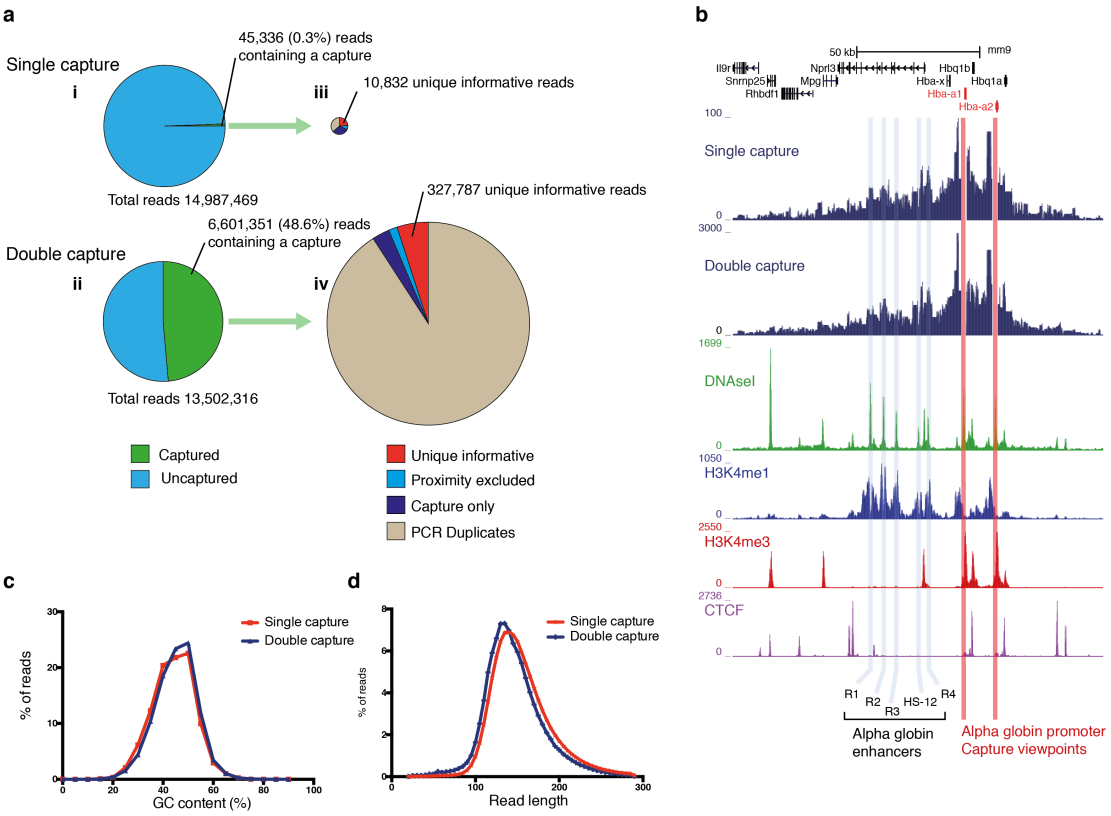


Figure 2.3 Comparison of single and double oligonucleotide capture

3C material generated from erythroid cells was captured using a single set of oligonucleotides designed to the alpha globin promoters (Supplementary Table). Since the two copies of the gene are virtually identical interaction profiles are generated from both genes simultaneously. After the first oligonucleotide capture step some of the material was sequenced using the Illumina MiSeq. The

remaining library was used as input for a second round of oligonucleotide capture and the resulting material was then sequenced.

a) Comparison of the enrichment (to scale) resulting from the single and double capture. (i) Single capture results in 5-20,000 fold enrichment but this only results in around 0.3% of the reads containing a sequence that maps to the captured fragment. (ii) Double capture increases the enrichment markedly; producing up to 3,000,000 fold enrichment. This dramatically increases the percentage of reads containing a restriction fragment that maps to the capture region from 0.3% to 48.6%. The number of unique interactions is increased around 30-fold following double capture (from 10,832 to 327,778) (iii & iv) because the library complexity now becomes the limiting factor.
b) Comparison of the raw informative interactions count per restriction enzyme fragment for single and double capture. The red vertical lines denote the location of captured viewpoints. The light blue lines highlight the five well described regulatory elements in the mouse (R1, R2, R3, R4 and HS-12). This shows that double capture does not alter the significantly alter the local interaction profile but it has 30-fold increased sensitivity.
c) Comparison of GC content of the informative reads between single and double capture. This shows that despite additional PCR amplification there is virtually no change in GC content in the reported reads.
d) Comparison of read length of reads that could be completely reconstructed from single and double capture, showing that there is little alteration in the distribution of the read lengths.

2.4.2 Reproducibility of data

The data generated were highly reproducible between different replicates (R2 >0.95) (Figure 2.4). The data below showing very little variability are from samples that were processed independently. It is a comparison of the read counts for each restriction fragment; the only processing that the data have undergone is to correct the count for the total number of reads in the track across the genome. I developed the protocol so that multiple samples could be

processed simultaneously, which makes it significantly easier to produce data with very little inter sample variability.

Importantly there was very little skewing caused by the size of the interacting fragments (Figure 2.4c). Interacting fragments below 35bp in size were underrepresented compared to the calculated distribution of restriction fragment sizes across the genome. This is principally due to difficulties mapping small fragments to the genome. However, above this size, the distribution matches the computational fragment size distribution (calculated by cutting the mm9 genome in silico using a custom perl script).

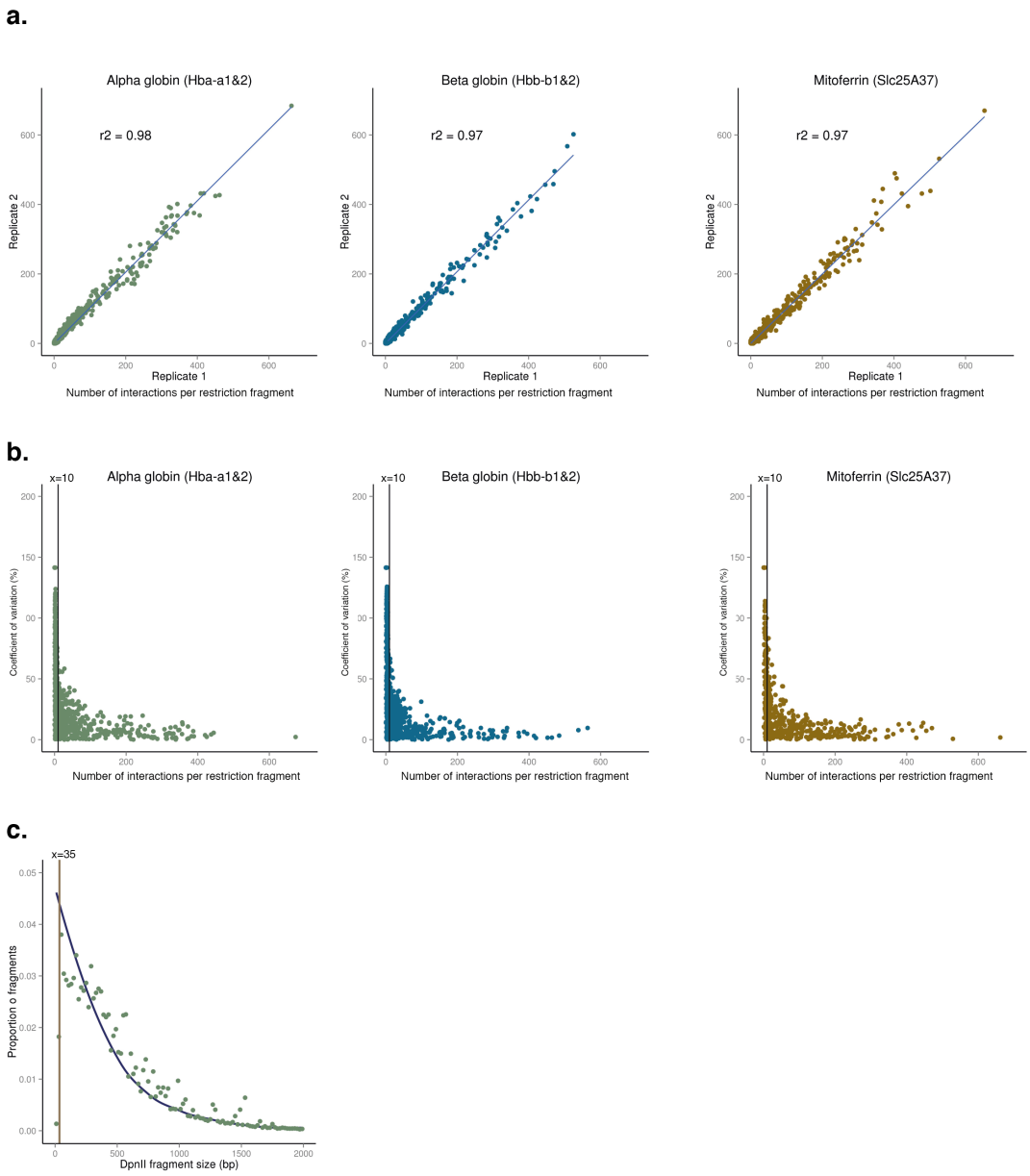


Figure 2.4 Reproducibility of NG-Capture-C

a. Comparison of two biological replicates from three different genes (Hba-1&2, Hbb1&2 and Slc25A37 (mitoferrin)) showing very high degrees of correlation. The data has been normalised to a total interaction count of 100,000 across the whole genome.
b. Plot of coefficient of variation against normalised number of unique interactions per restriction fragment. The vertical lines are plotted at 10 interactions showing that the CV drops below 50% when more than 10 unique interactions are detected per restriction fragment. Given that the track is scaled to contain a total of 100,000 interactions this means that we are able to reliably detect interactions present at 1 allele in 10,000.
c. Distribution of informative interactions by DpnII fragment size (green points) compared to the distribution of DpnII fragments across the genome (blue line). This shows that when the DpnII fragment is above 35bp in size (vertical line) the distribution of the informative interactions follows that of the distribution of fragments across the genome. (Data used from the Capture-C profile for the Hba-a1&2 promoters).

2.5 Comparison of NG Capture-C with other methods

NG Capture was performed at the alpha globin locus, which has been studied using several other techniques. The method compares very favorably with the available 4C data (Figure 2.5), which has significant gaps in the profile caused by an inability to amplify some of the larger restriction fragments.
It also produces much higher quality data than the original Capture-C method. The key statistic to judge the quality of Capture-C and other 3C-seq base methods is the number of interactions from each viewpoint once PCR duplicates have been removed. In the original method around 4000 unique read pairs were identified from the alpha globin promoters and when a library was sequenced following a single capture step between 4000-6500 unique read pairs were determined (Table 1). By comparison the new method this was over 100,000 in most experiments and as high as 327,000 when a single pair of oligos was used just to pull down the Hba promoter fragments (Table 1). The interaction profile is also more complete and much longer range interactions can be identified (see Chapter 5).

In addition the far higher numbers of read pairs allowed us to report the data in raw format i.e. the number of reads falling on each restriction fragment rather than using the moving windows used by 4C and the original Capture-C method (Figure 2.6). In addition because the vast majority of fragments were located clos to the ends of the fragments the resolution could be increased further by reporting the data to the half of the fragment, containing the midpoint of the read (Figure 2.5 & Figure 2.6).

Experimental conditions					
Material	Phenylhydrazine spleen	Phenylhydrazine spleen	Phenylhydrazine spleen	Ter119+ cells selected from phenylhydrazine spleen	Ter119+ cells selected from phenylhydrazine spleen
Genes	Hba-a1&2	Hba-a1&2	Hba-a1&2	Hba-a1&2, Hbb-b1&2, Slc25A37	Hba-a1&2, Hbb-b1&2, Slc25A37
Capture method	Single capture	Double capture	Double capture	Double capture	Double capture
Sequencing statistics					
Sequencing platform	MiSeq	MiSeq	MiSeq	MiSeq	MiSeq
Total number of reads:	14,987,469	13,502,316	16,227,848	14,209,264	12,152,335
Aligning sequences (total read pairs):	18,518,606	20,560,056	23,323,154	19,392,239	16,281,427
Non aligning sequences:	4,898,893	2,056,422	2,557,071	2,261,741	2,006,231
% Non aligning sequences	26.5	10.0	11.0	11.7	12.3
Reads with a capture:	41,702	6,510,497	8,359,826	5,461,616	4,094,576
% Containing a capture	0.3	48.2	51.5	38.4	33.7
Total number of non-duplicated reads:	13,926,025	1,842,568	3,018,770	6,754,595	6,374,133
Duplicated reads:	1,061,444	11,659,748	13,209,078	7,454,669	5,778,202
% Duplicates	7.1	86.4	81.4	52.5	47.5
% Reads containing a DpnII cut site	31.7	64.3	61.7	48.5	45.0
Number of unique interactions					
Hba-1	6,189	197,592	327,757	172,390	144,904
Hba-2	4,643	130,195	229,977	169,103	141,479
Hbb-b1				123,337	101,920
Hbb-b2				159,853	130,309
Slc25A37				121,810	104,857

Table 1 Single and double capture statistics

This shows data from 5 experiments with small numbers of target genes (range 1-3). The data compares single with double capture in the same samples. Note that in these experiments the key statistic is the number of unique interactions from each viewpoint. These depending on which target is used, as the oligonucleotides have different capture efficiencies but the main determinant is the library complexity.

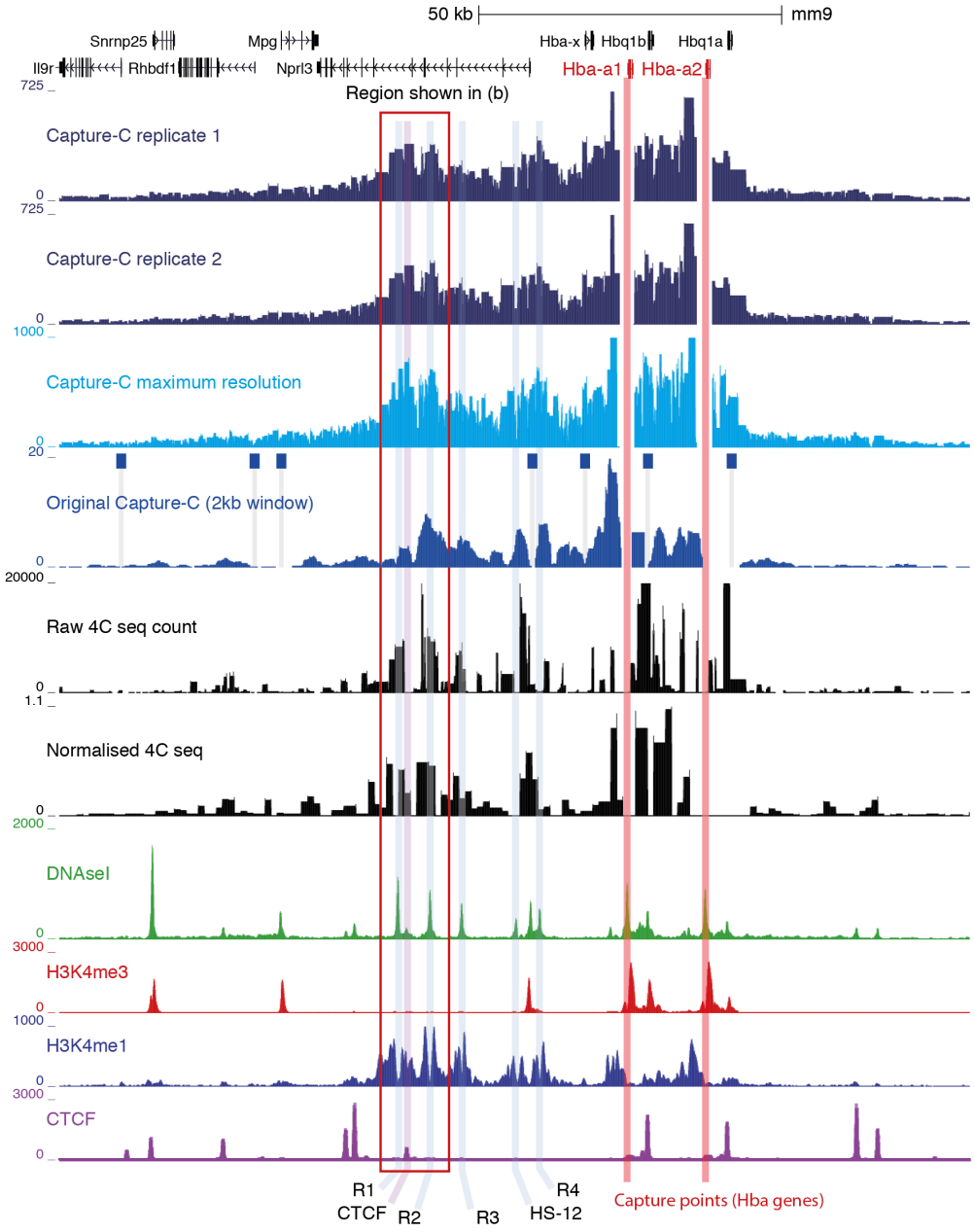


Figure 2.5 Comparison with 4C-seq

Comparison of two biological replicates from the alpha globin promoters (Hba-a1&2 highlighted in red) with the original Capture-C method (Hughes et al., 2014) and 4C-seq (van de Werken et al., 2012b). The top two tracks are from biological replicates showing the high degree reproducibility between samples.

The third track shows the maximum possible resolution for 3C-seq based experiments with interactions mapped to the half of the restriction fragment which contains the midpoint of the interacting read (see below for further details).

The original Capture-C experiment requires a 2kb moving window. The capture sites included in the experiment are denoted by blue boxes. The data from these sites has been excluded.

The raw 4C-seq profile was generated using the Capture-C analysis tools (normal resolution). The PCR duplicate filtering used for all Capture-C data cannot be performed because the amplicons from any particular location are identical which results virtually complete loss of the interaction profile. Below it is the published normalised 4C-seq profile showing the compensation for differences in PCR efficiency.

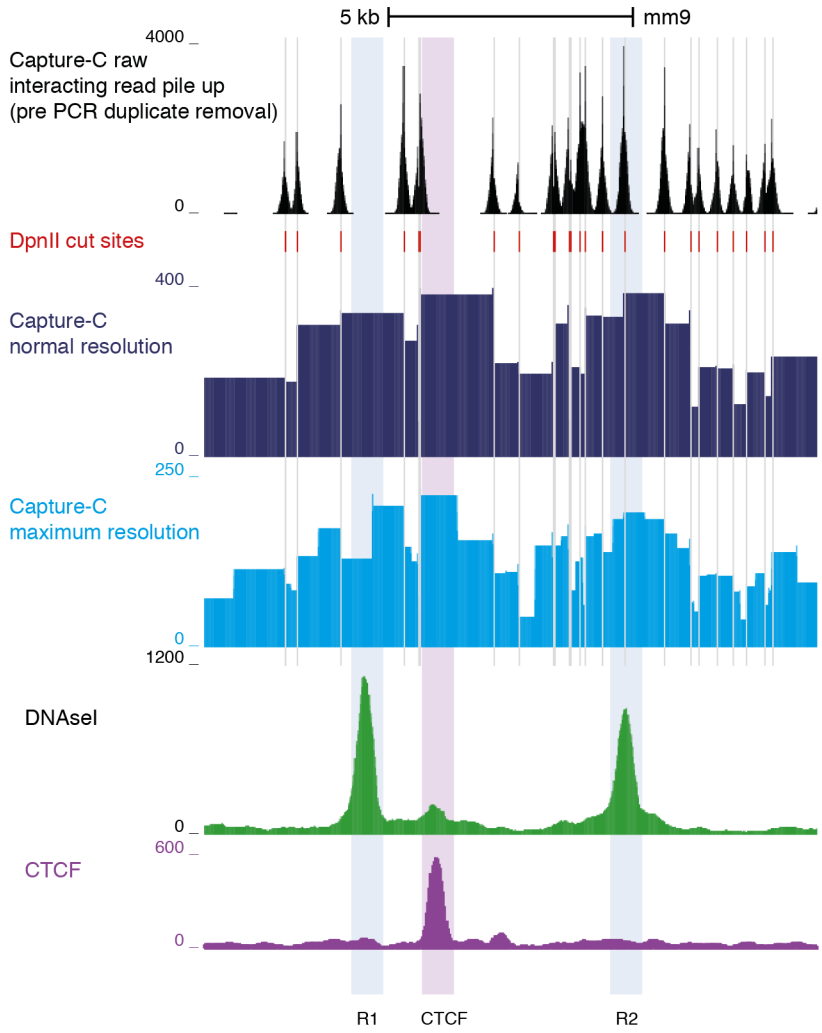


Figure 2.6 Comparison of Capture-C data mapped to normal and maximum resolution

This shows a close up on the two most important regulatory elements in the alpha globin locus (R1 and R2) (as shown by the red box in Figure 2.5). The top profile shows the raw pile up of interacting reads prior to PCR duplicate removal. Note that the interacting reads are almost entirely juxtaposed to the DpnII cut sites (the sharp peaks correspond to the DpnII cut sequence, where the reads on both sides of the junction overlap). Since the reads are so close to the ligation junction and for larger fragments virtually no interacting reads span the middle of the read they can be mapped to the whole restriction fragment (normal resolution) or two the half of the fragment, which contains the midpoint of the read. This potentially provides the maximum possible resolution for 3C experiments.

2.6 Detailed methods

2.6.1 3C library preparation

The protocol used for generation of the 3C libraries was similar to previous methods (Hagege et al., 2007; Hughes et al., 2013; Stadhouders et al., 2013) with minor modifications. 10^7 cells were crosslinked with 2% formaldehyde (10 minutes, room temperature); quenched with cold glycine; washed in phosphate buffered saline; resuspended in cold lysis buffer (tris 10mM, NaCl 10mM, NP40 0.2%, complete proteinase inhibitor (Roche) and snap frozen to -80^{0}C. Cells were thawed on ice, washed in dH_2O and Dounce homogenised on ice (x 40 strokes). Cells were then resuspended with 0.25% SDS and restriction enzyme buffer (1x NEBuffer DpnII) and incubated at 37°C for 1h at 1400rpm on a Comfort Thermomixer (Eppendorf) followed by a further incubation of 1h following the addition of Triton X100 (source final concentration 1.67%). An overnight digestion was performed using DpnII (NEB) at 500U /ml at 37^{0}C agitated at 1400 rpm (Comfort Thermomixer Eppendorf).

The digested chromatin was ligated overnight (Fermentas HC Ligase final concentration 10U/ml) at 16 degrees at 1400 rpm on the Thermomixer. The samples were then decrosslinked overnight at 65°C with Proteinase K (Roche) followed by a 30 minute incubation at 37°C with RNAse (Roche). Phenol/Chloroform extraction was then performed followed by an Ethanol precipitation and a wash with 70% Ethanol.

2.6.2 Quality assurance of 3C libraries

3C experiments require careful controls (Dekker, 2006). The quality of 3C libraries was therefore carefully assessed by gel electrophoresis and PCR.

2.6.2.1 Agaraose gel electrophoresis of 3C libraries

Agarose gel electrophoresis is a standard method for assessing digestion efficiency of 3C libraries. Gel electrophoresis was performed using mini gels (1% agarose in x1 TBE (Tris/Borate/EDTA) and Ethidium Bromide (0.001%)). Approximately equal quantities of the undigested control, digest control and

ligation products were loaded onto the gel with Loading Dye with 1kb ladder (Fermentas). The gels were run at 100V for approximately 90 minutes.

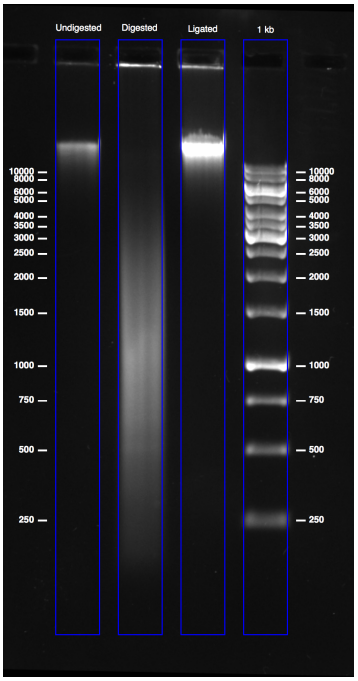


Figure 2.7 Gel electrophoresis to assess digestion efficiency

Note that the main peak of digestion is around 1kb.

2.6.2.2 Real time PCR assessment of digestion efficiency

A real time PCR assay was also developed to assess digestion efficiency. PCR primers and a Taqman probe was designed using PrimerBlast software (Ye et al., 2012). Two sets of primers and probes were used. One set, contained within the alpha globin restriction fragment, was used to control for the concentration of the library. The other spanned the DpnII restriction enzyme cut site, which allowed digestion to be quantified.

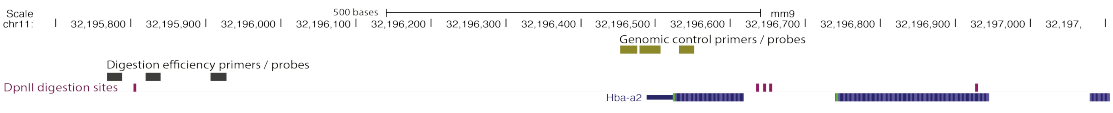


Figure 2.8 Schema of primers used to assess digestion efficiency

Primer sequences:

DpnII digestion control forward primer GTGTCACCAAAACCAGCTCA
DpnII digestion control reverse primer CCTGGAATCCTTTGGCTCAAG
DpnII digestion control Taqman probe GGGCAGCTAAGATGCAAGTC

Hba genomic control forward primer TGGAGGGCATATAAGTGCTACTTG
Hba genomic control reverse primer TGCTTTTGTCTTCCCCAGAGA
Hba genomic control Taqman probe TGCAGGTCCAAGACACTTCTGATTCTGACA

Real time PCR was performed with a total reaction volume of 25 ul (Taqman PCR master mix, Invitrogen) with primers and probes at 200 nM and 100 nM respectively. Real time PCR reactions were performed using the StepOnePlus System (lifetechnologies / Applied Biosystems). Digestion efficiency was calculated using log linear regression of standard curves (Microsoft excel) of genomic DNA for both the genomic and digestion primers.
A threshold of 70% digestion was used for the libraries to be used for analysis but digestion efficiencies were nearly always in excess of 80% when assessed by real-time PCR.

2.6.3 Discrepancies between digestion efficiencies

There was a consistent discrepancy between the digestion efficiency when measured by gel electrophoresis and real-time PCR. This may result from differences in chromatin architecture modifying the accessibility of chromatin to digestion by the restriction enzyme. Since our PCR primers were designed to assess the digestion efficiency at the alpha globin promoter, which is DNAseI hypersensitive in erythroid cells, they are likely to give a higher estimate of the digestion efficiency than the gel electrophoresis, which provides a more global estimate.

2.7 Addition of Illumina sequencing adaptors

Because PCR duplicates are excluded when analyzing Capture-C data it rapidly became clear that the library complexity was one of the key determinant of the number of unique informative interactions in NG Capture-C. Only four ligation junctions can be generated one cell from each viewpoint fragment (one from

either end from each allele). However, only around two informative ligation junctions can be obtained because around half of the sequenced reads that map in part to the capture restriction fragment either do not have an interacting partner or the interaction lies within the 1kb exclusion around the viewpoint. It is therefore important to maximize the amount of 3C library used for the experiment.

Many 3C based protocols recommend only using 500ng of starting material, which corresponds to the DNA from ~100,000 cells (in mouse and human each cell contains approximately 6 pg), so the maximum possible number of informative interactions would be less than 200,000 in total from each view-point if the downstream processes were completely efficient (which they are not).

To maximise the number of unique informative interactions defined, the NG Capture-C protocol was been optimized to use 3C library containing DNA from over eight times as many cells for the capture process. In order to increase library complexity the sequencing library preparation was usually performed in duplicate allowing a total of 10µg of 3C library to be used. Since most applications do not need several hundred thousand fold read depth across the genome, the available reagents for adding sequencing adaptors are not optimized for this purpose. The protocols were modified to minimize the losses by optimizing the SPRI bead clean up steps to use a ratio of 1.8:1 vol/vol with the product of the reaction and size selection steps were avoided. In addition it was found that increasing the percentage of alcohol in the washes to 80% improved the losses. It was also found that increasing the amount of adaptor added to the ligation reaction reduced the possibility of dA tailed sequences ligating to each other rather than the adaptors. Finally, it was found that changing the PCR kit in the final LM-PCR add the indices to the Agilent Herculase II kit markedly improved the yield from this reaction.

In order to maximize library complexity material was mixed from two parallel library preparations was mixed prior to capture. The amount of input into the hybridization reaction was also increased to 2µg from 500ng increasing the complexity 8-fold in comparison to the original Capture-C protocol. Performing the library preparation in parallel meant that there was excess material for the

capture process and this meant multiple captures could be performed from different view points on the same library. The concentration and profile of the fragment size was assessed after the ligation of adaptors and the indexing PCR, which allowed the library complexity to be monitored.

2.7.1 Sonication

5 µg of 3C library (as assessed by Qubit, broad range assay, lifetechnologies) was placed in a Covaris microTube). The material was sonicated to 200bp using a Covaris S220 focussed ultrasonicator with the following settings:

Duty cycle 10%

Intensity 5

Cycles per burst 200

6 cycles each of 60s (use the batch function)

Set mode: Frequency sweeping

Following sonication the material was cleaned using 180 µl Ampure XP SPRI beads. Two washes with 80% ethanol were performed.

The fragment size and concentration were measured initially using the Agilent bioanalyser (DNA 1000) and later the Agilent Tapestation (D1000).

2.7.2 Ligation of adaptors

End repair and dA tailing were performed using the NEBnext protocol (NEBnext DNA library Prep Reagent Set E6000S or E6040S). The main deviations to the protocol were the ratio of the SPRI beads to material being purified was always 1.8:1. This was to minimise the selection of larger fragments and DNA losses. Size selection, with lower ratios of SPRI beads, was not necessary because the oligonucleotide hybridization and bead cleanup steps were very effective at removing primers and adaptor dimers. NEBnext adaptors (NEBnext Multiplex Oligos for Illumina (E7335 or E7500S)) were ligated to the material using the NEBnext reagents and the USER enzyme step was performed to generate adaptor forks.

2.7.3 LM-PCR for addition of Truseq adaptors

Indexed NEBnext primers were added to the adaptor ligated DNA using the Agilent Herculase II PCR kit, which was found to be superior to the NEBnext reagents. Indices were chosen to match the Nimblegen HE blocking oligos and to satisfy the Illumina rules for acceptable indices for low-plex libraries. These PCRs were performed in duplicate.

Material from both pre and post PCR was assessed using the Agilent bioanalyser or tapestation.

2.7.4 Reconstitution of capture oligonucleotides

Capture oligonucleotides were individually reconstituted to a concentration of 2.89 μM (a x10 stock solution was often used). This allowed capture libraries to be made by simply mixing different oligonucleotides with a 1:1 stoichiometry (a total of 13 pM in 4.5 μl was used in the hybridization reactions irrespective of the number of oligonucleotides). The oligonucleotides are over 1 million fold in excess in the hybridization reaction even when 40 viewpoints are being captured. Contamination from different sets of oligonucleotides occurred during the manufacturing process on one occasion and great care was used when reconstituting the oligonucleotides in trying to avoid this.

2.7.5 Hybridisation with capture oligonucleotides

The hybridisation reaction was prepared in a 1.5ml lobind Eppendorf tube. This included the following components:

1-2 μg sonicated 3C library with sequencing adaptors

1000 pM Nimblegen HE universal blocking oligo

1000 pM Nimblegen HE indexed blocking oligos (corresponding to the indices used in the experiment)

5 μg Cot DNA from the appropriate species (to prevent secondary capture from repetitive sequences)

These were dried in a vacuum centrifuge (50 °C) until all of the water had evaporated. 10.5 μl hybridisation buffer and 3 μl component A (Nimblegen Hybridisation and Wash Buffer Kit) were added to the Eppendorf tube and the mixture was carefully mixed with a pipette and vortexed to reconstitute the dried DNA. This mixture was then heated to 95 °C for 10 minutes to denature

the DNA. 4.5 µl biotinylated capture library was concurrently heated to 47 °C in a 0.2 ml PCR tube. After 10 minutes at 95 °C the mixture containing the 3C library was vortexed and centrifuged briefly, to collect the fluid in the bottom of the tube, before being transferred to the PCR tube containing the biotinylated capture library. The hybridisation reaction was incubated at 47 °C for 64-72h.

2.7.6 Extraction of captured DNA

Streptavidin beads were initially prepared with two washes with bead wash buffer (Nimblegen Hybridisation and Wash Buffer Kit). The hybridisation reaction was then added to the prepared beads and incubated at 47 °C for 45 minutes. Following this 100 µl wash buffer I heated to 47 °C was added to the sample. This was mixed quickly by vortexing and the supernatant was separated and discarded from the beads using a magnetic rack. 200 µl stringent wash buffer heated to 47 °C was then added to the beads and a 5 minute incubation was performed at 47 °C. A further wash step was performed with another 200 µl hot stringent wash buffer. Following this the supernatant was discarded and 200 µl room temperature wash buffer I was added. The samples were then vortexed for 2 minutes. The streptavidin beads were again separated using a magnetic stand and the supernatant discarded. 200 µl room temperature wash buffer II was then added and the samples vortexed for 1 minute. Again the samples were placed on a magnetic stand and the supernatant discarded. Finally 200 µl room temperature wash buffer III was added and the samples were vortexed for 30 seconds.

3 Data analysis

A set of tools for design and analysis of Capture-C experiments was developed. An online tool to generate oligonucleotide design for multiple targets was written by Prof Jim Hughes and can be found at http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi. I wrote all of the scripts for analysis of Capture-C data although some minor modifications were made by Jelena Telenius to make them suitable for public release. These are available via github (https://github.com/telenius/captureC/releases).

3.1 Outline of data analysis.

The initial analysis of these data involved multiple steps, as follows (Figure 3.1):

1. The raw data are taken in FASTQ format.
2. Initially, the paired end reads are reconstructed into single sequences using the central area of overlap to align the sequences.
3. Each read is split *in silico* using the restriction enzyme recognition sequence. This splits the reads into its component restriction fragments and the read name defined in the FASTQ file is used to link sets of fragments from the same read.
4. Reads are mapped to the genome using standard aligner (bowtie). Reads that do not contain a sequence that maps inside the captured viewpoint restriction fragment are discarded.
5. PCR duplicates are excluded by removing reads that are not unique (based on the genomic position of the sonicated ends).
6. Reads are mapped to the restriction fragments. Interactions are only reported when one component of a read pair maps completely within a captured fragment and the other maps outside all of the capture fragments and all the proximity exclusion regions in the experiment.

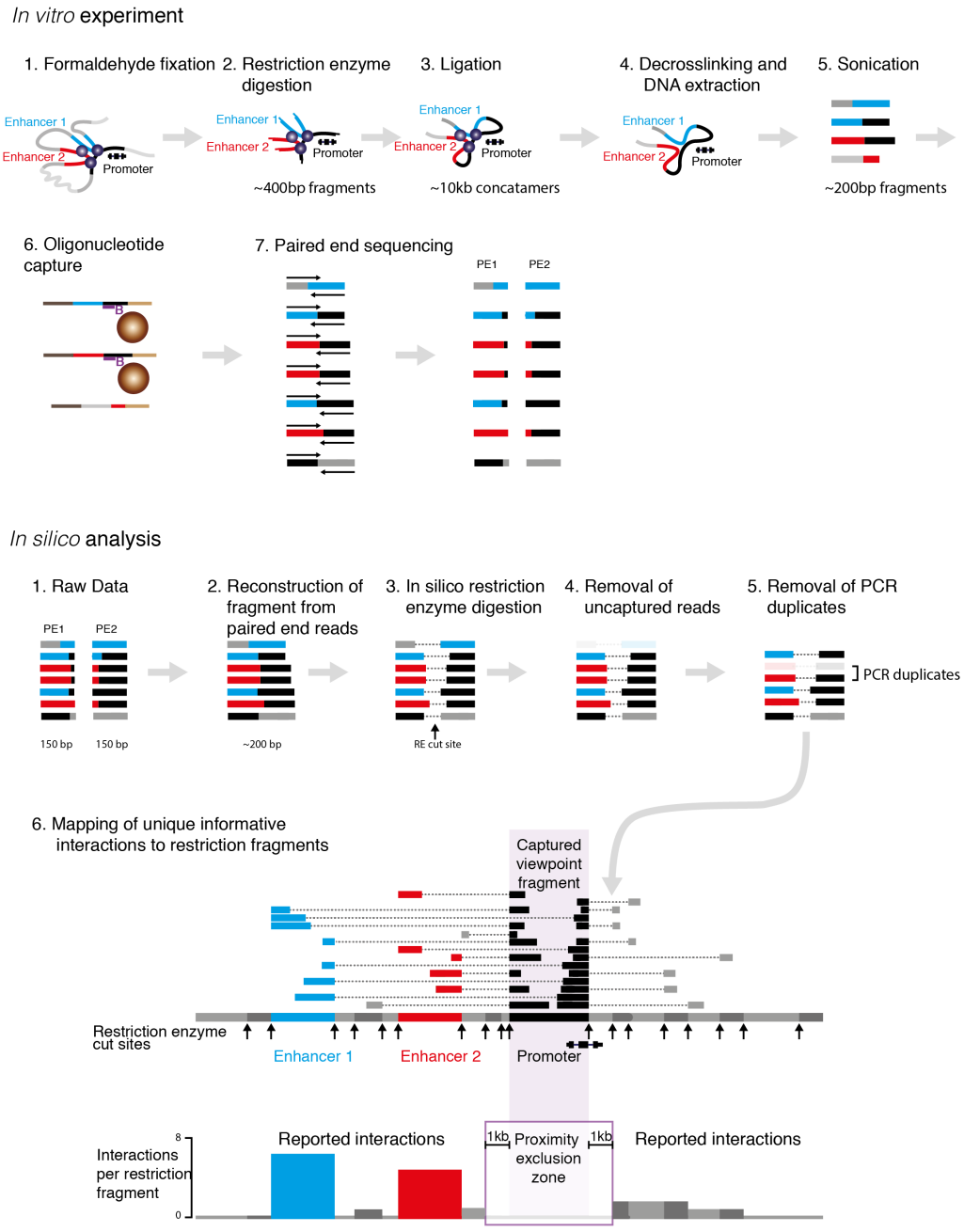


Figure 3.1 Outline of initial data analysis

3.1.1 Reconstruction of reads

Initially the adaptor sequences are removed from the reads in the raw FASTQ files using Trim_galore (a wrapper tool around Cutadapt and FastQC; Babraham Institute). This is particularly necessary when using longer reads for sequencing because when the sequenced read is longer than the original fragment of DNA

the adaptor on the end is also sequenced. These adaptor sequences need to be removed otherwise the sequence does not align to the genome properly.
For experiments that were sequenced using the Illumina MiSeq it was possible to use 150bp paired end sequencing (i.e. 150bp was sequenced from each end – 300bp in total). It was possible for 95% of the DNA sequences to be completely reconstructed using the central area of overlap that was sequenced from both ends because the material was sonicated into 200bp fragments, which were then sequenced with 300bp reads (150bp paired end). The paired end reads were reconstructed into single reads, where possible using FLASH with interleaved output settings (Magoc and Salzberg, 2011).
This allowed us to determine that the ligation event occurred at the restriction enzyme cut sequence (which is normally inferred in most 3C seq based experiments).

3.1.2 In silico restriction enzyme digestion

This was performed by the script dpnII2E.pl. This script takes in a FASTQ file format and splits the reads at a specific restriction enzyme cut sequence. The reads are reconstructed to include the maximum restriction enzyme sequences to aid later alignment i.e. the sequences on the ends contain one cut sequence whereas the middle fragments have a sequence on both ends.

Thus the fragment:
XXXXX-GATC-YYYYY-GATC-ZZZZZ

Would be split into three fragments as follows:
XXXXX-GATC
GATC-YYYYY-GATC
GATC-ZZZZZ

The code also ensures that the quality scores are correctly included and adds to the read name specified by the aligner so that the fragments can be identified back to the original read later.

3.1.3 Alignment to the genome

The resulting FASTQ file of read fragments was then aligned using bowtie1 (using P1, M2, best and strata settings). M2 (which allows reads to be aligned where the sequence occurs twice in the genome) was used in these experiments because both the Hba and Hbb genes have duplicates in the murine genome (using the more stringent M1 settings would have caused all of the reads mapping to these promoters to be discarded).

Composite fragments that result from non-specific ligation will not be mapped to the genome by bowtie 1 and are therefore discarded at this point.

3.1.4 Main analysis script

The resulting SAM file is then analysed with the main script CCanalyser2.pl. The flow diagram for this complex script is depicted in Figure 3.2.

This classifies the sub-fragments as either being: a) "capture" if they are contained within the capture fragment; b) "proximity exclusion" if they are inside the defined proximity exclusion coordinates but outside the capture coordinates or c) "reporter" if they are outside of all of the capture and proximity exclusion regions in the entire experiment. The proximity exclusion zones are normally set at 1kb on either side of the captured viewpoint fragment so that undigested material cannot be falsely reported. This is done to prevent undigested material being reported as interacting and to prevent interactions being falsely reported from fragments that could be captured by two different oligonucleotides.

PCR duplicates were excluded by removing reads that had the same start and end coordinates of each sub-fragment. For long-range *cis* and *trans* analysis more stringent filtering was performed so that the start and stop coordinates of the interacting read also had to be unique. This more stringent filter was used to remove PCR duplicates because occasionally sequencing errors in the captured restriction fragment allowed PCR duplicates to appear unique. Unique interactions were only reported when the read was unique and there were one or more "reporter" and a single "capture" sub-fragment defined from a single read.

Due to the depth of the sequence data obtained following double capture, unique interactions can be reported for each individual restriction fragment or half fragment (Figure 2.6), which is the highest possible resolution for such experiments; there is no requirement to integrate data by using a moving window. This script can map the reads either to the whole restriction enzyme fragment or, to give the maximum resolution possible, they can be mapped to the half fragment based on the mid point of the read and restriction fragment.

The script is also capable of creating SNP specific tracks, in which a specific base has to be present at a specific position in the capture fragment for the data to be included.

The script outputs data into a variety of different formats including BedGraph, GFF and WIG. In addition it generates a track hub for the UCSC genome browser.

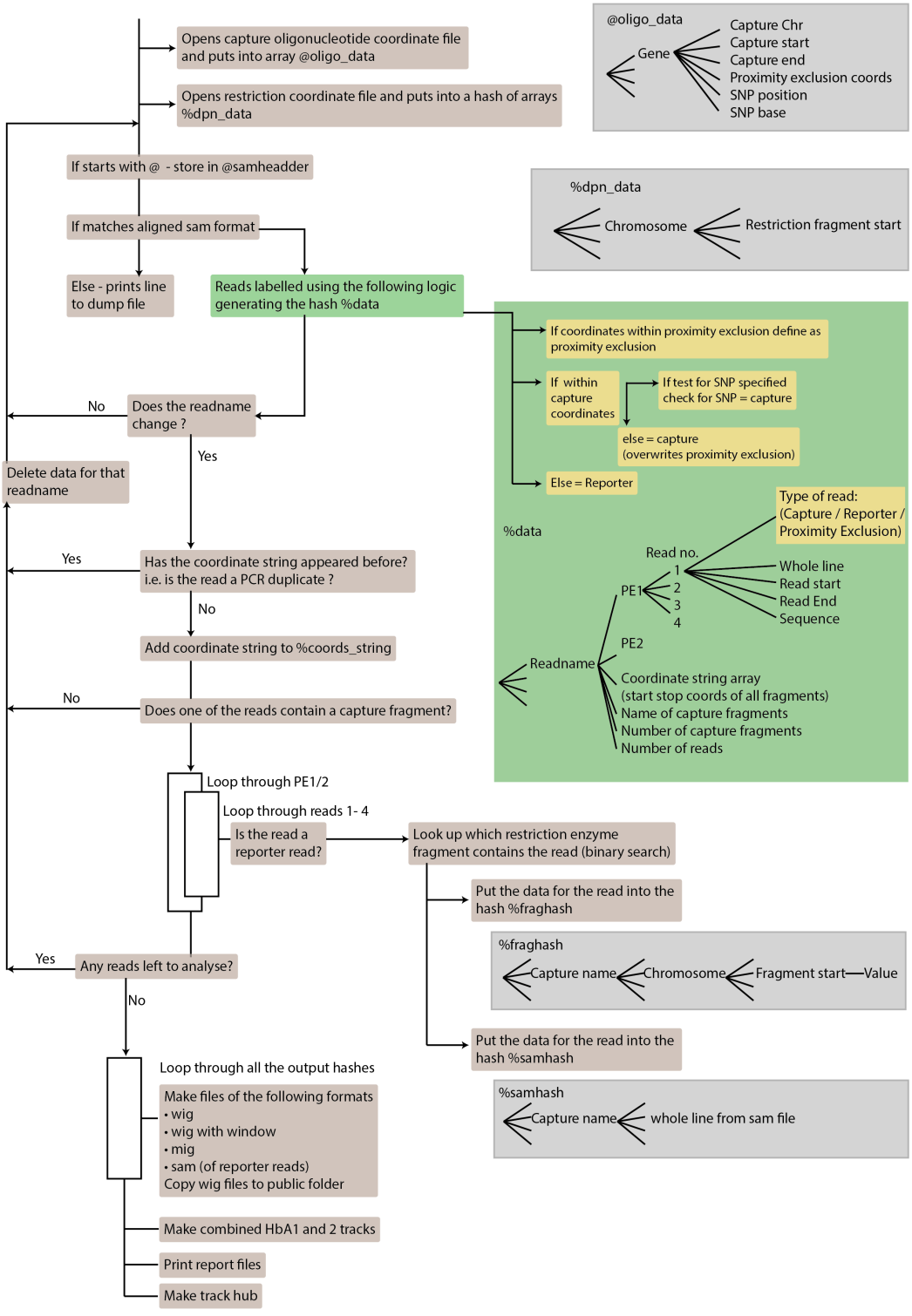


Figure 3.2 Script flow diagram (CCanalyser2.pl)

Script flow diagram for the main analysis script (CCanalyser2.pl). The script requires 3 input files:

1. A file of the input coordinates of the capture oligonucleotides it needs to in the following 9 column tab separated format. (1. Name of capture

(avoid spaces in the names please); 2. Chromosome of capture fragment 3. Start coordinate of capture fragment; 4. End of capture fragment; 5. Chromosome of proximity exclusion; 6. Start coordinate of proximity exclusion; 7. End coordinated of proximity exclusion; 8. Position of SNP; 9. Base of SNP)

2. A file of all of the restriction enzyme fragments in the genome being used (made using a script called dpngenome.pl)
3. A sam file generated by bowtie1 with the pre processing described above

The script makes extensive use of complex data structures in perl. In particular it uses hashes, which are a perl specific data structure that allows the programmer to add multiple values to each element in the hash. Outlines of the data structures are shown in the grey boxes.

The script labels each fragment as either being 'Capture', 'Proximity exclusion' or Reporter using the logic shown in the green box. The name of the captured target is used so that each captured target is handled separately and genome wide profiles are created for each captured target. The proximity exclusion zones and capture sites of all targets are removed from all of the profiles to exclude peaks that could be caused by oligonucleotide capture from two targets.

3.1.5 Downstream processing

The data are then filtered to remove regions with problematic mappability due to copy number differences (Kowalczyk et al., 2012). Further filtering was performed to remove mis-mapped reads from the proximity exclusion region. This was done by using BLAT to map the sequence of the proximity exclusion zone back to the genome so that reads, which can potentially misalign to other parts of the genome, could be removed. This was important for analysis of alpha globin, because there is a third copy of the gene on chromosome 17 in the mouse, which appears to have interactions due to mis-mapping.

3.1.6 Normalisation of data

Further analysis was performed in R. The data were normalized using the total count of unique interactions across the genome. Usually this was scaled to a total of 100,000 interactions across the genome. This allowed subtraction of one track from another, which was also done using R.

Code was written to take in the data table outputs from R and output these into a track hub for the UCSC genome browser.

3.1.7 DESeq2 analysis

When there were multiple replicates the data were also analysed using common approaches for the differential analysis of NGS count based data, such as the Bioconductor package DEseq2 (Love et al., 2014) (See Chapter 4 Multiplexed Capture-C). The analysis in R including the analysis by DESeq2 was automated using in Perl using the module Statistics::R. This allowed data from hundreds of interaction profiles to be analysed rapidly.

3.1.8 Circos

Genome wide heat maps and looping glyphs were produced using Circos (e.g. Figure 5.6). Scripts were written in Perl to process the data into the correct format for input into Circos.

4 Multiplexed Capture-C

The substantial increase in signal afforded by next generation Capture-C allows multiple 3C libraries (e.g. from different cell types) to be indexed and pooled prior to capture. This greatly increases the throughput of the assay, and importantly allows biological replicates and different experimental conditions to be processed and analysed together. This removes some sources of technical variation between experiments and enables the comparison of profiles. This, in combination with the removal of over amplified PCR products produces an assay that is highly reproducible across biological replicates even when considering very weak long-range cis interactions.

To demonstrate the power of the NG Capture-C approach we scaled up to a 35 gene design and increased the number of samples analysed in a single experiment, capturing seven pooled indexed libraries in a single assay. The pooled libraries were generated from three independent samples of mouse ES cells and four samples of erythroid cells (two from C57bl6 mouse strain and two from F1 C57bl6/CBA crosses). The efficiency that resulted from the double capture step, allowed us to sequence these 245 interaction profiles using a single Illumina Hi-seq run (177 million reads).

4.1 Identification of tissue-specific regulatory elements by direct comparison of chromosome conformation in different cell types

The reproducibility of profiles generated from co-captured libraries enabled us to compare differences in regulatory interactions between different cell types. We observe that a gene promoter will interact with the surrounding genome to some extent whether it is active or inactive, therefore by determining the differences between interaction profiles of different cell types it is possible to reveal the interactions that are specific to a given tissue.

4.2 Identification of erythroid specific regulatory elements using ES cells as a control

After normalization of the profiles of the erythroid-specific globin genes and mitoferrin 1 from erythroblasts (active) and ES cells (inactive), subtraction of these profiles highlights the known regulatory elements of the genes with much greater precision than is possible with only the active profile.

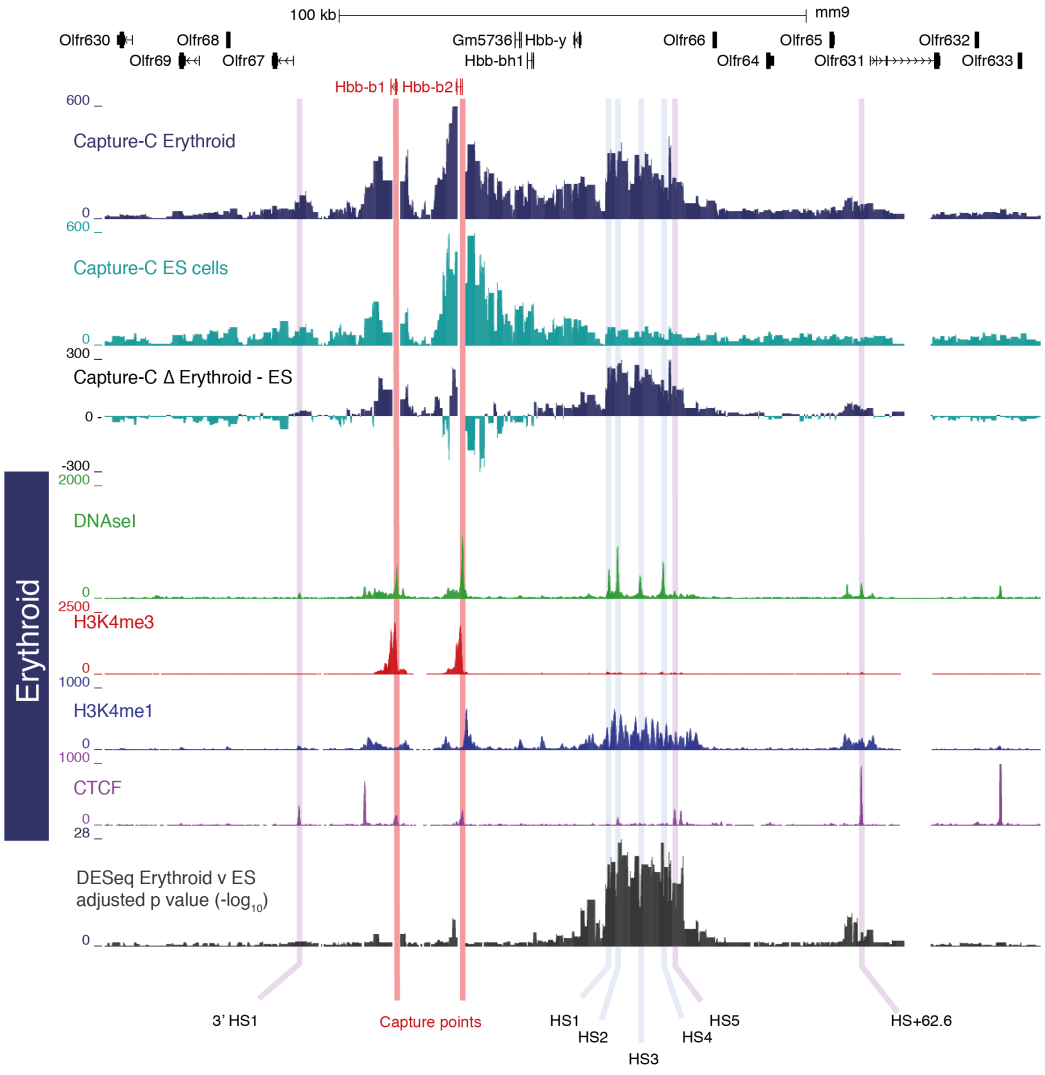


Figure 4.1 Capture-C at the beta globin locus

Capture-C data at the beta globin (Hbb-b1&2) locus. There are two copies of the gene and so the interaction profiles are a composite of the profiles from the two promoters. The Δ Capture-C profile clearly picks out interactions with the very well described regulatory elements (HS1, HS2, HS3, HS4 and HS5). In addition there are clear interactions with the CTCF site at HS+62.6 and an additional erythroid specific site just upstream of this. The DESeq analysis highlights these

regulatory elements precisely with p values below 10-20 over multiple adjacent fragments.

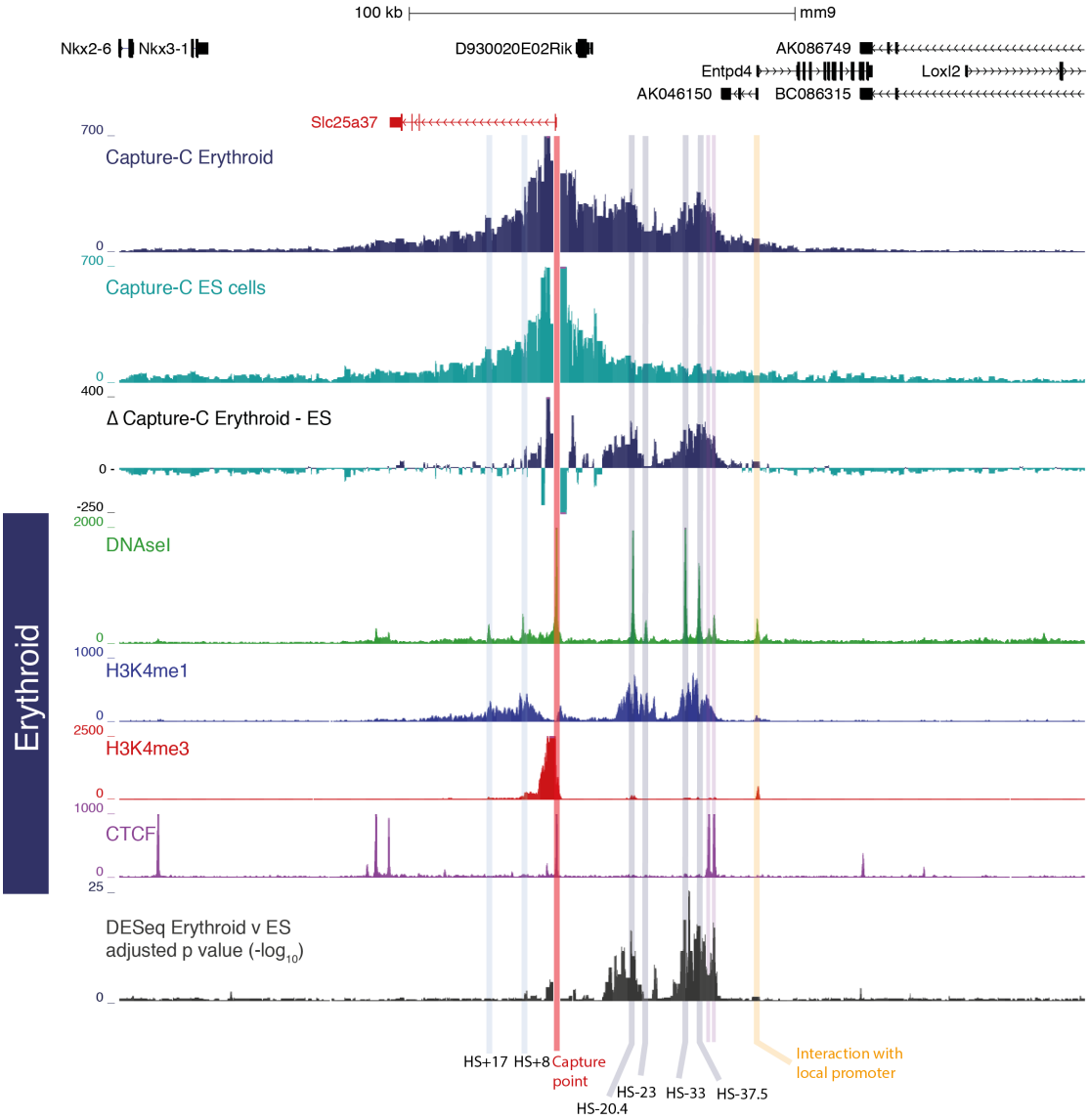


Figure 4.2 Capture-C from the mitoferrin 1 (Slc25A37) gene promoter.

Capture-C from the mitoferrin (Slc25A37) gene promoter demonstrating interactions with the previously identified regulatory elements (Hughes et al., 2014). Note that the differential Capture-C track specifically highlights the two blocks of active regulatory elements and that the DESeq analysis only highlights these interactions as being statistically significant.

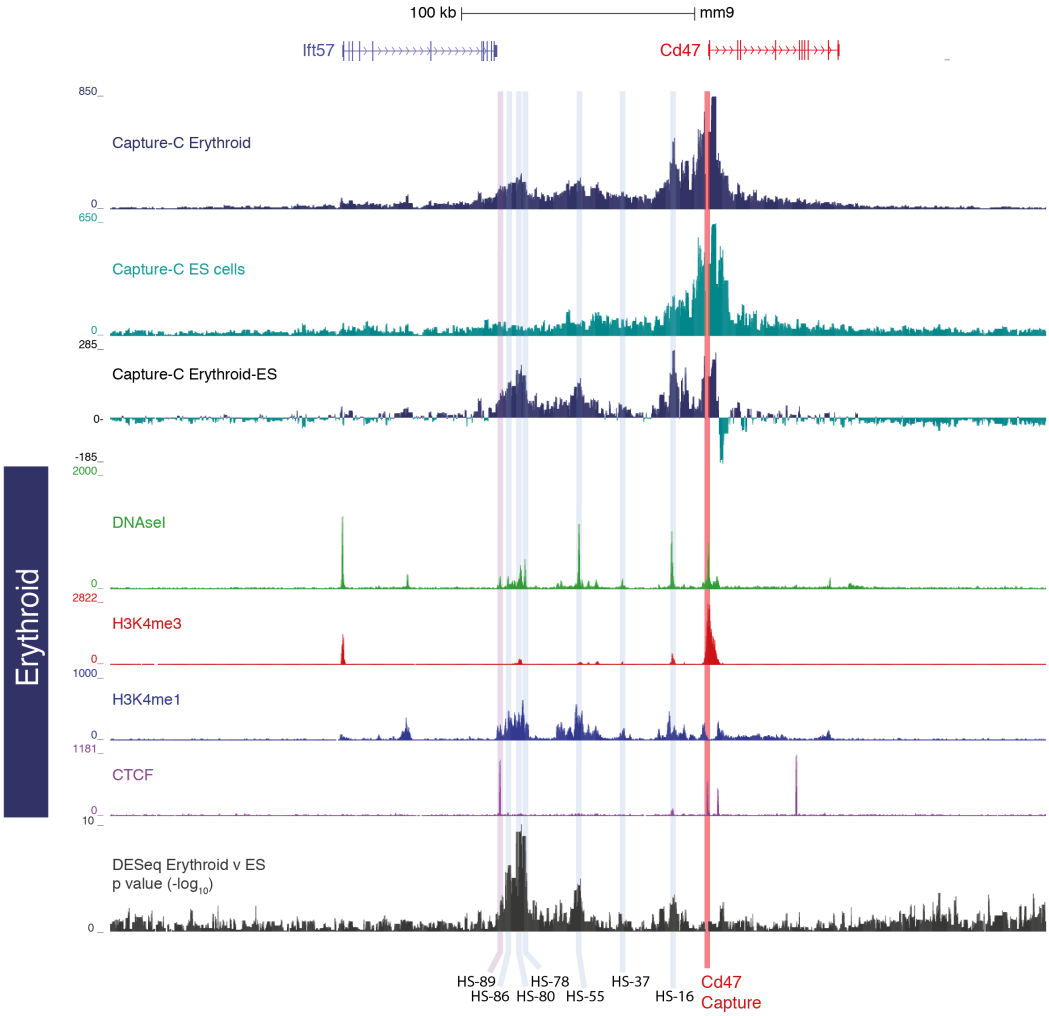


Figure 4.3 Capture-C data from the Cd47 promoter

Capture-C data from the Cd47 promoter showing erythroid specific interactions with 7 elements over 128kb. The -log10 p value track was generated using DESeq2 and shows the p value for each restriction fragment comparing differences between erythroid and ES cells. Interestingly comparison of the off and on states shows that there is little change in the interaction profile over the gene body itself.

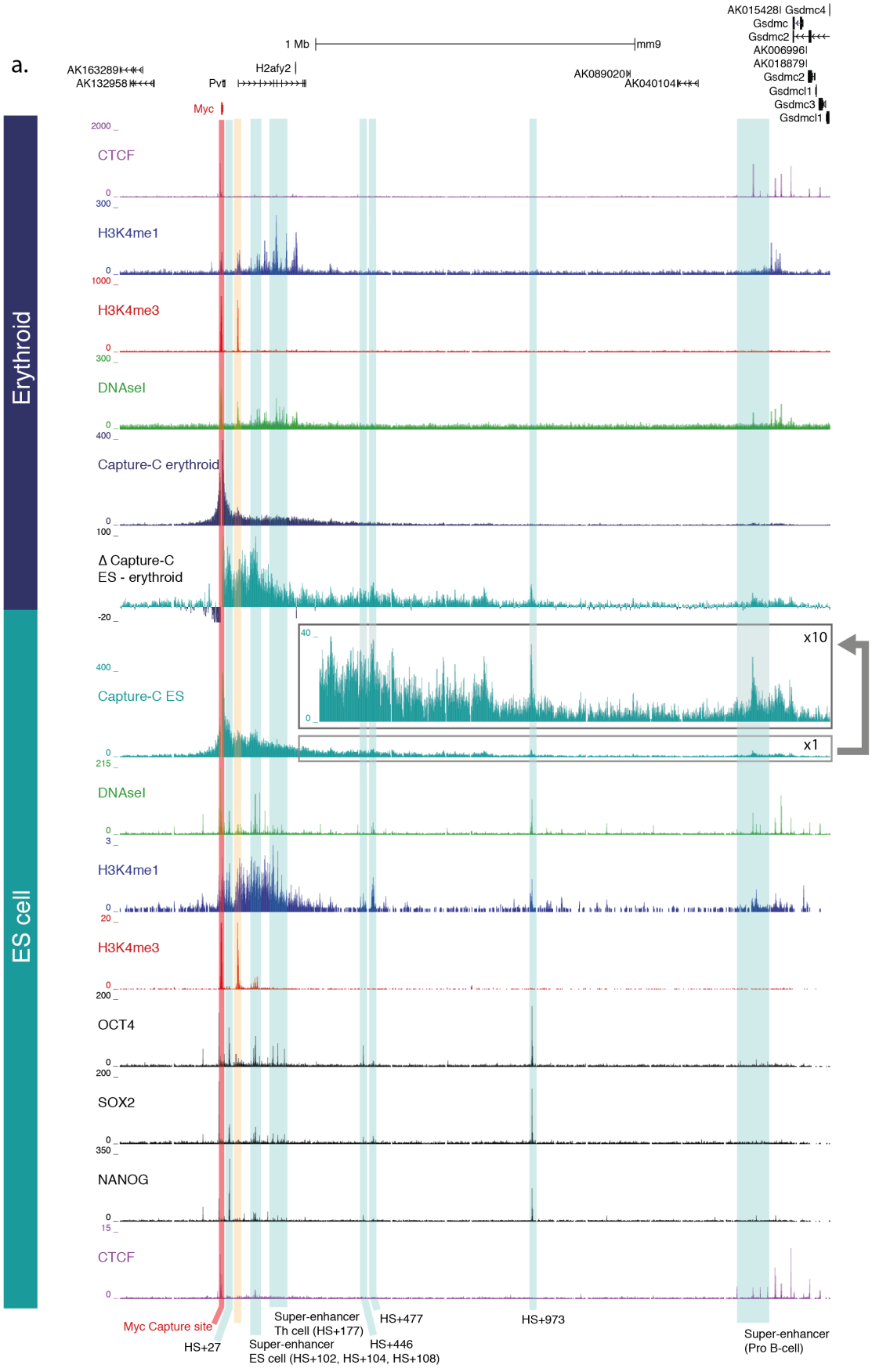
4.2.1 Interaction with the gene body on activation of transcription

Interestingly, changes in the interaction profile were not generally seen over the body of the gene when comparing the profiles between ES cells and erythroblasts at large genes such as mitoferrin and Cd47 (see above). This is in contrast to previous studies, which have proposed that active genes form 'gene-

loops' between the promoter and end of the gene, which facilitate the reloading of polymerase to the promoter and prevent the production of aberrant ncRNAs (Tan-Wong et al., 2012).

4.3 Identification of ES cell specific regulatory elements

This type of subtractive analysis was also performed for ES specific genes (Pou5f1 (Oct4), Sox2, Klf4, Myc and Nanog) using the erythroid data as the inactive profile. Impressively, at loci such as Myc we demonstrate interactions with regulatory elements over 1Mb from the promoter (Figure 4.4).


a.
1 Mb
mm9
AK163289
AK132958
H2afy2
PvuII
AK089020
AK040104
AK015428 Gsdmc4
Gsdmc
Gsdmc2
AK006996
AK018879
Gsdmc2
Gsdmc1
Gsdmc3
Gsdmc1
Myc
Erythroid
ES cell
CTCF
H3K4me1
H3K4me3
DNAseI
Capture-C erythroid
Δ Capture-C
ES - erythroid
Capture-C ES
x10
x1
DNAseI
H3K4me1
H3K4me3
OCT4
SOX2
NANOG
CTCF
Myc Capture site
HS+27
Super-enhancer
Th cell (HS+177)
Super-enhancer
ES cell (HS+102, HS+104, HS+108)
HS+477
HS+446
HS+973
Super-enhancer
(Pro B-cell)

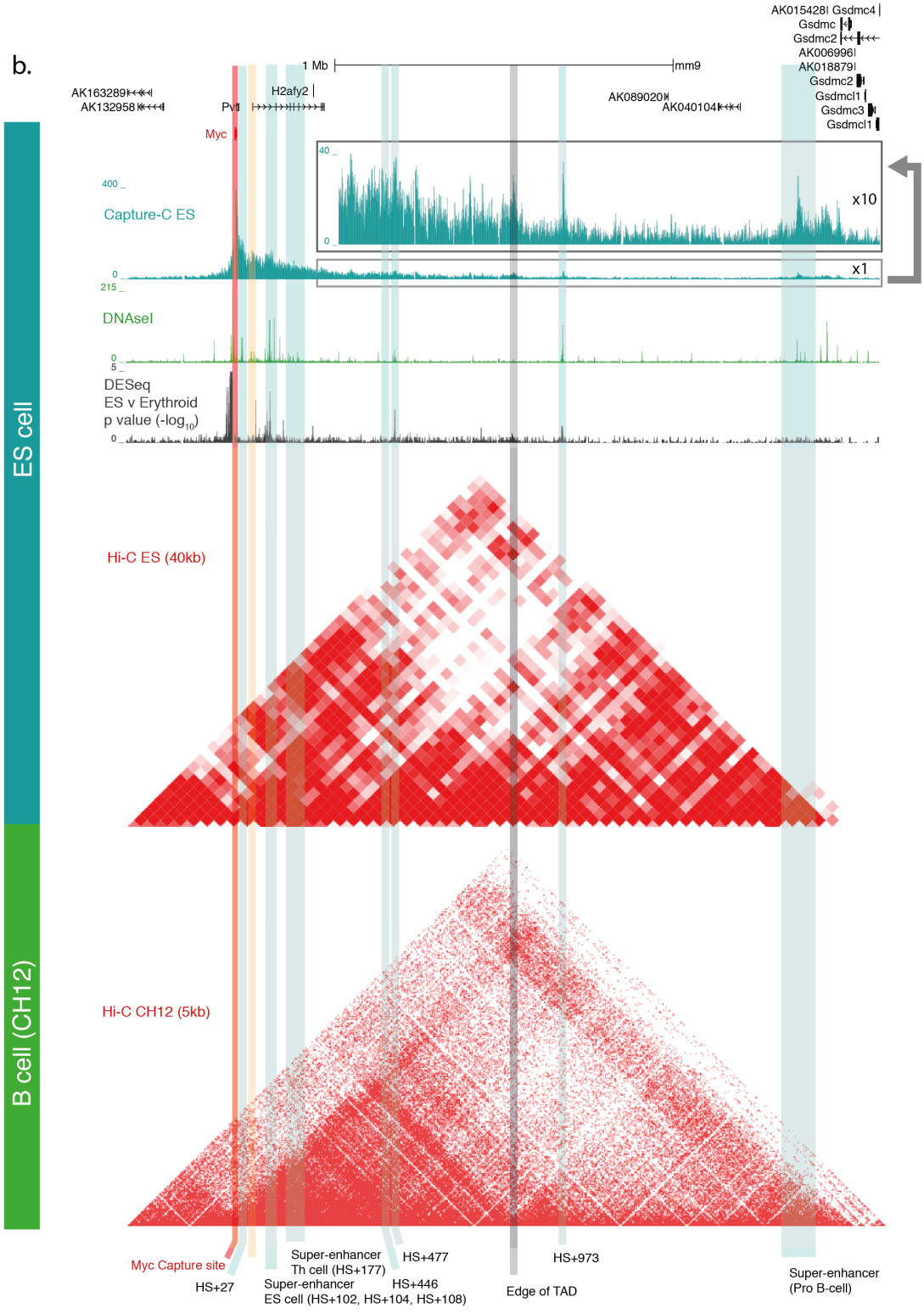


Figure 4.4 Capture-C at the Myc locus.

a. Shows the comparison between erythroid and ES capture C data from the promoter of the Myc gene. The inset box shows the ES cell specific data with a x10 fold change in scale on the Y axis. This complex locus has three different

annotated superenhancers (Whyte et al., 2013). Interestingly not only does the promoter interact with the ES cell specific super -enhancer but it also interacts with a super-enhancer called in Th cells as well as one described in Pro-B cells, which is well over 1Mb away from the gene. In addition there is a very well localised interaction with an ES cell specific enhancer at HS+973.

Panel b. shows the -log10 adjusted p value showing statistically significant differences between the erythroid and ES cell profiles (calculated by DESeq2). Below is a comparison with between NG Capture-C and two Hi-C data sets from ES cells at 40kb resolution and CH12 cells (a B cell derived line) at 5kb resolution, which is one of the highest resolution Hi-C datasets available at present. These data show that the TAD containing the promoter does not contain all of the regulatory elements. Our data and the high resolution Hi-C data sets show clear interactions extending outside of the TAD with more distal regulatory elements.

Erythroid data: Hughes et al., 2014. ES data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340). Hi-C data obtained from Yue Lab (http://www.3dgenome.org) ES cells (Dixon et al., 2012), Hi-C CH12 cells (Rao et al., 2014)

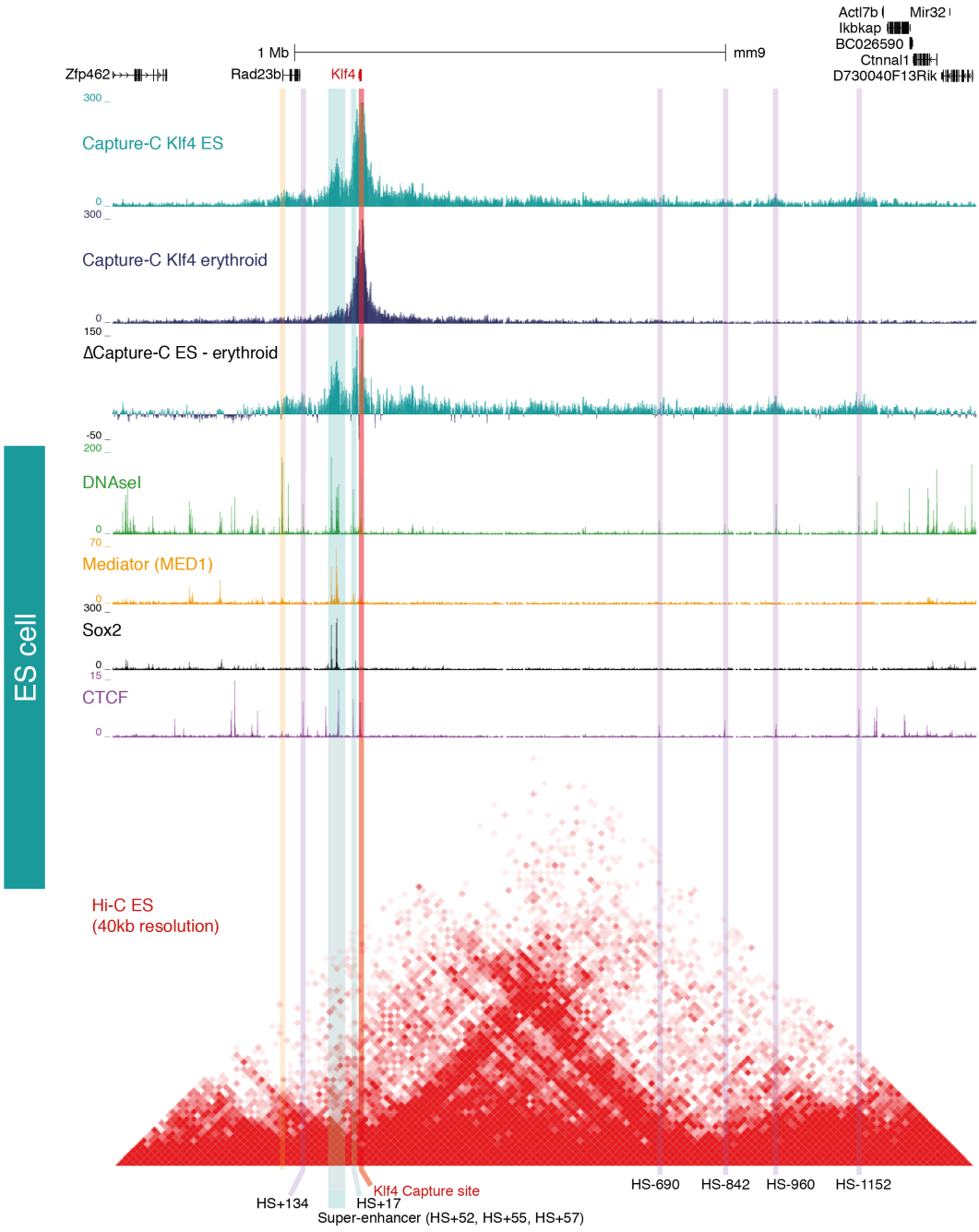


Figure 4.5 Capture-C at the Klf4 locus in ES and erythroid cells.

There is clear interaction with the previously identified ES cell specific group of enhancers between HS+52 and HS+57. In addition there are interactions extending over 1Mb on the other side of the gene with sites bound by CTCF. Interestingly the Hi-C data confirms the domain of interaction between the Klf4 promoter and the CTCF sites in the middle of the gene desert.

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259),

Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340). Hi-C data Dixon et al., 2012.

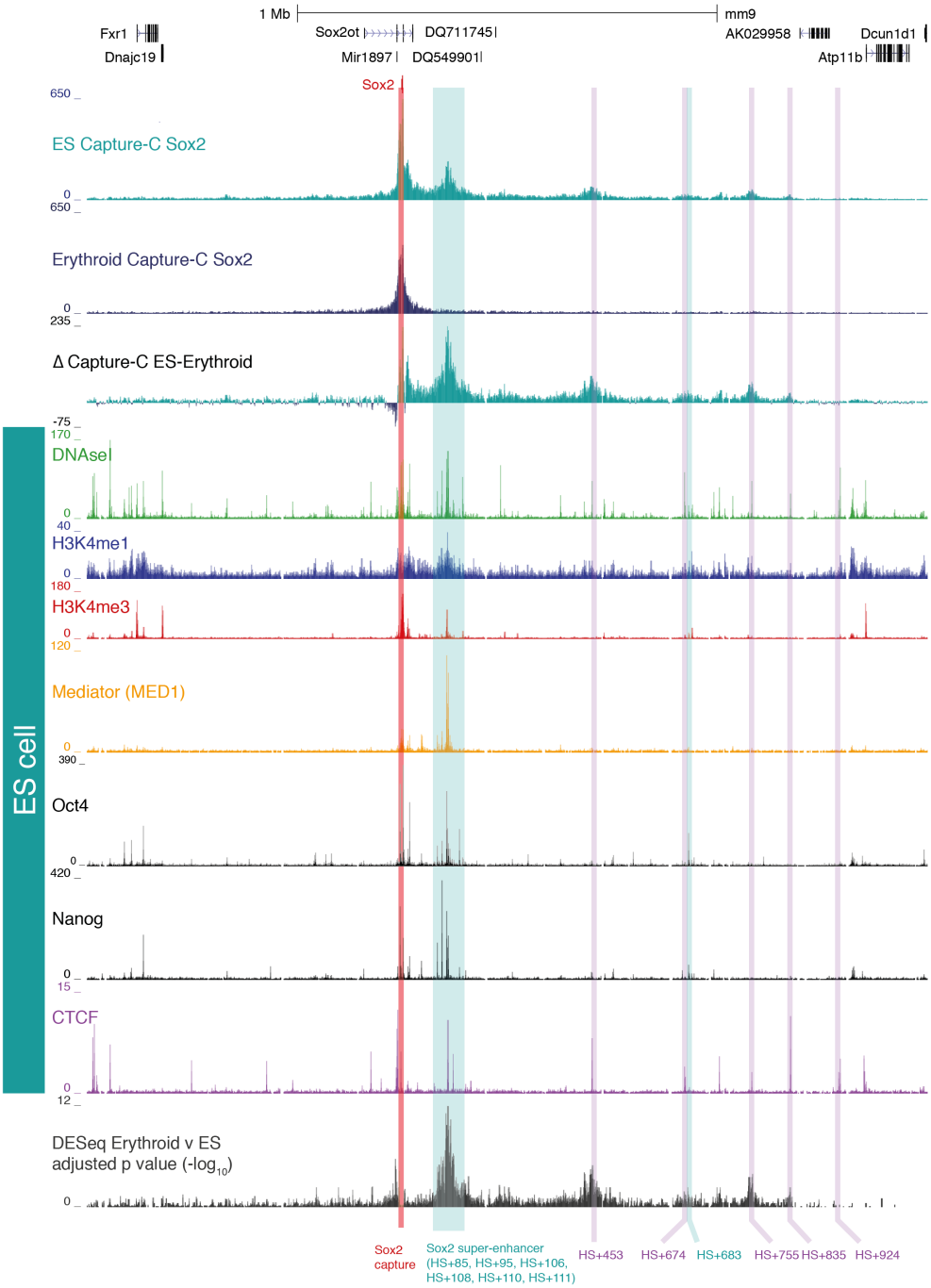


Figure 4.6 Capture-C from the Sox2 promoter in ES and erythroid cells.

This shows that in addition to interactions with the ES cell superenhancer identified by Whyte et al., 2013 the gene has several other cell type specific interactions extending nearly 1Mb away from the promoter. These tend to be with CTCF sites (highlighted in purple) although there is an additional element potential regulatory element bound by Nanog and Oct4 (HS+683).
Analysis using DESeq2 shows that there are significant differences in the interaction profile between erythroid and ES cells extending over 750kb from the promoter.

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340)

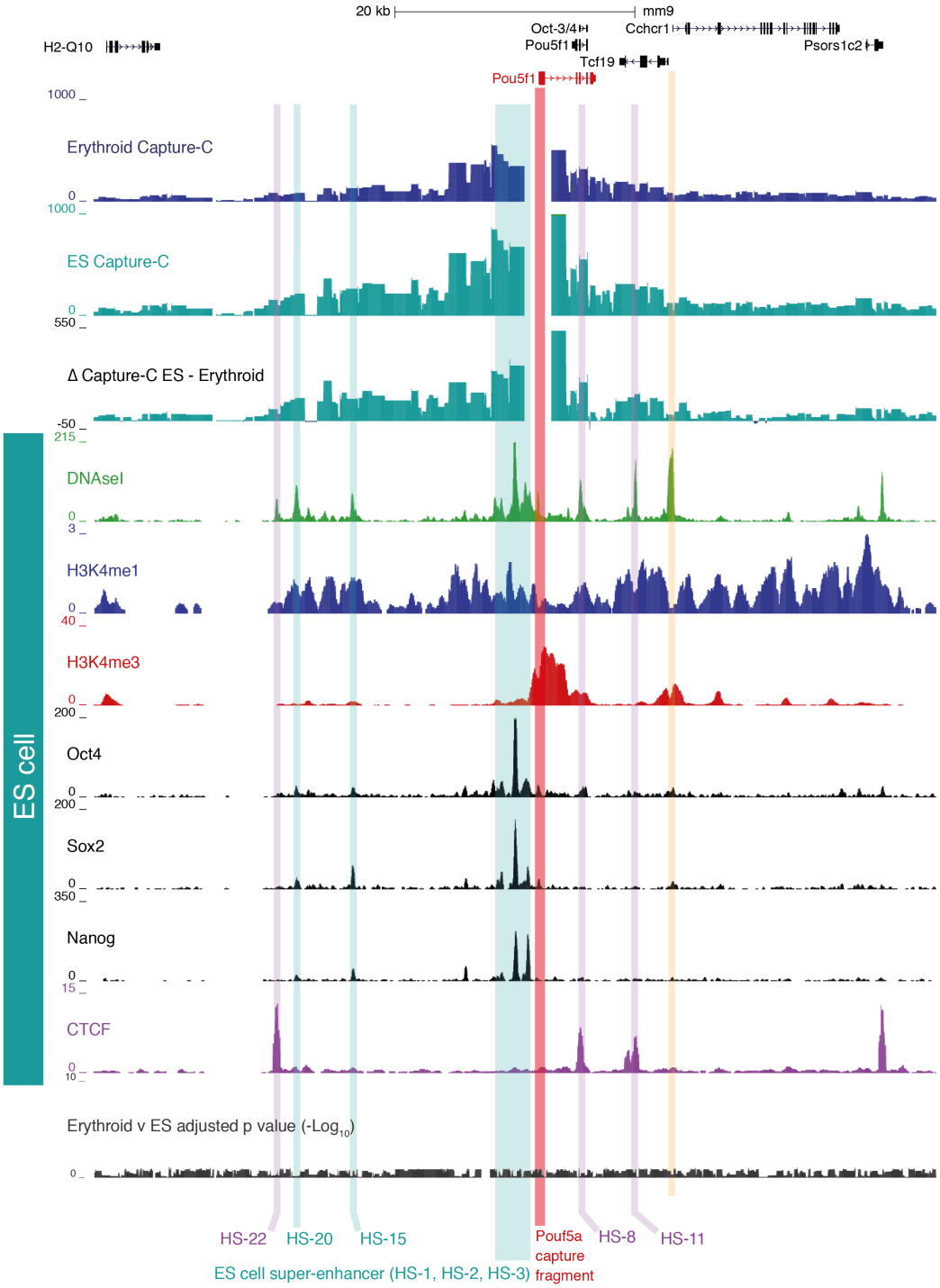


Figure 4.7 Capture-C from the Pou5f1 (Oct4) promoter.

There are clear interactions with the regulatory elements adjacent to the promoter, which meet criteria for a super-enhancer in ES cells. In addition the domain of interaction extends a further 22kb upstream to a CTCF binding site at

HS-22. There are interactions with two further potential regulatory elements in this domain at HS-15 and HS-20 that are DNAseI hypersensitive and bound by ES cell specific transcription factors. The differential analysis using DESeq dose not reveal any statistically significant differences between the interaction profiles in erythroid and ES cells at this locus because the shape of the interaction profile in erythroid cells is very similar to that of ES cells. This highlights a potential limitation of this approach to detect significant changes in the interaction profile, when the regulatory elements are close to the gene of interest. However, given that the theoretical interaction frequency between two particles joined by a string is related to the inverse cube of the distance between them it is highly likely that elements within 20kb of a gene will interact strongly irrespective of whether the gene is active or not.

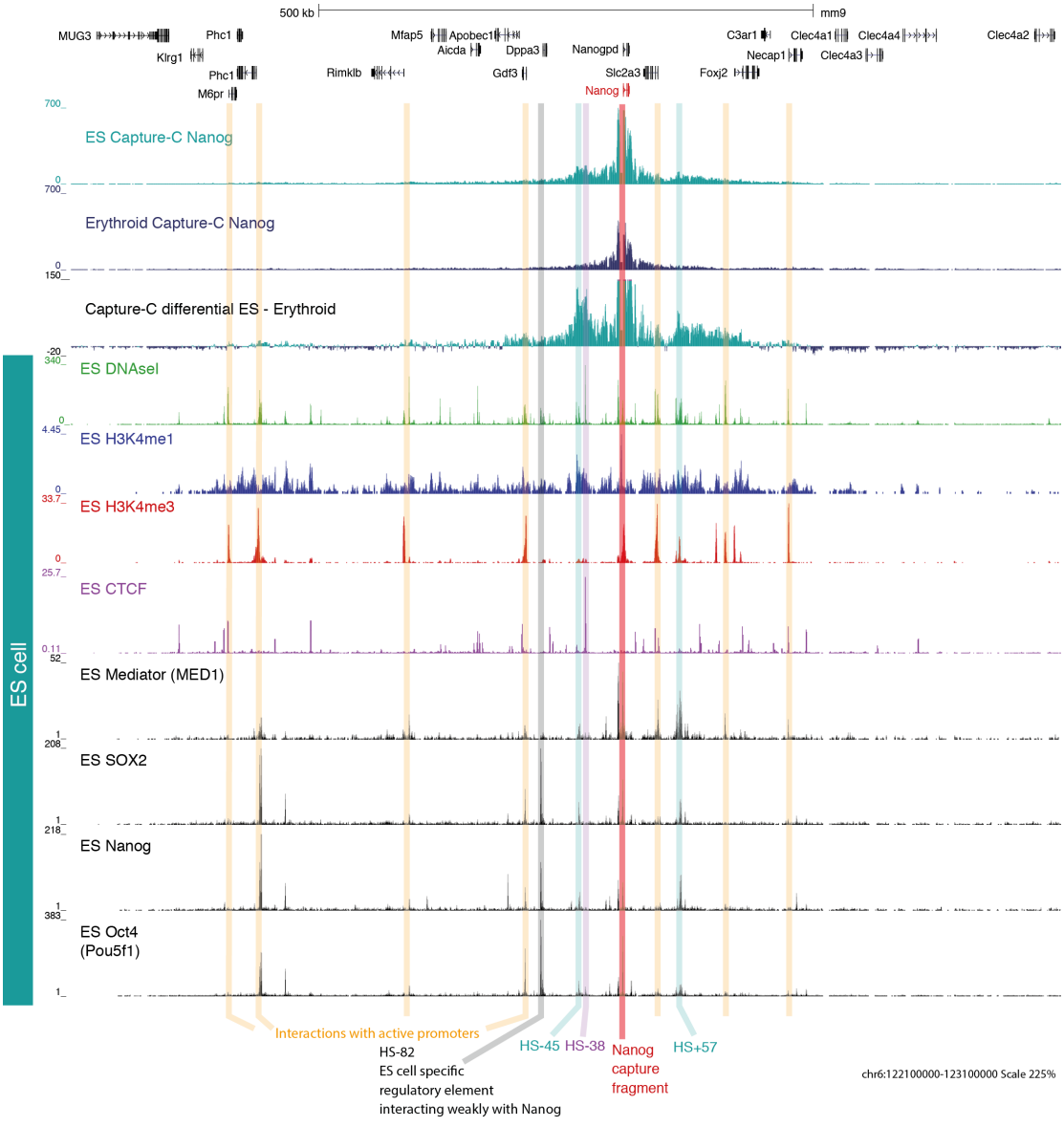


Figure 4.8 Capture-C from the promoter of Nanog

Capture-C from the promoter of Nanog showing interactions with a regulatory element at HS-45. There is an additional regulatory element (HS-82) that is bound strongly by Oct4, Nanog and Sox2 in ES cells, however, this only interacts weakly with the promoter compared to the other elements and we would predict that it is less likely to be important in regulating the gene than the other elements.

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340).

4.3.1 Identification of active elements in both cell types simultaneously

This type of analysis also uncovered fine details of tissue-specific regulation of genes that are active in both cell types. For example, the PNPO gene encodes Pyridoaxime 5'-phosphate Oxidase which is a rate limiting enzyme in the metabolism of vitamin B6 to produce pyridoxal 5'-phosphate; an essential cofactor in the heme synthetic pathway. This gene is upregulated in mouse erythroid cells under the influence of an erythroid-specific regulatory element HS-26 (Figure 4.9) (Hughes et al., 2014). Comparison of co-capture data from ES and erythroid data highlight precisely and specifically the element known to increase expression in erythroid tissues at a resolution sufficient to distinguish it from the promoter of Cdk5rap3, which is located only ~1kb away.

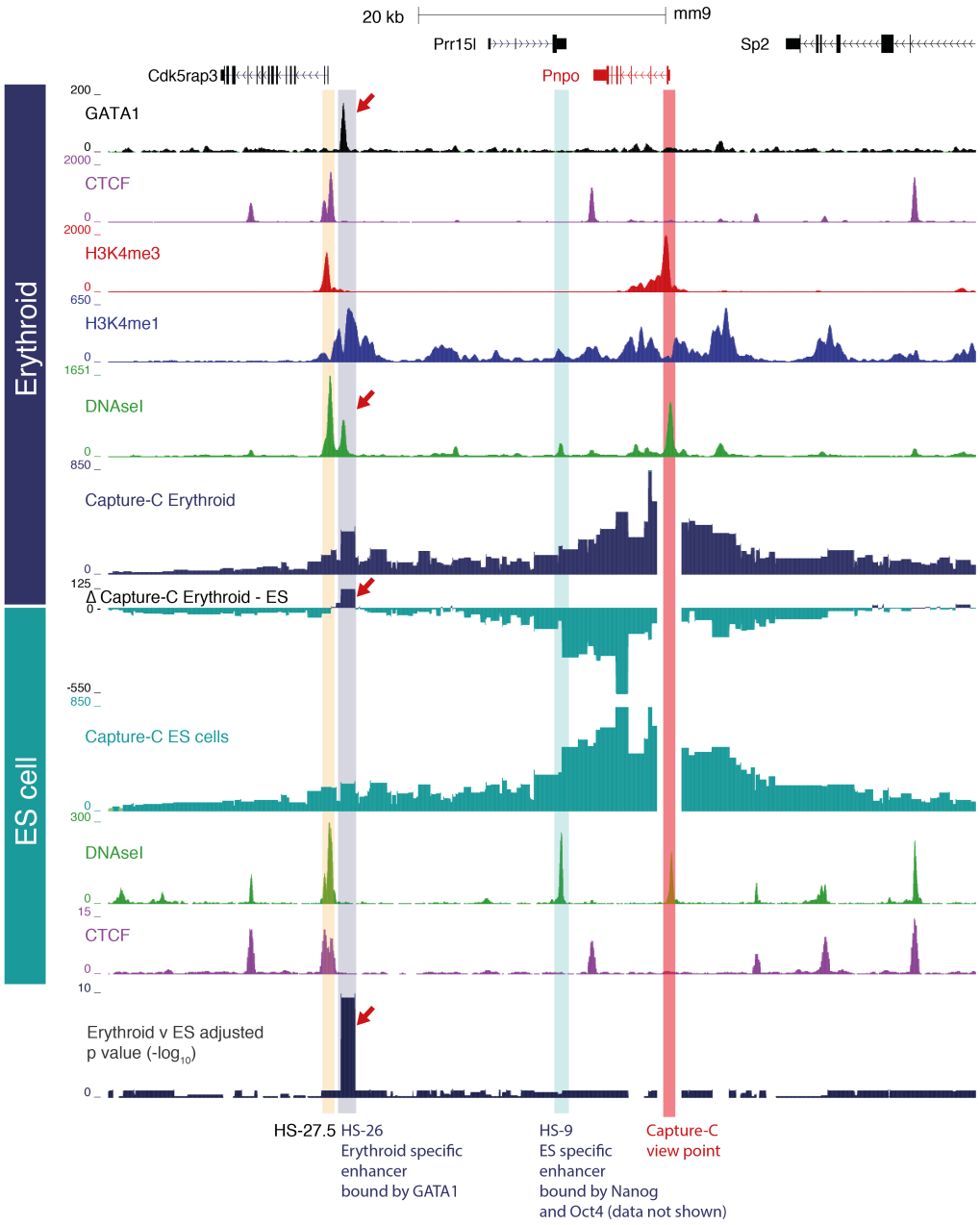


Figure 4.9 Capture-C at the Pnpo (pyridoxine 5'-phosphate oxidase) locus.

Capture-C at the Pnpo (pyridoxine 5'-phosphate oxidase) locus. This gene is essential for vitamin B(6) metabolism and is ubiquitously expressed. However, vitamin B(6) is important in haem synthesis and the gene is therefore upregulated in erythroid tissues. This is highly likely to be due to a regulatory element (HS-26), which becomes DNAseI sensitive in erythroid cells and is bound by GATA1. The Δ Capture-C clearly and very specifically pinpoints this element (red arrow heads), which would otherwise appear to be simply regulating the promoter of the adjacent gene Cdk5rap3. Analysis with DESeq2

clearly shows that the interaction frequency with this regulatory element is significantly different between ES and erythroid cells (adjusted p value = 10-9.58).
Erythroid data (Hughes et al., 2014)
ES cell data: DNAseI-seq (ENCODE UW); CTCF (LICR GSM918748)

Of interest, these subtractive analyses not only showed new interactions in the cell type under investigation but also identified new patterns of interaction in the cell type used for comparison. For example analysis of the Tal1 locus revealed one pattern of interactions in ES cells and another in erythroid cells; these cells acting as reciprocal controls for each other (Figure 4.10). This protocol therefore opens up a new way of comparing interactions in different cell types.

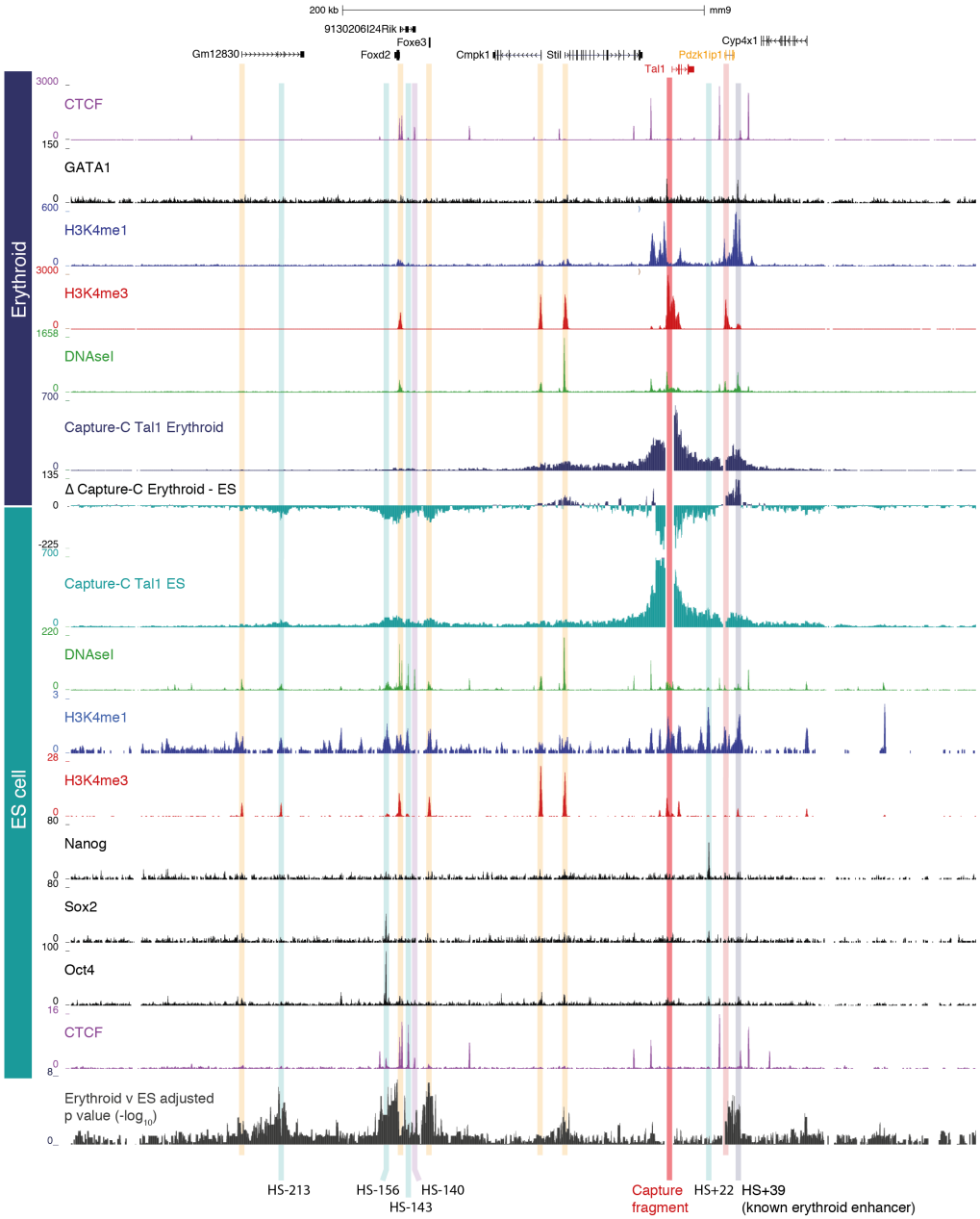


Figure 4.10 Capture-C from the Tal1 gene in ES and erythroid cells.

The Capture-C data helps to describe the potential regulatory elements at this complex locus. Comparative analysis between ES cells and erythroid cells was undertaken using DESeq2. There is a clear interaction with the previously described enhancer at HS+39 (Zhou et al., 2013, Blood http://dx.doi.org/10.1182/blood-2013-02-483875). In ES cells there are cell type specific interactions with a group of DNAseI sites around 150kb upstream of the gene. This includes the promoter of Foxe3, which is H3K4 trimethylated in ES cells and a potential regulatory element bound by Sox2 and Oct4 (HS-156).

Erythroid data: Hughes et al., 2014.

ES cell data: DNAseI-seq (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); CTCF (LICR GSM918748); MED1 (Young lab GSM1038259), Sox2 (Young lab GSM1082341), Nanog (Young lab GSM1082342), Oct4 (Young lab GSM1082340).

4.4 Conclusions

It is possible multiplex both viewpoints and samples simultaneously. This allows contrasting samples to be analysed rapidly, with very little experimental variation. This combined with the increased sensitivity of NG Capture-C allows the read count to be reported for each individual fragment across the genome rather than having to use windowing to improve signal strength is required by 4C and the original Capture-C data set. This opens up new possibilities for data analysis because comparison of data sets can be to look for changes in the interaction profile can be achieved using standard tools for comparison of NGS count based data, such as DESeq2.

The analysis of these data sets by DESeq2 very specifically highlights areas of differential interaction between the 'on' and 'off' states and it does so more accurately than other distance modeling approaches (Davies et al., 2015).

5 Analysis of long range cis and trans interactions

Previous Capture C data did not readily identify very long-range chromosome interactions or interactions in *trans*. The increased sensitivity of NG capture C enabled us to investigate such interactions and, importantly, evaluate their relative strength compared to local interactions.

At present there is good evidence to support functional interactions between enhancers and promoters up to about 1-2Mb, at loci such as Myc and Sonic Hedgehog. However, there is no consensus as to whether functional interactions can occur over much longer range in *cis* although some investigators have proposed that functional interactions can occur at these distances (Noordermeer et al., 2011a). In addition there is also debate as to whether functional interactions can be seen in trans. The majority of data from 3C based techniques shows that trans interactions are rare. However, there are a number of studies that conclude that interactions occur in trans between enhancers and promoters and that these are functional (Markenscoff-Papadimitriou et al., 2014; Osborne et al., 2004).

5.1 NG Capture-C defines previously unidentified interactions at the alpha globin locus

The increased sensitivity of NG Capture-C shows clear, reproducible, interactions with three loci neighboring the alpha globin locus. The strongest interactions are seen with the promoter of the gene *Cpbe4* and a cluster of enhancers upstream of this gene. Long range interactions are also seen with *Chac2* and the *Npm1* gene, which is an important gene recurrently mutated in AML and other myeloid neoplasms. These interactions are not present in Capture-C from the *Hba-a1&a2* promoters in ES cells. They are also present in the interactions from the R1 and R2 viewpoints (Figure 5.1).

None of these were domains of interaction were readily identifiable by either 4C seq or the first generation Capture-C.



Figure 5.1 Interactions between alpha globin and surrounding genes

5.2 Hi-C data at the alpha globin locus.

In collaboration with Job Dekker's laboratory Hi-C was performed on the same erythroid cells that were used for the Capture-C analysis.

Ter119+ cells were obtained from the spleen of C57Bl/6 mice treated with phenylhydrazine and fixed with 1% formaldehyde in accordance with the Dekker lab protocol. Three biological replicates were analysed. This fixed material was sent to Job Dekker's laboratory for Hi-C library preparation and the material was sequenced with one Illumina HiSeq run for each replicate. The data were analysed by the Dekker lab's pipeline.

These data confirm the interactions that were seen using Capture-C between the alpha globin genes and *cpbe4*, *chac2* and *npm1.* However, the Hi-C data does not readily demonstrate the interactions between the alpha globin enhancers and promoters. Despite using 3 HiSeq runs the data has been reported using 40kb bins, which makes it impossible to distinguish the functional interactions with the enhancers from the strong interactions that occur universally between the promoters and enhancers.

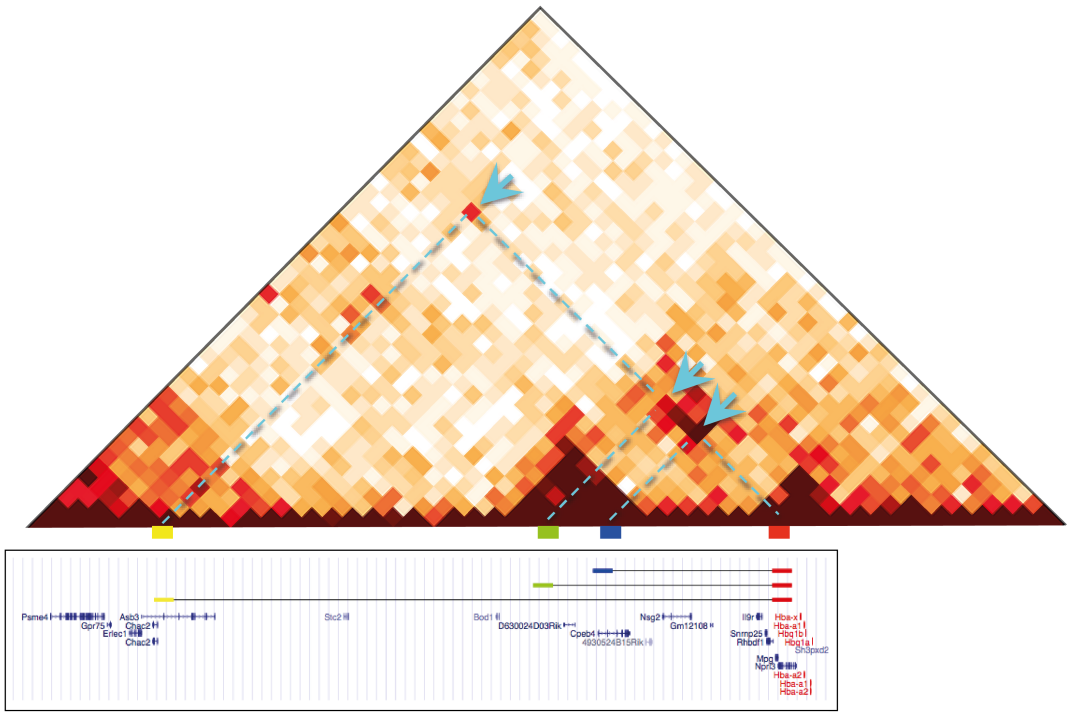


Figure 5.2 Hi-C data from erythroid cells at the alpha globin locus

5.2.1 Possible functional effects of long range interactions in *cis*

Interestingly preliminary analysis of RNA levels using nanostring nCounter technology from mice with deletions of two most important enhancers (R1 and R2) suggests that the expression of some of these genes is slightly reduced when these elements are deleted. This raises the possibility that blocks of strong enhancers can modulate the expression of many genes within several megabases.

5.3 Long range interactions are present in *cis*

Analysis of interaction frequencies across the whole length of the chromosome containing a captured region of interest shows that interactions with the entire chromosome are not easily seen when viewed on the same scale as interactions with the more local regulatory elements. However, reproducible, low level (<100 fold) *cis*-interactions are detected with many other active regions of the chromosome.

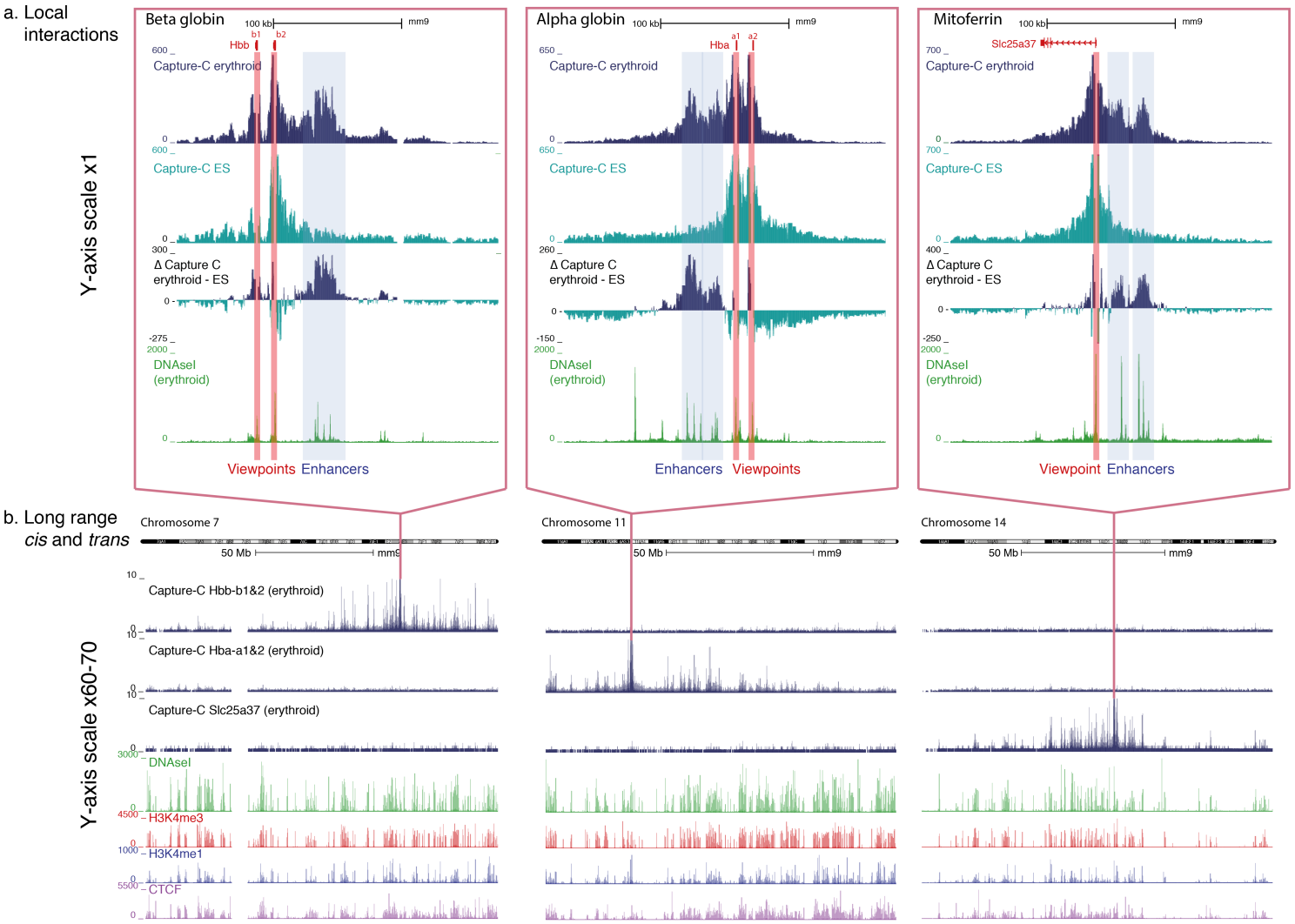


Figure 5.3 Comparison of local with long range cis and trans interactions

a. Top panel shows the normalized mean Capture-C profile at the alpha globin, beta globin and mitoferrin loci in erythroid (n=4) and ES cells (n=3). These data were generated along with the profiles for another 32 gene promoters simultaneously from a single capture reaction (making a total of 245 interaction profiles from one oligonucleotide capture reaction). The captured viewpoint fragments are highlighted in red and the interactions with the well known enhancers as annotated by DNAseI hypersensitivity are highlighted in blue. The differential track (ΔCapture-C) shows that interactions with the local erythroid enhancers are clearly and specifically increased in erythroid cells when the genes are active. The Y-axis denotes the mean number of unique interactions per restriction fragment, scaled to a total of 100,000 interactions genome-wide.
b. This shows the whole chromosome in *cis* for the above genes (beta globin - chr7; alpha globin – chr11; mitoferrin – chr14). Note the 60-70-fold change in scale on the Y-axis compared to the top panel. Much weaker interactions can be seen spreading out from the viewpoint. These correlate with active promoters and enhancers as well as CTCF sites as determined by DNAseI hypersensitivity and ChIP-seq for H3K4me3, H3K4me1 and CTCF.

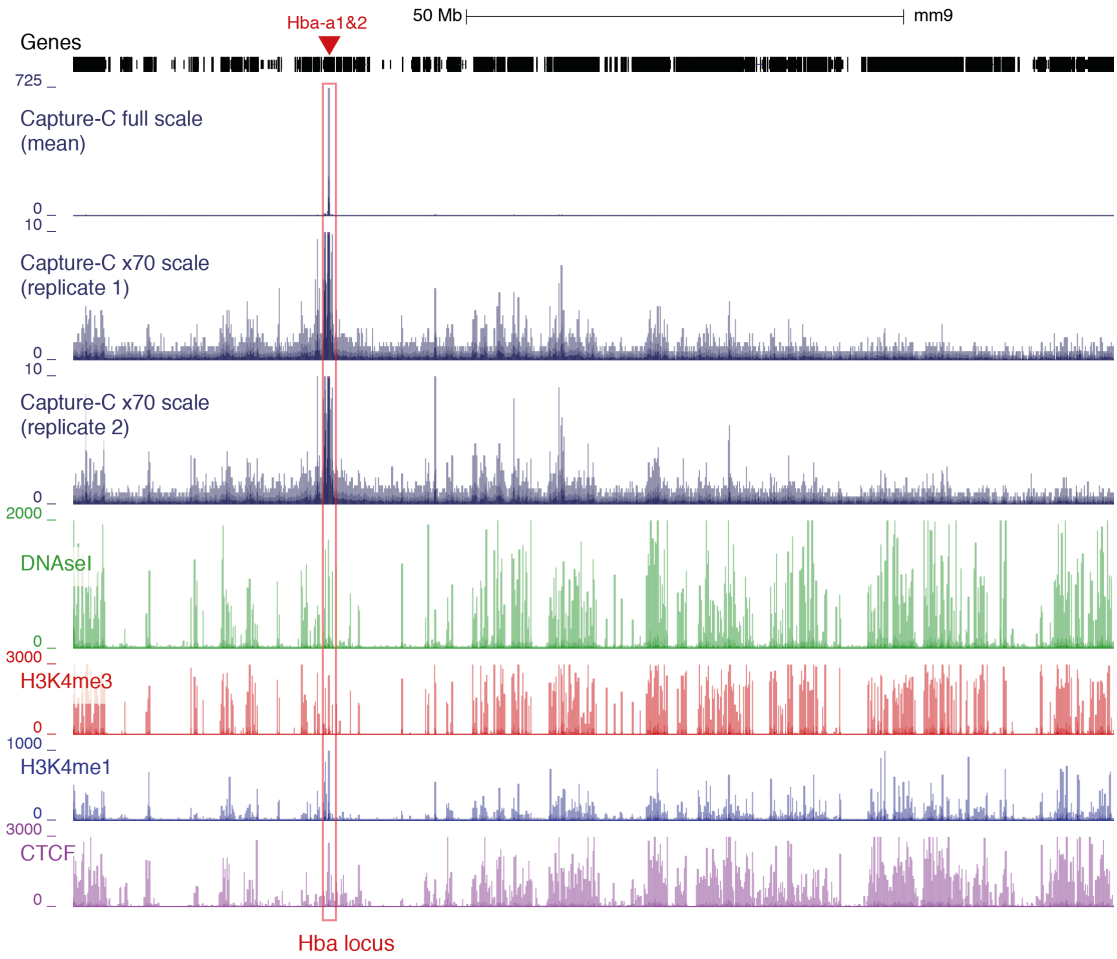


Figure 5.4 Capture-C from the alpha globin promoters - whole chromosome in cis

Capture-C from the alpha globin promoters (Hba-a1&2) in erythroid cells showing the whole of chromosome 11. The top track shows the data with the same scale on the Y axis as used for Figure 5.4. On the lower two tracks the scale has been changed to show that there are reproducible interactions in cis over several megabases, however, these data points are still reported per DpnII fragment. These correlate with gene rich regions, active enhancers and promoters as well as CTCF sites.

5.4 *Trans* interactions

Similar patterns of general interactions can also been seen in *trans* but these are a further 10 fold weaker than the long-range *cis* interactions (Figure 5.5, Figure 5.6). The patterns of *trans* interactions become visible when the threshold for any interaction is reduced to fewer than 250 interactions per 100kb and these interactions have similar distributions independent of the gene promoter used as the view point (Figure 5.6) and are correlated with gene density and the number of active elements, promoters, enhancers and CTCF sites (as annotated by DNAseI hypersensitivity and ChIP-seq for H3K4me3, H3K4me1 and CTCF respectively) (Figure 5.7). This is of particular interest in the case of the Hba and Hbb genes as they have been reported to interact with each other in erythroid cells. Some have suggested that these interactions are frequent (Osborne et al., 2004) whereas others have shown them to be rare (Noordermeer et al., 2011a). Now, with the sensitivity and robust quantitation provided by the NG Capture-C approach, the *trans*-interaction between these two genes appears to be rare (~1000 fold less that local *cis*-interactions) and on the same scale as those with most other active regions of the genome.

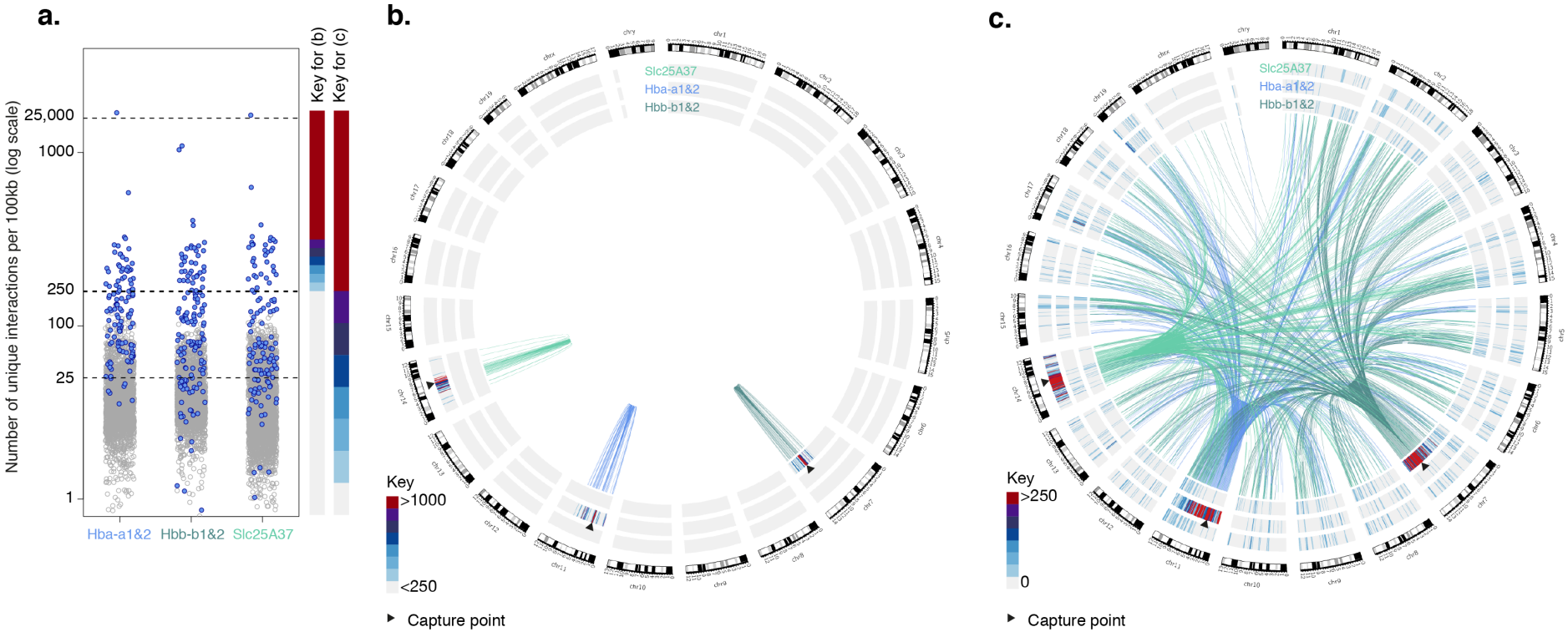


Figure 5.5 Cis and trans interactions from alpha globin, beta globin and mitoferrin

a. Scatter plot of numbers of interactions in 100kb bins from three viewpoints. Cis and trans interactions are denoted by blue and grey circles respectively.
b. Genome wide heatmap (each gene promoter is shown in a separate ring) and looping glyphs of interactions from Mitoferrin (Slc25A37), Hba-a1&2 and Hbb-b1&2. A cut off of 250 interactions per 100kb has been used to exclude very weak infrequent interactions. This shows that there are no interactions in trans with any of the gene promoters with a frequency of 100 fold less than those of the strong functional local interactions.
c. Genome wide heat map and looping glyphs of interactions from the same viewpoints as in a., however, no cut off has been used to exclude very infrequent interactions. This shows that there are very weak trans from all of the gene promoters. Interestingly these weak contacts have similar distributions in trans irrespective of the viewpoint used.

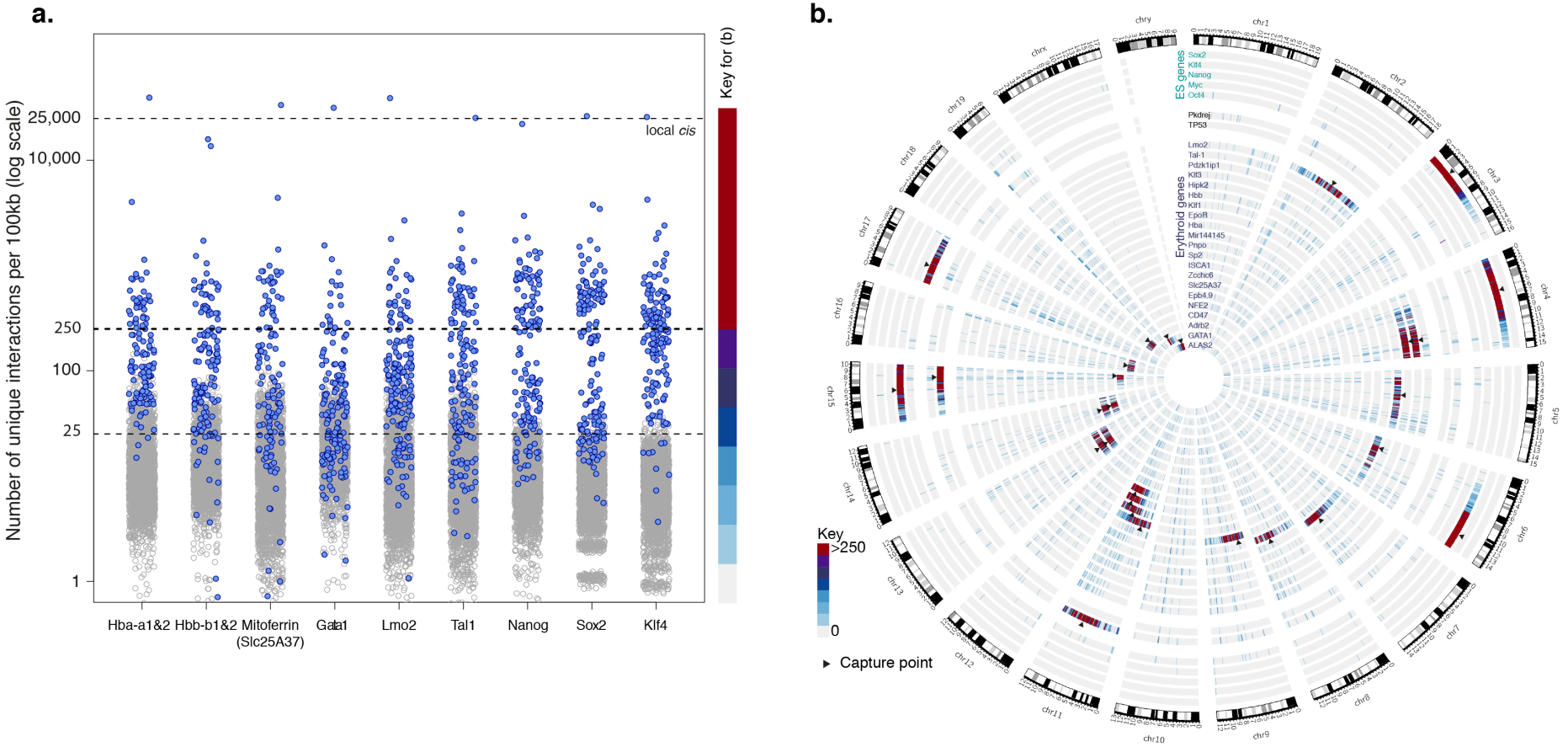


Figure 5.6 Cis and trans interactions from 30 gene promoters

a. Scatter plot of the total informative interaction count across the genome in 100kb bins. Cis interactions are in blue and trans in grey. For all of the genes the local interactions around the viewpoint are around 100-fold stronger than the weak long range cis interactions and 1000-fold more frequent than trans interactions.

b. Genome wide heat map of interactions from 28 genes. Each ring represents the profile from a single viewpoint, with the captured viewpoint being denoted by the black arrowheads. The heat map shows that strong interactions (red) are only found in cis from all 28 genes. In addition weak trans interactions can be seen and these tend to have similar distributions from active genes irrespective of the viewpoint.

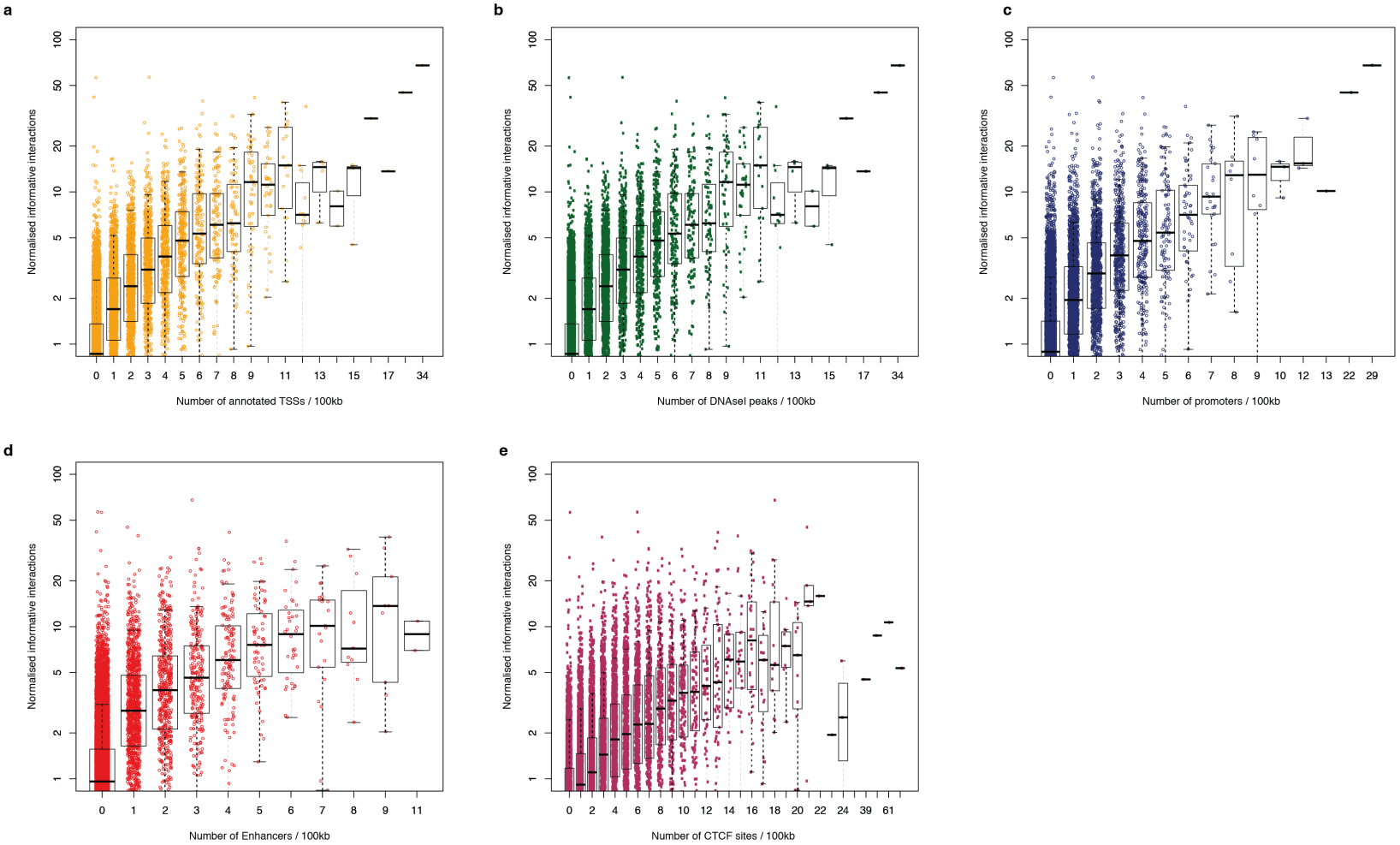


Figure 5.7 Analysis of trans interactions from the alpha globin (Hba-a1&2) promoters

The genome was divided into 100kb bins and the number of a. transcription start sites; b. DNAseI hypersensitive peaks; c. enhancers (as determined by DNAseI and H3K4me1); d. promoters (as determined by DNAseI and H3K4me3) and e. CTCF peaks was counted in each bin. There is a correlation between all of these factors and the number of normalised informative interactions.

5.5 Conclusions

This increased sensitivity of NG Capture-C allows us to detect interactions between the alpha globin promoters and the promoters of several neighboring genes that have not previously been identified using 4C or the initial Capture-C protocols. There is a possibility that these interactions are preliminary functional as data from knockout mice with deletions of the alpha globin enhancers suggests that alpha globin enhancer knockout may reduce the transcription of these genes as well as alpha globin.

We can also detect very weak interactions in over 10s of Mb in *cis* as well as interactions in *trans* that are reproducible between different samples. Interestingly the interactions coincide with other active areas of the genome. This is in keeping with Hi-C data, which shows compartmentalization of the genome into active and inactive areas (Sexton et al., 2012). However, the frequency of these interactions is an order of magnitude lower than those detected with the alpha globin enhancers and we there for do not think that these are likely to be functional. It is important to have an assay with this level of sensitivity because it allows us to be confident that all functional interactions are detected (within the limitations of 3C technology). In addition, detection and quantification of these weak yet reproducible interactions allows us to refute the idea that has been proposed that functional interactions occur between active genes on different chromosomes such as the alpha and beta globin loci (Osborne et al., 2004). Importantly although we have also found evidence of physical interactions between regulatory elements in *trans* and the promoters of the genes that we have studied, none of these interactions are comparable in frequency with the interactions with known regulatory elements in cis and there

are no single or small clusters of elements that appear to be interacting preferentially with any individual gene. In fact where trans interactions are detectable they appear to be independent of the promoter used as the viewpoint (Figure 5.6). Therefore, we think it is unlikely that regulatory elements control any of the genes that we have studies through interchromosomal contacts.

6 Capture-C for reduced cell numbers

6.1 Aims

It is important to investigate the effects of genetic variation in the correct cell type, so the ability to interrogate primary cells is critical. However, cell number is often a limiting factor when using primary tissues, particularly if pure populations of FACS sorted cells are required. At present most 3C based techniques typically use tens of millions of cells, making it impractical to analyse many cell types. For this reason we wanted to adapt the NG Capture-C protocol so that it could be used for small numbers of cells.

6.2 Adaptation of the NG Capture-C protocol for reduced cell numbers

6.2.1 3C library preparation

The 3C library preparation protocol was adapted for lower cell numbers so that the entire 3C library preparation could be performed in a single Eppendorf tube. Cells were initially fixed in 2% formaldehyde for 10 minutes at a concentration of 1 million cells /ml. A nuclear preparation was made using the standard lysis buffer. The cells were snap frozen using dry ice and ethanol and stored at -80°C. In an initial experiment to determine whether the protocol could be performed on reduced numbers of cells aliquots of fixed material samples containing 100,000, 500,000, 1 million and 3 million cells were taken from a sample containing 10 million cells. Subsequently, cells of different stages of differentiation were sorted by FACS and were fixed with as few as 150,000 cells in each sample.

After being defrosted the cells were centrifuged and the lysis buffer carefully removed. Following this the cells were Dounce homogenized in 1 ml PCR grade water (Sigma). This step can be omitted without markedly altering the digestion efficiency if there are concerns about losing the pellet with very small numbers of cells. The cells were then centrifuged and resuspended in a variable volume of water depending on the cell number and the volumes of the reagents were scaled to take into account the changes in cell number (see table below). The libraries were prepared in Eppendorf LoBind tubes to minimize DNA losses.

Number of cells	Volume to water to resuspend cells	Restriction enzyme buffer mix			Triton X 100	DpnII
		Water	SDS (20%)	RE buffer		
1 x 10^7	200 µl x3	414 µl	10 µl	80 µl	66 µl	10/10/10 µl
3 x 10^6	200 µl	414 µl	10 µl	80 µl	66 µl	10/10/10 µl
1 x 10^6	100 µl	207µl	5 µl	40 µl	33 µl	5/5/5 µl
5 x 10^5	50 µl	104 µl	2.5 µl	20 µl	16 µl	2.5/2.5/2.5 µl
1 x 10^5	50 µl	104 µl	2.5 µl	20 µl	16 µl	2.5/2.5/2.5 µl

The ligation reactions were performed by adding the ligation buffer and water to the Eppendorf tubes as outlined in the table below.

Number of cells	Water	T4 DNA ligase buffer
3 x 10^6 (700 µl)	500 µl	133 µl
1 x 10^6 (400 µl)	250 µl	72 µl
5 x 10^5 (200 µl)	125 µl	36 µl
1 x 10^5 (200 µl)	125 µl	36 µl

The decrosslinking, proteinase K and RNAse steps were also adjusted depending on cell number as follows.

Number of cells	Proteinase K	RNAse
3 x 10^6	5 µl	10 µl
1 x 10^6	3 µl	5 µl
5 x 10^5	2 µl	2.5 µl
1 x 10^5	2 µl	2.5 µl

Finally the DNA extraction was also adjusted. The large volume DNA precipitation step was avoided and a much smaller volume was used to minimize losses during the DNA precipitation.

Number of cells	Phenol / Chloroform	Chloroform	Ethanol precipitation	Ethanol wash	Water for reconstituting
3 x 10^6	1.3 ml	4 ml	Ethanol 100% 9 ml* Water 2 ml NaOAc (3M) 300 µl (15 ml Falcon)	3 ml 70% ethanol (15 ml Falcon)	100 µl
1 x 10^6	700 µl	700 µl	Split into 2 1.5 ml Eppendorfs Ethanol 100% 1000 µl NaOAc 40 µl Glycogen 1 µl	1 ml 70% ethanol	30 µl
5 x 10^5	350µl	350µl	Ethanol 100% 900 µl NaOAc 36 µl Glycogen 1 µl	1 ml 70% ethanol	20 µl
1 x 10^5	350µl	350µl	Ethanol 100% 900 µl NaOAc 36 µl Glycogen 1 µl	1 ml 70% ethanol	20 µl

Using the above protocol it was possible to generate 3C libraries with small numbers of cells. The concentration of the libraries was assessed using the Qubit (lifetechnologies) allowing the losses during the 3C library preparation process to monitored. This showed that the losses were constant on a per cell basis when the cell numbers were reduced, with approximately 50% of the input material being lost. Thus around 3pg 3C library was produced per cell of input material.

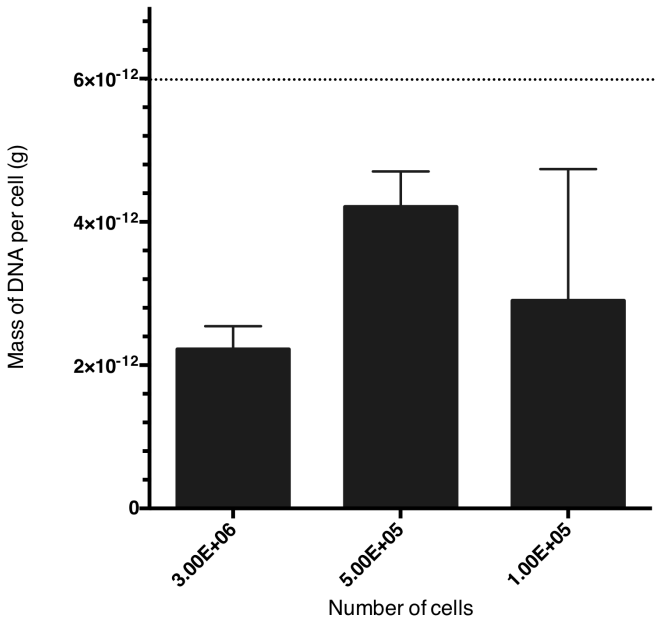


Figure 6.1 Mass of DNA in the 3C library from different cell numbers

The mass of DNA extracted from the 3C preparation corrected for the number of cells used in the assay. The dotted line is drawn at 6pg / cell, the theoretical DNA content of a cell. Reducing the number of cells used does not significantly reduce the mass of DNA extracted per cell, if anything the protocol improves the efficiency of the 3C library preparation.

6.2.2 Library quality assurance

In order to minimize DNA losses the undigested and digestion controls were not performed. Instead the ligation reaction was tested using PCR for digestion efficiency. This is possible because the probability that a fragment will ligate back to it's original partner is small (~5%). Interestingly, if one assumes that the probability that it will ligate to other ligation junctions with a similar or lower probability than back to its original partner this suggests that each fragment has more than 20 potential ligation partners.

It is safe to use the 3C product for assessing digestion efficiency since it will have a tendency to underestimate the digestion efficiency. Using the 3C libraries it was found that the digestion efficiency was not altered by reducing the number cells.

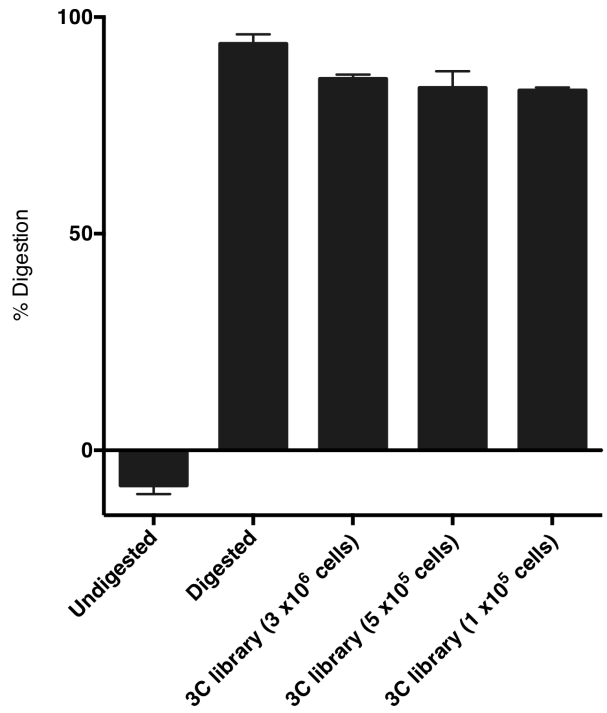


Figure 6.2 Assessment of digestion efficiency with reduced cell numbers

Undigested and digested controls are included from samples in which cell numbers were not limiting. The samples from the 3C libraries have undergone a ligation step but this does not markedly reduce the apparent digestion efficiency.

6.2.3 Ligation of adaptors and addition of indices

When more than 1ug of material was obtained the normal protocol was followed for the addition of sequencing adaptors and indices. However, when less than 1ug was available end repair and dA tailing were performed using the NEBnext Ultra DNA library prep kit. This is designed for adding adaptor sequences to small quantities of DNA and has a maximum input of 1ug. It's main advantage over the standard NEBnext protocol for adding adaptors is that it only contains one SPRI bead clean up step (rather than four), which reduces the loss of material prior to PCR amplification.

The main deviations from the NEBnext Ultra protocol were that the ratio of the SPRI beads to material being purified was always 1.8:1. This was to minimise the selection of larger fragments and minimise DNA losses. Size selection, with lower ratios of SPRI beads, was not necessary because the oligonucleotide

hybridization and streptavidin bead cleanup steps remove adaptor dimers and excess primers very effectively.

Again the Agilent Herculase II reagents were used to PCR amplify the material. However, additional cycles of PCR were used to adjust for the reduced available DNA and prevent loss of library complexity by increasing the amount of material available for input into the oligonucleotide capture reaction.

Number of cells	
5 x 10^5	10 cycles
1 x 10^5	12 cycles

6.2.4 Oligonucleotide capture and sequencing

A double oligonucleotide capture was performed from three genes (alpha globin, beta globin and mitoferrin) without any adjustments for the reduced cell numbers and the differently indexed samples were sequenced using the Illumina MiSeq.

6.3 Results

Reducing the cell numbers to 100,000 resulted in a reduction in the number of informative interactions from an average of 137,000 when cell number was not limiting to 19,000. Given that there is theoretical maximum of four interactions from each cell (one from each end of the viewpoint fragment on each allele) there is a substantial reduction in the loss of complexity with fewer cells. This is likely to be a consequence of the optimization of the reagents for adding sequencing adaptors to reduced numbers of cells.

The interaction profiles with the known enhancers at the alpha and beta globin loci were virtually unchanged compared to the traces where the cell numbers were not limited although the long-range weak interactions became more difficult to determine reproducibly.

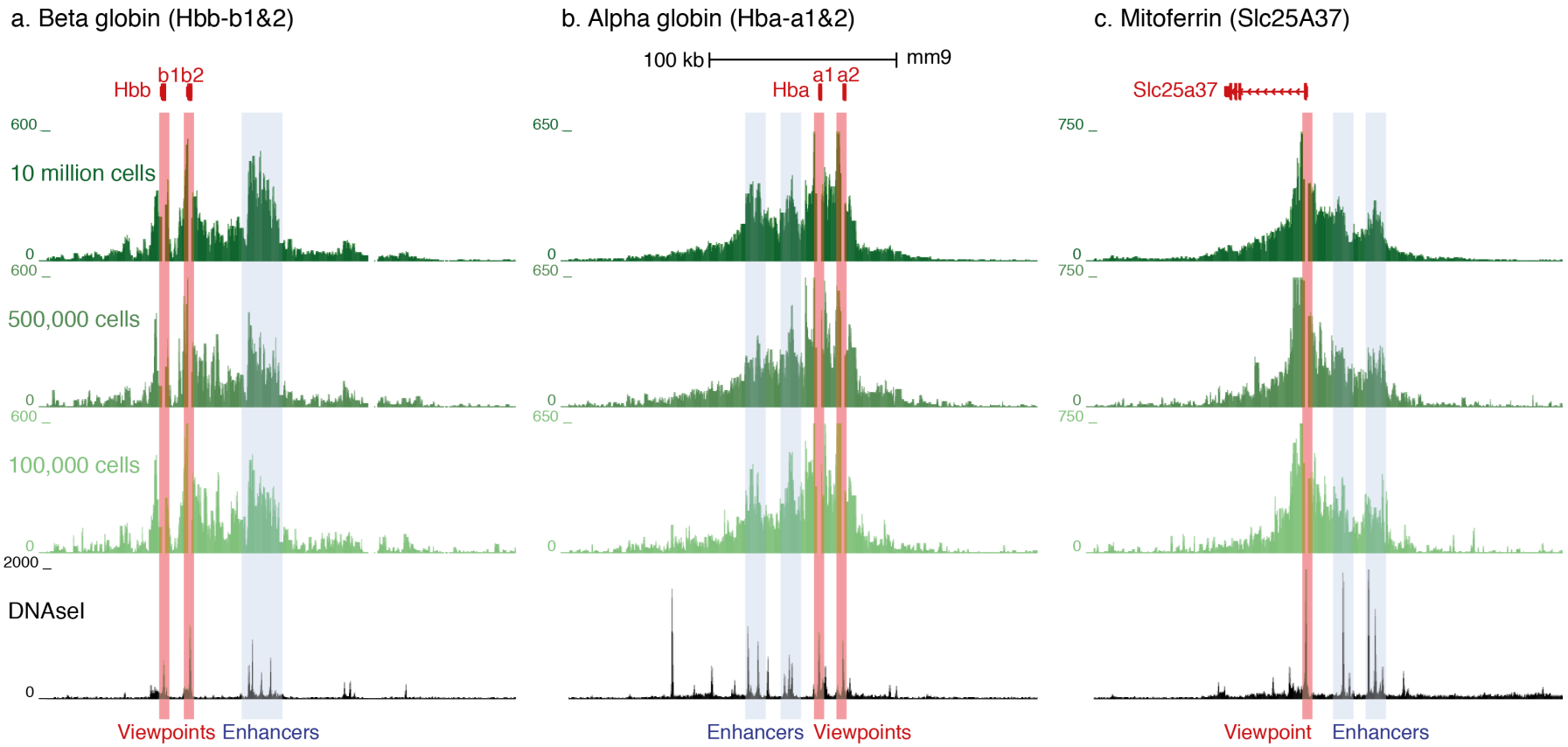


Figure 6.3 Robust interaction profiles generated from small numbers of cells

Comparison of capture-c profiles from the promoters at the beta globin (Hbb-b1&2) (a), alpha globin (*Hba-a1&2*) (b) and mitoferrin (*Slc25A37*) (c) loci (highlighted in red) generated from different numbers of cells ranging from 10 million to 100,000. The plots show the number of interactions per restriction enzyme fragment for the mean of two biological replicates, normalized to 100,000 total informative interactions genome-wide. Despite the 100-fold reduction in cell number, the interactions with the main regulatory elements (highlighted in blue) can still be clearly detected at all three genes.

6.3.1 Limitations from using reduced cell numbers

Longer range interactions become more difficult to resolve using smaller numbers of cells. However, even using small numbers of cells for input NG Capture-C compares very well with 4Cseq and the original Capture-C method.

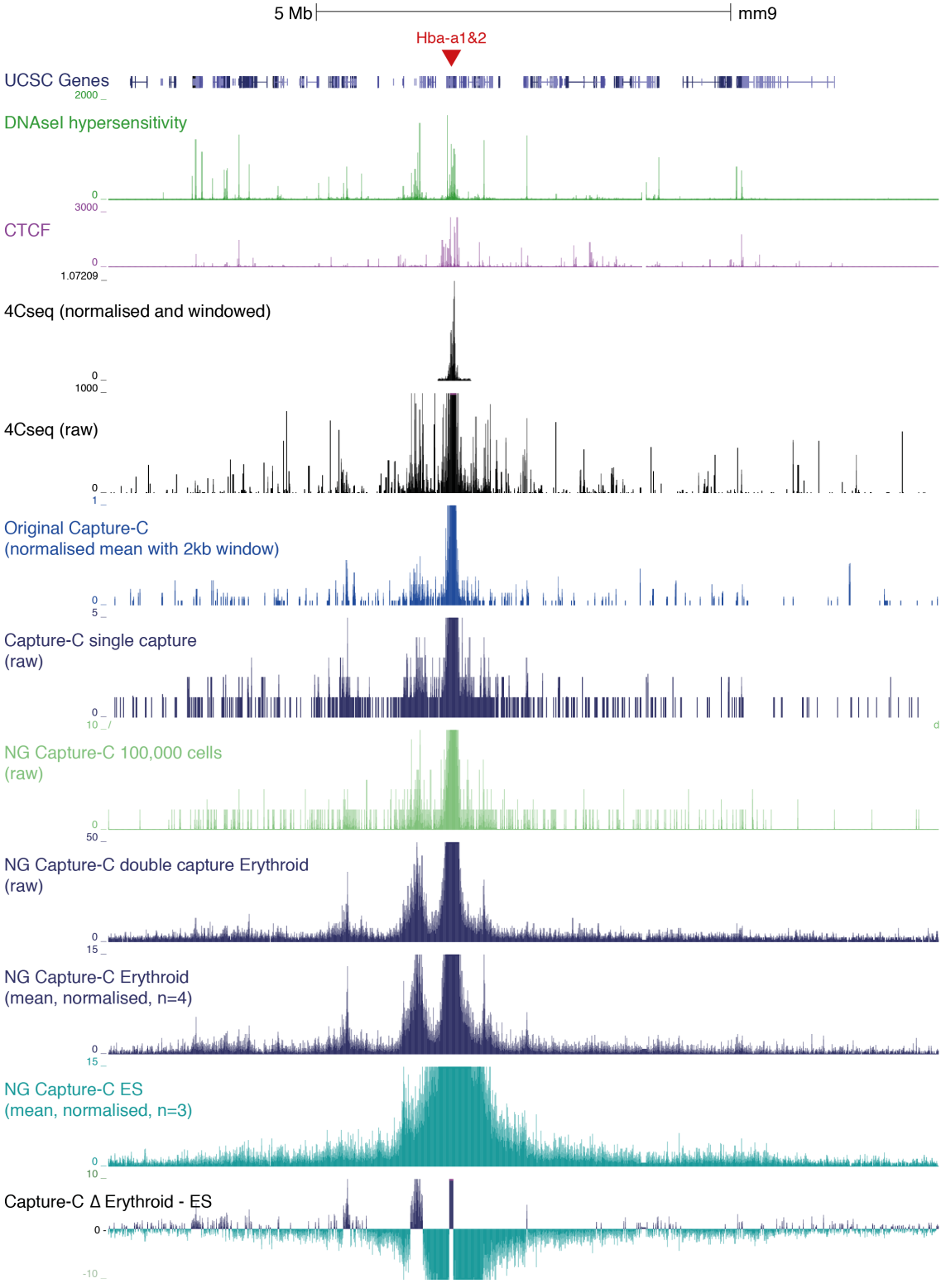


Figure 6.4 Comparison of NG Capture-C with different numbers of cells with other methods

Comparison of different methods with NG Capture-C with unlimited and reduced cell numbers in the 10Mb region surrounding the alpha globin locus. Interactions become more difficult to detect over 1Mb from the promoter with single capture and when cell numbers are reduced. Two 4Cseq tracks are included. The first shows the data as analysed by the de Laat group (van de Werken et al., 2012b), which is limited to the region immediately surrounding the gene (as can be normalized using BACs). The second shows the raw data as analysed by the Capture-C pipeline.

6.4 Determination of the interaction profile during cellular differentiation.

The reduction in cell numbers allowed sorted cells from different stages of erythroid differentiation to be analysed.

6.4.1 Sorting of cells at different stages of erythroid differentiation.

Cells were obtained directly from the livers of E14.5 wild type C57Bl6 mice. Two biological replicates were analysed, each containing the mixed cells from three fetal livers. The cells were stained for Ter119, CD71, CD117 and Hoechst (a live/dead marker) and were sorted using the FACS aria (BD biosciences). Two populations of cells were collected simultaneously. The alpha globin expressing intermediate erythroblast population was sorted based Ter119+ CD71+ and these cells were nearly all CD117-. The early population from the fetal liver was sorted based on CD117+, CD71+, Ter119- antigens following the strategy used by the ENCODE consortium.
https://genome.ucsc.edu/ENCODE/protocols/cell/mouse/EPC_Stam_protocol.pdf).

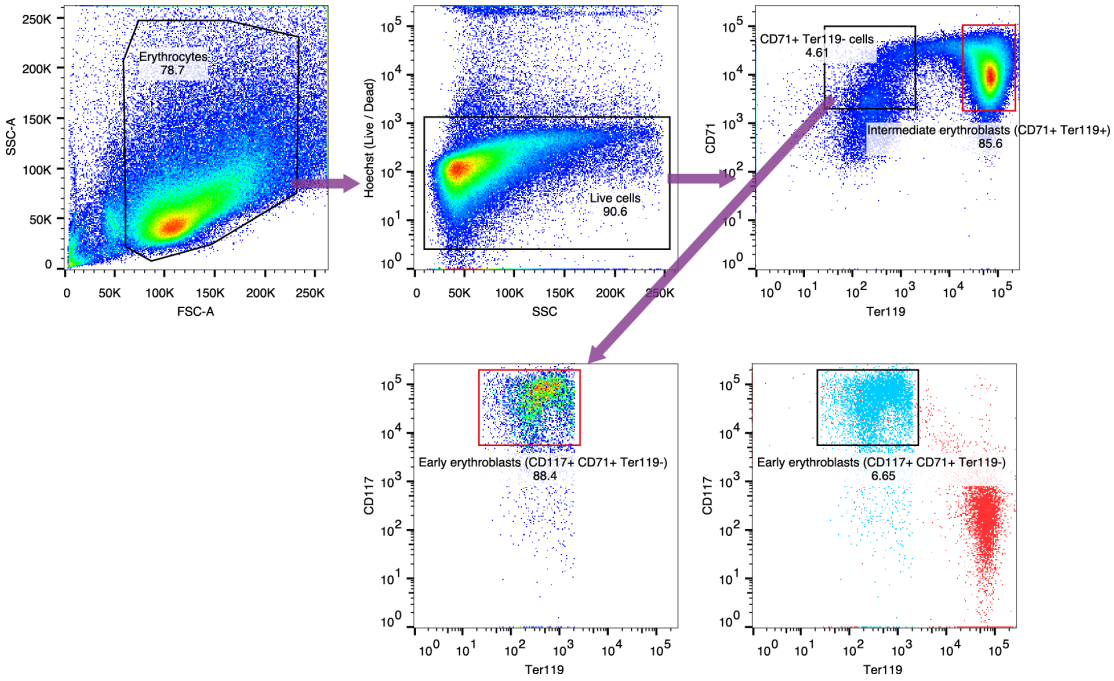


Figure 6.5 Gating strategy for sorting early and intermediate erythroblasts.

An initial gate was used to select cells of an appropriate size based on forward and side scatter. A second gate on FSC-A was used to remove duplets. Dead cells were then gated out based on Hoechst and the cells were then gated into two groups based on Ter119 and CD71 expression. The early CD71+ Ter119- cells were also gated for the CD117+ fraction. The final panel shows that selection based on CD71 and Ter119 clearly determines the expression of CD117 to a large extent.

6.4.2 Capture-C on cells from different stages of differentiation

Capture-C was performed initially from the alpha globin, beta globin and mitoferrin (Slc25A37) gene promoters in the sorted early and late erythroid cells. These data show that prior to the onset of transcription the interactions between the enhancers and promoters are markedly diminished. However, there are still interactions present when the early profile is compared with that from ES cells.

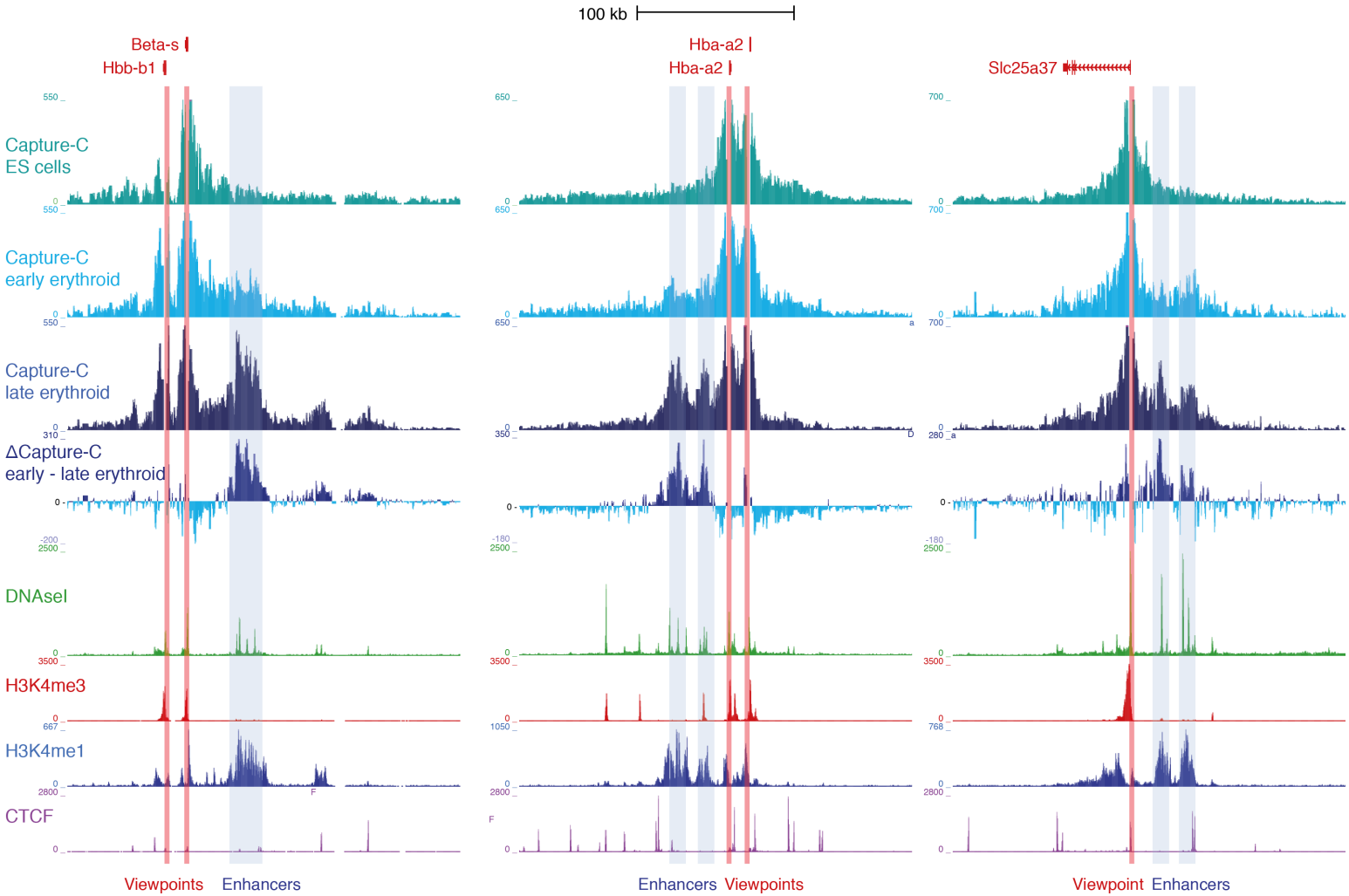


Figure 6.6 Changes in the Capture-C profile during erythroid differentiation

Capture-C was performed from the alpha and beta globin and mitoferrin gene promoters on cells sorted for different stages of erythroid differentiation. The early CD117+ CD71+ Ter119- cells do not express alpha globin and here the interactions with the enhancers are clearly diminished compared to the later Ter119+ expressing cells. However, some interactions are clearly beginning to occur between the enhancers and promoters compared to ES cells, in which the genes are not expressed.

6.5 Conclusions

These data appear to show that the interaction profile is neither absent or completely established prior to the onset of transcription. However, these data need to be interpreted with some caution. We wanted to determine whether the interactions between the enhancers and the promoters were the key final trigger that initiates transcription or whether the compartment is set up prior to gene expression. In order to answer this question very pure populations of late erythroid cells need to be isolated in which globin transcription is absent but just about to be initiated. Contamination with non erythroid cells or early stages of erythroid differentiation will produce an intermediate picture between 'on and off'.

It is difficult to sort completely pure populations of erythroid precursors because there are no specific positive cell surface markers that are exclusively present on early erythroid populations (Pronk et al., 2007). The current methods for isolation of early erythroid cells are dependent on negative selection and staining with markers that are positive in other early populations of cells. Although cells were sorted based on a well recognized SOP for isolation of mouse fetal liver cell subsets from the ENCODE project, a potential problem with these data is that CD71 can also be expressed by earlier non-erythroid progenitors (some of the earlier B cell lineages, for example). Although it is possible to remove B cells with a lineage cocktail, earlier non-erythroid stages such as PreGM would not be removed with this strategy. It is not clear to what extent these populations are found in fetal liver at day E14.5 but clearly these could alter the interaction profile.

Purification of populations of specific erythroid precursors can be attempted using the strategy outlined by Pronk et al., 2007. However, this is a much more complex FACS panel, which uses adult bone marrow and it would be very challenging to obtain sufficient numbers of early erythroid cells using this approach for analysis by Capture-C.

Therefore we can conclude that some of the interactions are set up prior to the initiation of transcription but because we cannot be certain about the purity of the cells used in this experiment it is possible that the reduction in the interactions detected could be due to contamination of the cells by either early erythroid cells or other lineages expressing CD71.

7 Allele Specific Capture-C

SNPs underlying GWAS traits are frequently present on one allele but not the other and may affect the regulatory interactions of the affected allele. Therefore, it would be of great value to generate allele-specific interaction tracks. This has not previously been attempted on Capture-C because it requires considerable read depth but we thought that due to the greatly improved depth of signal provided by the NG Capture-C protocol it should be possible.

7.1 Production of allele specific tracks

When a SNP was located close (within 120bp) to the restriction enzyme cut site of the captured fragment it was possible to create allele specific tracks. In order to do this Capture-C was performed on F1 crosses between C57Bl/6 and CBA/J mice. Allele specific tracks were generated by separating the read pairs based on the presence or absence of a SNP in the captured viewpoint fragment (Figure 7.1). Thus independent interaction profiles could be generated from each allele (Figure 7.2).

This could only be done when the data was high quality with more than 20,000 unique interactions since it reduces the number of interactions in the track approximately 8-fold. This is because half of the reads are discarded from the end of the fragment that does not contain the SNP. In addition, depending on where the SNP is located up to half of the reads at the end that contains the SNP may be discarded if it is not contained within the read and the depth is halved again because one track is determined from each allele.

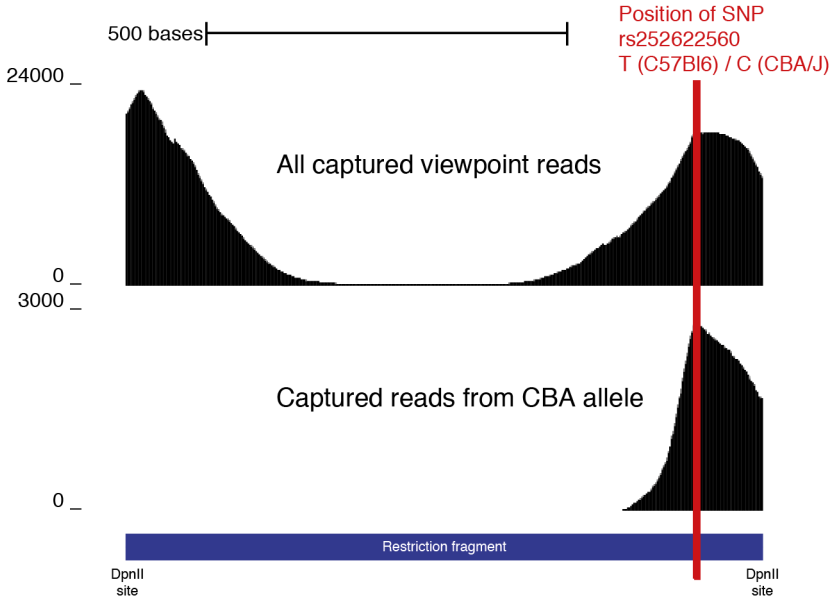


Figure 7.1 SNP specific interaction profiles

This shows a read density plot of the reads mapping to the captured restriction fragment (the Tal-1 promoter fragment is shown). SNPs under the captured promoter allowed us to generate allele specific interaction profiles in F1 crosses between C57BL/6 and CBA/J mice. In the example locus the SNP rs252622560 has been used to separate interaction profiles from the two different alleles.

Note the asymmetry of the capture profiles in the top profile. The left hand capture site has a sharper peak and the reads do not extend as far into the fragment on this side. In comparison the peak is more rounded on the right hand side. This was present at all of the capture fragments we investigated. Extensive checking of the bioinformatics was undertaken but the same skew was present in the raw data and so we concluded that this is likely to be due to the synthesis of the oligonucleotides. The oligonucleotides are synthesized in a left to right direction and so the sequence quality will be higher for the left hand end of the oligos, making the capture more efficient on the left hand side of the capture fragment.

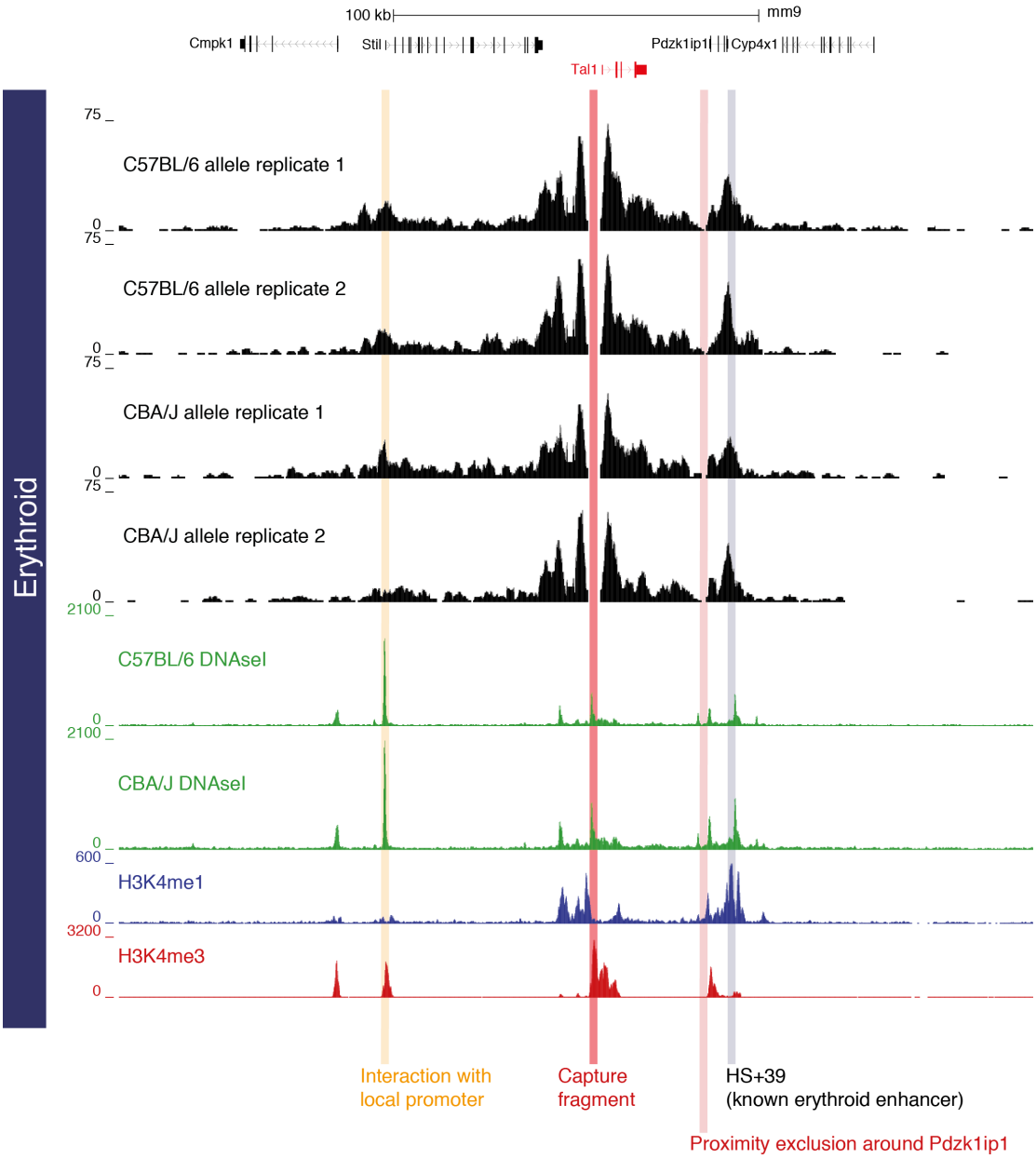


Figure 7.2 Allele specific tracks at the Tal-1 locus.

This shows allele specific Capture-C data (with a 2kb moving window) at the Tal-1 locus in F1 erythroid cells (C57Bl6 / CBA/J crosses). Since the regulation at this locus is the same in both strains the profiles from the different alleles are very similar. The interaction profiles themselves were very similar simply because none of the SNPs in the locus are of likely functional importance.

7.1.1 Analysis of the phase of SNPs in allele specific tracks

Although the interaction profiles at these loci were very similar it was possible to demonstrate that these genuinely were allele specific tracks by analyzing the

phase of the SNPs in *cis*. This was achieved by initially determining all of the positions across the genome where the C57Bl/6 and CBA/J reference genomes have differential SNPs using the data generated by the Sanger Mouse Genomes Project (Keane et al., 2011; Yalcin et al., 2011). When these SNPs were found at capture sites they were used to generate allele specific tracks. This was done by adding a subroutine into the main script that allowed the sequenced base to be identified at any position in the read. This allowed the 'capture' fragments to be classified based on a particular base pair being present at a defined position in addition to the normal criteria, which are based on whether both ends of the read lie within the captured restriction fragment. The script that generates the allele specific tracks generates a SAM file of all of the reads that interact with the viewpoint. Importantly SAM file format retains the sequence of the read and these were analysed to determine the whether the SNPs in the interacting reads were in phase with the capture fragment.

It was found that over 95% of the strain-specific SNPs in *cis* are in phase with the captured strain specific SNP (Figure 7.3). Thus interactions with the sister chromatid are relatively rare (not much above the level found with the other chromosomes in *trans*).

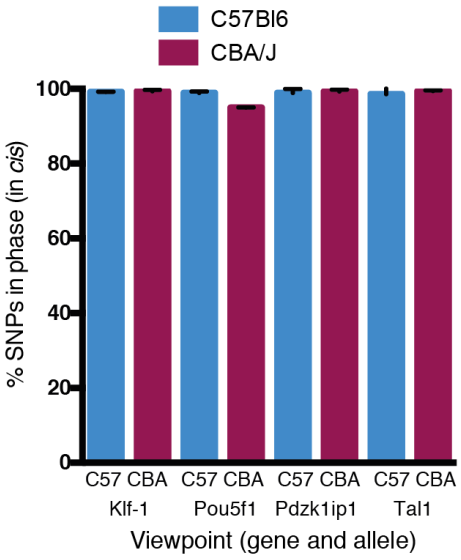


Figure 7.3 Phase of SNPs with the viewpoint fragment.

This shows a graphical representation of the % of SNPs in phase in the interacting reads compared with the strain of the captured allele in *cis*. This

demonstrates that the chromosome predominately interacts with itself in *cis* rather than its sister chromatid.

7.1.2 Generation of separate tracks from Hba-a1 and 2

This type of analysis also applies to non-allelic SNPs such as those that distinguish the tightly *cis* linked, parologous *Hba-a1* and *Hba-a2* genes. This allowed us to identify and quantitate specific interactions with each of these genes. The two Hba genes in mouse are almost exact copies which differ at only a few base positions, one of which lies near the 5' capture point for this gene (Fig. 4c). This allows separate interaction profiles for the two non-allelic *Hba-a1* and *Hba-a2* genes to be generated (Figure 7.4). The 5' *Hba-a1* interacts with the proximal regulatory elements (HS-12 and R4) more frequently than *Hba-a2* interacts with them. Interestingly, *Hba-a2* has very similar interactions compared to *Hba-a1*, with the R1 and R2 regulatory elements, which are thought to have stronger enhancer function than the other elements (Anguita et al., 2002; Bernet et al., 1995).

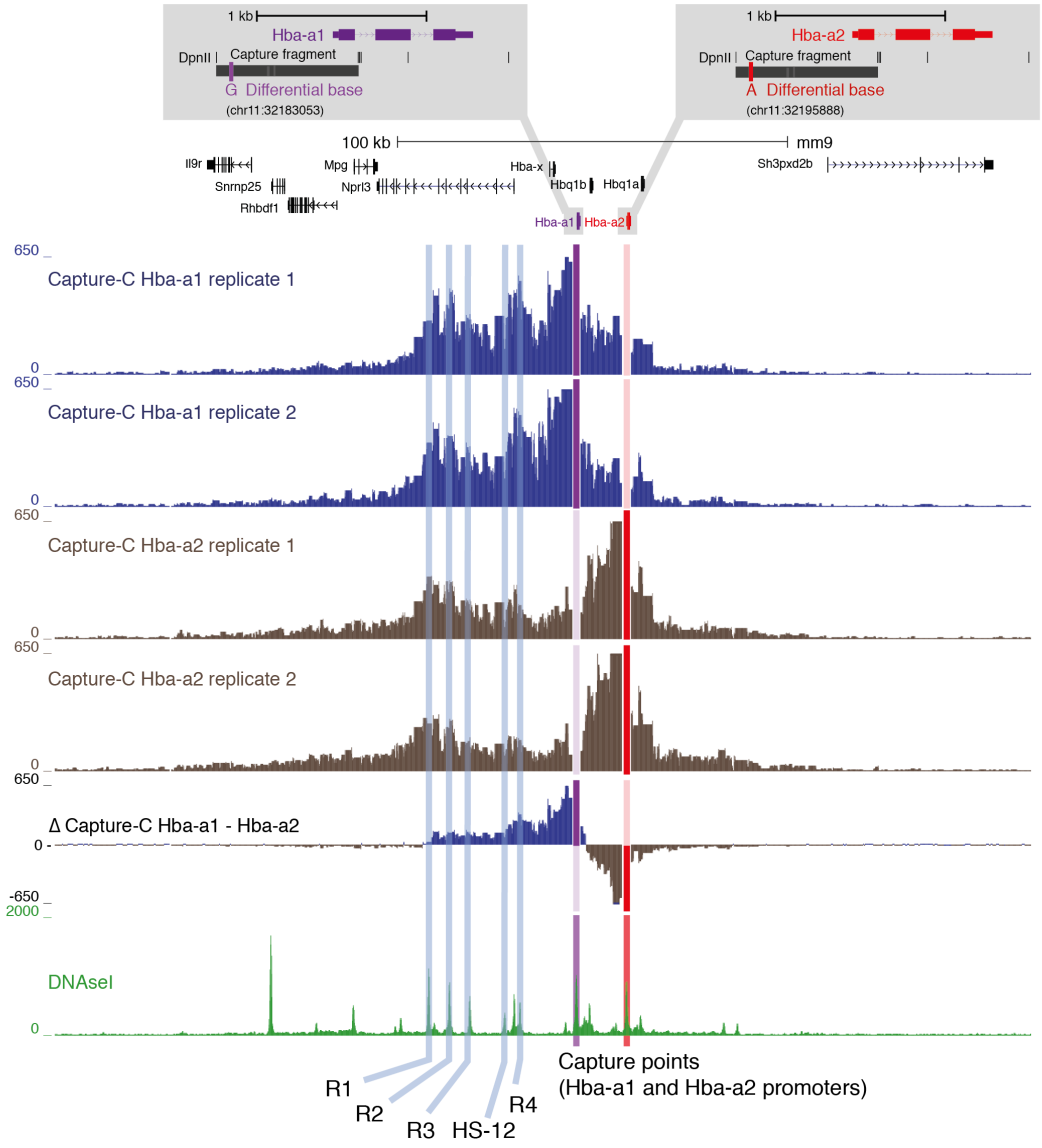


Figure 7.4 Specific interaction profiles for *Hba-a1* and *Hba-a2* parologous genes generated using SNP specific Capture-C.

A single nucleotide difference between the two genes allows generation of specific tracks (see inset). *Hba-a1* is the more active of the two genes, producing around 70% of the total mRNA. Comparison of the two biological replicates shows that the SNP specific profiles are highly reproducible. The Δ Capture-C track shows the difference of the mean *Hba-a1* and *Hba-a2* profiles. This reveals that that the *Hba-a1* gene preferentially interacts with the enhancers, particularly proximal HS-12 and R4 elements. The *Hba-a2* gene interacts much more strongly with the chromatin between the two genes and interestingly it interacts with the most distal enhancer (R1) to a very similar degree to the *Hba-a1* gene.

Together these data show that using NG Capture C it is possible to analyse allele specific interactions and *cis*-interactions between a regulatory element and two or more duplicated non-allelic paralogues.

7.2 Conclusions

We have shown that chromosomes preferentially interact with themselves rather than their sister chromatid. This implies that interactions in *cis* are dictated by the physical properties of the chromatin fiber rather than the sequence specific binding of transcription factors since the pattern of transcription factor binding would be very similar on the sister chromatid.

The ability to generate allele specific interaction profiles is important because it should aid researchers to unravel the functional effects of polymorphisms uncovered by GWAS. At present only NG Capture-C is able to delineate allele specific profiles because it requires very high levels of sequence depth because only around 10-20% of the data can be used to generate the allele specific track.

8 Analysis of a single locus from multiple viewpoints by Capture-C

8.1 Aims

The alpha globin locus is one of the best characterized loci in the human and mouse genomes. We therefore wanted to investigate the interaction profiles from multiple viewpoints in the locus to try and build up a detailed picture with a view to trying to understand how the interaction profile differs from different features; hoping ultimately that this would provide insight into what determines the pattern of interactions.

In addition, although the physical interactions between enhancers and promoters appear to coincide with gene expression, at present the mechanisms that lead to the changes in the interaction profile with the onset of transcription are unclear. An important question in the field is whether the enhancer promoter interactions are purely dependent on the binding of transcription factors and co-factors at the enhancer elements or whether the interaction profile is independent of the binding of transcription at the enhancers. Our laboratory has made mice with individual and combined deletions of the enhancer elements.

8.2 Experimental methods

In order to generate interaction profiles from different view points in the same locus the oligonucleotide capture had to be repeated five times using different capture oligonucleotides. This had to be done because of the difficulties in interpreting the interaction profile from one viewpoint around the sites of capture by other oligonucleotides in the same experiment. This is because sequences that can be captured by two different oligonucleotides are likely to be preferentially enriched, which is likely to skew the interactions towards regions of which can be captured by two or more oligonucleotides. The interactions within +/- 1kb from the restriction fragment of all of the capture sites in the experiment were removed from all of the interaction profiles. However, if one wanted to report this data it would be extremely important to control for the

different capture efficiencies of the different oligonucleotides. This is a complex system to model when one considers that not only do the different oligonucleotides have different intrinsic capture efficiencies but it is common for the oligonucleotides to only partially capture sequence. This is because the fragments are sonicated to 200bp; on average half of the sequence will come from the capture site and the capture oligonucleotides are 120bp long. The capture efficiency will be dictated by the sequence of the region being captured but also the length of the sequence available to be captured by the oligonucleotide so the capture efficiency is likely to be variable for at least half of the captured fragments. A further level of complexity arises for fragments that can hybridise to two biotinylated capture oligonucleotides. Since the capture oligonucleotides have a combined length that is greater than the fragment being captured it is highly likely that one if not both oligonucleotides will only be partially capturing the sequence so the capture efficiency will virtually be unique for each individual fragment being captured.

To circumvent this problem multiple rounds of oligonucleotide capture from different viewpoints were performed on the same material. This was achieved relatively rapidly by manufacturing enough material for five capture experiments at the outset by sonicating and adding indexing adaptors simultaneously to two aliquots of 3C libraries. This material was mixed prior to the first capture to increase the library complexity and reduce the noise between the different experiments.

8.3 Capture-C from the alpha globin enhancers

Capture oligonucleotide probes were designed to the enhancer elements at the alpha globin locus. Initially the interaction profile from the R1 and R2 enhancers was determined (Figure 8.1). This shows a very similar profile of interaction from each enhancer. There is clear interaction with the alpha globin promoters and there are quite strong interactions with the other enhancer elements (R3, R4 and HS-12). In fact the profile is very similar to the profile from the viewpoint of the promoters, indicating that the enhancers and promoters are likely to have similar domains of interaction.

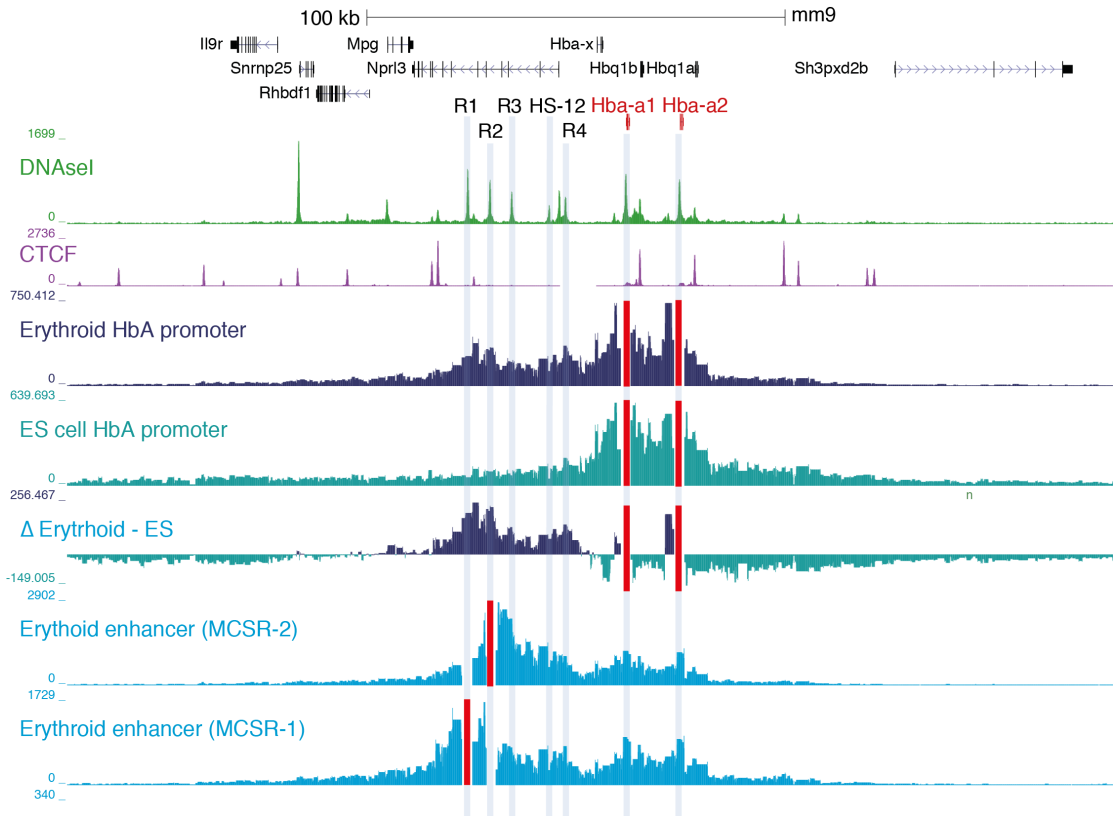


Figure 8.1 Interaction profile from the alpha globin enhancers

8.4 Capture-C from the CTCF sites within the alpha globin locus

The alpha globin locus contains several CTCF and cohesin binding sites. The locus contains two strong CTCF binding sites at the promoters of Hbq1a and Hbq1b and there is a weak CTCF site in between the R1&2 enhancers. In addition, the locus is flanked, both upstream and downstream, by two pairs of strong CTCF peaks. We therefore went on to determine the interaction profiles from the CTCF binding sites in the locus (Figure 8.2). This work was undertaken jointly with Lars Hanssen a DPhil candidate in the Higgs Lab, whose project is focused on investigating the role of CTCF at the alpha globin locus.

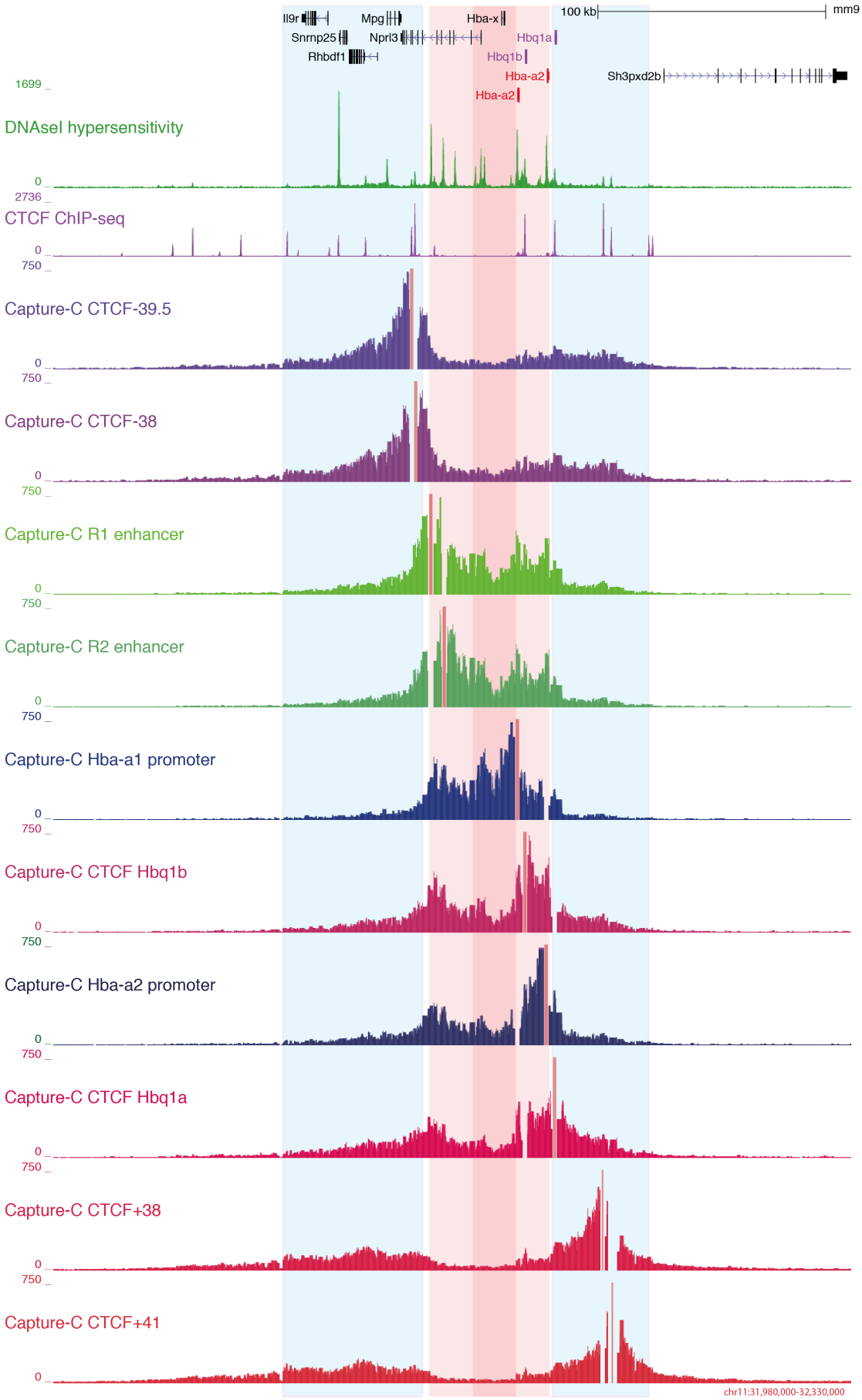


Figure 8.2 Capture-C from multiple viewpoints within the alpha globin locus

This figure shows Capture-C profiles from the two most important enhancer elements (R1 and R2) and six CTCF sites from within the alpha globin locus. In addition data is shown from the two alpha globin promoters (Hba-a1 and Hba-a2), which was generated using a differential SNP between the two paralogues (as described in the previous chapter). The data were generated from Ter119+ cells from two biological replicates of the spleen from phenylhydrazine treated mice. Sufficient sonicated 3C library with ligated sequencing adaptors was generated for five separate oligonucleotide capture experiments (Hba-a1&2; R1&R2 enhancers; and three separate experiments for the CTCF sites). This allowed virtually complete profiles to be generated across the locus from each viewpoint (only one other oligonucleotide capture site had to be excluded from each profile).
The shaded boxes denote two partially mutually exclusive domains of interaction (blue and red). The domain containing the *Hba-a1&2* genes has a possible internal sub-domain shown in darker red.

The interaction profile from the CTCF sites is remarkably different to that from enhancers and promoters. In particular the profile from the HS-38 and HS-39.5 sites is dramatically different to that of the two strong enhancers R1&2, which are only located 6kb away. These two CTCF sites are do not interact with the enhancers and only interact weakly with the promoters. These two CTCF sites also interact upstream with the *Mpg*, *Rhbdf1* and *Il9r* genes. In addition there are relatively strong interactions with the other CTCF sites flanking the alpha globin genes (at HS+38 and HS+40). The profile from these two downstream CTCF sites shows these changes even more clearly. Here very infrequent interactions are seen with the enhancers and promoters of the alpha globin genes and there is a relatively strong but broad domain of interaction extending from the CTCF sites at HS-38 to beyond the *Il9r* gene.
Capture-C was also performed from the two strong CTCF binding sites close to the alpha globin genes (located at the promoters of *Hbq1b* and *Hbq1a*). The site at *Hbq1b* (which is juxtaposed between the *Hba-a1* and *Hba-a2* promoters) shows strong interaction with a peak CTCF binding between the R1 and R2 enhancers as well as relatively strong interactions with the other enhancer

elements at the locus. In addition, by comparison with the promoters this CTCF sites appear to have a slightly broader domain of interaction that includes the up and downstream CTCF sites. The site at Hbq1a, which is downstream of the Hba-a2 gene has a similar profile to the other downstream CTCF sites (HS+38 and HS+40) and appears to have infrequent interactions with the enhancers.

8.5 Capture-C on alpha globin enhancer knockout mice

8.5.1 Methods

In order to investigate the effects of the enhancer knockout on the interaction profile erythroid cells had to be obtained from mouse fetal liver since the most severely affected R1&2 double knockout does not survive beyond around E14. Cells were harvested from the fetal liver of E12.5 mice and were expanded in culture for 5 days. Early erythroid cells were purified using MACS columns to obtain ter119 negative fraction. These cells were put back into culture for a further 18h with a higher concentration of erythropoietin. The cells were fixed using 2% formaldehyde and Capture-C was performed as outlined in Chapter 1. Capture-C was performed *Hba-a1* and *Hba-a2* promoters were used as viewpoints on cells from fetal liver cultures in triplicate from both the WT and the R1&2 double enhancer knockout.

8.5.2 Results

The profiles from the WT fetal liver cultured cells and the phenylhydrazine treated spleen WT are extremely similar and confirm that there is an approximately 70kb compartment of strong interactions, which contains the alpha globin enhancers.

The R1R2 and R2R3 double enhancer knockouts show that deletion of the enhancers results in reduced interactions across the whole locus compared to the wild-type (Figure 8.3). This was not simply with the deleted enhancer sequences themselves but also with the other more proximal regulatory elements. Importantly when these data were analysed using DESeq2, there were no other sites across the genome in which significant differences in the interaction profiles were seen between the knockout and WT samples (Figure 8.4).

This is a somewhat surprising result because the interaction profiles are normalized to the total number of reads in the track so if there were no changes in the interactions with the more proximal enhancers the normalized data would be likely to show increased interactions with the more proximal enhancers.

This indicates that the interactions between the promoter and the individual enhancer elements are interdependent to some extent. When the R1&2 knockout profile is compared to that of ES cells, in which the *Hba-a1&2* genes are minimally transcribed, it is clear that the interactions over the whole locus are not completely abrogated by the double enhancer knockout. In particular, there is a prominent residual interaction with a CTCF site juxtaposed between the R1 and 2 elements. Thus the interactions between the enhancers and promoters are partially but not entirely mediated by binding of transcription factors at the enhancers. In addition there appears to be some degree of interdependence between the different enhancer elements such that interaction between one element and the promoter increases the interactions between the others and the promoter.

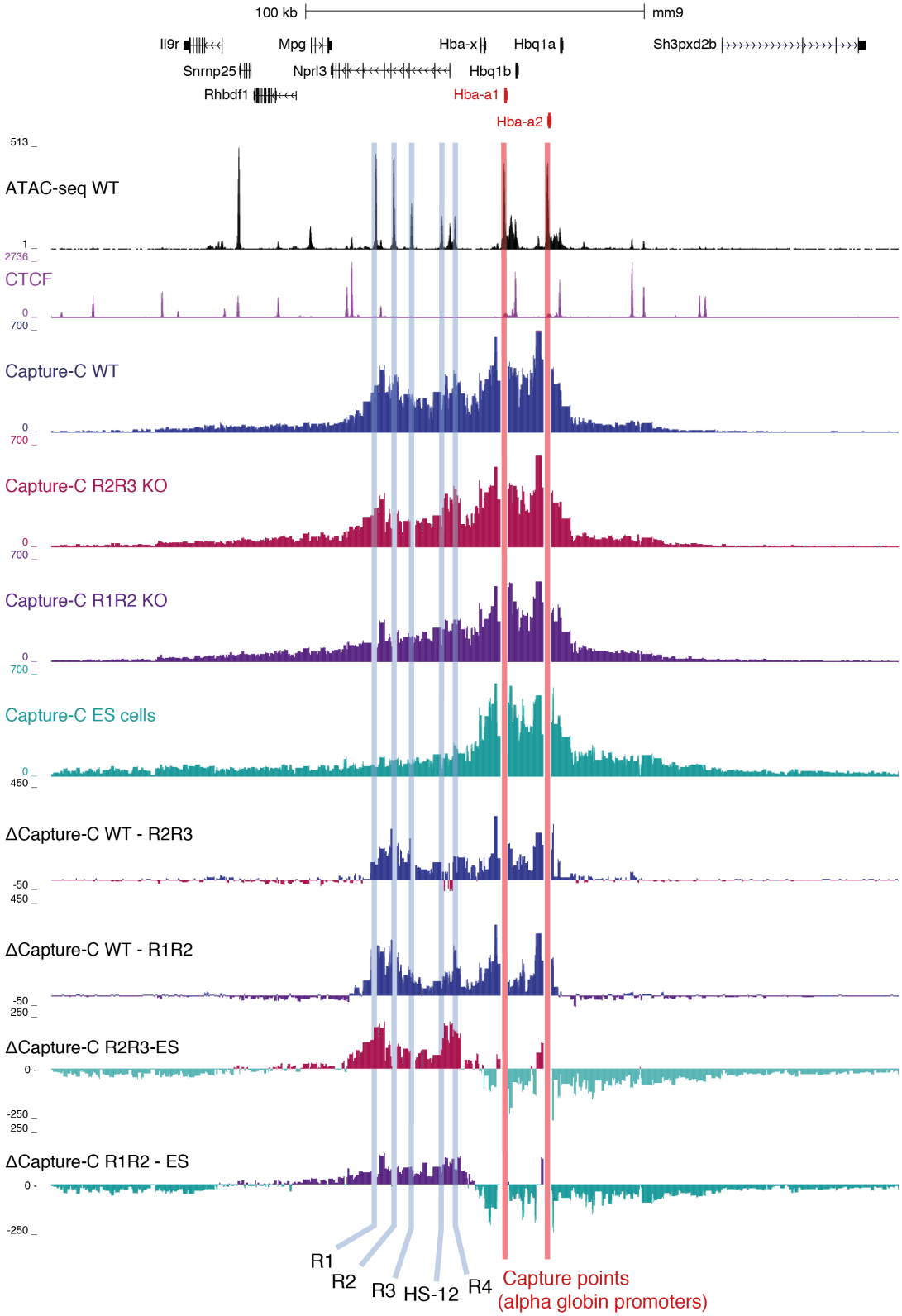
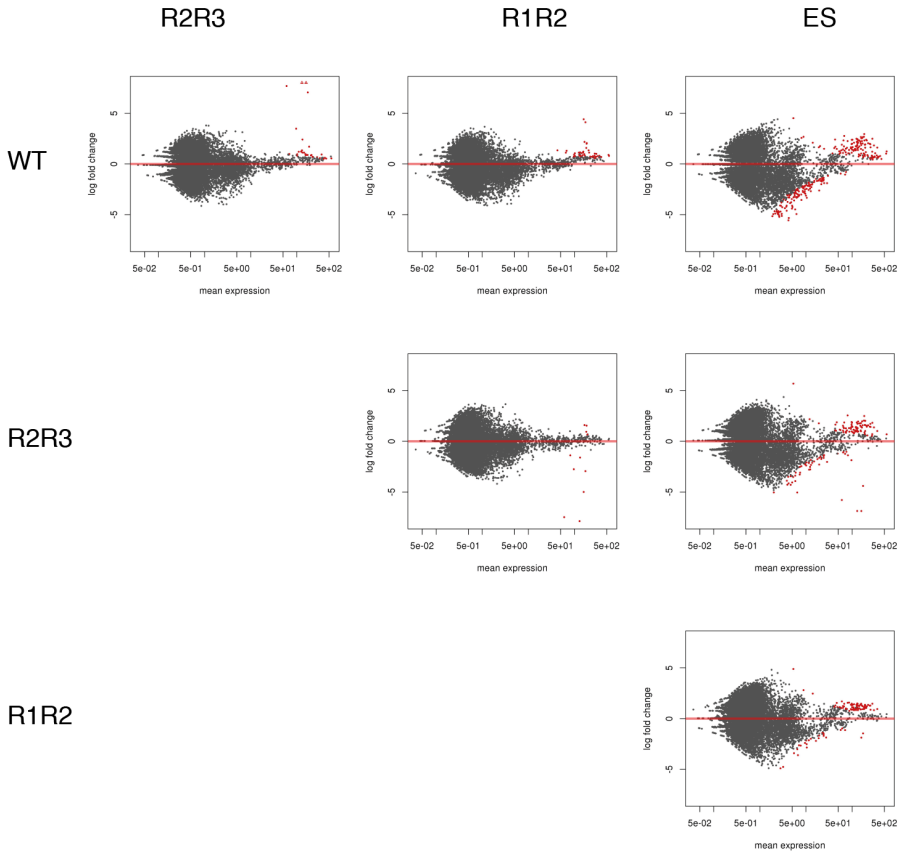


Figure 8.3 Effects of enhancer knock-out on the Capture-C interaction profile

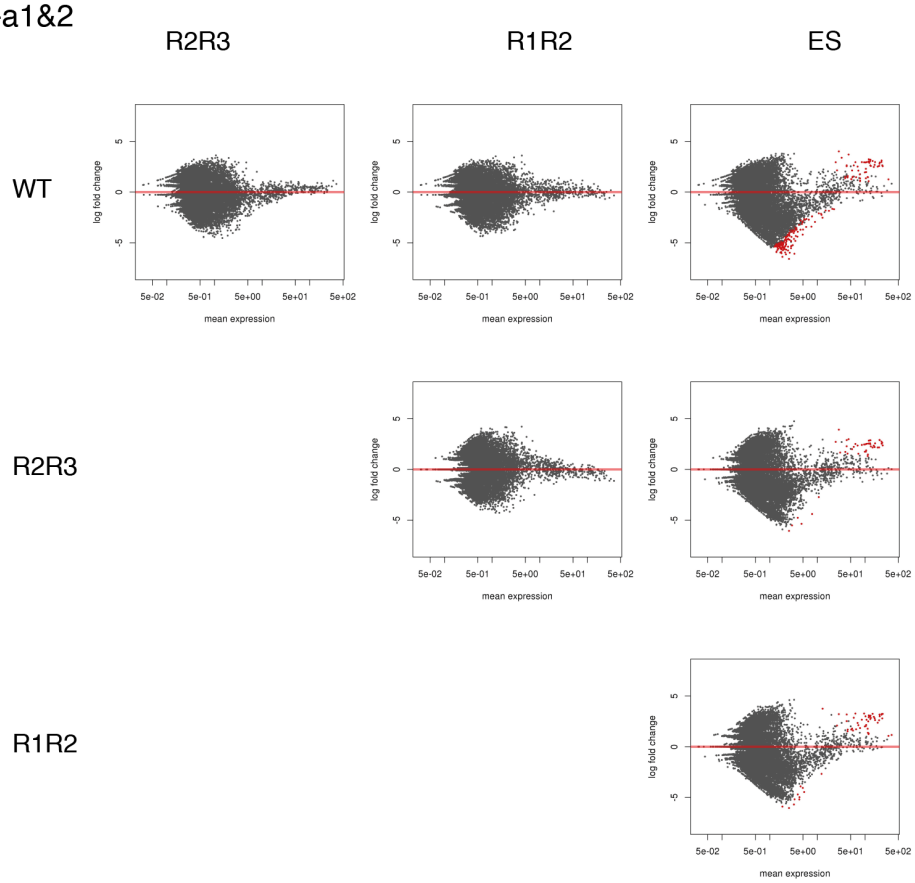
Comparison of the interaction profiles from the *Hba-a1&2* promoters in WT, R1R2 KO, R2R3KO and ES cells (E14). NG Capture-C was performed using the Hba-a1&2 promoters as a viewpoint (since the genes are virtually identical this

represents a composite interaction profile from both promoters). The Δ Capture-C profiles show the subtraction of the one profile from another. These show that enhancer KO results in diminished interactions between the enhancers and the promoters of *Hba-a1&2* compared to WT. However, the interaction profile is clearly still present in the R2R3 KO. In the R1R2 KO interactions are also clearly present between the promoter and enhancers, relative to ES cells. In particular, there appear to be conserved interactions with the CTCF site between the R1 and R2 enhancers.

Hba-a1&2



Hbb-a1&2



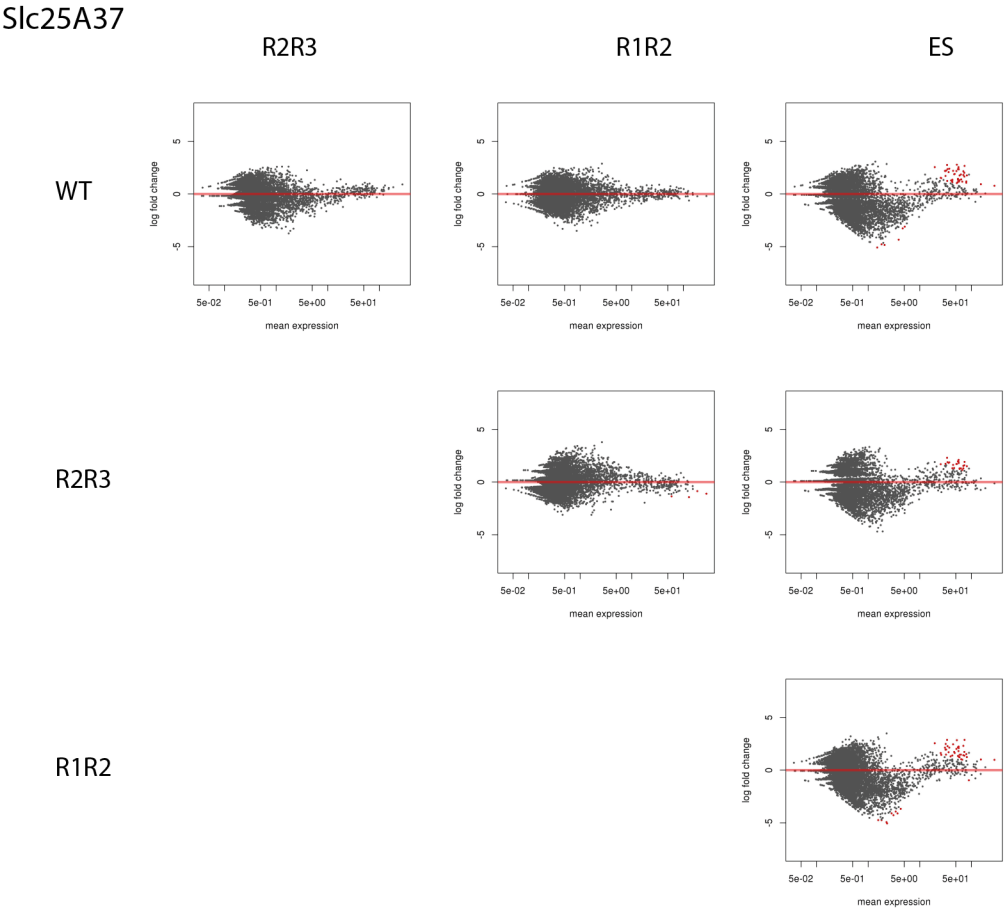


Figure 8.4 Statistical analysis of the Enhancer KO datasets.

This figure shows MA plots comparing Capture-C data in WT, R1R2 KO, R2R3 KO and ES cells at the alpha globin (Hba-a1&2), beta globin (Hbb-b1&2) and mitoferrin (Slc25A37) loci. Raw interactions counts per restriction fragment from three biological replicates from each condition were analysed using DESeq2, which is a Bioconductor package extensively used for analysis of next generation sequencing data sets, particularly RNA seq data. It uses a model of the negative binomial distribution, which applies to count based data sets such as Capture-C data in which one only sees positive interactions (there are no negative data points denoting an absence of interaction). MA plots (an form of Bland-Altman plot showing log fold change plotted against mean expression) were generated using R to compare each of the data sets in each of the genes. These show that there are significant differences (denoted by red data points) between ES cells and the WT and enhancer knock out cells at all of the loci. Importantly, however, significant differences were only found between WT and enhancer knockout at the alpha globin locus. At the beta globin and mitoferrin

loci no significant differences were found between WT and the enhancer knockout.

8.5.3 Conclusions

These data are not completely in keeping with the active chromatin hub model (Tolhuis et al., 2002), in which specific interactions occur between the regulatory elements and promoters (Figure 8.5).

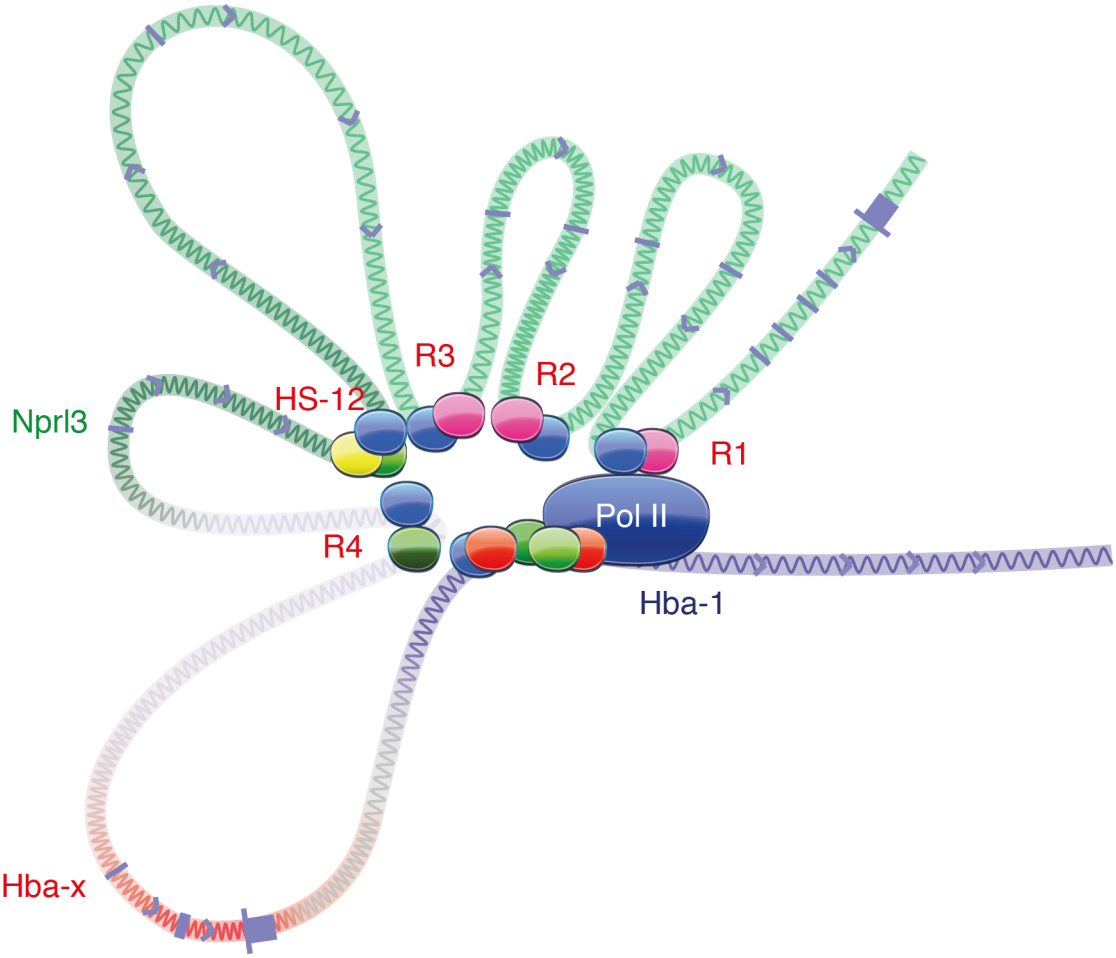


Figure 8.5 Active chromatin hub model

The profiles seem to show that the chromatin has much broader domains of interaction than can be explained by protein-protein interactions between the transcription factors at the enhancers and the transcription machinery at the promoter. However, within these domains there are probably more specific interactions between the enhancers and the promoters. The CTCF sites seem to

be found at the boundaries of these domains although this is not exclusively the case. In addition it appears possible for sub domains of stronger interaction to occur as is seen when contrasting the profiles from the *Hba-a1* and *Hba-a2* genes.

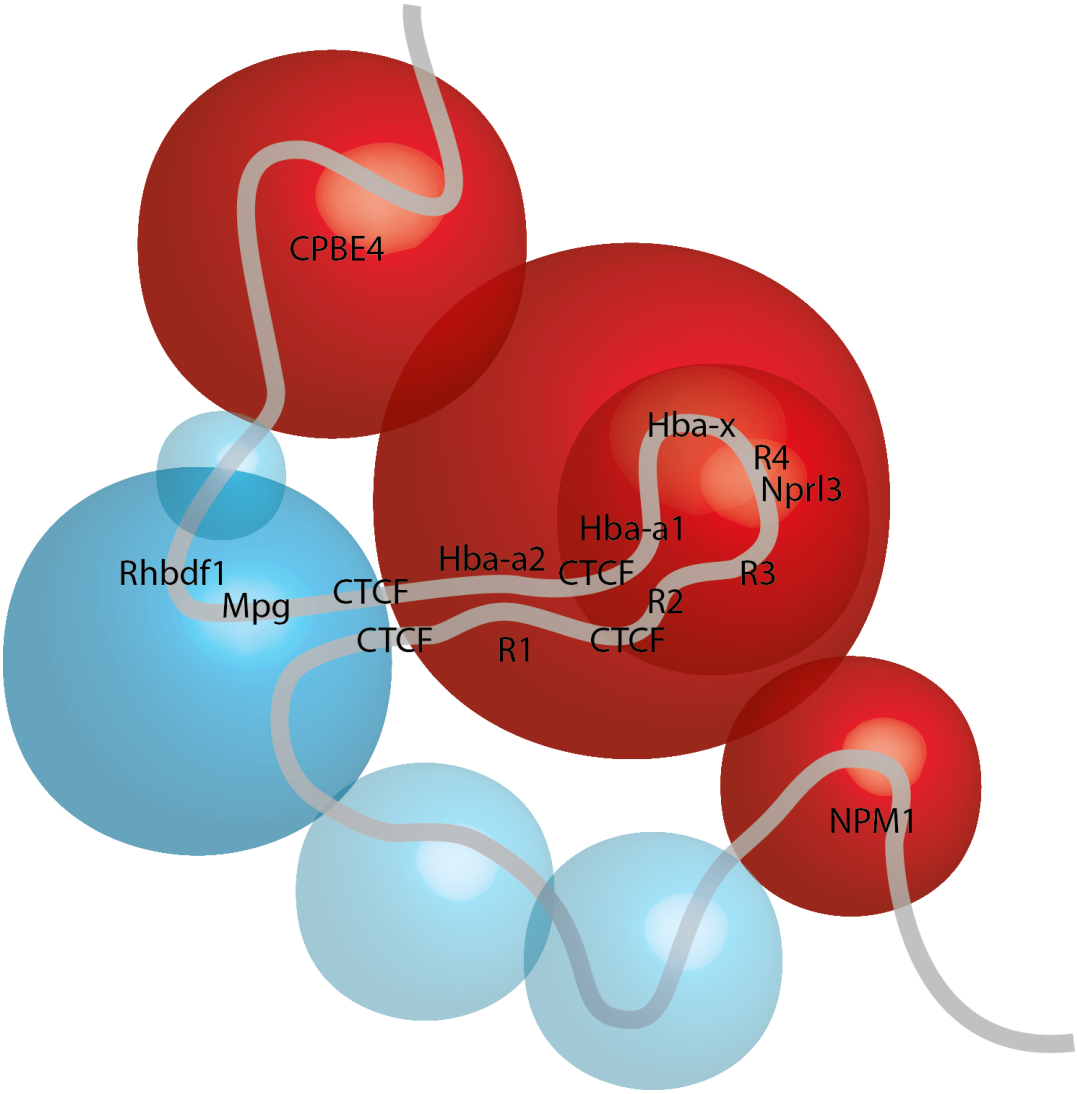


Figure 8.6 Domain model of interactions around the alpha globin locus

Here the red spheres represent domains containing active elements, which can interact weakly with one another and the blue spheres contain inactive elements that also tend to interact with each other.

Brackley et al., have modeled the chromatin fiber at the alpha globin locus and compared the model with our experimentally derived data (Brackley et al., 2016). This shows that the chromatin has multiple potential conformations that are transiently stable and the experimentally determined profiles are a

composite of these. Again this implies that the experimentally determined profiles that we generate by 3C represent probability distributions of the chromatin interactions.

The findings from the knockout mice show that the interaction profile is only partially dependent on the enhancer sequences, so it is unlikely to be mediated wholly by protein-protein interactions between the transcription factors at the enhancers and the transcription machinery at the promoter. In addition we have observed that deletion of pairs of enhancers reduces the interactions with the remaining enhancers at the locus.

This provides a potential explanation for the spatial clustering of enhancers into locus control regions or super-enhancers. It is not that super-enhancers have a singular composite function per se, it is simply that strong enhancers augment the interactions between their local chromatin environment and the promoters of the genes they interact with. This generates a compartment of interaction between the chromatin surrounding the enhancer and the promoter, which results in a favorable location for other enhancers to evolve. Thus nascent enhancers are more likely to evolve in these regions and consequentially senescent enhancers will also be colocalise in these clusters.

9 Development of the use of NanoString technology to analyse 3C libraries

9.1 Aims

We wanted to develop NanoString technology to analyse 3C libraries because we felt that it might have significant advantages over methods of 3C library analysis. In particular, it has the potential to be the only 3C based method that does not involve any PCR amplification or next generation sequencing of the libraries and the inherent biases associated with these. It is therefore likely to measure the relative proportions of interactions in the library more quantitatively than other existing methods. In addition NanoString technology is able to analyse 12 samples simultaneously in less than one day allowing multiple samples to be analysed quickly. Another significant advantage is that there is no need for complex bioinformatics in order to analyse NanoString data.

9.2 Overview of NanoString technology

NanoString nCounter technology was initially developed to quantify messenger RNA transcripts (mRNA) (Geiss et al., 2008) and it has been used to count DNA molecules in copy number variation (CNV) assays (Ruderfer et al., 2013) and ChIP experiments (Ram et al., 2011). Nanostring technology uses reporter probes, which comprise a specific barcode of six different fluorophores attached to a RNA oligonucleotide complementary to half of the sequence of interest and a biotinylated capture probe complementary to the other half of the target sequence. Following a hybridisation reaction the samples undergo washes capturing from the reporter and subsequently the capture probes so that only molecules hybridised to both moieties are retained. The trimeric complexes are then immobilised in an electric field using the biotin moiety on the capture probe and the reporter oligonucleotides are counted by imaging the barcoded fluorophores.

9.3 Adaption of NanoString technology for analysis of 3C libraries

9.3.1 Experimental design

This technology was adapted to quantify the interactions in 3C libraries by designing capture oligonucleotides to the ends of the restriction fragment of the alpha globin promoters, which we used as the viewpoint for the assay. 120 reporter oligonucleotides were designed to the restriction enzyme fragments across the alpha globin locus including fragments containing all the previously identified regulatory elements and gene promoters and multiple control regions. This design allows quantification of the ligation junctions present in the 3C library between the alpha globin promoter and the restriction fragments to which reporter probes had been designed.

A custom perl script was written to produce the sequences of all of the possible restriction site combinations with the alpha globin promoters for 2Mb across the alpha globin locus. The strand of the probes had to be carefully taken into account in the design of the sequences for oligonucleotide design (the top strand of the left hand end and the bottom strand of the other end of the fragment were used). The team at Nanostring technologies produced sequences of all of the possible oligonucleotide combinations across the locus using their pipeline. We then further filtered these sequences using our own filters for repetitive sequences.

9.3.2 Sensitivity of the Nanostring ncounter system

Nanostring quotes that the system has a sensitivity of 0.1-0.5 fM, which corresponds to a limit of detection of between 2,000 and 10,000 molecules (Geiss et al., 2008). This means that in order to be able to detect a ligation junction present that was present at 1 part in 1000 in the 3C library 1 million cells, corresponding to around 6ug material, would need to be used. It was therefore important to maximize the input material use the optimal NanoString settings to maximize the sensitivity of the assay.

9.3.3 Preparation of 3C libraries

3C libraries were prepared using the four cutter restriction enzyme DpnII as described in Chapter 2. The 3C libraries were sonicated to 200bp using a Covaris S220 Focussed ultrasonicator (6 cycles of 60s: duty cycle 10%; intensity 5; cycles per burst 200). A vacuum centrifuge (50 °C) was used to concentrate the libraries down to a volume of 5ul. The experiment was initially done with two biological replicates in erythroid and ES cells at 3 different DNA concentrations (5ug, 2ug and 500ng).

9.3.4 Hybridisation with NanoString probes

The libraries were initially denatured by heating to 95 °C for 10 minutes. Subsequently they were mixed with the NanoString code set and incubated at 65C for 18h to allow the capture and reporter probes to anneal to the 3C template.

Following the hybridization reaction the material was loaded onto the NanoString cartridge using the nCounter Prep Station. The nCounter Prep Station performs sequential capture steps first from the capture probe and then using the reporter probe so that only molecules that contain both the capture and reporter probe are loaded onto the cartridge. For this step the high sensitivity protocol was used to maximize the number of molecules that bind to the surface of the cartridge and the signal to noise ratio.

The cartridge was then analysed by the nCounter Digital Analyser on Maximum resolution settings to maximize the number of molecules counted by the digital analyser.

9.3.5 Effects of impurities in 3C libraries

Initially it was found that increasing the amounts of 3C library to the levels that had been calculated would be necessary for the detection of ligation junctions lead to a dramatic loss of the NanoString signal including from the internal positive controls (Figure 9.1).

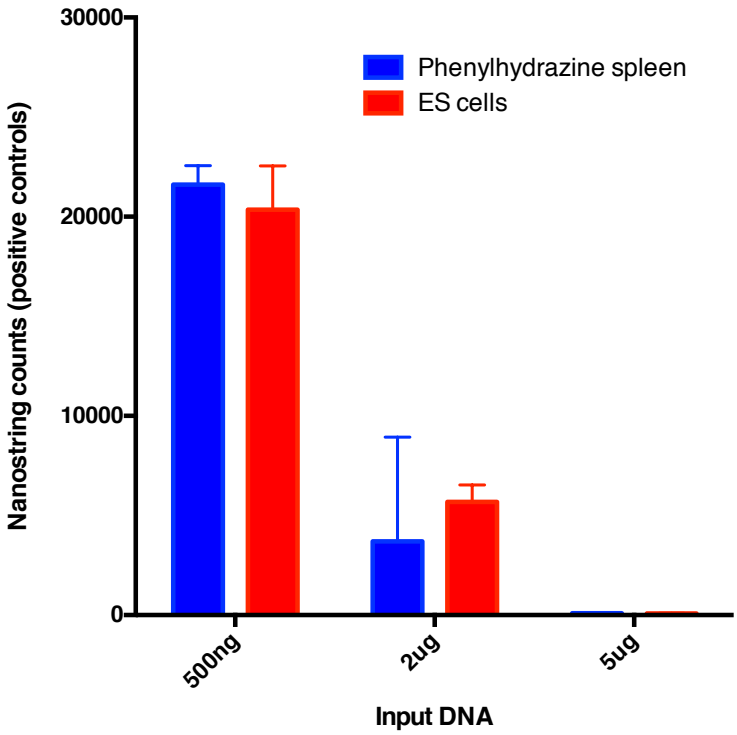


Figure 9.1 Effects of contaminants in the 3C library on NanoString counts

The NanoString ncounter system uses several spiked positive controls. The counts from these controls are clearly diminished by increasing the amount of 3C library used as an input.

3C libraries often contain large amounts of dithiothreitol (DTT). This is present at 10mM in the T4 DNA ligase buffer and precipitates with the DNA during the ethanol precipitation. Although the 3C library protocol performs the precipitation in 35ml ethanol in an attempt to reduce DTT contamination this does not completely remove it from the libraries. One of the problems encountered when adapting NanoString technology to analysis of 3C libraries was that the material for analysis to be reduced to a volume of 5ul for the hybridization reaction with the NanoString probes. This was done using a vacuum centrifuge and so the contaminants in the 3C library were also concentrated.

A SPRI bead clean up step was therefore introduced following sonication of the 3C library, prior to the vacuum centrifugation. This purified the library sufficiently that the contaminants it did not affect the NanoString internal positive controls.

9.4 Results

9.4.1 NanoString can determine an accurate interaction profile

The NanoString nCounter count was highly reproducible between technical replicates generated from both erythroid cells (phenylhydrazine treated mouse spleen) and ES cells using different amounts of input DNA (Figure 9.3).

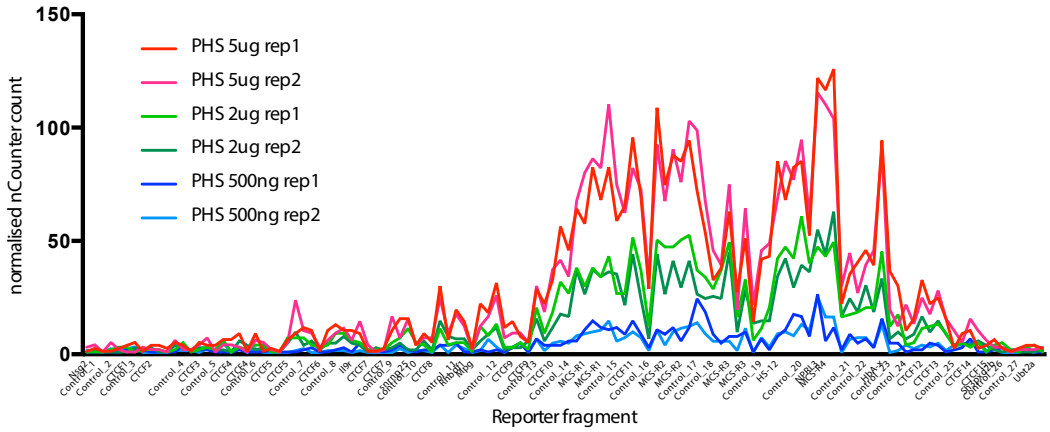


Figure 9.2 Replicates of C-nanostring from erythroid cells

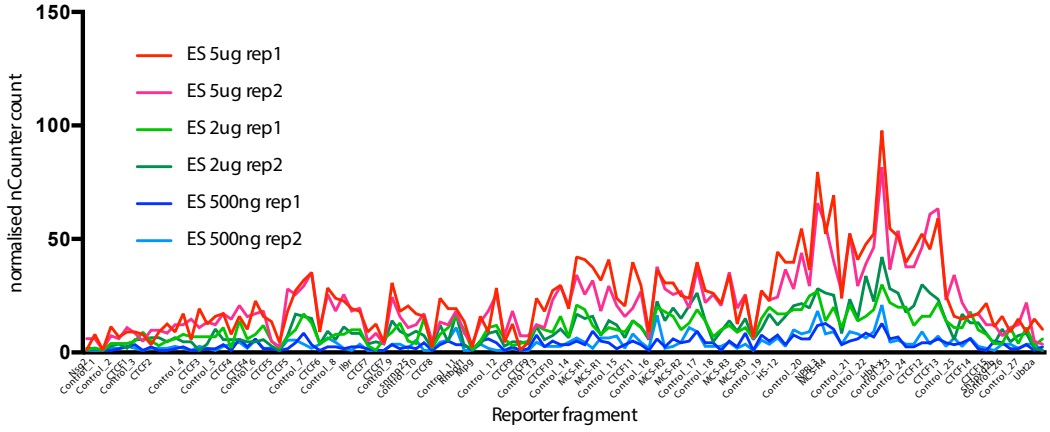


Figure 9.3 Replicates of C-nanostring from ES cells

There was a clear linear relationship between the quantity of input 3C library and the nCounter count (Figure 9.4). This indicates that the nCounter reliably correlates with the number of ligation junctions in the 3C library.

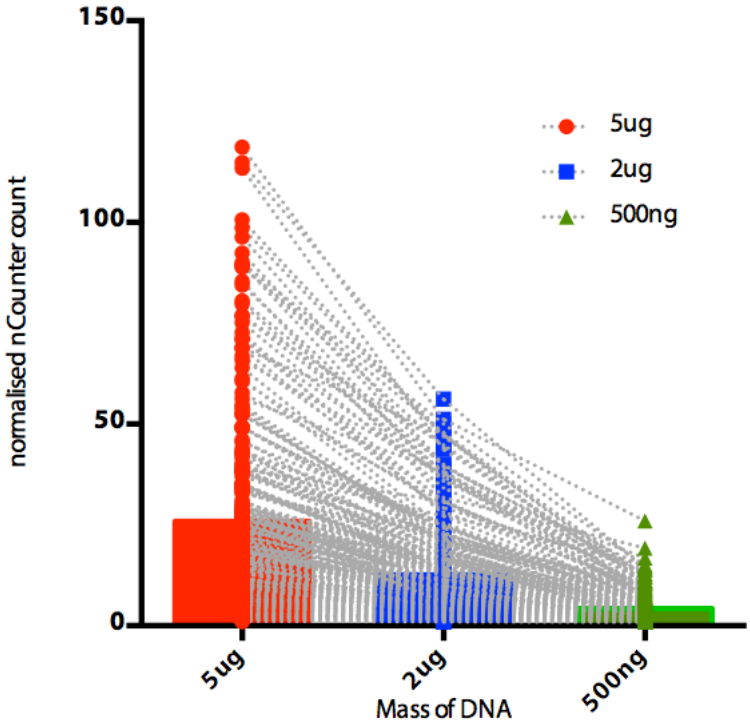


Figure 9.4 Combined strip chart and barograph of the relationship between nCounter count and amount of input 3C library

Individual data points from each replicate and probe are plotted. The grey dotted lines link data points from the same sample and reporter probe. The bar graphs represent the mean counts. There is a near linear relationship between the quantity of 3C library used and the nCounter count.

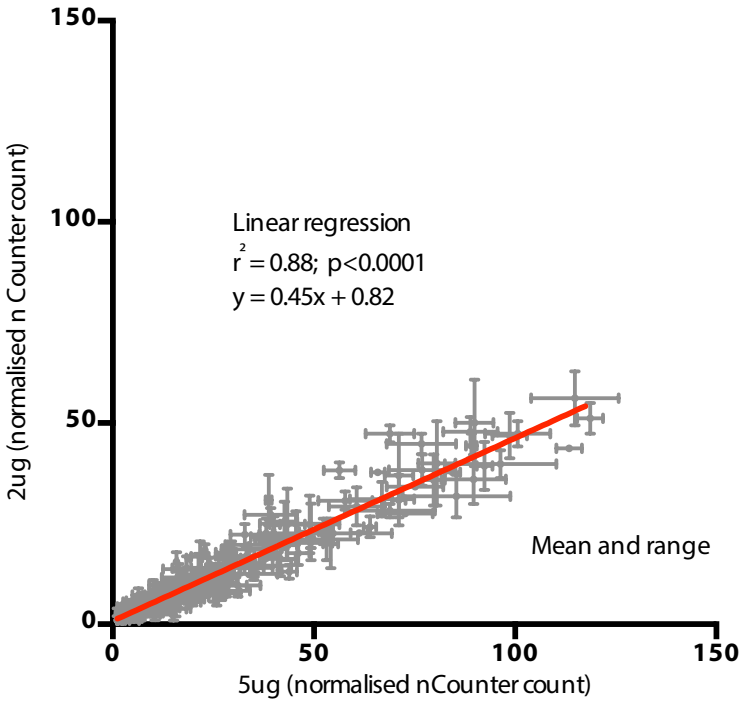


Figure 9.5 Scatter plot comparing when 2µg and 5µg input 3C library are used

The bars on the scatter plot represent the two replicates at each concentration. The red line shows the linear regression line. There is a close correlation between the normalized nCounter count using 2µg and 5µg input (R2 = 0.88). In addition there gradient of the linear regression line is close to what would be expected if there was a perfect relationship between the number of ligation junctions present in the input material and the nCounter count. The formula for the regression line is y =0.45x +0.82 compared y = 0.5x would be expected if the relationship were perfect.

9.4.2 Comparison of C-NanoString to Capture-C

The nanostring assay replicates the previously well-defined interaction profile at the alpha globin locus in erythroid cells (Figure 9.6 & Figure 9.7). The interactions with the enhancers are largely absent in mouse embryonic stem cells, which do not express alpha globin.

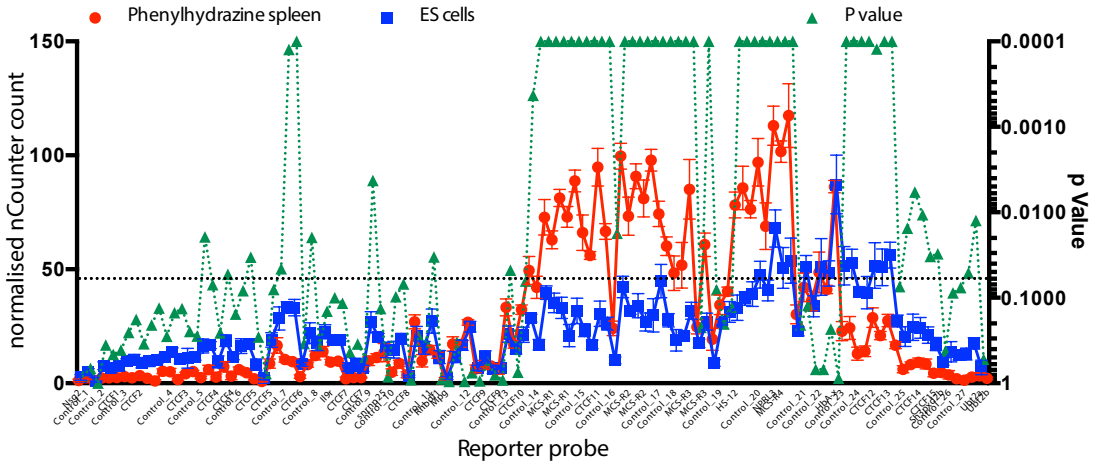


Figure 9.6 Statistical analysis of Erythroid and ES cell replicates

The interaction profile across the alpha globin locus in erythroid cells derived from the spleen of a phenylhydrazine treated mouse (red) and ES cells (blue). Multiple t-tests were used to compare the two profiles (these were corrected for multiple comparisons using the Holm-Sidak method (GraphPad Prism 6)). The p value for the comparison of the on and off states is shown in green (Right y axis). These show that the erythroid cells interact more strongly than ES cells with the enhancers of the alpha globin genes. Interestingly the interactions outside of the alpha globin locus appear to be stronger in ES cells and these reach statistical significance at several points.

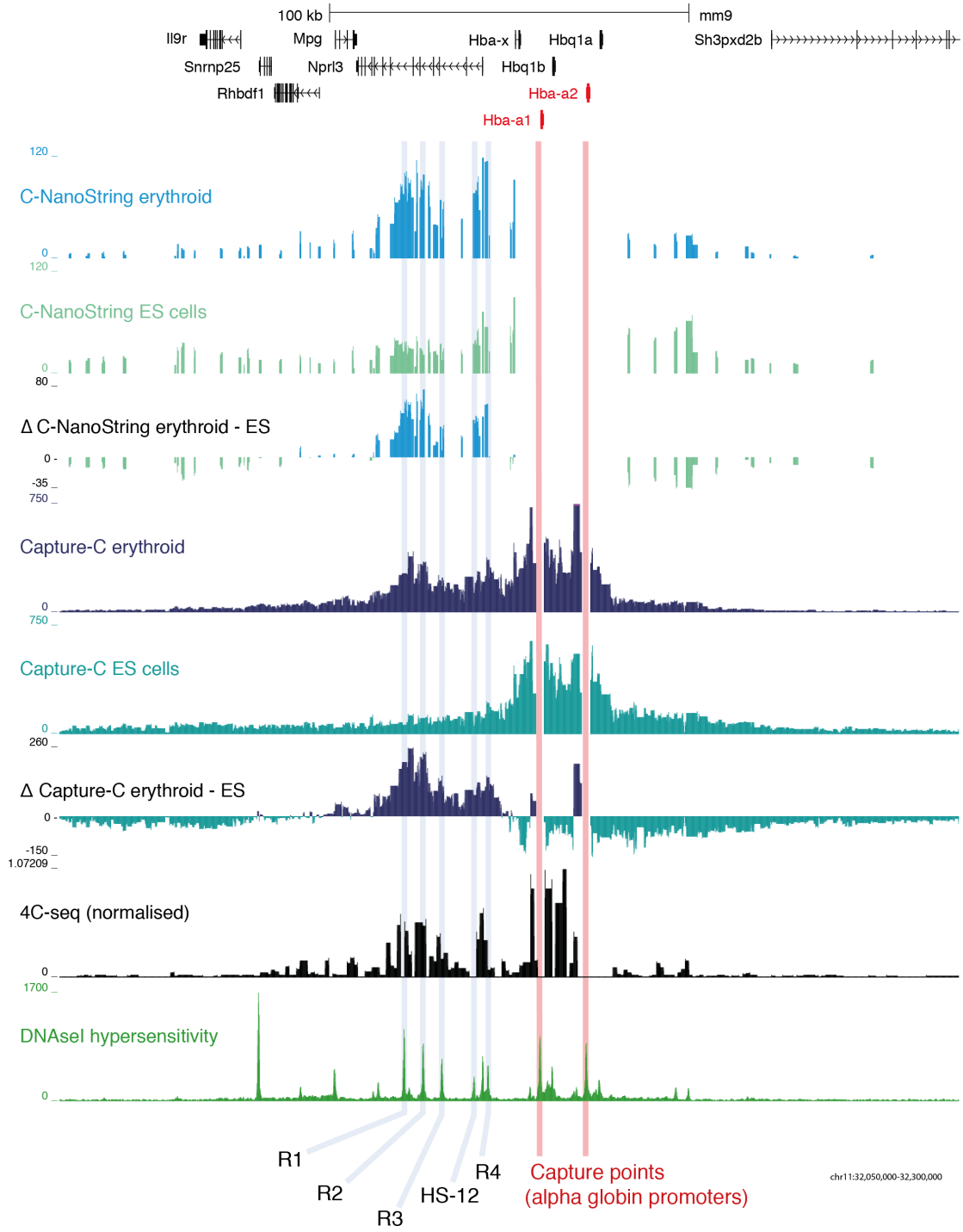


Figure 9.7 Comparison of C-NanoString with Capture-C and 4-C seq

This figure plots the C-NanoString data using the UCSC genome browser. Clear interactions are seen with the regulatory elements in the erythroid cells that are not present in the ES cells. Clearly the n-counter system cannot provide as comprehensive a profile as Capture-C but it is possible to place hundreds of probes to interrogate the locus in considerable detail.

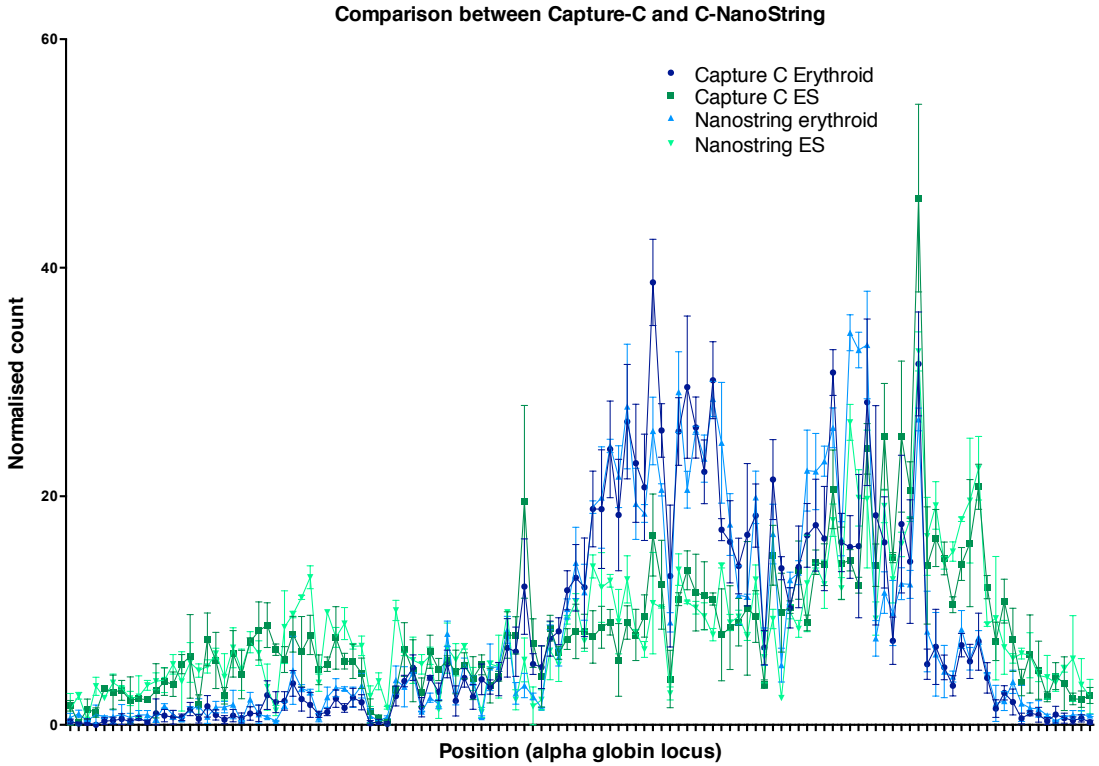


Figure 9.8 Comparison between Capture-C and C-NanoString

The figure shows the mean (± SD) count normalized to the total count across the all the restriction fragments analysed by C-NanoString for both Capture-C and C-NanoString. The Capture-C profile shows data from the 4 replicates of erythroid cells and 3 ES cell replicates. The NanoString data are derived from both the 5µg and 2µg input data sets corrected for the quantity of input 3C library.

There is close correlation between the ligation junction frequencies determined by Capture-C and C-NanoString.

Friedman's test for multiple non-parametric comparisons shows that there are statistically significant differences between the different cell types but there are no significant differences between the Capture-C and Nanostring profiles within the same cell type.

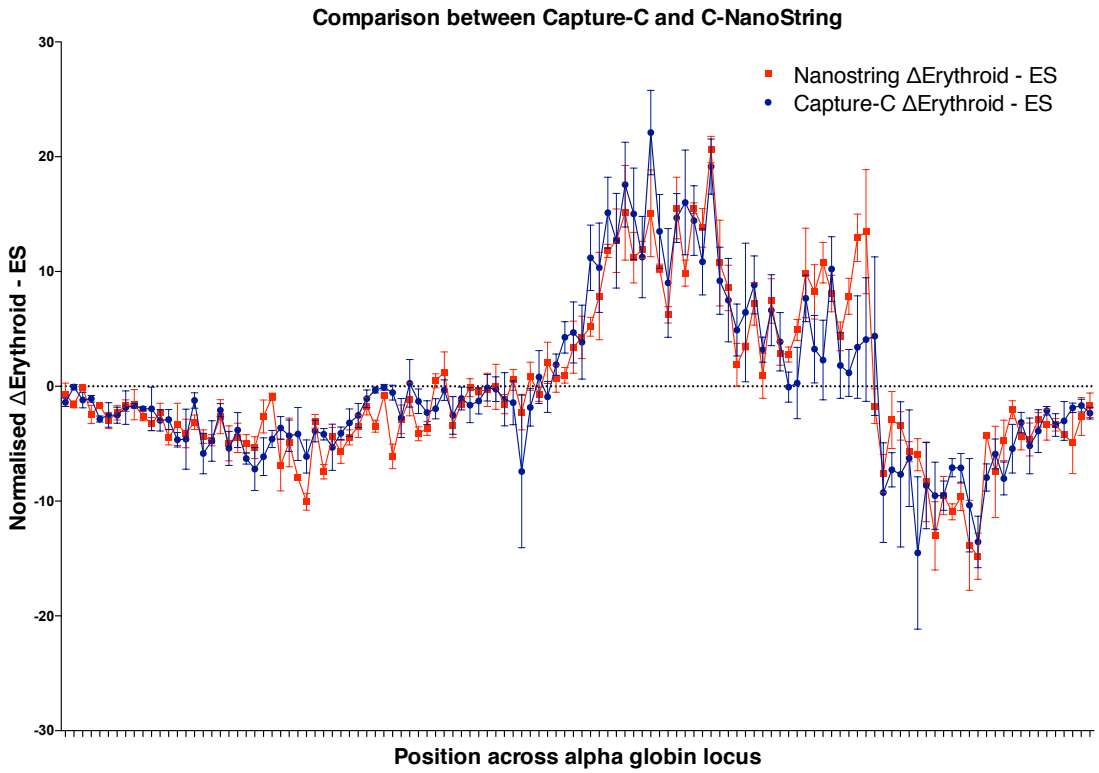


Figure 9.9 Comparison between Capture-C and C-NanoString

This graph compares the differential profiles determined by Capture-C and C-NanoString from erythroid and ES cells from the alpha globin promoter.

9.4.3 Conclusions

Nanostring technology can be adapted to analyse 3C libraries accurately and reproducibly. This technology is potentially superior to current techniques if one wishes to analyse the interactions in multiple samples (12 samples can be analysed simultaneously) rapidly over 50-300 different sites. It would be well suited for laboratories who do not have extensive experience with high throughput sequencing based experiments particularly because the data analysis is much more straightforward than in sequencing based experiments.

There is an excellent correlation between Capture-C and nanostring data. This shows that Capture-C can quantify the ligation junctions present in 3C libraries accurately. Nanostring analysis is PCR free and likely to be highly quantitative, since it is based on counting molecules and there are no exponential amplification steps. Furthermore the biases resulting from nanostring can Capture-C based quantification of 3C libraries are likely to differ. Areas of

repetitive sequence are likely to be over represented in nanostring counts because there is a potential for crosshybridisation and in Capture-C repetitive sequences are likely to be underrepresented because bowtie will discard sequences that can align to the genome in more than two places. Therefore the two techniques corroborate each other.

10 C-trap and C-minion

10.1 Aims

These techniques have been developed to investigate which interactions occur simultaneously when a single regulatory element interacts with a promoter (Figure 10.1). This has the potential to differentiate between an active chromatin hub model (de Laat and Grosveld, 2003) and a sequential model for activation of enhancers (Andrey et al., 2013).

When 3C libraries are made the ligation reaction results in a large concatamer (over 10kb in size, Figure 2.7). The strongest interactions are represented by ligation junctions present at levels of around 0.5-1% in a 3C library generated by a four cutter restriction enzyme. It is therefore very common for the other restriction fragments to become ligated between the regulatory element and the promoter. It is possible to PCR amplify these intervening 'trapped' fragments using 3C primers designed to the regulatory element and the promoter.

10.2 C PCR-trap

10.2.1 Methods

3C libraries were amplified using primers designed to the ends of the DpnII fragments containing the alpha globin promoters and the R2 enhancer, which is one of the important enhancer elements in the mouse alpha globin locus (Hughes et al., 2005). Since interactions can occur between either end of each restriction fragment, primers were designed to both ends of each fragment. Careful attention had to be paid to strand, when designing these primers due to the rotational symmetry of the restriction fragment cut sites (Figure 10.1).

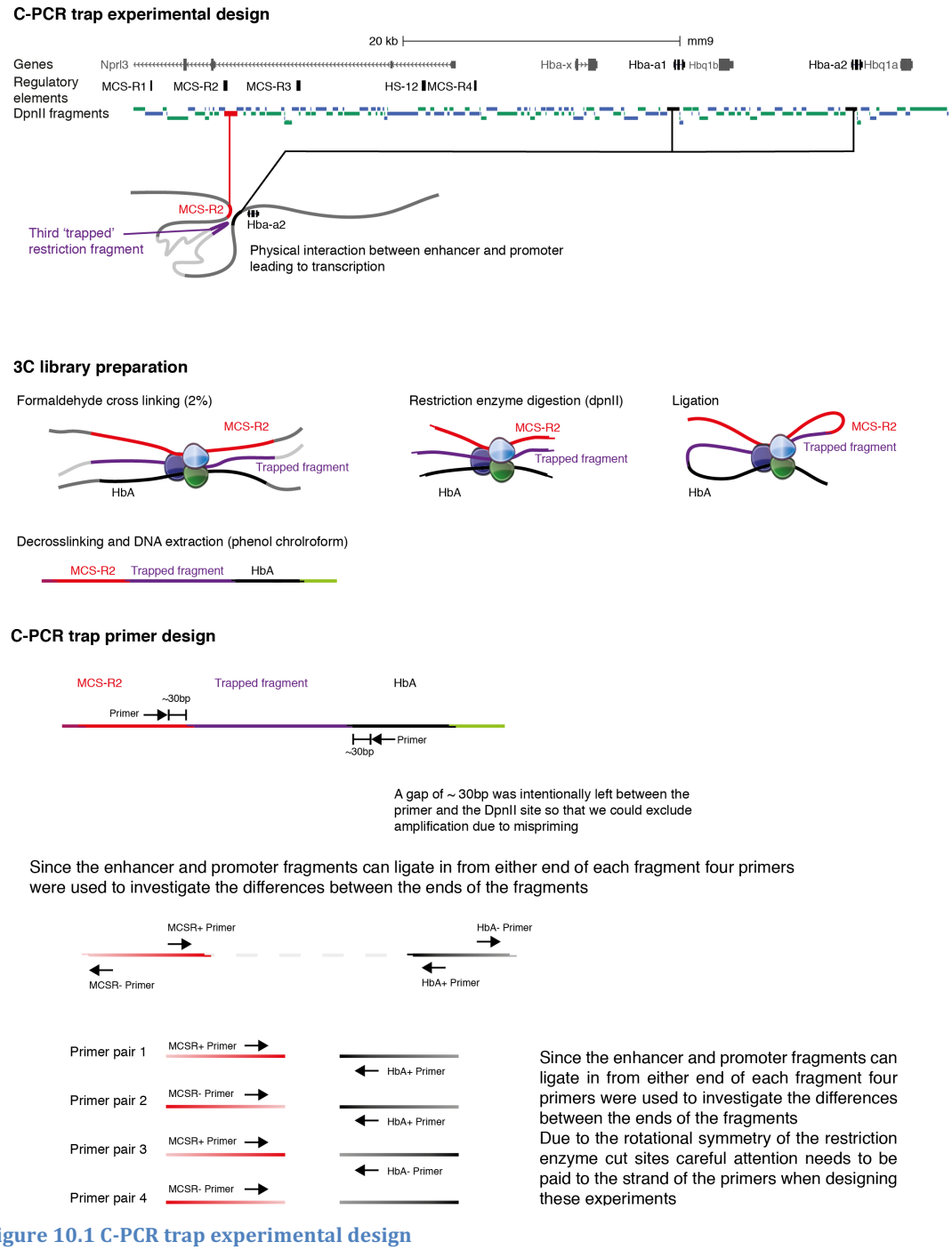


Figure 10.1 C-PCR trap experimental design

C-PCR trap aims to identify interactions that occur at the same time as when the alpha globin promoters interacts the R2 regulatory element. In conventional 3C library preparation elements often become ligated between any two fragments o interest (in this case we have used the alpha globin promoter and the R2 enhancer sequences). In order to allow these sequences to be sequenced they were amplified using PCR with primers designed to then amplify from the alpha globin and R2 fragments. The primers were designed around 30bp back from

the restriction enzyme cut site so that mispriming could be excluded bioinformatically. Since interactions can occur from either end of each fragment, primers were designed to each end of the restriction fragment and the strand of the primers was carefully selected to reflect the rotational symmetry of the restriction enzyme cut sites. These primers were combined to make four pairs of primers so that all the permutations of ligation junctions between the two fragments could be amplified.

The PCR reaction was optimized to use as much 3C material as possible and to amplify large GC rich fragments (Advantage GC2 polymerase and GC melt were used). This resulted in amplification of sequences up to and above 2kb trapped between the alpha globin and R2 fragments (Figure 10.2). Controls included ES cells (in which the alpha globin promoter does not interact significantly with the R2 enhancer); undigested DNA from mouse erythroid cells and digested but unligated chromatin. Amplification occurred in ES cells but to a much lesser degree than in erythroid cells. The other three controls did not did not amplify significantly.

PCR setup

Water	21µl
5x GC PCR buffer	10µl
GC melt	5µl
50x dNTPmix	5µl
50x Advantage GC2 polymerase	1µl
Primers	2µl
DNA template (500ng)	10µl

PCR conditions

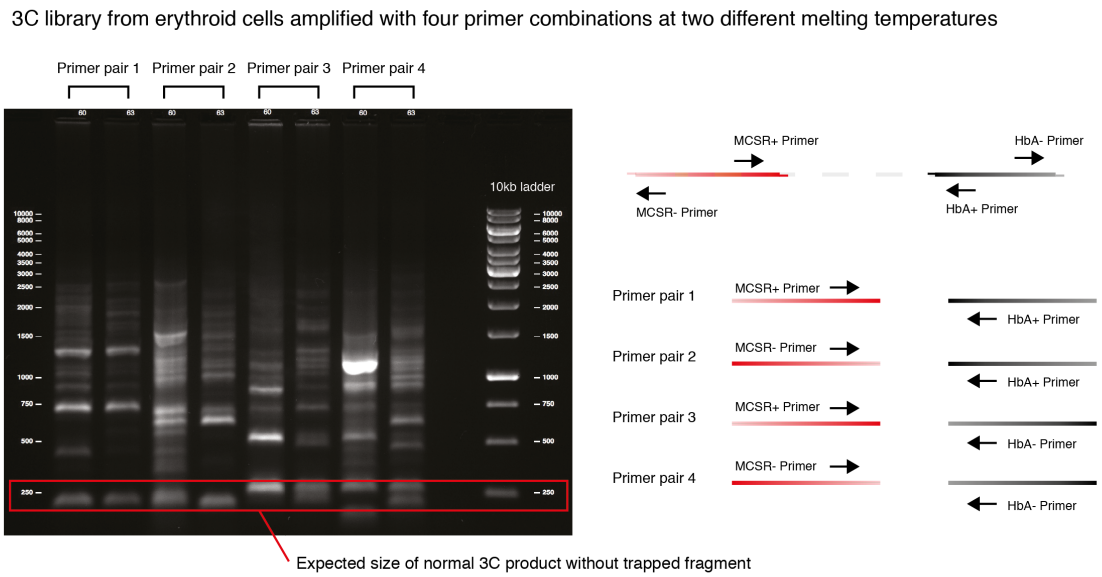
Lid temperature 100°C	
94 °C	3 min
35 Cycles	
94 °C	30 s

60-70°C gradient	3 min
60-70°C gradient	3 min

PCR primer sequences

Globin +	GGCTCAAGTGAATCTGCTAGC
Globin -	AGACAAAAGCAACATCAAGGCTG
R2 +	AATGAGGCAGAGTTTAATGGGAATG
R2 -	TATTACTCAGCAGTCCTAGTGGC

Following PCR amplification the material was sonicated to ~300bp using a Covaris sonicator. It was then sent to the Wellcome Trust Centre for Human Genetics for high throughput sequencing. A single lane of Illumina HiSeq 2000 was used to sequence the four independently indexed samples 100bp paired end.



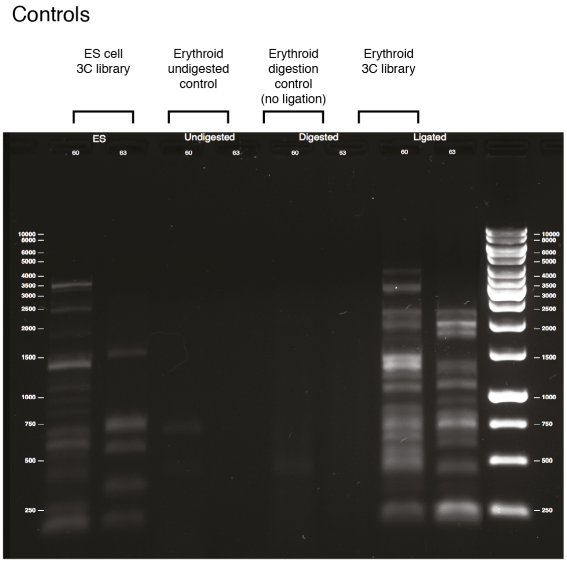


Figure 10.2 Agarose gel electrophoresis of PCR products from C-PCRtrap experiments

1% agarose gel electrophoresis with ethidium bromide of PCR products from erythroid 3C libraries using the different primer combinations at two different annealing temperatures (60°C and 63°C). This shows amplification of fragments up to 2.5kb with all of the combinations. The red box highlights the PCR product of the expected size if the promoter and enhancer ligate directly to one another.

1% agarose gel electrophoresis with ethidium bromide of PCR products from controls. These included ES cells in which the alpha globin gene is not expressed. This shows that there are still some interactions between the promoter and the

R2 enhancer but these are infrequent compared to erythroid cells. The other controls included undigested DNA from mouse erythroid cells and digested but unligated chromatin, which did not show significant amplification.

10.2.2 Results

These data were analysed using a very similar strategy to Capture-C except that it required exclusion zones around both the R2 enhancer and the alpha globin promoters. In addition no filtering for PCR duplicates could be performed because the 3C library is sonicated after the PCR reaction.

The data are shown in Figure 10.3. These data appear to show that when the R2 enhancer and alpha globin promoter interact they also interact simultaneously with the other regulatory elements and gene promoters in the locus. This interaction profile is similar to the 4C-seq profile from the alpha globin promoters. However, the profile is incomplete by comparison to the NG Capture-C profile from the alpha globin promoters. The gaps in the profile with the PCR trap experiment and 4C-seq data almost certainly occur because some fragments fail to amplify efficiently despite the attempts to optimize the amplification of long GC rich sequences.

Although the interaction profile varied depending on which combination of primers were used, it is difficult to know whether this is due to genuine differences in the interactions from the ends of the alpha globin and R2 fragments or whether it is due to differences in PCR efficiency between the different primer combinations.

The other problem with this experiment is that it is difficult to know how many ligation junctions are contributing to the data set.

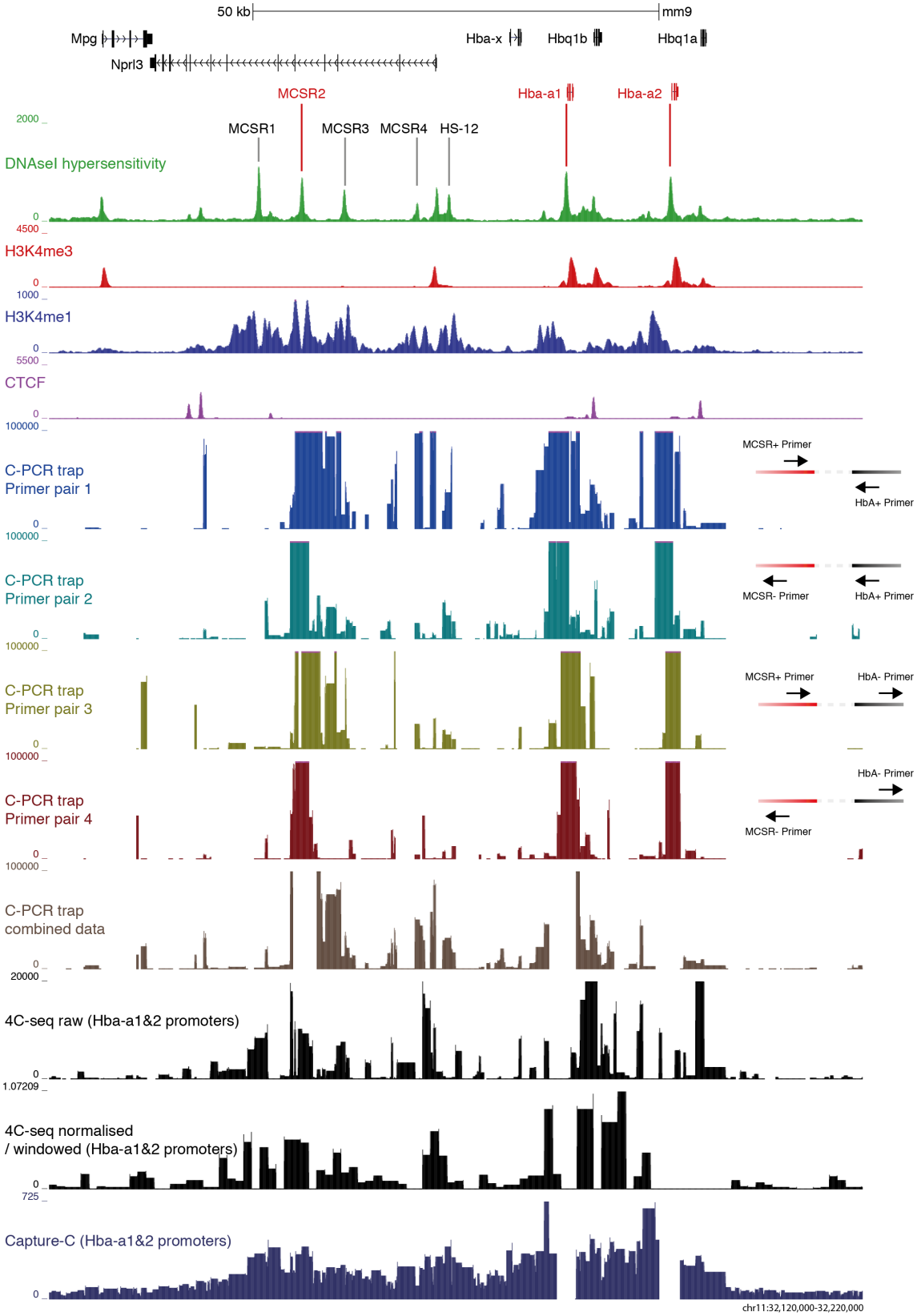


Figure 10.3 C-PCR trap results

3C library generated from erythroid cells was PCR amplified using primers designed to amplify from the restriction fragments containing the regulatory element R2 and the Hba-a1&2 promoters (which are virtually identical). Four

pairs of PCR primers (Primer pairs 1-4) were used to amplify the material so that all of the trapped interactions between either end of restriction fragments could be sequenced. The orientation of the primers is shown on the right hand side of the figure. The C-PCR trap combined data shows a composite profile from all of the primer pairs and this includes a bioinformatic exclusion zone around the primer sites. For comparison 4C-seq data (van de Werken et al., 2012b) and NG Capture-C data are also included. The 4C-seq shows two tracks: the top track shows raw data as a mapped to the individual restriction fragments (analysed using the same pipeline as the C-PCR trap and Capture-C tracks) and the second track shows normalized 4C-seq data.

10.3 C-MinION

10.3.1 Aim

The MinION is under development by Oxford Nanopore Technologies. It uses nanopore technology to sequence DNA and is capable of producing very long reads (up to 40kb). As part of the MinION access program we thought that it would be interesting to use this technology to sequence the libraries generated by PCR amplification from the promoter and enhancer as described above for the C-trap experiment. The aim being to use the very long reads that the MinION is capable of generating to define whether the enhancer elements interact simultaneously or whether there are more transient interactions.

10.3.2 Methods

3C libraries generated from erythroid cells were PCR amplified using the Primer Pair 1 for amplification of sequences from the Hba-a1&2 promoters and R2 enhancer.

Addition of the MinION adapters and operation of the MinION sequencer was performed by Helena Ayubb (research assistant in the Higgs Lab), following the protocols outlined by Oxford Nanopore Technologies.

The MinION adapter contains a hairpin structure (Figure 10.4). This allows the read to be sequenced initially as it travels through the pore in one direction and then the opposite strand is sequenced in the reverse direction after the hairpin structure has travelled through the pore (Jain et al., 2015).

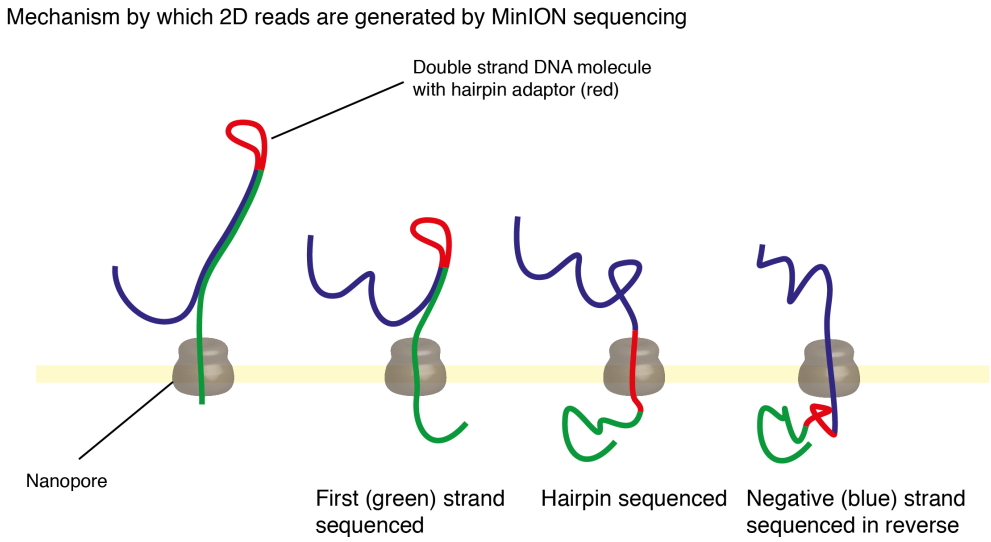


Figure 10.4 Sequencing of 2D reads by MinION

An adaptor with a hairpin is ligated to the double stranded DNA molecule being sequenced. The first strand is initially sequenced. The hairpin adaptor then passes through the pore followed by the negative strand, which is sequenced in reverse order relative to the first strand.

10.3.3 Data analysis

The raw MinION data were initially processed by the cloud based analytics platform Metrichor. The data were converted to FASTA format using poretools (https://github.com/arq5x/poretools). Only high quality 2D reads were included in the analysis.

Analysis of the data from this point is problematic because despite using the highest quality 2D reads the sequence is not sufficiently reliable to allow the restriction sites to be identified robustly from the sequence.

The reads were initially filtered for those containing part of the sequence between the PCR primer and the restriction enzyme cut site. A custom script was used to allow greater flexibility and less stringent filtering than is allowed by standard tools such as BLAT. This searched for matches from the first 100bp of the read with all the possible 6mers from the sequence between the PCR primer and the restriction enzyme cut site for both the enhancer and promoter sequences. The reads were then split into 35mer fragments with a 5bp walk

across the read so that each read was split into several hundred fragments. These 35bp sequences were then mapped to the genome using BLAT using non-stringent settings to allow for the poor sequence quality (minScore=30 -minIdentity=70 maxIntron=4000 tileSize=11).

The psl file produced by BLAT was parsed and the data from each 35bp fragment was mapped to DpnII fragments collapsed and a potential interaction profile across the genome could be generated of the number of reads mapping to each restriction fragment.

I then attempted to reconstruct the read from its component restriction fragments. In order to do this the best 6 matches to restriction fragments were determined for each read. This was done by the counting the number of times a 35bp fragment mapped to each restriction fragment. The top six restriction fragments were then sorted by the order in which they appeared in the original read. These data were outputted in a format for the Washington University Epigenome browser, which can display looping glyphs and a heat map of the read position was generated using R.

These data were also analysed using lastall, which is an aligner designed to map long reads with poor sequence quality. Unfortunately this did not yield useful results. This is probably because reads with multiple restriction fragments had to be used, which prevented accurate long read mapping.

10.3.4 Results

The long reads showed that sequences from within the enhancer compartment of the alpha globin locus were often interposed between the R2 enhancer and the alpha globin promoters (Figure 10.5 & Figure 10.6). In addition often more than the reads mapped in multiple places across the block of enhancers as can be seen in Figure 10.6.

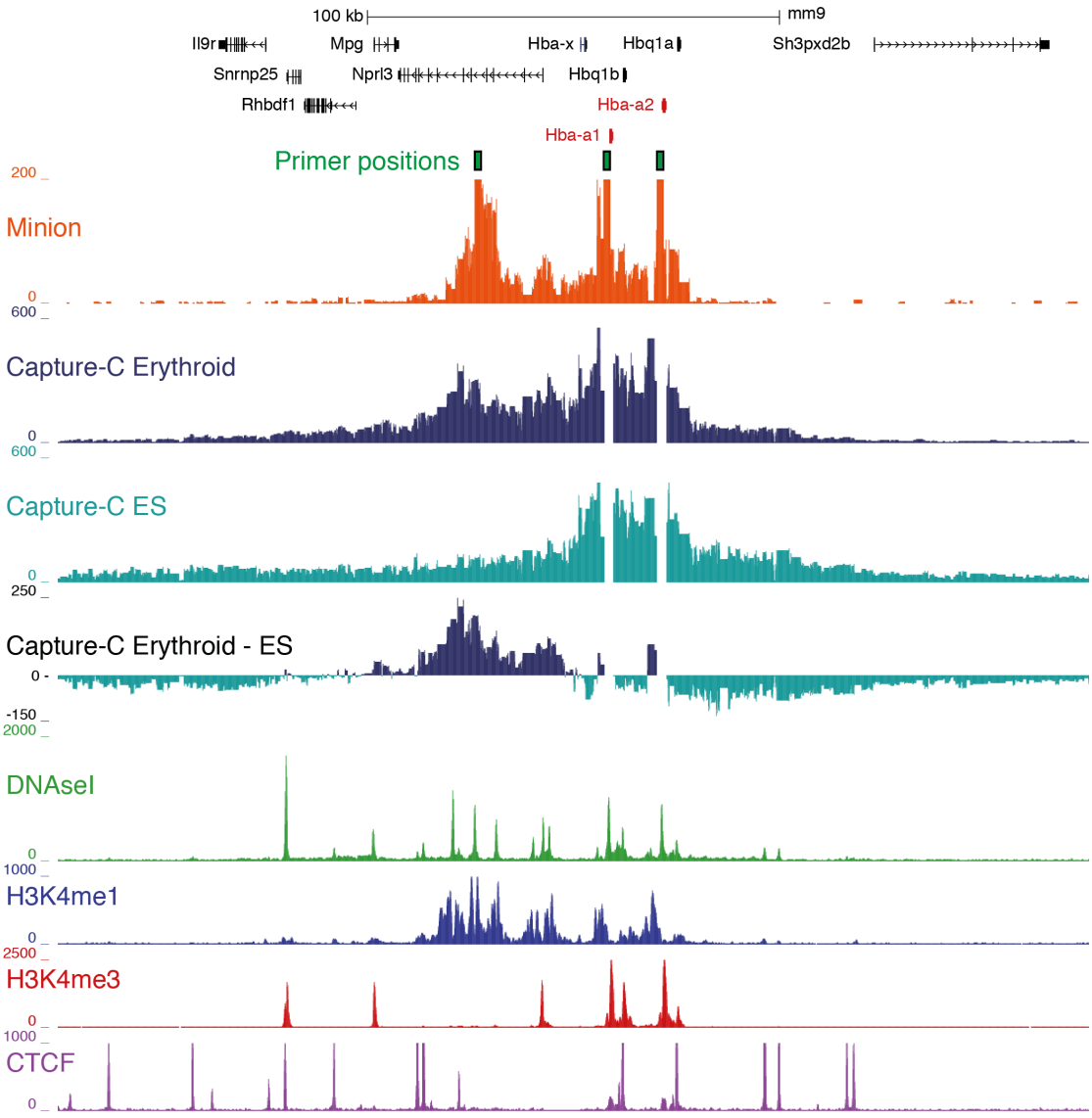


Figure 10.5 UCSC genome browser MinION profile at the alpha globin locus

The top track shows the profile of the collapsed reads across the alpha globin locus. The material was PCR amplified using primers designed to amplify from the R2 enhancer and the Hba-a1&2 promoters (the position of the primers is denoted by green boxes).

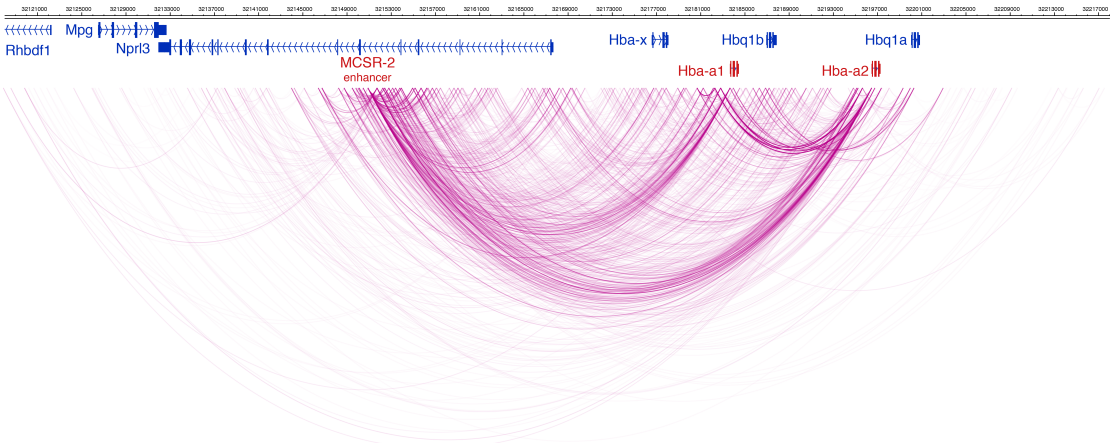


Figure 10.6 Looping Glyph of MinION data (Washington University Epigenome Browser)

The best matching restriction fragments were placed in order along the read and the interactions between adjacent fragments in the read were mapped to the genome. The darker colored of the loops indicate higher frequencies of the interaction.

In an attempt to try and define this further I went on to try and determine how many restriction fragments were contained in each read and where these mapped to across the genome. This was challenging because the poor read quality of minion sequencing compared to conventional next generation sequencing prevented the normal tools for genome alignment being used. However using the BLAT based strategy outlined above I was able to reconstruct reads (Figure 10.7).

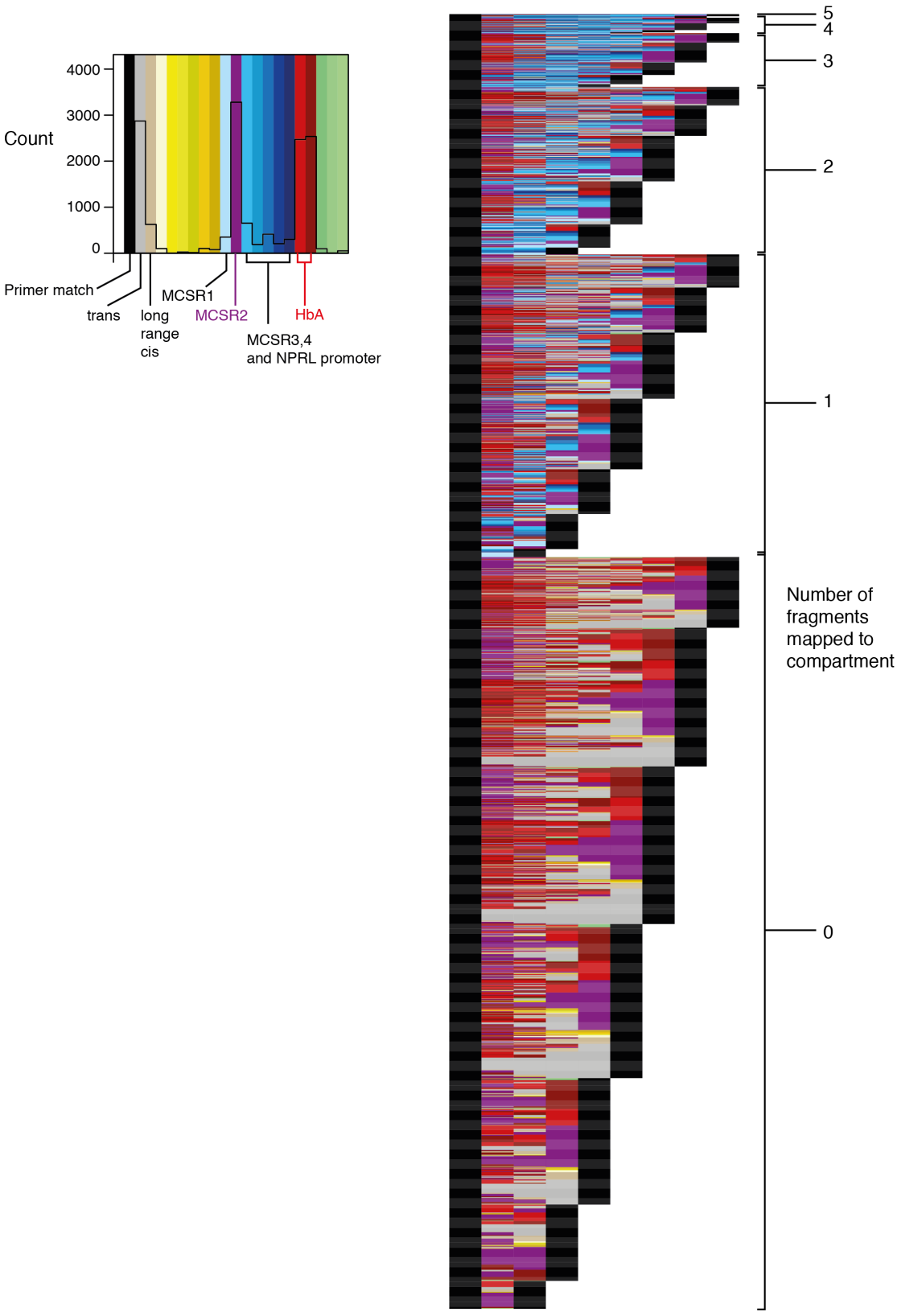


Figure 10.7 Map of reconstructed MinION reads

This shows the reconstruction of the reads by restriction fragment. Each of the coloured bars indicates a mapping to a restriction fragment. Each read was

allowed to map up to 7 restriction fragments and these are arranged so that they represent the order in which they occur in the read. These are colour coded by the position in the genome (as denoted by the inset key). The black bars on the outside of the reads indicate that the read matched the sequence between the primer and the DpnII cut site of either the R2 or Hba-a1&2 fragment using the initial unstringent custom filtering. The inner bands are derived from mapping by BLAT. In the majority of reads the initial mapping is confirmed by mapping by BLAT to the R2 and Hba-a1&2 restriction fragments. This is represented by a purple bar (R2) on one side of the read and a red bar (Hba-a1&2) on the other. The other positions denote the position in the genome, with the blue shades representing restriction fragments within the compartment containing the enhancer elements. The reads are primarily sorted by the number of times the read maps to the compartment and secondarily ordered by the number of restriction fragments in the read.

These data show that around 40% of the reads map at least once to another restriction fragment within the compartment in addition to the alpha globin promoter and R2 enhancer. However, these data need to be interpreted with some caution because the poor sequence quality is likely to skew the results towards reducing the proportion of reads mapping to the alpha globin locus. This is because sequencing errors that lead mismapping of data will almost exclusively cause reads from inside the alpha globin locus to map elsewhere in the genome and it is very unlikely that reads originally from elsewhere in the genome will wrongly map to the alpha globin locus.

10.4 Conclusions

These data show that it is possible to use the new nanopore based sequencing methods to analyse 3C libraries. The data from both experiments show that multiple restriction fragments can ligate to one another within the compartment, supporting the idea that multiple enhancers physically interact to activate transcription. However, because the libraries were PCR amplified prior to analysis it is difficult to be confident about quantifying the frequency of the

interactions. This is because it is impossible to know the degree of PCR duplication and the complexity of the underlying data. Although it should be possible to define whether fragments are unique or not using very long reads (the order and orientation of the fragments would provide a fingerprint), unfortunately the read quality of minion sequencing prevents this at present.

11 Discussion

11.1 Technical discussion

The 3C based methods are capable of linking genes and regulatory elements via their required physical contacts, but to do so robustly they need to be performed at high levels of resolution and signal. Existing methods such as 4C-seq are capable of doing this but can only address single ROIs at a time. At the other end of the spectrum, Hi-C can generate genome wide all vs all interaction profiles, at relatively high resolution, but this requires massive expenditure and expertise. Therefore, existing 3C methods have not become a ubiquitous tool and remain limited to specialist laboratories.

Capture-C was developed to try to address this problem and is capable of generating hundreds of high-resolution interaction profiles in a single assay. However, aspects of the original assay also limited its general utility, in particular the inflexibility of the SureSelect enrichment platform, which meant the original approach, while suitable for hundreds of genes, was limited in its scalability and throughput of samples.

Next Generation Capture-C was developed to generate a completely scalable assay. Changes in library design and methodology resulted in huge increases sensitivity over the original assay, such that it is capable of producing genome-wide interaction profiles of unprecedented sensitivity. The use of either individual or pools of capture probes provides reagents for many hundreds of reactions even at the lowest levels of synthesis. Importantly designs can be expanded as required simply by mixing further probes with existing pools, with designs of up to 20 loci robustly interrogated on "bench top" NGS sequencing platforms such as the Illumina MiSeq (15-20M reads).

11.1.1 Comparison of 3C based methods

11.1.1.1 Resolution

The resolution of 3C based experiments is largely dictated by two factors. First the choice of digestion enzyme (6-cutter v 4-cutter restriction enzyme v MNase v DNAse) dictates the spacing of ligation junctions and therefore the theoretical

maximum resolution (Figure 11.1). Secondly, most techniques have needed to use moving windows / binning of data points to improve the signal strength.

Capture-C uses 4-cutter restriction enzymes, which cut every 256bp on average, whereas the majority of other techniques (3C, 5C (Sanyal et al., 2012) and Hi-C) apart from 4C have nearly exclusively used 6-cutter restriction enzymes, which cut at ~4kb intervals. The use of 4 cutter restriction enzymes, provides a 16-fold greater theoretical maximum resolution. However, because 4-cutter restriction enzymes create libraries with 256 fold greater complexity (the complexity increases with the square of the cut site density) they are significantly more difficult to analyse. Further increases in resolution can be achieved by changing the enzymes used. Micro-C, the MNase based method, has a higher potential resolution than the restriction enzyme based techniques and this has the advantage of cutting at regular intervals in a non sequence dependent fashion, but to date this has only been performed in yeast (Hsieh et al., 2015). DNAse has also been used in Hi-C like experiments (Ma et al., 2015). This is interesting because it has the potential to generate libraries with single base pair resolution since it is not dependent on recognizing a specific sequence and ligation junctions could theoretically form between any sequences across the genome. Sonication also has this property in that it could potentially generate libraries in which ligation junctions can occur between any base pair within the genome. However, for DNAse Hi-C and sonication (ChIAPET) the actual resolution achieved by these methods is lower than has been generated by using 4 cutter restriction enzymes.

3C has only been performed with 6-cutter generated libraries because the increase in cut site density reduces the concentration of any individual ligation junction (by ~ 250 fold) to the point that it becomes difficult to detect strong interactions by RT-PCR. Since 5C requires oligonucleotides to be designed to both the viewpoint and reporter, it requires significant resources even when 6-cutter restriction enzymes are used (5826 oligos to were required cover 1% of the genome (Sanyal et al., 2012)). Thus to date, it has not been attempted with 4-cutter enzymes. Although Hi-C can be performed using 4-cutter restriction it requires an increase in the need for sequencing in line with the increase in library complexity. For large mammalian genomes massive sequencing needs to

be used to get sufficient sequence depth for it to be worthwhile to use a 4-cutter enzyme. Recent Hi-C experiments (Rao et al., 2014) have produced data using 1kb bins but this required a gargantuan effort; with the analysis of several billion read pairs. Clearly this is way beyond the capabilities of most laboratories and even with billions of read pairs the number of informative interactions from any individual viewpoint are around 100-fold lower than in next generation Capture-C.

However, the major determinant of resolution in most 3C-seq based experiments is the bin size used to view the data. Windowing or binning improves the signal strength and reduces biases, allowing a meaningful interaction profile to be determined, when raw signal is weak. Hi-C experiments were initially reported using 40kb bins but this has now been improved down to 1kb bin sizes. Interestingly DNAse Hi-C used 5kb bins so although the library has the potential to generate very high-resolution profiles in practice it does not increase resolution at present. However, it is likely that using enzymes that are not dependent on a specific recognition site will be key to improving the resolution of these assays in the future.

4C-seq has made extensive use of 3C libraries generated by 4-cutter restriction enzymes but the data needs windowing to be applied to generate a robust signal. This was also true of the original Capture-C data set, which also required a 2kb window to be applied. Although it improves signal strength, windowing has a number of disadvantages. The profile becomes skewed by the restriction enzyme cut site density (spurious peaks can appear where there are high numbers of restriction cut sites). It also smoothens the profile of the original signal, obscuring the quality of the underlying data and reducing resolution. Finally it modifies the data making it difficult to understand what scale is being used and it makes it difficult to use statistical methods for comparing data sets.

The high signal depth and lack of biases provided by NG-Capture-C allowed the raw interaction count per fragment to be used. Furthermore, the depth of signal provide by NG-Capture-C has allowed us to further double the potential resolution by enabling us to report interactions based on the half of the fragment that the read maps to (Figure 2.6).

11.1.1.2 Sensitivity

3C has a relatively limited sensitivity. Even with 6-cutter generated 3C libraries the positive and negative controls are often only separated by a ΔCT of around 2 for strong enhancers, such as those with the alpha globin LCR, and the positive interactions often require in the order of 32 PCR cycles to be detected. 3C is therefore insensitive to many of the long range interactions defined by Capture-C, 4C and Hi-C.

ChIAPET generates data of very low read depth from any particular viewpoint because chromatin immunoprecipitiation is very inefficient. In ChIAPET three read pairs are often used to define significant interactions (it is thought that this corresponds to a false discovery rate of <0.05) (Fullwood et al., 2009). In contrast, the strong functional interactions in our data sets are defined by tens of thousands of unique read pairs.

It is difficult to define the sensitivity of 5C because the levels of PCR duplication cannot be determined. However, analysis of the 5C data of the alpha globin locus from the Dekker group shows that this technique also has limited sensitivity for weak long range interactions. Long range interactions that we can detect reproducibly with Capture-C at both the alpha and beta globin loci fail to be detected by 5C.

Hi-C datasets have great sensitivity for detecting long range interactions because it is possible to combine data from many restriction fragments to define robust interactions on a chromosomal scale. However, the number of interactions determined from any individual restriction fragment are around 100-fold lower than in 4C-seq or NG Capture-C even in the recent Hi-C experiments that produced data at 1kb resolution using billions of read pairs (Rao et al., 2014). Therefore Hi-C is a relatively insensitive method for defining the fine details of the interactions occurring within TADs.

4C has excellent sensitivity when it is done carefully, allowing very long-range interactions to be detected. In order to generate high-quality 4C profiles, it is necessary to meticulously perform multiple PCRs to generate material potentially from a million or more cells (Simonis et al., 2007). However, it is impossible to differentiate between PCR duplicates and unique ligation junctions in conventional 4C experiments, so it is very difficult to determine the sensitivity

of the assay accurately and this is likely to vary considerably between different experiments.

Recently an variant of 4C has been developed that uses unique molecular identifiers (UMIs) (Schwartzman et al., 2016), to allow PCR duplicates to be collapsed. This is achieved by introducing a sonication step before a nested ligation-mediated amplification. These data show that the PCR reactions are fairly inefficient, in that only around 5-10,000 unique reads are amplified from each target for each microgram of 3C library used. In addition these authors compared their data to 4C data and found 4C to have similar levels of complexity so this result probably applies to all forms of 4C, in which the levels of PCR duplication cannot be determined.

NG Capture-C also has excellent sensitivity from a single point of reference and it has the advantage, like Hi-C, of being able to determine how many unique read pairs are contributing to the signal to accurately assess the quality of the data. In the best Capture-C experiments over 200,000 unique reads were identified from the viewpoints, which is considerably higher than UMI-4C suggests is possible with 4C. In addition Capture-C is much more efficient when using small numbers of cells because the reagents for addition of sequencing adaptors have been optimised for low DNA input, primarily for ChIP-seq and other NGS technologies that rely on generating NGS libraries from very little material. In the best experiments 30% of the cells from small cell number preparations provided informative ligation junctions (by contrast the figure was around 3% for 4C). The high sensitivity combined with the comparatively few artifacts of the technique allow weak long-range cis and even *trans* interactions to be quantified.

11.1.1.3 Biases

Although all of the methods have biases some of the methods are much more prone than others. ChIA-PET in particular is vulnerable to bias. The initial ChIP can introduce bias towards detecting interactions between sequences that bind the protein targeted by the antibody used for the immunoprecipitation (Simonis et al., 2007). Thus studies that show that sequences bound by the protein being

immunoprecipitated interact with each other need to be interpreted with caution.

4C also needs to be carefully normalized for potential biases resulting from the use of inverse PCR to amplify the interacting fragment from the desired viewpoint fragment (Stadhouders et al., 2013). The raw interaction profile is dependent on the GC content and size of the interacting fragment; some large GC rich fragments simply fail to amplify.

By comparison, Hi-C has relatively little biases, but it is still affected by systematic biases relating to the distance between restriction sites, GC content and sequence uniqueness and several methods have been developed to attempt to correct these (Hu et al., 2012; Imakaev et al., 2012; Yaffe and Tanay, 2011).

Although Capture-C also uses significant numbers of cycles of PCR this generates relatively little PCR bias compared to 4C because the material is sonicated and sequencing adaptors are ligated prior to any PCR amplification. This means that the size of the fragment being amplified is independent of the size of the restriction fragment, removing this significant bias. In addition PCR duplicates can be identified based on the unique ends generated by random sonication, which removes bias resulting from differential PCR amplification. In addition the PCR uses the P5 and P7 sequences in the sequencing adaptors, which is a highly optimized PCR used for all Illumina based next generation sequencing (the same reaction is used for the bridge amplification on the flow cell).

C-nanostring was developed to allow 3C libraries to be assayed without any PCR steps or the biases associated with this. It should therefore generate an interaction profile with very little bias compared to the other techniques. One of the potential biases associated with C-nanostring is that the hybridization efficiencies might differ between one reporter oligonucleotide and another.

11.1.1.4 Multiplexed detection of interactions

3C has the lowest ability to detect interactions simultaneously in that only a single viewpoint to interacting partner can be detected with each PCR reaction. Thus only a handful of interactions can be determined with 3C and it is therefore largely limited to confirming interactions with known regulatory elements.

C-nanostring is able to determine the interactions from a single viewpoint to several hundred different interacting fragments and is therefore only suitable for investigating a single locus at a time.

5C is able to massively multiplex the detection of ligation junctions by using a combination of forward or reverse primers that contain a T7 or T3c sequence respectively at one end and the sequence adjacent to the ligation junction at the other. The primers are hybridised to the 3C library and then amplified using the T7 and T3c sequences. Generally only a forward or reverse primer can be designed to each fragment and it is only possible to detect interactions between forward and reverse combinations (forward – forward and reverse – reverse interactions cannot be detected). In addition about 1/3 of the ligation junctions have to be excluded because the sequence not amenable for specific primer design. Therefore with an alternating forward / reverse design only about 1/9 of the possible interactions can be detected. In addition this requires considerable expenditure. For example, 5826 oligos were require to cover 1% of the genome for the pilot ENCODE project and only two replicates of three different cell lines were reported (Sanyal et al., 2012).

4C-seq has the ability to generate high resolution, 4-cutter profiles of interaction from a single viewpoint to the whole genome. Only a single viewpoint can be interrogated in each assay with conventional 4C and the PCR requires some degree of optimization for each viewpoint used. Since different primers need to be used for each viewpoint it is difficult to do internal controls, in contrast to Capture-C, where it is possible to investigate the effects of alterations at one locus and use several other loci and or cell types as controls. The recently developed UMI-4C allows 20-50 viewpoints to be assayed simultaneously (Schwartzman et al., 2016).

Capture-C is able to generate interaction profiles to the whole genome from up to 450 viewpoints in a single assay (Hughes et al., 2014). The major limitation with Capture-C is that it is difficult to normalize regions that can be captured by two different oligonucleotides and this makes a tiling approach challenging to analyse. Having said that tiling approaches have been attempted without using normalization between the capture efficiencies of the different oligonucleotides

used but this approach is unlikely to provide very reliable data. Hi-C is able to determine genome wide profiles from all viewpoints in the genome.

11.1.1.5 Multiplexing Samples

The investigation of gene regulation is not only limited by the number of genes or elements that can be interrogated, but also by the number of replicates, conditions, cell types and genetic variants that these loci can be assayed in. Until the development of NG Capture-C, all of the 3C based methods could only analyse a single sample at a time. NG Capture-C allows for the simultaneous capture of multiple samples in a single reaction, greatly increasing the throughput and economy of the method as well as the reproducibility and comparability of the data derived from replicates and different experimental conditions. This is the first time that it has been possible to determine the interaction profiles for multiple samples simultaneously. It has allowed us to determine the interaction profiles in two cell types simultaneously and to use the profile from one cell type as a control for the other. C-nanostring is able to analyse 12 samples simultaneously allowing multiple technical and biological replicates to be analysed rapidly.

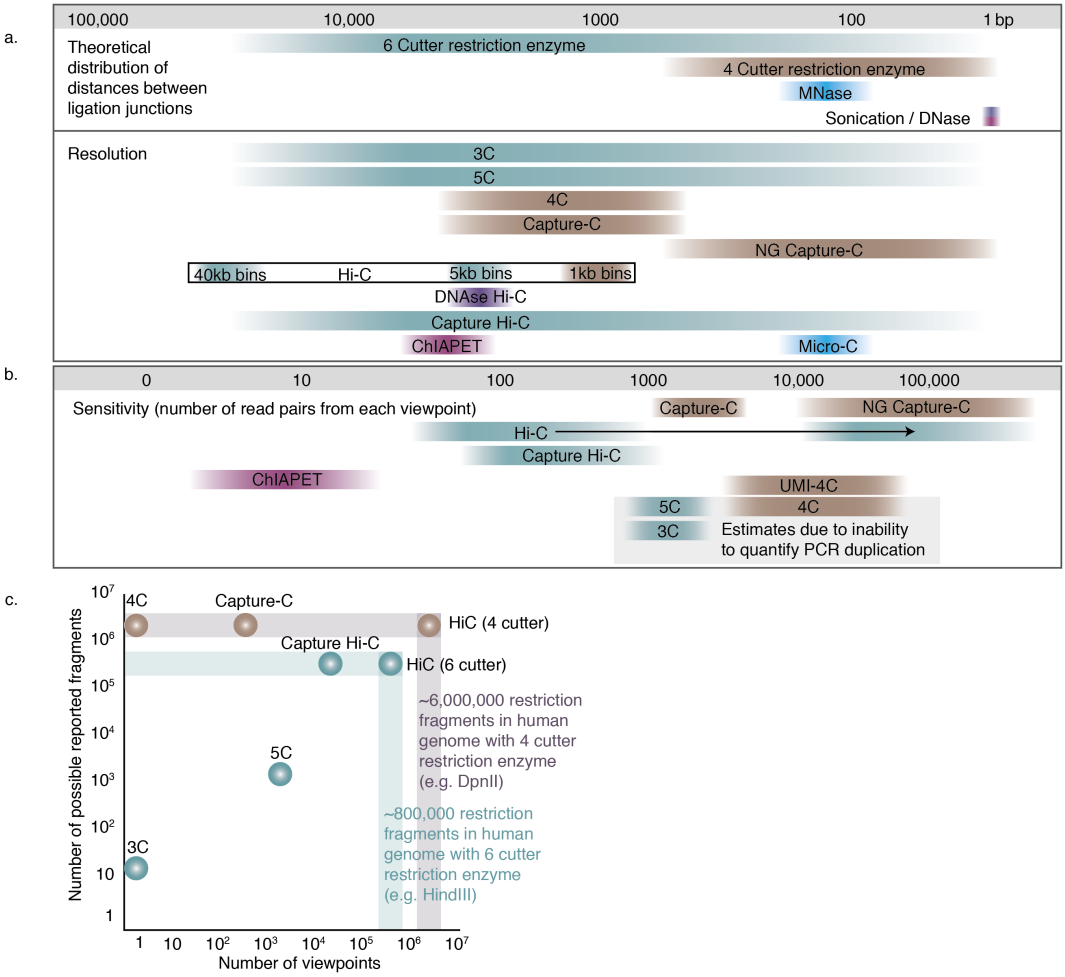


Figure 11.1 Comparison of sensitivity and resolution between techniques

The maximum resolution is determined by the frequency of cutting of the chromatin, which is usually dictated by the choice of restriction enzyme (6 bp recognition vs 4 bp recognition), MNAse (Micro-C), sonication (ChIAPET) or DNAse (DNAse Hi-C) or DNAse. For restriction enzymes and MNAse, which have relatively fixed cut sites the resolution is determined by the fragment sizes. When the methods can cut the DNA at any site in the genome as is possible with DNAse or sonication then it is theoretically possible to generate data at single base pair resolution. When the data is reported for each individual fragment then the cut site density of the restriction enzyme dictates the resolution. However, the majority of 3C based techniques need to use windowing / binning of the data to generate meaningful interaction profiles because this improve the sensitivity and reduces biases at the expense of a reduction in resolution.

b. Comparison of the sensitivity of the different techniques. The sensitivity is largely determined by the number of reads from each individual viewpoint

although techniques such as Hi-C, which undertake an all vs all approach can improve the sensitivity markedly by combining data from multiple view points.

c. Comparison of the multiplexing ability of the different techniques. 4C, Capture-C and Hi-C are all capable of reporting genome wide interactions, however the number of viewpoints differs radically between the methods. Importantly NG Capture-C can analyse multiple samples simultaneously.

11.1.2 Limitations of Capture-C

The main limitation in our approach to Capture-C is that we exclude all capture regions from all of the tracks generated from a single capture reaction, which precludes us from using a tiling approach (Kolovos et al., 2014). We think it is important to correct the data for the variability in capture efficiency between different oligonucleotides and this is challenging because the reported fragments are all unique and potentially partially captured by two different oligonucleotides.

To date we have only tested the double capture protocol on a relatively small targets. It is therefore unclear at what point the double capture step becomes obsolete, if at all. The enrichment afforded by a double capture step will fall with increasing numbers of oligonucleotides. However, with larger numbers of oligonucleotide targets the efficiency of the capture will become increasingly important because it is likely that the available sequencing depth will become limiting rather than the complexity in the library.

One of the other potential problems with increasing the number of oligonucleotides is that the probability that an oligonucleotide capable of massive off target capture increases. Particularly since the oligonucleotides will be capable of forming secondary structures in solution and therefore capturing unintended sequences.

The resolution of capture-C is also limited by the cut site density of restriction enzymes and a possible approach to improving this would be to combine the technique with libraries generated by micrococcal nuclease.

11.1.3 Limitations of C-nanostring

The main limitation of C-nanostring is that it is only able to determine the interaction profiles from a single viewpoint to up to a few hundred other viewpoints across the genome. As we have shown this is sufficient when the locus is well described but it would make the analysis of a poorly understood locus somewhat difficult.

11.2 General Discussion

11.2.1 Insights gained from Next Generation Capture-C

NG Capture-C has given us the most detailed interaction profiles of many genes to date but what insights has this provided? First, although the interaction profiles appear often show specific increases in interaction with the exact location of the regulatory elements (or promoter if using the enhancer as the viewpoint) these are superimposed on quite broad domains of interaction that extend over many kilobases. This is not clearly visible with 4C-seq because it does not generate as complete a profile as NG Capture-C; the gaps due to inefficient PCR appear to make the interactions appear discrete (Figure 2.5). The PCR biases inherent to 4C-seq seem to show that there are much more distinct regions of interaction although careful inspection of the profile shows that these do not exactly coincide with the regulatory elements and there appear to be a lack of interactions with larger fragments. This is an important observation because it implies that the interaction profile is not simply due to the protein-protein interactions between the transcription factors binding at the enhancers and the preinitiation complex at the promoter. The fact that the domain of interaction with the enhancers is significantly broader than the transcription factor binding sites suggests that when an enhancer interacts with the promoter the surrounding chromatin also interacts.

One of the other observations made with NG Capture-C is that most active promoters and enhancers interact with other active elements along the chromosome. Interestingly these are also much broader domains of interaction than one would expect from simple transcription factor binding. Again this

suggests that the interaction profile is not simply due to interactions between different transcription factors as is suggested by the active chromatin hub model. It seems more likely that the interactions are partially mediated by transcription factor binding but other factors also play an important role. Our data would be in keeping with a model in which other modifications of the chromatin (such as such as histone modifications) also contribute to the interaction profile.

11.2.2 NG Capture-C suggests that there is cooperation between enhancer elements

The analysis of the interaction profiles from the R1&2 knockout mice showed that when the distal enhancer elements are removed the interactions with the more proximal elements are also significantly reduced. This suggests that there is a degree of cooperation between adjacent enhancer elements. One would expect that if one enhancer element interacted strongly with the promoter that this would also result in significant concentration of any other adjacent regulatory elements with respect to the promoter.

This might in part explain why enhancers often cluster together to form locus control regions (or superenhancers). Although this implies that colocalised enhancers function is not independent of each other it does not necessarily mean that enhancers function as a group. Individual knockout of the different enhancer elements at the alpha globin locus has shown that there is considerable heterogeneity between the different enhancers. Several of the enhancers appear not to be required for the transcription of alpha globin. A possible explanation of this phenomenon is that nascent enhancers would be more likely to flourish and evolve if they are located adjacent to other strong enhancers because they would benefit from being drawn in towards the promoter. In addition enhancer elements that colocalise with strong enhancers are likely to inherit similar histone modifications from the adjacent enhancers. In addition it is thought that enhancer elements evolve rapidly and thus senescent enhancers would also be likely to be found adjacent to newer enhancers that have taken over the majority of the function.

11.2.3 Interaction profiles from different viewpoints within the alpha globin locus revels different compartments of interaction

The interaction profiles from the different viewpoints within the alpha globin locus highlight different compartments around the genes. First the Hba-a1&2 genes and the enhancers seem to occupy a central compartment. Interestingly, the interaction profile from the CTCF sites on the outside of the locus shows very little interaction with the neighboring promoters and enhancers despite the very close proximity of these regions and there is then clear cut interaction with the other side of the locus. This is clear cut evidence that there are different compartments of interaction and that different regulatory elements can have very different interaction profiles despite being in close proximity along the linear chromatin fiber.

The interaction profile of the SNP separated tracks possibly suggests that further compartmentalization occurs, with the Hba-a1 gene forming a sub-compartment with the proximal enhancers and Nprl3 promoter. Again the Hba-a2 gene seems to bypass interaction with these proximal enhancers in favor of interacting with the more distal enhancers, which, incidentally, seem to have a much greater influence on transcription of the gene.

This creates an alternative model to the active chromatin hub in which a locus is made up of several compartments such that elements within the same compartment preferentially interact with each other and elements in different compartments have a tendency not to interact to the same extent. It is seems unlikely that these compartments are fixed given the number of long range cis and trans interactions (~40% of interactions are outside the compartment). The compartment probably represents an area of increased probability of interaction rather than an enclosed physical space. The data from the minion suggest that it is not uncommon for fragments from other chromosomes to invade into the compartment containing the alpha globin promoters and enhancers and that it was relatively uncommon for the fragment to be made up purely of fragments from solely within the locus.

It seems likely that active compartments might temporarily merge, hence the peaks of interaction with the neighboring genes Cpbe4, Chac2 and Npm1. Alternatively several different conformations may be preferentially stable but at

much lower probability than the strong interaction between the alpha globin promoter and the enhancers. For example, it is possible that the alpha enhancers might occasionally stably interact with the promoter of Cpbe4 forming an alternative conformation to the normal compartment.

11.2.4 Locus control regions (or super-enhancers) may evolve due to enhancer driven compartments of interaction with gene promoters

The data from knockout mice models shows that deletion of a subset of the enhancers within a LCR or superenhancer reduces the interactions between the other enhancers and the promoter. Thus strong enhancers have the ability to facilitate the interaction between other enhancers in their vicinity and promoters. This generates a compartment of interaction between the chromatin surrounding the enhancer and the promoter. This provides a potential evolutionary explanation for the spatial clustering of enhancers because the domain of interaction results in a favorable environment for the evolution of other enhancers. Nascent enhancers are more likely to evolve in these regions and consequently senescent enhancers will also co-localize in these clusters. Thus, the observation that enhancers cluster may reflect the local evolutionary consequences of a strong enhancer producing a chromatin compartment that extends to its cognate promoter.

11.2.5 Distal enhancer positioning may occur to allow separate evolution of gene regulation from gene sequence

Tethering two structures to each other results in massive concentration of each with respect to each other. In a very simple model in which the two moieties are connected by a simple string this changes with the inverse cube of the distance between the two. In vivo measurements suggest that this effect is even more pronounced and the relationship is probably more closely related to the inverse fourth power of the distance between the two (Lucas et al., 2014). Thus two adjacent sequences of DNA will interact with each other very strongly but this effect reduces dramatically with increasing distance (every doubling of the distance between the two results in an at least 8 fold reduction in the concentration of the two moieties with respect to each other).

If one looks at the physical distances involved, it is remarkable that the enhancer and promoter should be separated by such large amounts of intervening DNA. The physical distance that two sequences are separated by is highly dependent on the conformation of the DNA. In B form DNA each base has a length of about 0.34 nm so an enhancer that is 100 kb away from the promoter will be constrained to within a 34 μm radius, which is several times larger than most haemopoietic cells. Of course this is does not occur in eukaryotic cells because DNA is largely wrapped around histones, which compacts the DNA around 6 fold. With this degree of compaction a 100kb segment would constrain the enhancer within a radius of around 5μm, which is still significantly larger than the nuclei of many haemopoietic cells. It is only when chromatin is compacted into to levels seen in the 30nm fiber and above (by another 10 fold or more) that the diffusion of an enhancer 100kb away from the promoter is constrained to a volume smaller than the nucleus. However, there is considerable debate whether the 30nm fiber occurs under physiological conditions (Song et al., 2014) within the nucleus and even if this were to occur it would be likely it would prevent transcription factor and polymerase binding.

It is likely that these large physical distances significantly hamper the ability of an enhancer to induce gene expression. This has two potential consequences. First, one might expect that genes, which require very high levels of gene expression (such as the globin genes), would not be regulated by very distant enhancers. Secondly, it is likely that a considerable selective advantage needs to be conferred by regulation of genes by distal enhancers compared to elements close to or at the promoter to offset the difficulties resulting from the large physical distances separating the enhancers from the promoter.

The benefits conferred by long-range enhancers are currently unclear. My hypothesis is that a selective advantage is conferred by distancing of enhancers from the promoter because this increases the chances of homologous recombination between the enhancer and promoter during meiosis. Thus it is beneficial if the regulatory sequences are in linkage disequilibrium with promoter because this allows the regulation a gene to evolve independently of the gene itself. The distal regulatory elements of a gene could be swapped with

the other parental allele so that in the subsequent generation the enhancers from one parent would regulate the promoter and coding sequence from the other.

The molecular process behind reciprocal recombination events during meisosis is a highly regulated pathway and the crossing over (CO) sites are clustered in narrow regions (~1-2kb) called hotspots. These hotspots are highly enriched for a 13-mer motif (Myers et al., 2008) that is bound by the zinc finger protein PRDM9 (Baudat et al., 2010). Variable number tandem repeats (VNTRs) and other repetitive are associated with increased rates of homologous recombination (Verstrepen et al., 2005).

For the genes such as alpha and beta globin the locus control regions are located relatively close to the promoter but recombination hotspots are found between the enhancers and promoters so that they are in different blocks of linkage (Smith et al., 1998). At the alpha locus a VNTR is located between the enhancers and the promoter (Jarman and Higgs, 1988), which coincides with the blocks of linkage disequilibrium.

Interestingly, when the regulatory elements are spread over very large distances as is seen at Myc, not only are the clusters of regulatory elements are located in different blocks of linkage disequilibrium, with the promoter, the different blocks of tissue specific enhancers appear to be in different blocks of linkage from each other. This potentially allows different patterns of tissue specific gene regulation to evolve without the need for all of the enhancers to evolve simultaneously on the same allele and therefore in the same lineage of organisms. It is likely that this would significantly increase the rate of evolution of complex organisms because it would allow gene expression in different tissues to evolve more rapidly. Another example is found at the HOXD loci where two different sets of enhancers are found on either side of the gene, which independently control development of the fingers and forearm (Andrey et al., 2013). This configuration potentially allows the forearm and finger length to evolve independently of each other. If these enhancers were located close to one another then these characteristics would be irrevocably linked and would have to evolve in tandem.

The converse is also probably true. Genes that require ubiquitous and highly conserved expression should not benefit from distal regulation. Here it would be a selective advantage to regulate the gene only using the promoter and adjacent

sequences. Generally 'housekeeping' genes are thought to be regulated by sequences at the promoter.

Thus genes that only need to be expressed at low levels, such as the transcription factors, which control important processes that would lead to a selective advantage if they are able to evolve different patterns of transcription rapidly, such as those controlling body pattern (e.g. HOXD and SHH), will use different blocks of enhancers in different cell types and these will tend to be located in different blocks of linkage from each other. Genes that require tissue specific regulation, and but which require high levels of transcription would use distant enhancers but not the extremely large distances seen at loci such as HOXD and SHH and these tend to have specific hot spots of recombination located between the enhancer and the promoter. Finally, it would be disadvantageous for the transcription to be controlled by rapidly evolving enhancers for genes that require robust, ubiquitous transcription that is critical for cellular function. Here there will be a tendency for transcription to be controlled by elements close to the promoter.

12 References

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science *340*, 1234167.

Andrulis, E.D., Neiman, A.M., Zappulla, D.C., and Sternglanz, R. (1998). Perinuclear localization of chromatin facilitates transcriptional silencing. Nature *394*, 592-595.

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.

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. (2001). Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics *78*, 73-82.

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.

Bau, 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 alpha-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology *18*, 107-114.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836-840.

Bernet, A., Sabatier, S., Picketts, D.J., Ouazana, R., Morle, F., Higgs, D.R., and Godet, J. (1995). Targeted inactivation of the major positive regulatory element (HS-40) of the human alpha-globin gene locus. Blood *86*, 1202-1211.

Brackley, C.A., Brown, J.M., Waithe, D., Babbs, C., Davies, J., Hughes, J.R., Buckle, V.J., and Marenduzzo, D. (2016). Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome biology *17*, 59.

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. Nature methods *10*, 1213-1218.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327-339.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C.*, et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics *38*, 626-635.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics *46*, 1311-1320.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.
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.
Cullen, K.E., Kladde, M.P., and Seyfred, M.A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. Science *261*, 203-206.
Danino, Y.M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. Biochimica et biophysica acta.
Davies, J.O., 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. Nature methods.
de Kok, Y.J., Vossenaar, E.R., Cremers, C.W., Dahl, N., Laporte, J., Hu, L.J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R.A., Parnes, L.S., *et al.* (1996). Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. Human molecular genetics *5*, 1229-1235.
de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499-506.
de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *11*, 447-459.
de Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., Welling, M., *et al.* (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227-231.
Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nature methods *3*, 17-21.
Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.
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.
Dostie, J., and Bickmore, W.A. (2012). Chromosome organization in the nucleus - charting new territory across the Hi-Cs. Current opinion in genetics & development *22*, 125-131.
Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research *16*, 1299-1309.
Duttke, S.H., Lacadie, S.A., Ibrahim, M.M., Glass, C.K., Corcoran, D.L., Benner, C., Heinz, S., Kadonaga, J.T., and Ohler, U. (2015). Human promoters are intrinsically directional. Molecular cell *57*, 674-684.

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS genetics *4*, e1000039.
Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H.*, et al.* (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58-64.
Geiss, G.K., Bumgarner, R.E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D.L., Fell, H.P., Ferree, S., George, R.D., Grogan, T.*, et al.* (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. Nature biotechnology *26*, 317-325.
Groschel, S., Sanders, M.A., Hoogenboezem, R., de Wit, E., Bouwman, B.A., Erpelinck, C., van der Velden, V.H., Havermans, M., Avellino, R., van Lom, K.*, et al.* (2014). A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell *157*, 369-381.
Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nature protocols *2*, 1722-1733.
Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F.*, et al.* (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nature genetics *43*, 630-638.
Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S.*, et al.* (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature methods *6*, 283-289.
Higgs, D.R., and Wood, W.G. (2008). Long-range regulation of alpha globin gene expression during erythropoiesis. Current opinion in hematology *15*, 176-183.
Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108-119.
Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J.S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics *28*, 3131-3133.
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. Proceedings of the National Academy of Sciences of the United States of America *102*, 9830-9835.
Hughes, J.R., Lower, K.M., Dunham, I., Taylor, S., De Gobbi, M., Sloane-Stanley, J.A., McGowan, S., Ragoussis, J., Vernimmen, D., Gibbons, R.J.*, et al.* (2013). High-resolution analysis of cis-acting regulatory networks at the alpha-globin locus. Philosophical transactions of the Royal Society of London Series B, Biological sciences *368*, 20120361.
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. Nature genetics *46*, 205-212.
Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature methods *9*, 999-1003.

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. Nature methods *12*, 351-356.
Jarman, A.P., and Higgs, D.R. (1988). A new hypervariable marker for the human alpha-globin gene cluster. American journal of human genetics *43*, 249-256.
Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.
Juven-Gershon, T., and Kadonaga, J.T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. Developmental biology *339*, 225-229.
Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M.*, et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289-294.
Kieffer-Kwon, K.R., Tang, Z., Mathe, E., Qian, J., Sung, M.H., Li, G., Resch, W., Baek, S., Pruett, N., Grontved, L.*, et al.* (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell *155*, 1507-1520.
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.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. American journal of human genetics *76*, 8-32.
Kolovos, P., van de Werken, H.J., Kepper, N., Zuin, J., Brouwer, R.W., Kockx, C.E., Wendt, K.S., van, I.W.F., Grosveld, F., and Knoch, T.A. (2014). Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. Epigenetics & chromatin *7*, 10.
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.
Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human molecular genetics *12*, 1725-1735.
Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M.*, et al.* (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66-72.
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.
Liu, Z., and Garrard, W.T. (2005). Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases. Molecular and cellular biology *25*, 3220-3231.
Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein

which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743-1753.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.
Lucas, J.S., Zhang, Y., Dudko, O.K., and Murre, C. (2014). 3D trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. Cell *158*, 339-352.
Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A.*, et al.* (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nature methods *12*, 71-78.
Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957-2963.
Markenscoff-Papadimitriou, E., Allen, W.E., Colquitt, B.M., Goh, T., Murphy, K.K., Monahan, K., Mosley, C.P., Ahituv, N., and Lomvardas, S. (2014). Enhancer interaction networks as a means for singular olfactory receptor expression. Cell *159*, 543-557.
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.
Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.
Mukherjee, S., Erickson, H., and Bastia, D. (1988). Detection of DNA looping due to simultaneous interaction of a DNA-binding protein with two spatially separated binding sites on DNA. Proceedings of the National Academy of Sciences of the United States of America *85*, 6287-6291.
Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nature genetics *40*, 1124-1129.
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.
Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948-953.
Noordermeer, D., Branco, M.R., Splinter, E., Klous, P., van Ijcken, W., Swagemakers, S., Koutsourakis, M., van der Spek, P., Pombo, A., and de Laat, W. (2008). Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. PLoS genetics *4*, e1000016.
Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R.H., and de Laat, W. (2011a). Variegated gene expression caused by cell-specific long-range DNA interactions. Nature cell biology *13*, 944-951.
Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011b). The dynamic architecture of Hox gene clusters. Science *334*, 222-225.
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.

Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature reviews Genetics *15*, 234-246.
Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W.*, et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nature genetics *36*, 1065-1071.
Parker, S.C., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Program, N.C.S.*, et al.* (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences of the United States of America *110*, 17921-17926.
Pasquali, L., Gaulton, K.J., Rodriguez-Segui, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Moran, I., Gomez-Marin, C., van de Bunt, M.*, et al.* (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nature genetics *46*, 136-143.
Pfeiffer, W., and Zachau, H.G. (1980). Accessibility of expressed and non-expressed genes to a restriction nuclease. Nucleic acids research *8*, 4621-4638.
Pronk, C.J., Rossi, D.J., Mansson, R., Attema, J.L., Norddahl, G.L., Chan, C.K., Sigvardsson, M., Weissman, I.L., and Bryder, D. (2007). Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. Cell stem cell *1*, 428-442.
Rach, E.A., Yuan, H.Y., Majoros, W.H., Tomancak, P., and Ohler, U. (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome biology *10*, R73.
Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M.*, et al.* (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628-1639.
Rao, S.S., 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.
Reddy, K.L., Zullo, J.M., Bertolino, E., and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature *452*, 243-247.
Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A.*, et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods *4*, 651-657.
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. Proceedings of the National Academy of Sciences of the United States of America *105*, 8309-8314.
Ruderfer, D.M., Chambert, K., Moran, J., Talkowski, M., Chen, E.S., Gigek, C., Gusella, J.F., Blackwood, D.H., Corvin, A., Gurling, H.M.*, et al.* (2013). Mosaic copy number variation in schizophrenia. European journal of human genetics : EJHG *21*, 1007-1011.
Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., *et al.* (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome research *25*, 582-597.
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. Nature methods *13*, 685-691.
Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458-472.
Simonis, M., Klous, P., Homminga, I., Galjaard, R.J., Rijkers, E.J., Grosveld, F., Meijerink, J.P., and de Laat, W. (2009). High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. Nature methods *6*, 837-842.
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). Nature genetics *38*, 1348-1354.
Simonis, M., Kooren, J., and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. Nature methods *4*, 895-901.
Smith, R.A., Ho, P.J., Clegg, J.B., Kidd, J.R., and Thein, S.L. (1998). Recombination breakpoints in the human beta-globin gene cluster. Blood *92*, 4415-4421.
Solovei, I., Kreysing, M., Lanctot, C., Kosem, S., Peichl, L., Cremer, T., Guck, J., and Joffe, B. (2009). Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell *137*, 356-368.
Song, F., Chen, P., Sun, D., Wang, M., Dong, L., Liang, D., Xu, R.M., Zhu, P., and Li, G. (2014). Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science *344*, 376-380.
Spilianakis, C.G., and Flavell, R.A. (2004). Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nature immunology *5*, 1017-1027.
Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes & development *20*, 2349-2354.
Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.J., Wendt, K.S., Grosveld, F., van Ijcken, W., *et al.* (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nature protocols *8*, 509-524.
Stamatoyannopoulos, J. (2016). Connecting the regulatory genome. Nature genetics *48*, 479-480.
Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science *338*, 671-675.
Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science *346*, 1238-1242.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular cell *10*, 1453-1465.
Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome research *14*, 62-66.
van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E., and de Laat, W. (2012a). 4C technology: protocols and data analysis. Methods in enzymology *513*, 89-112.
van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A*., et al.* (2012b). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nature methods *9*, 969-972.
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. The EMBO journal *26*, 2041-2051.
Vernimmen, D., Marques-Kranc, F., Sharpe, J.A., Sloane-Stanley, J.A., Wood, W.G., Wallace, H.A., Smith, A.J., 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.
Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nature genetics *37*, 986-990.
Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R*., et al.* (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome research *22*, 1680-1688.
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics *10*, 57-63.
Weintraub, H., and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science *193*, 848-856.
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.
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.
Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics *43*, 1059-1065.
Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellaker, C., Goodstadt, L., Nicod, J., Bhomra, A*., et al.* (2011). Sequence-based characterization of structural variation in the mouse genome. Nature *477*, 326-329.
Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics *13*, 134.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556-559.
Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908-921.
Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U.*, et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature genetics *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.