Chapter 1

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.

1.1 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.

1.1.1 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.

1.2 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 **??**. 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.

1.3 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 **??**). Such observations give greater weight to the need to understand the organisation of the human genome in health and disease. 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.

My hypothesis at the start of this PhD was that chromosomal translocation partner genes specifically associate at transcription factories. This was supported by evidence showing the recruitment of *Myc* to pre-existing transcription factories containing transcribing *Igh* alleles (Osborne et al., 2007). There has been some dispute to this finding since; Hakim *et al.* used 4C-seq to characterise the associations of *Myc* and *Igh* in cultured B cells with and without AID, a protein involved in somatic hypermutation (Hakim et al., 2012). They found the frequency of translocations correlated with genomic association frequencies in the absence of AID, but that AID translocation hotspots did not. Rocha *et al.* published a study using similar methodologies that came to the the opposite conclusion; that off target AID activity does correlate with genomic association (Rocha et al., 2012). However, Rocha *et al.* did not find *Myc* to be the most frequent association partner with *Igh*. It is of note that both of these studies investigated cultured B cells undergoing steady state transcription, whereas Osborne *et al.* used *ex-vivo* cells stimulated for five minutes (Osborne et al., 2007). *Myc* is an immediate-early gene, meaning that it undergoes rapid activation and transcription after activation. As such, the cells analysed by Osborne *et al.* will have been undergoing a synchronised wave of transcriptional activity and may exhibit different patterns of

nuclear organisation to those studied by Hakim and Rocha.

In my studies I investigated whether association between translocation partner genes may be a general phenomenon. If so, e4C analysis would reveal association between proto oncogenes in healthy cells. *BCR* and *MLL* have very different portfolios of translocation partners, the former only found partnered with a handful of genes and the latter more promiscuous with over 100 partners found to date (discussed in Section **??**). I expected the e4C *trans* association profiles to show association with all of these genes, and for their association to correlate with translocation frequency. In general, the profiles that we discovered were not so simple. *BCR* did show strong association with *ABL1*, but not in order of magnitude that we expected. Instead of the specific gene - gene associations, *BCR* was found to associate with the entire telomeric arm of chromosome 9. These results suggest that while association may affect the frequency of chromosomal translocations, it is unlikely the be the sole cause for partner selection. Certainly, our view of nuclear organisation is becoming increasingly complex.

An interesting observation described in this thesis is the presence of strong *BCR* association in the vicinity of the *ABL1* gene. Such a "bystander" effect has a number of implications. For example, bystander genes may account for risk factors seen in genome wide association studies (GWAS) which are not yet understood. To modify deleterious gene associations, the behaviour of adjacent loci causing the contact may need to be changed rather than the translocation partner genes themselves.

With the increasing capability of molecular biology techniques able to probe nuclear structure, we will soon be able to investigate how the activity of other genes can affect proto oncogenes on a more general scale.

Bibliography

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.

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.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner: 2002, 'Capturing chromosome conformation'. *Science* **295**(5558), 1306–1311.

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.

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.

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.

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.

Hakim, O., W. Resch, A. Yamane, I. Klein, K.-R. Kieffer-Kwon, M. Jankovic, T. Oliveira, A. Bothmer, T. C. Voss, C. Ansarah-Sobrinho, E. Mathe, G. Liang, J. Cobell, H. Nakahashi, D. F. Robbiani, A. Nussenzweig, G. L. Hager, M. C. Nussenzweig, and R. Casellas: 2012, 'DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes'. *Nature*.

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.

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.

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.

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.

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.

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.

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.

Rocha, P., M. Micsinai, J. Kim, S. Hewitt, P. Souza, T. Trimarchi, F. Strino, F. Parisi, Y. Kluger, and J. Skok: 2012, 'Close Proximity to Igh Is a Contributing Factor to AID-Mediated Translocations'. *Molecular Cell* **47**(6), 873–885.

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.

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.

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.

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.

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.

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.