1 Introduction

1.1 The discovery of human chromosomes

Human chromosomes were first observed by Flemming and Arnold in the 1880's with the correct number of chromosomes in a human cell finally identified as 46 in 1956 (Ford and Hamerton 1956; Tijo and Levan 1956). Human chromosomes are morphologically distinct, and were easily classified into 7 groups (a to g) according to their size and relative positions of the centromeric constriction (Patau 1960). In 1968 Torbjorn Caspersson developed a method of staining human metaphase chromosomes reproducibly to give a distinctive pattern of dark and light bands along their lengths (Caspersson et al. 1968; Caspersson et al. 1972). This distinctive banding pattern along with the relative sizing and centromeric positions are still used in general karyotyping analysis and the identification of chromosome rearrangements today.

1.2 Chromosome rearrangements

The development of chromosome banding allowed much more detailed analyses of chromosomes such that structural rearrangements were rapidly identified. Structural chromosome abnormalities including translocations, deletions, amplifications and inversions arise from recombination between chromosomes or the misrepair of chromosome breaks. These chromosome rearrangements are often identified by cytogenetic techniques, initially by studying the patient's karyotype using chromosome size and banding patterns, and more recently using microarrays. Chromosome rearrangements have been associated with abnormal phenotypes; either by directly disrupting gene structures, modifying gene regulation or alteration of the copy number.

1.2.1 Chromosome translocations

One of the first translocations identified was the pathogenic somatic translocation implicated in Chronic Myeloid Leukaemia (Rowley 1973). The t(9;22)(q34;q11)

reciprocal translocation produces the characteristic Philadelphia marker chromosome, which disrupts chromosome 9 within the Abelson murine leukaemia viral oncogene homologue 1 (*ABL*) gene and chromosome 22 within the breakpoint cluster region (*BCR*) gene. The resulting gene fusion product, BCR-ABL, inhibits DNA repair and accelerates cell division.

Constitutional translocations arise very early in development; either as the result of an abnormal gamete or possibly after abnormal fertilisation or early embryo formation and affect every cell within a body. Two classes of constitutional translocations exist: balanced (with no accompanying net gain or loss of material) and unbalanced. Robertsonian translocations are a specific class of unbalanced translocation resulting in the formation of a dicentric derivative chromosome after the joining of 2 acrocentric chromosomes (parts of the acrocentric arms are lost). Both balanced reciprocal translocations and Robertsonian translocations are stable through mitosis (Figure 1.1).

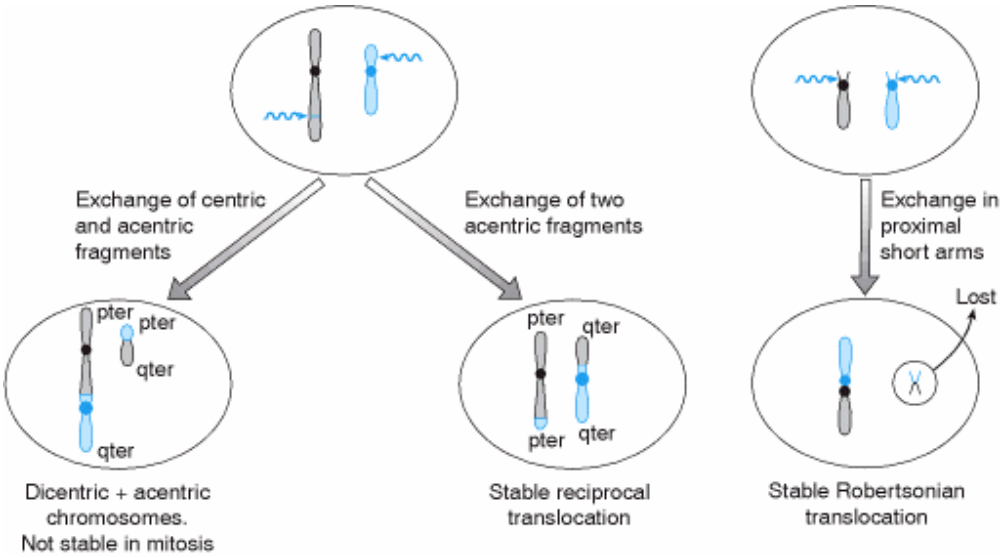

Figure 1.1 The origins of translocations (Strachan and Read 1999). Robertsonian translocations are a result of exchanges between the proximal short arms of the acrocentric chromosomes 13, 14, 15, 21 and 22. Both centromeres are present but function as one, so the derivative chromosome is stable.

The most common recurrent non-Robertsonian constitutional translocation is t(11;22)(q23;q11). It is associated with an increased frequency of spontaneous abortion and a 10 fold increase in risk of breast cancer within carriers. The 11q23 breakpoints fall within an AT-rich repeat (Edelmann et al. 1999) and the 22q11 breakpoints fall within a Low Copy Repeat (LCR22) (Funke et al. 1999) suggesting that the mechanism by which the translocation occurs is the occurrence of double strand breaks within palindromic sequences leading to illegitimate recombination events (Edelmann et al. 2001).

1.2.1.1 Frequency of translocations

It is estimated that 1 in 200 live births have an apparently balanced translocation and 1 in 500 have an unbalanced translocation (Jacobs et al. 1992). The frequency of *de novo* reciprocal apparently balanced translocations is approximately 1 in 2000 (Warburton 1991). Many carriers of a balanced translocation appear phenotypically normal and balanced translocations often go unnoticed until segregation of an unbalanced form of the translocation results in recurrent miscarriages or offspring with congenital abnormalities. Only 25% of gametes produced from a cell with a reciprocal translocation will contain copies of normal chromosomes, a further 25% will carry both derivative chromosomes and the other 50% will carry a partial monosomy for one chromosome and a partial trisomy for the other chromosome which may be lethal (Figure 1.2).

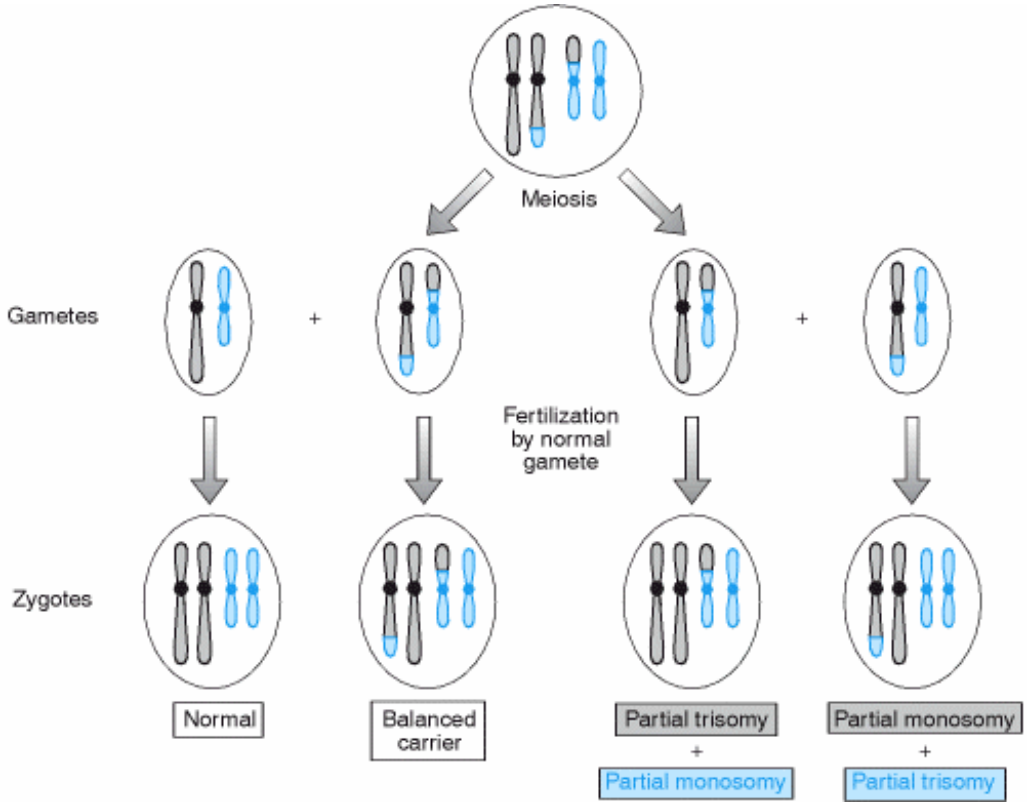

Figure 1.2 Results of meiosis in a carrier of a balanced translocation (Strachan and Read 1999).

The mapping and sequencing of chromosome breakpoints will improve our understanding of the mechanisms underlying chromosomal rearrangements and their involvement in disease.

To date, the sequence across the junctions of less than 50 non-recurrent constitutional translocations have been published (Table 1.1). Increasing the pool of data by developing efficient methods of breakpoint mapping may help to elucidate the mechanism underlying the formation of chromosomal rearrangements.

Translocation	Reference
t(X;21)(p21;p12)	Bodrug et al., 1987
t(X;2)(p21;q37)	Bodrug et al., 1991
t(X;4)(p21;q35)	
t(X;1)(p21;p34)	Cockburn, 1991
t(X;4)(p21.2;q31.22)	Giacalone and Francke, 1992
t(4;22)(q12;q12.2)	Arai, Ikeuchi and Nakamura, 1994
t(2;22)(q14;q11.21)	Budarf et al., 1995
t(X;5)(p21;q31.1)	van Bakel et al., 1995
t(X;9)(p21.1;q34.3)	Toriello et al., 1996
t(21;22)(p12;q11)	Holmes et al., 1997
t(X;8)(p22.13;q22.1)	Ishikawa_Brush et al., 1997
t(17;22)(q11.2;q11.2)	Kehrer-Sawatzki et al., 1997
t(6;7)(q16.2;p15.3)	Krebs et al., 1997
t(8;17)(p11.2;p13.3)	Kurahashi et al., 1998
t(1;10)(p22;q21)	Roberts, Chernova and Cowell, 1998
t(2;19)(q11.2;q13.3)	Yoshiura et al., 1998
t(6;12)(q16.2;q21.2)	Ikegawa et al., 1999
t(1;6)(p22.1;q16.2)	Holder, Butte and Zinn, 2000
t(1;8)(q21.1;q22.1)	Matsumoto et al., 2000
t(1;11)q42.1;q14.3)	Millar et al., 2000
t(12;22)(q24.1;q13.3)	Bonaglia et al., 2001
t(1;19)(q21.3;q13.2)	Nothwang et al., 2001
t(9;11)(p24;q23)	Willett-Brozick et al., 2001
t(7;16)(q11.23;q13)	Duba et al., 2002
t(1;8)(p34.3;q21.12)	McMullan et al., 2002
t(2;8)(q31;p21)	Spitz et al., 2002
t(2;8)(q31;p21)	Sugawara et al., 2002
t(6;13)(q21;q12)	Vervoort et al., 2002
t(7;22)(p13;q11.2)	Hill et al., 2003
t(6;11)(q14.2;q25)	Jeffries et al., 2003
t(4;22)(q35.1;q11.2)	Nimmakayalu et al., 2003
t(X;7)(p11.3;q11.21)	Shoichet et al., 2003
t(1;7)(q41;p21)	David et al., 2003
t(1;22)(p21.2;q11)	Gotter et al., 2004
t(3;8)(p14.2;q24.2)	Rodriguez-Perales et al., 2004
t(2;6)(q24.3;q22.31)	Bocciardi et al., 2005
t(4;17)(q28.3;q24.3)	Velagaleti et al., 2005
t(1;7)(p22;q32)	Borg et al., 2005
t(4;15)(q27;q11.2)	Schule et al., 2005
t(4;15)(q22.3;q21.3)	Klar et al., 2005
t(9;11)(q33.1;p15.3)	Tagariello et al., 2006
t(6;17)(p21.31;q11.2)	Mansouri et al., 2006
t(5;14)(q21;q32)	Haider et al., 2006
t(17;22)(q21.1;q12.1)	Gribble et al., 2007
t(2;7)(q37.1;q36.3)	
t(11;17)(p13;p13.1)	
t(2;7)(q37.1;q21.3)	Bocciardi et al., 2007

Table 1.1 Summary of published non-recurrent constitutional translocations.

1.2.1.2 Associated phenotypic risk

A study into the outcome of 377,357 amniocenteses at birth showed that the risk of a *de novo* apparently balanced translocation having an associated phenotypic abnormality is 6.1% (Warburton 1991). The risk of congenital abnormalities amongst carriers of balanced translocations is double that seen amongst individuals with normal karyotypes (Warburton 1991). Analysis of rearrangement breakpoints in relation to the banding pattern observed along Giemsa banded metaphase chromosomes showed that 84% of breakpoints occurred in Giemsa negative, gene rich regions of the genome (Warburton 1991; Niimura and Gojobori 2002).

A recent study investigated the hypothesis that translocation breakpoints in normal individuals were simple and did not disrupt genes (Baptista et al. 2005). The breakpoints in 13 phenotypically normal individuals with apparently balanced translocations were mapped by FISH. At the resolution of the study (approximately 150Kb) the breakpoints were seen to directly disrupt a gene in 2 patients and possibly disrupt a gene in a further 8 patients, the significance of which remained undetermined. An additional observation was made that the translocation breakpoints in phenotypically normal patients were not accompanied by additional imbalance in contrast to the breakpoints in phenotypically abnormal patients which are often seen to be accompanied by cryptic imbalances. FISH and PCR studies have reported that 8 out of 30 apparently balanced reciprocal translocations are more highly rearranged than identified by G-banding (Kumar et al. 1998; Astbury et al. 2004; Patsalis et al. 2004). We have similarly shown that 6 out of 10 apparently balanced translocations as described by G-banding were more complex when analysed by DNA microarray analysis (Gribble et al. 2005). A different study using microarrays and FISH has shown that of the 4 translocation patients studied, all were more complex than originally thought and that in 3 of the patients the

abnormal phenotype was postulated to be associated with the additional imbalance (Ciccone et al. 2005). These studies have led to the suggestion that the genetic diagnosis of constitutional rearrangements should routinely include molecular karyotyping by array based whole genome screening in order to detect submicroscopic imbalances (Vermeesch et al. 2007).

Many published translocation breakpoints have shown that the direct disruption of a gene can lead to an associated phenotype (Bhalla et al. 2004; Bocciardi et al. 2005; Klar et al. 2005) but it is also thought that translocation breakpoints can cause an affect on genes several Kb away. This position effect phenomenon was initially described in *Drosophila* and yeast and was first documented in humans in 1995. A study into a patient with a t(4;11)(q22;p13) translocation showed that the translocation breakpoints did not directly disrupt the *PAX6* gene to which the aniridia phenotype was associated in patients with deletions of the same region (Fantes et al. 1995). The translocation breakpoint was mapped approximately 125-185Kb distal to the *PAX6* gene in the patient with aniridia, suggesting that the translocation was still exhibiting an effect on the gene and the patient's phenotype. Other examples of position effect arising from a chromosome rearrangement include a t(2;8)(q31;p21) translocation which affects the *HOXD* gene 60Kb away from the chromosome 2 breakpoint in a patient with Mesomelic Dysplasia and vertebral defects (Spitz et al. 2002) and a t(6;11)(q14.2;q25) translocation affecting the *B3GAT1* gene which lies 299Kb centromeric to the chromosome 11 breakpoint in a patient with psychosis (Jeffries et al. 2003). The longest range position effect observed to date was observed in a patient with a t(4;7;8;17) translocation. The patient presented with Campomelic Dysplasia which has been attributed to the *SOX9* gene on chromosome 17q24.3. The chromosome 17 breakpoint was found to fall 1.3Mb downstream of the gene (Velagaleti et al. 2005).

The position effect phenomenon arises from the disruption of *cis*-acting regulatory elements such as promoters, enhancers and silencers which can be directly altered, distanced from the gene they influence, or brought into proximity of a gene not normally under their control when a chromosome undergoes a rearrangement. These elements have been observed as far away as 1.1Mb from the gene they regulate, as in the case of the SOX9cre1 element which was identified upstream of the *SOX9* gene (Bien-Willner et al. 2007).

1.2.2 Deletions, duplications, inversions

Regions of repeat sequence have been implicated in the mechanism of rearrangement such as deletions, duplications and inversions as summarised in Figure 1.3. Homologous regions within the genome are thought to align, allowing recombination to occur between the sequences, resulting in rearrangements.

A study of 14.677 conceptions has estimated the frequency of unbalanced structural rearrangements at 1 in 460 (Jacobs et al. 1992) and a study of 377,357 amniocentesis estimated the rate of inversions to be 1 in 10,000 with a 9.4% risk of an associated congenital abnormality (Warburton 1991).

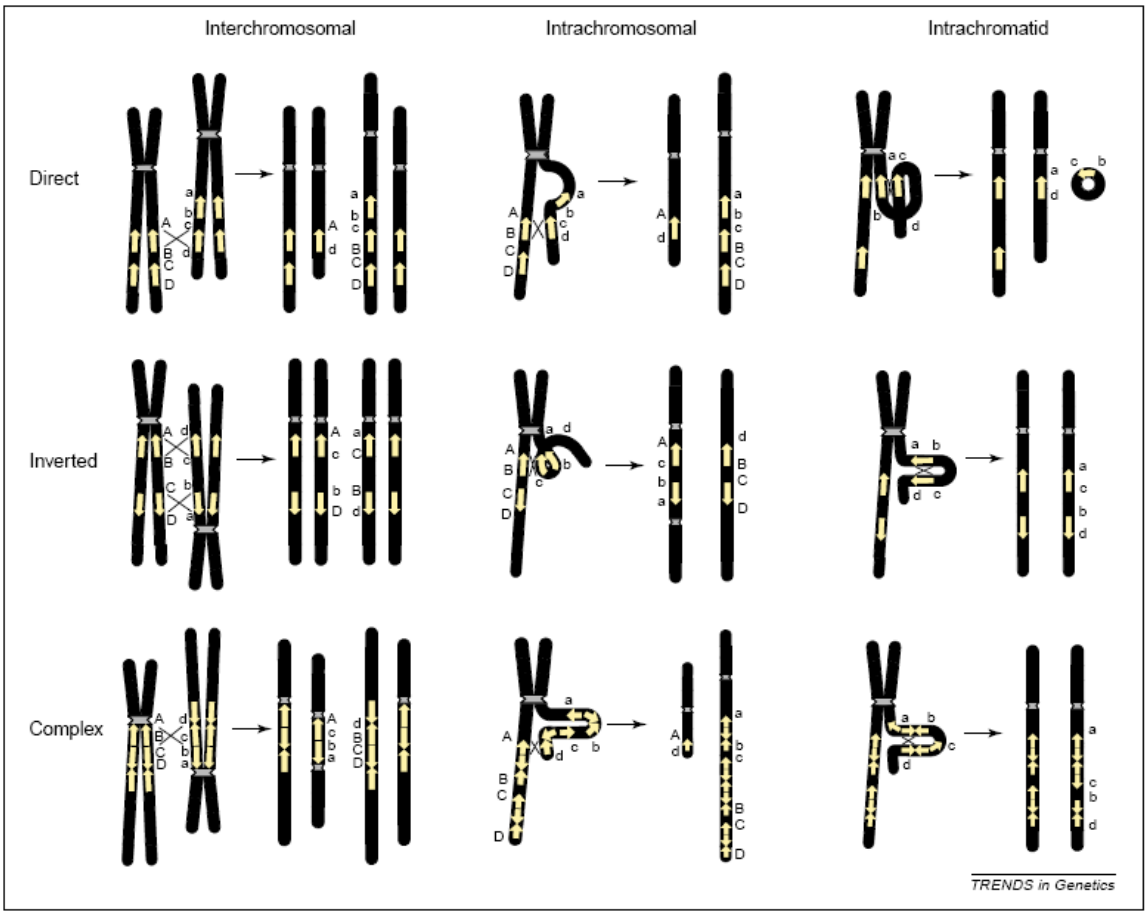

Figure 1.3 Schematic representation of homologous recombination based mechanisms for genomic rearrangements (Stankiewicz and Lupski 2002). Yellow arrows depict the orientation of regions of homology.

Estimates of the repeat content of the genome range from approximately 5-50%. One physical mapping study using 1,243 randomly selected BAC clones found that 5.4% hybridised to more than one chromosomal location (Cheung et al. 2001). Bioinformatic analysis of the human genome draft sequence estimated the repeat content to be approximately 50% (Lander et al. 2001). Repeatmasker analysis of NCBI Build 36 of the human genome reference sequence reveals that current calculations estimate that 48.9% of the genome is repeat sequence (data extracted from http://genome.ucsc.edu/). The discrepancy between the estimates

of repeat sequence can be explained by the resolution afforded by the analysis techniques. Physical analysis of the genome by FISH has a limited resolution and relies on a region of hybridisation being detected visually. Bioinformatic tools such as Repeatmasker will consider repeats of any size ranging from several Megabases (e.g. segmental duplications) to only a few basepairs (e.g. simple repeats) which would not be detected visually by FISH.

1.2.2.1 Phenotypic effect of deletions, duplications and inversions

Changes in the number of copies of an expressed gene may alter a patient's phenotype unless dosage compensation is observed. Loss of a copy of a gene may result in haploinsufficiency, or an increase in the number of genes may result in overexpression. All these changes refer to loss or gain of a functional gene with its associated regulatory elements. However, the breakpoints associated with deletions, duplications and inversions may also disrupt a gene as discussed in section 1.2.1.2.

1.2.2.2 Genomic disorders

The term "genomic disorders" was coined to describe a group of recognised diseases that arise due to rearrangements in the genome which are mediated by the genomic architecture. Well studied examples include Sotos syndrome, Charcot-Marie-Tooth disease type 1A (CMT1A), hereditary neuropathy with liability to pressure palsies (HNPP), Smith Magenis Syndrome (SMS) and dup(17)(p11.2p11.2) syndrome.

Sotos syndrome is a genomic disorder characterised by childhood overgrowth, mental retardation and specific craniofacial features associated with haploinsufficiency of the *NSD1* gene on 5q35 either arising from deletion of the gene or a mutation. Analysis of the Sotos syndrome region revealed two complex mosaic low copy repeats, either side of the *NSD1* gene. The proximal repeat of approximately 390Kb and the distal repeat of approximately 429Kb show an

overall homology of 98.5%. Studies have shown that the deletions associated with Sotos syndrome are likely to be the result of rearrangements between these low copy repeats (Kurotaki et al. 2005; Visser et al. 2005).

CMT1A and HNPP are dysmyelinating peripheral neuropathies resulting from altered dosage of the *PMP22* gene caused by reciprocal duplication and deletion events on 17p12. These recurrent rearrangements of a 1.4Mb genomic region are mediated by proximal and distal low copy repeats which are 24Kb in length with 99% homology (Reiter et al. 1997; Shaw et al. 2004)

SMS is a syndrome associated with mental retardation and multiple congenital abnormalities whilst dup(17)(p11.2p11.2) syndrome sometimes results in milder forms of mental retardation. The 4Mb and 5Mb deletions associated with SMS and the reciprocal duplications associated with dup(17)(p11.2p11.2) syndrome are mediated by 3 LCRs (256Kb proximal, 241Kb middle, 176Kb distal) with 98% homology (Smith et al. 1986; Potocki et al. 2000; Park et al. 2002).

In all these cases, it is believed that the rearrangements arise because of illegitimate homologous recombination via the repeat structures identified. Because of this mechanism, the disorders are seen to be recurrent within the human population, although the exact breakpoints of the rearrangements are observed to vary slightly. For example, a study of patients with CMT1A found that 19 out of 24 breakpoints fell within a 741bp region within the CMT1A repeat (Lopes et al. 1998).

1.2.3 Analysis of chromosome rearrangements

Many techniques have been applied to the analysis of chromosome rearrangements. Conventional cytogenetic techniques such as G-banding of metaphase chromosomes are generally applied for the initial identification of rearrangements, but more recently microarray based techniques have been

developed. Analysis of a chromosome's size and constitution by flow cytometry can also be used to identify large aberrations. Rearrangement breakpoints can be investigated at increased resolution using FISH, somatic cell hybrids, array painting and custom-made libraries.

1.2.3.1 Metaphase chromosome banding analysis

Many stains can be used to analyse banding patterns along metaphase chromosomes. For example, the fluorescent stain DAPI preferentially binds to AT rich DNA resulting in regions of dark and light staining along the chromosome corresponding to GC and AT rich regions (Lin et al. 1977). Conventional G-banding uses Giemsa to produce a banding pattern. Analysis of this pattern in relation to the human genome reference sequence showed that Giemsa positive bands (dark) were AT rich and gene poor and Giemsa negative bands (light) were GC rich and gene rich (Niimura and Gojobori 2002). Analysis of chromosome banding patterns can identify chromosome rearrangements and localise the breakpoints to within approximately 3Mb depending on the quality and length of the prepared chromosomes (Lichter et al. 2000).

1.2.3.2 FISH

The development of fluorescent *in situ* hybridisation (FISH) has increased the resolution with which chromosome rearrangements can be studied and the completion of the Human Genome Project has provided an invaluable resource of fully sequenced large insert clones enabling scientists to study human chromosomes in more detail more easily. Once a chromosome rearrangement has been identified by banding analysis, large insert clones can be selected across the region of interest, fluorescently labelled and hybridised to metaphase chromosomes (Bauman et al. 1980). Analysis of the signals obtained from the hybridisation of subsequent clones along the chromosome can identify clones which span the rearrangement breakpoints.

1.2.3.3 Fluorescence activated chromosome sorting

Fluorescence activated chromosome sorting (FACS) was developed in 1979 as a technique for the separation of individual human chromosomes (Gray et al. 1979). Two fluorochromes were used for chromosome staining; Hoechst 33258 which binds to AT rich DNA and Chromomycin A3 which binds to GC rich DNA. After staining, chromosomes can be resolved by their DNA content and relative size with the exception of chromosomes 9-12 which remain as a single peak due to their similar size and basepair constitutions (Figure 1.4). Rearrangements such as reciprocal translocations which alter the size and basepair constitution of chromosomes can be detected using flow cytometry.

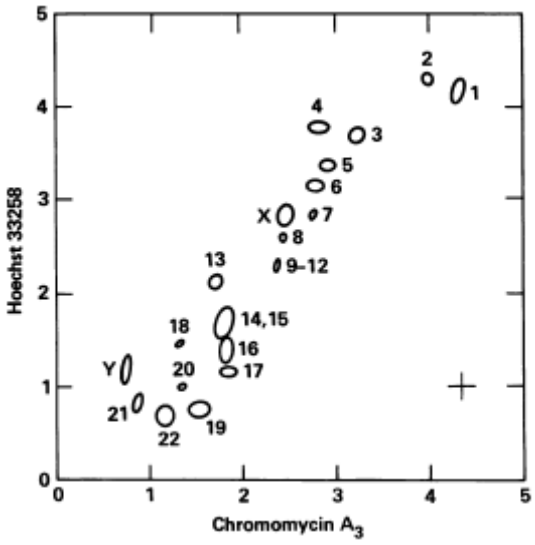

Figure 1.4 Statistical summary of peaks seen for the resolution of human chromosomes when stained with Hoechst 33258 and Chromomycin A3 for a normal 46,XY karyotype (Langlois et al. 1982). Chromosomes 9, 10, 11 and 12 do not resolve into discrete peaks.

Chromosome sorting technologies can be used for the isolation of derivative chromosomes and to improve the purity of template DNA for the analysis of chromosome rearrangements.

1.2.3.4 Somatic cell hybrids

Somatic cell hybrids have been used to isolate derivative chromosomes from patients with chromosome translocations. The derivative chromosomes can be digested to create a restriction digest map for Southern blotting.

1.2.3.5 Southern blotting

Southern blotting has been used to identify fragments of DNA that contain translocation breakpoints. Fragments of DNA from the digested derivative chromosomes isolated by somatic cell hybrids, or digested genomic DNA from the patient can be screened using probes designed from breakpoint spanning clones identified by FISH. These fragments can then be cloned, sequenced and analysed in comparison with the human genome reference sequence to identify the breakpoint position (Vervoort et al. 2002).

1.2.3.6 Comparative genomic hybridisation

The principle of comparative genomic hybridisation (CGH) was developed to analyse the genomic changes within cancer cell lines in direct comparison with the DNA of normal individuals (Kallioniemi et al. 1992). DNA from a test cell line labelled with one fluorochrome, and a reference DNA labelled with a second fluorochrome were simultaneously hybridised to metaphase chromosomes in the presence of Cot1 DNA to block repetitive sequences. Analysis of the changes in the ratios of intensity of the 2 fluorochromes revealed regions of amplification and deletion within the 2 samples at a resolution of approximately 10Mb with subsequent studies increasing the resolution to approximately 3Mb (Kirchhoff et al. 1999).

In order to increase the resolution of amplification and deletion screening within the genome of interest further, genomic microarrays were developed and used as a target for CGH, replacing metaphase chromosomes (Solinas-Toldo et al. 1997).

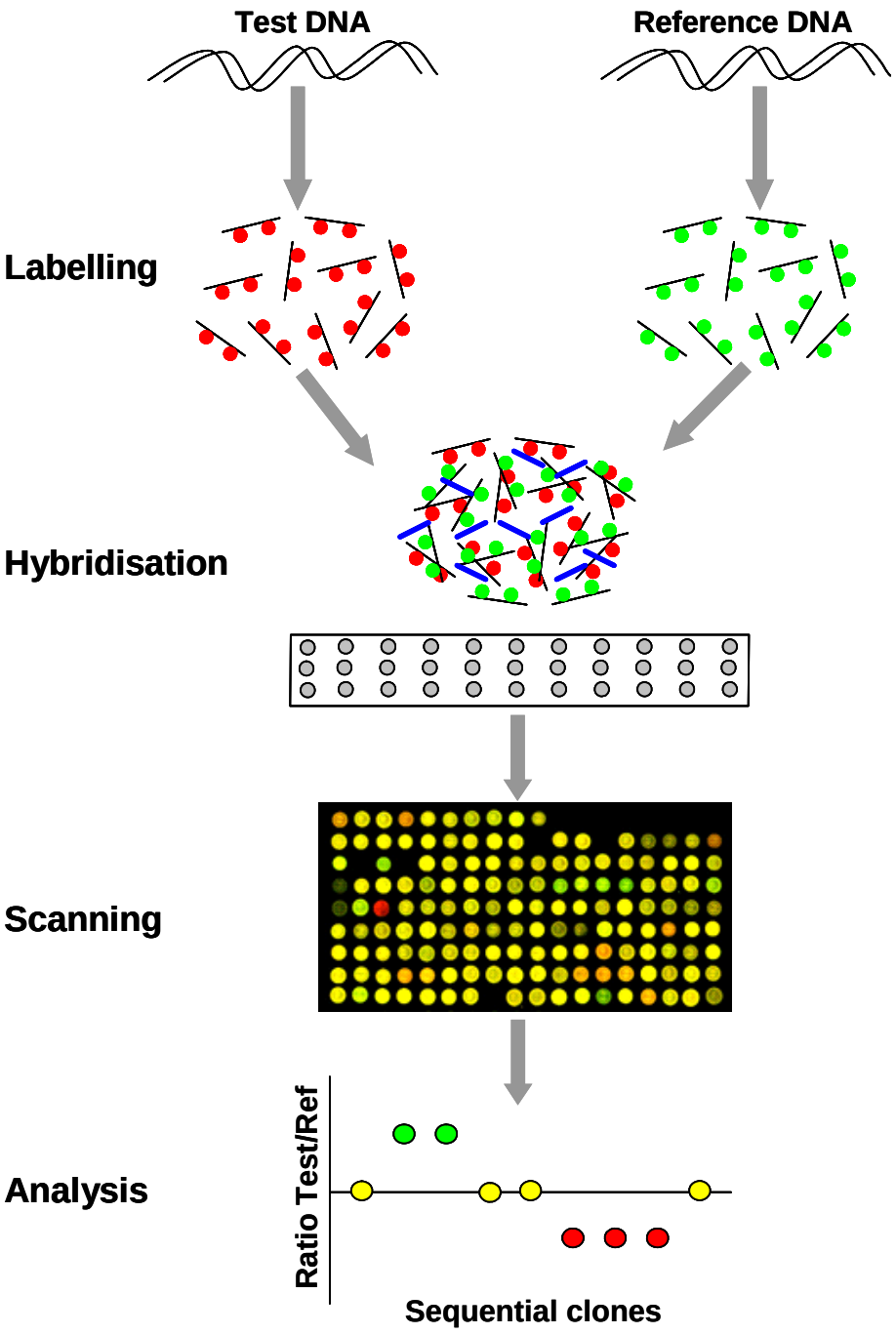

Figure 1.5 The principles of array CGH. The test (red) and reference (green) DNAs are differentially labelled and hybridised to a genomic microarray slide in the presence of Cot1 DNA (blue). Amplifications and deletions are identified by differences in the fluorochrome ratios.

During an array CGH experiment, patient genomic DNA and reference DNA are differentially labelled and competed with Cot1 DNA. The mixture is then hybridised to the target DNA spotted on the microarray slide to identify any genomic imbalances between the 2 samples. A change in the ratio of fluorescence intensities of the clones indicates a copy number change (Figure 1.5).

The advent of genomic microarrays has meant that a patient's genome can be studied more rapidly and at a greater resolution than was previously possible by FISH analysis with the resolution of the genomic microarray limited only by the size and spacing of the probes (clones, PCR products or oligonucleotides) selected during its design.

The most frequently used DNA microarrays to date have a resolution of approximately 1-1.4Mb (Pollack et al. 1999; Snijders et al. 2001; Fiegler et al. 2003a; Vissers et al. 2003; Schoumans et al. 2004; Shaw-Smith et al. 2004). These arrays provide genome-wide scans at relatively low resolution. To further refine rearrangement breakpoints, higher resolution arrays can be used. Arrays at tiling path resolution have recently been created with the ability to map chromosome breakpoints within a spanning clone in a single experiment (Ishkanian et al. 2004; Fiegler et al. 2007). The large insert clones used to create these arrays provide a resolution of approximately 150Kb.

To generate genomic clone microarrays with further increased resolution, alternative library resources are available. A library of fosmid clones was created and end sequenced as part of the human genome project providing a validation method for the sequence data generated (IHGSC 2004). This library had on average an 8 fold coverage of the euchromatic regions of the human genome providing a high level of redundancy. Clones can be selected from this publicly available resource to cover a genomic region of interest and create a targeted

custom-made microarray with a high level of redundancy and at a higher resolution than that afforded by large insert clone microarrays.

Custom-made microarrays using alternative targets to genomic clones have been created to investigate particular genomic regions of interest at increased resolution. Microarrays constructed from PCR products were used to analyse genomic imbalances across the Neurofibromatosis 2 gene. PCR products ranging in size from 150-650bp were pooled affording a resolution of 23Kb (Mantripragada et al. 2003). Use of PCR products spotted as individual targets onto microarrays slides has subsequently increased the resolution of PCR product arrays (Dhami et al. 2005). By amplifying genomic sequence using PCR primer pairs tagged with an 8bp universal sequence and performing a second round of amplification using this universal sequence to attach an amino group, each product was able to be spotted individually onto the microarray, allowing the resolution of the microarray to be as small as the PCR product.

A single nucleotide polymorphism (SNP) microarray was first developed to analyse sequence polymorphisms within genomes (Chee et al. 1996). 135,000 oligonucleotide probes covering the 16.6Kb human mitochondrial genome were synthesised and spotted onto an array, revolutionising the resolution with which genomes could be studied.

Oligonucleotide based microarrays have since been developed to investigate imbalance within genomes, so increasing the resolution of array CGH. A study using a microarray with 60-mer oligonucleotides showed that when compared with CGH using BAC arrays the oligonucleotide arrays were able to detect amplifications with a higher accuracy and greater special resolution (Carvalho et al. 2004).

Microarrays using oligonucleotides directly synthesised onto microarray slides have been used to probe the genomes of 5 cancer patients at a resolution of 6Kb (Selzer et al. 2005). In the same study, custom-made oligonucleotide microarrays at higher resolution were created to map the breakpoints of the rearrangements identified by the whole genome screen to intervals as low as 50bp. NimbleGen Systems, Inc. routinely use oligonucleotide microarrays with a median probe spacing of 713bp for whole genome screening. Custom-made oligonucleotide microarrays can also be generated with a median probe spacing from 10bp.

The power of array CGH for the identification of rearrangements involved in disorders has been demonstrated by a study of 290 individuals with mental retardation (Sharp et al. 2006). A genomic BAC microarray was created to cover regions of the genome suspected of instability due to flanking segmental duplications. Four individuals were identified with the same 500Kb deletion of 17q21.31 and subsequent analysis of these patients using a targeted oligonucleotide microarray with a mean spacing of 1 probe every 131bp revealed a minimal critical region of 478Kb.

1.2.3.7 Array painting

Array CGH is a powerful tool for identifying unbalanced chromosome rearrangements, but it is unable to detect balanced rearrangements. Array painting is an adaptation of the array CGH technique which overcomes this limitation. The genomic DNA used in CGH experiments is replaced with flow sorted derivative chromosomes which are differentially labelled and hybridised to the microarray to map translocation chromosome breakpoints (Fiegler et al. 2003b) (Figure 1.6).

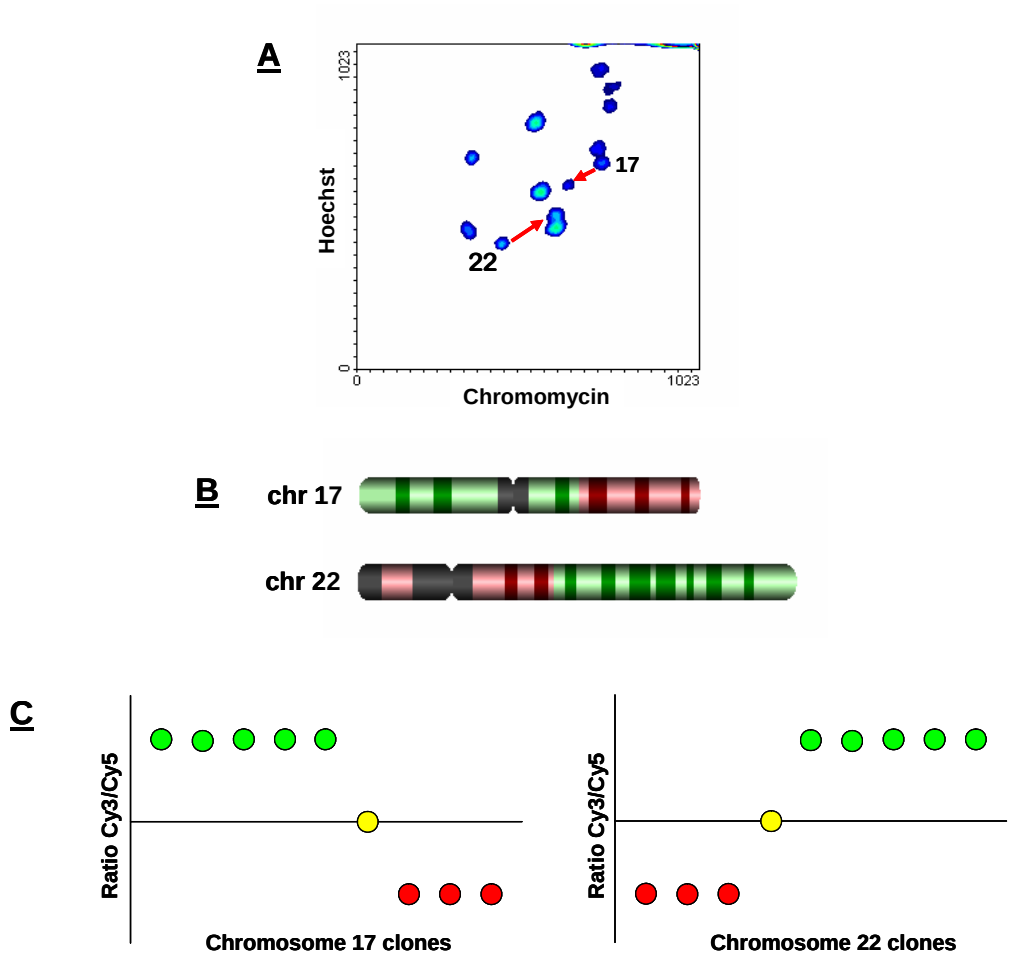

Figure 1.6 ***Schematic showing the process of array painting for a patient with a t(17;22)(q21.1;q12.2) translocation. A Partial flow karyogram showing the shift between chromosome 17 and derivative chromosome 17 and also chromosome 22 and derivative chromosome 22 (indicated by the red arrows); B Ideogram showing the translocation after differential labelling of the flow sorted derivative chromosomes; C Array painting analysis of the Cy3/Cy5 ratios. The breakpoint region is identified by a shift from a low to high ratio or vice versa. A spanning clone (yellow) will show an intermediate ratio.***

Array painting in conjunction with array CGH can map the breakpoints of reciprocal translocations and detect any additional imbalance within the patient's genome.

1.2.3.8 Custom-made libraries

As discussed in section 1.2.1.2, apparently balanced reciprocal translocations are often more complex than initially identified by G-banding, highlighting the need for alternative methodologies of breakpoint mapping that do not rely heavily on knowledge of the sequence surrounding the breakpoints.

The use of libraries has been instrumental to the mapping and sequencing of the human genome. Large insert clone libraries were constructed; YAC clones (Traver et al. 1989; Albertsen et al. 1990), PAC clones (Ioannou et al. 1994), BAC clones (Kim et al. 1996; Osoegawa et al. 2001), cosmid clones (Wood et al. 1992) and fosmid clones (IHGSC 2004). Generation and screening of a custom-made library using DNA from a translocation patient relies only on limited sequence information surrounding the breakpoints.

During the generation of a library, DNA is fragmented to the appropriate size, ligated into a vector, packaged into a suitable host and cloned as summarised in Figure 1.7. Many different vector and host systems exist, such as yeast artificial chromosomes (YACs), P1 derived artificial chromosomes (PACs), bacterial artificial chromosomes (BACs), fosmids, cosmids and plasmids capable of containing inserts ranging in size from approximately 1Mb to under 5Kb.

Many factors affect the decision of which vector and host system is used. YAC libraries were initially used in cloning studies as they were capable of cloning large fragments of DNA up to 1Mb, however they have been shown to be prone to rearrangement. The choice of cloning vector used is often dependent upon the resolution of the library required; fosmid or cosmid clones contain inserts of approximately 40Kb and plasmid clones contain inserts less than 15Kb in size.

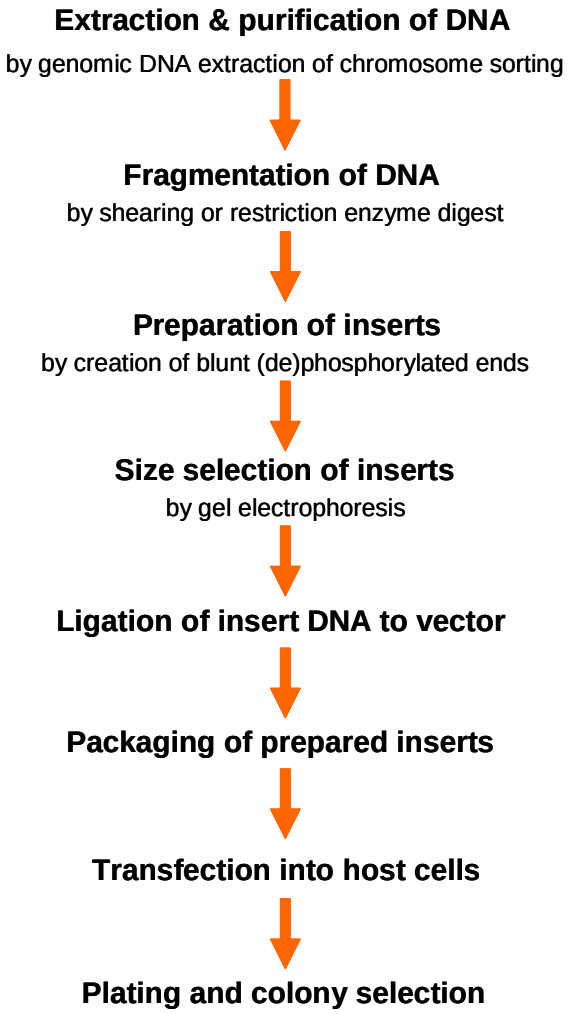

Figure 1.7 Schematic showing the key stages in the production of a custom-made library. Precise stages required are dependant upon the type of vector and host.

1.2.3.9 Sequence analysis

The completion of the human genome reference sequence announced in 2004 provided an invaluable resource for the investigation of chromosome rearrangements at sequence level (IHGSC 2004). Approximately 99% of the euchromatic regions of the genome had been sequenced, with the data made publicly available in genome browsers such as UCSC (http://genome.ucsc.edu/)

and Ensembl (http://www.ensembl.org/Homo_sapiens/index.html). In addition to the reference sequence, these databases provided details of genomic clones and their relative positions available for genomic research. These websites also allowed researchers to download sequence and interrogate regions of interest for fully annotated genomic structures including genes, repeat elements, polymorphic regions and regions of conservation. Multiple web-based tools also exist for the analysis of sequence for genomic motifs (Appendix A1). The analysis of genomic sequence with relation to rearrangement breakpoints and their influence in disease is enhanced by the availability of such tools which enable scientists to link genomic aberrations to phenotypes, therefore identifying candidate disease genes.

1.2.4 Public databases of constitutional rearrangements

In general, constitutional non-recurrent rearrangements are investigated by clinical and research groups around the world, with little or no contact. In an effort to enhance communication between these groups, web-based databases have been developed such as the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (Decipher; http://www.sanger.ac.uk/PostGenomics/decipher/), the Gross Rearrangement Breakpoint Database (GRaBD; http://archive.uwcm.ac.uk/uwcm/mg/grabd/) and the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA; http://agserver01.azn.nl:8080/ecaruca/ecaruca.jsp). The aim of these databases is to co-ordinate research into rare phenotypes and their underlying rearrangements with the goal of pooling research and possibly identifying the causative genes involved such as in the 17q21.3 microdeletion syndrome (Shaw-Smith et al. 2006).

1.2.5 Normal variation within the human genome

The publication of the human genome reference sequence by the Human Genome Project provided a huge resource with which to compare the genomes

of patients with clinical abnormalities. However, differences are seen between the genomes of apparently normal individuals and it is believed that this heterogeneity within the human genome might partly account for the huge amount of phenotypic diversity seen between individuals. Two initial studies into this variation using a 1Mb large insert clone microarray and an oligonucleotide microarray comparing 55 and 20 individuals respectively, revealed variations ranging from 100Kb to 2Mb often encompassing known genes (Iafrate et al. 2004; Sebat et al. 2004). A further study using the end sequence reads from a fosmid library created from an alternative individual to the human genome reference sequence source revealed 297 structural variation sites greater than 8Kb (Tuzun et al. 2005). A more recent study into this copy number variation (CNV) using 270 individuals from 4 different populations showed that 12% of the genome was identified as being polymorphic (Redon et al. 2006). The data produced by this study has been incorporated into the UCSC and Ensembl web browsers, providing an invaluable tool for the comparison of patient data with known polymorphisms.

CNV data is vital in the study of the effect of chromosomal rearrangements on phenotype. Once a rearrangement has been discovered, it must be determined whether it is causal to the phenotype or is a polymorphism with no phenotypic effect. Analysis of the CNV data for a particular genomic region will identify common polymorphisms with no phenotypic effect. In addition, analysis of a proband's parental DNA will establish whether the event is *de novo* or familial and likely to be causal to the aberrant phenotype.

Whilst a CNV itself might not be instrumental in directly causing a disease phenotype, it might predispose. One study of 30 patients with Thrombocytopenia-Absent Radius (TAR) Syndrome showed that all patients had a deletion at 1q21.1 with a minimal critical region of approximately 200Kb containing 12 genes (Klopocki et al. 2007). Only 25% of these deletions were found to be *de novo*,

with the remaining deletions inherited maternally or paternally from phenotypically normal parents. These results led to the hypothesis that although the syndrome was associated with the deletion, an additional event was required for presentation of the phenotype, suggesting that the region at 1q21.1 is a susceptibility locus.

1.3 Mechanisms behind chromosomal rearrangements

Cellular DNA continuously undergoes a process of breakage and repair which can lead to chromosome rearrangements after the mis-repair of double strand breaks.

1.3.1 Double strand breaks

Double strand breaks can occur along DNA as a result of environmental factors such as ultra-violet and ionising radiation or as a result of normal metabolic function such as degradation by the topoisomerase type II protein, Spo11. Spo11 uses its catalytic tyrosine to attack the phosphodiester backbone of DNA, creating a covalent bond between itself and the 5’ end of the break.

After formation of double strand breaks in the DNA a cell is able to maintain its integrity by either of 2 cellular processes which occur during the cycle of a normal cell; Homologous recombination and non-homologous end joining. Homologous recombination is a high fidelity mechanism which typically occurs without any alteration in the genetic sequence as it relies on a region of almost identical sequence to act as template. However, non-homologous end joining is often accompanied by the loss or addition of nucleotides as the free strands of DNA are altered during the repair of the DNA strands.

Evidence suggests that the mechanism behind the repair of double strand breaks in recurrent rearrangements is homologous recombination as the breakpoints for

these genomic disorders tend to fall within repeat regions. For non-recurrent rearrangements, no mechanism has so far been definitively identified.

1.3.2 Homologous recombination

Recombination was first observed in *D. melanogaster* where it was seen that blocks of genes from homologous chromosomes could be exchanged during a crossing over event generating increased genetic diversity (Morgan 1910). Current models of homologous recombination rely on two sequences with a high degree of homology aligning, followed by the formation of a double strand break and the subsequent re-forming of the phosphodiester bonds to form uninterrupted DNA strands.

Recombination is thought to occur via the formation of an intermediate structure known as the Holliday Junction (Holliday 1964). There are currently 2 proposed methods for the formation of this structure and 2 alternatives for its resolution.

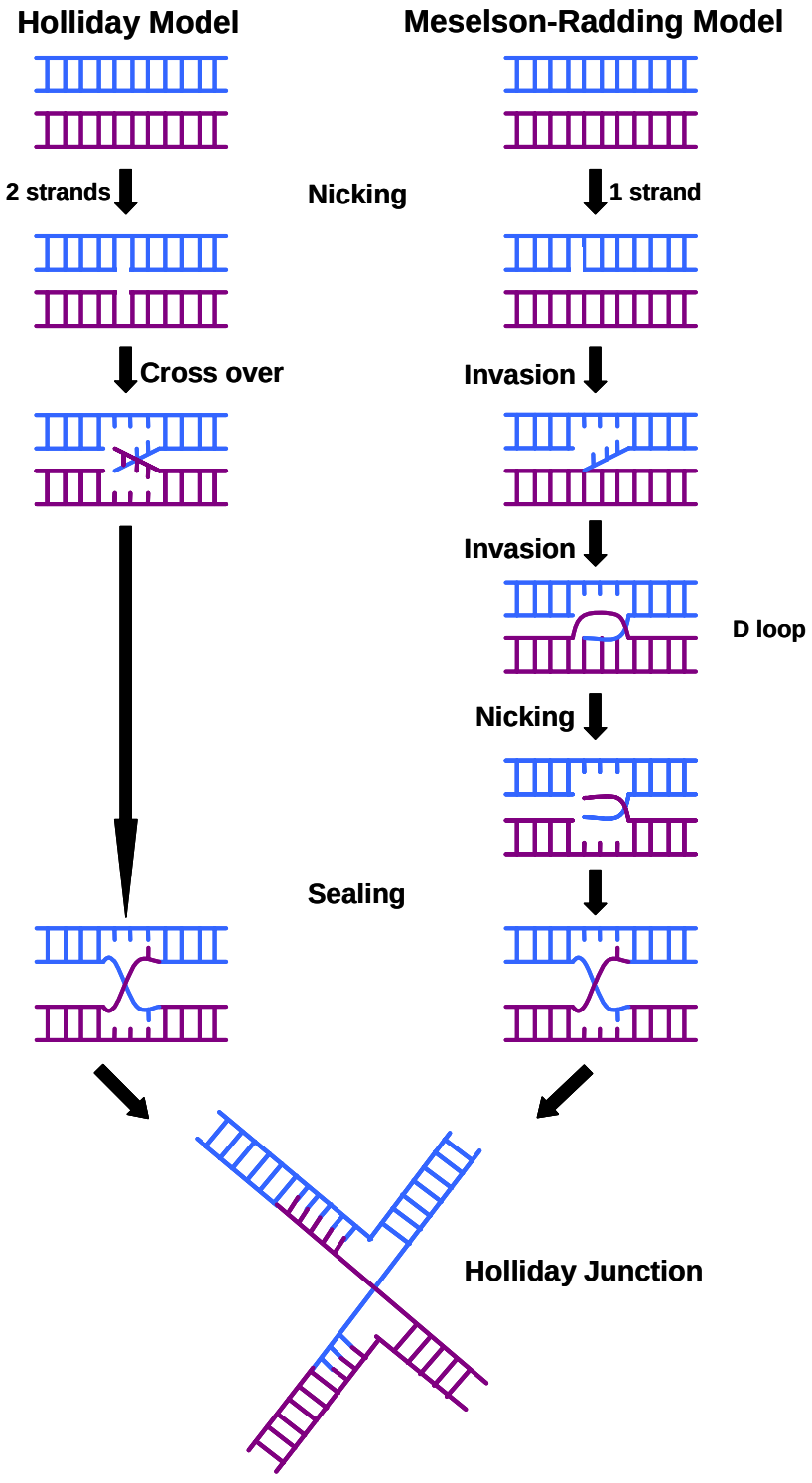

Figure 1.8 Comparison between the Holliday and Meselson-Radding models for formation of the Holliday Junction intermediate structure during recombination.

For the formation of the Holliday junction, the first method proposes that two homologous sequences are aligned and a nick is made in one strand of each promoting strand exchange which occurs at the site of the nicks (Holliday 1964). In the second model, only one nick is needed creating a 5' phosphate end which invades the homologous sequence and causes strand displacement (creating a D-loop) (Meselson and Radding 1975). Both models are compared in Figure 1.8.

The length of homologous sequence present has been shown to have a big impact on recombination within *E.coli* and yeast and regions of mis-match within the homologous regions also have an effect (Watt et al. 1985; Hua et al. 1997). Homologous recombination in somatic mammalian cells is believed to require a minimum of 134-232bp of uninterrupted homology (Waldman and Liskay 1988) and meiotic homologous recombination in humans is believed to require a minimum of 337–456bp (Reiter et al. 1998).

For the resolution of the Holliday junction, the first method proposes that all four strands are cut at the crossover site resulting in recombinant chromosomes. The second method proposes that the Holliday structure rotates at the crossover site and two strands are cut (if the original un-nicked strands are cut, recombinants are formed but if the original nicked strands are cut, recombinant chromosomes are not formed) (Figure 1.9).

Variations in the rates of recombination within the human genome have revealed hotspots of recombination although the exact reason for this remains unknown (Crawford et al. 2004; McVean et al. 2004). It is estimated that approximately 80% of recombination occurs within 10-20% of the sequence with an average of one hotspot every 50Kb along the genome (Myers et al. 2005). Known regions of meiotic recombination hotspots have been collated from the literature in an effort

to improve our understanding of the mechanisms behind this phenomenon (Nishant et al. 2006).

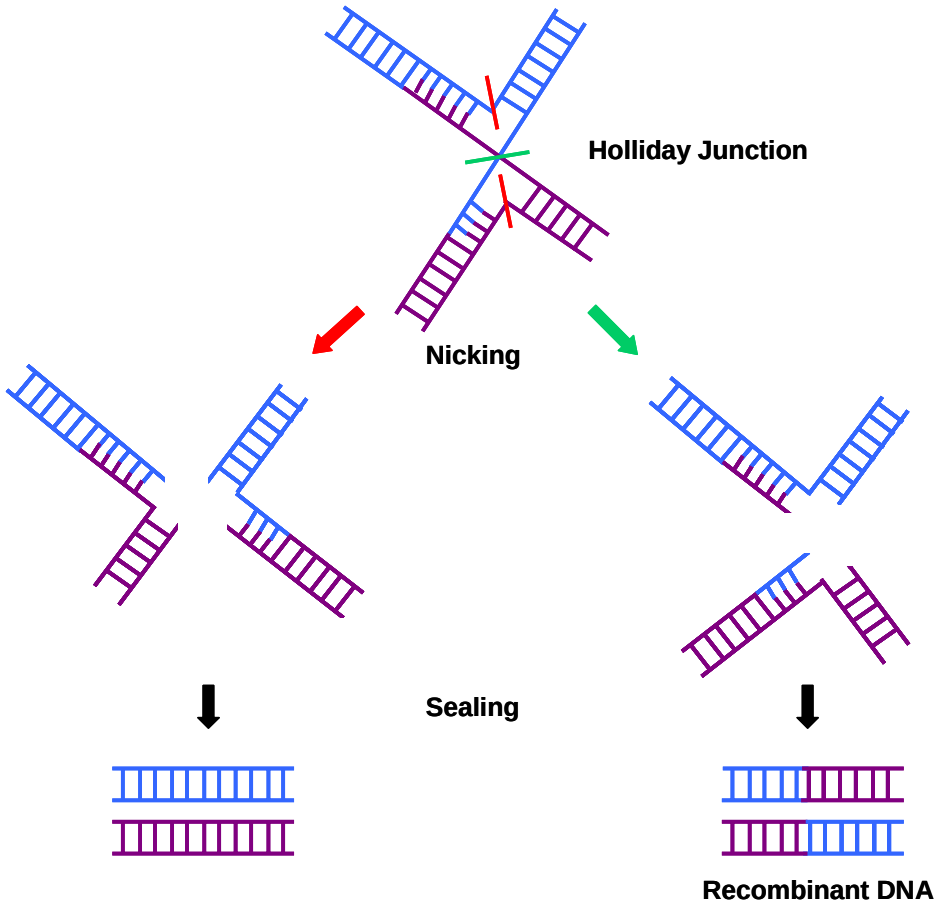

Figure 1.9 Schematic showing the resolution of the Holliday junction during the process of homologous recombination. If the nicking occurs in the previously unaffected strands and no strand invasion has occurred (not depicted on diagram), no recombinant DNA is observed (red). If the 2 strands that were nicked during formation of the Holliday junction are nicked again, the result is recombinant DNA (green).

1.3.3 Non-homologous end joining

When double strand breaks occur in non repetitive regions of the genome and a region of homology cannot be found to act as template for the repair, then the cell relies on non-homologous end joining. Typically the broken ends will be

incompatible and the rejoining will require nucleases to remove nucleotides and/or polymerases to fill in nucleotides, resulting in the characteristic gain or loss of nucleotides at the repaired junctions.

1.3.4 Genome architecture associated with rearrangements

Certain genetic diseases are the result of recurrent chromosomal rearrangements at particular regions of the genome. Analysis of these rearrangements and use of the human genome reference sequence across these regions has identified certain features that may be associated ranging from recombinogenic sequence motifs and repeat sequences to secondary DNA structures that leave the DNA prone to double strand breaks.

1.3.4.1 Chi sequences

The Chi element is known to be a mediator of prokaryotic recombination. The RecBCD enzyme in E.coli recognises the Chi sequence motif GCTGGTGG and binds, promoting recombination (Smith et al. 1981). In humans a Chi-like sequence has been observed with the consensus; GC[A/T]GG[A/T]GG (Krowczynska et al. 1990).

1.3.4.2 Translin associated motifs

Translin motifs have been associated with the translocation breakpoints in many lymphoid malignancies. The consensus sequence observed (ATGCAG and GCCC[A/T][G/C][G/C][A/T]) are identified by the translin protein which binds to the DNA and is believed to be involved in chromosomal translocations (Aoki et al. 1995).

1.3.4.3 Topoisomerase I and II sites

Topoisomerases catalyse DNA unwinding. Topoisomerase I cuts a single strand whilst Topoisomerase II cuts both strands. Topoisomerase II sites also function as chromosome scaffold attachment sites – a chromosome configuration that brings distant DNA sequences into close proximity. The Topoisomerase I

consensus sequence is [A/T][G/C][A/T]T (Been et al. 1984) and there are 3 Topoisomerase II recognition sequences;

Vertebrate [A/G]N[T/C]NNCNNG[T/C]NG[G/T]TN[T/C]N[T/C]

Drosophila GTN[T/A]A[C/T]ATTNATNNG

Invertebrate [T/C][A/C]CNTAC[C/G][C/T]CC[T/G][T/C][T/C]TNNC

(Sander and Hsieh 1985; Spitzner and Muller 1988; Kas and Laemmli 1992).

1.3.4.4 Immunoglobulin heptamers

The immunoglobulin gene locus undergoes recombination which results in variation within the antibody response system. The immunoglobulin recognition sequence consists of a heptamer with the consensus GATAGTG or CACAGTC separated from a nonamer with the consensus sequence ACAAAAACC (Sakano et al. 1979; Chen et al. 1989).

1.3.4.5 Mini satellite sequences

The discovery of hypervariable mini satellite DNA sequences and their involvement in human polymorphisms lead to the suggestion that they might be involved in chromosomal rearrangements (Wyman and White 1980). They are comprised of short tandemly repeated DNA fragments with a core sequence of GGGCAGGC[A/G]G (Jeffreys et al. 1985; Jeffreys 1987).

1.3.4.6 Purine/pyrimidine tracts

Purine (A/G) and pyrimidine (T/C) tracts have been observed to be recombinogenic (Boehm et al. 1989; Majewski and Ott 2000) and a study of 96 chromosome translocations revealed that around breakpoints alternating purine/pyrimidine tracts (2-30bp) were underrepresented, polypurine tracts (2-23bp) were overrepresented and polypyrimidine tracts (2-44bp) were overrepresented (Abeysinghe et al. 2003).

1.3.4.7 SINE elements

Short Interspersed Nuclear Elements (SINE elements) are transposable repeat elements which are typically 100-400bp long (Smit 1996). There are 2 major classes; Alu repeats and the MIR family (Mammalian-wide interspersed repeat).

The Alu repeat is the most abundant sequence in the human genome with an estimated 1 million copies accounting for just over 10% (Lander et al. 2001). The full Alu repeat is approx 280bp long flanked by direct repeats of 6-18bp. Alu repeats have a relatively high GC content (56%) and are reported to preferentially be found in the pale bands of G-banded chromosomes (Korenberg and Rykowski 1988).

MIR repeats have approximately 450,000 copies comprising 2.5% of the human genome (Lander et al. 2001).

1.3.4.8 LINE elements

Long interspersed nuclear elements (LINE elements) are typically 6-8Kb in length and the estimated 868,000 copies present in the human genome are responsible for approximately 21% of its composition (Lander et al. 2001). There are 3 classes of LINE elements; L1, L2 and L3. They are a type of transposable element encoding a reverse transcriptase and a DNA-nick-looping enzyme, allowing them to move about the genome autonomously.

1.3.4.9 Long terminal repeats

Long terminal repeats (LTRs) account for approximately 8% of the human genome (Lander et al. 2001). Many LTRs contain viral promoter, enhancer and polyadenylation signals and have functional consequences (Landry and Mager 2002).

1.3.4.10 Class II transposons

Class II Transposons move by excision and reintegrate into the genome without using an RNA intermediate. They are characterized by terminal inverted repeats of 10-500bp in length.

1.3.4.11 Palindromes

Palindromic DNA can lead to unstable single stranded hairpin or double stranded cruciform structures which are prone to double strand breaks. The DNA then illegitimately recombines to form the translocated chromosomes. In particular, the motif TTTAAA has been shown to bend DNA at sites of recombination (Singh et al. 1997).

1.3.4.12 Segmental duplications

Segmental duplications or Low Copy Repeats (LCRs) are regions of the genome that have been duplicated and range from 1-200Kb in length (Lander et al. 2001). These regions of homologous sequence have been implicated in the aetiology of several genomic disorders such as Sotos syndrome (Kurotaki et al. 2005; Visser et al. 2005), Charcot-Marie-Tooth disease type 1A and hereditary neuropathy with liability to pressure palsies (Shaw et al. 2004) and Smith-Magenis syndrome and dup(17)(p11.2p11.2) syndrome (Smith et al. 1986; Potocki et al. 2000). The inter- and intra- chromosomal segmental duplications within the human genome greater that 10Kb are summarised in Figure 1.10.

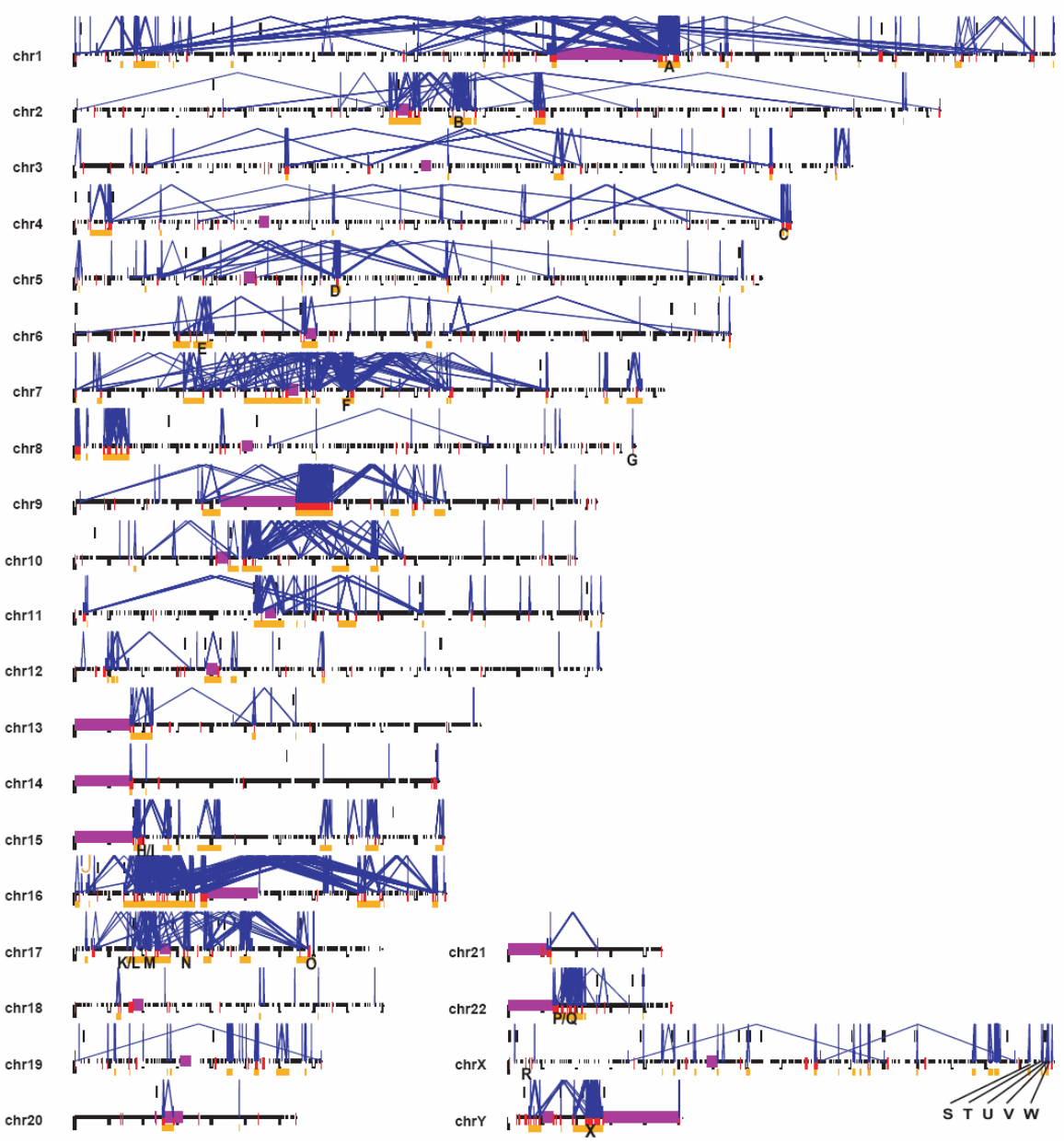

Figure 1.10 ***Patterns of interchromosomal (red) and intrachromosomal (blue) segmental duplications of ≥10Kb with ≥95% homology across the human genome. Colour coding; Purple - Areas not sequenced as part of the human genome project (Acrocentric chromosome arms, heterochromatic regions, centromeres) and orange - unique regions of the genome 50Kb-10Mb in size encompassed by intrachromosomal duplications. 24 regions labelled A to X are known hotspots for genomic rearrangement (Full details in Appendix A2) (Bailey et al. 2002).***

1.4 Viral Integration into the human genome

Human chromosomes are also susceptible to rearrangement by integration from viral genomes. Viruses have the ability to invade human cells, particularly in immunocompromised patients, and have been seen to integrate into the human genome. A well characterised example of targeted viral integration is the adeno-associated virus (AAV). The virus has 2 terminal repeats of 145bp and evidence suggests that viral insertions are in a tandem head to tail orientation via these repeats (Cheung et al. 1980). More recently, this integration has been shown to be a targeted event with integration into the distal portion of chromosome 19q (Samulski et al. 1991). Generation of sequence across and subsequent analysis of the viral-cellular junctions revealed a 2-3bp homology between the viral and cellular DNAs.

1.4.1 The Human Herpes Virus

The human herpes virus (HHV) is ubiquitous in the human population and normally exists as an extra chromosomal element in the host (Kondo et al. 1991; Luppi et al. 1999). There are 8 members in the human herpes virus family; HHV-1 to HHV-8. HHV-6 is a lymphotropic virus which has been identified as a causative agent in the illness exanthum subitum, also known as Roseola infantum, whose patients present with rash, fever and febrile convulsions (Kondo et al. 1991). The data suggested that the viral DNA may persist within cells in a latent form before being reactivated. HHV-6 exists as two variants; HHV-6A and HHV-6B which are highly conserved genetically with the amino acid sequence of genes sharing 75-95% homology (Lindquester et al. 1996).

HHV-6 was initially isolated from peripheral blood leukocytes in 6 patients with lymphoproliferative disorders and 4 healthy donor samples (Salahuddin et al. 1986). Subsequent investigations isolated HHV-6 from HIV-infected patients (Downing et al. 1987; Tedder et al. 1987; Lopez et al. 1988), immunocompromised patients following organ transplantation (Ward et al. 1989)

and from healthy individuals (Pietroboni et al. 1988; Saxinger et al. 1988; Okuno et al. 1989; Harnett et al. 1990). The human herpes virus has the ability to cause disease in immuno-compromised patients, such as transplant recipients and is thought to be a cofactor in the progression of HIV disease (Salahuddin et al. 1986). Studies have shown that up to approximately 95% of the global population have been infected (Aberle et al. 1996) with infection usually occurring early in life (Briggs et al. 1988; Knowles and Gardner 1988).

1.4.2 Evidence of integration and inheritance

A study of 3 patients using Pulsed field gel electrophoresis and Southern blot analysis showed that the virus was attached to high molecular weight DNA and the authors hypothesised that this DNA was human chromosomal DNA (Luppi et al. 1993). Further investigations into these 3 patients by FISH showed that the virus integrated into 17p13 in all 3 cases (Torelli et al. 1995), later confirmed to be close to or within the telomeric sequence of 17p (Morris et al. 1999).

A further study into the prevalence of HHV-6 in healthy individuals using quantitative PCR revealed that 36% (9 out of 25) of the volunteers studied exhibited a viral burden in their blood (Clark et al. 1996). One of these volunteers showed a consistently high burden over the 10 months of the study, a suggested result of integration into the human genome leading to the conclusion that the HHV-6 genome has integrated into the genome of approximately 3% of the British population.

More recently evidence of inheritance of the integrated viral genome has been obtained. In one study, FISH identified integration of the HHV-6 virus into 1q44 of a female patient. Subsequent analysis of the patient's offspring and grandchildren revealed that her son and granddaughter both showed HHV-6 integration at the same 1q44 locus (Daibata et al. 1998). An additional study

identified a family in which the mother carried HHV-6 integration at 22q13, the father at 1q44, and their daughter at both locations (Daibata et al. 1999).

1.4.3 The structure of the HHV-6B Virus

Two strains of the HHV-6B genome have been fully sequenced, Z29 (NCBI accession number; NC_000898) and HST (NCBI accession number; AB021506). The Z29 strain was fully characterised by restriction digest patterns using BamHI, ClaI, HindIII and SalI (Pellett et al. 1990). This data was subsequently verified and the restriction digest fragments cloned to provide a valuable resource for investigations into HHV-6B infection (Lindquester et al. 1996). The HHV-6B virus is a linear genome of approximately 160Kb consisting of a central segment of approximately 141Kb of unique sequence flanked by two direct repeats (termed DR_L and DR_R) which can vary in length from 10 to 13Kb (Figure 1.11). These repeats containing 2 smaller blocks of sequence $(GGGTTA)_n$ similar to that found at the telomeres of human chromosomes, $(TTAGGG)_n$ (Moyzis et al. 1988; Thomson et al. 1994) and it is postulated that this might be the mechanism by which the virus is able to integrate into the human genome.

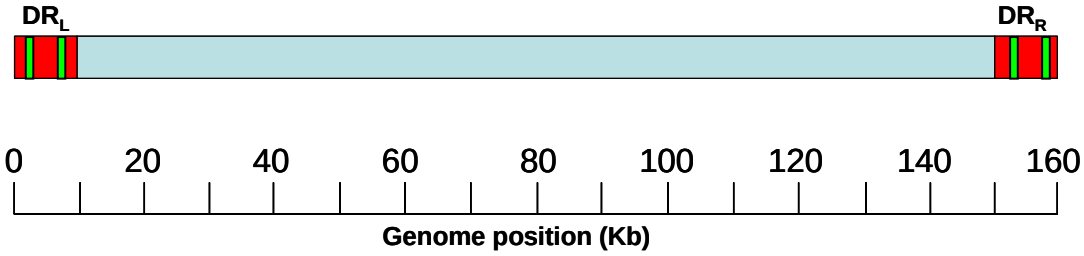

Figure 1.11 Schematic of the HHV-6 genome showing unique sequence (blue) of the viral gnome flanked by direct repeats (red) at the left and right ends of the genome (DR_L and DR_R). These direct repeats each contain two regions of repeats (green) similar to the human telomeric repeat sequence.

1.5 Aims of this thesis

1.5.1 Use and adaptation of current techniques for the mapping of rearrangement breakpoints in three patients with abnormal phenotypes

The aim of this study is to map and sequence 3 apparently balanced reciprocal translocations in patients with learning difficulties and/or physical developmental delay with a goal of identifying the underlying cause of the phenotype and the mechanism mediating the rearrangements. To do this I have implemented molecular cytogenetic tools such as FISH and genomic microarrays combined with flow sorting technologies (array painting) followed by PCR to amplify junction fragments.

1.5.2 Development of techniques for the investigation of rearrangement breakpoints

As more translocations are mapped, it has become increasingly apparent that the sequence surrounding the breakpoints may harbour previously undetected abnormalities (Kumar et al. 1998; Astbury et al. 2004; Patsalis et al. 2004; Ciccone et al. 2005; Gribble et al. 2005). These abnormalities may hamper the use of PCR in the refining of translocation breakpoints and the amplification of breakpoint junction fragments. An alternative method to array painting and STS PCR mapping is the generation of a custom-made library derived from the patient's flow sorted derivative chromosomes and the selection of clones chimeric for both donor chromosome sequence either side of the breakpoint. By creating a fosmid library, approximately 40Kb of sequence around the breakpoint can be obtained and studied for changes compared to the reference sequence. A library approach may also be beneficial to obtain sequence across a breakpoint if other methods fail to sufficiently refine it. To facilitate the mapping of breakpoints I will develop a custom-made fosmid library approach taking advantage of chromosome sorting technology to enrich the library for derivative chromosome material. Once established, the benefits of a library approach can be compared

to methods such as FISH, PCR and microarrays, including commercially available tools.

1.5.3 Bioinformatic analysis of the genomic architecture surrounding genomic rearrangement breakpoints

The generation of sequence across constitutional rearrangement breakpoints and analysis of its composition may reveal clues as to the mechanism behind the rearrangement. Genome browsers and web-based tools will be used to search for structures known to be involved in other rearrangements and to search for motifs common to the breakpoints and compare this data with the analysis performed on the published constitutional translocation breakpoints listed in Table 1.1.

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

In this project I aim to utilise currently available tools and develop and apply new methods for the investigation of translocation breakpoints to the investigation into viral integration sites (Figure 1.12). I plan to characterise the integration site in 3 patients believed to have integrated Human Herpes Virus-6 genomes. I will investigate the chromosomal location of the viral integration using FISH and whether the complete virus is integrated using comparative genomic hybridisation microarray analysis. By generating a custom-made fosmid library and identifying clones that contain the integration sites, information about the sequence at the integration sites can be obtained, perhaps giving an indication as to the mechanism by which the virus integrates.

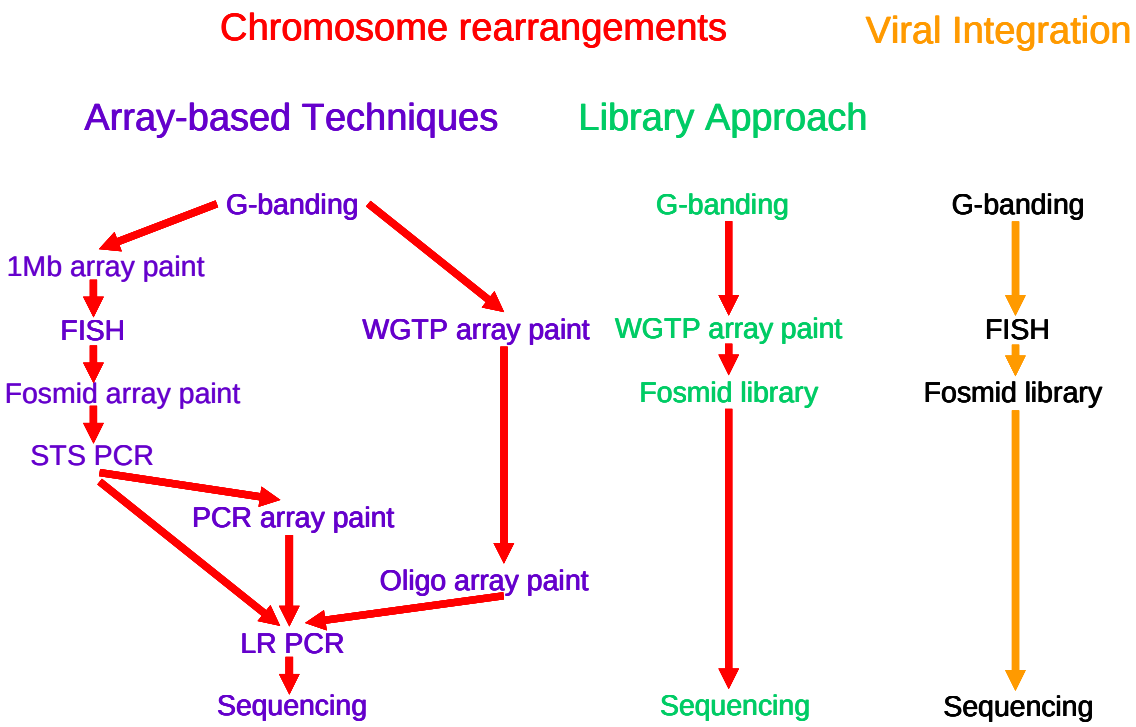

Figure 1.12 Summary of the proposed techniques for the analysis of rearrangement breakpoints and viral integration sites in the human genome.