Fusion genes in breast cancer

Elizabeth M. Batty

Clare College, University of Cambridge

A dissertation submitted to the University of Cambridge in candidature for the degree of Doctor of Philosophy

November 2010

Declaration

This dissertation contains the results of experimental work carried out between October 2006 and October 2010 in the Department of Pathology, University of Cambridge. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been submitted whole or in part for any other qualification at any other University.

Summary

Fusion genes in breast cancer

Elizabeth Batty

Fusion genes caused by chromosomal rearrangements are a common and important feature in haematological malignancies, but have until recently been seen as unimportant in epithelial cancers. The discovery of recurrent fusion genes in prostate and lung cancer suggests that fusion genes may play an important role in epithelial carcinogenesis, and that they have been previously under-reported due to the difficulties of cytogenetic analysis of solid tumours. In particular, breast cancers often have complex, highly rearranged karyotypes which have proved difficult to analyse using classical cytogenetic techniques.

The aim of this project was to search for fusion genes in breast cancer by using high-resolution mapping of chromosome rearrangements in breast cancer cell lines. Mapping the chromosome rearrangements was initially done using high-resolution DNA microarrays and fluorescence in-situ hybridisation, but moved to high-throughput sequencing as it became available. Interesting candidate genes identified from the mapped chromosome rearrangements were investigated on a larger set of cell lines and primary tumours.

The complete karyotypes of two breast cancer cell lines were constructed using a combination of microarrays, fluorescence microscopy, and high-throughput sequencing. A number of potential fusion genes were identified in these two cell lines. Although no expressed fusion genes were found, the complete karyotypes gave insight into the number and mechanisms of chromosome rearrangement in breast cancer, and identified interesting candidate genes which may be of importance in tumourigenesis. Two genes which were fused in other breast cancer cell lines, *BCAS3* and *ODZ4,* were disrupted by chromosome rearrangements and identified as interesting candidate genes in tumorigenesis.

A bioinformatic pipeline to process high-throughput sequencing data was set up and validated, and shown to more accurately predict fusion genes than other methods, and can be used to investigate further cell lines and tumours for recurrent fusion genes. The pipeline was used to analyse data from 3 other breast cancer cell lines and predict chromosomal rearrangements and fusion genes, several of which were found to be expressed. Of the fusions predicted in the cell line ZR-75-30, 7 expressed fusion genes were identified, and may have functional significance in breast cancer.

Acknowledgements

I would like to thank all my colleagues and collaborators, especially Karen and Susie for all their support in the lab, and Carole for advising me. I acknowledge the contributions of Karen Howarth, Suet-Feung Chin, Ina Schulte, Jess Pole, Susanne Flach, Scott Newman, and Kevin Howe to the work in this thesis.

I am grateful to Clare College, the Medical Research Council, and especially Breast Cancer Campaign for their financial support.

Thanks to my family and friends, especially my parents for their continued support, and my housemates for their continual provision of coffee.

My final thanks go to my supervisor, Paul, for providing support, inspiration and enthusiasm for my project for four years, and for giving me the chance to work in his lab.

Contents

Declaration ii

Summary iii

Acknowledgements iv

List of Figures ix

List of Tables xii

Abbreviations xiii

Chapter 1 – Introduction

1.1 Introduction 2
1.2 Genes and pathways altered in cancer 2
1.2.1 How many genetic changes are needed to progress to cancer? 4
1.2.2 The role of genomic instability 8
1.3 Breast cancer 9
1.3.1 Genes commonly altered in breast cancer 9
1.3.2 Breast cancer susceptibility genes 10
1.4 Classification of breast cancer 11
1.4.1 Breast cell lineages 12
1.5 Cytogenetics of breast cancer 13
1.5.1 Mechanisms of chromosome amplification 17
1.6 Chromosome translocations 20
1.7 Research techniques 25
1.7.1 Karyotyping 25
1.7.2 Fluorescence in-situ hybridisation 25
1.7.3 Chromosome painting and spectral karyotyping 26
1.7.4 Comparative genomic hybridisation 27
1.7.5 Array comparative genomic hybridisation 27
1.7.6 Sequencing 28
1.7.7 Cell lines 29
1.8 Hypothesis 30

Chapter 2 – Materials and methods

2.1 Cell culture 32
2.1.1 Sources 32

2.1.2 Culturing cells 32
2.2 RNA extraction 33
2.3 Protein extraction 33
2.4 DNA extraction 34
2.5 cDNA synthesis 34
2.6 PCR 34
2.7 Real-time PCR 35
2.8 Sequencing 35
2.9 Flow sorting of chromosomes 35
2.9.1 Genomiphi amplification of sorted chromosomes 36
2.10 Metaphase preparations from cell lines 37
2.10.1 Metaphase spreads 37
2.11 BAC growth and DNA extraction 38
2.12 FISH probes and hybridisation 38
2.12.1 Nick translation 38
2.12.2 Hybridisation 39
2.12.3 Detection 39
2.13 Array painting 40
2.13.1 Labelling 40
2.13.2 Hybridisation 40
2.13.3 Washing 41
2.13.4 Scanning 41
2.14 Western blotting 41
2.14.1 Blotting 41
2.14.2 Detection 42
2.15 High-throughput sequencing 42

Chapter 3 – Rearrangements of *BCAS3* in breast cancer

3.1 Introduction 44
3.2 Results 45
3.2.1 High-resolution array painting 45
3.2.2 FISH mapping of the derivative chromosome 51
3.2.3 Fine mapping and sequencing of the breakpoint 55
3.2.4 *BCAS3* in other cell lines and tumours 58
3.3 Discussion 63

Chapter 4 – The complete karyotype of HCC1806

4.1 Introduction 66
4.1.1 Previous work 67
4.2 Results 69
4.2.1 High-resolution breakpoints from SNP6.0 arrays 69
4.2.2 Determining the breakpoints 70

4.2.3 Comparison of array painting and the SNP6.0 arrays 71
4.2.4 Identification of previous undetectable copy number changes 75
4.2.5 Assembly of the complete karyotype 76
4.2.6 Discrepancies between array painting and whole genome SNP6.0 array 77
4.2.7 Systematic search for fusion genes 84
4.2.8 New fusion genes identified by high-resolution arrays 87
4.2.9 Fusion genes caused by small deletions 89
4.2.10 Fusion genes caused by tandem duplications 93
4.3 Discussion 98

Chapter 5 – The complete karyotype of MDA-MB-134 obtained using array painting and high-throughput sequencing

5.1 Introduction 101
5.1.1 Previous work 102
5.2 Results 103
5.2.1 Assembling a complete karyotype of MDA-MB-134 103
5.2.2 Array painting of rearranged chromosomes 110
5.2.3 High-throughput sequencing 122
5.2.4 Paired-end read high-throughput sequencing of MDA-MB-134 125
5.2.5 Detection of structural variants 125
5.2.6 Validation of structural variants 127
5.2.7 Sequencing of structural variant junctions 133
5.2.8 Potential fusion genes found by high-throughput sequencing 137
5.2.9 *ODZ4* as a potential fusion gene 150
5.3 Discussion 154

Chapter 6 – Bioinformatics of high-throughput sequencing of breast cancer

6.1 Introduction 159
6.2 Alignment of high-throughput sequencing reads 162
6.2.1 Calling normal and aberrant reads 164
6.2.2 Processing mate-pair data 166
6.3 Clustering and structural variant calling 168
6.4 Fusion gene prediction 174
6.4.1 Validation of predicted fusion genes 178
6.5 Copy number variation 181
6.5.1 Correcting for mappability 181
6.5.2 Correcting for GC content 181
6.5.3 Segmentation and copy number analysis 186
6.6 Discussion 188

Chapter 7 – Discussion

7.1 How prevalent are fusion genes in breast cancer? 191
7.2 How important are fusion genes in breast cancer? 192
7.3 How can fusion genes in breast cancer be found? 195
7.4 Mechanisms of chromosome rearrangement in breast cancer 197
7.5 Future directions 198
7.5.1 *ODZ4* 198
7.5.2 High-throughput sequencing and bioinformatic analysis 199
7.6 Conclusion 203

References 204

Appendix 1 - Primers used 223

Appendix 2 – BACs used 232

Appendix 3 – Manufacturers and suppliers 233

Appendix 4 – Bioinformatic pipeline scripts 235

Appendix 5 – Documentation for bioinformatic pipeline 254

List of figures

Chapter 1 – Introduction

Figure 1.1 Breakage-fusion-bridge cycles 19

Chapter 3 – Rearrangements of *BCAS3* in breast cancer

Figure 3.1 Array painting 47
Figure 3.2 Position of the chromosome 17 48
Figure 3.3 Breaks in *BCAS3* in MCF7 and HCC1806 48
Figure 3.4 Position of the deletion on chromosome 7 50
Figure 3.5 Position of the breakpoint on chromosome 8 50
Figure 3.6 Possible orientations of chromosome fragments 51
Figure 3.7 FISH for orientation of chromosome 7 fragment 52
Figure 3.8 FISH for possible inversion on chromosome 7 54
Figure 3.9 Tiling path array of chromosome 7 56
Figure 3.10 Junction sequence for the 7;17 junction 58
Figure 3.11 Final orientation of chromosome fragments in the der(7)t(8;7;17) 58
Figure 3.12 FISH for *BCAS3* on a SUM52 metaphase spread 59
Figure 3.13 RT-PCR on *BCAS3* exons in cell liens 60
Figure 3.14 Location of BAC probes used for TMA FISH 61
Figure 3.15 Western blot of *BCAS3* 62

Chapter 4 – The complete karyotype of HCC1806

Figure 4.1 SKY karyotype of HCC1806 66
Figure 4.2 Flow karyotype of HCC1806 chromosomes 68
Figure 4.3 Segmentation of SNP6.0 arrays by circular binary segmentation 71
Figure 4.4 Breakpoint duplication from SNP6.0 arrays 73
Figure 4.5 Related derivative chromosome in HCC1806 75
Figure 4.6 A deletion identified from the SNP6.0 array 76
Figure 4.7 Array painting of chromosome 2 in fraction A compared to SNP6.0 array 79
Figure 4.8 Array painting of chromosome 2 in fraction G compared to SNP6.0 array 81
Figure 4.9 FISH to confirm chromosome 2 amplification in fraction G 82
Figure 4.10 FISH to confirm chromosome 2 breakpoint in fraction G 83
Figure 4.11 Fusion gene caused by translocation 85
Figure 4.12 Readthrough fusion gene caused by translocation 85
Figure 4.13 Primer design for amplification of fusion products 86
Figure 4.14 PCR for fusion genes caused by translocations in HCC1806 88
Figure 4.15 Fusion gene caused by intrachromosomal deletion 90
Figure 4.16 PCR for fusion genes caused by small deletions 91
Figure 4.17 Fusion genes caused by tandem duplications 1 94
Figure 4.18 Fusion genes caused by tandem duplications 2 95

Figure 4.19 Fusion genes caused by tandem duplications 3 96
Figure 4.20 PCR for fusion genes caused by tandem duplications 97

Chapter 5 – The complete karyotype of MDA-MB-134 obtained using array painting and high-throughput sequencing

Figure 5.1 SKY karyotype of MDA-MB-134 101
Figure 5.2 Schematic of chromosome 8 and 11 amplification 103
Figure 5.3 Flow karyotype of MDA-MB-134 chromosomes 105
Figure 5.4 Reverse chromosome painting of chromosome fractions A and C 107
Figure 5.5 Reverse chromosome painting of chromosome fraction 14 109
Figure 5.6 Array painting of MDA-MB-134 fraction C 111
Figure 5.7 Array painting of MDA-MB-134 fraction 14 112
Figure 5.8 FISH to show orientation of der(15)t(15;17) 114
Figure 5.9 FISH to show orientation of der(16)t(16;18) 115
Figure 5.10 Array painting of MDA-MB-134 fraction A 117
Figure 5.11 Comparison of array painting and whole genome SNP6.0 data 119
Figure 5.12 Position of amplifications on chromosome 8 120
Figure 5.13 Position of deletions on chromosome 8 121
Figure 5.14 Position of amplifications on chromosome 11 122
Figure 5.15 Illumina high-throughput sequencing 124
Figure 5.16 FISH to confirm chromosome X translocation 132
Figure 5.17 Circos plot of structural variants 135
Figure 5.18 Circos plot of structural variants on chromosomes 8 and 11 136
Figure 5.19 Structure of the 8;11 amplicon 137
Figure 5.20 Fusion gene PCR strategy 141
Figure 5.21 PCR to detect fusion transcripts 1 142
Figure 5.22 PCR to detect fusion transcripts 2 144
Figure 5.23 PCR to detect readthrough fusion transcripts 145
Figure 5.24 The two predicted *KLHL35* fusions 147
Figure 5.25 PCR to detect internal gene deletions 149
Figure 5.26 *ODZ4* breakpoints in breast cancer cell lines 151
Figure 5.27 PCR for *ODZ4* in cell lines 153
Figure 5.28 Real-time PCR for *ODZ4* in cell lines 154
Figure 5.29 Comparison of segmented copy number 156

Chapter 6 – Bioinformatics of high-throughput sequencing of breast cancer

Figure 6.1 High-throughput sequencing pipeline steps 161
Figure 6.2 Small-insert sequencing libraries 165
Figure 6.3 Construction of mate-pair libraries 167
Figure 6.4 Clustering of reads to call structural variants 1 169
Figure 6.5 Clustering of reads to call structural variants 2 169
Figure 6.6 Structural variant calling from small-insert libraries 171
Figure 6.7 Structural variant calling from mate-pair library 172
Figure 6.8 Fragment size distribution from a mate-pair library 174

Figure 6.9 Fusion gene prediction from small-insert library 176
Figure 6.10 Fusion gene prediction from mate-pair library 177
Figure 6.11. Reads in a window versus GC percentage of the window 183
Figure 6.12 GC bias in copy number plots 184
Figure 6.13 Correction of copy number data for GC bias 185
Figure 6.14 Segmented copy number plot showing small region of amplification 187

Chapter 7 – Discussion

Figure 7.1 Artefactual versus real structural variants 201
Figure 7.2 Improvements to structural variant calling 202

List of tables

Chapter 2 – Materials and methods

Table 2.1 Origin of cell lines 32
Table 2.2 Cell line growth conditions 33
Table 2.3 Extended chromosome preparations 37

Chapter 4 – The complete karyotype of HCC1806

Table 4.1 Potential fusion genes caused by translocations and large deletions 87
Table 4.2 Potential fusion genes caused by small deletions 90
Table 4.3 Potential fusion genes cause by tandem duplications 93

Chapter 5 – The complete karyotype of MDA-MB-134 obtained using array painting and high-throughput sequencing

Table 5.1 Categories of structural variants in MDA-MB-134 126
Table 5.2 Predicted structural variants in MDA-MB-134 128
Table 5.3 Validated structural variants in MDA-MB-134 131
Table 5.4 Exact breakpoints and homology of structural variants in MDA-MB-134 134
Table 5.5 Predicted gene fusions from high-throughput sequencing 139
Table 5.6 Predicted readthrough gene fusions from high-throughput sequencing 140
Table 5.7 Predicted internal gene deletions from high-throughtput sequencing 148

Chapter 6 – Bioinformatics of high-throughput sequencing of breast cancer

Table 6.1 Predicted fusion genes in the cell line ZR-75-30 179
Table 6.2 Predicted fusion genes in the paired cell lines VP229 and VP267 180

Chapter 7 – Discussion

Table 7.1 Fusion genes currently known in breast cancer 194

Abbreviations

API	application programming interface
ATCC	American Type Culture Collection
BAC	bacterial artificial chromosome
BSA	bovine serum albumin
CGH	comparative genomic hybridisation
DAPI	4'6-diamidino-2-penylindole
DMEM	Dulbecco's Modified Eagle medium
DMSO	dimethyl sulphoxide
DOP-PCR	degenerate oligonucleotide polymerase chain reaction
FBS	foetal bovine serum
FISH	fluorescence in-situ hybridisation
ITS	insulin-transferrin-selenium supplement
LB	Luria Bertani
MAQ	Mapping and Assembly with Qualities
M-FISH	multiplex fluorescence in-situ hybridisation
MMTV	mouse mammary tumour virus
PBS	phosphate buffered saline
PMT	photomultiplier tube
RPMI	Roswell Park Memorial Institute
SKY	Spectral karyotyping
SSC	sodium chloride sodium citrate
SST	sodium chloride sodium citrate 0.05% Tween 20
SV	structural variant
TE	tris-EDTA

Chapter 1

Introduction

1.1 Introduction

Cancer is caused by the accumulation of genetic changes in genes which control cell death and proliferation, but the number of changes which are necessary to progress to malignancy, which genes or pathways they affect, and the different mechanisms of genetic change are a subject of much debate.

1.2 Genes and pathways altered in cancer

Hanahan and Weinberg (2000) described six key processes which must be deregulated in the cell for progression to malignancy, and suggest that the large numbers of genes implicated in cancer represent different ways to evade the anti-cancer defences of the cell. As I am primarily interested in breast cancer, I have looked at how breast tumours may exhibit genetic changes which contribute to the deregulation of these six processes.

The first process which must be overcome is the dependence on external growth factors to signal the cell to proliferate. This can be overcome by the cell generating its own growth factors, or by altering the requirements of the growth factor receptors and their downstream pathways. An example of this process in breast cancer is the amplification and overexpression of the *ERBB2* receptor in breast cancer (Slamon et al., 1987), which may act by making the cell hypersensitive to small amounts of growth factors.

The cell must also ignore the antiproliferative signals which attempt to block cell proliferation. At the G1 – S phase transistion of the cell cycle, the cell decides whether to continue to proliferate, or whether to stop dividing and become quiescent or differentiate. The *RB* tumour suppressor gene is important in the control of this transition, and changes in the *RB* pathway in breast cancer may remove the block on cell proliferation. *RB* expression is lost in 20-35% of breast tumours (Bosco and Knudsen, 2007). In tumours which retain *RB* expression, it may be inactivated by phosphorylation by cyclin/CDK complexes, and the amplification of *CCND1* in breast cancer may contribute to aberrant phosphorylation. Estrogen also upregulates the promoter of *CCND1*, and anti-estrogenic therapies may act by inihibiting cell cycle progression (Foster et al., 2001).

Tumour cells must also evade apotosis, either by deregulating the machinery which senses the signals which trigger apoptosis, or by turning off the pathways which respond to these signals and cause the cell to die. *TP53* is a sensor of DNA damage, and upregulates other pro-apoptotic genes. In breast cancer, the *TP53* gene is one of the two most commonly mutated genes, and its downstream targets are also commonly mutated (Pharoah et al., 1999).

The cell may also have to evade the signals which limit their multiplicative potential and become immortal. As the telomeres of chromosomes become shorter with each cell division, telomere length acts as a break on unlimited replication, as the telomeres become lost and the ends of the chromosomes fuse, usually leading to cell death. Tumours often evade this check by upregulating the expression of telomerase - in breast cancer, telomerase activity was found in over 90% of breast tumours (Hiyama et al., 1996).

For a tumour to progress and grow to greater size, it must recruit new blood vessels to supply oxygen and nutrients to the tumour, and it must deregulate the mechanisms which control angiogenesis and vasculogenesis in the cell, by upregulating the inducers of angiogenesis or downregulating the suppressors. Whether this is achieved by alteration of the genes involved in angiogenesis pathways, or whether angiogenesis is upregulated in the tumour as part of the natural response to hypoxia is unclear, and molecules which regulate angiogenesis are produced not only by the cancerous cell but by the normal cells surrounding the tumour (Carmeliet and Jain, 2000). Regardless of whether the mechanism is genetic or a part of normal homeostatic processes, pro-angiogenic genes are a common drug target in cancer (Banerjee et al., 2007). In breast cancer, the amplification and overexpression of *ERBB2* can increase angiogenesis and expression of *VEGF* (Kumar and Yarmand-Bagheri, 2001), and *VEGF* inhibitors are a target of antiangiogenic agents in breast cancer (Banerjee et al., 2007).

The final barrier to tumour progression is to acquire the capability for invasion and metastasis. The genetic changes which lead to invasion and metastasis are not well known, and one explanation for the difficulty in understanding these changes is that genetic changes which specifically lead to invasion and metastasis do not exist. Bernards and Weinberg (2002) argue that a change which leads to metastasis, unlike the changes which lead to tumour growth and immortality, does not confer an advantage to the primary tumour and would remain rare. This implies that the changes which enable metastasis are already present in the tumour, and confer some early selective advantage as well as the ability to metastasize later in tumour progression. This is supported by evidence from gene expression profiling, which shows that breast primary tumours and their distant metastases show similar expression patterns, suggesting that the dominant clone in the primary tumour may have already acquired the capability for metastasis long before it occurs (Weigelt et al., 2003). A further argument suggests that metastasis occurs by the chance event of a malignant cell escaping into the vasculature and finding a site suitable for growth, without any additional genetic changes (Edwards, 2002). However, this argument remains controversial, and further research is needed to give a definitive answer.

1.2.1 How many genetic changes are needed to progress to cancer?

It was noted as early as 1957 that tumours progressed through different stages to metastasis in a stepwise manner (Foulds, 1957), and that this was consistent with a tumour which emerged from a single cell and progressed by acquiring more genetic changes. Although early modelling of the age distribution of cancer suggested that four to twelve mutational events would be necessary to cause cancer (Armitage and Doll, 1954), this assumed that the mutational events were independent. Later work suggests that a better model is a two-stage theory of carcinogenesis, where the first stage gives the cell a selective advantage, and makes it more likely to accumulate the mutation or mutations necessary to progress to a second stage (Armitage and Doll, 1957). A detailed analysis of this two-stage model suggests that three rate-determining events were needed for cancer to arise (Stein and Stein, 1990).

This model also suggests that while most cancers follow this two-stage pattern, some cancers, such as retinoblastoma, have single-stage kinetics. This was explained by work on the occurrence of retinoblastoma by Knudson, which suggested that a “two-hit” model was in effect, and both copies of the *RB* gene must be lost for retinoblastoma to occur. A predisposition to retinoblastoma was due to an inherited mutation which knocked out one copy of the gene, requiring only a somatic mutation in the other copy to cause cancer rather than two mutations in the same cell (Knudson, 1971). The germline mutation of *RB* serves as the first event, and a further somatic mutation allows the tumour to progress to the second stage.

Further evidence for a multi-step model of carcinogenesis comes from studies of colorectal carcinoma by Fearon and Vogelstein (1990). They used the model system of colorectal tumours, which were known to progress from benign adenomas to carcinomas to metastatic disease, to suggest that tumours began with a mutation in a single cell, which acquired more mutations as the disease progressed, and that while the genetic changes often occurred in a similar order in different tumours it was the overall number of mutations which was important for progression. Their early estimate, based on evidence from colorectal cancer progression as well as mathematic models of cancer progression, was that three to seven hits were necessary for malignancy. This figure was based on a few known mutations such as *KRAS* and *TP53*, and low-resolution data which could only identify loss of heterozygosity over whole chromosome arms as a putative mechanism for deletion of a particular tumour suppressor gene (Vogelstein and Kinzler, 1993).

Finding the mutations important for tumour initiation and progression is more difficult due to the presence of non-functional “passenger” mutations which occur by chance and are carried through successive rounds of clonal expansion. This problem is especially prevalent in genome-wide studies which do not look at specific candidate genes found by functional studies or linkage analysis. Sjöblom et al. (2006) found that the number of coding mutations in a series of breast tumours and cell lines is higher than the background rate of mutation would suggest, and statistical methods are needed to distinguish important “driver” mutations from passengers by finding genes which are found mutated in a higher proportion of tumours than would be expected by chance. Using this method, Sjöblom et al. predict that up to 20 of the somatic mutations found in breast tumours may be driving mutations, with another 80 mutations which are passengers, a much higher figure than previously reported, although these calculations rely on an accurate assessment of the background rate of mutation (Forrest and Cavet, 2007), and may suffer from a high false discovery rate (Getz et al., 2007). A study of mutations in a greater number of tumour types but looking only at protein kinase mutations showed a wide variation in the mutation rates between tumour types, suggesting that a background rate of mutation would be difficult to estimate (Greenman et al., 2007). Greenman et al. do not predict a number of driver mutations per tumour, but estimate that 119 of their 518 sequenced genes contain a driving mutation, leading them to a similar conclusion to Sjöblom et al. - the number of driver mutations is larger than previously estimated. Mathematical modelling supports the experimental evidence and suggests that this large number of driving mutations indicates that while there are certain common pathways which are mutated and give a large selective advantage, such as *TP53* and *APC*, the majority of driving mutations will confer only a small selective advantage, and that the stochastic nature of these mutation contributes to the heterogeneity of cancer (Beerenwinkel et al., 2007).

Recent studies of complete cancer genomes are consistent with the idea that there are many driving mutations. The earliest complete breast cancer sequence was of a metastasis and reported 32 non-synonymous mutations in coding sequences, none of which were in genes reported as candidate cancer genes by Sjöblom et al (Shah et al., 2009). 11 of the 32 mutations could be found in the primary tumour, and 6 of these mutations were present at low levels in the primary tumour, suggesting heterogeneity of somatic mutations in the primary tumour. The only complete coding sequences of both a breast primary tumour and metastasis which is so far complete reports 50 mutations in coding regions, with a ratio of synonymous to non-synonymous mutations similar to that which would be expected by chance, suggesting that the majority of coding mutations are not strongly selected for and are not driving mutations (Ding et al., 2010). The complete sequence of a melanoma cell line (Pleasance et al., 2010a) found 292 coding mutations, with a similar lack of selection for non-synonymous mutations, and over 33,000 mutations in non-coding regions of the genome, and similar figures were obtained for a small-cell lung cancer with 134 coding mutations and over 22,000 non-coding mutations (Pleasance et al., 2010b).

Although studies have focused on the number of mutations needed to progress to cancer, genes can be altered by other mechanisms such as copy-number alteration. A study of copy-number changes focusing only on major copy-number changes (defined as deletion of all copies, or amplification to >11 copies) showed on average 17 genes altered by major copy number changes per tumour (Leary et al., 2008). Stephens et al. (2009) generated the most comprehensive study of somatic rearrangements in 24 breast cancer samples and found large differences in the number and type of rearrangements present, from a tumour with only a single rearrangement to tumours with hundreds of tandem duplications present in the genome. The rearrangements were enriched for those affecting genes, although it is not clear whether this is due to the rearrangements being selected for, or due to the mechanism of genome rearrangements favouring coding regions.

These studies are beginning to uncover the number and depth of the changes present in cancer genomes, but the complete picture is still not clear. The most comprehensive study of rearrangements in breast cancer yet published estimates they detected only

50% of the changes present in each sample (Stephens et al., 2009), and studies of somatic mutation may miss small indels, which were difficult to detect with the alignment methods used in these studies (Li and Durbin, 2009). To observe the overall picture of the number of changes needed for progression will require an integrated analysis combining mutation, copy-number alteration, identification of fusion genes and epigenetic changes to identify key pathways altered in cancer (Teschendorff and Caldas, 2009).

1.2.2 The role of genomic instability

Cells must accumulate a number of genetic changes in order to progress to cancer; the question of whether these changes can arise given the normal human mutation rate or whether there must be an underlying genetic instability to explain the number of mutations is still subject to debate. It has been suggested that genomic instability may not be a requirement for tumour development, but a secondary effect of mutations whose primary effect is to protect against apoptosis (Bodmer, 2008) or carcinogens (Bardelli et al., 2001). Even if genomic instability is not a requirement, carcinogenesis may proceed more quickly when the genome is unstable, especially if the number of changes required for progression is large and the genomic instability arises early (Beckman and Loeb, 2006).

Tumours demonstrate a number of mechanisms of genomic instability, which affect the genome at different levels, ranging from single nucleotide changes to rearrangement of chromosome and the gain and loss of chromosome arms and whole chromosomes. Patterns of genomic instability were first seen in colorectal tumours, where a small proportion of tumours with a near-normal karyotype display microsatellite instability, due to defects in genes in the mismatch repair pathways. Other colorectal tumours have an aneuploid karyotype and show loss and gain of whole chromosomes as well as loss of heterozygosity (Lengauer et al., 1997).

Many breast tumours display genomic instability. 47% of breast tumours have aneuploid karyotypes (Teixeira et al., 2002), while *BRCA1* and *BRCA2* can suppress genome instability, and *BRCA1* and *BRCA2*-deficient cells exhibit chromosomal instability (Venkitaraman, 2002).

1.3 Breast cancer

Breast cancer exhibits considerable molecular heterogeneity, with many different genes associated with the disease and few recurring mutations, and unlike other common epithelial tumours, no single pathway has emerged as the dominant pathway in breast cancer tumorigenesis.

1.3.1 Genes commonly altered in breast cancer

Studies of somatic mutations in breast cancer support the model that there are few commonly mutated genes, and many genes which are mutated much less frequently. Two genes stand out as often mutated in breast cancer across all subtypes: *TP53* and *PIK3C*A.

PIK3CA has been reported to be mutated in 8-40% of breast tumours, and may be a relatively early event in tumorigenesis (Miron et al., 2010). The mutations cluster around exon 9 and exon 20, and result in increased kinase activity (Samuels et al., 2005). Wood et al. (2007), in a screen of coding mutations in 20,000 genes, found mutations in a number of genes in pathways involved in *PIK3CA* signalling. These mutations are often mutually exclusive, suggesting that only one mutation is needed to disrupt the pathway sufficiently to drive tumourigenesis (Velculescu, 2008).

TP53 is a tumour suppressor gene mutated in 20-40% of breast tumours (Pharoah et al., 1999). It plays an important role in the cellular response to stress, and acts by inducing cell cycle arrest and apoptosis. Most cancers have lost *TP53* activity by point mutation, with few deletions or frameshifts (Vousden and Lu, 2002). The mutant forms of *TP53* are often more stable than the wild-type and found at high levels in the cell, and may act as dominant-negative inhibitors when they form complexes with the wild-type protein. Tumours with high expression of *HER2* and accumulation of *TP53* have considerably decreased overall survival (Yamashita et al., 2004). Similarly to *PIK3CA*, even in *TP53* wild-type tumours, regulators and targets of *TP53* are often mutated – *MDM2*, which stabilises *TP53* and is downregulated in response to stress, is amplified in up to 6% of breast tumours (Al-Kuraya et al., 2004). Tumours with mutations in the breast cancer susceptibility genes *BRCA1* and *BRCA2* are more likely to have *TP53* mutations (Greenblatt et al., 2001), and show a different spectrum of mutations than in sporadic cancers, suggesting that the inactivation of particular functions of *TP53* may be important in *BRCA*-deficient tumours (Venkitaraman, 2002).

1.3.2 Breast cancer susceptibility genes

Known breast cancer susceptibility alleles can be divided into three classes, based on the penetrance of the alleles (Turnbull and Rahman, 2008). *BRCA1* and *BRCA2* are high-penetrance genes, and carriers of a mutant allele have a greater than tenfold increased risk of breast cancer. *BRCA1* and *BRCA2* are involved in the DNA damage response, and many of the disease-associated mutations result in loss of function (Gudmundsdottir and Ashworth, 2006). Between them these two genes represent 15-20% of the excess familial risk. *TP53* is also mutated in Li-Fraumeni syndrome, which gives a high risk of developing several types of cancer, but the number of families with Li-Fraumeni Syndrome is rare and account for only a small part of the increased familial breast cancer risk.

Four alleles are known which give a 2-4X relative risk of breast cancer and are classed as intermediate-penetrance alleles. All four genes (*CHEK2* (Meijers-Heijboer et al., 2002), *BRIP1* (Seal et al., 2006)*, ATM* (Renwick et al., 2006) and *PALB2* (Rahman et al., 2007)) are involved in the DNA damage response, and have roles in the same pathways as *BRCA1* and *BRCA2.* Eight low-penetrance variants that give a relative risk of <1.5 are currently known from genome-wide associate studies (Easton et al., 2007; Cox et al., 2007; Stacey et al., 2007). Most of these variants do not lie within protein-coding genes, and it is not known how they cause increased breast cancer risk.

1.4 Classification of breast cancer

Breast cancer appears to be a heterogenous disease which shows wide variation in gene expression, point mutations and structural variation. Tumours can be classified based on histopathological grade, immunohistochemical staining, and lately gene expression profiling, which groups tumours into subtypes based on gene expression levels, and may distinguish between histologically similar tumours which are molecularly different (Rouzier et al., 2005). Gene expression profiling suggests that the different subtypes of breast cancer vary widely, harbouring different gene alterations and responding differently to therapy, and the different subtypes may even be distinct diseases (Herschkowitz et al., 2007).

Sørlie et al. (2001) carried out gene expression profiling and used the results to cluster tumours. This approach grouped tumours into two classes largely based on ER status, and each class was divided into three subtypes. Of the ER negative tumours, the ERBB2+ subgroup shows high expression of *ERBB2* and other genes present in the *ERBB2* amplicon, while the basal subgroup expresses basal-type cytokeratins, laminins and fatty acid binding proteins, and the normal-like subgroup shows high expression of basal epithelial genes and low expression of luminal epithelial genes. The ER positive/luminal tumours were split into at least two subgroups, with luminal A tumours showing the highest expression of the ER-related genes, and the luminal B subtype could be further separated into luminal B and luminal C by expression in the luminal C group of a set of genes of unknown function but which are also highly expressed in the basal and ERBB2+ subtypes. The basal and ERBB2+ subtypes were also correlated with poor prognosis and mutations in *TP53*.

Subsequent studies have replicated some of the initial classification, and further molecular classifications have been suggested. The divide between ER positive and ER negative tumours is consistent and the two groups have distinct gene expression profiles (Gruvberger et al., 2001). The luminal A and luminal B subgroups have been found to have differences in proliferation, histological grade and prognosis, with luminal B having a poorer prognosis (Weigelt et al., 2010), but the initial separation of luminal B into two further subgroups is not always repeated in subsequent studies. This suggests that the distinction between the different luminal subgroups is less clear than the divide between other subgroups, and the luminal group represents a continuum of gene expression which can be arbitrarily divided into different subgroups (Wirapati et al., 2008). In the ER negative category, the basal-like and ERBB2+ classes are highly reproducible (Rouzier et al., 2005), but the normal-like category may be an artefact of high normal tissue contamination of tumours (Parker et al., 2009). Other groupings of ER negative tumours have been suggested, such as the 'claudin-low' subtype, which shows low expression of the genes involved in cell-cell adhesion (Herschkowitz et al., 2007), and a subtype showing low genomic instability, discovered using integrated gene expression and copy number profiling (Chin et al., 2007).

1.4.1 Breast cell lineages

A question underlying the classification of breast cancers is whether the different subtypes reflect a difference in the cell types which give rise to them, or whether the subtypes are independent of the cell of origin.

The human mammary epithelium probably contains two general lineages, the luminal cells and the myoepithelial cells (Stingl and Caldas, 2007). A population of undifferentiated basally-positioned cells may represent the mammary gland stem cell, and gives rise to progenitor cells which may be multilineage, or produce luminal or myoepithelial cells only. These stem and progenitor cells are thought to be important as the initial cells which give rise to tumours, as any mutation in the progenitor will be passed to the daughter cells, which will acquire further mutations through subsequent rounds of cell division (Cairns, 2002).

Some evidence for tumours arising from different progenitor cells has been found. Cell lines with a luminal gene expression pattern show no cells with basal characteristics, suggesting that a luminal progenitor cell gave rise to the tumour (Stingl and Caldas, 2007). Tumours induced using the same combination of oncogenes gave rise to tumours with different phenotypes, suggesting the cell type of the precursor influences the type of tumour produced (Ince et al., 2007).

Although the exact number of breast cancer subtypes varies between the approaches, it is clear that a number of different subtypes exist, with different gene expression and patterns of chromosomal rearrangement, and few genetic changes have been found which are common to all subtypes. Whether this is due to a difference in the cell of origin, or whether the tumours originate from the same cell type but follow a different mutational path is not yet clear, but there is some evidence to suggest that breast cancer is not one disease but a set of heterogenous diseases which arise from the same tissue, and further research into the development of the mammary gland may help to determine which of the two possibilities is correct.

1.5 Cytogenetics of breast cancer

Classical cytogenetic analysis of breast cancer is difficult due to the technical difficulties of obtaining good karyotypes. The results are often dependent on the culture techniques used (Teixeira et al., 2002), may be biased towards those malignant tumours that divide better in culture. The karyotypes produced are often complex and difficult to interpret. A further difficulty arises from the heterogeneity of individual tumours, as analysis of a single sample may not give an accurate picture of the tumour karyotype.

The studies of the cytogenetics of breast tumours which have attempted to overcome these technical difficulties show very heterogenous karyotypes within breast cancers, ranging from near-diploid with few chromosome alterations, to tumours with complex highly-rearranged karyotypes (Teixeira et al., 2002). Among the near-diploid tumours, certain rearrangements were often present as the sole chromosome aberration, such as deletion of 3p13-14, which may delete the candidate tumour suppressor gene *FHIT*, and a der(1;16)(p10;q10) rearranged chromosome.

While the karyotype of an individual tumour provides only information on the state of the karyotype at that time and not the evolutionary history of the tumour, studies of large numbers of tumours and the different clones within a tumour provide an overall picture of the karyotypic evolution of breast cancer. By looking at the number of chromosomes and the number of rearrangements across a series of breast cancers, Dutrillaux et al*.* (1991) suggest that chromosomes are lost early on, due to whole chromosome loss and unbalanced rearrangements, followed by endoreduplication of the whole genome and further chromosome loss and rearrangement. The presence of hyperploid sidelines in near-diploid tumours supports this pathway, as does the trend towards increasing number of rearrangements as chromosome number decreases, confirmed by Texiera et al. (2002). Although many tumours follow this pathway, it is not the only way for a breast cancer karyotype to evolve, as seen by the presence of near-diploid tumours which have not follow the pathway of chromosome loss and endoreduplication.

Higher-resolution array CGH studies have validated the results from earlier cytogenetic studies. Regions of the genome which are commonly gained, lost, and amplified in breast cancer can be defined at higher resolution than chromosome banding provides, and a number of recurrent amplicons have been found, including regions on 8p12, 11p13, and 17q21, as well as regions of frequent low copy number gain or loss (Fridlyand et al., 2006).

A set of copy number subtypes have been defined according to patterns and frequency of copy number change, although different studies have produced slightly different subtypes. One subtype includes those with the simplest karyotype of gain of 1q and loss of 16q, which occur in ER positive tumours with low histological grade, and were seen by Fridyland et al. (2006). A second subtype includes low-level gains and losses with occasional peaks of amplification, and are seen by both Fridyland et al. (2006) and Hicks et al. (2006), who found 60% of tumours fall into this "simplex" subtype. Tumours with complex karyotypes with frequent gains and losses, in which few regions of the genome are present at normal copy number, were found in both studies, and were termed "sawtooth" tumours by Hicks et al. (2006). The tumours tend to be ER negative and have a significantly worse outcome than tumours in the other subtypes (Fridlyand et al., 2006). Hicks et al. also identify a fourth group of tumours displaying a "firestorm" pattern of amplification, with clustered, narrow peaks of high-level amplification in a relatively simple karyotype. 11q and 17q are among the regions most often found in "firestorm" amplifications.

Much research has focused on finding the genes which drive amplifications in breast cancer. A well-studied example is the amplification of 17q12, which leads to overexpression of the *ERBB2* gene, which is correlated with poor prognosis, and has been successfully targeted for chemotherapy (Järvinen and Liu, 2003). The targets of other amplicons are less clear.

Amplification of 17q23 is found in up to 20% of breast tumours (Bärlund et al., 2000) and is associated with poor prognosis. While 17q23 is gained in a number of different tumour types, high-level amplification is only seen in breast tumours (Andersen et al., 2002). The consensus region of amplification covers over 5Mb and a number of genes have been implicated as the drivers of amplification, including *APPBP2*, *RAD51C, THRAP1,* and *PPM1D* (Bärlund et al., 2000; Monni et al., 2001; Lambros et al., 2010), mainly by correlation of mRNA overexpression with copy number gain. It is possible that no single gene is the driver of amplification, but rather that a number of genes at the core of the amplicon are the target of amplification (Parssinen et al., 2007), and that high-level amplification is required for significant overexpression. An alternate hypothesis looks for the minimal region of amplification on high resolution arrays and suggests a minimal region of 250Kb centred around the microRNA *mir-21* (Haverty et al., 2008).

8p11-12 is another region of common amplification in breast cancer, found in 10-25% of tumours (Garcia et al., 2005). Early studies suggested *FGFR1* as a candidate to be the driving oncogene in this region (Ugolini et al., 1999), but while *FGFR1* is overexpressed at the mRNA level when amplified, additional FGFR1 protein is not seen in cell lines with amplification, and inhibition of *FGFR1* does not slow the growth of cell lines (Ray et al., 2004). Refinement of the minimal region of amplification using higher resolution array CGH suggests that *FGFR1* is outside the minimal region, and suggests a 1.5Mb region of minimal amplification with *ZNF703*, *ERLIN2*, *BRF2* and *RAB11FIP1* as candidate driving oncogenes (Garcia et al., 2005). Other studies have suggested that rather than one simple amplicon, the 8p11-12 region contains four separate regions of amplification, one of which overlaps with the 1.5Mb region of Garcia et al. (Gelsi-Boyer et al., 2005), which is contradicted by the findings of Haverty et al. (2008) who used high-resolution Affymetrix arrays to narrow the region of minimal amplification to 400Kb containing, among other genes, *BRF2*, *RAB11FIP1*, and *ZNF703*.

11q13 is found amplified in around 19% of breast cancers, but it is rarely found amplified alone, and is often found co-amplified with other commonly amplified regions such as 8p12 and 17p12 (Letessier et al., 2006). *CCND1* is the best supported candidate driving oncogene, as it is within the most frequently amplified region, and *CCND1* overexpression promotes mammary tumours in mice (Wang et al., 1994).

Co-amplification of different regions in the same tumour suggests that genes in different amplicons may collaborate to drive tumourigenesis. *FGFR1*/*CCND1* co-amplification results in poorer prognosis than when the genes are amplified separately, as does *ERBB2*/*MYC* co-amplification (Cuny et al., 2000). The amplified regions are often physically associated and arranged in complex structures (Paterson et al., 2007). Although no correlation between genes amplified on 8p and 11q has been found, expression of *CCND1* on 11q13 may induce expression of *ZNF703* on 8p12 (Kwek et al., 2009). Another hypothesis is that a translocation between chromosomes 8 and 11 is an early event which is then amplified, and that a fusion gene at the translocation junction may be the driving event (Paterson et al., 2007), but as yet this hypothesized fusion gene has not been found, and the amplicons are often physically separated.

1.5.1 Mechanisms of chromosome amplification

A number of mechanisms have been proposed to explain how chromosome amplification occurs. If the amplification is at a distant site to the original gene, the proposed mechanism involves duplication of the gene and excision of the duplicated copy, which replicates extrachromosomally and reintegrates into the DNA at a different site (Schwab, 1999). A common example of this method of replication is the *MYCN* locus in neuroblastoma, where double minute chromosomes containing multiple copies of the *MYCN* locus can be seen in tumours. The amplified copies of *MYCN* in cell lines are more often found as homogenously staining regions, which are more common in cell lines than tumours (Benner et al., 1991), but the site of integration of the *MYCN* amplification is never at the locus on chromosome 2 where *MYCN* is normally found (Schwab et al., 2003; Storlazzi et al., 2010).

For amplifications where the extra material resides where the single-copy gene would normally be found, different mechanisms of amplification has been proposed, of which the breakage-fusion-bridge cycle is the best known (Schwab, 1999). The initiating event is a double chromatid break which is repaired to form a fusion (Figure 1.1). During anaphase this forms a bridge between sister chromatids, which must be broken to allow cell division to continue. If the break is not in the same place as the original break and fusion occurred, the daughter products will have either a duplication or a deletion. Further rounds of this breakage-fusion-bridge cycle will result in further amplification of material. The signature of this mechanism is that the amplified genes are inverted (Schwab, 1999).

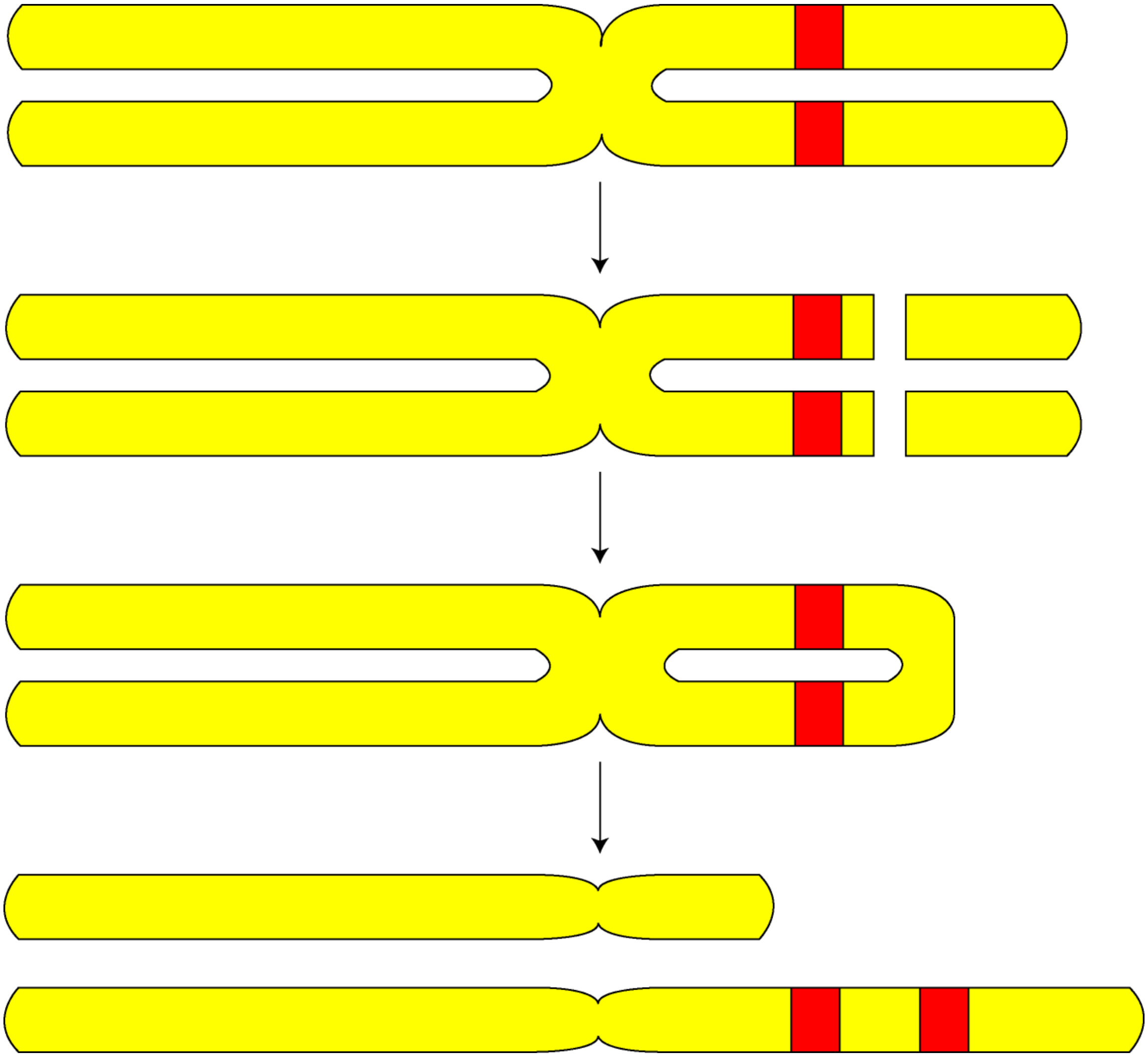


Figure 1.1. Breakage-fusion-bridge cycles as a mechanism of oncogene amplification. A break occurs in both chromatids, which fuse and resolve unequally to give two possible daughter products, one with a deletion and one with a duplication. The red gene will be duplicated and one of the copies will be inverted.

Another mechanism for chromosome rearrangement is the replication fork stalling and template switching (FoSTeS) model proposed by Lee et al. (2007). This occurs when the lagging single strand of DNA during replication forms a secondary structure, blocking the progress of the replication fork, which then switches to another template with microhomology. Depending on the position of the other replication fork, this can cause a duplication, but does not explain high-level amplifications, unlike the breakage-fusion-bridge cycle which will continue until the chromosome acquires a telomere (Hastings et al., 2009).

1.6 Chromosome translocations

Chromosome aberrations have been seen in cancer for many years. Abnormal segregation of chromosomes was seen by Boveri in the in the early 1900s and proposed to be a cause of malignancy, and in the 1950s it was shown that nearly all tumour cell lines had chromosome aberrations (Rowley, 2001) but these chromosome abnormalities were assumed to be a result of chromosome instability as the events did not appear to recur in different tumours from the same origin, and not an important event in their own right.

The first recurrent chromosome abnormality in a human cancer was found in 1960, with the discovery of the Philadelphia chromosome in chronic myeloid leukaemia. The nature of the translocation was not discovered until chromosome banding techniques showed it was a reciprocal translocation of chromosome 22 to chromosome 9 , which produced a fusion of the *BCR* locus on chromosome 22 to the *ABL* tyrosine kinase on chromosome 22 (Shtivelman et al., 1985). This creates a fusion of the two genes which leads to mRNA and protein containing domains from both *BCR* and *ABL*, under the control of the *BCR* promoter. The fused protein product was shown to have tyrosine kinase activity and to induce leukaemia when expressed in mouse bone-marrow cells (Daley et al., 1990). Imatinib, a specific inhibitor of the *BCR-ABL* fusion, is used to treat leukaemia patients

(Deininger et al., 2005). The *BCR-ABL* fusion is also found in acute lymphoblastic leukaemia, with a different breakpoint which includes less of the BCR protein in the fused product (Hermans et al., 1987), with even greater tyrosine kinase activity than in CML. The success of the *BCR-ABL* fusion as an effective therapeutic target linked to a specific cancer led to a search for other chromosomal aberrations which could produce similar specific therapeutic targets.

Although the *BCR-ABL* fusion was the first to be seen cytogenetically, the genes involved in another recurrent translocation in cancer were discovered first. The karyotypes of cells taken from Burkitt's lymphoma showed an extra band on chromsome 14, while one was missing from the end of chromosome 8 (Zech et al., 1976). When the oncogene *MYC* was located to chromosome 8, it was shown that the coding region of *MYC* was juxtaposed with the promoter region of the immunoglobulin heavy chain (*IGH*) (Dalla-Favera et al., 1982; Taub et al., 1982). This does not create a direct fusion of the two genes as is the case for *BCR-ABL*, but changes the expression levels and pattern of *MYC*, which leads to tumorigenesis (ar-Rushdi et al., 1983). Translocations causing fusions between oncogenes and members of the immunoglobulin family are a hallmark of B-cell lymphomas, and have been discovered in mantle cell lymphoma (*CCND1-IGH*) and follicular lymphoma (*BCL2-IGH*) at high frequency, and at lower frequencies in other lymphomas (Kuppers, 2005).

Subsequently, hundreds of other recurrent (present in 1% or more cases) gene fusions of both types have been discovered in common haematological malignancies (Mitelman et al., 2007), although fusions which produce a fusion protein are more common than promoter insertions (Rowley, 2001).

Until recently, fusion genes caused by recurrent chromosome aberration were thought to be a feature of haematological and soft tissue cancers, but not of solid tumours, where other mechanisms such as deletion and point mutation were thought to be more important, and recurrent chromosomal aberrations were rarely seen. This may be due to tissue-specific mechanisms causing chromosome rearrangements, such as recombination in haematopoetic progenitor cells which then give rise to leukaemias (Albertson et al., 2003), but there is some evidence that this stems from the difficulties of performing cytogenetic analysis on epithelial tumours rather than from a lack of recurrent gene fusions. Additionally, the prevalence of recurrent rearrangements in haematological malignancies may have been overestimated due to selection bias for patients reported in the literature due to a cytogenetic abnormality (Mitelman et al., 2005). The actual proportion of malignancies with recurrent rearrangements may be as high in epithelial tumours as in haematological malignancies, but represent a large number of rare rearrangements without the common rearrangements seen in leukaemias (Mitelman et al., 2004).

There are several technical difficulties which make it more difficult to find fusion genes in solid tumours. Karyotyping solid tumours is more difficult due to poor chromosome morphology, and the karyotypes are often so complex they cannot be characterized completely (Mitelman et al., 2004). Obtaining metaphases is difficult as the carcinoma cells may not divide, and contaminating normal cells or minor clones may grow better than the dominant clone (Persson et al., 1999). Further evidence that the lack of reported fusion genes was a technical artefact and not a difference between the two types of tumour is that the fusion genes which were reported in rare epithelial cancers often included the same genes involved in haematological fusions. The *ETV6-NTRK3* fusion found in secretory breast cancer (Tognon et al., 2002) is also seen in congenital fibrosarcoma (Knezevich et al., 1998) and acute myeloid leukaemia (Eguchi et al., 1999).

Some fusion genes were known in rare epithelial cancers. Fusions of *RET* are found in papillary thyroid cancer, with the most common fusion being caused by inversions of chromosome 10 which fuse the 3’ end of *RET* to 5’ portion of *H4* although there are a number of other 5’ partners. Fusions of *NTRK1* to a number of partners (Alberti et al., 2003) and an *AKAP9-BRAF* fusion produced by an intrachromosomal inversion have also been reported (Ciampi et al., 2005). The *AKAP9-BRAF* fusion was found more often in radiation-induced cancers while sporadic thyroid carcinoma often carried a *BRAF* point mutation, and *RET* fusions are also more common in radiation-induced cancers, suggesting the mechanisms of gene activation are linked to environmental factors. A fusion of *BRAF* and *KIAA1549*, which shows constitutive kinase activity, has also been found in 66% of pilocytic astrocytomas, a common paediatric brain tumour (Jones et al., 2008).

The first recurrent fusion to be discovered in a common epithelial cancer was the fusion of *TMPRSS2* to members of the *ETS* transcription factor family in prostate cancer (Tomlins et al., 2005). Previously, fusion genes had been found using cytogenetics or by transfection assays, but this fusion was discovered using a bioinformatic approach, working on the assumption that a gene fusion should result in overexpression of an oncogene in a subset of cases, and that the two genes involved in the fusion would both be overexpressed. Applying this Cancer Outlier Profile Analysis to prostate cancer gave two strong candidates, *ERG* and *ETV1*, and overexpression of these two genes was mutually exclusive, suggesting they played a similar role in prostate cancer development. 5' RACE on *ERG* and *ETV1* transcripts showed fusions to the prostate-specific androgen-sensitive gene *TMPRSS2*. *TMPRSS2-ETV4* fusions have also been found (Tomlins et al., 2006). Fusions of *TMPRSS2* and members of the *ETS* family have been found in up to 80% of prostate cancers (Tomlins et al., 2007). Although *TMPRSS2* fusions appeared to be driving the majority of *ERG* overexpression, many prostate cancers had *ETV1* overexpression without a fusion to *TMPRSS2*, and *ETV1* was found fused to a number of different 5’ partners, including the housekeeping gene *HNRPA2B1*, the androgen-induced gene *SLC45A3*, and the androgen-repressed gene *c15orf21* (Tomlins et al., 2007). These 5’ partners do not contribute coding sequence to the fusion, but place *ETV1* under the control of promoter and enhancer elements of distant genes.

Prostate cancer fusions demonstrate another reason why fusion genes may have been difficult to find in common epithelial cancers. The *BCR-ABL* fusion is atypical in being present in the majority of cases of CML (Mitelman et al., 2005), and fusions in other cancers are often found at lower frequency. Additionally, fusion genes which involve the same gene fused to a range of partners are well known, such as the fusions of *EWS* to multiple members of the *ETS* family in Ewing's sarcoma (Arvand and Denny, 2001), but the *ETV1* fusions in prostate involve a number of 5' partners from different gene families, including both androgen-induced and androgen-repressed genes, prostate-specific partners and ubiquitously expressed genes, and looking for fusions which involve such a wide range of partners with no commonality could be more difficult than looking for fusions which involve genes from the same family or pathway.

A fusion between *EML4* and *ALK* was subsequently discovered in around 10% of non-small cell lung cancer by searching a retroviral cDNA library for inserts which would transform mouse fibroblast cells (Soda et al., 2007). *ALK* was already known to form fusions with *NPM* in anaplastic lymphoma (Morris et al., 1994), and the kinase domain is retained in both fusions. A *KIF5B-ALK* fusion has also been found in NSCLC and shows transforming potential (Takeuchi et al., 2009). The *EML4-ALK* fusion was also found by a study of phosphorylation of tyrosine kinases in NSCLC, which also found a fusion of *SLC34A2* to *ROS* (Rikova et al., 2007).

Gene fusions have been discovered in breast cancer cell lines and tumours but so far none have been shown to be recurrent. The cell line MDA-MB-175 has a fusion of *ODZ4* to *NRG1* (Liu et al., 1999), and *FHIT* has a fusion to *MACROD2* in BrCa-MZ-02 (Popovici et al., 2002), although this is associated with a lack of FHIT protein rather than a fusion product. Howarth et al. (2008) found two fusion genes in a study of three cell lines, *TAX1BP1-AHCY*, and *RIF1-PKD1L1.* Recent high-throughput sequencing studies have found a number of other fusions – Hampton et al. (2009) found four expressed fusion genes in MCF7, including the previously discovered *BCAS4-BCAS3* fusion (Bärlund et al., 2002), and suggest that the fusions may be suppressing wild-type expression of the genes by dominant-negative effects. Stephens et al. (2009) found 21 expressed fusion genes in a study of 24 breast cancer cell lines and tumours, none of which were recurrent.

Although the prevailing view has previously been that fusion genes in epithelial cancers are rare and unimportant, this is being increasingly challenged by the number of fusion genes which are being identified in common tumours. It is likely that the fusion genes in epithelial cancers are not like the common, recurrent fusions found in leukaemia, but will involve individually rarer fusions which act on genes in the same pathway, or fusions where the fusion partners differ but all have the same effect on the important gene in the fusion, and to find these rarer fusions will require large-scale studies of cancer genetics which can only be achieved through high-resolution microarray and sequence analysis.

1.7 Research techniques

1.7.1 Karyotyping

Early karyotype analysis was hampered by an inability to accurately distinguish different chromosomes. The discovery that treatment with trypsin and staining with Giemsa produced a banding pattern which allowed all human chromosomes to be identified enabled the detection of structural aberrations such as translocations, deletions, and inversions. However, even high-resolution G-banding could only provide a resolution of ~3Mb at best, with ~6Mb being more usual, and any aberrations which did not have a clear banding pattern were impossible to identify (Smeets, 2004).

1.7.2 Flurorescence in-situ hybridisation

Fluorescence in-situ hybridisation (FISH) is a method of visualising the location of DNA using a fluorescently-labelled probe which binds to the target DNA. The probe consists of genomic DNA which hybridizes to the target region of the genome. The DNA is either directly labelled with a fluorophore, or a reporter molecule such as biotin is incorporated into the DNA and fluorescently-labelled antibodies are used to visualize it. It was developed as an alternative to the visualization of nucleic acids by radiolabelled probes, as fluorescent labelling offers better resolution and are safer to use, and multiple fluorophores can be used to visualize more than one sequence at a time (Levsky and Singer, 2003). FISH was first performed in 1982 (Van Prooijen-Knegt et al., 1982), with a probe hybridized to metaphase chromosomes.

Metaphase FISH has a resolution of around 3Mb (Raap, 1998). An advantage of FISH over traditional cytogenetics is that it can be performed on interphase nuclei, with a resolution of up to 100kb, and fibre-FISH using DNA fibres attached to a slide can resolve probes down to 1kb apart (Ersfeld, 2004). This allows FISH to be used to resolve even small-scale genomic rearrangements.

1.7.3 Chromosome painting and spectral karyotyping

The development of chromosome flow sorting allowed whole human chromosomes to be amplified, labelled and used as FISH probes. Spectral karyotyping (Schröck et al., 1996) is a technique which uses combinations of different fluorophores to label all 24 human chromosomes. The combination of fluorophores present at each pixel of the image is measured using an interferometer and used to classify each chromosome. SKY and the similar M-FISH technique can easily identify the chromosomes involved in translocations, including small pieces of chromosomes and homogenously staining regions which cannot be identified by G-banding. However, the resolution of SKY is around ~10Mb and translocations smaller than this cannot be identified, and small chromosome pieces can be misidentified due to overlap of fluoroscence. As SKY is based on chromosome painting, internal deletions, duplications, and inversions cannot be detected.

1.7.4 Comparative genomic hybridisation

Comparative genomic hybridization (Kallioniemi et al., 1992) is another technique for using fluorescently-labelled DNA to determine tumour karyotypes. Tumour and normal reference DNA are labelled with two different fluorophores, and hybridized together to a normal metaphase spread. The amount of labelled DNA which binds to each locus is relative to the abundance of the locus in the two samples, and deletions and amplifications change the ratio of the two signals at each locus. By analysing the ratio of the two signals at each locus, the patterns of gain and loss along each chromosome can be plotted.

1.7.5 Array comparative genomic hybridisation

Array comparative genomic hybridisation improves the resolution of CGH by hybridizing the labelled reference and tumour DNA to a microarray of DNA probes and measuring the ratio for each probe in one experiment. Initial experiments used BAC clones (Pinkel et al., 1998), but later arrays have used smaller inserts from cosmids and fosmids, and modern array CGH uses short oligonucleotides, with the limit of resolution being determined by the spacing of the oligonucleotides on the array. Oligonucleotides can also be designed to avoid repeats, reducing the noise caused by hybridisation to repetitive regions (Beaudet and Belmont, 2008). The sensitivity of oligonucleotide hybridisation also allows them to be used for large-scale SNP calling. Oligonucleotides are designed specifically to hybridise to the different SNP alleles (Kennedy et al., 2003), and can be used to find areas of uniparental disomy and loss of heterozygosity. Bignell et al. (2004) demonstrated the use of arrays originally designed to detect SNPs to detect genotype and copy number at once. High-density commercial arrays such as the Affymetrix SNP6.0 array include up to 2 million probes for simultaneous genotyping and detection of copy number aberrations at high resolution.

1.7.6 Sequencing

A limitation of array-based techniques for mapping chromosome rearrangements is that the sequences which are juxtaposed at the breakpoints are not known. Sequencing-based approaches overcome this limitation by using a paired-end approach to sequence both sides of a breakpoint.

End-sequence profiling was used to map chromosome rearrangements by creating BAC libraries from a genome and sequencing from the ends of the BACs to identify rearrangements, as the end sequences of a BAC containing a chromosome rearrangement will align to the genome in the wrong position or orientation (Volik et al., 2003). Copy number can also be determined from the density of the end-sequences across the genome, although the resolution is determined by the size of the BAC fragments.

High-throughput paired-end sequencing uses the same principle as end-sequence profiling but the sequenced DNA fragments are much smaller and give a correspondingly higher resolution than end-sequence profiling (Campbell et al., 2008). At high levels of coverage, sequencing can also be used to identify point mutations and small insertions and deletions in the genome (Pleasance et al., 2010a); (Pleasance et al., 2010b).

Transcriptome sequencing can be used to find the consequences of chromosome rearrangement such as fusion genes or internal rearrangements by finding transcripts which align to two different genes, and may be produced by a genomic rearrangement. Transcriptome sequencing may find fusion genes which would not be detected by genome sequencing, as they are produced by read-through transcripts produced from neighbouring genes, such as the *SLC45A3-ELK4* fusion found in prostate cancer which has no detectable DNA rearrangement (Maher et al., 2009).

1.7.7 Cell lines

Cell lines derived from tumours are commonly used in the laboratory to overcome the difficulties with the use of primary tumours. Cell lines offer unlimited material for study, are free of stromal contamination, and can be replaced from fresh stocks if they become contaminated (Burdall et al., 2003). Common concerns about the use of cell lines include problems of genetic drift, the use of 'false' cell lines contaminated with other cell lines (MacLeod et al., 1999), and cell lines which are not from the supposed tissue of origin, such as MDA-MB-435, commonly thought to be a breast cancer cell line which is in fact derived from the M14 melanoma cell line (Rae et al., 2006). Furthermore, as breast cell lines are often derived from post-treatment metastases or pleural effusions, not primary breast tumours, they may not be representative of the disease as a whole but model primarily the later-stage aggressive disease.

A study of the HCC series of breast cancer cell lines showed excellent concordance between primary tumours and the cell lines established from them, including cell morphology, ploidy, expression of ER and PR, and loss of heterozygosity (Wistuba et al., 1998). There was also no correlation between the length of time in culture of the cell lines and the concordance with the primary tumour, and studies of colorectal and ovarian cancer cell lines have found that a stable karyotype is maintained over many generations (Roschke et al., 2002).

Comparisons between CGH on cancer cell lines and primary tumours showed that the chromosome gains and losses found in the cell lines are a good model for those found in real tumours (Greshock et al., 2007). Some specific rearrangements are more often found in cell lines, such as loss of chromosome 18 (Neve et al., 2006) and amplification of the *MYC* locus (Greshock et al., 2007), and the subset of breast cancer with simple 1q/16 rearrangements is under-represented in cell lines. In general, breast cancer cell lines recapitulate events commonly found in primary tumours, such as patterns of high-level amplification (Neve et al., 2006), and the pattern of chromosome loss followed by endoreduplication known to occur in many breast tumours (Dutrillaux et al., 1991) is recapitulated in breast cancer cell lines (Morris et al., 1997).

1.8 Hypothesis

The importance of fusion genes in leukaemias and lymphomas has been known for many years, but the importance and the prevalence of fusion genes in solid tumours has been underestimated due to the difficulty of finding them. Using breast cancer cell lines as a model, the aim of this project was to investigate chromosomal rearrangements and find any fusion genes which may occur, and to investigate the recurrence and importance of any fusion genes in other cell lines and tumours. This involved mapping all the chromosomal rearrangements in breast cancer cell lines using high-resolution techniques, which can be validated against our existing knowledge of the rearrangements, and use the resulting karyotypes to provide insight into the cytogenetics of breast cancer.

Chapter 2

Materials and methods

2.1 Cell culture

2.1.1 Sources

The origin of the cell lines used is given in Table 2.1.

Cell line	Supplier	Reference
MDA-MB-134	O'Hare	Cailleau et al., 1978
HCC1143	ATCC	Gazdar et al., 1998
HCC1806	ATCC	Gazdar et al., 1998
HCC2218	ATCC	Gazdar et al., 1998
VP229/VP267	McCallum	McCallum and Lowther, 1996
ZR-75-30	O'Hare	Engel et al., 1978
HB4a	O'Hare	Stamps et al., 1994

Table 2.1. Origin of cell lines. O'Hare: cell lines were a kind gift from Professor MJ O'Hare, (LICR/UCL Breast Cancer Laboratory, University College Medical School, London, UK).

HB4a is a cell line derived from normal human breast epithelium by immortalization with SV40 large T-antigen (Smeets et al., 1994) which has been shown by gene expression profiling to have similar gene expression to normal human breast epithelium (Git et al,. 2008).

2.1.2 Culturing cells

Ampoules of cells frozen in liquid nitrogen were thawed at 37°C, centrifuged to remove residual DMSO, and resuspended in warm culture medium in a 25cm^2 flask. Once the adherent cells were confluent they were washed with 2ml of Versene, then 2ml of Versene with trypsin (0.5mg/ml except for HCC1806 which was 1mg/ml) was added and incubated at 37°C for 2 – 5 minutes until cells detached. 2ml of media was added to the flask, and the cell suspension was centrifuged at 1600g for 3 minutes to pellet the cells. The cells were resuspended in the appropriate volume of media for the new flask (6ml for a T75, 12ml for a T150).

All cells were grown with 100U/ml penicillin and 100µg/ml streptomycin and cultured at 37°C with

5% CO_2, except MDA-MB-134 which was cultured in 7.5% CO_2.

Cell line	**Growth type**	**Medium**	**Additives**
MDA-MB-134	Adherent	50:50 DMEM-F12	15% FBS
HCC1143	Adherent	RPMI	10% FBS
HCC1806	Adherent	RPMI	10% FBS
HCC2218	Suspension	RPMI	10% FBS
VP229/VP267	Adherent	MCDB-201	2% FBS + 1% ITS
ZR-75-30	Adherent	50:50 DMEM-F12	10% FBS + 1% ITS
HB4a	Adherent	50:50 DMEM-F12	10% FBS

Table 2.2. Cell line growth conditions.

To freeze cells, the cells were pelleted as above, and resuspended in 1.5ml of media with 10% DMSO in 2ml cryotubes. Tubes were frozen slowly at -80°C and stored in liquid nitrogen.

2.2 RNA extraction

The media was changed 12 hours before harvesting cells at 70% confluence. For adherent cell lines, 7.5ml (for a T75) or 15ml (for a T150) Trizol reagent (Invitrogen) was added and the cells left at room temperature for at least 5 minutes before the cells were harvested using a cell scraper and transferred to a Falcon tube. For suspension cells, the cells were centrifuged at 1600g for 3 minutes and the supernatant removed, and the pellet resuspended in the appropriate volume of Trizol. 1.5ml of chloroform was added, and the cells were vortexed and centrifuged at 2000g at 4°C for 15 minutes. The top layer was retained and mixed with 4ml of isopropanol, and centrifuged at 2000g at 4°C for 15 minutes. The pellet was washed twice in 70% ethanol, and either stored under ethanol at -80°C, or resuspended in 200µl of RNase-free water for immediate use.

2.3 Protein extraction

Cells were trypsinised and the pellet washed with PBS, then lysed by adding 1ml of RIPA buffer (50mM Tris HCl (pH 8), 150 mM NaCl, 1% NP-40 (v/v), 0.5% sodium deoxycholate (w/v), 0.1% SDS (w/v), 0.5 mM EDTA, Complete Protease Inhibitor Cocktail (Roche, used according to the manufacturer's instructions)) and mixing well. Cells were placed on ice for 20 minutes, then centrifuged at 16000g at 4°C for 10 minutes, and the supernatant retained.

2.4 DNA extraction

Cells were trypsinised and 1ml of DNAzol reagent (Invitrogen) added, and the cells were lysed with a P1000 pipette. 0.5ml of 100% ethanol was added and mixed by inversion, and left at room temperature for 3 minutes. The precipitated DNA was spooled around a pipette tip and transferred to a clean tube, and washed twice with 1ml of 95% ethanol. The DNA was resuspended in 250µl of water and quantified on the NanoDrop spectrophotometer.

2.5 cDNA synthesis

The DNA-free kit (Ambion) was used to remove DNA contamination. 10µg of total RNA extracted as above was treated with 1µl of rDNase I and the rDNase was removed with the DNAse Removal Reagent . First strand cDNA synthesis was performed using the SuperScript III First-Strand Synthesis Kit (Invitrogen). 5µg of DNase-treated RNA was mixed with 50ng of random hexamer primers and 1µl of 10mM dNTPs and incubated at 65°C for 5 minutes, then cooled on ice. 2µl of 10X RT buffer, 4µl of 25mM $MgCl_2$, 2µl of 0.1MDTT, 40U RNaseIN (Promega) and 200U SuperScript III were added, incubated at 25°C for 10 minutes then 50°C for 50 minutes, and the reactions were stopped by incubating at 85°C for 5 minutes. The cDNA was stored at -20°C until needed.

2.6 PCR

Primers to amplify specific regions of genomic DNA or cDNA were designed using Ensembl (www.ensembl.org) and Primer3 (Rozen and Skaletsky, 2000) (http://frodo.wi.mit.edu/primer3/).

All standard PCR (for target regions under 2kb) was carried out using HotMaster Taq polymerase (VWR) with the following reaction mix: 2.5µl of 10X HotMaster buffer, 1µl of 10mM dNTPs, 1µl each of 100mM forward and reverse primer, 1U Taq, 1µl of 50ng/µl DNA in a 25µl volume. Cycling conditions were 95°C for 5 minutes, then 35 cycles of 95°C for 30 seconds, 60°C for 30 seconds, 72°C for 1 minute per kb of target, and finally 72°C for 10 minutes. Long range PCR for targets up to 10kb was performed using Elongase polymerase mix (Invitrogen) and a reaction mix containing 1µl 10mM dNTPs, 1µl each of 10µM forward and reverse primer, 2µl of 50ng/µl DNA, 1U of Elongase polymerase mix and 10µl of a mix of buffer A and B optimised for the best Mg^{2+} concentration , made up to 50µl. Cycling conditions were 94°C for 30 seconds, followed by 35 cycles of 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 1 minute per kb of target, and finally 68°C for 10 minutes. 10µl of product was visualised on a 0.8-2% agarose gel with 0.05% ethidium bromide.

2.7 Real-time PCR

Gene-specific primers were designed as above. The reaction was carried out in a 10µl volume containing 1X SybrGreen PCR Master Mix (Applied Biosystems), 1µl each of 2.5mM forward and reverse primer, and 1µl of 50ng/µl cDNA. Cycling was carried out using an ABI Prism 7900HT RT-PCR machine (Applied Biosystems) and the cycling conditions were 50°C for 2 minutes and 95°C for ten minutes, then 40 cycles of 95°C for 15 seconds, 60°C for one minute, and a final dissociation step of 95°C for 15 seconds and 60°C for 15 seconds. Primer pair efficiency was calculated using a standard curve from cDNA dilutions, and primers with an amplification efficiency of 1.8 or higher were used. In the experiments in Chapter 3, *GAPDH* was used as a control cDNA, and expression of the cDNA of interest was normalised to the value of *GAPDH* in each cell line. In the experiments in Chapter 5, 3 genes (*GAPDH, UBC,* and *RPL13a*) were used as control cDNAs, and expression of the cDNA of interest was normalised to the mean of all three genes.

2.8 Sequencing

PCR products under 1kb were purified using the QIAquick PCR Purification Kit (Qiagen) according to manufacturer's instructions and capillary sequencing of the products was performed by the DNA Sequencing Facility, Department of Biochemistry. Longer PCR products were first cloned in a pCR-XL-TOPO vector using the TOPO XL PCR Cloning Kit (Invitrogen). The plasmid DNA containing the insert was extracted using the HiSpeed Plasmid Midi-Prep Kit (Qiagen) and sequenced as above.

2.9 Flow sorting of chromosomes

Flow sorting of chromosomes was performed according to standard methods (Ng and Carter, 2006). The cells were subcultured 1:2 the day before sorting to synchronise the cells. Colcemid was added to a final concentration of 0.1µg/ml 6 hours before the cells were harvested. Adherent cells were harvested by banging the flask, and the medium removed to a 50ml Falcon tube. The cells were pelleted by centrifuging at 250g for 5 minutes, the pellet resuspended in 5ml of PBS, and incubated at room temperature for 10 minutes.

Cell swelling was monitored by mixing 10µl of cell suspension with 10µl of Turk's solution. The cell suspension was spun down at 250g for 5 minutes and resuspended in 1-3ml of polyamine isolation buffer. After incubation on ice for 10 minutes and vortexing for 20 seconds, a small sample of the preparation was stained with propidium iodide (5mg/ml) and observed under a fluorescence microscope to see whether the chromosomes had clumped together, and vortexed until the chromosomes were free. The chromosome suspension was transferred to a 15ml tube and centrifuged for 1 minute at 173g, and the supernatant transferred to a fresh 15ml tube. Hoechst 33258, $MgSO_4.7H_20$, and Chromomycin A3 were added to the suspension to a final concentration of 1µg/ml, 10mM and 80µg/ml respectively, and incubated at 4°C overnight. The next day the suspension was centrifuged for 2 minutes at 250g and the supernatant removed to a new tube. Sodium sulphite to a final concentration of 250mM was added one hour before the chromosomes were sorted. Aliquots of 500 or 2000 chromosomes were sorted on a MoFlo (Cytomation Bioinstruments) and analysed using Summit software (Beckman Coulter) to count the number of events in each chromosome fraction.

2.9.1 Genomiphi amplification of sorted chromosomes

Aliquots of sorted chromosomes (volume ~20µl) were precipitated by adding 0.5µl Pellet Paint co-precipitant (Merck), 1.5µl of 2.5M sodium acetate pH 5.5, and 50µl ethanol, incubating at -20°C overnight, and centrifuging at 16000g for 20 minutes at 4°C. The pellet was washed in 70% ethanol and air-dried, then resuspended in 1µl TE. Chromosomes were amplified using the GenomiPhi DNA Amplification Kit (GE Healthcare) according to the manufacturer's protocol and purified using MicroSpin G50 columns.

2.10 Metaphase preparation from cell lines

The cells were split the day before sorting to synchronise the cells. For standard metaphase preparations, colcemid was added to a final concentration of 0.1µg/ml 20 hours after the cells were split and incubated for 90 minutes. To produce extended chromosome preparations, cells were treated with BrdU, EtBr and colcemid:

Cell line	**BrdU (40µg/ml)**	**EtBr(5µg/ml)**	**Colcemid(0.1µg/ml)**
HCC1806	16.5 hours	1.5 hours	None
MDA-MB-134	20 hours	1.5 hours	0.75 hours

After incubation, cells were trypsinised and centrifuged at 1600g for 3 minutes to pellet the cells. The supernatant was removed, leaving ~500µl of medium on the pellet, and the cells were resuspended using a P1000 pipette. 20ml of 0.075M KCl warmed to 37°C was added drop by drop to the cells with agitation, and the cells incubated at 37°C for 15 minutes. 10-20 drops of ice cold freshly prepared 3:1 fix (3 parts methanol to 1 part acetic acid) were added, and the cells centrifuged at 1600g for 3 minutes. The supernatant was removed, again leaving ~500µl of medium on the pellet, and the cells were resuspended using a P1000 pipette to ensure no cell clumps were left. 20ml of 3:1 fix was added drop by drop with agitation, and the cells were incubated on ice for 5 minutes, then centrifuged at 1600g for 3 minutes. The supernatant was removed leaving ~500µl of fix on the pellet, and the cells were resuspended using a P1000 pipette. The fixation step was repeated once more with 3:1 fix, and then with 3:2 fix (3 parts methanol to 2 parts acetic acid). The ~500µl of metaphase suspension was transferred to a 2ml Eppendorf tube and made up to 2ml with 3:2 fix, and stored at -20°C for at least 24 hours before use.

2.10.1 Metaphase spreads

100µl of water was placed on a glass microscope slide, and 10-20µl of metaphase suspension was dropped onto the slides from a height of ~40cm. The slides were checked under a light microscope for the presence of metaphases and the area containing metaphases was marked with a diamond pen. The slides were dehydrated by incubating for 3 minutes each in 70, 90 and 100% ethanol at room temperature, and the slides allowed to air-dry before incubating at 37°C overnight. Slides were stored at -20°C until needed.

2.11 BAC growth and DNA extraction

Bacterial artificial chromosomes were stored as glycerol stabs at -80°C. They were streaked onto LB agar with 20µg/ml chloramphenicol or 25µg/ml kanamycin and grown overnight at 37°C. A single colony was grown for 6-8 hours at 37°C with shaking in 5ml LB media with 20µg/ml chloramphenicol or 25µg/ml kanamycin . 1ml of this colony was added to 100ml of LB/chloramphenicol or LB/kanamycin media and grown overnight at 37°C with shaking. The 100ml cultures were centrifuged at 3000g for 15 minutes and the BAC DNA was extracted using the HiSpeed Plasmid Midi-Prep Kit (Qiagen) according to the manufacturer's instructions. The DNA was precipitated by adding 1/10 volume of 3M sodium acetate pH 5.2 and 2 volumes of 100% ethanol before incubating overnight at 20°C. The DNA was centrifuged at 15000g for 30 minutes, and the pellet was air-dried and resuspended in 50µl of TE buffer and incubated at 65°C for two hours. The DNA concentration was determined on the NanoDrop spectrophotometer.

2.12 FISH probes and hybridisation

2.12.1 Nick translation

Input material was BAC DNA, Genomphi-amplified sorted chromosomes, or sorted chromosomes amplified by 3 rounds of DOP-PCR. Sorted normal human chromosomes were provided by Patricia O'Brien and Professor Malcolm Ferguson-Smith, Department of Veterinary Medicine, University of Cambridge. 500ng – 1µg DNA was nick translated in a 25µl reaction volume containing 2.5µl of nick translation buffer, 1.9µl of low-C dNTPs (0.1M dATP, dGTP, dTTP, 0.03M dCTP), 0.7µl of labelled dUTPs , 0.7µl of DNAse I and 1µl of DNA polymerase I. dUTPs were labelled with Digoxigenin-11, Spectrum Orange, or Biotin. The reaction was incubated at 14°C for 2 hours. 4µl of the reaction was run on a 1.5% agarose gel to check the product, and the reaction was stopped with 2.5µl of EDTA and incubated at 65°C for ten minutes.

2.12.2 Hybridisation

7µl of each probe or chromosome paint was precipated overnight at -20°C with 3µl of CoT-1 DNA, 1µl of glycogen and 300µl of ethanol. The probe mixture was centrifuged at 15000g for 30 minutes and the pellet allowed to air dry before being resuspended in 20µl of hybridization buffer (50% DI formamide, 10% dextran sulphate, 1X Denhardt's solution (Sigma), 2XSSC, 23mM Na_2HPO_4, 17mM NaH_2PO_4) and left at 37°C for 30 minutes. Finally, the mixture was incubated at 70°C for ten minutes, cooled on ice for 2 minutes, and incubated at 37°C for one hour.

Prepared metaphase spreads were incubated overnight at 37°C. The slides were denatured for 1 minute in denaturation solution (70% deionised formamide, 2XSSC) heated to 70°C and placed in ice-cold 70% ethanol for 5 minutes. The slides were washed in a series of 70, 90 and 100% ethanol for 3 minutes each and allowed to air-dry before incubation at 37°C for ten minutes.

18µl of the probe mixture was pipetted onto a slide and covered with a clean coverslip, then sealed with rubber cement. The slides were placed in a humid box and hybridized at 37°C overnight.

2.12.3 Detection

The rubber cement was removed from the slides with tweezers and the coverslips removed by soaking in 2xSSC. The slides were washed twice for 5 minutes each in a solution of 50% formamide and 1xSSC at 42°C and twice for 5 minutes each in a solution of 1xSSC at 42°C, then once for 5 minutes in a solution of 4xSST. The slides were blocked with 100µl 3% BSA in 4xSST for 30 minutes and washed briefly in 4xSST. Antibody layers were prepared by adding 1µl of antibody per slide to 200µl of 1% BSA in 4xSST, incubating in the dark for 10 minutes, and centrifuging for 10 minutes at 15000g. Digoxigenin labelled probes were detected with sheep FITC anti-digoxigenin. Diotin-labelled probes were detected with a layer of Cy5-labelled streptavidin, a layer of biotinylated anti-streptavidin (Vector Laboratories), and a final layer of Cy5-labelled streptavidin. Each antibody layer was incubated for 30 minutes at 37°C and the slides were washed 3 times in 4xSST with 0.5% BSA. After all antibody layers were complete, 20µl of Vectashield with DAPI (Vector Laboratories) was placed on a clean coverslip, and the slide is inverted onto the coverslip and allowed to dry in the dark before being sealed with nail varnish. The slides were analysed on a Nikon Eclipse E800 Fluorescence microscope using Cytovision software (Applied Imaging) and stored at 4°C while not in use.

2.13 Array painting

1Mb genomic arrays produced by the Cancer Research UK DNA Microarray Facility and tiling path arrays (a gift from Dr K. Ichimura, as described in Ichimura et al., 2006) were used for array painting. Genomiphi-amplified chromosomes were labelled with Cy3 and reference DNA (a pool of normal female DNA) was labelled with Cy5. Custom NimbleGen arrays used for high-resolution array painting were designed by Dr Karen Howarth and hybridised by Roche-NimbleGen. The whole-genome array CGH data was produced by Dr Graham Bignell and the Cancer Genome Project, Wellcome Trust Sanger Institute, using the human SNP6.0 array (Affymetrix).

2.13.1 Labelling

Labelling was carried out using a BioPrime Labelling Kit (Invitrogen). 450ng of Genomphi-amplified sorted chromosome DNA was mixed with 60µl 2.5X Random Primer Solution in a 150µl reaction. The DNA was heated to 100°C for ten minutes and cooled on ice, and 15µl 1X dNTPs, 1.5µl Cy3 or Cy5 labelled dCTP (Amersham) and 3µl exo-Klenow polymerase were added while on ice. The reaction was incubated at 37°C overnight and stopped with 15µl of stop buffer. The unincorporated nucleotides were removed with a Micro-Spin G50 cleanup column (Amersham) and the Cy3/Cy5 incorporation was measured on the NanoDrop.

2.13.2 Hybridisation

The Cy3 and Cy5 labelled DNA was mixed with 80µl human CoT1 DNA, 44µl 3M sodium acetate (ph 5.2), 6µl yeast tRNA and 1ml 100% ethanol, mixed well, and precipitated at -20°C overnight. A hybridisation chamber was prepared by adding a strip of Whatman paper soaked in 2xSSC/20% formamide solution. The precipitated DNA was spun for 15 minutes at 15000g to pellet the DNA, the pellet was washed with 500µl 80% ethanol, and the supernatant was removed. The dry pellet was resuspended in 50µl array hybridisation buffer (50% formamide, 10% dextran sulphate, 0.1% Tween 20, 2X SSC, 10mM Tris buffer pH 7.4) pre-heated to 70°C. The sample was denatured for 10 minutes at 70°C, then incubated for one hour at 37°C in the dark. The sample was pipetted onto the array slide and covered with a coverslip, then placed in the prepared hybridiation chamber at 37°C for 24 hours.

2.13.3 Washing

The slide was washed in PBS with 0.05% Tween 20 to remove the coverslip, then washed in a fresh solution of PBS/0.05% Tween 20 for 10 minutes at room temperature with shaking. The slide was transferred to 1X SSC/50% formamide pre-heated to 42°C and incubated for 30 minutes at 42°C with shaking, then washed in fresh PBS/0.05% Tween 20 for 10 minutes at room temperature with shaking. The slide was dried by centrifugation at 750g for 2 minutes, and stored in a dark box until ready to scan.

2.13.4 Scanning

The slides were scanned on an Axon 4000B scanner using GenePix Pro 6.0 software (Molecular Devices), using constant PMT gain settings of 1000 for the Cy5 channel and 800 for the Cy3 channel. The median signal (minus background) for each channel for each probe was used, and any probe where the signal in the Cy5 channel was not twice the signal from the Drosophila control probes was rejected. The $\log_2$ ratio of the test (Cy3) to the reference (Cy5) signal was plotted and used to call chromosome losses and gains.

2.14 Western blotting

2.14.1 Blotting

Total protein extracts prepared as above were mixed 1:1 with β-mercaptoethanol and denatured at 99°C for 2 minutes then cooled on ice. 10µl of each sample was loaded onto a 12% Tris-acetate gel (Invitrogen) and run at 125V for 90 minutes. The gel was trimmed and soaked in transfer buffer, and transferred onto a PVDF membrane for 2 hours at 250mA at 4°C. The membrane was blocked overnight in blocking buffer (1X TBS with 0.1% Tween 20 and 5% milk powder) at 4°C.

2.14.2 Detection

The membrane was washed 3 times for 5 minutes each in 1X TBS/0.1% Tween 20. Primary antibody (diluted 1:1000-1:10000 in blocking buffer) was added and incubated for at least 1 hour, and washed 3 times for 5 minutes each in 1X TBS/0.1% Tween 20. Secondary antibody (diluted 1:10000 in blocking buffer) was added and incubated for at least 1 hour at room temperature. Detection was performed using the ECL Plus Western Blotting Detection System (GE Healthcare) according to manufacturer's instructions.

2.15 High-throughput sequencing

DNA sequencing libraries were prepared by Drs Jessica Pole and Ina Schulte in the lab from genomic DNA extracted as above or from Genomiphi-amplified DNA using Paired-End DNA Sample Prep Kit or Mate-Pair Library Prep Kit (Illumina) according to the manufacturer's instructions. Sequencing was performed on an Illumina GAIIx sequencer at the Cancer Research UK Cambridge Research Institute.

Chapter 3

Rearrangements of *BCAS3* in breast cancer

3.1 Introduction

Fusion genes are well-known in haematological malignancies. Recurrent fusion genes was first discovered in chronic myeloid leukaemia (Shtivelman et al., 1985) and Burkitt's lymphoma) (Dalla-Favera et al., 1982; Taub et al., 1982), and fusion genes which lead to fusion proteins have subsequently been found in many other common haematological malignancies (Mitelman et al., 2007). Some recurrent fusion genes have been identified in solid tumours, such as the *TMPRSS2-ERG* fusion in prostate cancer (Tomlins et al., 2005), but at the start of my project, only 3 fusion genes had been found in breast cancer – a fusion of *ODZ4* to *NRG1* in the cell line MDA-MB-175 (Liu et al., 1999), a fusion of *FHIT* to a cDNA later identified as *MACROD2* in BrCa-MZ-02 (Popovici et al., 2002), and a fusion of *BCAS4* to *BCAS3* in MCF7 (Bärlund et al., 2002). All of these fusions were found in cell lines, and none were known to be recurrent. However, the recurrent fusions of *TMPRSS2* to members of the *ETS* transcription factor family had recently been shown in prostate cancer (Tomlins et al., 2005), suggesting that recurrent gene fusions are present in epithelial cancers, and that there may be recurrent fusions in breast cancer which had yet to be discovered.

Work by Dr Karen Howarth in the lab had mapped many of the chromosome rearrangements in three breast cancer cell lines to look for fusion genes. Two fusion genes had been found in the breast cancer cell line HCC1806, *RIF1-PKD1L1* and *TAX1BP1-AHCY* (Howarth et al., 2008). The SKY karyotype of this cell line includes a der(7)t(8;7;17) chromosome, and array painting of this derivative chromosome to a 1Mb array showed a break in *BCAS3* which was joined to part of chromosome 7. A break in *BCAS3* was interesting as it was already known to take part in a fusion in breast cancer, and a recurrent fusion would be one of the very few recurrent fusions in common epithelial cancers.

The *BCAS4-BCAS3* fusion was discovered by investigating chromosome amplification in the cell line MCF7. This line shows amplification of 17q23 and 20p13, which are two of the commonly amplified regions in breast cancer, with amplification of 17q23 found in around 20% of breast cancers, and 20p13 in between 12 and 39% of cancers (Bärlund et al., 2002). Expression analysis of the genes in the amplicon in MCF7 identified an expressed sequence tag as the most overexpressed transcript in this region (Monni et al., 2001).

Further investigation of this novel EST was performed, and showed that the full-length cDNA (now called *BCAS3,* for breast carcinoma amplified sequence 3), was not only overexpressed in MCF7 but fused to another novel gene on 20p13, *BCAS4* (Bärlund et al., 2002). The *BCAS4-BCAS3* fusion joins exon 1 of *BCAS4* to exons 23 and 24 of *BCAS3*, and alters the open reading frame of *BCAS3*, resulting in a truncated protein which ends after 21bp of *BCAS3* exon 23.

Other studies of *BCAS3* have suggested a possible functional role in breast carcinogenesis. High expression of *BCAS3* in breast cancer has been associated with tamoxifen resistance (Gururaj et al., 2006), and its expression is induced by estrogen receptor alpha (Gururaj et al., 2007).

As it was part of one of the few known fusions in breast cancer, I decided that the break in *BCAS3* was worthy of further investigation. To determine whether *BCAS3* was part of a fusion gene in HCC1806, I first needed to determine the other breakpoints on the derivative chromosome to high resolution to determine the potential fusion partner. I would then investigate any possible *BCAS3* fusion in HCC1806, and look for any signs of a recurrent break or fusion in other breast cancer cell lines and primary tumours.

3.2 Results

3.2.1 High-resolution array painting

At the start of my project, most of the breakpoints in HCC1806 were known only to low resolution based on a 1Mb BAC array, which had 3000 probes spaced at roughly 1Mb intervals. For most breakpoints, this was not high enough resolution to determine the exact gene at the breakpoint. To find the breakpoints at higher resolution and determine which genes were broken, the der(7)t(8;7;17) was hybridized to a custom Nimblegen array (array designed by Dr Karen Howarth). The Nimblegen array was designed around the breakpoints already known from 1Mb and tiling path array painting, and covers small regions at high resolution. Figure 3.1 shows the 1Mb array painting for the three chromosomes, and the high-resolution Nimblegen array results for the regions around the breakpoints. On chromosome 17, the breakpoint was between 56,164,500 and 56,172,000bp, which is within *BCAS3* as expected, between exons 4 and 5 (Figure 3.2). This retains the 3’ end of the gene, which is the same end of the gene as is retained in the known MCF7 fusion (Figure 3.3).

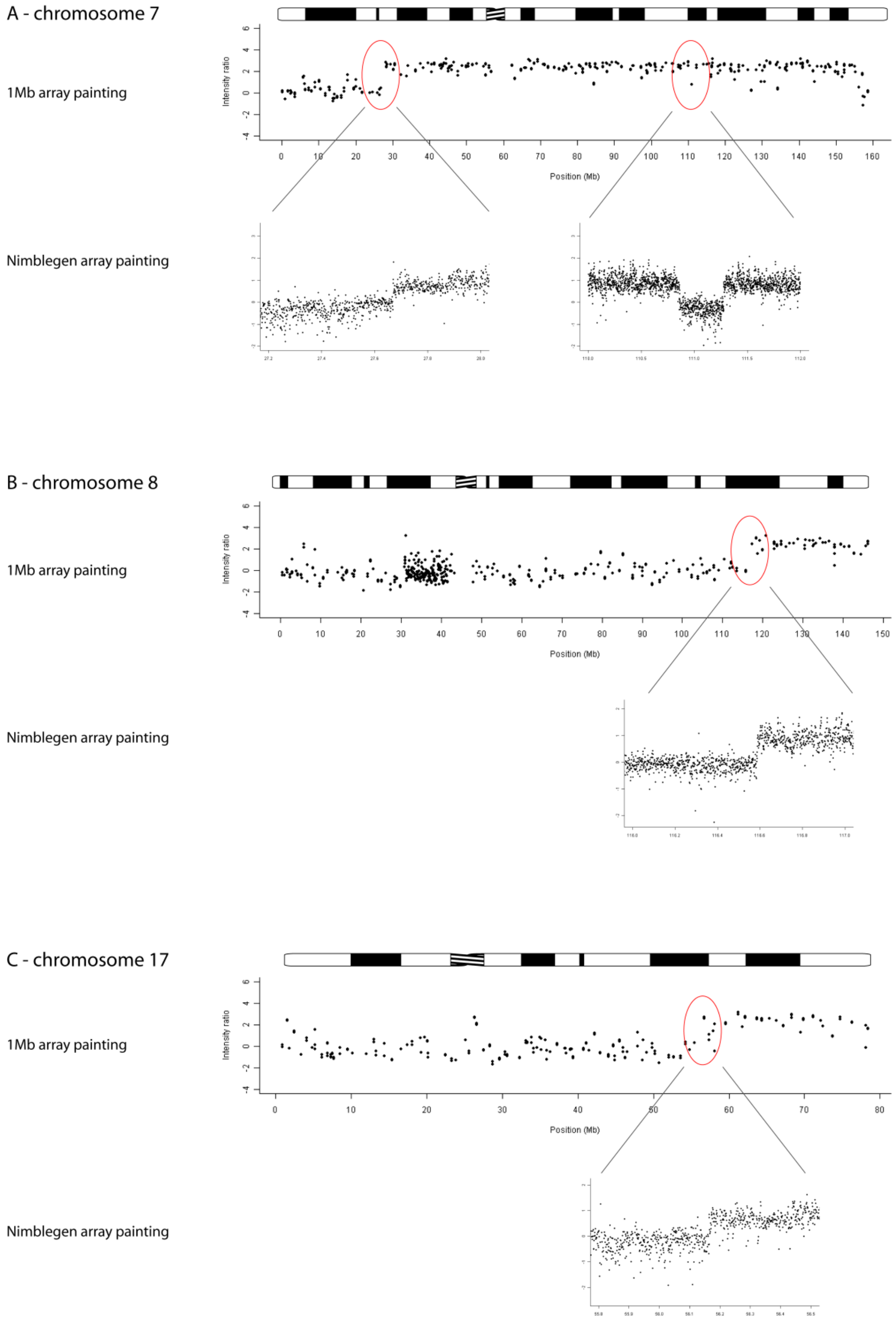


Figure 3.1. 1Mb array painting and Nimblegen array painting for the der(7)t(8;7;17) chromosome from HCC 1806. A – chromosome 7, showing a break at 27,670,000bp and a 400kb deletion between 110,858,500 and 111,278,500bp. B – chromosome 8, showing a break at 116,586,500bp. C – chromosome 17, showing a break at 56,170,000bp.

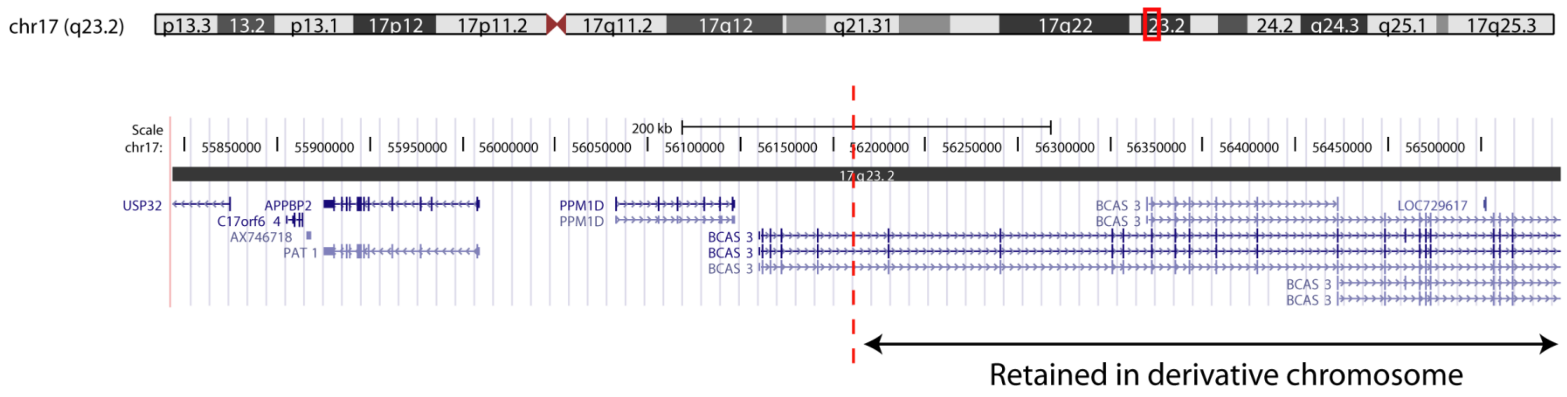


Figure 3.2. Position of breakpoint on chromosome 17 in der(7)t(8;7;17). The breakpoint is between exons 4 and 5 of BCAS3, and the 3' end of the gene is retained in the derivative chromosome.

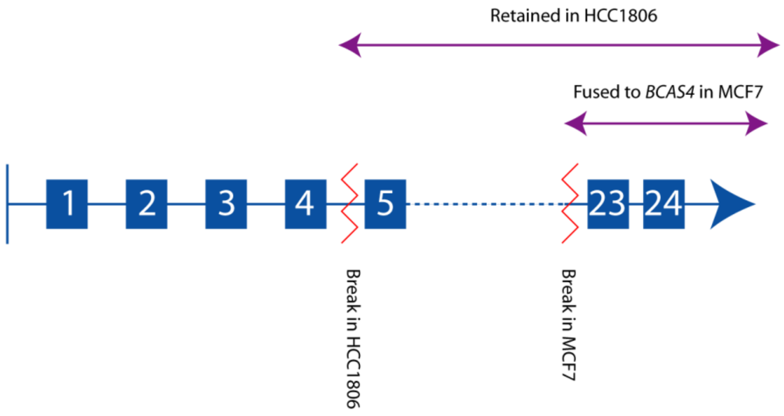


Figure 3.3. Diagram of the breaks in *BCAS3* in MCF7 and HCC1806, showing the exons retained in each cell line.

On chromosome 7, there was a breakpoint between 27,669,500 and 27,671,500bp, which is just outside the promoter of *HIBADH* and 50kb from the gene *TAX1BP1*, and a deletion between 110,858,500 and 111,278,500bp which deletes part of *IMMP2L* and *DOCK4* (Figure 3.4). On chromosome 8, the breakpoint was between 116,585,500 and 116,588,500bp, which is in the gene *TRPS1* (Figure 3.5). (See Chapter 4 for investigation of the *TRPS1-TAX1BP1* fusion.)

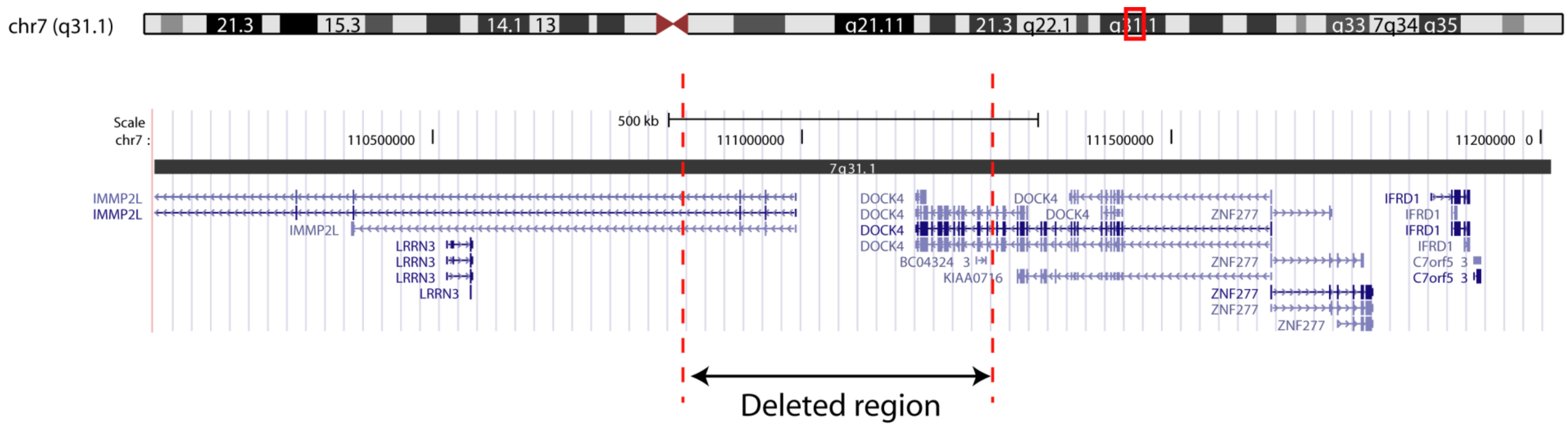


Figure 3.4. Position of deletion on chromosome 7 in der(7)t(8;7;17). The deletion removes part of *DOCK4* and *IMMP2L.*

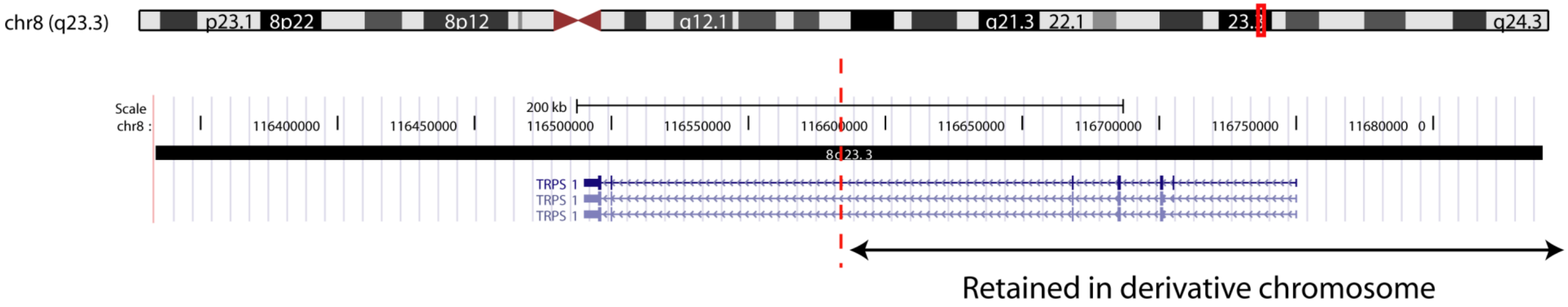


Figure 3.5. Position of breakpoint on chromosome 8 in der(7) t(8;7;17). The breakpoint is withing the large intron of *TRPS1.*The 5' end of the gene is retained in the derivative chromosome.

3.2.2 FISH mapping of the derivative chromosome

Although the individual breakpoints were known to high resolution, the arrangement of the chromosomes was not completely resolved, and the possible fusion partners of *BCAS3* could not be determined from the array data alone. Chromosome 7 was known to be joined to chromosomes 8 and 17 from the SKY karyotype, but only one breakpoint on chromosome 7 was seen from the array painting. One possibility was that chromosome 7 was fused to another chromosome at or very close to the telomere, which would not be seen on the array painting as there were few probes in regions near the telomere (Figure 3.6A). A second possibility was that the small deletion was not an interstitial deletion but part of a more complex rearrangement involving an inversion and a deletion, which would mean one of the breakpoints from the deletion was joined to another chromosome (Figure 3.6B). The orientation of the chromosome 7 fragment with respect to the other 2 chromosomes was not clear, and from the array painting it could not be determined whether the known breakpoint was joined to chromosome 17 or chromosome 8.

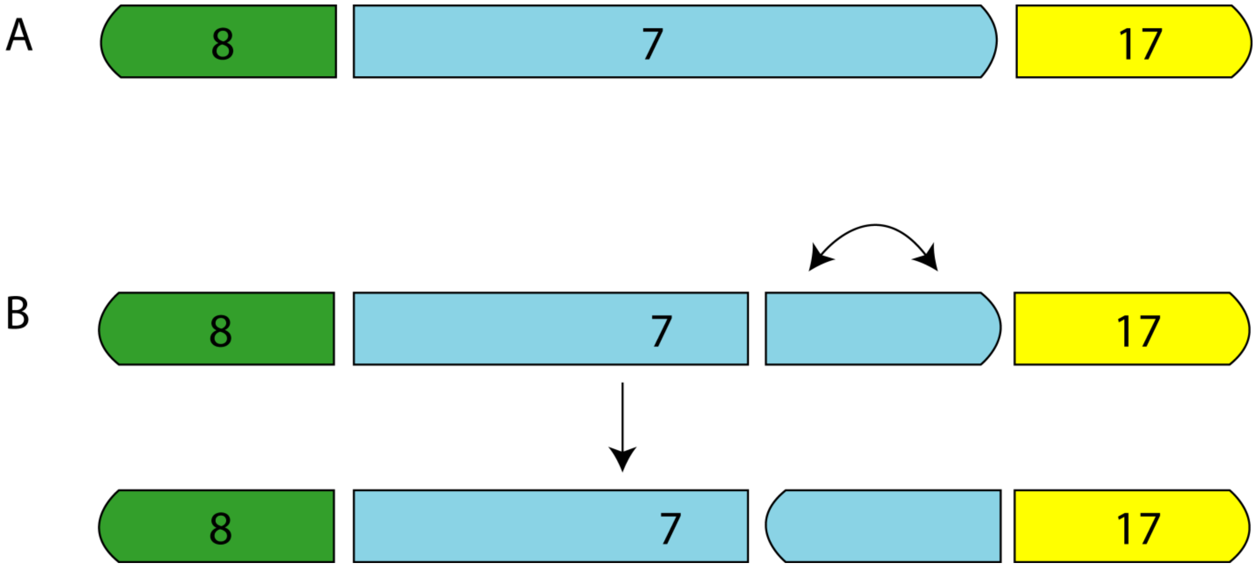


Figure 3.6. Possible orientations of the chromosome fragments in the der(7)t(8;7;17) chromosome based on 1Mb array painting. A – the unknown break on chromosome 7 could be a near-telomeric fusion. B – the unknown break could be part of a more complicated rearrangement involving inversion and a deletion, possibly the known deletion on chromosome 7.

To determine the arrangement of the chromosome fragments a series of FISH experiments was performed. A probe close to the known breakpoint on chromosome 7 was hybridised to a HCC1806 metaphase with chromosome 17 paint (Figure 3.7). The probe did not co-localise with chromosome 17, but was found at the other end of the derivative chromosome, showing that the known chromosome 7 breakpoint was joined to chromosome 8.

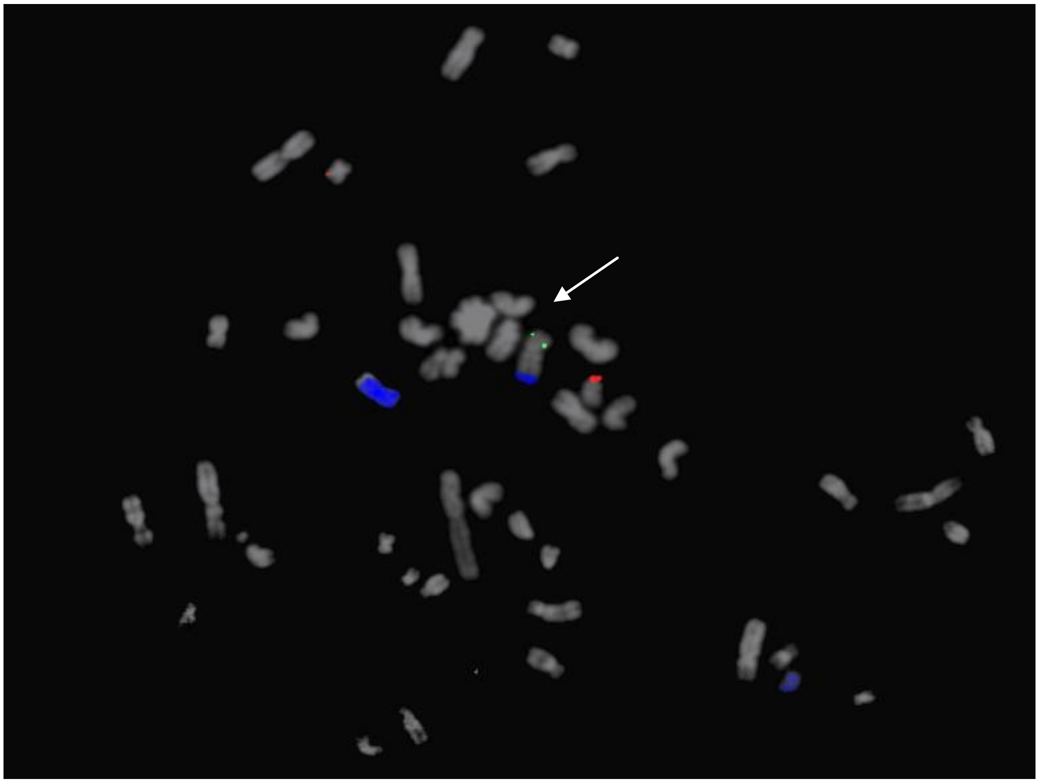


Figure 3.7. FISH on HCC1806 metaphase to determine orientation of chromosome 7 fragment. Chromosome 17 paint labeled with Spectrum Orange is shown in blue. RP4-781A18 on chromosome 7 is shown in green. RP4-781A18 is near the known breakpoint on chromosome 7 at 27.7Mb, and is present on the opposite end of the derivative chromosome from the chromosome 17 paint, showing that the known breakpoint is joined to chromosome 8 and not chromosome 17. The red signal is a mismapped probe.

To determine whether there had been a deletion and inversion, a FISH experiment was designed with probes near the telomere and the deletion on chromosome 7 (Figure 3.8). The probe near the telomere was not seen on the derivative chromosome, indicating that there had been a deletion or breakpoint near the telomere. The probe next to the deletion was present but was not close to the chromosome 17 fragment, indicating that there had not been the inversion on chromosome 7 hypothesized in Figure 3.6B.

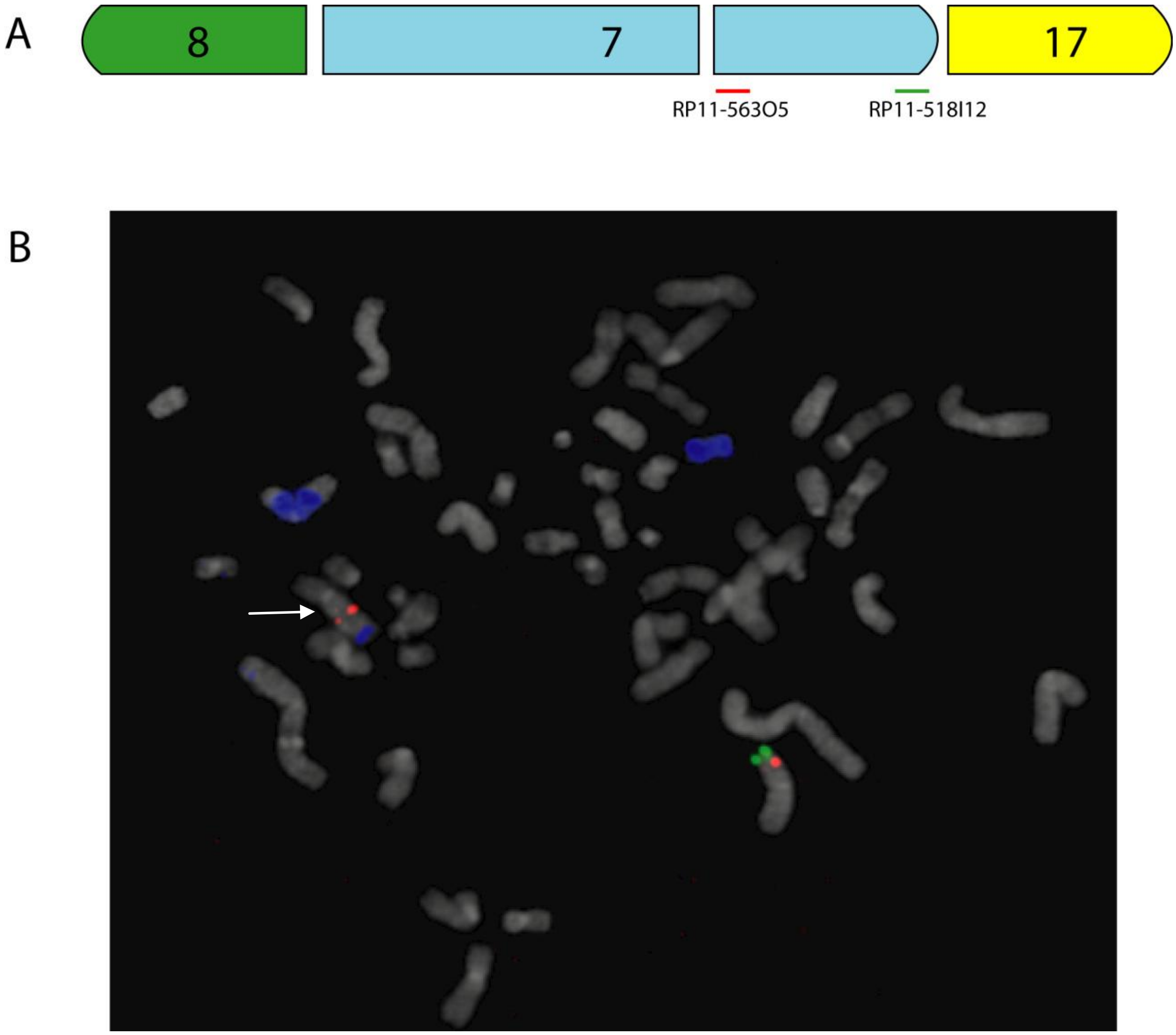


Figure 3.8. FISH to investigate a possible inversion on chromosome 7 in the der(7)t(8;7;17). A – the chromosome fragments known to be in the derivative chromosome in some orientation, and the location of the FISH probes used on chromosome 7. RP11-518I12 is close to the telomere of chromosome 7, and RP11-563O5 is next to the small deletion at 111Mb. B – FISH on HCC1806 metaphase. Spectrum orange chromosome 17 paint is shown in blue. RP11-518I12 is shown in red, and RP11-563O5 is shown in green. The chromosome in the lower right shows the normal arrangement of the green and red probes on the telomere of chromosome 7. On the derivative chromosome the green telomeric probe is absent, indicating a deletion near the telomere. The red probe at 111Mb is in the expected position and not juxtaposed with the chromosome 17 paint, indicating that there has not been an inversion.

3.2.3 Fine mapping and sequencing of the breakpoint

The results of the FISH experiments suggested that the chromosome 17 fragment was joined to the chromosome 7 fragment near the telomere of chromosome 7, but that some material had been lost from the telomeric region. The der(7)t(8;7;17) had previously been hybridised to a chromosome 7 tiling path array by Dr Karen Howath, but the loss of material from the telomere had not been detected. The analysis of the tiling path arrays was designed to reduce noise by routinely removing all probes which did not meet a set threshold for signal relative to a set of Drosophila control probes on the array, and re-analysis of the chromosome 7 tiling path showed that a number of probes near the telomere of chromosome 7 had been removed during this noise reduction process as the signal for the normal reference DNA fell below the signal threshold. When there probes were included, there was a deletion at the telomere which had been previously overlooked (Figure 3.9A) and that the breakpoint was between 155,560,009 and 155,669,524bp. There were no genes at this breakpoint, and the nearest gene was *SHH* (chromosome 7, 155,288,319-155,297,728), which was over 300kb from the breakpoint, suggesting that *BCAS3* was not part of a gene fusion (Figure 3.9B).

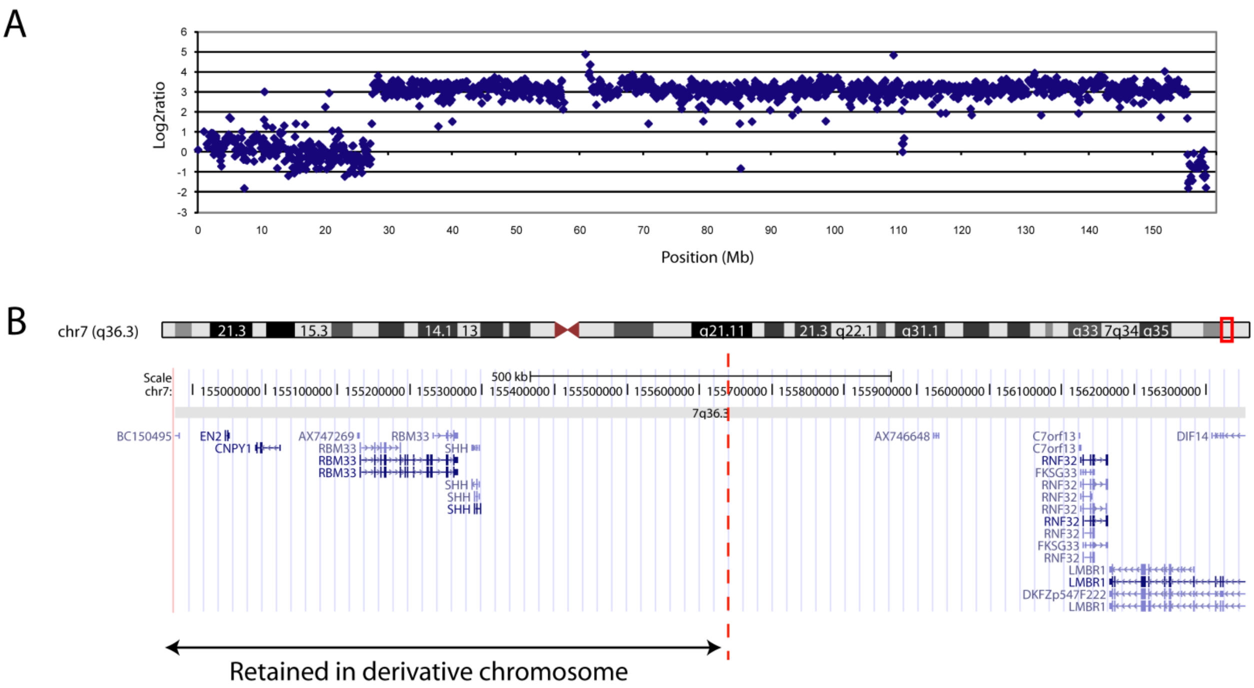


Figure 3.9. A - Tiling path array for chromosome 7 of the der(7)t(8;7;17). As well as the previously detected breakpoint at 27Mb, a further breakpoint at 155.5Mb can be seen. B – diagram of the deleted region on chromosome 7.

To confirm the result of this mapping which *BCAS3* was not fused to another gene, I cloned and sequenced the breakpoint junction. Using the breakpoint positions from the tiling path array painting, and from whole genome SNP6 arrays which had become available since the start of the project (SNP6.0 data kindly provided by Dr Graham Bignell and colleagues at the Sanger Institute Cancer Genome Project, later published as Bignell et al., 2010), the breakpoint on chromosome 7 could be mapped to between 155,709,517 and 155,715,189bp. The breakpoint on chromosome 17 was already known to be between 56,164,500 and 56,172,000bp. Primers were designed at 1kb intervals in the breakpoint regions, and long range PCR using combinations of these primers was carried out to amplify a junction product. A product of around 2kb was obtained using primers from 155,713,659bp on chromosome 7 and 56,166,019bp on chromosome 17. The product was cloned and sequenced to show the exact breakpoint was at 155,714,224bp on chromosome 7 and 56,165,019bp on chromosome 17, with a 1bp overlap between the sequences (Figure 3.10). The final arrangement of the der(7)t(8;7;17) chromosome is shown in Figure 3.11.

AAGGAGATGAGACACATCTGGTGAACACAGGTGACAGACATGGAGAAGTGAAAATGCTGTACCAAAAT
ATATTCTTCAATATGGTATTAAATATGGTATTTTAACAAATTTCTACGTATTAAATATTAATAGCATGATTT
GAGATCAGGAGAGCTAAGGTACATTATGCTAAATAACATTAAGGTAGAATAGTGAGACCAATATAGACT
GAATCATTCATTCATCAATTTATTCATTCAACAAGCATCTCTTTGGATTAACTATCATTTATTGAGTGCCAA
TTATTATATACTATCAAATATACAATTATACATTTAACAAATATACAATTTATATATTGTTGCCTGGTTAGA
TAAATATGTTATTAACCTTATTTTAAAACGAAACTCAGATTTAGTAAATTTGTATAGCTAATAAGCATANT
CCATTTTCTTTTCTACTA

Figure 3.10. Junction sequence for 7;17 junction of the der(7)t(8;7;17) chromosome. The bases highlighted in blue are from chromosome 7, 155,714,171 to 155,714,224bp on the positive strand. The bases highlighted in red are from chromosome 17, 56,165,019 to 56,165,424bp on the positive strand. The base highlighted in green is a 1bp overlap between the sequences. The 12 bases shown in black are a 12bp insertion into chromosome 17.

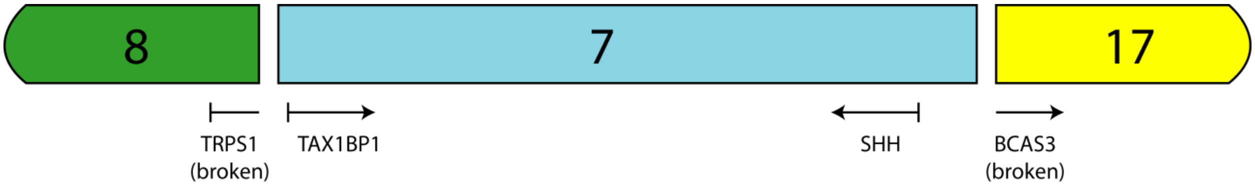


Figure 3.11. The correct arrangement of chromosome fragments in the der(7)t(8;7;17) chromosome in HCC1806. The nearest genes to the breakpoints are shown. Although genes are broken at the breakpoints on chromosome 8 and 17, they are joined to non-genic regions on chromosome 7, and no fusions can be found.

3.2.4 *BCAS3* in other cell lines and tumours

Microarray data from Chin et al. (2007) suggested that there was also a break in *BCAS3* in the breast cancer cell line *SUM52*. This was confirmed using FISH probes upstream and downstream of *BCAS3*, and showed that there was one extra copy with just the 5' end of the gene retained, and 5 extra copies of the 3' end of the gene (Figure 3.12).

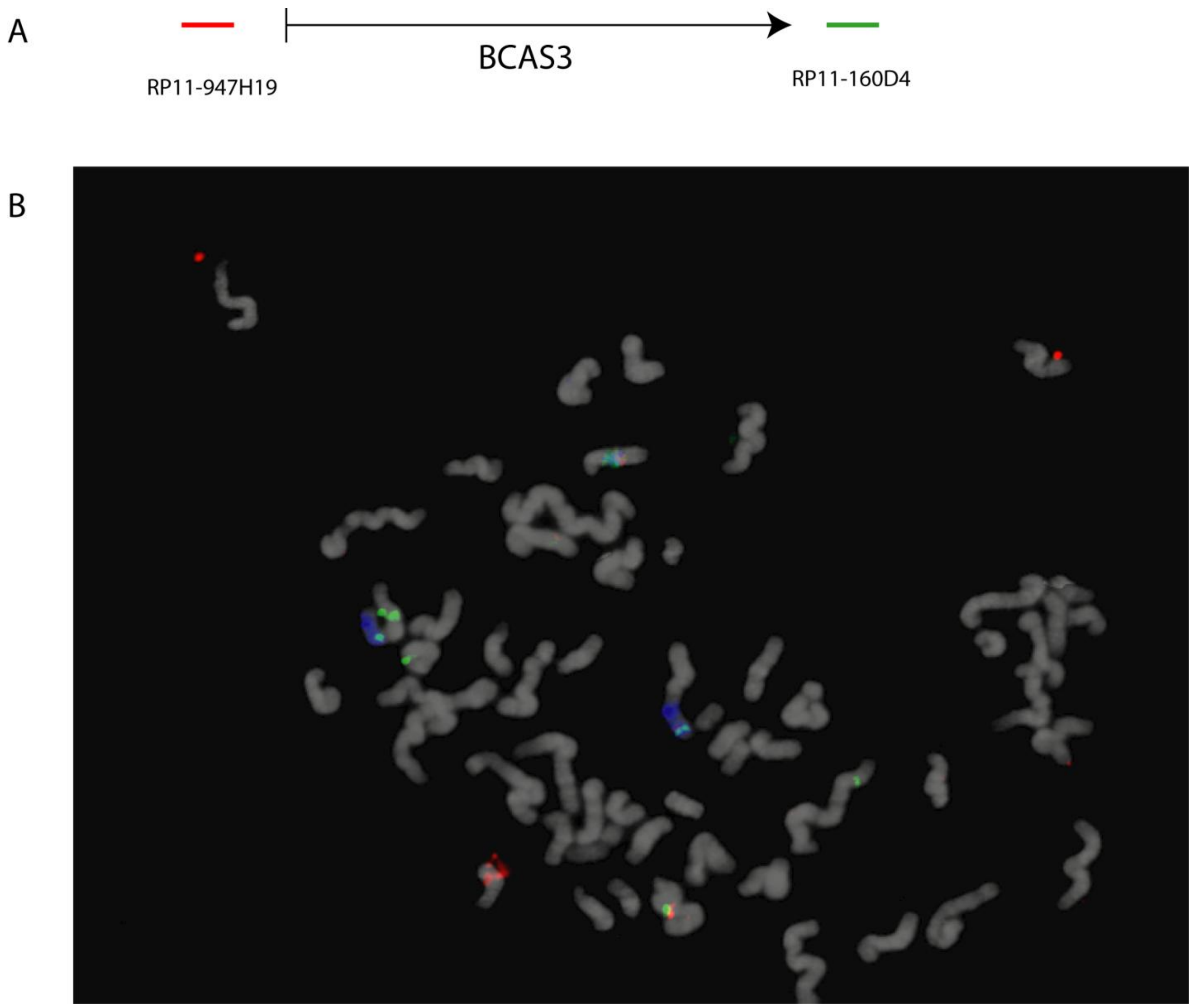


Figure 3.12. FISH with probes 5’ and 3’ to *BCAS3* on a SUM52 metaphase. A – diagram showing probe location. RP11-947H19 is located 100kb upstream of the start of *BCAS3*, and RP11-160D4 is located 30kb downstream of the end of *BCAS3*. B – results of hybridization to SUM52 metaphase. RP11-947H19 is shown in red and RP11-160D4 is shown in green.There are two intact copies of *BCAS3*, with five copies where only the 3’ end is retained (individual green signals), and one copy where only the 5’ end is retained (individual red signals).

To further investigate whether *BCAS3* was fused in any of the cell lines, I performed real time PCR using 3 sets of primers from the beginning, middle and end of the gene to look for any cell lines which differentially expressed part of the gene. The results are shown in Figure 3.13. The normal human breast cell line HB4a was used as a control. HMT3552 is another normal human breast cell line.

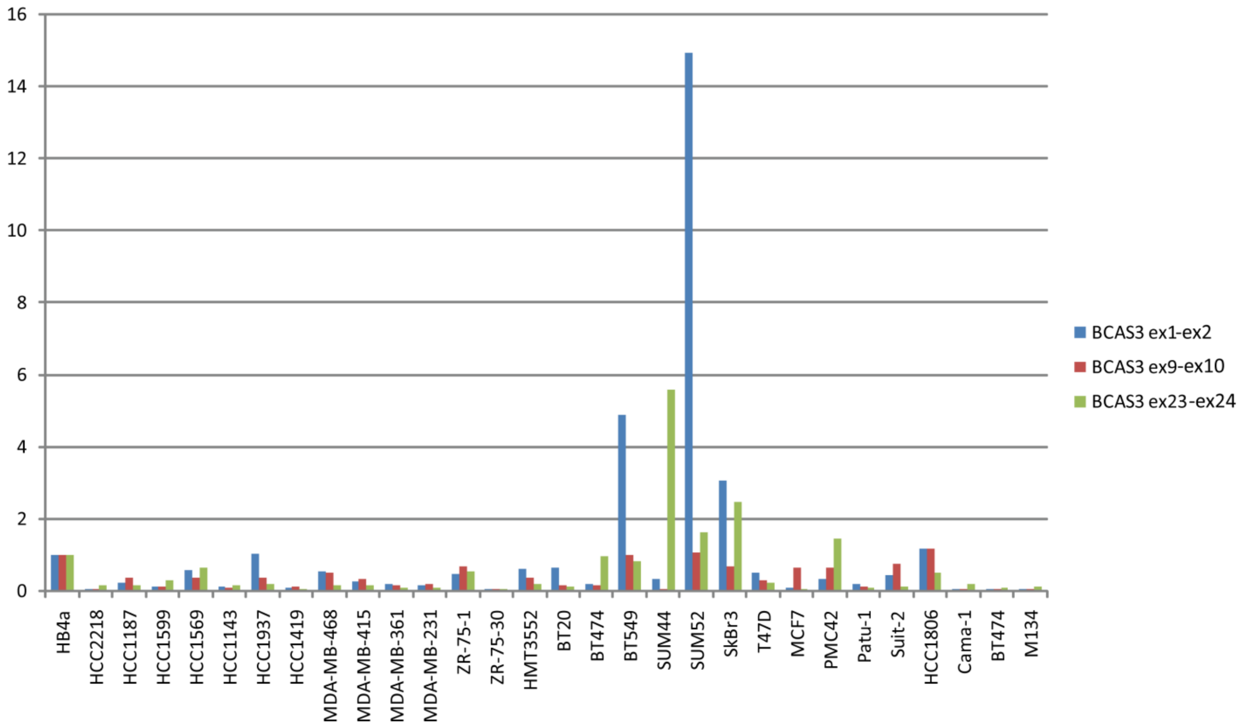


Figure 3.13. Real time PCR for 3 different exons in BCAS3 on a panel of cell lines. All the values are normalized to HB4a, a normal breast epithelial cell line.

There are four lines which show twofold or higher overexpression of any part of *BCAS3* relative to HB4a: SUM52, BT549, SUM44, and SkBr3. SUM52 shows fifteen times higher expression of the first exon of *BCAS3* compared to HB4a, with lower expression of the two primer pairs in exons 9 and 23. There are more copies of the 3' end than the 5' end of *BCAS3* in SUM52, but this result suggests that the extra copies of the 3' end are not affecting expression of the gene. Whole genome SNP6.0 data for BT549 does not show any chromosome rearrangements in *BCAS3*. The highest resolution array data available for SUM44 and SkBr3 is a custom 30k Agilent array (data kindly provided by Dr Suet-Feung Chin), which shows no unbalanced rearrangements in *BCAS3* in either cell line at the resolution of the array. HCC1806 does not show overexpression of any exons of *BCAS3*. Notably, MCF7, which has the original *BCAS4-BCAS3* fusion gene, does not show overexpression of *BCAS3*. Overexpression of the primer pair in exon 23 would be expected as it is found in the *BCAS4-BCAS3* fusion transcript.

As *BCAS3* was broken in multiple cell lines, I decided to look for breaks in a set of tumours. 6 probes were chosen, 3 overlapping probes each upstream and downstream of *BCAS3.* Each set of probes was pooled and hybridized to a normal metaphase to ensure they gave a single strong signal, and then hybridized to a tissue microarray containing cores from 141 breast tumours (tissue microarray kindly provided by Dr Suet-Feung Chin, who also performed the hybridization) (Figure 3.14).

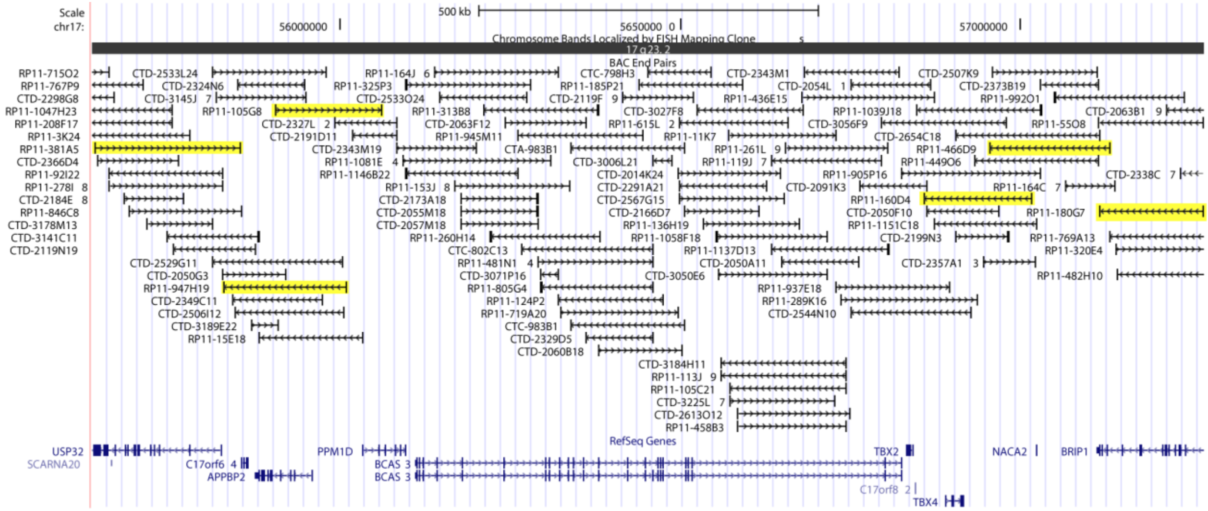


Figure 3.14. Locations of BAC probes used for TMA FISH. The upstream and downstream probes are just outside the *BCAS3* gene, and the three probes on each side overlap to give a single signal on an interphase nuclei. The upstream probes are RP11-105G8, RP11-381A5, and RP11-947H19. The downstream probes are RP11-160D4, RP11-466D9, and RP11-180G7.

Each of the tumour cores was scored according to number of signals seen for each pool of probes. Of the 141 tumours, 107 could be successfully scored. 97 of the cores showed a normal result, with overlapping signals from the two sets of probes. A further 7 tumours showed amplification of both sets of probes, indicating the whole of *BCAS3* was amplified. Only 3 tumours showed split probes, indicating that there was a break in the gene, and all 3 tumours had extra copies of the 5' end of *BCAS3* only. No tumours showed isolated signals from only the 3' end of the gene.

A Western blot was performed to analyse the BCAS3 protein, initially in 3 cell lines, HB4a, HCC1806, and MCF7 (Figure 3.15). Unfortunately the only available antibody to BCAS3 gave multiple nonspecific bands and could not be used for any further analysis.

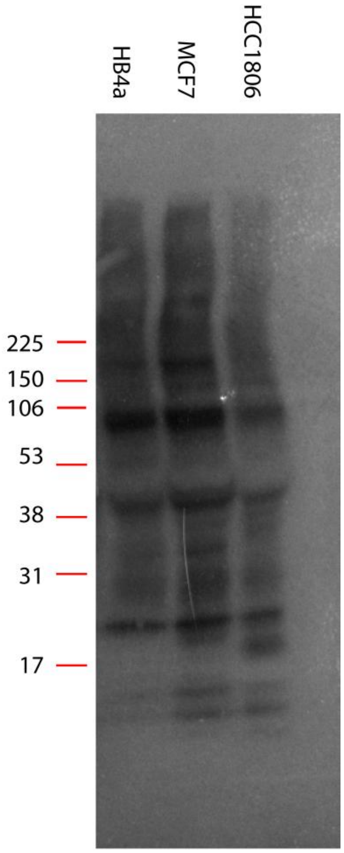


Figure 3.15. A Western blot of BCAS3 on 3 cell lines. HB4a is immortalized normal breast epithelium, HCC1806 and MCF7 have known breaks in BCAS3. Multiple non-specific bands were observed in all cell lines. The BCAS3 protein is 101kDa.

3.3 Discussion

Although *BCAS3* is broken in HCC1806, it does not form part of a gene fusion. The break in *BCAS3* on the der(7)t(8;7;17) removes the 5’ end of the gene and may inactivate it, but there are three complete copies of *BCAS3* present on other chromosomes in HCC1806, so the inactivation of one copy is unlikely to have an effect, and the expression of *BCAS3* in HCC1806 is not decreased compared to the normal breast cell line. The nearest gene to the broken copy of *BCAS3* is *SHH*, but there are no known enhancer elements near the breakpoint which could affect *SHH* expression.

In total, 8.2% of the tumours showed amplification of *BCAS3*, close to the figure of 9.4% reported by Bärlund et al. (2002). This is consistent with the knowledge that around 20% of breast tumours show amplification of 17q23, as *BCAS3* is just outside the minimal region of amplification as defined by Pärssinen et al. (2007) and would not be expected to be amplified in all tumours showing 17q23 amplification. The 3 tumours showing amplification of the 5’ end of *BCAS3* also support this interpretation – as *BCAS3* is a large gene just outside the minimally amplified region, some breaks in this gene would be expected, and may represent cases where the amplification ends inside *BCAS3*. As it was the 3’ end of *BCAS3* which was fused to *BCAS4* in MCF7, if a similar fusion was present in any of the tumour on the tissue microarray, split signals which retained the 3’ end of *BCAS3* would be expected, and that was not seen in any of the tumours.

Analysis of the tiling path array for chromosome 7 showed a deletion at the telomere which had not previously been detected, as the standard filtering procedure had removed the deleted probes. A section of the chromosome where a number of probes had been lost at any region other than the telomeres would be noticed, but a small deletion near the telomeres was not immediately seen. This suggests that the standard protocol developed for array quality control may cause other real breakpoints to be missed, particularly telomeric breakpoints.

Although *BCAS3* appeared to be a good candidate for a recurrent fusion gene in breast cancer, it does not appear that it is important as a fusion gene. Subsequent work by Dr Ina Shulte has found a fusion of the 5’ end of *BCAS3* to *HOXB9* in the cell line ZR-75-30, but this proved to be out of frame. While there does not appear to be an important recurrent fusion of *BCAS3*, it is recurrently broken, both in cell lines and tumours which show amplification of 17q, and in cell lines which do not show chromosome 17 amplification. This may be simply due to chance, as it is a large gene which may be broken by chance, especially as it is near the edge of a common amplicon, or it may be that overexpression of a truncated form of *BCAS3* including the 5’ end can affect *BCAS3* activity.

Chapter 4

The complete karyotype of HCC1806

4.1.1 Introduction

The first cell line I completely analysed for chromosome rearrangements was HCC1806. HCC1806 is described as a cell line derived from an acantholytic squamous carcinoma of the breast (Gazdar et al., 1998). It has a heavily rearranged hyper-diploid karyotype (Figure 4.1), with a median of 51 chromosomes and no normal copies of chromosomes 2, 3, 4, 6, 7, 10, 12, 14, and 21. It is ER, PR and HER2 negative, and has a deletion of the potential tumour suppressor gene *FHIT* (Sevignani et al., 2003).

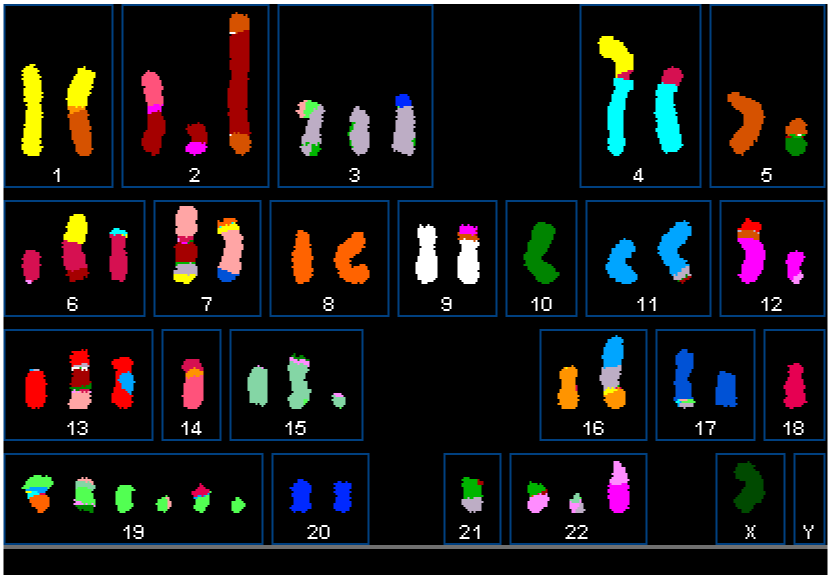


Figure 4.1. SKY karyotype of HCC1806 (Mira Grigorova, unpublished). A typical metaphase is shown – the consensus karyotype of HCC1806 is 51(49-53), X, -X, 1x1, der(1;5)(p10;q10), -2, der(2)t(2;5;2)dup(2), del(2)t(2;12), der(2?)t(2;14), der(3)del(3), der(3)t(3;22)(p12;?), der(3)t(3;20)(p12;?), der(3)t(3;19), der(4)t(4;6)(p15;p12), der(4)t(1;4)(q11;p15), 5x1, der(5;10)t(p10;p10), der(6)t(4;6)(p15;p12), der(6)t(1;6p), der(6)del(6)(q10-qter), -7, der(7?)t(2;7), der(7)t(8;7;17), 8x1, der(8)del(8)(p12-pter), 9x1, der(9)t(9;12)(p21; p12?), -10, der(10)t(6;10)(?;p11), i(10q), 11x1, der(11)t(3;11), -12, der(12)t(12;13)(p12;?), der(12)t(12;22)(q13-14;q13), der(12)del(12)(q13-qter), 13x1, der(13)t(13;2;7), der(13)t(13;11;13), -14, der(14)t(6;14)(?;p11.2), 15x1, isodic(15)t(15;10), der(15)del(15), 16x1, der(16)t(16q11.1;3p11;11p11-pter), 17x1, i(17q)t(3;17)/i((17q)t(3/;17;15), 18x1, 19x1, der(19)t(8;19), der(19)t(18;19), der(19)t(22;19;22), der(19)t(7;19;10), der(19)del(19), 20x2, -21, der(21)t(3;21), 22x1, der(22)t(21;22), der(22)t(12;22)(q13;q13)

Previous work analysed all the breakpoints at low resolution using 1Mb array painting, with higher-resolution tiling path and custom oligonucelotide arrays used to analyse primarily the balanced breakpoints (Howarth et al., 2008). HCC1806 was chosen as a good cell line for this analysis as it had a small chromosome number, making it easier to flow sort each chromosome, and it had a large number of reciprocal balanced translocations, which we were specifically interested in. Many of the unbalanced breakpoints had not been analysed at higher resolution than the 1Mb array painting, which gives breakpoints to a resolution of 3-4Mb, which is normally not enough to identify the genes which are broken and any fusions or rearrangements which may result. Using higher-resolution whole genome array CGH from the Affymetrix SNP 6.0 platform (Bignell et al., 2010), I aimed to complete the karyotype of HCC1806 to a high resolution, including the deletions and amplifications too small to be seen on previous arrays, and to investigate all the possible gene fusion events resulting from rearrangements.

4.1.2 Previous work

Previous work on HCC1806 in the lab was carried out by Dr Karen Howarth. Flow karyotypes of a chromosome preparation from HCC1806 cells were used to separate the chromosomes and each aberrant chromosome was hybridized separately to a 1Mb array. Flow sorting produced 51 separate chromosome fractions, labelled A to o (Figure 4.2) (Howarth et al., 2008).

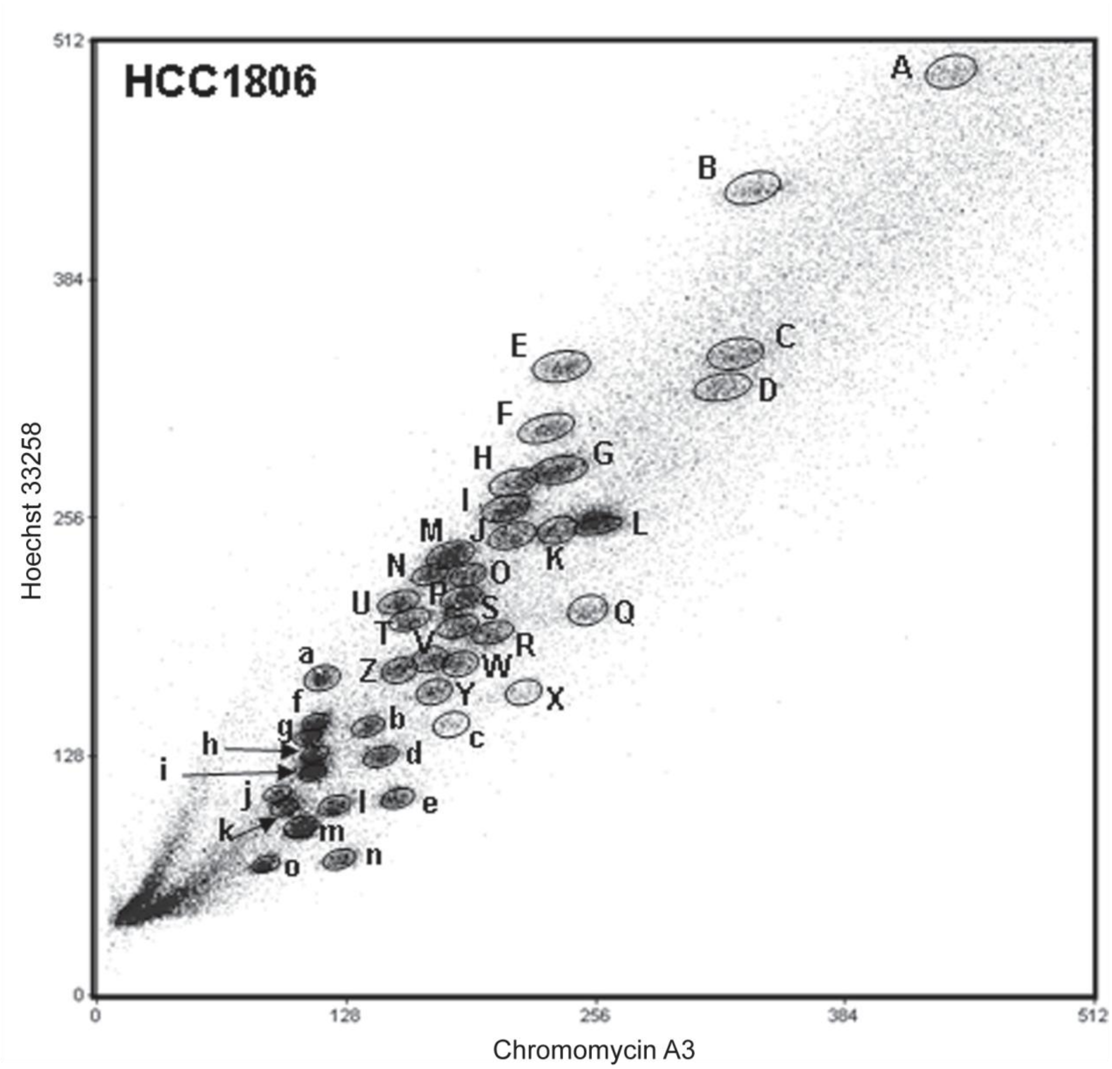


Figure 4.2. Flow karyotype of HCC1806 chromosomes. The chromosomes are sorted by separating the different fractions based on the staining intensity of Hoechst 33258 and Chromomycin A3. The 51 fractions are labelled A to o (Howarth et al., 2008).

A number of the fractions contained more than one derivative chromosome, as they co-localize on the flow karyotype due to the two derivative chromosomes being approximately the same size and with similar GC composition. There is one case in which the two derivative chromosomes in the same fraction contained pieces of the same chromosome, meaning that the mapped breaks could not be assigned to a single chromosome, but in all other cases the two co-sorted chromosomes did not contain pieces of the same chromosome. As the chromosome pieces involved in each derivative chromosome were known from the SKY karyotype, the breaks could be assigned to one of the derivative chromosomes.

Each fraction was labelled and hybridized to a 1Mb array; as 3 consecutive probes were considered necessary to call a change in copy number, the breakpoints could be mapped to a resolution of around 1Mb but rearrangements smaller than 3Mb would not affect 3 consecutive probes and were not identified. In total, 1Mb array painting revealed 93 breakpoints in HCC1806, of which 21 rearrangements were balanced to 1Mb resolution. Tiling path arrays which had probes every ~100Kb were available for chromosomes 6, 7 and 22, and 14 breakpoints on chromosomes 6, 7, and 22 were mapped to higher resolution using these arrays. A further 22 breakpoints, including all the balanced breakpoints, were mapped using custom Nimblegen oligonucleotide arrays designed to give probes every 200bp in the breakpoint regions. It was not practical to map all the breakpoints on the custom Nimblegen arrays, so the balanced breaks were prioritized as they would not be detected using whole genome array CGH even at high resolution.

Many of the breaks that were mapped to high-resolution with the Nimblegen arrays were within genes. Two fusion products were identified, both at balanced rearrangements: *TAX1BP1-AHCY* and *RIF1-PKD1L1* (Howarth et al., 2008).

4.2 Results

4.2.1 High-resolution breakpoints from SNP6 arrays

A total of 71 unbalanced breakpoints were not mapped at a high enough resolution by Howarth et al. (2008) to identify the genes broken. To map these breakpoints, I used whole genome array CGH data from the Affymetrix SNP6.0 platform, kindly provided by Dr Graham Bignell and colleagues from the Cancer Genome Project, Wellcome Trust Sanger Institute (later published as (Bignell et al., 2010)**.** This data was used to map at high resolution all the unbalanced breakpoints in HCC1806 which were previously known only to 1Mb or tiling path resolution, and to confirm the mapping previously performed using the Nimblegen arrays.

The Affymetrix SNP array 6.0 includes over 1,800,000 25bp probes. 900,000 probes detect SNPs, 200,000 copy number probes detect regions of copy number variation, while a further 700,000 probes are evenly spaced across the genome. This gives a median probe separation of 700bp. In addition, the SNP probes provide information on the genotype, allowing determination of regions of heterozygosity.

4.2.2 Determining the breakpoints

The provided segmentation of the whole genome SNP6.0 array using circular binary segmentation (Venkatraman and Olshen, 2007) was unreliable and missed several known breakpoints (Figure 4.3). Instead, the breakpoints were estimated by eye. 259 possible unique breakpoints were identified, without reference to the array painting in order to prevent selection bias towards already-known breakpoints. The size of the estimated interval containing the breakpoint varied according to the number of probes in the region of interest and the noise around the breakpoint, but the median size was 20kb.

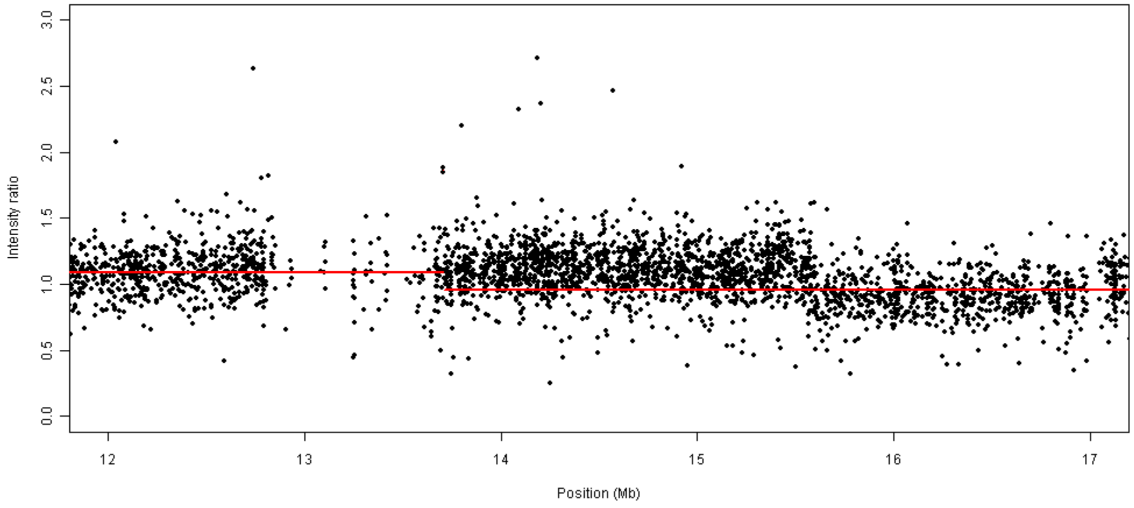


Figure 4.3. Incorrect segmentation of SNP6.0 arrays by circular binary segmentation. A plot of the SNP6.0 array for part of chromosome 1 with segmentation performed by circular binary segmentation. The position of the break is known from array painting to be at 15.5Mb, which is incorrectly assigned by the SNP6.0 segmentation.

4.2.3 Comparison of array painting and the SNP6.0 array

The whole genome SNP6.0 array was matched to our existing array painting data. 75 of the 259 breakpoints corresponded to breaks already identified by array painting. Comparison of the breakpoint regions called by eye on the SNP6 data with breakpoints which were known to tiling path or oligonucleotide array resolution showed that the SNP6 regions always agreed with the previous data, suggesting that the breakpoints called by eye are reliable.

The majority of the balanced breakpoints could not be seen as a copy number change on the whole genome SNP6.0 data, as they were balanced to the resolution of the array. There were three exceptions where the breakpoints could be seen as they were not perfectly balanced: the balanced breaks at 16p21.1 and 3p21.1 in the chromosome fractions L and I, and the balanced break at 7p15 in the fractions L and M. These breaks appeared perfectly balanced to 1Mb resolution, but with the higher resolution whole genome SNP6.0 data a small copy number gain of between 100-200kb could be seen at the position of each breakpoint (Figure 4.4). This gain could be caused by a duplication of material at the breakpoint, with the region of copy number gain being present on both derivative chromosomes, or it could represent a small duplication on one of the products, or an unrelated duplication on another copy of that chromosome. Subsequent work by Dr Karen Howarth confirmed using FISH and PCR that the duplicated material is present at the breakpoint on chromosomes from both fractions, and does not represent a tandem duplication on one of the translocation products.

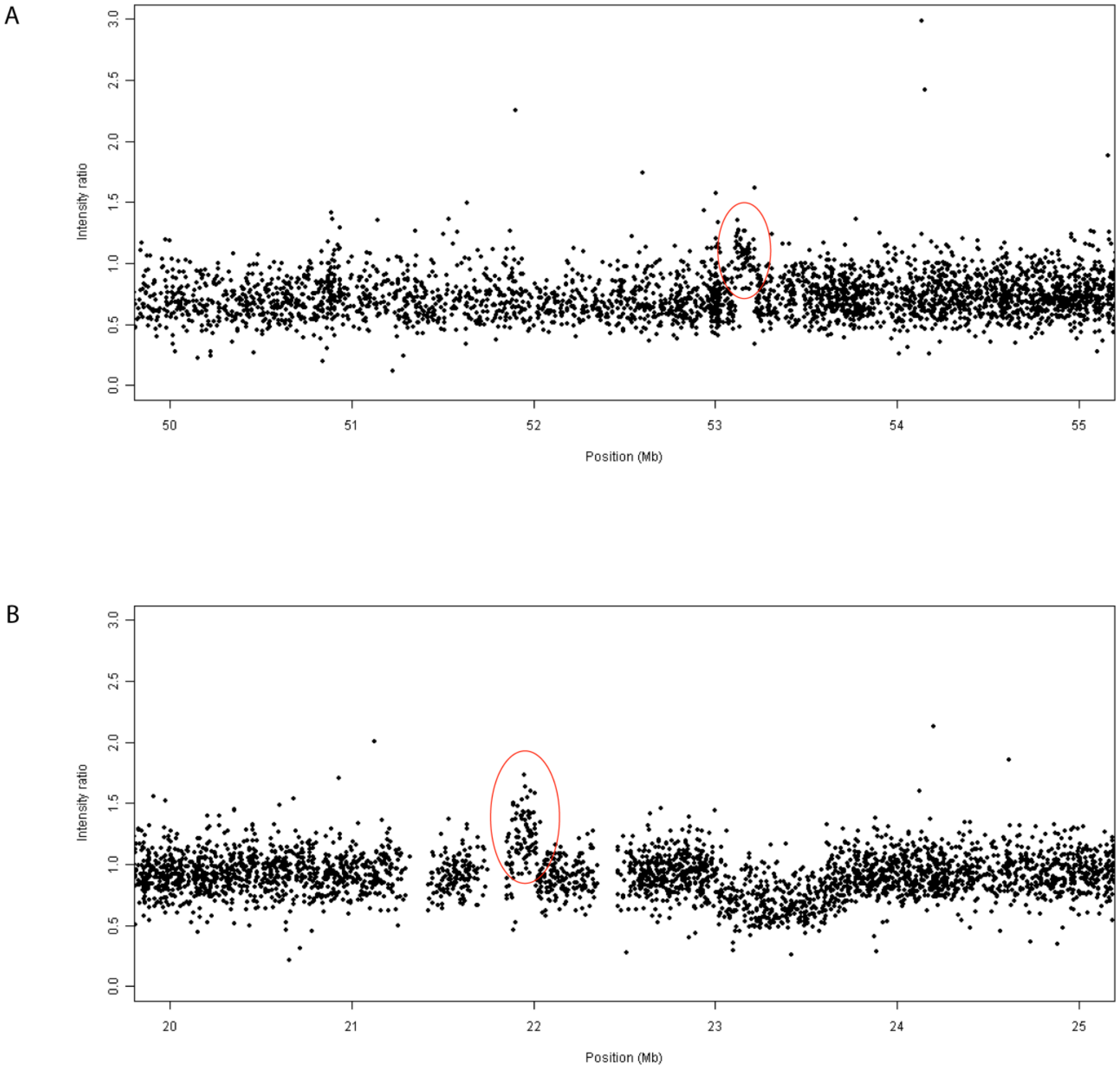


Figure 4.4. Breakpoint duplications from SNP6.0 arrays. The plots show the SNP6.0 array for HCC1806. A – part of chromosome 16, B – part of chromosome 3. The copy number changes marked in red are at the location of a balanced translocation, and represent duplicated sequences present in both products of the reciprocal translocation.

There was one other balanced break which can be seen on the whole genome SNP 6.0 data, which is the balanced break at 6p22. This break was found in three chromosome fractions (B, V and Z). The chromosome fragment from 6pter to 6p22 was found in two fractions (V and Z) and the reciprocal fragment was only found in fraction B. This gave an extra copy of one side of balanced break, which can be seen as a copy number step on the whole genome SNP6.0 array.

The only unbalanced breaks seen in the array painting data which were not present on the whole genome SNP6.0 array were the chromosome 15 and 17 breaks in chromosome fraction Q. Chromosome fraction Q contains a der(17)t(3;17;15) which is likely to be a further rearrangement of the der(17)t(3;17) chromosome in chromosome fraction X, as the chromosome 3 and 17 breaks were in the same locations (Figure 4.5). The absence of a copy number step corresponding to the der(17)t(3;17;15) in the SNP6 array data may represent a difference between the sample of HCC1806 used in our experiments and that used for the whole genome SNP6.0 array, suggesting the der(17)t(3;17;15) rearrangement may have occurred in culture, or it is possible that the breaks are actually balanced and that the reciprocal fragments have been lost in our sample and retained in the sample used for the SNP6.0 array. The predicted copy number from the SNP6.0 array is consistent with two copies of the der(17)t(3;17) and no copies of the der(17)t(3;17;15).

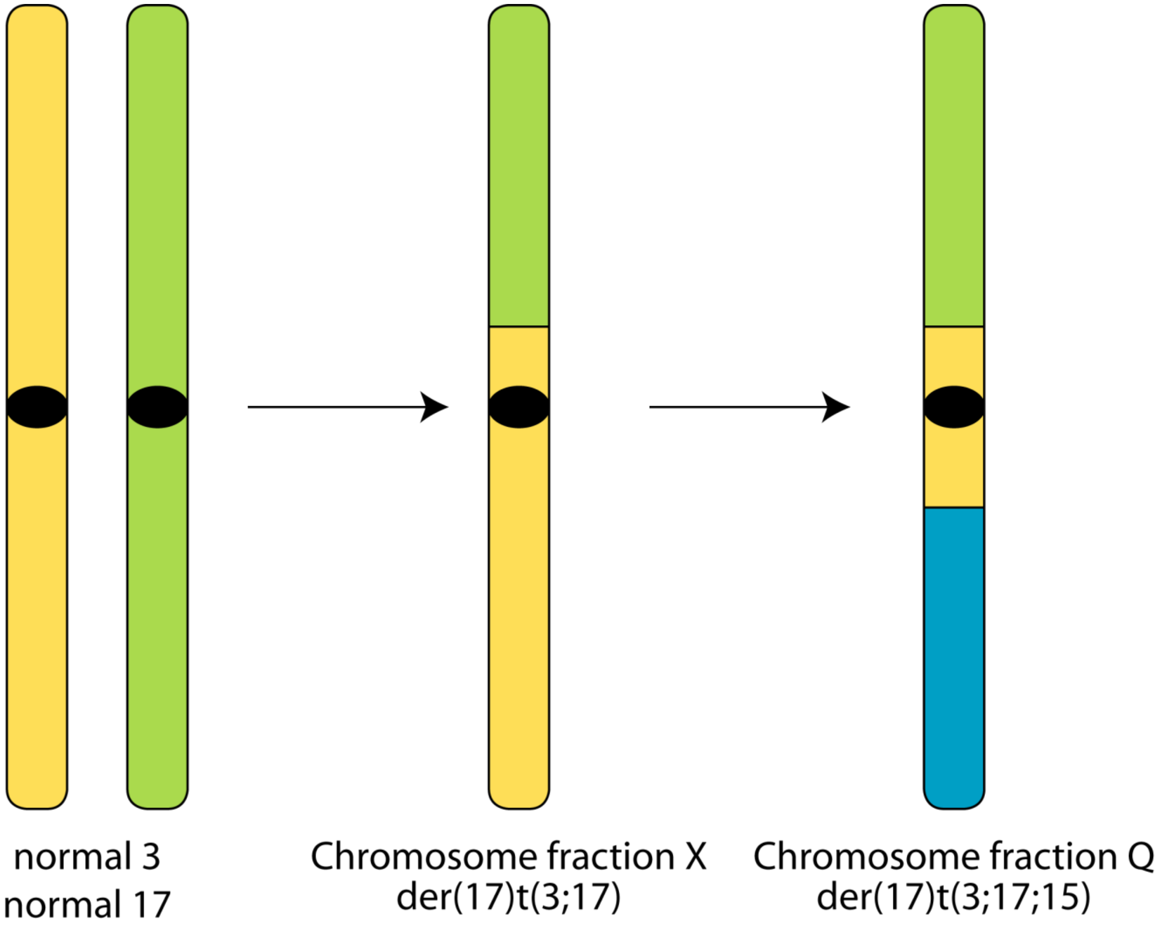


Figure 4.5. Related derivative chromosomes in HCC1806. The der(17)t(3;15;17) in chromosome fraction Q shares breaks on chr17 and chr3 with the der(17)t(3;17) in chromosome fraction X, and is assumed to be a further rearrangement of the same chromosome. The SNP6.0 copy number is consistent with their sample of HCC1806 having two copies of the der(17)t(3;17) and no copies of the der(17)t(3;17;15).

4.2.4 Identification of previously undetectable copy number changes

The breakpoints identified on the whole genome SNP6.0 array were used to find gains and losses which were not identified on the 1Mb array painting. As three consecutive clones at the same level were considered necessary to be sure of a break on the 1Mb array painting, any copy number gains or losses which spanned three probes or fewer would not have been called as a breakpoint from the 1Mb array. The exact resolution of the array depends on the exact spacing of the probes in that region, but any gains or losses under 5Mb are likely to have been overlooked on the 1Mb array painting. I identified previously-undescribed gains and losses by looking for any increase or decrease in copy number where neither of the boundaries were a known translocation and the region was under 5Mb. The SNP6 array showed 24 copy number gains, with a size range from 91kb to 2.11Mb and a median size of 1.01Mb, and 23 copy number losses ranging from 103kb to 4.5Mb with a median size of 577kb. An example of a loss found on chromosome 9 which was not called by array painting can be seen in Figure 4.6.

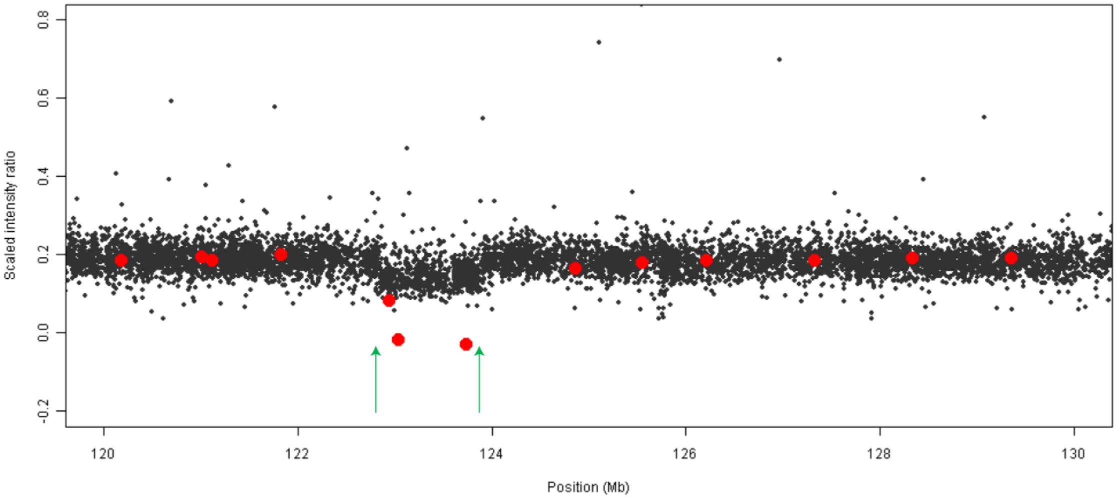


Figure 4.6. Example of a deletion identified from the SNP6.0 array. Whole genome SNP6.0 data for part of chromosome 9 is shown in gray, with the 1Mb array painting for chromosome fraction R overlaid in red. A deletion can be seen between the arrows at around 123 and 124Mb, which was not detected as 3 consecutive probes were not called as deleted. The sequence of the BAC at the left hand edge of the deletion probably overlaps the edge of the breakpoint.

4.2.5 Assembly of the complete karyotype

Using the SKY karyotype, array painting, and whole genome SNP6.0 array, a complete picture of the derivative chromosomes was constructed. Several assumptions were made in assembling the karyotype. First, I assumed that telomere fusions would be rarer than non-telomere fusions, and so two broken chromosomes are likely to join at the breakpoints rather than at the telomeres. A fusion at the telomeres would also leave two broken ends without telomeres. I further assumed that the karyotype that involves the fewest chromosome rearrangements to produce a derivative chromosome is correct, and a break that appears at the same location in two derivative chromosomes is joined to the same other chromosome, as it is more likely that the break has arisen once and undergone further rearrangement than for the same break to have arisen twice. For example, chromosome fraction O contains a der(20)t(3;20) chromosome, and chromosome fraction M contains a der(20)t(3;20;7) chromosome, and as the breaks on chromosome 3 and chromosome 20 are in the same locations in both derivative chromosomes I assumed they were joined to each other in both chromosomes and the der(20)t(3;20;7) had undergone a further translocation with chromosome 7. As the higher resolution CGH allowed the breakpoint intervals to be called to higher resolution this assumption is likely to be correct.

4.2.6 Discrepancies between array painting and whole genome SNP6.0 array

After the whole genome SNP6.0 was matched to the array painting, while the overall agreement was good there were still some breakpoints which could not be accounted for as either known breakpoints from the array painting, or small gains and losses that would be too small to see on the array painting due to the low resolution. I investigated some of these discrepancies in order to determine if they represented a true discrepancy between the two data sources.

One of the discrepancies I investigated was extra breaks on chromosome 2 (Figure 4.7). HCC1806 does not have a normal copy of chromosome 2, but there are 6 derivative chromosomes which contain pieces of chromosome 2, which were sorted into 5 different chromosome fractions. When the whole genome SNP6.0 data and the array painting were compared for chromosome 2, there were several breaks which were not part of small rearrangements and did not agree with any of the breaks previously called from the 1Mb array painting.

One such break was seen as a step down in copy number on the SNP6.0 data at 75Mb. This region of chromosome 2 was present on only the der(2)t(5;2;5) chromosome found in fraction A. A closer inspection of the 1Mb array painting data showed that the break at 75Mb appeared to be present on the array but had been missed (Figure 4.7). This may be due to the magnitude of the changes seen in array painting, as the change in log2 ratio between the regions of chromosome 2 which are not present in the derivative and those which are present at one copy is much greater than the shift between one and two copies present.

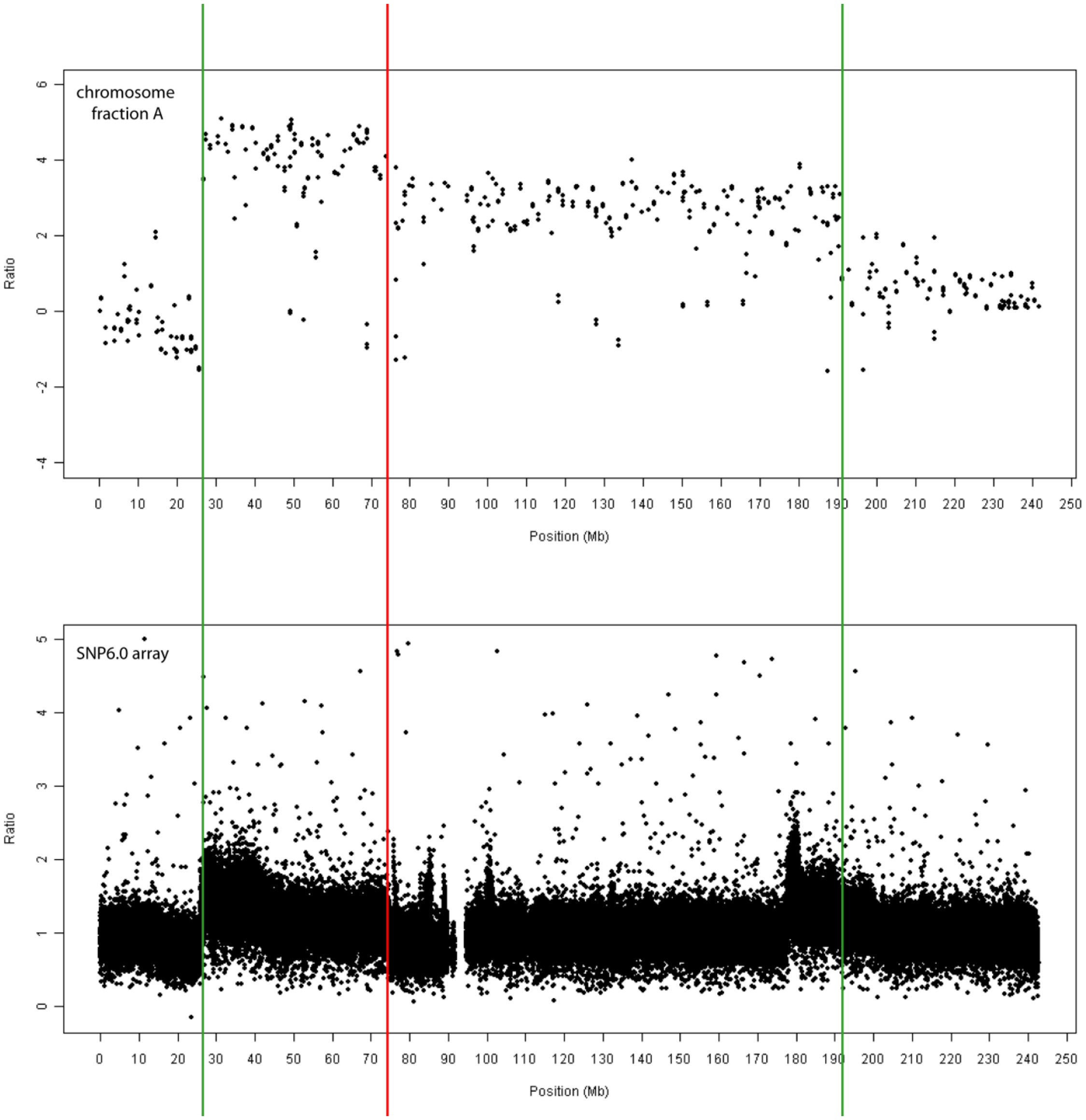


Figure 4.7. 1Mb array painting of chromosome 2 in chromosome fraction A (above) compared to the whole genome SNP6.0 array for chromosome 2 (below). The green lines show the breakpoints which were originally called from the 1Mb array painting and the matching breakpoints in the SNP6.0 data. The red line at 75Mb marks an additional breakpoint which was called from the SNP6.0 data and not previously known, and shows that the breakpoint can be found in the 1Mb array painting and was overlooked due to the smaller shift in hybridisation intensity from 1 to 2 copies than from 0 to 1 copy.

Another discrepancy involved several breakpoints on the q arm of chromosome 2 (Figure 4.8). In addition to a breakpoint at 150Mb known from the array painting, there was a break at 178Mb, a small amplification between 178 and 180Mb, and a break at 200Mb. The only chromosomes containing this region of chromosome 2 were the der(2)t(5;2;5) found in chromosome fraction A, and a der(7)t(2;7) found in chromosome fraction G.

On closer inspection of the 1Mb array painting, the extra breaks on the q arm of chromosome 2 could be seen to be at least one extra copy of chromosome 2 in chromosome fraction G from 178Mb to the qter (Figure 4.8). It was shown by FISH that there is an extra copy of that region of chromosome 2, with an amplification of the 178-180Mb region on one copy only (Figure 4.9). This amplification was seen on the array painting, but as there were only two probes in this region, it was not called as a copy number change. A further FISH experiment showed that although it is unclear from the array painting whether there is a break at 200Mb, there is an extra copy of the region between 180Mb and 200Mb on both chromosomes (Figure 4.10).

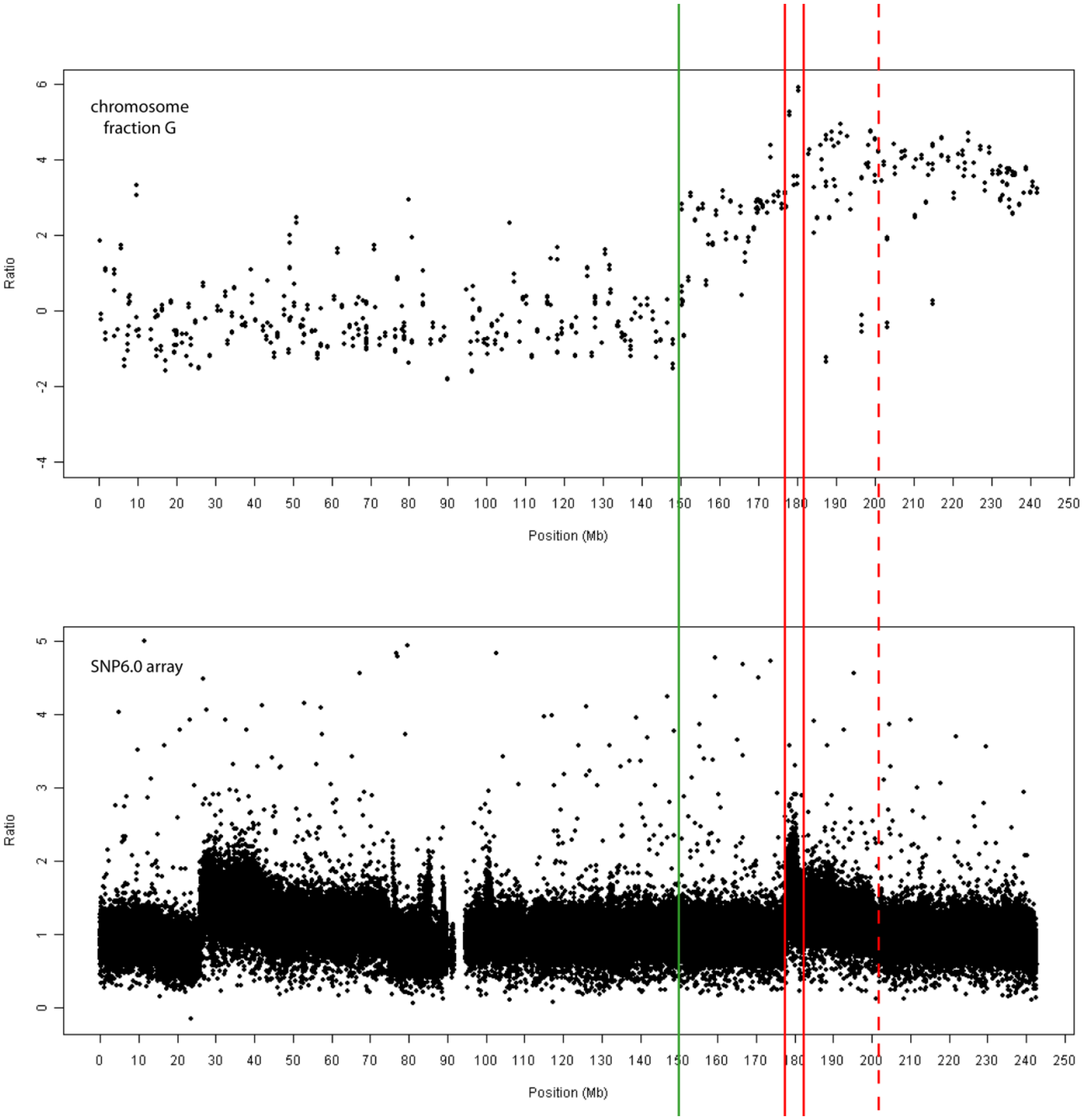


Figure 4.8. 1Mb array painting of chromosome 2 in chromosome fraction G compared to the whole genome SNP6.0 array for chromosome 2. The green line marks the breakpoint which was called from the 1Mb array painting. This break is balanced so not break is seen on the whole genome SNP6.0. The solid red lines show breaks which were seen on the SNP6.0 array but not previously seen on the 1Mb array painting, but they can been seen as a smaller shift on the array painting and may have been overlooked. The dashed red line shows a breakpoint on the SNP6.0 array which does not seem to have an associated shift on the 1Mb array painting.

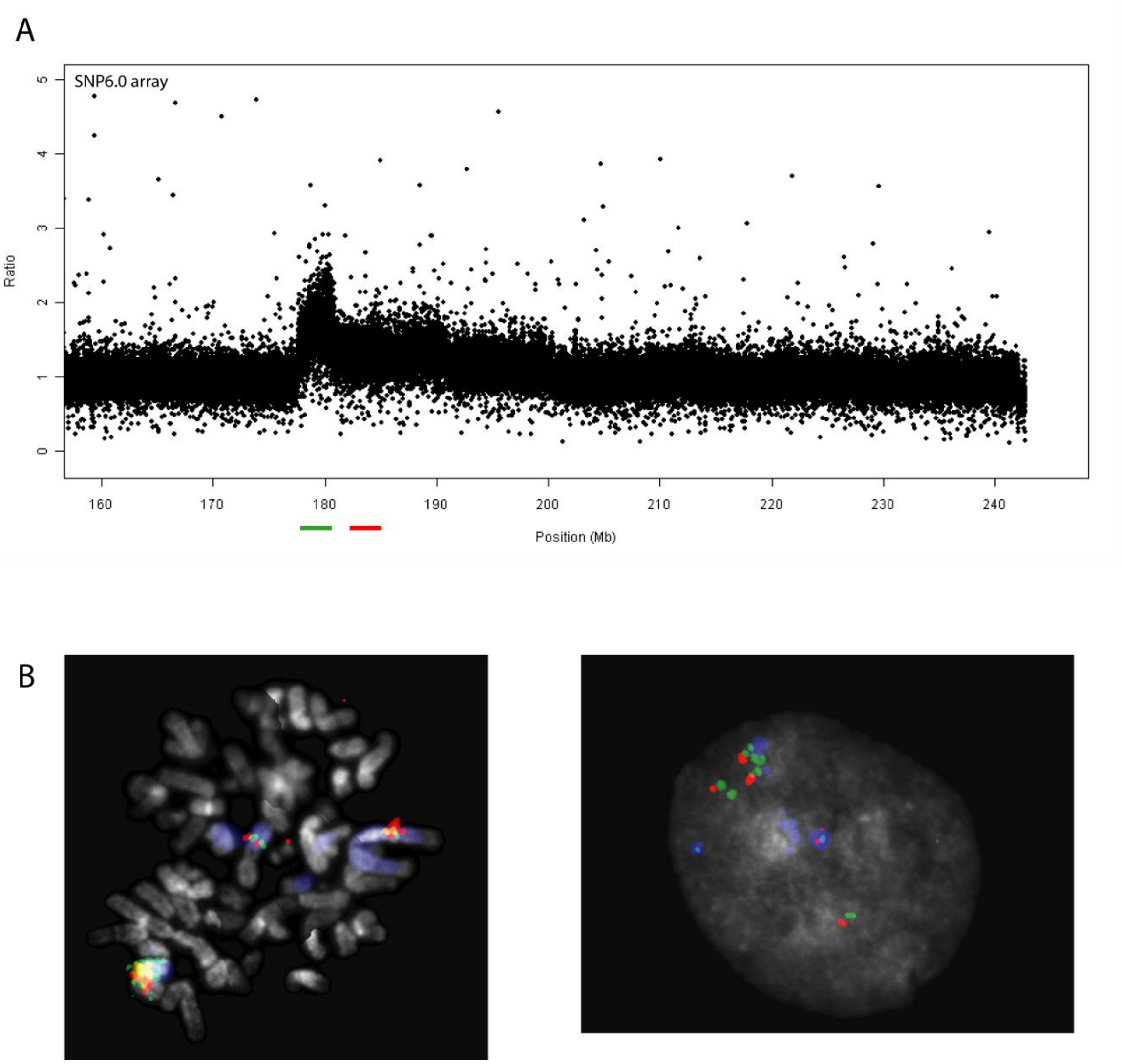


Figure 4.9. FISH to confirm breakpoint and amplification of extra chromosome 2 fragment in chromosome fraction G. A – whole genome SNP6.0 array for a portion of chromosome 2, with the two BAC probes used for FISH marked in red and green. B - FISH on interphase and metaphase nuclei shows chromosome 2 paint in blue, BAC RP11-65L3 in green and BAC RP11-67G7 in red. The FISH shows that there are 2 chromosomes with paired red and green signals, which are the known chromosomes from fractions A and G, and a chromosome with a single red signal and multiple green signals, which is an extra chromosome also present in fraction G and has an amplification of the region with the green probe.

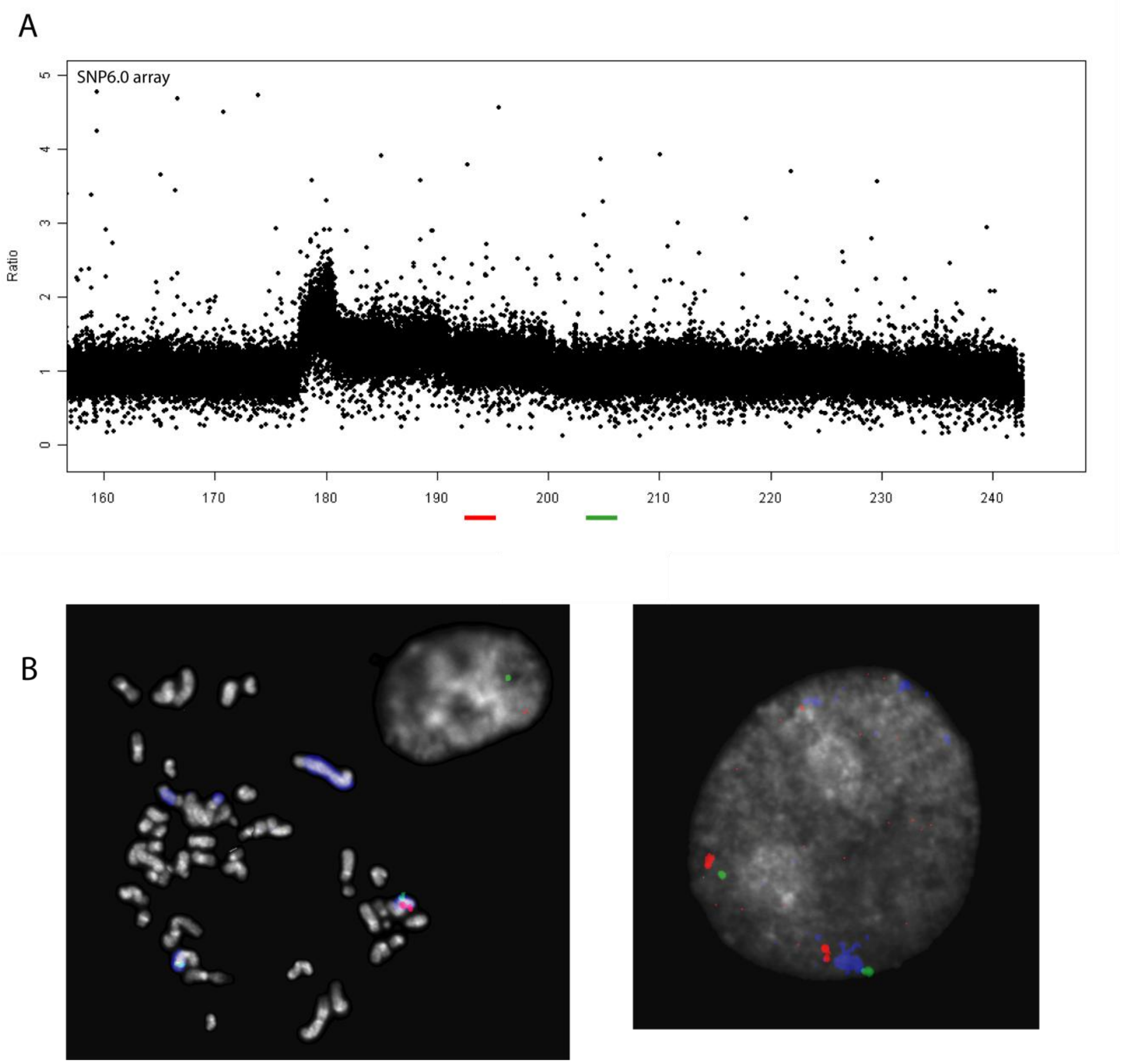


Figure 4.10. FISH to investigate extra chromosome 2 fragment in chromosome fraction G. A – whole genome SNP6.0 array for a portion of chromosome 2, with the two BAC probes used for FISH marked in red and green. B - FISH on interphase and metaphase nuclei shows chromosome 2 paint in blue, BAC RP11-15J24 in green and BAC RP11-59L22 in red. The FISH shows that there are 2 chromosomes that show two red signals and one green signal.

4.2.7 Systematic search for fusion genes

By assembling the complete karyotype of HCC1806, all the breakpoints were known to high resolution, and, in most cases, which breakpoints were joined together. It was possible that there was extra complexity at the breakpoints, such as a balanced inversion which would not be detected using either of the array platforms. With higher resolution array CGH, the genes at many of the unbalanced breakpoints are now known where previously there were several candidate genes, although there are still several breakpoints where the breakpoint cannot be mapped to a single gene.

By identifying the genes which are broken at each breakpoint, possible gene fusions could be predicted.

Fusion genes can be produced by chromosome translocations in two main ways. They can produce fusion genes directly by breaking a gene on each chromosome, which form a fusion gene on the translocated chromosome involving the 5’ end of one gene and the 3’ end of the second gene (figure 4.11). They can also cause a fusion product when only one gene is broken by removing the transcription termination and poly(A) addition site of the gene, which causes transcription to continue into an intact downstream gene and produce a fused transcript (figure 4.12). I refer to these as “readthrough” fusions. The *TAX1BP1-AHCY* fusion previously identified in HCC1806 (Howarth et al., 2008) is a readthrough fusion.

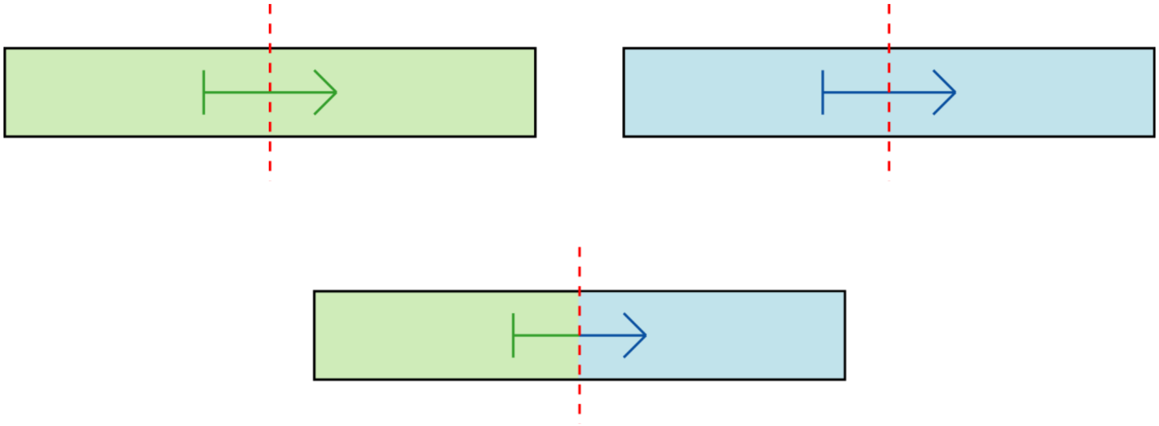


Figure 4.11. Translocation between two chromosomes directly forming a fusion gene. This produces a fusion gene with the 5' end of the green gene and the 3' end of the blue gene. If this is a balanced, reciprocal translocation, the reciprocal fusion gene may also be present.

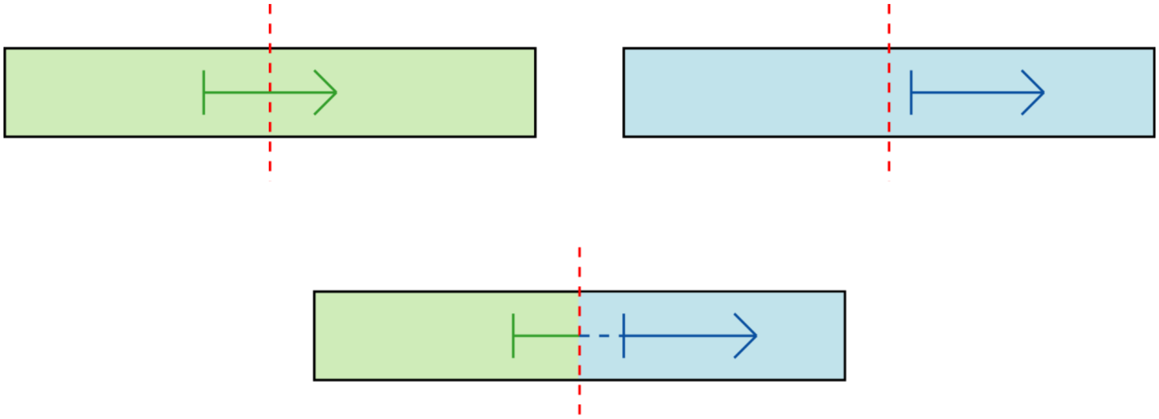


Figure 4.12. Translocation between two chromosomes where only one gene is broken to form a readthrough fusion. With the transcription end site of the green gene removed, this could produce a fusion containing the 5' end of the green gene and the whole of the blue gene (apart from the first exon, which would not normally have a splice acceptor site). This could alter the regulation of the blue gene, as it is under the control of a different promoter.

To investigate the fusions, PCR primers were designed according to the example in Figure 4.13. A pair of primers was designed to each gene to test expression of the gene in HCC1806 and in the normal breast cell line HB4a. By using a combination of the forward and reverse primers from different primer pairs, any fusion transcript would give a product only in HCC1806, and would not be present in the normal cell line.

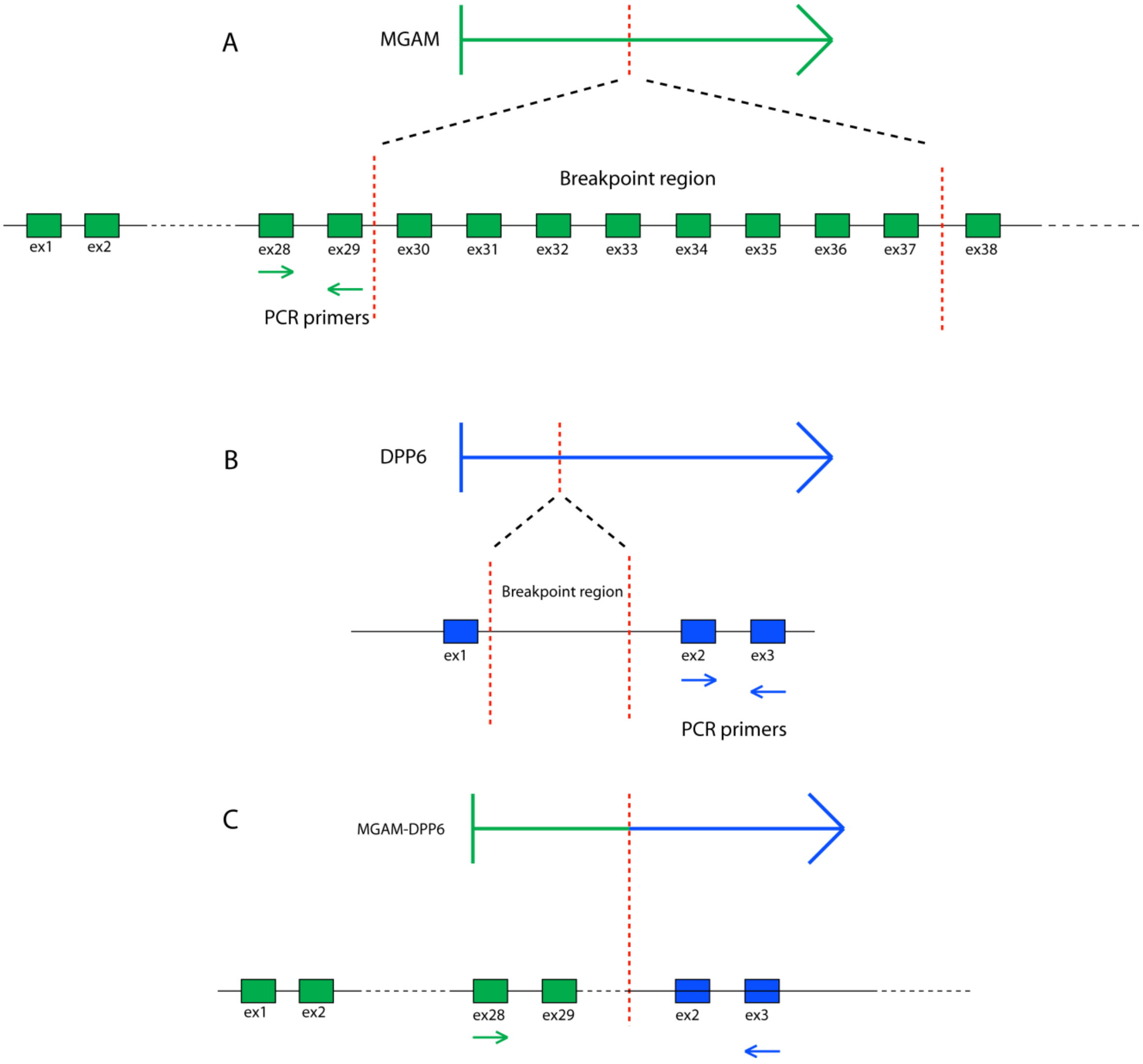


Figure 4.13. Primer design for amplification of fusion products on cDNA. A - The MGAM gene is broken in HCC1806 between exon 29 and exon 38 (breakpoint region defined by red dotted lines). PCR primers were designed between exon 28 and exon 29. B - The DPP6 gene is broken between exons 1 and 2. PCR primers were designed between exon 2 and exon 3. C - The hypothetical MGAM-DPP6 fusion protein would include the 5' exons from MGAM and the 3' exons from DPP6. By using the forward primer from MGAM and the reverse primer from DPP6, a product will only be produced if the fusion transcript is present, and a normal cell line not containing the translocation can be used as a control.

4.2.8 New fusions identified by high-resolution arrays

7 new candidate fusion genes were identified from higher resolution mapping of breakpoints (Table 4.1).

5' gene	Chromosome	3' gene	Chromosome	Chromosome fraction
FOXP4	6	HSP90	4	B, E
MGAM	7	DPP6	7	G
TMTC4	13	SUGT1L1	13	J, P
LMO1	11	NAG	2	K
TRPS1	8	TAX1BP1	7	L
CST4	20	EPHA3	3	M, O
BC022036	9	STAB2	12	R

Table 4.1. Potential fusion genes caused by translocations and large deletions. Chromosome fractions are defined in Howarth et al. (2008).

The results of the PCR for the fusion genes caused by translocations and large deletions are shown in Figure 4.14. No fusion transcripts were amplified; the genes *LMO1*, *HSP90* and *CST4* showed expression in HB4a but not in HCC1806, indicating that expression has been lost, either by disruption of the gene at a breakpoint, or by some other mechanism such as promoter methylation. *LMO1* is known to be recurrently translocated in T-cell leukaemia in the common t(11;14)(p13;q11) translocation (Boehm et al., 1991), and is thought to play a role in leukaemogenesis (Tremblay et al., 2010). Many of the genes did not show expression in either HB4a or HCC1806, and the lack of a fusion product may be due to lack of expression of the 5' gene.

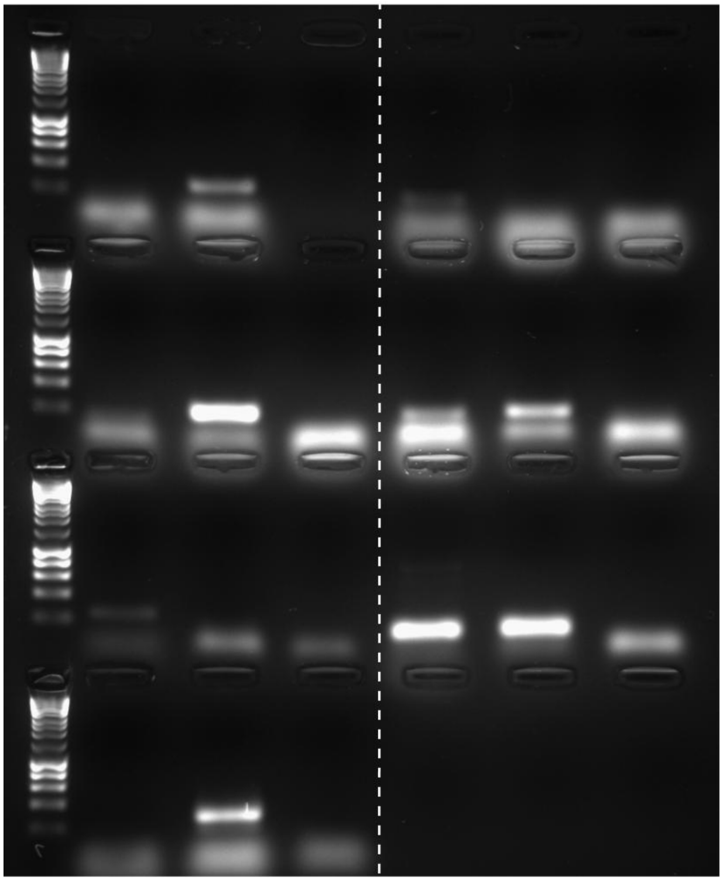


Figure 4.14. PCR for fusion genes caused by translocations in HCC1806. All PCRs were carried out using HCC1806 cDNA.

Row	Column	Primers
1	1	Ladder
1	2	*FOXP4*
1	3	*HSP90*
1	4	*FOXP4/HSP90*
1	5	*MGAM*
1	6	*DPP6*
1	7	*MGAM/DPP6*
2	1	Ladder
2	2	*LMO1*
2	3	*NAG*
2	4	*LMO1/NAG*
2	5	*CST4*
2	6	*EPHA3*
2	7	*CST4/EPHA3*

Row	Column	Primers
3	1	Ladder
3	2	*BC022036*
3	3	*STAB2*
3	4	*BC022036/STAB2*
3	5	*TRPS1*
3	6	*TAX1BP1*
3	7	*TRPS1/TAX1BP1*
4	1	Ladder
4	2	*SUGT1L1*
4	3	*TMTC4*
4	4	*SUGT1L1/TMTC4*
4	5	Negative control

4.2.9 Fusion genes caused by small deletions

Using the whole genome SNP6.0 array data, 23 small deletions which were not previously picked up by the 1Mb array painting were identified. Some of these small deletions were at regions known to have copy number variation in the normal population (Redon et al., 2006), and were assumed to be found in the germline, but many of the deletions remove parts of genes and could potentially produce fusion products, as shown in Figure 4.15. The 12 possible fusion genes are shown in Table 4.2. The results of the PCR for fusion genes caused by small deletions are shown in Figure 4.16. No fusion transcripts were amplified.

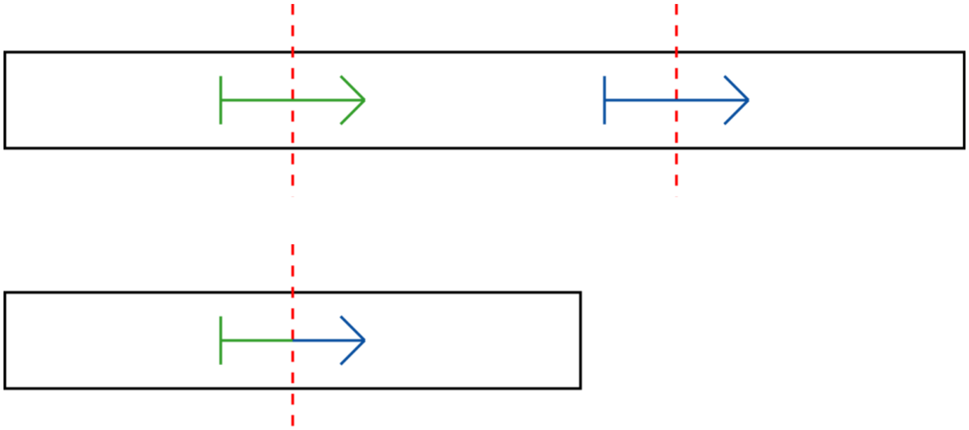


Figure 4.15. An intrachromosomal deletion can cause a fusion gene between the 5’ end of the green gene and the 3’ end of the blue gene. The other ends of the genes are lost.

Gene 1	Gene 2	Chromosome
DISC1	KIAA1383	1
HK2	REG3G	2
GPR39	MGAT5	2
NAP5	BC045801	2
MTDH	VPS13B	8
CTNLN	ADAMTS1L1	9
C5	TTLL1	9
HCCA2	OR52B2	11
SBF2	GALNTL4	11
USP31	ERN2	16
ATAD5	SUZ12	17
CHEK2	PITPNB	22

Table 4.2. Potential fusions from small deletions

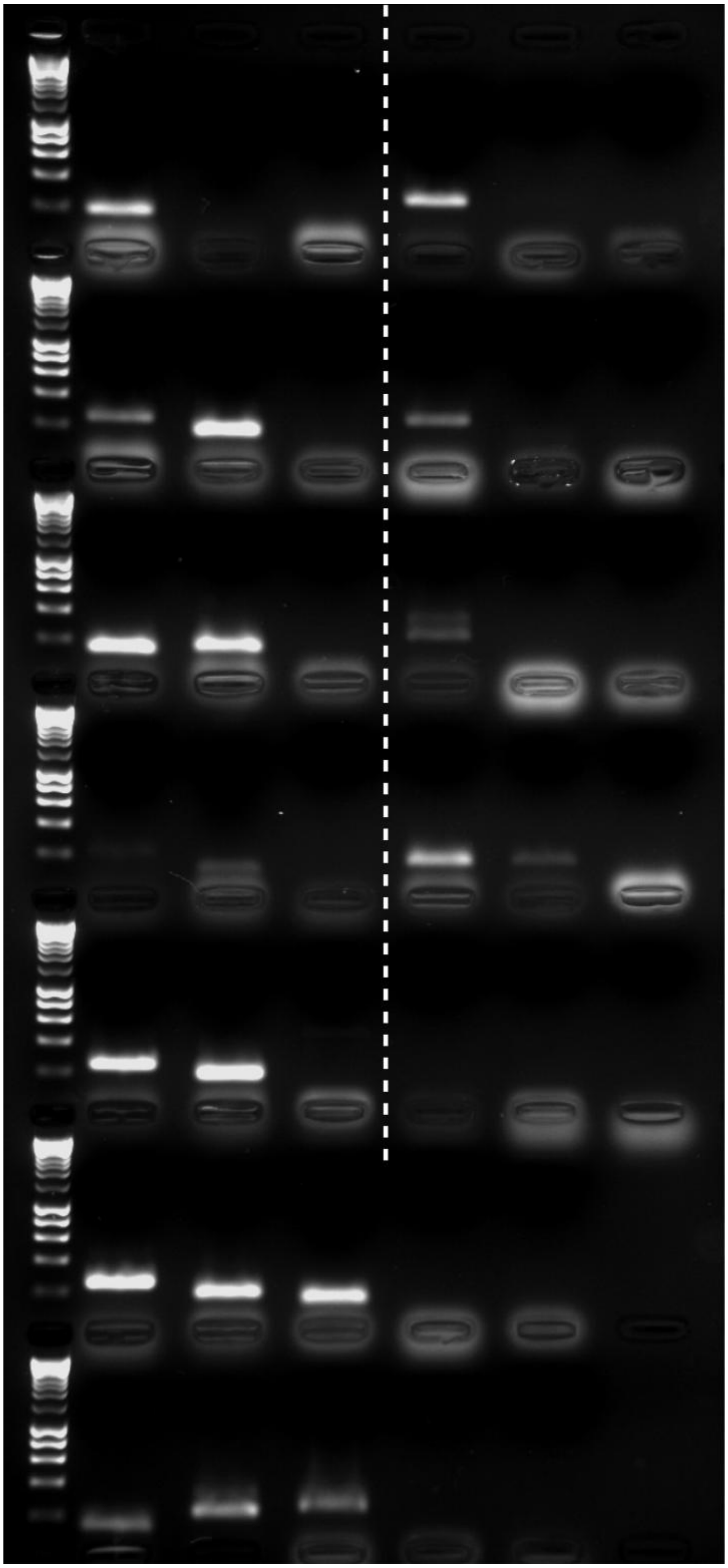


Figure 4.16. PCR for fusions caused by small deletions

Key for Figure 4.16:

Row	Column	Primers
1	1	Ladder
1	2	DISC1
1	3	KIAA1383
1	4	DISC1/KIAA1383 fusion
1	5	HK2
1	6	REG3G
1	7	HK2/REG3G fusion
2	1	Ladder
2	2	GPR39
2	3	MGAT5
2	4	GPR39/MGAT5 fusion
2	5	NAP5
2	6	BC045801
2	7	NAP5/BC045801 fusion
3	1	Ladder
3	2	MTDH
3	3	VPS13B
3	4	MTDH/VPS13B fusion
3	5	CTNLN
3	6	ADAMTS1L1
3	7	CTNLN/ADAMTS1L1 fusion

Row	Column	Primers
4	1	Ladder
4	2	C5
4	3	TTLL1
4	4	C5/TTLL1 fusion
4	5	HCCA2
4	6	OR52B2
4	7	HCCA2/OR52B2 fusion
5	1	Ladder
5	2	SBF2
5	3	GALNTL4
5	4	SBF2/GALNTL4 fusion
5	5	UPS31
5	6	ERN2
5	7	UPS31/ERN2 fusion
6	1	Ladder
6	2	ATAD5 exon 6
6	3	ATAD5 exon 14
6	4	SUZ12
6	5	ATAD5 exon 6/SUZ13 fusion
6	6	ATAD4 exon 14/SUZ12 fusion
7	1	Ladder
7	2	PITPNB
7	3	CHEK2 exon 2
7	4	CHEK2 exon 10
7	5	PITPNB/CHEK2 exon 2 fusion
7	6	PITPNB/CHEK2 exon 10 fusion

4.2.10 Fusion genes caused by tandem duplication

Small duplications may also produce fusion genes (Jones et al., 2008). 23 small duplications were seen in the SNP6 array CGH. These may be tandem duplications or they may be an insertion of an extra copy elsewhere in the genome. As it could not be determined from the array CGH which of these possibilities was correct, it was assumed for the purposes of predicting fusion genes that all of these duplications were tandem duplications. The potential fusion genes would then depend on the orientation of the genes at the breakpoint and the location and orientation of the inserted fragment. The possibilities are shown in Figures 4.17-4.19 and the predicted fusion genes are shown in Table 4.3. Figure 4.20 shows the PCR results. In one case, there were 3 possible genes for one end of a fusion due to a poorly-resolved breakpoint, and all 3 were tested. No fusion transcripts were found.

Gene 1	Gene 2	Chromosome
EPHB2	MYOM3	1
EPHB2	FUSIP1	1
EPHB2	PNRC2	1
MYOM3	EPHB2	1
FUSIP1	EPHB2	1
PNRC2	EPHB2	1
AFF3	BC156887	2
BC156887	AFF3	2
c6orf105	PHACTR1	6
PHACTR1	c6orf105	6
LAMA2	ARHGAP18	6
ARHGAP18	LAMA2	6
CATSPERB	TC2N	14
SMURF2	CCDC46	17
GPC3	HS6ST2	23

Table 4.3. Potential fusion genes resulting from small tandem duplications

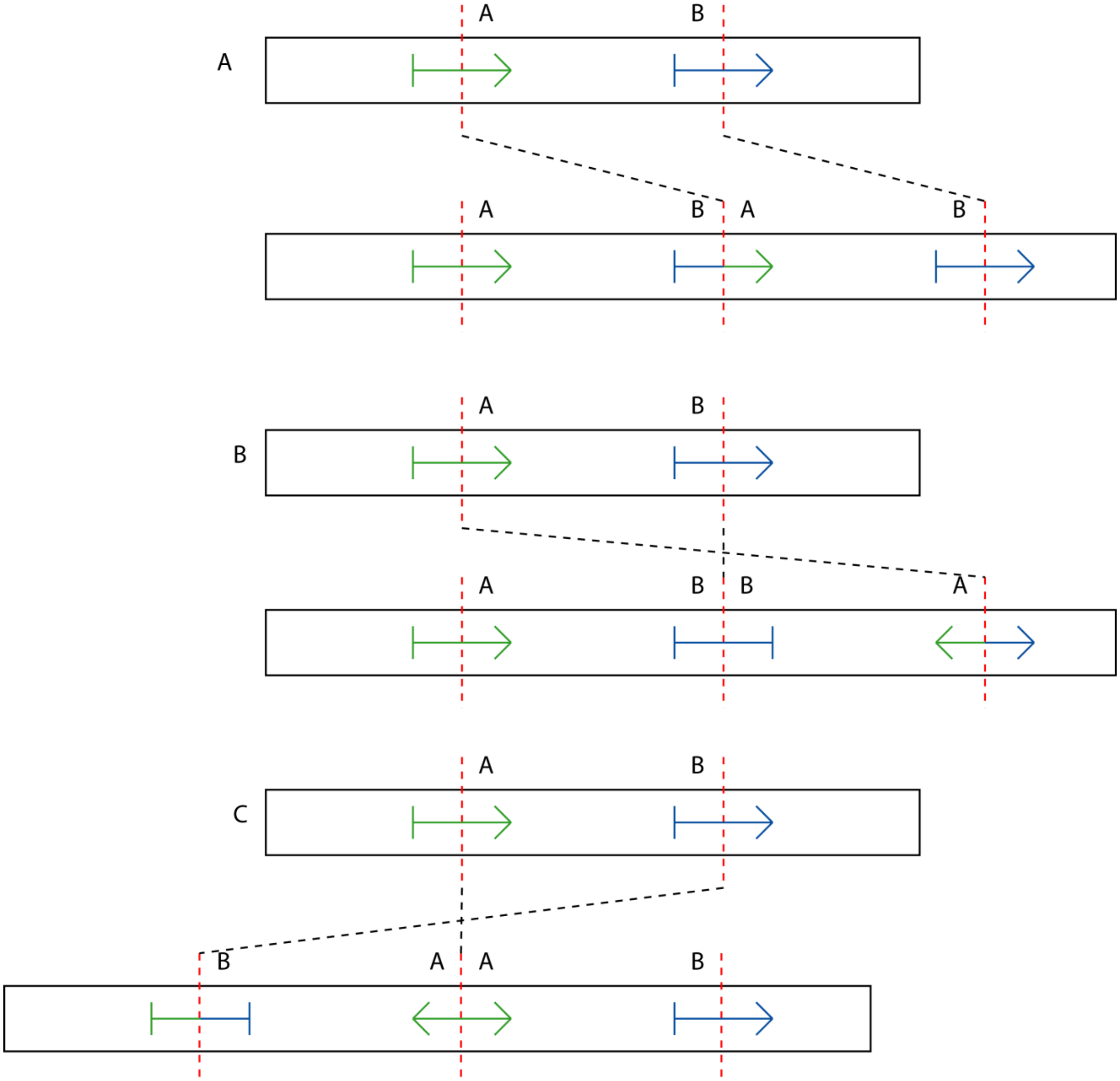


Figure 4.17. The possible fusion genes produced by a tandem duplication which breaks two genes on the same strand.A - a head-to-tail duplication which produces a fusion of the 5’ end of the blue gene to the 3’ end of the green gene. B - a head-to-head duplication which produces no fusion products. C - the other possible head-to-head duplication which also produces no fusion products.

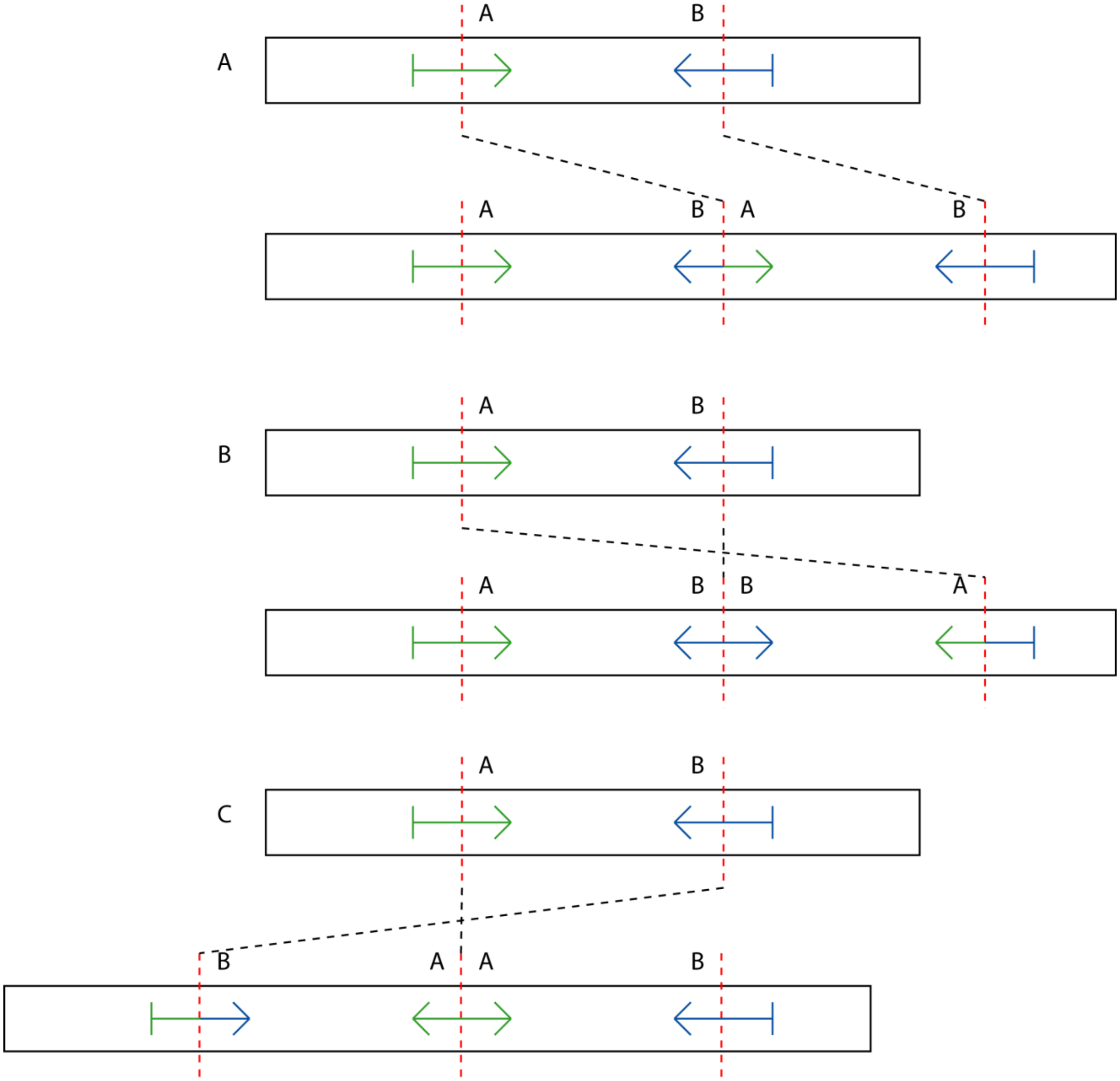


Figure 4.18. The possible fusion genes produced by a tandem duplication which breaks two genes on opposite strands, duplicating the 3’ ends of both genes. A - a head-to-tail duplication which produces no fusion product. B - a head-to-head duplication which produces a fusion of the 5’ end of the green gene with the 3’ end of the blue gene. C - the other possible head-to-head duplication which produces a fusion of the 5’ end of the blue gene and the 3’ end of the green gene.

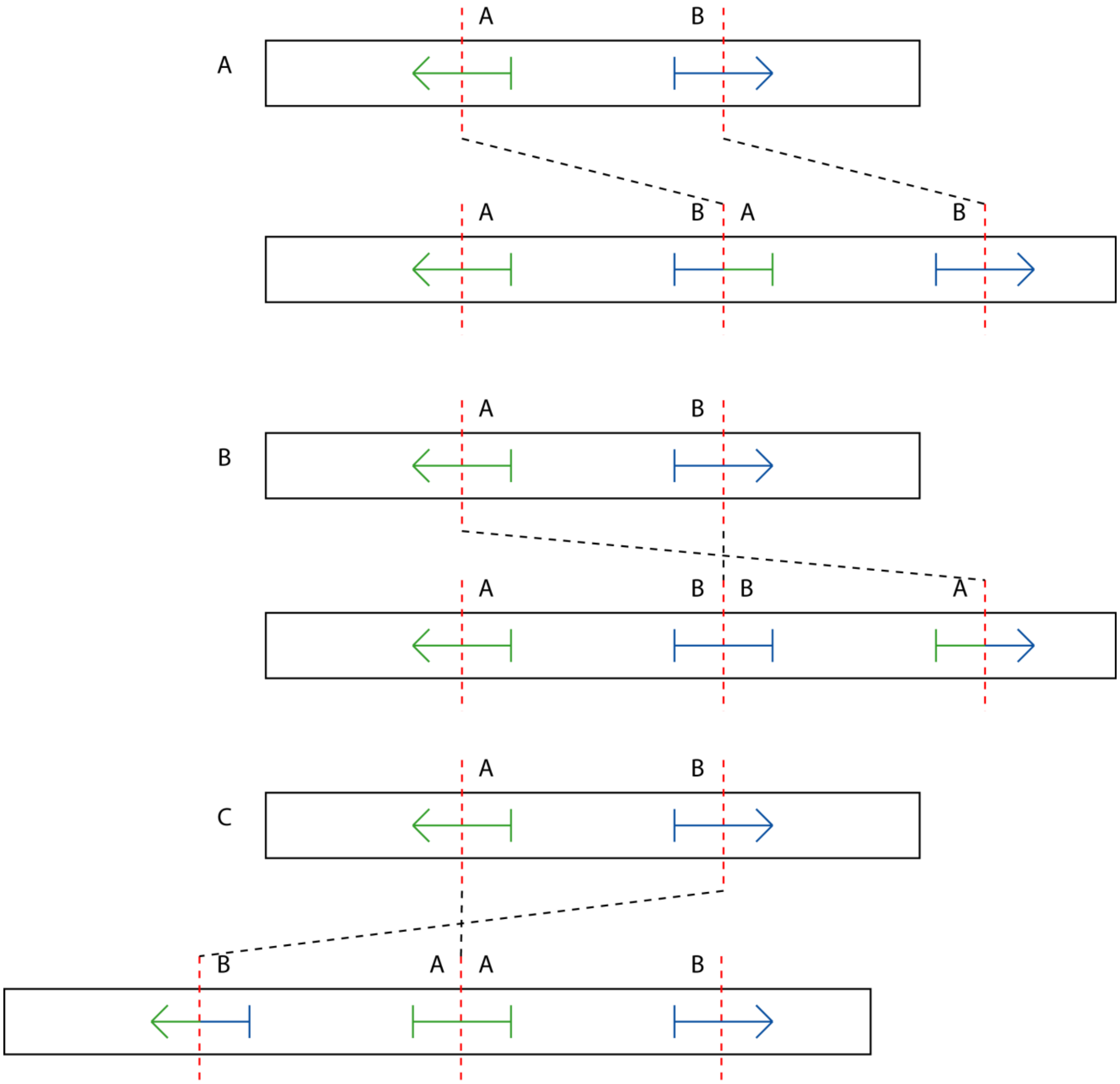


Figure 4.19. The possible fusion genes produced by a tandem duplication which breaks two genes on opposite strands, duplicating the 5’ ends of both genes. A - a head-to-tail duplication which produces no fusion product. B - a head-to-head duplication which produces a fusion of the 5’ end of the green gene with the 3’ end of the blue gene. C - the other possible head-to-head duplication which produces a fusion of the 5’ end of the blue gene and the 3’ end of the green gene.

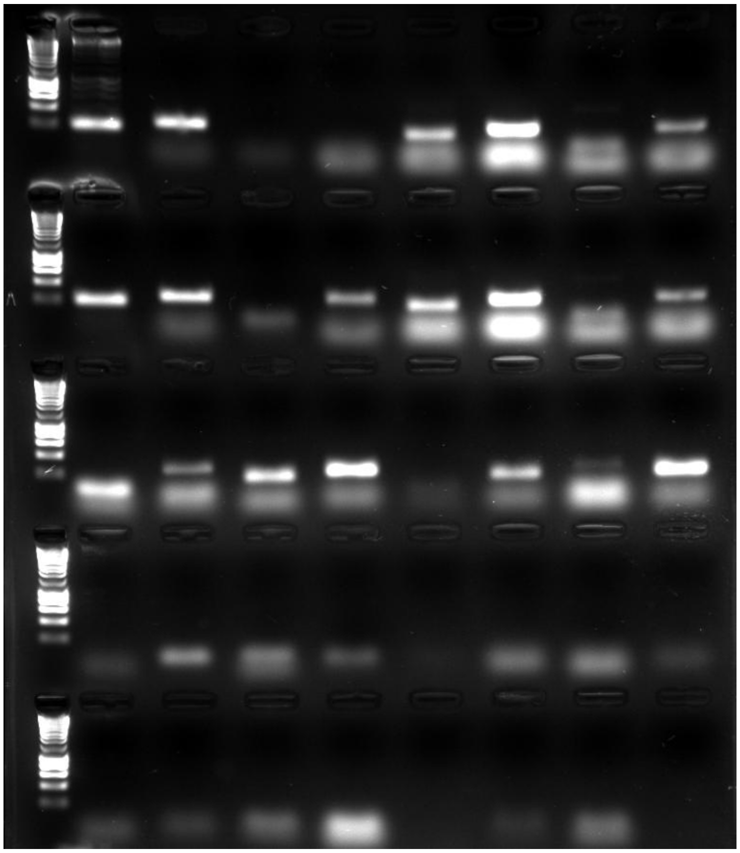


Figure 4.20. PCR for fusions produced by tandem duplications. All PCR was on HCC1806 cDNA.

Row	Column	Primers
1	1	Ladder
1	2	EPHB2 pair a
1	3	EPHB2 pair b
1	4	MYOM3 pair a
1	5	MYOM3 pair b
1	6	FUSIP1
1	7	PNRC2
1	8	AFF3 pair a
1	9	AFF3 pair b
2	1	Ladder
2	2	BC156887 pair a
2	3	BC156887 pair b
2	4	c6orf105 pair a
2	5	c6orf105 pair b
2	6	PHACTR1 pair a
2	7	PHACTR1 pair b
2	8	LAMA2 pair a
2	9	LAMA2 pair b
3	1	Ladder
3	2	ARHGAP18 pair a
3	3	ARHGAP18 pair b
3	4	CATSPERB
3	5	TC2N
3	6	SMURF2
3	7	CCDC46
3	8	GPC3
3	9	HS6ST2
4	1	Ladder
4	2	EPHB2/MYOM3
4	3	EPHB2/FUSIP1
4	4	EPHB2/PNRC2
4	5	MYOM3/EPHB2
4	6	FUSIP1/EPHB2
4	7	PNRC2/EPHB2
4	8	AFF3/BC156887
4	9	BC156887/AFF3
5	1	Ladder
5	2	c6orf105/PHACTR1
5	3	PHACTR1/c6orf105
5	4	LAMA2/ARHGAP18
5	5	ARHGAP18/LAMA2
5	6	CATSPERB/TC2N
5	7	SMURF2/CCDC46
5	8	GPC3/HS6ST2

4.3 Discussion

HCC1806 is described as derived from an acantholytic squamous cell carcinoma of the breast (Gazdar et al., 1998). Acantholytic squamous cell carcinoma is a rare form of breast cancer, which accounts for only 0.05% of all breast neoplasms. Although acantholytic squamous cell carincomas are rare, a small percentage of invasive ductal carcinomas show regions of squamous cell metaplasia (Fisher et al., 1983), so HCC1806 may be a rare variant of a true ductal carcinoma. CGH studies which have been carried out on small numbers of acantholytic squamous cell carcinomas suggest that they show some chromosome rearrangements which are characteristic of both breast cancer, and squamous cell tumours from other regions (Aulmann et al., 2005).

The combination of array painting and whole genome SNP6.0 array data made it possible to identify potential fusion genes caused by translocations which could not have been identified using one method alone. The high-resolution SNP6.0 data identified the genes at breakpoints, but the array painting was needed to determine which breaks are found together on a derivative chromosome and may be joined to each other. Some of this information could be inferred from the SKY karyotype, but the problems of resolution and overlap of chromosomes at breakpoints make it difficult to determine the exact arrangement from the SKY data alone.

No further fusion transcripts could be detected in HCC1806. This could be a true negative result and reflect that there are few fusion genes in this cell line and the two known fusions are the only fusion genes. The number of fusion genes found in cell lines and tumours in the Stephens et al. study (2009) ranged from zero to eleven, suggesting that if HCC1806 really has only two fusion genes then this is within the range found in breast cancers. However, this study did not look for readthrough fusion genes, and may underestimate the number of fusions in each cell line (see chapter 6 for details of fusions not found in the cell line HCC1187).

Alternatively, there may be other fusion genes present in HCC1806 which were not found using the methods I have employed. Although these methods have been successfully used to find fusion genes before, including the two known fusions in HCC1806, they may miss fusion genes which are expressed at very low levels which cannot be detected using standard PCR. The assumption is that fusion genes which are barely expressed are unimportant, but they may act as dominant negative inhibitors of one gene in the fusion even at low levels. Fusion genes that have unusual splicing patterns and do not include any of the exons tested by PCR would also be missed, as would some fusion genes that included novel exons. Most of the genomic junctions at breakpoints in HCC1806 have not been sequenced. It is possible that the breakpoints are more complex than they appear and may contain 'genomic shards' (Bignell et al., 2007; Campbell et al., 2008), small pieces of DNA which have been inserted at the breakpoint. These pieces are often smaller than a kilobase and would not be seen on the SNP6.0 array, but in rare cases could affect any fusion gene produced by a chromosome rearrangement. Another possibility is that there is an inversion at the breakpoint, like the known inversion at a breakpoint in the breast cancer cell line T47D (Pole et al., 2006), which are copy number neutral and would not be identified using microarrays.

Chapter 5

The complete karyotype of MDA-MB-134 obtained using array painting and high-throughput sequencing

5.1 Introduction

MDA-MB-134 is a breast cancer cell line derived from a pleural effusion obtained from a patient with metastatic breast cancer (Cailleau et al., 1974). It is a hypodiploid line with a median of 44 chromosomes. There is a subclone which has endoreduplicated and subsequently lost chromosomes, with a median of 66 chromosomes. The main rearrangements are two copies of a large marker chromosome with amplification of chromosome 8 and 11, and the der(15)t(15:17) and der(18)t(16:18) translocations (Figure 5.1) (Davidson et al., 2000).

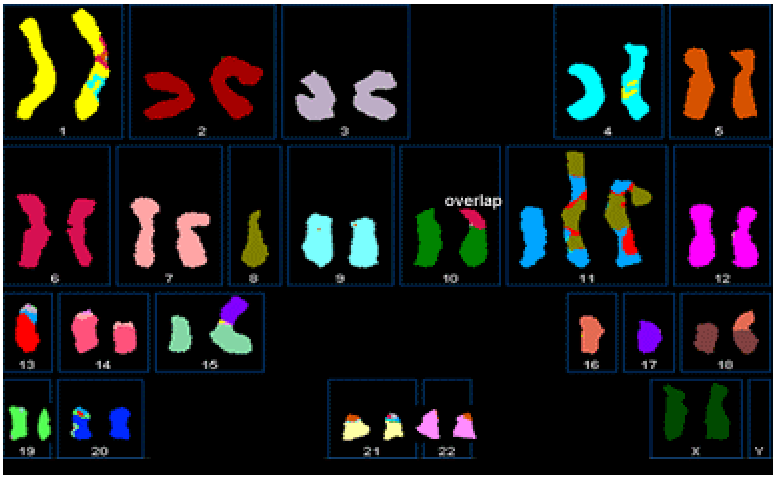


Figure 5.1. SKY karyotype of MDA-MB-134 (Davidson et al., 2000).

The aim of the work was to map all the chromosome rearrangements in MDA-MB-134 to high resolution, and search for gene fusions or other rearrangements which affect gene expression and function. Chromosome 8 and 11 amplifications are common in breast cancer (Lafage et al., 1992; Lemieux et al., 1996; Bautista and Theillet, 1998; Paterson et al., 2007), and 8p12 and 11q13 are found co-amplified in 8.2% of cases (Letessier et al., 2006). MDA-MB-134 is a model for chromosome 8 and 11 amplification and has a simple karyotype with few other translocations, suggesting that the rearrangements in the amplicon are important driving events in carcinogenesis, and they will be easier to analyse in a cell line with a low level of other rearrangements.

5.1.1 Previous work

The large homogenously staining region of the marker chromosome was shown to contain sequences from chromosomes 8p11-12 and 11q13 arranged in a complex structure (Lafage et al., 1992). Microdissection of the hsr and hybridization to normal metaphases suggested that sequences from 8p12 and 11q13 were co-amplified, with a block of amplified DNA from 8q24 between the co-amplified regions (Guan et al., 1994). Further FISH suggested a rearrangement between the centromere of chromosome 11 and the juxtacentromeric region of chromosome 8 (Lemieux et al., 1996). A mostly complete copy of 8q forms the short arm of the marker chromosome, with a deletion of MYC, and there is no amplification of the copy of MYC located between the co-amplified regions.

Array CGH shows amplification of 8p12 and 11q13 with few other copy number changes. The amplicon on 8p12 is large and spans 7Mb of chromosome 8 (from 34.7 to 41.5Mb) and includes FGFR1 and ZNF703, while the chromosome 11 amplicon is smaller and contains two separate regions of amplification, one covering CCND1 and the other containing *EMSY* and *GARP* (Paterson et al., 2007). FISH shows that the amplicon has a complex interdigitated structure, with two blocks of 8p12 and 11q13 amplification separated by a region from 8q, and the chromosome with the amplification is present in two copies (Figure 5.2). Overlapping signals from chromosome 8 and 11 are seen, but the complete arrangement of the amplicon could not be derived using FISH.

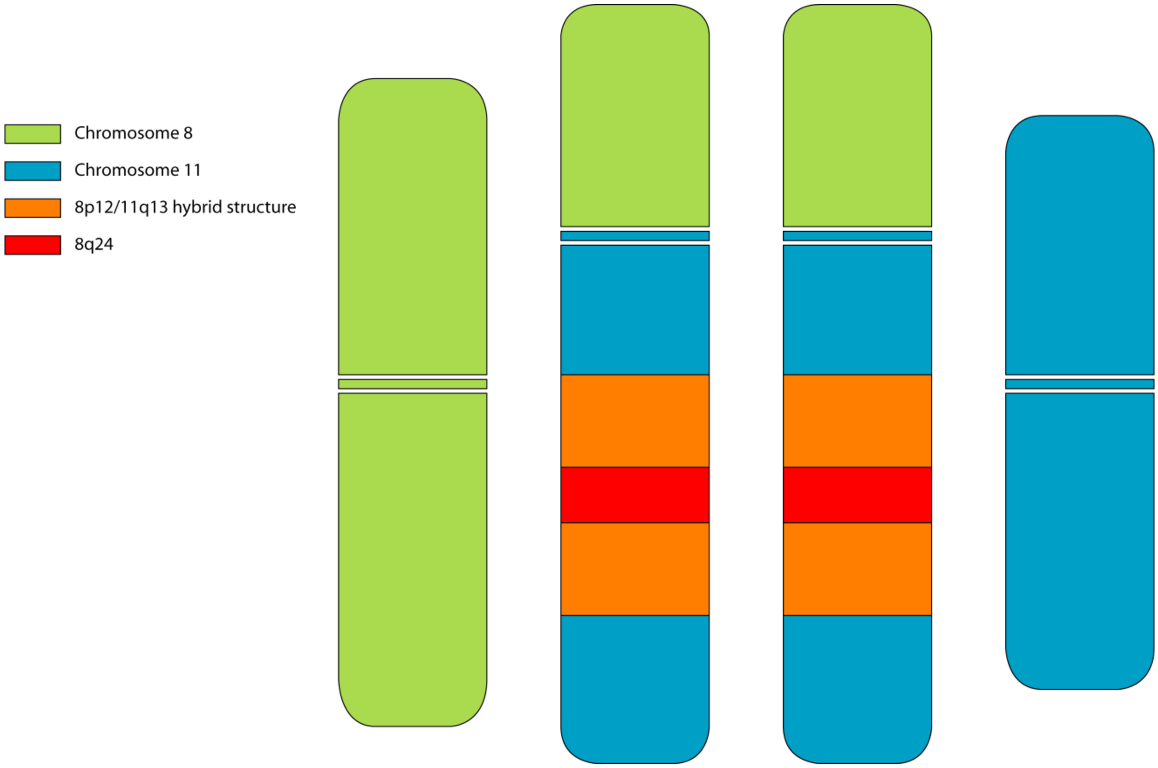


Figure 5.2. Schematic of the chromosome 8 and 11 amplifications in MDA-MB-134. There is a single copy of normal 8 and 11, and two copies of the der(11) marker chromosome with two regions of intermingled 8p12/11q13 separated by a single copy of material from 8q24. (Adapted from Paterson et al. 2007).

5.2 Results

5.2.1 Assembling a complete karyotype of MDA-MB-134

The aim was to use array painting to characterize the chromosomal rearrangements which could be seen on the SKY karyotype. The SKY karyotype suggested there were three rearranged chromosomes – the der(15)t(15:17), der(18)t(16:18), and two copies of the der(8)t(8;11), which previous work suggested were identical to the resolution of FISH and SKY and would be in the same position in the flow karyotype.

Comparison of the chromosome fractions of MDA-MB-134 to the chromosomes sorted from a normal human cell line showed six fractions in an abnormal position on the flow sort (Figure

5.3). These fractions may contain the rearranged chromosomes, or they may be outlying regions of the normal chromosome fractions, as the fractions are not as tightly sorted as the normal comparison. Fractions A and B were both collected as it was not possible to determine which was the der(8)t(8;11) and which was the normal chromosome 1 from their position on the flow sort. Fraction F was collected as, while it is common for two different homologues of chromosome 21 to form separate fractions, as can be seen in the normal flow sort (Figure 3A), it could not be determined from the flow sort whether it was a different homologue of chromosome 21 or one of the rearranged chromosomes. Fractions C, D and E were collected as they were in an unexpected position, and may be either rearranged chromosomes or outliers from the normal fractions.

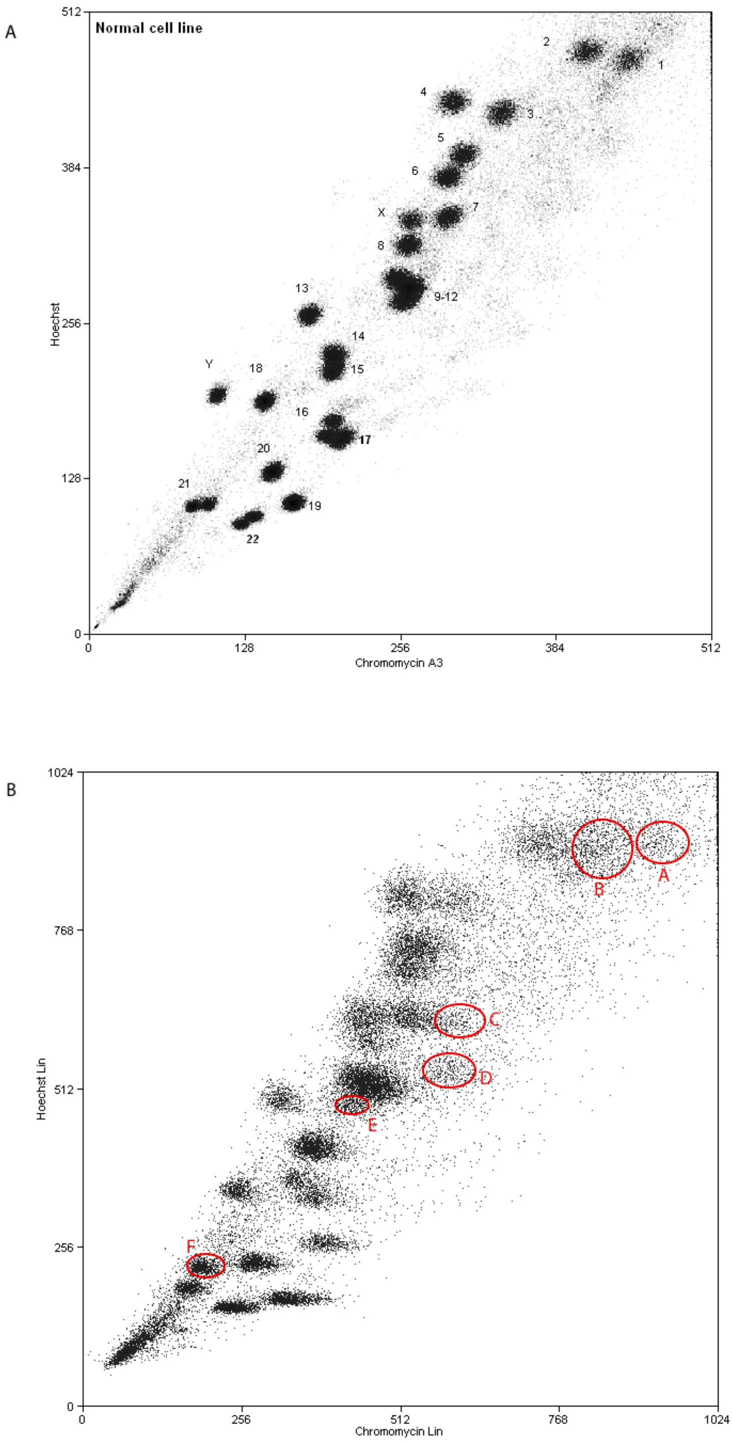


Figure 5.3. Flow karyotyping of abnormal chromosomes in MDA-MB-134A: Flow karyotype of chromosomes from the normal cell line GM11321B with chromosome fractions labelled (normal karyotype courtesy of Bee Lin Ng, Wellcome Trust Sanger Institute). B: Flow karyotype of chromosomes from MDA-MB-134. The six labelled fractions look to be in an abnormal position on the graph compared to the normal chromosomes, and may be the fractions containing chromosomes with translocations.

To determine which of the six candidate fractions contained the rearranged chromosomes, they were reverse chromosome painted to normal metaphases. The sorted chromosomes were amplified using the Genomphi DNA Amplification Kit, and hybridised to normal metaphase spreads.

Four of the candidate chromosome fractions contained normal chromosomes. Fraction B was chromosome 1, fraction D and fraction E were outlying regions of the fractions for chromosomes 9-12 (which co-localise) and chromosome 7 respectively, and fraction F was an extra fraction for 21 caused by the different homologues of the chromosome sorting into separate fractions. Fractions A and C contained two of the expected derivative chromosomes – fraction A was the large t(8;11) chromosome and fraction C was the der(15)t(15;17) (Figure 5.4).

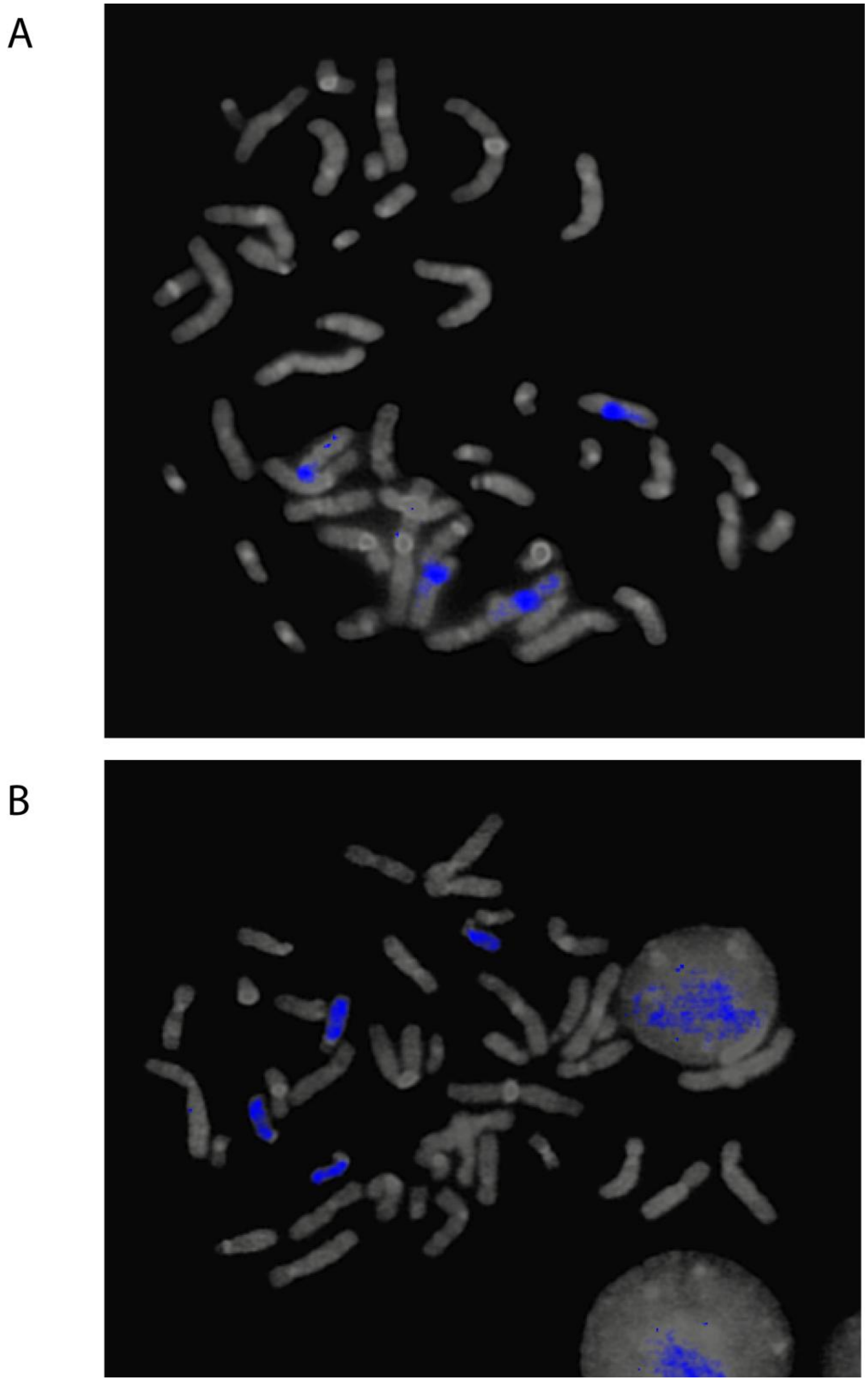
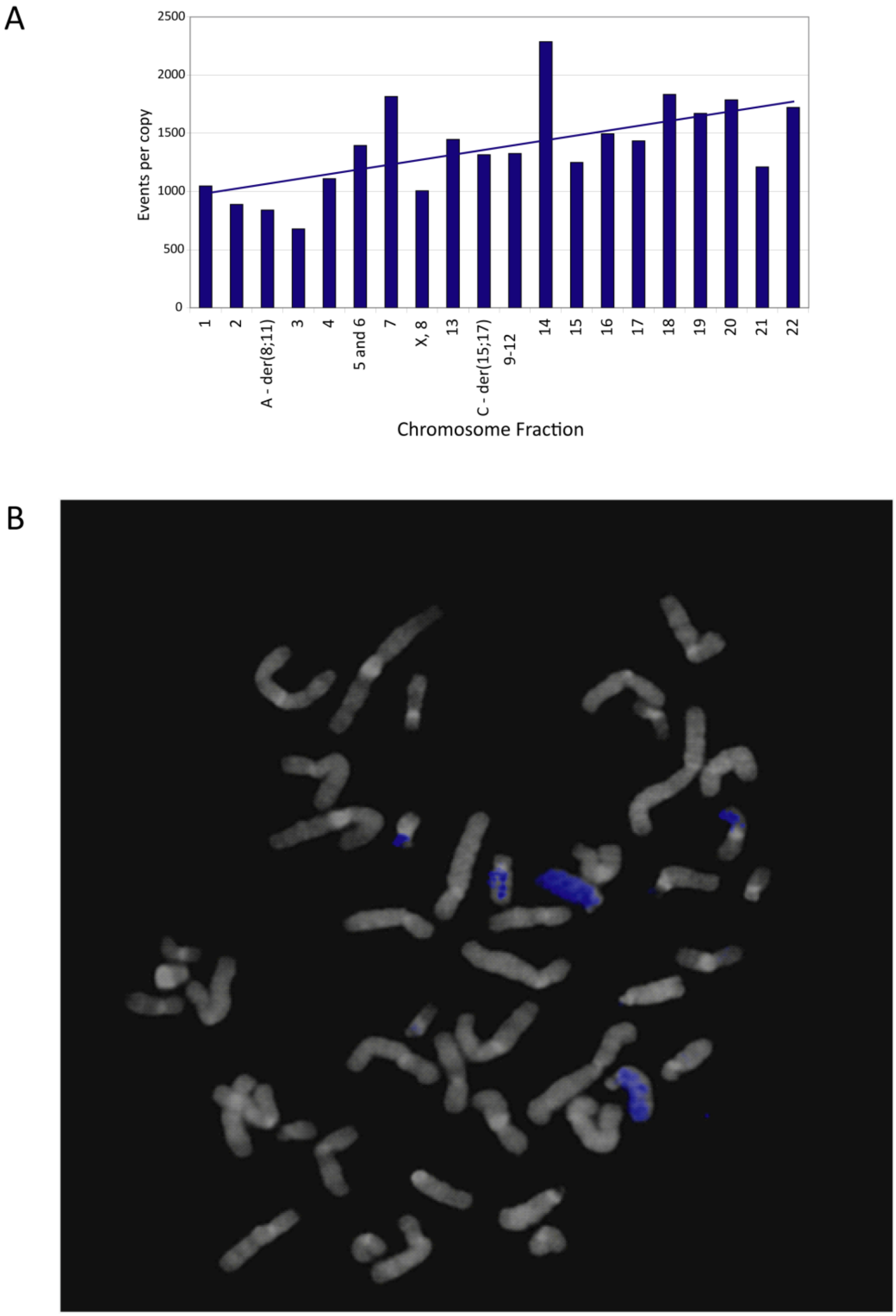


Figure 5.4. Reverse chromosome painting of sorted chromosome fractions to normal (DRM) metaphases. A – chromosome fraction A, showing signal on chromosomes 8 and 11. B – chromosome fraction C showing signal on chromosomes 15 and 17.

The der(18)t(16;18) was not found in any of the candidate abnormal fractions. It was possible that the size of the rearranged chromosome caused it to co-localise on the flow karyotype with a normal chromosome. If this was the case, the count of the number of events in that chromosome fraction would be higher than expected, as three rather than two chromosomes. Figure 5.5A shows the count of events in each chromosome fraction. The trend towards more events in the fractions for the shorter chromosomes is due to more of the shorter chromosomes being retained during the preparation for flow sorting, but even accounting for that trend the fraction for chromosome 14 showed an unusually high number of events for a fraction which should contain 2 chromosomes – over 2200 events when 1400 would be expected. This suggested that the der(18)t(16;18) chromosome may be contained in the chromosome 14 fraction, and reverse chromosome painting showed that it hybridized to the whole of chromosome 14, the p arm of chromosome 16, and the q arm of chromosome 18 (Figure 5.5B).

Figure 5.5. Locating the der(16)t(16;18) chromosome in MDA-MB-134. A **-** a graph showing the counts of each chromosome fraction in MDA-MB-134 during chromosome sorting, showing an more than the expected number of chromosomes in the chromosome 14 fraction . The trendline shows that more of the longer chromosomes are lost during the preparation for sorting. B - reverse painting of the chromosome 14 fraction to normal (DRM) metaphases, showing signal on chromosomes 14, 16, and 18.


A
2500
2000
1500
1000
500
0
Events per copy
1
2
A - der(8;11)
3
4
5 and 6
7
X, 8
13
C - der(15;17)
9-12
14
15
16
17
18
19
20
21
22
Chromosome Fraction
B

5.2.2 Array painting of rearranged chromosomes

The breakpoints on the three rearranged chromosomes were further mapped by array painting of the sorted chromosome fractions. The arrays used were 1Mb BAC arrays with 3,439 probes spread across the genome, giving a probe approximately every megabase, and 90 probes at higher density between 30.9 and 41.4Mb on chromosome 8, giving a probe on average every 120Kb across this region. Amplified sorted chromosomes were labeled and hybridized to arrays against labeled normal female DNA, and the ratio of the signals was used to find regions which were present or absent in the sorted chromosome compared to normal. The aim of using array painting instead of whole genome CGH was that the breakpoints could be unambiguously assigned to the rearranged chromosome as only the chromosomes present in each chromosome fraction are hybridized to the array.

As reverse chromosome painting showed that two of the sorted fractions contained only the derivative chromosome, while the t(16;18) co-localised with chromosome 14, which is not known to be involved in the rearrangement, the breakpoints could be unambiguously assigned to a particular derivative chromosome.

Array painting of the fraction containing the der(15)t(15;17) showed the breakpoints on each chromosome to be centromeric (Figure 5.6). The t(15;17) is formed of 15q joined to 17q. The array has no probes on the p arm of chromosome 15, but the whole of the q arm was retained, so the breakpoint is likely centromeric.

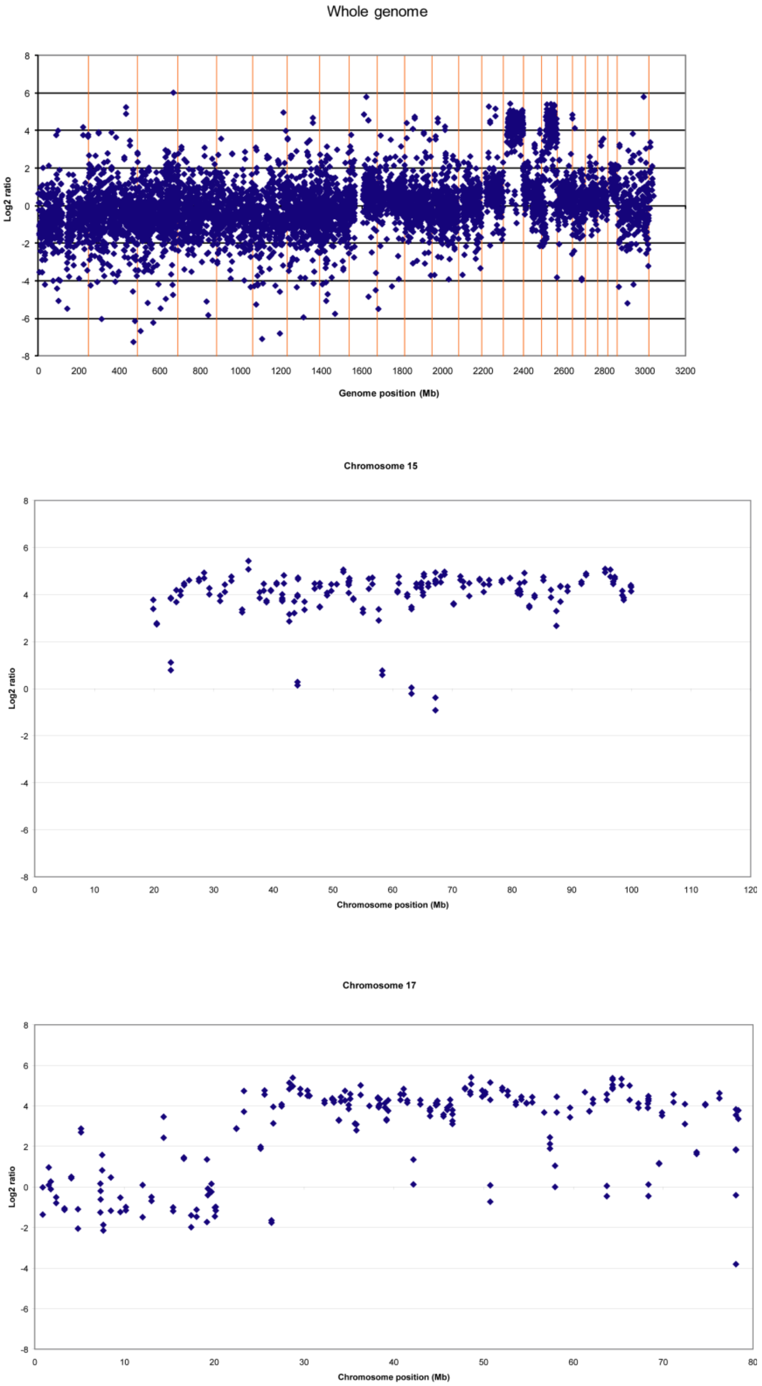


Figure 5.6. Array painting of MDA-MB-134 chromosome fraction C. Plots show the $\log_2$ ratio of the hybridization of the test chromosome fraction against a normal reference genome versus the distance along the genome or chromosome. From top to bottom: whole genome plot, chromosome 15, chromosome 17. The array has no probes on the p arm on chromosome 15. The array is known to have a number of misidentified BACs, which probably account for the probes which do not show the expected signal. Each probe is duplicated on the array, and both copies are plotted separately on these graphs.

Array painting of the fraction containing the der(18)t(16;18) (as well as chromosome 14) showed that the breakpoint on chromosomes 16 and 18 was also centromeric (Figure 5.7).

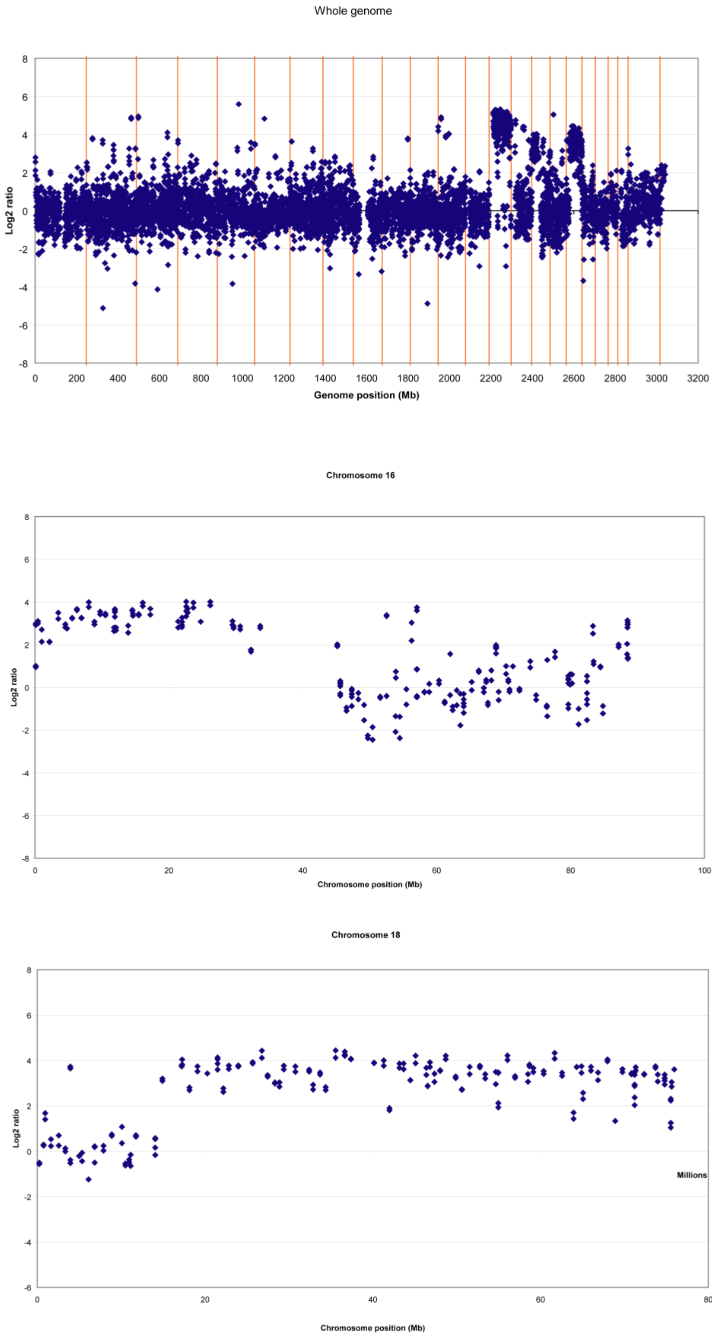


Figure 5.7. Array painting of MDA-MB-134 fraction 14. Plots show the $\log_2$ ratio of the hybridization of the test chromosome fraction against a normal reference genome versus the distance along the genome or chromosome. From top to bottom: whole genome plot, chromosome 16, chromosome 18. The whole genome plot shows signal from chromosome 14, as it co-localises with the derivative chromosome during flow sorting and cannot be separated.

Array painting showed that the translocations involve whole chromosome arms but could not determine whether the chromosome arms were joined at centromeres, telomeres, or had more complicated rearrangements which did not affect the copy number. To determine the orientation of the translocated fragments, FISH was performed using probes near the centromeres of the chromosome fragments (Figures 5.8 and 5.9).The derivative chromosomes were joined at the centromeres in both the 15;17 and 16;18 translocations, and did not show a telomeric fusion.

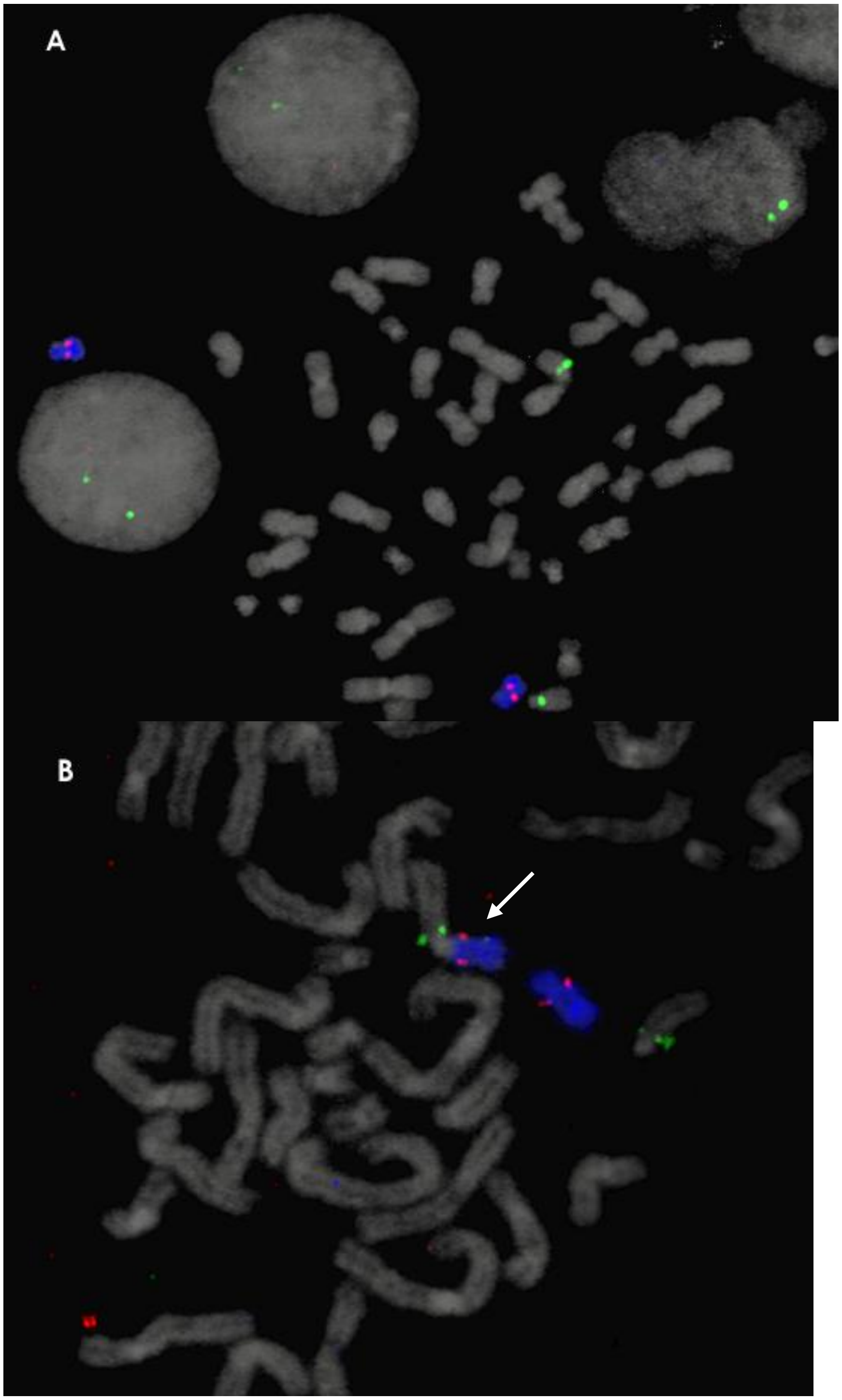


Figure 5.8. FISH to show orientation of chromosome fragments. Chromosome 17 paint labelled with Spectrum Orange (blue). Probe on 15q12 is RP11-570N16, labelled with Digoxygenin (green). Probe on 17q11.2 is RP11-403E9, labelled with Biotin (red). A is a normal (DRM) metaphase, B is an MDA-MB-134 extended chromosome preparation metaphase. Arrow shows the der(15)t(15;17) chromosome is formed of 15q and 17q chromosome fragments joined at the centromeres. There is no telomeric fusion.

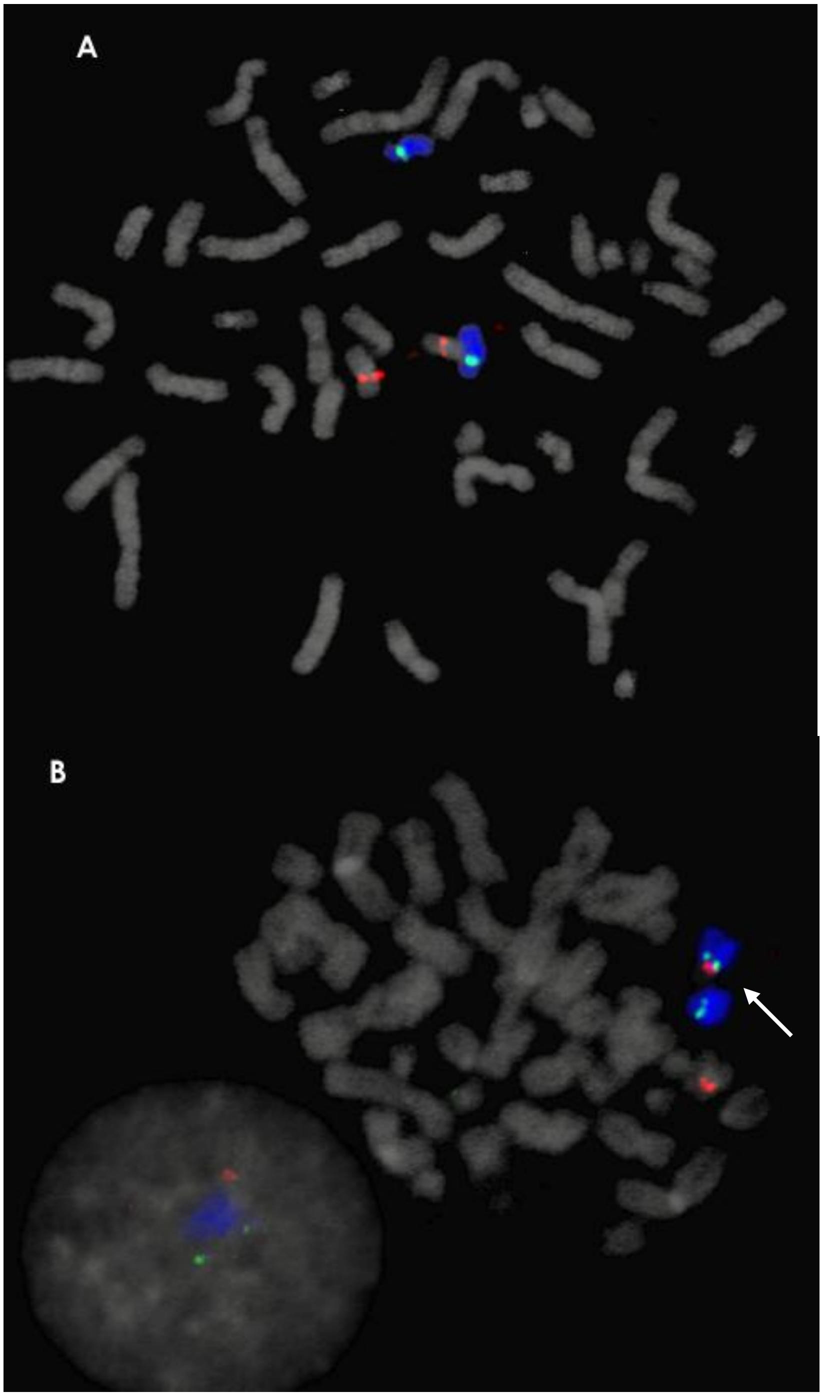


Figure 5.9. FISH to show orientation of chromosome fragments. Chromosome 18 paint labelled with Spectrum Orange (blue). Probe on 18q11.1 is RP11-280C8 labelled with Digoxygenin (green). Probe on 16p11.2 is RP11-2C24 labelled with Biotin (red). A is a normal (DRM) metaphase, B is an extended MDA-MB-134 metaphase. The green signal is on 18q near the centromere, and the red signal is on 16q near the centromere.

Array painting of the t(8;11) derivative chromosome allowed the amplified regions and breakpoints to be determined to higher resolution than through previous FISH and microarray studies (Figure 5.10). Using the criterion that a minimum of two adjacent probes showing a change in hybridisation ratio are needed to confirm a gain or loss, the 8p amplification appeared to have two separate regions of gain, from 30,945,723-31,288,495Mb and from 34,631,328-40,796,500Mb. The positions were taken from the start and end points of the BAC probes which are gained, but as the probes were spaced at megabase intervals the breakpoints could only be determined approximately. There was an extra copy of the whole of 8q, with two deletions between 109,137,354-122,721,235Mb and 134,255,043-139,307,409Mb.

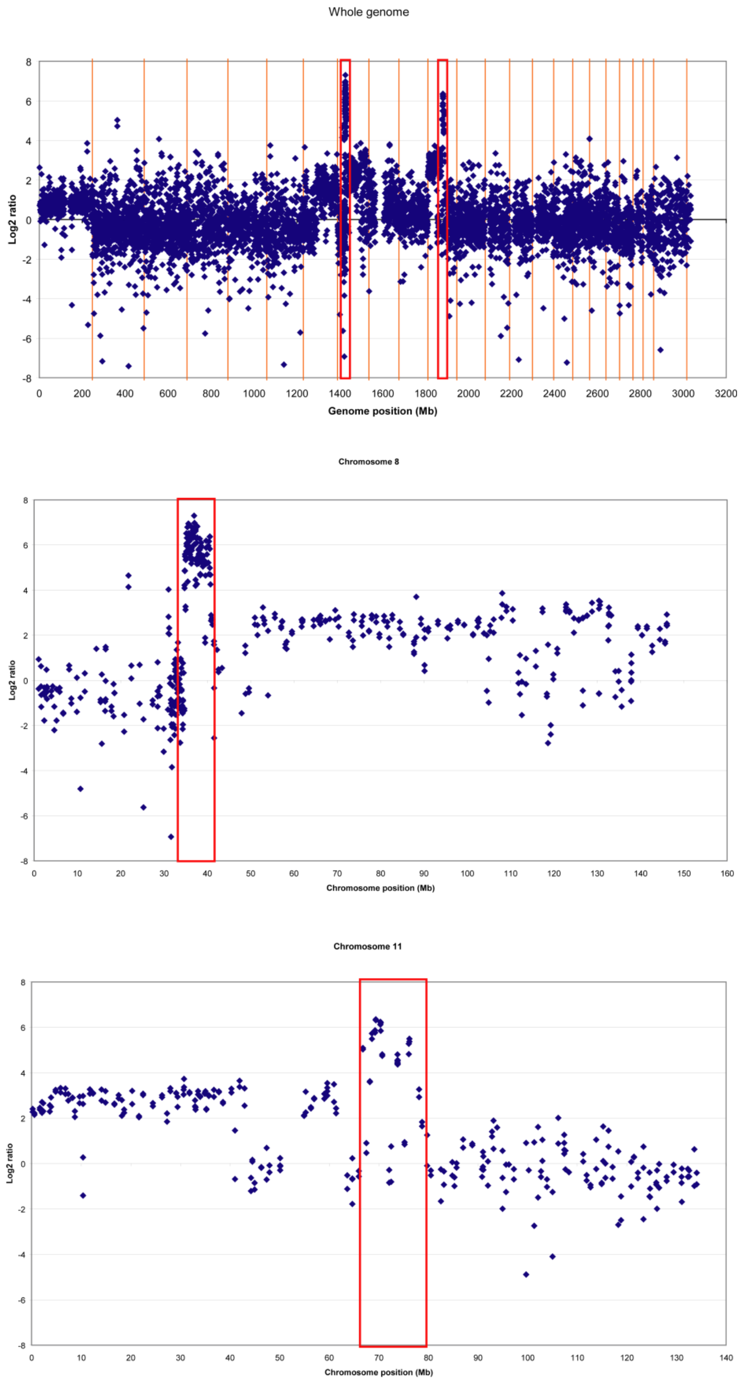


Figure 5.10. Array painting of MDA-MB-134 fraction A. Plots show the $\log_2$ ratio of the hybridization of the test chromosome fraction against a normal reference genome versus the distance along the genome or chromosome. From top to bottom: whole genome plot, chromosome 8, chromosome 11. The signal from chromosome 1 and chromosome 7 seen in the whole genome plot is contamination of the flow-sorted chromosome fraction. The boxes mark the amplicons on chromosomes 8 and 11.

Chromosome 11 showed a gain of most of 11p from the start to 42,018,028Mb, and the q arm had gained a region between the centromere and 62,403,825Mb. The amplicon on 11q showed a higher $\log_2$ ratio of signals, indicating more copies of this region had been gained, and the amplicon could be divided into two regions, 68,278,585-70,789,798Mb, and 73,676,966-78,791,818Mb. There may also be further rearrangements within the amplicon which could not be confirmed at this resolution, as two adjacent probes could not be called as gained or lost.

The extent of the amplification was subsequently confirmed using data from Affymetrix SNP6.0 arrays (Bignell et al., 2007), which has probes on average every 700bp and allows breakpoints to be called to much higher resolution. The low-resolution array painting agreed with the SNP6 data (Figure 5.11), except for the small amplicon around 31Mb on chromosome 8, which was called from two BAC probes and may be due to mismapped BACs . However, further rearrangements suggested by single probe changes on the 1Mb array could be confirmed on the higher resolution array, including a further high-level amplification on chromosome 8 between 21,375,101 and 21,983,001Mb. This amplification includes FGF17, which is overexpressed in prostate cancer and associated with poor prognosis (Heer et al., 2004). A summary of the chromosome rearrangements found on chromosomes 8 and 11 is shown in figures 5.12 to 5.14.

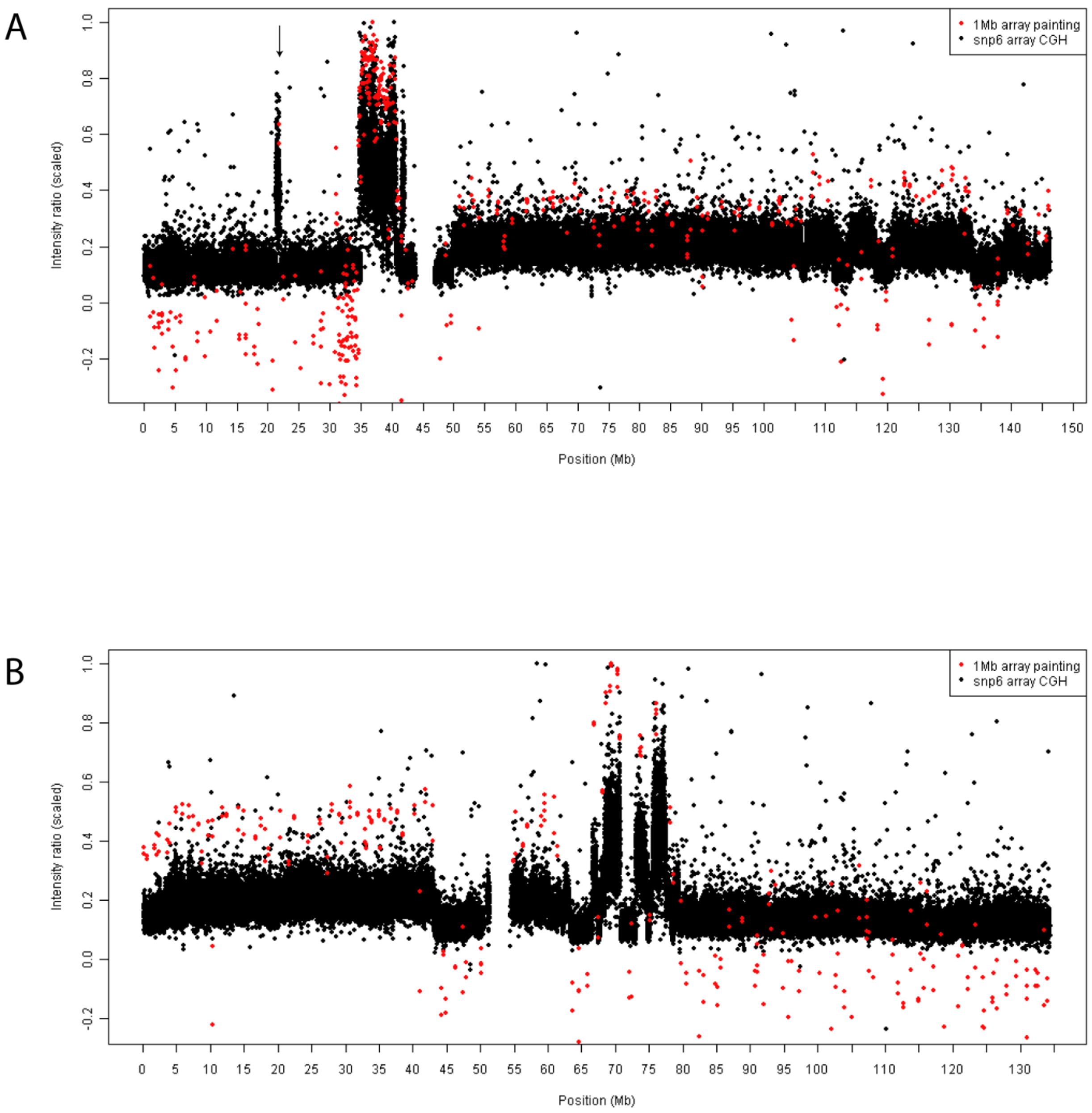


Figure 5.11. Comparison of 1Mb array painting data (in red) and whole-genome SNP6.0 data (in black). Data has been scaled to allow comparison of the two data sets. A – MDA-MB-134 chromosome 8. The two arrays agree on the extent of the 8p amplification, and there is an additional amplification at 21.3-21.9Mb (indicated by the arrow) which is represented on the 1Mb array by a single BAC. Each BAC is present in duplicate on the array, so the 2 points in the amplification represent the same BAC. B – MDA-MB-134 chromosome 11, showing agreement between the two arrays on the amplification.

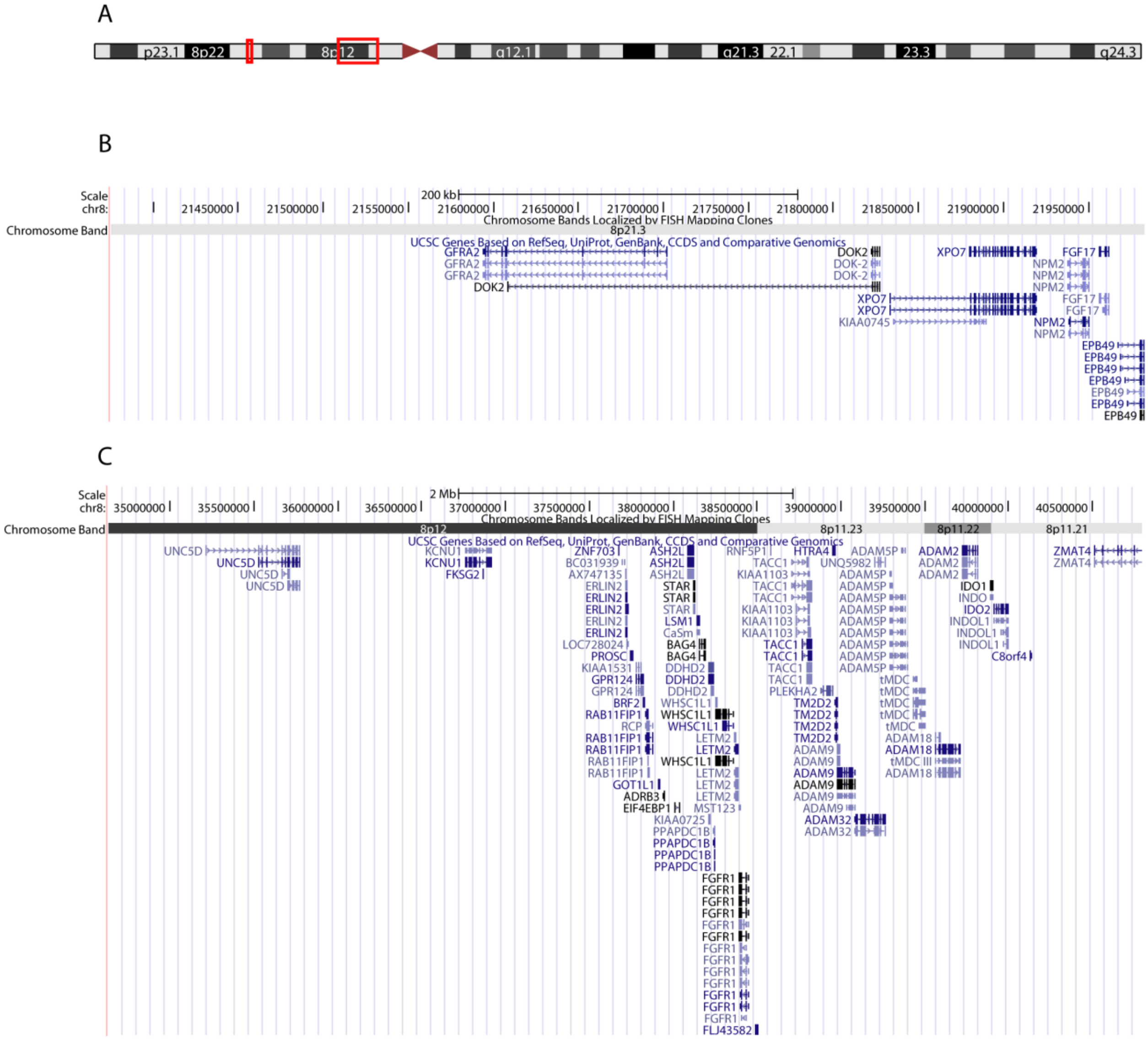


Figure 5.12. Amplifications on chromosome 8 in MDA-MB-134 found using array painting and confirmed by SNP6.0 arrays. A – Ideogram of chromosome 8. The red boxes mark the amplified regions found in the t(8;11) chromosome in MDA-MB-134. B – the genes found in the smaller amplified region 21.3-21.9Mb. C – the genes found in the larger amplified region 34.6-40.7Mb, including ZNF703 and FGFR1.

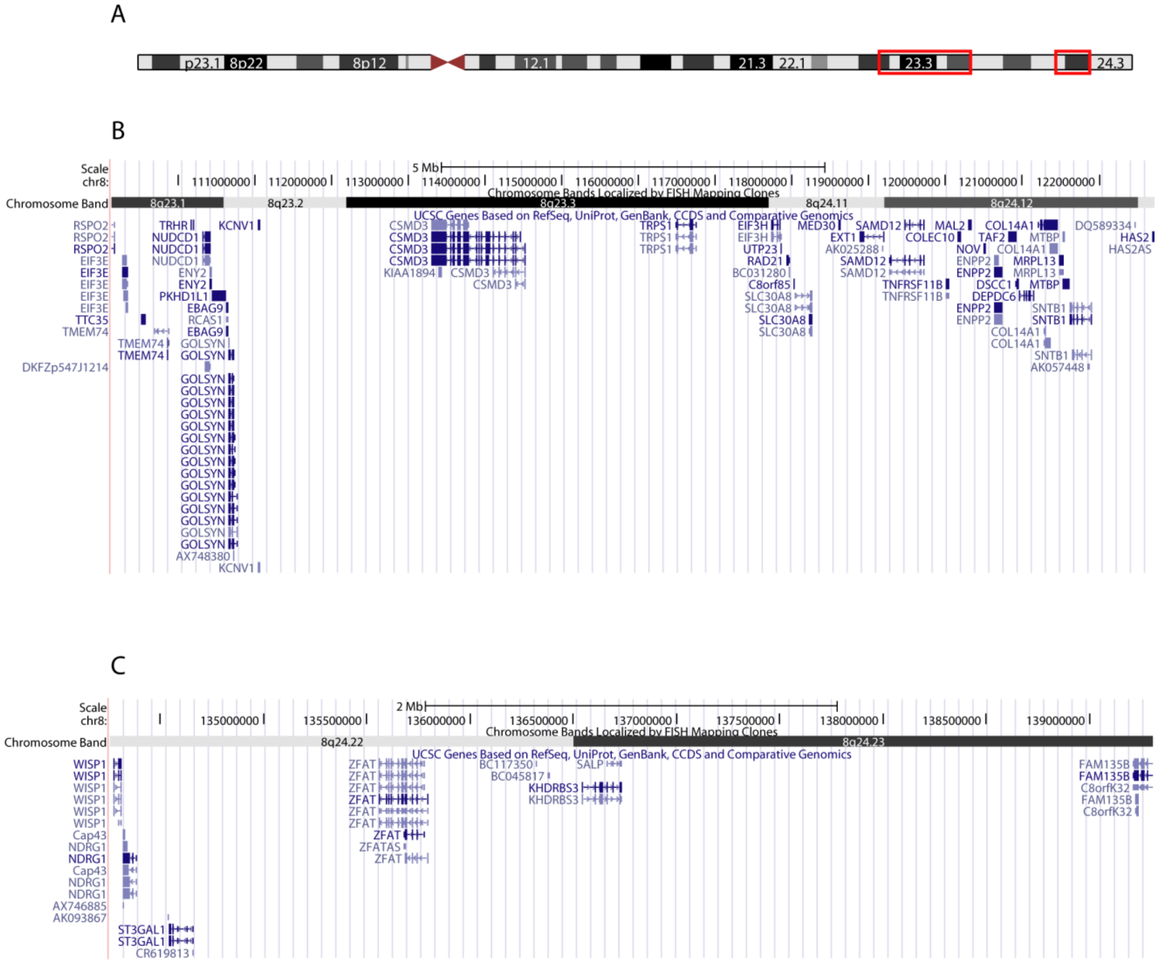


Figure 5.13. Deletions on chromosome 8 in MDA-MB-134 found using array painting and confirmed by SNP6.0 arrays. A – Ideogram of chromosome 8. The red boxes mark the deletions found in the t(8;11) chromosome in MDA-MB-134. B – the genes found in the larger deleted region 109.1-122.7Mb. C – the genes found in the smaller deleted region 134.2-139.3Mb.

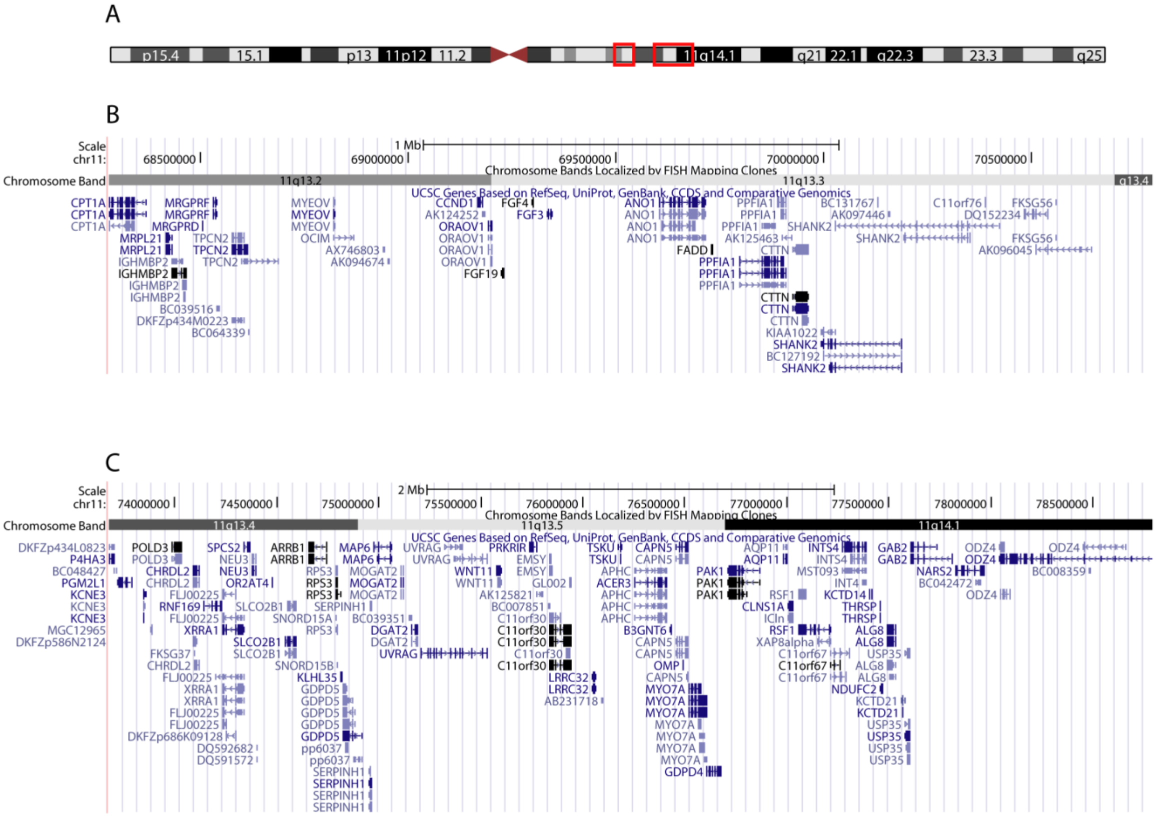


Figure 5.14. Amplifications on chromosome 11 in MDA-MB-134 found using array painting and confirmed by SNP6.0 arrays. A – Ideogram of chromosome 11. The red boxes mark the amplifications found in the t(8;11) chromosome in MDA-MB-134. B – the genes found in the larger amplified region, 68.3-70.8Mb, including CCND1. C – the genes found in the smaller amplified region, 73.7-78.8Mb.

5.2.3 High-throughput sequencing

Array-based approaches to mapping the amplifications allowed me to resolve the positions of the breakpoints to high resolution, but could not tell which breakpoints are joined together, which is essential to find rearrangements which may cause fusion genes.

High-throughput paired-end sequencing overcomes many of the limitations of array-based mapping of structural variation. It can be used to map many of the structural variants in a cell line in a single experiment, depending on the sequence coverage obtained. High-throughput sequencing gives millions of short sequence reads from across the genome in one experiment by sequencing small fragments of the genome in a massively parallel process. Paired-end sequencing gives reads from both ends of the fragment, which can be aligned to a human reference genome. Any fragments which contain a genomic rearrangement can be easily identified as they will cause a change in the size or orientation of the fragment relative to the reference genome, and can be easily identified. As the fragment straddles the rearrangement, both sides of the rearrangement can be identified, as well as the orientation of the DNA. Identification of rearrangements is not dependent on copy number changes, allowing balanced rearrangements to be identified.

For paired-end sequencing using the Illumina GAII platform, genomic DNA is fragmented and size-selected for the desired fragment size, which can be up to 800bp (Figure 5.15). Adaptors are ligated to each end of the fragments and amplified with 20 cycles of PCR. A sample from the fragment library is placed onto the flow cell where the adaptors adhere to a ‘lawn’ of primers. Each fragment is amplified on the flow cell surface to produce a cluster of identical single-stranded DNA strands. For each cycle of sequencing, four fluorescently-labelled nucleotides and DNA polymerase are added to the flow cell. Each nucleotide has a reversibly-blocked 3’-OH group so that only one base is incorporated at each step. The flow cell is imaged, then the nucleotides are unblocked and another round of sequencing can take place. Each base pair is called from the images with an associated quality score. For paired-end sequencing, each cluster is re-amplified and a second round of sequencing proceeds from the adaptor ligated to the other end of the fragments. This gives paired reads where the first read is from one end of the fragment, and the second read from the opposite end of the fragment (Mardis, 2008).

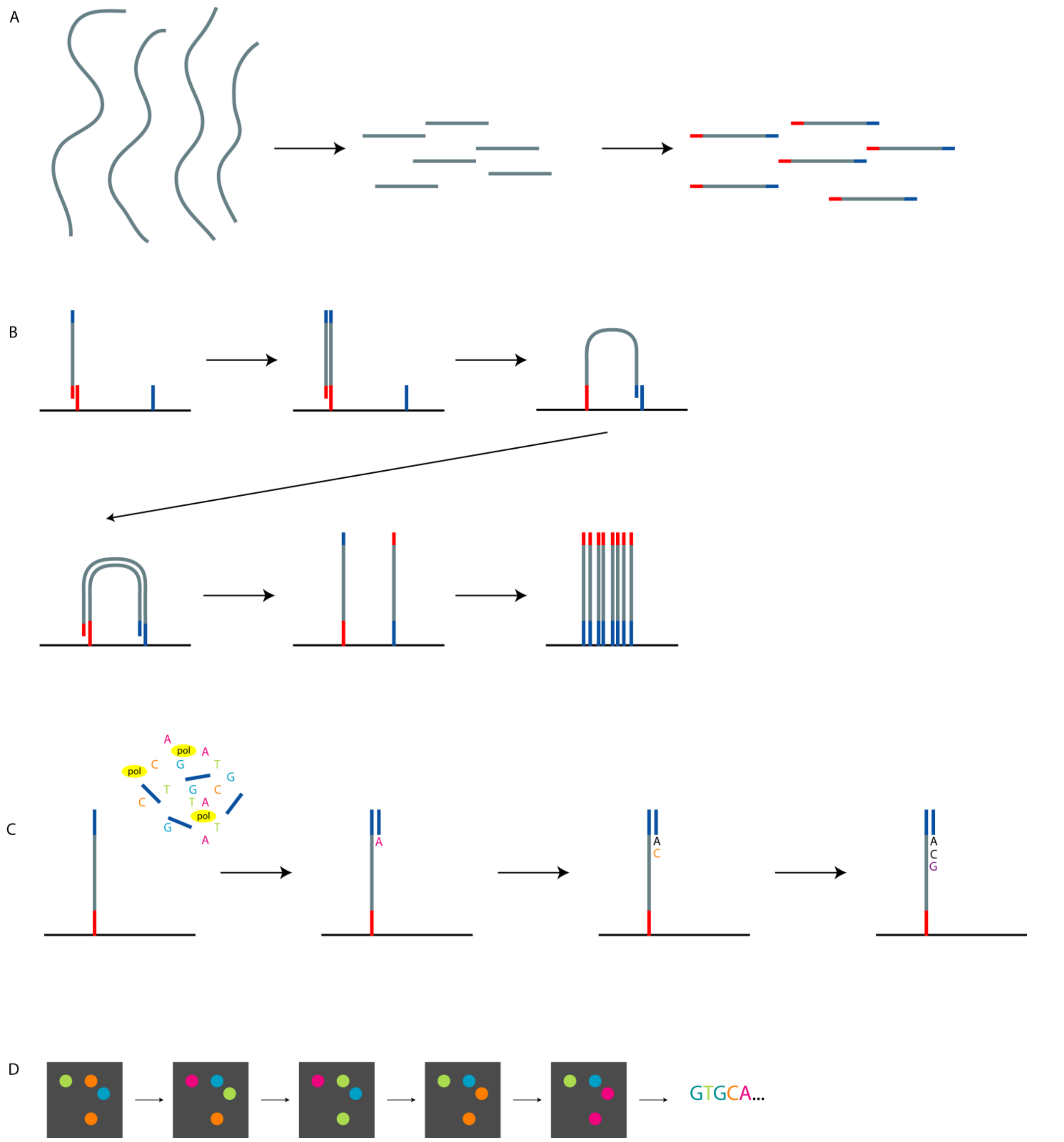


Figure 5.15. The Illumina high-throughput sequencing strategy. A – library preparation. The genomic DNA is fragmented and size-selected to give a population of small fragments, which have specific adaptors ligated to either end. B – cluster formation. Single-strand DNA fragments are bound to the flow cell surface, and cycles of bridge amplification create clusters of up to a million fragments. C – DNA polymerase and labelled nucleotides are added to the flow cell, and one fluorescently-labelled base is incorporated. The remaining nucleotides and polymerase are washed off, and an image of the whole flow cells is taken. The 3'-OH block and fluorescent label are removed from the incorporated nucleotide, and further rounds of synthesis take place. D – The images produced from the flow cell are used to call the base pairs incorporated into each fragment.

5.2.4 Paired-end read high-throughput sequencing of MDA-MB-134

A library of genomic fragments from MDA-MB-134 was prepared by Dr Jessica Pole and 18 million paired-end reads were sequenced using the Illumina GAII sequencer by the Genomics Facility of the Cancer Research UK Cambridge Research Institute. 37bp was sequenced from each end of the ~450bp genomic fragments and aligned to the human reference genome by the Bioinformatics Facility. Pairs where either end did not map uniquely to the genome were discarded, and are likely to be fragments produced from repeat regions where the sequences match multiple regions of the genome. For a set of reads which were exact duplicates of each other only one read was retained, as these were thought to be either PCR duplicates caused by amplification of the same fragment during the PCR amplification step of the library preparation, or optical duplicates caused by the same cluster being read as two clusters during the imaging step of sequencing. (See Chapter 6 for more detail of the bioinformatics used to process the sequencing data.)

I analysed the 12 million uniquely-mapping non-duplicated paired reads left after this filtering process to find structural variants and copy number changes. Although each paired read is only 74bp of sequence, a rearrangement anywhere in the fragment between the reads will be detected from the end sequences, giving around 1X diploid genome coverage. At this level of coverage, around 25% of the single-copy rearrangements in the genome would be detected as we require 2 independent reads to support any rearrangement. The amplified regions will represent proportionally more of the fragment library, as they are at a higher copy number and there are 2 copies of the der(8)t(8;11) chromosome, so the coverage will be higher in the amplified regions and more of the rearrangements in these regions will be detected.

5.2.5 Detection of structural variants

The 12 million reads were analysed for paired reads which appeared to suggest a structural variants in the genome of MDA-MB-134. 47,446 paired reads were called as possible reads across structural variants as they mapped to an unexpected location or orientation and were sequenced from fragments which contain a genomic rearrangement. To reduce the number of false positive structural variants which were called due to biological or bioinformatic artefacts, such as sequencing errors, chimeric fragments created during the library preparation process, or misaligned reads, two reads were required to call a structural variant from the possible reads. 679 structural variants supported by 2 or more independent paired reads were predicted, using 2,234 of the possible reads. The remaining reads not used to support a structural variant were presumed to be artefacts or reads where only a single read supporting a structural variant was found.

The structural variants were divided into 5 categories based on the probable type of rearrangement which could be inferred from the reads (Table 5.1). (See Chapter 6 for details of how the types of structural variant were inferred.)

52 variants mapped to regions of known copy number variation found in the human population, based on the data in Conrad et al. (2010). There is no matching normal cell line for MDA-MB-134 which would confirm these are germline rearrangements, but it is likely that they are not somatic rearrangements, and they were removed from further analysis.

Category of structural variant	Number
Interchromosomal translocation	16
Deletions larger than 10kb	6
Deletions between 1kb and 10kb	73
Deletions smaller than 1kb	485
Insertion	15
Inversion	14
Inverted tandem repeat	3

Table 5.1. Structural variants in MDA-MB-134, after removal of structural variants in regions of common copy number variation but before any further filtering

Of the 612 remaining structural variants, over 90% were intrachromosomal rearrangements under 10Kb, and 80% were under 1Kb. It is likely that some of the small rearrangements were false positives caused by the thresholds chosen to find structural variants. Any read pair with a fragment size larger than 3 standard deviations from the median was called as a possible abnormal read, which will incorrectly call some reads as aberrant when they are in the expected 0.3% of reads which fall outside this size threshold, and any region with two reads which fall into the tail of the distribution may be called as a small deletion. Raising the threshold to call an abnormal read would reduce the number of false positives but remove some real small deletions.

5.2.6 Validation of structural variants

As my interest was primarily in the structure of the amplicon, I decided to concentrate on validating the rearrangements between chromosomes and the intrachromosomal rearrangements larger than 10kb, which amounted to 42 structural variants. The reads supporting the predicted structural variants were re-aligned to the reference genome using BLAT (Kent, 2002), which is a more sensitive but slower alignment tool, and reports multiple possible alignments while the faster Maq alignment reports only the best possible alignment that it has found, which may not always be correct. 7 of the predicted structural variants were removed as the reads were in repeat regions including centromeres, or had other good alignments which suggested the pair were a normal read and not a structural variant.

The 35 selected structural variants (Table 5.2) were validated using PCR. As the median size of the fragments in the library was 454bp, the position of the breakpoints was known to ~500bp resolution. Primers were designed to amplify the breakpoints in genomic DNA from MDA-MB-134. A pool of normal human female DNA was used as a control.

Type of structural variant	Supporting reads	Chromosome	Breakpoint region start	Breakpoint region end	First read strand	Chromosome	Breakpoint region start	Breakpoint region end	Second read strand
Deletion	2	8	32,799,124	32,799,329	+	8	32,810,825	32,811,014	-
Deletion	12	8	34,902,273	34,902,554	+	8	35,015,339	35,015,607	-
Deletion	3	8	39,350,844	39,350,989	+	8	39,506,397	39,506,530	-
Deletion	2	8	106,144,206	106,144,375	+	8	139,083,092	139,083,275	-
Deletion	3	11	63,284,923	63,285,214	+	11	79,554,718	79,554,997	-
Deletion	4	11	76,836,485	76,836,687	+	11	77,033,372	77,033,540	-
Deletion	2	X	52,908,480	52,908,487	+	X	55,695,862	55,695,898	-
Inversion	4	7	70,058,671	70,058,799	+	7	70,076,473	70,076,605	+
Inversion	2	7	70,064,185	70,064,483	-	7	70,076,820	70,077,102	-
Inversion	7	8	36,017,686	36,017,889	+	8	36,548,255	36,548,465	+
Inversion	16	8	41,650,073	41,650,447	-	8	42,088,880	42,089,260	-
Inversion	2	8	41,773,851	41,774,164	+	8	133,032,657	133,032,959	+
Inversion	4	11	63,285,646	63,286,004	-	11	79,551,488	79,551,832	+
Inversion	2	11	63,289,773	63,289,783	+	11	78,702,800	78,702,823	+
Inversion	5	11	66,697,611	66,697,896	-	11	70,224,114	70,224,384	-
Inversion	2	11	70,656,229	70,656,305	-	11	76,410,016	76,410,063	+
Inversion	3	11	70,765,883	70,765,964	+	11	77,329,765	77,329,838	+
Inversion	5	11	73,044,828	73,045,170	-	11	77,572,905	77,573,291	+
Inversion	2	11	74,811,880	74,812,025	-	11	78,266,004	78,266,155	+
Inversion	5	11	74,812,741	74,812,993	-	11	76,838,302	76,838,571	-
Inversion	3	11	76,418,909	76,419,129	-	11	76,984,735	76,984,919	-
Inversion	2	16	21,501,819	21,501,842	+	16	22,617,924	22,617,940	+
Inversion	2	16	33,148,642	33,148,677	-	16	33,201,522	33,201,815	+
Translocation	2	2	41,905,754	41,905,954	+	4	66,096,540	66,096,761	-
Translocation	3	8	42,088,497	42,088,709	+	11	68,454,146	68,454,386	-
Translocation	2	8	124,072,313	124,072,368	-	X	136,263,026	136,263,099	-
Translocation	11	11	69,633,071	69,633,448	+	8	38,665,641	38,665,965	+
Translocation	8	11	70,783,301	70,783,583	+	8	21,981,233	21,981,523	+
Translocation	9	11	74,001,145	74,001,470	+	8	36,546,068	36,546,354	-
Translocation	7	11	74,001,293	74,001,424	-	8	36,548,943	36,549,081	-

Translocation	7	11	74,819,515	74,819,840	+	8	21,375,639	21,375,999	-
Translocation	25	11	75,535,691	75,536,039	-	8	34,784,458	34,784,809	-
Translocation	2	11	76,086,980	76,087,112	+	8	39,756,523	39,756,660	-
Translocation	3	12	106,727,070	106,727,245	+	7	110,840,421	110,840,624	-
Translocation	5	17	63,921,789	63,921,813	+	8	35,511,422	35,511,471	-

Table 5.2. Table 2. Predicted structural variants larger than 10Kb in MDA-MB-134. Structural variants have been filtered to remove common copy number variants and variants which had a normal mapping suggested by BLAT alignment. The read strand refers to the alignment of the reads in the read pairs to the genome – the expected strands for a normal read pair were the first read on the positive strand and the second read on the negative strand. (See Chapter 6 for more details on the expected strands for paired-end reads.)

Of the 35 selected variants, 4 variants could not be validated by PCR as no bands were produced using two different primer pairs designed to amplify these regions. Of the 31 remaining variants, 7 produced a PCR product using the control normal female DNA as well as MDA-MB-134 DNA, and were presumed to be common polymorphisms.

The 24 validated variants not present in the pooled normal DNA are shown in Table 5.3. All but one of the rearrangements involve chromosomes 8 and 11. This was the expected result based on the low numbers of copy number changes seen elsewhere in the genome on the Affymetrix SNP6.0 and array painting data, and the low sequence coverage of the genome. Using copy number data from the high-throughput sequencing, a further 9 unbalanced rearrangements were detected by segmentation which have no paired reads supporting them. 5 of these changes were single-copy deletions, and the other 4 rearrangements were small gains. These rearrangements may be missed due to low copy number, or because they fall in repeat regions and the reads spanning the junction cannot be aligned (see Chapter 7 for further discussion of the problems of finding junctions which fall in repeat regions).

The one validated structural variant which did not involve chromosomes 8 and 11 suggested an 8;X translocation. The break on chromosome X was around 136,263,000, and a break at this location can be seen on a copy number plot generated from the normal paired-end reads (figure 5.16A). Chromosome painting showed that that 20Mb of distal Xq has been translocated onto one copy of the marker chromosome (figure 5.16B). This rearrangement was not seen in the SKY karyotype, although it could have been missed as it is a small piece of chromosome X or may have been thought to be caused by chromosome overlap. It was also not present in the SNP 6.0 data, suggesting the rearrangement is present in our sample of MDA-MB-134 and may have been a late event in the evolution of the cell line.

Type of structural variant	Supporting reads	Chromosome	Breakpoint region start	Breakpoint region end	Strand	Chromosome	Breakpoint region start	Breakpoint region end	Strand
Deletion	12	8	34,902,273	34,902,554	+	8	35,015,339	35,015,607	-
Deletion	2	8	106,144,206	106,144,375	+	8	139,083,092	139,083,275	-
Deletion	3	11	63,284,923	63,285,214	+	11	79,554,718	79,554,997	-
Deletion	4	11	76,836,485	76,836,687	+	11	77,033,372	77,033,540	-
Inversion	7	8	36,017,686	36,017,889	+	8	36,548,255	36,548,465	+
Inversion	16	8	41,650,073	41,650,447	-	8	42,088,880	42,089,260	-
Inversion	2	8	41,773,851	41,774,164	+	8	133,032,657	133,032,959	+
Inversion	4	11	63,285,646	63,286,004	-	11	79,551,488	79,551,832	+
Inversion	5	11	66,697,611	66,697,896	-	11	70,224,114	70,224,384	-
Inversion	2	11	70,656,229	70,656,305	-	11	76,410,016	76,410,063	+
Inversion	3	11	70,765,883	70,765,964	+	11	77,329,765	77,329,838	+
Inversion	5	11	73,044,828	73,045,170	-	11	77,572,905	77,573,291	+
Inversion	2	11	74,811,880	74,812,025	-	11	78,266,004	78,266,155	+
Inversion	5	11	74,812,741	74,812,993	-	11	76,838,302	76,838,571	-
Inversion	3	11	76,418,909	76,419,129	-	11	76,984,735	76,984,919	-
Translocation	3	8	42,088,497	42,088,709	+	11	68,454,146	68,454,386	-
Translocation	2	8	124,072,313	124,072,368	-	X	136,263,026	136,263,099	-
Translocation	11	11	69,633,071	69,633,448	+	8	38,665,641	38,665,965	+
Translocation	8	11	70,783,301	70,783,583	+	8	21,981,233	21,981,523	+
Translocation	9	11	74,001,145	74,001,470	+	8	36,546,068	36,546,354	-
Translocation	7	11	74,001,293	74,001,424	-	8	36,548,943	36,549,081	-
Translocation	7	11	74,819,515	74,819,840	+	8	21,375,639	21,375,999	-
Translocation	25	11	75,535,691	75,536,039	-	8	34,784,458	34,784,809	-
Translocation	2	11	76,086,980	76,087,112	+	8	39,756,523	39,756,660	-

Table 5.3. Validated structural variants larger than 10Kb in MDA-MB-134. All variants were validated by PCR and sequencing.The read strand refers to the alignment of the reads in the read pairs to the genome – the expected strands for a normal read pair were the first read on the positive strand and the second read on the negative strand.

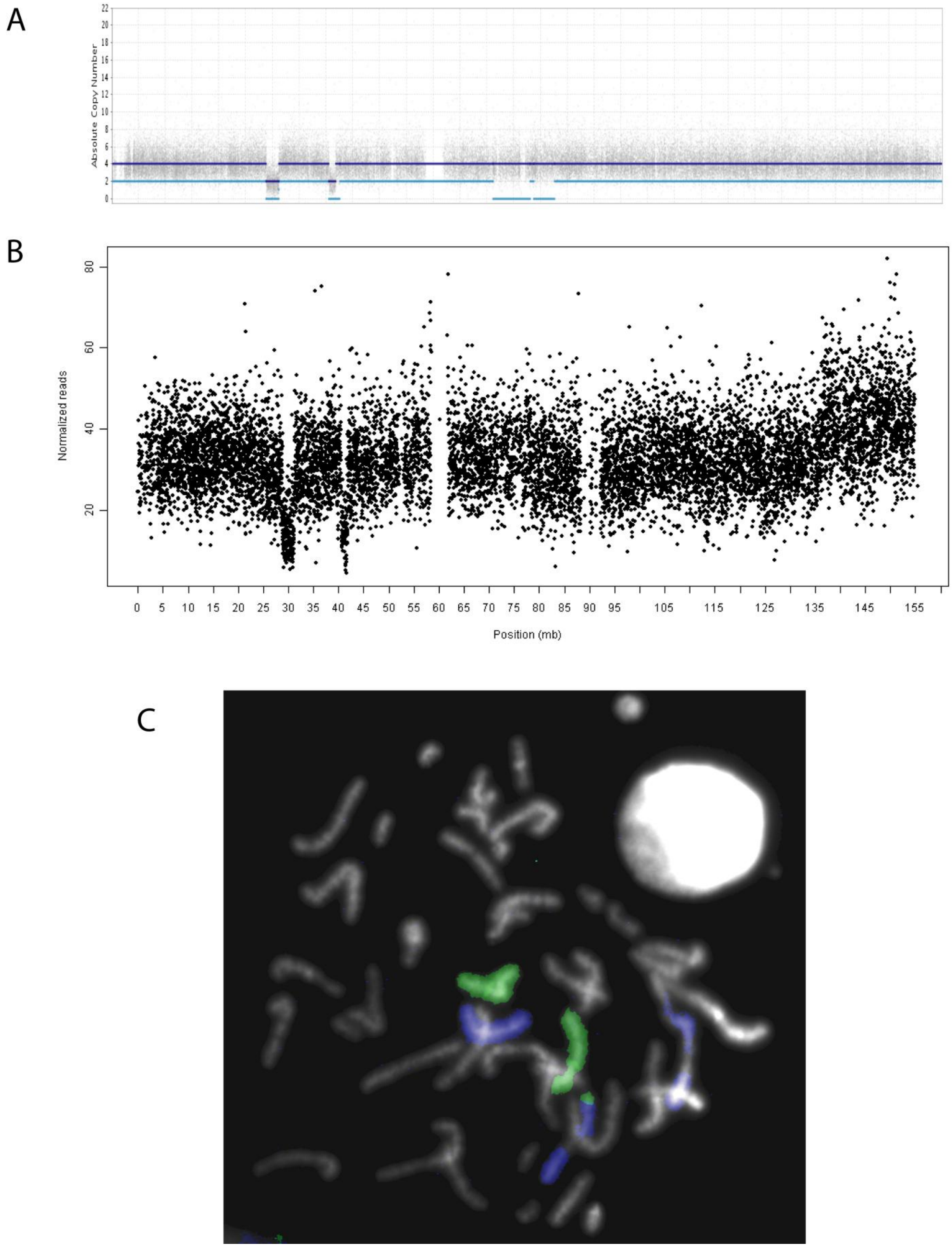

A
Absolute Copy Number
0
2
4
6
8
10
12
14
16
18
20
22
B
Normalized reads
20
40
60
80
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 105 115 125 135 145 155
Position (mb)
C

Figure 5.16. FISH to investigate an unexpected structural variant between chromosome 8 and chromosome X. A – copy number plot from whole-genome SNP6.0 array for chromosome X of MDA-MB-134, showing no copy number step on the q arm. B – copy number plot from high-throughput sequencing for chromosome X of MDA-MB-134, showing a copy number step around 136Mb. C - FISH confirming 8;X translocation in MDA-MB-134. Chromosome 8 spectrum orange-labelled paint is blue, chromosome X FITC-labelled paint is green. One normal 8 and two normal X chromosomes can be seen, along with two der(11) marker chromosomes, one of which shows a translocation with chromosome X.

5.2.7 Sequencing of structural variant junctions

All of the validated variants were Sanger sequenced to confirm their positions and to find the exact sequence at the junctions. All the breakpoints were found in the expected positions, with the exact breakpoints being within a library insert size of the position of the reads spanning the breakpoint (Table 5.4).

Figures 5.17-5.19 show all the validated structural variants in the genome, plotted against the copy number data obtained from sequencing. Many of the breakpoints match the copy number steps, although there are several structural variants which do not appear to be associated with a copy number change, such as the junction between 74,811,827 and 78,266,347Mb on chromosome 11 which only shows a copy number step on one side of the junction, and these may be balanced breakpoints where there is no copy number change to be detected.

There are also copy number steps which are not associated with a structural variant. This could be due to lack of coverage, as few breaks at low copy number would be detected at the current coverage levels in MDA-MB-134, or they may represent breaks in repeat regions, as if the region is highly repetitive any sequence from that region will be rejected as they will be a perfect match to more than one location in the genome.

Of the junction sequences of the 24 variants, 9 showed no homology at the breakpoint, 14 showed homology of between 1 and 4bp, 1 variant had 13bp of homology, and 1 showed an insertion of 1bp at the breakpoint (Table 5.4).

Type of structural variant	Chromosome	Breakpoint	Chromosome	Breakpoint	Overlap/insertion	Length	Sequence
Translocation	11	74,819,892	8	21,375,631	Overlap	1	GT
Translocation	11	69,633,489	8	38,666,031	Overlap	1	T
Translocation	8	42,088,756	11	68,454,008	None	0	None
Translocation	11	74,001,087	8	36,548,912	Overlap	4	AGGT
Inversion	11	76,418,750	11	76,984,724	Overlap	1	A
Translocation	8	124,072,213	X	136,262,830	Overlap	4	CCCT
Translocation	11	70,783,703	8	21,981,574	Overlap	2	GT
Translocation	11	74,001,597	8	36,546,065	Insertion	1	T
Translocation	11	75,535,675	8	34,784,454	Overlap	3	TTG
Translocation	11	76,087,332	8	39,756,483	None	0	None
Inversion	11	74,811,827	11	78,266,347	Overlap	1	C
Inversion	11	74,812,675	11	76,838,258	Overlap	1	T
Deletion	11	76,836,832	11	77,033,295	None	0	None
Deletion	8	34,902,627	8	35,015,270	None	0	None
Inversion	8	36,017,999	8	36,548,613	Overlap	1	T
Inversion	8	41,650,072	8	42,088,882	Overlap	2	GT
Inversion	8	41,774,211	8	133,033,042	None	0	None
Deletion	8	106,144,619	8	139,083,096	None	0	None
Deletion	11	63,285,330	11	79,554,711	None	0	None
Inversion	11	66,697,564	11	70,224,058	Overlap	1	T
Inversion	11	70,655,988	11	76,410,171	None	0	None
Inversion	11	70,766,233	11	77,329,957	Overlap	13	TTCTTTTTGGAGA
Inversion	11	73,044,824	11	77,573,265	None	0	None
Inversion	11	63,285,642	11	79,551,886	Overlap	3	AAA

Table 5.4. The exact breakpoints and junction homology of the 24 validated structural variants in MDA-MB-134.

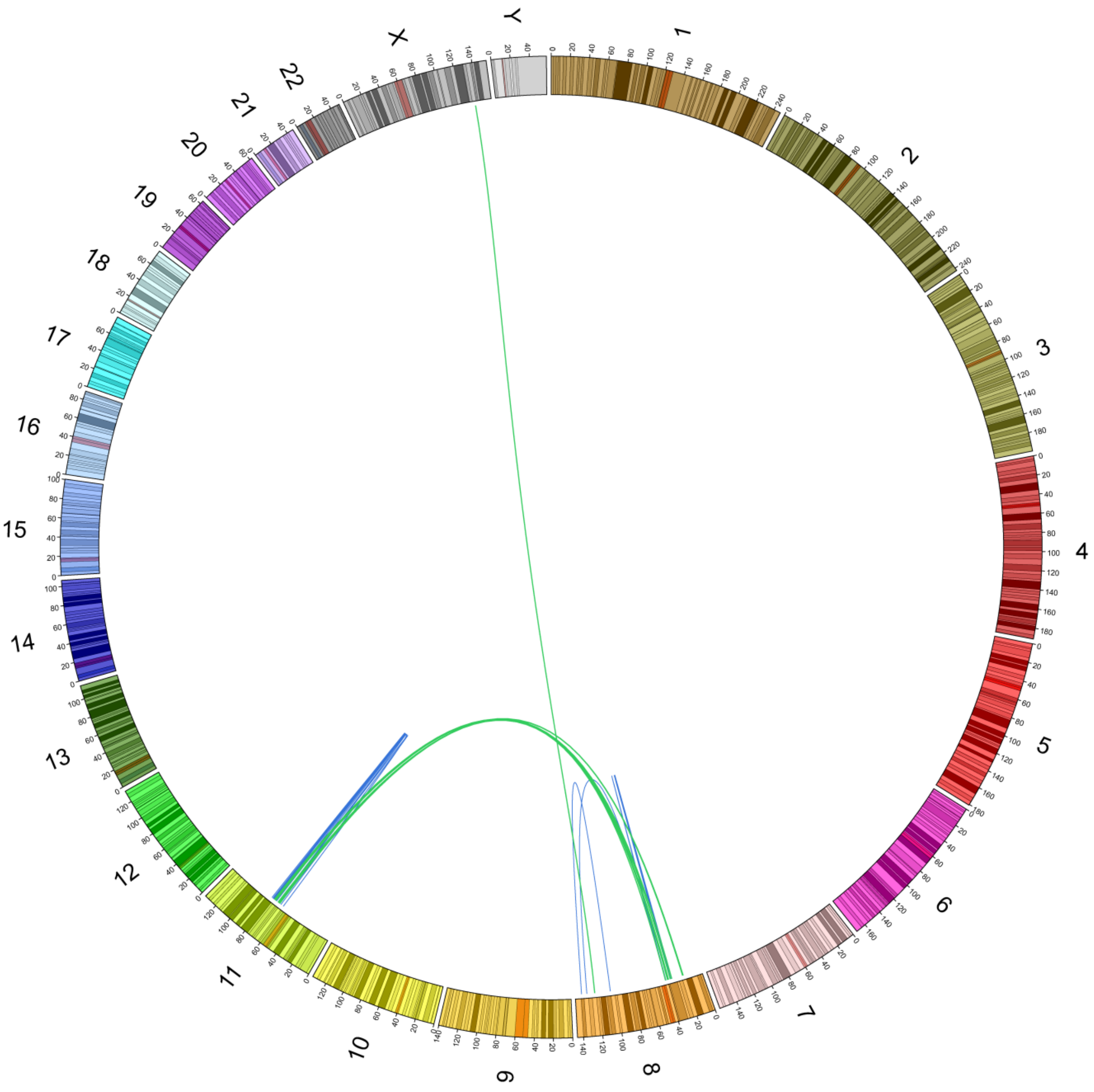


Figure 5.17. Structural variants in MDA-MB-134 plotted on a circular genome. Interchromosomal rearrangements are plotted in green, while intrachromosomal rearrangements are shown in blue. The plot was generated using the Circos software (Krzywinski et al., 2009).

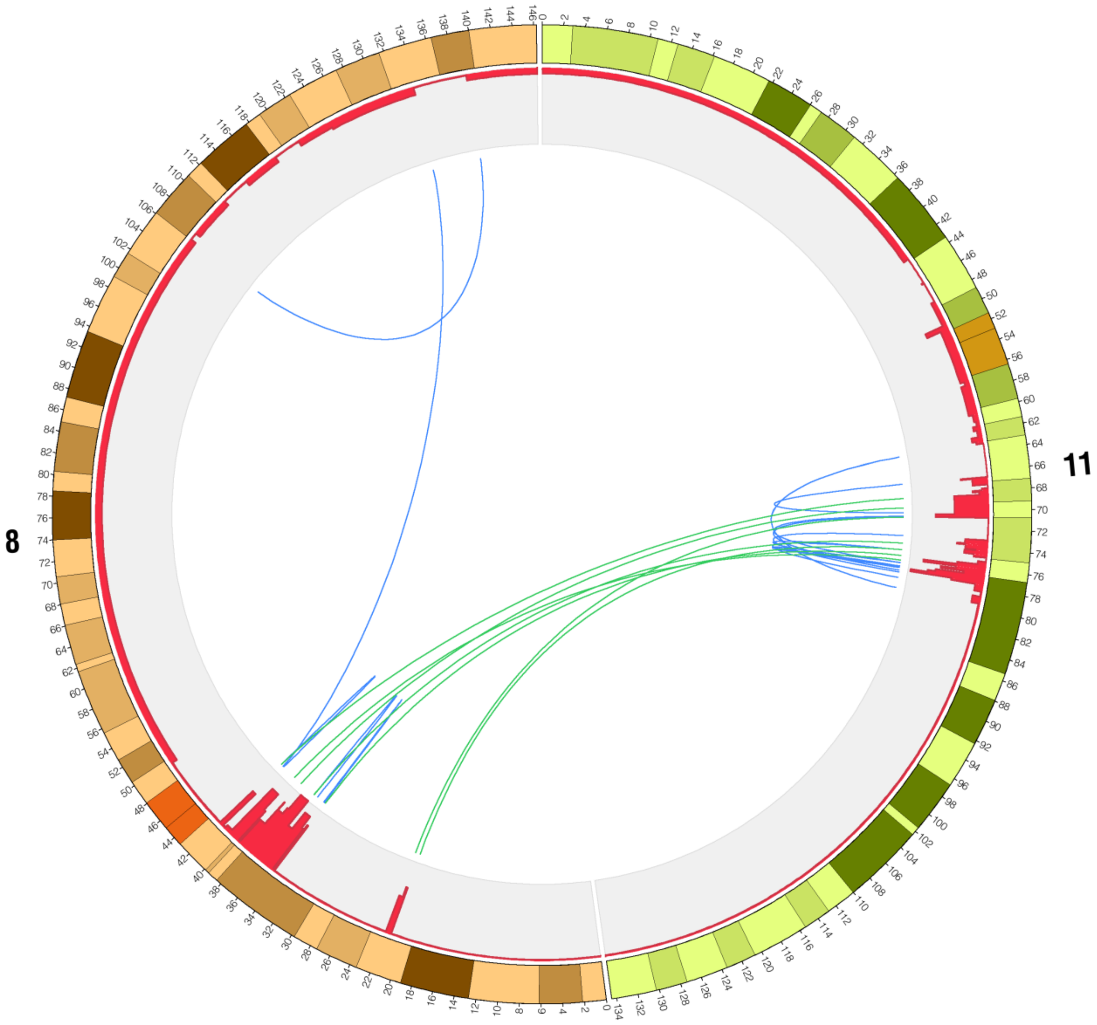


Figure 5.18. Structural variants in MDA-MB-134 showing chromosomes 8 and 11 only. The blue lines mark intrachromosomal rearrangements, and the green lines and interchromosomal rearrangements. The histogram in red shows copy number segments predicted using the DNACopy program. The figure was generated using Circos (Krzywinski et al., 2009).

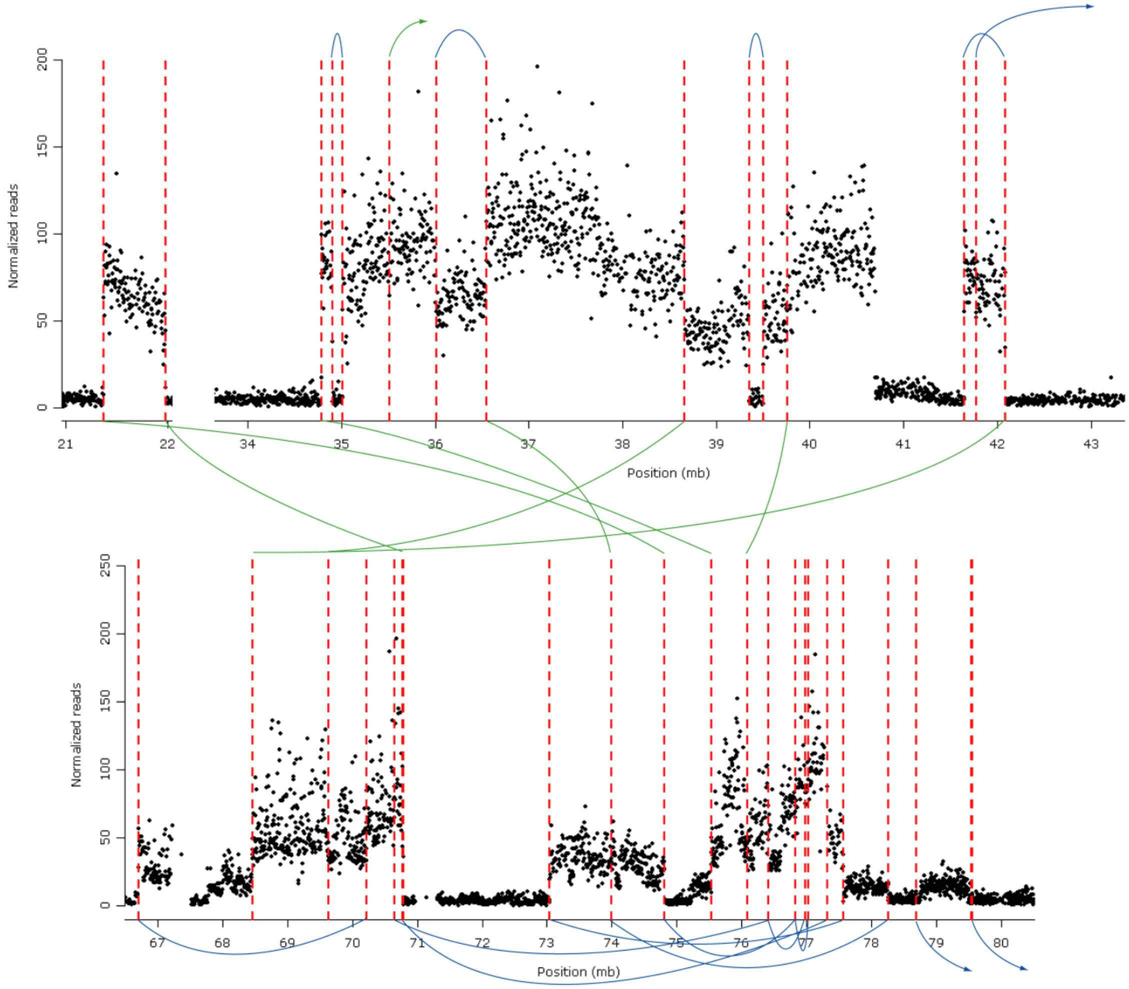


Figure 5.19. Structure of the 8;11 amplicon in MDA-MB-134. The plots show loess-corrected copy number from high-throughput sequencing, with the upper plot showing the amplified regions of chromosome 8, and the lower plot showing the amplified regions of chromosome 11. The dotted red lines mark the breakpoints of validated structural variants, with the blue and green lines showing the interchromosomal and intrachromosomal rearrangements. (See Chapter 6 for details of how the loess correction of copy number was performed.)

5.2.8 Potential fusion genes found by high-throughput sequencing

The potential fusion genes at each breakpoint could be predicted using high-throughput sequencing, as the breakpoints could be determined to high enough resolution that the genes at both sides of the breakpoint could be identified.

All of the structural variants were used to test for potential fusion genes, readthrough fusion genes, and internal exon deletions in MDA-MB-134. This included the small rearrangements which were not selected for PCR validation, as only a few of these rearrangements affected exons. The majority of the structural variants did not affect genes, or were small rearrangements which only rearranged introns.

4 fusion genes were predicted (Table 5.5), of which 3 were part of the 8;11 amplicon. Primer pairs were designed which would amplify each gene separately, to see if it was expressed, and in combination would amplify a fusion product (Figure 5.20). The results of the PCR to detect fusion genes are shown in Figures 5.21 and 5.22. No fusion products were detected, but three genes were expressed in MDA-MB-134 but not in the human immortalized breast epithelial cell line HB4a. These three genes, *ODZ4, SHANK2,* and *UNC5D*, are all in the amplified regions on chromosome 8 and 11.

8 readthrough fusions were predicted (Table 5.6), of which 6 were part of the 8;11 amplicon. No readthrough fusions were found. The results of the PCR are shown in Figure 5.23. *KLHL35*, which also forms a potential fusion gene with *ODZ4*, forms a potential readthrough fusion with *AQP11*, but no expression of this readthrough was detected, and it appears to be a separate event to the rearrangement which causes *KLHL35* and *ODZ4* to be fused (Figure 5.24).

14 structural variations were predicted to cause deletions of 1 or more exons from a gene (Table 5.7). All these deletions are small deletions under 10Kb, and in contrast to the predicted fusion genes and readthrough fusions, none of the rearrangements are part of the 8;11 amplicon. PCR primers were designed to either side of the deletion to test for shorter transcripts in MDA-MB-134, which would indicate a possible deletion of exons (Figure 5.25). No such transcripts were detected.

Type of structural variant	Read	Chromosome	Breakpoint region start	Breakpoint region end	Read strand	Genes	Predicted fusion
Insertion	First read	11	74811880	74812069	-	KLHL35	5' of KLHL35 into 3' of ODZ4
	Second read	11	78266004	78266199	+	ODZ4	
Translocation	First read	17	63921789	63921857	+	ARSG	5' of ARSG into 3' of UNC5D
	Second read	8	35511422	35511515	-	UNC5D	
Inversion	First read	8	41773851	41774208	+	ANK1	5' of EFR3A into 3' of ANK1
	Second read	8	133032657	133033003	+	EFR3A	
Inversion	First read	11	66697611	66697940	-	FBXL11	5' of SHANK2 into 3' of FBXL11
	Second read	11	70224114	70224428	-	SHANK2	

Table 5.5. Gene fusions in MDA-MB-134 predicted from structural variants called from paired-end sequencing. The read strand refers to the alignment of the reads in the read pairs to the genome.

Type of structural variant	Read	Chromosome	Breakpoint region start	Breakpoint region end	Read strand	Broken gene	Readthrough partner	Predicted fusion gene
Translocation	First read	11	70783301	70783627	+	EPB49	SHANK2	EPB49 is broken and may read through into SHANK2
	Second read	8	21981233	21981567	+			
Translocation	First read	11	76086980	76087156	+	ADAM2	LRRC32	ADAM2 is broken and may read through into LRRC32
	Second read	8	39756523	39756704	-			
Insertion	First read	11	70656229	70656349	-	ACER3	NADSYN1	ACER3 is broken and may read through into NADSYN1
	Second read	11	76410016	76410107	+			
Translocation	First read	12	106727070	106727289	+	IMMP2L	PRDM4	IMMP2L is broken and may read through into PRDM4
	Second read	7	110840421	110840668	-			
Deletion	First read	7	30501892	30501945	+	GGCT	NOD1	GGCT is broken and may read through into NOD1
	Second read	7	30502611	30502711	-			
Inversion	First read	11	74812741	74813037	-	KLHL35	AQP11	KLHL35 is broken and may read through into AQP11
	Second read	11	76838302	76838615	-			
Inversion	First read	11	74812741	74813037	-	PAK1	SERPINH1	PAK1 is broken and may read through into SERPINH1
	Second read	11	76838302	76838615	-			
Inversion	First read	8	41650073	41650491	-	ANK1	AP3M2	ANK1 is broken and may read through into AP3M2
	Second read	8	42088880	42089304	-			

Table 5.6. Readthrough gene fusions in MDA-MB-134 predicted from structural variants called from paired-end sequencing. The read strand refers to the alignment of the reads in the read pairs to the genome.

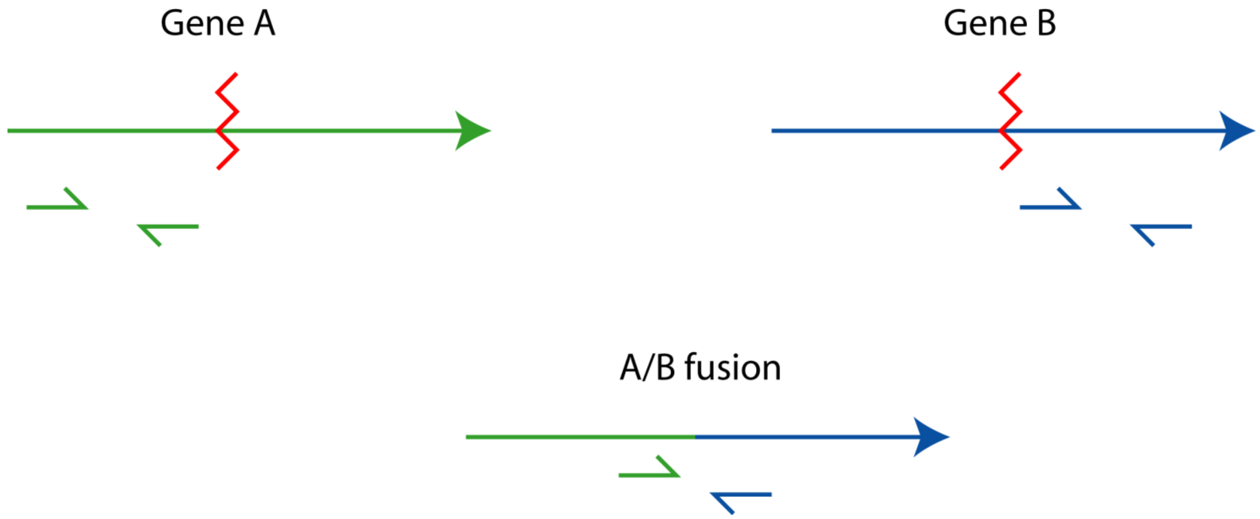


Figure 5.20. Fusion gene PCR strategy. The green and blue arrows represent genes, while the red jagged line represents a breakpoint in the gene. PCR primers were designed which would amplify the normal cDNA from each gene, and in combination would amplify only a fusion product.

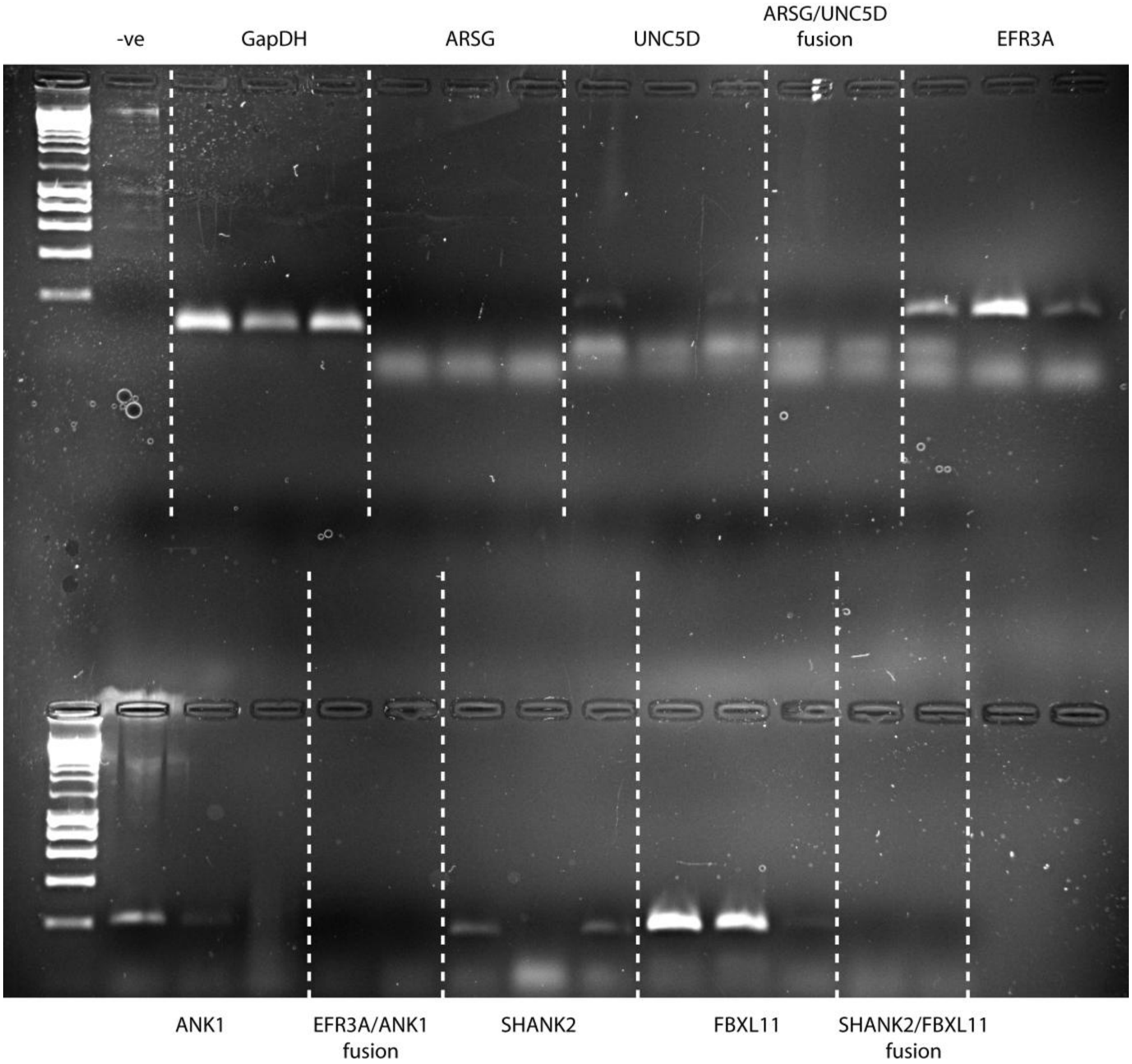


Figure 5.21. PCR to detect fusion transcripts in MDA-MB-134. No fusion transcripts were detected. See next page for key to PCR reactions.

Upper row				Lower row		
Well	**Primers**	**cDNA**		**Well**	**Primers**	**cDNA**
1	GapDH	None		1	ANK1	Reference
2	GapDH	Reference		2	ANK1	HB4a
3	GapDH	HB4a		3	ANK1	MDA-MB-134
4	GapDH	MDA-MB-134		4	EFR3A/ANK1	HB4a
5	ARSG	Reference		5	EFR3A/ANK1	MDA-MB-134
6	ARSG	HB4a		6	SHANK2	Reference
7	ARSG	MDA-MB-134		7	SHANK2	HB4a
8	UNC5D	Reference		8	SHANK2	MDA-MB-134
9	UNC5D	HB4a		9	FBXL11	Reference
10	UNC5D	MDA-MB-134		10	FBXL11	HB4a
11	ARSG/UNC5D	HB4a		11	FBXL11	MDA-MB-134
12	ARSG/UNC5D	MDA-MB-134		12	SHANK2/FBXL11	HB4a
13	EFR3A	Reference		13	SHANK2/FBXL11	MDA-MB-134
14	EFR3A	HB4a				
15	EFR3A	MDA-MB-134				

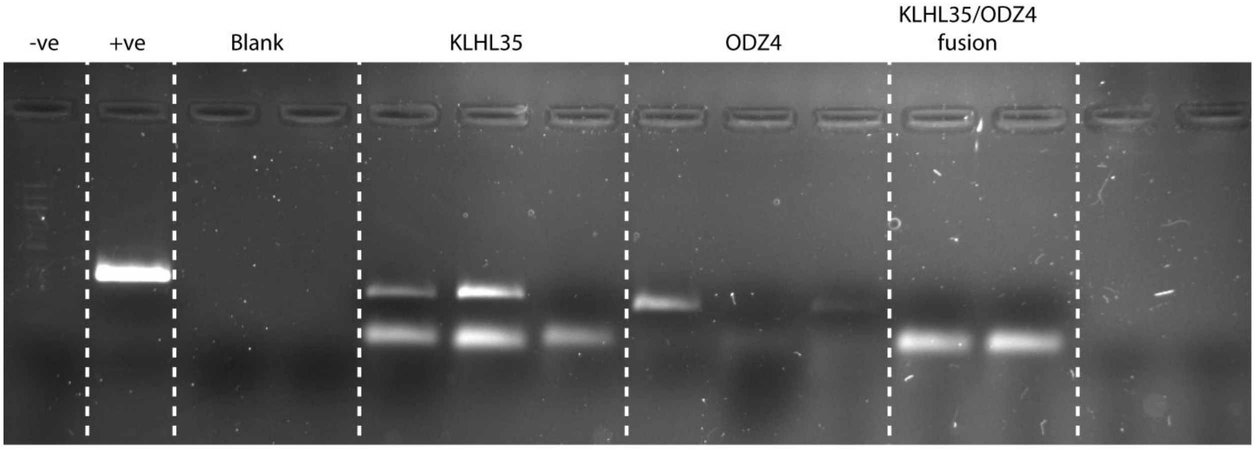


Figure 5.22. PCR to detect fusion transcripts in MDA-MB-134. No fusion transcripts were detected.

Well	Primers	cDNA
1	GapDH	None
2	GapDH	Reference
3	Blank	Blank
4	Blank	Blank
5	KLHL35	Reference
6	KLHL35	HB4a
7	KLHL35	MDA-MB-134
8	ODZ4	Reference
9	ODZ4	HB4a
10	ODZ4	MDA-MB-134
11	KLHL35/ODZ4	HB4a
12	KLHL35/ODZ4	MDA-MB-134

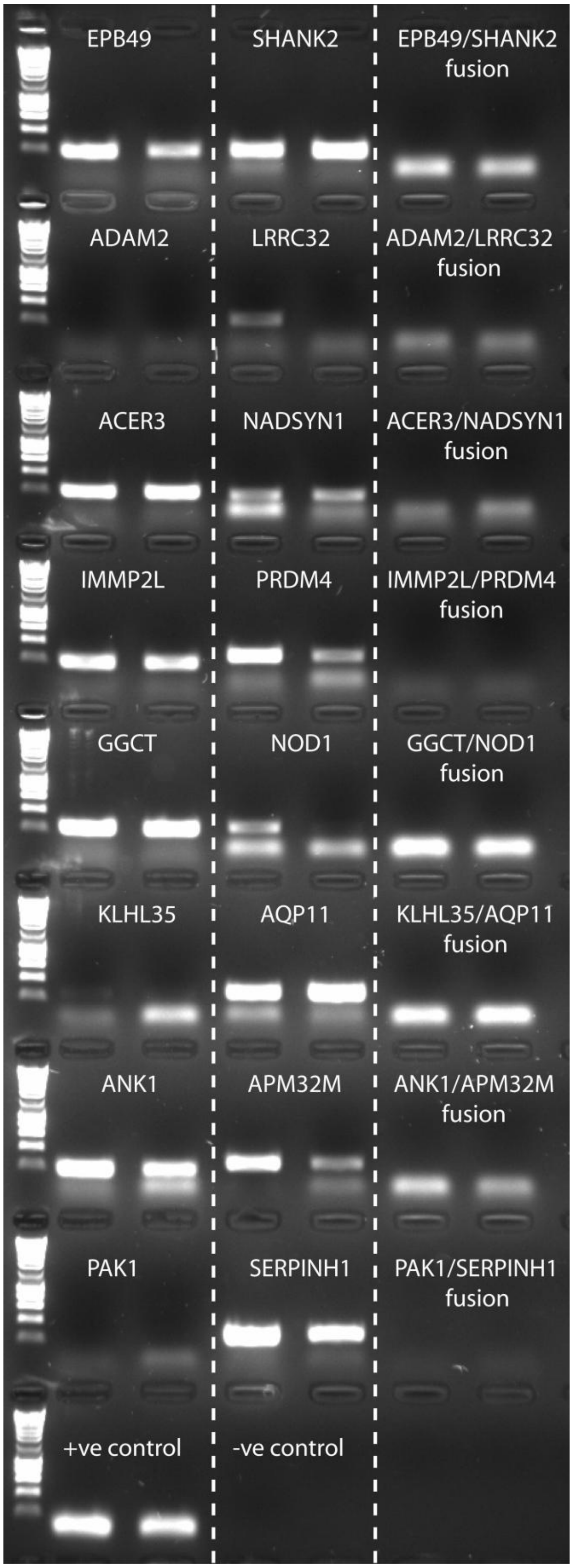


Figure 5.23. PCR to test for predicted readthrough fusions in MDA-MB-134. A key to the wells can be found on the following page. No fusion transcripts were detected.

Row	Column	Primers	cDNA
1	1	EPB49	Reference
1	2	EPB49	MDA-MB-134
1	3	SHANK2	Reference
1	4	SHANK2	MDA-MB-134
1	5	EPB49/SHANK2	Reference
1	6	EPB49/SHANK2	MDA-MB-134
2	1	ADAM2	Reference
2	2	ADAM2	MDA-MB-134
2	3	LRRC32	Reference
2	4	LRRC32	MDA-MB-134
2	5	ADAM2/LRRC32	Reference
2	6	ADAM2/LRRC32	MDA-MB-134
3	1	ACER3	Reference
3	2	ACER3	MDA-MB-134
3	3	NADSYN1	Reference
3	4	NADSYN1	MDA-MB-134
3	5	ACER3/NADSYN1	Reference
3	6	ACER3/NADSYN1	MDA-MB-134
4	1	IMMP2L	Reference
4	2	IMMP2L	MDA-MB-134
4	3	PRDM4	Reference
4	4	PRDM4	MDA-MB-134
4	5	IMMP2L/PRDM4	Reference
4	6	IMMP2L/PRDM4	MDA-MB-134

Row	Column	Primers	cDNA
5	1	GGCT	Reference
5	2	GGCT	MDA-MB-134
5	3	NOD1	Reference
5	4	NOD1	MDA-MB-134
5	5	GGCT/NOD1	Reference
5	6	GGCT/NOD1	MDA-MB-134
6	1	KLHL35	Reference
6	2	KLHL35	MDA-MB-134
6	3	AQP11	Reference
6	4	AQP11	MDA-MB-134
6	5	KLHL35/AQP11	Reference
6	6	KLHL35/AQP11	MDA-MB-134
7	1	ANK1	Reference
7	2	ANK1	MDA-MB-134
7	3	APM32M	Reference
7	4	APM32M	MDA-MB-134
7	5	ANK1/APM32M	Reference
7	6	ANK1/APM32M	MDA-MB-134
8	1	PAK1	Reference
8	2	PAK1	MDA-MB-134
8	3	SERPINH1	Reference
8	4	SERPINH1	MDA-MB-134
8	5	PAK1/SERPINH1	Reference
8	6	PAK1/SERPINH1	MDA-MB-134
9	1	GapDH	Reference
9	2	GapDH	MDA-MB-134
9	3	GapDH	Reference

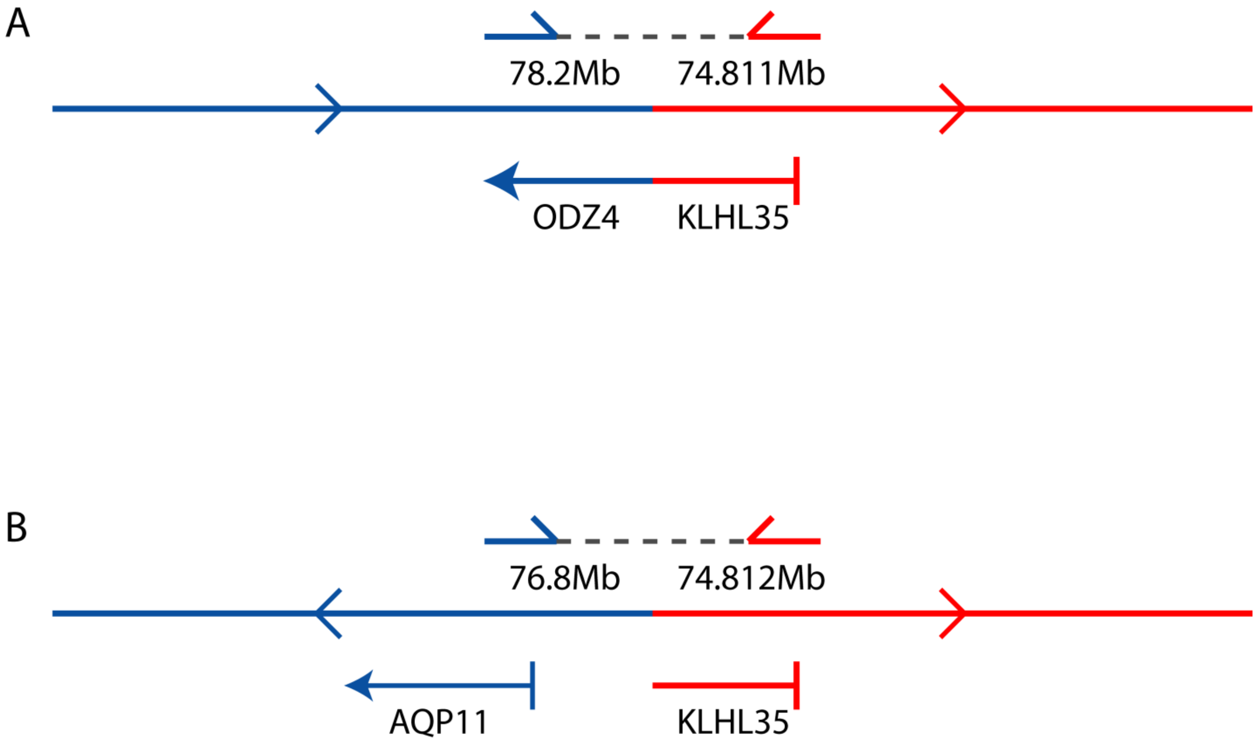


Figure 5.24. Two possible *KLHL35* fusions predicted in MDA-MB-134. A – fusion predicted between *KLHL35* and *ODZ4*. B – readthrough fusion predicted between *KLHL35* and *AQP11.*

Type of structural variant	Chromosome	Deletion start	Deletion end	Gene
Deletion	11	411136	411904	ANO9
Deletion	11	3081376	3082009	OSBPL5
Deletion	11	7673249	7674173	OVCH2
Deletion	11	92742021	92743103	CCDC67
Deletion	12	131707143	131707992	P2RX2
Deletion	15	87669090	87669745	POLG
Deletion	15	91316598	91317352	CHD2
Deletion	16	1387181	1387880	C16orf28
Deletion	17	37743398	37744175	STAT3
Deletion	19	55820546	55821085	SYT3
Deletion	2	85746607	85747410	SFTPB
Deletion	20	62066274	62068265	ZNF512B
Deletion	6	158468017	158469427	SERAC1
Deletion	7	157080659	157081216	PTPRN2

Table 5.7. Internal gene deletions in MDA-MB-134 predicted from deletions called from paired-end sequencing.

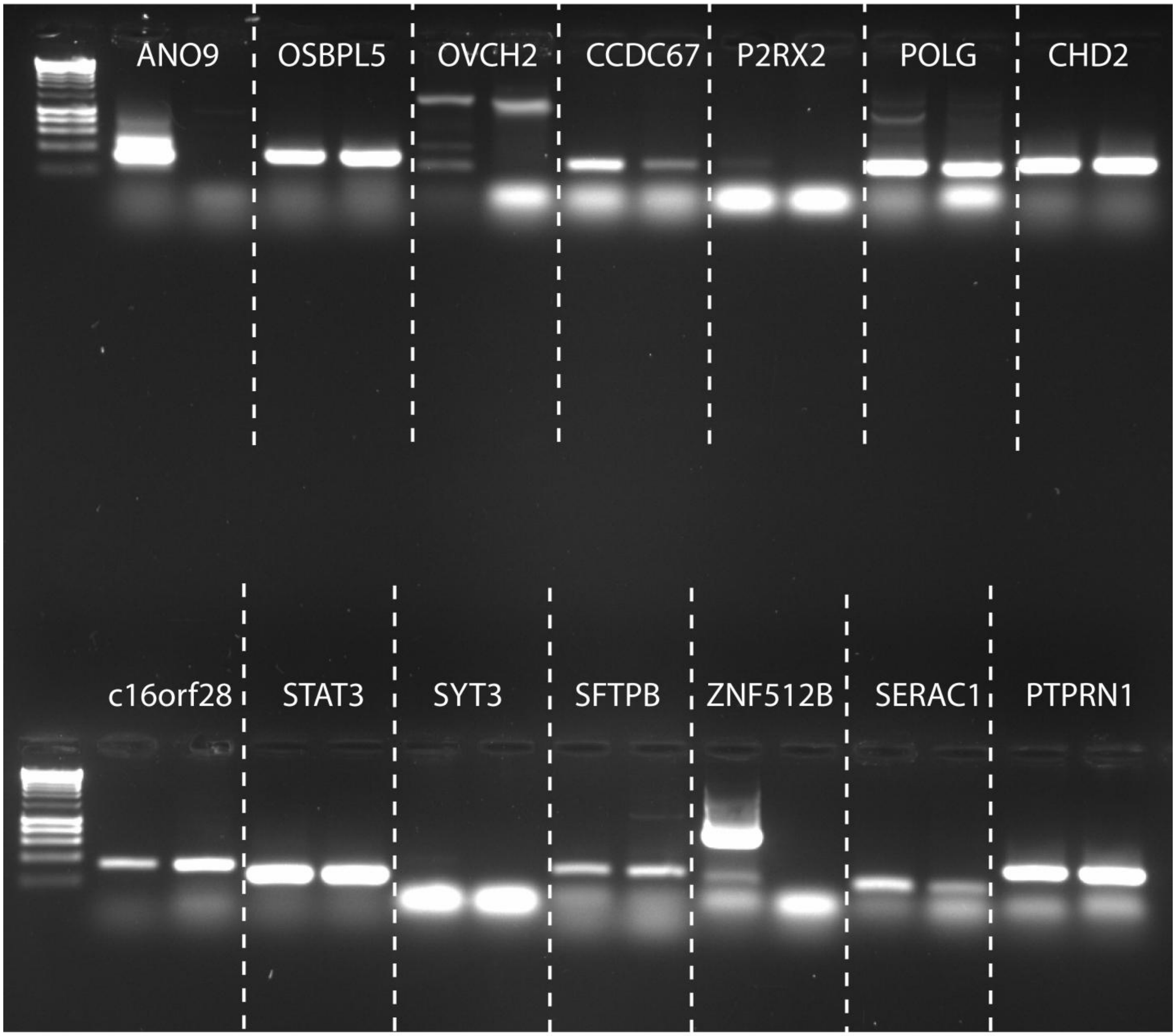


Figure 5.25. PCR to look for internal deletions of genes in MDA-MB-134. Each PCR was performed using genes spanning the deletion. The first PCR in each pair is on universal reference cDNA, and the second is on MDA-MB-134 cDNA to look for a different size product, which would indicate a possible transcript with exons deleted.

5.2.9 ODZ4 as a potential fusion gene

The candidate fusion gene *KLHL35-ODZ4* was of particular interest. *ODZ4* (also known as *DOC4* or *TEN4*) was part of one of the few gene fusions known before the start of my project, as it is fused to *NRG1* in the breast cancer cell line MDA-MB-175 (Liu et al., 1999). *ODZ4* is part of the teneurin family of signaling molecules, which can function as transmembrane receptors and also transcription factors (Tucker and Chiquet-Ehrismann, 2006). *ODZ1* overexpression has been implicated in mouse mammary tumorigenesis by MMTV-insertion experiments (Theodorou et al., 2007). *ODZ4* has not been proposed as one of the important genes in the 11q amplicon in breast cancer as it is outside the minimal region of amplification. From the SNP6.0 data from the Cancer Genome Project (Bignell et al., 2010), breaks in *ODZ4* can be seen in the cell lines HCC1599, MRK-nu-1, and MCF7. The genome of MCF7 has been mapped by paired-end sequencing (Hampton et al., 2009), but no fusions of *ODZ4* were found. A diagram of the known breaks in *ODZ4* is shown in Figure 5.26.

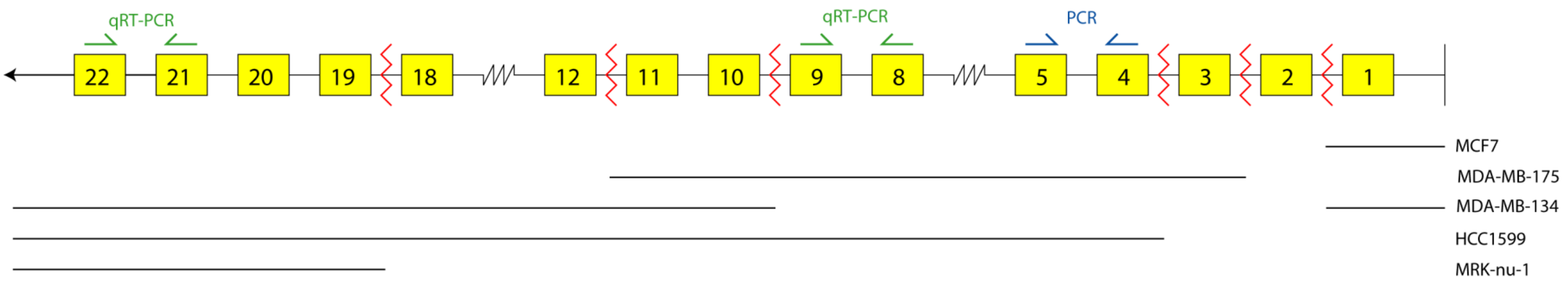


Figure 5.26. *ODZ4* breakpoints in breast cancer cell lines. The black lines mark the parts of the gene retained in the five cell lines with *ODZ4* breaks. Breakpoint locations were taken from whole-genome SNP6.0 data (Bignell et al., 2010) except for MDA-MB-134 where the breakpoints were determined by high-throughput sequencing. The green and blue arrows show the primer pairs used for standard and qRT-PCR. Exons 3-11 of *ODZ4* are involved in a fusion gene in MDA-MB-175.

ODZ4 was seen to be expressed in MDA-MB-134 but not in the normal immortalized breast cell line HB4a. To investigate whether *ODZ4* was expressed in other breast cancer cell lines, a primer pair designed to amplify exons 4 and 5 was tested on a panel of cell lines. In this semi-quantitative assay, out of 28 breast cancer cell lines, 5 showed expression of *ODZ4* (Figure 5.27). *ODZ4* was also expressed in the non-cancer breast cell line HMT3552, but not detectably in HB4a. Of the cell lines with known breaks, MDA-MB-134 and MDA-MB-175 showed expression while HCC1599 and MCF7 did not. No data was available for MRK-nu-1. HCC1419, HCC1500, and PMC42 also showed expression of *ODZ4*.

Quantitative real-time PCR was performed on *ODZ4*. Primers were designed to amplify exon 8, which is part of the fusion gene in MDA-MB-175, and exon 21, which is not part of the fusion gene, and tested on a panel of cell lines. The results are shown in figure 5.28, with expression compared to the universal human reference cDNA, containing cDNA from a pool of different cell lines.

HB4a and HMT3552 both express ODZ4 at low but detectable levels. The lack of expression seen in HB4a using semi-quantitative methods was likely to be due to the expression levels being at the limit of detection for standard PCR. The cell lines BT-474, HCC1500, PMC42 and SUM52 showed a low level of upregulation of both exons of *ODZ4* compared to the control cell line HB4a. Semi-quantitative PCR suggested that HCC1500 would have a higher expression level than was seen by quantitative PCR, which may be due to the standard PCR using different samples of cDNA which had not been normalized to any housekeeping genes.

MDA-MB-134 and MDA-MB-175 showed higher overexpression of *ODZ4* than any other cell lines. MDA-MB-175 was the only cell line which showed higher expression of exon 8 than exon 21, and this may be due to the fusion of *ODZ4*, which includes exon 8 but not exon 21.

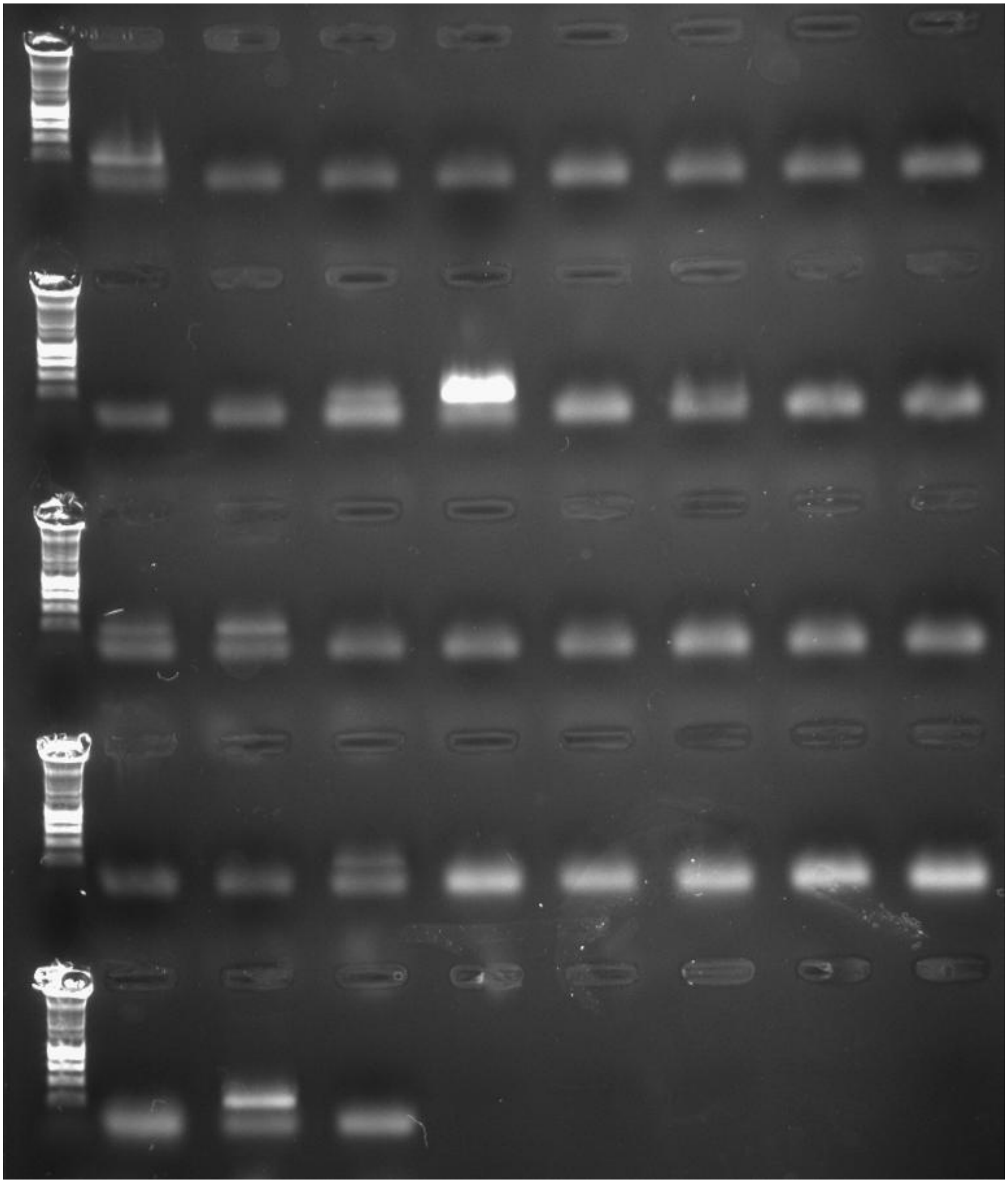


Figure 5.27. PCR using primers for ODZ4 exons 4-5 on a panel of breast cancer cell lines.
Row 1: HMT3552 (positive), HB4a, SUM44, SUM52, BT20, BT474, BT549, MCF7
Row 2: HCC38, HCC1143, HCC1419 (positive), HCC1500 (positive), HCC1569, HCC1599, HCC1806, HCC1937.
Row 3: MDA-MB-134 (positive), MDA-MB-175 (positive), MDA-MB-231, MDA-MB-361, MDA-MB-415, MDA-MB-468, ZR-75-1, ZR-75-30.
Row 4: VP229, VP267, PMC42 (positive), Patu-1, Suit-2, Mia-paca-2, DU4475, SKBr3
Row 5: T47D, universal reference cDNA (positive control), negative control.

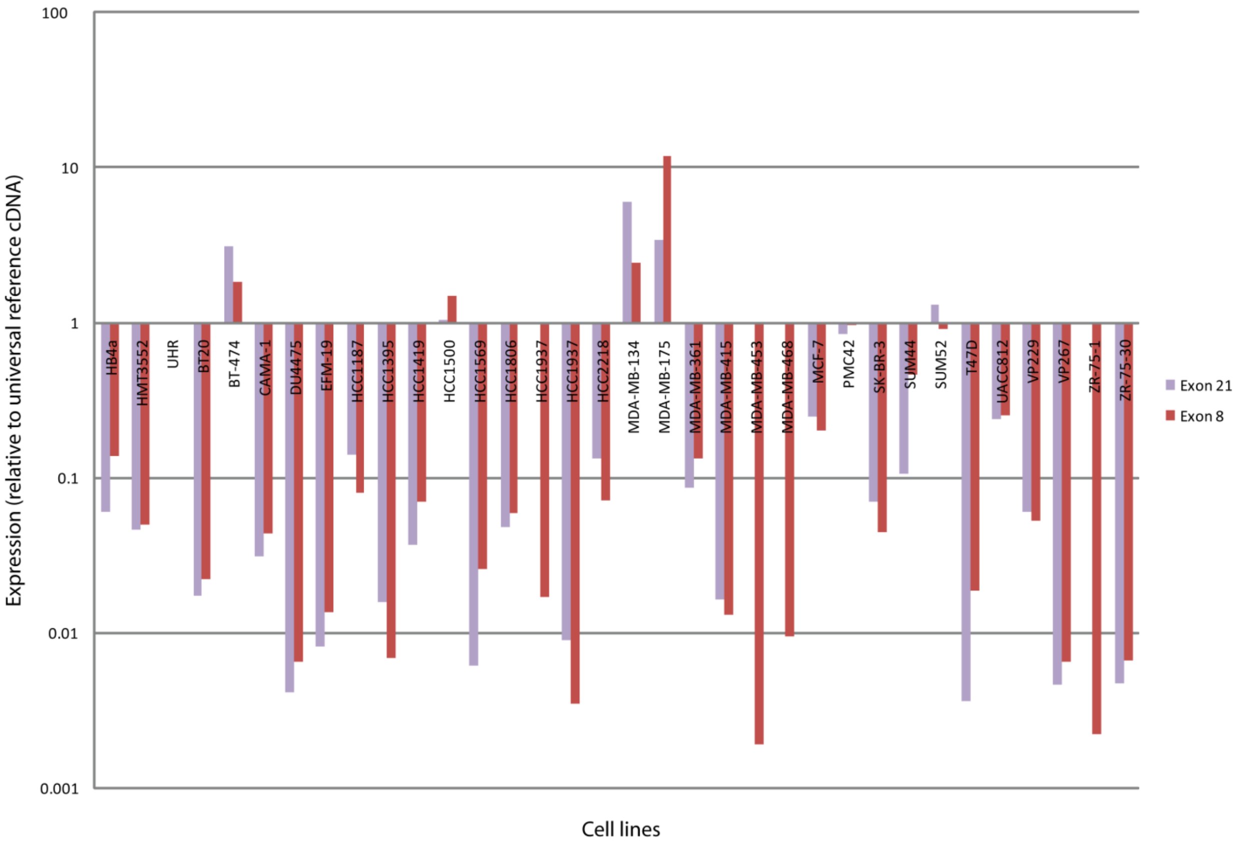


Figure 5.28. Real-time PCR on two sets of primers in *ODZ4*. Expression for each cell line was normalised to three housekeeping genes (*GAPDH*, *UBC*, and *RPL13a*), and the expression for each cell line was plotted relative to the expression of the universal reference cDNA.

5.3 Discussion

The karyotype of MDA-MB-134 has now been investigated using low-resolution array painting, high-resolution SNP6.0 arrays, and high-throughput sequencing. A comparison of the three methods shows good concordance between the three methods for determining the structure of amplicons.

The 1Mb array painting offers the lowest resolution data, but it is the only method which allows the structure of the amplified chromosome to be determined directly from the chromosome without the possibility of breaks on other copies of the chromosome being included as part of the amplification. While it appears that the breaks on chromosomes 8 and 11 in MDA-MB-134 are confined to the two amplified chromosomes, in cell lines and tumours with a more complicated pattern of amplification, spread across multiple different derivative chromosomes, whole-genome approaches could lead to confusion when trying to assemble complicated amplicons. The higher resolution data also suggests that requiring 3 consecutive clones gained or lost to call a copy number change in the 1Mb array painting was conservative, as there were regions with 1 or 2 clones showing a copy number change which appears to be real.

A limitation of microarray-based approaches is that it is difficult to quantify regions of high and low copy number in one experiment without the probes at high copy number becoming saturated (Williams and Thomson, 2010). High-throughput sequencing should be unaffected by this problem as the copy number is obtained directly from the number of reads. A comparison of the copy number data obtained from the SNP6.0 array and from high-throughput sequencing suggests that the structure of the amplicon has more copy number changes than are called from the SNP6.0 data as the array becomes saturated at high copy number (Figure 5.29).

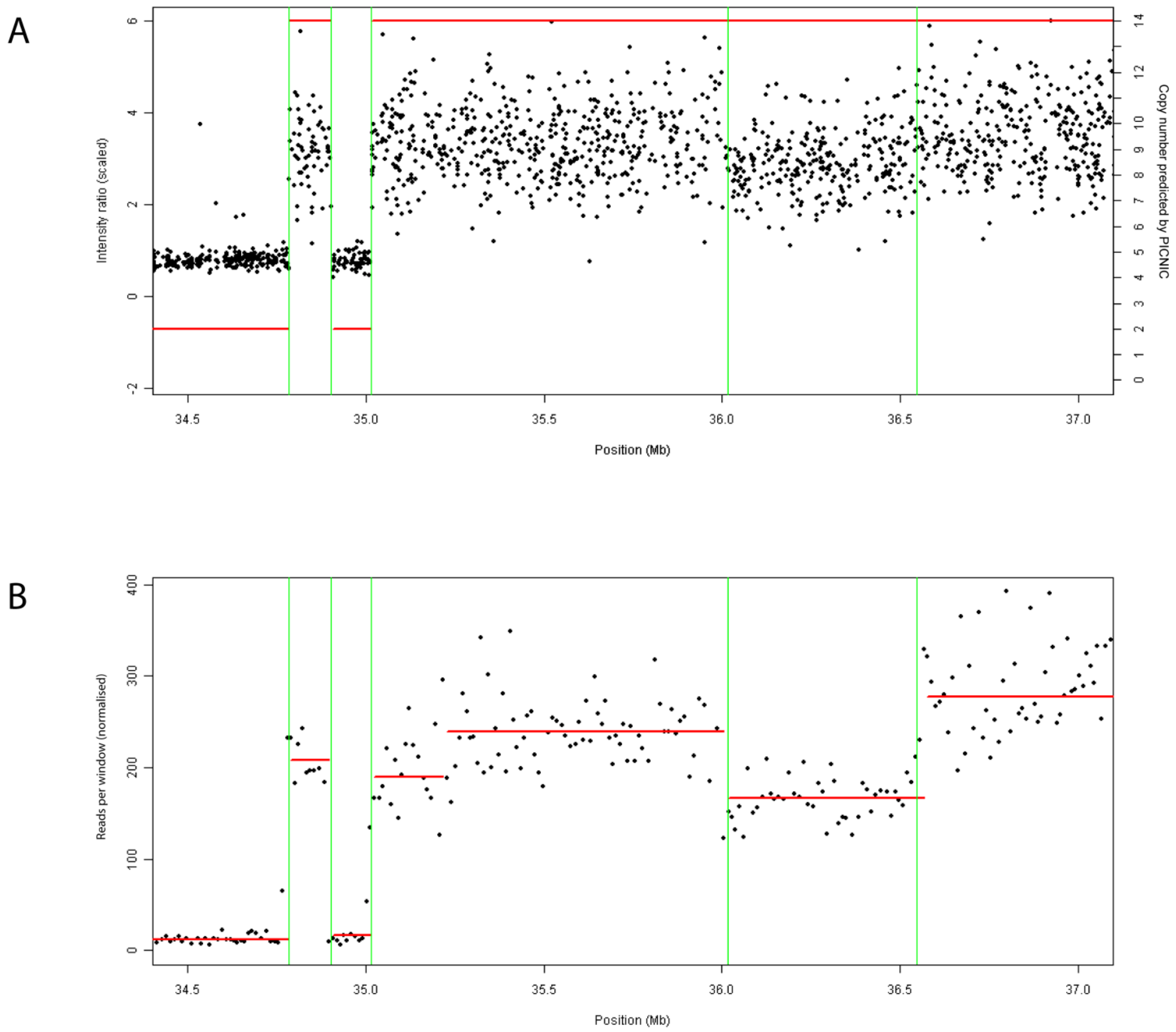


Figure 5.29. Comparison of segmented copy number for MDA-MB-134 chromosome 8. A – SNP6.0 array data, with segments predicted by the PICNIC algorithm (Greenman et al., 2010) shown in red. B – Illumina sequencing copy number data for the same region, with segments predicted by DNACopy shown in red. The vertical green lines mark known breakpoints confirmed by PCR and capillary sequencing. In A, the amplification appears to be 7-fold, with a shift from 2 copies to 14 copies (the SNP6.0 data is taken from the endoreduplicated sideclone of MDA-MB-134, which is why distal 8p is present at 2 copies rather than 1). In B, the amplification appears to be 16-fold.

The search for fusion genes in MDA-MB-134 was disappointing, as 12 fusion genes were predicted and none were found to be expressed. This was particularly disappointing as two of the predicted fusions involved genes known to be fused in epithelial cancers. An expressed in-frame fusion of *SHANK2* is known in a melanoma cell line (Berger et al., 2010). The potential fusion partner in MDA-MB-134 is *EPB49*, also known as dematin, which encodes a cytoskeletal protein which has been implicated in prostate tumorigenesis. Truncated forms of *EPB49* have a dominant negative effect and cause cytoskeletal abnormalities and altered cell shape (Lutchman et al., 1999), raising the possibility that a fusion of *EPB49* could also act as a dominant negative inhibitor of the wild-type protein.

ODZ4 was another known fusion partner predicted to be fused in MDA-MB-134. Although no fusion transcript was detected, *ODZ4* was substantially upregulated in MDA-MB-134 compared to the normal control cell lines. *ODZ4* lies outside the highly amplified region, which suggests another mechanism causes overexpression rather than an increase in expression due to increased copy number. *ODZ4* is also overexpressed in BT-474, for which high-resolution SNP6.0 data is also available (data provided by the Cancer Genome Project (Bignell et al., 2010)). Although there are rearrangements on chromosome 11, there is no copy number change in *ODZ4*. SNP6.0 data is not available for PMC42, HCC1500, or SUM52, the other cell lines which overexpress *ODZ4*, but data from a custom oligonucleotide array containing 30,000 probes suggests that SUM52 may have an amplification of *ODZ4*, while PMC42 and HCC1500 do not show a copy number change. (Data kindly provided by Dr Suet-Feung Chin, Cancer Research UK Cambridge Research Institute.) This suggests that multiple mechanisms may cause *ODZ4* expression, including copy number gain and expression as part of a fusion gene. Additionally, PMC42 has been suggested as a good model of normal breast epithelium despite originating from a breast cancer, as it has similar mRNA and miRNA expression profiles to HB4a (Git et al., 2008), but it has higher expression of *ODZ4* than the cell lines derived from normal breast epithelium.

Chapter 6

Bioinformatics of high-throughput sequencing of breast cancer

6.1 Introduction

The primary aim of the high-throughput sequencing I performed on cancer cell lines was to identify structural variants in the cancer genome. I chose to perform paired-end sequencing, which sequences a short read from either end of a fragment of a known size, rather than single end sequencing, which produces only one sequence read from each piece of fragmented DNA. Paired-end sequencing is a better method for detecting structural variants than single-end sequencing, as sequencing from either end of a DNA fragment will detect a rearrangement occurring anywhere in the fragment. This gives a much higher level of coverage for rearrangements for the same amount of actual sequence from a single-end read, where rearrangements will only be detected if there is a read across the rearrangement.

Finding the structural variants in a tumour genome usually relies on comparing the tumour genome to a reference genome by aligning all the sequences from the tumour to the reference genome and detecting any variation. This is easier than assembling the tumour genome from scratch as it is difficult to assemble a human genome from current short sequence reads (Medvedev et al., 2009). Paired-end reads from a structural variant in the cancer genome will produce an identifiable signature depending on the type of rearrangement, and these signatures can be detected and used to find the underlying structural rearrangement. The read coverage across the genome can also be used to detect unbalanced structural variants, as a change in copy number will cause a change in the number of reads across that region of the genome.

Although the idea of sequencing the end of genomic fragments to find structural variants has been used in the past (Volik et al., 2003), this was limited by the capacity of capillary sequencing and produced small numbers of sequences from large DNA fragments. High-throughput sequencing produces orders of magnitude more sequence reads than previous approaches, and while some bioinformatics software was available to analyse high-throughput sequencing data, at the time this project was started there was no software available which would use paired-end sequencing data to predict structural variants and copy number changes, and any effect they had on genes. I developed a bioinformatic analysis pipeline using a combination of available software packages and some custom software which would process the raw sequence data and generate copy number and structural variation information. The steps in the pipeline are shown in Figure 6.1.

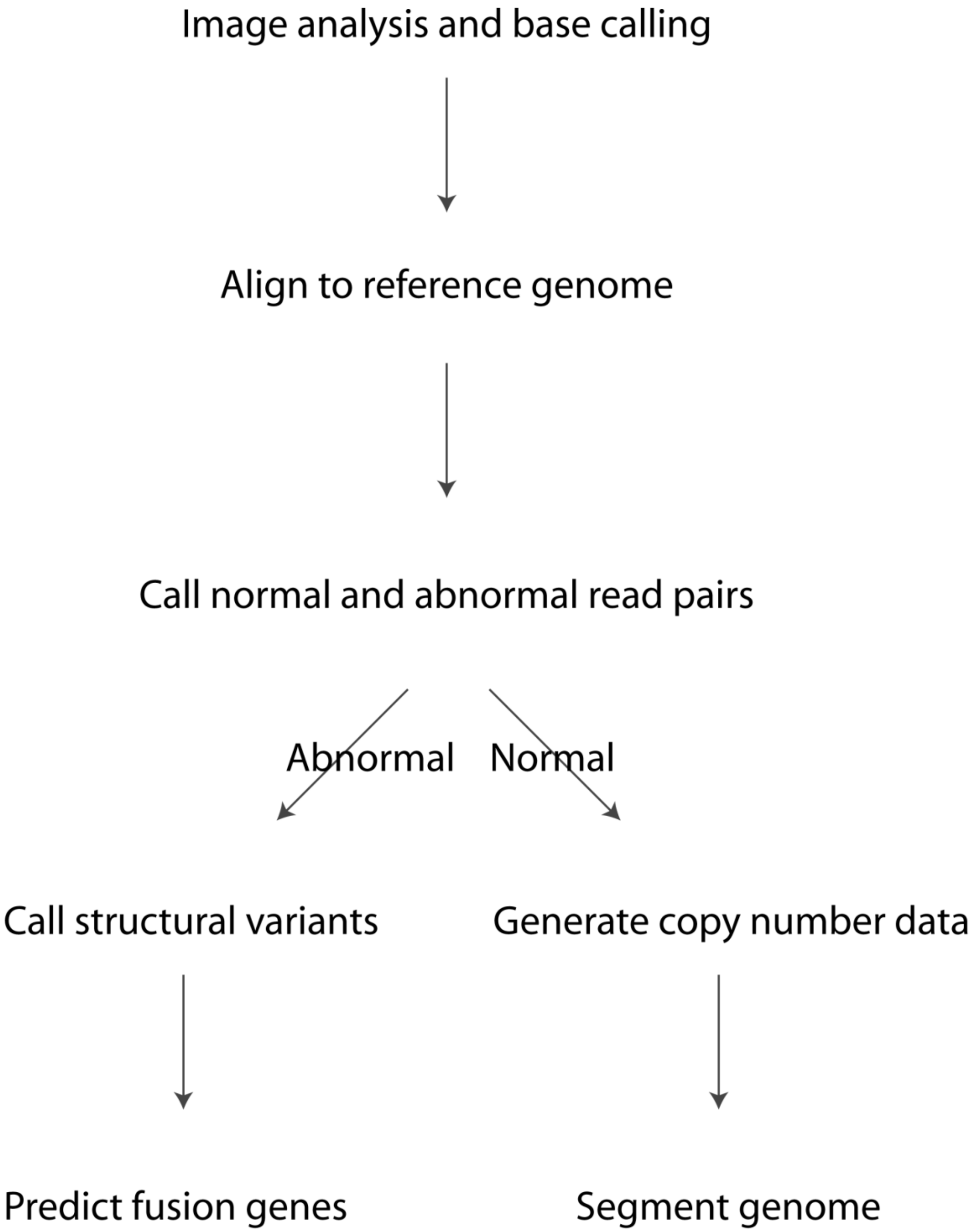


Image analysis and base calling
Align to reference genome
Call normal and abnormal read pairs
Abnormal
Normal
Call structural variants
Generate copy number data
Predict fusion genes
Segment genome

Figure 6.1. Outline of the pipeline used to process high throughput sequencing data

6.2 Alignment of high-throughput sequencing reads

The process of DNA alignment for high-throughput sequencing involves comparing a short sequence read to a reference genome and determining the most likely position within the genome that produced the sequence. Sequence alignment algorithms specifically designed for high-throughput sequencing perform differently to those optimized for other applications, as they can make certain assumptions – for instance, we can assumed that all matches will be perfect or near-perfect, which is not true of alignment algorithms designed to find alignments between different species (Flicek and Birney, 2009). Sequence alignment in general involves a trade-off between the sensitivity of the alignment and the speed of the algorithm, as a faster algorithm may not find a more distant alignment, or may misalign some sequences. Designing algorithms specifically for high-throughput sequencing allows them to perform quickly enough to cope with a high volume of data without too much of a trade-off in accuracy.

To get the raw sequences, the image analysis (FIRECREST) and base calling modules (BUSTARD) of the standard Illumina pipeline were used. The sequence alignment program MAQ (Li et al., 2008) was then used to align the paired-end reads. MAQ was one of the first algorithms developed specifically to handle high volumes of short sequence reads, and adapts the seed and extension method used by earlier algorithms. BLAST is an early example of this approach – short exact hits or 'seeds' within the sequence are found, and then each seed is extended into a longer alignment (Altschul et al., 1990). This allows for fast searching by initially searching for only the short seed sequences and then searching for longer alignments only where a short match has already been found, narrowing the search space. PatternHunter (Ma et al., 2002) improved this method through the use of non-contiguous seeds, which were shown to increase the sensitivity of matches as they are less affected by a mutation at a single base pair. MAQ uses the first 28bp of each to read to create six non-contiguous seeds, which increases the match sensitivity, and speeds up the searching by creating six hash tables, one for each seed. The hash table is created by taking the base pairs at each position in the seed and using a hash function to generate an integer value based on them. The reads can then be ordered according to this integer and grouped together in memory. The same hash function is then applied to a 28bp subsequence of the reference sequence, which also generates an integer which can be used to look up the indexed reads which gave the same integer. If a hit is found then the match is extended beyond 28bp to see if it matches the whole length of the sequence. This is repeated over the six hash tables for all possible 28bp subsequences of the reference genome to find the best match.

To find the best match for a sequence in the genome, MAQ searches for the ungapped match with the lowest mismatch score, which is defined as the quality scores of the bases that are mismatched. For greater speed, MAQ only considers hits with two or fewer mismatches in the first 28 positions of the read. Each alignment is assigned a quality score, which is a measure of the probability that the alignment reported by MAQ is the correct alignment. For each quality score, Q:

$$Q = -10log_{10}\mathrm{Pr}\,\{read\ is\ incorrectly\ mapped\}$$

A quality score of 30 indicates a 1 in 1000 chance that the read is incorrectly mapped.

If MAQ finds multiple alignments with equally good quality scores, it will return one alignment at random and give a quality score of 0, allowing these alignments to be easily identified and filtered out. I used a quality threshold of 35 to decide whether to retain a read pair for further processing, which removes low quality alignments which may be misaligned, and also removes any reads that do not have a single unique match as non-uniquely mapping reads will give spurious results if used for structural variant calling.

Multiple read pairs that map to identical positions in the genome are likely to be duplicates of the same fragment created during the PCR amplification step. After the reads had been aligned with MAQ, the first step was to identify reads that were likely to be PCR duplicates and remove all but one of the identical read pairs.

6.2.1 Calling normal and aberrant reads

A normal read was defined as a pair of reads which aligned to the genome with the expected size and orientation (Figure 6.2). The expected size was determined from the size distribution of the DNA fragments in the library. Anything within 3 standard deviations of the median was considered inside the normal range, as the library size distribution is approximately normally distributed. The expected orientation is with the read aligned to the positive strand in a lower position on the chromosome, and the read aligned to the negative strand in a higher position on the chromosome.

Any read pair that did not meet the criteria for a normal read was called as an aberrantly mapping read. Aberrantly mapping reads are candidates to be reads from DNA fragments that contain a structural difference from the reference human genome.

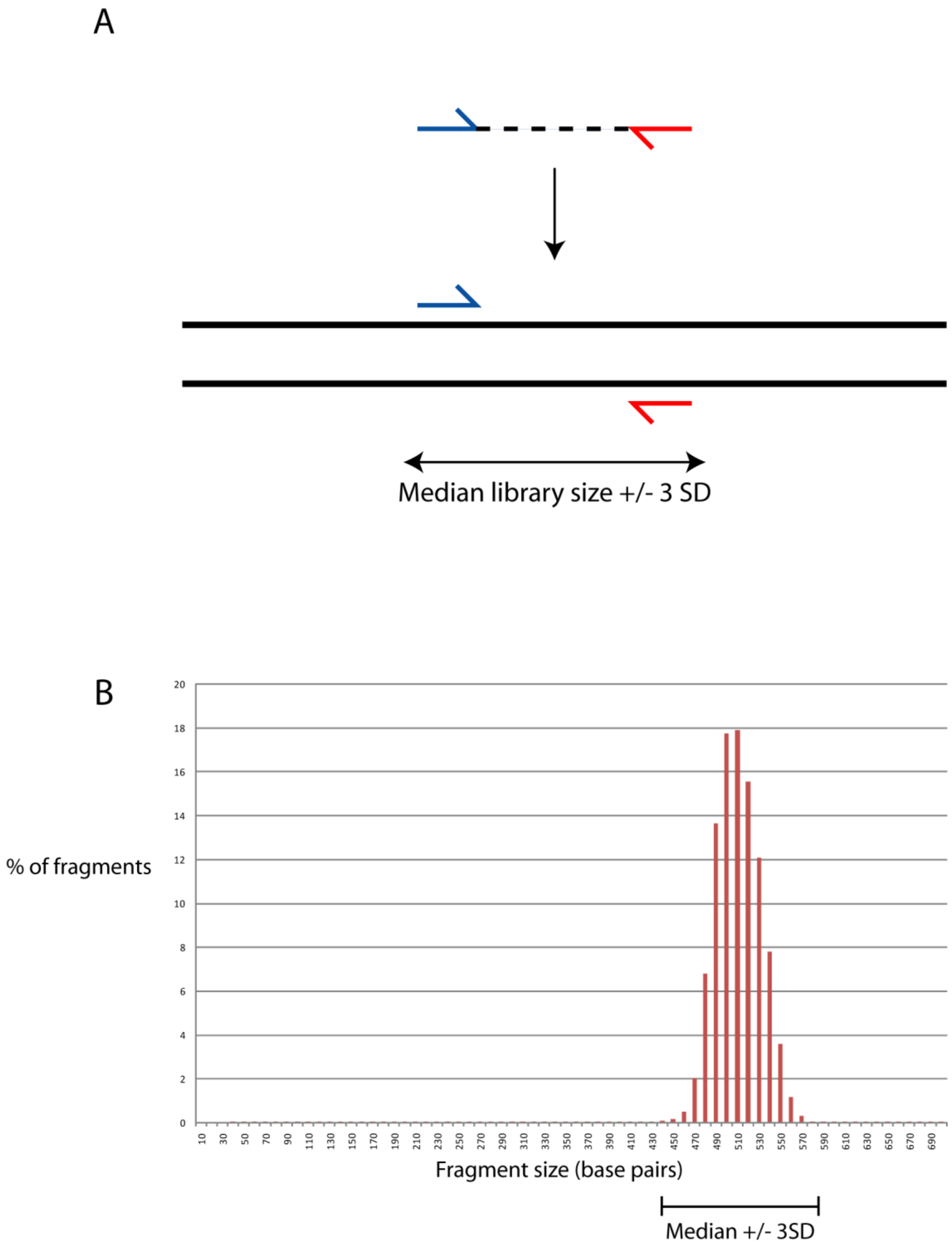


Figure 6.2. A - a normal read pair aligned to the reference genome. The blue read is in the lower position on the chromosome and aligns to the positive strand, while the red read is in the higher position on the chromosome and aligns to the negative strand. B - the library size distribution of a typical small insert library showing the range of fragment sizes which are considered as normal fragments. (Example taken from ZR-75-30 cell line library.)

6.2.2 Processing mate-pair data

There are two types of paired-end library. Standard small-insert libraries have a fragment size of up to 800bp. Mate-pair libraries use longer fragments which are circularized, ligated, and fragmented a second time to produce a library with a larger insert size than can be achieved using standard small-insert libraries.

Mate-pair libraries need to be processed differently, as a normal read from a mate-pair library will have a different size and orientation than a standard small-insert library (Figure 6.3). In a mate-pair library, the paired reads are taken from the ends of a longer DNA fragment. As long DNA fragments cannot be directly sequenced on the Illumina platform, longer fragments are circularized and the junction labelled with biotin. A smaller fragment containing the ligated junction is cut out and these junction fragments are pulled out using avidin-labelled beads. The junction fragments are similar in size to the fragments from a standard small-insert library and are sequenced in the same way as a standard library. As the DNA has been circularized, when the sequence reads from the ends of the small fragments are aligned to the genome, they will align in the opposite orientation to the normal fragments of a small-insert library (Figure 6.3). The size range for a normal fragment is also larger, as the size distribution of the fragments in a mate-pair library is wider than for a small insert library.

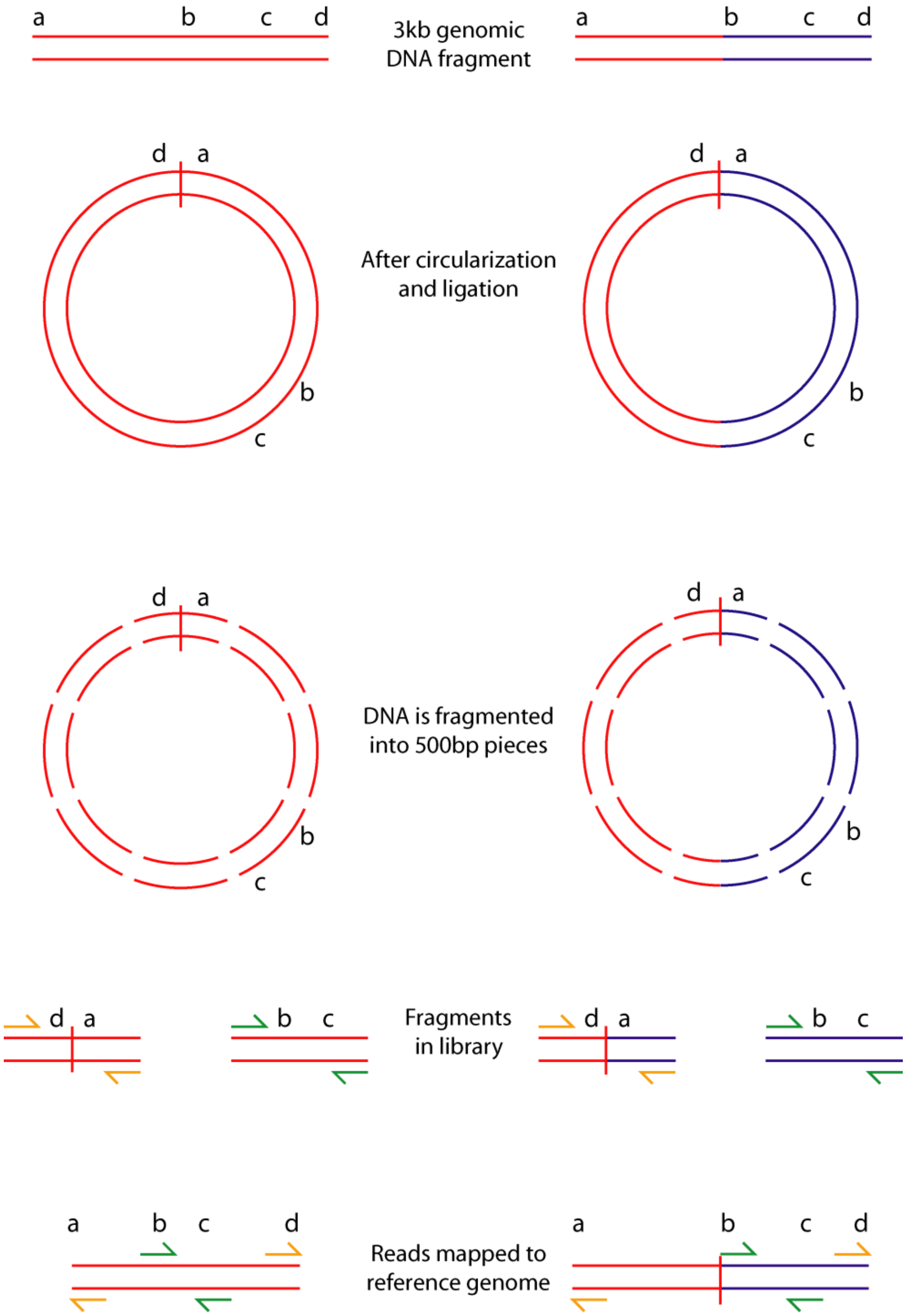

a
b
c
d
3kb genomic DNA fragment
a
b
c
d
d
a
b
c
After circularization and ligation
d
a
b
c
d
a
b
c
DNA is fragmented into 500bp pieces
d
a
b
c
d
a
b
c
Fragments in library
d
a
b
c
a
b
c
d
Reads mapped to reference genome
a
b
c
d

Figure 6.3. Construction of mate-pair libraries. After circularization, ligation and fragmentation of the libraries, there are two populations of short fragments in the library. The desired small fragments have reads shown as yellow arrows. A population of unwanted fragments will also be present in the library due to imperfect selection for the biotin-labelled fragments. The reads from these fragments are shown in green, and have opposite orientations to the desired reads.

6.3.1 Clustering and structural variant calling

To call a structural variant, two or more independent high-quality reads which supported the structural variant were needed in order to minimize artefactual predictions, on the basis that chimeras produced during the library preparation or misaligned reads are much less likely to occur twice in the same location than two real reads across a structural variant.

To find multiple read pairs which covered the same structural variant, the reads were clustered (Figure 6.4). The read pairs were first sorted so that the first read in the pair came first in the genome, and then the read pairs were sorted into order by chromosome, position, and strand they aligned to. If there were multiple read pairs where all the first reads in the pairs were on the same chromosome, aligned to the same strand, and the distance between them was less than the upper limit of the library size range, they were clustered together. The clustering process was repeated for the second reads in the pairs. Only read pairs where both of the reads from all the pairs were in the same regions are used to call the structural variants (Figure 6.5). These reads are assumed to be spanning the same breakpoint, and the true position of the breakpoint lies no further than the upper limit of the library insert size from the position of the read in the cluster which was furthest from the breakpoints.

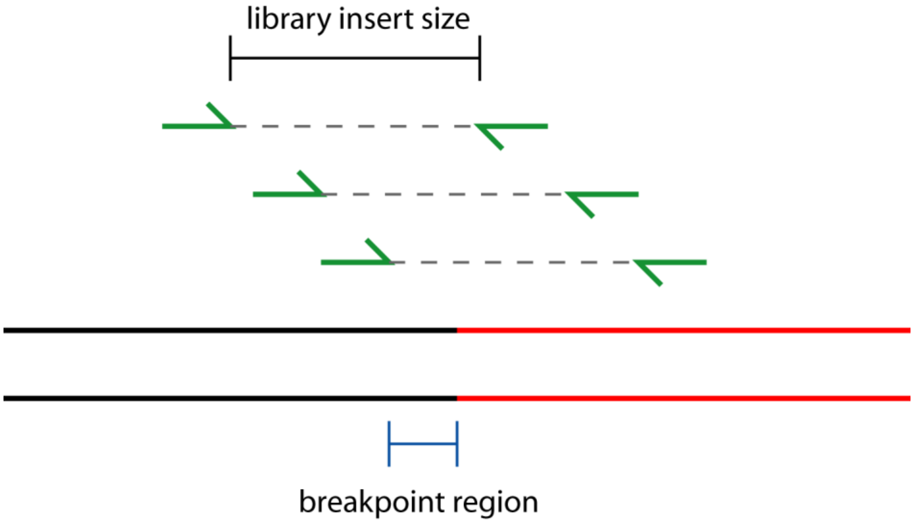


Figure 6.4. Clustering of reads to call structural variants. Paired reads are clustered if they map to the same strand and the difference in their positions is smaller than the maximum library insert size. The position of the breakpoint must lie less than the maximum library insert size from the end of the read furthest to the breakpoint.

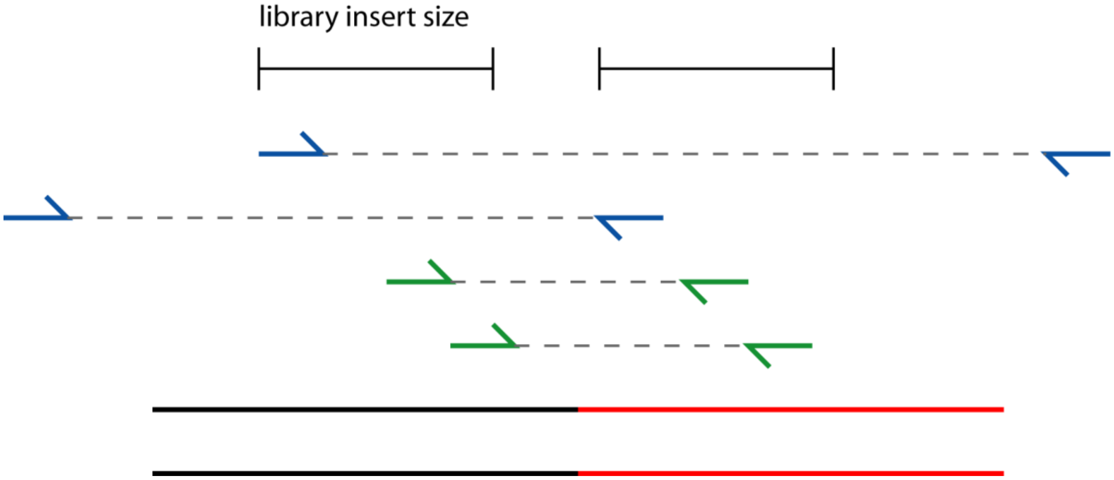


Figure 6.5. Clustering of reads to call structural variants. Only the reads shown in green support this structural variant, as both reads in the pair support the variant. The reads shown in blue would not be used to support the structural variant as only one of the reads in a pair supports the structural variant.

The structural variants were classified based on read strands and positions (Figure 6.6), and named based on the most likely chromosomal rearrangement which can be inferred from the reads.

A DIF is called when the two reads in a pair map to different chromosomes. It is most likely to be an interchromosomal translocation or insertion. A DEL is called when the two reads in a pair map further apart than the maximum library size, and the read which maps to the lower position is on the positive strand, and the read in the higher position is on the negative strand. It is most likely to be a deletion, but it could also be an insertion of material normally found at a higher position on the chromosome, with no loss of DNA. An INS is called when the read which maps to the lower position is on the negative strand, and the read which maps to the higher position is on the positive strand. It is mostly likely to represent an insertion. Head-to-tail tandem duplications will also be in this class, and as the paired reads cover only one side of the insertion it cannot be distinguished from a larger insertion. An INV is called when both reads in the pairs map to the same strand, indicating one side of the breakpoint has been inverted. An ITR is called when both reads in the pair map to the same position and the same strand, and it is a special variant of the INV class caused by a head-to-head tandem duplication.

Just as the strands for a normal read pair are reversed for mate-pair libraries, the expected strands when calling structural variants must also be reversed (Figure 6.7). A probable deletion is called from a read pair where the read in the lower position is aligned to the negative strand, and the read in the higher position is aligned to the positive strand. A probable insertion is called from a read pair where the read in the lower position is aligned to the positive strand, and the read in the higher position is aligned to the negative strand. Probable inversions and inverted tandem repeats will have both reads aligned to the same strand.

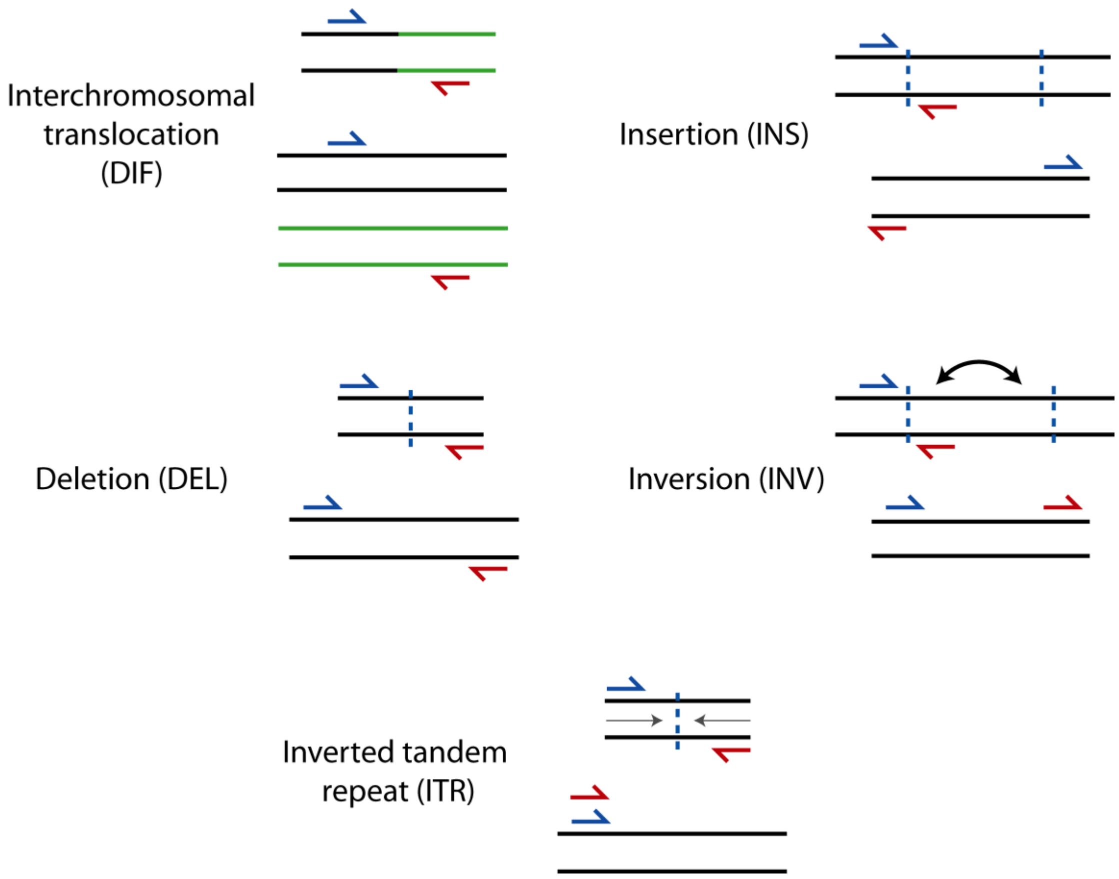

Interchromosomal translocation (DIF)
Insertion (INS)
Deletion (DEL)
Inversion (INV)
Inverted tandem repeat (ITR)

Figure 6.6. Structural variant calling of small-insert library. For each type of structural variant, the upper diagram shows the arrangement of the abnormal chromosome, and the lower diagram shows how the read pair would align to the reference genome.

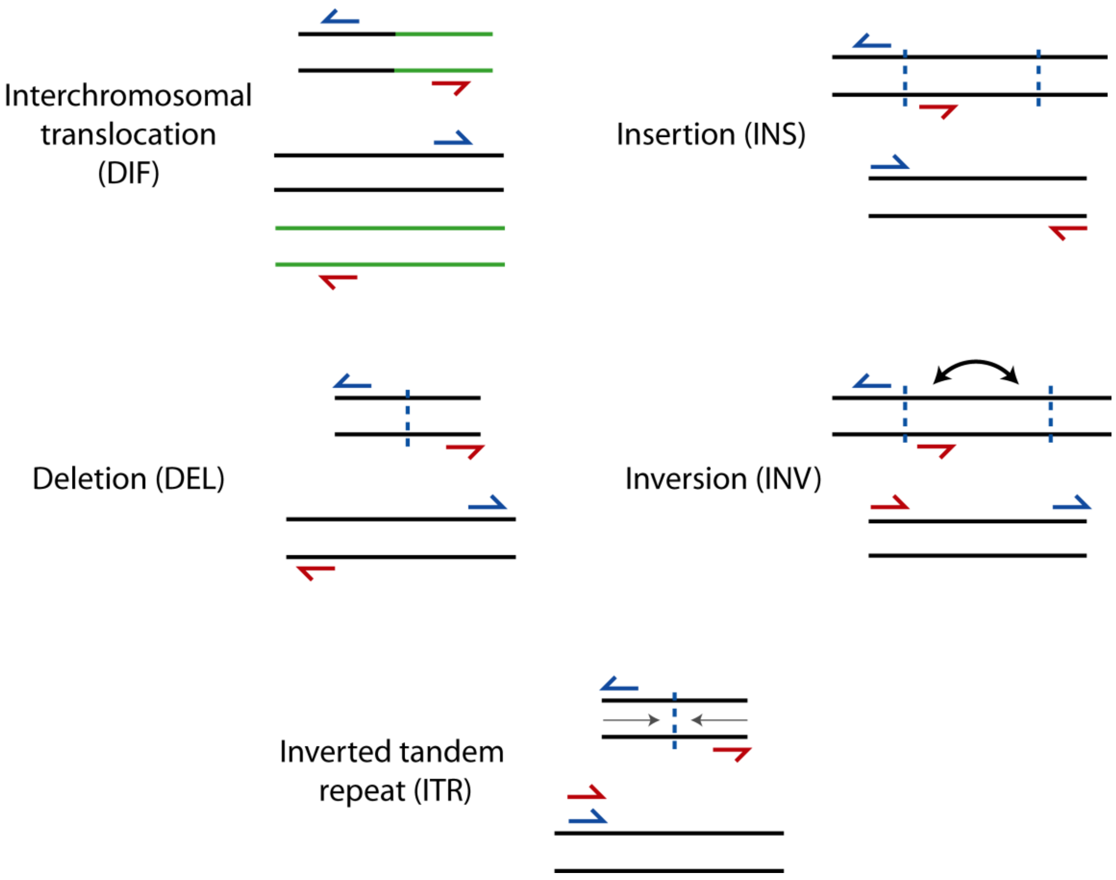

Interchromosomal translocation (DIF)
Insertion (INS)
Deletion (DEL)
Inversion (INV)
Inverted tandem repeat (ITR)

Figure 6.7. Structural variant calling from mate-pair library. For each type of structural variant, the upper diagram shows the arrangement of the abnormal chromosome and the reads produced from a mate-pair library, and the lower diagram shows how the read pair would align to the reference genome.

Structural variant calling is more difficult from mate-pair libraries because the biotin selection step of the library preparation is not perfect, and some of the unbiotinylated fragments will also be retained. These small fragments, which do not cross the circularization junction, will be retained after size selection (Figure 6.3), and they can be seen as a second peak in a graph of the library size distribution (Figure 6.8). These fragments are similar to the fragments produced from a small-insert library, and will align to the same strands as for a small-insert library – the read in the lower position will align to the positive strand, and the read in the higher position will align to the negative strand. This gives the small fragments the same read orientation as insertions in the mate-pair library, and they could cause a spurious insertion to be called. Requiring two or more hits to call a structural variant reduces the likelihood of calling these insertions, as the small fragments make up a smaller percentage of the library than the desired junction fragments, and the probability of getting two reads across the same region is lowered. The read pairs which appear to be from the small fragments can also be filtered out – this will remove a small number of real insertions along with the false positives. As much larger numbers of small insertions were called from unfiltered mate-pair libraries than from small-insert libraries, it was likely that many of them were spurious insertions, and the small fragments were filtered out before any further analysis was done.

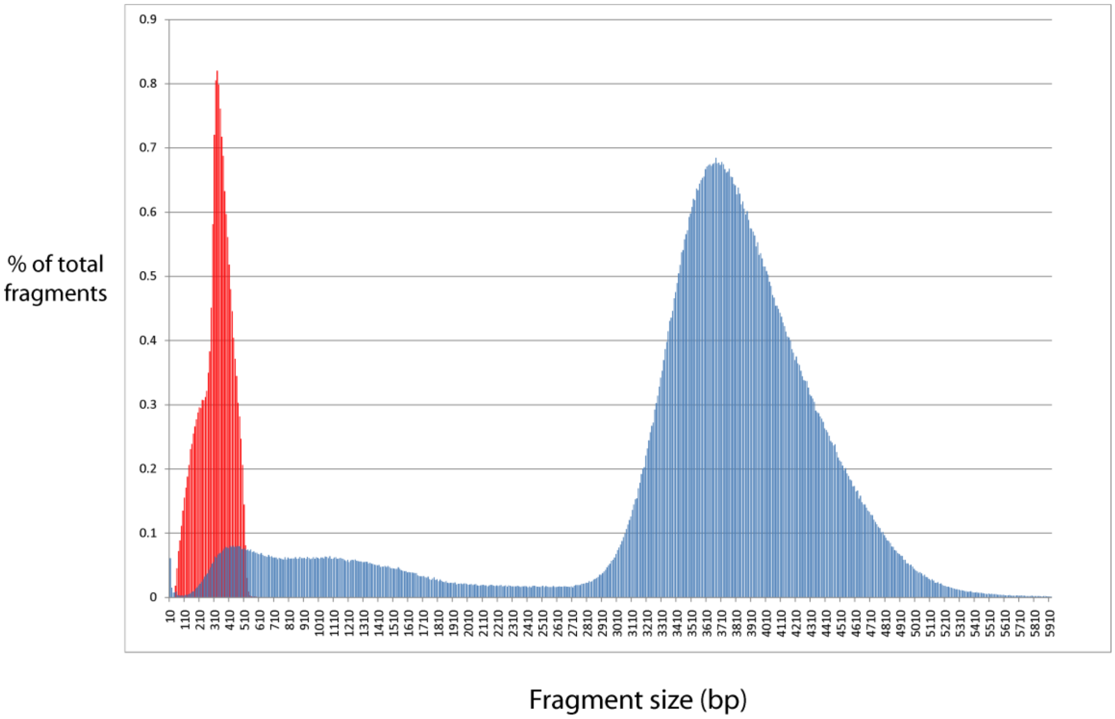


Figure 6.8. Fragment size distribution from the mate-pair library of the cell line ZR-75-30. The histogram shows the percentage of fragments of each size. The blue bars show the size distribution for the fragments where the reads map in the expected orientation for the biotinylated fragments from a mate-pair library. The red bars how the size distribution of the smaller non-biotinylated fragments which were not removed during library preparation.

6.4 Fusion gene prediction

Once a list of well-supported structural variants was produced from the short sequences, the next step was to look for any potential fusion genes which may result from these structural variants.

The list of structural variants was first checked against a list of known human copy number variations (Conrad et al., 2010) and any which matched were considered to be normal human variation and not acquired in the tumour.

The set of structural variants with known common copy number variations removed was used to predict potential fusion genes. The fusions were predicted computationally at the DNA level by using the Ensembl Application Programming Interface to retrieve all the genes which overlap the breakpoint region of a structural variant, and predicting whether a fusion transcript could be formed based on the direction of the reads and the strands of the genes (Figures 6.9 and 6.10). As well as rearrangements where two genes are broken and fused, we considered rearrangements which break only one gene and remove the 3' end including the transcription stop site, which could result in transcription continuing into a downstream gene until it reaches the stop site of the downstream gene. As a breakpoint could break two genes on different strands, it is possible for more than one fusion gene to be predicted for each structural variant. As the breakpoints are not resolved to base-pair level, if the gene is broken inside an exon it was not possible to predict whether or not an in-frame transcript would be produced. Although it is theoretically possible to predict whether an intronic break will produce an in-frame fusion, alternate splicing and cryptic exons are found in fusion genes (Howarth et al., 2008), which complicates the prediction of fusions.

Structural variants can also cause genes to be internally rearranged, deleting or duplicating exons. As the majority of structural variants are small deletions entirely within introns, which are expected to have no effect on the coding sequence of the gene, the number of exons deleted is also returned by the script, so that the deletions which may affect genes can be easily prioritised over those which do not delete any coding regions.

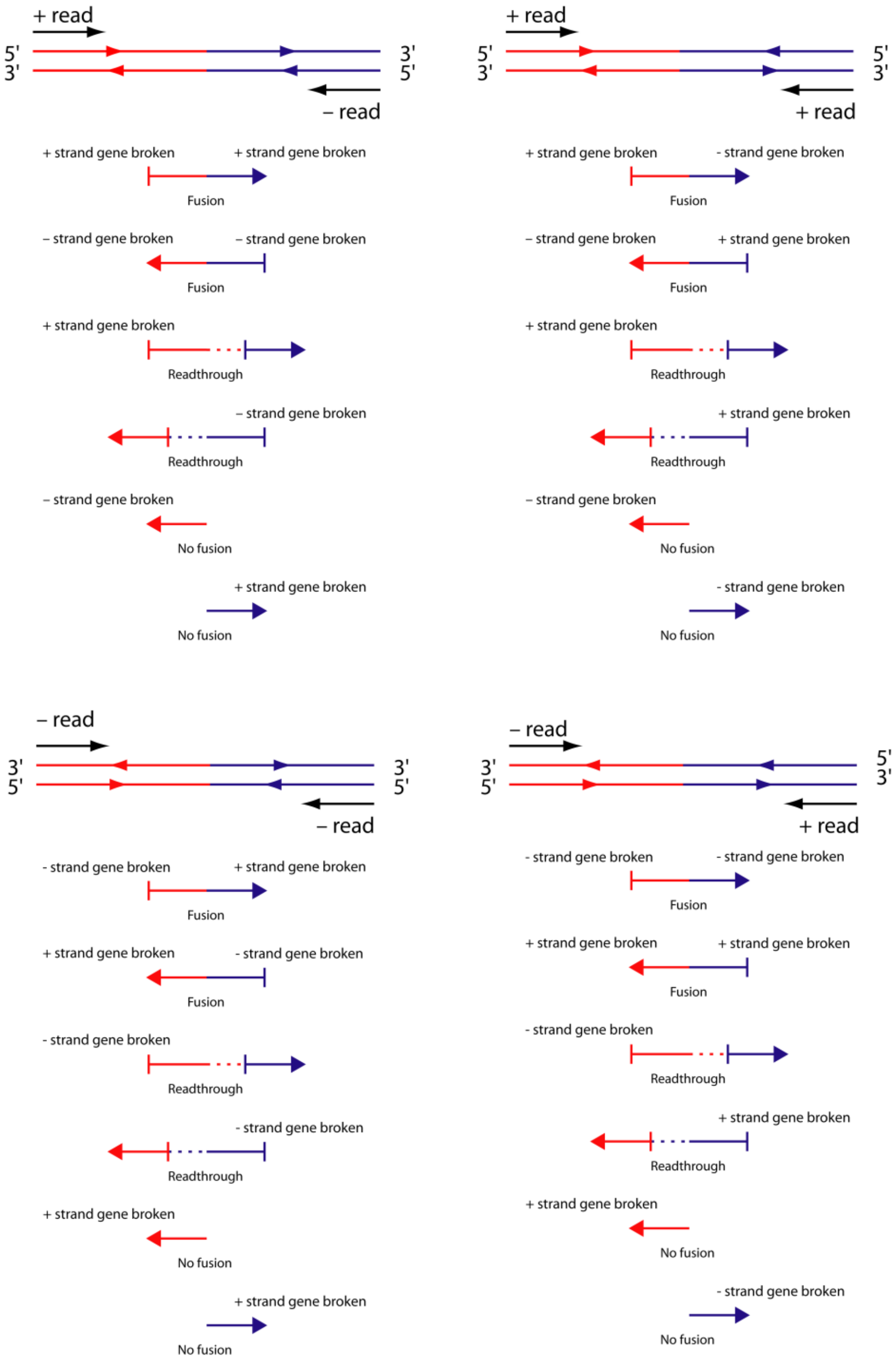


Figure 6.9. Fusion gene prediction from small-insert library. The strands of the read alignments allow the orientation of the chromosomes and genes to be determined, and predict whether a gene fusion will be formed. Depending on the orientation of the genes, a fusion gene may be predicted, there may be no possible fusion, or there may be a potential readthrough fusion. As a single breakpoint may break two genes on different strands, more than one possible gene fusion or readthrough is possible at each junction.

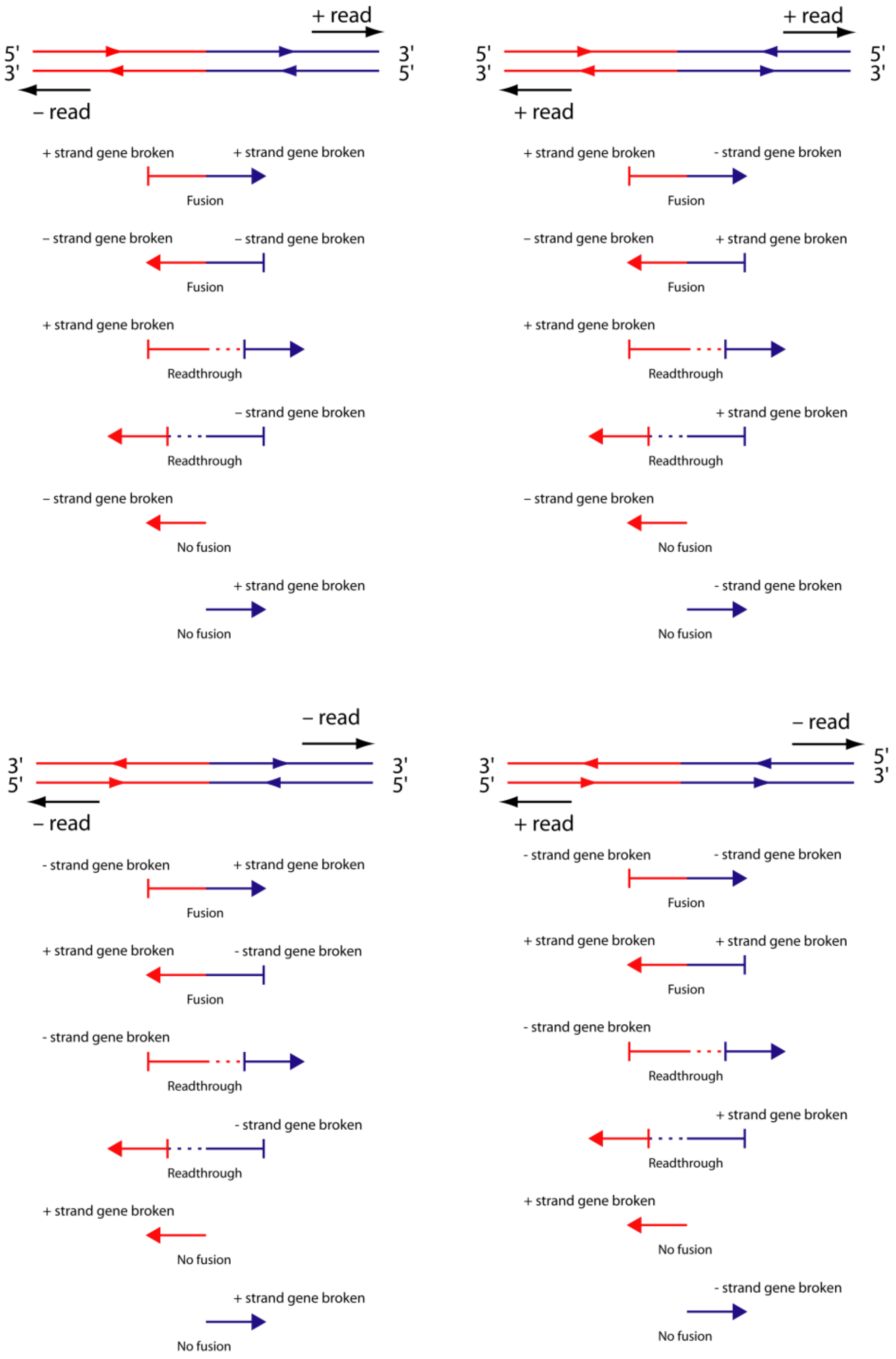


Figure 6.10. Fusion gene prediction from mate-pair library. The gene prediction is similar to for the small-insert library , but the orientations of the chromosome and genes are reversed relative to the aligned strands of the reads.

6.4.1 Validation of predicted fusion genes

To validate the computational prediction of fusion genes, I used independent structural variation data sets from the cell lines MCF7 (Hampton et al., 2009) and HCC1187 (Stephens et al., 2009a). In their paper, Hampton et al. predicted that MCF7 had 10 possible in-frame fusions caused by breakpoints occurring in the introns of 2 genes, and they found that 4 of these were present at the cDNA level. I took their published data set, which contained all the structural variants they found in MCF7, and put the data into my fusion gene prediction script. The script successfully found all 10 fusions that Hampton et al. predicted, as well as 4 other potential fusions, 3 internal deletions and 41 potential readthrough fusions.

Stephens et al. (2009) found a number of fusions and internal rearrangements in the cell line HCC1187. They found 6 expressed fusion genes, 2 of which were in-frame fusions and 4 out-of-frame fusions. I put their data set of structural variants into my fusion gene prediction script, and all 6 of the expressed fusions were found, along with 5 other predicted fusions not mentioned in their paper. At least one of the fusions I predicted (*PUM1-TRERF1*) is known to be expressed (Dr Karen Howarth, unpublished). They also looked for internal gene rearrangements and exon deletions, and found 5 expressed in-frame internal gene rearrangements and 7 internal rearrangements which were not checked for expression. All of these internal rearrangements were predicted computationally by my script, along with 8 other internal gene rearrangements. One of the 8 rearrangements not found by Stephens et al. (2009) is a tandem duplication of *RAD51L1*, which is known to be expressed (Susanne Flach, unpublished work in the lab). This validation shows that my method of fusion prediction not only successfully finds known fusion genes in other data sets, but predicts fusion genes which have been missed by other methods of fusion gene prediction.

The fusion prediction script was then used to predict fusion genes using a data set of paired end reads from cell lines and tumour samples. In the cell line ZR-75-30, 20 potential fusion genes were predicted. In collaboration with Dr Ina Schulte I looked for expression of the fusion genes and found 7 expressed fusion genes, of which 3 produced in-frame transcripts (Table 6.1) .

Type of structural variant	**5' Gene**	**3' Gene**	**Expressed?**	**In-frame?**
Insertion	*PCTK3*	*NFASC*	No	No
Translocation	*GRIP2*	*BCL11A*	No	No
Inversion	*PREX2*	*TSNARE1*	No	No
Translocation	*HYLS1*	*TIMM23*	No	No
Translocation	*STRN3*	*PLCE1*	No	No
Translocation	*FSIP1*	*BAZ2A*	No	No
Translocation	*CBX3*	*c15orf57*	No	No
Translocation	*TRAPPC9*	*STARD3*	No	No
Translocation	*NDRG1*	*HOXB4*	No	No
Translocation	*TRAPPC9*	*SPAG5*	No	No
Translocation	*PPM1D*	*TRAPPC9*	No	No
Inversion	*TAOK1*	*CA10*	No	No
Insertion	*SSH2*	*PLXDC1*	No	No
Inversion	*ZMYM4*	*OPRD1*	Yes	No
Translocation	*COL14A1*	*SKAP1*	Yes	Yes
Translocation	*APPBP2*	*PHF10L1*	Yes	Yes
Inversion	*TAOK1*	*PCGF2*	Yes	Yes
Deletion	*UPS32*	*CCDC49*	Yes	No
Inversion	*BCAS3*	*HOXB9*	Yes	No
Deletion	*TIAM1*	*NRIP1*	Yes	Yes

Table 6.1. Predicted fusion genes in the ZR-75-30 cell line.

In the paired cell lines VP229 and VP267, taken from a patient at different stages of disease, there were 27 fusions which were predicted to be in both cell lines. 3 were found to be expressed, of which 2 were out of frame, and 1 was in frame (Scott Newman and Susanne Flach, unpublished) (Table 6.2).

Type of structural variant	**5' Gene**	**3' Gene**	**Expressed?**	**In-frame?**
Insertion	*NRG3*	*c10orf11*	No	No
Translocation	*GRIK1*	*CPXM2*	No	No
Deletion	*OR52N1*	*TRIM5*	No	No
Insertion	*NRG3*	*SAMD8*	No	No
Translocation	*PDLIM1*	*ZBBX*	Yes	No
Inversion	*ACADSB*	*ADAM12*	No	No
Translocation	*PDLIM1*	*TNIK*	No	No
Inversion	*FAM125B*	*SPTLC1*	Yes	No
Translocation	*NRG3*	*GRIP1*	No	No
Translocation	*DLG5*	*KCNMB2*	No	No
Translocation	*MYNN*	*NRG3*	No	No
Inversion	*AL356155.1*	*SORCS1*	No	No
Inversion	*ROR2*	*NEK6*	No	No
Deletion	*ZFAND2a*	*c7orf50*	No	No
Translocation	*CPLX1*	*DUSP14*	No	No
Inversion	*UBTD1*	*SLIT1*	No	No
Inversion	*PBX3*	*ROR2*	No	No
Translocation	*MDS1*	*KCNMA1*	Yes	Yes
Deletion	*FNBP1*	*FAM129B*	No	No
Inversion	*FAM129B*	*NEK6*	No	No
Insertion	*APOBEC3G*	*APOBEC3D*	No	No
Translocation	*DPY19L2*	*DPY19L2P2*	No	No
Inversion	*AATF*	*AC113211.2*	No	No
Translocation	*c10orf11*	*c17orf63*	No	No
Translocation	*c17orf63*	*c10orf11*	No	No
Translocation	*ADK*	*KCNMB2*	No	No
Translocation	*UBR4*	*ZFP37*	No	No

Table 6.2. Predicted fusion genes in both of the paired cell lines VP229 and VP267.

6.5 Copy number variation

The read pairs which map normally were used to find copy-number alterations. The number of sequence reads within a given interval which align to the genome should be proportional to the copy number, and the resolution is limited only by the read coverage across the genome. To get accurate copy-number from sequencing, the data must be corrected for the repeat content and GC percentage across the genome.

6.5.1 Correcting for mappability

Repetitive regions of the genome will have fewer aligned reads than unique regions as fragments produced from repeat regions during library preparation are likely to align perfectly to multiple regions of the genome, and will be discarded during data processing. To account for this 'mappability' of the genome, the start positions of all the genomic locations where a 35bp read would give a single match were simulated. This list of mappable starts was used to divide the genome up into windows which each contain the same number of mappable starts, giving windows of variable size. Highly repetitive regions of the genome give larger windows.

The sequence reads were binned into the windows across the genome, and the number of reads in each window should be proportional to the copy number.

6.5.2 Correcting for GC content

The library preparation protocols for the Illumina Genome Analyzer are known to introduce bias and produce greater numbers of reads from regions with high GC content (Dohm et al., 2008). This bias was reduced by lowering the melting temperature during the gel extraction of the library preparation protocol, which has been shown to considerably reduce the GC bias (Quail et al., 2008). To examine the GC content bias in our data, the GC percentage for each window was retrieved using the Ensembl Application Programming Interface, and plotted against the number of reads in each window (Figure 6.11). This showed a bias in our data at both ends of the scale, with fewer reads in windows with either a high or a low GC content. Although a bias towards fewer reads in regions of low GC content was previously seen (Dohm et al., 2008), the corresponding drop-off in read number at high GC was not seen by previous studies. This may be because the effect was masked by the larger GC bias produced during the library preparation, as they had not followed the improved protocol described above, or because their data was generated from bacterial genomes, and there were few regions with high enough GC content for the effect to be seen. This bias is also seen as a 'wave' in the uncorrected copy number data when plotted alongside the GC percentage of the genome (Figure 6.12).

The GC bias was noted to be similar to the GC 'wave' seen by Marioni et al. (2007) in array CGH data*,* and a similar method of loess correction was used to normalise the data. A loess curve was fitted to the a plot of GC content against the number of reads per window, and the values predicted from the loess curve were used to normalize the number of reads per window (Figure 6.13). This produced better copy number data as judged by eye.

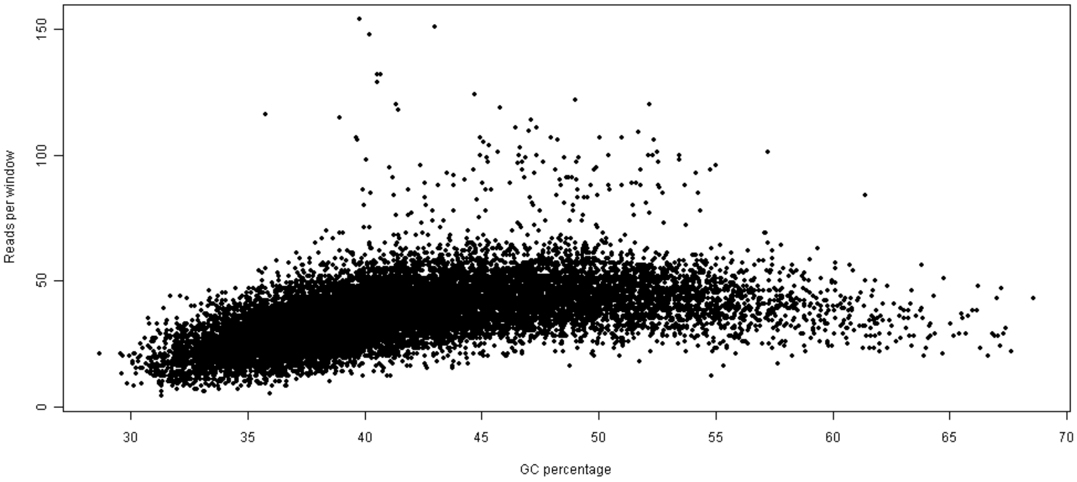


Figure 6.11. The number of reads in a window plotted against GC percentage of the window to show the bias against reads at both low and high GC percentages. Data shown is from MDA-MB-134, chromosome 1. Most of chromosome 1 is present in two copies, with a small region present at higher copy number.

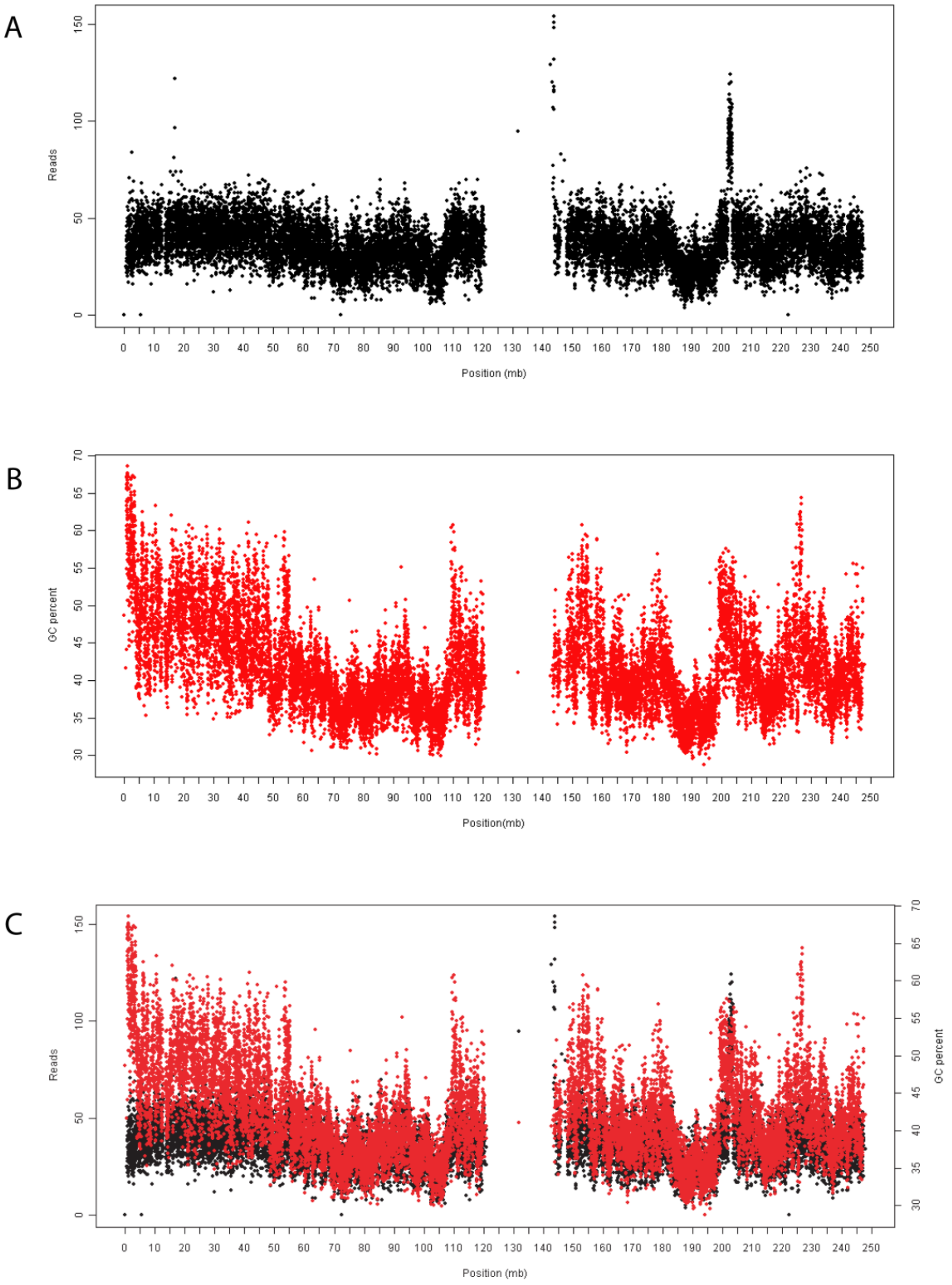


Figure 6.12. Copy number plots for MDA-MB-134 chromosome 1 showing GC content bias. A - the number of reads in each window across MDA-MB-134 chromosome 1, corrected for mappability but not GC percentage. B - the GC percentage of each window. C - the GC percentage and reads per window plotted on the same graph.

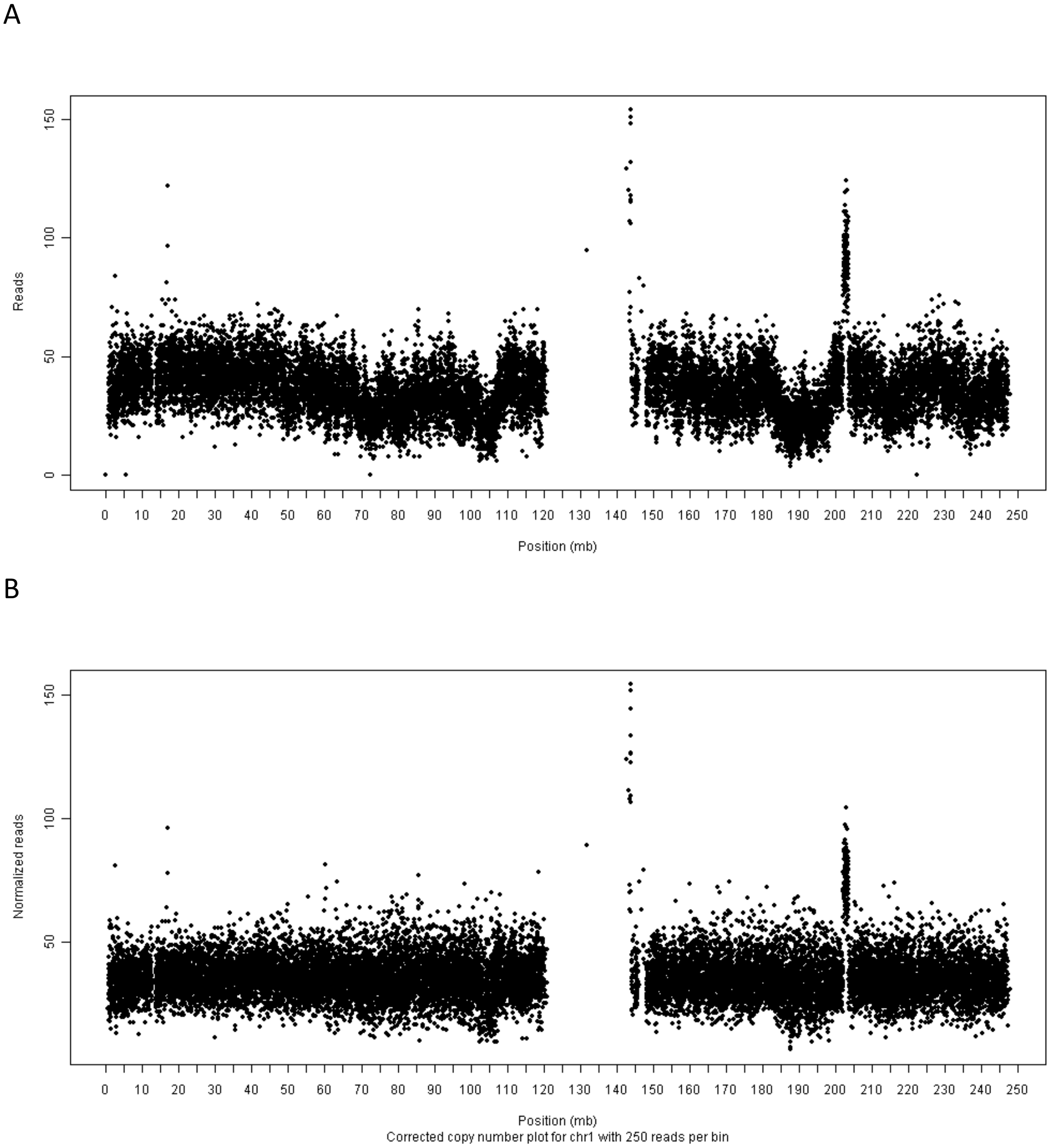


Figure 6.13. Correction of copy number data for GC content bias. A - Number of reads per window across MDA-MB-134 chromosome 1, uncorrected for GC content. B - Number of reads per window across MDA-MB-134 chromosome 1 after loess correction for GC content bias.

6.4.4. Segmentation and copy number analysis

Segmentation in the context of copy number data refers to the process of computationally determining breakpoints and copy number alterations from a dataset of copy number information. A number of segmentation methods have been previously developed, primarily for analysing array CGH data. The SegSeq algorithm (Chiang et al., 2009) is currently the only segmentation method designed specifically for high-throughput sequencing, but it relies on the use of a matched normal sample to call breakpoints, and was therefore unsuitable for my purposes. Instead I used DNAcopy (Olshen et al., 2004; Venkatraman and Olshen, 2007), which uses circular binary segmentation to call breakpoints. DNAcopy has been shown to perform better than other segmentation methods on both simulated and real CGH data (Lai et al., 2005; Willenbrock and Fridlyand, 2005).

DNAcopy allows the user to set different parameters which determine how the segmentation is performed, and allow the algorithm to be "tuned" for better performance on a particular dataset (for example, to reduce false positives from data with a low signal-to-noise ratio) (Lai et al., 2005). There is little available information on how to best choose parameters for segmentation, although a study on the GLAD algorithm concluded that the choice of parameters had minimal effect on their data set (Rigaill et al., 2008). DNAcopy had been previously used to segment breast cancer cell line CGH data using the default parameters (Venkatraman and Olshen, 2007), and I changed only one parameter from the default, which was to use an "undo" method to remove breakpoints detected due to local trends in the data, as is recommended by the authors of the method. The parameters were tested by comparing the segments produced using different parameter sets for chromosomes from MDA-MB-134 and comparing them to the segmentation produced by the PICNIC algorithm from Affymetrix SNP6 data. The chosen parameters correctly detected known copy number changes without adding extra segments which did not appear to be supported by the data. The exception is at the centromeres where additional copy number segments may be called due to incorrect mapping of reads to repeat regions. A known weakness of the circular binary segmentation method is an inability to detect small regions of copy number change in the middle of chromosomes (Olshen et al., 2004), but DNAcopy successfully segments the small amplification on MDA-MB-134 chromosome 8 (Figure 6.14).

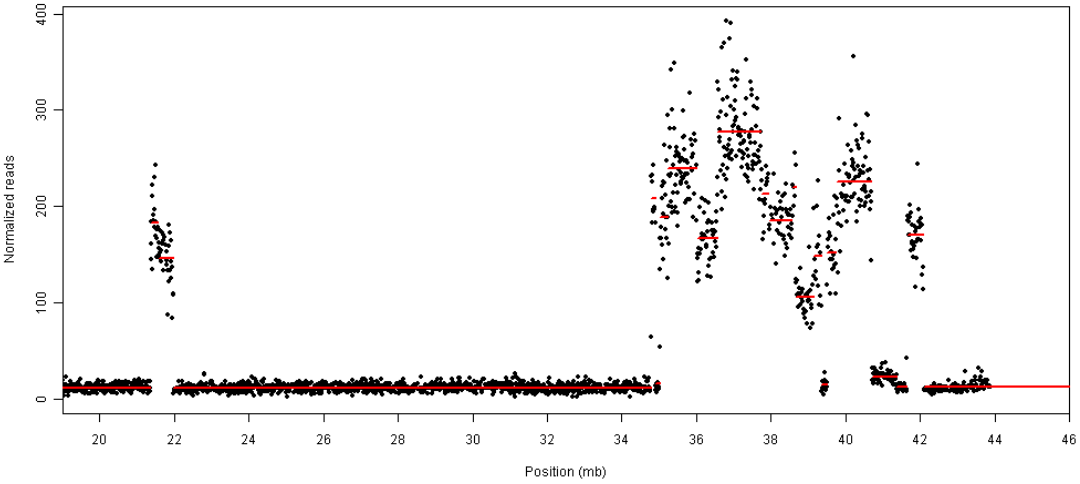


Figure 6.14. Segmented copy number plot for part of MDA-MB-134 chromosome 8. A weakness of circular binary segmentation is the inability to detect small regions of copy number change, but using DNACopy the small amplification at 21.8Mb is successfully segmented.

6.6 Discussion

The bioinformatic pipeline was validated by using data from a number of cell lines to assess how well it performed the analysis we required.

The structural variant calls from MDA-MB-134 have been the most thoroughly analysed (see Chapter 5 for details). My analysis prioritised the validation of translocations and large genome rearrangements, as they were more likely to affect genes. Of the 42 large rearrangements predicted by the pipeline, 7 were filtered out by re-aligning the reads using BLAT, 4 could not be validated by PCR, and 7 were also present in a pool of normal female DNA, leaving 24 which could be validated and sequenced, or around 57% of the predicted variants.

The 3 different ways that false positive variants were identified suggest there are multiple ways that the analysis could be improved. Improving the alignment steps of the pipeline is an easy way to remove the structural variants that are caused by misalignment. The variants that could not be validated by PCR are more difficult to address, as it is not possible to tell whether they are spurious structural variant calls, or whether they are true variants which fall in a region that is difficult to PCR. The variants that are also present in the normal female DNA do not represent a failure of the bioinformatic pipeline, as they are real variants which are present in the sequencing, but future sequencing projects will involve tumour genomes sequenced alongside the matched normal genome. Bioinformatic methods can be used to filter out any variants that appear in the normal genome. The smaller predicted structural variants have not been validated, so the number of false positives is unknown.

The fusion gene predictions were tested in cell lines by looking for expression of a fusion transcript. The number of predicted fusions which were expressed ranges from 0 out of 4 in MDA-MB-134, 7 out of 20 in ZR-75-30, and 3 out of 27 in VP229 and VP267. These figures are for gene fusions only, not including readthrough fusions or internal rearrangements which have not yet been completely tested in any cell line except MDA-

MB-134. The Stephens et al. (2009b) study found a higher proportion of their predicted fusion genes were expressed than in my study, but their predictions missed at least one expressed fusion gene, suggesting that my fusion prediction pipeline may predict more fusion genes which prove not to be expressed, but also find real fusion genes which other studies would miss. They also found that fewer of their predicted fusion genes were expressed in amplified regions, and all of these cell lines contain highly rearranged amplicons which may contribute to the lower number of expressed fusions.

The copy number segments which were predicted for MDA-MB-134 agree with both our previous knowledge of MDA-MB-134 from FISH and microarray studies (Paterson et al., 2007), and the breakpoints which are known to base pair resolution fall within the breakpoint regions predicted by DNACopy. Although the methods used to produce copy number segments are accurate, a disadvantage of the methods is that the breakpoints cannot be determined to a higher resolution than the size of the windows used to divide up the reads. As sequence coverage increases, a smaller window size can be used to refine the breakpoints, but a better method is to use a breakpoint calling technique which is not reliant on fixed window size. The SegSeq algorithm uses hidden Markov models to predict copy number change points, and even at relatively low coverage (~15million 36bp reads) it can predict breakpoints to within 1kb (Chiang et al., 2009). The rSW-seq algorithm, which uses a Smith-Waterman approach to map breakpoints, is another copy number prediction algorithm developed to avoid the use of windows and improve resolution. While both these algorithms identify breakpoints at higher resolution than my window-based approach, they both require normal sequences to compare against the tumour sequences, and they have so far been tested using cell line data which will not have the problems of stromal contamination found in tumour samples. Using a matched normal sample and calling the differences between the tumour and the normal sample also removes the problem of GC content bias, as the biases should be the same in both samples.

Chapter 7

Discussion

7.1 How prevalent are fusion genes in breast cancer?

Part of the central hypothesis of my thesis is that the prevalence of fusion genes in solid tumours has been underestimated, and my study of breast cancer cell lines supports this idea. At the start of my project, only 5 fusion genes had been found in breast cancer, all in cell lines – a fusion of *ODZ4* to *NRG1* in the cell line MDA-MB-175 (Liu et al., 1999), a fusion of *FHIT* to a cDNA later identified as *MACROD2* in BrCa-MZ-02 (Popovici et al., 2002), a fusion of *BCAS4* to *BCAS3* in MCF7 (Barlund et al., 2002)*,* and the *RIF1-PKD1L1* and *TAX1BP1-AHCY* fusions recently found in HCC1806 (Howarth et al., 2008). In HCC1806, which has two known fusion genes, I found no further fusion genes, and in MDA-MB-134 no fusion genes were found. However, the predictions of fusion genes I made in the cell lines ZR-75-30 and the paired cell lines VP267/VP229 using high-throughput sequencing data were validated and showed that ZR-75-30 has at least seven expressed fusion genes and VP267 and VP229 have three expressed fusions. This agrees with the recent data from Stephens et al. (2009), which found between zero and eleven expressed fusion genes per tumour or cell line. Stephens et al. estimate they would have found 50% of the rearrangements present in their samples, which suggests that there are further fusion genes they have missed due to low coverage of the genome, and this is likely to be true of the cell lines I have investigated by sequencing – MCF7, which has been investigated by both genomic sequencing (Hampton et al., 2009) and paired-end transcriptome analysis (Ruan et al., 2007), has 13 known expressed transcripts (Table 7.1).

One possible explanation for why no further fusion genes were found in HCC1806 is that fusion genes are more likely to be found at balanced rather than unbalanced breakpoints. HCC1806 was first selected for array painting partly because the SKY karyotype suggested that it had balanced translocations, and the two known fusion genes in HCC1806, *RIF1-PKD1L1* and *TAX1BP1-AHCY*, both involve balanced chromosome rearrangements. No further fusion genes were found when I investigated the unbalanced rearrangements in this cell line. It is probable that balanced rearrangements, which do not cause a gain or loss of material, are selected because they have an effect on genes in other ways, and balanced rearrangements may produce fusion genes more often than unbalanced rearrangements.

Another question is whether the lack of fusion genes found in MDA-MB-134 was due to technical difficulties of finding fusion genes, or whether it is a true result. The majority of the rearrangements found in MDA-MB-134 were in the amplicon, which was as expected based on the low-coverage sequencing, which was unlikely to find all the breakpoints in non-amplified regions. All of the 24 validated structural variants found by paired-end sequencing in MDA-MB-134 were part of the amplicon, and none of the 9 rearrangements outside the amplicon suggested by copy number changes were detected by paired-end sequencing. This is a lower number of rearrangements detects than might be expected, which may be due to the repetitive nature of the breakpoint regions (see further discussion below). No structural variants were found to suggest that MDA-MB-134 contains any balanced rearrangements, which may be more likely to produce fusion genes. Additionally, the complex nature of the amplifications in MDA-MB-134 may make it less likely that rearrangements in the amplicon will produce fusion genes, as the two fusion partners may be further rearranged by a nearby breakpoint, or may be missing the nearby DNA sequences needed to form a stable transcript (Stephens et al., 2009).

A hypothesis put forward in Paterson et al. (2007) was that a fusion gene formed from co-amplification of chromosomes 8 and 11 might drive this amplification. No fusion gene has been found in MDA-MB-134 to support this hypothesis. It is more likely that the driver of amplification is co-expression of two genes in the amplicon, as suggested by Kwek et al. (2009).

7.2 How important are fusion genes in breast cancer?

In contrast to the 5 known fusion genes at the start of my project, there are now 64 known fusion genes across a range of breast cancer cell lines and tumours (Table 7.1). Although increasing numbers of fusion genes have been found, only one fusion is though to be recurrent in breast cancer. The *EML4-ALK* translocation first reported in non-small cell lung cancer (Soda et al., 2007; Rikova et al., 2007) has also been reported as present in 2.4% of breast tumours (Lin et al., 2009), although this contradicts an earlier study which found that the transcript is specific to NSCLC and not found in breast tumours (Fukuyoshi et al., 2008).

5' Gene	3' Gene	In-frame?	Cell line or tumour	Source
ZMYM4	*OPRD1*	No	ZR-75-30	This study
COL14A1	*SKAP1*	Yes	ZR-75-30	This study
APPBP2	*PHF10L1*	Yes	ZR-75-30	This study
TAOK1	*PCGF2*	Yes	ZR-75-30	This study
UPS32	*CCDC49*	No	ZR-75-30	This study
BCAS3	*HOXB9*	No	ZR-75-30	This study
TIAM1	*NRIP1*	Yes	ZR-75-30	This study
PDLIM1	*ZBBX*	No	VP267/VP229	This study
FAM125B	*SPTLC1*	No	VP267/VP229	This study
MDS1	*KCNMA1*	Yes	VP267/VP229	This study
EML4	*ALK*	Yes	5 breast tumours	Lin et al., 2009
PLXND1	*TMCC1*	Yes	HCC1187	Stephens et al., 2009
RGS22	*SYCP1NM*	Yes	HCC1187	Stephens et al., 2009
EFTUD2	*KIF18B*	Yes	HCC1395	Stephens et al., 2009
ERO1L	*FERMT2*	Yes	HCC1395	Stephens et al., 2009
PLA2R1	*RBMS1*	Yes	HCC1395	Stephens et al., 2009
CYTH1	*PRPSAP1*	Yes	HCC1599	Stephens et al., 2009
NFIA	*EHF*	Yes	HCC1937	Stephens et al., 2009
STRADB	*noP58*	Yes	HCC1954	Stephens et al., 2009
INTS4	*GAB2*	Yes	HCC2157	Stephens et al., 2009
RASA2	*ACPL2*	Yes	HCC2157	Stephens et al., 2009
SMYD3	*ZNF695*	Yes	HCC2157	Stephens et al., 2009
ACBD6	*RRP15*	Yes	HCC38	Stephens et al., 2009
LDHC	*SERGEF*	Yes	HCC38	Stephens et al., 2009
MBOAT2	*PRKCE*	Yes	HCC38	Stephens et al., 2009
SLC26A6	*PRKAR2A*	Yes	HCC38	Stephens et al., 2009
SMF	*PPARGC1B*	Yes	HCC38	Stephens et al., 2009
RAF1	*DAZL*	Yes	PD3664a	Stephens et al., 2009
AC141586.2	*CCNF*	Yes	PD3670a	Stephens et al., 2009
SEPT8	*AFF4*	Yes	PD3670a	Stephens et al., 2009
ETV6	*ITPR2*	Yes	PD3688a	Stephens et al., 2009
KCNQ5	*RIMS1*	Yes	HCC1395	Stephens et al., 2009
HN1	*USH1G*	Yes	PD3693a	Stephens et al., 2009
AGPAT5	*MCPH1*	No	HCC1187	Stephens et al., 2009
CTAGE5	*SIP1*	No	HCC1187	Stephens et al., 2009
PLXND1	*TMCC1*	No	HCC1187	Stephens et al., 2009
SUSD1	*ROD1*	No	HCC1187	Stephens et al., 2009
EIF3K	*CYP39A1*	No	HCC1395	Stephens et al., 2009
IL6R	*ATP8B2*	No	HCC2157	Stephens et al., 2009

RBM14	*PACS1*	No	HCC2157	Stephens et al., 2009
FBXL18	*RNF216*	No	PD3670a	Stephens et al., 2009
ITPR2	*ETV6*	No	PD3688a	Stephens et al., 2009
GRB7	*PERLD1*	No	HCC2218	Stephens et al., 2009
HDAC11	*FBLN2*	No	PD3670a	Stephens et al., 2009
FGFR1	*ZNF703*	No	PD3690a	Stephens et al., 2009
SSH2	*SUZ12*	No	PD3693a	Stephens et al., 2009
RIF1	*PKD1L1*	Yes	HCC1806	Stephens et al., 2009
TAX1BP1	*AHCY*	Yes	HCC1806	Stephens et al., 2009
FHIT	*MACROD2*	Yes	BrCa-MZ-02	Popovici et al., 2002
BCAS4	*BCAS3*	Yes	MCF7	Barlund et al., 2002
ARGHEF2	*SULF2*	Yes	MCF7	Hampton et al., 2009
DEPDC1B	*ELOVL7*	Yes	MCF7	Hampton et al., 2009
RAD51C	*ATXN7*	Yes	MCF7	Hampton et al., 2009
SULF2	*PRICKLE2*	Yes	MCF7	Hampton et al., 2009
NPEPPS	*USP32*	Yes	MCF7	Hampton et al., 2009
ASTN2	*PTPRG*	Yes	MCF7	Hampton et al., 2009
BCAS3	*RSBN1*	Yes	MCF7	Hampton et al., 2009
ASTN2	*TBC1D16*	Yes	MCF7	Hampton et al., 2009
BCAS4	*PRKCBP1*	Yes	MCF7	Hampton et al., 2009
cXorf15	*SYAP1*	Yes	MCF7	Ruan et al., 2007
RPS6KB1	*TMEM49*	Yes	MCF7	Ruan et al., 2007
BRCC3	*FUNDC2*	Yes	MCF7	Ruan et al., 2007
NRG1	*ODZ4*	Yes	MDA-MB-175	Liu et al., 1999

Table 7.1. Fusion genes currently known to be expressed in breast cancer cell lines and tumours.

BCAS3 is the only other gene known to be recurrently fused in breast cancer, as although a fusion could not be found in HCC1806 (Chapter 3), the sequencing of ZR-75-30 discovered a fusion of *BCAS3* to *HOXB9*. However, this did not produce an in-frame product, and involved the 5' end of the gene in the fusion as opposed to the 3' end retained in the known fusion in MCF7. Similarly, investigation of *ODZ4* as a potential recurrent target of fusion did not find it fused in any cell line other than the previously-known MDA-MB-175 fusion.

Fusion gene recurrence has been previously used to determine whether gene fusions were important, but the approaches which were used in haematological malignancies may not be the best way to find important events in solid tumours. Already there is evidence from prostate cancer that while there are important recurrent fusions such as *TMPRSS2-ERG*, the same genes are found fused to different fusion partners, and some of the variant fusions have so far been found in only one case (Tomlins et al., 2007). There may be no recurrent fusion genes in breast cancer, but as the number of known fusion genes grows, we may see fusions involving the same genes fused to different fusion partners, and fusion genes involving different members of the same pathway which have the same effect on the cell.

7.3 How can fusion genes in breast cancer be found?

Two different approaches have been taken in the search for fusion genes in solid tumours. One approach looks for a fusion transcript and then locate the genomic rearrangement which produces the fusion. This approach has successfully found fusion genes by several different methods: using expression array data to look for genes which are overexpressed and potentially fused (Tomlins et al., 2005; Lin et al., 2009); using retroviral expression libraries to find novel transforming genes (Soda et al., 2007); using proteomic approaches to look at fusion proteins directly (Rikova et al., 2007); and more recently by using RNA-seq to find fusion genes across the whole transcriptome (Maher et al., 2009), a technique which will also find fusion genes that are not the result of a genomic rearrangement, such as readthrough fusions between two adjacent genes (Berger et al., 2010).

The second approach searches for genomic rearrangements, and looks for the fusion genes which may result from the genomic rearrangement. This is the approach I have taken.. Over the course of this study, the technology I used to map chromosome rearrangements in breast cancer moved from low-resolution array painting, to high-resolution whole genome arrays, and finally to high-throughput sequencing. All of these different approaches were used in turn to map the rearrangements in MDA-MB-134.

Array painting is useful to determine which chromosome fragments are present in a derivative chromosome, and hence discover which breakpoints are joined together. Prior to high-throughput sequencing, this was the only way to determine which chromosome breakpoints were joined together, and while paired-end sequencing can also determine the two sides of a breakpoint, it cannot assemble the whole derivative chromosome. High-throughput sequencing is also unable to detect rearrangements near the centromeres and telomeres by finding the sequences crossing the breakpoint, as the sequences produced from these repetitive regions cannot be unambiguously aligned to a reference genome. The der(15)t(15;17) and der(16)t(16;18) chromosomes in MDA-MB-134 are an example of a rearrangement which would not be detected using high-throughput sequencing. However, the loss of material caused by the unbalanced translocation would still be seen from copy number plots taken from high-throughput sequencing, even if the exact breakpoint is not known. It is likely that a breakpoint in the repetitive regions near the centromeres and telomeres is not affecting genes directly, but that the loss of material is the important event.

Although the array painting in this study was carried out on low-resolution 1Mb arrays, individual breakpoints can be mapped using high-resolution custom Nimblegen arrays (Gribble et al., 2007; Howarth et al., 2008), and higher-resolution arrays such as the SNP6.0 array could be used for array painting as well as for whole-genome array CGH. As the SNP6.0 array has probes to detect genotype as well as copy number, the genotypes of different derivative chromosomes could be compared to give information about karyotype evolution, as chromosomes which evolved from the same parental copy could be identified.

High-throughput sequencing has the potential to replace array-based methods of mapping chromosome rearrangements. As well as paired-end sequencing to provide information on both sides of a breakpoint, the sequences can be used to give high-resolution copy number information. Array painting could also be replaced with paired-end high-throughput sequencing, as it is possible to sequence sorted chromosomes individually (Chen et al., 2010). However, as well as the biases caused by the GC and repeat content of the data, there may be other, less obvious biases in the sequencing data which could affect copy number that have not yet been discovered.

As noted above, a potential problem with the use of high-throughput sequencing is aligning sequences to repetitive regions. Although many of the copy number steps seen in MDA-MB-134 had structural variants associated with them, there were copy number changes which did not have any associated structural variants. One possible explanation for this is low coverage, but breakpoints at comparable copy number to the missing breaks were found with multiple supporting reads, and relaxing the criteria for calling a structural variant to require only one read across the breakpoint did not find any potential structural variants associated with these breakpoints. Another possibility is that the breakpoints are in a region which is highly repetitive and although there are paired-end reads which span the breakpoint, none of these read pairs had a unique mapping to the genome and were discarded.

7.4 Mechanisms of chromosome rearrangement in breast cancer

Analysis of the MDA-MB-134 amplicon supports the breakage-fusion-bridge cycle model of amplicon formation. Many of the junctions involved in the amplicon are inversions, which are a signature of breakage-fusion-bridge cycles of amplification. An alternative model is the translocation-excision-amplification model (Van Roy et al., 2006), which proposes that the amplified regions are first excised from chromosomes and amplified as double minute chromosomes, which then re-integrate into the genome (Storlazzi et al., 2010). This model requires a translocation between chromosomes 8 and 11 in MDA-MB-134, which does appear to have occurred, but in the translocation-exicision-amplification model the sequences which were excised from the translocated chromosome re-integrate at a different site than they were excised from (Corvi et al., 1994), and the amplicon in MDA-MB-134 appears to have been amplified in place.

The microhomology at the breakpoints in MDA-MB-134 suggests two different mechanisms are involved. No homology or very short regions of microhomology, with small insertions at the junction, suggest non-homologous end-joining as a mechanism for double-strand break repair (Hastings et al., 2009), and this is seen in the majority of the junctions in MDA-MB-134. The longer region of homology seen at one breakpoint suggests the alternative pathway of microhomology-mediated end-joining is also operative in MDA-MB-134 (McVey and Lee, 2008). The mechanisms that control which method of repair is used are not well known, but studies in urothelial cancer suggest that cancer cells may preferentially use the more error-prone non-homologous end-joining even when there is sufficient microhomology for microhomology-mediated end-joining to be used (Windhofer et al., 2008). Alternatively, microhomology-mediated end-joining may be a "back-up" mechanism used only when non-homologous end-joining is unavailable (Lieber, 2010).

The karyotype of HCC1806 shows a number of tandem duplications. This could be an example of the particular "mutator phenotype" suggested by Stephens et al. (2009), where an unknown mechanism probably related to DNA damage repair produces large number of tandem duplications. HCC1806 is consistent with the observation that these mutator phenotype tumours are ER and PR negative, and do not have BRCA1 or BRCA2 mutations.

7.5 Future Directions

7.5.1 *ODZ4*

The high-throughput sequencing of MDA-MB-134 predicted that *ODZ4*, which is known to be fused to *NRG1* in the breast cancer cell line MDA-MB-175, may be fused to *KLHL35* in MDA-MB-134. Although there is no expression of the predicted fusion MDA-MB-134, *ODZ4* is overexpressed relative to the normal breast cell line, and it is more highly expressed than might be expected based on dosage effects, as while it is part of the amplified region on chromosome 11 it is at the outer edge of the amplicon, and is not at high copy number. This suggests that there is an alternate mechanism driving the overexpression of this gene other than extra copies. Further studies could investigate the mechanism of *ODZ4* overexpression, which may be unrelated to its presence in an amplified region, but could potentially be due chromosome rearrangements placing *ODZ4* under the control of a different promoter, or near an amplified enhancer. Mutations in *ODZ4* have also been found in pancreatic cancer (Yachida et al., 2010), suggesting that *ODZ4* may contribute to tumorigenesis by other mechanisms which do not cause overexpression.

ODZ4 has not previously been suggested as important in 11q13 amplification, but my studies suggest that it may be upregulated by a combination of amplification and other mechanisms, and a study which looks for candidate genes using the minimal region of amplification would not look at *ODZ4*. There are a number of other criteria which have been suggested as important when looking for the important genes in amplification, such as correlation with clinical outcome and analysis of biological activity using siRNA knockdowns (Santarius et al., 2010), and further studies of *ODZ4* could investigate its importance by methods other than looking at gene amplification and expression.

7.5.2 High-throughput sequencing and bioinformatic analysis

Methods for sequence alignment and analysis are constantly being improved, and since the bioinformatic pipeline described in this study was developed, it has been updated with a newer generation of alignment and analysis programs. Longer read lengths cannot be aligned by MAQ, and the primary alignment is now performed using BWA (Li and Durbin, 2009) with a more sensitive re-alignment step using Novoalign (www.novocraft.com), and using Picard for duplicate calling and to assess the depth of the library (picard.sourceforge.net).

Wet-lab validation of the results of high-throughput sequencing is expensive and time-consuming, and it is important to improve the bioinformatic analysis of the sequencing to remove as many false positive and misleading results as possible before any downstream analysis is done.

One area for improvement is in the initial sequence alignment. Sequence misalignment is one possible cause of artefactual structural variants, and several of the structural variant calls in MDA-MB-134 proved to be due to sequence misalignment. As sequence coverage increases, the number of misaligned reads will increase, and depending on the probability of a sequence being misaligned, the number of artefactual structural variants called may increase faster than the number of real structural variants (Figure 7.1). A way to decrease the problems of misalignment is to use a two-stage alignment process, as has been implemented in the latest version of the bioinformatic pipeline described in Chapter 6. The initial alignment step uses a fast but less sensitive alignment program to align the large numbers of reads produced, such as BWA (Li and Durbin, 2009) or Bowtie (Langmead et al., 2009), which would miss the true alignments of a small number of reads. The possible aberrant reads called from this first-pass alignment would be realigned using a more sensitive algorithm, such as BFAST which is specifically designed for sensitive alignment of short reads (Homer et al., 2009). As the number of aberrant reads is a small percentage of the total, it is possible to use a much slower algorithm to re-align the aberrant reads which would be impractical to use to align the total set of reads.

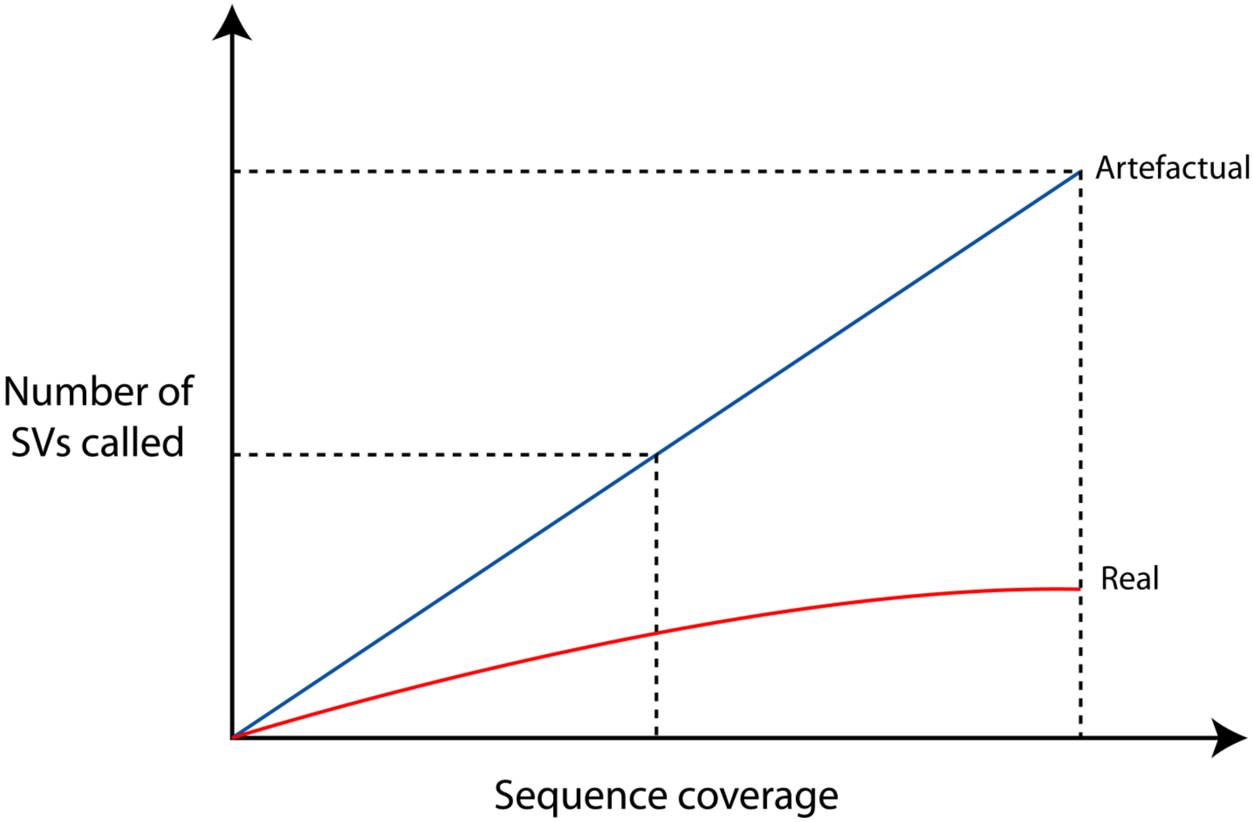


Figure 7.1. Graph to show how artefactual structural variants could increase at a greater rate than real structural variants.

There are also improvements that could be made to structural variant calling. The strategy I have used looks at each structural variant individually, and does not link variants together, such as two structural variants that are on either side of an insertion (Figure 7.2A). Linking structural variants together can help to identify regions of insertion or inversion, rather than just one side of an event, but if there is a further rearrangement which has not been detected then the linkage will be wrong, and so this method is more useful for small rearrangements where it is less likely that another variant between them has been missed (Medvedev et al., 2009).

Another improvement to structural variant calling is to use the sequences that span the breakpoint junction to identify the breakpoint to base pair resolution without PCR validation and sequencing (Figure 7.2B). A sequence that maps to two regions of the genome will not be aligned using current alignment algorithms, which do not look for split mapping reads because this would slow down the alignment. This strategy would not work with 37bp reads because they would produce too many possible alignments, but as sequence reads become longer they can be re-aligned to look for split mappings and find the breakpoint junctions. Split mapping of reads becomes more important as longer sequence reads are produced from small insert libraries, as more of each fragment is sequenced and there is more chance of a breakpoint falling within a read rather than in between the paired-end reads.

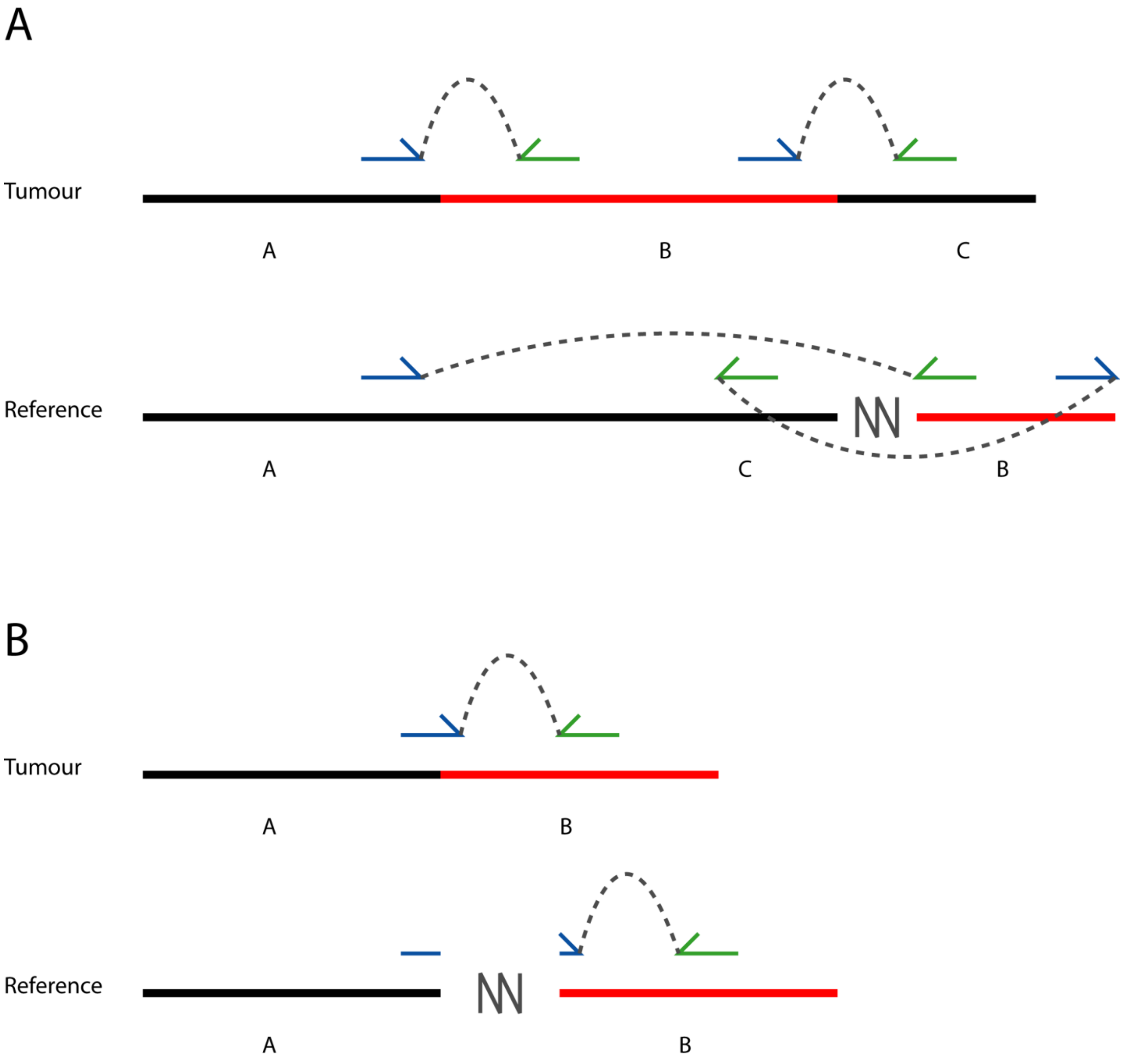


Figure 7.2. Improvements to structural variant calling. A –how two pairs of reads spanning the two sides of an insertion would align to the reference genome. B –how a read spanning a breakpoint would align to the reference genome in two locations. (Diagram adapted from Medvedev et al (2009)).

7.6 Conclusion

From the results in this study and others which have been published over the course of my project, it is clear that there are fusion genes in breast cancer. High-throughput genomic and transcriptome sequencing makes it easier to find fusion genes than by previous array and FISH based methods, and as the number of sequenced breast cancer genomes increases, the number of fusion genes found will increase as well. The importance of fusion genes in breast cancer has yet to be demonstrated, as recurrent gene fusions have yet to be found, and functional studies will be necessary to link fusion gene to carcinogenesis, but given the importance of fusion genes in other cancers it seems likely that fusion genes will also be important in breast cancer. As the number of known fusions increases, it will be easier to find pathways which are recurrently involved in fusion genes, and a combined analysis of fusion genes along with copy-number alterations, mutations and epigenetic modifications will provide an overall picture of the genes which are involved in breast cancer.

Alberti, L., Carniti, C., Miranda, C., Roccato, E., and Pierotti, M. A. (2003). RET and NTRK1 proto-oncogenes in human diseases. Journal of Cellular Physiology *195*, 168-186.

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W. (2003). Chromosome aberrations in solid tumors. Nat. Genet. *34*, 369-376.

Al-Kuraya, K., Schraml, P., Torhorst, J., Tapia, C., Zaharieva, B., Novotny, H., Spichtin, H., Maurer, R., Mirlacher, M., Köchli, O., et al. (2004). Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Res. *64*, 8534-8540.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

Andersen, C. L., Monni, O., Wagner, U., Kononen, J., Barlund, M., Bucher, C., Haas, P., Nocito, A., Bissig, H., Sauter, G., et al. (2002). High-throughput copy number analysis of 17q23 in 3520 tissue specimens by fluorescence in situ hybridization to tissue microarrays. Am. J. Pathol. *161*, 73-79.

Armitage, P., and Doll, R. (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Br. J. Cancer *11*, 161-169.

Armitage, P., and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. Br. J. Cancer *8*, 1-12.

ar-Rushdi, A., Nishikura, K., Erikson, J., Watt, R., Rovera, G., and Croce, C. M. (1983). Differential expression of the translocated and the untranslocated c-myc oncogene in Burkitt lymphoma. Science *222*, 390-393.

Arvand, A., and Denny, C. T. (2001). Biology of EWS/ETS fusions in Ewing's family tumors. Oncogene *20*, 5747-5754.

Aulmann, S., Schnabel, P. A., Helmchen, B., Dienemann, H., Drings, P., Otto, H. F., and Sinn, H. P. (2005). Immunohistochemical and cytogenetic characterization of acantholytic squamous cell carcinoma of the breast. Virchows Archiv *446*, 305-309.

Banerjee, S., Dowsett, M., Ashworth, A., and Martin, L. (2007). Mechanisms of Disease: angiogenesis and the management of breast cancer. Nat. Clin. Prac. Oncol. *4*, 536-550.

Bardelli, A., Cahill, D. P., Lederer, G., Speicher, M. R., Kinzler, K. W., Vogelstein, B., and Lengauer, C. (2001). Carcinogen-specific induction of genetic instability. Proc. Natl. Acad. Sci. U.S.A *98*, 5770 -5775.

Bärlund, M., Monni, O., Kononen, J., Cornelison, R., Torhorst, J., Sauter, G., Kallioniemi, O. P., and Kallioniemi, A. (2000). Multiple genes at 17q23 undergo amplification and overexpression in breast cancer. Cancer Res. *60*, 5340-5344.

Bärlund, M., Monni, O., Weaver, J. D., Kauraniemi, P., Sauter, G., Heiskanen, M., Kallioniemi, O. P., and Kallioniemi, A. (2002). Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes Cancer *35*, 311-317.

Bautista, S., and Theillet, C. (1998). CCND1 and FGFR1 coamplification results in the colocalization of 11q13 and 8p12 sequences in breast tumor nuclei. Genes Chromosomes Cancer *22*, 268-277.

Beaudet, A. L., and Belmont, J. W. (2008). Array-based DNA diagnostics: let the revolution begin. Annu. Rev. Med. *59*, 113-129.

Beckman, R. A., and Loeb, L. A. (2006). Efficiency of carcinogenesis with and without a mutator mutation. Proc. Natl. Acad. Sci. U.S.A 103, 14140 -14145.

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., and Nowak, M. A. (2007). Genetic progression and the waiting time to cancer. PLoS Comput. Biol. *3*, e225.

Benner, S. E., Wahl, G. M., and Von Hoff, D. D. (1991). Double minute chromosomes and homogeneously staining regions in tumors taken directly from patients versus in human tumor cell lines. Anticancer Drugs *2*, 11-25.

Berger, M. F., Levin, J. Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L. A., Robinson, J., Verhaak, R. G., Sougnez, C., et al. (2010). Integrative analysis of the melanoma transcriptome. Genome Res. *20*, 413-427.

Bernards, R., and Weinberg, R. A. (2002). A progression puzzle. Nature *418*, 823.

Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S., et al. (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. *17*, 1296-1303.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., et al. (2010). Signatures of mutation and selection in the cancer genome. Nature *463*, 893-898.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., et al. (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res. *14*, 287-295.

Bodmer, W. (2008). Genetic instability is not a requirement for tumor development. Cancer Res. *68*, 3558-3561.

Boehm, T., Foroni, L., Kaneko, Y., Perutz, M. F., and Rabbitts, T. H. (1991). The rhombotin family of cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations to human chromosomes 11p15 and 11p13. Proc. Natl. Acad. Sci. U. S. A. *88*, 4367-4371.

Bosco, E. E., and Knudsen, E. S. (2007). RB in breast cancer: at the crossroads of tumorigenesis and treatment. Cell Cycle *6*, 667-671.

Burdall, S. E., Hanby, A. M., Lansdown, M. R., and Speirs, V. (2003). Breast cancer cell lines: friend or foe? Breast Cancer Res. *5*, 89-95.

Cailleau, R., Olivé, M., and Cruciger, Q. V. (1978). Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. In Vitro *14*, 911-915.

Cailleau, R., Young, R., Olivé, M., and Reeves, W. J. (1974). Breast tumor cell lines from pleural effusions. J. Natl. Cancer Inst. *53*, 661-674.

Cairns, J. (2002). Somatic stem cells and the kinetics of mutagenesis and carcinogenesis. Proc. Natl. Acad. Sci. U. S. A. *99*, 10567-10570.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*, 722-729.

Carmeliet, P., and Jain, R. K. (2000). Angiogenesis in cancer and other diseases. Nature *407*, 249-257.

Chen, W., Ullmann, R., Langnick, C., Menzel, C., Wotschofsky, Z., Hu, H., Doring, A., Hu, Y., Kang, H., Tzschach, A., et al. (2010). Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. Eur. J. Hum. Genet. *18*, 539-543.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods *6*, 99-103.

Chin, S. F., Teschendorff, A. E., Marioni, J. C., Wang, Y., Barbosa-Morais, N. L., Thorne, N. P., Costa, J. L., Pinder, S. E., van de Wiel, M. A., Green, A. R., et al. (2007). High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol. *8*, R215.

Ciampi, R., Knauf, J. A., Kerler, R., Gandhi, M., Zhu, Z., Nikiforova, M. N., Rabes, H. M., Fagin, J. A., and Nikiforov, Y. E. (2005). Oncogenic AKAP9-BRAF fusion is a novel mechanism of MAPK pathway activation in thyroid cancer. J. Clin. Invest. *115*, 94-101.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., et al. (2010a). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704-712.

Corvi, R., Amler, L. C., Savelyeva, L., Gehring, M., and Schwab, M. (1994). MYCN is retained in single copy at chromosome 2 band p23-24 during amplification in human neuroblastoma cells. Proc. Natl. Acad. Sci. U.S.A. *91*, 5523-5527.

Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S., Reed, M. W. R., Pooley, K. A., Scollen, S., Baynes, C., Ponder, B. A. J., Chanock, S., et al. (2007). A common coding variant in CASP8 is associated with breast cancer risk. Nat. Genet. *39*, 352-358.

Cuny, M., Kramar, A., Courjal, F., Johannsdottir, V., Iacopetta, B., Fontaine, H., Grenier, J., Culine, S., and Theillet, C. (2000). Relating genotype and phenotype in breast cancer: an analysis of the prognostic significance of amplification at eight different genes or loci and of p53 mutations. Cancer Research *60*, 1077-1083.

Daley, G. Q., Van Etten, R. A., and Baltimore, D. (1990). Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science *247*, 824-830.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C., and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proc. Natl. Acad. Sci. U.S.A. *79*, 7824-7827.

Davidson, J. M., Gorringe, K. L., Chin, S., Orsetti, B., Besret, C., Courtay-Cahen, C., Roberts, I., Theillet, C., Caldas, C., and Edwards, P. A. W. (2000). Molecular cytogenetic analysis of breast cancer cell lines. Br. J. Cancer *83*, 1309-1317.

Deininger, M., Buchdunger, E., and Druker, B. J. (2005). The development of imatinib as a therapeutic agent for chronic myeloid leukemia. Blood *105*, 2640-2653.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999-1005.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105.

Dutrillaux, B., Gerbault-Seureau, M., Remvikos, Y., Zafrani, B., and Prieur, M. (1991). Breast cancer genetic evolution: I. Data from cytogenetics and DNA content. Breast Cancer Res. Treat. *19*, 245-255.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature *447*, 1087-1093.

Edwards, P. A. W. (2002). Metastasis: the role of chance in malignancy. Nature *419*, 559-560.

Eguchi, M., Eguchi-Ishimae, M., Tojo, A., Morishita, K., Suzuki, K., Sato, Y., Kudoh, S., Tanaka, K., Setoyama, M., Nagamura, F., et al. (1999). Fusion of ETV6 to neurotrophin-3 receptor TRKC in acute myeloid leukemia with t(12;15)(p13;q25). Blood *93*, 1355-1363.

Engel, L. W., Young, N. A., Tralka, T. S., Lippman, M. E., O'Brien, S. J., and Joyce, M. J. (1978). Establishment and characterization of three new continuous cell lines derived from human breast carcinomas. Cancer Res. *38*, 3352-3364.

Ersfeld, K. (2004). Fiber-FISH: fluorescence in situ hybridization on stretched DNA. Methods Mol. Biol *270*, 395-402.

Fearon, E. R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.

Fisher, E. R., Palekar, A. S., Gregorio, R. M., and Paulson, J. D. (1983). Mucoepidermoid and squamous cell carcinomas of breast with reference to squamous metaplasia and giant cell tumors. Am. J. Surg. Pathol. *7*, 15-27.

Flicek, P., and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. Nat. Methods *6*, S6-S12.

Forrest, W. F., and Cavet, G. (2007). Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers". Science *317*, 1500a.

Foster, J. S., Henley, D. C., Ahamed, S., and Wimalasena, J. (2001). Estrogens and cell-cycle regulation in breast cancer. Trends Endocrinol. Metab. *12*, 320-327.

Foulds, L. (1957). Tumor progression. Cancer Res. *17*, 355-356.

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B. M., Jain, A. N., et al. (2006). Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer *6*, 96.

Fukuyoshi, Y., Inoue, H., Kita, Y., Utsunomiya, T., Ishida, T., and Mori, M. (2008). EML4-ALK fusion transcript is not found in gastrointestinal and breast cancers. Br. J. Cancer *98*, 1536-1539.

Garcia, M. J., Pole, J. C. M., Chin, S., Teschendorff, A., Naderi, A., Ozdag, H., Vias, M., Kranjac, T., Subkhankulova, T., Paish, C., et al. (2005). A 1Mb minimal amplicon at 8p11-12 in breast cancer identifies new candidate oncogenes. Oncogene *24*, 5235-5245.

Gazdar, A. F., Kurvari, V., Virmani, A., Gollahon, L., Sakaguchi, M., Westerfield, M., Kodagoda, D., Stasny, V., Cunningham, H. T., Wistuba, I. I., et al. (1998). Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. Int. J. Cancer *78*, 766-774.

Gelsi-Boyer, V., Orsetti, B., Cervera, N., Finetti, P., Sircoulomb, F., Rouge, C., Lasorsa, L., Letessier, A., Ginestier, C., Monville, F., et al. (2005). Comprehensive profiling of 8p11-12 amplification in breast cancer. Mol. Cancer Res. *3*, 655-667.

Getz, G., Hofling, H., Mesirov, J. P., Golub, T. R., Meyerson, M., Tibshirani, R., and Lander, E. S. (2007). Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers". Science *317*, 1500b.

Git, A., Spiteri, I., Blenkiron, C., Dunning, M., Pole, J., Chin, S., Wang, Y., Smith, J., Livesey, F., and Caldas, C. (2008). PMC42, a breast progenitor cancer cell line, has normal-like mRNA and microRNA transcriptomes. Breast Cancer Res. *10*, R54.

Greenblatt, M. S., Chappuis, P. O., Bond, J. P., Hamel, N., and Foulkes, W. D. (2001). TP53 Mutations in Breast Cancer Associated with BRCA1 or BRCA2 Germ-line Mutations. Cancer Research *61*, 4092-4097.

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., et al. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostat. *11*, 164-175.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153-158.

Greshock, J., Nathanson, K., Martin, A. M., Zhang, L., Coukos, G., Weber, B. L., and Zaks, T. Z. (2007). Cancer cell lines as genetic models of their parent histology: analyses based on array comparative genomic hybridization. Cancer Res. *67*, 3594-3600.

Gribble, S. M., Kalaitzopoulos, D., Burford, D. C., Prigmore, E., Selzer, R. R., Ng, B. L., Matthews, N. S., Porter, K. M., Curley, R., Lindsay, S. J., et al. (2007). Ultra-high resolution array painting facilitates breakpoint sequencing. J. Med. Genet. *44*, 51-58.

Gruvberger, S., Ringnér, M., Chen, Y., Panavally, S., Saal, L. H., Borg, Å., Fernö, M., Peterson, C., and Meltzer, P. S. (2001). Estrogen receptor status in breast cCancer Is associated with remarkably distinct gene expression patterns. Cancer Res. *61*, 5979-5984.

Guan, X. Y., Meltzer, P. S., Dalton, W. S., and Trent, J. M. (1994). Identification of cryptic sites of DNA sequence amplification in human breast cancer by chromosome microdissection. Nat. Genet. *8*, 155-161.

Gudmundsdottir, K., and Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. Oncogene *25*, 5864-5874.

Gururaj, A. E., Holm, C., Landberg, G., and Kumar, R. (2006). Breast cancer-amplified sequence 3, a target of metastasis-associated protein 1, contributes to tamoxifen resistance in premenopausal patients with breast cancer. Cell Cycle *5*, 1407-1410.

Gururaj, A. E., Peng, S., Vadlamudi, R. K., and Kumar, R. (2007). Estrogen Induces Expression of BCAS3, a novel estrogen receptor-alpha coactivator, through proline-, glutamic acid-, and leucine-rich protein-1 (PELP1). Mol. Endocrinol. *21*, 1847-1860.

Hampton, O. A., Den Hollander, P., Miller, C. A., Delgado, D. A., Li, J., Coarfa, C., Harris, R. A., Richards, S., Scherer, S. E., Muzny, D. M., et al. (2009). A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res. *19*, 167-177.

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat. Rev. Genet. *10*, 551-564.

Haverty, P. M., Fridlyand, J., Li, L., Getz, G., Beroukhim, R., Lohr, S., Wu, T. D., Cavet, G., Zhang, Z., and Chant, J. (2008). High-resolution genomic and expression analyses of copy number alterations in breast tumors. Genes Chromosomes Cancer *47*, 530-542.

Heer, R., Douglas, D., Mathers, M. E., Robson, C. N., and Leung, H. Y. (2004). Fibroblast growth factor 17 is over-expressed in human prostate cancer. J. Pathol. *204*, 578-586.

Hermans, A., Heisterkamp, N., von Linden, M., van Baal, S., Meijer, D., van der Plas, D., Wiedemann, L. M., Groffen, J., Bootsma, D., and Grosveld, G. (1987). Unique fusion of bcr and c-abl genes in Philadelphia chromosome positive acute lymphoblastic leukemia. Cell *51*, 33-40.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K. E., Jones, L. P., Assefnia, S., Chandrasekharan, S., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. *8*, R76.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N. E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V., et al. (2006). Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res. *16*, 1465-1479.

Hiyama, E., Gollahon, L., Kataoka, T., Kuroi, K., Yokoyama, T., Gazdar, A. F., Hiyama, K., Piatyszek, M. A., and Shay, J. W. (1996). Telomerase activity in human breast tumors. J. Natl. Cancer Inst. *88*, 116-122.

Homer, N., Merriman, B., and Nelson, S. F. (2009). BFAST: An alignment tool for large scale genome resequencing. PLoS ONE *4*, e7767.

Howarth, K. D., Blood, K. A., Ng, B. L., Beavis, J. C., Chua, Y., Cooke, S. L., Raby, S., Ichimura, K., Collins, V. P., Carter, N. P., et al. (2008). Array painting reveals a high frequency of balanced translocations in breast cancer cell lines that break in cancer-relevant genes. Oncogene *27*, 3345-3359.

Ichimura, K., Mungall, A. J., Fiegler, H., Pearson, D. M., Dunham, I., Carter, N. P., Collins, V. P. (2006). Small regions of overlapping deletions on 6q26 in human astrocytic tumours identified using chromosome 6 tile path array-CGH. Oncogene *25*, 1261-1271.

Ince, T. A., Richardson, A. L., Bell, G. W., Saitoh, M., Godar, S., Karnoub, A. E., Iglehart, J. D., and Weinberg, R. A. (2007). Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. Cancer Cell *12*, 160-170.

Järvinen, T. A. H., and Liu, E. T. (2003). HER-2/neu and topoisomerase II alpha in breast cancer. Breast Cancer Res. Treat. *78*, 299-311.

Jones, D. T., Kocialkowski, S., Liu, L., Pearson, D. M., Bäcklund, L. M., Ichimura, K., and Collins, V. P. (2008). Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. Cancer Res. *68*, 8673-8677.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science *258*, 818-821.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., et al. (2003). Large-scale genotyping of complex DNA. Nat. Biotech. *21*, 1233-1237.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Res. *12*, 656-664.

Knezevich, S. R., McFadden, D. E., Tao, W., Lim, J. F., and Sorensen, P. H. (1998). A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma. Nat. Genet. *18*, 184-187.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U. S. A. *68*, 820-823.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639-1645.

Kumar, R., and Yarmand-Bagheri, R. (2001). The role of HER2 in angiogenesis. Semin. Oncol. *28*, 27-32.

Kuppers, R. (2005). Mechanisms of B-cell lymphoma pathogenesis. Nat. Rev. Cancer *5*, 251-262.

Kwek, S. S., Roy, R., Zhou, H., Climent, J., Martinez-Climent, J. A., Fridlyand, J., and Albertson, D. G. (2009). Co-amplified genes at 8p12 and 11q13 in breast tumors cooperate with two major pathways in oncogenesis. Oncogene *28*, 1892-1903.

Lafage, M., Pedeutour, F., Marchetto, S., Simonetti, J., Prosperi, M. T., Gaudray, P., and Birnbaum, D. (1992). Fusion and amplification of two originally non-syntenic chromosomal regions in a mammary carcinoma cell line. Genes Chromosomes Cancer *5*, 40-49.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics *21*, 3763-3770.

Lambros, M. B., Natrajan, R., Geyer, F. C., Lopez-Garcia, M. A., Dedes, K. J., Savage, K., Lacroix-Triki, M., Jones, R. L., Lord, C. J., Linardopoulos, S., et al. (2010). PPM1D gene amplification and overexpression in breast cancer: a qRT-PCR and chromogenic in situ hybridization study. Mod. Pathol. *23,* 1334-1345.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Leary, R. J., Lin, J. C., Cummins, J., Boca, S., Wood, L. D., Parsons, D. W., Jones, S., Sjöblom, T., Park, B., Parsons, R., et al. (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc. Natl. Acad. Sci. U.S.A. *105*, 16224 -16229.

Lee, J. A., Carvalho, C. M., and Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235-1247.

Lemieux, N., Apiou, F., Vogt, N., Malfoy, B., and Dutrillaux, B. (1996). Structural heterogeneity of hsr(11) in the MDA-MB-134 mammary carcinoma cell line. Cancer Genet. Cytogenet. *90*, 75-79.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1997). Genetic instability in colorectal cancers. Nature *386*, 623-627.

Letessier, A., Sircoulomb, F., Ginestier, C., Cervera, N., Monville, F., Gelsi-Boyer, V., Esterni, B., Geneix, J., Finetti, P., Zemmour, C., et al. (2006). Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers. BMC Cancer *6*, 245.

Levsky, J. M., and Singer, R. H. (2003). Fluorescence in situ hybridization: past, present and future. J. Cell. Sci. *116*, 2833-2838.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851-1858.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Lieber, M. R. (2010). NHEJ and its backup pathways in chromosomal translocations. Nat. Struct. Mol. Biol. *17*, 393-395.

Lin, E., Li, L., Guan, Y., Soriano, R., Rivers, C. S., Mohan, S., Pandita, A., Tang, J., and Modrusan, Z. (2009). Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. Mol. Cancer. Res. *7*, 1466-1476.

Liu, X., Baker, E., Eyre, H., Sutherland, G., and Zhou, M. (1999). γ -Heregulin: a fusion gene of DOC-4 and neuregulin-1 derived from a chromosome translocation. Oncogene *18*, 7110–7114.

Lutchman, M., Pack, S., Kim, A. C., Azim, A., Emmert-Buck, M., van Huffel, C., Zhuang, Z., and Chishti, A. H. (1999). Loss of heterozygosity on 8p in prostate cancer implicates a role for dematin in tumor progression. Cancer Genetics and Cytogenetics *115*, 65-69.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. Bioinformatics *18*, 440-445.

MacLeod, R. A., Dirks, W. G., Matsuo, Y., Kaufmann, M., Milch, H., and Drexler, H. G. (1999). Widespread intraspecies cross-contamination of human tumor cell lines arising at source. Int. J. Cancer *83*, 555-563.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., and Chinnaiyan, A. M. (2009). Transcriptome sequencing to detect gene fusions in cancer. Nature *458*, 97-101.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. *9*, 387-402.

Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T. D., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T., et al. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. *8*, R228.

McCallum, H. M., and Lowther, G. W. (1996). Long-term culture of primary breast cancer in defined medium. Breast Cancer Res. Treat *39*, 247-259.

McVey, M., and Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. *24*, 529-538.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat. Meth. *6*, S13-S20.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat. Genet. *31*, 55-59.

Miron, A., Varadi, M., Carrasco, D., Li, H., Luongo, L., Kim, H. J., Park, S. Y., Cho, E. Y., Lewis, G., Kehoe, S., et al. (2010). PIK3CA mutations in in situ and invasive breast carcinomas. Cancer Res.*70*, 5674-5678.

Mitelman, F., Johansson, B., and Mertens, F. (2004). Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat. Genet. *36*, 331-334.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer *7*, 233-245.

Mitelman, F., Mertens, F., and Johansson, B. (2005). Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. Genes Chromosomes Cancer *43*, 350-66.

Monni, O., Bärlund, M., Mousses, S., Kononen, J., Sauter, G., Heiskanen, M., Paavola, P., Avela, K., Chen, Y., Bittner, M. L., et al. (2001). Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc. Natl. Acad. Sci. U. S. A. *98*, 5711-5716.

Morris, J. S., Carter, N. P., Ferguson-Smith, M. A., and Edwards, P. A. (1997). Cytogenetic analysis of three breast carcinoma cell lines using reverse chromosome painting. Genes Chromosomes Cancer *20*, 120-139.

Morris, S., Kirstein, M., Valentine, M., Dittmer, K., Shapiro, D., Saltman, D., and Look, A. (1994). Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. Science *263*, 1281-1284.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J. P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515-27.

Ng, B. L., and Carter, N. P. (2006). Factors affecting flow karyotype resolution. Cytometry A *69*, 1028-1036.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160-1167.

Parssinen, J., Kuukasjarvi, T., Karhu, R., and Kallioniemi, A. (2007). High-level amplification at 17q23 leads to coordinated overexpression of multiple adjacent genes in breast cancer. Br. J. Cancer *96*, 1258-1264.

Paterson, A. L., Pole, J. C., Blood, K. A., Garcia, M. J., Cooke, S. L., Teschendorff, A. E., Wang, Y., Chin, S. F., Ylstra, B., Caldas, C., et al. (2007). Co-amplification of 8p12 and

11q13 in breast cancers is not the result of a single genomic event. Genes Chromosomes Cancer *46*, 427-439.

Persson, K., Pandis, N., Mertens, F., Borg, Å., Baldetorp, B., Killander, D., and Isola, J. (1999). Chromosomal aberrations in breast cancer: A comparison between cytogenetics and comparative genomic hybridization. Genes, Chromosomes and Cancer *25*, 115-122.

Pharoah, P. D., Day, N. E., and Caldas, C. (1999). Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. Br. J. Cancer *80*, 1968-1973.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. *20*, 207-211.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M., Ordonez, G. R., Bignell, G. R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M., Beare, D., Lau, K. W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184-190.

Pole, J. C. M., Courtay-Cahen, C., Garcia, M. J., Blood, K. A., Cooke, S. L., Alsop, A. E., Tse, D. M. L., Caldas, C., and Edwards, P. A. W. (2006). High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. Oncogene *25*, 5693-5706.

Popovici, C., Basset, C., Bertucci, F., Orsetti, B., Adelaide, J., Mozziconacci, M. J., Conte, N., Murati, A., Ginestier, C., Charafe-Jauffret, E., et al. (2002). Reciprocal translocations in breast tumor cell lines: cloning of a t(3;20) that targets the FHIT gene. Genes Chromosomes Cancer *35*, 204-218.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat. Methods *5*, 1005-10.

Raap, A. K. (1998). Advances in fluorescence in situ hybridization. Mutation Research *400*, 287-298.

Rae, J. M., Creighton, C. J., Meck, J. M., Haddad, B. R., and Johnson, M. D. (2006). MDA-MB-435 cells are derived from M14 Melanoma cells—a loss for breast cancer, but a boon for melanoma research. Breast Cancer Res. Treat. *104*, 13-19.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., et al. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat. Genet. *39*, 165-167.

Ray, M. E., Yang, Z. Q., Albertson, D., Kleer, C. G., Washburn, J. G., Macoska, J. A., and Ethier, S. P. (2004). Genomic and expression analysis of the 8p11–12 amplicon in human breast cancer cell lines. Cancer Res. *64*, 40-47.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444-454.

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., North, B., Jayatilake, H., Barfoot, R., Spanova, K., et al. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat. Genet. *38*, 873-875.

Rigaill, G., Hupe, P., Almeida, A., La Rosa, P., Meyniel, J., Decraene, C., and Barillot, E. (2008). ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics *24*, 768-774.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., et al. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell *131*, 1190-1203.

Roschke, A. V., Stover, K., Tonon, G., Schäffer, A. A., and Kirsch, I. R. (2002). Stable karyotypes in epithelial cancer cell lines despite high rates of ongoing structural and numerical chromosomal instability. Neoplasia *4*, 19-31.

Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., Cristofanilli, M., Anderson, K., Hess, K. R., Stec, J., Ayers, M., Wagner, P., et al. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical Cancer Res. *11*, 5678-5685.

Rowley, J. D. (2001). Chromosome translocations: dangerous liaisons revisited. Nat. Rev. Cancer *1*, 245-250.

Rozen, S., and Skaletsky, H. J. (2000). Primer3 on the WWW for general users and for biologist programmers. In Bioinformatics Methods and Protocols: Methods in Molecular Biology (Humana Press), pp. 365-386.

Ruan, Y., Ooi, H. S., Choo, S. W., Chiu, K. P., Zhao, X. D., Srinivasan, K. G., Yao, F., Choo, C. Y., Liu, J., Ariyaratne, P., et al. (2007). Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res. *17*, 828-838.

Samuels, Y., Diaz Jr., L. A., Schmidt-Kittler, O., Cummins, J. M., DeLong, L., Cheong, I., Rago, C., Huso, D. L., Lengauer, C., Kinzler, K. W., et al. (2005). Mutant PIK3CA promotes cell growth and invasion of human cancer cells. Cancer Cell *7*, 561-573.

Santarius, T., Shipley, J., Brewer, D., Stratton, M. R., and Cooper, C. S. (2010). A census of amplified and overexpressed human cancer genes. Nat. Rev. Cancer *10*, 59-64.

Schröck, E., Manoir, S. D., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M. A., Ning, Y., Ledbetter, D. H., Bar-Am, I., Soenksen, D., et al. (1996). Multicolor Spectral Karyotyping of Human Chromosomes. Science *273*, 494-497.

Schwab, M. (1999). Oncogene amplification in solid tumors. Semin. Cancer Biol. *9*, 319-325.

Schwab, M., Westermann, F., Hero, B., and Berthold, F. (2003). Neuroblastoma: biology and molecular and chromosomal pathology. Lancet Oncol. *4*, 472-480.

Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., et al. (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat. Genet. *38*, 1239-1241.

Sevignani, C., Calin, G. A., Cesari, R., Sarti, M., Ishii, H., Yendamuri, S., Vecchione, A., Trapasso, F., and Croce, C. M. (2003). Restoration of fragile histidine triad (FHIT) expression induces apoptosis and suppresses tumorigenicity in breast cancer cell lines. Cancer Research *63*, 1183-1187.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809-813.

Shtivelman, E., Lifshitz, B., Gale, R. P., and Canaani, E. (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature *315*, 550-554.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science *235*, 177-182.

Smeets, D. F. C. M. (2004). Historical prospective of human cytogenetics: from microscope to microarray. Clin. Biochem. *37*, 439-446.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561-566.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A. *98*, 10869-10874.

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., et al. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat. Genet. *39*, 865-869.

Stein, W. D., and Stein, A. D. (1990). Testing and characterizing the two-stage model of carcinogenesis for a wide range of human cancers. J. Theor. Biol. *145*, 95-122.

Stephens, P. J., McBride, D. J., Lin, M., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., et al. (2009a). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005-1010.

Stingl, J., and Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat. Rev. Cancer *7*, 791-799.

Storlazzi, C. T., Lonoce, A., Guastadisegni, M. C., Trombetta, D., D'Addabbo, P., Daniele, G., L'Abbate, A., Macchia, G., Surace, C., Kok, K., et al. (2010). Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. Genome Res *20*, 1198-1206.

Takeuchi, K., Choi, Y. L., Togashi, Y., Soda, M., Hatano, S., Inamura, K., Takada, S., Ueno, T., Yamashita, Y., Satoh, Y., et al. (2009). KIF5B-ALK, a novel fusion oncokinase identified by an immunohistochemistry-based diagnostic system for ALK-positive lung cancer. Clinical Cancer Research *15*, 3143-3149.

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., and Leder, P. (1982). Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc. Natl. Acad. Sci. U. S. A. *79*, 7837-7841.

Teixeira, M. R., Pandis, N., and Heim, S. (2002). Cytogenetic clues to breast carcinogenesis. Genes Chromosomes Cancer *33*, 1-16.

Teschendorff, A., and Caldas, C. (2009). The breast cancer somatic 'muta-ome': tackling the complexity. Breast Cancer Res. *11*, 301.

Theodorou, V., Kimm, M. A., Boer, M., Wessels, L., Theelen, W., Jonkers, J., and Hilkens, J. (2007). MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. Nat. Genet. *39,* 759-769.

Tognon, C., Knezevich, S. R., Huntsman, D., Roskelley, C. D., Melnyk, N., Mathers, J. A., Becker, L., Carneiro, F., MacPherson, N., Horsman, D., et al. (2002). Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell *2*, 367-376.

Tomlins, S. A., Laxman, B., Dhanasekaran, S. M., Helgeson, B. E., Cao, X., Morris, D. S., Menon, A., Jing, X., Cao, Q., Han, B., et al. (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature *448*, 595-599.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644-648.

Tomlins, S. A., Mehra, R., Rhodes, D. R., Smith, L. R., Roulston, D., Helgeson, B. E., Cao, X., Wei, J. T., Rubin, M. A., Shah, R. B., et al. (2006). TMPRSS2:ETV4 Gene Fusions Define a Third Molecular Subtype of Prostate Cancer. Cancer Res. *66*, 3396-3400.

Tremblay, M., Tremblay, C. S., Herblot, S., Aplan, P. D., Hébert, J., Perreault, C., and Hoang, T. (2010). Modeling T-cell acute lymphoblastic leukemia induced by the SCL and LMO1 oncogenes. Genes & Development *24*, 1093-1105.

Tucker, R. P., and Chiquet-Ehrismann, R. (2006). Teneurins: a conserved family of transmembrane proteins involved in intercellular signaling during development. Dev Biol *290*, 237-45.

Turnbull, C., and Rahman, N. (2008). Genetic predisposition to breast cancer: past, present, and future. Annu. Rev. Genomics Hum. Genet. *9*, 321-345.

Ugolini, F., Adélaïde, J., Charafe-Jauffret, E., Nguyen, C., Jacquemier, J., Jordan, B., Birnbaum, D., and Pébusque, M. J. (1999). Differential expression assay of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and Fibroblast Growth Factor Receptor 1 (FGFR1) as candidate breast cancer genes. Oncogene *18*, 1903-1910.

Van Prooijen-Knegt, A. C., Van Hoek, J. F., Bauman, J. G., Van Duijn, P., Wool, I. G., and Van der Ploeg, M. (1982). In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp. Cell Res. *141*, 397-407.

Van Roy, N., Vandesompele, J., Menten, B., Nilsson, H., De Smet, E., Rocchi, M., De Paepe, A., Pahlman, S., and Speleman, F. (2006). Translocation-excision-deletion-amplification mechanism leading to nonsyntenic coamplification of MYC and ATBF1. Genes Chromosomes Cancer *45*, 107-117.

Velculescu, V. E. (2008). Defining the blueprint of the cancer genome. Carcinogenesis *29*, 1087-1091.

Venkatraman, E. S., and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics *23*, 657-663.

Venkitaraman, A. R. (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell *108*, 171-182.

Vogelstein, B., and Kinzler, K. W. (1993). The multistep nature of cancer. Trends Genet. *9*, 138-141.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W. L., et al. (2003). End-sequence profiling: sequence-based analysis of aberrant genomes. Proc. Natl. Acad. Sci. U. S. A. *100*, 7696-701.

Vousden, K. H., and Lu, X. (2002). Live or let die: the cell's response to p53. Nat. Rev. Cancer *2*, 594-604.

Wang, T. C., Cardiff, R. D., Zukerberg, L., Lees, E., Arnold, A., and Schmidt, E. V. (1994). Mammary hyperplasia and carcinoma in MMTV-cyclin D1 transgenic mice. Nature *369*, 669-671.

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of Pathology *220*, 263-280.

Weigelt, B., Glas, A. M., Wessels, L. F. A., Witteveen, A. T., Peterse, J. L., and van't Veer, L. J. (2003). Gene expression profiles of primary breast tumors maintained in distant metastases. Proc. Natl. Acad. Sci. U. S. A. *100*, 15901-15905.

Willenbrock, H., and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics *21*, 4084-4091.

Williams, A., and Thomson, E. (2010). Effects of scanning sensitivity and multiple scan algorithms on microarray data quality. BMC Bioinformatics *11*, 127.

Windhofer, F., Krause, S., Hader, C., Schulz, W. A., and Florl, A. R. (2008). Distinctive differences in DNA double-strand break repair between normal urothelial and urothelial carcinoma cells. Mutat. Res. *638*, 56-65.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. *10*, R65.

Wistuba, I. I., Behrens, C., Milchgrub, S., Syed, S., Ahmadian, M., Virmani, A. K., Kurvari, V., Cunningham, T. H., Ashfaq, R., Minna, J. D., et al. (1998). Comparison of features of human breast cancer cell lines and their corresponding tumors. Clin. Cancer Res. *4*, 2931-2938.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108-1113.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114-1117.

Yamashita, H., Nishio, M., Toyama, T., Sugiura, H., Zhang, Z., Kobayashi, S., and Iwase, H. (2004). Coexistence of HER2 over-expression and p53 protein accumulation is a strong prognostic molecular marker in breast cancer. Breast Cancer Res. *6*, R24-30.

Zech, L., Haglund, U., Nilsson, K., and Klein, G. (1976). Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. Int. J. Cancer *17*, 47-56.

Appendix 1 –List of primers used for PCR and sequencing

Name	Sequence	Used for
BCAS3_56.164_fwd	AGATCGCTGGTAGCAAGGAA	BCAS3 junction fine mapping
BCAS3_56.164_rev	ACCAAGGTGGTAAGCAGCAT	BCAS3 junction fine mapping
BCAS3_56.165_fwd	GAGGTGGGAGAAATGCTTGA	BCAS3 junction fine mapping
BCAS3_56.165_rev	AGCCGAGCCATAAACTGAGA	BCAS3 junction fine mapping
BCAS3_56.166_fwd	CCCAGTAATGCGATCCTAGC	BCAS3 junction fine mapping
BCAS3_56.166_rev	CTGCTTGAGGCCAAGAGTTC	BCAS3 junction fine mapping
BCAS3_56.167_fwd	GCTCACTACAACCCCTGCTT	BCAS3 junction fine mapping
BCAS3_56.167_rev	AGAGGTGAGTGGATCGCTTG	BCAS3 junction fine mapping
BCAS3_56.168_fwd	TGTCTTTCAAGGGGTTGGTC	BCAS3 junction fine mapping
BCAS3_56.168_rev	AGCCTCAGCAAAAGAGCAAG	BCAS3 junction fine mapping
BCAS3_56.169_fwd	CTACAACCCTTGCCTCTTCG	BCAS3 junction fine mapping
BCAS3_56.169_rev	GAGGCCAACAAGCAGATCAC	BCAS3 junction fine mapping
chr7-155709k-fwd	CCTCTGCTTGCTGGTGTGTA	BCAS3 junction fine mapping
chr7-155709k-rev	CAGGTTGAGCACCACTGTGT	BCAS3 junction fine mapping
chr7-155710k-fwd	ACCCAAGCTCCCTTCTCTTC	BCAS3 junction fine mapping
chr7-155710k-rev	CTTCATGAAAGCATGCTGGA	BCAS3 junction fine mapping
chr7-155711k-fwd	AAGAGCCTGACACAGCCATT	BCAS3 junction fine mapping
chr7-155711k-rev	TCTGTTGTTGGTCAGCCTTG	BCAS3 junction fine mapping
chr7-155712k-fwd	TCTCCTTGTGAAGCGTGATG	BCAS3 junction fine mapping
chr7-155712k-rev	TCTGCTGGCTCAGTAGAGCA	BCAS3 junction fine mapping
chr7-155713k-fwd	AAGCCCCAGGACAAGAAAAT	BCAS3 junction fine mapping
chr7-155713k-rev	GTCAAGTCCGGGGGTAGATT	BCAS3 junction fine mapping
chr7-155714k-fwd	GCACCTTCTATGTGCCATCA	BCAS3 junction fine mapping
chr7-155714k-rev	TTTGAATCCCCAGTGTGTTG	BCAS3 junction fine mapping
chr7-155715k-fwd	TGCATTACACTGGAACGTGAA	BCAS3 junction fine mapping
chr7-155715k-rev	CGATGCCAACCCACTTATTA	BCAS3 junction fine mapping
chr7-cloning-fwd	GCAGAACACAAAATCACCACA	BCAS3 junction cloning and sequencing
chr7-cloning-rev	TCTCCTCAGGGATGTTAATGTATG	BCAS3 junction cloning and sequencing
chr17-cloning-fwd	TTTGCATTGTTGTGATAGGACAT	BCAS3 junction cloning and sequencing
chr17-cloning-rev	CTCCAGTGCATTTTGCCTTT	BCAS3 junction cloning and sequencing
BCAS3-ex23-fwd	GGATCCGGAACAGAACTTCA	BCAS3 real-time PCR
BCAS3-ex23-rev	TTGCTGGTACCTACGGGAAG	BCAS3 real-time PCR
BCAS3-ex1-fwd	ATTCCCCAAGAAGACCCAGT	BCAS3 real-time PCR
BCAS3-ex1-rev	TCCTGCAGAAAAGTCGAGGAG	BCAS3 real-time PCR
BCAS3-ex9-fwd	GGAGCCTGTGGAGACAACAT	BCAS3 real-time PCR
BCAS3-ex9-rev	CTGTGGATGGCAACATCATC	BCAS3 real-time PCR
USP31 rev	CAGAGCTAAGGTGCGAGAGC	HCC1806 small deletion fusions
ERN2 ex8 fwd	GACACAGCCACCCTCTTCTC	HCC1806 small deletion fusions
ERN2 ex8 rev	CTCGCTCTCCTGAGACTTGG	HCC1806 small deletion fusions

ERN2 ex19 fwd	CTTTATCGCCAGGCAAACAT	HCC1806 small deletion fusions
ERN2 ex19 rev	ACTCCTTCTCCAGCCAGTCA	HCC1806 small deletion fusions
ATAD5-ex6-fwd	AGCAGCTGATCCTGTCCCTA	HCC1806 small deletion fusions
ATAD5-ex6-rev	CAAATGCCACAAACAACACC	HCC1806 small deletion fusions
SUZ12-fwd	GAGGGGGTGGCAGTTACTC	HCC1806 small deletion fusions
SUZ12-rev	AGATCTGTGTTGGCTTCTCAAA	HCC1806 small deletion fusions
ATAD5-ex14-fwd	AGCACTTCCTCCCAAAACCT	HCC1806 small deletion fusions
ATAD5-ex14-rev	CAGCTCCAACGTCTTTGACA	HCC1806 small deletion fusions
PITPNB fwd	GGGCAGCTTTACTCTGTTGC	HCC1806 small deletion fusions
PITPNB rev	ATGCAGGCACTTTGCTCTTT	HCC1806 small deletion fusions
CHEK2 ex2 fwd	AACTCCAGCCAGTCCTCTCA	HCC1806 small deletion fusions
CHEK2 ex2 rev	TGTCCCTCCCAAACCAGTAG	HCC1806 small deletion fusions
CHEK2 ex10 fwd	CTGTTGGGACTGCTGGGTAT	HCC1806 small deletion fusions
CHEK2 ex10 rev	CGTAAAACGTGCCTTTGGAT	HCC1806 small deletion fusions
AFF3-a-fwd	AGAAGAGAGCTCCACGCTCA	HCC1806 tandem duplication fusions
AFF3-a-rev	GCTCCCGTTCCTTTTCTTTC	HCC1806 tandem duplication fusions
AFF3-b-fwd	TGAAGTCGTCTTCGGAAACC	HCC1806 tandem duplication fusions
AFF3-b-rev	ACTTTGCCAGGTGCTTGAAT	HCC1806 tandem duplication fusions
BC156887-a-fwd	GCAGAAGTGGGAGCCAAG	HCC1806 tandem duplication fusions
BC156887-a-rev	GTCCATGGTGGGAGGTGTC	HCC1806 tandem duplication fusions
BC156887-b-fwd	AGTTGGCAGCCATCAAAGTT	HCC1806 tandem duplication fusions
BC156887-b-rev	CAAGCCAGAGTTGGTCATCA	HCC1806 tandem duplication fusions
CATSPERB-a-fwd	CAGAGAAACGCTTTGCATGT	HCC1806 tandem duplication fusions
CATSPERB-a-rev	GGAGCAAGTCCTCCTGATGT	HCC1806 tandem duplication fusions
TC2N-b-fwd	CATTTGTGGTGCCCAAGTTT	HCC1806 tandem duplication fusions
TC2N-b-rev	AGCGTCGACTCAAATCAGGT	HCC1806 tandem duplication fusions
EPHB2-a-fwd	ATGCGGAAGAGGTGGATGTA	HCC1806 tandem duplication fusions
EPHB2-a-rev	TGTTGATGGGACAGTGGGTA	HCC1806 tandem duplication fusions
EPHB2-b-fwd	TGGTCTTCCTCATTGCTGTG	HCC1806 tandem duplication fusions
EPHB2-b-rev	CCGATCACCTGCTCAATTTT	HCC1806 tandem duplication fusions
FUSIP1-fwd	CACGTCTCTGTTCGTCAGGA	HCC1806 tandem duplication fusions
FUSIP1-rev	TCAGCATCACGAACATCCTC	HCC1806 tandem duplication fusions
MYOM3-a-fwd	CTGGGAGAGGACTGAGATCG	HCC1806 tandem duplication fusions
MYOM3-a-rev	AGCGAAGAGGGATCCAGAAC	HCC1806 tandem duplication fusions
MYOM3-b-fwd	GGACCCCAAAGACTCAGACA	HCC1806 tandem duplication fusions
MYOM3-b-rev	CCTGAGACGATGCAAGTCAA	HCC1806 tandem duplication fusions
PNRC2-fwd	ACTGGGTCCCTGTTTCCTTT	HCC1806 tandem duplication fusions
PNRC2-rev	CACAGTGCACACAACACGAG	HCC1806 tandem duplication fusions
LAMA2-a-fwd	CTCCTCCTTCTGCTGCTCTC	HCC1806 tandem duplication fusions
LAMA2-a-rev	TTGCTGATGTGCCTGTGACT	HCC1806 tandem duplication fusions
LAMA2-b-fwd	CCTGGAAACTGGATTTTGGA	HCC1806 tandem duplication fusions
LAMA2-b-rev	GCACTTGGTCTCCCATTGAT	HCC1806 tandem duplication fusions

ARHGAP18-a-fwd	CTCTCCAGTTCCCAGGGAGT	HCC1806 tandem duplication fusions
ARHGAP18-a-rev	CCTTTGCATGGCTGTTCC	HCC1806 tandem duplication fusions
ARHGAP18-b-fwd	CTGGAGATCCACAGGAAAGC	HCC1806 tandem duplication fusions
ARHGAP18-b-rev	TGCCTCGTCATCTCTTCCTT	HCC1806 tandem duplication fusions
HS6ST2-b-fwd	CCCGGTACTTGAGTGAGTGG	HCC1806 tandem duplication fusions
HS6ST2-b-rev	GCGGTTGTTGGCTAGATTGT	HCC1806 tandem duplication fusions
GPC3-a-fwd	GATGCTGCTCAGCTTGGACT	HCC1806 tandem duplication fusions
GPC3-a-rev	TCCATGTTCAATCGTGCTGT	HCC1806 tandem duplication fusions
SMURF2-a-fwd	CGAGACGAGAGGAGGGAAA	HCC1806 tandem duplication fusions
SMURF2-a-rev	GGATGTGCCGAGAGTCG	HCC1806 tandem duplication fusions
CCDC46-b-fwd	GCAGCTGGTAGAGCTTGGTC	HCC1806 tandem duplication fusions
CCDC46-b-rev	TTGGCCTTTTCCAGTGTCAT	HCC1806 tandem duplication fusions
c6orf105-a-fwd	TTGCACACCATTTTCCAAGA	HCC1806 tandem duplication fusions
c6orf105-a-rev	GTTGTTTTTGGCATGTGCAG	HCC1806 tandem duplication fusions
c6orf105-b-fwd	TTTTGGCATTCTGGATCCTC	HCC1806 tandem duplication fusions
c6orf105-b-rev	TACACCCAGGTACCCGTCTC	HCC1806 tandem duplication fusions
phactr1-a-fwd	GGCCAGGATCTCCTTTAACC	HCC1806 tandem duplication fusions
phactr1-a-rev	TCGCTTTTCTTCTTCCTCCA	HCC1806 tandem duplication fusions
phactr1-b-fwd	GGTCACCAAAGCAGGACCTA	HCC1806 tandem duplication fusions
phactr1-b-rev	GCCAGGGAGCTGGTGTATAA	HCC1806 tandem duplication fusions
IMMP2L-fwd	GGTCACATCTGGGTTGAAGG	HCC1806 large deletion fusion
IMMP2L-rev	TAAGCGCTCTGGAGGAAGAA	HCC1806 large deletion fusion
DOCK4-fwd	GGTGCTGAAGGCACAAGAAT	HCC1806 large deletion fusion
DOCK4-rev	CCAGACCCTTTGCTCTCTTG	HCC1806 large deletion fusion
c21.1	GGCAGCCGGTGAGGAGTTTGG	MDA-MB-134 genomic junction PCR and sequencing
c21.2	AGGCTGCCCCACAGAGACCC	MDA-MB-134 genomic junction PCR and sequencing
c26.1	GCTGGAGAGGCCGTTGTCTGG	MDA-MB-134 genomic junction PCR and sequencing
c26.2	CGACTGCAGTGAGGTCAGGCG	MDA-MB-134 genomic junction PCR and sequencing
c28.1	TGTGGCCATTCCCCTGCTGC	MDA-MB-134 genomic junction PCR and sequencing
c28.2	CCAAGGACAGCCCACGTGCC	MDA-MB-134 genomic junction PCR and sequencing
c35.1	GGCCCACTGTCCTTTAGCATGGC	MDA-MB-134 genomic junction PCR and sequencing
c35.2	CCTTTCTGTTCCCTTCCTCCTTCCC	MDA-MB-134 genomic junction PCR and sequencing
c8.1	AGAGAAGGGAAGTGGGTGTGGC	MDA-MB-134 genomic junction PCR and sequencing
c8.2	TTGCCTATGCTGTCTTTCTGTGACAA	MDA-MB-134 genomic junction PCR and sequencing
i228.1	AGGCCAGAGCCCAGGAGTGG	MDA-MB-134 genomic junction PCR and sequencing

i228.2	ACTGAGAGGGGGTGAACTGGGC	MDA-MB-134 genomic junction PCR and sequencing
i242.1	GAGAGCAGCCCCAGGGAGGG	MDA-MB-134 genomic junction PCR and sequencing
i242.2	GGCTTTACCATGTTGGCGTTGAATTGG	MDA-MB-134 genomic junction PCR and sequencing
nc1.1	TTTAACGCCTTTTGGTGTCC	MDA-MB-134 genomic junction PCR and sequencing
nc1.2	TGCTCCAGAGGTGTGAACAG	MDA-MB-134 genomic junction PCR and sequencing
nc2.1	GTGCTGACCTTCTGGTCCAT	MDA-MB-134 genomic junction PCR and sequencing
nc2.2	AGTCAGTCCATCCGGTGTTC	MDA-MB-134 genomic junction PCR and sequencing
nc3.1	TACCCTCTCAGGTGCTGTCC	MDA-MB-134 genomic junction PCR and sequencing
nc3.2	CAGACTACAGGGGCTGCAAT	MDA-MB-134 genomic junction PCR and sequencing
nc4.1	CCAAGTGCTCCTGTCCTCTC	MDA-MB-134 genomic junction PCR and sequencing
nc4.2	AATGGTTGACCAGGTTCTGC	MDA-MB-134 genomic junction PCR and sequencing
nc5.1	AGCGCCTGGTACACAAGAAT	MDA-MB-134 genomic junction PCR and sequencing
nc5.2	CACTCTTTGAATTGGCGTGA	MDA-MB-134 genomic junction PCR and sequencing
nc6.1	ACTCCCTGTTGTGGGAACAC	MDA-MB-134 genomic junction PCR and sequencing
nc6.2	GAAACCATCTGGTCCAGGAA	MDA-MB-134 genomic junction PCR and sequencing
nc7.1	TTTTAAGCCTGTCGGAAAAG	MDA-MB-134 genomic junction PCR and sequencing
nc7.2	TGGCCCTGAATACTTTTTGG	MDA-MB-134 genomic junction PCR and sequencing
ni1.1	TGGCCCTGAATACTTTTTGG	MDA-MB-134 genomic junction PCR and sequencing
ni1.2	TTTCTTTTGCCCCACTGTTC	MDA-MB-134 genomic junction PCR and sequencing
ni10.1	CTGGAGGTCTCTGCCAGTTC	MDA-MB-134 genomic junction PCR and sequencing
ni10.2	ACTGCTCCCTTCTTCCTTCC	MDA-MB-134 genomic junction PCR and sequencing
ni11.1	CCAGAGGCAGAGGACAGAAC	MDA-MB-134 genomic junction PCR and sequencing
ni11.2	AATAGGGGAATTGGGGTGAG	MDA-MB-134 genomic junction PCR and sequencing
ni12.1	GCCCAGCCAAAATAGATTCA	MDA-MB-134 genomic junction PCR and sequencing
ni12.2	CAGCTTGGACTCCCTGTGAT	MDA-MB-134 genomic junction PCR and sequencing
ni13.1	TTTTGGACACAGAGGGAAGG	MDA-MB-134 genomic junction PCR and

		sequencing
ni13.2	GAGTTTAGCGGCTCACACCT	MDA-MB-134 genomic junction PCR and sequencing
ni14.1	TTCAGCCATCTGGATTTTCC	MDA-MB-134 genomic junction PCR and sequencing
ni14.2	GGTTGCTTCCTGTGTTTGGT	MDA-MB-134 genomic junction PCR and sequencing
ni15.1	CACCACTGAGTCTGGAAGCA	MDA-MB-134 genomic junction PCR and sequencing
ni15.2	GTTTTGAAATGGGGGACCTC	MDA-MB-134 genomic junction PCR and sequencing
ni16.1	CACCTGTTCTCCCAAACGAT	MDA-MB-134 genomic junction PCR and sequencing
ni16.2	GGCAGAATGAAGTGGATTCAA	MDA-MB-134 genomic junction PCR and sequencing
ni17.1	GCCACACCAGAAGGTTGTTT	MDA-MB-134 genomic junction PCR and sequencing
ni17.2	CATCCACATCTGGAATGCTG	MDA-MB-134 genomic junction PCR and sequencing
ni18.1	TTCAGCGAGTAGGGCAGAGT	MDA-MB-134 genomic junction PCR and sequencing
ni18.1	TGTCTCCATCACCAGGAAAA	MDA-MB-134 genomic junction PCR and sequencing
ni19.1	CTTCTGCAGCTTTGGTCCAT	MDA-MB-134 genomic junction PCR and sequencing
ni19.2	GCTCCCTTCTCCATCCCTAC	MDA-MB-134 genomic junction PCR and sequencing
ni2.1	CATATTACTTTTGCTGAAGATTCTGA	MDA-MB-134 genomic junction PCR and sequencing
ni2.2	ACAACCACTGCAAACCATGA	MDA-MB-134 genomic junction PCR and sequencing
ni20.1	AGGGAGAGGAAAAGGGTCAG	MDA-MB-134 genomic junction PCR and sequencing
ni20.2	AACTCCCCACAAAGTTGCAC	MDA-MB-134 genomic junction PCR and sequencing
ni21.1	ACTTCAGCCCAGGAGTTCAA	MDA-MB-134 genomic junction PCR and sequencing
ni21.2	ACTCGCTTCCCGAAACACTA	MDA-MB-134 genomic junction PCR and sequencing
ni3.1	AGAGATGATCATGGGCAAGC	MDA-MB-134 genomic junction PCR and sequencing
ni3.2	TAGGCTGGCTTGGATTGC	MDA-MB-134 genomic junction PCR and sequencing
ni4.1	CTTCCTGTTTGGGAGTTGGA	MDA-MB-134 genomic junction PCR and sequencing
ni4.2	AGAGCCTGCATTTCTTGCAT	MDA-MB-134 genomic junction PCR and sequencing
ni5.1	TAAACAGACCCCACCCAGAG	MDA-MB-134 genomic junction PCR and sequencing
ni5.2	GCCATTTCCAGTTTCGATGT	MDA-MB-134 genomic junction PCR and sequencing

ni6.1	TAAGTGCAGTGGCTCACACC	MDA-MB-134 genomic junction PCR and sequencing
ni6.2	AGGAGTGGCATTCAATGGAG	MDA-MB-134 genomic junction PCR and sequencing
ni7.1	TGTGGCGAAGCTTAGAGGAT	MDA-MB-134 genomic junction PCR and sequencing
ni7.2	CAGAGAGGTCATGGTTGTGC	MDA-MB-134 genomic junction PCR and sequencing
ni8.1	GATGAGCAGAGGGGGTATCA	MDA-MB-134 genomic junction PCR and sequencing
ni8.2	ACTCAGCATACTGCCCCACT	MDA-MB-134 genomic junction PCR and sequencing
ni9.1	CCAGGCAGAATGAAGAAAGC	MDA-MB-134 genomic junction PCR and sequencing
ni9.2	AAGTGATCTGCCCACCTCAG	MDA-MB-134 genomic junction PCR and sequencing
c21nested1a	CAAACAGGGTAATCGGAGGA	MDA-MB-134 genomic junction PCR and sequencing
c21nested1b	TTTTCAACAGCGGAGTAGGC	MDA-MB-134 genomic junction PCR and sequencing
c21nested2a	CTTCCATCATGGTGATGTGC	MDA-MB-134 genomic junction PCR and sequencing
c21nested2b	TTGGCTGCTGAGTTTCTCCT	MDA-MB-134 genomic junction PCR and sequencing
c35nested1a	CCTTTAGCATGGCTTTCTGG	MDA-MB-134 genomic junction PCR and sequencing
c35nested1b	TGTCTGCAATGGGGACATTA	MDA-MB-134 genomic junction PCR and sequencing
c35nested2a	AATAATTGGCCATGCTCCTG	MDA-MB-134 genomic junction PCR and sequencing
c35nested2b	TTCCTCCTTCCCTTTTGGTT	MDA-MB-134 genomic junction PCR and sequencing
new-nc7-1	CGAGCCAGGTAAGGGATGT	MDA-MB-134 genomic junction PCR and sequencing
new-nc7-2	ACAGGGCTTTCCTGATCAAA	MDA-MB-134 genomic junction PCR and sequencing
new-ni1-1	CTGTGGTTGCCTGTCACCTA	MDA-MB-134 genomic junction PCR and sequencing
new-ni1-2	ACTGGGCTTTCCATTCACTG	MDA-MB-134 genomic junction PCR and sequencing
new-ni13-1	CAGCCTGTGCAAAACGAATA	MDA-MB-134 genomic junction PCR and sequencing
new-ni13-2	TTTGAGGCTGCAGTGAGCTA	MDA-MB-134 genomic junction PCR and sequencing
new-ni3-1	ACCATTAGTGTGGGCGAAAG	MDA-MB-134 genomic junction PCR and sequencing
new-ni3-2	GATTTCTTGGCTGGCTTGA	MDA-MB-134 genomic junction PCR and sequencing
new-ni5-1	TCTCCCACAAGCCTCTCACT	MDA-MB-134 genomic junction PCR and sequencing
new-ni5-2	TCCAGCAGTGACAACAGAGG	MDA-MB-134 genomic junction PCR and

		sequencing
UNC5D-fwd	CGAGAGCTCAGGTTTGAAGG	MDA-MB-134 fusions
UNC5D-rev	CTTCCCTTCCTTGTGGGTCT	MDA-MB-134 fusions
ARSG-fwd	GTCACCAGCACTGCCTTGTA	MDA-MB-134 fusions
ARSG-rev	AGGCCTTGTAACGCTCCAG	MDA-MB-134 fusions
EFR3A-fwd	ATCCAAAAGATGGCCTTGTG	MDA-MB-134 fusions
EFR3A-rev	CAGTGCCTCCATAGCAATCA	MDA-MB-134 fusions
ANK1-fwd	TGCTGCTACCAGCTTTCTGA	MDA-MB-134 fusions
ANK1-rev	TGTTCCCCTTCTTGGTTGTC	MDA-MB-134 fusions
SHANK2-fwd	AGACCATTGGGAGCTACGTG	MDA-MB-134 fusions
SHANK2-rev	GTACTCGAAGGCCGAGAGTG	MDA-MB-134 fusions
FBXL11-fwd	CCGGATCCAGACTTCACTGT	MDA-MB-134 fusions
FBXL11-rev	GGCTAAACTCGAGGCTGATG	MDA-MB-134 fusions
ANO1-fwd	GGCTCTGGTGCACTATGTGA	MDA-MB-134 internal rearrangements
ANO1-rev	GTACTCGACGCAGTTGCTGA	MDA-MB-134 internal rearrangements
OSBPL5-fwd	TCGGAGAGAGAGAACCCTGA	MDA-MB-134 internal rearrangements
OSBPL5-rev	CAGCTTCATGCGGCTGTA	MDA-MB-134 internal rearrangements
OVCH2-fwd	GAAGCTGCACTTCCCAGAAA	MDA-MB-134 internal rearrangements
OVCH2-rev	CCTTGTCACTGTAGTTTTCAGGAT	MDA-MB-134 internal rearrangements
CCDC67-fwd	GCCTAAAGGCTCAATTTTCCA	MDA-MB-134 internal rearrangements
CCDC67-rev	CCATTTAGTTGAGTTTGGTAACTTTGT	MDA-MB-134 internal rearrangements
P2RX2-fwd	CCCAAATTCCACTTCTCCAA	MDA-MB-134 internal rearrangements
P2RX2-rev	GGTGGTGCCATTGATCTTGT	MDA-MB-134 internal rearrangements
POLG-fwd	CAGCACCTTCCTGGACACC	MDA-MB-134 internal rearrangements
POLG-rev	CTGTTCGAGACAGTGCTTCCT	MDA-MB-134 internal rearrangements
CHD2-fwd	GCTCTTGCCAAAGGAACAAG	MDA-MB-134 internal rearrangements
CHD2-rev	TTCGGATTTCTCCCTTGATG	MDA-MB-134 internal rearrangements
c16orf28-fwd	GGCCATGATTGAGAAGATCC	MDA-MB-134 internal rearrangements
c16orf28-rev	GCATGTCGTTGCATTTTGTA	MDA-MB-134 internal rearrangements
STAT3-fwd	TCAGGATGTCCGGAAGAGAG	MDA-MB-134 internal rearrangements
STAT3-rev	CGTACTCCATCGCTGACAAA	MDA-MB-134 internal rearrangements
SYT3-fwd	GGTGGTGCTGGACAACCT	MDA-MB-134 internal rearrangements
SYT3-rev	CCGATCACCTCGTTGTGC	MDA-MB-134 internal rearrangements
SFTPB-fwd	CTAGGGCATTGCCTACAGGA	MDA-MB-134 internal rearrangements
SFTPB-rev	GCATACAGATGCCGTTTGAG	MDA-MB-134 internal rearrangements
ZNF512B-fwd	GTGTCCAAACTCAGGGTGCT	MDA-MB-134 internal rearrangements
ZNF512B-rev	GTGTCTGCACTCAGCTGGAA	MDA-MB-134 internal rearrangements
SERAC1-fwd	TCCTTCAGCAACAGTGGAAA	MDA-MB-134 internal rearrangements
SERAC1-rev	CATATTTCCAATGACACGCATTA	MDA-MB-134 internal rearrangements
PTPRN2-fwd	ACATGGAGGACCACCTGAAG	MDA-MB-134 internal rearrangements
PTPRN2-rev	GTCAGCATGACGATCACCAC	MDA-MB-134 internal rearrangements
EPB49-fwd	CCTCCCGAGATTCCAGTGT	MDA-MB-134 readthrough fusions

EPB49-rev	CGTGTTCCAGGAGGGAATAA	MDA-MB-134 readthrough fusions
SHANK2-fwd	TTGAGGAGAAGACGGTGGTC	MDA-MB-134 readthrough fusions
SHANK2-rev	GAAGTCCCCGGTCCTTAGTC	MDA-MB-134 readthrough fusions
POLD3-fwd	CAAATGGCTGAGCTATACACTAGG	MDA-MB-134 readthrough fusions
POLD3-rev	AACCTTGTGGCAGGAATGTC	MDA-MB-134 readthrough fusions
KCNU1-fwd	GCCATGTAAGAAGCCTCCAC	MDA-MB-134 readthrough fusions
KCNU1-rev	ACAGCTTCCAACAGGGTCAG	MDA-MB-134 readthrough fusions
ADAM2-fwd	CCAGTTGATTGGATTGACGA	MDA-MB-134 readthrough fusions
ADAM2-rev	CTTGAAAGGTTGCACCAACA	MDA-MB-134 readthrough fusions
LRRC32-fwd	GCTGCACAACACCAAGACAA	MDA-MB-134 readthrough fusions
LRRC32-rev	GCTGATCTCATTGGTGCTCA	MDA-MB-134 readthrough fusions
ACER3-fwd	TCAGCCAGCTCTGCTCTGAT	MDA-MB-134 readthrough fusions
ACER3-rev	GGCACAACCATGACCTCTCT	MDA-MB-134 readthrough fusions
NADSYN1-fwd	GCTACGGATGTTGGGATCAT	MDA-MB-134 readthrough fusions
NADSYN1-rev	GCAGGATCTTCCTGTTGAGG	MDA-MB-134 readthrough fusions
IMMP2L-fwd	AGTCACAAGGGTGGGTGAAA	MDA-MB-134 readthrough fusions
IMMP2L-rev	TGGTTCAAAAGCACCACATC	MDA-MB-134 readthrough fusions
PRDM4-fwd	TACCCTCACCTGGAGAGCAG	MDA-MB-134 readthrough fusions
PRDM4-rev	ATGAAGACTTTGGGCACCAT	MDA-MB-134 readthrough fusions
GGCT-fwd	GGAGGGATAGCCACCATTTT	MDA-MB-134 readthrough fusions
GGCT-rev	ATGGGGGAGCACTTTCGTA	MDA-MB-134 readthrough fusions
NOD1-fwd	GCGGCGATTACAGAAAACAT	MDA-MB-134 readthrough fusions
NOD1-rev	CTCTCAGCAGAAGGGCAATC	MDA-MB-134 readthrough fusions
KLHL35-fwd	TACGACCCCTTCTCCAACAC	MDA-MB-134 readthrough fusions
KLHL35-rev	GATGGTGTCCTCAAGGGAGA	MDA-MB-134 readthrough fusions
AQP11-fwd	AGCTTTGGCACTTTCGCTAC	MDA-MB-134 readthrough fusions
AQP11-rev	CCGGTGTTTTCCATATGAGG	MDA-MB-134 readthrough fusions
PAK1-fwd	AGTTACCACCTCCTGCCTCA	MDA-MB-134 readthrough fusions
PAK1-rev	CGAGCTACCGCTTCACTTTC	MDA-MB-134 readthrough fusions
SERPINH1-fwd	AGCAGCAAGCAGCACTACAA	MDA-MB-134 readthrough fusions
SERPINH1-rev	AGGACCGAGTCACCATGAAG	MDA-MB-134 readthrough fusions
ANK1-fwd	GAGCTGCAGTTCAGTGTGGA	MDA-MB-134 readthrough fusions
ANK1-rev	TCCAGCATGTTCACGATCTC	MDA-MB-134 readthrough fusions
AP3M2-fwd	AGAAAATGTGCCTCCGGTTA	MDA-MB-134 readthrough fusions
AP3M2-rev	TGGAAAACCATTGTCAAGCA	MDA-MB-134 readthrough fusions
ODZ4-ex1-fwd	CGCCGAGAGAGAGAGGAG	ODZ4 exons, real time
ODZ4-ex1-rev	ATCTCTCAGACACTTGGTCGG	ODZ4 exons, real time
ODZ4-ex4-fwd	GACTGAAATTACCATGTGTCCAAA	ODZ4 exons, real time
ODZ4-ex4-rev	AAGGAGAGGAAGCCTTACCG	ODZ4 exons, real time
ODZ4-ex6-fwd	CACCGAGCATGAAAACACTG	ODZ4 exons, real time
ODZ4-ex6-rev	TCCATTAACTCCCTGAACCG	ODZ4 exons, real time
ODZ4-ex8-fwd	GACATTGCAGGACAACCTCA	ODZ4 exons, real time

ODZ4-ex8-rev	ATCACCAGGGTACCCACTGA	ODZ4 exons, real time
ODZ4-ex11-fwd	GCCTCCCTCCTTCACATACA	ODZ4 exons, real time
ODZ4-ex11-rev	TTTGGATTCAGGAATCTGGC	ODZ4 exons, real time
ODZ4-ex18-fwd	AGGAGCTGGCTGTGACACTT	ODZ4 exons, real time
ODZ4-ex18-rev	TGATGTCCAGAGGGTTAGGG	ODZ4 exons, real time
ODZ4-ex26-fwd	GTGGTGGTGAAGGACCTTGT	ODZ4 exons, real time
ODZ4-ex26-rev	GTGGACAAGTTTGGGCTGAT	ODZ4 exons, real time
ODZ4-ex29-fwd	GAGACCTCCAGCAAGGATGA	ODZ4 exons, real time
ODZ4-ex29-rev	AACAGCTACTACATCGGGGC	ODZ4 exons, real time
ODZ4-ex21-fwd	AGCCCCAGACCTGTCCTATT	ODZ4 exons, real time
ODZ4-ex21-rev	TTGGACGCGTCAATTTCATA	ODZ4 exons, real time
GAPDH_RT_fwd	GCAAATTCCATGGCACCGT	GAPDH, real-time control
GAPDH_RT_rev	TCGCCCCACTTGATTTTGG	GAPDH, real-time control

Appendix 2 - BACs

BAC clones used for FISH. All positions are from the hg18/GrCH36 build of the human genome.

Name	Chromosome	Start position (bp)	End position (bp)
RP11-65L3	2	178,966,383	179,139,203
RP11-67G7	2	182,686,311	182,851,083
RP11-59L22	2	192,914,919	193,076,744
RP11-15J24	2	205,174,891	205,347,264
RP4-781A18	7	27,976,454	28,166,805
RP11-563O5	7	114,005,319	114,175,715
RP11-518I12	7	157,549,662	157,756,716
RP11-381A5	17	55,638,614	55,853,525
RP11-947H19	17	55,827,724	56,009,281
RP11-105G8	17	55,904,414	56,060,400
RP11-160D4	17	56,857,614	57,015,721
RP11-466D9	17	56,954,223	57,129,790
RP11-180G7	17	57,115,715	57,267,805

Appendix 3 – Manufacturers and suppliers

Reagent	Manufacturer/Supplier
anti-BCAS3 antibody	Gift from Dr Jason Carroll, CRUK Cambridge Research Institute, Cambridge, UK
BACs	Wellcome Trust Sanger Institute, UK/Invitrogen, Paisley, UK
BioPrime labelling kit	Invitrogen, Paisley, UK
Biotin dUTP	Roche Diagnostics, Basel, Switzerland
Biotinylated anti-streptavidin	Vector Laboratories Inc., Burlingame, CA, USA
Chloramphenicol	Sigma-Aldrich, Dorset, UK
Colcemid	Sigma-Aldrich, Dorset, UK
Complete Protease Inhibitor Cocktail	Roche Diagnostics, Basel, Switzerland
Cryotubes	Fisher Scientific, Loughborough, UK
Cy3-labelled dCTP	Amersham, Epsom, UK
Cy5-labelled dCTP	Amersham, Epsom, UK
Cy5-labelled streptavidin	Amersham, Epsom, UK
DAPI in Vectashield	Vector Laboratories Inc., Burlingame, CA, USA
Denhardt's Solution	Sigma-Aldrich, Dorset, UK
Dextran sulphate	Sigma-Aldrich, Dorset, UK
Digoxygenin-11 dUTP	Roche Diagnostics, Basel, Switzerland
DMEM-F12	GIBCO Technologies, Invitrogen, Paisley, UK
DMSO	Invitrogen, Paisley, UK
DNA polymerase I	Sigma-Aldrich, Dorset, UK
DNA-free Kit	Ambion, Applied Biosystems, Foster City, USA
DNAse I	Sigma-Aldrich, Dorset, UK
DNAzol reagent	Invitrogen, Paisley, UK
dNTPs	Invitrogen, Paisley, UK
ECL Plus Western Blotting Detection System	GE Healthcare, Buckinghamshire, UK
Elongase polymerase mix	Invitrogen, Paisley, UK
Eppendorf tubes	Starlab, Milton Keynes, UK
Ethanol	Sigma-Aldrich, Dorset, UK
Falcon tubes	Bibby Sterilin, Stone, UK
FBS	Sigma-Aldrich, Dorset, UK
FITC-labelled anti-digoxygenin	Roche Diagnostics, Basel, Switzerland
Formamide	VWR International, Lutterworth, UK
G50 MicroSpin columns	GE Healthcare, Buckinghamshire, UK
GenomiPhi Kit	GE Healthcare, Buckinghamshire, UK
HiSpeed Plasmid Midi-Prep Kit	Qiagen UK, Crawley, UK
HotMaster Taq	VWR International, Lutterworth, UK
Hyperladder I	Bioline, London, UK
Isopropanol	Invitrogen, Paisley, UK

ITS	Sigma-Aldrich, Dorset, UK
Kanamycin	Sigma-Aldrich, Dorset, UK
LB agar	Hutchison/MRC Centre Media Unit
LB broth	Hutchison/MRC Centre Media Unit
Mate-Pair Library Prep Kit	Illumina, San Diego, CA, USA
MCBD-201	GIBCO Technologies, Invitrogen, Paisley, UK
NaH2PO4	VWR International, Lutterworth, UK
NaHPO4	VWR International, Lutterworth, UK
NanoDrop spectrophotometer	Labtech International, Ringmer, UK
Paired-End DNA Sample Prep Kit	Illumina, San Diego, CA, USA
PBS	Hutchison/MRC Centre Media Unit
Pellet Paint	Merck KGaA, Darmstadt, Germany
Penicillin/streptomycin	GIBCO Technologies, Invitrogen, Paisley, UK
Pipette tips	Starlab, Milton Keynes, UK
QIAquick PCR Purification Kit	Qiagen UK, Crawley, UK
RIPA buffer	Hutchison/MRC Centre Media Unit
RNAseIN	Promega, Fitchburg, USA
RPMI-1640	GIBCO Technologies, Invitrogen, Paisley, UK
Rubber cement	Heffers Art and Graphics Shop, Cambridge, UK
Sodium acetate	Hutchison/MRC Centre Media Unit
Spectrum Orange dUTP	Vysis UK Ltd/Abbott Laboratories, Downers Grove IL, USA
SSC	Hutchison/MRC Centre Media Unit
SuperScript III First-Strand Synthesis Kit	Invitrogen, Paisley, UK
SYBR Green PCR Master Mix	Applied Biosystems, Foster City, USA
TE	Hutchison/MRC Centre Media Unit
Tissue microarrays	Dr Suet-Feung Chin, CRUK Cambridge Research Institute, Cambridge, UK
TOPO XL PCR Cloning Kit	Invitrogen, Paisley, UK
Tris-acetate pre-cast gel	Invitrogen, Paisley, UK
Trizol reagent	Invitrogen, Paisley, UK
Trypsin	GIBCO Technologies, Invitrogen, Paisley, UK
Tween 20	QbioGene, Livingston, Scotland
Versene	Hutchison/MRC Centre Media Unit
Yeast tRNA	Invitrogen, Paisley, UK

Appendix 4 – Bioinformatic pipeline scripts

A – Perl script to predict fusion genes

```
# Fusion Gene Prediction
# Takes the structural variant calls from sequencing and predicts possible
fusion genes, readthroughs, and internal gene rearrangements
# Uses the Ensembl API -
http://www.ensembl.org/info/docs/api/api_installation.html for installation
and necessary modules
# Liz Batty emb51@cam.ac.uk
# Last modified August 2010

use warnings;
use strict;
use Bio::EnsEMBL::Registry;
use Getopt::Long;

# set default values
my $matepair = 0;
my $strands = 0;
my $linecounter = 0;
my $columns = 0;
my $cnv = 0;
my $help = 0;
my $input = 'library.lanes.sv_calls.txt';
my $insertsize = 470;

#uncomment one of these to use either hg18 (may2009) or hg19 website in
hyperlinks
#my $ensembl_site = "may2009.archive.ensembl.org";
my $ensembl_site = "www.ensembl.org";


my( $type_of_sv,
$support_for_sv,
$node1_chr,
$node1_start,
$node1_end,
$node1_strand,
$extra_support_1,
$node2_chr,
$node2_start,
$node2_end,
$node2_strand,
$extra_support_2,
$node1_cnv,
$node2_cnv );

my $result = GetOptions ("matepair|m" => \$matepair,
                                    "strands|s" => \$strands,
                                    "insertsize|i=i" => \$insertsize,
                                    "columns|c" => \$columns,
                                    "cnv|n" => \$cnv,
                                    "help|h" => \$help);
if ($help) {
      print "Options are:\n--matepair\tUse for mate pair libraries where RF
is a normal read
--strands\tThe file uses -1 and 1 for strand directions
--insertsize\tUse to set the insert size of the library
--columns\tThe file has extra supporting read columns (see later version of
Kevin's script)
--cnv\tThe file has been checked for CNVs\n";
                  exit;
                  }

if (@ARGV) {
      $input = $ARGV[0];
            }

open (INPUT, $input)          || die print "failed to open input file $!";

# print correct header line to the output file
{
if ($columns == 0 && $cnv == 0) {
      print "Type of SV\tSV Support\tNode 1 chr\tNode 1 start\tNode 1
end\tNode 1 direction\tNode 2 chr\tNode 2 start\tNode 2 end\tNode 2
direction\tGene at node 1\tGene at node 2\tType of fusion\tDetails\n";
}

elsif ($columns == 1 && $cnv == 0) {
      print "Type of SV\tSV Support\tNode 1 chr\tNode 1 start\tNode 1
end\tNode 1 direction\tExtra support\tNode 2 chr\tNode 2 start\tNode 2
end\tNode 2 direction\tExtra support\tGene at node 1\tGene at node 2\tType of
fusion\tDetails\n";
      }

elsif ($columns == 0 && $cnv == 1) {
      print "Type of SV\tSV Support\tNode 1 chr\tNode 1 start\tNode 1
end\tNode 1 direction\tNode 2 chr\tNode 2 start\tNode 2 end\tNode 2
direction\tNode 1 CNVs\tNode 2 CNVs\tGene at node 1\tGene at node 2\tType of
fusion\tDetails\n";
      }

elsif ($columns == 1 && $cnv == 1) {
      print "Type of SV\tSV Support\tNode 1 chr\tNode 1 start\tNode 1
end\tNode 1 direction\tExtra support\tNode 2 chr\tNode 2 start\tNode 2
end\tNode 2 direction\tExtra support\tNode 1 CNVs\tNode 2 CNVs\tGene at node
1\tGene at node 2\tType of fusion\tDetails\n";
      }
}



#make a connection to the Ensembl database
my $registry = 'Bio::EnsEMBL::Registry';
$registry->load_registry_from_db(
      -host => 'ensembldb.ensembl.org',
      -user => 'anonymous'
);

# tells it we want to work with a slice of the human genome
my $slice_adaptor = $registry->get_adaptor( 'Human', 'Core', 'Slice' );
```

```
#read through the list of SVs
while(<INPUT>)
      {
            #chomp the newline, replace the + and - with 1 and -1, and read
into variable
            if ($strands == 0) {
                  $_=~s/\+/1/g;
                  $_=~s/\-/-1/g;
            }
            chomp $_;

            my $structural_variant = $_;

            my $has_sv_been_printed = 0;


            #split up the columns in the SV file
            if ($columns == 1 && $cnv == 0) {
                  ( $type_of_sv,
                        $support_for_sv,
                        $node1_chr,
                        $node1_start,
                        $node1_end,
                        $node1_strand,
                        $extra_support_1,
                        $node2_chr,
                        $node2_start,
                        $node2_end,
                        $node2_strand,
                        $extra_support_2 ) = split('\t', $_);
            }
            elsif ($columns == 1 && $cnv == 1) {
                  ( $type_of_sv,
                        $support_for_sv,
                        $node1_chr,
                        $node1_start,
                        $node1_end,
                        $node1_strand,
                        $extra_support_1,
                        $node2_chr,
                        $node2_start,
                        $node2_end,
                        $node2_strand,
                        $extra_support_2,
                        $node1_cnv,
                        $node2_cnv) = split('\t', $_);

            }
            elsif ($columns == 0 && $cnv == 1) {
                  ( $type_of_sv,
                        $support_for_sv,
                        $node1_chr,
                        $node1_start,
                        $node1_end,
                        $node1_strand,
                        $node2_chr,
                        $node2_start,
                        $node2_end,
                        $node2_strand,
                        $node1_cnv,
                        $node2_cnv) = split('\t', $_);

            }
            else {
                  ( $type_of_sv,
                        $support_for_sv,
                        $node1_chr,
                        $node1_start,
                        $node1_end,
                        $node1_strand,
                        $node2_chr,
                        $node2_start,
                        $node2_end,
                        $node2_strand, ) = split('\t', $_);

            }

            #skip header, if it is there
            next if ($_=~/^Type/);

            #skip LOPs
            next if ($type_of_sv eq 'LOP');

            #skip ITRs (newer equivalent of LOP)
            next if ($type_of_sv eq 'ITR');

            #skip over mitochondria and other haplotypes, etc
            next if ($node1_chr eq 'M' || $node1_chr eq 'MT' ||
$node1_chr=~/"Un"/ || $node2_chr eq 'M' || $node2_chr eq 'MT' ||
$node2_chr=~/"Un"/ );

            #mate pair reads have the opposite strand
            if ($matepair == 1) {

                  if ($node1_strand == 1) {
                        $node1_strand=~s/1/-1/g;
                        }
                  else {
                        $node1_strand=~s/-1/1/g;
                        }

                  if ($node2_strand == 1) {
                        $node2_strand=~s/1/-1/g;
                        }
                  else {
                  $node2_strand=~s/-1/1/g;
                  }
            }


            ########################################
            ## FIND THE GENES AT THE BREAKPOINTS ##
            ########################################
            my $which_node = 1;
            my ($node1_breaks_gene, $has_sv_been_printed_in_node1,
$node1_genearrayref) = get_broken_genes($has_sv_been_printed,
$structural_variant, $node1_chr, $node1_start, $node1_end, $which_node);
            $has_sv_been_printed = $has_sv_been_printed_in_node1;
            my @node1_genearray = @$node1_genearrayref;

            $which_node = 2;
            my ($node2_breaks_gene, $has_sv_been_printed_in_node2,
$node2_genearrayref) = get_broken_genes($has_sv_been_printed,
$structural_variant, $node2_chr, $node2_start, $node2_end, $which_node);
            $has_sv_been_printed = $has_sv_been_printed_in_node2;
            my @node2_genearray = @$node2_genearrayref;


            ###################################
            ## PREDICT ANY FUSIONS PRODUCED ##
            ###################################

            # check if there are broken genes at both nodes, ie fusion is
possible
            if ( $node1_breaks_gene == 1 && $node2_breaks_gene == 1) {
                  #use a loop to test all genes at node 1 against all genes
at node 2
                  foreach( @node1_genearray ) {

                        # split up the attributes of the gene from node one
                        my( $gene1_dbid,
                              $gene1_displayname,
                              $gene1_externalname,
                              $gene1_start,
                              $gene1_end,
                              $gene1_strand,
                              $gene1_stableid ) = split('\t', $_);

                  foreach( @node2_genearray ) {

                              # split up the attributes of the gene from node
two
                              my ($gene2_dbid,
                                    $gene2_display,
                                    $gene2_externalname,
                                    $gene2_start,
                                    $gene2_end,
                                    $gene2_strand,
                                    $gene2_stableid) = split('\t', $_);



                              # test if the genes are the same - ie it is an
internal deletion/duplication
                              if ( $gene1_stableid eq $gene2_stableid ) {

                                          print
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$gene1
_stableid\",
\"$gene1_externalname\"\)\t=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapien
s\/Gene\/Summary?g\=$gene2_stableid\", \"$gene2_externalname\"\)\tINTERNAL";

                                          undef my @exon_already_seen;

                                          if ($type_of_sv eq "DEL") { # only
check for exons which are deleted, not amplified/inverted

                                                # set boundaries of deletion
- node 2 may be before node 1
                                                my $deleted_slice;
                                                if ($node1_start <
$node2_start) {
                                                      $deleted_slice =
$slice_adaptor->fetch_by_region('chromosome', $node1_chr, $node1_start,
$node2_end);
                                                }
                                                else {
                                                      $deleted_slice =
$slice_adaptor->fetch_by_region('chromosome', $node1_chr, $node2_start,
$node1_end);
                                                }

                                                my $deleted_exons =
$deleted_slice->get_all_Exons();
                                                my $are_exons_deleted = 0;
                                                my $exoncounter = 0;
                                                my @exon_id_list;
                                                while ( my $exon = shift
@{$deleted_exons} ) {
                                                      $are_exons_deleted = 1;
                                                      my $stable_id     =
$exon->stable_id();
                                                      @exon_already_seen =
grep($stable_id, @exon_id_list);
                                                      if (@exon_already_seen)
{
                                                            $exoncounter++;
                                                      }
                                                      else {

      push(@exon_id_list, $stable_id);
                                                      }
                                                }

                                                if ($are_exons_deleted == 0)
{
                                                      print "\tNO EXONS
DELETED";
                                                }

                                                else {
                                                      print "\t$exoncounter
EXONS DELETED";
                                                }
                                          }
                                    }
```

```
                              #test for a head-on collision - ie two genes 
which could produce readthoughs
                              elsif ( $gene1_strand == $node1_strand && 
$gene2_strand == $node2_strand ) {
                                          get_run_through(\@node1_genearray, 
$node1_strand, $node2_strand, $node2_start, $node2_end, $node2_chr);
                                          get_run_through(\@node2_genearray, 
$node2_strand, $node1_strand, $node1_start, $node1_end, $node1_chr);

                              }

                              #test for a 3'to 5' fusion
                              elsif ( $gene1_strand != $node1_strand && 
$gene2_strand == $node2_strand ) {
                                    print 
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$gene1
_stableid\", 
\"$gene1_externalname\"\)\t=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapien
s\/Gene\/Summary?g\=$gene2_stableid\", \"$gene2_externalname\"\)\tFUSION\t5' 
of $gene2_externalname into 3' of $gene1_externalname";
                              }

                              #test for a 5' to 3' fusion
                              elsif ($gene1_strand == $node1_strand && 
$gene2_strand != $node2_strand) {
                                    print 
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$gene1
_stableid\", 
\"$gene1_externalname\"\)\t=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapien
s\/Gene\/Summary?g\=$gene2_stableid\", \"$gene2_externalname\"\)\tFUSION\t5' 
of $gene1_externalname into 3' of $gene2_externalname";
                              }

                              else {
                                    print 
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$gene1
_stableid\", 
\"$gene1_externalname\"\)\t=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapien
s\/Gene\/Summary?g\=$gene2_stableid\", \"$gene2_externalname\"\)\tNO FUSION 
PREDICTED";

                              }
                        }
                  }

                  if ( $has_sv_been_printed == 1 ) {
                        print "\n";
                        }
            }


            elsif( $node1_breaks_gene == 1 && $node2_breaks_gene == 0 ) {
                  #broken gene at node 1, want to find the gene it runs into
                  #node 2 is the test node
                  get_read_through(\@node1_genearray, $node1_strand, 
$node2_strand, $node2_start, $node2_end, $node2_chr);
```

```
                  if ($has_sv_been_printed == 1) {print "\n";};
            }

            elsif( $node1_breaks_gene == 0 && $node2_breaks_gene == 1 ) {
                  #broken gene at node 2, want to find the gene it runs into
                  #node 1 is the test node
                  get_read_through(\@node2_genearray, $node2_strand,
$node1_strand, $node1_start, $node1_end, $node1_chr);
                  if ($has_sv_been_printed == 1) {print "\n";};
            }


      $linecounter++;
      }

close INPUT;
print "Processing finished!\n";

sub get_broken_genes
{
      my $has_sv_been_printed = shift;
      my $structural_variant = shift;
      my $node_chr = shift;
      my $node_start = shift;
      my $node_end = shift;
      my $which_node = shift;

      #create a chromosome slice for the possible breakpoint region
      my $node_chromslice = $slice_adaptor->fetch_by_region('chromosome',
$node_chr, $node_end, $node_end+$insertsize);

      my $node_slicestart = $node_chromslice->start();
      my $node_sliceend   = $node_chromslice->end();
      my $node_genes = $node_chromslice->get_all_Genes();
      undef my @node_genearray;
      my $node_breaks_gene = 0;

      while (my $gene = shift @{ $node_genes } ) {

            #this loop tells it to output the break if it has not been done
before
            if ($has_sv_been_printed != 1) {

                  print "$structural_variant\t";
                  $has_sv_been_printed = 1;
            }

                  # call subroutine to get all the gene attributes and
returned them concatenated into the array
                  @node_genearray = get_gene_attributes($node_slicestart,
$node_sliceend, $which_node, $gene);
                  $node_breaks_gene = 1;
      }

      undef $node_genes;
      return ($node_breaks_gene, $has_sv_been_printed, \@node_genearray);
```

```
}


sub get_read_through
{
                  my $genearray_ref       = shift;
                  my $brokennode_strand   = shift;
                  my $testnode_strand     = shift;
                  my $testnode_start            = shift;
                  my $testnode_end        = shift;
                  my $testnode_chr        = shift;

                  # go and find nearest unbroken gene in correct orientation
                  foreach(@{$genearray_ref})
                  {

                        # split up the attributes of the gene from the broken
node
                        my ($brokengene_dbID,
                              $brokengene_display,
                              $brokengene_externalname,
                              $brokengene_start,
                              $brokengene_end,
                              $brokengene_strand,
                              $brokengene_stableid) = split('\t', $_);
                        #print "broken gene is $brokengene_externalname\n";
                        if ($brokennode_strand == $brokengene_strand) {

                              #fetch 1000bp near test node
                              my ($testslice_start, $testslice_end);

                              if ($testnode_strand == 1)
                              {

                                    $testslice_start = $testnode_start -
1000;
                                    $testslice_end = $testnode_start;
                              }
                              else
                              {
                                    $testslice_start = $testnode_end;
                                    $testslice_end = $testnode_end + 1000;
                              }

                              my $iteration_counter = 1;
                              my $found_a_gene = 0;
                              #only iterate 1000 times max- ie, will find a
readthrough within 1Mb of the break
                              until ($found_a_gene == 1 || $iteration_counter
== 1000)
                                    {

                                          my @nearbygenes =
get_nearby_genes($testnode_chr, $testslice_end, $testslice_end+$insertsize);

                                          foreach(@nearbygenes)
                                                {
                                                      my ($nearbygene_dbID,

      $nearbygene_display,

      $nearbygene_externalname,

      $nearbygene_start,
                                                            $nearbygene_end,

      $nearbygene_strand,

      $nearbygene_stableid) = split ("\t", $_);

                                                      if ($nearbygene_strand
!= $testnode_strand)
                                                      {


                                                            print
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$broke
ngene_stableid\",
\"$brokengene_externalname\"\)\t=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_s
apiens\/Gene\/Summary?g\=$nearbygene_stableid\",
\"$nearbygene_externalname\"\)\tREADTHROUGH\t$brokengene_externalname is
broken and may read through into $nearbygene_externalname
",$iteration_counter,"kb away";
                                                            $found_a_gene =
1;
                                                      }
                                                }

                                          if ($testnode_strand == 1)
                                          {
                                                $testslice_end =
$testslice_start;
                                                $testslice_start -= 1000;


                                          }
                                          else
                                          {
                                                $testslice_start =
$testslice_end;
                                                $testslice_end += 1000;
                                          }

                                          undef @nearbygenes;
                                          $iteration_counter++;

                                    }

                        }

                        else {
                              print
"=HYPERLINK\(\"http\:\/\/$ensembl_site\/Homo_sapiens\/Gene\/Summary?g\=$broke
ngene_stableid\", \"$brokengene_externalname\"\)\t\tNO READTHROUGH WITHIN
1Mb";
                        }
                  }

}


sub get_nearby_genes
{
      undef my @genearray;
      my $chr     = shift;
      my $start   = shift;
      my $end     = shift;
      my $chromslice = $slice_adaptor->fetch_by_region('chromosome', $chr,
$start, $end);
      my $genes = $chromslice->get_all_Genes();
      while ( my $gene = shift @{$genes} )
            {
                  my $dbID          = $gene->dbID();
                  my $display       = $gene->display_id();
                  my $externalname= $gene->external_name();
                  my $genestart   = $gene->start();   #positions are relative
to the slice, not the absolute chromosomal location
                  my $geneend       = $gene->end();
                  my $genestrand  = $gene->strand();
                  my $stable_id     = $gene->stable_id();
                  my $geneconcat = join("\t", $dbID, $display, $externalname,
$genestart, $geneend, $genestrand, $stable_id);
                  push (@genearray, $geneconcat);
            }

      return @genearray;
}

sub get_gene_attributes
{

            my $start   = shift;
            my $end     = shift;
            my $node    = shift;
            my $gene    = shift;

            my $dbID          = $gene->dbID();
            my $display       = $gene->display_id();
            my $externalname= $gene->external_name();
            my $genestart   = $gene->start();   #positions are relative to
the slice, not the absolute chromosomal location
            my $geneend       = $gene->end();
            my $genestrand  = $gene->strand();
            my $stable_id     = $gene->stable_id();

            my $geneconcat = join("\t", $dbID, $display, $externalname,
$genestart, $geneend, $genestrand, $stable_id);

            push (my @genearray, $geneconcat);
            return @genearray;
}
```

B – R script to window sequencing reads

```
# Makes copy number bins for Illumina copy number analysis
# Input is the for_cnv.txt file divided into chromosomes, and requires the
appropriate mappable_starts directory
# Liz Batty, last modified July 2010

#how many reads to put in a bin - more reads, larger bins
readsperbinlist <- c(100, 250, 500)

#include lanes in sample name
samplelist <- c("81777.a","83493.a","82674.a","938.a","946.a","81828.a")

for (sample in samplelist) {
      for(i in c(1:22,"X", "Y")) {
            print(paste("processing chromosome ",i," from ",sample,sep=""))
            #read in mappable starts for the chr
            chrstarts <-
read.table(file=paste("mappable_starts/chr",i,".mappable.starts", sep=""),
header=FALSE)

            #read the file of good reads for CNV analysis (per chromosome)
            cnvreads <-
read.table(file=paste(sample,"/",sample,".forcnv.chr",i,".txt", sep=""),
header=FALSE)


            for (readsperbin in readsperbinlist) {

                  #subset to find the bin edges, determined by readsperbin
                  #ie, if this were idealised genome, this gives bin sizes
which would give readsperbin reads in each
                  chrstarts.short <- chrstarts[seq(1, nrow(chrstarts),
readsperbin),2]

                  #adds a final bin to catch all the end of chr reads
                  bin.edges <-
c(0,chrstarts.short,chrstarts.short[length(chrstarts.short)]+1000000)

                  #cuts the file up according to the bin edges
                  res <- cut(
                        as.integer(cnvreads[cnvreads[,1]==i,2]),
                        breaks=bin.edges-0.1,
                        labels=round(bin.edges[1:(length(bin.edges)-
1)]+(bin.edges[2:length(bin.edges)] - bin.edges[1:(length(bin.edges)-1)])/2)
                  )

                  #write the results to output as a table
                  result <- table(res)

                  write.table(
                        result,
                        sep="\t",

      file=paste(sample,"/",sample,".chr",i,".raw_bincounts.",readsperbin,"re
adsperbin",sep=""),
                        row.names=FALSE,
                        col.names=FALSE,
                        quote=FALSE
                  )

            }
      }
}
```

C – Perl script to retrieve GC percentages

```
# GC percentage fetcher
# Takes the raw_bincounts from the copy number part of the pipeline, and gets
the GC percentage over those bins
# Liz Barrt last modified July 2010

use warnings;
use strict;
use Bio::EnsEMBL::Registry;

# set default values
my $linecounter = 0;
my $input = 0;

#make a connection to the Ensembl database
my $registry = 'Bio::EnsEMBL::Registry';
$registry->load_registry_from_db(
      -host => 'ensembldb.ensembl.org',
      -user => 'anonymous'
);

#input is copy number bins from copy_number_bins.R
my @files = <*raw_bincounts*>;

foreach(@files)
      {
            $input            = $_;

            my ($sample, $lanes, $chr, $bincount, $reads) = split(/\./,$_);
            my $output
      ="$chr.$reads.gcbins.txt";

            open (INPUT, $input)          || die print "failed to open input
file $!";
            open (OUTPUT, ">$output")                 || die print "failed to
open output file $!";

            print "Input file is $input\n";
            print "Output file will be saved as $output\n";


            # tells it we want to work with a slice of the human genome
            my $slice_adaptor = $registry->get_adaptor( 'Human', 'Core',
'Slice' );

            my $bin_start = 1;
            #read through the list of SVs
            while(<INPUT>)
            {
                        print "Processing line $linecounter of $chr\r";

                        #chomp the newline, replace the + and - with 1 and -
1, and read into variable
                        chomp $_;
                        #split up the columns in the SV file
                        my( $bin_end,
                              $reads ) = split('\s', $_);


                        my $short_chr = substr $chr, 3;
                        #print "bine end is $bin_end, reads is $reads, short
chr is $short_chr\n";

                        my $chromslice = $slice_adaptor-
>fetch_by_region('chromosome', $short_chr, $bin_start, $bin_end);

                        #print Dumper ($chromslice);
                        my $gc_count = $chromslice->get_base_count->{'%gc'};

                        print OUTPUT "$bin_start\t$bin_end\t$gc_count\n";
                        $bin_start = $bin_end;
                        $linecounter++;
            }

      close INPUT;
      close OUTPUT;
      }

print "Processing finished!\n";
```

D – R script to perform GC correction and segmentation of sequencing reads, and produce graphs and tables

```
# GC correction, segmentation and output of graphs from Solexa binned reads
# Liz Batty, last modified July 2010
# requires DNAcopy library


library(DNAcopy)

mb <- seq(0,2.5e8,5e6)
shortmb <- seq(0,250,10)
samplelist <- c("83493.a", "82674.a","938.a","81828.a","946.a")
readsperbinlist <- c(100,250,500)

for (sample in samplelist) {
      for(i in c(1:22,"X")) {
            print(paste("processing chromosome ",i," from ",sample,sep=""))

            for (readsperbin in readsperbinlist) {

                  # read in binned reads and reformat for DNAcopy
                  chrbincount <-
read.table(file=paste(sample,"/",sample,".chr",i,".raw_bincounts.",readsperbi
n,"readsperbin",sep=""), header=FALSE)
                  chrbincount[,3] <-
array(i,dim=c(length(chrbincount[,1]),1))
                  chrCNA <- CNA(chrbincount[,2], chrbincount[,3],
chrbincount[,1], data.type="logratio", sampleid=sample)

                  #smooth (remove outliers) and segment the data
                  smooth.chrCNA <- smooth.CNA(chrCNA)
                  segment.chrCNA <- segment(smooth.chrCNA, verbose=1,
undo.splits="sdundo", undo.SD=2)

                  #read in GC percentages (retrieved from Ensembl - see
gcpercent.pl) and add to smoothed data
                  chrgc <-
read.table(file=paste("gcpercent/chr",i,".",readsperbin,"readsperbin.gcbins.t
xt",sep=""), header=FALSE)
                  smooth.chrCNA[,4] <- chrgc[,3]
                  colnames(smooth.chrCNA) <- c("chr", "position", "reads",
"gc")
                  smooth.chrCNA[,3][smooth.chrCNA[,3]==0] = NA

                  print("Performing loess correction")

                  # perform loess over chromosome data and predict correction
factor, then correct data and plot
                  smooth.chrloess <- loess(smooth.chrCNA$reads ~
smooth.chrCNA$gc, span=0.3, loess.control(iterations=3))
                  smooth.chrCNA$loesspred <- predict(smooth.chrloess,
smooth.chrCNA$gc)
                  smooth.chrCNA$dist_from_median <-
(smooth.chrCNA$loesspred/median(smooth.chrCNA$reads, na.rm=TRUE))
                  smooth.chrCNA$reads[smooth.chrCNA$reads==NA] = 0
                  smooth.chrCNA$corrected <-
(smooth.chrCNA$reads*(1/smooth.chrCNA$dist_from_median))
                  smooth.chrCNA$mbposition <- smooth.chrCNA$position/1000000

      png(file=paste(sample,"/",sample,".chr",i,".",readsperbin,"readsperbin.
loesscorrected.png",sep=""), width=1000, height=500)

                  plot(smooth.chrCNA$mbposition, smooth.chrCNA$corrected,
                  pch=20,
                  xlab="Position (mb)",
                  ylab="Normalized reads",
                  sub=paste("Corrected copy number plot for chr",i," with
",readsperbin," reads per bin",sep=""),
                  xaxt="n",
                  #ylim=c(0,500)
                  )

                  axis(1,shortmb)
                  dev.off()


                  # redo the DNAcopy segmentation with corrected data
                  correctedCNA <- CNA(smooth.chrCNA$corrected,
smooth.chrCNA$chr, smooth.chrCNA$position, data.type="logratio",
sampleid=sample)

                  segment.correctedCNA <- segment(correctedCNA, verbose=1,
undo.splits="sdundo", undo.SD=1, alpha=0.005)

                  #print the segments to file
                  segmentmatrix <- as.matrix(segment.correctedCNA)
                  write.table(as.matrix(segment.correctedCNA$output),
sep="\t",
file=paste(sample,"/",sample,".chr",i,".",readsperbin,"readsperbin.correcteds
egments.seg",sep=""), row.names=FALSE, col.names=TRUE, quote=FALSE)
                  segmentmatrix <- as.matrix(segment.correctedCNA)
                  newsegments <- segmentmatrix[2,1]

                  #print segments formatted for circos histogram
                  hs <-
paste("hs",array(i,dim=c(length(segment.correctedCNA$output$loc.start))),
sep="")
                  circos <-
cbind(hs,segment.correctedCNA$output$loc.start,segment.correctedCNA$output$lo
c.end, segment.correctedCNA$output$seg.mean)
                  write.table(
                        circos,

      file=paste(sample,"/",sample,".chr",i,".",readsperbin,"readsperbin.segm
ents.circos",sep=""),
                        quote=FALSE,
                        sep="\t",
                        na="0",
                        row.names=FALSE,
                        col.names=FALSE
                  )
```

```
         #plot the segments produced from the corrected copy number
manually - easier to change plot than for standard DNA copy plots

   png(file=paste(sample,"/",sample,".chr",i,".",readsperbin,"readsperbin.
correctedsegments.png",sep=""), width=1000, height=500)
         plot(smooth.chrCNA$mbposition, smooth.chrCNA$corrected,
            pch=20,
            xlab="Position (mb)",
            ylab="Normalized reads",
            main=paste("Corrected segmentation plot for chr",i,"
with ",readsperbin," reads per bin",sep=""),
            xaxt="n",
            #ylim=c(0,500)
         )
         segment.correctedCNA$output$loc.start.mb <-
segment.correctedCNA$output$loc.start/1000000
         segment.correctedCNA$output$loc.end.mb <-
segment.correctedCNA$output$loc.end/1000000
         segments(segment.correctedCNA$output$loc.start.mb,
segment.correctedCNA$output$seg.mean, segment.correctedCNA$output$loc.end.mb,
segment.correctedCNA$output$seg.mean, lwd=2, col="red")
         axis(1,shortmb)
         dev.off()



         #print the corrected bin values to a GFF file
         gff <-
paste("chr",array(i,dim=c(length(smooth.chrCNA[,1]),1)),sep="")
         solexa <- array("solexa",dim=c(length(gff),1))
         samplelist <- array(sample,dim=c(length(gff),1))
         dots <- array(".",dim=c(length(gff),1))
         colorcol <- array(";color 000000",dim=c(length(gff),1))
         endpos <- smooth.chrCNA$position
         startpos <- endpos[-(length(endpos))]
         startpos <- append(startpos,1,0)
         gffwhole <- cbind(gff, solexa, samplelist, startpos,
endpos, smooth.chrCNA$corrected, dots, dots, colorcol)
         write.table(
            gffwhole,

   file=paste(sample,"/",sample,".chr",i,".",readsperbin,"readsperbin.corr
ected.gff",sep=""),
            quote=FALSE,
            sep="\t",
            na="0",
            row.names=FALSE,
            col.names=FALSE
         )
      }
   }
}
```

Appendix 5 – Pipeline documentation

Documentation produced for the bioinformatic pipeline described in Chapter 6.

Processing solexa sequencing data

General useful unix information

`cd` will change directory, like under Windows. To move up a directory, use `cd ..`

`ls` lists all the files in a directory.

`grep <pattern> <file>` searches for a particular string in a file. This is useful for finding the original reads from a large file. `grep -A 5 <pattern> <file>` will pull out a line and the 5 lines following it.

To read the help pages for a command, use `man command.`

If two commands are separated by `|` , the output of the first command is 'piped' into the second command.

If a command is followed by `> filename` , the output will end up in that file. This is used to string commands together, especially when the intermediate files would be very large.

`gunzip` unzips compressed files (extension .gz). By default this removes the compressed file and replaces it with the uncompressed one; to keep it, use `gunzip -c input.txt.gz > output.txt`

`ca`t concatenates files. If it is used with only one file, it will just output that file, so it is used as a quick way to pipe a file into some other command.

Many of the bioinformatics programs are only available under Unix. To access them, you can run a virtual machine inside OSX, which is slower than running them outside the virtual machine but does work, although the screen can be slow to respond. To use the virtual machine, run the program VirtualBox and start up the ubuntu install - the username is liz and the password is Liz. You can also install programs on OSX by compiling them yourself using the GCC compiler found in the Apple XCode developer’s tools, or getting Darwin/FinkCommander to install them from a Debian/Ubuntu package if one exists.

File formats

.sh files are bash files - essentially a list of unix commands which will be run in order. Variables which are passed to the .sh script are stored as $1, $2, $3, etc. The line `datadir=$1` reads the first variable into `datadir`, which will be used in the script whenever `${datadir}` is used.
.pl files are Perl scripts. If there is no input file specified, they use the file given as a command line argument.

awk/gawk is a language used to quickly manipulate text files.

.R is an R script.

.sam files are Sequence Alignment/Map files. This is the new standard format for aligned sequences, and is describe in detail here:
http://samtools.sourceforge.net/SAM1.pdf

.bam files (and associated .bai files) are compressed .sam files. They are not human-readable but they are much smaller than .sam files. To convert SAM to BAM (and vice-versa) requires the SAMtools utilities http://samtools.sourceforge.net/

To reach GroupDocs from a Mac, use cd /Volumes/Edwards/GroupDocs. To reach GroupDocs under Linux, it needs to be mounted with the command :

```
sudo mount -t cifs //datacentre/Edwards -o username=emb51,domain=h-mrc,password=password Documents
```

This will put GroupDocs in the folder Documents.

Alignment

This is the process of finding the best match in the reference genome for each read. This is usually done for us by the CRI.

The raw sequences come as FASTQ format files, with extension fq. Usually there are two files per lane, one for each read in the pair, and straight off the machine they are named for lane and read in the pair - s1_1.fq, s1_2.fq, s2_1.fq, etc

The fq format looks like

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The first line is the unique id - machine name and run, lane on the flow cell, tile, xy coordinates on the tile, and /1 or /2 to indicate read in a pair.
The second line is the sequence.

The third line is just a spacer.

The last line is the quality score for each base in the sequence. The quality scores are in Sanger format. Each character is the quality score + 33 encoded as the ASCII character for that number.

See http://en.wikipedia.org/wiki/FASTQ_format for more details.

Human reference genomes for build 18 and 19 are in the human reference genome directories. They are formatted as FASTA files, one entry per chromosome.

Alignment using MAQ

All the cell line data on GroupDocs was aligned using MAQ. MAQ is slow, does not do proper gapped alignment, and only supports reads up to 63bp, but it may be the only way to align mate pair reads. (It may be possible to align it with BWA/Bowtie if you tell it to expect the opposite orientation I haven't tested this, and it may not work for both populations in the mate pair library.) It aligns 2 million 37bp reads to the reference genome in about 2 hours.

See MAQ manual for details of alignment (http://maq.sourceforge.net/maq-manpage.shtml) but the basic alignment process is as follows:

1. Convert FASTQ files to MAQ's binary FASTQ format (.bfa)
2. Map reads in .bfa format to the reference sequence. This produces a .map file, which is not human-readable.
3. Output aligned reads as a mapview file.
4. This outputs the reads as a mapviewpair file, which is very similar to a mapview file but puts the two reads in a pair on one line, and adds an extra column for later de-duplication. This column is just the chromosome and position for the two reads separated by colons (eg, 1:1256980:1:1345980), and is used to remove PCR duplicates. NOTE: Some of the very early data does not have the final column on it. If the file does not have this column, use the add28.pl script.

Alignment using BWA

The current CRI pipeline aligns using BWA. This is much faster, does 108bp reads, and produces gapped alignment. It can align millions of reads in a few minutes. This is the alignment method used for the new tumour data.

BWA can be found at http://bio-bwa.sourceforge.net/bwa.shtml . General alignment procedure:

1. Index the reference genome: `bwa index -p prefix -a ls <in.db.fasta>`
2. Align with command `bwa aln` . Each end is aligned separately. See manual for options - I have not played around with anything but the defaults.
3. Generate SAM format output file `bwa sampe` .

NOTE: BWA is quite fiddly - in particular it dislikes some reference genomes and will give an error shortly after starting the alignment. Bowtie is a similar alignment program which has a much more user-friendly manual and may be more useful.

Alignment using novoalign

http://www.novocraft.com/main/index.php

Novoalign is much slower than either MAQ or BWA (a million reads takes ~16 hours) but is more accurate. In the current CRI pipeline it is used to realign reads called as aberrant in case they are misalignments from BWA.

Calling structural variants

This can be broken down into the different steps:

1. Generate stats on the file.
2. Remove PCR duplicates.
3. Call the abnormal reads.
4. (optional) Re-align the abnormal reads with a slower/better alignment program.
5. (optional) Remove the bad regions.
6. Cluster the reads and call structural variants.

Calling structural variants under the old pipeline

This pipeline was used for all of the cell line data. All the commands are listed in the file maq_paired_end_postprocessing.sh

1. Stats are generated using the `get_stats_from_mapviewpair.pl` script, which takes the .mapviewpair file as an input. The stats are used to find the upper limit of the library size.

2. The last column of the mapviewpair file is used to remove duplicates as it contains the chromosome and start position for the two reads in a pair, which should be unique for all pairs. The unix command uniq is used for this: `uniq -f 28` returns all the lines in a file which have a unique column 28. This will only remove duplicates which are adjacent to each other (so the file needs to be sorted first) and it will always return the first of the duplicates.

3. The abnormal reads are called using awk. The intention is to find all the pairs where the two reads are farther apart than the upper limit, or have the wrong strands for the reads (ie, FR is normal, FF, RR, and RF are abnormal).

The gawk command is:

```
gawk -v UPPER=${upper} '(and($6, 2) == 0 || $5 < -UPPER || $5 > UPPER)
&& $9 > 0'
```

This uses a bitwise and to query the mapviewpair flag – see below for an explanation of the bitwise flag.

4. In the old pipeline the reads are not re-aligned.

5. The bad regions are removed using the script `filter_mapview_by_region.pl` . This needs a file of bad regions in the form <chromosome> <start> <end>. The existing file is based on bad regions found by Susie and Jess. The current file is called `regions_to_mask.curated.txt,` and there is a version updated for the new genome build called `regions_to_mask.grch37.txt.`

6. The structural variants are called using sv_from_mapview.pl. The input should be the mapviewpair file with the abnormal read pairs in.

The script needs two options, `-insertmax <number>` which should be the upper library insert size, and `-mappingqualmin <number>` which is the minimum quality of read to use to call structural variants. `-bed` is optional and specifies a bed file for output as well as a plain text file.

Calling structural variants using the new pipeline

This is more complicated because Kevin's updated pipeline is more closely tied into the CRI cluster, but it is likely that the CRI will be able to produce most of the necessary files for us.The new tumour data has all been run through part of this pipeline, and we get the data already processed through step 4 . We have Kevin's handover documentation which explains a lot of this pipeline – secondary_pipeline.txt and structural_variation_pipeline.txt.

1. The stats are generated from the .bam file by a CRI script. The script is alignment_stats.pl, but it won't run without the cluster.

2. Duplicates are computed using samtools or Picard. Samtools is mentioned above. Picard is a similar sort of thing but uses a java command-line interface to manipulate .sam/.bam files. You can find it here: http://picard.sourceforge.net/

This produces a file called `<library>.dupreport.txt` which has statistics on the duplication rate in the library. The different statistics are explained here:
http://picard.sourceforge.net/picard-metric-definitions.shtml#DuplicationMetrics
Of particular interest is the Estimated Library Size, which estimates the number of unique molecules in the library and can give an idea of the depth of the library.

3. I do not have the exact command which produces aberrant reads under the new pipeline. However, it should be fairly simple to do this with gawk in the same way as the old pipeline. The two requirements are to pull out pairs where the reads are in the wrong orientation, and pairs where the two reads map further apart than the maximum library size.

4. The aberrant reads are re-aligned using Novoalign, which is more sensitive. We have been re-aligning against the whole genome, but also against the different haplotypes, which helps remove false positives. The file of interest is
`${Library}.bwa_aberrant_pairs.novoalign.processed.aberrants.namesorted.sam.`

5. The bad regions are not removed in the new pipeline at the CRI. I have been masking the list of bad regions, the centromeres, and 100kb from the telomeres, which removes a lot of SVs in those regions. Use the command `filter_sv_by _region.pl -regions regions_to_mask.grch37.txt library.aberrantpairs.sam > library.aberrantpairs.filtered.sam`

6. The structural variants are called using `sv_from_sam.pl` . This can find the structural variants in multiple libraries in one run. To tell the script what samples to expect, we construct a samplesheet, containing the library prefix, maximum insert size, and library name.

Eg, for the new tumour libraries:

```
CRIRUN_369:4     504   81823
CRIRUN_306:1     621   83539
```

The library prefix is the run number and lane, which can be found in the .sam files or the stats file, and tells the sv caller what sequences belong to which library.

A simple command to call SVs using sv_from_sam.pl:

```
cat
81823.bwa_aberrant_pairs.novoalign.processed.aberrants.namesorted.sam
83539.bwa_aberrant_pairs.novoalign.processed.aberrants.namesorted.sam
| sv_from_sam.pl -samplesheet samplesheet.txt > 81823-
83539.sv_calls.txt
```

This will concatenate the aberrant pairs from 81823 (a tumour) and 83539 (the matched normal) libraries. This file is then sent to the sv caller, which knows which sequences belong to which sample using the sample sheet. The output will have a final column showing how many reads support the SV in each library. Note that if you want 2 reads supporting a variant, the script does not care which library they come from, so it could be 1 from the tumour and 1 from the normal.

Further options:

`-mappingqualmin <number>` is the minimum map quality to use, default is 35.

`-edgepairs <number>` is how many reads needed to call an SV, default is 2.

`-clip <number>` will clip all alignments to the specified length, to deal with multiple read lengths in the same file.

`-outputreads <file>` will output all the reads which contribute to the SVs to a .sam file.

To filter the SVs, use the script sv_filter.pl . This returns only those SVs with more than N hits in the tumour and no hits in the normal.

```
perl sv_filter.pl  -tumour <name> -normal <name> -hits <number>
tumour-normal.svcalls.txt > filtered.svcalls.txt
```

Calling fusion genes from structural variants

To call fusion genes, a script retrieves the genes at each of the two nodes of a structural variant and predicts whether any of them are in the right orientation to form a fusion gene.

This relies on the Ensembl API to retrieve the genes. Information about the API can be found here: http://www.ensembl.org/info/data/api.html The script will only run if the Ensembl Perl modules are installed, as well as DBD::MySQL, Getopt::Long and BioPerl.

Different versions of the Ensembl Perl modules use different builds of the genome. To check which genome build you are using, use the ensembldbcheck.pl script, which will output the current versions, and also the coordinates for BCAS3 to check if it is Hg18 or Hg19/GrCH37.

The file has different options to cope with different formats of input SV file, as the columns have changed over time.
--columns: use if the file has extra support columns, as most of the later files do
--cnv: use if the file has been checked against a list of CNVs, see below
--insertsize <number>: max insert size of the library
--strands: use if the file has -1 and 1 as strands instead of + and -.
--matepair: use if the input is from a mate pair library – all it does is flip the strands of the reads

A typical command:

```
perl fusion_gene_prediction.pl –columns –insertsize 450
81823.abcd.sv_calls.txt > 81823.breaks.txt
```

The output is a text file, but has automatic hyperlinks for Excel.

There is a script which runs a simple check to see if the nodes of the SV overlap with a list of known CNV regions (taken from Conrad et al.) and adds an extra column. This is cnvcheck.pl and takes the SV file as input, it also requires the conradcnv.txt file with the CNVs in it. Use the –columns option if the SV file has the extra support columns.

Copy number pipeline

The input for the copy number calling pipeline is the library.for_cnv.txt file produced by the CRI alignment pipeline. This contains all the start positions of the reads. Currently there is no script to run this code all in one go, as it is difficult to run R from the command line in Windows, so I do it in stages and cut and paste the code into R.

1. Chop up the for_cnv file into the individual chromosomes. This is done using the cnvchopper.pl script

2. Run the code in copy_number_bins.R . This needs the directory mappable_starts, which was generated by Kevin and lists all the potential start positions of a read on the chromosome. This is used to divide up the genome into bins where we would expect the same number of reads – this is not as simple as dividing up the genome into equal size pieces, as fewer reads will map to repetitive regions. Then the actual reads are placed into the bins calculated from the genome. The output files are all named for the library, chromosome, and number of reads per bin used to calculate the genome bins.

To run this code, put the window sizes you want in the readsperbin list:

```
readsperbinlist <- c(100, 250, 500)
```

and the names of the libraries in the samplelist:

```
samplelist <- c("81777.a","83493.a","82674.a","938.a”)
```

250 is a good number for the bin size.

3. Retrieve the GC percentages for each bin across the genome. This only needs doing once for each bin size. The gcpercent directory contains the files for bin sizes 100 and 250.
 Any further GC percent data can be retrieved using the gcpercent.pl script. This will retrieve the GC percent data for any files with raw_bincounts in the name in the directory it is run from.

4. The code in binsize_graphs.R performs a loess correction on the data using the GC percentage, segments it with DNAcopy, and outputs some plots, the copy number segments as .seg files and also formatted for Circos plots, and a .gff file of the corrected data. It needs the DNACopy R library. Again, the readsperbinlist and the samplelist need to include the bin sizes used and the libraries to process.

SAM flags and bitwise and functions

Both the mapview and SAM formats use a bitwise flag. This is a way of encoding multiple pieces of information in a single number, by looking at the individual bits of the number as a binary number.

For instance, in the bitwise flag in a SAM file, the first three bits represent whether the read is pair, whether the read is mapped in a proper pair, or whether the read is unmapped.

Bit 1: read is in a pair
Bit 2: read is in a proper pair
Bit 3: read is unmapped

If all these things are true for a read, all three flags would be set. This could be represented in binary as 111, converting this to decimal gives us 1+2+4 = 7.

If the read is in a pair, but the pair is not a proper pair and the read is unmapped, the binary representation is 100, and in decimal this is 1.

(For another explanation, see here: http://seqanswers.com/forums/showthread.php?t=2301)

SAM encodes eleven different bits of information in a single field. These are described in the SAM specification, and there is a tool to decode them here: http://picard.sourceforge.net/explain-flags.html

The bitwise `and` function queries the individual bits. For instance, the code `and(<flag>, 4)` would compares the two things in the brackets , in this case the bitwise flag, and the 4 bit. If both these were set to 1 (or true), it would return 1. If the flag is false at the 4 bit, it will return 0. In this way we can select only the reads where the 4 bit of the flag is set to 1, ie all the mapped reads, and reject all the unmapped reads.