Chapter 1

Discussion

The role of three-dimensional organisation in the functional regulation of mammalian genomes is increasingly being recognised as being critical. In this thesis I describe how I have extended and further developed e4C, a technique based on chromosome conformation capture, to investigate the genome-wide association profiles of three proto-oncogenes in human CD34$^+$ cells and the GM12878 cell line.

1.1 The e4C methodology

The past ten years has seen a rapid adoption of 3C and 3C-based techniques. The slew of protocols to investigate the make up of 3C libraries has allowed us to investigate differences in association between different loci in an increasingly high throughput manner. The e4C sequencing technique discussed in this thesis represents another step towards the aim of understanding the contacts made by genes and how they may affect us in health and disease.

The data presented here is not without its flaws. The clearest example of this are the CD34$^+$ MLL e4C libraries. These libraries exhibit approximately ten-fold less coverage than the BCR bait e4C libraries for reasons currently unknown. It is of note that a recent paper by 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 transcriptional 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%) 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 lower complexity of the GM12878 ABL e4C library relative to the CD34$^+$ BCR or the higher complexity of the GM12878 MLL e4C libraries relative to the CD34$^+$ e4C library. Given extra time and funds it would be interesting to replicate my e4C libraries for all baits in both cell types and create more e4C libraries using different baits to elucidate a pattern.

In chapter three 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. Whilst this experiment resulted in captured sequences that suggested problems in the library preparation, I believe that the concepts behind the attempt are sound. With the benefit of hindsight I think that multiplexed e4C libraries prepared in parallel have the potential to greatly improve the standard of data produced by e4C. It seems likely that the non-homology in the random barcode sequence caused the problems in library preparation; using *NlaIII* adapters prepared using homologous oligos with a single barcode sequence should not cause the same problems and can still be used to detect identical ligation products found in more than one multiplexed library. In my hands up to eight e4C libraries could be prepared in parallel, so it would not be infeasible to multiplex many tens of libraries with the same bait on a single flow cell, greatly improving the coverage and resolution of the resulting data.

The technology behind many molecular biology techniques has undergone a radical transformation in the past twenty years, especially through the development of next generation sequencing techniques in a trend that seems set to continue. As the quantity of data produced by our experiments exceeds immediate analysis through observation the methods we use to process and digest that data become more important. The development of e4C sequencing described in this thesis acts as a paradigm of this story: 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; 4C scaled this technique to use microarray technology to study seven mouse chromosomes (Simonis et al., 2006) and later a representative selection of all mouse chromosomes (Schoenfelder et al., 2010); this thesis describes the use of e4C sequencing to produce many millions of sequencing reads across the genome. With greater quantities of data comes a need for more careful analysis, with a large enough dataset one can choose tests to demonstrate any hypothesis. To address this, 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 3C derived association data that have been published independently during the time that the analysis concepts in this thesis were developed. 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 (low library complexity in sequencing, restriction endonuclease site distribution) and suggest some solutions that differ from those proposed in this thesis (spiking in phiX library instead of Bareback processing, using running probes created over multiple restriction fragments instead of calculating the proportion of fragments observed). In this way, 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 based techniques. The trend so far has generally been one of individual loci investigation to global investigation with recent 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 one 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 for every fragment in the genome once, over half a billion sequence reads are needed ($752000^2 = 5.7\times10^{11}$). 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 dataset a HiC library would require over seven quadrillion reads (seven thousand billion, $5.7\times10^{11}\times 12500 = 7.125\times10^{15}$) requiring many thousands of runs with today's sequencing technology (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 levels of deep sequencing with the human genome, there is a role for e4C and other techniques able to interrogate a specific subsection of 3C libraries. As described in this thesis, e4C enriched 3C libraries for a specific bait, revealing an in-depth genome wide map of association for that sequence. 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 allow investigation of a myriad of micro-environments of the nucleus.

I believe that as 3C based techniques continue to gain interest and sequencing power, we will see a flourishing of studies investigating comparative and dynamic nuclear organisation. Comparative studies may yield new understanding of the differences in genomic organisation between different tissues, healthy and disease states, the evolution of genome structure and the heterogeneity of structure between single cells and populations. Studying the dynamics of genome organisations can teach us new insights into cell cycle progression, tissue differentiation, the processes driving organisation and the effect of pharmacological agents. As our understanding of the gross rules governing nuclear organisation increases we will be able to better understand how differences in organisation can affect biological function and how biological function can affect organisation. In the future I think that diagnostic tests for nuclear organisations may reach the clinic, and drugs able to modify organisation may become common place, 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 could be split into two genomic compartments using principal component analysis. The authors characterised the components as active and inactive compartments, the active showing looser compaction and a correlation with gene density, mRNA expression, DNAse sensitivity and active histone marks (Lieberman-Aiden et al., 2009). This observation of active and inactive compartments in the human and mouse genomes 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 and models of chromatin loops described in detail in Section **??**. Clustering of active regions gives 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 facttory. Such a model has a number implications for our understanding of nuclear function, partly in appreciation of the complexity of the nucleus but also through explaining the role of many nuclear factors in transcription. 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 well as the association of *MLL* with its translocation partner genes (Murmann et al., 2005; Gué et al., 2006; Cowell et al., 2012). Where these studies differ from the work described in this thesis is that they all use DNA- or RNA-FISH to measure association. As such, whilst they are able to show significantly enriched association in comparison to candidate negative control loci, they cannot describe this 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 suggests that chromosomal associations may play an important role in the formation of chromosomal translocations. 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 of localisation of the *MLL* gene to prevent these translocations. Whilst clinical applications such as this still are still some way from reaching the clinic, further development of our understanding of the biological processes governing disease initiating events is critical.

Bibliography

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.

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.

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.