

Spatial organisation of proto-oncogenes in human haematopoietic progenitor cells

A dissertation submitted for the degree of *Doctor of Philosophy*



UNIVERSITY OF
CAMBRIDGE



TRINITY HALL
CAMBRIDGE

Philip Ewels

21st September 2012

“If you want to increase your success rate, double your failure rate.”

Tom Watson

*“Science is a lot like rowing:
although you are relentlessly pushing forwards, it always feels like you’re going
backwards.”*

Adrian Leonard

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Signed:

Phil Ewels

Summary

The eukaryotic cell nucleus is a highly organised organelle, with distinct specialised sub-compartments responsible for specific nuclear functions. Within the context of this functional framework, the genome is organised, allowing contact between specific genomic regions and sub-compartments. Actively transcribing genes in both *cis* and *trans* co-associate at shared transcriptional sub-compartments called transcription factories. Remarkably, genes exhibit a preference to co-associate with certain other genes at the factories. I hypothesised that such preferred juxtaposition at transcription factories may impact the propensity for specific cancer-initiating chromosomal translocations to occur.

Here I have employed a ligation based proximity assay known as enriched 4C, coupled with high throughput sequencing to identify the genomic regions that spatially co-associate with the proto-oncogenes *MLL*, *ABL1* and *BCR* in human CD34⁺ haematopoietic progenitor cells and lymphoblastoid cell line GM12878. I find that the association profiles of these three genes show strong correlation to the binding profile of RNA Polymerase II and other active marks. This suggests that transcribed genes have a propensity to associate with other transcribed regions of the genome, consistent with studies showing that genes often co-associate at transcription factories. However, each gene also exhibits a unique repertoire of preferred associations with certain regions of the genome. Significantly, I find that the most frequent trans association of *BCR* is telomeric chromosome 9, encompassing its recurrent translocation partner gene *ABL1*. I use DNA-Fluorescence in-situ hybridisation to show that the maximal point of association lies near the highly expressed *SURF* cluster of genes, suggesting a mechanism for mediating the interaction.

My data supports a hypothesis that gene transcription has a direct role on genome organisation. I

suggest that preferred co-associations of genes at transcription factories may promote the occurrence of specific chromosomal translocations.

Acknowledgements

Acknowledgement of assistance

It is recognised by the Institute and the University that work leading to a research degree is frequently now carried out in groups and that you are likely to have availed yourself of the help of postdocs, technicians or other students etc. in your work to some extent. The Institute feels it is important for students to be open about the amount of help they have received during the course of their research in preparing their thesis.

1. Initial training in techniques and laboratory practice and subsequent mentoring
 - (a) Allen Chong, Dr. Cameron Osborne - Training in 3C, e4C, general molecular biology techniques
 - (b) Dr. Stefan Schoenfelder - Training in RNA-FISH
 - (c) Dr. Daniel Bolland - Training in RNA-FISH and DNA-FISH
 - (d) Dr. Geoff Morgan - Training in FACS
 - (e) Dr. Cameron Osborne, Christel Kreuger - Mentoring
2. Data obtained from a technical service provider (*e.g.* DNA sequencing, illustrations, simple bioinformatics information etc.)
 - (a) Kristina Tabbada - DNA Sequencing
3. Data produced jointly (*e.g.* where it was necessary or desirable to have two pairs of hands)
 - (a) Mayra Furlan-Magaril - Initial e4C work

- (b) Lauren Ferreira - DNA-FISH probe generation
4. Data / materials provided by someone else (*e.g.* one-off analysis, bioinformatics analysis)
- (a) Dr. George Follows, Dr. Kevin Jestice, Paul Boraks - CD34⁺ samples
 - (b) Dr. Karen Anderson - Test blood samples
 - (c) Maureen Hamon - SF9 cells for carrier e4C
 - (d) Sarah Toscano - SL2⁺ cells for carrier e4C
 - (e) Simon Andrews, Felix Kreuger - Initial bioinformatics processing

Nomenclature

3C Chromatin conformation capture

3C Chromosome Conformation Capture

4C Circularised chromosome conformation capture

ACH Active chromatin hub

ALL Acute Lymphoblastic Leukaemia

AML Acute Myeloid Leukaemia

BrUTP Bromouridine triphosphate

BSA Bovine serum albumin

ChIP Chromatin immuno-precipitation

ChIP Chromatin immunoprecipitation

CHO Chinese hamster ovary cells

CML Chronic myeloid leukaemia

CT Chromosome territory

DMEM Dulbecco's modified Eagle's medium

DMSO Dimethyl Sulfoxide

DNA-FISH DNA fluorescence in-situ hybridisation

DNA-FISH DNA fluorescence in-situ hybridisation

DSB Double strand break

e4C Enriched Chromosome Conformation Capture

EDTA Ethylene-diamine-tetraacetic acid

ES cell Embryonic stem cell

ESI Electron spectroscopic imaging

FBS Foetal bovine serum

FISH Fluorescence in-situ hybridisation

FRAP Fluorescence recovery after photobleaching

GAIx Illumina Genome Analyser IIx

GCSF Granulocyte colony-stimulating factor

GOAT Illumina General Oligo Analysis Tool

GWAS Genome-wide association study

HR Homologous recombination

HRP Horseradish peroxidase

HSC Haematopoietic stem cell

ICD Inter-chromosome domain

ICR Imprinting control region

LAD Lamina associated domain

LCR Locus control region

NAHR Non-allelic homologous recombination

NHEJ Non-homologous end joining

NPC Nuclear pore complex

OLB Illumina Off-Line Basecaller

PBS Phosphate buffered saline

PcG Polycomb protein complex

PCR Polymerase chain reaction

PCR Polymerase chain reaction

PCR Polymerase chain reaction

PE Ad 1.0 Illumina Paired End Adapter 1.0

PRE Polycomb response element

qPCR Quantitative PCR

qPCR Quantitative PCR

qPCR Quantitative real-time PCR

RNA TRAP RNA tagging and recovery of associated proteins

RPMI Roswell Park Memorial Institute medium

RTA Illumina Real Time Analysis

S2R+ Schneider's line 2 Drosophila cell line

SAGE Serial analysis of gene expression

SCS Illumina Sequence Control Software

SDS Sodium dodecyl sulfate

SMC Structural maintenance of chromosome proteins

SPRI Solid-phase reversible immobilization

SSA Single strand annealing

TCC Tethered chromosome capture

TE Buffer Tris-EDTA Buffer

TSA Trichostatin A

TSS Transcription start site

YACs Yeast artificial chromosome

Contents

1	Introduction	2
1.1	Chromatin	2
1.1.1	Histones	3
1.1.2	Histone modifications	3
1.1.3	Histone variants	5
1.2	Two-dimensional organisation	6
1.3	Chromosome Territories	8
1.3.1	CTs and transcription	9
1.3.2	Chromosome territory dynamics	9
1.3.3	Chromatin decondensation and the inter-chromosomal space	10
1.3.4	Chromosome territory intermingling	11
1.4	Nuclear compartmentalisation	11
1.4.1	Transcription factories	12
1.5	Chromatin interactions in three dimensions	16
1.5.1	Promoter and enhancer interactions	16
1.5.2	Chromatin hubs	17
1.5.3	Long range interactions	18
1.5.4	Interactions in <i>trans</i>	19
1.5.5	Global interaction maps	19

CONTENTS

CONTENTS

1.6	What drives nuclear organisation?	20
1.6.1	Transcription	20
1.6.2	CTCF	22
1.6.3	Cohesin	24
1.6.4	Tethering	25
1.6.5	Actin and myosin	27
1.6.6	Replication	27
1.6.7	Polycomb	28
1.6.8	The bigger picture	28
1.7	Chromosomal translocations	28
1.7.1	Formation of chromosomal translocations	29
1.8	Leukaemia	31
1.8.1	Chronic myeloid leukaemia	32
1.8.2	Mixed lineage leukaemia	33
1.9	Effect of nuclear organisation on translocation formation	37
1.9.1	Breakage first and contact first models	37
1.9.2	Chromosome territories and translocations	38
1.9.3	Transcription factories and translocations	39
1.10	Thesis overview	41
2	Materials and Methods	43
2.1	CD34 ⁺ cell handling	43
2.1.1	Peripheral blood collections	43
2.1.2	Leukapheresis collections	44
2.1.3	CD34 ⁺ cell separation	45
2.2	Cell culture	47
2.3	3C	48
2.3.1	Nuclei preparation and digestion	48

2.3.2	Ligation and purification	49
2.3.3	Digestion efficiency analysis	50
2.3.4	Detection of 3C products by qPCR	50
2.4	e4C	53
2.4.1	Primer extension and primary <i>NlaIII</i> digestion	53
2.4.2	Bait enrichment and secondary <i>NlaIII</i> digestion	53
2.4.3	PCR and germline removal	54
2.4.4	Gel extraction and second round PCR	55
2.4.5	e4C library quality control	55
2.5	Carrier e4C	56
2.6	DNA fluorescence <i>in-situ</i> hybridisation	56
2.6.1	Probe generation	56
2.6.2	DNA-FISH	61
2.6.3	Visualising signals using the Metacyte	66
3	Developing an assay for gene association	67
3.1	Introduction	67
3.1.1	Chromosome Conformation Capture	68
3.1.2	Enriched 4C	69
3.2	3C Restriction Enzyme Choice	70
3.3	Primer Design	73
3.3.1	Paired end sequencing or single end sequencing?	73
3.3.2	Bait specific primers	74
3.4	Preparation of CD34 ⁺ cells	75
3.4.1	CD34 ⁺ sorting purities	77
3.5	e4C with low cell numbers	78
3.5.1	ChIP e4C	78
3.5.2	Carrier e4C	79

CONTENTS**CONTENTS**

3.6	e4C with large cell numbers	80
3.7	Multiplexing e4C libraries	81
3.7.1	Multiplexed libraries	82
3.7.2	Cross over products	83
3.8	Increasing e4C library complexity	85
3.8.1	Barcoded <i>NlaIII</i> adapter	85
3.8.2	Multiplexing same-bait e4C libraries	86
3.8.3	Results of e4C modifications	87
4	Developing the analysis of e4C data	89
4.1	Introduction	89
4.2	Initial data handling	89
4.2.1	Bareback processing	90
4.2.2	Quality control	91
4.2.3	Sequence trimming	92
4.2.4	Sequence alignment	93
4.2.5	Importing into SeqMonk	94
4.3	e4C library biases	95
4.3.1	Potential fragment libraries	95
4.3.2	GC content bias and fragment length bias	96
4.4	Significance of single regions	98
4.5	<i>AseI</i> site distribution normalisation	98
4.5.1	<i>In-silico</i> testing	99
4.5.2	Standard scores	100
4.6	100
5	Initial e4C library analysis results	101
5.1	Introduction	101

5.1.1	Overview of e4C libraries	101
5.2	Concerning raw data	102
5.2.1	Duplicate reads	102
5.2.2	<i>AseI</i> fragment saturation	103
5.2.3	Library complexity	103
5.3	<i>cis</i> interaction profiles	104
5.3.1	Interaction frequency in <i>cis</i> declines as a function of linear separation	104
5.3.2	Specific interactions in <i>cis</i>	106
5.4	<i>BCR</i> , <i>ABL1</i> and <i>MLL</i> reside in an active nuclear compartment	107
5.4.1	Correlation with active epigenetic marks	108
5.4.2	Correlations between e4C libraries	110
5.5	Different genes have different preferred interaction partners	111
6	Preferential association of <i>BCR</i> with Chromosome 9	113
6.1	<i>BCR</i> preferentially co-associates with 9q34 in CD34 ⁺ cells	113
6.1.1	Single window testing	114
6.2	<i>ABL</i> e4C associations in GM12878 cells	114
6.3	Validation using publicly available datasets	117
6.3.1	Hi-C associations in GM06990 cells	117
6.3.2	Hi-C and TCC associations in GM12878 cells	118
6.4	Narrowing the window - the Surfeit Cluster	118
6.4.1	Candidate genes	119
6.5	Validation by 3C	119
6.5.1	Design	119
6.5.2	Results	119
6.6	Validation by DNA-FISH	120
6.6.1	DNA-FISH in GM12878 cells	120
6.6.2	DNA-FISH in CD34 ⁺ cells	120

CONTENTS**CONTENTS**

6.7 Discussion	120
7 Discussion	121
7.1 Introduction / Context	121
7.2 Limitations of the e4C data / Discussion	121
7.2.1 Coverage	121
7.2.2 Things I'd have done differently	121
7.3 Sample choice	122
7.4 Association between <i>BCR</i> and <i>ABL1</i>	122
7.4.1 Evidence for a bystander effect	122
7.5 A model for the formation of chromosomal translocations	122
7.5.1 MLL - topoisomerases?	122
7.6 Future perspectives	123
A Primers	124
B e4C Library Statistics	125
B.1 Numbers of reads per Chromosome	125
C e4C analysis scripts	130
C.1 e4C sequence processing and trimming	130
C.2 <i>In-silico</i> restriction fragment libraries	131
C.3 GC content and fragment length bias detection	131
C.4 Systematic bias correction	131
C.5 Restriction fragment distribution normalisation	131
C.5.1 inputminted method	131
C.5.2 L ^A T _E X include method	131
C.5.3 Lyx include child document method	131
C.5.4 Copy and paste method	131

CONTENTS**CONTENTS**

D External datasets used	134
D.1 Datasets used for active mark correlations	134
E Hi-C	136
F Other Perl scripts and R scripts used	137
G Web Tools	138
G.1 Publicly available tools I've written - Cytobands, Genome RE Sites, Sequences. Mention traffic.	138
G.2 Talk about contribution to FastQC	138
H Publications	139
Bibliography	140

List of Figures

1.1.1 Structure of the nucleosome	5
1.3.1 Chromosome Territories	8
1.4.1 Nascent RNA and transcription factories.	12
1.4.2 Klf1 specialised transcription factories	15
1.6.1 Putative structure of CTCF	22
1.7.1 DSB repair	30
1.8.1 Diagram of the <i>MLL</i> breakpoint cluster region	37
1.9.1 Correlation of CT intermingling and radiation induced translocation frequencies	39
3.1.1 Overview of the 3C methodology	68
3.1.2 Overview of the e4C methodology	70
3.2.1 <i>BamHI</i> e4C Libraries	71
3.2.2 Effect of Triton-X100 on <i>BamHI</i> digestion	72
3.2.3 Restriction Enzyme Tests	73
3.4.1 CD34 ⁺ Separation FACS Plots	77
3.5.1 CD34 ⁺ RT-PCR	79
3.5.2 <i>Drosophila melanogaster</i> S2R+ 3C Tests	80
3.7.1 Crossover reads between multiplexed e4C libraries	84
3.8.1 Barcoded <i>NlaIII</i> adapter	86

LIST OF FIGURES

4.2.1 Bareback overview	91
4.2.2 Representative e4C Library FastQC and FastQ Screen	92
4.2.3 Expected structure of reads in each e4C library	93
4.2.4 Library trimming and alignments	94
4.2.5 Bowtie alignment parameters	94
4.3.1 e4C library biases	96
4.5.1 <i>AseI</i> site distribution normalisation	99
4.5.2 <i>AseI</i> normalisation <i>in-silico</i> test.	100
5.2.1 e4C duplicate frequency	103
5.2.2 e4C <i>AseI</i> fragment saturation	104
5.2.3 e4C library read statistics	105
5.3.1 e4C <i>cis</i> association profiles	106
5.3.2 <i>BCR</i> e4C <i>cis</i> association	107
5.4.1 <i>BCR</i> e4C correlation with H3K4me1	109
5.4.2 e4C library correlations	110
5.5.1 <i>BCR</i> e4C normalisation against H3K4me1	112
6.1.1 <i>BCR</i> association with 9q3	114
6.1.2 Frequency of <i>in-silico</i> hits for single window in telomeric Chromosome 9	115
6.2.1 ABL e4C <i>trans</i> association profile	116
6.3.1 GM06990 Hi-C heat map	118

LIST OF FIGURES

List of Tables

1.1.1 Summary of known mammalian histone modifications	4
3.3.1 e4C sequencing primer properties	74
3.3.2 e4C primers	76
4.2.1 Illumina sequence processing statistics with Bareback.	91
4.3.1 <i>in-silico</i> potential AseI - NlaIII fragment library statistics	95
5.1.1 e4C Libraries	102
5.4.1 e4C: active epigenomic mark correlation scores	108
B.1.2 e4C reads by Chromosome	125
B.1.1 e4C library <i>AseI</i> fragment statistics	129
C.1.1 Library trimming statistics	132
C.1.2 e4C library read counts	133
D.1.1 Accession codes	135

Chapter 1

Introduction

The nucleus is a highly complex organelle responsible for the faithful replication and maintenance of the DNA template and regulation its transcriptional products. To achieve this, the contents of the nucleus are organised into compartments specialising in processes such as transcription and replication. Abnormalities in the organisation of the nucleus are often associated with diseases such as cancer. In this chapter I will discuss the current understanding of nuclear structure and organisation and how it may be involved in the initiation of oncogenic chromosomal translocations.

1.1 Chromatin

Chromatin is a general term used to describe DNA packaged around histone proteins to form nucleosomes, and the plethora of additional proteins that bind to them both. There are two principle forms of chromatin: euchromatin and heterochromatin. These were first described in the early twentieth century due to their differential staining within the nucleus (Heitz, 1928). Heterochromatin stains darkly because it remains highly condensed during interphase and often associated with the nuclear periphery. Constitutive heterochromatin consists of repetitive elements found in centromeres and telomeres, it plays a structural role within the genome and is highly compacted during interphase. Facultative heterochromatin is less compact and consists of inactive chromatin that can vary between

cell types as they differentiate. Heterochromatin is well known as being a repressive environment for gene expression. An example of this is the position-effect variegation seen in *Drosophila melanogaster* - a chromosomal inversion in the X chromosome characterised in the early 1930s can place the *white* gene close to pericentric heterochromatin, leading to the spreading of heterochromatin marks which silences the *white* gene and results in a change in eye colour (Vogel et al., 2009). In contrast, euchromatin is the site of most genic transcription (Chesterton et al., 1974), its looser compaction allowing cellular machinery access to the DNA, enabling the binding of transcription factors and the initiation of transcription.

The differences between heterochromatin and euchromatin lie within the proteins that they contain. Chromatin acts as a platform for proteins to bind to. Differences in histone modifications, histone variants, nucleosome packing and DNA modifications affect the accessibility and binding profile of the chromatin, and therefore affect how the DNA is interpreted by the cell.

1.1.1 Histones

To package DNA, the double helix is wrapped around an octomer of core histones: two H2A, two H2B, two H3 and two H4. 146 base pairs of DNA interact with these positively charged proteins to form the nucleosome, which is then bound by histone H1 with linker DNA to make a total of 166 base pairs (Fig 1.1.1, Davey et al., 2002). This packing forms the 10 nm fibre, often known as 'beads on a string' due to its appearance in electron micrographs. At their most basic level, histones function to compact DNA by counteracting the negative charge of the phosphorylated back bone.

1.1.2 Histone modifications

Core histones have flexible amino-terminal tails which extend outside of the nucleosome, and are accessible to proteins within the nucleoplasm. These tails can be post-translationally modified at a large number of residues - lysine (methylation, acetylation, ubiquitination, sumoylation, ADP-Ribosylation), arginine (methylation) as well as serine and threonine (phosphorylation). These modifications can affect the packing of chromatin (Wolffe and Hayes, 1999) as well as which proteins can bind. The

large number of combinatorial possibilities that result from these modifications have been dubbed the 'Histone Code' (Strahl and Allis, 2000). Different modifications are related to different chromatin states. For example, active promoters are typically enriched for di- and tri-methylation of histone 3 lysine 4 (H3K4) whilst inactive promoters are enriched for trimethylation at lysine residues 27 (H3K27me3) and 9 (H3K9me3) (See table 1.1.1 for a summary. For review, see Zhou et al., 2011).

Modification	Histone	Residue	Effects on transcription
Acetylation	H2A	K5	Activation
	H2B	K5, K12, K15, K20	Activation
	H3	K4, K14, K18, K23, K27	Activation
		K9	Histone deposition
	H4	K5, K12	Histone deposition
		K8, K16	Activation
Methylation	H3	K4, K79	Euchromatin
		K9, K27	Silencing
		R17	Activation
		K36	Elongation
	H4	R3	Activation
		K20	Silencing
Phosphorylation	H2A	S1, T119	Mitosis
	H2AX	S139	DNA repair
	H3	T3, S10, T11, S28	Mitosis
	H4	S1	Mitosis
Ubiquitination	H2A	K119	Silencing
	H2B	K120	Activation

Table 1.1.1 – Summary of known mammalian histone modifications. H, histone; K, lysine; R, arginine; S, serine; T, threonine. Adapted from Sadri-Vakili and Cha, 2006.

As chromatin immunoprecipitation (ChIP) has become a common laboratory technique, combined with microarray techniques (ChIP on chip) and next generation sequencing (ChIP-Seq), our understanding of how histone modifications affect chromatin biology on a genome-wide scale has advanced dramatically. Profiling chromatin types using multiple datasets covering a large number of histone modifications is sufficient to predict the identity and function of regions in the genome with a high degree of accuracy, revealing previously unknown enhancers (Heintzman et al., 2007; Ernst and Kellis, 2010; Hon et al., 2009). Ernst *et al.* used the genome-wide profiles of nine histone modifications in nine different cell types to define fifteen chromatin states, including promoters, enhancers, insulators and

transcribed regions (Ernst et al., 2011). They integrated data from genome-wide association studies (GWAS) and found numerous enhancer elements that coincide with disease associated mutations. Such genome-wide approaches can reveal the dynamics across different cell types and are powerful tools in understanding how the genome is interpreted in health and disease (see Section ??).

1.1.3 Histone variants

In addition to histone tail modifications, chromatin can be modified by the incorporation of histone variants. Canonical core histone genes are found in clustered repeat arrays within the genome, are transcribed during replication and are highly conserved between species. Histone variants are found as single genes spread through the genome and are subject to far greater diversity (Talbert and Henikoff, 2010).

CENP-A is a human variant of histone H3 which replaces the canonical histone in centromeric heterochromatin. It is a key factor in the establishment of the centromeres and kinetochores required for mitosis. The histone variant is incorporated with the help of a number of chaperone proteins, including HJURP, after replication of DNA has finished (Dunleavy et al., 2009; Foltz et al., 2009). CENP-A is essential for the formation of centromeres.

Another frequent histone variant found in humans is H3.3, which differs from canonical H3 by just four amino acids (Talbert and Henikoff, 2010). This histone variant is found within transcribed genes, promoters and regulatory elements, and is thought to be laid down during transcriptional elongation (Schwartz and Ahmad, 2005). Nucleosomes

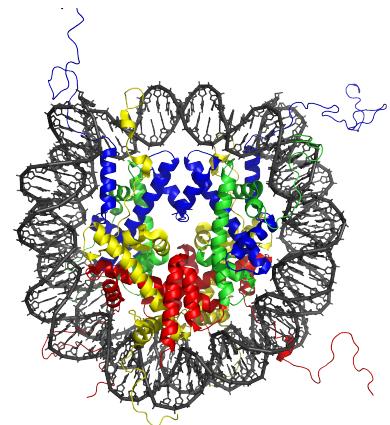


Figure 1.1.1 – Structure of the nucleosome. X-ray structure of a nucleosome core particle at a resolution of 1.9 Å. DNA can be seen wrapped around the core histones, which have flexible tails extending into the nuclear matrix. PDB structure 1KX5 (Davey et al., 2002).

containing H3.3 appear to be less stable than canonical nucleosomes, with a high turnover (Schwartz and Ahmad, 2005). It is possible that this increased turnover of the nucleosomal components helps to keep the chromatin open and accessible to the transcriptional machinery (Talbert and Henikoff, 2010).

Other core histone proteins also have variants, such as H2A.Z, a histone variant located on either side of the nucleosome free regions found at the transcriptional start sites of active genes as well as insulator regions (Zlatanova and Thakar, 2008). H2A.Z is able to promote the recruitment of RNA polymerase II to certain regions, such as the yeast *GAL1-10* genes, by mediating C-terminal interactions with the transcriptional machinery (Adam et al., 2001).

1.2 Two-dimensional organisation

Since the inception of cytogenetics, it has been known that not all chromosomes are the same. The development of chromosome banding techniques in the 1970s allowed detailed human karyotypes to be determined, complete with differential staining of regions within each chromosome (Caspersson et al., 1970). Banding assays can show regions of heterochromatin and euchromatin, highlighting the variation in characteristics across regions of the genome (Trask, 2002). In prokaryotes, genes are often found in cistrons and can be coexpressed in single polycistronic mRNAs. This type of linear organisation is not present in most higher eukaryotes, with the notable

exception of the nematode worm *Caenorhabditis elegans* (Blumenthal et al., 2002), though gene clusters resulting from tandem duplication are frequently found throughout mammalian genomes. Some specific examples of two-dimensional clustering have been shown: testes-specific genes in *Drosophila melanogaster* have been found in clusters more frequently than would be expected by chance (Boutanaev et al., 2002) and genes sharing transcription factors can be found in clusters in the yeast *Saccharomyces cerevisiae* (Janga et al., 2008).

The sequencing of the human genome allowed detailed analysis of GC content, gene density and repetitive sequence content (Lander et al., 2001). Versteeg et al. integrated a multitude of SAGE tag expression profiles from different cell types into the genomic map and built on earlier work defined regions of high transcriptional activity, called ridges (Versteeg et al., 2003; Caron et al., 2001). They found ridges to be gene-dense, highly transcribed, have a high GC content and low LINE repeat density. These features are based purely on the underlying sequence content and so do not vary amongst cell types. The different ridges and anti-ridges were found to contain different classes of genes, with weakly expressed genes clustering within anti-ridges and clusters of highly expressed housekeeping genes found predominantly in ridges (Versteeg et al., 2003).

Whilst the two-dimensional organisation of the genome cannot completely explain the degree of complexity found within the transcriptome, these studies demonstrate that the order of sequence within the genome is not entirely random and can affect the transcriptional control of genes.

1.3 Chromosome Territories

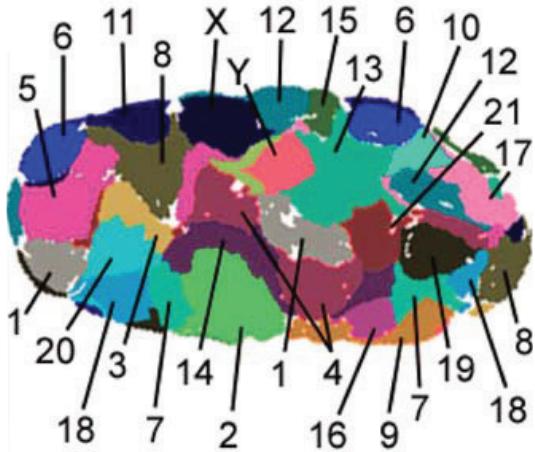


Figure 1.3.1 – Chromosome Territories. Image of a Human fibroblast cell in G0 with all chromosomes labelled using multi-colour FISH. Adapted from Bolzer *et al.* (Bolzer et al., 2005)

The interphase nucleus is a highly structured organelle. As chromosomes decondense after metaphase they retain some degree of structure, forming “chromosome territories” (CTs) (Cremer and Cremer, 2001). Circumstantial evidence for interphase organisation of chromosomes has existed for a long time, first suggested by Carl Rabl in 1885 (Rabl, 1885). Observations by Stack *et al.* using microscopy with giemsa-band staining suggested that chromosomes retained some degree of organisation during interphase (Stack et al., 1977), and in 1982

Cremer *et al.*, showed that interphase chromosomes occupy territories by studying the pattern of DNA damage in metaphase chromosomes after spot irradiation during interphase (Cremer et al., 1982). The subsequent development of chromosome paints, a method to visualise entire or part-chromosomes with fluorescence *in-situ* hybridisation (FISH), confirmed these findings (Schardin et al., 1985; Manuelidis, 1985; Bolzer et al., 2005).

As FISH techniques have developed, so too has the detail with which CT organisation can be studied. Several groups have shown that chromosome territory position within the nucleus is not random and correlates with chromosome size (Sun et al., 2000; Cremer et al., 2001; Bolzer et al., 2005), gene-density (Cremer et al., 2001; Croft et al., 1999) and replication timing (Ferreira et al., 1997; Visser et al., 1998) - those near the centre of the nucleus tend to be gene-rich, early replicating and small. CT positioning appears to be conserved through evolution (Tanabe et al., 2002) and is cell type specific (Parada et al., 2004; Kuroda et al., 2004).

1.3.1 CTs and transcription

It is known that for some but not all genes, positioning at the nuclear periphery correlates with reduced gene expression (Kosak et al., 2002; Dietzel et al., 2004; Zink et al., 2004). To investigate whether nuclear positioning can cause changes in transcriptional activity or is simply a consequence, three groups published studies which artificially tethered genomic regions to the inner nuclear membrane using *lac* operators (*lacO*) (Finlan et al., 2008; Reddy et al., 2008; Kumaran and Spector, 2008). Finlan *et al.* and Reddy *et al.* both observed a decrease in the transcriptional activity of the regions when tethered (Finlan et al., 2008; Reddy et al., 2008), an effect that was ablated when cells were treated with trichostatin A (TSA) to inhibit class I and class II histone deacetylases (Finlan et al., 2008). It should be noted that this effect does not appear to apply to all loci (Kumaran and Spector, 2008). Whilst these studies suggest that it is nuclear positioning that leads to transcriptional effects, a study by Croft *et al.* showed that inhibition of transcription causes a reversible change in CT position (Croft et al., 1999) and Bridger *et al.* showed a difference in CT positioning between proliferating and senescent human fibroblasts (Bridger et al., 2000). Large scale rearrangements of CTs have also been observed during cell differentiation (Stadler et al., 2004; Szczerba et al., 2009), likely due to changes in transcriptional profiles and chromatin remodelling.

This evidence suggests that there is a dynamic interplay between CT positioning and transcription - gross transcriptional patterns may drive the position of chromosomes within the nucleus, and those positions may in turn affect the transcription of the genes they contain.

1.3.2 Chromosome territory dynamics

The movement of chromatin loci within the nucleus is known to be largely due to local Brownian motion, or “constrained diffusion” and seems to be limited by attachment to a nuclear cytoskeleton, nucleoli and the inner nuclear membrane (Marshall et al., 1997; Chubb et al., 2002), though some studies have shown a mixture of local diffusion and larger, active movements (Vazquez et al., 2001). Chuang *et al.* studied Chinese hamster ovary (CHO) cells after stimulation with a transcriptional activator and analysed the movement of a reporter locus from the nuclear periphery to the interior

of the nucleus (Chuang et al., 2006). They found that chromosome movements happened in rapid unidirectional bursts, suggesting an active mechanism. This was supported by a later study in human fibroblast cells by Mehta *et al.* who demonstrated chromosome movement only 15 minutes after serum starvation, a process that was ablated by the inhibition of actin polymerisation or myosin activity (Mehta et al., 2010).

1.3.3 Chromatin decondensation and the inter-chromosomal space

In addition to the study of whole chromosome positioning, there has been a great deal of research into the positioning of individual sequences relative to their chromosome territory. Early studies suggested that transcribed genes were found at the surface of chromosome territories (Zirbel et al., 1993) leading to a model whereby transcriptionally inactive genes are buried within territories and expressed genes are able to contact transcriptional machinery in an inter-chromosome domain (ICD) (Cremer et al., 1993). This model gained support due to studies showing genes at the periphery of CTs (Kurz et al., 1996) and new techniques to visualise the ICD using microscopy (Bridger et al., 1998). A number of FISH studies showed genes moving away from their territories in large loops upon activation (Volpi et al., 2000; Mahy et al., 2002a; Chambeyron and Bickmore, 2004), suggesting that they may be recruited to the ICD for transcription.

Despite these case studies, looping is not a prerequisite for transcription; DNA-FISH studies have shown transcription within the volume of chromosome territories (Verschure et al., 1999; Mahy et al., 2002b). Osborne *et al.* showed that the actively transcribed *Uros* gene is more frequently outside the CT than the inactive gene *Fgfr2*, though this position was not necessary for transcription, suggesting that actively transcribed genes may preferentially locate to the surface of CTs, but that this alone is not sufficient to drive transcription (Osborne et al., 2004). As such, it has been suggested that the inter-chromosomal domain model should be renamed the inter-*chromatin* domain model, whereby chromosomes are invaginated with channels and subdivided into ~1 Mbp domains of chromatin (Cremer and Cremer, 2001).

1.3.4 Chromosome territory intermingling

A question that followed immediately from the discovery of loops extending from chromosome territories was that of chromosome intermingling. Mathematical modelling approaches capable of predicting intermingling volumes correlated with known rates irradiation induced DNA damage (Holley et al., 2002; Hlatky et al., 2002). This was later backed up with a study by Branco and Pombo, who studied thin cryosections of nuclei with high resolution light and electron microscopy; they found that there is significant intermingling of chromatin between chromosome territories, and that the extent of this intermingling correlated strongly with previously recorded irradiation induced translocation frequencies (Branco and Pombo, 2006). The degree of intermingling changed significantly for three chromosome pairs after transcription inhibition with α -amanitin, suggesting a role for specific transcription interactions in the organisation of the nucleus (Branco and Pombo, 2006).

In support of chromosome intermingling, the *HoxB* extra-chromosomal loops found to extend from the chromosome territory upon gene activation (Chambeyron and Bickmore, 2004) were found to make increased *trans* chromosomal interactions whilst looping out (Würtele and Chartrand, 2006), suggesting that the loops contact other chromosomes rather than occupying an empty inter-chromosomal space. A large number of inter-chromosomal contacts have been detected by recent genome-wide chromosome conformation studies, supporting the presence of chromosome intermingling (Lieberman-Aiden et al., 2009).

1.4 Nuclear compartmentalisation

Because the nuclear interior is devoid of membrane bound structures its organisation is defined by a dynamic equilibrium - a sum product of the many processes and requirements involved in nuclear biology. In addition to the non-random positioning of chromosomes, a number of proteins are also found in aggregates typically referred to as subnuclear compartments. The clustering of proteins into compartments increases the efficiency of biochemical processes and is predicted by the principle of molecular crowding (reviewed in Cook, 2002) and is a key feature in the organisation of the nucleus.

1.4.1 Transcription factories

One class of nuclear subcompartment which has come to light within the past thirty years is the transcription factory, seen as foci of hyper-phosphorylated RNA polymerase II scattered throughout the nucleus. The majority of genic transcription appears to take place at transcription factories (Jackson et al., 1993; Osborne et al., 2004; Eskiw et al., 2008), challenging the classical model of transcription found in many text books (discussed below).

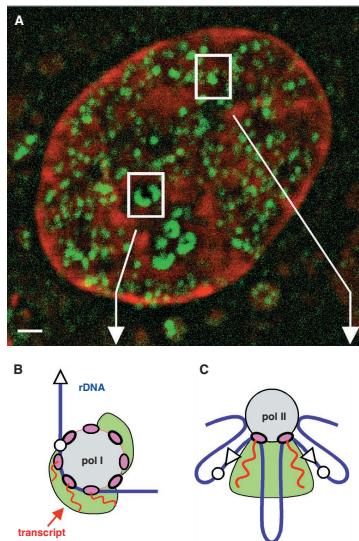


Figure 1.4.1 – Nascent RNA and transcription factories. (A) Transcription foci in HeLa cells, visualised with labelled Br-UTP in 100 nm cryosections. Nascent RNA (green) is concentrated in punctate foci. (B) Model for a nucleolar factory, showing a transcript with multiple polymerases generating a crescent shaped focus. (C) Model for a nucleoplasmic factory. Multiple transcribed regions each with a single polymerase generate a smaller cloud of nascent transcripts. Adapted from Cook et al., Science (2009) (Cook, 1999).

The first study to suggest that eukaryotic transcription does not take place with a processive RNA polymerase moving along a DNA template was by Jackson, McCreedy and Cook in 1981. They showed that nascent RNA transcripts labelled with [³H] uridine remained within the nucleus when loops of DNA were removed using a nuclease (Jackson et al., 1981). They went on to show that RNA polymerase II and active genes were also resistant to elution after chromatin digestion (Jackson and Cook, 1985). The term 'transcription factories' was coined by Jackson et al. in 1993. Fluorescence microscopy was used to label the incorporation of bromouridine triphosphate (BrUTP) into nascent RNA; discrete foci of nascent transcription could then be seen within the nucleus which did not form in the presence of the RNA polymerase II inhibitor α -amanitin (Jackson et al., 1993). Further studies showed that these foci contained RNA polymerase II along with many other components required for transcription (Iborra et al., 1996; Grande et al., 1997). An ultrastructural study by Eskiw et al. used correlative microscopy with both electron spectroscopic imaging (ESI) and fluorescence microscopy to study nuclei sections. ESI can distinguish nitrogen and phosphorous atoms without labelling, and fluorescence light microscopy can visualise transcription through labelling

BrUTP in nascent transcripts. The authors found that nascent RNA is almost always associated with the surface of large nitrogen-rich protein structures with a diameter of ~87 nm, comparable in size to that predicted for a transcription factory (Eskiw et al., 2008).

The discovery of transcription factories has demanded a new model for the action of RNA polymerase II (Cook, 1999). The revised model proposes that instead of RNA polymerase II freely diffusing to active genes and tracking along the gene body, genes are recruited to transcription factories and are pulled through a stationary polymerase. Such a model provides a better explanation for the mechanics of transcription - clustering of transcriptional activity may enable the cell to conduct transcription in a much more efficient manner; HeLa nuclei have a 1 μ M concentration of active RNA polymerase II, whereas the local concentration within transcription factories is closer to 1 mM (Cook, 2002). Additionally, a polymerase enzyme moving along a gene would rotate with the helix of the DNA wrapping the nascent transcript around the template. Genes pulled through static transcription factories would extrude their RNA transcripts into the nucleoplasm (Iborra et al., 1996; Cook, 1999), creating topological loops within the template DNA which may be removed through the activity of topoisomerases.

Jackson *et al.* went on to work on a quantitative analysis of transcription factories in HeLa cells calculating the number of active RNA polymerases, the number of transcription sites and the number of polymerases associated with each transcriptional unit (Jackson et al., 1998). They showed that each HeLa cell nucleus contain approximately 2400 transcription factories, each with approximately 30 active RNA polymerase II complexes (Jackson et al., 1998). Importantly, this study showed that there are more transcribing units than there are foci of transcription, suggesting that genes must colocalise to transcribe. The number of transcription factories varies a great deal amongst cell types, but the observation that genes colocalise within transcription factories has been confirmed by a number of different techniques. Osborne *et al.* showed that genes situated within transcription factories are actively transcribed, whereas those outside are not (Osborne et al., 2004). They went on to demonstrate that transcription is a discontinuous process with the frequency of nascent RNA transcription foci related to primary transcript RNA concentrations, suggesting that transcription occurs in bursts. Multiple

genes both in *cis* and in *trans* were seen to dynamically colocalise in transcription factories, supporting predictions that genes must share transcription factories (Osborne et al., 2004; Jackson et al., 1998).

A later paper by Osborne *et al.* showed that the immediate-early genes *Myc* and *Fos* are dynamically recruited to existing transcription factories within five minutes of B-cell stimulation, suggesting that the recruitment of genes to pre-existing transcription factories may be a method of transcriptional control (Osborne et al., 2007).

It is worth noting that the concept of genes being recruited to immobile transcription factories is not universally accepted, with some doubt over the resolution achievable by FISH and 3C studies (reviewed in Sutherland and Bickmore, 2009).

1.4.1.1 Specialised transcription factories

After the demonstration that genes share transcription factories, a number of groups postulated that specific genes may colocalise at a subset of transcription factories. The self-organising principle applied to transcription factories would predict this; just as the local concentration of active RNA polymerase II is elevated by the clustering of transcribing units, the local concentration of transcription factors that bind to those transcribed units will also be elevated (Cook, 2002). It is thought that genes diffusing to a transcription factory already engaged with other genes that share the same factors are more likely to engage and be transcribed themselves (Bartlett et al., 2006). Indeed, genes regulated by common transcription factors appear to cluster within specific chromosomes in yeast (Janga et al., 2008) and testis-specific genes are found clustered in *Drosophila* (Boutanaev et al., 2002) - genes clustered in linear sequence are more likely to associate in three dimensions.

Some evidence that such specialised transcription factories may exist within mammalian cells came from Osborne *et al.* (Osborne et al., 2007). *Myc* and *Igh* are commonly translocated in Burkitt's lymphoma and mouse plasmacytoma. They showed that approximately one quarter of actively transcribing *Myc* alleles shared a transcription factory with *Igh*, over double the rate of colocalisation found with the control genes tested. Using DNA-FISH they showed an overall reduction in spacing between *Myc* and *Igh* alleles upon B cell stimulation, suggesting that *Myc* alleles are being

specifically recruited to transcription factories containing transcribing *Igh* (Osborne et al., 2007). Xu and Cook later demonstrated that the transcription of plasmids transfected into cells clustered together at a handful of transcription factories. The plasmids were generated with one of four promoter types, one of three genes and one of three 3' regions and were found to segregate according to their promoter and the presence of an intron (Xu and Cook, 2008).

Perhaps the best evidence for the existence of specialised transcription factories came in a publication by Schoenfelder *et al.* in 2010 (Schoenfelder et al., 2010). They used a variety of techniques to investigate the nuclear localisation of erythroid genes. A genome-wide screen of genes associating with *Hba* and *Hbb* at transcription factories showed enrichment for genes with CACC motifs capable of binding the erythroid-specific transcription factor Klf1. Using immunofluorescence they found that nuclear Klf1 foci overlap with active RNA polymerase II foci, suggesting that a subset of transcription factories are enriched for Klf1 in erythroid tissues (Fig 1.4.2). Co-localising Klf1 dependent genes associated with these Klf1 specific transcription factories at a high frequency, and a number of these gene associations were lost in *Klf^{-/-}* knockout mice. These data suggest that a network of Klf1 specific transcription factories exist within mouse erythroid tissues, and that Klf1 specific genes are preferentially recruited to these sites (Schoenfelder et al., 2010). If such specialised transcription factories are a general feature in mammalian nuclei, they could be a key driving force in the organisation of the genome.

It is important to note that the evidence above does not describe *exclusively* specific transcription

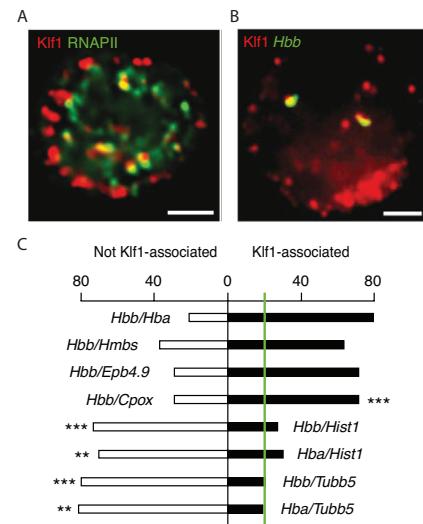


Figure 1.4.2 – Klf1 specialised transcription factories. (A) Immunofluorescence showing Klf1 (red) overlapping with a subset of RNA Polymerase II transcription factories (green). (B) RNA immuno-FISH showing transcribing *Hbb* alleles (green) co-localising with Klf1 (red). (C) Percentages of co-localising transcript pairs colocalisation with Klf1 associated transcription factories (right, black bars) and non-Klf1 associated factories (left, white bars). Expected level of association due to chance shown by green bar. Scale bars show in (A) and (B) represents 2 μm. Taken from Schoenfelder *et al.* (Schoenfelder et al., 2010).

factories, i.e. factories which are incapable of transcribing genes not controlled by Klf1. Indeed, although statistically significant, the gene associations detected suggest a population bias. Specialised transcription factories are certainly not required for the transcription of the eukaryotic genome, but may aid efficiency and affect nuclear organisation and function.

1.5 Chromatin interactions in three dimensions

1.5.1 Promoter and enhancer interactions

The classical model of gene promoter regulation began with the characterisation of the *lac* operon in *Escherichia coli* in 1961 (Jacob and Monod, 1961). Three genes required for the digestion of lactose are controlled by the binding of a repressor protein which blocks transcription by RNA polymerase. This repressor can form a tetramer and bind two region simultaneously, requiring a topological loop in the chromatin (Savageau, 2011).

A number of eukaryotic genes were later found to require distal enhancer elements (Banerji et al., 1981), prompting speculation that the looping of chromatin allowing direct interaction between sequences may be a common mechanism (Dorsett, 1999). Direct evidence for *in-vivo* chromatin interactions was provided in 2002 with two papers studying the mouse beta globin (*Hbb*) locus (Carter et al., 2002; Tolhuis et al., 2002). This locus contains four beta-like genes arranged in order of their expression through development. Approximately 50 kilobases upstream is a locus control region (LCR), containing multiple DNase hypersensitive sites (HS1-6) (Bender et al., 2000). Carter *et al.* used a novel technique called RNA TRAP (tagging and recovery of associated proteins) which localises horseradish peroxidase (HRP) to the site of nascent RNA production of a specific gene. The HRP catalyses biotinylation of nearby chromatin which can be purified and quantified by quantitative real-time PCR (qPCR). They showed a 15-fold enrichment of biotin over the HS2 region of the LCR when the actively transcribing *Hbb-b1* transcript was used for the HRP localisation, demonstrating that this region of the LCR is in very close proximity to the *Hbb-b1* gene (Carter et al., 2002).

Tolhuis *et al.* (Tolhuis et al., 2002) used the chromosome conformation capture technique (3C)

developed by Dekker *et al.* to study the organisation of chromatin within yeast (Dekker *et al.*, 2002) (for discussion of this technique see Section 3.1.1). They studied the same globin locus and also found evidence for interaction between the LCR hypersensitive regions and the *Hbb* genes. Importantly they showed that in mouse foetal brain tissue, where the *Hbb* locus is not expressed, the chromatin adopted a linear conformation without any looping (Tolhuis *et al.*, 2002). Palstra *et al.* went on to show that each gene contacts the LCR as it is expressed, supporting a system where the developmentally controlled genes must contact an enhancer within the LCR in order to be expressed (Palstra *et al.*, 2003). They showed that erythroid progenitor cells dedicated to the lineage but not yet expressing beta-like globin genes form a 'poised' structure, contacting the LCR but not engaging strongly with the HS2 enhancer (Palstra *et al.*, 2003).

1.5.2 Chromatin hubs

The beta globin locus has become a model system for the active chromatin hub (ACH) model (de Laat and Grosveld, 2003), describing a system where loops of chromatin containing elements capable of controlling the expression of genes are held in close three-dimensional space to the genes that they control. Other gene clusters have also been shown to behave in a similar manner, notably the T_H2 locus and *Hox* clusters.

The T_H2 locus control region is involved in the transcriptional control of cytokine genes IL4, 5 and 13 (Lee *et al.*, 2003). Spilianakis *et al.* showed that these genes cluster together in T cells, NK cells, B cells and fibroblasts, despite not being expressed in the last two cell types. In T cells and NK cells the genes additionally associate with the T_H2 LCR (Spilianakis and Flavell, 2004). The authors suggest that this mechanism allows the coordinate expression of the gene cluster in a controlled manner.

The *Hox* genes are master regulators of gene transcription and are responsible for the creation of vertebrate segments during development. *Hox* gene clusters *A* to *D* are transcribed in sequence as development of the embryo progresses and have been observed decondensing and looping out from their chromosome territories upon activation (Chambeyron and Bickmore, 2004). Noordermeer *et al.* studied the *Hox* clusters in three different mouse embryonic day 10.5 tissues: forebrain, anterior

trunk and posterior trunk (Noordermeer et al., 2011). They found that the *Hoxd* cluster formed a discrete domain in forebrain, where it is inactive. In anterior and posterior trunk the *Hoxd* cluster is active, but different genes are transcribed. In both tissues they found the cluster to form two distinct compartments correlating with an inactive and active regions. Using circularised chromosome conformation capture (4C), a technique based on 3C, they showed that genes move from the inactive to active compartment as they are activated, correlating with active histone marks (Noordermeer et al., 2011).

1.5.3 Long range interactions

As the role for distal enhancers has become more established, evidence has been uncovered for increasingly distant interactions. For example, Sharpe *et al.* developed a mouse model for preaxial polydactyly by random insertion of a reporter cassette (Sharpe et al., 1999) which they found to affect a *cis* regulatory site over a megabase upstream of the gene *Shh*, known to be important in the condition (Lettice et al., 2002). Lettice *et al.* went on to characterise this enhancer, which lies within an intron of the gene *Lmbr1* unrelated to the condition, demonstrating that 7q36 abnormalities found in patients with preaxial polydactyly disrupt this enhancer (Lettice et al., 2003). In 2005, Velagaleti *et al.* characterised breakpoints found in two patients with the skeletal malformation syndrome campomelic dysplasia (Velagaleti et al., 2005). They found the breakpoints corresponded to two different regulatory elements, one 1.1 Mb upstream of the target gene *SOX9* and one 1.3 Mb downstream. Kleinjan *et al.* used mouse models carrying yeast artificial chromosomes (YACs) to characterise multiple distal enhancers of the developmental control gene *Pax6* (Kleinjan et al., 2006). They found that as different enhancers were removed, expression of the gene was abolished in different tissues, suggesting a complex system of enhancer - promoter interactions driving the pattern of tissue-specific expression (Kleinjan et al., 2006).

1.5.4 Interactions in *trans*

Spilianakis *et al.* went on from characterising the intra-chromosomal interactions of the T_H2 LCR (described above) to show that the same locus forms inter-chromosomal interactions (Spilianakis *et al.*, 2005). Depending on the stimulus received, naïve T cells can differentiate into either TH1 or TH2 cells, defined by the expression of either IFN- γ or IL-4. Spilianakis *et al.* showed that the *Ifng* gene on chromosome 10 can interact with the T_H2 LCR on chromosome 11 to stimulate *Ifng* expression whilst inhibiting *IL4* expression. This interaction is the first interchromosomal interaction known to regulate gene expression (Spilianakis *et al.*, 2005).

Lomvardas *et al.* used 3C to demonstrate the association of an olfactory receptor gene enhancer made specific contacts to multiple other olfactory genes across the genome (Lomvardas *et al.*, 2006). Mouse dendrites can express one of approximately 1300 odorant receptor genes and Lomvardas *et al.* suggested that this enhancer-gene interaction was the mechanism responsible for the expression of that gene. However, it should be noted that deletion of this enhancer had little effect on the usage of olfactory genes outside of its cluster (Fuss *et al.*, 2007).

A number of other studies have shown specific interchromosomal contacts involved in a number of processes ranging from X-inactivation to genomic imprinting, showing that these interactions may play an important role in chromatin biology (reviewed in Schneider and Grosschedl, 2007).

1.5.5 Global interaction maps

Our understanding of the three-dimensional organisation of the genome has advanced in leaps and bounds during the last decade largely because of the development of the 3C method and its derivatives (for review, see (Osborne *et al.*, 2011)). There are a large number of 3C variants, but they can be grossly categorised into four classes based on how many loci can be interrogated in a single experiment: one-to-one (3C, quantitative 3C), one-to-all (4C, e4C, ACT), many-to-many (5C) and all-to-all (ChIA-PET, Hi-C, TCC).

The recent development of all-to-all methods has allowed the conformation of the entire genome to be probed in a single experiment. This approach has many advantages; such an unbiased approach

allows the detection of unexpected interactions and associations can be probed in parallel allowing a far higher rate of data collection. Whilst these techniques have had great impact on the field, they are currently limited by the depth of sequencing that is achievable with today's technology. To address this Sanyal *et al.* recently published a paper as part of the ENCODE project describing the interaction profiles of 628 transcription start sites (TSS) and 4535 surrounding fragments, representing approximately 1% of the genome (Sanyal et al., 2012). To achieve the resolution required for the robust detection of promoter-element interactions, Sanyal *et al.* used 5C, a many-to-many technique that uses a panel of oligonucleotides with common adapters to anneal to 3C products and create a library capable of being sequenced. They sequenced libraries from three ENCODE cell lines: K562, HeLa-S3 and GM12878. Only a small proportion of the looping interactions uncovered were shared between the three cell types, with approximately 60% of interactions being unique to a single cell line. The majority of TSS looping interactions could be classified as interacting with enhancer elements, promoters or regions bound by the structural protein CTCF. Looping interactions with enhancer elements were significantly enriched for actively expressed TSS, demonstrating the importance of three-dimensional chromatin contacts in the regulation of gene expression.

1.6 What drives nuclear organisation?

As our understanding of the structure of the nucleus evolves, an increasing number of structural features and patterns are being uncovered. Teasing apart correlation and causation to find the driving forces behind nuclear organisation is not an easy task and remains a major challenge within the field.

1.6.1 Transcription

The discovery of transcription factories has changed our view of nuclear organisation substantially. If genetic templates are mobile and transcription factories are fixed, then it maybe be possible to use the process of transcription as a tool to fold the genome into specific conformations.

Kimura *et al.* quantified the amount of stable RNA polymerase II in HeLa cells (Kimura et al., 1999) adding to the work by Jackson *et al.* demonstrating the stability of nascent transcripts and

Polymerase in the nucleus (Jackson et al., 1981; Jackson and Cook, 1985). Mitchell and Fraser later demonstrated that RNA polymerase II transcription factories remain in the absence of transcription, though gene association with factories is ablated if transcription initiation is inhibited (Mitchell and Fraser, 2008). These data, along with the observation that genes are recruited to pre-existing transcription factories upon activation (Osborne et al., 2004), support a model whereby transcription factories are attached to a relatively immobile nuclear substructure. This means that RNA polymerase II can act as a motor, dragging template chromatin through the nucleus as it is transcribed, powered by the removal of phosphate groups during RNA synthesis (Cook, 1999). Yin *et al.* measured the force produced by a single *E. coli* RNA polymerase using an immobilised enzyme transcribing a template bound to a polystyrene bead held by optical tweezers (Yin et al., 1995). RNA polymerase stalled when the force applied was greater than 14 piconewtons (pN), substantially more than kinesin or myosin, making RNA polymerase the most powerful biological motor known. Papantonis *et al.* demonstrated the potential of RNA polymerase to pull chromatin transcripts through the nucleus *in vivo* by using ³C to measure the change in association between regions of DNA after activation of the TNF α gene (Papantonis et al., 2010). They found that as the gene was transcribed, downstream regions of chromatin progressively came into contact with other transcribing regions at the transcription factory.

The role for transcription in the organisation of the genome is further supported by the pervasive nature of transcription. As much as 93% of genomic bases in the human genome are thought to be transcribed in at least one cell type (described in Clark et al., 2011), including many enhancers (Ling et al., 2005; Kim et al., 2010). In the recent ENCODE study of TSS interactions, Sanyal *et al.* found that enhancer elements looping to a TSS were significantly more likely to express enhancer RNAs (Sanyal et al., 2012), supporting a model that the transcription of a chromatin template could be responsible for the formation of chromatin loops.

An attractive model for larger scale genome organisation revolves around the transcription of housekeeping and tissue-specific genes. Lercher *et al.* have shown that genes with high expression in multiple tissue types have a propensity to be present in clusters within the genome (Lercher et al., 2002) and the mouse alpha globin locus has been shown to assemble at a transcription factory already

transcribing a cluster of housekeeping genes in erythroid cells (Zhou et al., 2006). Gavrilov *et al.* showed that these housekeeping genes were bound stably to the nuclear matrix and resistant to high salt extraction, whereas the alpha globin genes were less stable and could be eluted (Gavrilov et al., 2010). These studies may point to a model of the nucleus where persistently transcribed housekeeping genes are responsible for forming transcription factories, with tissue specific genes being recruited to these sites upon activation (Gavrilov et al., 2010). Such a system could have far reaching consequences, with chromosome conformation being determined by housekeeping gene hubs.

1.6.2 CTCF

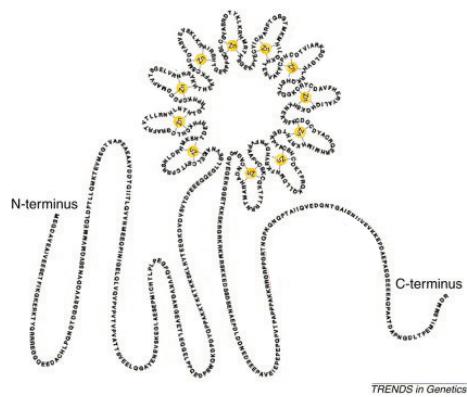


Figure 1.6.1 – Putative structure of CTCF.
Adapted from Ohlsson *et al.* 2001 (Ohlsson et al., 2001).

CTCF (CCCTC-binding factor) is a highly conserved DNA-binding protein with eleven zinc finger domains which binds the consensus sequence CCCTC as well as a range of variant sequences using combinations of different zinc fingers (Filippova et al., 1996) (Fig 1.6.1, reviewed in Ohlsson et al., 2001). The core sequence has extremely high sequence conservation between mouse, chicken and human and mice homozygous for the gene knockout exhibit early embryonic lethality (Splinter et al., 2006). The protein is expressed ubiquitously and misregulation by over-expression or RNAi knockdown have a wide range of effects (Torrano et al., 2005).

When CTCF was first isolated it was thought to be a transcriptional repressor for the *c-Myc* gene (Lobanenkov et al., 1990). Since then its proposed functions have included transcriptional activator (Vostrov and Quitschke, 1997), insulator and structural protein. Bell, West and Felsenfeld were the first to characterise the role of CTCF as an insulator in a study of the chicken beta-globin locus (Bell et al., 1999). They showed that CTCF binds the HS4 region of the β -globin LCR, a region previously shown to act as an insulator in a transgene enhancer-blocking assay (Chung et al., 1997).

One of the best characterised examples of CTCF acting as an enhancer is at the imprinted *H19/Igf2* locus. Four CTCF binding sites were found in the imprinting control region (ICR) which exhibit methylation-sensitive binding (Bell and Felsenfeld, 2000; Hark et al., 2000; Szabó et al., 2000). On the unmethylated maternal allele, CTCF binds the ICR and the downstream enhancer element stimulates expression of the *H19* gene. On the methylated paternal allele, CTCF binding is abrogated and the enhancer contacts the distal *Igf2* gene instead, causing expression of *Igf2* and not *H19*. This model is supported by 3C data showing parent-of-origin specific interactions between the enhancer and the two genes (Murrell et al., 2004; Kurukuti et al., 2006).

Genome wide studies of CTCF binding using ChIP-on-chip and ChIP-Seq have greatly developed our understanding of this complex protein. The model of CTCF acting as a global transcriptional activator or repressor has been largely abandoned; binding by CTCF is not able to predict transcriptional activity or correlate with specific classes of genes (Chen et al., 2008). In support of CTCF in the role of a global insulator, Xie *et al.* determined a large number of conserved CTCF binding sites using computational analysis and derived a dataset containing divergent pairs of genes with or without a CTCF separating them (Xie et al., 2007). Divergent gene pairs facing away from each other frequently show correlated expression patterns (Trinklein et al., 2004; Li et al., 2006) - as expected, gene pairs not separated by a CTCF site showed a propensity to be coexpressed, yet those with a CTCF site had a close to background level chance of coexpression (Xie et al., 2007). Not all CTCF binding sites appear to act as enhancer-blocking insulators however, as demonstrated by looping over CTCF sites in the human beta-globin locus (Tolhuis et al., 2002). The recent 5C study by Sanyal *et al.* found that 79% of long range interactions skip one or more CTCF binding sites, prompting the authors to suggest that additional factors may be needed for the insulator function of CTCF (Sanyal et al., 2012).

The development of the ChIA-PET technique has allowed an in depth interrogation of all CTCF bound chromatin interactions in mouse embryonic stem (ES) cells (Handoko et al., 2011). Handoko *et al.* demonstrated with this technique that CTCF acts in multiple roles with different sizes of loops. Correlative analysis using data about surrounding histone modifications allowed the authors to determine five distinct classes of CTCF-derived loops: active domains enriched for H3K4me1,

H3K4me2 and H3K36me3; repressive domains enriched for H3K9, H3K20 and H3K27 methylation; putative enhancer-promoter interactions; barrier insulator behaviour, separating active and inactive chromatin compartments; and finally, loops with no known correlation or function (Handoko et al., 2011). This study illustrates the complexity of CTCF binding within the genome and points to multiple roles for the protein involving chromatin organisation. This interaction has subsequently been shown to be dependent on protein-protein contacts between the SA2 subunit of cohesin and the C-terminal domain of CTCF (Xiao et al., 2011).

1.6.3 Cohesin

The structural maintenance of chromosome (SMC) proteins, also known as cohesin, form complexes required for sister chromatid cohesion and chromosome segregation in mitosis. Originally characterised in *Drosophila melanogaster* due to mutants having a defect in sister chromatid separation (Michaelis et al., 1997), cohesin is a multi subunit complex consisting of two large coiled-coil domain proteins (Smc1, Smc3) connected by a smaller dimer of Scc1 and Scc3. Cohesin complexes are thought to function in mitosis by holding sister chromatids together through topological looping (reviewed in Hudson et al., 2009).

A number of studies suggested that cohesin may play a role in the interphase nucleus as well as in mitosis, being implicated in gene regulation, recombination, repair and domain formation (reviewed in Hagstrom and Meyer, 2003). In 2008 three studies showed that the genome-wide binding profiles of cohesin subunits in mammalian cells correlates highly with that of CTCF (Parelho et al., 2008; Wendt et al., 2008; Rubio et al., 2008). Parelho *et al.* expressed a FLAG-tagged cohesin subunit of Rad21 in mouse lymphoid cell lines which was able to pull down the SMC1, SMC3 and SA1 cohesin subunits. They generated ChIP-chip libraries covering approximately 3% of the mouse genome and determined that cohesin binds DNase hypersensitive regions enriched for a motif highly similar to the CTCF consensus binding sequence (Parelho et al., 2008). Wendt *et al.* followed a similar path, using the HeLa cell line to create ChIP-chip libraries covering approximately 1% of the human genome for both SMC3 and CTCF (Wendt et al., 2008). Rubio *et al.* studied the binding partners of CTCF by

mass spectrometry and found that Scc3/SA1 was a key binding partner, leading them to create ChIP-chip libraries for CTCF and Scc3/SA1 in HBL100 cells (Rubio et al., 2008). All three studies came to the same conclusion - that the majority of CTCF and cohesin binding sites overlap in mammalian genomes, leading to a model where the sequence specific binding of CTCF is responsible for the targeting of the cohesin complex.

In these and investigations, cohesin has been implicated at many if not most of the loci described in the above sections. Stedman *et al.* and Wendt *et al.* both demonstrated that cohesin is enriched at the CTCF binding sites of the imprinted *H19/Igf2* locus in the same methylation- and parent of origin-sensitive manner and that this binding is ablated upon mutation of the CTCF sites (Stedman et al., 2008; Wendt et al., 2008). Enrichment of cohesin at this locus is also disrupted in mutants lacking the C-terminal domain of CTCF responsible for binding cohesin (Xiao et al., 2011) and the presence of cohesin is required for the three-dimensional conformation of the locus (Nativio et al., 2009). Cohesin has also been implicated in the control of the beta-globin locus (Wendt et al., 2008; Hou et al., 2010) and *Ifng* / T_H2 LCR (Parelho et al., 2008; Hadjur et al., 2009).

CTCF depletion does not affect the amount of cohesin bound to the genome, but rather the enrichment of cohesin at specific sites (Parelho et al., 2008; Wendt et al., 2008). Schmidt *et al.* showed a subset of cohesin bound regions independent of CTCF binding in MCF-7 cells which colocalise with ER- α binding (Schmidt et al., 2010). A similar association of cohesin and mediator / Nipb1 at promoters and enhancers in ES cells has been described (Kagey et al., 2010). These studies suggest an attractive hypothesis that multiple tissue-specific proteins may be able to target cohesin binding to specific sites to affect transcriptional profiles.

1.6.4 Tethering

The nucleus is a structured organelle which contains a number of architectural features such as nuclear pores, the inner nuclear membrane (nuclear lamina) and the nucleolus. Chromatin can bind to these regions in a specific nature leading to changes in nuclear organisation and gene expression.

As microscopy studies have advanced our understanding of how the genomic positioning of genes

can affect expression, it has became clear that association with the nuclear periphery generally correlates with gene silencing (Kosak et al. (2002); Dietzel et al. (2004); Zink et al. (2004); Finlan et al. (2008); Reddy et al. (2008); discussed in Section 1.3.1). In 2008, Guelen *et al.* used the DamID technique with lamin-B1 tethered to DNA adenine methyltransferase (Dam) to identify regions of the genome associated with the nuclear lamina in human lung fibroblasts (Guelen et al., 2008). They found that the fraction of lamina-associated chromatin on each chromosome correlated with known CT positioning preferences in fibroblasts. Their key finding was that chromatin-lamina associations existed as distinct regions, termed lamina associated domains (LADs). The domains range from 0.1 to 10 megabases in size and are enriched for chromatin marks associated with transcriptional repression: H3K27me3, H3K9me2, low H3K4me2, low RNA polymerase II, low gene expression and low gene density. LAD boundaries are enriched for CTCF binding and CpG island, suggesting a mechanism of association (Guelen et al., 2008). Shimi *et al.* showed in the same year that Lamins A and B form separate meshes on the inner nuclear membrane and relatively static structures within the nuclear matrix, hinting at the existence of highly complex micro-environments based on the binding of chromatin to lamin networks (Shimi et al., 2008). A subsequent study investigating cells with a mutation in the *LMNA* lamin gene has shown that the position, compaction and transcriptional activity of some lamin-associated regions are affected (Mewborn et al., 2010), suggesting that the many diverse conditions caused by lamin mutations may arise due to changes in chromatin structure and so gene expression.

Although LADs are associated with low gene expression, not all chromatin at the nuclear membrane is silenced. Early electron micrographs showed regions of less dense chromatin at nuclear pores (reviewed in Capelson and Hetzer (2009); Arib and Akhtar (2011)). This observation was recently validated with the observation of nuclear pore complexes (NPCs) contacting channels with heterochromatin using a new form of sub-diffraction limit light microscopy able to simultaneously image NPCs, lamins and chromatin (Schermelleh et al., 2008). Nuclear pore proteins have been found to be associated with active regions of chromatin and are present both at NPC and within the nucleoplasm (Vaquerizas et al., 2010; Kalverda et al., 2010).

1.6.5 Actin and myosin

The presence of nuclear actin has been debated for many years; its initial detection often labelled as artefacts due to the inability of phalloidin to stain actin fibrils within the nucleus (reviewed in Hofmann and de Lanerolle, 2006). Despite this skepticism, nuclear actin research has had a resurgence in recent years with studies linking filamentous actin to processes such as transcription (Hofmann et al., 2004) and nuclear export (Hofmann et al., 2001). McDonald *et al.* used fluorescence recovery after photobleaching (FRAP) microscopy to study nuclear actin in HeLa cells (McDonald et al., 2006). Treatment with latrunculin which inhibits the polymerisation of actin lead to a loss of a slow moving population of actin, suggesting the existence of polymeric actin in the nucleus. Within the cytoplasm, force is generated through interaction between bundles of filamentous actin and polymerised myosin II. The nucleus does not contain any myosin II, though an isoform of myosin I incapable of forming filaments has been detected (Pestic-Dragovich et al., 2000). Our understanding of how nuclear actin and myosin are involved in nuclear organisation is still in its infancy, yet is rapidly gaining traction as techniques are developed which allow us to probe and manipulate their behaviour within the nucleus without disrupting cytoplasmic processes.

1.6.6 Replication

Proliferating cells must replicate their genomes once per cell cycle and do so in only a few hours, despite their size. This is accomplished by simultaneously replicating many regions of the genome at shared sites of replication called replication factories (Jackson and Pombo, 1998; Ma et al., 1998). The genome is replicated in an organised manner; active genes are linked to early replication and inactive genes tend to replicate late (reviewed in Goren and Cedar, 2003). Clusters of replication foci that share replication factories continue to associate through multiple cycles of cell division (Jackson and Pombo, 1998). Many studies have shown correlation between behaviour in replication and genomic features or activity, though causative links are still lacking (reviewed in (Chakalova et al., 2005)).

1.6.7 Polycomb

Another nuclear subcompartment known to be involved in genomic organisation and looping is the polycomb body. The Polycomb protein complex (PcG) is involved in the directed silencing of regions at polycomb response elements (PREs), mediated by repressive histone modifications such as H3K27me3 (Cao et al., 2002), and are important for the maintenance of silencing of the *Hox* genes (reviewed in Pirrotta, 1998). Long range interactions and chromatin looping has been implicated in repression by PcG proteins Tiwari et al., 2008b;a.

1.6.8 The bigger picture

When seen as a whole, many of the models described above are not mutually exclusive. Many hold in common the presence of chromatin loops within the nucleus and describe domains of chromatin defined by epigenetic marks. It seems entirely plausible that the overall structure of the nucleus is determined by the combined result of many different processes directing specific interactions, driven by different processes and stabilised by different types of contact. Whilst this picture is still developing, it is clear that the nucleus is a highly dynamic and plastic organelle defined by millions of years of evolutionary pressures on every aspect of its content.

1.7 Chromosomal translocations

One of the major medical implications of genomic organisation is how it may relate to the formation of chromosomal translocations. Specific chromosomal rearrangements are frequently associated with certain cancer types, and can be highly predictive of patient prognosis. Furthermore, cancers involving different translocations can respond to treatments in different ways, paving the way for patient-specific treatment regimes (reviewed in Mitelman et al., 2007). Understanding the principles of chromosomal translocation formation is an important step in the development of novel treatments.

1.7.1 Formation of chromosomal translocations

Since the development of chromosome banding microscopy, it has been known that the genetic material within cancerous cells is frequently disrupted. Chromosomal translocations involve the rearrangement of genetic material between non-homologous chromosomes through the formation and aberrant repair of at least two double strand breaks (DSBs) that are situated on different chromosomes. The resulting product can be a straight swap (balanced translocations) or can result in deletions or even gain of material after malsegregation in mitosis (unbalanced translocations). Chromosomal translocations can involve multiple DSBs on different chromosomes resulting in hugely complex karyotypes, especially in cells predisposed to translocations due to defects in repair.

1.7.1.1 DSB repair

Double strand breaks can be caused by exogenous damage (ionising radiation, free radicals) and endogenous damage (physiological programmed DSBs). As many as 1 million DNA lesions are formed per cell per day (Alberts Lewis, Raff, Roberts, Walter, 2007) and can be highly deleterious to the cell, causing loss of genetic material, translocations and ultimately cell death if left unrepaired. Mammalian cells have an array of different DSB repair pathways that depend on the cellular context and type of lesion (Longhese et al., 2006).

Homologous recombination (HR, Fig 1.7.1B) repairs double strand breaks by using long regions of homology on undamaged sister chromatids or homologous chromosomes. HR is primarily active during cell replication and rarely leads to serious chromosomal rearrangements or translocations, although ‘crossing-over’ can sometimes occur, whereby two sections of homologous chromosomes switch to their partner chromosomes. This process is encouraged during meiosis with the formation of chiasmata, crossovers that are 10^4 to 10^5 times more frequent than in mitosis (Lee et al., 2009). During meiosis homologous chromosomes are preferentially used for HR to promote crossovers, but sister chromatids are more commonly used in mitosis to minimise variation (Schwacha and Kleckner, 1997; Haber, 2000). Non-allelic homologous recombination (NAHR) is a variation of HR characterised by the aberrant use of low-copy repeats during strand invasion. NAHR can result in duplications,

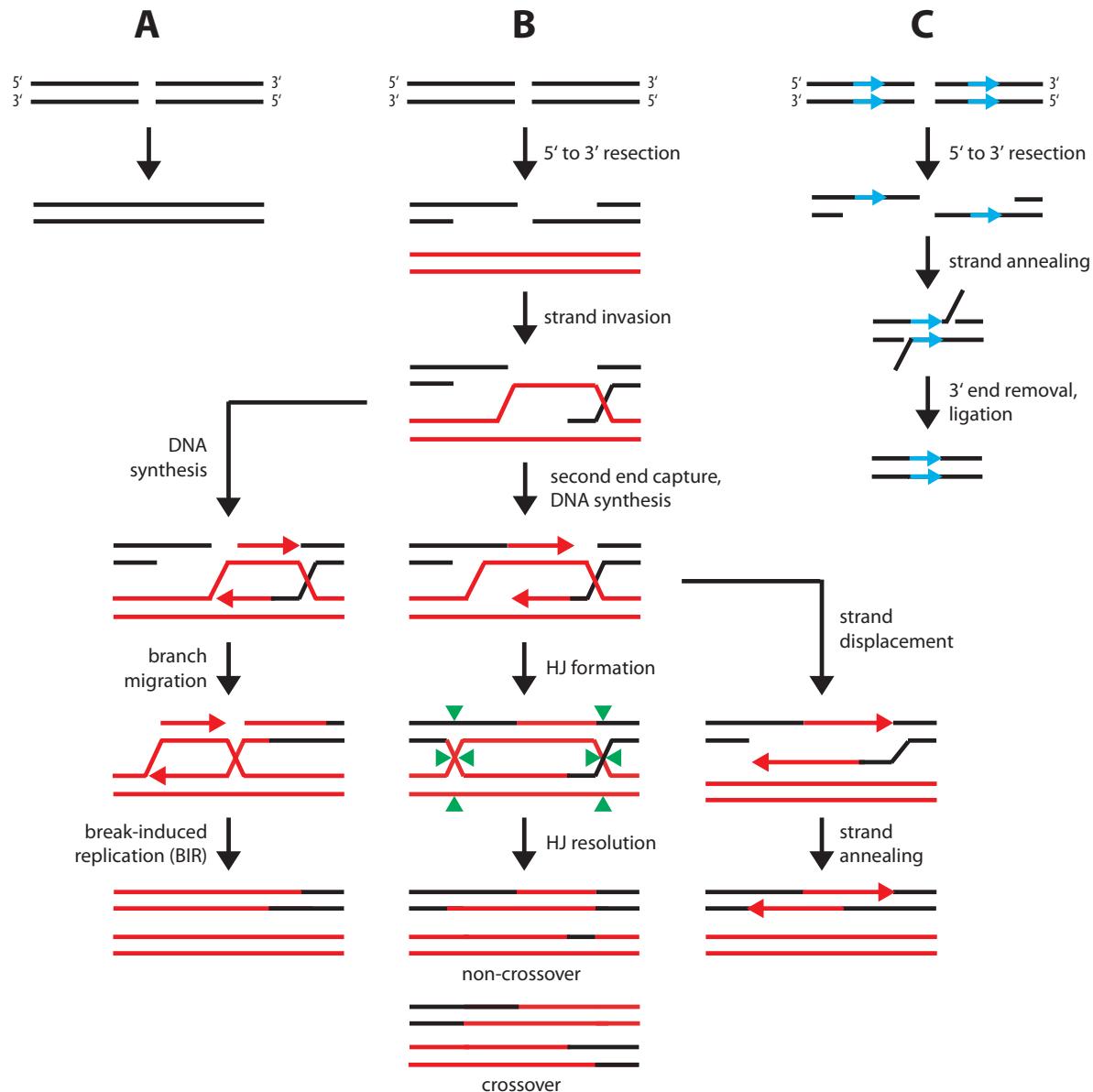


Figure 1.7.1 – DSB repair. Three double strand break repair pathways. **(A)** NHEJ (Non-Homologous End Joining) requires little or no sequence homology and joins any two nearby ends. **(B)** HR (Homologous Recombination) uses a strand of homologous DNA on a duplicated interphase chromosome or sister chromatid to reconstitute sequence before joining. Although it only requires a single double strand break to generate a crossover, the extensive region of homology needed makes errors rare. **(C)** SSA/NHEJ (Single Strand Annealing / Non-Homologous End Joining) searches for regions of homology, such as repeat sequences, and deletes any sequence between. HJ is Holliday junction. From Longhese *et al.*, 2006 (Longhese *et al.*, 2006).

inversions and translocations and is responsible for a number of spontaneous genetic disorders such as Potocki-Lupski syndrome (Potocki et al., 2007) and cancers involving recurring breakpoints (Darai-Ramqvist et al., 2008) (reviewed in Gu et al., 2008).

Single strand annealing (SSA, Fig 1.7.1C) creates a short region of single stranded DNA at the site of the DSB which is used to search for regions of microhomology, usually repeat sequences in the same orientation. These regions anneal and any excess single stranded tails are removed to allow the nicks to be ligated. SSA is prone to introducing deletions into the genome, and can form chromosomal translocations if two sequences with similar repeat regions are nearby, and both suffer simultaneous DSBs.

Non-homologous end joining (NHEJ, Fig 1.7.1A) is a last resort for the cell, whereby any two adjacent DSBs are ligated, needing little or no sequence homology. This pathway is extremely prone to errors as it is capable of joining any two strands of DNA irrespective of their identity.

A number of serious conditions exist due to mutations in genes key in the DSB repair pathways described above. Ataxia telangiectasia is an autosomal-recessive neurodegenerative disease caused by mutations in the *ATM* gene which is involved in NHEJ and HR (Beucher et al., 2009). Fanconi anaemia is a condition caused by mutations in one of a number of proteins involved in DSB recognition and repair which is associated with a high incidence of leukaemia and a number of congenital defects. Blooms syndrome is characterised by a excessive HR and genomic instability caused by mutations in the *BLM* gene. These, and other similar diseases, are indicative of the importance of DSB repair and the danger of genomic instability.

1.8 Leukaemia

Chromosomal translocations are frequently observed within cancer cells and can be one of the initiating events leading to oncogenesis. Leukaemias are one of the best studied cancer models due to the ease of accessibility of the affected cells; many of the best characterised cancer pathologies have been within leukaemias.

1.8.1 Chronic myeloid leukaemia

Chronic myeloid leukaemia (CML) is one of the best studied cancers, despite having a relatively low occurrence (Hehlmann et al., 2007). The term 'leukaemia' was based on CML patients in the mid 1800s (Geary, 2000), and it was the first cancer for which a direct causative genotype was found (Nowell and Hungerford, 1960) (see below). A number of standardised therapies exist for CML and up to 87% of patients achieve complete cytogenetic remission (Hehlmann et al., 2007), making it one of the success stories of cancer research.

1.8.1.1 The Philadelphia chromosome

This genetic abnormality, primarily associated with CML, was first the first consistent chromosomal abnormality found with any cancer. Identified in 1960 by Nowell and Hungerford (Nowell and Hungerford, 1960) they noticed the presence of an abnormal chromosome and called it the Philadelphia chromosome, or Ph chromosome, after the city in which it was discovered. The chromosome was initially thought to be the result of a deletion until chromosome banding enabled its identification as a product of a translocation (Rowley, 1973). The breakpoints within chromosomes 9 (cytogenetic band q34) and 22 (band q22) were subsequently identified as being within the *c-ABL* and *BCR* genes, respectively. The *c-ABL* gene (also known as *ABL1*) was named as such because of its similarity to the Abelson murine leukaemia virus gene *v-ABL* (Rowley, 2001). Less is known about the *BCR* gene, so named because it was found at the Breakpoint Cluster Region of chromosome 22.

The *ABL1* gene encodes a tyrosine kinase which is ubiquitously expressed in mammalian cells. Involved in cell cycle regulation, it is thought to be involved in several cell signalling pathways (Deininger et al., 2000). The *BCR* protein contains a serine-threonine kinase and has GTPase activity, but its function is not known (Deininger et al., 2000). The t(9;22)(q34;q11) translocation creates a fusion protein lacking the SH2 domain of *ABL1* which normally regulates its activity. This results in constitutively active tyrosine-kinase activity which drives oncogenesis (Deininger et al., 2000). The mechanism of translocation formation is yet to be characterised, though ionising radiation is known to be a risk factor (Tanaka et al., 1989; Corso et al., 1995). Identification and characterisation of the

BCR-ABL translocation and its fusion protein led to the development of a number of kinase inhibitors such as the drug *imatinib* (also known as *gleevec*), which achieve excellent success in the treatment of CML and a number of other cancers involving the t(9;22)(q34;q11) translocation (Hehlmann et al., 2007).

1.8.2 Mixed lineage leukaemia

Acute leukaemia is typically categorised as either Acute Lymphoblastic Leukaemia (ALL) or Acute Myeloid Leukaemia (AML) according to which lineage of blood cells are cancerous. ALL is defined by the uncontrolled proliferation of lymphoblasts, precursors to lymphocytes which differentiate into B cells, T cells and NK cells. ALL was estimated to account for approximately 12% of all leukaemia cases in the US in 2008 with around five and a half thousand cases occurring annually. It is the most common form of cancer in children aged under fourteen (Jemal et al., 2009).

AML is a cancer of the myeloid lineages of blood cells, responsible for approximately 30% of all cases of leukaemia in the US for 2008 with a higher mortality rate than ALL (66% and 27% mortality for AML and ALL, respectively) (Jemal et al., 2009). The incidence of AML increases with age; the median age of AML diagnosis between 2002 and 2006 in the US was 67 years, and the median age of mortality was 72 (Jemal et al., 2009).

The biology of leukaemia is not always as discrete as this classification however, some patients present with expansion of both lymphoid and myeloid lineages (Matutes et al., 1997). Two genetic abnormalities sometimes found in these patients include the t(9;22)(q34;q11) Philadelphia chromosome and structural changes in 11q23 (Matutes et al., 1997). Chromosomal translocations within band q23 of chromosome 11 have been implicated in both AML and ALL, all involve the gene *MLL* (also known as *ALL-1*, *Htrx*, *HRX*) identified by Ziemin-van der Poel et al. in 1991 (Ziemin-van der Poel et al., 1991). Translocations involving the *MLL* gene are found in over 70% of all infant leukaemias (Biondi et al., 2000) and approximately 10% of adult AML cases (Krivtsov and Armstrong, 2007). The translocation correlates with poor patient prognosis and is of high clinical interest (Chen et al., 1993).

In 2002, Armstrong *et al.* showed that acute lymphoblastic leukaemias containing a translocation within the *MLL* gene have a unique expression profile that is different to ALL and AML, and suggests an origin within a less committed progenitor cell which can produce cells in both the myeloid and lymphoid cell lineages (Armstrong *et al.*, 2002). They suggest that these leukaemias are substantially different from AML and ALL and deserve a new, distinct, class of leukaemia called Mixed Lineage Leukaemia (MLL). Further cytogenetic studies have supported this theory; leukaemic cells in MLL have been found to express cell surface antigens normally present on both myeloid and lymphoid cells such as CD14 and CD19 (Krivtsov and Armstrong, 2007).

The suggestion that MLL is initiated within a progenitor cell is in line with previous evidence showing that some cases of ALL and AML are initiated within undifferentiated haematopoietic stem cells (HSCs) (Sutherland *et al.*, 1996; Bonnet and Dick, 1997). HSCs are present within the CD34⁺ progenitor cell population, accounting for approximately 3% of normal human bone marrow and 0.3 to 0.5% of human cord blood mononuclear cells (Libura *et al.*, 2008). They are long lived and capable of self-renewal, differentiating into lineage restricted progenitors and eventually mature terminally differentiated white blood cells. HSCs are necessary for the long term maintenance of the haematopoietic system and are commonly used to repopulate bone marrow after myeloablative therapy, as well as in the treatment of a number of other disorders such as autoimmune, cardiac and vascular diseases (Burt *et al.*, 2008). Cancer stem cells are thought to be present in both leukaemias and solid tumours (Hamburger and Salmon, 1977) and HSCs are a probable founder population due to their ability to self-renew (Reya *et al.*, 2001). The concept of a small pool of cancer stem-cells driving the large heterogeneous pool of cancer cells has a number of implications for treatment - these are the cells that must be targeted for the efficient and long lasting cure of cancer (Reya *et al.*, 2001).

1.8.2.1 The MLL protein

The MLL protein is a H3K4 methyltransferase involved in the positive regulation of global gene regulation, including the maintenance of expression of the *Hox* genes (Yu *et al.*, 1995). *MLL* is required for embryonic haematopoiesis (Hess *et al.*, 1997) and adult bone marrow maintenance (Jude

et al., 2007). MLL is a mammalian homologue of the *Drosophila melanogaster* trithorax complex and is thought to bind DNA via an AT-hook domain (Zeleznik-Le et al., 1994) and a zinc finger domain (Birke et al., 2002). It is thought that the zinc finger domain targets MLL to unmethylated CpG island DNA (Birke et al., 2002) and ChIP studies have shown that MLL binds to a subset of transcribed genes (Milne et al., 2005). MLL binds promoters and gene bodies, associating tightly with RNA polymerase II (Milne et al., 2005).

All known MLL fusion proteins contain exons 8-13 of *MLL* and in-frame exons of a partner gene (Krivtsov and Armstrong, 2007). Fusion proteins always retain their AT-hook and zinc-finger CxxC motifs, which are essential for their transforming potential (Slany et al., 1998). The H3K4 methyltransferase domain of MLL is often lost in fusion proteins (Krivtsov and Armstrong, 2007), despite this the fusion proteins can drive constitutive expression of *HOXA9* and *MEIS1* which, if over-expressed together in the absence of a MLL fusion protein, give a similar phenotype (Zeisig et al., 2004).

1.8.2.2 The *MLL* gene

Translocations within the *MLL* gene are found in approximately 10% of all human leukaemias (Huret et al., 2001). There are 87 documented *MLL* translocation partners of which 51 have been characterised at the molecular level (Meyer et al., 2006); the five most frequent translocation partners, *AF4*, *AF9*, *ENL*, *AF10* and *AF6* account for approximately 80% of cases (Meyer et al., 2006). Translocations within the *MLL* gene usually occur within an 8.3 kb *BamHI* fragment known as the breakpoint cluster region (Gu et al., 1994). This region contains exons 5-11 as well as a number of repeat regions, notably eight direct *Alu* SINE repeats, five direct L1 and L2 LINE repeats and two MER elements, as well as a number of putative topoisomerase II binding sites and a SAR/MAR (Sung et al., 2006) (Fig 1.8.1). An internal promoter is present within the murine *Mll* breakpoint cluster region, correlating with etoposide-induced DSBs (Scharf et al., 2007), DNase I hypersensitive sites (Strissel et al., 1998) and histone modifications associated with transcription (Khobta et al., 2004).

1.8.2.3 Mechanisms of *MLL* translocation formation

A clue to how translocations form within the *MLL* gene comes from the observation that 11q23 translocations are especially prevalent in therapy-related leukaemias - secondary leukaemias that develop in patients after treatment for a primary cancer with topoisomerase II inhibitors (Krivtsov and Armstrong, 2007). Topoisomerases are found in all eukaryotic nuclei and are able to relieve supercoiling and promote chromosome disentanglement (Buck and Zechiedrich, 2004). They function by binding DNA, forming a transient double strand break and passing another strand of DNA through the gap, before ligating the DSB.

Topoisomerases are important in transcription, which creates supercoils as template DNA is processed through static RNA polymerase II enzymes (Liu and Wang, 1987). DNA topoisomerase II associates with gene promoters (Collins et al., 2001) and is required for the transcription of genes longer than 3 Kbp in yeast (Joshi et al., 2012). Topoisomerase II induced DSBs have been implicated in the regulation of certain genes via the assembly of transcription complexes and changes in chromatin structure (Ju et al., 2006).

DNA topoisomerase II inhibitors are commonly used as chemotherapeutic agents and work by decreasing the ligation rate, disrupting the cleavage / ligation equilibrium. This leads to an increase in DNA cleavage and an accumulation of DSBs, triggering the cell DNA damage response and leading to cell death by apoptosis (Burden and Osheroff, 1998). Topoisomerase II inhibitors are widely used chemotherapeutic agents, effective against a range of malignancies including small-cell lung cancer and gonadal tumours (Arnold and Whitehouse, 1981).

The *MLL* breakpoint cluster region contains a number of putative topoisomerase II binding sites suggesting a mechanism of topoisomerase inhibitor related translocations (Broeker et al., 1996). Libura *et al.* showed that CD34⁺ HSPCs exposed to etoposide formed 11q23 chromosomal aberrations typical of those seen in clinical samples (Libura et al., 2005). They went on to show that such exposure to etoposide increased the proliferative potential of the cells in a bone marrow graft assay using immune-deficient mice (Libura et al., 2008). This data supports a model whereby poisoned topoisomerase II may create double strand breaks within the *MLL* gene, allowing the NHEJ machinery

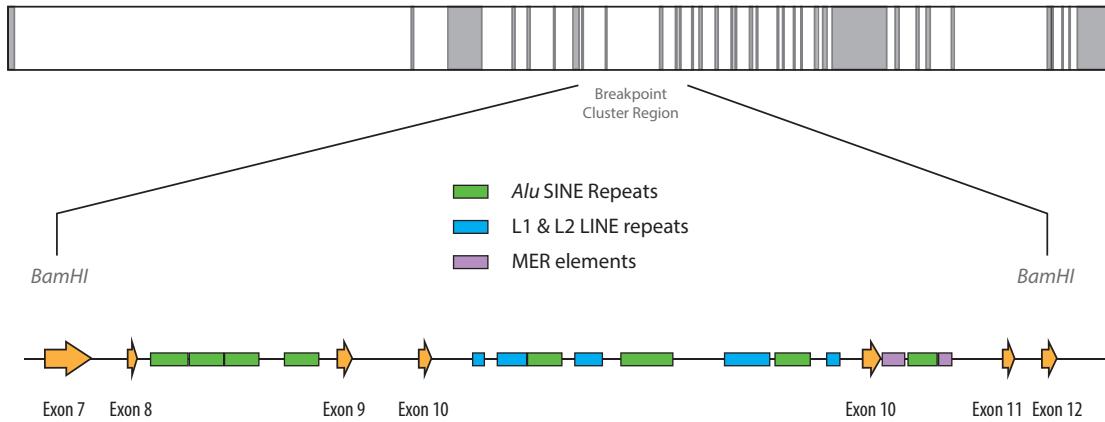


Figure 1.8.1 – Diagram of the *MLL* breakpoint cluster region.

to aberrantly repair the locus due to microhomology found in the nearby repetitive elements (Fig 1.8.1).

1.9 Effect of nuclear organisation on translocation formation

1.9.1 Breakage first and contact first models

For a chromosomal translocation to form, two double strand breaks must exist simultaneously and be adjacent in three-dimensional space. Two models have been described to explain how DSBs may meet in the nucleus – the breakage first model and the contact first model. The breakage first model states that DSBs can form anywhere and are able to freely diffuse in the nuclear space. They undergo large scale movement through the nucleus until they meet and are joined. The contact first model states that the two DSBs form in sequences already close to each other within the nuclear space, therefore large scale chromatin movements are not required for the two breaks to meet.

In support of the breakage first model, Aten *et al.* showed that DSBs formed simultaneously in HeLa nuclei clustered together (Aten et al., 2004). However, others have argued that this may be

the result of a higher degree of chromatin mobility along the ion beam trajectory used to generate the DSBs (Jakob et al., 2009). A larger body of evidence supports a contact first model; DSBs created with ultra-soft X-rays by Nelms *et al.* remained in a fixed position for several hours after the damage was caused (Nelms et al., 1998) and Jakob *et al.* used live cell microscopy to visualise proteins involved in DNA damage signalling and repair. After accumulation at sites of DNA damage caused by heavy ion impacts, the foci exhibited only a small degree of movement (Jakob et al., 2009). Soutoglou *et al.* showed a high degree of positional stability of DSB ends created by endonuclease digestion, with a greater degree of local diffusion seen in the absence of the Ku80 DNA-end binding protein (Soutoglou et al., 2007).

The contact first model of translocation formation has large implications within the field of nuclear organisation, as it requires specific chromosomal contacts within cells prior to translocation.

1.9.2 Chromosome territories and translocations

The large scale organisation of chromosome territories has been implicated in translocation frequency by a number of studies. Kozubek *et al.* showed that chromosomes 9 and 22 were found in the centre of nuclei more frequently than would be expected by chance in lymphocytes, T- and B-cells, HL60 cells and bone marrow cells (Kozubek et al., 1999). They used neutron irradiation to show that transfer of genetic material was much higher than that found with chromosome 8, which was situated towards the nuclear periphery (Kozubek et al., 1999). Parada *et al.* investigated the positions of chromosomes 12, 14 and 15 in a mouse lymphoma cell line and mouse splenocytes (Parada et al., 2002). They found that two translocated chromosomes preferentially paired together in the nucleus of the cell line as well as in normal cells not containing the translocation. (Parada et al., 2002). Parada *et al.* went on in a further study to examine the positioning of a larger range of chromosomes in a number of different tissue types (Parada et al., 2004). They found that chromosome pairing was tissue specific and correlated with the occurrence of tissue-specific translocation events (Parada et al., 2004). Kuroda *et al.* published a similar study in the same year, showing that association between chromosomes 12 and 16 varies through adipocyte differentiation. These two chromosomes are involved in a chromosomal translocation that can

lead to liposarcomas, thought to be initiated within pre-adipocytes (Kuroda et al., 2004).

In 2006, Branco and Pombo published a study where they developed a new technique known as cryo-FISH to enhance the resolution of chromosome territory detection (Branco and Pombo, 2006). This technique had sufficient resolution to study the degree of intermingling between chromosome territories and they found that the degree of intermingling between chromosome pairs correlated strongly with their propensity to form translocations when subjected to radiation (Fig 1.9.1, Branco and Pombo, 2006).

Interestingly, derivative chromosomes that result from balanced translocations affect the organisation of CTs within the nucleus (Harewood et al., 2010), raising the possibility that the global changes in gene expression observed after oncogenic translocation formation could be in part due to changes in genome organisation (Harewood et al., 2010).

1.9.3 Transcription factories and translocations

Mounting evidence has shown that in addition to CT pairing, the position of specific genes involved in chromosomal translocations are frequently found in close proximity (Neves et al., 1999; Roix et al., 2003). As discussed in Section 1.6.1, localisation at transcription factories can drive organisation within the genome. Osborne *et al.* used DNA- and RNA-FISH to study the localisation of the *Myc* proto-oncogene upon activation of B cells (Osborne et al., 2007). They showed that *Myc*, on mouse chromosome 15, is dynamically recruited to an existing transcription factory and that this transcription factory was preferentially occupied by *Igh*, on mouse chromosome 12. These observations of preferential association at transcription factories support to a model whereby genes sharing transcription factories are predisposed to the formation of chromosomal translocations (Osborne et al., 2007). Similar transcriptional associations have recently been observed for *MLL* and its frequent

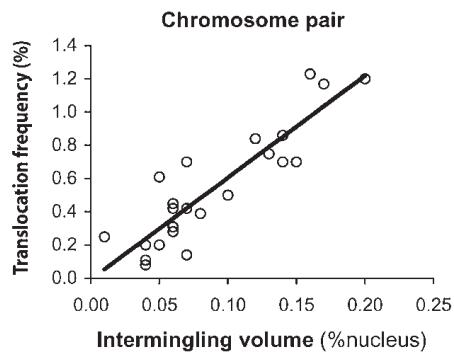


Figure 1.9.1 – Correlation of CT intermingling and radiation induced translocation frequencies. Adapted from Branco and Pombo, 2006 (Branco and Pombo, 2006).

translocation partners *AF4* and *AF9* (Cowell et al., 2012).

1.10 Thesis overview

Our understanding of nuclear architecture and organisation has developed rapidly in the last twenty years, in concert with the development of techniques allowing ever larger and less biased studies. Correlations have been found between the organisation of specific chromosomes and gene loci and their propensity to form chromosomal translocations which can lead to oncogenesis. A limitation of many of these studies is their scope: either they describe the gross organisation of chromosomes and large genomic features, or they describe the association of specific loci. Gross nuclear organisation is valuable in determining global rules and patterns of genomic organisation, but lacks the resolution needed to study specific loci. Studying the positions of individual genes is also useful, but limited to the study of anticipated associations, leading to candidate choice bias.

In this thesis I describe the further development of the e4C technique which allows the interrogation of the three-dimensional contacts made by a chosen locus in a relatively unbiased, genome-wide manner. I investigate the association profiles of three proto-oncogenes: *BCR*, *ABL1* and *MLL*. I do so in human CD34⁺ haematopoietic stem cells, the tissue though to be the site of initial chromosomal translocations for these genes, and the human lymphoblastoid cell line GM12878.

The e4C protocol was previously used in mouse foetal liver cells for analysis with microarrays (Schoenfelder et al., 2010). In chapter three I describe my modification of the protocol for use with limited numbers of human CD34⁺ cells and analysis with high-throughput sequencing technologies. This work demonstrates the potential of the technique to analyse the interaction profiles of any genomic locus, as well as a number of pitfalls which must be negotiated in its use.

Chapter four describes my subsequent development of analysis tools and techniques which I use to understand the sequencing data. At the time this work was carried out, such genome-wide interaction datasets were being produced for the first time and a great deal of exploration was required in order to find useful forms of analysis which could provide accurate and unbiased representations of the association data. In this chapter I describe a number of steps developed to normalise biases and analyse the data, several of which can be found in parallel studies published during the same period.

In chapter five I describe the initial analysis of association data from the three genes, which exhibit a common association with actively transcribed regions of the genome. This analysis is feasible due to the availability of a number of publicly accessible datasets describing the binding profiles of RNA polymerase, histone modifications and abundance of transcripts. The patterns I uncover support similar findings by other groups showing the separation of the genome into active and inactive compartments.

In chapter six I analyse the interaction of *BCR* with the telomeric region of chromosome 9, containing its recurrent translocation partner *ABL1*. I demonstrate that this interaction is the strongest in the genome and highly significant. Interestingly, the maximal point of association with *BCR* is not at the *ABL1* locus, but rather in three loci several megabases away. I describe my use of DNA-FISH to validate these interactions in CD34⁺ cells and GM12878 cells.

Unbiased yet specific studies of genomic interactions such as those described within this thesis help to demonstrate the importance of genomic organisation in the formation of chromosomal translocations and initiation of cancer. By furthering our understanding of how healthy cells are regulated, we may better understand how to prevent disease causing events.

Chapter 2

Materials and Methods

2.1 CD34⁺ cell handling

2.1.1 Peripheral blood collections

Mobilised CD34⁺ cells were collected from peripheral blood samples supplied by Dr. George Follows with the help of apheresis coordinator / specialist nurse Paul Boraks. These blood samples were taken with consent from the patient at the same time as clinical samples used to test the mobilisation efficacy of the GCSF stimulation. Blood samples were collected and immediately diluted into 5ml RPMI with heparin to prevent agglutination. Samples were transported from Addenbrookes Haematology Day Unit to the Babraham Institute.

2.1.1.1 Buffy coat isolation and fixation of cells

Samples were diluted with 2-4 volumes of PBS (Dulbecco's PBS; PAA) containing 2mM EDTA (ultra-pure; GIBCO). 15 ml Ficoll (Ficoll-Paque Premium; GE Healthcare) was added to empty 50 ml tubes, and 35 ml cell suspension was layered on top. Samples were centrifuged at 400x g for 40 minutes

at room temperature. The central buffy coat found at the interface between the plasma and Ficoll layer, containing primarily mononuclear cells, was carefully pipetted into a new labelled 50 ml tube. This was washed in PBS containing 2 mM EDTA and centrifuged at 300x g for 10 minutes at room temperature. Cell pellets were washed twice by resuspending in 50 ml MACS Buffer (D-PBS [PAA] with 0.5% BSA [Sigma] and 2 mM EDTA [Gibco]) and re-centrifuged at 200x g for 15 minutes at room temperature. Cell pellets were resuspend cells in 2 ml total volume MACS Buffer and transferred to tubes containing DMEM (PAA) with 10% FBS (PAA). Cells were then fixed with 2% formaldehyde (histology-grade, min 37% free from acid; Merck) for 10 minutes at room temperature with constant mixing. The fixation was quenched by adding glycine to a final concentration of 0.125 M. Fixed cells were centrifuged at 300x g for 10 mins at 4°C before being resuspend in 50 ml ice-cold MACS Buffer. Cells were counted using a haemocytometer.

2.1.2 Leukapheresis collections

Mobilised CD34⁺ cells were collected from leukapheresis collections supplied by Dr. George Follows and Dr. Kevin Jestice. These samples were made available after the harvesting of mobilised CD34⁺ cells from patients who had responded to GCSF stimulation exceptionally well, meaning that there was an excess of cells not required for clinical use. Leukapheresis samples were refrigerated in serum bags until collection from Addenbrookes National Blood Service, and transported to the Babraham Institute.

2.1.2.1 Cell washing and fixation of cells

Because white blood cells are already separated from the blood in the leukapheresis samples, buffy coat isolation is not necessary. The much higher number of cells to be processed is also a point of consideration in the processing of these samples.

The sample was aliquotted into 50 ml tubes and diluted with 2 - 4 volumes of PBS with 2 mM EDTA, to a final volume of 50 ml. Tubes were centrifuged at 300x g for 10 minutes at room temperature. Cells were washed twice by resuspending in 50 ml PBS with 2 mM EDTA and centrifuged

at 200x g for 15 minutes at room temperature. Each cell pellet was resuspended in 50 ml PBS with 2 mM EDTA and transferred to a 500 ml bottle of DMEM (PAA) supplemented with 10% FBS (PAA). Cells were fixed by adding 26 ml 37% formaldehyde was added to a final concentration of 2% (histology-grade, min 37% free from acid; Merck) and being placed on a rocker for 10 minutes at room temperature. The fixation was quenched by adding glycine to a final concentration of 0.125 M. Samples were centrifuged at 300x g for 10 mins at 4°C. The supernatant was removed and each cell pellet resuspended in 50 ml ice-cold MACS buffer. Cells were counted using a haemocytometer.

2.1.3 CD34⁺ cell separation

CD34⁺ cells were separated from the mixture of fixed buffy coat cells using either the Invitrogen Dynal CD34 Progenitor Cell Selection System or the Miltenyi MACS CD34 MicroBead Kit. See section 3.4 for the results of both separation methods.

2.1.3.1 Cell separation using the Invitrogen Dynal CD34 Progenitor Cell Selection System

Cell separation was carried out as per the manufacturer's instructions. Samples were centrifuged at 400x g for 8 mins at 4°C and resuspended in Sort Buffer (D-PBS [PAA] supplemented with 0.1% BSA [Sigma] and 2 mM EDTA [Gibco]) at a concentration of 4 x 10⁷ to 1 x 10⁸ per ml in 2 ml microcentrifuge tubes. Invitrogen Dynalbeads were aliquotted into a 2 ml microcentrifuge tube (100 µl per ml sample). 1 ml Sort Buffer was added and the tubes placed on a separation magnet. The supernatant was removed and the beads resuspended in 100 µl sort buffer per ml sample. The beads were then added to the samples and mixed, before incubating on a rotating wheel at 4°C for 30 minutes. 700 µl Sort Buffer was added to each tube before placing each tube on the separation magnet for at least 2 minutes. The supernatant was then removed and kept at 4°C to use as a negative control in the FACS analysis. The beads were then washed three times in 2 ml Sort Buffer, separating on the magnet for at least 1 minute each wash. The beads were resuspended in 100 µl Sort Buffer and 100

μ l DETACHaBEAD added per tube. Samples were incubated at room temperature on a shaker at 600 rpm for 45 minutes. 1.8 ml Sort Buffer was added to each tube before being placed on the separation magnet for at least 2 minutes. The supernatant was then transferred to a 15 ml tube and the beads washed three times with 500 μ l Sort Buffer, the supernatant being added to the 15 ml tube each time. An aliquot was taken to count the cells using a haemocytometer and the remaining sample was made up to 15 ml with Sort Buffer. 1 ml was removed into fresh tube for the FACS purity analysis. The remaining sample was centrifuged at 400x g for 8 minutes at 4°C. The supernatant was discarded and the cell pellet flash frozen in liquid nitrogen before storage at -80°C.

2.1.3.2 Cell separation using the Miltenyi MACS CD34 MicroBead Kit

Samples were centrifuged at 300x g for 10 mins at 4°C and resuspend in 300 μ l MACS buffer per 1 $\times 10^8$ cells. The following assumes a single aliquot of 1 $\times 10^8$ cells, and was scaled up as necessary. Some leukaapheresis samples had enough cells to warrant thousands of pounds worth of MACS beads, in these cases I used aliquots of greater than 1 $\times 10^8$ cells.

Cell separation was carried out as per the manufacturer's instructions. In brief, 100 μ l MACS FcR Blocking reagent was added to block non-specific binding of the MicroBeads before adding 100 μ l CD34 MicroBeads. Samples were mixed and refrigerated for 30 minutes at 4°C. 10 ml MACS Buffer was added and samples centrifuged at 300x g for 10 mins at 4°C. Samples were resuspend in 500 μ l MACS Buffer and applied to pre-wetted MACS MS columns held in an OctoMACS separator magnet at 4°C. Flow through was collected and stored at 4°C to use as a control in the later FACS analysis. Columns were washed three times with 500 μ l MACS Buffer and cells were eluted from their columns in 1 ml MACS Buffer directly into a second pre-wetted MACS MS column. Multiple samples were combined into a single column at this point. This second magnetic separation step greatly increases the purity of the separated cells. The column was washed three times with 500 μ l MACS buffer and cells eluted in 1 ml MACS Buffer. Purified cells were counted and made up to 15 ml with MACS buffer. A 1 ml aliquot was taken for later FACS purity analysis before the remaining sample was centrifuged at 300x g for 10 minutes at 4°C. The supernatant was discarded and the cell pellet flash

frozen in liquid nitrogen before storage at -80°C.

2.1.3.3 FACS CD34⁺ purity analysis

FACS analysis was used to quantify the percentage of cells staining positively for the CD34 antigen. Representative FACS plots can be seen in Figure 3.4.1.

The 1 ml aliquot taken after the magnetic sort was transferred into a 2 ml microcentrifuge tube, along with two aliquots of no more than 1 x 10⁶ cells was taken from the unbound magnetic sort. Cells were centrifuged at 300x g for 10 minutes at 4°C and resuspended in 100 µl in MACS buffer. 10 µl CD34-APC antibody was added to the purified sample and one of the two unbound controls, and all three samples were and incubated in the dark at 4°C for 10 minutes. Cells were diluted with 1.8 ml MACS buffer and centrifuged at 300x g for 10 mins at 4°C. Cell pellets were resuspend in 200 µl PBS and analysed using a BD FACSCalibur flow cytometer.

2.2 Cell culture

ENCODE cell lines GM12878 and GM06990 were grown in cell culture facilities at the Babraham Institute. Both cell lines were obtained from Coriell Cell Repositories and grown according to their recommendations. Cell lines were grown in RPMI 1640 with 2mM L-glutamine (PAA; E15-840), supplemented with 15% fetal bovine serum (PAA; A11-152) and 1x Penicillin / Streptomycin (PAA; P11-010). Separate bottles of culture were kept for each cell line. Cultures were kept in T25 flasks, stood upright with loose caps at 37 °C in 5% CO₂.

Cultures were passaged when cell densities reached 1 x 10⁶ cells / ml, typically every other day. Cell culture media was made up if required and warmed to 37 °C in a water bath. Cell cultures were removed from the incubators into a cell culture laminar flood hood, and resuspended using a 10 ml pipette. Cultures were transferred to a 50 ml tube and counted using a haemocytometer. Tubes were centrifuged at 300x g for 3 minutes at room temperature. Supernatants were discarded, and the cell pellet resuspended in fresh media to between 2 x 10⁵ and 5 x 10⁵ cells / ml. Resuspended cell solutions

were transferred to fresh T25 flasks and put back into the incubator.

Stocks of both cell lines were stored in liquid nitrogen, frozen in 1 ml aliquots of 5×10^6 cells in RPMI 1640 with 2mM L-glutamine (PAA; E15-840), supplemented with 20% fetal bovine serum (PAA; A11-152) and 6% Dimethyl Sulfoxide (DMSO - Sigma-Aldrich; #154938).

2.3 3C

Chromosome Conformation Capture (3C) is a ligation based proximity assay used to determine the physical association of sequences of DNA in within the nucleus Dekker et al. (2002) (Fig 3.1.1). The 3C protocol used was based on that published by Cope and Fraser Cope and Fraser (2009), with some modifications for using CD34⁺ cells and different restriction endonucleases. See Chapter 3 for more details about the development of this assay.

2.3.1 Nuclei preparation and digestion

CD34+ cells were fixed, sorted and flash frozen in cell pellets as described in Section 2.1. Cell pellets were thawed on ice and resuspended in 50 ml permeabilization buffer (10 mM Tris-HCl, 10 mM NaCl, 0.2% Igepal CA-360 [Sigma], 1 tablet complete EDTA-free protease inhibitor [Roche]). Cells were incubated on ice for 30 minutes on a rocker whilst an aliquot was taken to count the nuclei using a haemocytometer. Samples were centrifuged at 760x g for 5 minutes at 4°C and resuspended in 500 µl 1.2x NEB3 buffer (100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl₂, 1 mM Dithiothreitol, pH 7.9 [New England Biolabs]). Samples were transferred into a 1.5 ml microcentrifuge tube and centrifuged at 760x g for 5 min at 4°C. Nuclei pellets were resuspended in 500 µl fresh 1.2x NEB3 and SDS was added to a final concentration of 0.3% to further permeabilise the nuclei and remove protein that has not been crosslinked.. Nuclei were incubated at 37°C for 1 hour, shaking at 950 rpm. Triton-X100 was added to a final conentration of 1.8% to sequester the SDS and samples incubated at 37°C for 1 hour, shaking at 950 rpm. A 10µl aliquot was taken for later digestion efficiency analysis (See Section 2.3.3).

To digest the crosslinked chromatin 30 μ l high concentration *AseI* was added (1500 units [New England Biolabs]) and samples incubated at 37°C overnight, shaking at 950 rpm. An additional 10 μ l *AseI* (500 units) was added and incubated for a further 3 hours at 37°C, 950 rpm. Another 10 μ l aliquot was taken to check for digestion efficiency later and the digestion stopped by the addition of SDS to a final concentration of 1.6% before incubating at 65°C, 950 rpm for 25 minutes.

2.3.2 Ligation and purification

Samples were cooled to room temperature and added to 15 ml tubes containing 7 ml 1.1x T4 DNA ligase buffer (Made using 10x buffer [New England Biolabs] - 55 mM Tris-HCl pH 7.5, 11 mM MgCl₂, 1.1 mM ATP, 11 mM DTT, 27.5 μ g/ml BSA). To sequester the SDS, Triton-X100 was added to a final concentration of 1% and samples incubated at 37°C for 1 hour with mixing by inversion every 10 minutes. Samples were allowed to equilibrate in a 16°C in water bath before the addition of 2 μ l T4 DNA ligase (800 units [NEB]). Samples were incubated at 16°C for 4 hours, then at room temperature for 30 minutes. Proteinase K was added to a final concentration of 100 μ g/ml and samples incubated at 65°C overnight. Tubes were cooled to room temperature and RNase A added to a final concentration of 40 μ g/ml before incubation at 37°C for 1 hour. Samples were split into two and transferred into 15 ml Phase Lock Gel Tubes (Phase Lock Gel Light 15 ml [5 PRIME]). DNA was purified using a phenol / chloroform extraction followed by precipitation in ethanol and sodium acetate. DNA pellets were resuspended in 250 μ l Molecular Biology grade water (Qiagen).

3C samples were quantified using the Quant-iT PicoGreen assay (Invitrogen). Dilutions of 3C product in TE Buffer (200 mM Tris-HCl, 20 mM EDTA, pH 7.5) were measured in duplicate and compared to standard curves made using lambda DNA supplied within the Quant-iT PicoGreen kit. Fluorescence of the PicoGreen dye was assayed using a Cytofluor II multi-well plate reader (485 nm excitation wavelength , 530 nm emission wavelength [PerSeptive Biosystems]).

2.3.3 Digestion efficiency analysis

To determine the efficiency of the restriction enzyme digestion step in the 3C protocol, I used quantitative PCR. In brief, primers are designed that flank a known restriction enzyme site, and their performance in quantitative PCR is compared to a set of primers nearby that don't flank a restriction enzyme site. This is done for samples taken before and after the restriction enzyme digestion step.

5 µl aliquots are taken from the 3C samples, before the addition of the restriction enzyme and after the end of the digestion incubation. 500 µl of Proteinase K buffer (5 mM EDTA, pH 8.0; 10 mM Tris-HCl, pH 8.0; 0.5 % SDS) was added with 1 µl of 10 mg/ml Proteinase K (10 µg final). Samples were incubated at 65 °C for at least 30 minutes before being equilibrated at 37 °C. 1 µl of 1 mg/ml RNase A (1 µg final) was then added, and samples incubated for 2 hours at 37 °C. Samples were then phenol-chloroform extracted with ethanol precipitation and resuspended in 60 µl of molecular biology grade water.

Samples are now assayed for digestion efficiency using qPCR. I used SYBR Green Master (Roche - 04913850001), with the manufacturer's protocol. Restriction digestion efficiency is calculated using cycle thresholds, Ct values:

$$\% \text{Restriction} = 100 - \frac{100}{2^{(CtR-CtC)DIG-(CtR-CtC)UND}}$$

Where CtR is the Ct value of the primer pair spanning a restriction enzyme site, CtC is the control primer pair. DIG is the digested sample and UND is the pre-digestion sample. Typically, primers were designed for the region of interest (for example, the e4C bait region) as well as an inactive region of heterochromatin.

2.3.4 Detection of 3C products by qPCR

2.3.4.1 Primer design

Insert the gene of interest into the primer3 programme (gene has been exported from Ensembl).

Square brackets around the AseI / BglII site.

Define primer characteristics.

Primer size: 20 22 24

Primer Tm: 60 62 64

Primer GC%: 30 50 70

For one round PCR: product size ~ 150-230 bp

Difference Tm b/w primers = 1.5C

Number to return: 25 (minimum).

2.3.4.2 Equimolar mix

2.3.4.3 qPCR

Real-time PCR machine: ABI Prism Sequence Detection System

Initial step: 50°C, 2 min Denaturation: 95°C, 10 min 40 cycles: 95°C, 15 s

Dissociation Curve: 95°C, 1 min; 55°C, 30s; 95°C, 30s 60°C, 1 min

Reaction mixture (μ l):

6 H₂O

5 Template

0.75 Forward primer (stock 10 μ M, final 300 nM)

0.75 Reverse primer (stock 10 μ M, final 300 nM)

12.5 SYBRgreen 2x mix

25 Final volume

Protocol

1. Pipette out templates into plate (bottom of wells)
2. Make master mix
3. Pipette out MM (left of wells)
4. Put seal on plate

5. Spin briefly
6. Run qPCR

Using 0.2 pmol (200 fmol) equimolar mix

Mix dilutions:

1 x 10-1

2 x 10-2

4 x 10-3

8 x 10-4

1.6 x 10-4

3.2 x 10-5

3C DNA template is diluted to 25 ng/ μ l

5 μ l template / well (125 ng)

4 replicates per primer pair

2.3.4.4 Analysis

Make standard curves from controls for each primer pair - plot log Ct and log quantity of controls.

Log Ct of samples, work out log quantity from standard curve. $10^{(\log \text{Ct})}$, then average and work out SD.

Plot variance from mean across plate to check for evaporation

Plot values with SD

Check all dissociation curves

2.4 e4C

2.4.1 Primer extension and primary *NlaIII* digestion

12 µg of 3C material was defrosted and cleaned using *solid-phase reversible immobilization* (SPRI), as described in the manufacturer's protocol (Beckman & Coulter Agencourt AMPure XP+ beads). Samples were eluted in 252 µl molecular biology grade water. Six primer extensions were set up with 2 µg template 3C material, alongside a positive control primer extension using 1 µg of human genomic DNA, cut with EcoRI. 50 µl reaction volumes were made up containing capture DNA, 1x ThermoPol buffer (supplied with Vent), 200 µM each dNTPs, 10 pmol biotinylated bait enrichment primer and 2 units Vent (exo⁻) DNA polymerase (NEB). The primer extension was then run on a thermocycler with the following cycle parameters: 95 °C for 4 min, 61 °C for 2 min (primer specific; ~2.5°C below T_m of biotinylated primer), 72 °C for 10 min. Tubes were snap chilled on ice and then briefly pulsed in the microcentrifuge. The six sample tubes were pooled into two microcentrifuge tubes and cleaned using a SPRI cleanup. Samples were eluted in 80 µl molecular biology water (40 µl for the genomic DNA positive control). Samples were digested with 20 units of *NlaIII* (NEB) in 1x NEB4 buffer and 1x BSA (both supplied with *NlaIII*) for 3 hours at 37 °C. Samples were purified using SPRI beads and eluted in 50 µl of Molecular Biology water.

2.4.2 Bait enrichment and secondary *NlaIII* digestion

To enrich the e4C bait regions, bound with biotinylated primers from the extension step, the magnetic streptavidin Invitrogen Dynabeads kilobaseBINDER Kit (M-280) was used. Beads were aliquotted into microcentrifuge tubes, 10 µl (100 µg) per sample. The preservative was removed on a magnet and the beads resuspended in 50 µl binding buffer (provided in Dynabeads kilobaseBINDER kit). This was removed on the magnet and the beads resuspended in 50 µl fresh binding buffer. The primer extension material (50 µl) was added and the tubes incubated on a shaker set to 1200 rpm overnight at room temperature. The supernatant was then removed on a magnet and the beads washed twice in 100

μ l wash buffer, once in 100 μ l 1x TE Buffer and once in 50 μ l 1x NEB4 buffer with 1x BSA (NEB) before being resuspended in 50 μ l NEB4 with BSA.

A second digestion step was then carried out with the enriched 3C material. To the resuspended beads, 0.5 μ l (2.5 units) *NlaIII* was added and tubes were incubate for 2 hours at 37°C, shaking at 1200 rpm. The supernatant was then removed on the magnet and the beads washed twice in 100 μ l wash buffer, once in 100 μ l 1x TE Buffer and once in 50 μ l 1x NEB Ligase Buffer (supplied with NEB T4 ligase). Beads were resuspended in 50 μ l fresh 1x NEB Ligase Buffer heated to 55°C for 5 minutes on a heating block along side a 7.5 μ l aliquot of 100 μ M stock PE Ad 2.0 Nla adapter. Both were snap chilled on ice. Beads were placed on the magnet and the supernatant removed before being resuspended in 40 μ l ligation mixture (1x NEB Ligase Buffer; 200 pmol PE Ad 2.0 Nla adapter; 2000 units NEB T4 DNA ligase). Tubes were incubated at room temperature for 2 hours on a rotating wheel. The supernatant was removed on a magnet and the beads washed twice in 100 μ l wash buffer and twice in 100 μ l 1x TE Buffer.

2.4.3 PCR and germline removal

Beads were washed in PCR wash mix (1x Phusion HF Reaction Buffer (NEB) with 200 μ M each dNTPs) and split into a strip of 8 PCR tubes. 4 strip PCR tubes were used for half of the gDNA positive control and two tubes were used as no template controls. Using the magnet outside its housing, the PCR wash mix was removed and the beads resuspended in 50 μ l of PCR master mix - 1x Phusion HF Reaction Buffer (NEB); 200 μ M each dNTPs; 10 pmol PE Ad 1.0 + nested primer; 10 pmol PE Ad 2.0 Nla adaptor primer; 1 unit HF Phusion Pol II (NEB). The PCR was run on a thermocycler with the following program: 98 °C for 30 seconds; 35 cycles of 98 °C for 10 seconds, 65 °C for 30 seconds and 72 °C for 30 seconds; 72 °C for 5 minuntes; hold at 4 °C. The PCR tubes were pooled back into single sample tubes and the PCR supernatant transferred into a fresh microcentrifuge tube using a magnet. The beads were washed twice in 100 μ l wash buffer and once in 100 μ l 1x TE Buffer before storing at 4 °C as a backup. The PCR supernatant was cleaned using SPRI beads and eluted in 400 μ l Molecular Biology water. 5 μ l sample was kept back for the later gel.

To avoid sequencing religation events, which make up the majority of the e4C library, a digestion step is undertaken using a rare-cutting enzyme that recognises a sequence within the germline locus after the *AseI* site. Aliquots of each sample were taken and processed from this point forwards, taking 100 µl aliquots and keeping the remainder as a backup. Each aliquot was made up to 1x with NEB buffer for the relevant restriction enzyme and then ~65 units of restriction enzyme (*BspEI* for *MLL*, *BglII* for *BCR*). Samples were incubated at 37°C for 2 hours before being cleaned with SPRI beads and eluted in 100 µl of molecular biology grade water. A 1.5% gel was run with 5 µl of the digested gDNA sample and 5 µl of the undigested gDNA sample to check removal of germline bands.

2.4.4 Gel extraction and second round PCR

Half of the e4C sample was run on a 1.5% gel and stained in fresh Ethidium Bromide. Using a UV box to visualise the samples, a gel block was cut out for each sample from 250bp to 700bp. The samples were extracted from these gel blocks using the Qiagen Gel Extraction Kit, as per the manufacturer's instructions (without heating to resuspend). Samples were eluted in 30 µl Qiagen Elution Buffer.

To add the full illumina adapters for sequencing, a second PCR is used. 3 µl e4C sample is made up to a 50 µl PCR mix with 1x Phusion HF Reaction Buffer (NEB); 200 µM each dNTPs; 10 pmol PE PCR 1.0 Primer; 10 pmol PE PCR 2.0 Primer; 1 unit Phusion HF Taq (NEB). The Phusion two step PCR was then run: 98 °C for 30 seconds; 15 cycles of 98 °C for 10 seconds and 72 °C for 30 seconds; 72 °C for 5 minutes; hold at 4 °C. PCR products were cleaned using SPRI beads and eluted in 40 µl molecular biology grade water. To check the products, 10 µl was run on a 1% gel.

2.4.5 e4C library quality control

To test the e4C library before sequencing a panel of PCR reactions were done using two sense primers facing adjacent *AseI* recognition sites. Samples were quantified by qPCR with TaqMan probes and run on an Agilent 2100 Bioanalyzer by Kristina Tabbada of the Babraham Sequencing Facility. Library size and quantities were assessed before deciding to sequence each sample.

2.5 Carrier e4C

Due to the low numbers of CD34⁺ cells initially obtained from the peripheral blood collections, I worked on optimising the e4C protocol for use with low cell numbers. Because of the use of 'carrier chromatin' from other organisms, used to buffer loss of DNA, I named this protocol Carrier e4C.

2.6 DNA fluorescence *in-situ* hybridisation

Modified from D.Bolland original by I.Clay 23/09/08, remodified by D. Bolland 14/01/09. Originally from Thomas Ried lab, NIH, Bethesda and the Cremer lab (I. Solovei chapter in FISH book).

2.6.1 Probe generation

Directly-labeled DNA probes for in situ hybridisation

Nick translation with aminoallyl-dUTP followed by chemical-coupling with alexafluor reactive dye

Start with clean BAC DNA (use Dan's triple precipitation method) or pure plasmid or PCR product DNA prepared using a kit. Make sure DNA is in H₂O, not TE or EB buffer.

1. Add the following components on ice:

DNA to be labeled 1 ug

10 × NTB 5 µl

0.1M DTT (comes with superscript II/III) 5 µl

d(GAC)TP mix (0.5mM each) 4 µl

dTTP (0.5mM) 1 µl

aminoallyl-dUTP (0.5mM) 6 µl

DNA Polymerase I, 10 U/µl 1 µl

DNase I dilution (currently 1:30) 1 µl

H₂O to final volume 50 µl

It is best to scale this up to give pellets large enough to see and to make extra nick-translated DNA that can be coupled to different dyes at a later date. I have found doing a 4x reaction gives around 4ug nick-translated DNA and reasonable pellets. Labelling 1ug of this each time will give enough probe for 100 single spot slides (10ng probe per slide, 22 x 22mm coverslip, 10ul hyb. mix)

For my current batch of DNase I (Roche 04716728001) I have found that a 1:30-1:35 DNase I dilution gives a smear of optimal size after 2 hours at 16°C. However, this will need to be optimized for each new/different batch of DNase I. This can be done by running the above reaction but without nucleotides or DNA Polymerase I and titrating the DNase I dilution. This can range from 1:10-1:2000 so it may be necessary to do multiple tests. A smear of 100bp-800bp with a maximum at ~300bp (before heating) is ideal at this stage. It is then typical to use double the concentration in the actual NT reaction, although this may require some fine tuning. The DNase I will typically go off slowly over time.

Ensure DNA Polymerase is well within date as it is critical this enzyme works well. Dilute DNase I in 1x DNase 1 buffer.

2. Incubate at 16°C for 2h.
3. Run an aliquot of 1μl on a 2% agarose gel to check the size of the fragments. The optimal size for in situ hybridisation is 100-500 bp. Before heating (below) nick-translation reactions run slower in agarose gels (probably due to BSA in NTB). A smear of products from 1kb to 150bp with the peak around 200-300bp is suitable at this stage. If the size range is significantly larger than this (eg. 500bp-1.5kb), add another 1μl of a new DNase I dilution and place at 16°C for a further 15-60 min. Run another aliquot of 0.5μl on a 2% gel to check size again. Repeat until optimum size is achieved.

4. Heat to 75°C for 5 min to inactivate DNase I. If you run 1ul on a 2% gel at this stage the DNA smear will have shifted to 100-500bp. Clean up amine-modified DNA using Qiagen PCR purification kit. Elute in 200ul.

5. Precipitate amine-modified DNA immediately with 20μl NaOAc and 550μl 100% ethanol. Precipitate at -20°C for at least an hour (overnight is best). Centrifuge at 13-14k for 15 min at 4°C. Pellet will be small but visible if NT reaction was scaled up but tiny if a 1x reaction was done. Wash

pellet with 100 μ l 70% ethanol, spin, take off ethanol, allow pellet to air dry.

6. Resuspend pellet of ultraclean amine-modified DNA in 1.25 μ l H₂O per NT reaction. You will lose a lot of DNA in the two cleanup steps but these are necessary to remove all non-probe primary amines that would otherwise compete in the coupling reaction (below). If a single NT reaction was done then presume at this stage that you have ~1ug of NT DNA. If 4x NT reactions were done then resuspend in 5 μ l H₂O and quantify 1 μ l on the nanodrop. Amine-modified DNA can be stored at -20°C at this stage and used in future labeling reactions.

7. If using single-use dried pellets of amine reactive dye, take out of freezer and resuspend in 2 μ l anhydrous DMSO, vortex, spin down. Each tube of dye can be used to label 4-5ug of NT DNA but concentration of the dye in the coupling reaction should always be constant. Once reconstituted with DMSO the dyes rapidly lose reactivity so use immediately (cannot be stored for future use). Use the following table as a guide for probes per tube of reactive dye.

DNA amount	DNA volume	NaB buffer	Volume of dye	Total	Enough for
4 x 1ug	1.25ul each	0.75ul each	0.5ul each	2.5ul (1/4 reactions)	100 slides each
3 x 1.35ug	1.67ul each	1ul each	0.67ul each	3.34ul (1/3 reactions)	130 slide each
2 x 2ug	2.5ul each	1.5ul each	1ul each	5ul (1/2 reactions)	200 slides each
1 x 4ug	5ul	3ul	2ul	10ul (Full reaction)	400 slides

8. Heat amine-modified DNA to 95°C for 5 min in PCR block, snap cool on ice then add NaB labeling buffer (0.2M Sodium bicarbonate pH 8.3) on ice and mix/pulse down. Heating gives 10-20% enhanced labeling. Add amine-reactive dye to amine-modified DNA/NaB, vortex, spin down and incubate at room temperature in the dark for at least 1h.

Amine-reactive dyes are very sensitive to water and DMSO is hygroscopic so allow everything to warm to room temperature before opening tubes/resuspending. Always use fully anhydrous DMSO.

9. Add 40 μ l H₂O and use Qiagen PCR cleanup kit to purify the probe, do two washes of the column with PE. Elute in 40 μ l for 1ug, 60 μ l for 1.35ug, 100 μ l for 2ug or 200 μ l for 4ug probe. Liquid should be slightly orange/yellow for alexa488, pink for alexa555 or blue for alexa647.

10. Can be analysed on the nanodrop using the ‘proteins and labels’ setting to measure dye

incorporation and DNA concentration using Dan's probe analysis excel workbook. You can also run an aliquot (1 μ l should be enough) on a gel also to check size, yield etc. You can then analyse the gel on the phosphoimager to check incorporation and if there is any unincorporated dye. In the FISH, 10ng per single-spot slide/22x22mm coverslip works well with bright probes.

Reagents

10×NTB: 0.5 M Tris-HCl (or Trizma-HCl), pH 7.5, 50mM MgCl₂, 0.5 mg/ml nuclease-free BSA fraction V. Filter purify.

10×NTB 5 ml

1 M Tris-HCl, pH 7.5 2.5 ml

1 M MgCl₂ 0.25 ml

10 mg/ml BSA fraction V 250 μ l

Make to 5ml with nuclease-free water 2 ml

0.1M DTT – comes with Superscript II/III or make your own stock

DNA Polymerase I – NEB cat. no. M0209S from Stores £39.95

DNase I – Roche cat. no. 04716728001 from Stores £89.59

Sodium bicarb. (NaHCO₃) 250 g Sigma cat#S7277 £14.30

Stock solution 0.2 M, pH 8.3 MW 84.01 dissolve 100 mg in 3.95 ml water

Make single use aliquots (5-10 μ l)

Amino-allyl dUTP 1 mg Sigma cat#A0410 £92.10

Stock solution 20 mM MW 523.2 add 95.5 μ l of H₂O

Cheaper, alternate source: Ambion Cat# 8439 50mM solution (50 μ l) £108 or even cheaper

Fermentas R1101 50ul 50mM solution £84.88 through SSC

2 μ l amino-allyl dUTP (50mM stock)

198 μ l H₂O

Make 15 μ l aliquots

d(GAC)TP 10 x mix = (0.5 mM dGTP, dATP, dCTP)

2 μ l dATP (100mM stock)

2µl dCTP (100mM stock)

2µl dGTP (100mM stock)

394µl H2O

Make 10µl aliquots

dTTP 10 x 1µl dATP (100mM stock)

199µl H2O

Make 5µl aliquots

Succinimidyl ester-coupled dyes plus others

Alexa-fluor reactive dye decapacks (10 vials) for microarray applications – from Invitrogen/Molecular probes – dyes available;

Alexa fluor 488 (2nd brightest) A32750 £130

Alexa fluor 555 (1st brightest) A32756 £134

Alexa fluor 594 A32751 £113

Alexa fluor 647 (3rd brightest) A32757 £130

Alexa fluor 555+647 (2 x 10 vials). A32755 £239

Other dyes available – Alexa fluor 488 (TFP) 1mg A30005 £217

Tetrafluorophenyl (TFP) esters are an improvement over the succinimidyl ester (SE or NHS-ester) chemistry typically used to attach fluorophores or haptens to the primary amines of biomolecules. Both reactive chemistries produce the same strong amide bond between the dye or hapten and the compound of interest, but TFP esters are less susceptible to spontaneous hydrolysis during conjugation reactions. Alexa Fluor TFP esters are stable for several hours at the basic pH typically used for reactions – far outlasting succinimidyl esters. Many succinimidyl ester-coupled dyes available from molecular probes or GE life sciences – 1mg aliquots ~£200 – resuspend in 41.4µl anhydrous DMSO, store at -20°C in single use aliquots of 2µl in box with dessicant (eg. Silica gel).

(from old method)

10xNTS: 0.5 M Tris-HCl (or Trizma-HCl), pH 7.5, 0.1 M MgSO₄, 1 mM DTT, 0.5 mg/ml nuclease-free BSA fraction V

10×NTS 5 ml

1 M Tris-HCl, pH 7.5 2.5 ml

1 M Mg SO₄ 0.5 ml

100 mM DTT 50 µl

10 mg/ml BSA fraction V 250 µl

Make to 5ml with nuclease-free water 1.7 ml

2.6.2 DNA-FISH

2.6.2.1 Fixing cells in solution and cytopinning on to slides for DNA FISH

This is a variation of the Hi-C fixation that I have found works for DNA FISH in P20 male germ cells and fetal liver cells (so far). It should work for other cells. Can do variable numbers of cells in different tubes from microtubes to 50ml falcons. For tubes that fit in a microcentrifuge – spin live cells at 4500rpm for 2 min, fixed cells can be spun a little harder – 6000rpm for 2 min. For falcon tubes, centrifuge live cells at 1300rpm, 5 min, 4°C, fixed cells at 1500rpm, 10 min, 4°C. This protocol uses the Shandon Cytospin but the cells are fixed and a very slow speed is used – 3D nuclear structure appears to be largely preserved based on DAPI staining and how spherical they are.

1. Prepare a single-cell suspension of your cells in PBS on ice at 1x10⁶ cells per ml. How you do this will depend on the cell type but it's best to have a very good single cell suspension for FISH – filter through 40µm cell strainers, multiple PBS washes etc etc.
2. Add 37% formaldehyde (FA - Merck) to a final concentration of 2%. Eg. 40ml – 2.28ml, 20ml – 1.14ml, 10ml – 0.57ml, 1ml - 57ul etc)
3. Mix gently on a rocker for 10 min to fix. (10 min fixation works for P20 male germ cells and fetal liver so far – lower times may be needed for other cell types)
4. Quench formaldehyde with 1M glycine (final concentration 120mM – eg. 40ml – 6ml glycine, 20ml – 3ml glycine, 10ml – 1.5ml glycine, 1ml - 150ul etc.)
5. Mix gently on rocker for 5 min, then leave on ice for 15 min.

6. If any aggregations or debris have formed during the fixation then filter through 40um cell strainers to remove at this stage.
7. Top up tube with cold PBS, take some cells to count on a hemocytometer and spin down at 4°C.
8. Discard supernatant and carefully resuspend pellet in PBS at 0.5-3x10⁷ cells/ml (see note in 9). At this stage cells can be stored in the fridge for some time (?unknown at present?) in PBS with 0.05% sodium azide.
9. To cytospin, pipette 100ul of cells into a pre-assembled cytospin funnel/clip/card/slide assembly and spin at 300rpm for 3 min in the Shandon cytospin (Fatima/Wendy Dean). Use the standard polyprep poly-L-lysine slides from sigma. Ensure cells are fully mixed before pipetting into cytofunnel. The optimal number of cells to cytospin for optimal density will have to be determined for each cell type as it depends on cell size and probably many other factors. A good density for FISH is one where the cells are dense but not too many are touching/are not clumping too much. Some cell will be lost during the FISH.
10. As soon as the spin is finished quickly take out and disassemble the cytofunnel assembly and carefully extract the slide, being careful not to scrape the cells with the absorbent card. This is a race against time as the cells will be rapidly drying out since the PBS has now gone. Quickly place slides in a coplin jar containing 1x PBS. Can assess cell density by placing slides in PBS in a petri dish and viewing on TC microscope.
11. Continue with the standard DNA FISH protocol!

2.6.2.2 DNA FISH probe precipitation

1. Add to an eppendorf tube:

Probe DNA (eg. 10-20ng directly-labelled BAC)

1µl (1µg) Cot-1 DNA

1µl (9.7µg) salmon sperm DNA

2. Add Na-Acetate, 1/10 of the total volume of probe DNA mixture (above).

3. Add 100% ethanol, i.e., 2.5 x total volume of mixture of DNA + Na-Acetate.

4. Vortex, store tube at -20°C for 1 hour or overnight.
5. Centrifuge (13,000 rpm) the precipitated DNA at 4°C for 30 min.
6. Wash pellet with 70% Ethanol and spin again.
7. Carefully take off supernatant, allow pellet to dry.
8. Add 5µl deionized formamide (pH 7.0) and incubate tube at 37°C using a thermomixer, shaking at 300-500rpm for 30 min, or until probe has resuspended.
9. Add 5 µl Master Mix, vortex, centrifuge briefly. Your probe is now ready to apply. Make sure you mix again by pipetting well before loading onto coverslip

2.6.2.3 Day 1 – Prepare slides, apply probes

1. Cells that stick well to poly-L-lysine slides (eg. FACS/MACS sorted BM B cells, Thymocytes, Splenocytes) can be prepared as normal in 1x PBS.

- For anaemic spleens (AS) it is best to refilter the cell suspension through a 70µm sieve, as this removes larger debris. Several washes with PBS and refiltering through 40um sieve will help to eliminate crud and get a good single cell suspension, giving cells that stick better to the slides and are less prone to fall off during the pre-hyb steps.
- One AS in ~15 ml buffer will produce a good density of isolated cells in the final product.

Cells that do not stick well to poly-L-lysine can be fixed in solution using Dan's fixation in solution/cytospin methodology (see protocol at end) which is adapted from the Hi-C fixation, then cytospun onto slides.

2. Place 50-200µl cell suspension onto poly-L-lysine slides as you normally would, allow cells to settle for 2 min, gently place slides in tray with 4% PFA in PBS and fix for 10 min.

- Use home made 4% PFA in PBS or standard.
 - For a tray: 20ml 10x PBS, 50ml 16% FA, H₂O -> 200 ml
3. Quench in 155mM glycine for at least 10 min, at RT.
4. Permeabilize cells in 0.1% saponin/0.1% triton X-100 in PBS for 10 min, RT.
- For large coplin:

0.5 ml 20% Tx-100, 5 ml 2% Saponin, 1xPBS -> 100 ml

5. Wash twice in PBS for 5 min each.

6. Incubate for 20 min or longer in 20% glycerol/PBS at RT

- Precipitate probes during incubation

- 10ml 10x PBS, 20ml glycerol, H₂O -> 100ml

- If not using immediately, store slides in 50% glycerol/PBS at -20°C. Can be stored for at least several weeks, probably longer.

7. 3X freezing/thawing in liquid nitrogen.

- Ensure a good depth of liquid nitrogen so that the slides are well covered

- After freezing, lay on tissues to allow thawing.

- Once slide is totally thawed (opaque frozen glycerol no longer visible), place back in 20% glycerol/PBS before subsequent round of freezing.

- Listen for characteristic ‘popping’ sound after 3-4 seconds in liquid nitrogen before withdrawing slide. May not always be audible, if so withdraw after ~4 seconds.

- Process slides one by one, 15 slides can be done comfortably in rotation, i.e. thawing on tissues while the rest are frozen. Wait for all slides to be back in glycerol before starting the next round.

8. Rinse twice in PBS 5 min.

9. Incubate slides for 30 min in 0.1 N HCl at RT

- For large coplin: 840ul conc. HCl, H₂O -> 100ml

10. Equilibrate in PBS for 5 min.

11. RNAase A/T1 (Fermentas) treatment for 60 min with 100ug/ml Rnase A/T1 in 2x SSC at 37oC.

- Begin spin down of precipitated probe mixes at this point.

- Apply 100µl on a 50x22mm coverslip in a FISH box at 37°C

12. Rinse in 2x SSC 5 min then PBS 5 min.

13. Permeabilize in 0.5% saponin/0.5% triton X-100/PBS for 30 min at RT

- For large coplin:

2.5 ml 20% Tx-100, 25 ml 2% Saponin, 1xPBS -> 80 ml

14. Rinse twice with PBS, 5 min each.

15. Equilibrate in 50% formamide/2X SSC for at least 10 min (can mix with pipette to speed this step up)

- 1:1 4xSSC : Formamide

16. To apply probe, first place 10ul probe mix (see later) onto 22x22mm coverslip. Take corresponding slide out of jar, carefully wipe off excess liquid around cell spot(s) and invert coverslip onto cell spot(s)
– try to avoid bubbles and cells drying out. Seal with rubber cement ('fixogum' is best) and allow to dry completely. If using directly labeled probes protect from light at all subsequent stages.

17. Once rubber cement is dry place slides on hot plate set at 78°C for exactly 2 min. Cover with box to protect from light during denaturation.

18. Remove from hot plate and incubate 16+ hours at 37°C in a humidified chamber (eg. FISH box with water-soaked tissues).

Next Day - wash slides

19. Prepare water baths and coplin jars with solutions at correct temperatures.

- Check temperature of solutions not the water baths.

20. Take off rubber cement carefully, place in 2X SSC until coverslips loosen and slide off.

21. Wash in 50% formamide/2X SSC for 15 min at 45°C, once,

22. Then in 0.2X SSC at 63°C for 15 min, once

23. Then in 2X SSC at 45°C for 5 min, once

24. Equilibrate in 2X SSC at RT for 5 min.

25. If using only directly labeled probes proceed to step 27. For antibody detection (eg. hapten probes or immuno-FISH), block for 30 min at RT with 3% BSA in 2X SSC (can also use TSB from RNA FISH protocol)

- Apply on a coverslip and incubate in the same humidified FISH box as used for the hybridisation

26. Apply detection antibodies diluted in 3% BSA in 2X SSC or TSB, incubating at RT for 30-45 min in dark humidified box. 30 min is fine if doing multiple layers – longer time might be important

if doing just one layer.

27. Wash three times for 3 min each in 2X SSC/0.1% TX-100 or TSTX (from RNA FISH protocol) at RT after each detection.

28. Rinse in PBS for 5 min

29. Stain for 2 min with DAPI in 2X SSC at RT in coplin jar

• Jar and solution can be stored at 4°C (protected from light by wrapping in foil) for several weeks and reused

30. Wash in PBS for 10 min to destain DAPI.

31. Fix in 3.7% formaldehyde/PBS for 5 min to fix probes/antibodies in place (this gives a cleaner result and the slides should last longer).

32. Quench in 155mM glycine for at least 30 minutes at RT (longer is better – this quenches autofluorescence from the formaldehyde fixation).

33. Wash in PBS for 5 min to remove glycine.

34. Mount a coverslip with a drop of vectashield or slowfade gold (NO DAPI) on slide area with cells. Seal with nail varnish.

2.6.3 Visualising signals using the Metacyte

Chapter 3

Developing an assay for gene association

3.1 Introduction

To characterise the genome-wide association profiles of the proto-oncogenes *BCR*, *ABL1* and *MLL* I used e4C, a technique based on chromosome conformation capture (3C).

The e4C protocol was developed by Dr. Tom Sexton in the laboratory of Dr. Peter Fraser (Schoenfelder et al., 2010), extended from an earlier methodology called chromosome conformation capture on chip (4C, not to be confused with circularised chromosome conformation capture) (Simonis et al., 2006). Schoenfelder *et al.* used e4C to probe the interactions of the globin genes *Hba* and *Hbb* in mouse foetal liver cells. They used ChIP to purify ligation products associated with elongating RNA polymerase II and identified association partners by hybridising the e4C library to a microarray.

In this project I made a number of modifications to the protocol - I used a different cell type, different bait regions, I removed the ChIP step (see Section 3.5.1) and I adapted the protocol for use

with next generation sequencing analysis instead of microarrays. In this chapter I discuss that process, the challenges it presented and reasons for modifications that I made.

3.1.1 Chromosome Conformation Capture

Chromosome Conformation Capture (3C) is a ligation based proximity assay. Developed in 2002 by Dr Job Dekker and colleagues, it allows the physical association of chromatin to be interrogated (Dekker et al., 2002). 3C has become a key methodology in the field of nuclear organisation and the foundation for a range of variant techniques (reviewed in Osborne et al., 2011).

The 3C protocol begins with formaldehyde fixation used to preserve nuclear structure. A restriction endonuclease is used to cut the DNA, typically recognising a 6 base pair sequence cutting on average once every 1.2 Kb ($6^4 = 1296$). This results in “hairballs” of cross linked DNA and protein which are diluted into a large volume. A ligation reaction is run favouring ligations between sequences held together by formaldehyde linkages. These cross links are reversed and the DNA is purified, resulting in a 3C library (Fig 3.1.1). The frequency with which restriction fragments are found ligated together corresponds to the degree of their interaction within the nucleus.

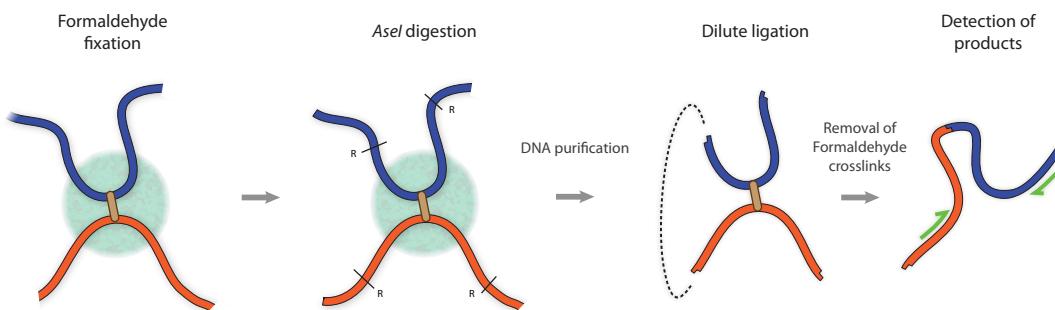


Figure 3.1.1 – Overview of the 3C methodology. R denotes the recognition sites for the first six-cutter restriction endonuclease.

Traditional 3C uses PCR with primers positioned adjacent to the restriction endonuclease recognition sites of two loci of interest. If a PCR product is detected, these two sequences were physically close within the nucleus. The relative strengths of interactions can be assessed by quantifying intensities

of bands on a gel or with RT-qPCR threshold values. In this way, interactions can be studied on a one-to-one basis, investigating the interaction between two candidate loci.

3.1.2 Enriched 4C

A 3C library contains fragments representing the interaction frequencies of all sequences in the genome. Enriched 4C (e4C) purifies all ligation products containing a specific bait (sequence of interest) by using a biotinylated primer with annealing and extension steps. Magnetic streptavidin beads then enrich ligation products containing the bait. A second restriction enzyme is used to cut the unknown sequence at a 4 base pair recognition site (approximately once every 256 base pairs: $4^4 = 256$), giving a region of single stranded DNA used to ligate an adapter. The resulting e4C library has known sequence at both ends of every fragment and can be amplified using PCR (Fig 3.1.2).

To use the e4C library with the Illumina Genome Analyser IIx (GAIIx) next generation sequencing platform, I incorporated a second PCR step using primers with overhanging tails. This adds the sequences needed for cluster generation with the Illumina GA IIx flow cells (Fig 3.1.2). Libraries were then purified and quantified using the Agilent 2100 Bioanalyzer and TaqMan qPCR before being loaded onto flow cells for cluster generation and sequencing. To avoid sequencing the known bait sequence with the standard single-end Illumina sequencing primer, a custom sequencing primer was used which binds to the bait region immediately before the first restriction enzyme cut site. A few remaining few base pairs were left between the sequencing primer and the cut site to use as a barcode region, allowing multiplexing of libraries on the a single flow cell lane (see Section 3.7). This sequence is also used to confirm the identity of the bait sequence. The remaining base pairs beyond the cut site are aligned to the genome and their distribution used to plot a genome-wide interaction profile for the bait sequence.

For discussion of e4C data analysis see Chapter 4.

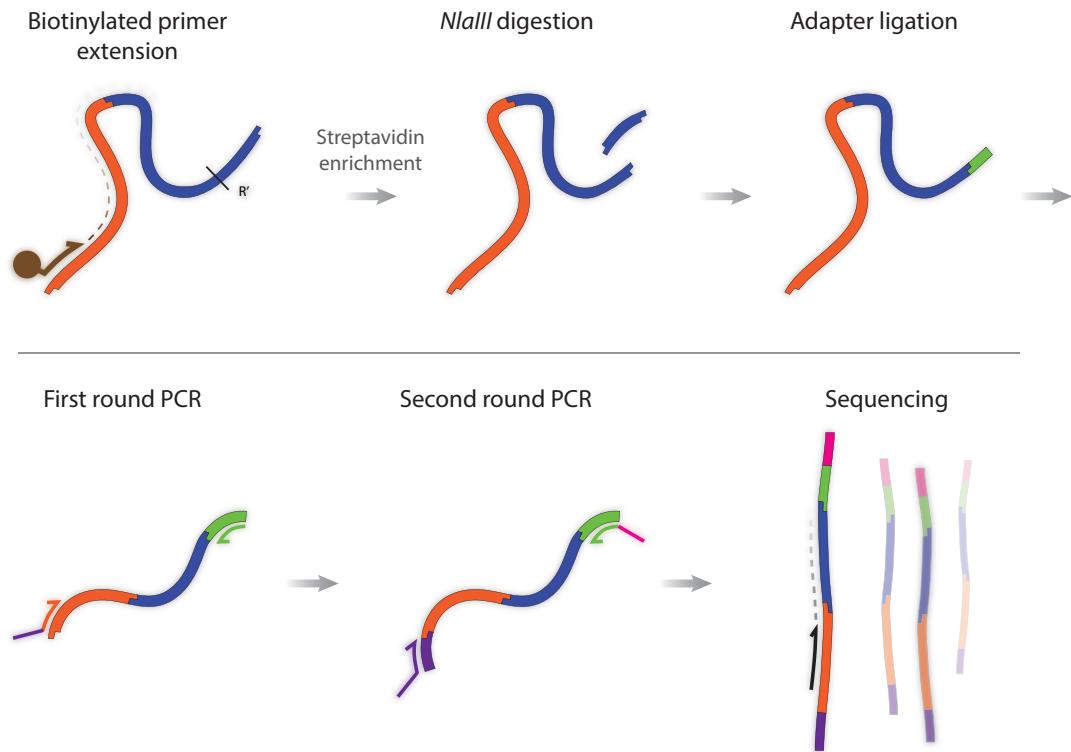


Figure 3.1.2 – Overview of the e4C methodology. R' denotes the recognition site for the second four-cutter restriction endonuclease. BC denotes the barcode region used to confirm the identity of the bait sequence and identify sequences when multiplexing libraries for sequencing.

3.2 3C Restriction Enzyme Choice

The first step in designing an e4C experiment is the choice of a bait region and six-cutting restriction endonuclease. This enzyme defines the 3C fragments to be ligated and the position of the bait sequence to be enriched.

The majority of chromosomal translocations within the *BCR* and *MLL* genes occur within tight breakpoint cluster regions. The *MLL* breakpoint cluster region is an 8.3 Kb region between exons 6 and 14 delimited by two *BamHI* restriction sites (Gu et al., 1994), making this a natural choice for the primary 3C restriction endonuclease.

After a number of test e4C runs using human buffy coat cells it became clear that the assay was not working to its full potential. When run on a gel e4C libraries should appear as a smear due to the

range of different restriction fragments ligated to the bait sequence (eg. Mouse *HindIII* e4C library, Fig 3.2.1). Bands are sometimes visible due to the germ line sequences being the most common ligation products, removed at a later state by digestion with a rare-cutting restriction endonuclease. In my human *BamHI* e4C libraries the germ line band was visible but little or no smear was present (Fig 3.2.1).

The lack of a smear could be the result of a number of factors - inefficient PCR amplification; loss of library through purification steps; inefficient enrichment of bait; low ligation efficiency in 3C or adapter ligation or poor digestion in initial 3C and e4C four cutter digestion steps. I ran a number of diagnostic tests to identify the cause, including a digestion efficiency assay which quantifies the amount of PCR product produced by primers spanning a restriction endonuclease recognition site (see Section 2.3.3 for method). This showed that only 44% of *BamHI* recognition sites were being cut in the 3C digestion step. Typically the digestion efficiency for this digestion step should be greater than 80% (Cope and Fraser, 2009).

The steps immediately before the *BamHI* digestion in the protocol are SDS treatment to permeabilise the fixed nuclei to aid access for the restriction enzyme, followed by treatment with Triton-X100 to quench the SDS. To test if the SDS was inhibiting *BamHI* digestion, I digested purified plasmid DNA after SDS treatment and a range of Triton-X100 concentrations (Fig 3.2.3 A). This showed that *BamHI* digestion is inhibited if the SDS is not adequately sequestered. To test the effect of an increased Triton-X100 concentration in fixed nuclei I prepared human buffy coat cells as described in the e4C

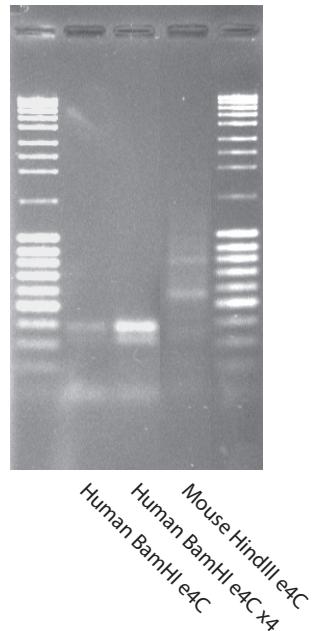


Figure 3.2.1 – *BamHI* e4C Libraries. Test e4C library prepared from human buffy coat cells using *BamHI* as the primary 3C restriction endonuclease. Lanes correspond to the test library, 4x the concentration of the same library loaded for extra visibility and a positive control mouse e4C library with *Myc* bait prepared with *HindIII*.

protocol (Section 2.3) and ran *BamHI* digestions after treatment with SDS and two Triton-X100 concentrations (Fig 3.2.3 B). The nuclei treated with the higher concentration of Triton-X100 showed a higher digestion efficiency as shown by the lower smear in (Fig 3.2.3 B), though this digestion is still not satisfactory for e4C.

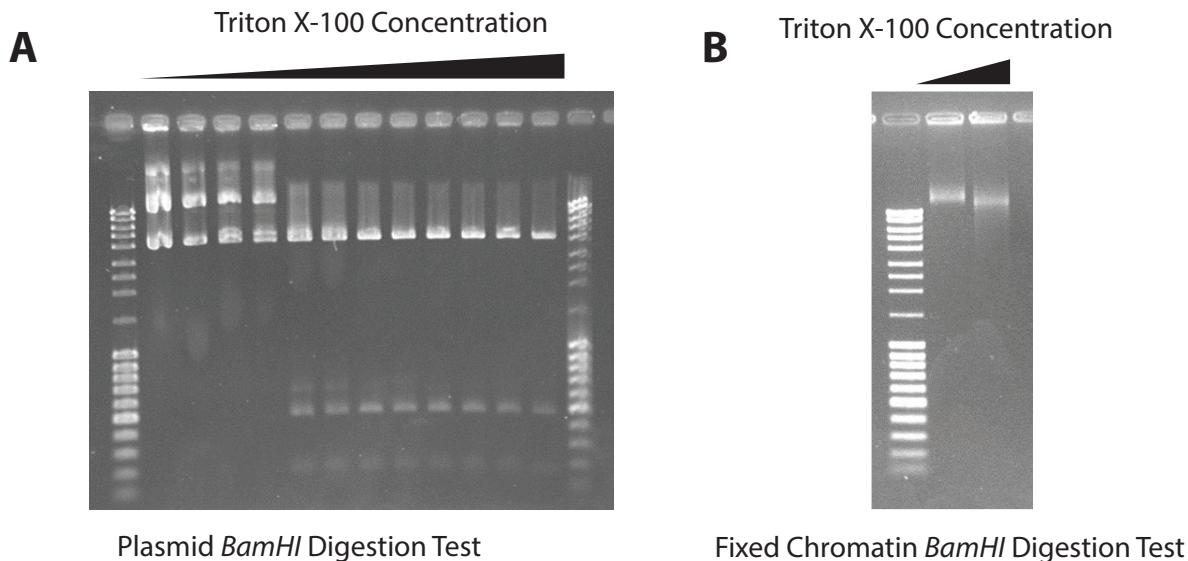


Figure 3.2.2 – Effect of Triton-X100 on *BamHI* digestion. *BamHI* digestion efficiency of purified plasmid DNA with increasing Triton-X100 concentrations after SDS treatment. *BamHI* can be seen to be inactive at the lower four Triton-X100 concentrations.

To compare the digestion efficiencies of *BamHI* against a number of other restriction enzymes commonly used for 3C in the literature, I ran another test 3C with human buffy coat cells varying the degree of formaldehyde fixation and trying the removal of the SDS / Triton-X100 steps (Fig 3.2.3 C). *BamHI* showed markedly improved digestion for nuclei fixed in 1% formaldehyde without SDS or Triton-X100, as shown by the extended smear in the bottom right gel. However, restriction enzymes *AseI*, *HindIII* and *PvuII* all showed far superior digestion efficiency in all conditions. I chose to use *AseI* for future e4C experiments because it produces suitable fragment sizes around the break point cluster regions of both *BCR* and *MLL*. Further digestion efficiency qPCR tests showed its efficiency at a far superior 81 - 99% of sites being cut. Completion of test e4C libraries using *AseI* showed a gel smear comparable to that shown in Fig 3.2.1 for the mouse *HindIII* e4C library (data not shown).

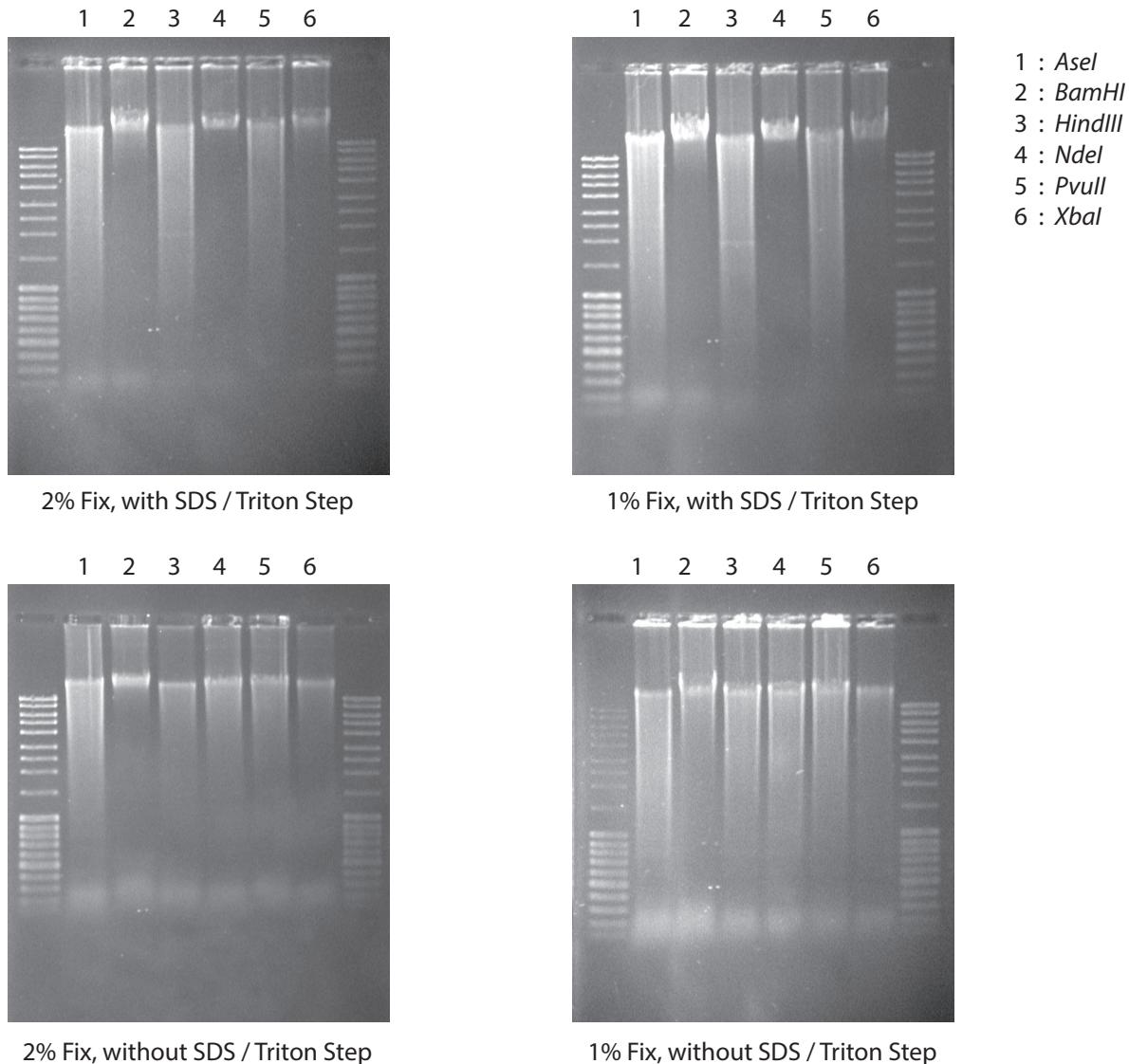


Figure 3.2.3 – Restriction Enzyme Tests. *BamHI* digestion efficiency of fixed human buffy coat nuclei with increasing Triton-X100 concentrations after SDS treatment. The lower smear at the higher Triton-X100 concentration indicates an increased digestion efficiency. **(C)** Digestion tests with six enzymes in four different conditions. Cells were fixed in either 1% or 2% formaldehyde as described in Materials and Methods: 3C (Section 2.3), with or without the pre-digestion SDS / Triton-X100 steps. Samples were treated with proteinase K, RNase A and cleaned with a phenol / chloroform extraction and precipitation, as described in Section 2.3. All samples shown were run on 1% gels with a Fermentas MassRuler™ DNA Ladder Mix; 80 - 10,000 bp.

3.3 Primer Design

3.3.1 Paired end sequencing or single end sequencing?

When I started my project, the intention was to use Illumina Paired End Sequencing to sequence the libraries. I designed primers and adapters so that the first read would be from the *NlaIII* adapter end of ⁷³

the e4C fragment and would give sequence information about the unknown, captured fragment. The second read of the paired end run would then go from the bait end of the fragment, and confirm the identity of the e4C bait region. However, Dr Cameron Osborne was sequencing mouse e4C libraries and ran into problems with this approach: read lengths were not sufficient to sequence through the entire captured region and into the bait sequence on the first round of sequencing, so we could not be certain that the first read was not from a ligation concatemer and not directly ligated to the bait sequence.

At this time, the sequences of the Illumina sequencing primers were made public, so we decided to design our own sequencing primers. This allowed us to use single end sequencing runs which start very close to the ligation junction of the e4C products, confirming the identity of the bait product but giving us enough sequence in the unknown partner to be aligned to the genome. Sequencing primers were designed to be as similar in their length, T_m and GC content as the official Illumina Sequencing Primer 1.0 (Table 3.3.1).

	Length (bp)	T_m	GC Content
Illumina Sequencing Primer 1.0	33	77.4 °C	51.5%
BCR Sequencing Primer	38	73.4 °C	36.8%
MLL Sequencing Primer	35	74.2 °C	34.3%
ABL Sequencing Primer	32	77.9 °C	46.9%
MLL-p1 Sequencing Primer	34	76.1 °C	41.2%
MLL-p2 Sequencing Primer	40	74.7 °C	42.5%

Table 3.3.1 – e4C sequencing primer properties. Custom primers designed to be as close to the official Illumina sequencing primer as possible in length, T_m and GC content.

3.3.2 Bait specific primers

A number of bait-specific primers need to be designed for each e4C library: an extension primer, a first round PCR primer with an overhanging tail consisting of half the Illumina Paired End Adapter 1.0 (PE Ad 1.0) and a sequencing primer complementary to the library bait sequence. The sequencing primer must anneal close to the restriction endonuclease cut site but leaving a barcode sequence of at least 3 base pairs. This bar code region is important - it allows us to multiplex e4C libraries (see

Section 3.7) and confirms that the sequenced fragment is the expected bait - it is not present in any primers so must result from original genomic sequence.

Primers unaffected by the bait region are also used: a first round PCR primer complementary to the *NlaIII* adapter and second round PCR primers to generate the full Illumina cluster generation sequences. These bind the *NlaIII* adapter and the overhanging sequence produced by the nested bait specific primer in the first round PCR (see Fig 3.1.2).

When designing these primers, a suitable bait fragment must be chosen; for my *MLL* and *BCR* e4C libraries I picked *AseI* fragments which lay within the breakpoint cluster regions (hg19 chr22:23,596,608-23,612,941 and chr11:118,354,804-118,356,061, respectively). Bait fragments should not be excessively large or small, as this can affect the ligation efficiency in the generation of the 3C library. One end of the bait fragment is used for the e4C, this must be free of repetitive DNA, and have enough room to design primers before the first *NlaIII* restriction site.

3.4 Preparation of CD34⁺ cells

It is thought that translocations involving *BCR* and *MLL* in leukaemia occur within the a progenitor cell compartment (Sutherland et al., 1996; Bonnet and Dick, 1997). The CD34 antigen is a marker for primitive blood- and bone marrow derived progenitor cells and anti-CD34 cell sorting is commonly used to study blood progenitor cells (Khobta et al., 2004; Libura et al., 2008).

Samples were collected from consenting leukaemia patients undergoing GCSF treatment to mobilise progenitor cells into the bloodstream. Care was taken to only take samples from patients with cancers not affecting haematopoietic cells. To best preserve the nuclear organisation of the nuclei I fixed them with formaldehyde as quickly as possible; immediately after a Ficoll separation to isolate the lymphoblasts but before sorting with anti-CD34⁺ beads. Finally, I analysed an aliquot of the purified cells using FACS (using an antibody against a different epitope of CD34) to determine the purity of the samples.

	Germline	TGTTTCTGCAGCACAGAGTTGGCAGAACCTCAGAACCTCTGTCTGTATTGCTTTGTCAGAAGCTTAGTACCA <u>GGAAATTAAT</u>
BCR	Biotinylated	^o TGTTTCTGCAGCACAGAG
	Nested	ACACTCTTCCCTACACGAGCCTCTCCGA<u>T</u>GTCTGTATTGCTTTGTCAG
	Sequencing	TGTCCTGTTATGCTTGTTCAGAAGCTTAGTACCA
	Germline	GAACAAATCACACCCATTGCGTTCACATTGCAA <u>CAATTTATAAT</u>
MLL	Biotinylated	^o TCACACCCATTGCGTTCA
	Nested	ACACTCTTCCCTACACGAGCCTCTCCGA <u>T</u> GTCTGTATTGCTTTGTCAG
	Sequencing	CGATCTGCCAACAGATAATAATGCAA <u>ATGCAAAT</u>
	Germline	A AAGGAGCCAGAGCAACTCAGGAATTGGTTATTGTTGGAGATTAGGG <u>GGAAATTATAAT</u>
MLL 1	Biotinylated	^o AAGGAGCCAGACAGTC
	Nested	ACACTCTTCCCTACACGAGCCTCTCCGA <u>T</u> GTCTGTATTGTTGGAGATTAGGG
	Sequencing	CGATCTGCCATTGAGATTAGGG <u>GAAATTATAAT</u>
	Germline	ACAGCAGCACCCRTCTGTTGACTAGAAGTTAGGCTGGTTGAATAATAAACCCCTAGGG <u>GTCTCGGTATTAAAT</u>
MLL 2	Biotinylated	^o ACAGCAGCACCCRTCTGTTGACTAGAAGTTAGGCTGGTTGAATAATAAACCCCTAGGG <u>GTCTCGGTATTAAAT</u>
	Nested	ACACTCTTCCCTACACGAGCCTCTCCGA <u>T</u> GTCTGTATTGTTGGAGATTAGGG
	Sequencing	CGATCTCCCCCTGAAACAGTGTTTATCAGTA <u>ACTAGTCTCGGTATTAAAT</u>
	Germline	CACTGCGCCTGGCTAGTTTCACTGGACA <u>ATACCTTGATGAAGGAGCAAAATPATTAAAT</u>
ABL	Biotinylated	^o TCACTGGACAA <u>TACCTTGATGA</u>
	Nested	ACACTCTTCCCTACACGAGCCTCTCCGA <u>T</u> GTCTGTATTGTTGGAGACGAAAA
	Sequencing	GCTCTTCCGATCTCTTGATGAAGGAGCAAA
	Nla ^{II} e4C adapter	5' §AGATGGAAAGGCGGTTCAGCAGGAATGCCGAG 3' 3' G*TACTCTAGCCTTCTGCCAA <u>G</u> GTCTACGGCCTACGGC 5'
Constant	PE Ad 1.0 Primer	
	PE Ad 2.0 Primer	
	Illumina sequencing	ACACTCTTCCCTACACGAGCCTCTCCGA <u>T</u> CT

Table 3.3.2 – e4C primers. Germline shows bait locus, barcode region in bold pink, *Ase*/*I* restriction site show underlined. Biotinylated primers have biotinylation moiety on 5' base. Nested primers are comprised of two sections - black is complementary to bait region for hybridisation, green is the tail region to incorporate the first half of the Illumina PE Ad 1.0. Sequencing primer shown in blue. Constant primers show *Nla*^{II}/e4C adapter (ordered as two oligos and hybridised before use) and second round PCR primers used to incorporate full Illumina PE adapters. Illumina sequencing primer shown as reference (not used for any e4C sequencing). ^o denotes a biotinylation moiety. § denotes a phosphorylation modification. * denotes a phosphorothioate modification.

3.4.1 CD34⁺ sorting purities

The first kit I used to separate the fixed CD34⁺ cells was the Invitrogen Dynal CD34 Progenitor Cell Selection System. I used this to process 17 blood samples. The FACS plots showed some material with low front scatter, likely to be platelets or debris resulting from the fix and sort (Fig 3.4.1). I used gating within the FACS analysis software to ignore these fragments. To determine CD34⁺ purity, I used an empirically determined threshold for anti-CD34⁺ fluorescence. This varied from 17% - 91%, with a median purity of 67%. 10 of 17 samples collected using the Invitrogen kit had to be discarded.

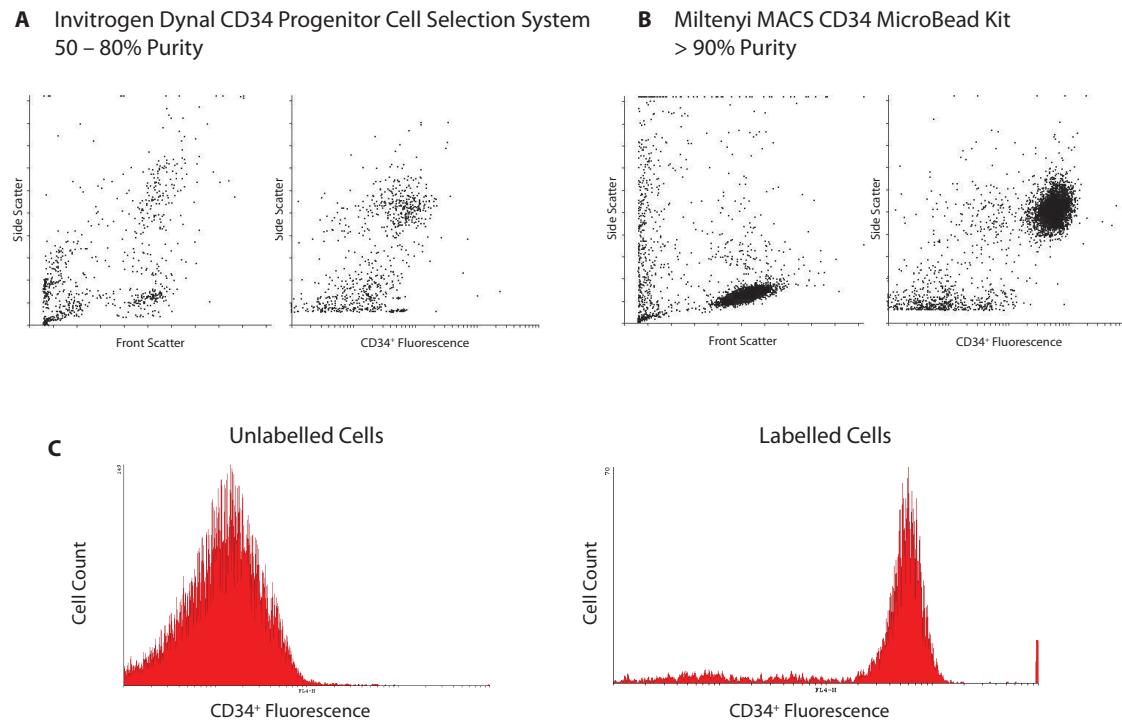


Figure 3.4.1 – CD34⁺ Separation FACS Plots. Diagnostic FACS plots after magnetic bead separation of fixed CD34⁺ cells using the (A) Invitrogen Dynal CD34 Progenitor Cell Selection System and the (B) Miltenyi MACS CD34 MicroBead Kit. Upper right quadrant of Front vs. Side Scatter plot in (A) shows heavy granulocyte contamination. (C) Histogram plots of anti-CD34 fluorescence on an unstained control and stained sorted cells. A threshold was set on the left of the stained peak to determine CD34⁺ purity. All FACS plots generated using a BD FACSCalibur flow cytometer with Miltenyi Biotec anti-CD34-PE antibodies.

I attempted a number of alterations to the cell separation protocol to improve this purity, one of which was the use of the Miltenyi MACS CD34 MicroBead Kit. This immediately gave much better

performance, with purities ranging from 90% - 97% with a median of 96% purity (typical FACS plot shown in Fig 3.4.1). All data described in this thesis is derived from cell collections using the Miltenyi kit with CD34⁺ cell purities above 92%.

3.5 e4C with low cell numbers

Initially, CD34⁺ cells were collected from peripheral blood supplied by Dr. George Follows in the Addenbrookes Haematology Day Unit, with the help of apheresis coordinator / specialist nurse Paul Boraks (see Section 2.1.1 for details). Samples were collected on a weekly basis before being fixed and sorted as described in Section 2.1. Typically, between 12 and 30 ml whole blood was received in each sample, giving a median final count of 3.6×10^5 CD34⁺ cells per collection. Typically, 3C experiments use 1×10^7 cells, equivalent to 28 collections pooled together. Between 2008 and 2009 I managed to collect and sort 23 peripheral blood samples, giving a total of 1.04×10^6 total cells, however 10 of these samples had to be discarded due to poor purity (see Section 3.4). This left me with 13 samples and a total of 7×10^6 cells (minimum 76% purity, average 91.6% purity). Because of this low cell number I worked on optimising the 3C protocol for use with low cell numbers. Some of this work was done with the help of Dr. Mayra Furlan-Magaril, who was a visiting student in our laboratory at the time.

3.5.1 ChIP e4C

I had originally hoped to incorporate a chromatin immunoprecipitation (ChIP) step to purify 3C complexes containing RNA polymerase II, enriching for chromatin interactions at transcription factories. This protocol was used with some success in my lab by Dr Cameron Osborne and Allen Chong, though the immunoprecipitation step meant that little 3C material was available for the e4C stage resulting in interaction datasets that were very low in complexity. This was offset to some degree by the specific nature of the hits and low signal to noise ratio (unpublished data).

Although the chromatin immunoprecipitation steps worked well in my hands with test mouse B

cell samples, I soon decided to remove this enrichment step in favour of keeping as much library DNA as possible due to the scarcity of my CD34⁺ samples. I showed by RT-PCR that my bait genes were expressed in CD34⁺ cells (Figure 3.5.1). All interactions may be relevant to the formation of chromosomal translocations, so any resulting interaction data is useful.

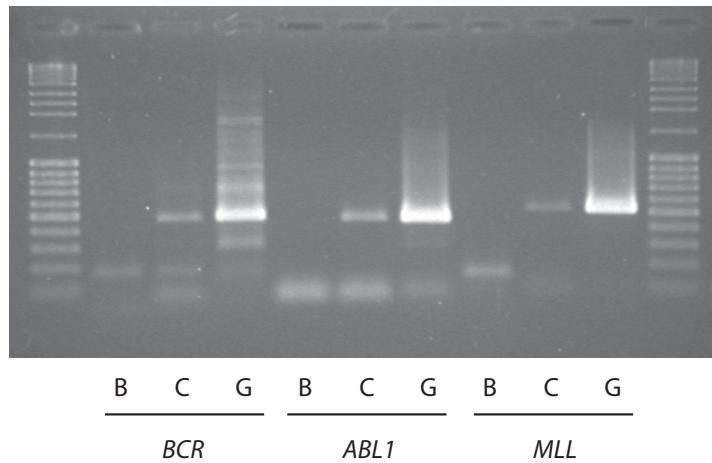


Figure 3.5.1 – CD34⁺ RT-PCR. B = Buffy Coat cells, C = CD34⁺ cells, G = genomic DNA control. CD34⁺ cells not treated with DNasel due to low cell numbers, hence genomic band contamination.

3.5.2 Carrier e4C

An established technique when using low cell numbers for methods such as ChIP and MeDIP is to use 'carrier' chromatin to buffer loss of sample DNA (O'Neill et al., 2006; Ficz et al., 2011). This is done by processing cells or DNA from another organism in with the sample, giving larger quantities of DNA to work with. Loss through non-specific binding to plastic such as tubes and pipette tips has less of an effect on the sample DNA and steps such as precipitations are aided by having larger pellets to work with.

I started using this technique with the Sf9 cell line derived from *Spodoptera frugiperda* (Fall Armyworm) obtained from Maureen Hamon at the Babraham Institute. Cells were processed in parallel with the sample as described in Section 2.3, up until the point of permeabilisation. Sf9 cells were then split into two aliquots of 1×10^7 cells, and 1×10^6 sample cells were added to one aliquot. Both samples were then processed as normal for the rest of the 3C protocol.

After this initial test run, I switched to using the Schneider's line 2 (S2R+) cell line derived from *Drosophila melanogaster*, grown with help from Dr. Sarah Toscano at the Babraham Institute. The reason for this change was because the genome of *Drosophila melanogaster* is known, allowing easy design of PCR primers to test for carrier DNA contamination and allowing identification of any products occurring as a result of a non-specific ligation events in the e4C.

Initial tests with the carrier e4C were promising, showing good recovery of sample DNA and successful 3C product detection using primers designed for the *Drosophila melanogaster* genome (Fig 3.5.2). However, due to the removal of the removal of the RNA polymerase II ChIP step (Section 3.5.1) and subsequent acquisition of much larger CD34⁺ cell collections (Section 3.6), the carrier e4C protocol was not needed.

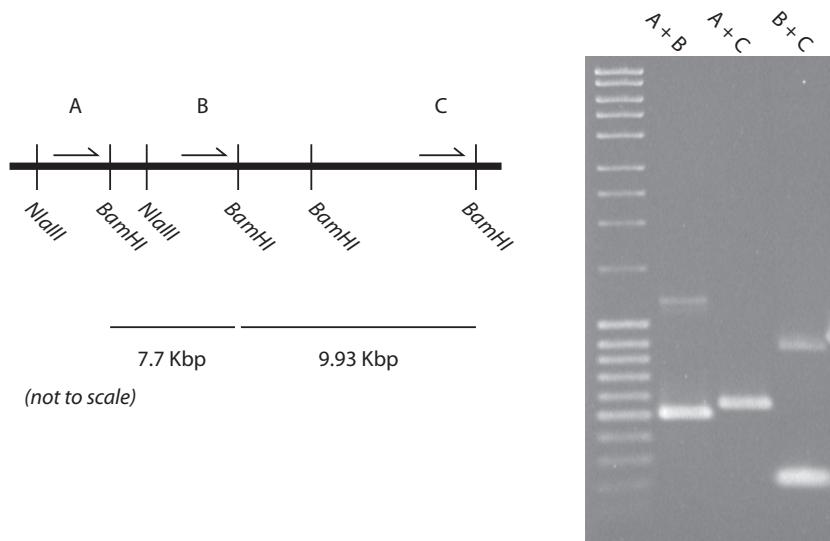


Figure 3.5.2 – *Drosophila melanogaster* S2R+ 3C Tests.

3.6 e4C with large cell numbers

In December 2009 I was able to obtain a leukapheresis sample from Dr. George Follows with the help of Kevin Jestice, the chief biomedical scientist within the Haematology Department at Addenbrookes Hospital, Cambridge. This sample came from a patient who had responded exceptionally well to G-

CSF treatment and had enough mobilised peripheral CD34⁺ cells that there was leukapheresis sample spare after use by the Haematology Department. I processed this sample as described in Section 2.1.2. The final yield of CD34⁺ cells from this single collection was 2.88×10^7 cells, over four times the number of cells collected from a year of weekly peripheral blood CD34⁺ sample collections. These collections also had the added benefit of coming from a single patient, so not risking any population variation we may see within the interaction datasets. In a subsequent two leukapheresis collections 3.99×10^7 and 1.4×10^8 CD34⁺ cells were collected, with an average purity of 95.3%.

The first CD34⁺ collection was used to generate a 3C library which was used as the basis for the e4C libraries. The second CD34⁺ collection was used to generate 3C material used in the 3C RT-qPCR validation (Section 6.5). The third CD34⁺ collection was used to generate *HindIII* 3C material by Alice Young, a fellow PhD student in my group. She went on to create Hi-C libraries with this material.

3.7 Multiplexing e4C libraries

Although the cost of next generation sequencing is rapidly decreasing, sequencing an Illumina library is an expensive process. In order to make libraries cheaper, multiplexing is a common practice. All multiplexing uses the same principle: Illumina libraries are created with a small number of base pairs constant within each sequence - a barcode. Multiple libraries with different barcodes are mixed and simultaneously sequenced in a single flow cell lane. After the sequences are retrieved from the sequencing run, the libraries are separated by identifying the barcode present at within each sequence. The greater the number of libraries which are multiplexed within a single lane, the fewer the reads retrieved for each library. The total number of returned sequences is often in excess of what is needed, as in the case of my e4C libraries where a great number of duplicates are returned, showing that the full depth of the e4C library has been sequenced.

A number of variants of multiplexing have been used with the Illumina Genome Analyser IIx and Illumina itself produces a kit for the production of multiplexed libraries which uses a short paired-

end read. Because I was sequencing my e4C libraries with custom sequencing primers, I was able to specify the sequence present at the start of each read. As such, I designed the sequencing primers to leave a few base pairs for sequencing before the *AseI* restriction site. This has the dual benefit of allowing the multiplexing of multiple e4C libraries with different baits, and ensuring that the sequenced products are not the result of unspecific primer binding; genomic sequence is used as the barcode which is not incorporated with any primers used.

3.7.1 Multiplexed libraries

3.7.1.1 *BCR* Run 1 and *MLL* Run 1

For the first sequencing runs with my e4C libraries, I multiplexed two libraries: a CD34⁺ *BCR* breakpoint region bait and a CD34⁺ *MLL* breakpoint region bait. These libraries are known as *BCR* Run 1 and *MLL* Run 1 for the rest of this thesis. Once the raw reads were available from the sequencing run, we separated the two e4C libraries *in-silico* (for details, see Section 4.2). Despite loading equimolar amounts of both libraries and sequencing primers onto the flow cell, the *MLL* library only returned 55,868 reads vs. the *BCR* library with 22,390,498, a 400-fold difference.

3.7.1.2 Reasons for the bias

There are several possible reasons for such a bias occurring with the multiplexed libraries. Quantitation of the libraries could be inaccurate with DNA present without Illumina sequencing adapters. Similarly, there could be a problem with the *MLL* library causing it to bind to the flow cell or create clusters inefficiently. Both of these seem unlikely however, as the e4C libraries were quantified using quantitative real-time PCR using fluorescent TaqMan probes which only fluoresce when bound to the Illumina adapter sequences. Additionally, the *MLL* Run 2 e4C library that was sequenced without multiplexing on its own lane generated clusters as expected by its concentration and returned the expected number of reads. Differences in the sequencing chemistry related to the sequencing primers could cause the *BCR* library to sequence more efficiently than the *MLL* library, but that would presumably not give a

greater number of *BCR* library clusters, generated using the same Illumina adapters for both libraries. The total number of reads returned from the multiplexed lanes were equivalent to those returned from single library lanes, however, indicating that there were more *BCR* library clusters than *MLL*.

3.7.1.3 Equalising the bias: *BCR* Run 2 and *MLL* Run 3

For my second multiplexed run, I changed the relative concentrations of *BCR* and *MLL* libraries and sequencing primers loaded on the flow cell in order to return an equal number of reads. This time, the *MLL* library returned 9,960,249 reads and the *BCR* library returned 12,616,122 reads, only a 1.2 fold difference.

3.7.2 Cross over products

Upon initial analysis of the multiplexed *BCR* Run 2 and *MLL* Run 3 e4C libraries, it became clear that a significant proportion of the reads present within the *MLL* library were crossover products from the *BCR* library. There is significant enrichment around the *BCR* bait locus (Figure 3.7.1A) with a familiar drop off interaction associated with *cis* linkage. Interestingly, an equivalent enrichment around the *MLL* bait locus was not visible in the *BCR* e4C library (Figure 3.7.1B). Crossover reads typically had few duplicates, whereas suspected genuine reads were present in high numbers. A similar effect was seen in mouse ChIP-e4C libraries prepared and sequenced using different baits (unpublished data).

We returned to the original raw crossover reads for verification and these crossover reads definitely start with the *MLL* barcode sequence before proceeding directly into the bait sequence, so there was no error in the *in-silico* separation of the libraries.

3.7.2.1 *MLL* Run 3 phiX

To confirm that the reads are definitely as a result of the multiplexing and not due to contamination during e4C library preparation, we spiked an aliquot of the *MLL* Run 3 e4C library into the phiX control lane on a subsequent Illumina sequencing run. Although competition with the standard

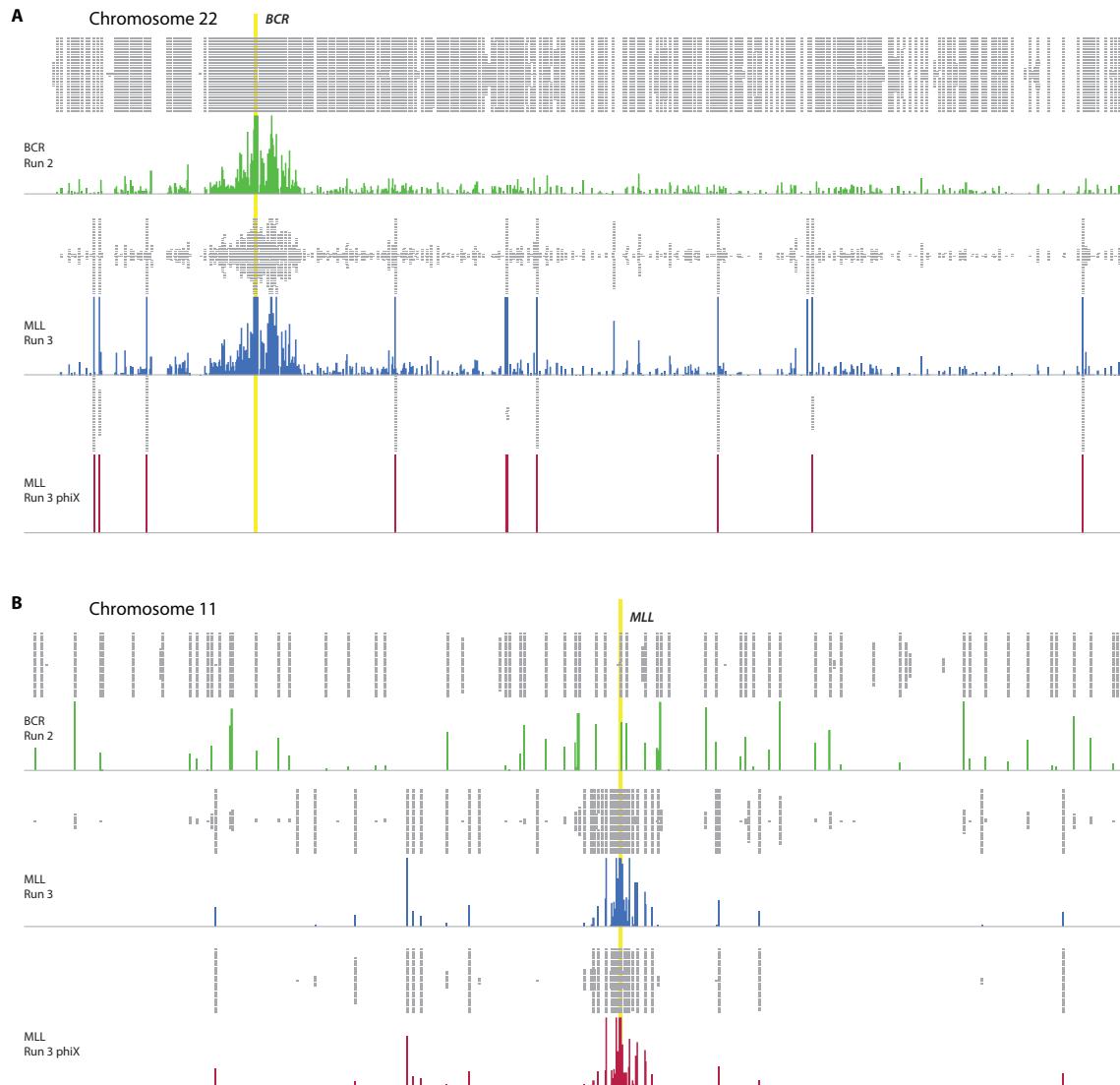


Figure 3.7.1 – Crossover reads between multiplexed e4C libraries. (A) 35 Mb region of chromosome 22, showing enrichment around the *BCR* bait for both the *BCR* and multiplexed *MLL* e4C libraries. There is no enrichment in the phiX *MLL* control library. (B) 35 Mb region of chromosome 11, showing enrichment around the *MLL* bait for the two *MLL* e4c libraries, but not the multiplexed *BCR* library. Bait regions shown by yellow bars.

Illumina sequencing primer meant that the majority of *MLL* library reads were of the bait sequence, enough reads were returned to carry out some analysis. As can be seen in Figure 3.7.1, the *MLL* phiX reads correlate precisely with peaks of large read numbers in the multiplexed *MLL* library but not any suspected crossover reads.

After removing the two bait chromosomes 11 and 22 and removing all duplicate reads, I analysed the hit *AseI* fragments between libraries. The multiplexed *MLL* library had a correlation of 0.656 with the multiplexed *BCR* library and a correlation of 0.36 with the *MLL* phiX library. If a threshold of 20 reads was set before removal of duplicate reads, these two correlations became 0.009 and 0.782, respectively. This demonstrates that only a small number of reads are crossing over between libraries, as can be seen in Figure 3.7.1. Unfortunately this threshold was chosen empirically and could lead to inaccurate analysis results. As such, the thresholded library was not used for further analysis.

3.7.2.2 Reasons for crossover reads

It seems likely that these crossover reads result from an error in the cluster identification during the Illumina sequencing process. We approached Illumina with our observations but received no satisfactory response. Because we could not address this problem without a known cause, we did not multiplex any further e4C libraries.

3.8 Increasing e4C library complexity

Due to the low coverage of the e4C libraries (see Section 7.2.1 for discussion), I tried to modify the e4C protocol to increase the diversity of the sequencing reads.

3.8.1 Barcoded *NlaIII* adapter

In an attempt to retain quantitative data about the frequency of 3C ligation products before the PCR steps, I designed a new *NlaIII* adapter which included 4 base pairs of unspecified sequence in the oligo to use as a barcode. The idea behind this new adapter was to use it in a paired end Illumina sequencing run. Read one would use the same custom sequencing primer and methodology as that described above for the single end sequenced e4C libraries. The second read would use the standard Illumina sequencing primer two, which would read through the barcode and into the unknown sequence. The barcode could be used to deconvolute quantitative information about the number of pre-PCR 3C

ligation products, and the unknown sequence would provide additional validation of the ligation product. Every pre-PCR 3C product should give two different barcodes from the two strands of random barcode, which could then be quantified to give ligation frequencies and greater coverage information. This approach should allow the e4C data to be treated quantitatively.

In order to ensure that the two single stranded oligos would hybridise to form a functional adapter, I added a 6 base pair GC-clamp after the barcode before the *NlaIII* sticky end (Fig 3.8.1). After the first round of PCR, the barcode will have a complementary sequence, so the initially different bases should not cause any problems in the sequencing.

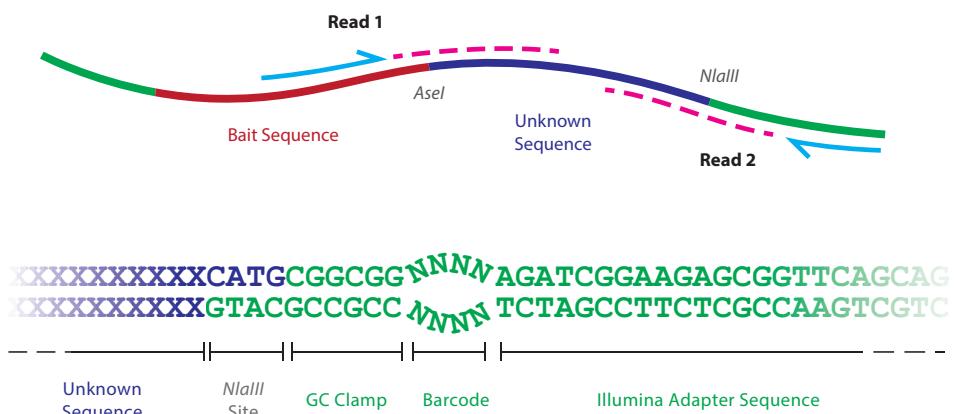


Figure 3.8.1 – Diagram showing design of the barcoded *NlaIII* adapter. Top panel shows an overview of a typical e4C library product. Read 1 confirms the identity of the bait region with a few bases of germ line sequence before reading through the *AseI* restriction endonuclease recognition site into the unknown ligated sequence. Read 2 reads the random adapter barcode before the GC clamp, *NlaIII* restriction site and unknown sequence.

3.8.2 Multiplexing same-bait e4C libraries

The high number of duplicates seen for every read in the e4C libraries (Section 5.2.1) suggests that there is a lack of complexity in the e4C libraries, and we were sequencing the full depth of the libraries. In an effort to increase the diversity in a sequencing run, I created eight e4C libraries using the *BCR* breakpoint region bait and the new *NlaIII* adapter described above. The e4C libraries were prepared as biological replicates from new CD34⁺ 3C material generated from a recent patient sample.

Whilst multiplexing e4C libraries with different bait regions in a single Illumina lane caused

crossover reads which skewed downstream analysis, any crossover reads between libraries sharing the same bait would not affect analysis. Combined with the new *NlaIII* adapter able to identify unique 3C ligation products, such multiplexing should not suffer the same problems as previous runs.

3.8.3 Results of e4C modifications

In order to test the e4C protocol modifications described above, I spiked a small amount of the multiplexed e4C libraries with the new *NlaIII* adapter into the phiX control lane on an Illumina GAIIX sequencing run. Despite loading a high concentration of the custom sequencing primer, a large proportion of the reads came from the standard Illumina sequencing primer and consisted of *BCR* bait sequence. 111,897 clusters were returned with the expected read 1. Of these clusters only 2,541 (2.27%) had a second paired end read with the expected structure (barcode, GC clamp, *NlaIII* site, unknown). The remaining reads were instead either *NlaIII* site > GC clamp > *NlaIII* site (4,394 - 3.93%); *NlaIII* site but no clamp (102,284 - 91.41%); or something else (2678 - 2.39%).

To try to find out why the paired end reads did not conform to the expected structure, we analysed all paired end reads from bait clusters irrespective of their sequencing primer in read 1. Over 2.7 million began with the *NlaIII* recognition site CATG. Of these, the majority had very similar sequences. Most of these were not mappable to the human genome and a number of the highly represented reads mapped to the *NlaIII* adapter. The presence of these reads suggests that there were problems with either the oligo hybridisation used to anneal the *NlaIII* adapter or the ligation reaction used to ligate the adapter to the 3C library. Both of these problems could be exacerbated by the presence of the non-homologous barcode region.

Due to the problems with the barcoded adapter, the second read from the paired end run was not reliable. The low numbers of reads from read 1 meant that this e4C sequencing run did not yield any useful information in its own right. Because of time and money constraints I did not continue the development of this modification. Had I the chance, I would use one *NlaIII* barcode per e4C library. This would hopefully remove problems caused by non-homologous sequence and although losing the ability to differentiate between individual ligation products it would keep the ability to differentiate

between separate e4C libraries multiplexed on a single flow cell. This approach would increase both coverage depth and give some degree of quantitative information in the association data.

Chapter 4

Developing the analysis of e4C data

4.1 Introduction

At the time that my first e4C libraries were sequenced the method was novel. Unlike RNA-Seq and ChIP-Seq which are established enough to have accepted analysis pipelines, e4C data analysis had not been developed. As such I spent a great deal of time discovering how to analyse the data in a way that is accurate and informative.

4.2 Initial data handling

All of the e4C libraries discussed were sequenced using the Illumina Genome Analyser IIx. Initial data processing up to the point of a SeqMonk library was done by the Babraham Bioinformatics team, primarily Dr. Felix Krueger and Dr. Simon Andrews, in discussion with myself.

4.2.1 Bareback processing

Illumina next generation sequencing machine such as the Genome Analyser IIx 'sequence by synthesis'. Libraries are hybridised to oligonucleotides on a chip and clusters are generated surrounding each library sequence with multiple rounds of amplification. Fluorescent bases are then added one by one, and the chip imaged each round. The colour information collected in the image can be used to determine the base being added to the cluster according to its colour, and so the sequence of each cluster is determined.

Central to this technique is the process of calling cluster locations, typically done by the Illumina Sequence Control Software (SCS) with Real Time Analysis (RTA) once the first fluorescent base is added. Once determined, the cluster positions are used for the remaining base pair calling. If the cluster density on the chip is large, then spots can start to merge. This isn't a problem with a typical Illumina library, as the different spots will usually be different colours, and so discernible from each other. However, every sequence within e4C libraries begins with a barcode region followed by the restriction enzyme recognition sequence. This lack of diversity within the first bases of the sequence can cause problems for the cluster calling; merged clusters may be called as one which can be thrown out by the purity filter because of its size, or rejected later when the sequence diverges and it starts to exhibit mixed fluorescence signals. This has the effect of a vastly reduced number of reads being processed (Fig 4.2.1). The same effect can happen in multiplexed libraries using barcodes to identify different samples on the same chip.

This problem was first experienced in our institute with Dr. Cameron Osborne's e4C libraries, and also strongly affected my sequencing runs. To overcome the initial lack of diversity, Dr. Felix Krueger of the Babraham bioinformatics department and Dr. Osborne developed a package called *Bareback* (barcode back-processing) (Krueger et al., 2011). Bareback uses the raw image files generated by the Genome Analyser IIx and moves the images taken during low diversity to the back of the stack by renaming the files. These are then analysed using the Illumina GOAT (General Oligo Analysis Tool) pipeline, now part of the Illumina OLB (Off-Line basecaller).

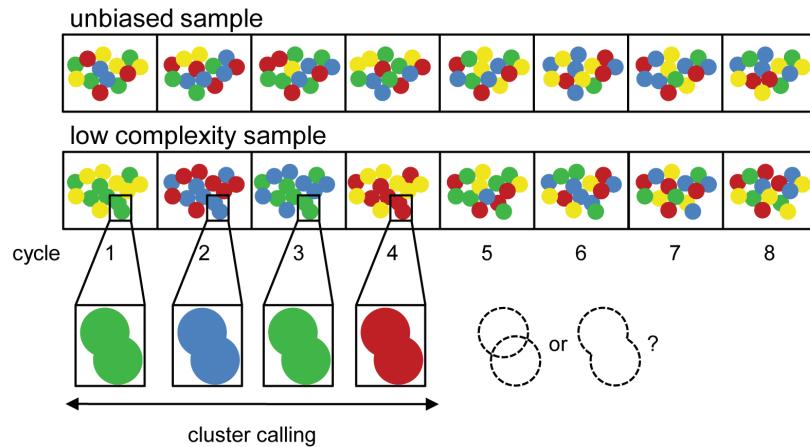


Figure 4.2.1 – Bareback overview. Diagram showing problematic cluster calling with samples containing low complexity in the initial sequencing cycles. Taken from Krueger et al. (2011)

Bareback processing greatly increased the number of sequence reads returned from my e4C libraries (Table 4.2.1), completely rescuing the ABL library with particularly dense clusters from which the standard SCS processing returned no reads.

	Illumina SCS processing	Bareback processing	Fold increase
BCR and MLL	14,409,580	22,167,823	1.54
MLL	13,147,751	16,704,073	1.27
BCR and MLL	18,434,991	23,251,377	1.26
MLL p1	4,690,956	27,868,955	5.94
MLL p2	272,927	30,632,135	112.24
ABL	0	33,157,523	∞
MLL phiX	24,349,038	26,284,182	1.08
BCR phiX			

Table 4.2.1 – Illumina sequence processing statistics with Bareback.

4.2.2 Quality control

Once sequences had been produced by the Bareback processing, the quality of the sequence data was assessed using two tools written by Dr. Simon Andrews of the Babraham bioinformatics department, *FastQC* and *FastQ Screen*. I was involved in development of FastQC version 0.9.3 (released 16/6/11) by contributing a new CSS theme to the report structure allowing simultaneous viewing of the overview

navigation and report results.

All e4C libraries passed the quality control steps without any cause for concern. Representative results are shown in Figure 4.2.2.

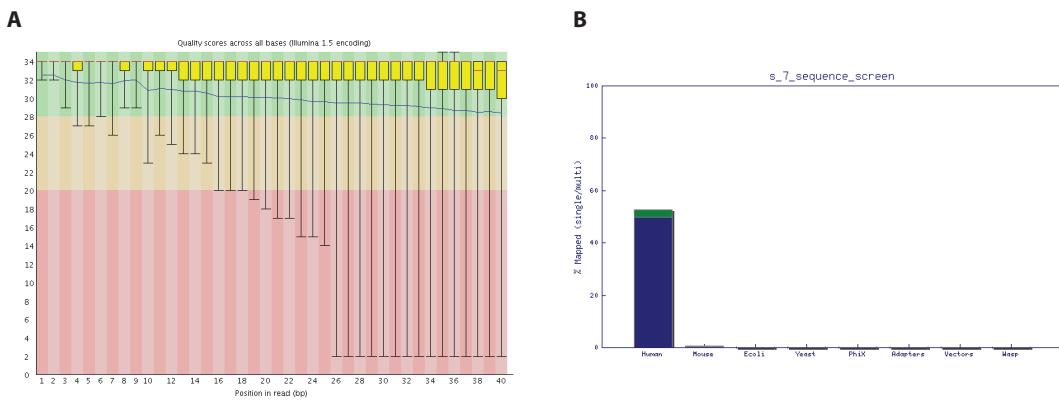


Figure 4.2.2 – Representative e4C Library FastQC and FastQ Screen. (A) FastQC - average sequence quality. (B) FastQ Screen - species alignment. Both statistics are for the second CD34⁺ BCR e4C sequencing run.

4.2.3 Sequence trimming

Before the sequenced e4C library reads can be analysed, they need to be aligned to the human genome. To do this the bait sequence must be removed up to the *AseI* recognition site. Over 97% of the sequences from each sequencing run started with an expected barcode sequence, with the exception of the *MLL* Run 3 phiX run which suffered from competition with the standard Illumina sequencing primer.

The expected structure of each e4C after the bait specific region is an *AseI* recognition site followed by unknown partner sequence (Fig 4.2.3). This partner sequence must also be processed to remove any unwanted sequence. Two such scenarios exist: the ligated fragment can be short due to a nearby *NlaIII* recognition site, in which case the read will run into the *NlaIII* site and on into the adapter. Secondly, due to the possibility of 3C concatemers forming from multiple *AseI* fragments ligated together, a second *AseI* site could be encountered followed by a third ligation product which we are not interested in. In both cases the sequence beyond the restriction endonuclease binding site

must be trimmed. A threshold is set to check the length of the sequence to be mapped after trimming and reads are discarded if they are too short.

To pre-process the e4C library reads, Dr. Felix Kreuger in the Babraham Bioinformatics group wrote a script in Perl, which I later modified and used myself (Appendix C.1). The script goes through the FastQ file and checks the first three base pairs of each sequence to see if it matches the barcode. If so, it trims them off and looks to see if the next six base pairs are an *AseI* recognition site (ATTAAT). Next, the script searches for *AseI* and *NlaIII* sites. If a second *AseI* site is found the line is rejected, if a *NlaIII* site is found the sequence is trimmed to that site. If the remaining sequence is less than 25 bp long it is rejected. The remaining content in the FastQ read, such as quality scores, are trimmed to the same length and the line is printed to the output file which is then aligned to the genome. The processing statistics from the e4C libraries can be seen in

Figure 4.2.4.

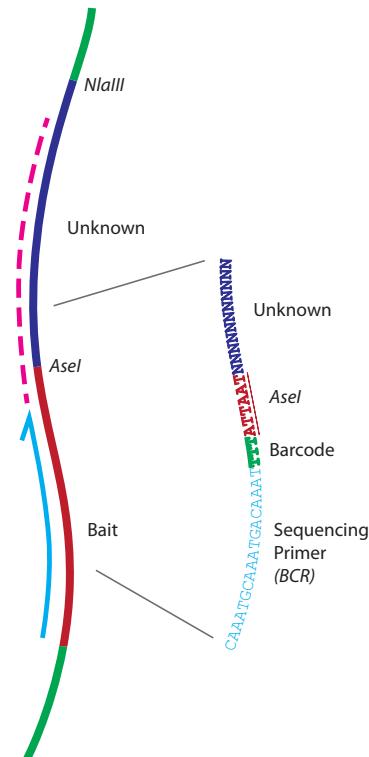


Figure 4.2.3 – Expected structure of reads in each e4C library.

4.2.4 Sequence alignment

The resulting trimmed reads were next passed through the genome alignment tool *Bowtie* (Langmead et al., 2009). Reads were aligned with 82% - 92% efficiency (Fig 4.2.4, Table C.1.1). Parameters were used as described in Fig 4.2.5. Reads not mapped to the genome are likely to have been removed from the strict alignment parameters which removed any reads with more than one alignment. Additionally, the aligned sequences were up to six base pairs shorter than regular Illumina GA IIx sequencing runs due to having the bait identifier removed.

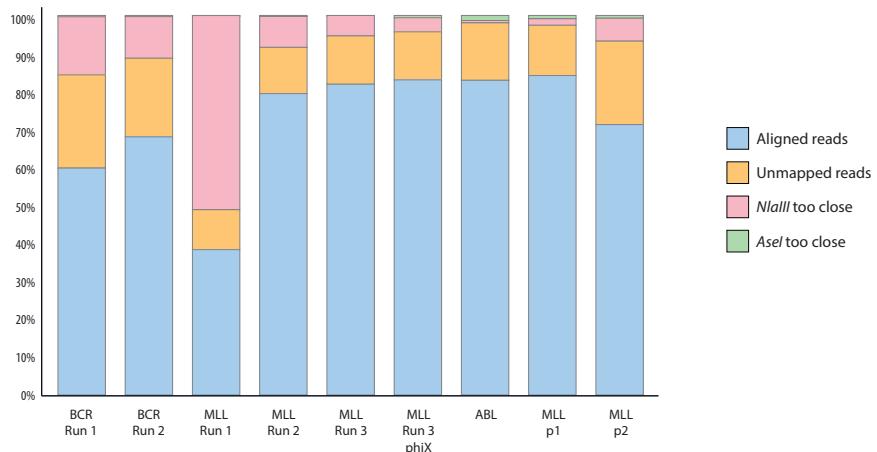


Figure 4.2.4 – Library trimming and alignments. Proportion of e4C libraries discarded due to Asel or NlalII sites being too close to the bait, or not being mapped to the genome / being suppressed due to multiple alignments. Raw numbers can be seen in Appendix Table C.1.

```
bowtie -q --phred64-quals -p 8 -m 1 hg19 <trimmed_reads.txt>
          <output_alignment.bowtie>
```

Figure 4.2.5 – Bowtie alignment parameters. -q: Input is in FastQ format. --phred64-quals: Correct interpretation of ASCII quality scores. -p 8: Use eight CPU cores for the alignment. -m 1: Ignore any read with more than one alignment. hg19: Name of the reference genome used for alignment. UCSC genome build hg19 was used.

4.2.5 Importing into SeqMonk

For analysis and quantitation of the e4C libraries, I used a program called *SeqMonk* (<http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>). *SeqMonk* has been developed by Dr. Simon Andrews, head of the Babraham Bioinformatics team, and is written in Java with a graphical user interface. It has been built from the ground up for analysis and visualisation of next generation sequencing data, and has grown extensively throughout the duration of my PhD.

SeqMonk works with aligned sequence reads, and can quantify them with probes - custom windows set across the genome. Probes are simply sets of paired genomic co-ordinates, and can be created over any feature (such as gene or restriction fragment) or as running windows. These probes are then used as bins within which reads can be quantified.

4.3 e4C library biases

In 2011, Dr Amos Tanay's group wrote a paper reanalysing a previously published Hi-C dataset (Yaffe and Tanay, 2011; Lieberman-Aiden et al., 2009). They showed biases present within the Hi-C data towards higher GC content and particular restriction fragment lengths (Yaffe and Tanay, 2011). To investigate whether my e4C interaction datasets were affected by the same biases, I wrote perl scripts to compare the characteristics of the observed *AseI* - *NlaIII* fragments versus all possible fragments.

4.3.1 Potential fragment libraries

A prerequisite for the analysis of these biases is the generation of an *in-silico* library of all potential fragments. The original perl script to generate these libraries was written by Marek Piatek, in the Babraham bioinformatics department. I have since re-written the script from scratch, as well as writing other scripts to including an online tool to generate lists of restriction enzyme recognition sites (Appendix C.2). I have made some of these available as online tools at <http://www.tallphil.co.uk/bioinformatics/> (Appendix G).

In principle, all of the scripts are similar - the genome is loaded into memory one chromosome at a time, and perl's *index* function is used to search for the restriction enzyme recognition sites. For *AseI* - *NlaIII* fragments, once an *AseI* site is found the next *NlaIII* site is searched for. If one is found before the next *AseI* site, the resulting fragment length is tested; if it is shorter than 35 base pairs or longer than 700 base pairs then it is discarded due to the gel extraction size selection used in the e4C protocol. The resulting library of fragments is then aligned using *Bowtie* and so filtered for unique mappability (Table 4.3.1).

	Fragments Removed	Total Fragments
All <i>AseI</i> - <i>NlaIII</i> fragments	0	2407346
Filter for length < 35bp	283205	2124141 (88.24%)
Filter for length > 700bp	74398	2049743 (85.15%)
Filter for unique mappability	153862	1895881 (78.75%)

Table 4.3.1 – *in-silico* potential *AseI* - *NlaIII* fragment library statistics.

4.3.2 GC content bias and fragment length bias

The *in-silico* library described above was imported into SeqMonk as an annotation track, and a probe created over each potential fragment. A binary value was assigned to each fragment by using *Minimum Coverage Depth* quantitation. Results were exported as an annotated probe report, containing the co-ordinates of every potential *AseI* - *NlaIII* fragment with a binary flag to indicate whether it was observed or not. I wrote a Perl script to process these datasets (Appendix C.3). The script fetches the genomic sequence for each set of co-ordinates and increments counters representing 5% GC content bins, according to the %GC content of the sequence. A second set of counters were also incremented if the sequence was observed. A modified version of this script was then used to calculate *AseI* - *NlaIII* fragment lengths.

The resulting counts were plotted and can be seen in Figure 4.3.1. It can be seen that the e4C libraries are biased towards greater %GC content and shorter fragments. These are likely to be experimental biases that come from the ligation, PCR and sequencing steps. It should be noted that the interaction libraries are enriched for GC rich regions (Section 5.4), so this bias may be a reflection of the specificity of the interaction data.

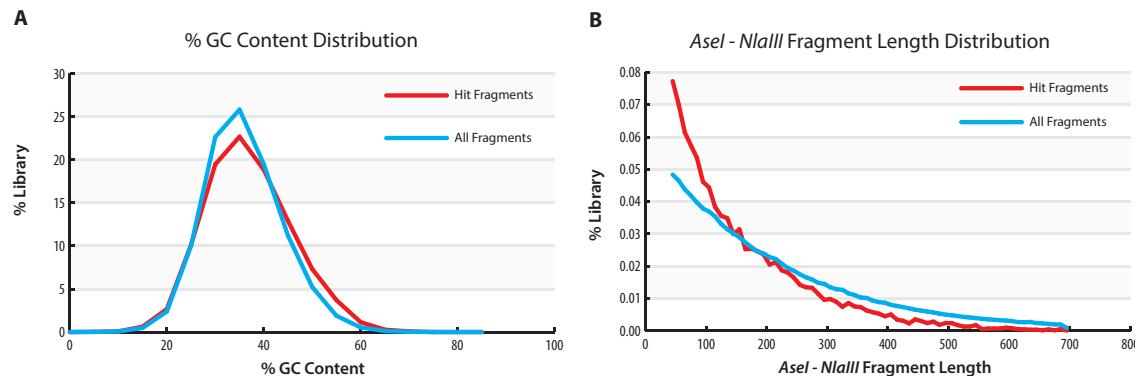


Figure 4.3.1 – e4C library biases. **(A) %GC content distribution.** GC content of each sequence calculated and quantified in 5% GC bins for all potential *AseI*-*NlaIII* fragments (blue) and all observed *AseI*-*NlaIII* fragments (red). An enrichment for higher GC content can be seen in the observed fragments. **(B) Fragment length distribution.** *AseI*-*NlaIII* fragments were quantified by their fragment length. An enrichment of shorter fragments can be seen in the observed *AseI*-*NlaIII* fragments (red).

4.3.2.1 Bias correction

To correct the e4C libraries for systematic biases that are due to GC content and fragment length, a library of all potential *AseI - NlaIII* fragments was created with associated 'expected probability' values. These were generated with a perl script which calculates a correction value for each bin of GC % and fragment length: (Appendix C.4)

$$C = \frac{\text{hit fragments in bin}}{\text{hit fragments}} / \frac{\text{all fragments in bin}}{\text{all fragments}}$$

Then, for every potential fragment, the appropriate correction values for %GC and fragment length were multiplied against the overall chance of any fragment being hit:

$$\rho_{frag} = \frac{\text{observed fragments}}{\text{potential fragments}} \times C_{\%GC} \times C_{fraglength}$$

This library of expected observation rates for every potential fragment was then used in single window testing (Section 4.4).

4.3.2.2 Validation

To check that the values produced by this script were reliable, I wrote a small Perl script to generate *in-silico* libraries using the correction values. The script loops through every potential fragment and outputs it as a hit if a random number between 0 and 1 is greater than or equal to the correction factor. This was repeated 5 times to generate a suitable number of reads. This *in-silico* library was then loaded into SeqMonk for visual inspection before being analysed for biases as described in Section 4.3.2. The resulting plot shows the *in-silico* library with identical biases to the observed library, validating the correction values.

4.4 Significance of single regions

To test the significance of the observed hits in a single region, I wrote a script that generates *in-silico* random libraries for defined regions of the genome. This process can be repeated many times and the number of *in-silico* fragment hits generated for each run compared to the number of observed reads in the real e4C library. A *p* value can be calculated for a single window using this approach by counting the number of times that the *in-silico* fragment hit counts are \geq the number of observed fragments hit for the region.

$$p = \frac{\sum(\text{in-silico runs} \geq \text{observed fragments})}{\sum(\text{in-silico runs} < \text{observed fragments})}$$

For an example of this analysis in use, see Section 6.1.1.

4.5 *AseI* site distribution normalisation

Because e4C uses restriction fragments to analyse sequence proximity, there is a chance that the unequal distribution of *AseI* sites across the genome could skew the results. To correct for this, I used a Perl script written by Dr. Simon Andrews in the Babraham Bioinformatics department. The script loads a list of potential fragments with binary flags indicating whether they are observed or not, prepared as described in Section 4.3.2. It creates 100 Kb rolling windows and counts the number of *AseI* fragments present. If there are greater than 5 fragments in a window, it calculates an ($\frac{\text{observed}}{\text{potential}}$) value and outputs this to the results file. If there are less than 5 fragments it outputs 0; calculating percentages with such small numbers of fragments can give uninformative values for regions with few restriction sites, such as centromeres.

To allow the analysis of these percentage values within SeqMonk, I wrote a Perl script to create a number of 'virtual reads' over each 100 Kb window according to the percentage score. These can then be imported into SeqMonk and quantified by whatever method is desired (Fig 4.5.1).

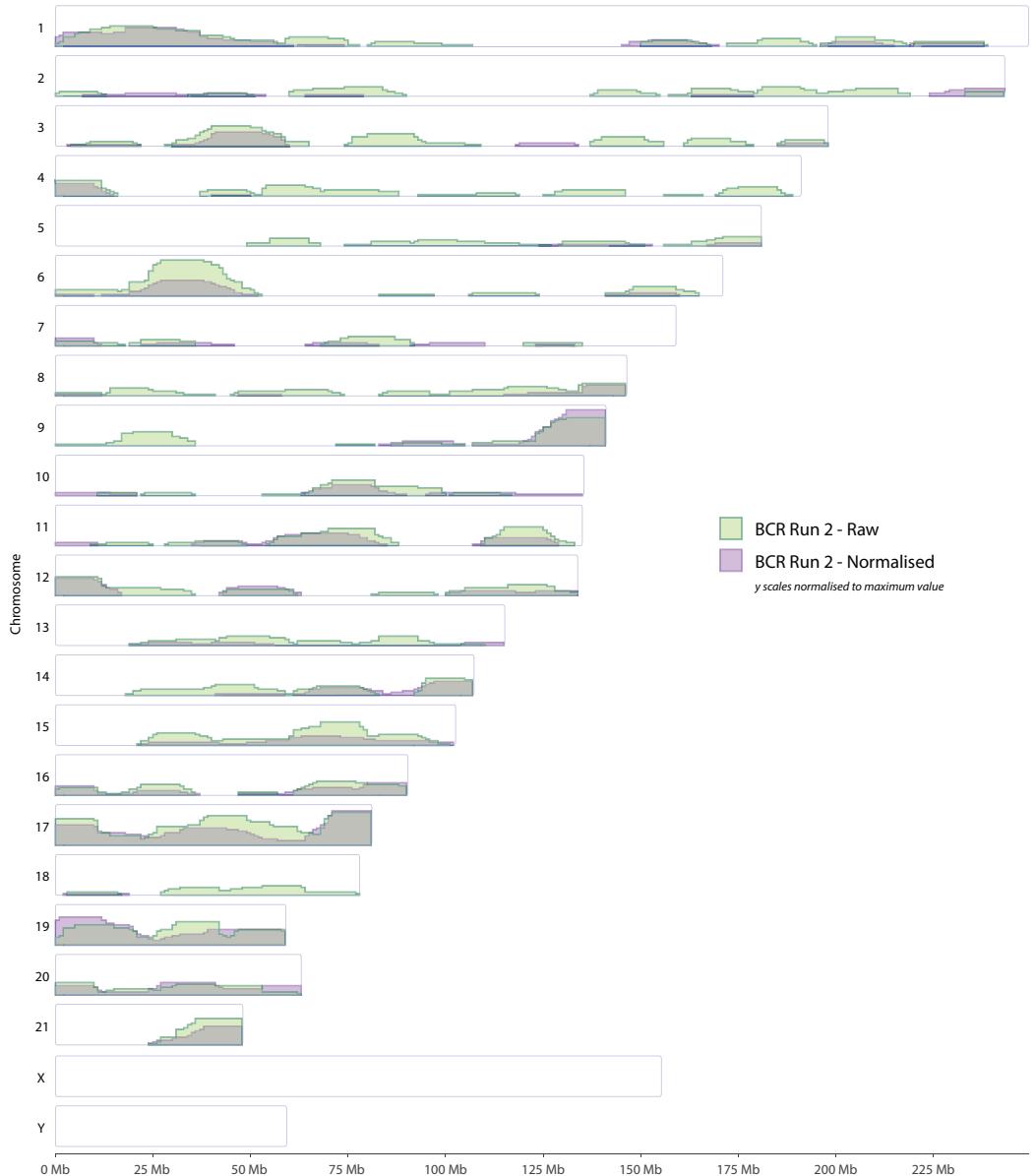


Figure 4.5.1 – *AseI* site distribution normalisation. Genome wide plots of BCR e4C Run 2 before and after *AseI* site distribution normalisation. Standard scores calculated and normalised against maximum value.

4.5.1 *In-silico* testing

To validate this normalisation, I wrote a Perl script which creates randomised virtual e4C libraries from the list of potential *AseI* - *NlaIII* fragments and loaded this into SeqMonk. I then ran the *AseI*

normalisation as described above and loaded the resulting normalised library back into SeqMonk. The randomised library had peaks across the genome whereas the normalised library looked almost perfectly flat (Fig 4.5.2) suggesting that my normalisation is capable of removing bias due to the unequal distribution of *AseI* sites.

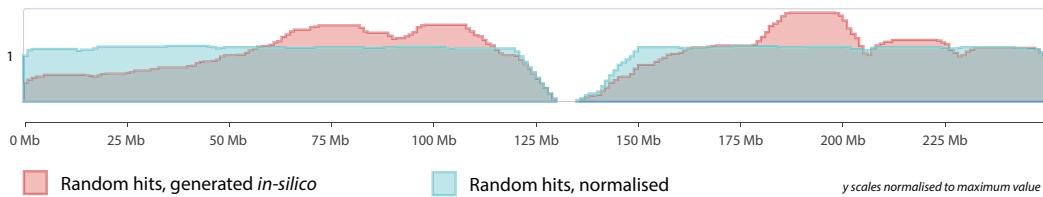


Figure 4.5.2 – *AseI* normalisation *in-silico* test. Chromosome 1 showing the *in-silico* e4C randomised library before and after normalisation for *AseI* site distribution. Read counts quantified and normalised against maximum value.

4.5.2 Standard scores

To allow the comparison of different e4C libraries with varying degrees of coverage, I typically converted and final quantification values to *standard scores*, also known as *Z scores*. A standard score represents how many standard deviations from the mean a single value lies within a dataset. I initially calculated these values by exporting data from SeqMonk and importing it into IBM SPSS Statistics. The ability to requantify probes with standard scores has since been built into SeqMonk.

4.6

Chapter 5

Initial e4C library analysis results

5.1 Introduction

In this thesis I investigate the hypothesis that proto-oncogenes involved in leukaemic chromosomal translocations may physically associate in human haematopoietic cells. In this chapter, I describe the initial analysis of e4C association profiles of *BCR*, *ABL1* and *MLL* in human CD34⁺ cells, the lymphoblastoid cell line GM12878. I study the association frequencies of the genes in *cis* as in *trans* and show that these genes interact with an active nuclear compartment.

5.1.1 Overview of e4C libraries

To give context to the e4C libraries discussed in this thesis see an overview in Table 5.1.1. In brief, I began by sequencing e4C libraries generated from CD34⁺ 3C material for *BCR* and *MLL* breakpoint region baits. Due to low read counts in the MLL 1 run (see Section 3.7.1.1) I sequenced a second MLL e4C library in its own lane (MLL 2). I sequenced a second multiplexed lane with further technical repeats of the *BCR* and *MLL* e4C libraries (BCR 2, MLL 3) and to investigate the crossover reads I

repeated sequencing of this MLL library in a phiX lane (MLL 3 phiX, see Section 3.7.2). To take advantage of plentiful cell numbers and available epigenomic datasets, I created e4C libraries for *ABL1* and *MLL* with the promoter *AseI* fragments in the GM12878 cell line and sequenced these three libraries in their own lanes.

	Name	Date Sequenced	Bait Position	Description
CD34 ⁺	BCR 1	June 2010	chr22: 23,596,613 - 23,612,935 Breakpoint region of <i>BCR</i> gene.	First e4C library sequenced, multiplexed with MLL 1.
	BCR 2	Oct 2010		Technical replicate of BCR 1, library generated from the same 3C material. Sequenced in multiplex with MLL 2
	BCR phiX	Mar 2011		Test of new multiplexing and barcode protocol. See Section 3.8.
	MLL 1	June 2010	chr11: 118,354,809 - 118,356,055 Breakpoint region of <i>MLL</i> gene	First e4C library sequenced, multiplexed with BCR 1.
	MLL 2	Sept 2010		Technical replicate of MLL 1, generated from the same 3C material. Sequenced in its own lane.
	MLL 3	Oct 2010		Technical replicate of MLL 1, generated from the same 3C material. Sequenced in multiplex with BCR 2.
	MLL 3 phiX	Dec 2010		Repeat sequencing of MLL 3, multiplexed in phiX lane.
GM12878	ABL	Dec 2010	chr9: 133,576,435 - 133,591,823	Promoter of <i>ABL1</i> gene
	MLL p1	Dec 2010	chr11: 118,304,771 - 118,311,258	Promoter of <i>MLL</i> gene (centromeric end of <i>AseI</i> fragment)
	MLL p2	Dec 2010	chr11: 118,304,771 - 118,311,258	Promoter of <i>MLL</i> gene (telomeric end of <i>AseI</i> fragment)

Table 5.1.1 – e4C Libraries. Chromosomal positions in Bait Position column describe the bait *AseI* fragment in UCSC genome build hg19.

5.2 Concerning raw data

5.2.1 Duplicate reads

The first thing that can be seen when viewing the e4C libraries is the degree of sequence duplication. The e4C protocol involves two rounds PCR resulting in multiple copies of each initial 3C product being sequenced. It is clear that some quantitative information is present within the numbers of reads found, as shown by the increased number of duplicates per unique read in near *cis* (< 5 Mb from the bait region) compared to *trans* (Fig 5.2.1, Appendix Table C.1.2). However, because each e4C fragment end is defined by a restriction enzyme cut site, duplicates from multiple 3C ligation events

are indistinguishable from duplicates generated by the PCR amplification. Because of this inability to distinguish biologically relevant duplicates from technical duplicates, I removed all non-unique reads from the data.

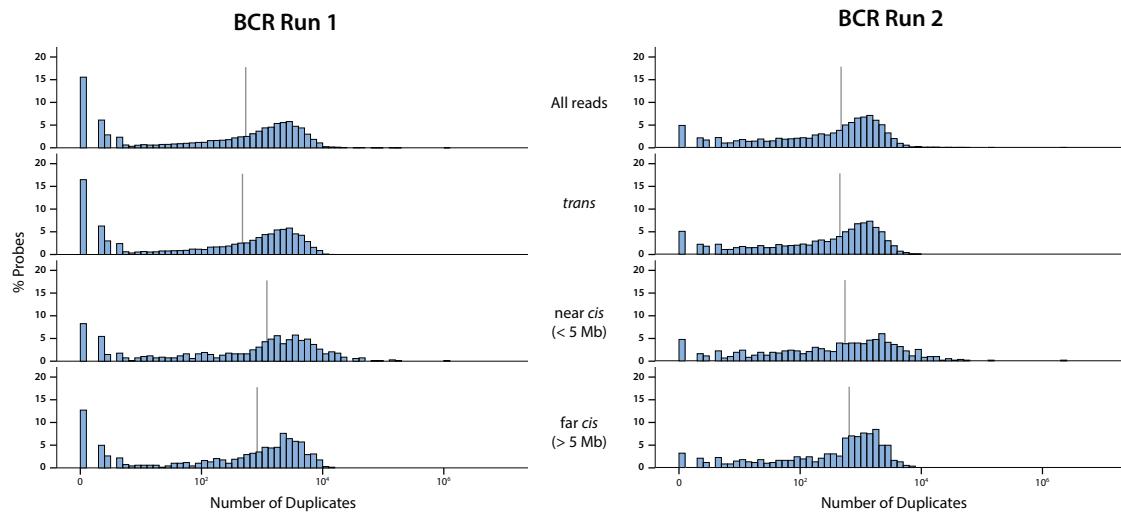


Figure 5.2.1 – e4C duplicate frequency. Histogram showing the frequency of duplicate counts per unique read for *BCR* e4C libraries. X axis shows the number of duplicates per unique read (\log_{10} count), Y axis shows percentage of that collection that the number of duplicates applies to. Grey bars show median number of duplicates.

5.2.2 *AseI* fragment saturation

Because the e4C assay is based off 3C libraries, the highest achievable resolution is dependant on the length of *AseI* fragments. A concern when analysing the *cis* data was that we could reach a point of saturation. To determine the degree of saturation I reduced all e4C reads to binary observations of *AseI* fragments, represented as a percentage of potential fragments (Fig 5.2.2, Appendix Table B.1.1).

5.2.3 Library complexity

The level of library complexity and coverage can be seen in Fig 5.2.3 and Appendix Table B.1.1. Two observations can be made: the detail in coverage is not very great in *trans*, meaning that association at the level of individual genes cannot be probed. Secondly, the two *BCR* breakpoint region bait libraries

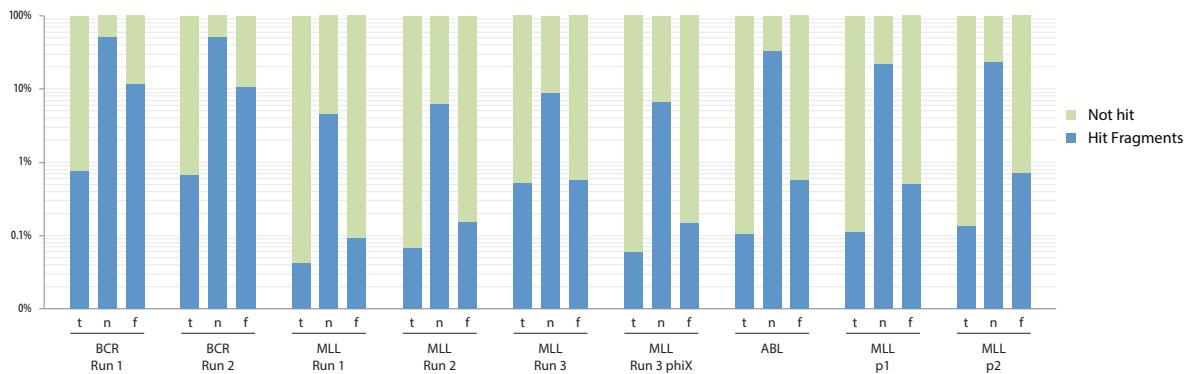


Figure 5.2.2 – e4C *AseI* fragment saturation. Degree of *AseI* fragment separation in *trans* (t), near *cis* (n, within 2.5 Mb either side of bait) and far *cis* (f, beyond 2.5 Mb either side of bait). Raw numbers can be seen in Table B.1.1.

have approximately ten fold greater complexity than the MLL, ABL, MLL p1 or MLL p2 baits. The reason for this difference in complexity is not known but is likely to be due to a primer-specific effect in bait enrichment, PCR amplification or sequencing.

5.3 *cis* interaction profiles

The linear separation of two DNA sequences is a highly constraining factor in the three dimensional organisation of chromatin. If two regions are on the same chromosome, they are physically linked and the maximum distance that they may separate by is constrained by the length of sequence separating them. As such, the linear separation of two sequences is the strongest predictor of three dimensional interaction. This effect can clearly be seen within all of the e4C libraries, with a large proportion of my libraries (29 - 45% raw reads) mapping to the *cis* chromosome (Fig 5.2.3, Appendix Table C.1.2). Sequence density variations are so large between the *cis* and *trans* chromosomes that I analyse them independently.

5.3.1 Interaction frequency in *cis* declines as a function of linear separation

To quantify *cis* association I plotted *cis* association data normalised against *AseI* fragment distribution in SeqMonk (100 Kb windows with a minimum of five fragments), summing the percentage of

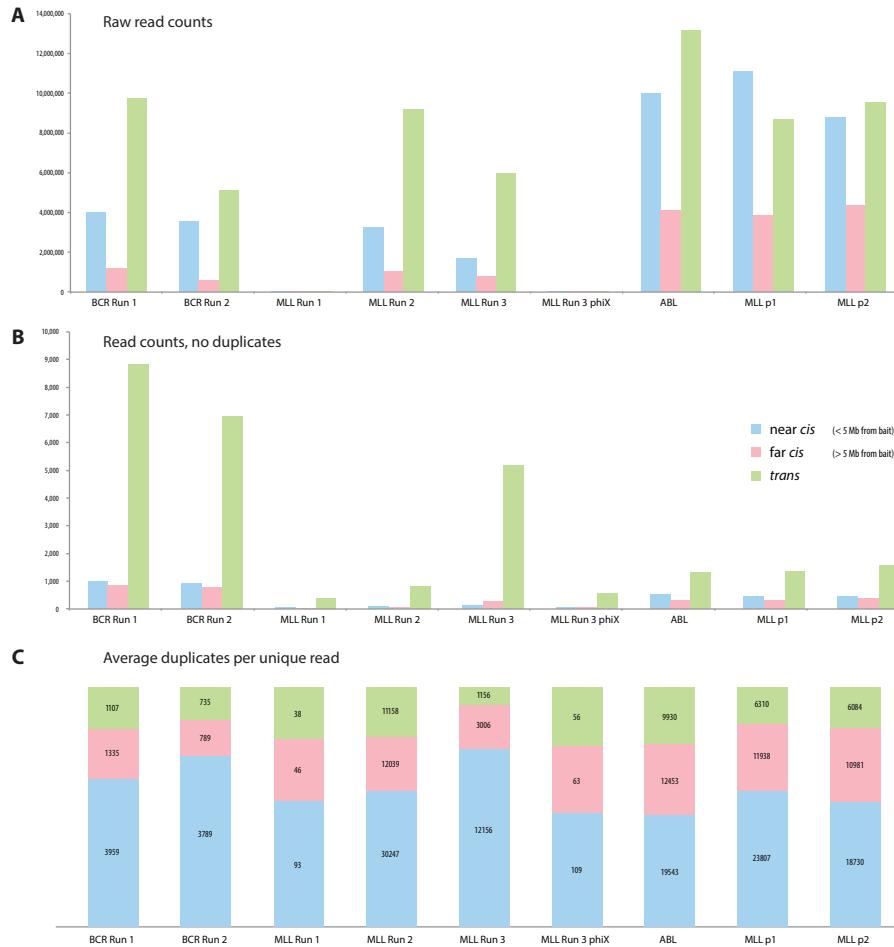


Figure 5.2.3 – e4C library read statistics. (A) **Aligned raw read counts** for near *cis*, far *cis* and *trans*. Reads within 5 Mb of the bait can be seen to represent a large proportion of all reads. Low read counts can be seen for MLL Run 1 (due to biased multiplexing, see Section 3.7.1.1) and MLL Run 3 phiX (due to competition with Illumina sequencing primer). Increased total read counts can be seen in the ABL, MLL p1 and MLL p2 libraries due to advances in achievable cluster densities on the Illumina GA IIx. (B) **Read counts after removal of duplicate reads.** The number of near *cis* reads can be seen to drop relative to (A), due to saturation of *Ase*/fragments near the bait. A large number of *trans* reads can be seen in MLL Run 3 due to crossover reads from the BCR Run 2 in multiplexing (see Section 3.7.2). This plot shows the ~10 fold difference in library complexity between the BCR bait and all other baits. (C) **Number of duplicates per unique read.** Profiles of duplication can be seen to be similar for all e4C libraries, despite highly variant total read counts. Near *cis* and far *cis* have far more duplicates per unique read than the *trans* hits. The number of duplicates per read is high for all hits.

fragments hit in ten megabase windows separated by one megabase (Fig 5.3.1). Large association peaks can be seen surrounding the bait regions (grey bars) as expected, falling off sharply ~ 2.5 Mb from the bait region on either side (the 5 Mb window I refer to as near *cis*).

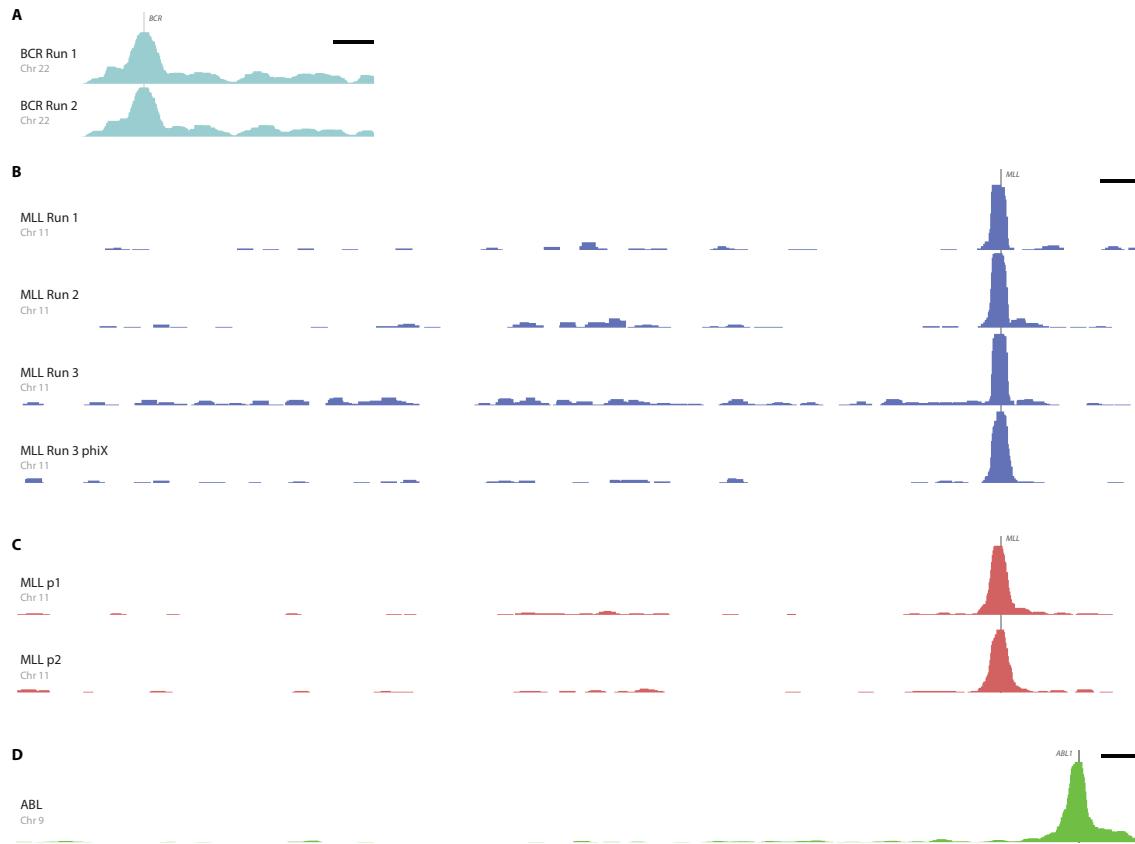


Figure 5.3.1 – e4C *cis* association profiles. *cis* interaction profiles generated using 100 kb sliding window fragment proportions (minimum 5) summed in 1 mb windows separated by 100 kb. Grey bars show position of bait genes. Black horizontal bars are 5 Mb. **(A)** Interaction profiles of the *BCR* breakpoint region bait in CD34⁺ cells. Non-repetitive region of chromosome 22 q arm shown. **(B)** Interaction profiles of the *MLL* breakpoint region bait on chromosome 11 in CD34⁺ cells. **(C)** Interaction profiles of the two *MLL* promoter baits on chromosome 11 in GM12878 cells. **(D)** Interaction profile of the *ABL1* promoter bait on chromosome 9.

5.3.2 Specific interactions in *cis*

In addition to the obvious peak of *cis* association common to all of the e4C association datasets, some *cis* bait-specific interactions can be seen (Fig 5.3.1, Fig 5.3.2). The ABL bait interacts with the telomeric region of chromosome 9 more strongly with the rest of the chromosome and the BCR bait can be seen to make a number of contacts in both replicates, including a shoulder centromeric to the bait and four clusters telomeric (Fig 5.3.1). To see *cis* interactions in more detail in the BCR datasets

with higher complexity, I plotted the fragment interaction percentages summed in ten fold smaller one megabase windows separated by 100 Kb (Fig 5.3.2 A). To ensure that any observed peaks are not artefacts caused by my method of *AseI* fragment distribution normalisation, I also plotted association using a different method of normalisation: sliding windows covering ten *AseI* fragments separated by two fragments (Fig 5.3.2 B).

Both normalisation techniques show several hits above the expected baseline in both replicates. To the centromeric (left) side of the *BCR* bait three distinct clusters of interaction can be seen.¹

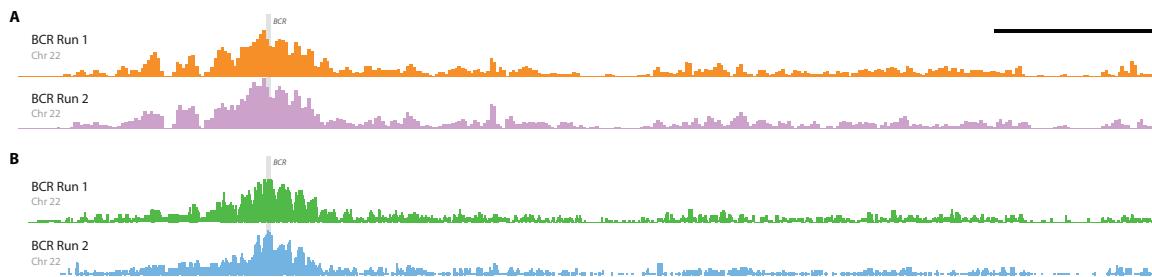


Figure 5.3.2 – *BCR* e4C *cis* association. *cis* interaction profile of the *BCR* bait on chromosome 22 in CD34⁺ cells. (A) Interaction shown as the proportion of *AseI* – *NlaIII* fragments hit within a 50 kb sliding window (minimum 5 fragments). (B) Interaction shown using a sliding windows of 10 *AseI* fragments separated by 2 fragments. Black horizontal bar is 5 Mb.

5.4 BCR, ABL1 and MLL reside in an active nuclear compartment

Upon plotting the e4C *trans* interaction profiles of the different bait genes, it was clear that they share a number of features. My bait genes are actively transcribed within CD34⁺ cells (Figure 3.5.1) and visual inspection of the interaction profiles appeared to share some features with previously described active regions (Versteeg et al., 2003), so I was interested in whether my e4C interaction libraries showed correlations with RNA polymerase II binding and other active chromatin marks.

¹What are these? Further discussion?

5.4.1 Correlation with active epigenetic marks

To investigate the correlation between e4C interaction frequency and chromatin marks, I analysed publicly available genome-wide datasets (Bernstein et al., 2010; Raha et al., 2010; Hu et al., 2011; See Appendix D.1 for NCBI GEO accession numbers). Gene density was calculated by counting genes labelled as 'protein-coding' in ensembl (Flicek et al., 2012). These datasets were quantified by their read counts and the resulting standard scores compared to the those from the *trans* e4C interaction data. Pearson's correlation coefficients were calculated using SeqMonk. The resulting correlation R scores are shown in Table 5.4.1.

		CD34 *					GM12878					
		BCR 1	BCR 2	MLL 1	MLL 2	MLL 3	MLL 3 Φ X	ABL	MLL p1	MLL p2		
Active Marks	Gene Density	0.707	0.741	0.228	0.201	0.284	0.258	0.614	0.503	0.577		
	RNA-Seq	0.700	0.723	0.216	0.240	0.294	0.249	0.635	0.533	0.613		
	DNase	0.672	0.682	0.281	0.387	0.218	0.376	0.547	0.491	0.605		
	H3K27ac	0.629	0.639	0.279	0.379	0.227	0.363	0.558	0.501	0.610		
	H3K36me3	0.607	0.622	0.263	0.333	0.232	0.348	0.549	0.483	0.622		
	H3K4me1	0.749	0.757	0.306	0.366	0.292	0.379	0.630	0.551	0.655		
	H3K4me3	0.636	0.659	0.263	0.345	0.198	0.329	0.548	0.463	0.583		
	DNase	0.718	0.731	0.275	0.351	0.268	0.386	0.589	0.547	0.648		
	RNA-Seq	0.688	0.713	0.242	0.241	0.280	0.241	0.672	0.553	0.637		
	RNAPII	0.743	0.760	0.291	0.316	0.275	0.320	0.659	0.573	0.654		
Inactive Marks	CD34 *	H3K9me3	0.064	0.080	0.032	0.185	0.055	0.157	0.034	0.016	0.132	
		H3K27me3	0.353	0.344	0.140	0.230	0.084	0.227	0.303	0.317	0.374	
	GM12878	H3K9me3	0.120	0.119	0.090	0.229	-0.020	0.158	0.099	0.115	0.210	
		H3K27me3	0.356	0.342	0.188	0.319	0.100	0.250	0.359	0.344	0.406	
	Control	GM12878	ChiP-Seq Input	0.254	0.247	0.136	0.164	0.212	0.127	0.016	0.087	0.096
			Colour scale	0		0.4		0.8				

Table 5.4.1 – e4C: active epigenomic mark correlation scores. Standard scores were calculated using quantified reads in ten megabase windows separated by one megabase. e4C bait *cis* chromosomes were removed from analysis. Gene density was calculated using protein coding genes only. Values shown in italics are comparisons across different cell types.

The strong correlation between bait interaction and active mark association score shows that the e4C bait genes are frequently interacting with active regions of the genome. The low scores seen with

the inactive histone modifications signify that there is not more interaction with these regions than could be expected by chance. The correlation of the shape of the profiles can be clearly seen when plotted genome wide (Fig 5.4.1).

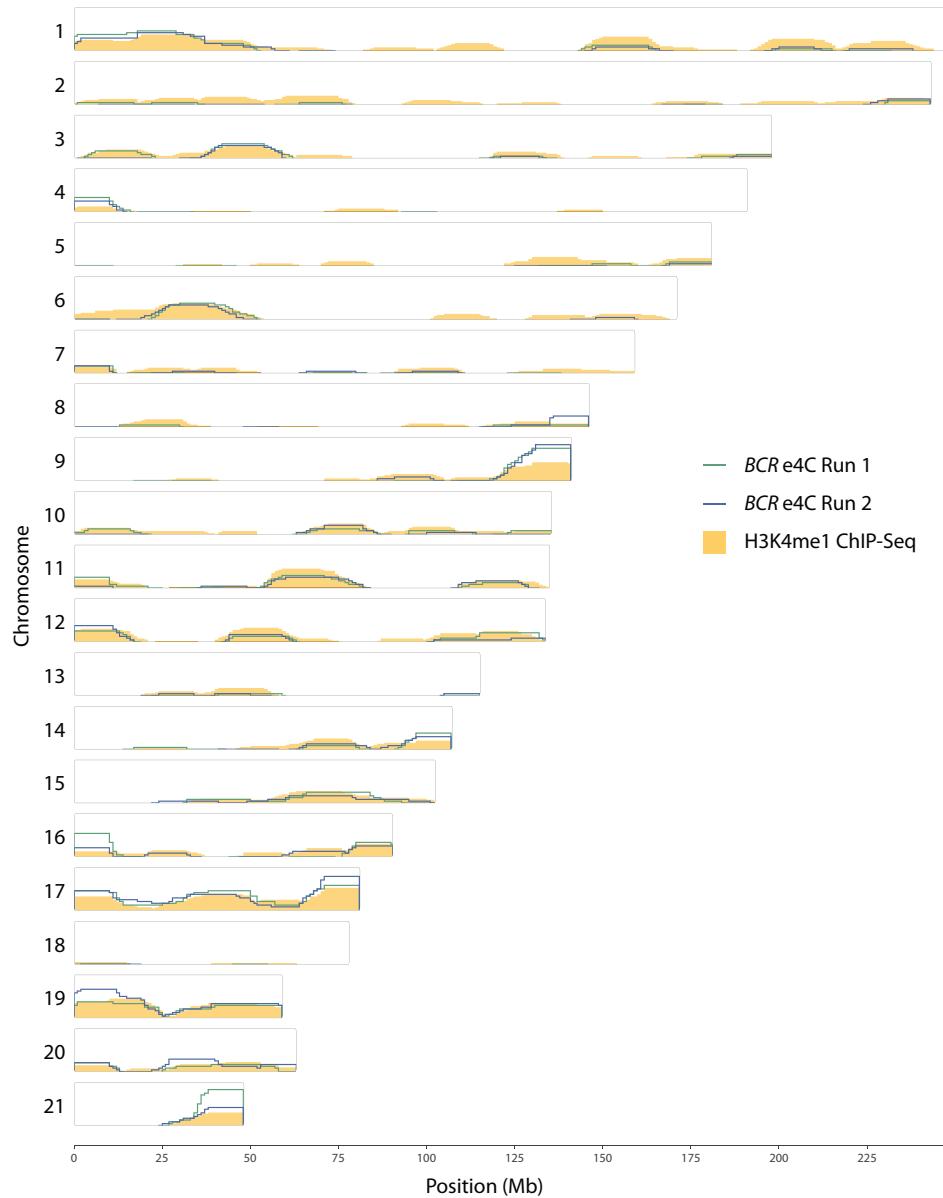


Figure 5.4.1 – BCR e4C correlation with H3K4me1. Genome wide association profiles of the two *BCR* e4C technical replicates (green and blue lines) normalised for *Ase1* distribution in 100 Kb sliding windows (minimum five fragments) and summed in ten megabase windows with one megabase separation and quantified with standard scores. H3K4me1 ChIP-Seq data quantified in the same windows with standard scores (orange fill) (Bernstein et al., 2010). Y scales of the datasets were adjusted to have equal maximal values to better show correlation. Negative standard scores hidden to aid visibility.

Many of the differences in correlation scores between the *BCR*, *MLL* and *ABL1* e4C datasets can be attributed to differences in e4C coverage; *ie.* the CD34⁺ *MLL* libraries do not have enough reads to show strong enrichment regions. It should be noted that these correlations are not dependent on cell type; this is consistent with the large scale organisation of ridges and anti-ridges in the genome (Lercher et al., 2002; Versteeg et al., 2003). The lack of correlation with inactive histone marks instead of anti-correlation is supported by the recently published ENCODE 5C dataset, where Sanyal *et al.* also observe a lack of enrichment for H3K27me3 (Sanyal et al., 2012).

5.4.2 Correlations between e4C libraries

As all of the bait e4C profiles associate with active marks to some degree, I was interested to see the correlation between the different datasets. Pearson's correlation coefficients were calculated as described above, though with both bait *cis* chromosomes removed. A number of interesting observations can be made from the resulting correlation scores: the higher data complexity and reproducible nature of the BCR technical replicates can be seen by their strong correlation; the effect of the crossover reads in the MLL Run 3 library can be seen by its correlation with the BCR profiles; MLL 1 and MLL 3 show good correlation which could be due to them both being multiplexed with BCR runs, MLL 1 had very low complexity which could explain its lack of correlation with BCR.

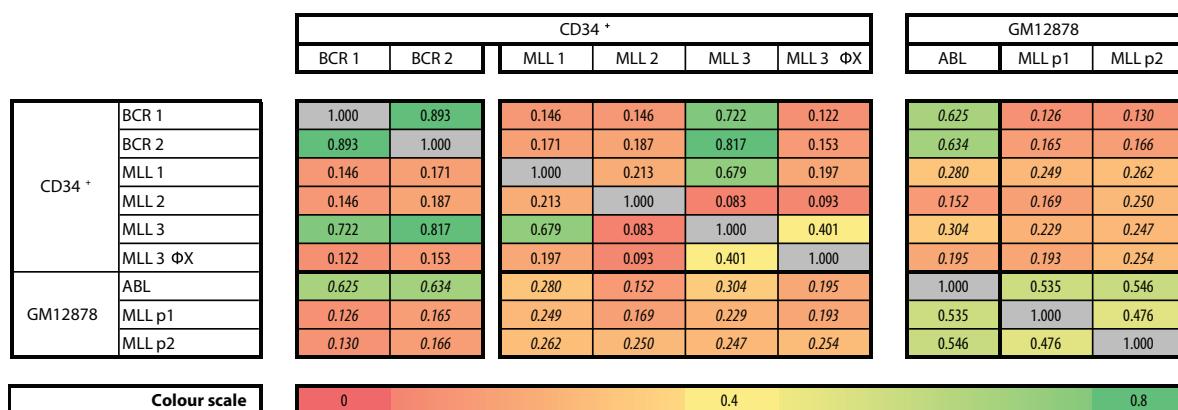


Figure 5.4.2 – e4C library correlations. Standard scores were calculated using quantified reads in ten megabase windows separated by one megabase. Both e4C bait *cis* chromosomes were removed from analysisValues shown in italics are comparisons across different cell types.

Interestingly, the GM12878 ABL dataset shows strong correlation with the CD34⁺ BCR datasets. This is unlikely to be due to differences in complexity as the GM12878 MLL p1 and MLL p2 which have similar read numbers and complexity do not show this correlation. This suggests that the correlation could indicate a matching of specific genome-wide contacts by the two loci caused by colocalisation within the nucleus. Using genome-wide association correlations to measure proximity instead of direct contact frequency as measured by 3C product reads is often used in Hi-C data analysis (Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011).

5.5 Different genes have different preferred interaction partners

To see if any bait-specific *trans* interactions were present within the datasets, I normalised the interaction profiles using the most correlated active marks: H3K4me1 for the CD34⁺ e4C libraries (Fig 5.5.1) and GM12878 RNA polymerase II ChIP-Seq for the GM12878 e4C libraries (data not shown). Standard scores from the epigenomic datasets were subtracted from the interaction library standard scores for each window position. This normalisation against has the effect of largely flattening the genome wide profile (Figure 5.5.1), as expected by the observed correlations (Table 5.4.1). Despite this, some noticeable peaks remain within the *BCR* datasets. The low coverage of the *MLL* and *ABL1* interaction datasets makes the detection of any specific local interaction regions difficult, and those observed unreliable. I discuss the main peak of the *BCR* datasets in Chapter 6.

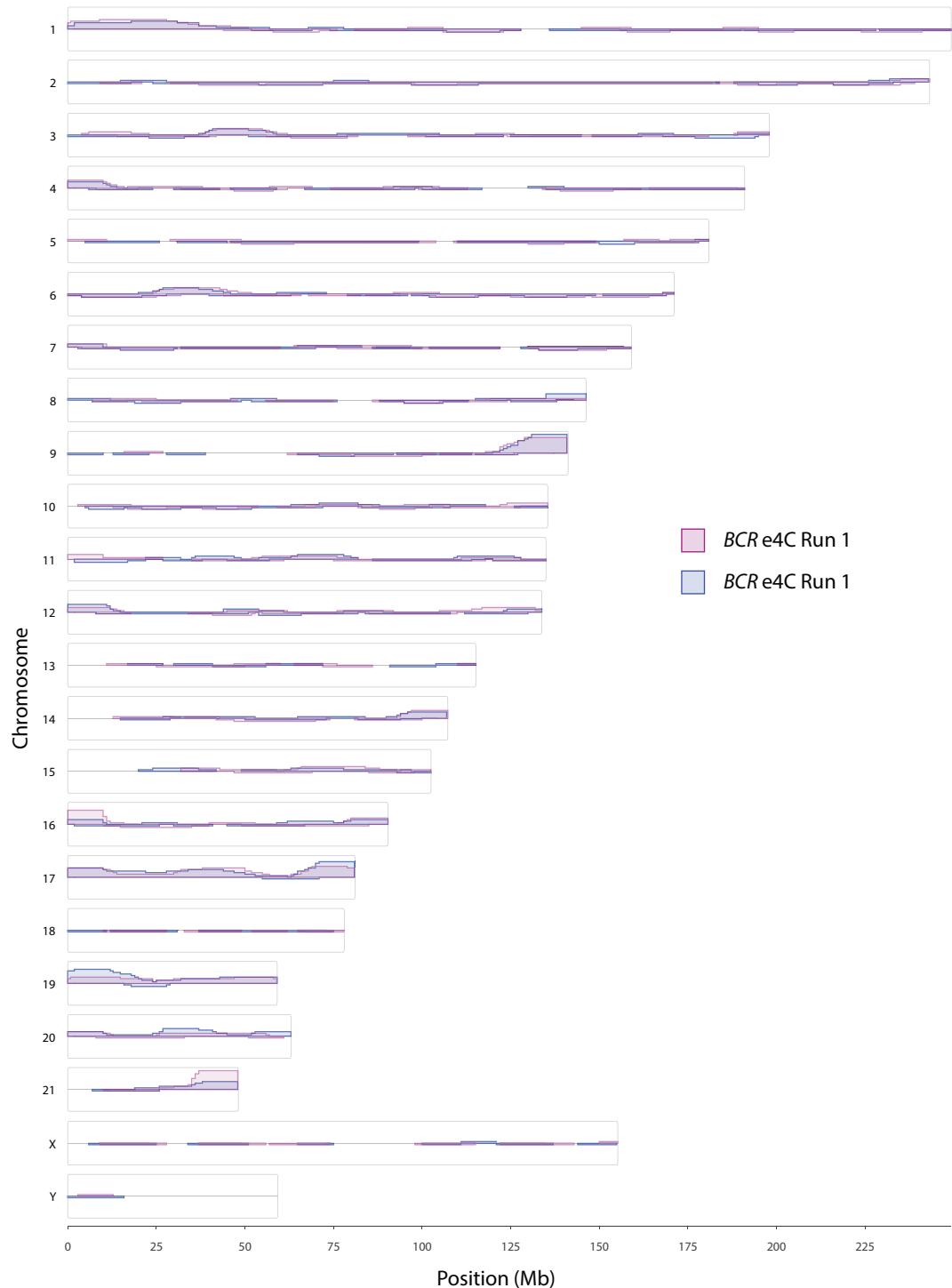


Figure 5.5.1 – *BCR e4C normalisation against H3K4me1.* Association profiles normalised against a H3K4me1 ChIP-Seq profile (Bernstein et al., 2010). Standard scores were calculated for the normalised association datasets and the ChIP-Seq dataset. Association datasets were normalised against the ChIP-Seq dataset by subtraction.

Chapter 6

Preferential association of *BCR* with Chromosome 9

6.1 *BCR* preferentially co-associates with 9q34 in CD34⁺ cells

The most prominent interaction seen in the two replicate *BCR* interaction profiles is on chromosome 9, cytogenetic band q34. When the two replicates are combined this interaction gives approximately 7 fold enrichment of association above the mean with a maximal standard score of 17.043. This is the strongest *trans* interaction that I found in any of my e4C association datasets.

Chromosome 9 q34 is a gene dense region containing a large number of highly transcribed genes. It contains the gene *ABL1*, a recurrent translocation partner with *BCR*. The t(9;22)(q34;q11) translocation is found in 95% of chronic myeloid leukaemia cases (Hehlmann et al., 2007) and was the first cytogenetic abnormality of its kind to be characterised (Nowell and Hungerford, 1960). For more discussion about the *BCR-ABL1* translocation, please see Section 1.8.1.

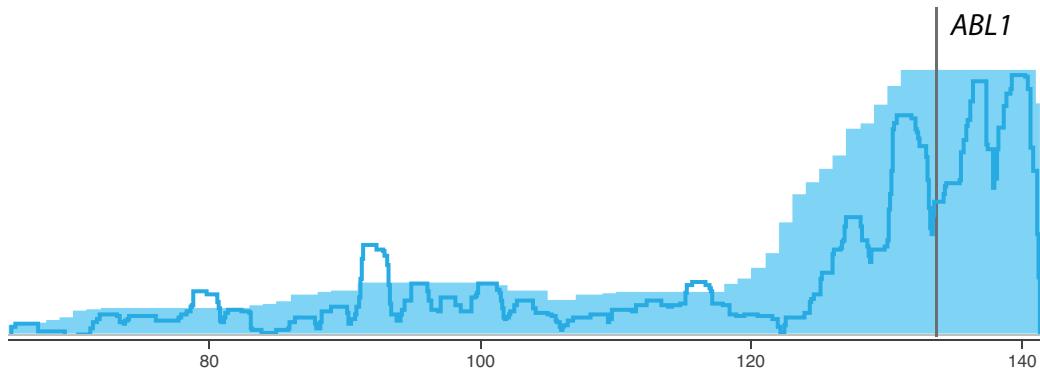


Figure 6.1.1 – BCR association with 9q34. Association of the combined *BCR* replicates, normalised for *Ase1* site distribution, across the q arm of chromosome 9 (centromere to the left, telomere to the right). Light blue fill shows quantification using ten megabase windows with one megabase separation. Darker blue line shows quantification using one megabase windows with 100 Kb separation. X axis shows chromosome position in megabases. *ABL1* is shown as a grey bar.

6.1.1 Single window testing

To confirm that this interaction is not an artefact of bias introduced through GC content or restriction fragment length, I used the single-window testing method described in Section 4.4. The script was run with one million loops to generate one million *in-silico* single window scores. The resulting frequency of each number of *in-silico* scores is plotted in Fig 6.1.2. The observed hit count for this region is clearly far above what would be expected from GC content and fragment length of the region. To generate a probability score for the number of observed fragments, I ran the script a second time for ten trillion iterations. None of the resulting scores were greater than or equal to the number of observed fragments, meaning that $p < 3 \times 10^{-13}$ for this region.

6.2 ABL e4C associations in GM12878 cells

If *BCR* preferentially contacts regions in telomeric chromosome 9 we should be able to detect reciprocal ligation products by using a bait region within this region. To investigate the interaction profile of this region I made e4C libraries with a bait in the *ABL1* promoter in GM12878 cells. This cell line is one of the top tier ENCODE project cell types and has numerous publicly accessible epigenomic datasets

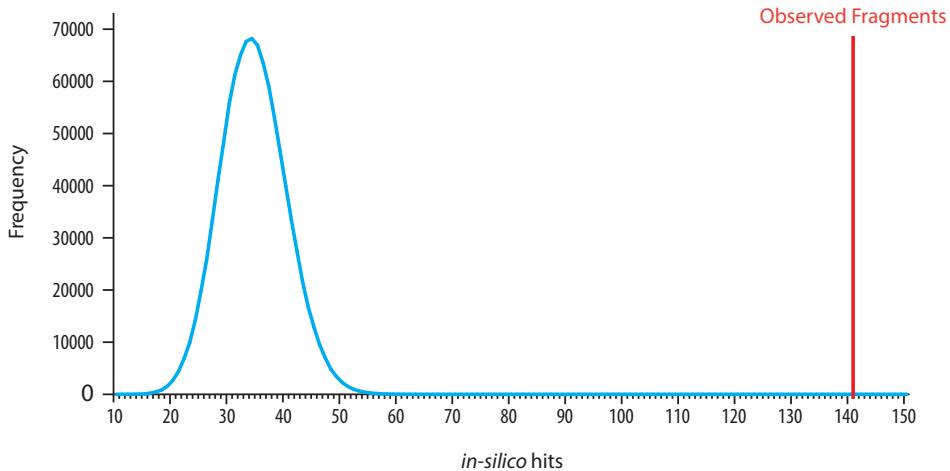


Figure 6.1.2 – Frequency of *in-silico* hits for single window in telomeric Chromosome 9.

available.

As described in chapter 5, the ABL e4C library did not reach the level of coverage of either of the BCR e4C libraries, with approximately five fold less unique reads. The ABL e4C library shows stronger correlation with the BCR libraries than any of the MLL libraries, suggesting that these two regions have similar genome-wide interactions. When visually inspected, the *trans* interaction plot of the e4C library has somewhat sharp peaks of association due to the lack of complexity. While the *BCR* locus on chromosome 22 is enriched, it is not the strongest interaction genome-wide. This could be due to the low coverage not providing enough resolution to pick out specific interactions, a lack of association between these regions in GM12878 cells, or it could be due to the lack of a direct specific interaction between the *BCR* and *ABL* genes.

When the BCR e4C libraries are combined and the *trans* association plotted with ten fold smaller one megabase windows separated by 100 Kb, a triplet peak can be seen to emerge over the 9q34 locus (Fig 6.1.1). The *ABL1* gene sits between two of these peaks, leading to the intriguing possibility that the *BCR-ABL1* interaction could be driven by the association of other nearby loci. The *ABL1* locus shows strong association with the entire 9q34 region, above the expected base level of *cis* interactions (Fig 5.3.1). These observations suggest that *BCR* and *ABL1* could be bystander genes brought together through specific interactions of different genes within the region. Such a model would explain the

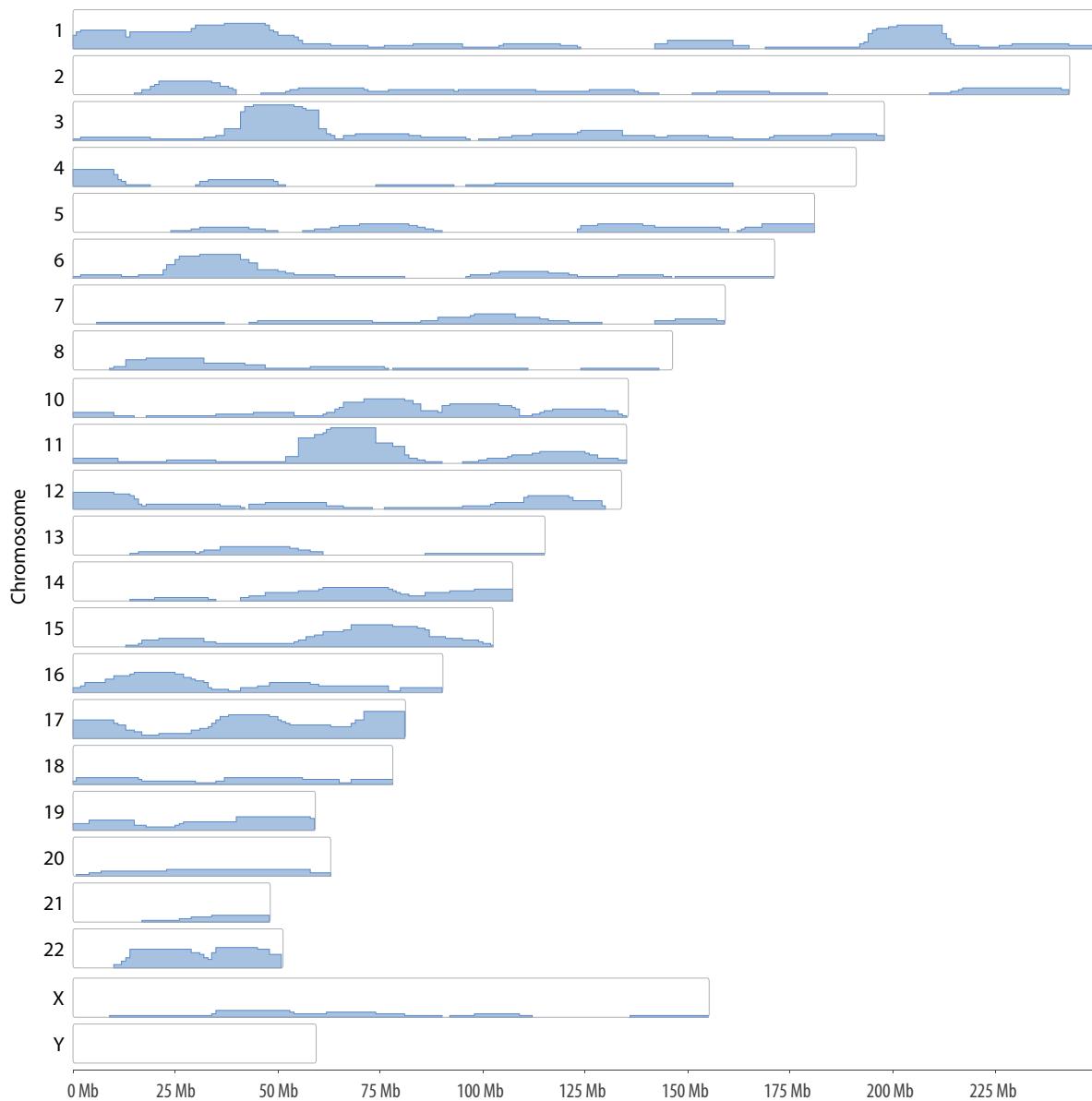


Figure 6.2.1 – ABL e4C *trans* association profile. Genome-wide plot of *ABL* e4C trans associations in GM12878 cells. Reads normalised for *Ase1* distribution with 100 Kb windows, minimum five fragments. Association scores summed in ten megabase windows separated by one megabase.

triplet peak of BCR association over the region, with ABL sitting in a trough above the base level of genome-wide *trans* association but not at the maximum point of association. It would also explain why I do not see specific *BCR-ABL1* interactions from the viewpoint of *ABL1*.

6.3 Validation using publicly available datasets

To gather additional evidence about the association between these regions, I made use of publicly available Hi-C and TCC datasets published at the time I was generating the e4C interaction data (Lieberman-Aiden et al., 2009; Kalhor et al., 2011).

6.3.1 Hi-C associations in GM06990 cells

In 2009 Dr Job Dekker's group, responsible for the first use of 3C to probe nuclear structure (Dekker et al., 2002), published a study using a new technique called Hi-C (Lieberman-Aiden et al., 2009). Hi-C is a technique based on 3C able to capture all interactions within the nucleus (all-to-all). In brief, 3C ligation products are enriched using the incorporation of a biotin moiety, adapters are added and the entire library sequenced. Hi-C libraries are incredibly complex due to the huge range of ligation events captured and require great sequencing power to achieve a useful interaction resolution.

The study by Lieberman-Aiden *et al.* investigated interactions in the ENCODE cell lines K562 and GM06990. Like GM12878 cells, the GM06990 cell line is a lymphoblastoid cell line. Lymphoblasts are multi-potent progenitor cells which later develop into the different lymphocyte cell types: NK cells, T lymphocytes and B lymphocytes. They are derived from CD34⁺ haematopoietic stem cells via common lymphoid progenitor cells. Initial analysis using the Hi-C heat map tool available at <http://hic.umassmed.edu> shows that telomeric chromosome 9 does interact strongly with two large regions of chromosome 22 with the strongest association slightly centromeric of *BCR* and either side of *ABL1* (Fig 6.3.1).

To analyse the interaction profiles in the Hi-C data in more detail, I took paired reads with one end within a one megabase region surrounding the *BCR* gene from the GM06990 *NcoI* Hi-C replicates and analysed the partner reads. I did not use the *HindIII* Hi-C replicates as they had four fold fewer reads (4395 for *HindIII* vs. 18388 for *NcoI*). I analysed the partner reads as if they were from an e4C experiment with the exception of normalisation for *NcoI* restriction site distribution. This is because the analysed reads are from the second paired end read rather than coming immediately adjacent to

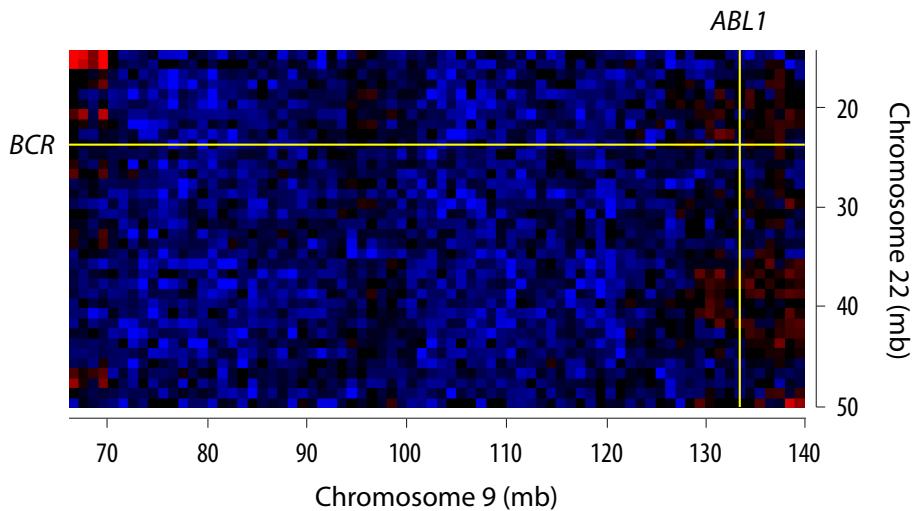


Figure 6.3.1 – GM06990 Hi-C heat map. Interaction heat map showing the q arm of chromosomes 9 and 22. Colours represent observed / expected ratios; blue colours are weak interactions, red are strong. Each block represents a one megabase window. Generated using the Hi-C data browser: <http://hic.umassmed.edu>, data from (Lieberman-Aiden et al., 2009).

the *NcoI* restriction site. This end is generated using sonication rather than the second digestion step used in e4C. As such, read counts can still be informative after removal of duplicates.

The Hi-C data has a lower signal to noise ratio, and so a flatter profile, but a peak of association over the same region of chromosome 9 is visible.

6.3.2 Hi-C and TCC associations in GM12878 cells

In 2011 a study by Kalhor *et al.* described a modified version of Hi-C called tethered chromosome capture (TCC) (Kalhor et al., 2011).

6.4 Narrowing the window - the Surfeit Cluster

To investigate the hypothesis that *BCR-ABL1* association could be driven by a bystander effect of nearby gene associations, I studied the genes present within the vicinity of *ABL1*. This region is very gene dense, containing approximately 50 to 80 genes per one megabase window.

6.4.1 Candidate genes

Description of data analysis (“narrowing the window”) with the 4C and Hi-C that led me to the Surfeit Cluster.

6.5 Validation by 3C

The results of the e4C demonstrate interaction frequencies. The analysis of the data makes a number of assumptions, one of which is that the results of the e4C library sequencing accurately represent the frequency of ligation products found within the original 3C libraries. There are a number of steps within the e4C library preparation which could affect this representation, not least the two PCR amplification steps which may preferentially amplify some ligation products irrespective of their initial frequency. Because of these potential biases the validation of interaction frequencies by traditional RT-qPCR 3C was a logical step to take. In addition to this validation, I hoped that the RT-qPCR 3C may provide a resolution of interaction frequencies at a superior resolution than the e4C, both in terms of quantitative interaction of individual fragments and in differences between fragments within close proximity.

6.5.1 Design

Primers for the qPCR were designed for a number of *AseI* fragments within the candidate genes described above, typically at both ends of fragments covering gene promoters. A control library was prepared by amplifying the fragments with genomic DNA, digesting with *AseI*, mixing in equimolar concentrations and randomly ligating.

6.5.2 Results

Results, explain why they’re not conclusive (interaction too weak to reliably quantitate using traditional 3C)

6.6 Validation by DNA-FISH

To validate the associations we see in the e4C and Hi-C data, we used DNA-Fluorescence *In-Situ* Hybridization (DNA-FISH). DNA-FISH is able to show the distance between genetic loci on a single-cell basis, and does not share any of the potential biases present in 3C technologies.

6.6.1 DNA-FISH in GM12878 cells

6.6.2 DNA-FISH in CD34⁺ cells

6.7 Discussion

Describe why this work is new, and different to what has already been shown with purely DNA-FISH
- eg. the Neves paper (not only do the genes co-localize, they are the top co-localization partners in the entire genome). Talk about the bystander effect (?).

Should this go in the discussion section instead?

Chapter 7

Discussion

7.1 Introduction / Context

Talk about how the stuff I've developed fits into wider context. Especially analysis - de Laat 4C analysis paper, Amos Tanay etc.

7.2 Limitations of the e4C data / Discussion

7.2.1 Coverage

The problems of low coverage in the libraries, and how this differs for the different bait regions. What we did to try to overcome this (using GM12878 cells, using multiple bait regions). Why this makes the data from the ABL1 and MLL baits have limited use.

7.2.2 Things I'd have done differently

Multiplexing similar libraries with barcodes (not wobble)

7.3 Sample choice

Why I'm using CD34+ cells - initiation of cancers in haematopoietic stem cells.

7.4 Association between *BCR* and *ABL1*

CML and Philadelphia Chromosome. BCR and ABL genes and proteins in normal context. Description of fusion protein and path to oncogenesis. Emphasise that lots is known about how the translocation is formed, but not why it is. Talk about DNA-FISH study looking at BCR-ABL-PMLa association through development.

7.4.1 Evidence for a bystander effect

7.5 A model for the formation of chromosomal translocations

Lay out the idea that non-random co-association at transcription factories is bringing together specific gene pairs, which in turn leads to an increased likelihood that they will form a chromosomal translocation which will go on to cause cancer. Talk about the order of events (are there other prerequisites, such as genomic instability?).

7.5.1 MLL - topoisomerases?

Go on to talk about the more detailed translocation mechanism regarding topoisomerase described by Libura et al., and how this ties in nicely with transcription and torsional stress on the DNA. Expand out to the bigger picture, and how this could be a common mechanism for all/many translocations.

7.6 Future perspectives

Go on to talk about implications - how this understanding can aid other areas of research and possibly lead to new treatments (eg. preventative drugs during topoisomerase inhibitor treatment). Talk about recent developments and where the field is going. Suggest useful future studies.

Appendix A

Primers

Appendix B

e4C Library Statistics

B.1 Numbers of reads per Chromosome

Table B.1.2 – e4C reads by Chromosome

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
1	Hits	886539	423842	1203	1228968	546017	2568	1265658	886650	931892
	Hits (dedup)	774	608	35	81	376	48	145	131	124
	AseI Frags	958 (0.90%)	833 (0.78%)	49 (0.05%)	88 (0.08%)	563 (0.53%)	71 (0.07%)	164 (0.15%)	141 (0.13%)	150 (0.14%)
2	Hits	702347	373986	2101	885876	593764	3120	865147	797590	735614
	Hits (dedup)	663	551	32	89	372	51	97	130	116
	AseI Frags	875 (0.70%)	755 (0.60%)	49 (0.04%)	109 (0.09%)	503 (0.40%)	71 (0.06%)	107 (0.09%)	140 (0.11%)	140 (0.11%)
3	Hits	789199	338050	651	465938	519108	2837	748604	896418	832899
	Hits (dedup)	661	520	26	51	326	39	101	91	130
	AseI Frags	794 (0.75%)	686 (0.65%)	43 (0.04%)	61 (0.06%)	453 (0.43%)	58 (0.05%)	120 (0.11%)	92 (0.09%)	144 (0.14%)
	Hits	671498	330078	2153	473346	443583	2299	685665	528751	581340

Numbers of reads per Chromosome

e4C Library Statistics

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
4	Hits (dedup)	600	458	37	63	331	50	67	91	103
	AseI Frags	728 (0.63%)	623 (0.54%)	56 (0.05%)	69 (0.06%)	474 (0.41%)	75 (0.06%)	83 (0.07%)	111 (0.10%)	128 (0.11%)
5	Hits	501640	280020	898	951755	443227	2453	431719	481299	740173
	Hits (dedup)	528	394	22	78	273	41	72	77	105
	AseI Frags	658 (0.67%)	525 (0.53%)	29 (0.03%)	91 (0.09%)	374 (0.38%)	58 (0.06%)	70 (0.07%)	83 (0.08%)	137 (0.14%)
6	Hits	494300	319686	1332	727194	367236	1890	495133	727477	471205
	Hits (dedup)	505	443	38	40	279	34	70	102	84
	AseI Frags	632 (0.69%)	589 (0.64%)	59 (0.06%)	47 (0.05%)	383 (0.42%)	51 (0.06%)	79 (0.09%)	112 (0.12%)	103 (0.11%)
7	Hits	476244	212755	749	688327	280344	1331	453340	624651	486517
	Hits (dedup)	431	331	21	57	212	31	48	79	78
	AseI Frags	519 (0.66%)	452 (0.57%)	33 (0.04%)	55 (0.07%)	285 (0.36%)	40 (0.05%)	67 (0.09%)	94 (0.12%)	94 (0.12%)
8	Hits	416367	266614	724	533817	203668	1200	437464	229088	415457
	Hits (dedup)	439	346	20	47	210	22	41	63	71
	AseI Frags	528 (0.70%)	456 (0.61%)	29 (0.04%)	53 (0.07%)	295 (0.39%)	35 (0.05%)	44 (0.06%)	77 (0.10%)	79 (0.11%)
9	Hits	455003	227442	457	210815	377310	2214	14089114	310032	407339
	Hits (dedup)	395	310	16	22	209	32	834	60	61
	AseI Frags	477 (0.82%)	396 (0.68%)	22 (0.04%)	22 (0.04%)	285 (0.49%)	48 (0.08%)	672 (1.15%)	63 (0.11%)	75 (0.13%)
10	Hits	458043	204292	247	413720	205262	964	671827	203188	319160
	Hits (dedup)	409	317	17	35	171	19	66	42	72
	AseI Frags	480 (0.77%)	421 (0.68%)	24 (0.04%)	32 (0.05%)	237 (0.38%)	29 (0.05%)	74 (0.12%)	47 (0.08%)	86 (0.14%)
11	Hits	422600	239318	6851	4307812	2510508	12945	892290	14981823	13145547
	Hits (dedup)	425	338	94	196	409	145	73	790	858
	AseI Frags	522 (0.83%)	462 (0.74%)	121 (0.19%)	185 (0.29%)	480 (0.76%)	186 (0.30%)	73 (0.12%)	623 (0.99%)	776 (1.24%)
12	Hits	587965	245711	713	930562	605475	2942	978019	480184	620858
	Hits (dedup)	452	355	20	43	211	29	74	66	84

e4C Library Statistics

Numbers of reads per Chromosome

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
13	AseI Frags	558 (0.86%)	476 (0.73%)	29 (0.04%)	50 (0.08%)	283 (0.43%)	37 (0.06%)	77 (0.12%)	71 (0.11%)	105 (0.16%)
	Hits	334196	175512	558	257937	169930	854	139197	505106	392592
	Hits (dedup)	326	268	20	39	189	21	34	64	64
14	AseI Frags	411 (0.70%)	378 (0.65%)	31 (0.05%)	43 (0.07%)	270 (0.46%)	32 (0.05%)	39 (0.07%)	75 (0.13%)	78 (0.13%)
	Hits	322578	187374	271	58770	36651	164	627703	191766	273517
	Hits (dedup)	310	282	10	22	165	11	51	47	43
15	AseI Frags	380 (0.85%)	374 (0.83%)	17 (0.04%)	24 (0.05%)	230 (0.51%)	17 (0.04%)	60 (0.13%)	54 (0.12%)	49 (0.11%)
	Hits	373427	188154	392	190630	130416	722	732461	277002	255223
	Hits (dedup)	322	272	9	20	148	21	67	55	49
16	AseI Frags	389 (1.08%)	354 (0.98%)	12 (0.03%)	23 (0.06%)	208 (0.58%)	29 (0.08%)	82 (0.23%)	48 (0.13%)	61 (0.17%)
	Hits	226640	168317	169	73883	296458	1645	533661	221188	169482
	Hits (dedup)	216	197	3	11	139	14	50	34	43
17	AseI Frags	277 (0.97%)	272 (0.95%)	3 (0.01%)	14 (0.05%)	200 (0.70%)	18 (0.06%)	52 (0.18%)	45 (0.16%)	56 (0.20%)
	Hits	512135	299241	586	367490	68173	344	1349407	264994	507019
	Hits (dedup)	417	336	13	16	202	13	86	54	64
18	AseI Frags	503 (1.95%)	463 (1.79%)	16 (0.06%)	23 (0.09%)	303 (1.17%)	18 (0.07%)	90 (0.35%)	64 (0.25%)	74 (0.29%)
	Hits	258583	134215	23	149524	158156	895	253479	116454	145111
	Hits (dedup)	250	180	6	24	132	22	38	33	32
19	AseI Frags	306 (0.75%)	239 (0.58%)	9 (0.02%)	29 (0.07%)	190 (0.46%)	33 (0.08%)	44 (0.11%)	38 (0.09%)	35 (0.09%)
	Hits	255305	163418	34	82337	74237	305	287777	135796	353231
	Hits (dedup)	212	203	4	14	111	8	33	17	43
20	AseI Frags	249 (1.84%)	272 (2.02%)	6 (0.04%)	15 (0.11%)	159 (1.18%)	11 (0.08%)	30 (0.22%)	17 (0.13%)	50 (0.37%)
	Hits	201666	141846	706	191977	101117	642	298607	288105	346280
	Hits (dedup)	217	187	7	16	123	13	28	37	26
	AseI Frags	273 (1.22%)	263 (1.18%)	11 (0.05%)	20 (0.09%)	173 (0.77%)	22 (0.10%)	34 (0.15%)	39 (0.17%)	32 (0.14%)
	Hits	211560	96444	473	213499	106146	516	354888	54673	143351

Numbers of reads per Chromosome

e4C Library Statistics

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
22	Hits (dedup)	190	132	7	18	80	9	17	16	20
	AseI Frags	230 (1.24%)	185 (1.00%)	10 (0.05%)	23 (0.12%)	112 (0.60%)	15 (0.08%)	19 (0.10%)	18 (0.10%)	21 (0.11%)
	Hits	5151402	4148254	71	24387	138754	585	280056	251658	142409
X	Hits (dedup)	1727	1466	19	7	810	10	22	23	19
	AseI Frags	1625 (17.03%)	1531 (16.05%)	29 (0.30%)	7 (0.07%)	1019 (10.68%)	14 (0.15%)	23 (0.24%)	23 (0.24%)	27 (0.28%)
	Hits	199693	90259	559	106708	105365	547	246552	234679	217705
Y	Hits (dedup)	220	152	12	33	96	13	39	63	57
	AseI Frags	252 (0.32%)	204 (0.26%)	15 (0.02%)	36 (0.05%)	131 (0.17%)	20 (0.03%)	51 (0.07%)	85 (0.11%)	66 (0.08%)
	Hits	19003	6048	1	6	17151	212	2	19	4959
	Hits (dedup)	23	11	1	1	15	4	2	1	5
	AseI Frags	32 (0.26%)	15 (0.12%)	2 (0.02%)	2 (0.02%)	16 (0.13%)	5 (0.04%)	2 (0.02%)	2 (0.02%)	7 (0.06%)

		All Fragments	Hit Fragments	% Hit
BCR Run1	<i>trans</i>	1425596	11,031	0.77%
	near <i>cis</i>	1261	653	51.78%
	far <i>cis</i>	8279	972	11.74%
BCR Run2	<i>trans</i>	1425596	9,693	0.68%
	near <i>cis</i>	1261	654	51.86%
	far <i>cis</i>	8279	897	10.83%
MLL 1	<i>trans</i>	1372354	583	0.04%
	near <i>cis</i>	1418	64	4.51%
	far <i>cis</i>	61364	57	0.09%
MLL 2	<i>trans</i>	1372354	936	0.07%
	near <i>cis</i>	1418	89	6.28%
	far <i>cis</i>	61364	96	0.16%
MLL 3	<i>trans</i>	1372354	7146	0.52%
	near <i>cis</i>	1418	126	8.89%
	far <i>cis</i>	61364	354	0.58%
MLL 3 phiX	<i>trans</i>	1372354	807	0.06%
	near <i>cis</i>	1418	95	6.70%
	far <i>cis</i>	61364	91	0.15%
ABL	<i>trans</i>	1376812	1484	0.11%
	near <i>cis</i>	1011	340	33.63%
	far <i>cis</i>	57313	332	0.58%
MLL p1	<i>trans</i>	1372354	1539	0.11%
	near <i>cis</i>	1418	314	22.14%
	far <i>cis</i>	61364	309	0.50%
MLL p2	<i>trans</i>	1372354	1879	0.14%
	near <i>cis</i>	1418	340	23.98%
	far <i>cis</i>	61364	437	0.71%

Table B.1.1 – e4C library *Ase1* fragment statistics. Near *cis*: within 2.5 megabases either side of bait gene. Far *cis*: beyond 2.5 megabases either side of bait gene.

Appendix C

e4C analysis scripts

C.1 e4C sequence processing and trimming

Not sure if this should go in as I didn't write it?

C.2 *In-silico* restriction fragment libraries

C.3 GC content and fragment length bias detection

C.4 Systematic bias correction

C.5 Restriction fragment distribution normalisation

C.5.1 `inputminted` method

C.5.2 `LATEX` include method

C.5.3 `Lyx` include child document method

C.5.4 Copy and paste method

	Reads starting with correct barcode	Reads with second <i>Ase1</i> site (discarded)	Reads with <i>NlaIII</i> site (too short to be mapped)	Reads with <i>NlaIII</i> site (long enough to be mapped)	Reads with no <i>NlaIII</i> site (long enough to be mapped)	Aligned reads
<i>BCR</i> Run 1	22390498	85448 (0.4%)	3823824 (17.1%)	2860849 (12.8%)	18203128 (81.3%)	14931033 (66.7%) (82.0%)
<i>BCR</i> Run 2	12616122	41823 (0.3%)	1493641 (11.8%)	1146963 (9.1%)	10932276 (86.7%)	9260876 (73.4%) (84.7%)
<i>MLL</i> Run 1	55868	15 (0.03%)	29279 (52.4%)	1819 (3.3%)	26129 (46.8%)	21923 (39.2%) (83.9%)
<i>MLL</i> Run 2	16442664	45012 (0.03%)	1388307 (8.4%)	662684 (4.0%)	14950490 (90.9%)	13535278 (82.3%) (90.5%)
<i>MLL</i> Run 3	9960249	1535 (0.02%)	557742 (5.6%)	450422 (4.5%)	9368936 (94.1%)	8498056 (85.3%) (90.7%)
<i>MLL</i> Run 3 phiX	56994	321 (0.6%)	1974 (3.5%)	1983 (3.5%)	48952 (85.9%)	44194 (77.5%) (90.3%)
<i>ABL</i> Run 1	32140483	434554 (1.4%)	201658 (0.6%)	988675 (3.1%)	31082375 (96.7%)	27117770 (84.4%) (87.2%)
<i>MLL</i> Promoter 1	27548835	243399 (0.9%)	485996 (1.8%)	1124315 (4.1%)	26290533 (95.4%)	23688930 (86.0%) (90.1%)
<i>MLL</i> Promoter 2	30333263	232741 (0.8%)	1917547 (6.3%)	1875757 (6.2%)	27748869 (91.5%)	22638880 (74.6%) (81.6%)

Table C.1.1 – Library trimming statistics. Percentage of possible sequences that were successfully aligned are shown in italics.

		Total reads	Near <i>cis</i> (\leq 5 megabases)	Far <i>cis</i> ($>$ 5 megabases)	<i>trans</i>
BCR Run 1	Raw	14931033	3986287 (26.7%)	1165115 (7.8%)	9779631 (65.5%)
	De-duplicated	10714	1007 (9.4%)	873 (8.1%)	8834 (82.5%)
	Av. dups / read (<i>per 10⁶ reads</i>)	1394 (93)	3959 (265)	1335 (89)	1107 (74)
BCR Run 2	Raw	9260876	3542934 (38.3%)	605320 (6.5%)	5112622 (55.2%)
	De-duplicated	8657	935 (10.8%)	767 (8.9%)	6955 (80.3%)
	Av. dups / read (<i>per 10⁶ reads</i>)	1070 (116)	3789 (409)	789 (85)	735 (79)
MLL Run 1	Raw	21923	5030 (22.9%)	1821 (8.3%)	15072 (68.7%)
	De-duplicated	489	54 (11.0%)	40 (8.2%)	395 (80.8%)
	Av. dups / read (<i>per 10⁶ reads</i>)	45 (2045)	93 (4249)	46 (2077)	38 (1740)
MLL Run 2	Raw	13535278	3236378 (23.9%)	1071434 (7.9%)	9227466 (68.2%)
	De-duplicated	1023	107 (10.5%)	89 (8.7%)	827 (80.8%)
	Av. dups / read (<i>per 10⁶ reads</i>)	13231 (978)	30247 (2235)	12039 (889)	11158 (824)
MLL Run 3	Raw	8498056	1701901 (20.0%)	808607 (9.5%)	5987548 (70.5%)
	De-duplicated	5589	140 (2.5%)	269 (4.8%)	5180 (92.7%)
	Av. dups / read (<i>per 10⁶ reads</i>)	1520 (179)	12156 (1430)	3006 (354)	1156 (136)
MLL Run 3 phiX	Raw	44194	9054 (20.5%)	3891 (8.8%)	31249 (70.7%)
	De-duplicated	700	83 (11.9%)	62 (8.9%)	555 (79.3%)
	Av. dups / read (<i>per 10⁶ reads</i>)	63 (1429)	109 (2468)	63 (1420)	56 (1274)
ABL Run 1	Raw	27304920	10005936 (36.6%)	4121827 (15.1%)	13177157 (43.8%)
	De-duplicated	2170	512 (23.6%)	331 (15.3%)	1327 (61.2%)
	Av. dups / read (<i>per 10⁶ reads</i>)	12583 (461)	19543 (1953)	12453 (3021)	9930 (754)
MLL Promoter 1	Raw	23693130	11118007 (46.9%)	3868012 (16.3%)	8707111 (36.7%)
	De-duplicated	2171	467 (21.5%)	324 (14.9%)	1380 (63.6%)
	Av. dups / read (<i>per 10⁶ reads</i>)	10913 (461)	23807 (2141)	11938 (3086)	6310 (725)
MLL Promoter 2	Raw	22727268	8803028 (38.7%)	4359567 (19.2%)	9564673 (42.1%)
	De-duplicated	2439	470 (19.3%)	397 (16.3%)	1572 (64.5%)
	Av. dups / read (<i>per 10⁶ reads</i>)	9318 (410)	18730 (2128)	10981 (2519)	6084 (636)

Table C.1.2 – e4C library read counts. Percentages in brackets show proportion of total number of reads. Italicised numbers in brackets show the duplicates normalised to the total number of reads in that region for easier comparison between libraries.

Appendix D

External datasets used

D.1 Datasets used for active mark correlations

Accession codes for publicly available datasets:

External datasets used

Datasets used for active mark correlations

Name	NCBI GEO Accession Code
CD34+ Cell Datasets	
RNA-Seq	GSM651554
DNase	GSM595917
H3K27ac	GSM772870, GSM772885, GSM772894
H3K36me3	GSM486705, GSM486714
H3K4me1	GSM486707, GSM486708
H3K4me3	GSM486709, GSM486711
H3K9me3	GSM537663
H3K27me3	GSM537649
ChIP-Seq Input	GSM822292
GM12878 Cells	
GM12878 RNAP II	GSE19550
Other Cell Types	
K562 RNAP II	GSM325933
HeLa RNAP II	GSM320734
NB4 RNAP II	GSM325935

Table D.1.1 – Accession codes

Appendix E

Hi-C

Appendix F

Other Perl scripts and R scripts used

Appendix G

Web Tools

G.1 Publicly available tools I've written - Cytobands, Genome RE Sites, Sequences. Mention traffic.

G.2 Talk about contribution to FastQC

Appendix H

Publications

Meet the neighbours: tools to dissect nuclear structure and function

Osborne CS, **Ewels PA**, Young AN

Briefings in Functional Genomics (2011) 10(1), 11-7

References

Bibliography

- Adam, M., F. Robert, M. Laroche, and L. Gaudreau: 2001, 'H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions.'. *Molecular and cellular biology* **21**(18), 6270–9.
- Alberts Lewis, Raff, Roberts, Walter, J.: 2007, *Molecular Biology of the Cell*. Garland Science, 5 edition.
- Arib, G. and A. Akhtar: 2011, 'Multiple facets of nuclear periphery in gene expression control.'. *Current opinion in cell biology* **23**(3), 346–53.
- Armstrong, S. a., J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer: 2002, 'MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.'. *Nature genetics* **30**(1), 41–7.
- Arnold, A. M. and J. M. Whitehouse: 1981, 'Etoposide: a new anti-cancer agent'. *Lancet* **2**(8252), 912–915.
- Aten, J. A., J. Stap, P. M. Krawczyk, C. H. van Oven, R. A. Hoebe, J. Essers, and R. Kanaar: 2004, 'Dynamics of DNA double-strand breaks revealed by clustering of damaged chromosome domains'. *Science* **303**(5654), 92–95.
- Banerji, J., S. Rusconi, and W. Schaffner: 1981, 'Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.'. *Cell* **27**(2 Pt 1), 299–308.
- Bartlett, J., J. Blagojevic, D. Carter, C. Eskiw, M. Fromaget, C. Job, M. Shamsher, I. F. Trindade, M. Xu, and P. R. Cook: 2006, 'Specialized transcription factories.'. *Biochemical Society symposium* **181**(73), 67–75.
- Bell, a. C. and G. Felsenfeld: 2000, 'Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene.'. *Nature* **405**(6785), 482–5.
- Bell, a. C., a. G. West, and G. Felsenfeld: 1999, 'The protein CTCF is required for the enhancer blocking activity of vertebrate insulators.'. *Cell* **98**(3), 387–96.
- Bender, M. a., M. Bulger, J. Close, and M. Groudine: 2000, 'Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region.'. *Molecular cell* **5**(2), 387–93.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Bernstein, B. E., J. a. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. a. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. a. Thomson: 2010, 'The NIH Roadmap Epigenomics Mapping Consortium.'. *Nature biotechnology* **28**(10), 1045–8.
- Beucher, A., J. Birraux, L. Tchouandong, O. Barton, A. Shibata, S. Conrad, A. a. Goodarzi, A. Krempler, P. a. Jeggo, and M. Löbrich: 2009, 'ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2.'. *The EMBO journal* **28**(21), 3413–27.
- Biondi, A., G. Cimino, R. Pieters, and C. H. Pui: 2000, 'Biological and therapeutic aspects of infant leukemia'. *Blood* **96**(1), 24–33.
- Birke, M., S. Schreiner, M. P. Garcia-Cuellar, K. Mahr, F. Titgemeyer, and R. K. Slany: 2002, 'The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation'. *Nucleic Acids Res* **30**(4), 958–965.
- Blumenthal, T., D. Evans, C. D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W. L. Chiu, K. Duke, M. Kiraly, and S. K. Kim: 2002, 'A global analysis of *Caenorhabditis elegans* operons.'. *Nature* **417**(6891), 851–4.
- Bolzer, A., G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer: 2005, 'Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes.'. *PLoS biology* **3**(5), e157.
- Bonnet, D. and J. E. Dick: 1997, 'Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell'. *Nature Medicine* **3**(7), 730–737.
- Boutanaev, A. M., A. I. Kalmykova, Y. Y. Shevelyov, and D. I. Nurminsky: 2002, 'Large clusters of co-expressed genes in the *Drosophila* genome.'. *Nature* **420**(6916), 666–9.
- Branco, M. R. and A. Pombo: 2006, 'Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations.'. *PLoS biology* **4**(5), e138.
- Bridger, J. M., S. Boyle, I. R. Kill, and W. a. Bickmore: 2000, 'Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts.'. *Current biology : CB* **10**(3), 149–52.
- Bridger, J. M., H. Herrmann, C. Münkel, and P. Lichter: 1998, 'Identification of an interchromosomal compartment by polymerization of nuclear-targeted vimentin.'. *Journal of cell science* **111** (Pt 9), 1241–53.
- Broeker, P. L., H. G. Super, M. J. Thirman, H. Pomykala, Y. Yonebayashi, S. Tanabe, N. Zeleznik-Le, and J. D. Rowley: 1996, 'Distribution of 11q23 breakpoints within the MLL breakpoint cluster region in de novo acute leukemia and in treatment-related acute myeloid leukemia: correlation with scaffold attachment regions and topoisomerase II consensus binding sites'. *Blood* **87**(5), 1912–1922.
- Buck, G. R. and E. L. Zechiedrich: 2004, 'DNA disentangling by type-2 topoisomerases'. *Journal of Molecular Biology* **340**(5), 933–939.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Burden, D. A. and N. Osheroff: 1998, 'Mechanism of action of eukaryotic topoisomerase II and drugs targeted to the enzyme'. *Biochimica et Biophysica Acta* **1400**(1-3), 139–154.
- Burt, R. K., Y. Loh, W. Pearce, N. Beohar, W. G. Barr, R. Craig, Y. Wen, J. a. Rapp, and J. Kessler: 2008, 'Clinical applications of blood-derived and marrow-derived stem cells for nonmalignant diseases.'. *JAMA : the journal of the American Medical Association* **299**(8), 925–36.
- Cao, R., L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang: 2002, 'Role of histone H3 lysine 27 methylation in Polycomb-group silencing.'. *Science (New York, N.Y.)* **298**(5595), 1039–43.
- Capelson, M. and M. W. Hetzer: 2009, 'The role of nuclear pores in gene regulation, development and disease.'. *EMBO reports* **10**(7), 697–705.
- Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, R. van Asperen, K. Boon, P. a. Voûte, S. Heisterkamp, a. van Kampen, and R. Versteeg: 2001, 'The human transcriptome map: clustering of highly expressed genes in chromosomal domains.'. *Science (New York, N.Y.)* **291**(5507), 1289–92.
- Carter, D., L. Chakalova, C. S. Osborne, Y.-f. Dai, and P. Fraser: 2002, 'Long-range chromatin regulatory interactions in vivo.'. *Nature genetics* **32**(4), 623–6.
- Caspersson, T., L. Zech, and C. Johansson: 1970, 'Analysis of human metaphase chromosome set by aid of DNA-binding fluorescent agents.'. *Experimental cell research* **62**(2), 490–2.
- Chakalova, L., E. Debrand, J. a. Mitchell, C. S. Osborne, and P. Fraser: 2005, 'Replication and transcription: shaping the landscape of the genome.'. *Nature reviews. Genetics* **6**(9), 669–77.
- Chambeyron, S. and W. a. Bickmore: 2004, 'Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription.'. *Genes & development* **18**(10), 1119–30.
- Chen, C. S., P. H. Sorensen, P. H. Domer, G. H. Reaman, S. J. Korsmeyer, N. A. Heerema, G. D. Hammond, and J. H. Kersey: 1993, 'Molecular rearrangements on chromosome 11q23 predominate in infant acute lymphoblastic leukemia and are associated with specific biologic variables and poor outcome'. *Blood* **81**(9), 2386–2393.
- Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng: 2008, 'Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.'. *Cell* **133**(6), 1106–17.
- Chesterton, C. J., B. E. Coupar, and P. H. Butterworth: 1974, 'Transcription of fractionated mammalian chromatin by mammalian ribonucleic acid polymerase. Demonstration of temperature-dependent rifampicin-resistant initiation sites in euchromatin deoxyribonucleic acid.'. *The Biochemical journal* **143**(1), 73–81.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Chuang, C.-H., A. E. Carpenter, B. Fuchsova, T. Johnson, P. de Lanerolle, and A. S. Belmont: 2006, 'Long-range directional movement of an interphase chromosome site.'. *Current biology : CB* **16**(8), 825–31.
- Chubb, J. R., S. Boyle, P. Perry, and W. a. Bickmore: 2002, 'Chromatin motion is constrained by association with nuclear compartments in human cells.'. *Current biology : CB* **12**(6), 439–45.
- Chung, J. H., a. C. Bell, and G. Felsenfeld: 1997, 'Characterization of the chicken beta-globin insulator.'. *Proceedings of the National Academy of Sciences of the United States of America* **94**(2), 575–80.
- Clark, M. B., P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon, J. S. Rozowsky, M. B. Gerstein, C. Wahlestedt, Y. Hayashizaki, P. Carninci, T. R. Gingeras, and J. S. Mattick: 2011, 'The reality of pervasive transcription.'. *PLoS biology* **9**(7), e1000625; discussion e1001102.
- Collins, I., A. Weber, and D. Levens: 2001, 'Transcriptional consequences of topoisomerase inhibition.'. *Molecular and cellular biology* **21**(24), 8437–51.
- Cook, P. R.: 1999, 'The Organization of Replication and Transcription'. *Science* **284**(5421), 1790–1795.
- Cook, P. R.: 2002, 'Predicting three-dimensional genome structure from transcriptional activity.'. *Nature genetics* **32**(3), 347–52.
- Cope, N. F. and P. Fraser: 2009, 'Chromosome conformation capture.'. *Cold Spring Harbor protocols* **2009**(2), pdb.prot5137.
- Corso, A., M. Lazzarino, E. Morra, S. Merante, C. Astori, P. Bernasconi, M. Boni, and C. Bernasconi: 1995, 'Chronic myelogenous leukemia and exposure to ionizing radiation—a retrospective study of 443 patients.'. *Annals of hematology* **70**(2), 79–82.
- Cowell, I. G., Z. Sondka, K. Smith, K. C. Lee, C. M. Manville, M. Sidorchuk-Lesthurge, H. A. Rance, K. Padgett, G. H. Jackson, N. Adachi, and C. a. Austin: 2012, 'Model for MLL translocations in therapy-related leukemia involving topoisomerase II β -mediated DNA strand breaks and gene proximity.'. *Proceedings of the National Academy of Sciences of the United States of America* **109**(23), 8989–94.
- Cremer, M., J. von Hase, T. Volm, A. Brero, G. Kreth, J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer: 2001, 'Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells.'. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **9**(7), 541–67.
- Cremer, T. and C. Cremer: 2001, 'Chromosome territories, nuclear architecture and gene regulation in mammalian cells.'. *Nature reviews. Genetics* **2**(4), 292–301.
- Cremer, T., C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn: 1982, 'Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments.'. *Human genetics* **60**(1), 46–56.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Cremer, T., A. Kurz, R. Zirbel, S. Dietzel, B. Rinke, E. Schrock, M. Speicher, U. Mathieu, A. Jauch, P. Emmerich, H. Scherthan, T. Ried, C. Cremer, and P. Lichter: 1993, 'Role of Chromosome Territories in the Functional Compartmentalization of the Cell Nucleus'. *Cold Spring Harbor Symposia on Quantitative Biology* **58**(0), 777–792.
- Croft, J. A., J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore: 1999, 'Differences in the localization and morphology of chromosomes in the human nucleus'. *Journal of Cell Biology* **145**(6), 1119–1131.
- Darai-Ramqvist, E., A. Sandlund, S. Müller, G. Klein, S. Imreh, and M. Kost-Alimova: 2008, 'Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions.'. *Genome research* **18**(3), 370–9.
- Davey, C. a., D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond: 2002, 'Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution.'. *Journal of molecular biology* **319**(5), 1097–113.
- de Laat, W. and F. Grosveld: 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**(5), 447–59.
- Deininger, M. W., J. M. Goldman, and J. V. Melo: 2000, 'The molecular biology of chronic myeloid leukemia.'. *Blood* **96**(10), 3343–56.
- Dekker, J., K. Rippe, M. Dekker, and N. Kleckner: 2002, 'Capturing chromosome conformation'. *Science* **295**(5558), 1306–1311.
- Dietzel, S., K. Zolghadr, C. Hepperger, and A. S. Belmont: 2004, 'Differential large-scale chromatin compaction and intranuclear positioning of transcribed versus non-transcribed transgene arrays containing beta-globin regulatory sequences.'. *Journal of cell science* **117**(Pt 19), 4603–14.
- Dorsett, D.: 1999, 'Distant liaisons: long-range enhancer-promoter interactions in Drosophila.'. *Current opinion in genetics & development* **9**(5), 505–14.
- Dunleavy, E. M., D. Roche, H. Tagami, N. Lacoste, D. Ray-Gallet, Y. Nakamura, Y. Daigo, Y. Nakatani, and G. Almouzni-Pettinotti: 2009, 'HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres.'. *Cell* **137**(3), 485–97.
- Ernst, J. and M. Kellis: 2010, 'Discovery and characterization of chromatin states for systematic annotation of the human genome.'. *Nature biotechnology* **28**(8), 817–25.
- Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shores, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein: 2011, 'Mapping and analysis of chromatin state dynamics in nine human cell types.'. *Nature* **473**(7345), 43–9.
- Eskiw, C. H., A. Rapp, D. R. F. Carter, and P. R. Cook: 2008, 'RNA polymerase II activity is located on the surface of protein-rich transcription factories.'. *Journal of cell science* **121**(Pt 12), 1999–2007.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Ferreira, J. a., G. Paolella, C. Ramos, and A. I. Lamond: 1997, 'Spatial organization of large-scale chromatin domains in the nucleus: a magnified view of single chromosome territories.'. *The Journal of cell biology* **139**(7), 1597–610.
- Ficz, G., M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. a. Hore, C. J. Marques, S. Andrews, and W. Reik: 2011, 'Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation.'. *Nature* **473**(7347), 398–402.
- Filippova, G. N., S. Fagerlie, E. M. Klenova, C. Myers, Y. Dehner, G. Goodwin, P. E. Neiman, S. J. Collins, and V. V. Lobanenkov: 1996, 'An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.'. *Molecular and cellular biology* **16**(6), 2802–13.
- Finlan, L. E., D. Sproul, I. Thomson, S. Boyle, E. Kerr, P. Perry, B. Ylstra, J. R. Chubb, and W. A. Bickmore: 2008, 'Recruitment to the nuclear periphery can alter expression of genes in human cells'. *PLoS Genet* **4**(3), e1000039.
- Fliceck, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovčová, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle: 2012, 'Ensembl 2012.'. *Nucleic acids research* **40**(Database issue), D84–90.
- Foltz, D. R., L. E. T. Jansen, A. O. Bailey, J. R. Yates, E. a. Bassett, S. Wood, B. E. Black, and D. W. Cleveland: 2009, 'Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP.'. *Cell* **137**(3), 472–84.
- Fuss, S. H., M. Omura, and P. Mombaerts: 2007, 'Local and cis effects of the H element on expression of odorant receptor genes in mouse.'. *Cell* **130**(2), 373–84.
- Gavrilov, A. a., I. S. Zukher, E. S. Philonenko, S. V. Razin, and O. V. Iarovaia: 2010, 'Mapping of the nuclear matrix-bound chromatin hubs by a new M3C experimental procedure.'. *Nucleic acids research* **38**(22), 8051–60.
- Geary, C. G.: 2000, 'The story of chronic myeloid leukaemia.'. *British journal of haematology* **110**(1), 2–11.
- Goren, A. and H. Cedar: 2003, 'Replicating by the clock.'. *Nature reviews. Molecular cell biology* **4**(1), 25–32.
- Grande, M. a., I. van der Kraan, L. de Jong, and R. van Driel: 1997, 'Nuclear distribution of transcription factors in relation to sites of transcription and RNA polymerase II.'. *Journal of cell science* **110** (Pt 1), 1781–91.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Gu, W., F. Zhang, and J. R. Lupski: 2008, 'Mechanisms for human genomic rearrangements.'. *PathoGenetics* **1**(1), 4.
- Gu, Y., H. Alder, T. Nakamura, S. A. Schichman, R. Prasad, O. Canaani, H. Saito, C. M. Croce, and E. Canaani: 1994, 'Sequence analysis of the breakpoint cluster region in the ALL-1 gene involved in acute leukemia'. *Cancer Research* **54**(9), 2327–2330.
- Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel: 2008, 'Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.'. *Nature* **453**(7197), 948–51.
- Haber, J. E.: 2000, 'Partners and pathways repairing a double-strand break.'. *Trends in genetics : TIG* **16**(6), 259–64.
- Hadjur, S., L. M. Williams, N. K. Ryan, B. S. Cobb, T. Sexton, P. Fraser, A. G. Fisher, and M. Merkenschlager: 2009, 'Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus.'. *Nature* **460**(7253), 410–3.
- Hagstrom, K. a. and B. J. Meyer: 2003, 'Condensin and cohesin: more than chromosome compactor and glue.'. *Nature reviews. Genetics* **4**(7), 520–34.
- Hamburger, A. W. and S. E. Salmon: 1977, 'Primary bioassay of human tumor stem cells'. *Science* **197**(4302), 461–463.
- Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. H. Lee, C. Ye, J. L. H. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. Sung, Y. Ruan, and C.-L. Wei: 2011, 'CTCF-mediated functional chromatin interactome in pluripotent cells.'. *Nature genetics* **43**(7), 630–8.
- Harewood, L., F. Schütz, S. Boyle, P. Perry, M. Delorenzi, W. a. Bickmore, and A. Reymond: 2010, 'The effect of translocation-induced nuclear reorganization on gene expression.'. *Genome research* **20**(5), 554–64.
- Hark, a. T., C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman: 2000, 'CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus.'. *Nature* **405**(6785), 486–9.
- Hehlmann, R., A. Hochhaus, and M. Baccarani: 2007, 'Chronic myeloid leukaemia.'. *Lancet* **370**(9584), 342–50.
- Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. a. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren: 2007, 'Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.'. *Nature genetics* **39**(3), 311–8.
- Heitz, E.: 1928, 'Das heterochromatin der Moose.'. *I Jahrb Wiss Botanik* **69**, 762–818.
- Hess, J. L., B. D. Yu, B. Li, R. Hanson, and S. J. Korsmeyer: 1997, 'Defects in yolk sac hematopoiesis in Mll-null embryos'. *Blood* **90**(5), 1799–1806.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Hlatky, L., R. K. Sachs, M. Vazquez, and M. N. Cornforth: 2002, 'Radiation-induced chromosome aberrations: insights gained from biophysical modeling.'. *BioEssays : news and reviews in molecular, cellular and developmental biology* **24**(8), 714–23.
- Hofmann, W., B. Reichart, a. Ewald, E. Müller, I. Schmitt, R. H. Stauber, F. Lottspeich, B. M. Jockusch, U. Scheer, J. Hauber, and M. C. Dabauvalle: 2001, 'Cofactor requirements for nuclear export of Rev response element (RRE)- and constitutive transport element (CTE)-containing retroviral RNAs. An unexpected role for actin.'. *The Journal of cell biology* **152**(5), 895–910.
- Hofmann, W. A. and P. de Lanerolle: 2006, 'Nuclear actin: to polymerize or not to polymerize.'. *The Journal of cell biology* **172**(4), 495–6.
- Hofmann, W. a., L. Stojiljkovic, B. Fuchsova, G. M. Vargas, E. Mavrommatis, V. Philimonenko, K. Kysela, J. a. Goodrich, J. L. Lessard, T. J. Hope, P. Hozak, and P. de Lanerolle: 2004, 'Actin is part of pre-initiation complexes and is necessary for transcription by RNA polymerase II.'. *Nature cell biology* **6**(11), 1094–101.
- Holley, W. R., I. S. Mian, S. J. Park, B. Rydberg, and A. Chatterjee: 2002, 'A model for interphase chromosomes and evaluation of radiation-induced aberrations.'. *Radiation research* **158**(5), 568–80.
- Hon, G., W. Wang, and B. Ren: 2009, 'Discovery and annotation of functional chromatin signatures in the human genome.'. *PLoS computational biology* **5**(11), e1000566.
- Hou, C., R. Dale, and A. Dean: 2010, 'Cell type specificity of chromatin organization mediated by CTCF and cohesin.'. *Proceedings of the National Academy of Sciences of the United States of America* **107**(8), 3651–6.
- Hu, G., D. E. Schones, K. Cui, R. Ybarra, D. Northrup, Q. Tang, L. Gattinoni, N. P. Restifo, S. Huang, and K. Zhao: 2011, 'Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1.'. *Genome research* **21**(10), 1650–8.
- Hudson, D. F., K. M. Marshall, and W. C. Earnshaw: 2009, 'Condensin: Architect of mitotic chromosomes.'. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **17**(2), 131–44.
- Huret, J. L., P. Dessen, and A. Bernheim: 2001, 'An atlas of chromosomes in hematological malignancies. Example: 11q23 and MLL partners.'. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **15**(6), 987–9.
- Iborra, F. J., a. Pombo, D. a. Jackson, and P. R. Cook: 1996, 'Active RNA polymerases are localized within discrete transcription "factories" in human nuclei.'. *Journal of cell science* **109** (Pt 6), 1427–36.
- Jackson, D. and P. Cook: 1985, 'Transcription occurs at a nucleoskeleton.'. *The EMBO Journal* **4**(4), 919.
- Jackson, D., A. Hassan, R. Errington, and P. Cook: 1993, 'Visualization of focal sites of transcription within human nuclei.'. *The EMBO Journal* **12**(3), 1059.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Jackson, D. A., F. J. Iborra, E. M. Manders, and P. R. Cook: 1998, 'Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei.'. *Molecular biology of the cell* **9**(6), 1523–36.
- Jackson, D. A., S. J. McCready, and P. R. Cook: 1981, 'RNA is synthesized at the nuclear cage'. *Nature* **292**(5823), 552–555.
- Jackson, D. a. and a. Pombo: 1998, 'Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells.'. *The Journal of cell biology* **140**(6), 1285–95.
- Jacob, F. and J. Monod: 1961, 'Genetic regulatory mechanisms in the synthesis of proteins'. *Journal of Molecular Biology* **3**(3), 318–356.
- Jakob, B., J. Splinter, M. Durante, and G. Taucher-Scholz: 2009, 'Live cell microscopy analysis of radiation-induced DNA double-strand break motion.'. *Proceedings of the National Academy of Sciences of the United States of America* **106**(9), 3172–7.
- Janga, S. C., J. Collado-Vides, and M. M. Babu: 2008, 'Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes.'. *Proceedings of the National Academy of Sciences of the United States of America* **105**(41), 15761–6.
- Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun: 2009, 'Cancer statistics, 2008.'. *CA: a cancer journal for clinicians* **58**(2), 71–96.
- Joshi, R. S., B. Piña, and J. Roca: 2012, 'Topoisomerase II is required for the production of long Pol II gene transcripts in yeast.'. *Nucleic acids research* **40**(16), 1–9.
- Ju, B.-G., V. V. Lunyak, V. Perissi, I. Garcia-Bassets, D. W. Rose, C. K. Glass, and M. G. Rosenfeld: 2006, 'A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription.'. *Science (New York, N.Y.)* **312**(5781), 1798–802.
- Jude, C. D., L. Climer, D. Xu, E. Artinger, J. K. Fisher, and P. Ernst: 2007, 'Unique and independent roles for MLL in adult hematopoietic stem cells and progenitors.'. *Cell stem cell* **1**(3), 324–37.
- Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. a. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. a. Young: 2010, 'Mediator and cohesin connect gene expression and chromatin architecture.'. *Nature* **467**(7314), 430–5.
- Kalhor, R., H. Tjong, N. Jayathilaka, F. Alber, and L. Chen: 2011, 'Genome architectures revealed by tethered chromosome conformation capture and population-based modeling.'. *Nature biotechnology* **30**(1), 90–98.
- Kalverda, B., H. Pickersgill, V. V. Shloma, and M. Fornerod: 2010, 'Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm.'. *Cell* **140**(3), 360–71.
- Khobta, A., C. Carlo-Stella, and G. Capranico: 2004, 'Specific histone patterns and acetylase/deacetylase activity at the breakpoint-cluster region of the human MLL gene'. *Cancer Research* **64**(8), 2656–2662.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Kim, T.-K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. a. Harmin, M. Laptevitz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg: 2010, 'Widespread transcription at neuronal activity-regulated enhancers.' *Nature* **465**(7295), 182–7.
- Kimura, H., Y. Tao, R. G. Roeder, and P. R. Cook: 1999, 'Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure.' *Molecular and cellular biology* **19**(8), 5383–92.
- Kleinjan, D. a., A. Seawright, S. Mella, C. B. Carr, D. a. Tyas, T. I. Simpson, J. O. Mason, D. J. Price, and V. van Heyningen: 2006, 'Long-range downstream enhancers are essential for Pax6 expression.' *Developmental biology* **299**(2), 563–81.
- Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher, and H. Singh: 2002, 'Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development'. *Science* **296**(5565), 158–162.
- Kozubek, S., E. Lukásová, A. Marecková, M. Skalníková, M. Kozubek, E. Bártová, V. Kroha, E. Krahulcová, and J. Slotová: 1999, 'The topological organization of chromosomes 9 and 22 in cell nuclei has a determinative role in the induction of t(9;22) translocations and in the pathogenesis of t(9;22) leukemias.' *Chromosoma* **108**(7), 426–35.
- Krivtsov, A. V. and S. a. Armstrong: 2007, 'MLL translocations, histone modifications and leukaemia stem-cell development.' *Nature reviews. Cancer* **7**(11), 823–33.
- Krueger, F., S. R. Andrews, and C. S. Osborne: 2011, 'Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling.' *PloS one* **6**(1), e16607.
- Kumaran, R. I. and D. L. Spector: 2008, 'A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence.' *The Journal of cell biology* **180**(1), 51–65.
- Kuroda, M., H. Tanabe, K. Yoshida, K. Oikawa, A. Saito, T. Kiyuna, H. Mizusawa, and K. Mukai: 2004, 'Alteration of chromosome positioning during adipocyte differentiation.' *Journal of cell science* **117**(Pt 24), 5897–903.
- Kurukuti, S., V. K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. Zhao, V. Lobanenkov, W. Reik, and R. Ohlsson: 2006, 'CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2.' *Proceedings of the National Academy of Sciences of the United States of America* **103**(28), 10684–9.
- Kurz, A., S. Lampel, J. E. Nickolenko, J. Bradl, A. Benner, R. M. Zirbel, T. Cremer, and P. Lichter: 1996, 'Active and inactive genes localize preferentially in the periphery of chromosome territories.' *The Journal of cell biology* **135**(5), 1195–205.
- Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, a. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor,

BIBLIOGRAPHY

BIBLIOGRAPHY

- C. Raymond, M. Rosetti, R. Santos, a. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, a. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Cleo, N. Carter, a. Coulson, R. Deadman, P. Deloukas, a. Dunham, I. Dunham, R. Durbin, L. French, D. Graham, S. Gregory, T. Hubbard, S. Humphray, a. Hunt, M. Jones, C. Lloyd, a. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, a. Mungall, R. Plumb, M. Ross, R. Showkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. a. Marra, E. R. Mardis, L. a. Fulton, a. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendt, K. D. Delehaunty, T. L. Miner, a. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, a. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. a. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, a. Fujiyama, M. Hattori, T. Yada, a. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, a. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, a. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, a. Madan, S. Qin, R. W. Davis, N. a. Federspiel, a. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. a. Evans, M. Athanasiou, R. Schultz, B. a. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. a. Bailey, a. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. a. Jones, S. Kasif, a. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, a. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, a. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, a. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, a. Felsenfeld, K. a. Wetterstrand, a. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowski: 2001, 'Initial sequencing and analysis of the human genome.'. *Nature* **409**(6822), 860–921.
- Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg: 2009, 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.'. *Genome biology* **10**(3), R25.
- Lee, G. R., P. E. Fields, T. J. Griffin, and R. a. Flavell: 2003, 'Regulation of the Th2 cytokine locus by a locus control region.'. *Immunity* **19**(1), 145–53.
- Lee, P. S., P. W. Greenwell, M. Dominska, M. Gawel, M. Hamilton, and T. D. Petes: 2009, 'A fine-structure map of spontaneous mitotic crossovers in the yeast *Saccharomyces cerevisiae*'. *PLoS genetics* **5**(3), e1000410.
- Lercher, M. J., A. O. Urrutia, and L. D. Hurst: 2002, 'Clustering of housekeeping genes provides a unified model of gene order in the human genome.'. *Nature genetics* **31**(2), 180–3.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Lettice, L. A., S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff: 2003, 'A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.'. *Human molecular genetics* **12**(14), 1725–35.
- Lettice, L. a., T. Horikoshi, S. J. H. Heaney, M. J. van Baren, H. C. van der Linde, G. J. Breedveld, M. Joosse, N. Akarsu, B. a. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S. W. Scherer, P. Heutink, R. E. Hill, and S. Noji: 2002, 'Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly.'. *Proceedings of the National Academy of Sciences of the United States of America* **99**(11), 7548–53.
- Li, Y.-Y., H. Yu, Z.-M. Guo, T.-Q. Guo, K. Tu, and Y.-X. Li: 2006, 'Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance.'. *PLoS computational biology* **2**(7), e74.
- Libura, J., D. J. Slater, C. a. Felix, and C. Richardson: 2005, 'Therapy-related acute myeloid leukemia-like MLL rearrangements are induced by etoposide in primary human CD34+ cells and remain stable after clonal expansion.'. *Blood* **105**(5), 2124–31.
- Libura, J., M. Ward, J. Solecka, and C. Richardson: 2008, 'Etoposide-initiated MLL rearrangements detected at high frequency in human primitive hematopoietic stem cells with in vitro and in vivo long-term repopulating potential.'. *European journal of haematology* **81**(3), 185–95.
- Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. a. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. a. Mirny, E. S. Lander, and J. Dekker: 2009, 'Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome'. *Science* **326**(5950), 289–293.
- Ling, J., B. Baibakov, W. Pi, B. M. Emerson, and D. Tuan: 2005, 'The HS2 enhancer of the beta-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a cis-linked globin promoter.'. *Journal of molecular biology* **350**(5), 883–96.
- Liu, L. and J. Wang: 1987, 'Supercoiling of the DNA template during transcription'. *Proceedings of the National Academy of Sciences* **84**(20), 7024.
- Lobanenkov, V. V., R. H. Nicolas, V. V. Adler, H. Paterson, E. M. Klenova, A. V. Polotskaja, and G. H. Goodwin: 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**(12), 1743–53.
- Lomvardas, S., G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland, and R. Axel: 2006, 'Interchromosomal interactions and olfactory receptor choice.'. *Cell* **126**(2), 403–13.
- Longhese, M. P., D. Mantiero, and M. Clerici: 2006, 'The cellular response to chromosome breakage.'. *Molecular microbiology* **60**(5), 1099–108.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Ma, H., J. Samarabandu, R. S. Devdhar, R. Acharya, P. C. Cheng, C. Meng, and R. Berezney: 1998, 'Spatial and temporal dynamics of DNA replication sites in mammalian cells.'. *The Journal of cell biology* **143**(6), 1415–25.
- Mahy, N. L., P. E. Perry, and W. a. Bickmore: 2002a, 'Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH.'. *The Journal of cell biology* **159**(5), 753–63.
- Mahy, N. L., P. E. Perry, S. Gilchrist, R. a. Baldock, and W. a. Bickmore: 2002b, 'Spatial organization of active and inactive genes and noncoding DNA within chromosome territories.'. *The Journal of cell biology* **157**(4), 579–89.
- Manuelidis, L.: 1985, 'Individual interphase chromosome domains revealed by in situ hybridization'. *Hum Genet* **71**(4), 288–293.
- Marshall, W. F., A. Straight, J. F. Marko, J. Swedlow, A. Dernburg, A. Belmont, A. W. Murray, D. A. Agard, and J. W. Sedat: 1997, 'Interphase chromosomes undergo constrained diffusional motion in living cells.'. *Current biology : CB* **7**(12), 930–9.
- Matutes, E., R. Morilla, N. Farahat, F. Carbonell, J. Swansbury, M. Dyer, and D. Catovsky: 1997, 'Definition of acute biphenotypic leukemia.'. *Haematologica* **82**(1), 64–6.
- McDonald, D., G. Carrero, C. Andrin, G. de Vries, and M. J. Hendzel: 2006, 'Nucleoplasmic beta-actin exists in a dynamic equilibrium between low-mobility polymeric species and rapidly diffusing populations.'. *The Journal of cell biology* **172**(4), 541–52.
- Mehta, I. S., M. Amira, A. J. Harvey, and J. M. Bridger: 2010, 'Rapid chromosome territory relocation by nuclear motor activity in response to serum removal in primary human fibroblasts'. *Genome Biol* **11**(1), R5.
- Mewborn, S. K., M. J. Puckelwartz, F. Abuisneineh, J. P. Fahrenbach, Y. Zhang, H. MacLeod, L. Dellefave, P. Pytel, S. Selig, C. M. Labno, K. Reddy, H. Singh, and E. McNally: 2010, 'Altered chromosomal positioning, compaction, and gene expression with a lamin A/C gene mutation.'. *PloS one* **5**(12), e14342.
- Meyer, C., B. Schneider, S. Jakob, S. Strehl, A. Attarbaschi, S. Schnittger, C. Schoch, M. W. J. C. Jansen, J. J. M. van Dongen, M. L. den Boer, R. Pieters, M.-G. Ennas, E. Angelucci, U. Koehl, J. Greil, F. Griesinger, U. Zur Stadt, C. Eckert, T. Szczepanski, F. K. Niggli, B. W. Schäfer, H. Kempinski, H. J. M. Brady, J. Zuna, J. Trka, L. L. Nigro, A. Biondi, E. Delabesse, E. Macintyre, M. Stanulla, M. Schrappe, O. a. Haas, T. Burmeister, T. Dingermann, T. Klingebiel, and R. Marschalek: 2006, 'The MLL recombinome of acute leukemias.'. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **20**(5), 777–84.
- Michaelis, C., R. Ciosk, and K. Nasmyth: 1997, 'Cohesins: chromosomal proteins that prevent premature separation of sister chromatids.'. *Cell* **91**(1), 35–45.
- Milne, T. A., Y. Dou, M. E. Martin, H. W. Brock, R. G. Roeder, and J. L. Hess: 2005, 'MLL associates specifically with a subset of transcriptionally active target genes'. *Proceedings of the National Academy of Sciences of the United States of America* **102**(41), 14765–14770.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Mitchell, J. a. and P. Fraser: 2008, 'Transcription factories are nuclear subcompartments that remain in the absence of transcription.'. *Genes & development* **22**(1), 20–5.
- Mitelman, F., B. Johansson, and F. Mertens: 2007, 'The impact of translocations and gene fusions on cancer causation.'. *Nature reviews. Cancer* **7**(4), 233–45.
- Murrell, A., S. Heeson, and W. Reik: 2004, 'Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops.'. *Nature genetics* **36**(8), 889–93.
- Nativio, R., K. S. Wendt, Y. Ito, J. E. Huddleston, S. Uribe-Lewis, K. Woodfine, C. Krueger, W. Reik, J.-M. Peters, and A. Murrell: 2009, 'Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus.'. *PLoS genetics* **5**(11), e1000739.
- Nelms, B. E., R. S. Maser, J. F. MacKay, M. G. Lagally, and J. H. Petrini: 1998, 'In situ visualization of DNA double-strand break repair in human fibroblasts'. *Science* **280**(5363), 590–592.
- Neves, H., C. Ramos, M. G. Da Silva, A. Parreira, and L. Parreira: 1999, 'The nuclear topography of ABL, BCR, PML, and RARalpha genes: evidence for gene proximity in specific phases of the cell cycle and stages of hematopoietic differentiation.'. *Blood* **93**(4), 1197–1207.
- Noordermeer, D., M. Leleu, E. Splinter, J. Rougemont, W. De Laat, and D. Duboule: 2011, 'The Dynamic Architecture of Hox Gene Clusters'. *Science* **334**(6053), 222–225.
- Nowell, P. and D. Hungerford: 1960, 'A minute chromosome in human chronic granulocytic leukemia'. *Science* **132**, 1497–501.
- Ohlsson, R., R. Renkawitz, and V. Lobanenkov: 2001, 'CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease.'. *Trends in genetics : TIG* **17**(9), 520–7.
- O'Neill, L. P., M. D. VerMilyea, and B. M. Turner: 2006, 'Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations.'. *Nature genetics* **38**(7), 835–41.
- Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. a. Mitchell, S. Lopes, W. Reik, and P. Fraser: 2004, 'Active genes dynamically colocalize to shared sites of ongoing transcription.'. *Nature genetics* **36**(10), 1065–71.
- Osborne, C. S., L. Chakalova, J. a. Mitchell, A. Horton, A. L. Wood, D. J. Bolland, A. E. Corcoran, and P. Fraser: 2007, 'Myc dynamically and preferentially relocates to a transcription factory occupied by Igf.'. *PLoS biology* **5**(8), e192.
- Osborne, C. S., P. A. Ewels, and A. N. C. Young: 2011, 'Meet the neighbours: tools to dissect nuclear structure and function.'. *Briefings in functional genomics* **10**(1), 11–7.
- Palstra, R.-J., B. Tolhuis, E. Splinter, R. Nijmeijer, F. Grosveld, and W. de Laat: 2003, 'The beta-globin nuclear compartment in development and erythroid differentiation.'. *Nature genetics* **35**(2), 190–4.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Papantonis, A., J. D. Larkin, Y. Wada, Y. Ohta, S. Ihara, T. Kodama, and P. R. Cook: 2010, 'Active RNA polymerases: mobile or immobile molecular machines?'. *PLoS biology* **8**(7), e1000419.
- Parada, L. a., P. G. McQueen, and T. Misteli: 2004, 'Tissue-specific spatial organization of genomes.'. *Genome biology* **5**(7), R44.
- Parada, L. A., P. G. McQueen, P. J. Munson, and T. Misteli: 2002, 'Conservation of relative chromosome positioning in normal and cancer cells.'. *Current biology : CB* **12**(19), 1692–7.
- Parelho, V., S. Hadjur, M. Spivakov, M. Leleu, S. Sauer, H. C. Gregson, A. Jarmuz, C. Canzonetta, Z. Webster, T. Nesterova, B. S. Cobb, K. Yokomori, N. Dillon, L. Aragon, A. G. Fisher, and M. Merkenschlager: 2008, 'Cohesins functionally associate with CTCF on mammalian chromosome arms.'. *Cell* **132**(3), 422–33.
- Pestic-Dragovich, L., L. Stojiljkovic, A. A. Philimonenko, G. Nowak, Y. Ke, R. E. Settlage, J. Shabanowitz, D. F. Hunt, P. Hozak, and P. de Lanerolle: 2000, 'A myosin I isoform in the nucleus.'. *Science (New York, N.Y.)* **290**(5490), 337–41.
- Pirrotta, V.: 1998, 'Polycombing the genome: PcG, trxG, and chromatin silencing.'. *Cell* **93**(3), 333–6.
- Potocki, L., W. Bi, D. Treadwell-Deering, C. M. B. Carvalho, A. Eifert, E. M. Friedman, D. Glaze, K. Krull, J. a. Lee, R. A. Lewis, R. Mendoza-Londono, P. Robbins-Furman, C. Shaw, X. Shi, G. Weissenberger, M. Withers, S. a. Yatsenko, E. H. Zackai, P. Stankiewicz, and J. R. Lupski: 2007, 'Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype.'. *American journal of human genetics* **80**(4), 633–49.
- Rabl, C.: 1885, 'Über Zelltheilung'. *Morphologisches Jahrbuch* **10**, 214–330.
- Raha, D., Z. Wang, Z. Moqtaderi, L. Wu, G. Zhong, M. Gerstein, K. Struhl, and M. Snyder: 2010, 'Close association of RNA polymerase II and many transcription factors with Pol III genes.'. *Proceedings of the National Academy of Sciences of the United States of America* **107**(8), 3639–44.
- Reddy, K. L., J. M. Zullo, E. Bertolino, and H. Singh: 2008, 'Transcriptional repression mediated by repositioning of genes to the nuclear lamina.'. *Nature* **452**(7184), 243–7.
- Reya, T., S. J. Morrison, M. F. Clarke, and I. L. Weissman: 2001, 'Stem cells, cancer, and cancer stem cells.'. *Nature* **414**(6859), 105–11.
- Roix, J. J., P. G. McQueen, P. J. Munson, L. a. Parada, and T. Misteli: 2003, 'Spatial proximity of translocation-prone gene loci in human lymphomas.'. *Nature genetics* **34**(3), 287–91.
- Rowley, J. D.: 1973, 'Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.'. *Nature* **243**(5405), 290–3.
- Rowley, J. D.: 2001, 'Chromosome translocations: dangerous liaisons revisited.'. *Nature reviews. Cancer* **1**(3), 245–50.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Rubio, E. D., D. J. Reiss, P. L. Welcsh, C. M. Disteche, G. N. Filippova, N. S. Baliga, R. Aebersold, J. a. Ranish, and A. Krumm: 2008, 'CTCF physically links cohesin to chromatin.'. *Proceedings of the National Academy of Sciences of the United States of America* **105**(24), 8309–14.
- Sadri-Vakili, G. and J.-H. J. Cha: 2006, 'Mechanisms of disease: Histone modifications in Huntington's disease.'. *Nature clinical practice. Neurology* **2**(6), 330–8.
- Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker: 2012, 'The long-range interaction landscape of gene promoters'. *Nature* **489**(7414), 109–113.
- Savageau, M. a.: 2011, 'Design of the lac gene circuit revisited.'. *Mathematical biosciences* **231**(1), 19–38.
- Schardin, M., T. Cremer, H. D. Hager, and M. Lang: 1985, 'Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories.'. *Human genetics* **71**(4), 281–7.
- Scharf, S., J. Zech, A. Bursen, D. Schraets, P. L. Oliver, S. Kliem, E. Pfitzner, E. Gillert, T. Dingermann, and R. Marschalek: 2007, 'Transcription linked to recombination: a gene-internal promoter coincides with the recombination hot spot II of the human MLL gene'. *Oncogene* **26**(10), 1361–1371.
- Schermelleh, L., P. M. Carlton, S. Haase, L. Shao, L. Winoto, P. Kner, B. Burke, M. C. Cardoso, D. a. Agard, M. G. L. Gustafsson, H. Leonhardt, and J. W. Sedat: 2008, 'Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy.'. *Science (New York, N.Y.)* **320**(5881), 1332–6.
- Schmidt, D., P. C. Schwalie, C. S. Ross-Innes, A. Hurtado, G. D. Brown, J. S. Carroll, P. Flicek, and D. T. Odom: 2010, 'A CTCF-independent role for cohesin in tissue-specific transcription.'. *Genome research* **20**(5), 578–88.
- Schneider, R. and R. Grosschedl: 2007, 'Dynamics and interplay of nuclear architecture, genome organization, and gene expression.'. *Genes & development* **21**(23), 3027–43.
- Schoenfelder, S., T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. a. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C.-L. L. Wei, Y. Ruan, J. J. Bieker, and P. Fraser: 2010, 'Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells'. *Nat Genet* **42**(1), 53–61.
- Schwacha, a. and N. Kleckner: 1997, 'Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway.'. *Cell* **90**(6), 1123–35.
- Schwartz, B. E. and K. Ahmad: 2005, 'Transcriptional activation triggers deposition and removal of the histone variant H3.3.'. *Genes & development* **19**(7), 804–14.
- Sharpe, J., L. Lettice, J. Hecksher-Sorensen, M. Fox, R. Hill, and R. Krumlauf: 1999, 'Identification of sonic hedgehog as a candidate gene responsible for the polydactylous mouse mutant Sasquatch.'. *Current biology : CB* **9**(2), 97–100.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Shimi, T., K. Pfleghaar, S.-i. Kojima, C.-G. Pack, I. Solovei, A. E. Goldman, S. a. Adam, D. K. Shumaker, M. Kinjo, T. Cremer, and R. D. Goldman: 2008, 'The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription.'. *Genes & development* **22**(24), 3409–21.
- Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat: 2006, 'Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)'. *Nat Genet* **38**(11), 1348–1354.
- Slany, R. K., C. Lavau, and M. L. Cleary: 1998, 'The oncogenic capacity of HRX-ENL requires the transcriptional transactivation activity of ENL and the DNA binding motifs of HRX.'. *Molecular and cellular biology* **18**(1), 122–9.
- Soutoglou, E., J. F. Dorn, K. Sengupta, M. Jasin, A. Nussenzweig, T. Ried, G. Danuser, and T. Misteli: 2007, 'Positional stability of single double-strand breaks in mammalian cells.'. *Nature cell biology* **9**(6), 675–82.
- Spilianakis, C. G. and R. a. Flavell: 2004, 'Long-range intrachromosomal interactions in the T helper type 2 cytokine locus.'. *Nature immunology* **5**(10), 1017–27.
- Spilianakis, C. G., M. D. Lalioti, T. Town, G. R. Lee, and R. a. Flavell: 2005, 'Interchromosomal associations between alternatively expressed loci.'. *Nature* **435**(7042), 637–45.
- Splinter, E., H. Heath, J. Kooren, R.-J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. de Laat: 2006, 'CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.'. *Genes & development* **20**(17), 2349–54.
- Stack, S. M., D. B. Brown, and W. C. Dewey: 1977, 'Visualization of interphase chromosomes.'. *Journal of cell science* **26**, 281–99.
- Stadler, S., V. Schnapp, R. Mayer, S. Stein, C. Cremer, C. Bonifer, T. Cremer, and S. Dietzel: 2004, 'The architecture of chicken chromosome territories changes during differentiation.'. *BMC cell biology* **5**(1), 44.
- Stedman, W., H. Kang, S. Lin, J. L. Kissil, M. S. Bartolomei, and P. M. Lieberman: 2008, 'Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators.'. *The EMBO journal* **27**(4), 654–66.
- Strahl, B. D. and C. D. Allis: 2000, 'The language of covalent histone modifications.'. *Nature* **403**(6765), 41–5.
- Strissel, P. L., R. Strick, J. D. Rowley, and N. J. Zeleznik-Le: 1998, 'An in vivo topoisomerase II cleavage site and a DNase I hypersensitive site colocalize near exon 9 in the MLL breakpoint cluster region'. *Blood* **92**(10), 3793–3803.
- Sun, H. B., J. Shen, and H. Yokota: 2000, 'Size-dependent positioning of human chromosomes in interphase nuclei'. *Biophysical Journal* **79**(1), 184–190.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Sung, P. a., J. Libura, and C. Richardson: 2006, 'Etoposide and illegitimate DNA double-strand break repair in the generation of MLL translocations: new insights and new questions.'. *DNA repair* **5**(9-10), 1109–18.
- Sutherland, H. and W. a. Bickmore: 2009, 'Transcription factories: gene expression in unions?'. *Nature reviews. Genetics* **10**(7), 457–66.
- Sutherland, H. J., A. Blair, and R. W. Zapf: 1996, 'Characterization of a hierarchy in human acute myeloid leukemia progenitor cells'. *Blood* **87**(11), 4754–4761.
- Szabó, P., S. H. Tang, a. Rentsendorj, G. P. Pfeifer, and J. R. Mann: 2000, 'Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function.'. *Current biology : CB* **10**(10), 607–10.
- Szczerbal, I., H. A. Foster, and J. M. Bridger: 2009, 'The spatial repositioning of adipogenesis genes is correlated with their expression status in a porcine mesenchymal stem cell adipogenesis model system'. *Chromosoma* **118**(5), 647–663.
- Talbert, P. B. and S. Henikoff: 2010, 'Histone variants—ancient wrap artists of the epigenome.'. *Nature reviews. Molecular cell biology* **11**(4), 264–75.
- Tanabe, H., F. A. Habermann, I. Solovei, M. Cremer, and T. Cremer: 2002, 'Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications'. *Mutation Research* **504**(1-2), 37–45.
- Tanaka, K., M. Takechi, J. Hong, C. Shigeta, N. Oguma, N. Kamada, Y. Takimoto, A. Kuramoto, H. Dohy, and T. Kyo: 1989, '9;22 translocation and bcr rearrangements in chronic myelocytic leukemia patients among atomic bomb survivors.'. *Journal of radiation research* **30**(4), 352–8.
- Tiwari, V. K., L. Cope, K. M. McGarvey, J. E. Ohm, and S. B. Baylin: 2008a, 'A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations.'. *Genome research* **18**(7), 1171–9.
- Tiwari, V. K., K. M. McGarvey, J. D. F. Licchesi, J. E. Ohm, J. G. Herman, D. Schübeler, and S. B. Baylin: 2008b, 'PcG proteins, DNA methylation, and gene repression by chromatin looping.'. *PLoS biology* **6**(12), 2911–27.
- Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat: 2002, 'Looping and interaction between hypersensitive sites in the active beta-globin locus'. *Mol Cell* **10**(6), 1453–1465.
- Torrano, V., I. Chernukhin, F. Docquier, V. D'Arcy, J. León, E. Klenova, and M. D. Delgado: 2005, 'CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells.'. *The Journal of biological chemistry* **280**(30), 28152–61.
- Trask, B. J.: 2002, 'Human cytogenetics: 46 chromosomes, 46 years and counting.'. *Nature reviews. Genetics* **3**(10), 769–78.
- Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers: 2004, 'An abundance of bidirectional promoters in the human genome.'. *Genome research* **14**(1), 62–6.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Vaquerizas, J. M., R. Suyama, J. Kind, K. Miura, N. M. Luscombe, and A. Akhtar: 2010, 'Nuclear pore proteins nup153 and megator define transcriptionally active regions in the Drosophila genome.'. *PLoS genetics* **6**(2), e1000846.
- Vazquez, J., A. S. Belmont, and J. W. Sedat: 2001, 'Multiple regimes of constrained chromosome motion are regulated in the interphase Drosophila nucleus.'. *Current biology : CB* **11**(16), 1227–39.
- Velagaleti, G. V. N., G. a. Bien-Willner, J. K. Northup, L. H. Lockhart, J. C. Hawkins, S. M. Jalal, M. Withers, J. R. Lupski, and P. Stankiewicz: 2005, 'Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia.'. *American journal of human genetics* **76**(4), 652–62.
- Verschure, P. J., I. van Der Kraan, E. M. Manders, and R. van Driel: 1999, 'Spatial relationship between transcription sites and chromosome territories.'. *The Journal of cell biology* **147**(1), 13–24.
- Versteeg, R., B. D. C. van Schaik, M. F. van Batenburg, M. Roos, R. Monajemi, H. Caron, H. J. Bussemaker, and A. H. C. van Kampen: 2003, 'The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.'. *Genome research* **13**(9), 1998–2004.
- Visser, a. E., R. Eils, A. Jauch, G. Little, P. J. Bakker, T. Cremer, and J. a. Aten: 1998, 'Spatial distributions of early and late replicating chromatin in interphase chromosome territories.'. *Experimental cell research* **243**(2), 398–407.
- Vogel, M. J., L. Pagie, W. Talhout, M. Nieuwland, R. M. Kerkhoven, and B. van Steensel: 2009, 'High-resolution mapping of heterochromatin redistribution in a Drosophila position-effect variegation model.'. *Epigenetics & chromatin* **2**(1), 1.
- Volpi, E. V., E. Chevret, T. Jones, R. Vatcheva, J. Williamson, S. Beck, R. D. Campbell, M. Goldsworthy, S. H. Powis, J. Ragoussis, J. Trowsdale, and D. Sheer: 2000, 'Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei.'. *Journal of cell science* **113** (Pt 9), 1565–76.
- Vostrov, A. A. and W. W. Quitschke: 1997, 'The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation.'. *The Journal of biological chemistry* **272**(52), 33353–9.
- Wendt, K. S., K. Yoshida, T. Itoh, M. Bando, B. Koch, E. Schirghuber, S. Tsutsumi, G. Nagae, K. Ishihara, T. Mishiro, K. Yahata, F. Imamoto, H. Aburatani, M. Nakao, N. Imamoto, K. Maeshima, K. Shirahige, and J.-M. Peters: 2008, 'Cohesin mediates transcriptional insulation by CCCTC-binding factor.'. *Nature* **451**(7180), 796–801.
- Wolffe, a. P. and J. J. Hayes: 1999, 'Chromatin disruption and modification.'. *Nucleic acids research* **27**(3), 711–20.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Würtele, H. and P. Chartrand: 2006, 'Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology.'. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **14**(5), 477–95.
- Xiao, T., J. Wallace, and G. Felsenfeld: 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**(11), 2174–83.
- Xie, X., T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, and E. S. Lander: 2007, 'Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.'. *Proceedings of the National Academy of Sciences of the United States of America* **104**(17), 7145–50.
- Xu, M. and P. R. Cook: 2008, 'Similar active genes cluster in specialized transcription factories.'. *The Journal of cell biology* **181**(4), 615–23.
- Yaffe, E. and A. Tanay: 2011, 'Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture'. *Nature Genetics* **43**(11), 1–9.
- Yin, H., M. D. Wang, K. Svoboda, R. Landick, S. M. Block, and J. Gelles: 1995, 'Transcription against an applied force.'. *Science (New York, N.Y.)* **270**(5242), 1653–7.
- Yu, B. D., J. L. Hess, S. E. Horning, G. A. Brown, and S. J. Korsmeyer: 1995, 'Altered Hox expression and segmental identity in Mll-mutant mice'. *Nature* **378**(6556), 505–508.
- Zeisig, B. B., T. Milne, M.-P. García-Cuéllar, S. Schreiner, M.-e. Martin, U. Fuchs, A. Borkhardt, S. K. Chanda, J. Walker, R. Soden, J. L. Hess, and R. K. Slany: 2004, 'Hoxa9 and Meis1 are key targets for MLL-ENL-mediated cellular immortalization.'. *Molecular and cellular biology* **24**(2), 617–28.
- Zeleznik-Le, N. J., A. M. Harden, and J. D. Rowley: 1994, '11q23 translocations split the "AT-hook" cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene'. *Proceedings of the National Academy of Sciences of the United States of America* **91**(22), 10610–10614.
- Zhou, G.-L., L. Xin, W. Song, L.-J. Di, G. Liu, X.-S. Wu, D.-P. Liu, and C.-C. Liang: 2006, 'Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes.'. *Molecular and cellular biology* **26**(13), 5096–105.
- Zhou, V. W., A. Goren, and B. E. Bernstein: 2011, 'Charting histone modifications and the functional organization of mammalian genomes.'. *Nature reviews. Genetics* **12**(1), 7–18.
- Ziemin-van der Poel, S., N. R. McCabe, R. Espinosa 3rd, H. J. Gill, Y. Patel, A. Harden, P. Rubinelli, S. D. Smith, M. M. LeBeau, J. D. Rowley, and E. al.: 1991, 'Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias'. *Proceedings of the National Academy of Sciences of the United States of America* **88**(23), 10735–10739.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Zink, D., M. D. Amaral, A. Englmann, S. Lang, L. a. Clarke, C. Rudolph, F. Alt, K. Luther, C. Braz, N. Sadoni, J. Rosenecker, and D. Schindelhauer: 2004, 'Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei.'. *The Journal of cell biology* **166**(6), 815–25.
- Zirbel, R. M., U. R. Mathieu, A. Kurz, T. Cremer, and P. Lichter: 1993, 'Evidence for a nuclear compartment of transcription and splicing located at chromosome domain boundaries.'. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **1**(2), 93–106.
- Zlatanova, J. and A. Thakar: 2008, 'H2A.Z: view from the top.'. *Structure (London, England : 1993)* **16**(2), 166–79.