Spatial organisation of proto-oncogenes in human hæmatopoietic progenitor cells

This dissertation is submitted for the degree of *Doctor of Philosophy*

Philip Ewels

"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

I'd like to start by thanking my supervisor, Cameron Osborne, for inviting me into his group and letting me carry out my PhD in his laboratory despite my complete lack of ability to discuss anything about football.

The Osborne lab has been a tight knit group throughout my PhD - I'd like to thank Alan Chong for teaching me my baby steps, Mayra Furlan-Magaril for helping me through some of the most difficult parts of my project, Lauren Ferreira for being a great friend as well as an excellent pair of hands, Alice Young for inspiring me to keep trying and Cam for putting up with my appalling writing style.

The Osborne ~~lab~~ *posse*.

Being a small lab I relied heavily on those across the hall, especially Stefan Schoenfelder, Daniel Bolland, Chris Eskiw, Louise Harewood, Joana Martins and Mayra. Notably there was some extreme speed-reading of this thesis for which I am very grateful. I'd like to thank the Babraham bioinformatics group for getting so good at hiding their despair at seeing my face pop around the corner, in particular Simon Andrews who has had a large role in shaping what will hopefully become a career in bioinformatics of my own. Fellow PhD students including but not limited to Nicola Stead, Catherine Moir, the other "BI Newbies", Mike Stubbington, Luke Edelman and Andrew Dimond have all become good friends without who I would have lost my sanity long ago.

A big part of my life in Cambridge has been Trinity Hall, its MCR and the mighty THBC. I'd like to say a huge thanks to those who were part of this life with me - Sophie *"Beast"* Machin, Rachel Linn, *"Little"* Phil Maltas, *"French"* Stef Jacquot, Luch *"I hate rowing"* Harland-Lang and everyone else who made it such a memorable few years. You're all cheeky devils. Row Hall!

I think it would be rude not to thank my family, who are ultimately responsible for my love of science and exploration. What with weddings and babies, these four years have represented big changes for all of us and I look forward to spending more time with all of you in the near future!

My PhD has truly shaped who I am and taught me what a life in science involves. I owe a lot to many people for that privilege.

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 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

Contents

List of Abbreviations

3C Chromosome Conformation Capture

4C Circularised chromosome conformation capture

ACH Active chromatin hub

ALL Acute Lymphoblastic Leukaemia

AML Acute Myeloid Leukaemia

BAC Bacterial artificial chromosome

BrUTP Bromouridine triphosphate

BSA Bovine serum albumin

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

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

GAIIx Illumina Genome Analyser IIx

GCSF Granulocyte colony-stimulating factor

gDNA Genomic DNA

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

PE Ad 1.0 Illumina Paired End Adapter 1.0

PRE Polycomb response element

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 revsersible 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

List of Figures

List of Tables

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 with carmine acetic acid within the nucleus (Heitz, 1928). Heterochromatin stains darkly with giemsa stains 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 can control how the sequences within are used.

1.1.1 Histones

To package DNA, the double helix is wrapped around an octamer 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.

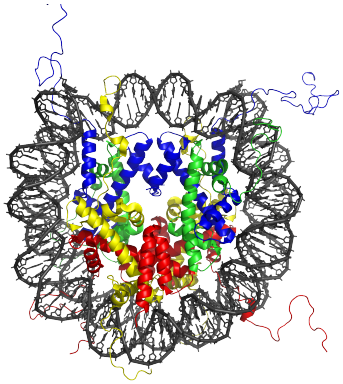


Figure 1.1.1 – Structure of the nucleosome. X-ray structure of a nucleosome core particle at a resolution of 1.9Å. DNA can be seen wrapped around the core histones, which have flexible tails extending into the nuclear matrix. PDB structure 1KX5 (Davey et al., 2002).

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 differences between 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 containing H3.3 appear to be less stable than canonical nuc-

leosomes, 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

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). Oberservations 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).

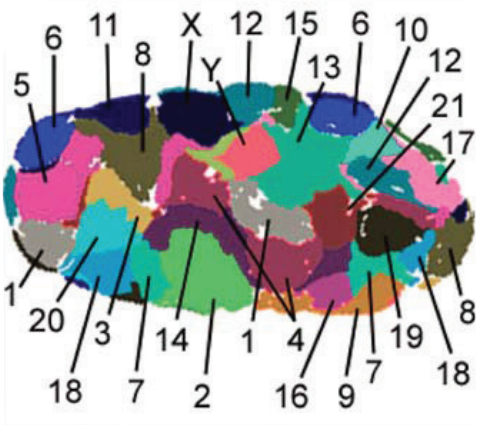


Figure 1.3.1 – Chromosome Territories. Image of a Human fibroblast cell in G0 with all chromosomes labelled using multi-colour FISH. Adapted from Bolzer *et al.* (Bolzer et al., 2005)

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; Szczerbal 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 poly-

merase II scattered throughout the nucleus. The vast majority of genic transcription appears to take place at transcription factories (Jackson et al., 1993; Osborne et al., 2004; Ragoczy et al., 2006; Eskiw et al., 2008; Schoenfelder et al., 2010), challenging the classical model of transcription found in many text books (discussed below).

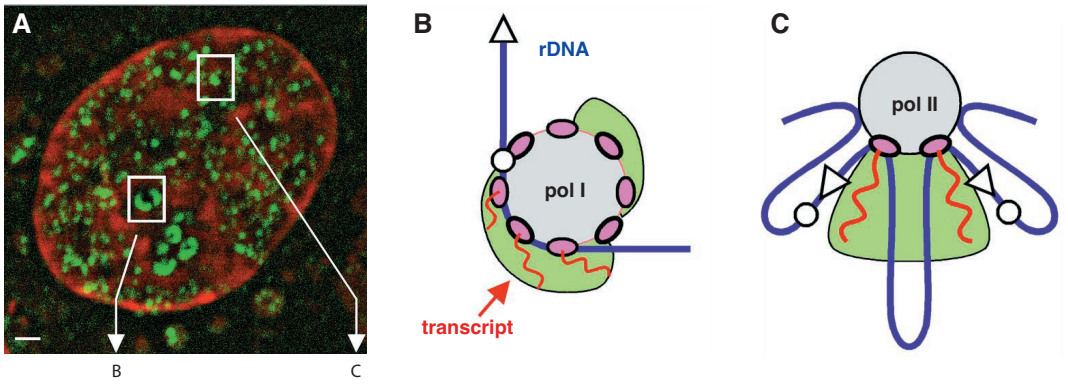


Figure 1.4.1 – Nascent RNA and transcription factories. (A) Transcription foci in HeLa cells, visualised with labelled Br-UTP in 100 nm cryosections. Nascent RNA (green) is concentrated in punctate foci. (B) Model for a nucleolar factory, showing a transcript with multiple polymerases generating a crescent 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 [^{3}H] 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 μM 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 approximately 90% of actively transcribed genes are associated with transcription factories (Osborne et al., 2004). They demonstrated 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.

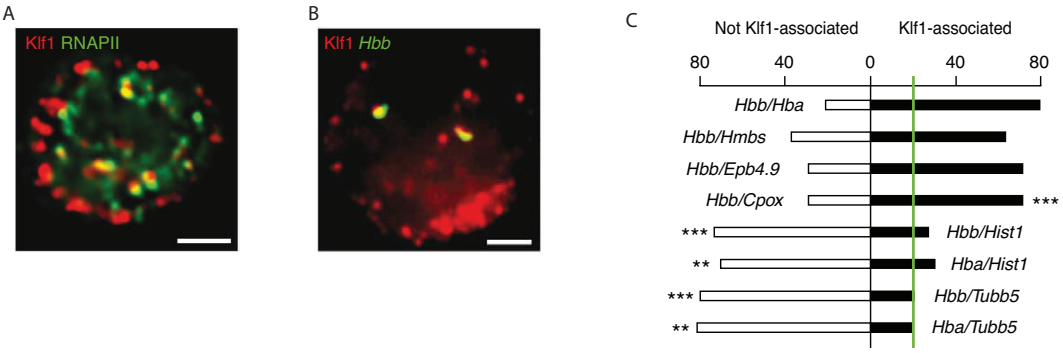


Figure 1.4.2 – Klf1 specialised transcription factories. (A) Immunofluorescence showing Klf1 (red) overlapping with a subset of RNA Polymerase II transcription factories (green). (B) RNA immuno-FISH showing transcribing *Hbb* alleles (green) co-localising with Klf1 (red). (C) Percentages of co-localising 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).

It is important to note that the evidence above does not describe *exclusively* specific transcription 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 embyro 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. Spiliankis *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 it's 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 TNFα 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 though 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 held in transcription factories, with tissue specific genes being recruited to these sites upon activation (Mitchell and Fraser, 2008; 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

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).

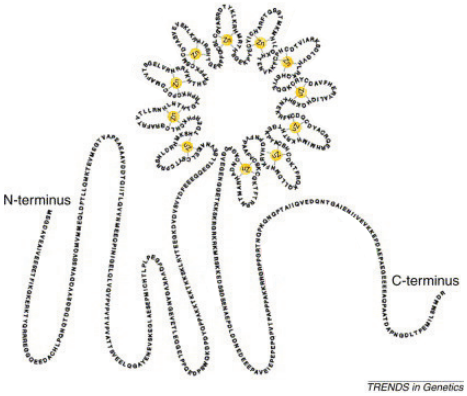


Figure 1.6.1 – Putative structure of CTCF. Adapted from Ohlsson *et al.* 2001 (Ohlsson et al., 2001).

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 an 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 be-

haviour, 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 a 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). Two papers from the Belmont and Bridger groups recently demonstrated active reorganisation of chromosome territory positioning which was dependent on the action of actin and myosin 1 (Chuang et al., 2006; Mehta et al., 2010).

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 replicated 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.

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

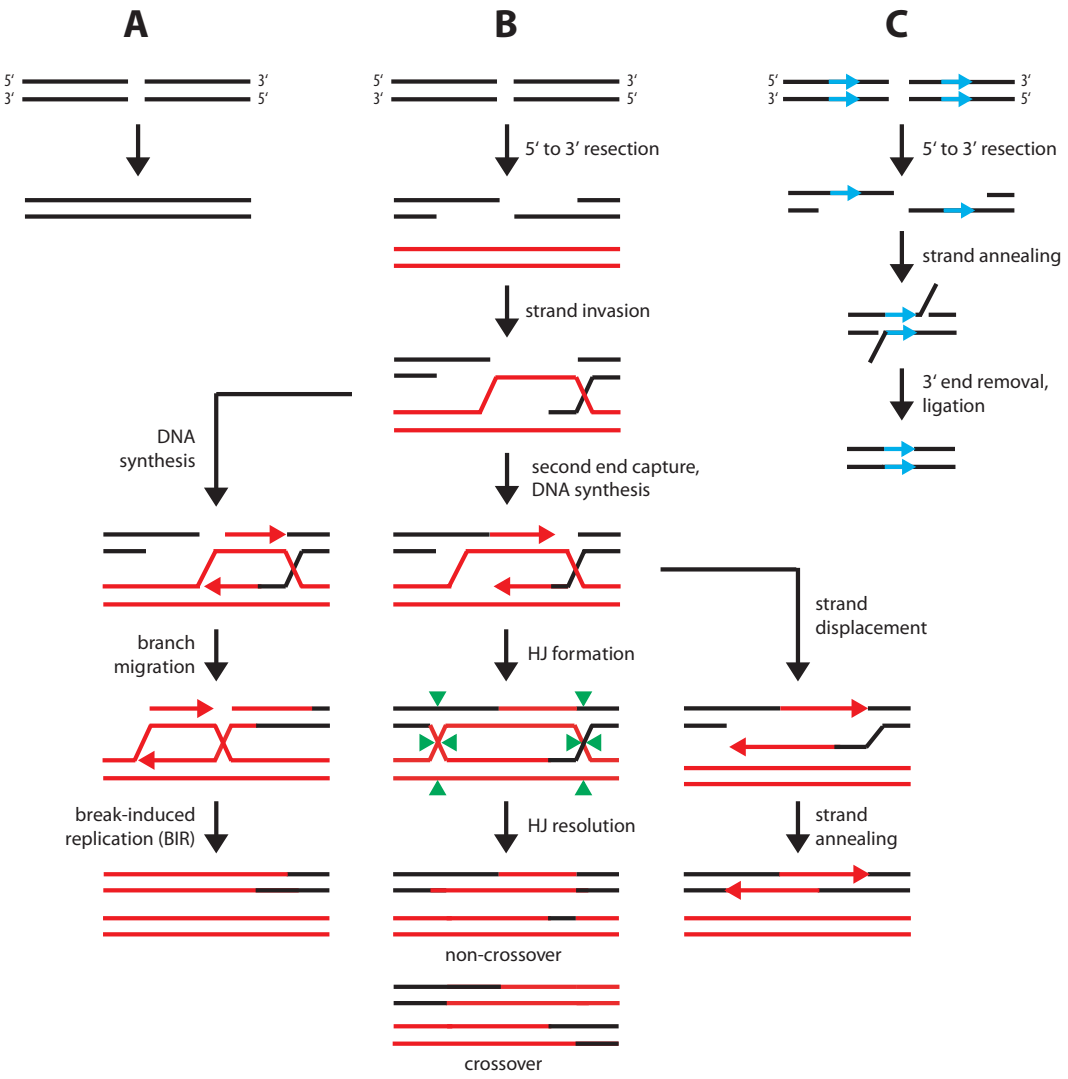


Figure 1.7.1 – **DSB repair.** Three double strand break repair pathways. **(A)** NHEJ (Non-Homologous End Joining) requires little or no sequence homology and joins any two nearby ends. **(B)** HR (Homologous Recombination) uses a strand of homologous DNA on a duplicated interphase chromosome or sister 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).

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, 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).

1.8.2.1 Leukaemia stem cells

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).

1.8.2.2 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).

1.8.2.3 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).

1.8.2.4 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).

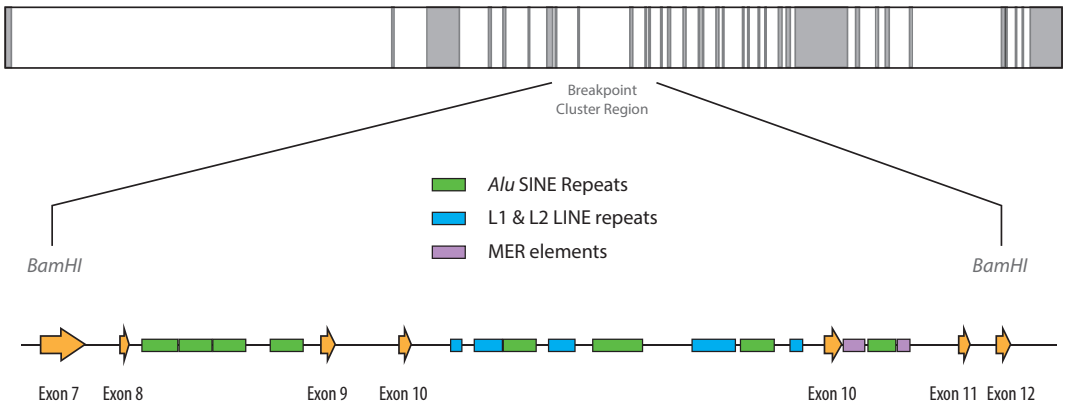


Figure 1.8.1 – Diagram of the *MLL* breakpoint cluster region.

The *MLL* breakpoint cluster region contains a number of putative topoisomerase II binding sites suggesting a mechanism of topoisomerase inhibitor related translo-

cations (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 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). The 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 chromomes 12 and 16 varies through adipocyte differentiation. These two chromosomes are involved 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).

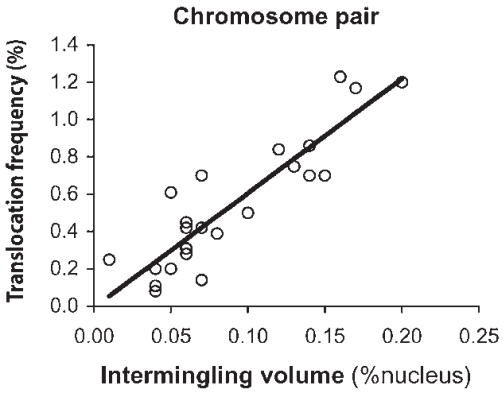


Figure 1.9.1 – Correlation of CT intermingling and radiation induced translocation frequencies. Adapted from Branco and Pombo, 2006 (Branco and Pombo, 2006).

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 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 chro-

mosomes 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 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 aliquoted 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 10^7 to 1 x 10^8 per ml in 2 ml microcentrifuge tubes. Invitrogen Dynalbeads were pipetted 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 µl 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 µl 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 µl MACS buffer per 1 x 10^8 cells. The following assumes a single aliquot of 1 x 10^8 cells, and was scaled up as necessary. Some leukapheresis samples had enough cells to warrant thousands of pounds worth of MACS beads, in these cases I used aliquots of greater than 1 x 10^8 cells.

Cell separation was carried out as per the manufacturer's instructions. In brief, 100 µl MACS FcR Blocking reagent was added to block non-specific binding of the MicroBeads before adding 100 µl 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 µl 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 µl 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 µl 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 10^6 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% foetal 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 ºC in 5% CO_2.

Cultures were passaged when cell densities reached 1 x 10^6 cells / ml, typically every other day. Cell culture media was made up if required and warmed to 37 ºC in a water bath. Cell cultures were removed from the incubators into a cell culture lam-

inar 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 10^5 and 5 x 10^5 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 x 10^6 cells in RPMI 1640 with 2mM L-glutamine (PAA; E15-840), supplemented with 20% foetal 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. GM12878 cells were fixed in 2% formaldehyde and quenched in glycine and washed in PBS twice. Cell pellets were thawed on ice and resuspended in 50 ml permeabilisation 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 $MgCl_2$, 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 cross linked.. Nuclei were incubated at 37°C for 1 hour, shaking at 950 rpm. Triton-X100 was added to a final concentration 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 cross linked chromatin 30 µl high concentration *AseI* was added (1500 units [New England Biolabs]) and samples incubated at 37°C overnight, shaking at 950 rpm. An additional 10 µl *AseI* (500 units) was added and incubated for a further 3 hours at 37°C, 950 rpm. Another 10µl 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 $MgCl_2$, 1.1 mM ATP, 11 mM DTT, 27.5 µg/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 µl 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 µg/ml and samples incubated at 65°C overnight. Tubes were cooled to room temperature and RNase A added to a final concentration of 40 µg/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 µl 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 ºC for at least 30 minutes before being equilibrated at 37 ºC. 1 µl of 1 mg/ml RNase A (1 µg

final) was then added, and samples incubated for 2 hours at 37 ºC. 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:

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

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

Quantitative analysis of relative nuclear association strengths can be done with real-time quantitative PCR of 3C products (RT-qPCR 3C). The SYBRgreen dye binds to double stranded DNA, and can be used to track the concentration of PCR products throughout the amplification. By comparing the PCR cycle at which a threshold concentration is reached (C_t), the relative starting concentrations of 3C ligation products can be calculated.

2.3.4.1 Primer design

RT-qPCR primers were designed surrounding both ends of restriction fragments of interest. Primers were designed using primer3 (Rozen and Skaletsky, 2000). Primers were designed with the following parameters: primer size - minimum 20

bp, ideal 22 bp, optimal 24 bp; primer T_m - minimum 60 °C, optimal 62 °C, maximum 64 °C; primer GC% - minimum 30%, optimal 50%, maximum 70%; product size ~ 150-230 bp; difference in T_m between primers - 1.5 °C. Chosen primers were analysed using Primera to check for secondary structure and primer dimer potential. Primers were then blasted using ensembl to check for unique binding. Primers were ordered from Sigma-Aldrich as dehydrated and deionised oligos before being resuspended in molecular biology grade water [Qiagen] at a concentration of 100 μM for stock and a 10 μM working dilution.

2.3.4.2 Equimolar mix

An equimolar mix is used with RT-qPCR 3C to standardise primer efficiencies and quantify products using a standard curve. Primer pairs were used to amplify genomic DNA using Qiagen HotStar Taq. Products were cleaned using a Qiagen PCR cleanup kit and run on a 1% agarose gel to confirm sizes and clean bands. Products were quantified using a Invitrogen Quant-iT™ PicoGreen ® dsDNA Kit. Molar concentrations were calculated and the PCR products mixed in equimolar amounts (0.2 pmol each) and digested with 150 units *HindIII* [New England Biolabs] in 90 μl volume for 2 hours at 37 °C. Mix was then incubated at 65 °C for 30 minutes to heat-kill the restriction endonuclease. Samples were cooled and spun down before being added to a 15 ml tube with 800 units T4 ligase [New England Biolabs], 700 μl 10x T4 ligase buffer and 6208 μl molecular biology grade water. Sample was incubated at 16 °C for four hours and then room temperature for 30 minutes. Equimolar mix was stored at 4 °C. Mix was tested by using combinations of 3C primers in a PCR using Qiagen HotStar Taq and running products on a gel to check size of bands.

2.3.4.3 qPCR

RT-qPCR 3C was done using an ABI Prism Sequence Detection System. A standard curve was created using the 200 fmol equimolar mix with the following dilutions: 1×10^{-1}, 2×10^{-2}, 4×10^{-3}, 8×10^{-4}, 1.6×10^{-4}, 3.2×10^{-5} and a no template control. Four replicates of 3C template were used, 125 ng each. Reaction mixtures were set up to a final volume of 25 µl with 6 µl H_2O, 5 µl 3C material, 0.75 µl forward and reverse primers (final 300 nM) and 12.5 µl SYBRgreen 2x mix [ABI]. qPCR was run with an step of 50 °C for 2 minutes and 95 °C for 10 minutes followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 1 min. Upon completion of amplification a dissociation curve was run at 95 °C for 1 minute, 55 °C for 30 seconds, 95 °C for 30 seconds and 60 °C for 1 minute.

2.3.4.4 Analysis

Analysis of RT-qPCR 3C data was done with a combination of the ABI Prism Sequence Detection System (to calculate C_t values) and Microsoft Excel. Standard curves were created for each primer pair by plotting $\log_{10}$ C_t against $\log_{10}$ concentration. A linear fit was calculated from the plot and used to derive concentrations for 3C sample templates from their C_t values. The mean and standard calculations were calculated from the four replicates. Dissociation curves were visually inspected to check for non-specific amplification and 3C products were run on a 1% agarose gel to check for bands.

2.4 e4C

e4C is a ligation based proximity assay based on chromosome confirmation capture (Dekker et al., 2002; Simonis et al., 2006; Schoenfelder et al., 2010). The protocol

I used changed a great deal during my PhD, the final method used to generate the majority of e4C association data presented in this thesis is described below.

2.4.1 Primer extension and primary *NlaIII* digestion

12 µg of 3C material was defrosted and cleaned using *solid-phase reversible immobilisation* (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 Dynalbeads kilobaseBINDER Kit (M-

280) was used. Beads were aliquoted 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 Dynalbeads 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 µl wash buffer, once in 100 µl 1x TE Buffer and once in 50 µl 1x NEB4 buffer with 1x BSA (NEB) before being resuspended in 50 µl NEB4 with BSA.

A second digestion step was then carried out with the enriched 3C material. To the resuspended beads, 0.5 µl (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 µl wash buffer, once in 100 µl 1x TE Buffer and once in 50 µl 1x NEB Ligase Buffer (supplied with NEB T4 ligase). Beads were resuspended in 50 µl fresh 1x NEB Ligase Buffer heated to 55°C for 5 minutes on a heating block along side a 7.5 µl aliquot of 100 µM stock PE Ad 2.0 NlaIII adapter. Both were snap chilled on ice. Beads were placed on the magnet and the supernatant removed before being resuspended in 40 µl ligation mixture (1x NEB Ligase Buffer; 200 pmol PE Ad 2.0 NlaIII 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 µl wash buffer and twice in 100 µl 1x TE Buffer.

2.4.3 PCR and germ-line removal

Beads were washed in PCR wash mix (1x Phusion HF Reaction Buffer (NEB) with 200 µM 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 µl of PCR master mix - 1x Phusion HF Reaction Buffer (NEB); 200 µM each dNTPs; 10 pmol PE Ad 1.0 + nested primer; 10 pmol PE Ad 2.0 *NlaIII* 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 minutes; 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 µl wash buffer and once in 100 µl 1x TE Buffer before storing at 4 °C as a backup. The PCR supernatant was cleaned using SPRI beads and eluted in 400 µl Molecular Biology water. 5 µl sample was kept back for the later gel.

To avoid sequencing re-ligation 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 germ-line 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 germ-line 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 DNA fluorescence *in-situ* hybridisation

DNA fluorescence *in-situ* hybridisation (DNA-FISH) is a microscopy technique using fluorescently labelled DNA probes to hybridise to and identify genomic sequences within fixed cell nuclei. I used a technique developed by Dr Daniel Bolland which was based on work by Dr Ieuan Clay and originally came from the laboratories of Dr Thomas Ried and Dr Thomas Cremer.

2.5.1 BAC preparation

The DNA used to create the DNA-FISH probes was generated using bacterial artificial chromosomes (BACs). BACs were ordered from Invitrogen and grown on agar plates with chloramphenicol. A single BAC colony was picked into 3 ml of LB containing chloramphenicol and incubated at 37˚C with shaking for 8 hours. Cultures were diluted $\frac{1}{500}$ to $\frac{1}{1000}$ in 100 ml of LB with chloramphenicol according to growth and placed in a large conical flask before incubation for 16 hours at 37˚C with shaking.

Cells were centrifuged in 50 ml tubes at 3500 rpm, 4˚C for 10 minutes. Cell pellets were resuspended in 4ml of buffer P1 [Qiagen] before adding 8ml buffer P2 [Qiagen] and mixing by inversion. Cells were incubated at room temperature for 5 minutes. 8ml buffer P3 [Qiagen] was added and mixed by inversion then left on ice for 5 minutes. Solutions were centrifuged at 4500 rpm for 10 minutes. Supernatant was filtered through pre-wetted filter papers placed inside clean glass funnels, held in centrifuge tubes. 0.7 volumes of isopropanol was added and BAC DNA allowed to precipitate for 10 minutes on ice.

DNA was centrifuged at 20,000x *g* for 15 minutes at 4˚C. The supernatant was discarded and the pellet resuspended in 1 ml 70% ethanol before being transferred to 1.5ml microcentrifuge tubes. Tubes were centrifuged at 13,000 rpm for 5 minutes before aspiration of the ethanol. Pellets were allowed to dry briefly at room temperature and resuspended in 400 µl molecular biology grade water [Qiagen]. DNA was precipitated with 100 µl of 4M NaCl and 540 µl 13% PEG-8000 solution (autoclaved). Tubes were centrifuged for 1 minute at 13,000 rpm and the pellet washed with 500 µl 70% EtOH. Tubes were again spun at 13,000 rpm for 1 minute and the pellete allowed to air dry briefly.

Pellets were resuspend in 250 µl TE buffer and added to 250 µl Phenol:chloroform:isoamyl alcohol (25:24:1, pH 8). Tubes were vortexed and centrifuged before removal of the aqueous phase into a fresh microcentrifuge tube. DNA was precipitated with 0.1 volume of 3M NaOAc (pH 5.2) and 2.5 volumes of 100% ethanol. Tubes were spun at 13,000 rpm for 10 minutes. DNA was washed with 500 µl 70% ethanol and then spun again 13,000 rpm for 1 minute. Ethanol was removed and the pellet allowed to air dry pellet before being resuspended in 100 µl molecular biology grade water [Qiagen]. DNA was quantified using a NanoDrop machine.

2.5.2 Probe generation

I used directly-labeled DNA probes for the FISH, generated with nick translation using aminoallyl-dUTP followed by chemical-coupling with an Invitrogen Alexa Fluor® reactive dye.

To generate nick translated aminoallyl-dUTP labelled probes I added 4 µg BAC DNA in H_2O to 20 µl 10 × NTB 5 µl (for 5 ml - 2.5 ml 1 M Tris-HCl, pH 7.5, 0.25 ml 1 M $MgCl_2$, 250 µl 10 mg/ml BSA fraction V, 2 ml nuclease-free water), 20 µl 0.1M DTT [invitrogen], 16 µl d(GAC)TP mix (0.5mM each), 4 µl 0.5 mM dTTP, 24

µl 0.5mM aminoallyl-dUTP, 40 units DNA Polymerase I [NEB], 4 µl 1:30 dilution DNase I [Roche] (dilution tested empirically for each new tube) and H_2O to a final volume 200 µl.

Tubes were incubated at 16 °C for 2 hours. A 1 µl aliquot was run on a 2% agarose gel to check the size of the fragments - the optimal size is a smear of products from 1kb to 150bp with the peak around 200-300bp. Samples were heated to 75 ˚C for 5 minutes to inactivate the DNase I. Amine-modified DNA was cleaned using a Qiagen PCR purification kit, eluted in 200 µl molecular biology grade H_2O [Qiagen] and precipitated with 20 µl NaOAc and 550 µl 100% ethanol at -20 ˚C for at least an hour (typically overnight). Samples were centrifuged at 13,500 rpm for 15 minutes at 4 Pellets were washed with 100 µl 70% ethanol, spun and dried at room temperature. Pellets were then resuspended in 5 µl molecular biology grade H_2O. Amine-modified DNA was stored at -20 ˚C.

For labelling, Invitrogen Alexa Fluor® reactive dyes were warmed to room temperature and resuspended in 2 µl anhydrous DMSO. One tube was used for 4 µg aminoallyl labelled DNA in 5 µl with 3 µl 0.2 M $NaHCO_3$ buffer (pH 8.3 [Sigma]) to make a total volume of 10 µl (tubes were split to label different probes on occasion, but ratios kept the same). Amine-modified DNA was heated to 95˚C for 5 and snap cooled on ice before addition. Reaction mixture was vortexed, spun down and incubated at room temperature in the dark for at least 1 hour.

To remove unincorporated dye, 40 µl molecular biology grade H_2O was added and probes cleaned using a Qiagen PCR cleanup kit, washing the column twice with PE. Probes were eluted in 200 µl H_2O (less if the dye was split). Probes were analysed on a Thermo Scientific NanoDrop using the 'proteins and labels' setting to measure dye incorporation and DNA concentration and stored at -20 ˚C.

Before use, 20 ng of probe was added to 1 µg Cot-1 DNA [Roche] and 9.7 µg

salmon sperm DNA [Roche] before precipitation with sodium acetate and 100% ethanol at -20 °C overnight. Tubes were spun, washed in 70% ethanol and dried. Pellets were resuspended in 5 μl deionized formamide (pH 7.0 [Sigma]) and incubated at 37 °C, shaking at 300 - 500 rpm, for at least 30 min. 5 μl 50% dextran sulfate in in 2×SSC was added before vortexing and centrifugation (all quantities in final paragraph are for one slide).

2.5.3 Slide preparation

$CD34^+$ cells and GM12878 cells were fixed in 2% formaldehyde and sorted as described in Sections 2.1 and 2.2. Cell pellets were resuspended in PBS and counted. 100 μl containing approximately 1×10^5 cells was pipetted into into an assembled cytospin funnel [Thermo Scientific] and polypro poly-L-lysine slide [Sigma] and spin at 300 rpm for 3 minutes in the Shandon cytospinner. Slides were placed into a coplin jar containing PBS. Cells were permeabilised in 0.1% saponin and 0.1% triton X-100 in PBS for 10 minutes at room temperature. Slides were washed in PBS for 5 minutes twice and stored in 50% glycerol and PBS at -20 ˚C for at least two days.

2.5.4 Probe hybridisation and washing

Slides were removed from -20 ˚C storage and placed in 20% glycerol and PBS at room temperature for 5 minutes. Nuclei were permeabilised using three freeze - thaw cycles in liquid nitrogen; slides were placed in the liquid N_2 for 3 - 4 seconds or until a characteristic 'popping' noise was heard. Slides were laid on tissue to thaw and placed back in 20% glycerol and PBS. After three repeats of this freeze - thaw cycle slides were washed in PBS for 5 minutes twice. Slides were then

incubated in 0.1 N HCl at room temperature for 30 minutes. Slides were washed in PBS for 5 minutes before the addition of 100 µl of 100 µg/ml RNase A [Roche] in 2×SSC on a coverslip. Slides were incubated horizontally at 37 ˚C for 60 minutes. Slides were washed in 2×SSC for 5 minutes and then PBS for 5 minutes. Slides were permeabilised in 0.5% saponin with 0.5% triton X-100 in PBS for 30 minutes at room temperature. Slides were washed in PBS for five minutes, twice. Slides were then equilibrated in 50% formamide and 2×SSC for at least 10 minutes.

10 µl of probe mix was pipetted onto a 22×22 mm coverslip. Slides were taken out of the formamide one by one, excess liquid was wiped off and coverslips with probe mixes were inverted and placed onto the cell spot. Slides were sealed with Fixogum rubber cement [Marabu] and heated to 78 ˚C for 2 minutes on a hot plate. Slides were then incubated at 37 ˚C in a humidified chamber overnight.

Slides were removed from the humidified chamber the following day and had the rubber cement removed. Slides were placed in 2×SSC to remove coverslips and washed in 50% formamide with 2×SSC for 15 minutes at 45 ˚C. Slides were then washed in 0.2×SSC at 63 ˚C for 15 minutes and then 2×SSC at 45 ˚C for 5 minutes. Slides were equilibrated back to room temperature in 2×SSC for five minutes and rinsed in PBS for 5 minutes. Slides were then stained in 80 ng/ml DAPI and 2×SSC for 2 minutes at room temperature. Slides were washed in PBS for 10 minutes and fixed in 3.7% formaldehyde and PBS for 5 minutes.Slides were quenched in 155 mM glycine for at least 30 minutes before being washed in PBS for 5 minutes. Slides had 64×22 mm coverslips mounted with a drop of Vectashield [Vector labs].

2.5.5 Visualising signals

Slides were analysed using a MetaSystems MetaCyte at 100× magnification. Nuclei with greater than two signals in either channel or no signals in either channel were discarded. A custom perl script was used to calculate the shortest *BCR* - target loci distance for both *BCR* foci. Target loci could be counted twice if one locus is closest to both *BCR* loci.

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 genome-wide associations of the globin genes *Hba* and *Hbb* in mouse foetal liver cells. They used ChIP to purify ligation products associated with hyper-phosphorylated 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 4 Kb ($4^6 = 4096$). This results in "hairballs" of cross linked DNA and protein which are diluted into a large volume. A ligation reaction is carried out which favours 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 association within the nucleus.

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 associations can be assessed by quantifying intensities of bands on a gel or with

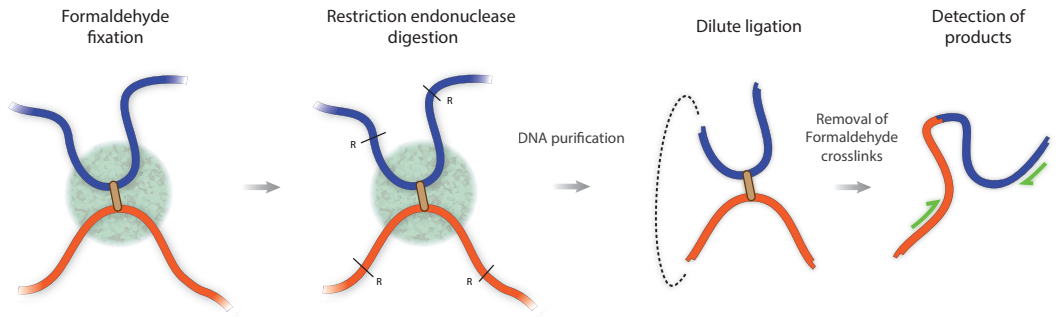


Figure 3.1.1 – Overview of the 3C methodology. R denotes the recognition sites for the first six-cutter restriction endonuclease.

RT-qPCR threshold values. In this way, chromatin contacts can be studied on a one-to-one basis, investigating the association between two candidate loci.

3.1.2 Enriched 4C

A 3C library contains fragments representing the association 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

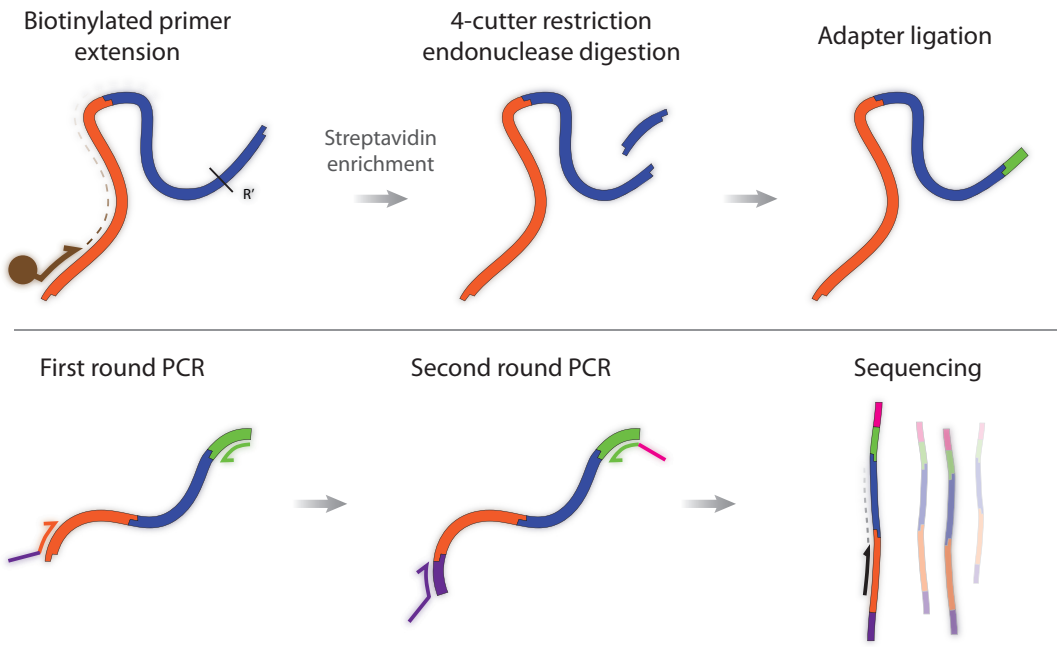


Figure 3.1.2 – **Overview of the e4C methodology.** R' denotes the recognition site for the second four-cutter restriction endonuclease. BC denotes the barcode region used to confirm the identity of the bait sequence and identify sequences when multiplexing libraries for sequencing.

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 3.8). 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 association profile for the bait sequence.

For discussion of e4C data analysis see Chapter 4.

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 series 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

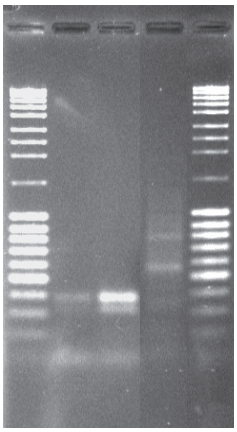


Figure 3.2.1 – *BamHI* e4C Libraries. Test e4C library prepared from human buffy coat cells using *BamHI* as the primary 3C restriction endonuclease. 1 - Human *BamHI* e4C test library. 2 - Human *BamHI* e4C library at 4x concentration. 3 - Mouse *HindIII* e4C library positive control prepared with a *Myc* bait.

this digestion step should be greater than 80% (Cope and Fraser, 2009).

The steps immediately before the *BamHI* digestion in the protocol include an 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 (1.8% to 5% Triton-X100, Fig 3.2.2 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 protocol (Section 2.3) and ran *BamHI* digestions after treatment with SDS and two Triton-X100 concentrations (Fig 3.2.2 B). The nuclei treated with the higher concentration of Triton-X100 showed a higher digestion efficiency, though still unsatisfactory for e4C.

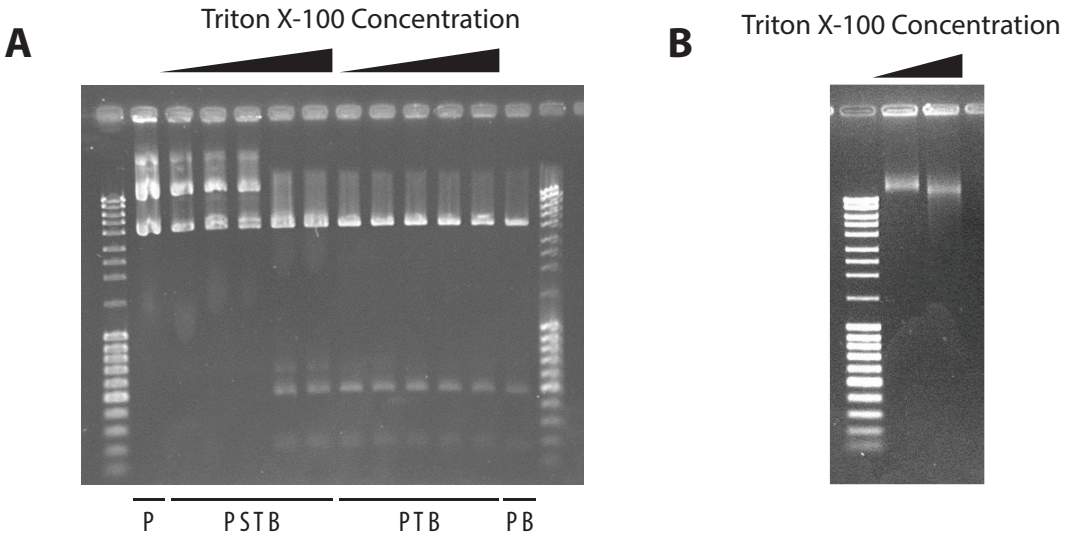
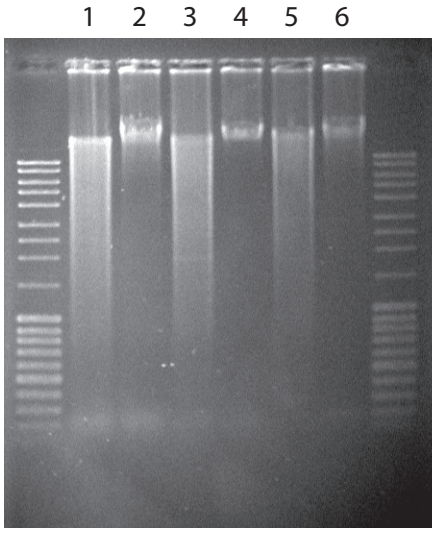


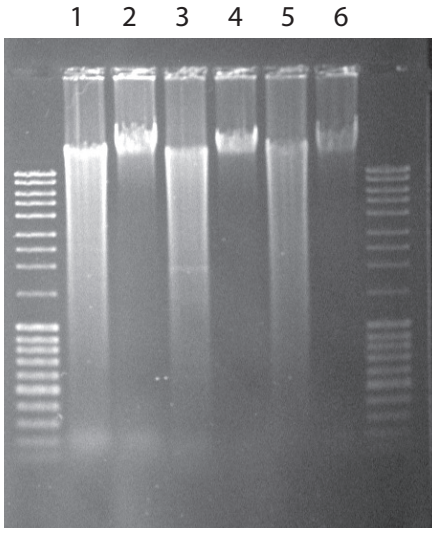
Figure 3.2.2 – Effect of Triton-X100 on *BamHI* digestion. (A) *BamHI* digestion efficiency of purified plasmid DNA with increasing Triton-X100 concentrations after SDS treatment. P, S, T and B show presence of Plasmid, SDS, Triton-X100 and *BamHI*. **(B)** - Digestion of crosslinked chromatin with *BamHI* after treatment with SDS and Triton-X100 (3.8%, 5%).

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). *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 (Fig 3.2.3).

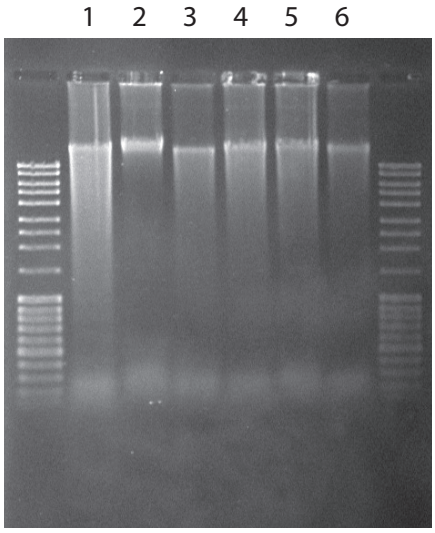
I chose to use *AseI* for future e4C experiments because it produces suitable fragment sizes around the break point cluster regions of both *BCR* and *MLL*. Further digestion efficiency qPCR tests showed its efficiency at a far superior 81 - 99% of sites being cut. Completion of test e4C libraries using *AseI* showed a gel smear comparable to that shown in Fig 3.2.1 for the mouse *HindIII* e4C library (Fig 3.2.4).



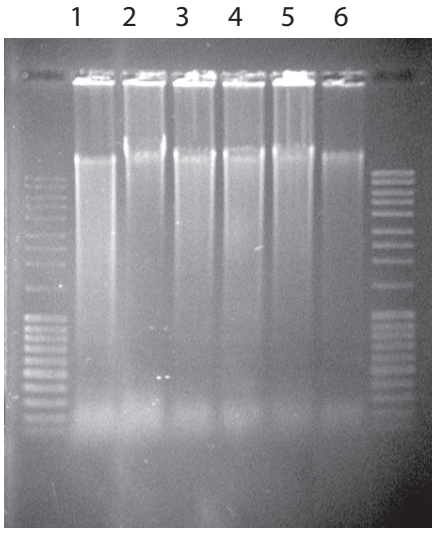
2% Fix, with SDS / Triton Step



1% Fix, with SDS / Triton Step



2% Fix, without SDS / Triton Step



1% Fix, without SDS / Triton Step

1 : *AseI* 3 : *HindIII* 5 : *PvuII*
2 : *BamHI* 4 : *NdeI* 6 : *XbaI*

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.

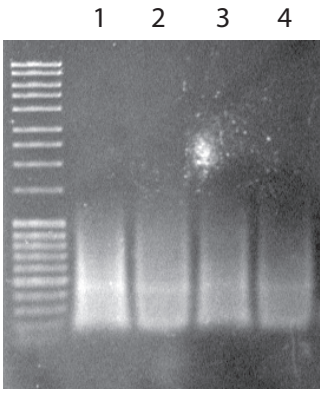


Figure 3.2.4 – **Gel of AseI e4C libraries.** 1% agarose gel of e4C libraries generated using *AseI* with *BCR* and *MLL* baits. 1 - Sample 1 *AseI* e4C, no SPRI. 2 - Sample 1 *AseI* e4C after SPRI. 3 - Sample 2 *AseI* e4C, no SPRI. 4 - Sample 2 *AseI* e4C after SPRI. Samples 1 and 2 were libraries generated in parallel from $CD34^+$ material.

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 the e4C fragment and would give sequence information about the unknown, captured fragment. The second read of the paired end run would sequence 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.

To circumvent this problem, I switched to using single end sequencing. To avoid sequencing the known bait region, I used custom sequencing primers that anneal to the bait region adjacent to the *AseI* ligation junction (Fig 3.3.1). This site was chosen to leave enough known sequence to confirm the identity of the bait product while returning long enough reads in the unknown partner to be aligned to the genome. The sequences of the Illumina sequencing primers had recently became publicly available, allowing me to design our custom sequencing primers with similar in length, T_m and GC content to the official Illumina Sequencing Primer 1.0 (Table 3.3.1).

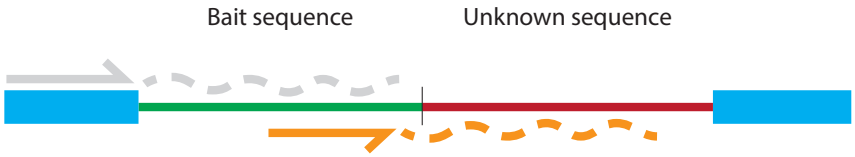


Figure 3.3.1 – Custom sequencing primer design. Illumina adapters shown in light blue, bait sequence in green and unknown sequence in red. Traditional sequencing primer and read on top in grey, custom sequencing primer annealing to bait sequence below in orange.

	Length (bp)	T_m	GC Content
Illumina Sequencing Primer 1.0	33	77.4℃	51.5%
BCR Sequencing Primer	38	73.4℃	36.8%
MLL Sequencing Primer	35	74.2℃	34.3%
ABL Sequencing Primer	32	77.9℃	46.9%
MLL-p1 Sequencing Primer	34	76.1℃	41.2%
MLL-p2 Sequencing Primer	40	74.7℃	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 3.8) 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

BCR	Germline	TGTTTCCTGCAGCACAAGAGGTTGGCAGAGCAACCTCAGAACCTCTTCTGTCTCTGTTTATGCTTTTGTTCAGAAGCTTAGTACCA**GGA**ATTAAT
	Biotinylated	ºTGTTTCCTGCAGCACAAGAG
	Nested	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCTCTGTTTATGCTTTTGTTCAG
	Sequencing	TGTCTCTGTTTATGCTTTTGTTCAGAAGCTTAGTACCA
MLL	Germline	GAACAAAATCACACCCTATTTGCCTTCACATTTGCAACAGATAATAATGCAAATGACAAAT**TTT**ATTAAT
	Biotinylated	ºTCACACCCTATTTGCCTTCA
	Nested	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAACAGATAATAATGCAAATGA
	Sequencing	CCGATCTGCAACAGATAATAATGCAAATGACAAAT
MLL 1	Germline	AAAGGAGCCAGAGCAAGTCAGAATTTGGTTTATTGTTGGAGATTTTAGGGCGAA**AAT**ATTAAT
	Biotinylated	ºAAAGGAGCCAGAGCAAGTCA
	Nested	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTTTATTGTTGGAGATTTTAGGG
	Sequencing	CGATCTGGTTTATTGTTGGAGATTTTAGGGCGAA
MLL 2	Germline	ACAGCAGCACCCTTTCTGTTGACTAGAAGTTAGGCTGTTGTAATAATAAACCCCTGAAACAGTGTGTTTATCAGTAACTAGTCT**TCGTTT**ATTAAT
	Biotinylated	ºACAGCAGCACCCTTTCTGTT
	Nested	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCCTGAAACAGTGTGTTTATC
	Sequencing	CGATCTCCCCTGAAACAGTGTGTTTATCAGTAACTAGTCT
ABL	Germline	CCACTGCGCCTGGCTAGTTTTCACTGGACAATACCTTGATGAAGCAGCAAAA**ATT**ATTAAT
	Biotinylated	ºTTCACTGGACAATACCTTGATGA
	Nested	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTTGATGAAGCAGCAAAAA
	Sequencing	GCTCTTCCGATCTCCTTGATGAAGCAGCAAAA
Constant	*NlaIII* e4C adapter	5′ §AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′ 3′ G*TACTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
	PE Ad 1.0 Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
	PE Ad 2.0 Primer	AAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
	Illumina sequencing	ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Table 3.3.2 – e4C primers. Germline shows bait locus, barcode region in bold pink, *AseI* 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 *NlaIII* 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.

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 smaller than 1 Kb or larger than 30 Kb as very small or very large fragment sizes can affect the ligation efficiency in the generation of the 3C library (van de Werken et al., 2012; Yaffe and Tanay, 2011). One end of the bait fragment is used for the e4C, this must be free of repetitive DNA and have enough sequence in which to fit a primer before the first *NlaIII* restriction site.

3.4 Preparation of CD34$^+$ cells

It has been proposed 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). The CD34 protein is a glycoprotein thought to be involved in cell-cell adhesion (reviewed in Furness and McNagny, 2006).

Samples were collected from consenting patients suffering from lymphoid leukaemias or multiple-myeloma, undergoing GCSF treatment to mobilise progenitor cells into the bloodstream. These cancers affect blood cells which are downstream of CD34$^+$ haematopoietic stem cells in the blood lineages, meaning that their CD34$^+$ cells are healthy. Any translocations affecting my genes of interest would be very clear in analysis: e4C associations show characteristic *cis* profiles which would be

spread across two chromosomes if a translocation were present. Such profiles are present in a publicly available HiC dataset from K562 cells which have a highly rearranged karyotype (Lieberman-Aiden et al., 2009), but are not present within my data.

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.

3.4.1 $CD34^+$ isolation

The first kit I used to isolate 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 due to low purity.

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%.

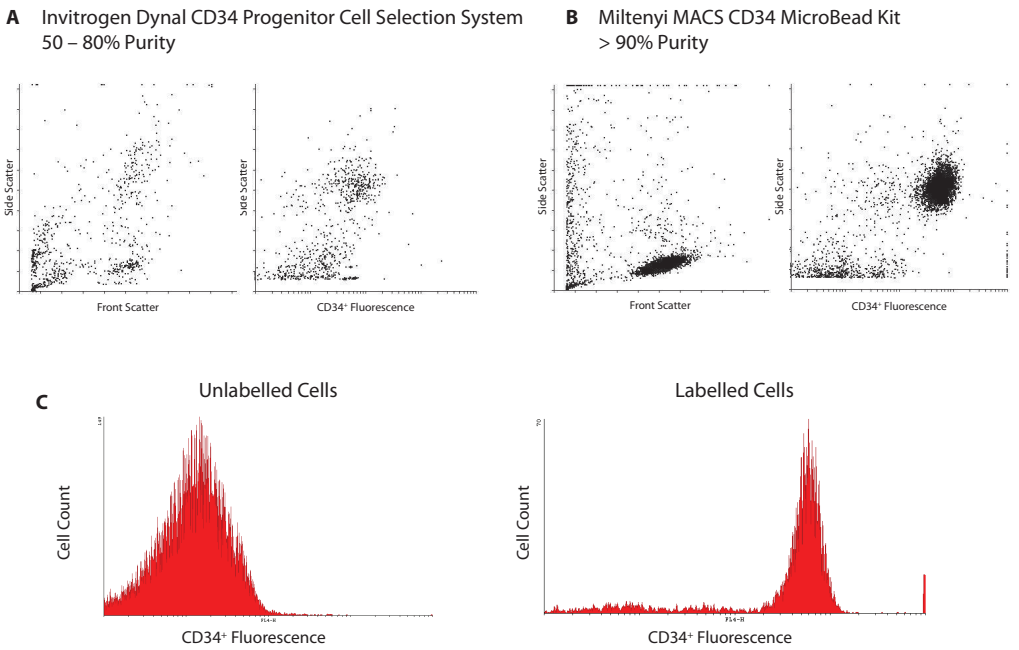


Figure 3.4.1 – CD34⁺ Separation FACS Plots. Diagnostic FACS plots after magnetic bead separation of fixed CD34⁺ cells using the **(A)** Invitrogen Dynal CD34 Progenitor Cell Selection System and the **(B)** Miltenyi MACS CD34 MicroBead Kit. Upper right quadrant of Front vs. Side Scatter plot in **(A)** 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.

3.5 e4C with low cell numbers

Initially, CD34⁺ cells were collected from peripheral blood supplied by Dr. George Follows in the Addenbrookes Haemotology 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 x 10^5 CD34⁺ cells per collection. Typically, 3C experiments use 1 x 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 x 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 x 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 intended to incorporate a chromatin immunoprecipitation (ChIP) step to purify 3C complexes containing RNA polymerase II, enriching for chromatin associations 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 association 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 *BCR* and *ABL1* are expressed in $CD34^+$ cells at a high level, and *MLL* at a low level (Figure 3.5.1). All associations may be relevant to the formation of chromosomal translocations, so association data outside transcription factories is still useful.

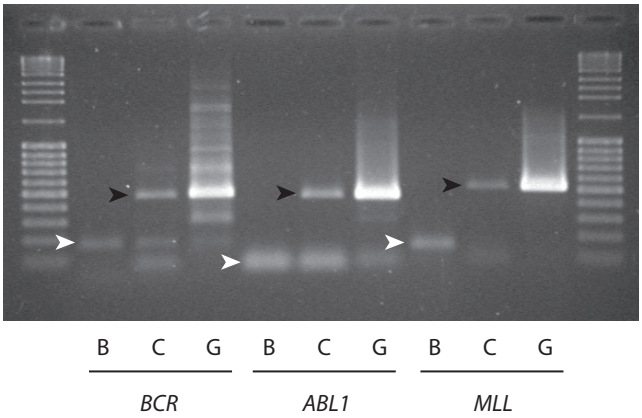


Figure 3.5.1 – CD34⁺ RT-PCR. B = Buffy Coat cells, C = CD34⁺ cells, G = genomic DNA control. Genomic DNA marked with black arrows, RNA bands marked with white arrows. CD34⁺ cells were not treated with DNaseI due to low cell numbers and limited sample, hence genomic band contamination.

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 achieved by spiking in 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 x 10^7 cells, and 1 x 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.

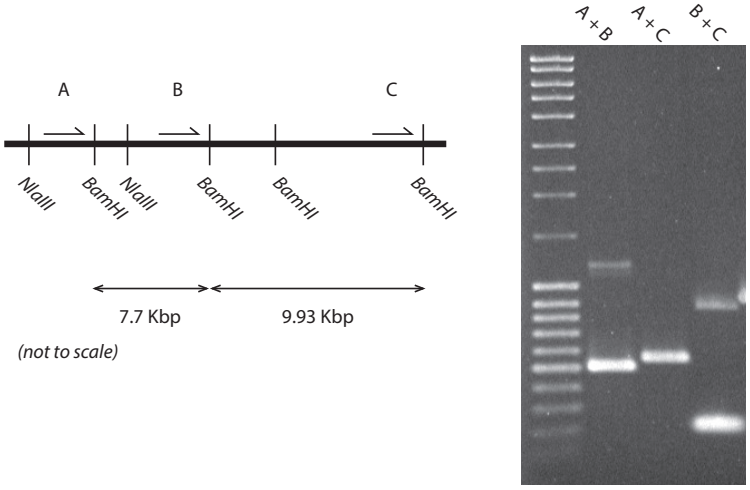


Figure 3.5.2 – *Drosophila melanogaster* S2R+ 3C Tests. PCR was performed on S2R+ 3C material using combinations of the primers shown. Lower bands are the expected sizes, higher bands are likely to be caused by concatemer ligation products containing additional 3C fragments between those being tested.

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 x 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 association datasets. In a further two leukapheresis collections 3.99 x 10^7 and 1.4 x 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 *BCR* e4C library preparation

A typical example of e4C library preparation is described below, detailing the results of the experiments to create the BCR Run 2 e4C library (chosen due to clear gels). This preparation came in several steps: the collection, fixation, sorting and

flash freezing of CD34$^+$cells, the creation of 3C material from that cell pellet and finally the generation of an e4C library from an aliquot of that 3C material.

CD34$^+$ cell collection is described in Section 3.4. The patient identifier was `Ref 2593P`, undergoing treatment for T-cell lymphoma. I collected 2.88×10^7 CD34$^+$ cells from a total count of 7.88×10^9 white blood cells. FACS analysis showed a CD34$^+$ purity of 96.39% (Fig 3.7.1, Table 3.7.1).

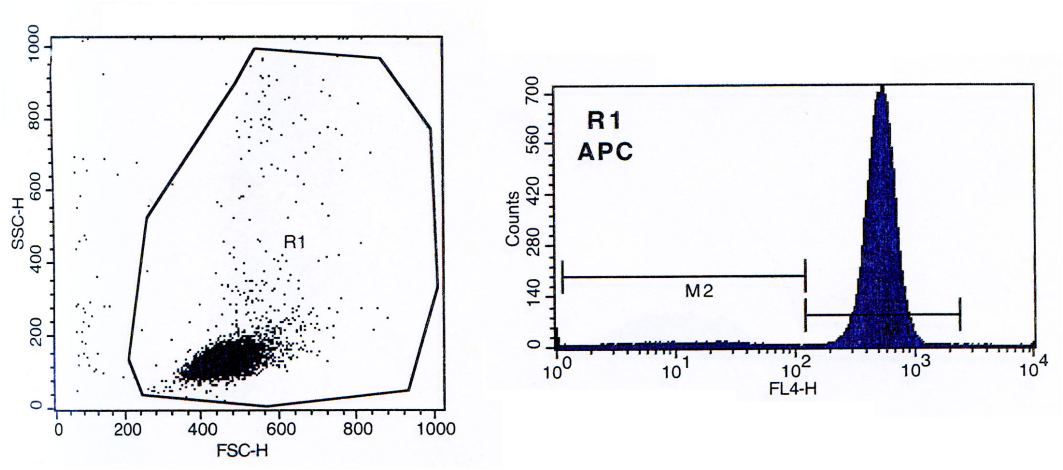


Figure 3.7.1 – CD34$^+$ FACS analysis plot.

Marker	Events	% Gated	% Total
All	50000	100.00	98.68
M1	47478	94.96	93.71
M2	2416	4.83	4.77

Table 3.7.1 – CD34$^+$ FACS analysis analysis.

CD34$^+$ 3C material was generated as described in the protocol in Section 2.3. PicoGreen analysis showed the yield to be 130 μg, 75% efficiency (assuming a hypothetical potential yield of 172.8 μg calculated by number of cells × 6 pg DNA per nucleus). I ran 1 μg of 3C material on a 1% gel which looked good, despite some distortion due to salt remaining after the ethanol precipitation (Fig 3.7.2).

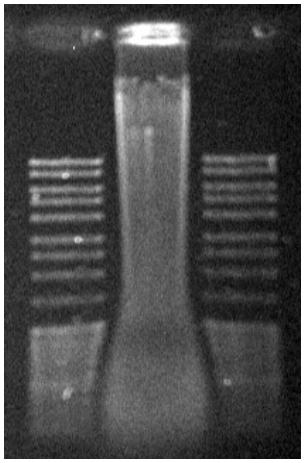


Figure 3.7.2 – CD34⁺ 3C library gel. 1% agarose gel showing 3C library smear. Distortion due to presence of salt.

CD34$^+$ e4C 3C material was cleaned using SPRI to remove salt contamination and processed according to the method described in Section 2.4. Genomic DNA (gDNA) control was used as a positive control for germline bands and a no template control (NTC) was used as a negative control. Figure 3.7.3 A shows the BCR e4C library after the first round PCR and germline band digestion. It can be seen that the strong germline band is digested in the gDNA control and that non-specific primer dimers are reduced in the NTC with SPRI (Fig 3.7.3 A).

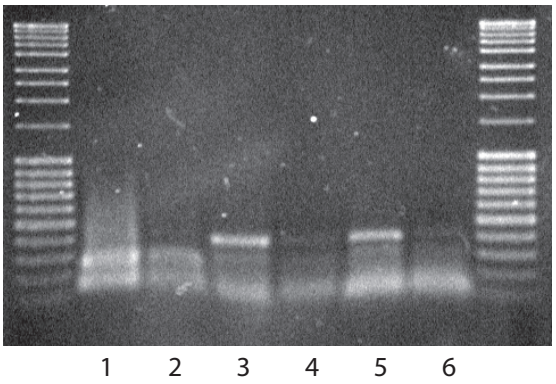


Figure 3.7.3 – BCR e4C libraries after first round PCR. 1 - BCR CD34⁺ e4C. 2 - gDNA control. 3 - gDNA before germline removal. 4 - NTC. 5 - gDNA before germline removal or SPRI. 6 - NTC before SPRI. gDNA = genomic DNA control. NTC = no template control.

To size select fragments of a length suitable for sequencing, the e4C library was purified using a gel extraction with a band from approximately 250 bp to 700 bp (Fig 3.7.4).

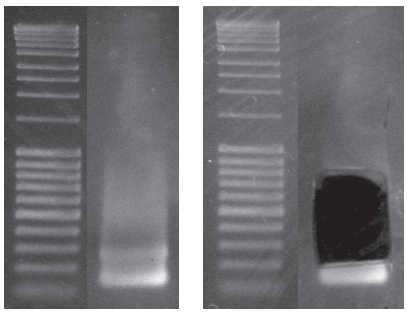


Figure 3.7.4 – CD34+ e4C library gel extraction. 1% agarose gels showing e4C library gel before and after gel extraction. Several empty lanes were run between the ladder and library to avoid contamination which have been removed from this image for clarity.

After the gel extraction the second round PCR was prepared to add the full Illumina sequencing adapters (Fig 3.7.5). The smear was the correct size for sequencing so I submitted the sample to the Babraham Sequencing Facility. Kristina Tabbada then quantified and sequenced the sample on a Illumina Genome Analyser IIx. Data analysis is described in Chapter 4.

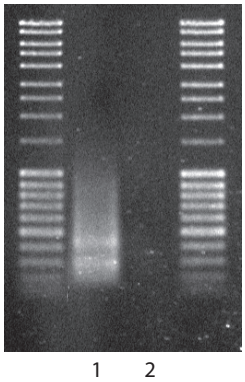


Figure 3.7.5 – Completed BCR e4C library. 1 = BCR e4C library. 2 = Second round PCR no template control after SPRI cleanup.

3.8 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 methods have been used with the Illumina Genome Analyser IIx and Illumina itself produces a kit for the production of multiplexed libraries which uses an additional read from adapter 2. 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.

3.8.1 Multiplexed libraries

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. Primers for both baits were used together through the e4C library preparation. Once the raw reads were available from the sequencing run, we separated the two e4C libraries *in-silico* (for details, see Section 4.1). The *MLL* library only returned 55,868 reads vs. the *BCR* library with 22,390,498, a 400-fold difference. This difference is likely to be due to a variation in primer efficiencies causing the *BCR* bait ligation products to be preferentially amplified above the level of the *MLL* bait products.

For my second multiplexed run, I prepared the e4C libraries separately in parallel and mixed them immediately before sequencing in equimolar amounts. 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.

3.8.2 Crossover products

Initial analysis of the *MLL* Run 3 e4C library showed much higher complexity than the previous two *MLL* libraries (Section 5.2.3). Upon closer inspection it became clear that a large number of hits in the *BCR* Run 2 e4C, multiplexed on the same lane, were present with the *MLL* barcode. These "crossover products" were characterised by low duplicate counts compared to what appear to be genuine *MLL* e4C reads (under 20 duplicates *vs.* 2,000 - 40,000 duplicates).

e4C libraries exhibit strong association in close *cis* linked chromatin due to physical linkage (Section 5.3). When viewing the region of chromosome 22 containing the *BCR* bait characteristic bait enrichment can be seen in the *MLL* e4C

library (Figure 3.8.1 A), a feature that could only be caused by e4C ligation products with a BCR bait. Interestingly, the crossover reads seem to be unidirectional; no *cis* bait enrichment around the *MLL* bait locus is seen in the BCR e4C library (Figure 3.8.1 B).

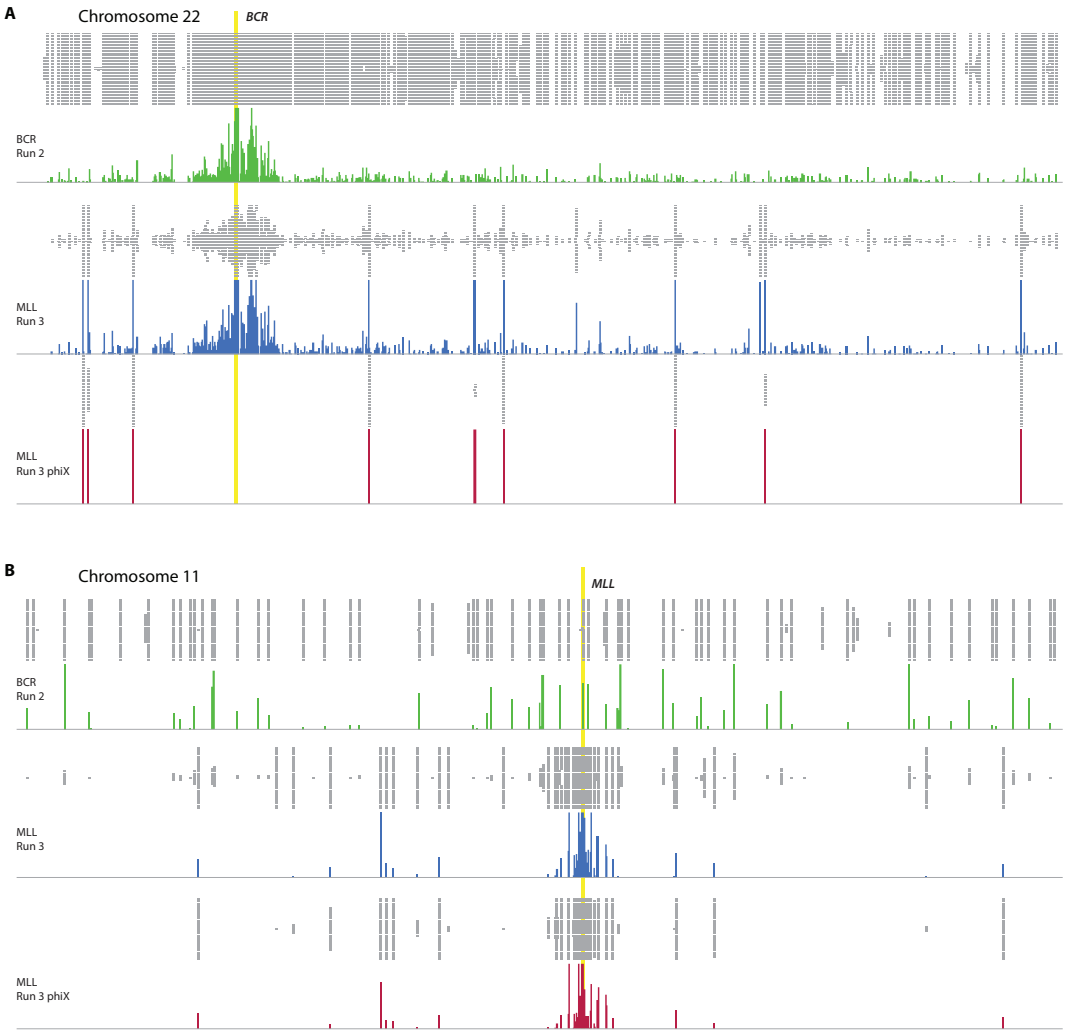


Figure 3.8.1 – Crossover reads between multiplexed e4C libraries. **(A)** 35 Mb region of chromosome 22, showing enrichment around the *BCR* bait for both the *BCR* and multiplexed *MLL* e4C libraries. There is no enrichment in the phiX *MLL* control library. **(B)** 35 Mb region of chromosome 11, showing enrichment around the *MLL* bait for the two *MLL* e4c libraries, but not the multiplexed *BCR* library. Bait regions shown by yellow bars.

To rule out ruling errors in the *in-silico* processing of the libraries I examined the

original sequencing data before processing to confirm that the crossover sequences began with the *MLL* barcode sequence. The *MLL* barcodes were present exactly as would be expected for an *MLL* e4C library product.

3.8.2.1 *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. The presence of the standard Illumina sequencing primer for the phiX library resulted in the majority of *MLL* library reads just containing bait sequence (Fig 3.3.1), however enough reads from the custom primer were returned for analysis. As can be seen in Figure 3.8.1, 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 3.8.1. Unfortunately this threshold was chosen empirically and could lead to inaccurate analysis results. As such, the thresholded library was not used for further analysis.

3.8.2.2 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 never came to a satisfactory explanation for the effect. Because we could not address this problem without a known cause, we did not multiplex any further e4C libraries.

3.9 Increasing e4C library complexity

Due to the low coverage of the e4C libraries (see Section 5.2), I attempted to modify the e4C protocol to increase the diversity of the sequencing reads.

3.9.1 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 3.9.1). 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.

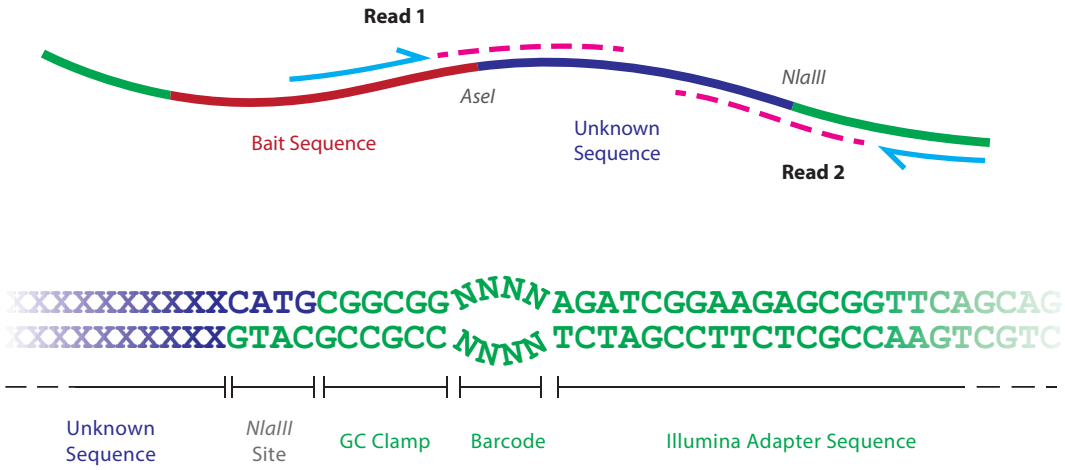


Figure 3.9.1 – Diagram showing design of the barcoded *NlaIII* adapter. Top panel shows an overview of a typical e4C library product. Read 1 confirms the identity of the bait region with a few bases of germ line sequence before reading through the *AseI* restriction endonuclease recognition site into the unknown ligated sequence. Read 2 reads the random adapter barcode before the GC clamp, *NlaIII* restriction site and unknown sequence.

3.9.2 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 second patient sample.

While 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.

3.9.3 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. 111,897 reads were returned, of which only 2,541 (2.27%) had a second paired end read with the expected structure (barcode, GC clamp, *NlaIII* site, unknown). The structure of the remaining reads were 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 attempt to determine why the paired end reads did not conform to the expected structure, I 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 should remove problems caused by non-homologous sequence in the barcode. Although this would lose 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 coverage depth and give semi-quantitative information in the association data.

3.10 Discussion

The past ten years has seen a rapid adoption of 3C and 3C-derived techniques. The slew of protocols to investigate the composition of 3C libraries has allowed us to investigate differences in association between loci in an increasingly high throughput manner. The e4C sequencing technique discussed in this chapter represents a further step towards the high throughput investigation of genome-wide chromatin associations.

The data presented here is not without its flaws. The clearest example of this is the poor quality of the $CD34^+$ *MLL* e4C libraries. These libraries exhibit approximately ten-fold less coverage than the BCR bait e4C libraries for reasons currently unknown (Section 5.2.3).

In a recent paper Cowell *et al.* attempted to use traditional 3C to measure contact frequencies between the *MLL* gene and its translocation partner *AF9* without success (Cowell et al., 2012). Although the study was able to demonstrate tran-

scriptional co-association between these genes using RNA-FISH, they were unable to detect any products by 3C. The authors attribute this to the low overall proportion of the cell population exhibiting a close interaction (2-3% cells) and the possibility that the large size of transcription factories may not allow efficient cross-linking of the fragments (Cowell et al., 2012).

It is an interesting possibility that the *MLL* locus may not be amenable to analysis by 3C based techniques, perhaps due to local inaccessibility of the chromatin. However, this explanation does fully explain the variation in complexity seen between my libraries: the low complexity of the GM12878 ABL e4C library or the difference in complexity between the GM12878 and $CD34^+$ *MLL* e4C libraries (Section 5.2.3). Given extra time and funds, it would be interesting to replicate my e4C libraries for all baits in both cell types to examine the influence of cell type and bait fragments.

In Section 3.8, I describe my efforts to further develop the e4C method to improve library complexity through the incorporation of a random barcode sequence to track individual ligation products and the multiplexing of e4C libraries prepared in parallel. While this experiment resulted in captured sequences that suggested problems in the library preparation, I believe that the concepts behind the attempt are sound. Multiplexed e4C libraries prepared in parallel have the potential to greatly improve the standard of data produced by e4C. It seems likely that the lack of homology in the random barcode sequence caused the problems in library preparation, using *NlaIII* adapters prepared using homologous oligos with a single barcode sequence may not cause the same problems and could still be used to detect identical ligation products found in multiple multiplexed libraries. Up to eight e4C libraries can be prepared in parallel without difficulty, so it would be feasible to multiplex many tens of libraries with the same bait on a single flow cell, greatly

improving the coverage of the resulting data.

Chapter 4

Developing the analysis of e4C data

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 developed a method to analyse the data in a way that is accurate and informative.

4.1 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.

4.1.1 Bareback processing

The Illumina next generation sequencing technologies including the Genome Analyser IIx (GA IIx) determine the order of nucleotides by using fluorescently tagged nucleotides: "sequencing 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 4.1.1). 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

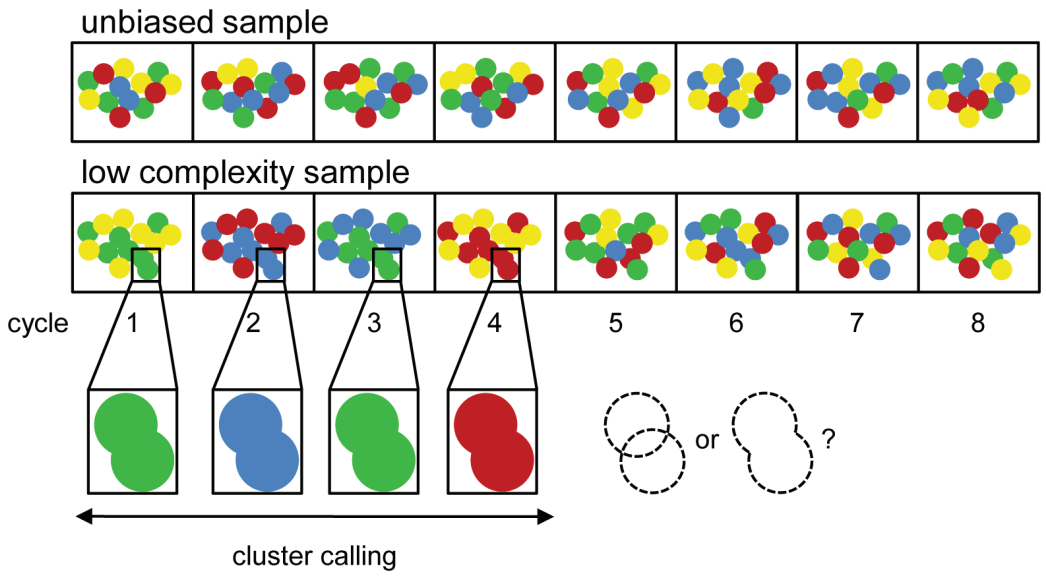


Figure 4.1.1 – Bareback overview. Diagram showing problematic cluster calling with samples containing low complexity in the initial sequencing cycles. Taken from Krueger et al. (2011).

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).

Bareback processing greatly increased the number of sequence reads returned from my e4C libraries (Table 4.1.1), 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 4.1.1 – Illumina sequence processing statistics with Bareback. Number of sequences returned from the standard Illumina SCS processing pipeline and from the bareback processing pipeline. Numbers shown for analysis of identical sequencing images.

4.1.2 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 cascading style sheet (web page 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 4.1.2.

4.1.3 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.

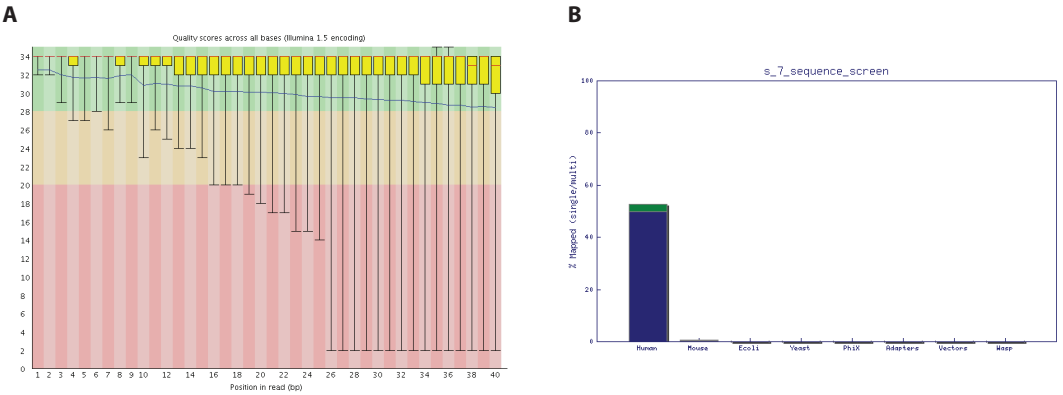


Figure 4.1.2 – Representative e4C Library FastQC and FastQ Screen. (A) FastQC - average sequence quality. (B) FastQ Screen - species alignment. Both statistics are for the second CD34⁺ BCR e4C sequencing run.

The expected structure of each e4C after the bait specific region is an *AseI* recognition site followed by unknown partner sequence (Fig 4.1.3). 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

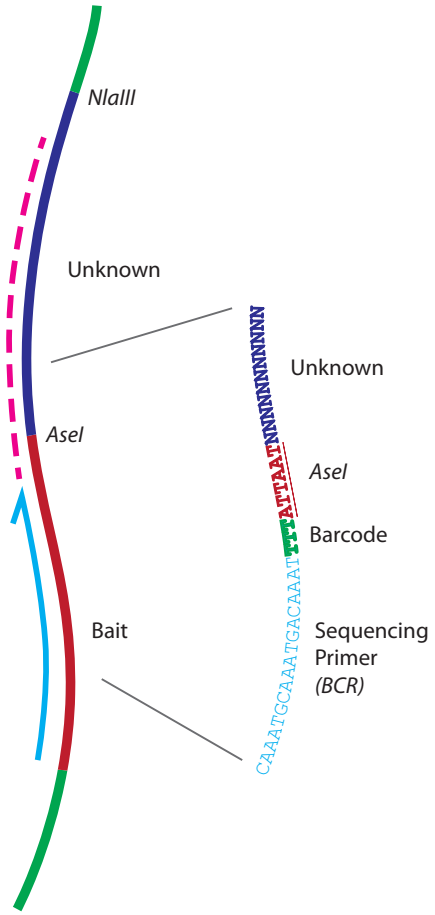


Figure 4.1.3 – Expected structure of reads in each e4C library.

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 Krueger in the Babraham Bioinformatics group wrote a script in Perl, which I later modified and used myself (Appendix A.2.4). The script loops through the FastQ file and attempts to match the first three base pairs against expected barcodes. If so, the barcode is trimmed and the next six base pairs are matched against the *AseI* recognition site (`ATTAAT`). Next, the script searches for *AseI* and *NlaIII* sites in the unknown sequence. If a second *AseI* site is found, the sequence 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 4.1.4.

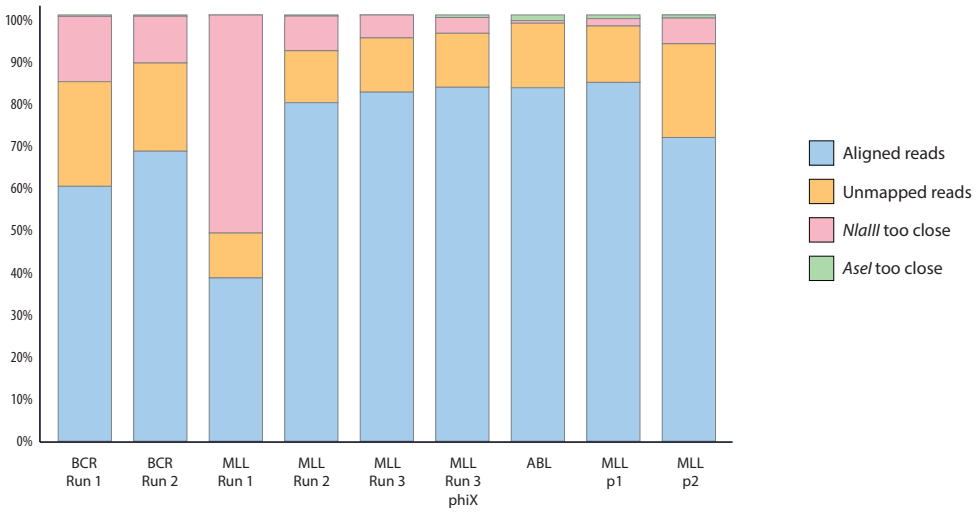


Figure 4.1.4 – Library trimming and alignments. Proportion of e4C libraries discarded due to AseI or NlaIII sites being too close to the bait, or not being mapped to the genome / being suppressed due to multiple alignments. Raw numbers can be seen in Appendix Table A.2.4.

4.1.4 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 4.1.4, Table A.2.4). Parameters were used as described in Fig 4.1.5. 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.

```
bowtie -q --phred64-quals -p 8 -m 1 hg19 <input> <output>
```

Figure 4.1.5 – 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.

4.1.5 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". Probes are 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.

4.2 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.

4.2.1 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 including an online tool to generate lists of restriction enzyme recognition sites (Appendix A.3.2). I have made some of these available as online tools at `http://www.tallphil.co.uk/bioinformatics/` (Appendix A.4).

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 4.2.1).

	Fragments Removed	Total Fragments
All *AseI – NlaIII* 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 4.2.1 – *in-silico* potential AseI - NlaIII fragment library statistics.

4.2.2 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 A.3.3). 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 incrememnted 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 4.2.1. It can be seen that the e4C libraries are biased towards greater %GC content and shorter fragments. This bias towards increased GC content and shorter fragment lengths

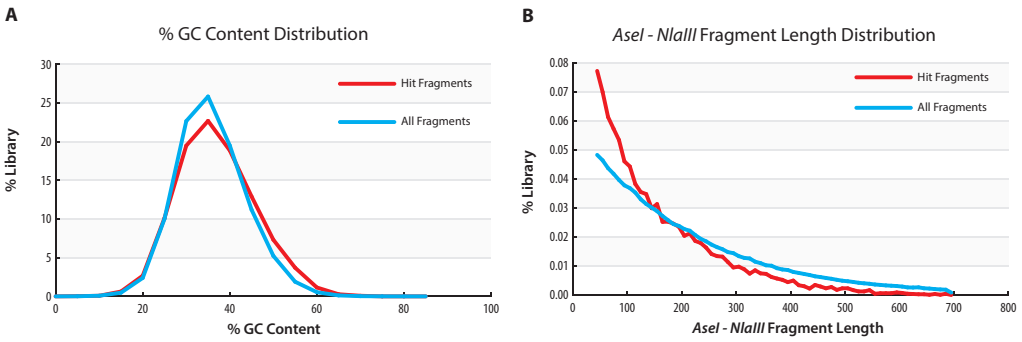


Figure 4.2.1 – e4C library biases. (A) %GC content distribution. GC content of each sequence calculated and quantified in 5% GC bins for all potential *AseI-NlaIII* fragments (blue) and all observed *AseI-NlaIII* fragments (red). An enrichment for higher GC content can be seen in the observed fragments. **(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).

is similar to observations made by Yaffe and Tanay in HiC data (Yaffe and Tanay, 2011). These are likely to be experimental biases that come from the digestion, 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.

4.2.2.1 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 A.3.4)

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

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:

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

This library of expected observation rates for every potential fragment was then used in single window testing (Section 4.3).

4.2.2.2 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 4.2.2. The resulting plot showed the *in-silico* library to have identical biases to the observed library, with the same GC and fragment length profiles plotted with a red line in Figure 4.2.1.

4.3 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.

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

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

4.4 *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 4.2.2. 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 $(\frac{observed}{potential})$ 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 4.4.1).

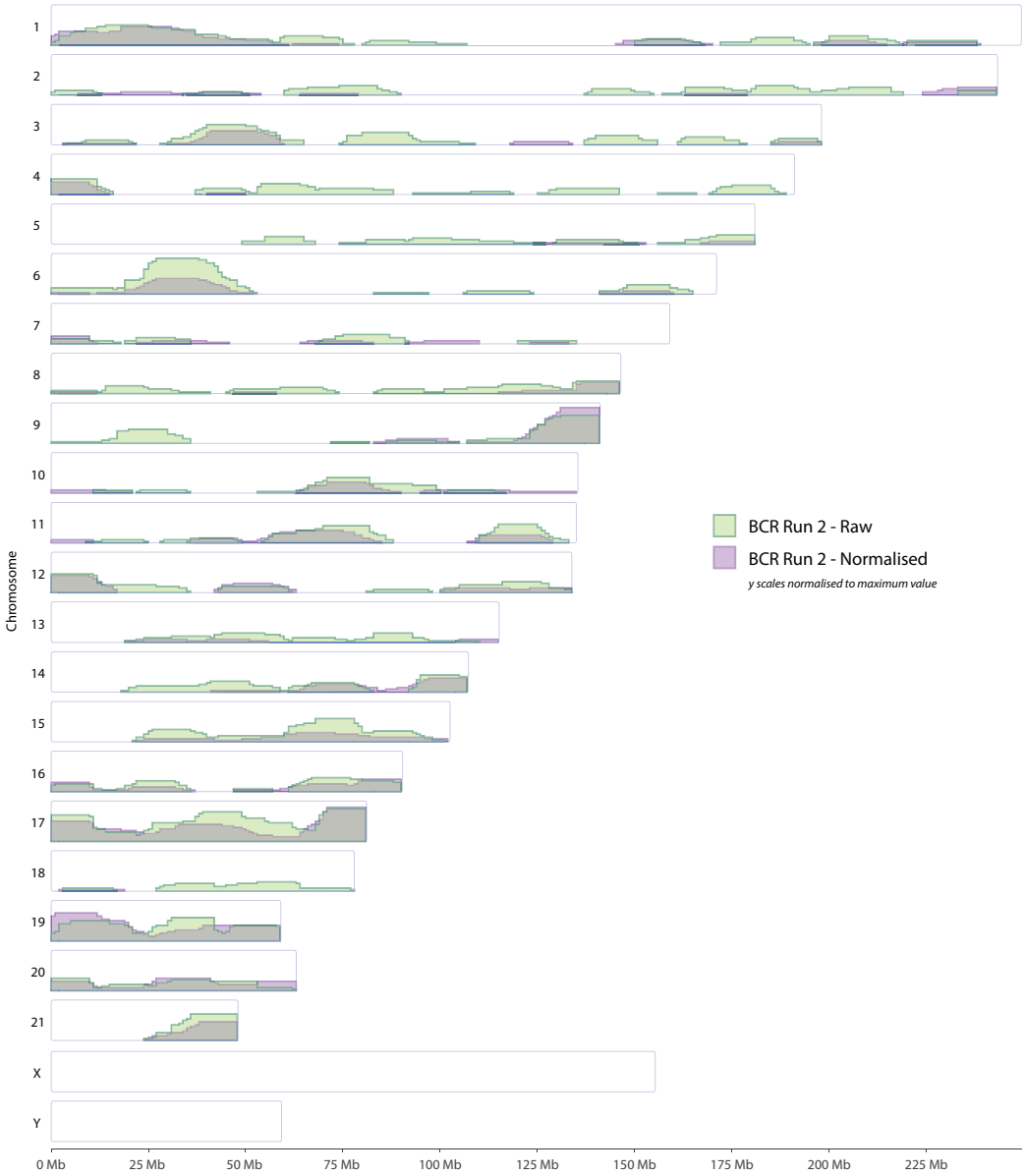


Figure 4.4.1 – *AseI* site distribution normalisation. Genome wide plots of BCR e4C Run 2 before and after AseI site distribution normalisation. Standard scores calculated and normalised against maximum value.

4.4.1 *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 4.4.2) suggesting that my normalisation is capable of removing bias due to the unequal distribution of *AseI* sites.

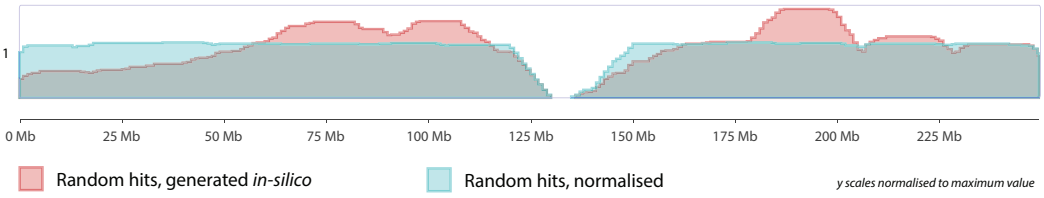


Figure 4.4.2 – *AseI* normalisation *in-silico* test. Chromosome 1 showing the *in-silico* e4C randomised library before and after normalisation for *AseI* site distribution. Read counts quantified and normalised against maximum value.

4.4.2 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.

4.5 Discussion

In this chapter I discuss my development of new analysis methodologies to extract, normalise and quantify e4C sequencing association data.

Because e4C sequencing is a novel technique under active development, custom bioinformatics tools must be written to process the raw sequence reads into a form that can be used for analysis. Large sequencing datasets are prone to systematic bias which can affect coverage (Cheung et al., 2011; Yaffe and Tanay, 2011). I describe a method to normalise variation across the genome in *AseI* cut site distribution. GC content and 3C fragment lengths have also been shown to bias 3C- derived datasets (Yaffe and Tanay, 2011) and I describe a method to create a normalised matrix of expected fragment hit probabilities. The use $\frac{observed}{expected}$ enrichment is a common method for analysing association data (Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011; Kalhor et al., 2011). Developing methods to test these normalisation techniques *in-silico* is a valuable exercise and helps to identify any problems in their development.

A good control experiment would be to sequence an "input" e4C library, created by removing formaldehyde cross-links prior to the ligation step. This control would include all biases introduced by the methodology and could be used to normalise experimental libraries. Unfortunately, such control libraries would need to be sequenced for each different bait fragment to control for variations in primer efficiencies and represents a significant expenditure.

In depth understanding of the processes generating the data is needed for proper analysis. In this chapter I describe how I chose to analyse association by quantification of the proportion of *AseI* fragments observed. This avoids any bias caused by differences in the efficiency of amplification between ligation products. Such binarisation of association data in *trans* analysis has been used by other groups for this reason (van de Werken et al., 2012).

Chapter 5

Initial e4C library analysis results

5.1 Introduction

In this thesis I investigate the hypothesis that proto-oncogenes involves 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* and in *trans* and demonstrate 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 3.8.1) 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 3.8.2). 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	Bait Position	Description
BCR 1	June 2010	chr22: 23,596,613 - 23,612,935 Breakpoint region of *BCR* gene.	First e4C library sequenced, multiplexed library preparation with MLL 1. $CD34^+$ cells.
BCR 2	Oct 2010		Technical replicated of BCR 1, library generated from the same 3C material. Sequenced in multiplex with MLL 2. $CD34^+$ cells.
BCR phiX	Mar 2011		Test of new multiplexing and barcode protocol. See Section 3.9. $CD34^+$ cells.
MLL 1	June 2010	chr11: 118,354,809 - 118,356,055 Breakpoint region of *MLL* gene	First e4C library sequenced, multiplexed library preparation with BCR 1. $CD34^+$ cells.
MLL 2	Sept 2010		Technical replicate of MLL 1, generated from the same 3C material. Sequenced in its own lane. $CD34^+$ cells.
MLL 3	Oct 2010		Technical replicate of MLL 1, generated from the same 3C material. Sequenced in multiplex with BCR 2. $CD34^+$ cells.
MLL 3 phiX	Dec 2010		Repeat sequencing of MLL 3, multiplexed in phiX lane. $CD34^+$ cells.
ABL	Dec 2010	chr9: 133,576,435 - 133,591,823	Promoter of *ABL1* gene. GM12878 cells.
MLL p1	Dec 2010	chr11: 118,304,771 - 118,311,258	Promoter of *MLL* gene (centromeric end of *AseI* fragment). GM12878 cells.
MLL p2	Dec 2010	chr11: 118,304,771 - 118,311,258	Promoter of *MLL* gene (telomeric end of *AseI* fragment). GM12878 cells.

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

5.2 Concerning raw data

5.2.1 Duplicate reads

Visual inspection of the e4C library sequence data within SeqMonk quickly reveals a large 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 A.2.5).

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 and analysed binary fragment observations instead of sequence read counts.

5.2.2 *AseI* fragment saturation

Because the e4C assay uses 3C libraries at its core, 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 A.2.2).

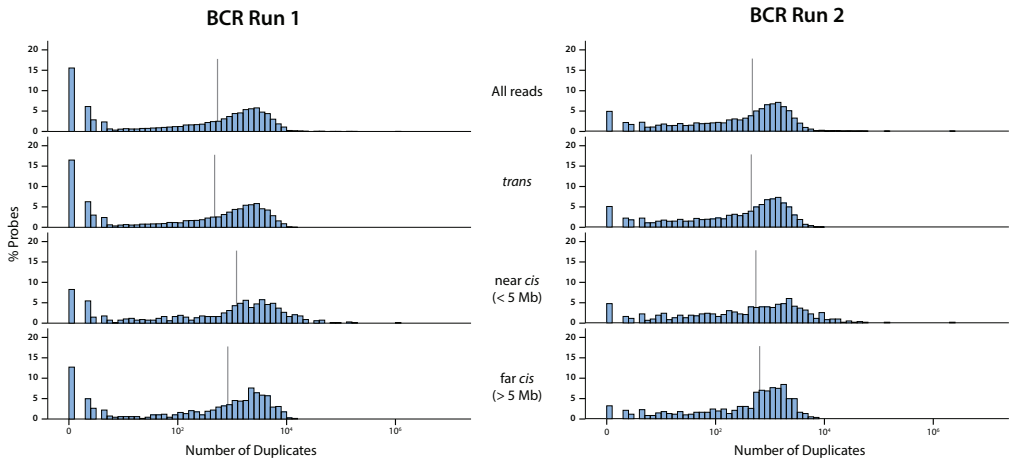


Figure 5.2.1 – e4C duplicate frequency. Histogram showing the frequency of duplicate counts per unique read for *BCR* e4C libraries. X axis shows the number of duplicates per unique read ($\log_{10}$ count), Y axis shows percentage of that collection that the number of duplicates applies to. Grey bars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

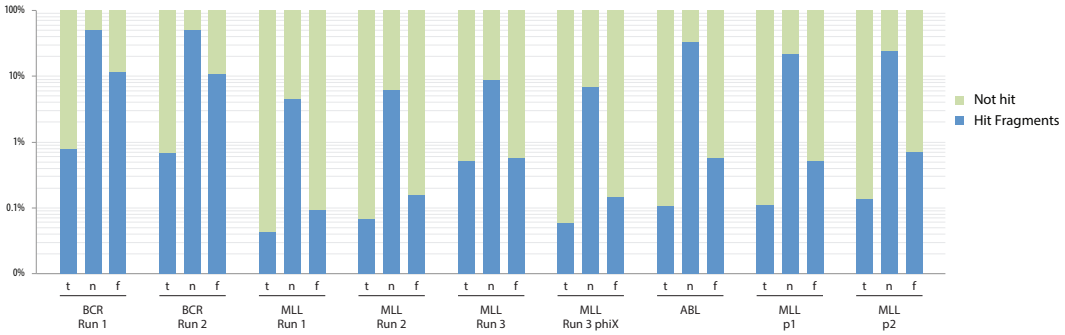


Figure 5.2.2 – e4C *AseI* fragment saturation. Degree of AseI fragment separation in *trans* (**t**), near *cis* (**n**, within 2.5 Mb either side of bait) and far *cis* (**f**, beyond 2.5 Mb either side of bait). Raw numbers can be seen in Table A.2.2.

5.2.3 Library complexity

The level of library complexity and coverage can be seen in Fig 5.2.3 and Appendix Table A.2.2. Figure 5.2.3 A shows the aligned raw read counts for near *cis*, far *cis* and *trans*. Reads within near *cis* (5 Mb window surrounding the bait) can be seen to represent a large proportion of all reads due to physical linkage in *cis*. Low read counts can be seen for MLL Run 1 (due to biased multiplexing, see Section 3.8.1) and MLL Run 3 phiX (due to competition with Illumina sequencing

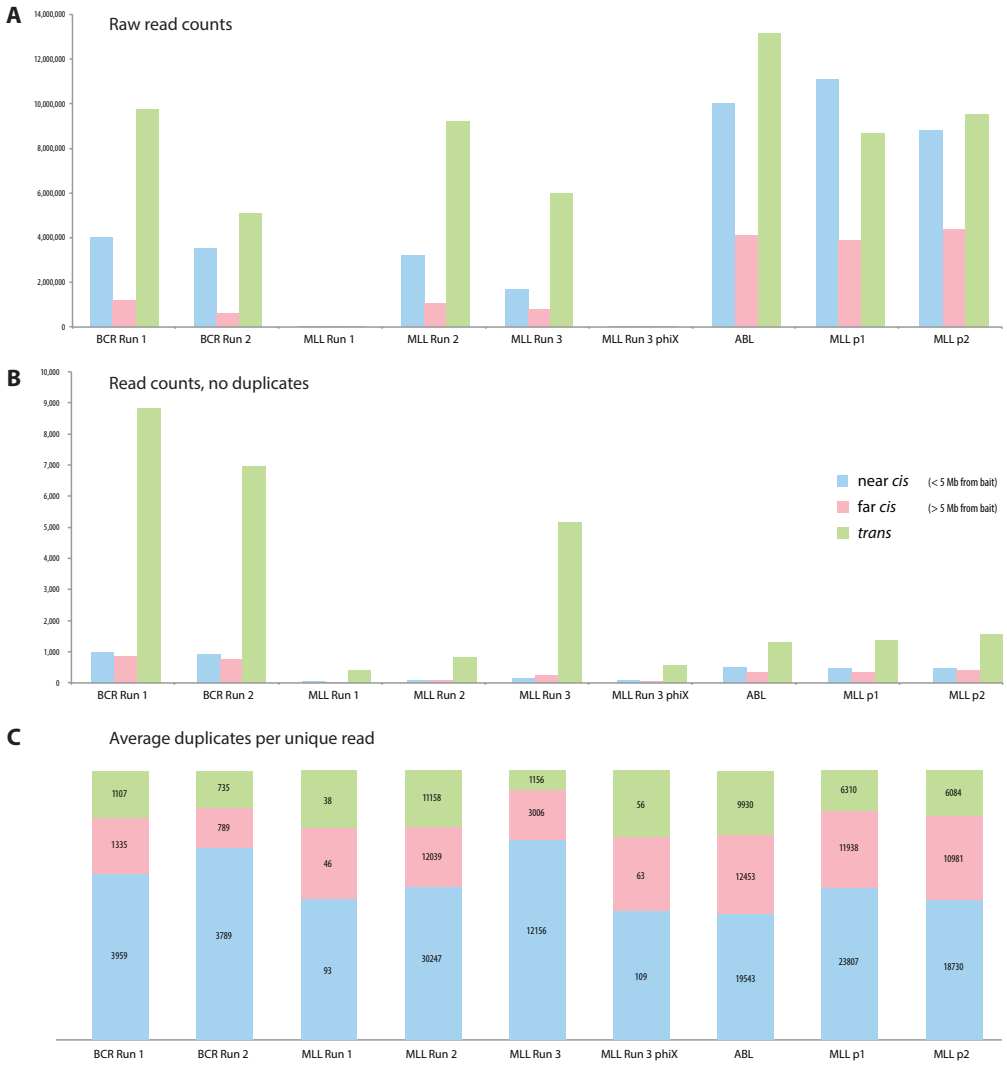


Figure 5.2.3 – e4C library read statistics. (A) Aligned raw read counts for near *cis*, far *cis* and *trans*. (B) Read counts after removal of duplicate reads. (C) Number of duplicates per unique read. Numbers were calculated in SeqMonk and plotted in Microsoft Excel and Adobe Illustrator.

primer). Increased total read counts can be seen in the ABL, MLL p1 and MLL p2 libraries due to advances in the Illumina GA IIx sequencing chemistry allowing greater cluster densities and so number of sequenced reads.

Figure 5.2.3 B shows the read counts for the same libraries after removal of duplicate reads. The number of near *cis* reads can be seen to drop relative to Figure

5.2.3 A, due to saturation of *AseI* 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 3.8.2). This plot shows the ~ 10 fold difference in library complexity between the BCR bait and all other baits.

Figure 5.2.3 C shows the 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 many more duplicates per unique read than the *trans* hits. The number of duplicates per read is high for all e4C libraries, showing that they are being sequenced to saturation.

Two general observations can be made from these plots: 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 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. This hypothesis is supported by the preferential amplification of the BCR bait in the first e4C library sequenced, where BCR and MLL baits were multiplexed in the same tubes for e4C library preparation (Section 3.8).

5.3 *cis* association 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 in-

teraction. 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 A.2.5). Sequence density variations are so large between the cis and trans chromosomes that I analyse them independently. This observation is in line with other reports in the literature (Yokota et al., 1995; Dekker et al., 2002; Lieberman-Aiden et al., 2009).

5.3.1 Association 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 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*). A similar observation of biphasic *cis* association within this range has been made using DNA-FISH and HiC (Yokota et al., 1995; Lieberman-Aiden et al., 2009), though other studies investigating local *cis* conformation show abrupt drops in association much closer to the bait (Tolhuis et al., 2002; Noordermeer et al., 2011).

5.3.2 Specific associations in *cis*

In addition to the obvious peak of *cis* association common to all of the e4C association datasets, some *cis* bait-specific associations 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

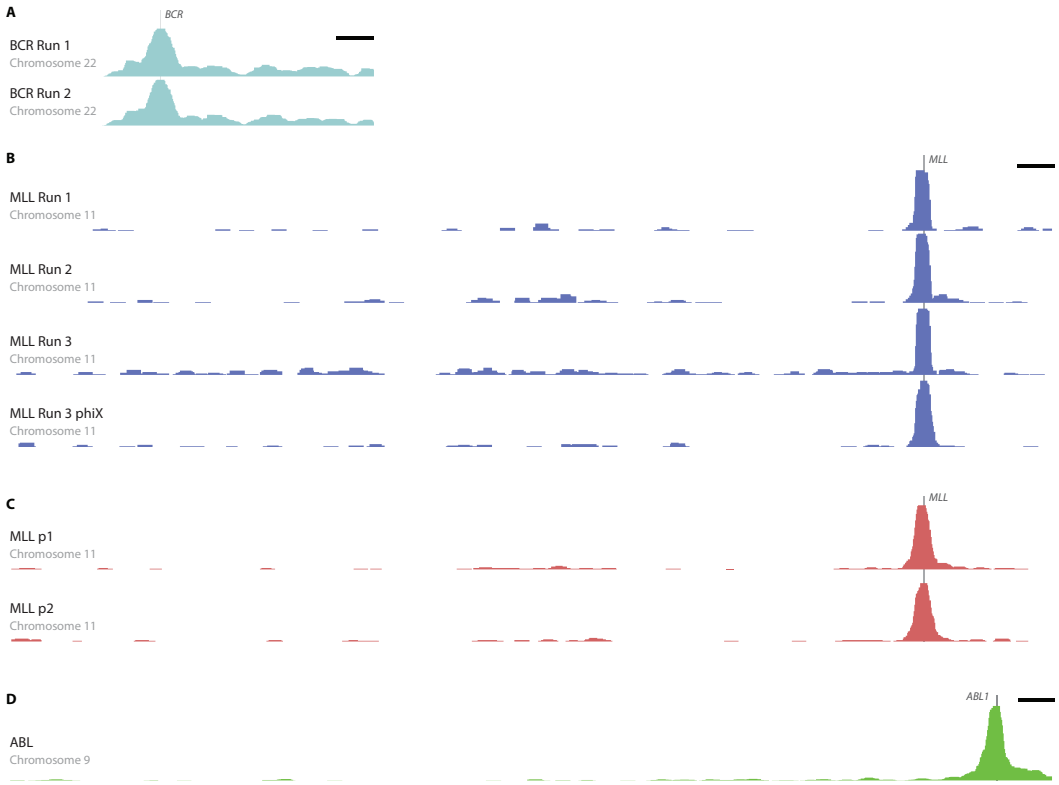


Figure 5.3.1 – e4C *cis* association profiles. Profiles generated using 100 kb sliding window fragment proportions (minimum 5 frags) summed in 1 mb windows separated by 100 kb. Grey bars show position of bait genes. Black horizontal bars are 5 Mb. Association profiles of the **(A)** *BCR* bait in CD34$^+$ cells, **(B)** *MLL* bait CD34$^+$ cells, **(C)** *MLL* promoter baits in GM12878 cells and **(D)** *ABL1* bait in GM12878 cells.

of contacts in both replicates, including a shoulder centromeric to the bait and four clusters telomeric (Fig 5.3.1).

To see *cis* associations in more detail in the BCR datasets with higher complexity, I plotted the fragment association 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

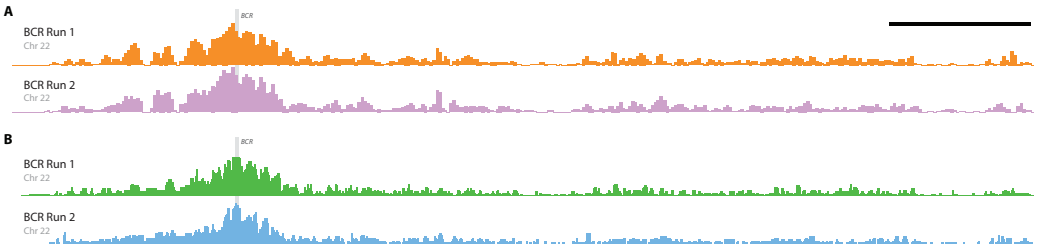


Figure 5.3.2 – *BCR* e4C *cis* association. *cis* association profile of the *BCR* bait on chromosome 22 in CD34$^+$ cells. **(A)** Association shown as the proportion of *AseI* – *NlaIII* fragments hit within a 50 kb sliding window (minimum 5 fragments). **(B)** Association shown using a sliding windows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

both replicates. To the centromeric (left) side of the *BCR* bait three distinct clusters of association can be seen.

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

Upon plotting the e4C *trans* association 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 association profiles appeared to share some features with previously described active regions (Versteeg et al., 2003), so I was interested in whether my e4C association libraries showed correlations with RNA polymerase II binding and other active chromatin marks.

5.4.1 Correlation with active epigenetic marks

To investigate the correlation between e4C association 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 A.2.3 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 association 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 ΦX	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
	CD34 +	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*
	GM12878	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
		H3K27ac	*0.584*	*0.588*	*0.222*	*0.295*	*0.215*	*0.216*	0.554	0.458	0.544
		H3K36me3	*0.598*	*0.607*	*0.259*	*0.324*	*0.229*	*0.335*	0.575	0.507	0.635
		H3K4me1	*0.660*	*0.661*	*0.298*	*0.350*	*0.261*	*0.329*	0.616	0.540	0.648
		H3K4me3	*0.550*	*0.557*	*0.252*	*0.331*	*0.182*	*0.290*	0.522	0.447	0.553
	K562	RNAPII	*0.773*	*0.794*	*0.305*	*0.239*	*0.401*	*0.355*	*0.653*	*0.527*	*0.641*
	HeLa	RNAPII	*0.770*	*0.799*	*0.178*	*0.253*	*0.225*	*0.279*	*0.660*	*0.543*	*0.629*
	NB4	RNAPII	*0.729*	*0.767*	*0.174*	*0.267*	*0.201*	*0.321*	*0.622*	*0.504*	*0.631*
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. See Appendix A.2.3 for NCBI GEO accession numbers

The strong correlation between bait association 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 association with these regions that 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).

Many of the differences in correlation scores between the *BCR*, *MLL* and *ABL1*

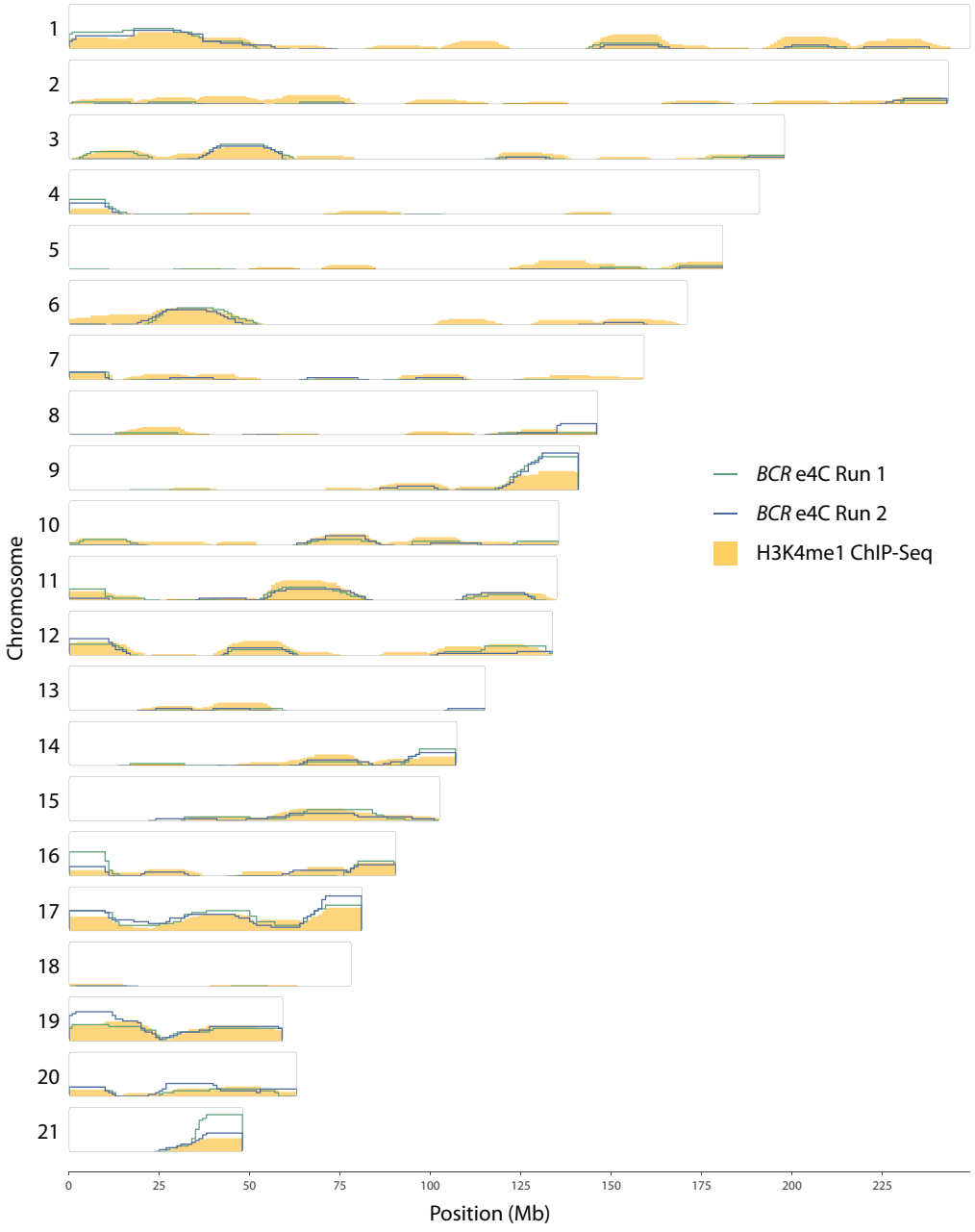


Figure 5.4.1 – *BCR* e4C correlation with H3K4me1. Genome wide association profiles of the two *BCR* e4C technical replicates (green and blue lines) normalised for *AseI* distribution in 100 Kb sliding windows (minimum five fragments) and summed in ten megabase windows with one megabase separation and 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.

e4C datasets can be attributed to differences in e4C coverage. The CD34$^+$ *MLL* libraries do not have enough reads to show strong enrichment regions. Although the difference between the CD34$^+$ *MLL* library correlations with active and inactive marks are difficult to see in Table 5.4.1 the majority of the correlation R scores are greater for active marks than inactive, a pattern that can be seen more clearly if the colour scale is adjusted.

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

Due to the association of the bait e4C profiles with active marks, I was interested to see the correlation scores 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.

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

		CD34 +						GM12878		
		BCR 1	BCR 2	MLL 1	MLL 2	MLL 3	MLL 3 ΦX	ABL	MLL p1	MLL p2
CD34 +	BCR 1	1.000	0.893	0.146	0.146	0.722	0.122	*0.625*	*0.126*	*0.130*
	BCR 2	0.893	1.000	0.171	0.187	0.817	0.153	*0.634*	*0.165*	*0.166*
	MLL 1	0.146	0.171	1.000	0.213	0.679	0.197	*0.280*	*0.249*	*0.262*
	MLL 2	0.146	0.187	0.213	1.000	0.083	0.093	*0.152*	*0.169*	*0.250*
	MLL 3	0.722	0.817	0.679	0.083	1.000	0.401	*0.304*	*0.229*	*0.247*
	MLL 3 ΦX	0.122	0.153	0.197	0.093	0.401	1.000	*0.195*	*0.193*	*0.254*
GM12878	ABL	*0.625*	*0.634*	*0.280*	*0.152*	*0.304*	*0.195*	1.000	0.535	0.546
	MLL p1	*0.126*	*0.165*	*0.249*	*0.169*	*0.229*	*0.193*	0.535	1.000	0.476
	MLL p2	*0.130*	*0.166*	*0.262*	*0.250*	*0.247*	*0.254*	0.546	0.476	1.000
Colour scale		0			0.4					0.8

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.

MLL p1 and MLL p2 which have similar read numbers and complexity do not show this correlation. This correlation could indicate a matching of specific genome-wide association profiles by the two loci due to bait colocalisation within the nucleus. Using genome-wide association correlations to measure proximity instead of direct contact frequency as measured by 3C product reads has been used for Hi-C data analysis in the literature (Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011; Kalhor et al., 2011).

5.5 Different genes have different preferred association partners

To investigate bait-specific *trans* associations within the datasets, I normalised the association 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 association 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* association datasets makes the detection of any specific local association regions difficult, and those observed unreliable. I discuss the main peak of the *BCR* datasets in Chapter 6.

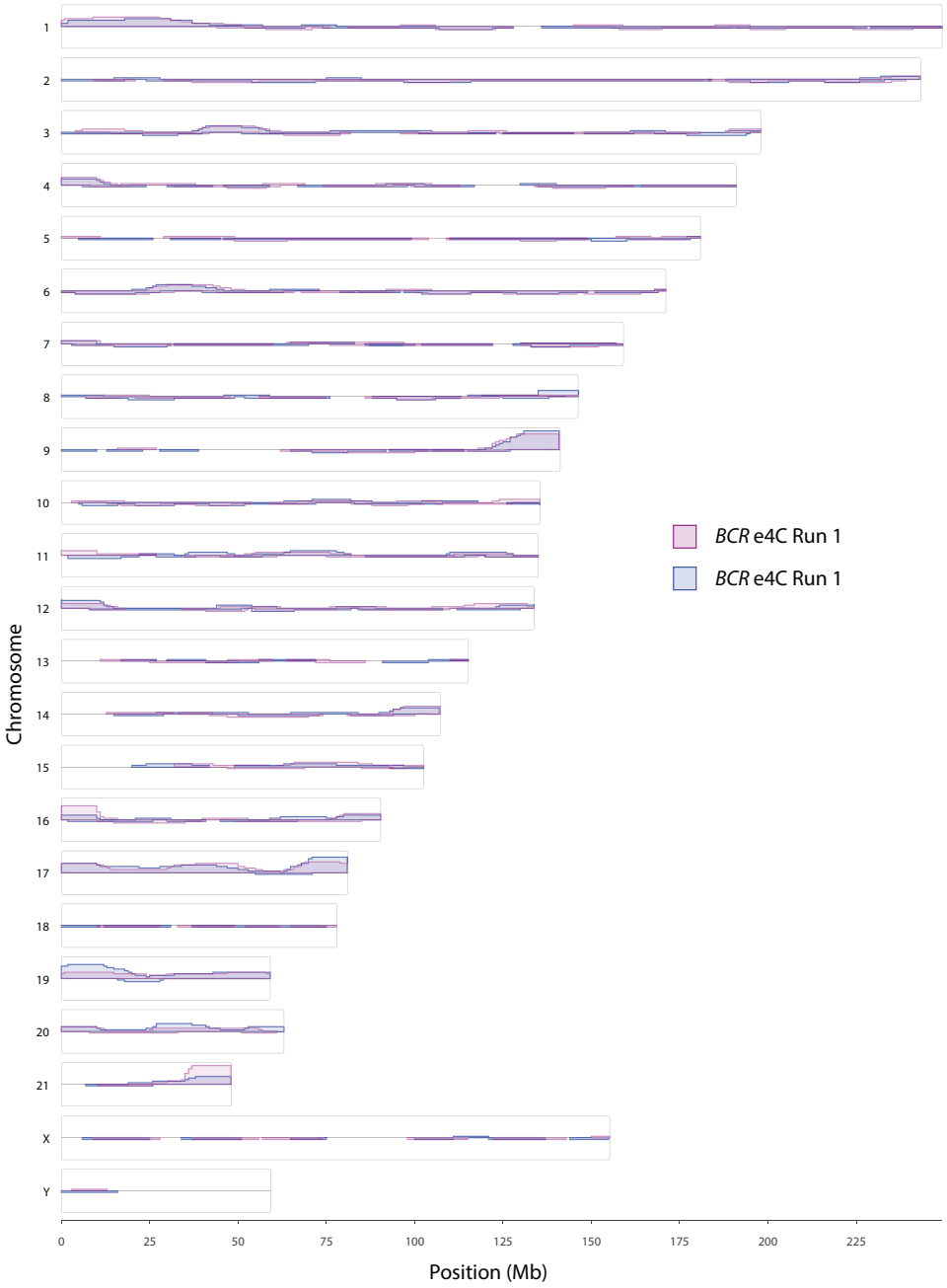


Figure 5.5.1 – *BCR* e4C normalisation against H3K4me1. Association profiles normalised against a H3K4me1 ChIP-Seq profile (Bernstein et al., 2010). Standard scores were calculated for the normalised association datasets and the ChIP-Seq dataset. Association datasets were normalised against the ChIP-Seq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

5.6 Discussion

In this chapter I describe the initial analysis performed on all of my e4C association libraries. I detail the problems encountered with the high level of duplicate sequences and low complexity encountered in the libraries and the common features observed between association profiles in *cis*. I discuss the specific associations seen within the e4C data in *cis* and demonstrate that the more complex and informative association libraries correlate with active epigenetic marks, suggesting that these genes reside within an active genomic compartment. I go on to describe specific associations seen in *trans* beyond that expected by genome-wide profiles of active epigenetic marks.

My observations of *cis* profiles are similar to those seen in other association studies (Simonis et al., 2006; Lieberman-Aiden et al., 2009; Sanyal et al., 2012), demonstrating the universal principle that *cis* linkage is the most dominant force in determining chromatin contacts. This concept has a number of implications for genomic organisation, such as evolutionary pressure on sequences required to co-associate within the nucleus to cluster within regions of the genome. An example where this evolutionary pressure is manifested is the clustering of highly expressed housekeeping genes throughout the genome (Lercher et al., 2002).

The correlation of bait gene association with regions enriched for active epigenetic marks is supported by a number of studies published during the course of my PhD (Lieberman-Aiden et al., 2009; Kalhor et al., 2011; Yaffe and Tanay, 2011; Simonis et al., 2006). I discuss the implications of this finding in more detail in Chapter 7.

Chapter 6

Preferential association of *BCR* with Chromosome 9

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

The most prominent association observed in the two replicate *BCR* association profiles is on chromosome 9, cytogenetic band q34. When the two replicates are combined this association gives approximately 7 fold enrichment of association above the mean with a maximal standard score of 17.04. This is the strongest *trans* association that I found in any of my e4C association datasets.

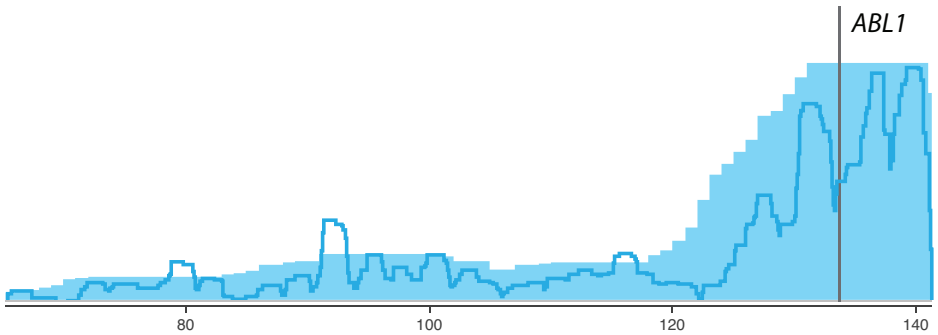


Figure 6.1.1 – *BCR* association with 9q34. Association of the combined *BCR* replicates, normalised for *AseI* site distribution, across the q arm of chromosome 9. Association calculated using ten megabase windows with one megabase separation (light blue fill) and one megabase windows with 100 Kb separation (ark blue line). X axis shows chromosome position in megabases. *ABL1* is shown as a grey bar.

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.

6.1.1 Single window testing

To confirm that this association is not an artefact of bias introduced through GC content or restriction fragment length, I used the single-window testing method described in Section 4.3. The script was run with one million loops to generate one million *in-silico* single window scores. The the resulting frequency of each number of *in-silico* scores is plotted in Figure 6.1.2.

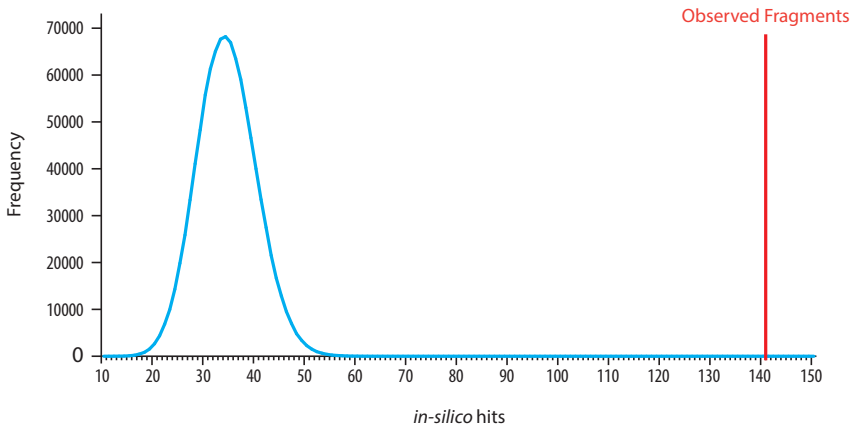


Figure 6.1.2 – Frequency of *in-silico* hits for single window in telomeric Chromosome 9. Result of *in-silico* random e4C libraries accounting for GC% and fragment length bias. X axis shows the number of fragments hit per interaction, Y axis shows the frequency with which that number was observed. Actual number of fragments observed in the *BCR* e4C shown as a red line.

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 $p < 3x10^{-13}$ 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 association profile of this region, I made e4C libraries with a bait in the *ABL1* promoter in GM12878 cells. This cell line is a tier 1 ENCODE project cell type and has numerous publicly accessible epigenomic datasets 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 fewer unique reads. The ABL e4C library shows stronger correlation with the BCR libraries than any of the MLL libraries (Table 5.4.2), suggesting that these two regions have similar genome-wide associations. When visually inspected, the *trans* association plot of the e4C library has somewhat sharp peaks of association due to the lack of complexity (Fig 6.2.1). While the *BCR* locus on chromosome 22 is enriched to a significant level ($p < 0.01$), it is not the strongest association genome-wide. This could be due to the low coverage not providing enough resolution to pick out specific associations, a lack of association between these regions in GM12878 cells, or it could be due to the lack of a direct specific association 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

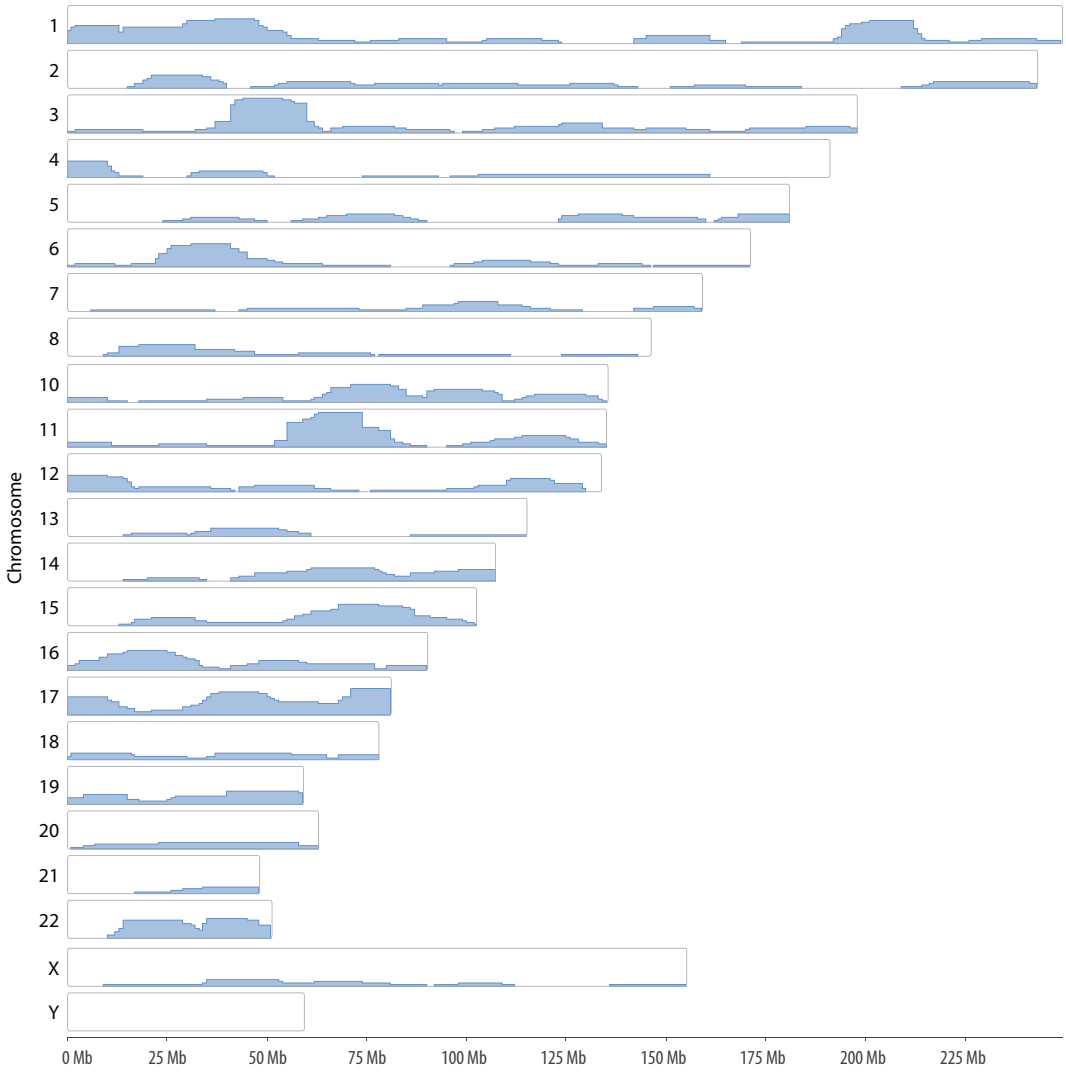


Figure 6.2.1 – ABL e4C *trans* association profile. Genome-wide plot of *ABL* e4C trans associations in GM12878 cells. Reads normalised for *AseI* distribution with 100 Kb windows, minimum five fragments. Association scores summed in ten megabase windows separated by one megabase.

of these peaks, leading to the intriguing possibility that the *BCR-ABL1* association 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* associations (Fig 5.3.1). These observations suggest that *BCR* and *ABL1* could be bystander genes brought together through associations of different genes within the region. Such a model would explain the 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 might also explain why I do not see specific *BCR-ABL1* associations 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 association 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 associations 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 association resolution.

The study by Lieberman-Aiden *et al.* investigated associations 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 inter-

act 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).

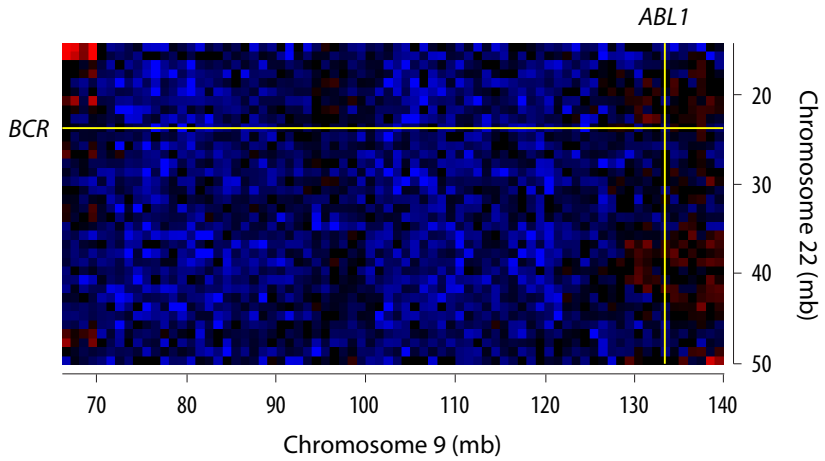


Figure 6.3.1 – GM06990 Hi-C heat map. Association heat map showing the q arm of chromosomes 9 and 22. Colours represent observed / expected ratios; blue colours are weak associations, red are strong. Each block represents a one megabase window. Generated using the Hi-C data browser: http://hic.umassmed.edu,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

To analyse the association 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 normalisation for *NcoI* restriction site distribution.

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 (Fig 6.3.2).

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). This study was done in GM12878

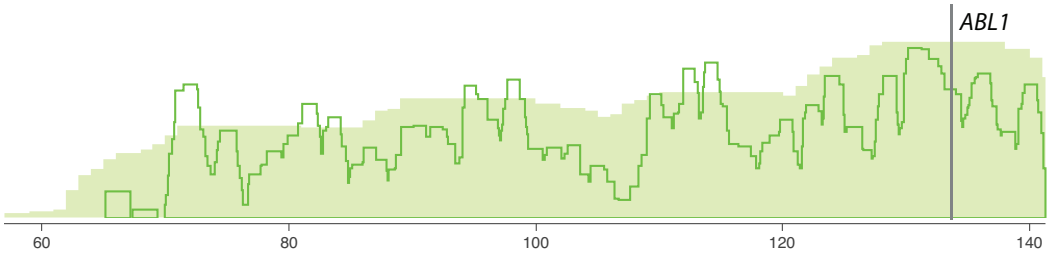


Figure 6.3.2 – GM06990 Hi-C *BCR* locus association with chromosome 9. Association of reads within one megabase of the *BCR* gene, normalised for *NcoI* site distribution, across the q arm of chromosome 9 (centromere to the left, telomere to the right). Light green fill shows quantification using ten megabase windows with one megabase separation. Darker green 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.

cells and the Hi-C dataset was sequenced to a greater depth. The authors claim that TCC, a modified version of Hi-C which tethers proteinaceous 3C complexes to solid phase beads before ligation, reduces non-specific ligation events and so removes background noise from the association data (Kalhor et al., 2011).

As with with the Hi-C data described above, genome-wide profiles of specific loci can be generated from this data. I plotted the association of a one megabase window surrounding the *BCR* gene on chromosome 9 after normalisation against *HindIII* restriction fragment distribution (Fig 6.3.3). A strong association can be seen over telomeric chromosome 9 using ten megabase windows. This splits into smaller peaks when the window size is dropped to one megabase, similar to the profile of the e4C association.

6.4 Narrowing the window of associations

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

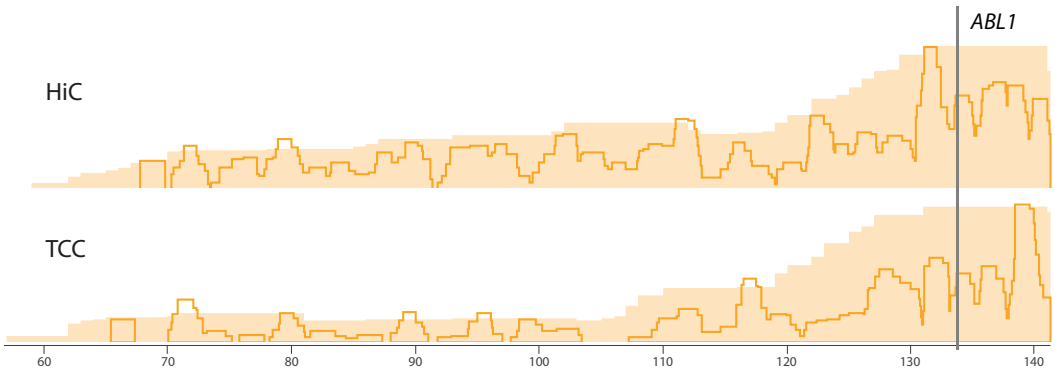


Figure 6.3.3 – GM12878 HiC and TCC *BCR* locus association with chromosome 9. Association of reads within one megabase of the *BCR* gene, normalised for *HindIII* site distribution, across the q arm of chromosome 9 (centromere to the left, telomere to the right). Light orange fill shows quantification using ten megabase windows with one megabase separation. Darker orange 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.

to 80 genes per one megabase window. Using RNA-Seq and RNA pol II ChIP-Seq datasets (Appendix A.2.3 for NCBI GEO accession numbers) I selected highly expressed genes and generated 4C-like association datasets for them. Taking this approach instead of studying two-dimensional heatmaps gives better resolution and allows finer control of association visualisations.

6.4.1 The Surfeit Cluster

The consensus association for *BCR* across the datasets is that there are three major peaks of association in chromosome 9 q34. The *ABL1* gene sits between the two most centromeric of these. The closest of these peaks contains the *Surfeit* cluster of genes. This is a cluster six housekeeping genes not related in sequence contained within 45 Kb. The *SURF1* - *SURF5* genes alternate in orientation (Huxley and Fried, 1990) and *SURF1* and *SURF2* share a bidirectional promoter (Lennard et al., 1994). *SURF3* (also known as *RPL7A*) encodes the 60S ribosomal protein L7a and is co-transcribed with the small nucleolar RNA genes U24, U36a, U36b, and U36c

(Giallongo et al., 1989; Nicoloso et al., 1996). The cluster is highly conserved from mouse and chicken to human (Duhig et al., 1998).

I hypothesised that a cluster of genes with such high expression may be responsible for the organisation of the surrounding chromatin and could be driving the observed association with chromosome 22. Generating 4C like association profiles from Hi-C and TCC reads in the *Surfeit* cluster shows strong associations over the region of chromosome 22 containing the *BCR* gene (Fig 6.4.1).

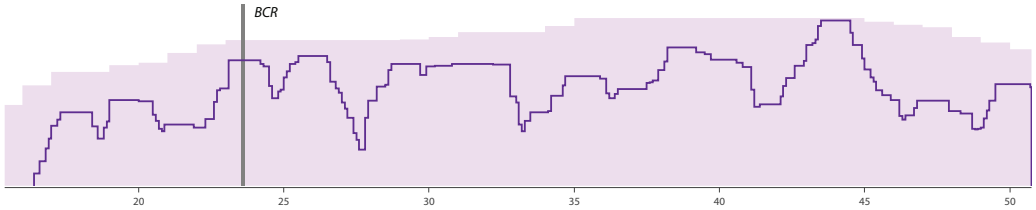


Figure 6.4.1 – GM06990 HiC *Surfeit* locus association with chromosome 22. Association of reads in a one megabase surrounding the *Surfeit* cluster, normalised for *NcoI* site distribution, across chromosome 22 (centromere to the left, telomere to the right). Light purple fill shows quantification using ten megabase windows with one megabase separation. Darker purple line shows quantification using one megabase windows with 100 Kb separation. X axis shows chromosome position in megabases. *BCR* is shown as a grey bar.

6.5 Validation by 3C

The results of the e4C indicate association frequencies. The analysis of the data makes a number of assumptions, namely 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, notably the two PCR amplification steps which may preferentially amplify some ligation products irrespective of their initial frequency. Because of these potential biases traditional RT-qPCR 3C was a logical method to validate the association. Additionally, the RT-qPCR 3C may

provide a resolution of association frequencies at a superior to the e4C, both in terms of quantitative association 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 *HindIII* fragments within the promoters of candidate genes described above. Four primers were used for each fragment, two either side of both restriction sites. 3C material prepared with *HindIII* was used to be compatible with co-worker's experiments and to avoid any biases that may be introduced by *AseI* digestion. An equimolar mix library to control for varying primer efficiencies was prepared by amplifying the fragments with genomic DNA, digesting with *HindIII*, mixing in equimolar concentrations and randomly ligating as described in Dekker (2006); Cope and Fraser (2009).

I interrogated three *HindIII* fragments covering the *Surfeit* locus, two fragments covering the *ABL1* promoters 1a and 1b. For use as a negative control I used a *HindIII* fragment covering the promoter of *SMC2*, an expressed gene 27 Mb centromeric of *ABL1* with low *BCR* e4C association.

I ran the 3C RT-qPCR using GM12878 3C material prepared with *HindIII*. I used GM12878 cells because the method uses a lot of 3C material and I wanted to do any optimisation required using a plentiful supply of cells.

6.5.2 Results

I ran two biological replicates of the 3C RT-qPCR with four technical replicates per primer pair. To calculate 3C product concentration I used a standard curve for each primer pair generated from six concentrations of the randomly ligated control

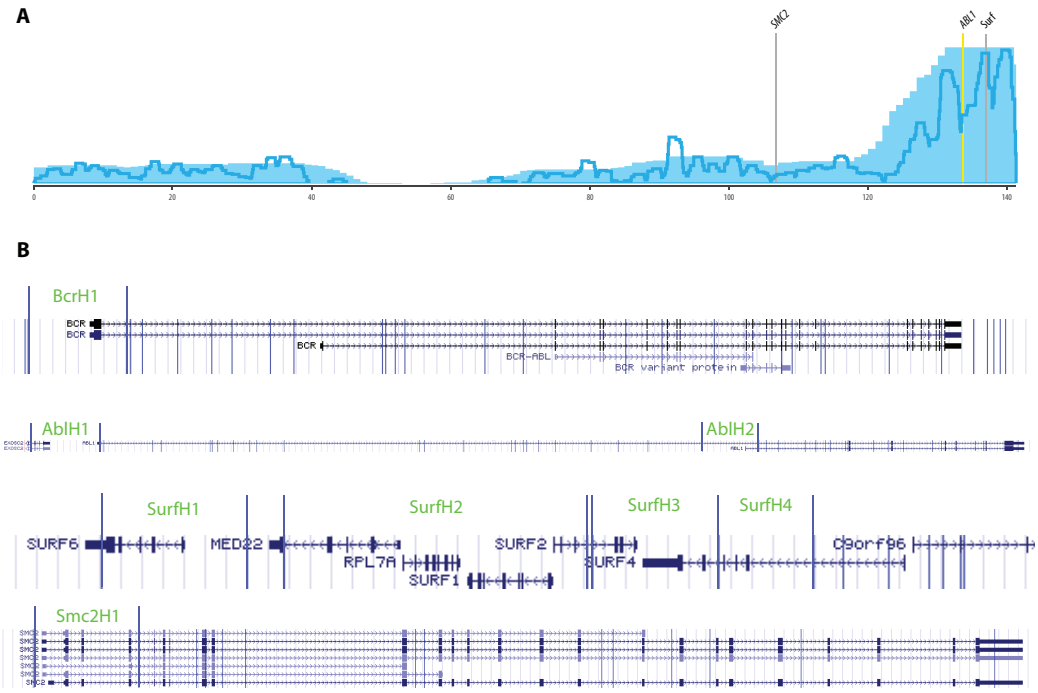


Figure 6.5.1 – 3C RT-qPCR Primer Locations. **(A)** BCR e4C association over telomeric chromosome 9 (shading - ten megabase windows, line - one megabase windows). **(B)** Screenshots of candidate genes taken from UCSC. *HindIII* restriction sites shown as blue lines, named fragments labelled in green text. Not shown at equal scale.

library and a no template control. To reach a final value for each primer pair I averaged the concentrations of the four technical replicates and plotted this with a standard deviation (Fig 6.5.2).

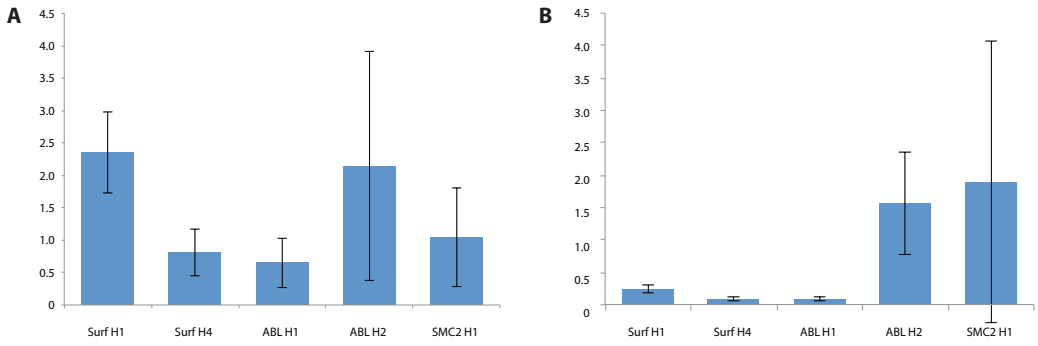


Figure 6.5.2 – 3C qPCR Results. Two repeats **(A)** and **(B)** showing average values for four replicates per sample in 3C RT-qPCR. Y values are concentration in μM, calculated from standard curve for each primer pair. Error bars show standard deviations.

Unfortunately, as can be seen in Fig 6.5.2, 3C ligation product concentrations were not reproducible between biological or technical replicates. To check for non-specific amplification I plotted a dissociation curve of samples from 58 °C to 85 °C. The dissociation curve for the standard curve samples looked good, however the 3C samples and lowest standard curve concentration showed very small peaks at varying temperatures, throwing into question the validity of the qPCR C_t values. To look for specific amplification visually, I ran the 3C qPCR samples and the two least concentrated standard curve samples on a gel (Fig 6.5.3). This shows specific bands in the more concentrated standard curve lanes for all replicates, but a mixture of specific bands, non-specific smears and empty lanes for the 3C samples and lowest concentration of standard curve (Fig 6.5.3). From the 8 lanes for each primer pair, Surf H1 appears to have one band, Surf H4 eight bands, ABL H1 seven bands, ABL H2 four bands and SMC2 two very weak bands.

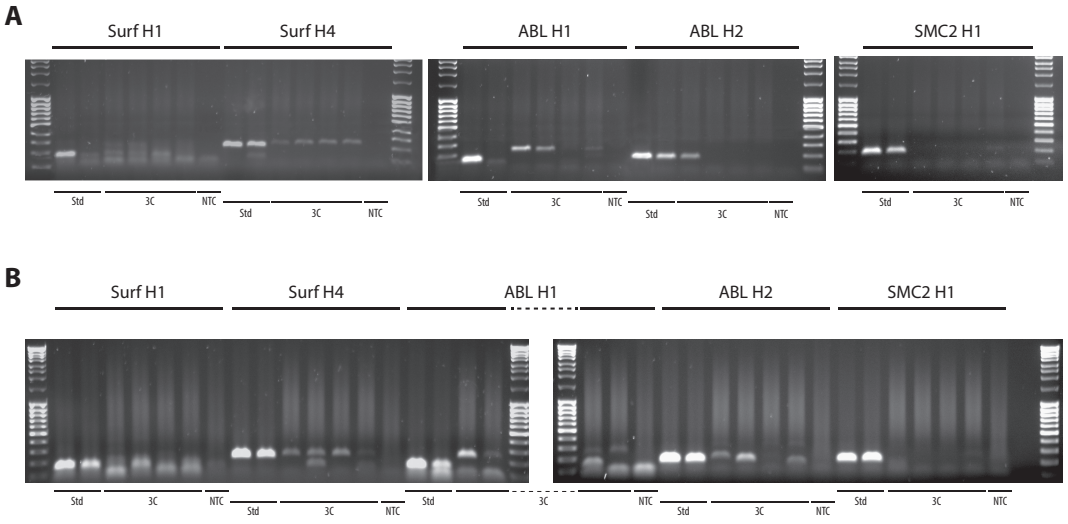


Figure 6.5.3 – Gel showing 3C qPCR Products. 1% agarose gel showing 3C RT-qPCR products, stained with Ethidium Bromide. Std = Standard curve products, 1.28 and 0.25 μM. 3C = four 3C RT-qPCR duplicates. NTC = No template control.

Empirically, this shows that the *Surfeit* cluster and *ABL1* have stronger associations with *BCR* gene than *SMC2*. The presence and absence of bands in replicates

demonstrates that these ligation products are very rare, due to the weak nature of *trans* contacts. This very low level of 3C ligation events and non-specific smearing due to lack of template means that 3C RT-qPCR cannot be used to accurately discern the relative association frequencies of these loci.

6.6 Validation by microscopy

Microscopy techniques are an excellent method of validating e4C data. Because the protocols are fundamentally different, they do not suffer the same biases as 3C-derived techniques, and are able to give information on a single-cell basis instead of just cell populations.

6.6.1 RNA-FISH and DNA-FISH

Initially, I attempted to use RNA-Fluorescence *In-Situ* Hybridisation (RNA-FISH) for this validation to show only genes that are being actively transcribed. RNA-FISH has excellent resolution as it only labels nascent transcripts, giving small sharp foci. However, after a great deal of setup and optimisation I switched to DNA-Fluorescence *In-Situ* Hybridisation (DNA-FISH). Because human blood products are a category II sample, they must be killed by fixation before being brought to the main laboratory. This change in the RNA-FISH protocol caused large problems and I was never able to collect any meaningful data.

DNA-FISH labels regions of genomic DNA using fluorescently labelled bacterial artificial chromosomes (BACs). It is a substantially less fickle technique than RNA-FISH due to the relative stability of DNA and the larger labelling target; a BAC is typically around one hundred kilobases in length whereas RNA-FISH probes are designed in non-repetitive intron sequences, usually a few hundred

base pairs long. Additionally, DNA-FISH labels two genomic loci in every nuclei, whereas RNA-FISH only labels nascent transcripts which may be far fewer in number.

6.6.2 DNA-FISH study design

For use in the DNA-FISH I used BAC probes covering target genes directly labelled with fluorescent moieties (555 nm for *BCR*, 488 nm for other genes). I used the genes tested in the 3C RT-qPCR: *BCR*, *ABL1*, *Surfeit* cluster and the *SMC2* negative control. I also used two new loci, *SPTAN1* and *QSOX2*, which are found at the apex of the centromeric and telomeric triplet peaks of association in chromosome 9. For use as a negative control in the DNA-FISH I also used *HLA-DMA*, a highly expressed gene on chromosome 6 within a region enriched for active marks but not displaying a specific association with *BCR* in the e4C (Table 6.6.1, Fig 6.6.1).

Gene	BAC ID	Chromosomal Position
BCR	CTD-2571K3	chr22:23,455,715-23,679,062
ABL1	RP11-83J21	chr9:133,652,008-133,828,473
Surfeit	RP11-152J3	chr9:136,206,288-136,374,469
SPTAN1	RP11-589E16	chr9:131,241,008-131,450,093
QSOX2	RP11-83N9	chr9:138,981,971-139,136,889
SMC2	RP11-989F24	chr9:106,791,249-106,969,527
HLA-DMA	RP11-629J17	chr6:32,449,530-32,684,058

Table 6.6.1 – DNA-FISH BACs. Table showing the BAC name and chromosomal position for each DNA-FISH probe used.

6.6.3 DNA-FISH analysis

Slides were analysed with the MetaSystems MetaCyte microscope and analysis software and a classifier set up able to identify DNA-FISH spots accurately. I exported data from the MetaCyte workstation with the three-dimensional co-ordinates

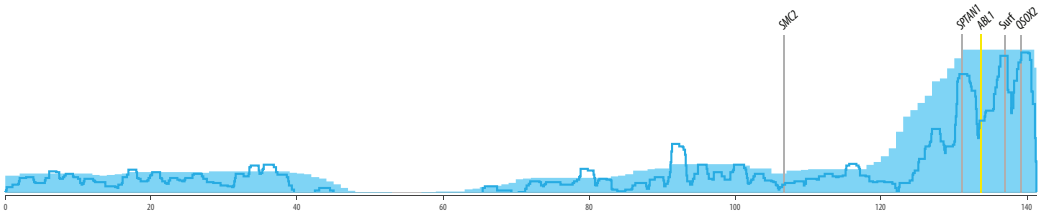


Figure 6.6.1 – DNA-FISH probe locations. Location of DNA-FISH probes on chromosome 9. e4C association profile with BCR using ten megabase windows with one megabase separation shown in light blue solid colour; association using one megabase windows with 100 Kb separation shown as solid blue line. HLA-DMA probe is located on chromosome 6. Identities and exact locations of BAC probes can be found in Table 6.6.1.

of each spot identified within each nuclei. Using a perl script written by Dr Felix Krueger in the Babraham Bioinformatics department I then calculated the distance between each pair of spots. I filtered this data in Microsoft Excel to give a list of signals corresponding to the shortest pair distance between each *BCR* signal and any target signal. I imported this data into GraphPad Prism for statistical analysis, using a one-way ANOVA test with a Tukey's range test to calculate significance and create box plots. The exception to this method was for the $CD34^+$ Run 2 which only had two samples, where I used .

6.6.4 DNA-FISH in $CD34^+$ cells

For the DNA-FISH with human $CD34^+$ cells I was able to make two sets of biological replicate slides. Cell numbers were limiting so I was not able to test all probe combinations for each sample. With slides from the first sample I tested the association of *BCR* (555 nm, red) against the *HLA-DMA*, *SMC2*, *ABL1* and *Surf* probes (Fig 6.6.2, Run 1). *BCR* was found significantly closer to *ABL1* and *Surf* than the negative control *SMC2* ($p < 0.001$), validating the observation of increased association over chromosome 9 q34 seen in the e4C (block colour ten megabase window analysis, Fig 6.6.1). The median association of *BCR* with the control probe

HLA-DMA was found to be slightly closer than *ABL1* and *Surf* in this experiment ($p < 0.1$, Fig 6.6.2). This gene was chosen as a control due to it's association with *BCR* in the e4C, so this result is not entirely unexpected. The reason for including it as a control was because the e4C data suggested that this association was no stronger than would be expected by the activity of the region. The expression of *HLA-DMA* is very strong in CD34$^+$ cells, so *BCR* may associate specifically with this gene below the level of resolution achieved in the e4C.

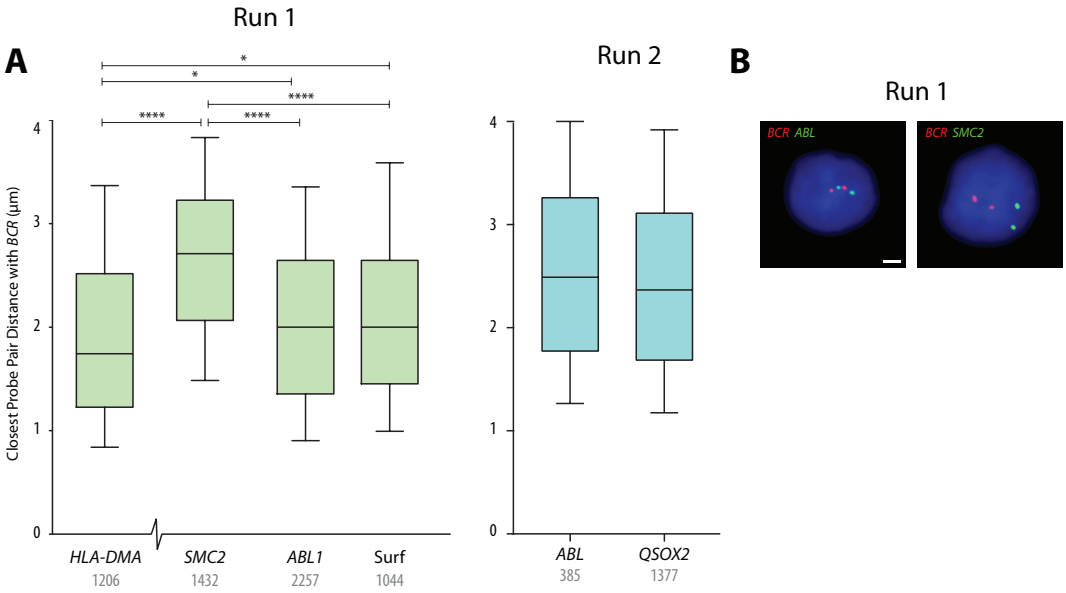


Figure 6.6.2 – CD34$^+$ DNA-FISH results. **(A)** Box plots showing target gene separation with *BCR* in CD34$^+$ cells. Nuclei with zero or greater than two signals in either channel were ignored. Central line shows median, box edges show 10th and 90th percentiles, whiskers show 5th and 95th percentiles. **(B)** Representative DNA-FISH images in CD34$^+$ nuclei. Scale bar is 2 µm.

The second biological replicate experiment was not as successful as the first, with a number of slides not reaching the required number of cells for analysis: any slides with less than 200 signals were discarded. I was able to record association of *BCR* with *ABL1* and *QSOX2*. The association with *QSOX2* was slightly tighter than with *ABL1*, though this margin was not significant (Fig 6.6.2).

6.6.5 DNA-FISH in GM12878 cells

I performed two replicate experiments with GM12878 cells using two batches of slides prepared from independent cell collections. In the first experiment I tested *BCR* (555 nm, red) in combination with the *HLA-DMA*, *SMC2*, *ABL1* and *Surf* genes (Fig 6.6.3). A similar pattern was observed to that seen in CD34$^+$ cells (Fig 6.6.2), with *BCR* exhibiting significantly closer association with *ABL1*, *Surf* and *HLA-DMA* ($p < 0.001$, Fig 6.6.3).. One difference was that the association of *BCR* with *ABL1* and *Surf* was closer than that with HLA-DMA ($p < 0.1$, Fig 6.6.3).

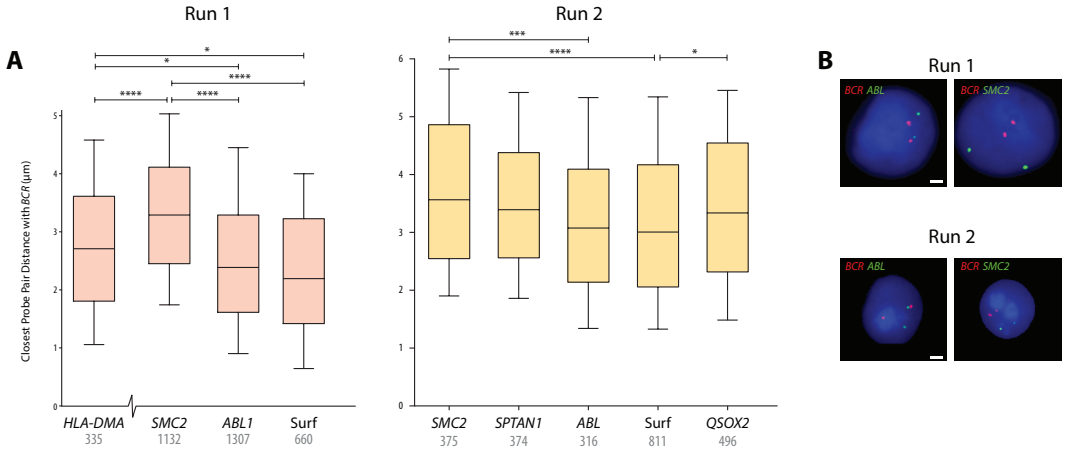


Figure 6.6.3 – GM12878 DNA-FISH results. (A) Box plots showing target gene separation with *BCR* in GM12878 cells. Nuclei with zero or greater than two signals in either channel were ignored. Central line shows median, box edges show 10th and 90th percentiles, whiskers show 5th and 95th percentiles. **(B)** Representative DNA-FISH images in GM12878 nuclei. Scale bar is 2 µm.

6.7 Discussion

In this chapter I describe data from my CD34$^+$ cell e4C with *BCR* bait showing that the strongest *trans* association with *BCR* is in chromosome 9 band q34 I demonstrate that this association is also present in Hi-C and TCC libraries generated from GM06990 and GM12878 cells and that this association is reciprocal. I go

on to discuss analysis to identify a candidate region driving the association and experiments designed to validate it using RT-qPCR 3Cand DNA-FISH.

This data shows a large scale enrichment of association between two regions of chromosomes 9 and 22 containing the genes *BCR* and *ABL1*. This association is specific and significant and validated by observations in both RT-qPCR 3C (empirical gel band counting, Fig 6.5.3) and DNA-FISH. The data suggests that there may be specific loci within these regions driving the association and causing *BCR* and *ABL1* to associate through a bystander effect, however the identification of specific loci with variations in association within chromosome 9 q34 was not conclusive. This could could be due to a number of reasons: the candidate loci chosen for the validation may have not been the genes driving the association; the variations in association between loci in this region may be too small to detect, or the variations shown in the e4C data may not be accurate. Chromatin associations in *trans* are significantly weaker than those in *cis*, making quantitative measurements difficult.

The observation that the most significant trans association of *BCR* is the region of chromosome 9 containing *ABL1* is an important finding. I discuss the implications of this in Chapter 7.

Chapter 7

Discussion

The three-dimensional organisation of the nucleus has a direct impact on the functional regulation of mammalian genomes (discussed in Chapter 1). In this thesis I describe how I have extended and further developed e4C, a technique based on chromosome conformation capture. I have used e4C to investigate the genome-wide association profiles of the proto-oncogenes *BCR*, *ABL1* and *MLL* in human $CD34^+$ cells and the GM12878 cell line.

In this chapter I discuss the broader context and implications of the work presented in this thesis.

7.1 The e4C methodology

The technology behind many molecular biology techniques has undergone a radical transformation in the past twenty years. The capacity of next generation sequencing techniques has rapidly increased, in a trend that seems set to continue. As the

quantity of data produced by experiments makes analysis through direct observation impossible, the methods we use to process and digest data become increasingly important.

The development of e4C sequencing described in this thesis acts as a paradigm of this progression: the original 3C technique is analysed through the visualisation of bands on a gel (Dekker et al., 2002); RT-qPCR 3C improves this by accurately quantifying the amplification (Hagège et al., 2007); 4C scaled this technique to use microarray technology (Simonis et al., 2006; Schoenfelder et al., 2010) and this thesis describes the use of e4C sequencing to produce many millions of sequencing reads across the genome.

With such large datasets, analysis methodologies must be systematic and as free of bias as possible. It is interesting to note the convergent approaches in techniques used to analyse association data found in the literature. In 2012 a methods paper was written by the de Laat laboratory describing the 4C-seq protocol and analysis (van de Werken et al., 2012). They describe many of the same problems and suggest some solutions that differ from those proposed in this thesis. For example, to overcome problems with cluster calling due to bait region similarity in sequencing they recommend spiking in phiX library, instead of Bareback processing. To normalise restriction endonuclease site distribution they use running probes created over multiple restriction fragments instead of calculating the proportion of fragments observed in fixed window sizes. Future studies using similar techniques can review a variety of analysis methodologies and choose those that suit their data best.

7.1.1 Future directions

The further development of e4C represents just one of many 3C derived techniques. These methodologies have generally scaled up from the investigation of individual loci to global all-to-all techniques such as HiC (Lieberman-Aiden et al., 2009) and TCC (Kalhor et al., 2011), able to investigate all genomic associations within a cell population in a single experiment. On initial inspection one might think that these techniques therefore supersede those before them, however they still suffer from an Achilles' heel: sequencing depth. Assuming a 3C library generated with a restriction endonuclease recognising a six base-pair site such as *HindIII*, *BglII* or *AseI*, there are approximately seven and a half thousand fragments in the human genome ($\frac{3.08\times10^9}{4^6}$ =752000). In order for a single association event to be recorded between every fragment in the genome, over half a billion sequence reads are needed ($752000^2 = 5.7 \times 10^{11}$). Quantitative analysis of association frequencies requires a great many more reads per fragment. For the level of complexity seen in the BCR e4C datasets described in this thesis a HiC library would require over seven thousand billion reads ($5.7 \times 10^{11} \times 12500 = 7.125 \times 10^{15}$), a number that would require thousands of runs with today's sequencing technology (the Illumina HiSeq 2500 is capable of three billion single end reads per run).

Recent all-to-all association studies investigating smaller genomes such as the yeast *Saccharomyces cerevisiae* and fruit fly *Drosophila melanogaster* give an indication to the potential of these technologies when capable of reaching full sequencing depth (Duan et al., 2010; Sexton et al., 2012). However, until sequencing technologies are able to reach similar depths of sequencing with the human genome, there is a role for e4C and other techniques able to interrogate specific subsets of 3C libraries. As described in this thesis, e4C enriches 3C libraries for a specific

bait, revealing an in-depth genome wide map of association for that fragment. Other approaches exist such as ChIP-e4C (Schoenfelder et al., 2010) and ChIA-PET (Fullwood et al., 2009) which use chromatin immunoprecipitation to enrich for associations taking place in concert with proteins of interest. I believe that similar techniques will continue to flourish to allowing the investigation of a myriad of micro-environments within the nucleus.

A number of as yet unexplored avenues exist within comparative and dynamic nuclear organisation. Comparative studies may yield new understanding of the differences between different tissues, healthy and disease states, the evolution of genome structure and heterogeneity between single cells and populations. Studying the dynamics of genome organisation can give us new insights into cell cycle progression, tissue differentiation, the processes driving nuclear organisation and the effect of pharmacological agents. As our understanding of the gross rules governing these processes increases we will be able to better understand how differences in nuclear organisation can affect biological function, and how biological function can affect organisation. It is not unreasonable to expect diagnostic tests based on nuclear organisations to reach the clinic in the future, along with drugs able to modify organisation, especially as preventative measures.

7.2 An active nuclear compartment

The data discussed in chapter four suggests that the proto-oncogenes *BCR*, *ABL1* and *MLL* reside within an active nuclear compartment, defined by the presence of active epigenetic marks. The existence of such active and inactive nuclear compartments has been suggested by a number of recent studies. Lieberman-Aiden *et al.* used HiC to investigate the nuclear organisation of GM06990 and K562 cells.

They proposed that if two genomic loci are nearby in three-dimensional space, they will have similar genome wide association profiles. They plotted intrachromosomal heat maps of Pearson correlation matrices and observed a stark plaid pattern, which was split into two genomic compartments using principal component analysis. The authors characterised the active component as showing looser compaction and correlating with gene density, mRNA expression, DNAse sensitivity and active histone marks (Lieberman-Aiden et al., 2009). This observation has been replicated by other groups (Yaffe and Tanay, 2011; Kalhor et al., 2011; Zhang et al., 2012) and is supported by earlier studies showing associations between active loci (Simonis et al., 2006; Schoenfelder et al., 2010) and domains of inactive chromatin (Guelen et al., 2008).

The existence of active and inactive compartments within the genome is compatible with the observation of transcription occurring at fixed transcription factories, as well as models of chromatin loops described in Section 1.6. Clustering of active regions suggests a model of genomic organisation whereby the transcriptional activity of genomic loci can be controlled by the adjustment of their position in the nucleus. As inactive genes are stimulated by external factors, they can be epigenetically remodelled allowing escape from a repressive environment and recruitment to a transcription factory (Chambeyron and Bickmore, 2004; Osborne et al., 2004). Future studies into the dynamics and control of these compartments will surely elucidate finer detail in the mechanisms by which mammalian gene expression is controlled.

7.3 Proto-oncogene associations

In chapter six I describe the co-association of *BCR* with chromosome 9 band q34, the region containing the t(9;22)(q34;q11) translocation partner gene *ABL1*. The association of these two loci has been studied before (Kozubek et al., 1997; Lukásová et al., 1997; Neves et al., 1999; Kozubek et al., 1999; Schwarz-Finsterle et al., 2005), as has the association of *MLL* with its translocation partner genes (Murmann et al., 2005; Gué et al., 2006; Cowell et al., 2012). These studies differ from the work described in this thesis by their use of FISH to measure association. While they are able to show significantly enriched association in comparison to candidate control loci, they cannot describe the association in the context of all genomic contacts. The BCR bait e4C data in this thesis suggests that the *BCR* : chr9 q34 association is the strongest *trans* association made by the *BCR* locus in the entire genome.

This data supports the hypothesis that chromosomal associations may play an important role in the formation of chromosomal translocations (Section 1.9). Understanding the process of translocation formation is important for the continued development of cancer treatments. For example, if therapy-related leukaemias involving the *MLL* gene are caused by topoisomerase induced DSBs during transcription, in the future we may see precautionary drugs able to modify the activity or localisation of the *MLL* gene. The greater our understanding of biological processes, the greater our ability to control them to prevent and cure disease.

Appendix A

Appendices

A.1 Primers

A.1.1 RT-qPCR Primers

Table A.1.1 – RT-qPCR 3C Primers

Name	Sequence
hSurf H1-1	ACCTCAGGGTTGTCCTCTGTTC
hSurf H1-2	GACCTTCAGACCCAGGGAAAG
hSurf H1-3	ATGCAATCATCATCCATTCTGG
hSurf H1-4	GAAGGAACTGGCCAGACTTTGT
hSurf H2-1	TCAGGTTTTGTTCCTGAGAACG
hSurf H2-2	AAGCCTTTGCGAAGCTAATGAC
hSurf H2-3	ACAGTGTGCATGGTTGAGAAGG
hSurf H2-4	TCAAAAGCCCGATACTTCCCTA
hSurf H3-1	CAGGAAGAGGCTGGAAGTCCT
hSurf H3-2	TCTGTCCATTTCCCTCCTTAGC
hSurf H3-3	TGCAACATCCTCTTCAGAACCT
hSurf H3-4	GGTTTGAGTAGACTTCTCTGGGTCT

hSurf H4-3	GTAAACACGGCGAGCACATAAA
hSurf H4-4	TGAACCATATCCTGCTTGATGG
hBCR H1-3	TAGTGTGGTTCTTGCGATCTGG
hBCR H1-4	ACTCTCTGCCCTCCAACTTCTG
hABL H1-1	AGGTGCAGCTGTCTCTCTTCCT
hABL H1-2	GCAACCTCGTACAAGAAAAGCA
hABL H1-3	GAGATGCAGCGAATGTGAAATC
hABL H1-4	CACAAAAACATTGCAGTGTGGA
hABL H2-1	CCAGGGGCCTAATAAGGAAGAG
hABL H2-2	GCAGGCCACTCACTCTCTATGA
hABL H2-3	CAGTGTTTATTGACGCCCACTC
hABL H2-4	TTGCTTTGACAGCTAGGCTGAG
hSMC2 H1-1	CTCTCTGGCCCCAAGAAGTACA
hSMC2 H1-2	GCTGGCTGTGGCTTACTTTTCT
hSMC2 H1-3	GTCCGGCCATCTGTTTAGAAAT
hSMC2 H1-4	TGAGATGCTCCCATCGACTTTA
hZcchc6 H1-1	TGTGGCAGGAGATATACGCAGT
hZcchc6 H1-2	GGGGACCATATCGTAATTGCTC
hZcchc6 H2-1	CACATTGTTGAATAATACCCTTCC
hZcchc6 H2-2	CAAAAGTTTGCATGTTTTTGCT
hZcchc6 H2-3	GGGCACCGAGAAATTAGTTTTG
hZcchc6 H2-4	AAAGGGATCGTGACAAACAGGT

A.2 e4C Library Statistics

A.2.1 Numbers of e4C reads per Chromosome

Table A.2.1 – 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%)

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
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%)
4	Hits	671498	330078	2153	473346	443583	2299	685665	528751	581340
	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%)

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
11	Hits	422600	239318	6851	4307812	2510508	12945	892290	14981823	13145547
	Hits (dedup)	425	338	94	196	409	145	73	790	858
	Asel 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
	Asel 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%)
13	Hits	334196	175512	558	257937	169930	854	139197	505106	392592
	Hits (dedup)	326	268	20	39	189	21	34	64	64
	Asel 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%)
14	Hits	322578	187374	271	58770	36651	164	627703	191766	273517
	Hits (dedup)	310	282	10	22	165	11	51	47	43
	Asel 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%)
15	Hits	373427	188154	392	190630	130416	722	732461	277002	255223
	Hits (dedup)	322	272	9	20	148	21	67	55	49
	Asel 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%)
16	Hits	226640	168317	169	73883	296458	1645	533661	221188	169482
	Hits (dedup)	216	197	3	11	139	14	50	34	43
	Asel 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%)
17	Hits	512135	299241	586	367490	68173	344	1349407	264994	507019
	Hits (dedup)	417	336	13	16	202	13	86	54	64
	Asel 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%)
18	Hits	258583	134215	23	149524	158156	895	253479	116454	145111
	Hits (dedup)	250	180	6	24	132	22	38	33	32
	Asel 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%)

		BCR Run 1	BCR Run 2	MLL Run 1	MLL Run 2	MLL Run 3	MLL Run 3 PhiX	ABL1	MLL_1	MLL_2
19	Hits	255305	163418	34	82337	74237	305	287777	135796	353231
	Hits (dedup)	212	203	4	14	111	8	33	17	43
	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%)
20	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%)
21	Hits	211560	96444	473	213499	106146	516	354888	54673	143351
	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%)
22	Hits	5151402	4148254	71	24387	138754	585	280056	251658	142409
	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%)
X	Hits	199693	90259	559	106708	105365	547	246552	234679	217705
	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%)
Y	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%)

A.2.2 *AseI* Fragment Statistics

Near *cis*: within 2.5 megabases either side of bait gene. Far *cis*: beyond 2.5 megabases either side of bait gene.

Table A.2.2 – e4C library *AseI* fragment statistics.

		All Fragments	Hit Fragments	% Hit
BCR Run1	*trans*	1425596	11,031	0.77%
	near *cis*	1261	653	51.78%

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

A.2.3 Datasets used for active mark correlations

Accession codes for publicly available datasets:

Table A.2.3 – Accession codes.

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

A.2.4 Library trimming statistics

Percentage of possible sequences that were successfully aligned are shown in italics.

Table A.2.4 – Library trimming statistics.

	Reads starting with correct barcode	Reads with second *AseI* site (discarded)	Reads with *NlaIII* site (too short to be mapped)	Reads with *NlaIII* site (long enough to be mapped)	Reads with no *NlaIII* site (long enough to be mapped)	Aligned reads
***BCR* Run 1**	22390498	85448 (0.4%)	3823824 (17.1%)	2860849 (12.8%)	18203128 (81.3%)	14931033 (66.7%) *(82.0%)*
***BCR* Run 2**	12616122	41823 (0.3%)	1493641 (11.8%)	1146963 (9.1%)	10932276 (86.7%)	9260876 (73.4%) *(84.7%)*
***MLL* Run 1**	55868	15 (0.03%)	29279 (52.4%)	1819 (3.3%)	26129 (46.8%)	21923 (39.2%) *(83.9%)*
***MLL* Run 2**	16442664	45012 (0.03%)	1388307 (8.4%)	662684 (4.0%)	14950490 (90.9%)	13535278 (82.3%) *(90.5%)*
***MLL* Run 3**	9960249	1535 (0.02%)	557742 (5.6%)	450422 (4.5%)	9368936 (94.1%)	8498056 (85.3%) *(90.7%)*
***MLL* Run 3 phiX**	56994	321 (0.6%)	1974 (3.5%)	1983 (3.5%)	48952 (85.9%)	44194 (77.5%) *(90.3%)*

***ABL* Run 1**	32140483	434554 (1.4%)	201658 (0.6%)	988675 (3.1%)	31082375 (96.7%)	27117770 (84.4%) *(87.2%)*
***MLL* Promoter 1**	27548835	243399 (0.9%)	485996 (1.8%)	1124315 (4.1%)	26290533 (95.4%)	23688930 (86.0%) *(90.1%)*
***MLL* Promoter 2**	30333263	232741 (0.8%)	1917547 (6.3%)	1875757 (6.2%)	27748869 (91.5%)	22638880 (74.6%) *(81.6%)*

A.2.5 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.

Table A.2.5 – e4C library read counts.

		Total reads	**Near *cis* ($\leq$ 5 megabases)**	**Far *cis* ($>$ 5 megabases)**	***trans***
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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	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 *(per 10^6 reads)*	9318 *(410)*	18730 *(2128)*	10981 *(2519)*	6084 *(636)*

A.3 e4C analysis scripts

A.3.1 Bait processing

This script was originally written by Dr Felix Krueger of the Babraham Bioinformatics department.

```
    ,
#!/usr/bin/perl
use warnings;
use strict;

### This script intends to analyse Phil's FastQ files and sort the sequences
    according to the first 3 bp barcode information. The experiment was a 4C
### analysis with two different baits which were ligated to prey sequences using
    an AseI digest. Therefore all of the sequences should look like:
### XXX ATTAAT NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN whereby XXX is the bait-
    specific barcode information, ATTAAT is the AseI site, and NNNN is the prey
### sequence.

my %baits = (
        MLL => {
              name => 'MLL_bait',
          sequence => 'TTT',
         },
        BCR => {
          name => 'BCR_bait',
          sequence => 'GGA',
```

```
18          }
19        );
20 my $barcode_detected = 0;
21 my $count = 0;
22 my $asei_counter = 0;
23 my $nla_counter = 0;
24 my $double_asei_counter =0;
25 my $nla_too_short_counter = 0;
26 my $sequence_in_file = shift @ARGV;
27 my $results;
28
29 my @bait=();
30 foreach my $bait (keys %baits){
31   # resetting all counters
32   ($double_asei_counter,$nla_too_short_counter,$count,$asei_counter,$nla_counter,
         $barcode_detected) = (0,0,0,0,0);
33   $results = '';
34   @bait = ();
35   @bait = split (//,$baits{$bait}->{sequence});
36   print "length of barcode sequence in $baits{$bait}->{name} is ",scalar@bait," bp
         \t";
37   print join (" ",@bait),"\n";
38   $results = create_filehandles($baits{$bait}->{name});
39
40   read_file ();
41   warn "Total number of sequences processed: $count\n\n";
42   warn "The correct barcode ",@bait," was detected in $barcode_detected cases\n";
43   warn "Total number of prey sequences containing an AseI and an NlaIII site which
          are getting too short (hence being discarded): $nla_too_short_counter\n";
44   warn "Total number of prey sequences containing two  AseI sites (hence being
         discarded): $double_asei_counter\n";
45   warn "Total number of prey sequences containing a AseI site as well as an NlaIII
          site which are still large enough to map: $nla_counter\n";
46   warn "Total number of prey sequences with an intact AseI site printed out (
         including long enough NlaIII sequences): $asei_counter\n\n";
47 }
48
49 sub create_filehandles {
50   my $bait_name = shift;
51   my %filehandles;
52   my $outfile = "${bait_name}_seqs.txt";
53   open ($filehandles{isbait},'>',$outfile) or die $!;
54   warn "Writing bait-side output to $outfile\n";
55   return \%filehandles;
56 }
57
58 sub read_file {
59   open (my $SEQ,$sequence_in_file) or die "Can't read sequence: $!";
60   while (1) {
61     my $seq = read_next_sequence ($SEQ);
62     last unless ($seq);
63     #   last if ($count > 1000000);
64     ++$count;
65     if ($count % 2500000 == 0) {
66       warn "Processed $count sequences\n";
67     }
68     process_sequence($seq);
69   }
70 }
71 sub read_next_sequence {
72   my ($fh) = @_;
73   my $seq;
74   # First line should be the id
75   my $id_line = <$fh>;
```

```
76    return undef unless ($id_line);
77    $seq->{id} = $id_line;
78    chomp $seq->{id};
79
80    # Next line is the actual sequence
81    $seq->{seq} = <$fh>;
82    chomp $seq->{seq};
83
84    # next line is like first line just with a + in the start
85    $seq->{third_line} = <$fh>;
86    chomp $seq->{third_line};
87
88    # finally we get the quality string
89    $seq->{qual} = <$fh>;
90    chomp $seq->{qual};
91
92    return $seq;
93  }
94
95
96  sub process_sequence {
97    my $seq = shift;
98    # $seq is a sequence which contains all 4 lines of the FastQ file. None of the
          lines contains a \n.
99    if (is_bait($seq)) {
100     ++$barcode_detected;
101     process_bait_end($seq);
102   }
103 }
104
105 sub is_bait {
106   my $seq = shift;
107
108   my @seq = split(//,$seq->{seq});
109   my @quality = split(//,$seq->{qual});
110
111   my $n_count = 0;
112   my $match_count = 0;
113   my $mismatch_count = 0;
114
115   for my $index (0..$#bait) {
116     last if ($index > $#seq);
117
118     if ($seq[$index] eq 'N') {
119       ++$n_count;
120       next;
121     }
122
123     if ($seq[$index] eq $bait[$index]) {
124       ++$match_count;
125     }
126     else {
127       ++$mismatch_count;
128     }
129   }
130
131   # barcode sequence is a perfect bait match
132   if ($match_count == 3) {
133     return 1;
134   }
135   else{
136     return 0;
137   }
138 }
```

```
139
140 sub process_bait_end {
141   my $seq = shift;
142
143   # Check for the presence of an AseI site
144   if ($seq->{seq} =~ /^(\w{3})(ATTAAT\w+)$/ ){
145     my $barcode = $1;
146     my $downstream_sequence = $2;
147     my $offset = length($seq->{seq})-length($downstream_sequence);
148
149     # kicking the sequence out if it contained another AseI site
150     if ($downstream_sequence =~ /^ATTAAT\w*ATTAAT/){
151       ++$double_asei_counter;
152       return;
153     }
154
155
156     # Recognition site for NlaIII is: CATG (CATG #cut)
157     if ($downstream_sequence =~ /^(.*CATG)/ ){
158       $downstream_sequence = $1;
159       # discarding very short fragments
160       if (length($downstream_sequence) < 25){
161   # print "$downstream_sequence too short!\n";
162   ++$nla_too_short_counter;
163   return;
164       }
165
166       # otherwise we count how many fragments do contain both AseI and NlaIII
167       ++$nla_counter;
168       ++$asei_counter;
169     }
170
171     # does contain an intact AseI site but no NlaIII site
172     elsif ($downstream_sequence !~ /CATG/ ){
173       ++$asei_counter;
174     }
175     else {
176       die "Either there is an NlaIII site or not!\n";
177     }
178     # print "$barcode\t$downstream_sequence\t$seq->{seq}\n";
179     # Print out the sequence downstream of the to a FastQ file
180     print {$results->{isbait}} $seq->{id},"\n";
181     print {$results->{isbait}} $downstream_sequence,"\n";
182     print {$results->{isbait}} $seq->{third_line},"\n";
183     print {$results->{isbait}} substr($seq->{qual},$offset,length
          $downstream_sequence),"\n";
184   }
185 }
```

A.3.2 *In-silico* restriction fragment libraries

```
,
1 #!/usr/bin/perl
2 use warnings;
3 use strict;
4 ############################################################
5 # Name: Potential Hits (4C)                                #
6 # Authors: Phil Ewels                                      #
7 # Version 1.0 - 08/04/2011                                 #
8 # -------------------------------------------------------- #
9 ############################################################
10
```

```
11 ##########################
12 ## CONFIGURATION OPTIONS ##
13 ##########################
14
15 # Path to chromosome fasta files. Replace chromosome number with %s
16 my $fn_base = 'D:\Genome Sequences\Human\chromFa\chr%s.fa';
17 #my $fn_base = 'D:\Genome Sequences\Human\GRCh37\Homo_sapiens.GRCh37.55.dna.
       chromosome.%s.fa';
18
19 # Path to output file
20 my $output = 'AseI_NlaIII_fragments.txt';
21
22 # Chromosomes to use. Default is (1..22,'X','Y') - other options are 'MT' etc.
23 my @chromosomes = (1..22,'X','Y');
24
25 # First restriction site to use (3C digestion - assumes palindromic 6 cutter)
26 my $re_search1 = 'ATTAAT';
27
28 # Second restriction site to use (4C digestion)
29 my $re_search2 = 'CATG';
30
31 # Size selection parameters (in base pairs)
32 my $minsize = 10;
33 my $maxsize = 700;
34
35 ##################################
36 ## END OF CONFIGURATION OPTIONS ##
37 ##################################
38
39 open (OUT,'>',$output) or die $!;
40
41 # go through each chromosome
42 foreach my $chromosome (@chromosomes) {
43   my $filename = sprintf($fn_base, $chromosome);
44   open (IN,$filename) or die "Can't read file: $!";
45   warn "Starting Chromosome $chromosome ($filename)\n";
46   my $sequence = '';
47   $_ = <IN>; # Remove fasta header
48   while (my $line = <IN>) {
49     chomp ($line);
50     $sequence .= $line;
51   }
52
53   # Search for restriction enzyme sites (greedy regex)
54   while ($sequence =~ /$re_search1(.*?)$re_search1/g) {
55     # Kick out fragments with lots of N's (centromeres and telomeres)
56     next if (index ($1,'NNNNN') >=0);
57     # Add fragment to array, with half of restriction site added back on either
           end
58     # Not using real site so we can revcomp and join without modifying (if AseI =
           AT^TAAT, fragment is AAT...ATT)
59     my $fragment = substr($re_search1,2,4).$1.substr($re_search1,0,2);
60
61     # Only proceed if we have a NlaIII site (otherwise ignore fragment)
62     if ($fragment =~ /CATG/i) {
63
64       # Get sequence from start of string to first NlaIII site
65       $fragment =~ /(^.+CATG?).*$/i;
66       #print OUT "$chromosome_name\t$start\t$end\t$length_fragment\t$string\n";
67       #warn "$chromosome\t".pos($sequence)."\t".(pos($sequence)+length("${1}CATG")
             )."\t+\t".length("${1}CATG")."\t${1}CATG\n";
68       if(length($fragment) > $minsize && length($fragment) < $maxsize) {
69         print OUT "$chromosome\t".pos($sequence)."\t".(pos($sequence)+length("${1}
               CATG"))."\t+\t".length("${1}CATG")."\t${1}CATG\n";
```

```
70        }
71
72        #  Get sequence from last NlaIII site to end of string
73        $fragment =~ /^.+CATG(.+$)/i;
74        #warn "$chromosome\t".pos($sequence)."\t".(pos($sequence)+length("${1}CATG")
              )."\t-\t".length("${1}CATG")."\t${1}CATG\n";
75        #sleep(1);
76        if(length($fragment) > $minsize && length($fragment) < $maxsize) {
77          print OUT "$chromosome\t".pos($sequence)."\t".(pos($sequence)+length("${1}
                CATG"))."\t-\t".length("${1}CATG")."\t${1}CATG\n";
78        }
79      }
80      # Move array pointer back six so that we don't skip a fragment
81      pos($sequence) -= 6;
82    }
83    close IN;
84  }
```

A.3.3 GC content and fragment length bias detection

The fragment length bias script is very similar to the GC bias and so not printed here due to space constraints.

```
    ,
1  #!/usr/bin/perl
2  use warnings;
3  use strict;
4  use Math::Round;
5  ############################################################
6  # Name: GC Bias                                            #
7  # Author: Phil Ewels                                       #
8  # Version 1.0 - 04/11/2011                                 #
9  ############################################################
10
11 # Set input file by command line
12 my ($input) = @ARGV;
13 if (!defined $input) {
14   die "Usage is GC_bias.pl [input file]\n";
15 }
16
17 # Get genome into a hash
18 warn "Loading Genome...\n";
19 my $fn_base = 'D:\Genome Sequences\Human\chr%s.fa';
20 my @chromosomes = (1..21,'X','Y');
21 my %chrom;
22 foreach my $chromosome (@chromosomes) {
23   my $filename = sprintf($fn_base, $chromosome);
24   open (IN,$filename) or die "Can't read file: $!";
25   $_ = <IN>; # Remove fasta header
26   while (my $line = <IN>) {
27     chomp ($line);
28     $chrom{$chromosome} .= uc($line); # Make everything upper case
29   }
30   warn "Chromosome $chromosome loaded...\n";
31 }
32
33 # Load annotated report
34 open (IN,$input) or die "Can't read file: $!";
35 $_ = <IN>; # File Header
36
37 my %all_frags;
38 my %hit_frags;
```

```
39
40 while (my $line = <IN>) {
41   chomp ($line);
42   my @field = split(/\t/, $line);
43   # $field[1] = Chromosome
44   # $field[2] = Start
45   # $field[3] = End
46   if (exists $chrom{$field[1]}) { # Check that we have the chromosome for this
         fragment
47     my $seq = substr($chrom{$field[1]},$field[2],($field[3] - $field[2] + 1));
48     my $at = ($seq =~ tr/AaTt//);
49     my $gc = ($seq =~ tr/GgCc//);
50     $at = (int($at)) ? $at : 0; # make sure that vars are numeric
51     $gc = (int($gc)) ? $gc : 0;
52 #   print "$seq\n$at\t$gc\n\n"; sleep(1);
53     if(($gc + $at) > 0){ # kick out empty strings
54       my $ratio = $gc / ($gc + $at);
55 #     print "$at\t$gc\t$ratio\n";
56       $ratio = nearest(0.05, $ratio); # rounds to the nearest 5%
57 #     print "$at\t$gc\t$ratio\n\n"; sleep(1);
58       $all_frags{$ratio} += 1; # count for all fragments
59       if($field[11] eq '1.0') { # count only for hit hits
60         $hit_frags{$ratio} += 1;
61       }
62     }
63   }
64 }
65
66 my $all_num_frags = 0;
67 my $hit_num_frags = 0;
68 while ((my $key, my $value) = each(%all_frags)){
69   $all_num_frags += $value;
70 }
71 while ((my $key, my $value) = each(%hit_frags)){
72   $hit_num_frags += $value;
73 }
74 print "\n\n$all_num_frags\t$hit_num_frags\n\n";
75
76 # Open output file
77 open (OUT, '>',$input."_GCbias_output.txt") or die "Can't read file: $!";
78
79 print OUT "% GC\tAll Frags\tHit Frags\t% All \t% Hit \t% Hit / % All\n";
80 foreach my $key (sort keys %all_frags) {
81   my $percent_all = 0;
82   my $percent_hit = 0;
83   my $num_hit = 0;
84   my $gc_percent = $key * 100;
85   $percent_all = $all_frags{$key}/$all_num_frags;
86   if (exists $hit_frags{$key}) {
87     $num_hit = $hit_frags{$key};
88     $percent_hit = $hit_frags{$key}/$hit_num_frags;
89   }
90   my $hit_all = $percent_hit / $percent_all;
91   $hit_all = ($hit_all == 0) ? 1 : $hit_all;
92   $percent_all = sprintf("%.5f", $p
```

A.3.4 Systematic bias correction

```
,
1 #!/usr/bin/perl
2 use warnings;
3 use strict;
```

```
use Math::Round;
############################################################
# Name: GC Bias                                            #
# Author: Phil Ewels                                       #
# Version 1.0 - 04/11/2011                                 #
############################################################

####################
####   SETUP    ####
####################
my $GC_binsize = 5/100;
my $AN_fraglength_binsize = 10;



# Set input file by command line
my ($input) = @ARGV;
if (!defined $input) {
  die "Usage is GC_bias.pl [input file]\n";
}

# Get genome into a hash
warn "Loading Genome...\n";
my $fn_base = 'D:\Genome Sequences\Human\chr%s.fa';
my @chromosomes = (1..21,'X','Y');
my %chrom;
foreach my $chromosome (@chromosomes) {
  my $filename = sprintf($fn_base, $chromosome);
  open (CHROMOSOMES,$filename) or die "Can't read file: $!";
  $_ = <CHROMOSOMES>; # Remove fasta header
  while (my $line = <CHROMOSOMES>) {
    chomp ($line);
    $chrom{$chromosome} .= uc($line); # Make everything upper case
  }
  warn "Chromosome $chromosome loaded...\n";
}

# Load annotated report
open (IN,$input) or die "Can't read file: $!";
$_ = <IN>; # File Header
my %all_frags_percent;
my %hit_frags_percent;
my %all_frags_length;
my %hit_frags_length;

########################################
####   WORK OUT CORRECTION FACTORS    ####
########################################
warn "\nCalculating statistics";
my $counter = 0;
while (my $line = <IN>) {
  $counter++;
  chomp ($line);
  my @field = split(/\t/, $line);
  my $chr = $field[1]; # Chromosome
  my $start = $field[2]; # Start
  my $end = $field[3]; # End
  my $hit = ($field[11] eq '1.0') ? 1 : 0;
  if (exists $chrom{$chr}) { # Check that we have the chromosome for this fragment
    ########## GC CONTENT
    my $seq = substr($chrom{$chr},$start,($end - $start + 1));
    my $at = ($seq =~ tr/AaTt//);
    my $gc = ($seq =~ tr/GgCc//);
    $at = (int($at)) ? $at : 0; # make sure that vars are numeric
```

```
68      $gc = (int($gc)) ? $gc : 0;
69      if(($gc + $at) > 0){ # kick out empty strings
70        my $ratio = $gc / ($gc + $at);
71        $ratio = nearest($GC_binsize, $ratio);
72        $all_frags_percent{$ratio} += 1; # count for all fragments
73        if($hit) { # count only for hit hits
74          $hit_frags_percent{$ratio} += 1;
75        }
76      }
77
78      ######### FRAGMENT LENGTH
79      my $length = $end - $start + 1; # +1 because of genomic co-ordinates, eg. pos
            4 to 5 is 2 bases, 5-4 = 1
80      $length = nearest($AN_fraglength_binsize, $length);
81      $all_frags_length{$length} += 1; # count for all fragments
82      if($hit) { # count only for hit hits
83        $hit_frags_length{$length} += 1;
84      }
85    }
86    if($counter % 200000 == 0) { warn "$counter lines analysed\n"; }
87  }
88  warn "\nStatistics calculated.";
89
90  ##### Total number of fragments processed
91  my $all_num_frags = 0;
92  my $hit_num_frags = 0;
93  while ((my $key, my $value) = each(%all_frags_percent)){
94    $all_num_frags += $value;
95  }
96  while ((my $key, my $value) = each(%hit_frags_percent)){
97    $hit_num_frags += $value;
98  }
99  warn "\n\nNumber of Fragments Processed\nTotal: $all_num_frags\tHit:
        $hit_num_frags\n\n";
100
101 ##### Calculate GC correction values
102 warn "Caculating GC correction values\n";
103 my %GC_correction;
104 foreach my $key (sort keys %all_frags_percent) {
105   my $percent_all = 0;
106   my $percent_hit = 0;
107   $GC_correction{$key} = 1;
108   if (exists $hit_frags_percent{$key}) {
109     $percent_all = $all_frags_percent{$key} / $all_num_frags;
110     $percent_hit = $hit_frags_percent{$key} / $hit_num_frags;
111     $GC_correction{$key} = $percent_hit / $percent_all;
112   }
113 }
114 ##### Optional - print output GC correction values
115 warn "Printing GC correction values\n";
116 open (GC_OUT, '>',$input."_GC_correction_factors.txt") or die "Can't read file: $!
        ";
117 foreach my $key (sort keys %GC_correction) {
118   print GC_OUT ($key*100)."%\t".$GC_correction{$key}."\n";
119 }
120
121 ##### Calculate AseI - NlaIII fragment length correction values
122 warn "Caculating AseI - NlaIII fragment length correction values\n";
123 my %ANlength_correction;
124 foreach my $key (sort {$a<=>$b} keys %all_frags_length) {
125   my $percent_all = 0;
126   my $percent_hit = 0;
127   $ANlength_correction{$key} = 1;
128   if (exists $hit_frags_length{$key}) {
```

```
    $percent_all = $all_frags_length{$key} / $all_num_frags;
    $percent_hit = $hit_frags_length{$key} / $hit_num_frags;
    $ANlength_correction{$key} = $percent_hit / $percent_all;
  }
}

##### Optional - print output frag length correction values
warn "Printing AseI - NlaIII fragment length correction values\n";
open (FL_OUT, '>',$input."_fragLength_correction_factors.txt") or die "Can't read
    file: $!";
foreach my $key (sort {$a<=>$b} keys %ANlength_correction) {
  print FL_OUT "$key\t".$ANlength_correction{$key}."\n";
}

#######################################
##### OUTPUT
#######################################
##### Print final correction value for each fragment
warn "\nPrinting final correction value file\n\n";
my $starting_chance = $hit_num_frags / $all_num_frags;
open (OUT, '>',$input."_correction_values.txt") or die "Can't read file: $!";
# reset input file pointer
seek(IN, 0, 0);
# loop through input file again
$counter = 0;
while (my $line = <IN>) {
  $counter++;
  chomp ($line);
  my @field = split(/\t/, $line);
  my $chr = $field[1]; # Chromosome
  my $start = $field[2]; # Start
  my $end = $field[3]; # End
  my $chance = $starting_chance;
  if (exists $chrom{$chr}) { # Check that we have the chromosome for this fragment
    ########## GC CONTENT
    my $seq = substr($chrom{$chr},$start,($end - $start + 1));
    my $at = ($seq =~ tr/AaTt//);
    my $gc = ($seq =~ tr/GgCc//);
    $at = (int($at)) ? $at : 0; # make sure that vars are numeric
    $gc = (int($gc)) ? $gc : 0;
    if(($gc + $at) > 0){ # kick out empty strings
      my $ratio = $gc / ($gc + $at);
      $ratio = nearest($GC_binsize, $ratio);
      # Multiple the chance by the correction value for this ratio
      $chance = $chance * $GC_correction{$ratio}
    }

    ######### FRAGMENT LENGTH
    my $length = $end - $start + 1; # +1 because of genomic co-ordinates, eg. pos
        4 to 5 is 2 bases, 5-4 = 1
    $length = nearest($AN_fraglength_binsize, $length);
    # Multiple the chance by the correction value for
```

A.3.5 Mock *in-silico* e4C library generation

```
,
#!/usr/bin/perl
use warnings;
use strict;
use Math::Random::MT::Perl qw(srand rand);
use IO::File;
############################################################
```

```
# Name: Single Window p Value                                    #
# Author: Phil Ewels                                             #
# Version 1.0 - 04/11/2011                                       #
#################################################################

####################
####   SETUP    ####
####################
my $repeats = 5;

# Set input file by command line
my ($corrections_file) = @ARGV;
if (!defined $corrections_file) {
  die "Usage is single_window.pl [corrections file]\n";
}

# Open output filehandles
my @output;
for my $i (0 .. ($repeats-1)) {
  push(@output, IO::File->new('> library_'.($i+1).'.txt'));
}

# Load corrections
warn "Loading correction factors..\n\n";
my $counter = 0;
open (CORRECTIONS, $corrections_file) or die "Can't read file: $!";
while (my $line = <CORRECTIONS>) {
  $counter++;
  chomp ($line);
  my @field = split(/\t/, $line)
```

A.3.6 Restriction site search

```
,
#!/usr/bin/perl
use warnings;
use strict;
#################################################################
# Name: Single RE Fragments                                     #
# Author: Phil Ewels                                            #
# Version 1.0 - 05/05/2011                                      #
# ------------------------------------------------------------- #
#################################################################

#================================
#==== CONFIGURATION OPTIONS ====
#================================

# Set by command line (optional)
my ($output,$re_search) = @ARGV;
if (defined $output && !defined $re_search) {
  die "Usage is single_RE_fragments.pl [output file] [search string]\nLeave blank
      to use defaults\n";
} elsif(!defined $output) {
  # Path to output file
  $output = 'AseI_fragments.txt';
  # Restriction site to use
  $re_search = 'ATTAAT';
  warn "Using file defaults: search string = $re_search, output = $output\n";
} else {
  warn "Using command line variables: search string = $re_search, output = $output
      \n";
```

```
}

# Path to chromosome fasta files. Replace chromosome number with %s
my $fn_base = 'D:\Genome Sequences\Human\chr%s.fa';
#$fn_base = 'D:\Genome Sequences\Human\GRCh37\Homo_sapiens.GRCh37.55.dna.
    chromosome.%s.fa';
warn "Looking for Genome Sequences in $fn_base\n\n";

# Chromosomes to use. Default is (1..22,'X','Y') - other options are 'MT' etc.
my @chromosomes = (1..22,'X','Y');


#==================================
#== END OF CONFIGURATION OPTIONS ==
#==================================

open (OUT,'>',$output) or die $!;

# go through each chromosome
foreach my $chromosome (@chromosomes) {
  my $filename = sprintf($fn_base, $chromosome);
  open (IN,$filename) or die "Can't read file: $!";
  warn "Starting Chromosome $chromosome ($filename)\n";
  my $sequence = '';
  $_ = <IN>; # Remove fasta header
  while (my $line = <IN>) {
    chomp ($line);
    $sequence .= uc($line); # Make everything upper case
  }

  my ($offset, $lastpos, $pm) = (0, 1, "+");
  my $pos = index($sequence, $re_search, $offset);
  while ($pos != -1) {
    # Kick out fragments with lots of N's (centromeres and telomeres)
    #next if ( index( substr($sequence,$lastpos,($pos-$lastpos)),'NNNNN') >=0);
    # Too slow...

    print OUT $chromosome."\t". # Chromosome Name
         $lastpos."\t". # Position Start
         $pos."\t". # Position Finish
         $pm."\n"; # Arbitrary orientation to help visulaisation
    if ($pm eq "+") { $pm = "-"; } else { $pm = "+"; }
    $lastpos = $pos + length($re_search);
    $offset = $pos + length($re_search);
    $pos = index($sequence, $re_search, $offset);
  }

  close IN;
}
```

A.3.7 Restriction fragment distribution normalisation

This script was originally written by Dr Simon Andrews of the Babraham Bioinformatics department.

```
    ,
#!/usr/bin/perl
use warnings;
use strict;

my ($infile,$outfile,$window,$min_count) = @ARGV;

```

```
unless (defined $min_count) {
  die "Usage is make_windows.pl [input file] [output file] [window size] [minimum
      count]\n";
}

open (IN,$infile) or die "Can't read $infile: $!";

if (-e $outfile) {
  print "$outfile already exists.  Overwrite [Y/N]?";
  my $answer = <STDIN>;
  if ($answer !~ /^y/i) {
    die "Exiting...\n";
  }
}

open (OUT,'>',$outfile) or die "Can't write to $outfile: $!";

# This is the store of lines we're currently dealing with
my @current_lines;

# This is the chromosome we're dealing with
my $last_chromosome = '';

# Record the header
my $header = <IN>;
chomp $header;

my @header = split(/\t/,$header);

for my $l (0 .. $#header) {
  warn "header[$l] = $header[$l]\n";
}

print OUT join("\t",qw(Chr Start End Count),@header[10..$#header]),"\n";

while (<IN>) {

  chomp;
  my @sections = split(/\t/);

  # Check for putting out a status update
  if ($sections[1] ne $last_chromosome) {
    warn "Processing Chromosome $sections[1]\n";
    $last_chromosome = $sections[1];
  }

  if (@current_lines) {
    # Check if we can add this to the existing set

    if ($current_lines[0]->[1] eq $sections[1]
      and int ($current_lines[0]->[2] / $window) == int ($sections[3] / $window))
          {
      push @current_lines,\@sections;
    }
    else {
      check_valid_group(@current_lines);
      @current_lines = (\@sections);
    }

  }
  else {
    push @current_lines, \@sections;
  }

```

```
}

sub check_valid_group {

  my @sections = @_;

  if (@sections < $min_count) {
    return;
  }

  # If we're keeping this then we need to work out
  # the position of the window and record the
  # number of hits and the average measure for
  # each sample

  my $window_start = $window * int($sections[0]->[2]/$window);
  my $window_end = $window_start + $window;

  my @line = ($sections[0]->[1],$window_start,$window_end, scalar @sections);

  for my $index (10..$#header) {
    my $average;
    foreach my $probe (@sections) {
      $average += $probe->[$index];
    }
    $average /= @sections;
    push @line, $average;
  }
  print OUT join("\t",@line),"\n";

}
```

A.4 Web Tools

I am a keen advocate in the publication and sharing of bioinformatics tools. Some of the scripts that I wrote during my thesis could be useful for others, so I published them on my personal blog, at `http://www.tallphil.co.uk/bioinformatics`

These scripts are not printed here due to their length.

A.4.1 Sequences

Frustrated by the lack of a quick and simple tool to get reverse compliment sequences during primer design, I created a web page which uses JavaScript to generate common derivations of short genomic sequences. The tool can be seen online at `http://www.tallphil.co.uk/bioinformatics/sequences`

A.4.2 Genome RE Sites

I found myself frequently running a script I wrote to generate lists of restriction endonuclease recognition site co-ordinates. To make this tool generally available I adapted it to work through a web page with Human and Mouse genomes, plus all lis-

ted New England Biolabs restriction endonucleases. The tool can be seen at `http://www.tallphil.co.uk/bioinformatics/genome_re_sites`

A.4.3 Cytobands

To batch convert cytogenetic bands (*eg*. 9q34) into chromosomal co-ordinates (*eg*. chr:130,300,000-141,213,431) I wrote a script that uses data from the UCSC table browser. The script automatically recognises the format of the input and returns both cytogenetic band and chromosomal co-ordinates. The tool can be seen at `http://www.tallphil.co.uk/bioinformatics/cytobands`

A.4.4 FastQC

FastQC is a tool written by Dr Simon Andrews of the Babraham bioinformatics department. It analyses fastQ sequence files and produces a report describing a number of metrics that can be used to assess sequence quality. The reports are given as a HTML web page, divided into sections. I re-wrote the CSS (cascading style sheet) styles for the report to produce a static navigation bar on the side of the page which aided use of the report. The styles respond to the size of the device being used and are moved for small screens. FastQC can be found at `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

A.5 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

Bibliography

Adam, M., F. Robert, M. Larochelle, 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.

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.

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.

Cheung, M.-S., T. a. Down, I. Latorre, and J. Ahringer: 2011, 'Systematic bias in high-throughput sequencing data and its correction by BEADS.'. *Nucleic acids research* **39**(15), e103.

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. Sidorczuk-Lesthuruge, H. A. Rance, K. Padget, G. H. Jackson, N. Adachi, and C. a. Austin: 2012, 'Model for MLL translocations in therapy-related leukemia involving topoisomerase IIβ-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.

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.: 2006, 'The three 'C' s of chromosome conformation capture: controls, controls, controls.'. *Nature methods* **3**(1), 17–21.

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.

Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble: 2010, 'A three-dimensional model of the yeast genome.'. *Nature* **465**(7296), 363–7.

Duhig, T., C. Ruhrberg, O. Mor, and M. Fried: 1998, 'The human Surfeit locus.'. *Genomics* **52**(1), 72–8.

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. Shoresh, 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.

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.

Flicek, 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. Vandrovcova, 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.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan: 2009, 'An oestrogen-receptor-alpha-bound human chromatin interactome'. *Nature* **462**(7269), 58–64.

Furness, S. G. B. and K. McNagny: 2006, 'Beyond mere markers: functions for CD34 family of sialomucins in hematopoiesis.'. *Immunologic research* **34**(1), 13–32.

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.

Giallongo, a., J. Yon, and M. Fried: 1989, 'Ribosomal protein L7a is encoded by a gene (Surf-3) within the tightly clustered mouse surfeit locus.'. *Molecular and cellular biology* **9**(1), 224–31.

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.

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.

Gué, M., J.-S. Sun, and T. Boudier: 2006, 'Simultaneous localization of MLL, AF4 and ENL genes in interphase nuclei by 3D-FISH: MLL translocation revisited.'. *BMC cancer* **6**, 20.

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 pathwaysrepairing 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.

Hagège, H., P. Klous, C. Braem, E. Splinter, J. Dekker, G. Cathala, W. de Laat, and T. Forné: 2007, 'Quantitative analysis of chromosome conformation capture assays (3C-qPCR).'. *Nature protocols* **2**(7), 1722–33.

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.

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.

Huxley, C. and M. Fried: 1990, 'The mouse surfeit locus contains a cluster of six genes associated with four CpG-rich islands in 32 kilobases of genomic DNA.'. *Molecular and cellular biology* **10**(2), 605–14.

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.

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.

Kim, T.-K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. a. Harmin, M. Laptewicz, 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.

Kozubek, S., E. Lukásová, L. Rýznar, M. Kozubek, A. Lisková, R. D. Govorun, E. a. Krasavin, and G. Horneck: 1997, 'Distribution of ABL and BCR genes in cell nuclei of normal and irradiated lymphocytes.'. *Blood* **89**(12), 4537–45.

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, 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. Clee, N. Carter, a. Coulson, R. Deadman, P. Deloukas, a. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, 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. Shownkeen, 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. Wendl, 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. Szustakowki: 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.

Lennard, A., K. Gaston, and M. Fried: 1994, 'The Surf-1 and Surf-2 genes and their essential bidirectional promoter elements are conserved between mouse and human.'. *DNA and cell biology* **13**(11), 1117–26.

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.

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.

Lukásová, E., S. Kozubek, M. Kozubek, J. Kjeronská, L. Rýznar, J. Horáková, E. Krahulcová, and G. Horneck: 1997, 'Localisation and distance between ABL and BCR genes in interphase nuclei of bone marrow cells of control donors and patients with chronic myeloid leukaemia.'. *Human genetics* **100**(5-6), 525–35.

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. Kempski, 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.

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.

Murmann, A. E., J. Gao, M. Encinosa, M. Gautier, M. E. Peter, R. Eils, P. Lichter, and J. D. Rowley: 2005, 'Local gene density predicts the spatial position of genetic loci in the interphase nucleus.'. *Experimental cell research* **311**(1), 14–26.

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.

Nicoloso, M., L. H. Qu, B. Michot, and J. P. Bachellerie: 1996, 'Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs.'. *Journal of molecular biology* **260**(2), 178–95.

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 Igh.'. *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.

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, 'Uber Zelltheilung'. *Morphologisches Jahrbuch* **10**, 214–330.

Ragoczy, T., M. a. Bender, A. Telling, R. Byron, and M. Groudine: 2006, 'The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation.'. *Genes & development* **20**(11), 1447–57.

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.

Rozen, S. and H. Skaletsky: 2000, 'Primer3 on the WWW for general users and for biologist programmers'. *Methods in Molecular Biology* **132**, 365–386.

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.

Schwarz-Finsterle, J., S. Stein, C. Grossmann, E. Schmitt, H. Schneider, L. Trakhtenbrot, G. Rechavi, N. Amariglio, C. Cremer, and M. Hausmann: 2005, 'COMBO-FISH for focussed fluorescence labelling of gene domains: 3D-analysis of the genome architecture of abl and bcr in human blood cells.'. *Cell biology international* **29**(12), 1038–46.

Sexton, T., E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli: 2012, 'Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome'. *Cell* pp. 1–15.

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.

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.

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.

van de Werken, H. J. G., P. J. P. de Vree, E. Splinter, S. J. B. Holwerda, P. Klous, E. de Wit, and W. de Laat: 2012, *4C Technology: Protocols and Data Analysis.*, Vol. 513. Elsevier Inc., 1 edition.

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.

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.

Yokota, H., G. van den Engh, J. E. Hearst, R. K. Sachs, and B. J. Trask: 1995, 'Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus.'. *The Journal of cell biology* **130**(6), 1239–49.

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.

Zhang, Y., R. McCord, Y.-J. Ho, B. Lajoie, D. Hildebrand, A. Simon, M. Becker, F. Alt, and J. Dekker: 2012, 'Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations'. *Cell* pp. 1–14.

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.

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.