7 Discussion

7.1 Methods of rearrangement breakpoint mapping

7.1.1 Breakpoint mapping using microarray based techniques

Apparently balanced translocations in 3 phenotypically abnormal patients were mapped using a variety of microarray based strategies prior to sequencing. The breakpoints were mapped by array painting onto a 1Mb microarray followed by FISH analysis of intermediate BAC clones prior to this study. The breakpoints were mapped further by array painting onto custom-made fosmid microarrays followed by STS PCR mapping prior to LR PCR for the amplification of junction fragments and subsequent sequencing. In the case of patient t(2;7)(q37.3;p15.1), STS PCR failed to refine the chromosome 7 translocation breakpoint further than array painting onto the custom-made fosmid microarray so alternative strategies of array painting onto custom-made PCR product microarrays and custom-made oligonucleotide microarrays with increased resolution were adopted.

The advancement of genomic microarray technology during the course of this project has increased the resolution from approximately 1Mb to under 50bp, increasing the speed and accuracy with which rearrangement breakpoints can be mapped. Our preferred strategy of translocation breakpoint mapping is currently array painting onto a Whole Genome Tile Path microarray followed by array painting onto a custom-made oligonucleotide microarray, thus enabling design of primers for amplification of the rearrangement junction fragments by PCR after just two hybridisations. In the future, the increasing resolution of microarrays may allow this to be reduced to a single hybridisation on one microarray. Patient t(2;7)(q37.3;p15.1) was used to directly compare these methodologies (Figure 7.1).

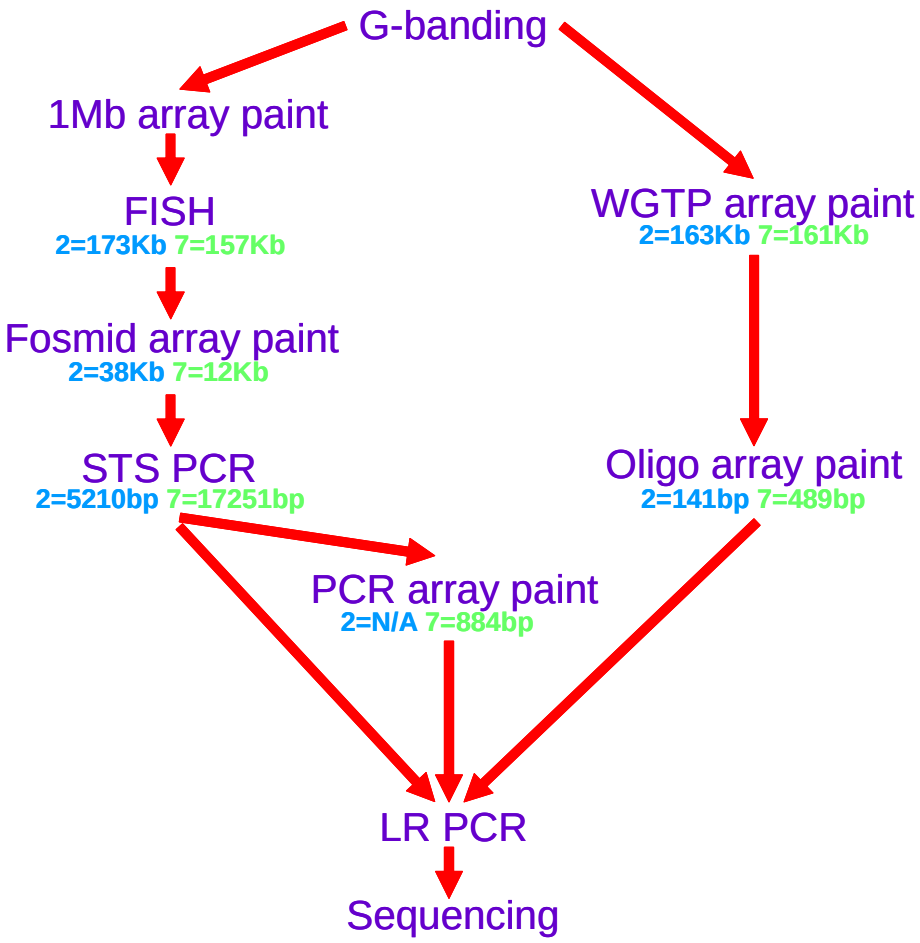

Figure 7.1 Direct comparison of strategies used for translocation breakpoint mapping using patient t(2;7;)(q37.3;p15.1) as a test case. The mapped intervals for the chromosome 2 breakpoint are shown in blue and the chromosome 7 breakpoint in green as defined by each technique.

The quantitative nature of array painting was demonstrated by further analysis of the array painting data generated by the custom-made fosmid microarrays. Interpolation of the data predicted that the chromosome 2 breakpoint fell at 236,556,071bp and chromosome 7 breakpoint fell at 30,984,039bp. This prediction was 7,492bp away from the actual breakpoint for chromosome 2 and 436bp away from the chromosome 7 breakpoint. This data emphasised the need for the generation of microarrays with a high redundancy of clones rather than

minimal tiling path coverage to enable the prediction of breakpoint positions from the ratios obtained from overlapping clones.

Full analysis of the chromosome 7 breakpoint region in patient t(2;7)(q37.3;p15.1) showed the complicated nature of the genome (Figure 7.2).

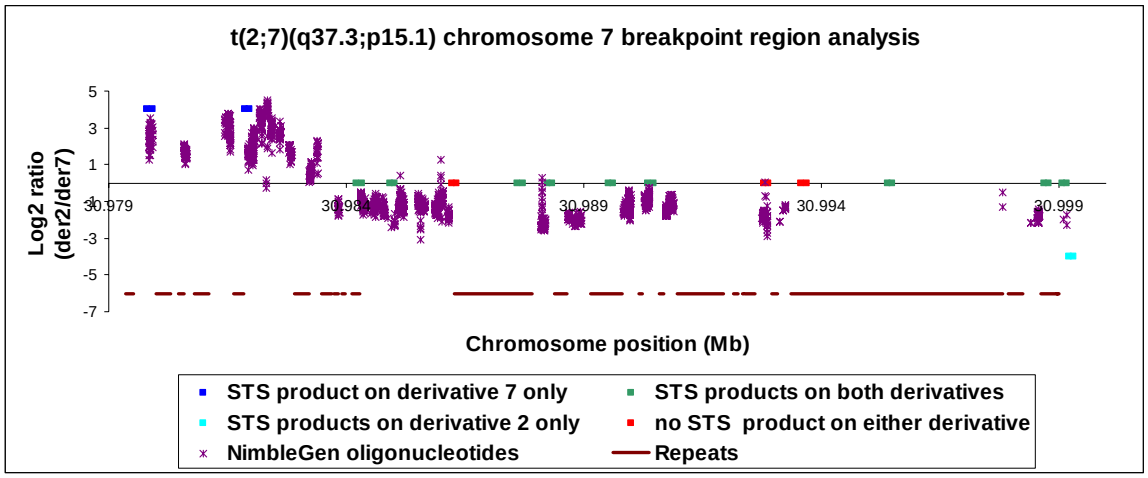

Figure 7.2 Analysis of STS PCR and oligonucleotide microarray array painting results for patient t(2;7)(q37.3;p15.1) with repeat content of surrounding region.

Only oligonucleotides considered to be unique within the human genome were selected for generation of the microarray resulting in incomplete coverage across the breakpoint region. As expected, closer inspection of the sequence revealed that gaps in the oligonucleotide coverage corresponded to known repeat elements annotated in the USCS web browser. STS PCR analysis across this region produced conflicting results due to the repetitive nature of the sequence. This complexity may have impeded whole genome amplification by DOP PCR performed prior to STS PCR mapping, resulting in incomplete amplification of the breakpoint region.

The techniques for translocation breakpoint mapping were also used to successfully map and sequence across a duplication junction in patient t(2;7)(q37.3;p15.1) with an additional duplication of 3p26.3. LR PCR primers were designed to amplify across a junction resulting from a simple tandem duplication. If these primers had failed to amplify a fragment, further investigation by FISH onto extended chromatin fibres using clones from the breakpoint regions as probes would have been required to resolve the orientation of the duplication prior to PCR amplification.

7.1.2 Translocation breakpoint mapping using custom-made fosmid libraries

The mapping of translocation breakpoints by array painting and PCR relies on the assumption that the rearrangement is simple. Translocations which are accompanied by cryptic imbalances or inversions around the breakpoints will be harder to map by these methods and in these instances, alternative strategies of breakpoint mapping would be required. For this reason we investigated the generation and screening of a custom-made library for breakpoint mapping as summarised in Figure 7.3. Custom-made libraries have previously been used to identify aberrations in a breast cancer cell-line (Volik et al. 2003). A BAC library created with 0.37 fold coverage was end sequenced and the data mapped back onto the human genome reference sequence to identify rearrangements within the cell line.

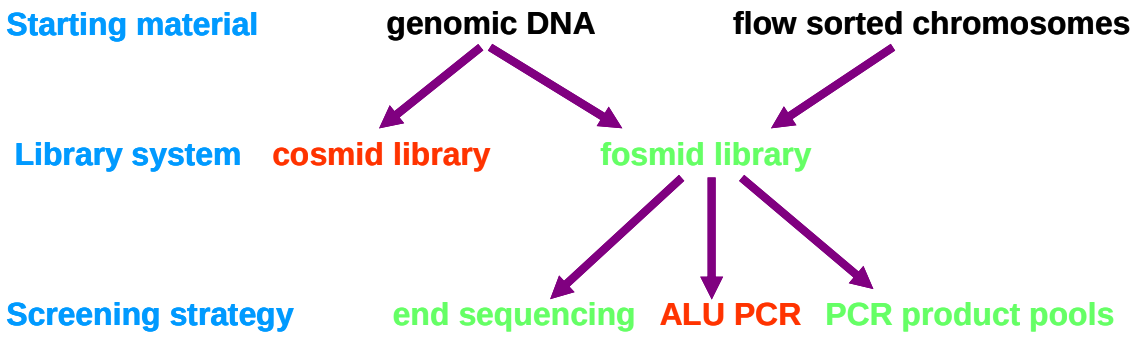

Figure 7.3 Schematic summarising the approaches used to generate and screen a custom-made library. Techniques in green were successful.

Using genomic DNA extracted from a patient derived cell-line, 2 protocols for the production of a custom-made library were investigated concurrently; a cosmid library and a fosmid library. The fosmid library proved to be a successful method, but the cosmid library failed to produce any colonies. The cosmid library technique was not investigated further but the lack of cosmid clones may have been due to technical issues such as loss of template material and reagents that were not optimal. Subsequently, flow sorted derivative chromosomes were used as starting material for the fosmid library in order to increase the purity and coverage of the library created.

A fosmid library was created from flow sorted derivative chromosomes from patient t(7;13)(q31.3;q21.3) to test the feasibility of this approach. All clones from the library were end sequenced and screened using radiolabelled Alu PCR products and a pool of radiolabelled PCR products. Alu PCR product screening did not successfully identify any translocation spanning clones in the test libraries. This may have been due to the lack of amplification of human derived DNA using the primers specified. Whilst bands in the range of 100 to 1000bp were detected by gel electrophoresis after temperature cycling, *in silico* PCR analysis using the human genome reference sequence failed to predict any amplification under 3Kb (anything larger than 2Kb would be outside the limits of the PCR cycling conditions). PCR product pool screening of the fosmid library identified spanning clones in both libraries. End sequence screening identified a spanning clone in the derivative chromosome 7 library and mapped the breakpoint to a 130Kb region between clones in the derivative chromosome 13 library. In addition, the end sequence data obtained provided further information about the coverage of the library along the derivative chromosomes. The 2 successful strategies of screening by end sequencing and by radiolabelled PCR product pools have advantages and disadvantages which are detailed in Table 7.1.

	Advantage	Disadvantage
End sequencing	Automated process	Cost
	Additional data generated	
PCR product pool	Quick results (4 days)	Radiation hazards

Table 7.1 Summary of major advantages to screening by end sequencing or by radiolabelled PCR product pools.

Analysis of the full sequence data generated from the chimeric clones showed the breakpoint positions were consistent with those mapped by array painting and PCR.

After the assessment of production and screening methods using patient t(7;13)(q31.3;q21.3) as a test case, a library was generated from a patient with a previously uncharacterised t(2;6)(q21.1;q25.1) translocation. The translocation breakpoints were initially mapped by array painting flow sorted derivative chromosomes onto a whole genome tile path microarray to identify spanning BAC clones. The breakpoint spanning fosmid clones from the custom-made library were successfully identified using radiolabelled PCR product pools generated from the spanning BAC clone information. The fosmid clones were isolated and sequenced allowing the identification of the precise breakpoint positions and confirming the simple nature of the rearrangement. It was noted that 7bp were deleted from chromosome 2 and 3bp from chromosome 6. Analysis of the extra sequence data (approximately 40Kb for each clone) shared 100% homology with the human genome reference sequence.

7.1.3 Application of techniques developed for the mapping of rearrangement breakpoints to the investigation of viral integration in the human genome

One of the aims of this thesis was to demonstrate that the techniques developed for the mapping of translocation breakpoints could also be applied to other rearrangements such as viral integration. A custom-made fosmid library was generated using flow sorted chromosomes from a patient carrying integrated

HHV-6 DNA. Whilst the generation and screening of the library has not, so far, identified clones chimeric for human and viral DNA and therefore spanning the breakpoint, further screening with alternative probes may do so. For example, using a human telomere probe in conjunction with probes generated towards the ends of the viral genome should identify chimeric clones as it is possible that the telomeric like repeat present in the virus enables the virus to integrate into the human telomeres by homologous recombination. A library with a greater depth of coverage along the chromosome and a pool of viral probes with more comprehensive tiling along the HHV-6 genome could also be investigated.

In addition, conventional techniques could also be applied to resolve the integration sites; Southern blots of the chromosome containing the integrated virus could be screened using human telomere and viral probes to isolate a chimeric fragment. Once a clone or fragment has been identified, full sequencing would generate data across the integration site. However, if as suspected, the integration occurs in the human telomere ($(TTAGGG)_n$) due to the telomeric like repeat present in the virus ($(GGGTTA)_n$) then it may be impossible to determine the exact integration site as long stretches of these repeats appear identical.

7.2 Bioinformatic analysis of genomic rearrangement breakpoints

Chromosome rearrangements have been associated with abnormal phenotypes; either by direct disruption of gene structures, modification of gene regulation or alteration of the copy number. Bioinformatic analysis of the regions surrounding the rearrangements may identify genes that may be causal to a phenotype in addition to genomic structures or motifs which may help to elucidate a mechanism behind the rearrangement.

7.2.1 Genotype to phenotype correlations

7.2.1.1 Identification of non-pathological variation

Analysis of human genomes is being conducted with increasing resolution. With this increase in resolution, greater complexity is found within the genome, but it is important to determine whether the variation seen between individuals is normal or pathological.

Genome-wide studies have identified large regions of variability within the human genome (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005). A recent study estimated that approximately 12% of the human genome was subject to copy number variation (CNV) (Redon et al. 2006). These regions of variation were identified in studies using apparently clinically normal individuals and are therefore assumed to be non-pathogenic. However, it should be noted that patients were described as clinically normal at the time of assessment and this assessment does not allow for extremely mild pathogenic phenotypes, or pathology that may develop later in life. All regions of variation identified by these studies have been summarised in the Database for Genomic Variants (http://projects.tcag.ca/variation/). In addition to large structural variation, single nucleotide polymorphisms (SNPs) are believed to be responsible for phenotypic variation. An investigation into the significance of SNPs and CNVs using expression level analyses of approximately 15,000 transcripts revealed that SNPs were responsible for 83.6% and CNVs 17.7% of the genetic variation detected in gene expression studied with little overlap (1.3%) (Stranger et al. 2007).

CNVs have dosage implications for the genes that they affect. A dosage sensitive gene within a region that is duplicated may lead to overexpression and a dosage sensitive gene in a region that is deleted may result in haploinsufficiency for that gene. For example, Charcot Marie Tooth disease (CMT1A) and hereditary

neuropathy with liability to pressure palsies (HNPP) are believed to be the result of overexpression and haploinsufficiency of the *PMP22* gene in duplications and deletions of chromosome 17p11.2 respectively (Patel et al. 1992; Chance et al. 1993). In order to ascertain whether a CNV is pathogenic, it is important to determine whether the variation is *de novo* or inherited.

Whole genome screening of patient t(2;7)(q37.3;p15.1) on a 1Mb resolution microarray revealed a duplication at 3p26.3. Subsequent CGH analysis on a custom-made fosmid microarray, followed by a custom-made oligonucleotide microarray refined the breakpoints sufficiently to allow direct amplification of the junction fragment. Sequencing of the junction fragment located the distal breakpoint within a 260bp region from 1,756,993 to 1,757,252bp and the proximal breakpoint within a 258bp region from 3,614,132 to 3,614,389bp. Analysis of the parental DNAs revealed that the same duplication was present in the phenotypically normal father, but not in the mother. Interrogation of the CNV data available in Ensembl revealed a known 289Kb CNV locus from 2,181,272 to 2,470,417bp. It was therefore unlikely that the duplication was pathogenic as the impact of an increase in copy number of genes within the duplicated region did not appear to be significant.

7.2.1.2 Gene disruption

Investigation into annotated genes at the breakpoint regions of the 3 patients studied revealed that 3 of the 6 translocation breakpoints directly disrupted a gene, providing candidate genes for 2 of the 3 patient phenotypes. Verification of the ensuing altered protein could be obtained through Western blotting and confirmation that the translocation is directly responsible for the phenotype could be achieved using expression studies to determine whether the truncated proteins are expressed. Murine knockout models could also be generated for comparative phenotype analysis.

In addition to chromosome rearrangements, mutations may have an effect on a patient's phenotype. In a case where a gene is not dosage sensitive, a translocation may not have an effect unless the other allele is affected in some way, for example, a simple mutation. In these cases, further analysis of the genes in the breakpoint regions would be required, for example resequencing of the gene exons for mutation detection.

It has been noted that translocation breakpoints in phenotypically normal patients are generally not accompanied by additional imbalance in direct contrast to the breakpoints in phenotypically abnormal patients which are often seen to be accompanied by cryptic imbalances (Baptista et al. 2005).

The risk of an apparently balanced reciprocal translocation having an associated congenital abnormality was estimated at 6.1% (Warburton 1991). However, the risk of a complex balanced rearrangement having an associated phenotype has been estimated at 23% (14 out of 60 cases studied) (Madan et al. 1997), indicating that the risk of an associated congenital abnormality rises with increasing complexity of the rearrangement.

The genotype to phenotype correlation for patient t(3;11)(q21;q12) is particularly complex. The patient is one of monochorionic, monoamniotic twins who by G-banding analysis carry the same translocation. Patient t(3;11)(q21;q12) presented with a congenital duodenal obstruction, complex congenital heart disease and facial dysmorphism whilst the sibling was clinically normal. As the twins are identical, it is unlikely that the translocation is responsible for the phenotype. The phenotypic difference between the twins could be due to discordant imprinting. Cases have been recorded where genetically identical twins present with different phenotypes as observed in Beckwith-Wiedermann syndrome (BWS). A study of 5 identical twins with discordant phenotypes

identified differences in the imprinting of the *KCNQ1OT1* gene as responsible for the variations observed in clinical diagnoses (Weksberg et al. 2002).

An alternative explanation for the clinical differences seen between patient t(3;11)(q21;q12) and her twin is X inactivation. Skewed X inactivation has been shown to be the cause of discordant phenotypes in a pair of twins where one twin presented with Duchenne Muscular Dystrophy and the other was clinically normal (Richards et al. 1990).

7.2.2 Mechanisms underlying genomic rearrangements

The sequence across the translocation junctions in all 3 patients investigated in Chapter 3 was characteristic of a mechanism of non-homologous end joining as there was a lack of homologous sequence between both the breakpoint regions and small numbers of bases were deleted and/or duplicated and/or inserted in all 6 of the breakpoints sequenced. In addition, analysis of the sequence surrounding the breakpoints did not identify any repeat structures or sequence motifs that were common to all the breakpoints. In patient t(2;7)(q37.3;p15.1) where a 19bp stretch of bases was inserted at the breakpoint, analysis of the sequence found no homology to the human genome reference sequence. A 41bp insertion of mitochondrial DNA had previously been identified at the derivative 9 junction in a patient with a t(9;11)(p24;q23) (Willett-Brozick et al. 2001) however, alignment of the 19bp insertion sequence in patient t(2;7)(q37.3;p15.1) to human mitochondrial DNA (Accession number NC_001807) failed to find any homology.

The mechanism behind the duplication seen in patient t(2;7)(q37.3;p15.1) was, in direct contrast, believed to be homologous recombination due to the Alu repeats found at both breakpoints. In fact, the precise duplication breakpoints could not be determined owing to the 97% homology between these repeats.

7.2.2.1 Predisposition

It is believed that certain genomic architecture can lead to a predisposition for disease. For example, in Williams Beuren Syndrome, 90% of patients have a 1.5Mb deletion at 7q11.23 where the breakpoints fall within low copy repeats (Nickerson et al. 1995; Bayes et al. 2003). Further analysis of the heredity of the disease revealed that 5% of the general population carry an inversion of the region between these repeat segments and in the next generation 33% of the offspring carry the deletion associated with Williams Beuren syndrome, suggesting that the inversion predisposes for the deletion (Osborne et al. 2001). This phenomenon is also believed to be true for Sotos syndrome; In a study of Sotos syndrome within the Japanese population, it was found that 14/14 fathers and 8/9 mothers of children carrying the associated 1.9Mb microdeletion carried a heterozygous inversion of the interval between the LCRs responsible in the paternally and maternally derived chromosomes respectively (Visser et al. 2005). More recently, a 900Kb inversion has been seen to predispose for a 500-650Kb deletion encompassing the *MAPT* locus at 17q21.31 in 3 patients with mental retardation and dysmorphic features (Shaw-Smith et al. 2006).

7.2.2.2 Occurrence of rearrangements

There are approximately 30-50,000 recombination hotspots representing 3% of the human genome equating to 1 every 50-100Kb. At least 80% of all recombination events occur in these short regions. A positive relationship between GC content and recombination has been noted with recombination rates found to be lower within genes (McVean et al. 2004).

Within the nucleus, chromosomes exist in discrete compartments known as chromosome territories where the most gene rich chromosomes (chromosomes 17,19 and 22) congregate at the centre of the nucleus whereas the more gene poor chromosomes gravitate towards the nuclear periphery (Boyle et al. 2001). A study of more than 10,000 constitutional translocations revealed that the larger

chromosomes are generally more involved in translocations, perhaps due to the length of DNA available for rearrangement (Bickmore and Teague 2002). It was also noted that the most gene dense chromosomes are less likely to be involved in translocations.

These studies may suggest that the human genome protects itself against rearrangements, however it must be considered that most studies have been conducted using patients rather than aborted conceptions so that the study group might be biased and that the majority of rearrangements that disrupt genes are lethal.

7.3 Recent developments and the future of breakpoint mapping

An alternative to the use of flow sorted chromosomes for array painting is the use of microdissected chromosomes for hybridisation to the microarrays. One study of 5 patients with apparently balanced reciprocal translocations used 4-6 microdissected derivative chromosomes for DOP PCR amplification prior to hybridisation to a 1 Mb resolution microarray (Backx et al. 2007). This procedure successfully mapped all 10 translocation breakpoints to 1Mb intervals.

International public efforts to sequence the human genome cost approximately $3 billion but the race is now on to sequence an individual's genome for $1000 (Service 2006). New sequencing technologies are being developed and the cost of sequencing is decreasing. One such development is the generation of sequence using 454 pyrosequencers (www.454.com). These machines fragment the genome into 300-500bp segments which are attached to beads, denatured and amplified in the presence of luciferase which fluoresces every time a nucleotide is added to the chain. By analysing the light flashes and the corresponding nucleotide present on each bead, the systematic growth and subsequent sequence can be tracked electronically. This 454 technology has been used to conduct a study of structural variation within the genomes of 2

individuals (Korbel et al. In press). The first was the female used to create a fosmid library to verify the human genome reference sequence as part of the Human Genome Project (IHGSC 2004) and more recently used to originally assess copy number variation by end sequencing (Tuzun et al. 2005). The second individual was a member of the group used to detect CNVs by array CGH (Redon et al. 2006). Genomic DNA from both individuals was fragmented to 3Kb and the paired ends of the fragments (equating to 2.1x and 4.3x coverage of the genomes) were sequenced and aligned to a reference genome to assess the level of structural variation. This technology enabled a 3Kb resolution scan of the genome to be performed with an average resolution of breakpoint assignment of 644bp (Korbel et al. 2007).

A more recent advancement of this method is to array the fragmented DNA directly onto microarray slides where they can be amplified *in situ* (www.illumina.com). Addition of all 4 fluorescently labelled nucleotides allow each cycle of the sequencing reaction to occur simultaneously, and the fluorescence of each incorporated base to be detected. This technology is capable of sequencing across long stretches of homopolymers.

Analysis of the read depths obtained using these sequencing technologies will provide data on any copy number variation within a genome. For example, a region showing amplification within a genome will exhibit a higher read depth than at a region with a normal copy number. These technologies could potentially be used to sequence derivative chromosomes in patients carrying rearrangements; direct comparison of the data generated with the human genome reference sequence would identify any differences between the 2 genomes including translocations, inversion and copy number differences. These strategies of breakpoint mapping would be of particular use in cases with complex rearrangements.

7.4 Conclusions

The advancement of chromosome analysis has improved the speed and resolution with which chromosome rearrangements can be studied. By increasing the efficiencies of methodologies for the mapping and subsequent sequencing of rearrangement breakpoints the number of published sequenced constitutional translocation breakpoints will be expanded, so that a comprehensive study of the breakpoint regions may help to elucidate the mechanism causing the rearrangement. The data so far suggests that the mechanism by which these non-recurrent rearrangements occurs is non-homologous end joining, in direct contrast to recurrent rearrangements which are mediated by homologous recombination.

This thesis has developed strategies for the rapid analysis of rearrangement breakpoints and viral integration sites in the human genome. New, higher resolution microarrays have been applied to the technique of translocation breakpoint mapping and the generation and screening of custom-made fosmid libraries have been successfully used to identify the breakpoints in a balanced reciprocal translocation. The technique of breakpoint mapping using custom-made libraries enables the mapping of more complex rearrangements where analysis by microarrays and PCR might fail. This thesis also demonstrates that these techniques are applicable to the investigation of other genomic rearrangement studies such as viral integration into the human genome.