The Structure and Evolution of Breast Cancer Genomes

Scott Newman

Clare College, University of Cambridge

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

February 2011

Declaration

This dissertation contains the results of experimental work carried out between October 2007 and December 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

The Structure and Evolution of Breast Cancer Genomes

Scott Newman

Chromosome changes in the haematological malignancies, lymphomas and sarcomas are known to be important events in the evolution of these tumours as they can, for example, form fusion oncogenes or disrupt tumour suppressor genes. The recently described recurrent fusion genes in prostate and lung cancer proved to be iconic examples as they indicated that important gene fusions are found in the common epithelial cancers also. Breast cancers often display extensive structural and numerical chromosome aberration and have among the most complex karyotyes of all cancers. Genome rearrangements are potentially an important source of mutation in breast cancer but little is known about how they might contribute to this disease.

My first aim was to carry out a structural survey of breast cancer cell line genomes in order to find genes that were disrupted by chromosome aberrations in “typical” breast cancers. I investigated three breast cancer cell lines, HCC1187, VP229 and VP267 using data from array painting, SNP6 array CGH, molecular cytogenetics and massively parallel paired end sequencing. I then used these structural genomic maps to predict fusion transcripts and demonstrated expression of five fusion transcripts in HCC1187, three in VP229 and four in VP267.

Even though chromosome aberrations disrupt and fuse many genes in individual breast cancers, a major unknown is the relative importance and timing of genome rearrangements compared to sequence-level mutation. For example, chromosome instability might arise early and be essential to tumour suppressor loss and fusion gene formation or be a late event contributing little to cancer development.

To address this question, I considered the evolution of these highly rearranged breast cancer karyotypes. The VP229 and VP267 cell lines were derived from the same patient before and after therapy-resistant relapse, so any chromosome aberration found in both cell lines was probably found in the common *in vivo* ancestor of the two cell lines. A large majority of structural variants detected by massively parallel paired end sequencing, including three fusion transcripts, were found in both cell lines, and therefore, in the common ancestor. This probably means that the bulk of genome rearrangement pre-dated the relapse.

For HCC1187, I classified most of its mutations as earlier or later according to whether they occurred before or after a landmark event in the evolution of the genome - endoreduplication (duplication of its entire genome). Genome rearrangements and sequence-level mutations were fairly evenly divided between earlier and later, implying that genetic instability was relatively constant throughout the evolution of the tumour. Surprisingly, the great majority of inactivating mutations and expressed gene fusions happened earlier. The non-random timing of these events suggests many were selected.

Acknowledgements

Thanks to members of the Edwards group past and present and especially to Paul Edwards, himself, for his sound advice and good humour over the past few years.

I am also grateful to Clare College and the Medical Research Council for their financial support.

Most of all, thanks also to my patient and understanding wife, Star who has supported me throughout my Master's and Ph.D studies. And last but not least, I need to thank my daughter, Ellinore. Her impending birth made me hasten the speed at which I was writing.

Contents

List of figures

List of tables

Abbreviations

API	application programming interface
ATCC	American Type Culture Collection
BAC	bacterial artificial chromosome
BSA	bovine serum albumin
BWA	Burrows Wheeler alignment
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
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
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
MLE	maximum likelihood estimator
PBS	phosphate buffered saline
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. Cancer

Cancer comprises a large number of diseases that can affect every tissue of the body and can afflict people at all ages. In 2006 cancer caused about 13% of all human deaths (Watson et al., 2006)

First and foremost, cancer is a disease of uncontrolled cell division. Under normal circumstances somatic cells divide, quiesce or die when appropriate but when a cell becomes cancerous it, and its progeny, divide uncontrollably, eventually forming a tumour. Often early cancers are in the form of benign, encapsulated lesions confined to a single tissue and many of these pre-malignant lesions do not represent a danger to health. Some benign lesions, however, acquire an ability to invade surrounding tissues and eventually spread to distant areas of the body - a process known as metastasis. The majority of cancer-related deaths are caused by metastatic lesions.

Secondly, cancer is an evolutionary process and evolution is stepwise mechanism driven by mutations in DNA. Nowell (1976) suggested that an initiating event causes a cell to divide inappropriately and the uncontrolled and error-prone process of cell division leads to the accumulation of genetic alterations (Nowell, 1976). This facilitates the “continual selective outgrowth of variant sub-populations of tumour cells with a proliferative advantage” (Bell, 2010, p.231). In subsequent years Nowell's view has been proven and now we regard each tumour as “... the outcome of a process of Darwinian evolution occurring among cell populations within the micro-environments provided by the tissues of a multicellular organism“ (Stratton et al., 2009, p.719). Thus, cancer represents a cell's regression to a state of self-interest. Rather than obeying instructions from its surrounding environment for the good of the organism's germ line a cancer cell “decides” the best way to perpetuate its genes is to divide regardless of the interests of the organism.

Central to the process of evolution is mutation. Mutations may be caused by chemical carcinogens, such as tobacco smoke, radiation or viral insertion into the genome but can also arise from random errors in DNA replication or repair. But regardless of the source of mutation the net result is production of gene variants which allow their host cell to either

survive and reproduce better or to die. A cancer cell must, therefore, acquire mutations allowing it to survive better and reproduce more within its habitat and eventually to move expand into other bodily habitats.

Molecular biology has identified several characteristics that cells acquire when they evolve towards this self-interested state: the so-called 'Hallmarks of Cancer.' (Hanahan and Weinberg, 2000). These traits are: i) Self-sufficiency in growth signals ii) insensitivity to anti-growth signals iii) evasion of apoptosis iv) limitless replicative potential v) sustained angiogenesis, where appropriate and vi) the ability to invade surrounding tissues and metastasise (Hanahan and Weinberg, 2000). The hallmarks have become a popular lens through which to view cancer evolution as they effectively split a large question, “what causes cancer?” into a series of smaller ones.

The hallmarks are far from immutable, and debate, to varying extents, exists over each. For example, liquid tumours do not require a blood supply and are, by definition, metastatic. Others have suggested phenotypes such as chromosome instability and escape from senescence should be considered hallmarks as well (Shay and Roninson, 2004; Negrini et al., 2010). More hallmarks probably remain to be described but, for now, the ongoing challenge of cancer research is to identify the genetic changes that alter the specific cellular processes necessary for cancer to develop.

1.2. Breast Cancer

1.2.1. Susceptibility Alleles

Breast cancer accounts for approximately 20% of all cancers in Western Europe and the USA. Five to ten percent of breast cancers show clear inheritance through families where mutations in *BRCA1* and *BRCA2* genes confer a highly penetrant disease risk (Miki et al., 1994; Wooster et al., 1995). Variants of four other genes, *CHEK2*, *BRIP1*, *ATM* and *PLAB2* confer a 2-4 fold relative risk of breast cancer and are classed as intermediate penetrance alleles (Ripperger et al., 2009). These mutations probably contribute in large part to early cancer development but whether they, themselves, cause a growth advantage or just

facilitate further mutation remains a contentious issue (Sieber et al., 2002; Skoulidis et al., 2010).

Li Fraumeni syndome and Cowden disease families, carrying mutations in *TP53* and *PTEN* receptively, also have an increased risk of developing breast, as well as other cancers (Li and Fraumeni, 1969; Liaw et al., 1997). Various other DNA-repair related syndromes involving *STK11, PTEN, CDH1, NF1* and *NBN* genes also increase the risk of breast cancer (Mavaddat et al., 2010) but due to the rarity of these syndromes, their overall contribution to the population burden of breast cancer is small.

Risk within the general population is modulated by several common gene variants including *FGFR2*, *TNRC9*, *MAP3K1, LSP1* and *RAD51L1* (Easton et al., 2007; Thomas et al., 2009). Several non-coding SNPs also confer susceptibility and probably play a role in the regulation of other cancer relevant genes (Wright et al., 2010). The moderate contribution to disease risk by SNPs in individuals means that most breast cancers are considered sporadic with no precise genetic or environmental cause determined.

1.2.2. Breast Cancer Histology

A defining feature of breast cancer is its heterogeneity. Breast cancers have distinct histopathological features, genetic and genomic variability, so are now considered by some as collection of diseases arising in the same organ rather than a single disease (Vargo-Gogola and Rosen, 2007). The challenge of identifying causative mutations in sporadic breast cancer has, therefore, been particularly difficult as hundreds of genes are mutated or rearranged and thousands of genes are differentially expressed between tumour subtypes.

Histologically, breast cancers are classified into several categories: the most advanced pre-invasive breast cancers are either lobular carcinoma in situ (LCIS) and ductal carcinoma in situ (DCIS). Invasive lesions are subdivided into tubular carcinoma (2%), medullary carcinoma (5%), lobular carcinoma (10%) and ductal carcinoma (80%) (Watson et al., 2006). Breast cancers are also staged according to whether they express the

oestrogen receptor (ER), progesterone receptor (PR) and *HER2/ERBB2* receptor. Breast tumours are further classified at diagnosis based upon size, lymph node status and metastasis, and degree of differentiation. Currently the combination of the above factors forms a model of risk and dictates treatment strategy.

1.2.3. Gene expression patterns

In addition to this histological heterogeneity, breast cancers are also heterogeneous on a molecular level. Messenger RNA profiling studies have revealed that breast cancer may have five or more, subtypes (Sotiriou et al., 2003; Sørlie et al., 2001; Weigelt et al., 2010b). Early studies by Perou et al. (2000) and Sørlie et al. (2001) showed two main and clinically relevant classes, based on ER status. ER-positive tumours can be further divided in luminal A and B. ER-negative tumours can be divided into basal epithelial-like *ERBB2*-over-expressing and normal-breast-like groups (Perou et al., 2000; Sørlie et al., 2001).The luminal subtypes display high levels of ER-activated genes. Luminal A tumours express lower levels of proliferative genes and are usually of low histological grade and have an excellent prognosis. Luminal B cancers tend to be of a higher grade, express higher levels of proliferative genes and have a significantly worse prognosis (Weigelt et al., 2010a). The ER-negative tumours appear to be more heterogeneous. *ERBB2* tumours express genes associated with the *ERBB2* pathway and like basal tumours, which express basal-like cytokeratins, laminins and fatty acid binding proteins, have an aggressive clinical behaviour. The clinical significance of normal breast-like tumours has yet to be determined.

Later studies have always reproduced the ER-positive/ER-negative classification, but subdivisions of these groups have sometimes varied between studies. For example, the luminal A and B classification seems robust, but some studies have proposed a sub-division of the luminal B group that is not always reproducible. Likewise, a significant number of *HER2*-amplified tumours are ER-positive. Weigelt et al., (2010) concluded that “despite the numerous publications describing this molecular taxonomy, it remains a working model in development and not a definitive classification system, given that further molecular subtypes have been and may be identified” (Weigelt et al., 2010a, p.267).

1.2.4. Developmental Hierarchy of Breast Cells

The developmental hierarchy of breast epithelium is not well understood but it is probable that the various subtypes of breast cancer arise in distinct cell types (Stingl and Caldas, 2007). A major question underpinning breast cancer classification is whether the different subtypes do indeed arise from different stem cells or commuted progenitors or if the molecular heterogeneity of breast cancer represents multiple evolutionary routes taken by a common cell of origin.

Mammary epithelium is composed of two lineages: an inner layer of luminal cells and an outer sheath of myoepithelial cells. A current model is that a single stem cell resides at the top of both luminal and myoepithelial hierarchies. This stem cell population splits into committed luminal progenitor and bipotent progenitors. The luminal progenitors produce luminal and alveolar epithelium and the bipotent progenitors giving rise to the ductal and myoepithelial components (Stingl, 2009). If cancers arise in committed progenitors rather than from the multipotent stem cell population this would provide a rationale for investigating separate molecular subtypes of breast cancer as if they were separate diseases (Cairns, 2002; Krivtsov et al., 2006).

1.3. Mutations that cause cancer

Classically, somatic mutations that confer a growth advantage could be classified in one of two categories: oncogenes and tumour suppressor genes. Oncogenic mutations are dominantly acting gains of function whereas tumour suppressor mutations are recessive losses of function. But as we have learned more about cancer biology, this simple classification is becoming less clear. For example, some have suggested a 'gatekeeper' and 'caretaker' subdivision for tumour suppressor genes (Levitt and Hickson, 2002).

The first naturally occurring, cancer-causing sequence change in humans was a G>T substitution resulting in a change from glycine to valine in codon 12 of the *HRAS* gene in bladder cancer (Parada et al., 1982; Tabin et al., 1982). Since the 1980s, the list of cancer

genes has grown to a current 427 according to one recent estimate and is only likely to get bigger (Futreal et al., 2004). An important feature of this "cancer gene census" is that genes can be disrupted by numerous different mechanisms in including point mutation, chromosome aberration and epigenetic changes.

1.3.1. Sequence-level changes

Oncogenic mutations typically alter the amino acid sequence of a protein, causing constitutive or inappropriate activation. Examples of this process are point mutations commonly observed in the *RAS* family of proto-oncogenes. *K-RAS* mutations are common in lung, pancreas and colon carcinomas whereas *N-RAS* mutations are often found in haematological malignancies such as acute myeloid leukaemia and *H-RAS* discussed above. The majority of mutations in these genes are in codon 12 and cause constitutive activation of the signal-transduction function of the *RAS* protein (Bos, 1989).

Loss of function can also be achieved by changes to the DNA sequence, for example by generating a stop codon within the reading frame of a gene, as is commonly observed in the tumour-suppressor gene, *APC*. DNA methylation at the CpG islands of gene promoters can also silence gene expression and aberrant methylation can cause tumour-suppressor gene silencing. Methylation at the promoters of various cancer-relevant genes, including *p16(INK4a), APC, BRCA1* and *CDH1* has been described (Esteller et al., 2001).

1.3.2. Sequence-level changes in breast cancer

The first unbiased mutation screens of the breast cancer exome took place in 2007 based on the CCDS database, representing 18,191 distinct genes. The authors sequenced all the coding exons in eleven breast cancer cell lines and showed that an average breast cancer cell line accumulates around 90 mutant alleles in its lifetime. (Sjöblom et al., 2006; Wood et al., 2007). The majority of alterations were single-base pair substitutions (92.7%), with 81.9% resulting in missense changes, 6.5% resulting in stop codons, and 4.3% resulting in alterations of splice sites or untranslated regions. The remaining somatic mutations were insertions, deletions, or duplications (7.3%).

The screen re-identified mutations in cancer genes such as *TP53* and *BRCA1* but the striking outcome of the study was, however, that the large majority of mutations were in genes not previously linked to cancer (Sjöblom et al. 2006; Wood et al. 2007). The authors used several criteria to estimate if a gene had mutated at a rate above background (see section 1.5.2.1), so was likely to have been selected. It was then clear that there was a large number of low prevalence and previously uncharacterised candidate cancer genes in breast cancer.

Recent whole genome sequencing studies have confirmed this view of the genomic landscape of breast tumours. Shah et al (2009) achieved >43-fold sequence-level coverage of an ER-positive metastatic lobular breast cancer obtained from a pleural effusion nine years after the initial diagnosis. Prior to surgery, radiotherapy was used to treat the primary tumour and in the intervening 9 years before relapse, the patient was treated with tamoxifen and an aromatase inhibitor. The authors identified 32 somatic non-synonymous coding mutations. When they went back to the primary tumour, from 9 years earlier, five of the 32 mutations were prevalent in the DNA of the primary tumour at high frequencies, six at lower frequencies and 19 could not be detected. It should be noted that the authors could not search effectively for indels, so the true number of mutations in their tumours was likely to be higher. Ding et al (2010) used a similar approach for a basal-like breast cancer, its metastasis and a xenograft of the metastasis. A total of 50 sequence-level somatic mutations were validated in at least one of the samples. Of these validated point mutations and small indels, 48 were detectable in all three tumours. The metastasis was significantly enriched for 20 mutations and one large deletion shared by the primary tumour.

When Shah et al. (2010) looked for recurrence of their 32 point mutations in 192 breast tumours, none had identical mutations and only three of the tumours had point mutations in one of the same genes. Furthermore, none of the 32 genes were previously identified as candidate cancer genes described by Wood et al. (2007). The authors concluded by "... predict[ing] that the key features of this landscape—a few gene mountains interspersed with many gene hills—will prove to be a general feature of most solid tumours ... This view of cancer is consistent with the idea that a large number of mutations, each associated

with a small fitness advantage, drive tumour progression" (Wood et al., 2007, p.1108).

1.3.3. Changes to chromosome structure

Catalogues of sequence-level somatic coding mutations are now growing rapidly, but this source of mutation only gives us a partial view of how genes can be disrupted in cancer (Wood et al., 2007; Forbes et al., 2010). The first loss of function mutation to be described, for example, was of the Retinoblastoma protein and was mediated through a chromosomal deletion (Knudson et al., 1976; Knudson, 1993).

Clonal chromosome abnormalities, often acquired during carcinogenesis, are visible down the microscope as gains, deletions, inversions and translocations and are a feature of nearly all cancers (Mitelman et al., 1997, 2007; Heim and Mitelman, 2009). Chromosome abnormalities contribute to carcinogenesis by four mechanisms: transcriptional deregulation of proto-oncogenes, duplication of proto-oncogenes (or regions containing them), deletion or interruption of tumour suppressor genes and the formation of chimeric fusion genes.

An increasing number of chromosome abnormalities are now recognised as important diagnostic and prognostic factors. As of 2007, a total of 11,500 articles had been published on clonal cytogenetic abnormalities and of the 427 consensus cancer genes about 70% were discovered at or near chromosome break points (Mitelman, 2000; Mitelman et al., 1997, 2007; Futreal et al., 2004). The vast majority of recorded abnormalities have been described in haematological malignancies as individual cases often contain only a single derivative chromosome and the break points can be mapped by G-banding or techniques such as fluorescence *in situ* hybridization (FISH).

Carcinoma genomes contain more structural variation than the average leukaemia and this has made it very difficult to identify chromosome breakpoints. In contrast to leukaemia etc. previous research in common epithelial cancers has had to focus on sequence-level mutations and genomic gains and losses. As a result, we simply do not know much about the contribution of gene deletions and fusions in such cancers. Some have even

considered chromosome rearrangements unimportant in carcinomas (Vogelstein and Kinzler, 2004). This conclusion probably reflects a lack of data rather than any fundamental difference between chromosome aberrations found carcinomas and those of haematological cancers (Mitelman et al. 2007).

1.3.4. The cytogenetics of breast cancer

In breast cancer, genomic rearrangements are so diverse it is debatable whether breast cancer cytogenetics really exists as it does for leukaemia, lymphoma and sarcoma (Stingl and Caldas, 2007). Past studies have noted the karyotypic alterations were clearly non-random but also very heterogeneous (Teixeira et al., 2002; Teixeira, 2006); the majority of primary breast tumours showing a complex pattern of chromosomal gain, loss and translocation. While breakpoints or regions of gain are often recurrent at cytogenetic resolution, further investigation at the gene-level has often been difficult as the breakpoints show some heterogeneity and additional rearrangements probably also take place (Paterson et al., 2007; Pole et al., 2008).

Observed complexity is so high that only the boldest cytogenetic features have been described in past G-banding and R-banding studies. Using standard cytogenetic methods, several recurrent karyotypic abnormalities have been identified: The most frequent chromosome rearrangements were del(3)(p12~10p14~21), der(1;16)(q10;p10) and i(1)(q10) found in 13, 12 and 9% of cases respectively (Teixeira et al. 2002; Teixeira 2006). The chromosomes 8, 1, 17, 16 and 20 were commonly involved in translocations, t(8;11), t(1;16) and t(5;17) being the most frequent.

Spectral karyotyping (SKY) of metaphase chromosomes in primary tumours and cell lines allowed previously unidentifiable origins of marker chromosomes to be resolved (Adeyinka et al., 1998; Kytölä et al., 2000; Davidson et al., 2000). This allowed us to form a fuller idea of the extent of rearrangement in many breast cancer genomes (although most intrachromosomal events could not be identified by this method). Interestingly, many balanced rearrangements were uncovered by SKY (Davidson et al. 2000) and as there is no (or very little) loss of genetic material in a balanced translocation junction, gene

changes at the chromosome breakpoints are more likely to be important events than those at unbalanced translocation breaks.

1.3.5. Tumour Suppressor Gene Deletion

Loss of 17p is one of the most common cytogenetic events in cancer. The tumour suppressor *P53*, located on 17p is most likely to be the critical gene lost here. Many other such losses can be observed in cancer genomes including focal deletions of *RB1* and *p16* (Knudson et al., 1976; Murphree and Benedict, 1984; Baker et al., 1989).

Recurrent losses are observed in breast cancer genomes also, most frequently, -1p, -8p, -11q and -16q. As regional variations in copy number cause considerable changes in the expression of large numbers of genes (Pollack et al., 2002) the difficulty in breast cancer is finding single genes driving cancer progression amongst a heterogeneous background (Chin et al., 2007). These difficulties are illustrated by 8p deletion in breast and other cancers.

Numerous studies have shown that distal 8p is lost in breast and other tumours (Pole et al., 2006). Increasingly high resolution approaches revealed that the chromosome break often falls within 8p12, within or just proximal to the gene *NRG1*, a plausible breast cancer gene as it is a ligand for the *ERBB2/ERBB3* heterodimer (Falls, 2003; Pole et al., 2008; Cooke et al., 2008). Subsequent systematic investigations of 8p identified the gene *NRG1* as recurrently broken in 6% of primary breast cancers and ovarian cancers (Adélaïde et al., 2003; Huang et al., 2004) and not expressed in a high proportion of breast tumours due to methylation at its promoter (Chua et al., 2009). But still the possibility remains that loss of *NRG1* is not the driving force behind 8p aberrations. Some tumours have breaks distal to *NRG1*, making some hypothesise that the true driver is somewhere distal of 8p12 or that multiple tumour suppressors are found on 8p (Cooke et al., 2008). Also, the *ODZ4-NRG1* fusion in the MDA-MB-175 cell line was the first fusion gene described in breast cancer and seems to encode a pro-proliferative secreted protein, meaning that *NRG1* has, in theory, oncogenic potential (Schaefer et al. 1997; X Liu et al. 1999; X Z Wang et al. 1999).

1.3.6. Oncogene Amplification

Oncogenes can achieve over-expression by increase in their genomic copy number (Pollack et al., 2002; Chin et al., 2007). Amplified oncogenes can be observed in karyotypes as double-minute (DM) chromosomes or homogeneous staining regions (HSRs). Their mechanism of formation is contentious but in each case, up to hundreds of copies of amplified regions can be present (Schwab 1998). DMs are acentric minichromosomes that exist episomally but can also integrate back into the genome (Storlazzi et al. 2010). HSRs are segments of chromosomes that lack any banding pattern and can encompass large regions of amplified genomic DNA. Three gene families *MYC*, *ERBB* and *RAS* are amplified in various tumour types and are clearly important factors in cancer (Santarius et al. 2010).

Several amplifications are recurrently observed in breast cancer: the best known being 17q12, which harbours the *HER2* (*ERBB2*) gene. *HER2* is highly amplified in approximately 20% of cases and defines a prognostically important subclass of breast cancers (Slamon et al., 1987, 1989, 2001). Other recurrent amplifications are on chromosomes 8, 11, 12, 17 and 20, bounding known and postulated breast cancer oncogenes such as *BRF2, ASH2L, CCND1, EMSY, NCOA3, MYBL2* and *STK6* (Garcia et al., 2005; Hughes-Davies et al., 2003). Co-amplification of 8p and 11q is often observed in breast cancer (Paterson et al., 2007). The gene(s) driving this amplification are not known but an explanation is synergistic up-regulation of *CCND1* and *ZNF703* (Kwek et al., 2009).

1.3.7. Gene Fusion

A major advance in cancer cytogenetics came with the description of the Philadelphia chromosome, an abnormal marker in the karyotypes of leukaemic cells (Nowell and Hungerford, 1960; Nowell, 1962). Chromosome banding then identified the Philadelphia chromosome as being derived from a translocation of chromosomes 9 and 22 (Rowley, 1973). Eventually the break points were mapped as t(9;22)(q34;q11) (Heisterkamp et al., 1985) and the *BCR-ABL* fusion oncogene was described and was subsequently found in a high proportion of chronic myeloid and other leukaemias (de Klein et al. 1982; Druker et al.

2001). Most importantly, the over-active kinase *ABL* could be targeted with a non-specific kinase inhibitor and the iconic drug Glivec has provided a paradigm for targeted therapy (Druker et al., 2001).

Examples of oncogenic fusion genes now abound in leukaemias, lymphomas and sarcomas and are often important diagnostic and prognostic features. In 2005 a recurrent fusion gene was found in a common epithelial cancer when Tomlins at al. used a bioinformatic approach to find gene fusions of ETS-family transcription factors to the *TMPRSS2* gene in approximately 70% of prostate cancers (Tomlins et al., 2005). In 2007, Soda et al. used a transformation assay to identify *EML4-ALK* gene fusions in non-small-cell lung cancer and went on to show the fusion was present in approximately 7% of patients (Soda et al., 2007). Many more oncogenic fusion proteins probably remain to be found in the complex genomes of epithelial cancers.

Most gene-fusions seem to fall into several general classes (although there are some examples which do not clearly fit into one of these categories such as the *BCL2* (antiapoptotic) fusions in B-cell leukaemias).

i) Activation of receptor tyrosine kinases
ii) Activation of intracellular kinases
iii) Formation of chimeric transcription factors or chromatin modifiers

1.3.7.1. Receptor Tyrosine Kinases

The *RET–CCDC6* fusion gene in the papillary thyroid carcinoma was the first recurrent genetic change caused by a chromosome aberration in an epithelial cancer and nine subsequent fusions involving *RET* were described. Approximately 40% of thyroid carcinomas are now known to carry one of these chimeric genes (Pierotti et al., 1992). *RET* is a receptor tyrosine kinase, and upon ligand binding the receptor dimerises and transphosphorylates the cytoplasmic tail of its neighboring molecule. The phosphorylated tail can recruit SH2 and SH3 containing cytoplasmic effector proteins, such as *Shc* and *Grb2* to activate mitogenic pathways. A common molecular mechanism leading to

activation of *RET* occurs in all cases. Fusion *RET* oncoproteins can dimerise and transphosporylate independent of ligand, so constitutively activate mitogenic pathways (Alberti et al., 2003).

1.3.7.2. Intracellular Kinases

Fusions of intracellular kinases has been observed in members of the *RAS* signal transduction pathway notably due to tandem duplication. Both *BRAF* and *RAF1* have found to be fused to *KIAA1549* and *SRGAP3* respectively in pilocytic astrocytomas. Both fusions include the kinase domains, and show elevated kinase activity (Jones et al., 2008a, 2009).

1.3.7.3. Transcription factors and Chromatin Modifiers

Gene fusions can form chimeric transcription factors. The translocation t(15;17)(q22;q21) in acute promyelocytic leukemia (PML) fuses the *PML* gene (15q22) with the retinoic acid receptor alpha gene (*RARA*) gene. The PML protein contains a RING finger DNA binding domain and *RARA* encodes the retinoic acid alpha-receptor. The *PML-RARA* fusion protein may confer altered DNA-binding specificity to the *RARA* ligand complex and *PML-RARA* gene fusion provided another target for therapy in the form of the retinoid, all-trans retinoic acid (Huang et al., 1988).

Chimeric chromatin modifiers can also affect gene transcription, for example through fusions of the *Mixed Lineage Leukaemia* gene, *MLL*. The *MLL* protein is a histone methyltransferase found within complexes that regulate transcription via chromatin remodelling and is the target of translocations in leukaemais. Specifically, *MLL* methylates histone H3 lysine 4, and regulates gene expression including multiple *HOX* genes. The numerous leukaemogenic *MLL* chimeras have lost methyltransferase activity and most *MLL* -translocated leukaemias appear to have increased expression of *HOX* and other target genes (Krivtsov and Armstrong, 2007).

1.3.8. Gene fusions in breast cancer

Prior to 2007 only four fusion genes had been described in tumours of the breast: *ETV6–NTRK3*, *ODZ4–NRG1*, *BCAS3-BCAS4* and *TBL1XR1–RGS17* (Wang et al., 1999; Hahn et al., 2004; Tognon et al., 2001; Mitelman et al., 2007). *ETV6–NTRK3*, the only recurrent fusion gene, is specific to secretory breast carcinoma, which is rare and atypical. In 2008, Howarth et al. used an array-based approach to finely map chromosome breakpoints for all balanced breaks within 3 breast cancer cell lines: HCC1187, HCC1806 and ZR-75-30. Breaks in genes *CTCF*, *EP300/p300* and *FOXP4* were observed as well as two gene fusions between *TAX1BP1-AHCY* and *RIF1-PKD1L1* reported (Howarth et al., 2008).

Subsequent studies based around massively parallel paired end sequencing have identified a large cache of, as far as we know, non-recurrent fusion genes in breast cancer cell lines and primary tumours – discussed below (Hampton et al., 2008; Zhao et al., 2009; Stephens et al., 2009). These systematic investigations support the idea that chromosome rearrangements play an important role in breast cancers and one of their modes of action is to fuse genes. There are potentially many recurrently disrupted genes at chromosome breakpoints in breast cancer but to date most have remained elusive (Edwards, 2010).

1.3.9. The complex structure of breast cancer genomes

The mutational burden of genes disrupted at the sequence-level and at the chromosome level is likely to be similar in breast cancer. Recent studies based around massively parallel paired end sequencing have confirmed this fact and add detailed maps of structural variation to the developing picture of the breast cancer genome (Hampton et al., 2008; Stephens et al., 2009). Massively parallel paired end sequencing has also been applied at the transcript level and can detect fusion transcripts (Maher et al., 2009; Chinnaiyan et al., 2009; Zhao et al., 2009).

Hampton et al. (2008) performed a structural survey of the widely-used MCF-7 breast cancer cell line using massively parallel paired end sequencing. They observed 157 somatic breakpoints and 79 known or predicted genes were found at translocation

breakpoints. Ten events were predicted to fuse the reading frames of disparate genes and four could be detected at the transcript level.

Stephens et al. (2009) recently employed a similar approach to discover genes disrupted and fused at chromosome breakpoints in 24 breast cancers (9 cell lines and 15 primary tumours). The authors showed, as did Hampton et al. (2008), that structural variants in breast tumour genomes contribute many hundreds of mutations to the overall total, and furthermore, that genes can be mutated by mechanisms we have not yet fully appreciated such as tandem duplication and internal rearrangement. For cell lines the median number of rearrangements per sample was 101 and ranged from 58 to 245 and for tumours the median was 38 and ranged from 1 to 231 (Campbell et al., 2008b; Stephens et al., 2009).

Past studies have shown that translocations can occur between spatially proximal areas of the genome (Roix et al., 2003; Osborne et al., 2007) so it follows that the majority of rearrangements should be small and intrachromosomal. This is indeed what Stephens et al. (2009) observed; 85% of rearrangements were within the same chromosome and less than 2Mb apart. Approaches such as SKY and array CGH probably would not have been able to identify them as many were balanced and most were below the resolution of this technique.

Many of the Stephens et al. (2009) rearrangements fell within genes, many fusion genes were predicted and several were expressed. An important observation is that breast cancers can express several fused genes. The study described 21 potentially functional novel fusion genes, most of unknown function but several within known cancer genes such as *ETV6* and *EHF*, although none were shown to be recurrent.

Given the small size of many of the rearrangements, many fell entirely within genes and in some cases this affected the exon structure of the transcript. Novel isoforms resulting from rearrangements were detected for oncogenes such as *RUNX1* but also in well-characterised tumour suppressor genes such as *RB1,* APC and *FBXW7*. Therefore, it is possible oncogenic activation or tumour suppressive loss of function was achieved by structural rearrangement of the open reading frames of these genes. It is interesting to

consider that Sanger sequencing studies would not have detected any mutations as the coding exons were all intact. It may be possible, therefore, that genes such as *APC* and *RUNX1*, considered unimportant in breast cancer, may, in fact, be relevant to breast cancer (Newman and Edwards, 2010; Edwards, 2010).

1.4. Questions for post-genome cancer research

In 2008-2010 we have seen first fully sequenced cancer genomes from breast cancer, chronic myeloid leukaemia, lung cancer and malignant melanoma (Ley et al., 2008; Shah et al., 2009; Ding et al., 2010; Pleasance et al., 2010b, 2010a; Lee et al., 2010; Mardis et al., 2009) and these studies have rewritten our perception of the number of mutations a cancer genome contains. It is now clear that tumour genomes accumulate thousands of mutations, a hundred or so of which are in the coding exons of genes. But in addition to coding mutations, structural changes to the genome can contribute approximately as many mutations as changes to the coding sequence (Hampton et al. 2008; Stephens et al. 2009). In any given tumour type there are hundreds of infrequently mutated genes and only a few frequently mutated ones. This results in large inter-tumour genetic heterogeneity but additional complexity exists as tumours themselves contain many competing sub-populations resulting in intra-tumour heterogeneity also (Anderson and Matsuno, 2006; Cooke et al., 2010b, 2010a; Navin et al., 2010).

The way ahead for many in the field seems clear: to define more cancer genomes at the sequence and structural levels as “[l]arge sample sets will have to be analysed to distinguish infrequently mutated cancer genes from genes with random clusters of passenger mutations” (Stratton et al., 2009, p. 721). With this in mind the international cancer genome consortium intends to sequence 500 genomes from each of the common cancers within the next few years (Hudson et al., 2010). But now that we might finally know what cancer genomes look like, the next question is how do we decide which of the many mutations are important?

1.4.1. What types of mutations are needed to cause cancer?

Modern thinking of cancer biology centres around biological pathways. Pathways turn signals into cellular responses, for example, in *Drosophila*, secreted wingless (*wnt*) signalling molecules cause specific cells to divide and differentiate into wing halteres (Sharma and Chopra, 1976). Cellular responses, in general, work through pathways so it is reasonable to assume that each of the hallmarks of cancer are achieved by inactivation, or alternatively, inappropriate activation of one or more biological pathway. Some have gone as far to say that all the complexity observed in tumour genomes affects no more than 20 biological pathways (Wood et al., 2007). The canonical *wnt* pathway is used here as an example as it is conserved between species, controls many events surrounding morphology, proliferation, motility and cell fate in embryogenesis and its aberrant signalling has been observed in several human cancers (Klaus and Birchmeier, 2008).

Secreted *Wnt* proteins bind extracellular domains of *Frizzled* family of receptors, this causes activation of *Dishevelled* (*Dsh*). When *DSH* is activated it can inhibit a second protein complex that includes *axin*, *glycogen synthase kinase-3* (*GSK-3*), and *adenomatous polyposis coli* (*APC*). The function of the axin/GSK-3/APC complex is to promote proteolytic degradation of another intracellular signalling molecule, *β-catenin*. Hence, when the complex is inhibited, cytoplasmic *β-catenin* is stabilised. Upon stabilisation, *β-catenin* enters the nucleus and binds *TCF/LEF* family transcription factors and promotes expression of specific genes linked with cellular proliferation (Polakis, 2000).

Disregulation of a pro-growth signalling pathway is a hallmark of cancer so we would expect to see mutations in members of this pathway in cancer, and indeed we do: Over-production of *Wnt-1*, the secreted signalling molecule, causes mammary tumours in the mouse (Nusse and Varmus, 1982) Loss of function of *APC*, one of the most famous examples of a tumour suppressor genes, is well described in colon cancer (Polakis, 1997). Activating mutations in *β-catenin* can also be observed in colon cancer as well as melanoma (Morin et al., 1997). Mutations in the *AXIN1* gene have been reported in human hepatocellular carcinomas (Satoh et al., 2000).

We can see that mutations within the same pathway can be oncogenic gains of function (*WNT1, β-catenin*) or tumour suppressive losses of function (*APC, axin*) but all have the eventual effect of dis-regulating the pathway, increasing net proliferation. Thus a series of isolated mutations can tell a common story.

Pathway-centred analysis of cancer mutations is beginning to bear fruit. For example, 12 core processes or pathways appear to be deregulated in pancreatic tumours, but each by slightly different mechanisms (Jones et al., 2008c). In breast cancer, Wood et al. (2007) observed several mutations in *PIK3CA*, a known oncogene in breast cancer. Not all tumours had PIK3CA mutations but some had mutations in interacting and related genes *GAB1, IKBKB, IRS4, NFKB1, NFKBIA, NFKBIE, PIK3R1, PIK3R4,* and *RPS6KA3.* This begins to implicate the PI3K pathway in general as well as links with *nuclear factor kappa B (NF-κB*) signalling in breast tumorigenesis.

The challenge we now face is predicting the effect, if any, a newly-discovered mutation has on a pathway, when the pathway or molecule in question is poorly characterised.

1.4.2. How many mutations are required for cancer to develop?

One of the largest unanswered questions in cell biology is, how many mutations are required to cause cancer? Many attempts have been made to answer this crucial question but one of the earliest models, suggesting six to seven rate limiting events, has remained the pre-genomic era's typical estimate for the number of necessary mutations in cancer (Armitage and Doll, 1954, 1957). Various mathematical models arguing for and against higher and lower numbers of mutations have been proposed (Tomlinson et al., 1996, 2002; Rajagopalan et al., 2003) but virtually all of these estimates come from the pre-genomic era. Until recently, sufficient data did not exist to allow people to address this crucial question based on observation rather than theory and extrapolation. Recent analyses based around the breast cancer 'mutatomes' have suggested much higher, even as many as fifty, selected mutations in epithelial cancer genomes (Beerenwinkel et al., 2007; Teschendorff and Caldas, 2009).

1.4.2.1. Drivers versus Passengers

Central to this question is the drivers versus passengers problem (Stephens et al., 2005; Greenman et al., 2007): As our knowledge of cancer genomes increases, distinguishing selected 'driving' mutations from large numbers of unselected ‘passenger’ mutations is becoming a major challenge. A driver mutation has conferred growth advantage on the cancer cell, so has been positively selected. A passenger mutation has not been selected: it has offered no growth advantage but any non-functional mutation that occurs in a cell with driver mutations will be carried along by cell division and reach fixation in the population, just as a driver would.

Greenman et al. (2007) made one of the first attempts to address this problem experimentally. The authors assumed that to be a driver mutation, there must be an effect on protein structure, and therefore function, whereas synonymous mutations, as they do not alter protein structure, cannot be selected. By sequencing 518 kinase genes in 210 cancers including breast, lung, colorectal, gastric, testis, ovarian, renal, melanoma, glioma and acute lymphoblastic leukaemia Greenman et al. (2007) were able to compare the relative proportions of synonymous and non-synonomous mutations.

There appeared to be an excess of non-synonymous mutations compared with the expected number given the background synonymous mutation rate. Over one thousand somatic mutations were detected, 921 of which were single base substitutions. Of these substitutions, Greenman et al. (2007) estimated 763 (95% confidence interval, 675–858) were passenger mutations. This left an estimated 158 driver mutations (95% confidence interval, 63–246). This equates to less than one driving kinase mutation per sample, which is not particularly surprising. But nevertheless, these data suggested that the number of driving mutations may be somewhat higher than traditional estimates. If we assume that the kinome is a reasonable model of how genes, in general, mutate within a cancer genome (Greenman et al. corrected for highly mutated genes such as *KRAS*) we might reasonably conclude that between 7 and 27% of *all* point mutations are likely to be driving events.

The first estimates for the number of driving mutations based on unbiased screening in breast cancer were attempted by Wood et al. (2007). Whole exome mutation screening of 11 breast cancers revealed that breast tumours accumulate, on average, 90 mutant genes. To estimate the probability that any given mutation was a passenger, the authors first established the background mutation rate in the genome from previously published data. They then factored in gene size and varying frequencies of different base substitutions. The authors were then able to estimate if a given gene was mutated more than chance would predict. The resulting candidate cancer (CAN) genes had more than 90% probability of having undergone a selected mutation. From 22 tumours, 11 each from breast and colon tumours, they identified 280 CAN-genes. In the average breast tumour, there were 14 CAN genes and this number can be equated to the lower estimate of drivers in that tumour. These estimates did not consider non-coding regions, RNA transcripts (including micro RNAs) or structural variation, so the true number of driving mutations could be considerably higher.

It may be reasonable to expect that driving cancer genes would be mutated in a high proportion of tumours and unselected passengers to be mutated at much lower frequency. Certain cancers appear to be “addicted” to certain oncogenes (Jonkers and Berns, 2004), for example, CML and *ABL1* or pancreatic cancer and *KRAS*. But, as the above studies show, the ‘genomic landscape’ of breast cancer contains only a small number of frequently mutated genes: 50% have mutations in *TP53*, 30% have mutations in *PI3KCA*, 20% have mutations in *CDH1* (Teschendorff and Caldas, 2009; Forbes et al., 2010). Hundreds of other genes are found mutated in much lower numbers and some of these must also be driving mutations.

1.4.2.2. Driving mutations caused by chromosome aberrations

It is probable that structural variation in the breast cancer genome is analogous to the sequence-level mutational landscape of breast cancer. Recurrent features such as *ERBB2* amplification are observed with high frequency but many hundreds of chromosome breakpoints occur at much lower frequency (Chin et al., 2007).

We also know that a typical breast cancer can express multiple fusion genes, but we do not know how many are driver events. Stephens et al. (2009) concluded most gene-fusions were not selected events. As the mechanisms that form chromosome translocations are thought to be random, the authors calculated that 2% of rearrangements would have generated an in-frame fusion gene by chance compared with 1.6% of predicted fusion genes in their data. Even if this observation is true, it is very different from saying gene fusions *do not* contribute any driving mutations in breast cancer. As we have seen above, even rare mutations can contribute driving events in cancer.

But interestingly, if one compares the number of expressed in-frame fusion genes to the number of expressed out-of-frame fusion genes, the ratio is approximately 1:1. This is somewhat different from the 1:3 ratio we would expect by chance and implies a high proportion of in-frame gene fusions in breast cancer are selected events.

1.4.3. How should we deal with intra-tumour heterogeneity?

As our capacity to identify mutations, both structural and sequence-level, increases the drivers versus passengers problem becomes ever more apparent. This task is made more challenging by the intra-tumour heterogeneity at the sequence-level as well as the structural-level evident in many tumours (Jones et al., 2008b; Attard et al., 2009).

Currently, one of the best strategies for identifying the important mutations in a heterogeneous cancer is based around comparative lesion sequencing (Jones et al., 2008b; Shah et al., 2009). These studies have looked for mutations in primary tumours and their associated metastases. From this one can identify three classes of mutation: those common to the primary tumour and metastasis and those private to one or the other. The common mutations are clearly interesting as these would contain all the early events in tumour evolution. Some may even argue that all the events necessary for metastasis are in this group as well (Weiss et al., 1983; Bernards and Weinberg, 2002; Edwards, 2002). Private events in the primary tumour can be disregarded as they probably represent the heterogeneity of clonal sidelines. Private events in the metastasis are more interesting as, where applicable, they may contain a mutation that allowed drug-resistant relapse. But

interestingly, the vast majority of *apparently* private mutations found within metastases are also found in the primary tumour, but often at a much lower frequency (Shah et al. 2009).

A second strategy for circumventing heterogeneity is to study cell lines. Historically, breast cancer cell lines have been used as models as it is difficult to obtain primary tumours and perform, for example, cytogenetic studies on them. Cell lines are considered much less heterogeneous than primary tumours as they presumably represent the outgrowth of a single ancestral cell in culture. Genome-wide screens have consistently identified higher numbers of mutations in cell lines than primary tumours (Wood et al., 2007; Stephens et al., 2009) and this is probably because finding low-frequency mutations in a heterogeneous environment is more difficult.

Alternatively, it could represent evolution in culture, but some studies have indicated this source of mutation probably does not add a large number of mutations to the cell line genomes (Neve et al., 2006; Jones et al., 2008d). For example, Neve et al. (2006) compared early and late passage breast cancer cell lines and concluded they had not accumulated substantial new aberrations during culture. The authors went on to show that, broadly speaking, a panel of 51 breast cancer cell lines showed genomic gains and losses and transcriptional profiles similar to those found in a panel of breast tumours. Thus, cell lines present, in many cases, a relatively homogeneous view of late stage tumours. While undoubtedly their genomes contain many passenger events, these would only be the passenger events in the history of a single lineage rather than the sum total of passenger mutations of the multiple clones of a primary tumour.

1.4.4. What is the role of chromosome instability?

The epithelial cancers often have highly rearranged genomes but a major unknown is how this state of chromosome instability (CIN) contributes to carcinogenesis[1]. If, for example, a

1CIN has two interchangeable meanings in the literature. Some take it to mean an acquired acceleration in the rate of chromosome aberrations but others only mean that it describes the observed state of a rearranged genome. Hereafter, I refer to an acceleration in the rate of chromosome rearrangements as acquired CIN and the observed state of a rearranged genome as CIN alone.

breast cancer expresses five fusion genes (Hampton et al., 2008), then how many are likely to be selected events as opposed to random or late passenger events? The question for epithelial cancers is, then, what proportion of chromosome changes are driving events? Essentially it is the drivers versus passengers problem restated. If upwards of 14 driving events come from somatic change at the sequence-level (Wood et al., 2007), is it possible to say a similar number come from changes in genome structure, given that chromosome changes probably disrupt as many genes as sequence changes? In order to answer these questions we must first speculate on the state of CIN, its timing and whether it is an acquired characteristic.

1.4.4.1. The State of CIN

We know that chromosome rearrangements can inactivate tumour-suppressor genes and activate proto-oncogenes. When a single cytogenetic aberration is observed in a leukaemia, for example, it is usually clear that chromosome instability has contributed a driving mutation to that particular cancer. Most recurrent clonal rearrangements in leukaemia are likely to be early events in cancer development, as they are usually the sole cytogenetic abnormality in a cell. This view is epitomised by Ford et al. (1998) who showed that a *TEL-AML* fusion transcript and its specific genomic junction was present in monozygotic twins who both developed leukaemia. This means the translocation probably happened early in development *in utero* (Ford et al., 1998).

In contrast to these early “primary” events, late stage and relapsed leukaemias sometimes show “secondary” translocations. These later events are not seen with any degree of recurrence between cases and are never the sole abnormality in a karyotype (Johansson et al., 1996). It is probable, therefore, that secondary translocations are nearly all passenger events. In the case of breast cancer, where most tumours do not exhibit many recurrent chromosome aberrations that we know of, it is tempting to conclude that we are only observing many late, and therefore, secondary events (Johansson et al., 1996).

However, it is not clear if the primary/secondary classification of rearrangements can be applied to heterogeneous epithelial cancer genomes in such a regimented way. There are

probably multiple primary and secondary chromosomal events in breast cancer. For example, in individual breast cancers we can often see amplified *ERBB2*, loss of 8p and loss of 17p – three events which clearly contribute to carcinogenesis – in the same tumour (Chin et al., 2007).

More evidence in support of this view has come from studies on the intra tumour heterogeneity of epithelial cancer cell line karyotypes in culture. Muleris and Dutrillaux (1996) showed the rate of unstable rearrangement (rearranged chromosomes not transmitted to the next generation) was approximately the same in all colorectal cancer cell lines. But for one subtype of cell lines, termed ‘monosomic’ - as they tended to lose chromosomes - the number of stable rearranged chromosomes in karyotypes was much higher (Muleris and Dutrillaux, 1996). Roschke et al. (2002) observed that rearrangements in the NCI60 panel of cell lines tended to be peripheral to a 'core' karyotype and more normal chromosomes that were gained relative to the rearranged chromosomes (Roschke et al., 2002, 2003). Taken together, these studies imply that, if the majority of chromosome translocations were secondary, and therefore passenger events, they would be lost at random from the aneuploid genomes of epithelial cancers. This does not seem to be the case.

1.4.4.2. The timing of CIN

If chromosome rearrangement starts late, then it is possible that all the driving mutations a that cancer requires preceded it. This would imply that most chromosome rearrangements are secondary events. Of course, if CIN is an acquired phenotype it probably starts early and the opposite possibility might be true.

The classical model of the genetic progression to cancer comes from studies on benign adenomas and more advanced carcinomas of the large intestine. By comparing mutations and LOH of known cancer genes in large numbers of samples at each stage one can reconstruct a progression from adenoma to carcinoma driven by specific genetic changes. The model starts with loss of *APC* or activation of *β-cateinin* and proceeds with activation of *KRAS*, loss of *DCC/SMAD4/SMAD2* and loss of *p53* (Cho and Vogelstein, 1992; Baker

et al., 1989, 1990).

Within this progression the authors proposed the onset of chromosome instability was relatively late, coinciding more-or-less with loss of *TP53*. And there is now some *in vivo* evidence that *KRAS* mutant tumours need to lose *TP53* for chromosome instability to happen (Hingorani et al., 2005). This has lead to the view that most useful sequence-level mutations must precede the onset of CIN (Vogelstein and Kinzler, 2004).

As *TP53* gene mutations were relatively rare in adenomas but relatively common in carcinomas, they were thought to occur at the transition from benign to malignant growth (Baker et al., 1990). This study was not, however, based on comparative lesion sequencing but rather a pool of adenomas was compared to an unrelated pool of carcinomas. As the majority of adenomas do not progress to carcinoma an alternative explanation is that only the adenomas with *TP53* mutations can progress to become carcinomas. If this is the case we would expect to see chromosome instability much earlier during the progression to cancer.

There is evidence that the karyotypes of some colorectal adenomas, even at an early stage, display aneusomy and structural rearrangement (Bomme et al., 1998). In benign breast lesions including fibrocystic lesions from women with and without a known hereditary predisposition to breast cancer, fibroadenomas, phyllodes tumors, and papillomas, karyotypes often show rearrangement but of a lesser extent than is often present in breast carcinoma. Commonly described changes in breast cancer such as gain of 1q, interstitial deletion of 3p, and trisomies 7, 18, and 20 can also be observed. Interestingly, the frequency of chromosome abnormalities seems to correspond with risk of developing invasive mammary carcinoma (Lundin and Mertens, 1998). It is reasonable, then, to think that chromosome rearrangements can start early in epithelial cancers as chromosome rearrangements can be seen in precursor lesions (Fiche et al., 2000; Ottesen et al., 2000; Cerveira et al., 2006).

Beyond benign stages we can see a stepwise progression of karyotypic complexity in malignant tumours. It appears that the number of chromosome rearrangements in cancers

including breast increase with tumour grade (Magdelenat et al., 1992). Relapses always have some of the karyotypic features of primary tumours but often show additional rearrangements (Cooke et al., 2010). In breast cancer, there also appears to be a correlation between karyotypic complexity and ER and PR status, ER-negative and PR-negative tumours having more complex karytotypes. If one ascribes to the view that loss of oestrogen receptor expression is one of the final stages in the evolution of an ER-positive tumour, this serves as further evidence for stepwise acquisition of chromosome abnormalities (Magdelenat et al., 1992).

1.4.4.3. The Acquisition of CIN

Inherited diseases such as Ataxia Telangiectasia, Bloom syndrome, Fanconi Anaemia and Nijmegen Breakage Syndrome predispose individuals to various early onset cancers. Investigations into the mechanisms of these diseases have invariably led to DNA repair and spindle checkpoint defects such as in the *FANC* family of genes, *ATM* and *ATR* (Taylor, 2001). As individuals show increased chromosomal instability and breakage, can we conclude that acquired CIN is a driving force behind accelerated cancer development in these individuals? Furthermore, as sporadic cancers have highly rearranged genomes, is it possible to conclude that an accelerated rate of genome rearrangement contributes to these cancers also?

A classic illustration of the need for a “mutator phenotype” in familial cancers comes from the study of Familial Adenomatous Polyposis (FAP) and Hereditary Non-Polyposis Colorectal Cancer (HNPCC). In FAP, loss of function of the *APC* gene deregulates *β-cateinin*-mediated gene expression leading to increased proliferation (Nishisho et al., 1991). In HNPCC, mutations in mismatch repair genes such as *MSH1* and *MLH2* cause small repeat elements to expand and disrupt gene function (Bodmer, 2006). A prominent feature of FAP is chromosome rearrangement and numerical abnormality but HNPCC karyotypes appear to be more normal (Kinzler et al., 1991; Nishisho et al., 1991; Nowak et al., 2002). There is evidence to suggest a degree of mutual exclusivity in these mechanisms and this strengthens the debate in favour of an acquired mutator phenotype in familial cancer predisposition syndromes (Komarova et al., 2002).

The best evidence for the contribution of acquired CIN to breast cancer comes from hereditary cancer syndromes associated with mutations in the *BRCA1* or *BRCA2* genes (breast cancer early onset 1and 2). Familial mutations in these genes carry a highly penetrant risk of early onset cancers of the breast and ovary (Miki et al., 1994; Wooster et al., 1995). A single mutant allele passed through the germline predisposes the individual to cancer, but cells within the resultant tumour undergo somatic loss of heterozygosity usually due to a chromosomal rearrangement (Venkitaraman, 2007). *BRCA1* and *BRCA2* are tumour suppressor genes which help maintain genomic integrity through roles within the homologous recombination pathway (Chen et al., 1999)

In sporadic breast cancer, many genomes are rearranged in a way that appears comparable to the early-onset familial cancers (Grigorova et al., 2004) If CIN were an early, acquired, phenotype in sporadic cancers too, we could suppose that it contributed substantially to early cancer development as has been suggested previously (Komarova et al., 2002; Nowak et al., 2002; Rajagopalan et al., 2003).

Only anecdotal evidence for acquired CIN in sporadic breast cancer exists currently. For example, by looking at CGH profiles it is clear that there are several different types of tumour profiles: some are relatively 'quiet' some are highly rearranged, some are mostly tetraploid, some are mostly triploid, some have large regions of loss of herterozygosity, some do not and some have a high number of small tandem duplications (Fridlyand et al., 2006; Chin et al., 2007; Stephens et al., 2009; Bignell et al., 2010). This suggests there are a large number of evolutionary routes a genome can take towards cancer, thus, the genomic landscapes we observe are likely to be “a composite of selection and particular failures in genome surveillance mechanism(s).”(Fridlyand et al. 2006).

1.5. Techniques used and discussed in this thesis

1.5.1. Florescence *in situ* Hybridization (FISH)

Fluorescence *in situ* hybridization (FISH) is a standard molecular cytogenetic technique and is useful in defining the position of chromosome breaks. This technique, just like many

of the ones below described below relies on the propensity of denatured DNA to re-anneal with its complementary DNA sequence. In FISH, cellular DNA is denatured with heat or formamide and then left to re-anneal in the presence of a large excess of labeled probe DNA. The denatured chromosomal DNA then anneals with probe DNA. Excess probe is washed away leaving only hybridized probe to give a fluorescent signal *in situ*.

Usually, approximately 100kb fragments of the human genome, contained within bacterial artificial chromosomes (BACs), are used for FISH experiments. These BACs come from libraries developed to sequence the human genome, so a BAC representing virtually any region of the human genome is available. If BACs are used, a breakpoint can be defined to within the length of that BAC as the hybridization signal will “split”. As the average BAC is around 100kb, this is a relatively low-resolution way of defining breakpoints. Chromosome painting is an extension of FISH but instead of labelled BAC DNA being used as a probe, the DNA of an entire chromosome is used.

1.5.2. Spectral Karyotyping

Spectral Karyotyping (SKY) and a related technique multicolour FISH (M-FISH), is a method to simultaneously paint and visualize all chromosomes. This is especially useful when one is dealing with highly rearranged karyotypes. Most often, FISH experiments are done with three colours, flourescein (or some derivative), Cy-3 and Cy-5 as these dyes have sufficiently different emission spectra to be detected separately. Theoretically, one can use any combination of fluorophores for FISH so long as you are able to excite and detect at the appropriate wavelengths. SKY uses combinatorial labelling of flow sorted chromosomes to achieve this simultaneous visualization. SKY and M-FISH systems use seven different haptens in different combinations. For example Chromosome 1 might be labelled with hapten A, chromosome 2 with hapten B, chromosome 3 is labelled with haptens A and B, chromosome 4 with A and C etc. In M-FISH each fluorophore is excited and imaged separately. In SKY, all fluorophores are excited, producing a unique spectral profile, and imaged simultaneously. In each case, computer software assigns a pseudo-colour to each chromosome or part of a chromosome based in its unique combination of fluorescence (Speicher et al., 1996; Speicher and Carter, 2005)

1.5.3. Flow sorting of Chromosomes

Fluorescence Activated Cell Sorting (FACS) can define and separate populations of cells within a sample, but can also be used to separate chromosomes. A large number of cells are arrested in metaphase using a microtubule inhibitor such as colcemid. Condensed mitotic chromosomes are then separated from cell nuclei and other debris by centrifugation. This chromosome suspension is labelled with two DNA-binding dyes, Chromomycin A3 binds to G/C and Hoescht 33258 binds to A/T regions. When passed through the FACS machine's lasers, chromosomes fluoresce at an intensity proportional to their AT and GC content. This fluorescence intensity ratio can be plotted and a given population of chromosomes can then be gated by the machine and collected (Telenius et al., 1992). Performing this process with metaphase chromosomes from a normal sample provides us with the raw material to make chromosome paints and repeating the process with tumour cell line chromosomes allows us to investigate them using reverse painting and array painting (Arkesteijn et al., 1999; Fiegler et al., 2003).

1.5.4. Array CGH

Array CGH is commonly used to investigate gains and losses of genomic regions. Arrays made from genomic BAC clones have achieved resolutions relative to the size of DNA contigs used. For example Pole et al. (2006) used CGH to investigate regional deletions of chromosome 8 in breast cancer. They achieved 1 Mb resolution over chromosome 8 and used a tiling path of over 8p12 to achieve a resolution of 0.2 Mb. An 8p12 fosmid array (Pole et al., 2006; Cooke et al., 2008) achieved a 0.04Mb resolution. Small regions of the genome can be investigated to kilobase resolution using custom oligonucleotide arrays as used by Howarth et al (2008).

High-resolution genome-wide copy number analysis can be achieved using single-nucleotide polymorphism (SNP) arrays for CGH. These arrays were originally intended to detect SNP genotypes over the whole genome, but can also be used to assess DNA copy number changes (Bignell et al., 2004). The Affymetrix SNP 6.0 platform has approximately

500,000 SNP-specific oligos distributed over the whole genome. As one probe is found approximately every 6kb, regions of gain or loss can often be identified to the level of the exon when genes are broken. In addition, SNP arrays provide both copy number information as well as genotype status. This is useful in identifying regions of loss of heterozgosity and may also unravel the series of events that formed complex karyotypes.

1.5.5. Array Segmentation Algorithms

Several bioinformatic 'segmentation' algorithms exist to find copy number change points in array CGH data. Statistical methods such circular binary segmentation and hidden Markov models (Marioni et al., 2006; Venkatraman and Olshen, 2007) have been employed to define change points from the florescence intensity data of CGH probes, but no previous algorithm as factored in the SNP6 array's capacity to differentiate between SNP alleles. The PICNIC (Predicting Integral Copy Numbers In Cancer) algorithm (Greenman et al., 2010) identifies absolute copy number of *each allele* in any given region of genome. SNP combinations such as AA, AB and BB occur in diploid regions in a 1:1 ratio, while in triploid regions AAA, BBB, AAB, ABB regions are apparent by a 2:1 allele ratio and in tetraploid regions AABB (2:2), and AAAB and BBBA (3:1) regions are visible.

1.5.6. Array Painting

In array painting, flow sorted chromosomes are hybridized to microarrays allowing rapid identification of chromosome breakpoints (Fiegler et al., 2003; Howarth et al., 2008). For example, HCC1187 contains a derivative chromosome formed from a translocation of chromosomes 1 and 8, der(8)t(1;8). The derivative chromosome is flow sorted, amplified by degenerate oligo primed PCR (Telenius et al., 1992), labelled and then hybridized to a chromosome 1 tiling path array. The array will show hybridization up to the point at which chromosome 1 is broken. Traditional array comparative genomic hybridization cannot detect balanced chromosome aberrations as there is no net gain or loss of material, but as only individual flow-sorted chromosomes are hybridized in array painting, even balanced breaks can be detected, provided they are interchomosomal

1.5.7. Massively Parallel Paired End Sequencing

These third-generation sequencing approaches produce millions of sequencing reads in a single experiment (Metzker, 2010). Typically, these sequences have been short, around 37 base pairs, but longer reads are now possible. An important part of the sequencing process is the alignment of millions of short reads to the reference genome. In the past researchers have used bioinformatic tools such as BLAST and FASTA for high accuracy alignment but these methods are too computationally intensive for millions of separate queries. Custom algorithms such as MAQ (Mapping and Assembly with Quantiles) and BWA (Burrows-Wheeler Alignment) have been written specifically to address this trade off between accuracy and speed but as a result the capacity to incorporate mismatches into the alignment has diminished (Li et al., 2008; Li and Durbin, 2009).

Massively parallel sequencing experiments represent a trade off between the amount of sequence generated and cost (Bashir et al., 2008). Sampling any given region is a stochastic process so the probability of a sequencing read crossing a translocation, for example, increases with the amount of sequence data generated .e.g. Stephens et al. (2009) estimated that their approach detected, on average, 50% of the rearrangements in their 24 samples. The probability of finding a given proportion of rearrangements with a given amount of sequence data can be described by the Poisson distribution where y=number of events, λ = mean number of events:

$$\mathbf{P(Y=y) = (\lambda^{y} e^{-\lambda})/y!}$$

For example, if there are approximately 3 billion bases in the haploid genome and if a sequencing experiment generates 810 million single 37 base pair reads, this translates to 30 billion base pairs of sequence. This is equivalent to 10 fold coverage of the haploid genome. Since the genome coverage is the average number of times each base pair is hit, λ = 10, and Y is the number of hits we are looking for, for example two, we can calculate the proportion of events that will be sampled twice in the experiment:

$$\mathbf{P(Y=2) = (10^{2} e^{-10})/2!}$$

= 0.00227

So in this example, 0.227% of events will be sampled in the experiment twice. In sequencing experiments, however we want the number of events hit twice or more which is:

P(2 or more hits) = 1- (chance 1 hit) - (chance of zero hits)

= 1 – 0.00045 – 0.00005

= 0.9995

So in this experiment, 99.95% of events will, in theory, be hit twice or more.

Sequencing coverage can be described in two ways: sequence depth and physical coverage. Depth is measured by the average number of times an individual locus is sampled. For the draft human genome sequence, this figure was 10 fold but was based around relatively accurate Sanger sequencing data (Lander et al., 2001). For high-throughput sequencing approaches, the reads are typically shorter and the confidence in each base call is less (Meyerson et al., 2010). Recent sequencing projects have combated this by increasing the sequence depth to around forty-fold (Ley et al., 2008; Ding et al., 2010; Shah et al., 2009; Pleasance et al., 2010b, 2010a). This figure is based around haploid genome coverage, so for polyploid cancer genomes the true figure is less. This 'deep sequencing' approach is currently one of the fastest ways to find sequence-level mutations in cancer genomes but it is currently quite expensive (Meyerson et al., 2010).

Another popular use of next generation sequencing is to generate “paired end reads”. The paired end strategy has been used to detect rearrangements in several cancer genomes to date and has proven an effective strategy to find structural rearrangements (Volik et al., 2003; Campbell et al., 2008b; Stephens et al., 2009). In this approach, genomic DNA is fragmented into pieces of a known size range and sequencing reads are generated from both ends of the fragment. Each read is then aligned to the reference genome just as for the above single reads but it is the spatial relationship of one read to its partner that can indicate a structural variation. Most of the reads, when aligned back to the genome, are in

the correct orientation and at the expected distance from their partners. Structural variants are apparent when pairs of sequences map to different chromosomes (translocation or insertion), too far apart (most likely a deletion), too close together (small insertion) or the wrong orientation (inversion) or the wrong genome order (tandem duplication).

As the read pairs are physically linked (for example 500 base pairs apart) much higher coverage of the genome can be achieved with less sequencing. Recent, paired-end sequencing experiments have generated approximately 50 million paired reads (Campbell et al., 2008b; Stephens et al., 2009), this translates to 0.61-fold haploid sequence coverage, but if one considers physical coverage this figure increases to 8.4 fold meaning 99% of events would be hit twice or more. The distance between the paired reads can be extended to around 3kb by employing the 'mate pair' strategy. Here genome fragments are circularised, the joined region cut out and the ends of each fragment sequenced just as for regular paired end sequencing (Shendure et al., 2005). Mate pair sequencing was not used as part of this thesis but, as, in theory, this strategy can produce much higher physical coverage than short fragment end sequencing it is currently being investigated by other lab members.

1.7. The purpose of this thesis

1.7.1. Aim 1: Map chromosome rearrangements in breast cancer

Breast cancer genomes are among the most complex of the common cancers displaying extensive structural and numerical chromosome aberration. Relative to sequence-level mutations, little is known about the genes affected by chromosome changes in breast cancer. In this thesis, I define breast cancer genome structure with a view to answering two questions.

1) How many genes are disrupted by chromosome aberrations in a “typical” breast cancer?
2) Fusion genes are an important feature of several cancers. Do chromosome aberrations fuse any genes in breast cancer?

1.7.2. Aim 2: Investigate the relative timing of point mutations and chromosome aberrations

A major unknown is the relative importance and timing of genome rearrangements compared to sequence-level mutation. For example, chromosome instability might arise early and be essential to tumour suppressor loss, or alternatively, be a late event contributing little to cancer development.

By taking an evolutionary view of individual cancer genomes I address this question as “although complex and potentially cryptic to decipher, the catalogue of somatic mutations present in a cancer cell therefore represents a cumulative archaeological record of all the mutational processes the cancer cell has experienced throughout the lifetime of the patient” (Stratton et al., 2009, p.720). By looking at the structure and sequence of breast cancer genomes together it is possible to speculate on the relative timing of mutations within individual tumours.

Chapter 2

Materials and Methods

2.1. Reagents, Manufacturers and Suppliers

Reagent	Manufacturer/Supplier
BAC and fosmid clones	BACPAC CHORI, Oakland, USA
Biotin dUTP	Roche Diagnostics, Basel, Switzerland
BigDye Terminator v3.1 cycle sequencing kit	Applied Biosystems Ltd. Foster City, CA
Biotinylated anti-streptavidin	Vector Laboratories Inc., Burlingame, CA, USA
Chloramphenicol	Sigma-Aldrich, Dorset, UK
Colcemid	Sigma-Aldrich, Dorset, UK
Cryotubes	Fisher Scientific, Loughborough, UK
Cy3-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
DNAse I	Sigma-Aldrich, Dorset, UK
DNAzol reagent	Invitrogen, Paisley, UK
dNTPs	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
FITC-16dUTP	Roche Diagnostics, Basel, Switzerland
Formamide	VWR International, Lutterworth, UK
G50 MicroSpin columns	GE Healthcare, Buckinghamshire, UK
GenomiPhi Kit	GE Healthcare, Buckinghamshire, UK
GoGreen PCR master mix	Fermentas Life Sciences, York, UK
HiSpeed Plasmid Midi-Prep Kit	Qiagen UK, Crawley, UK
HotMaster Taq	VWR International, Lutterworth, UK
Hyperladder I	Bioline, London, UK
Human C0t-1 DNA	Invitrogen
Isopropanol	Invitrogen, Paisley, UK
Kanamycin	Sigma-Aldrich, Dorset, UK
LB agar	Hutchison/MRC Centre Media Unit
LB broth	Hutchison/MRC Centre Media Unit
MCBD-201	GIBCO Technologies, Invitrogen, Paisley, UK
Mixed bed resin beads	Sigma-Aldrich, Dorset, UK
Megabace formamide sequencing buffer	Applied Biosystems Ltd. Foster City, CA
Na_2HPO_4	VWR International, Lutterworth, UK
$NaHPO_4$	VWR International, Lutterworth, UK

NanoDrop spectrophotometer	Labtech International, Ringmer, UK
Nucleofast 96 PCR cleanup kit	Clonetech, Mountain View, CA
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
Propidium iodide	Invitrogen, Paisley, UK
QIAquick PCR Purification Kit	Qiagen UK, Crawley, UK
PolyPrep poly-L-lysine coated slide	Invitrogen, Paisley, UK
PyroMark Gold reagents	Biotage
RNAseIN	Promega, Fitchburg, USA
RPMI-1640	GIBCO Technologies, Invitrogen, Paisley, UK
S.N.A.P UV-Free Gel Purification Kit	Invitrogen, Paisley, UK
Sodium acetate	Hutchison/MRC Centre Media Unit
Spectrum Orange dUTP	Vysis UK Ltd/Abbott Laboratories, Downers Grove IL, USA
Spermidine	Invitrogen, Paisley, UK
Spermine	Invitrogen, Paisley, UK
Spin Miniprep kit	Qiagen UK, Crawley, UK
SSC	Hutchison/MRC Centre Media Unit
Steptavidin-sepherose beads	GE Healthcare
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
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

Table 2.1. Reagent manufacturers and suppliers

2.2. Common Solutions

Name	Constituents
20X SSC	3M NaCl, 0.3 M trisodium citrate, pH 7.0
1X PBS	140 mM NaCl, 2.5 mM KCl, 10 mM Na_2HPO_4, 1.75 mM KH_2PO_4, pH 7.4
TE	10mM Tris-HCl, 1mM EDTA, pH 8
Versene	140mM NaCl, 2.6 mM KCl, 9 mM Na_2HPO_4, 1.5mM KH_2PO_4, 600µM EDTA, 20mM hepes, 0.015% (v/v) phenol red, pH 7.5

Luria-Bertani (LB) broth	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% NaCl(w/v), pH 7.0.
Luria-Bertani (LB) agar	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% NaCl(w/v), 15 g/L agar, pH 7.0.

Table 2.2. Commonly used solutions

2.3. Cell Lines and Culture

A panel of cell lines was analysed as part of this thesis. Several were cultured by myself (Table 2.3) and several by current and past lab members. Some lines, for example AU565, were not grown in the lab and only SNP6 data generated by the Wellcome Sanger Centre was available (see section 2.7.1 for details).

Cell Line	Media	Source	Reference
BT-474	RPMI 1640, 10%FBS	ATCC	(Lasfargues et al., 1978)
EFM-19	RPMI 1640, 10%FBS	DSMZ	(Simon et al., 1984)
HCC1187	RPMI 1640, 10%FBS	ATCC	(Gazdar et al., 1998)
HCC1954	RPMI 1640, 10%FBS	ATCC	(Gazdar et al., 1998)
M62	DMEMF12, 10%FBS	Tyler-Smith	(Mathias et al., 1994)
MCF7	RPMI 1640, 10%FBS	Dr M.J. O'Hare	(Bacus et al., 1990)
MDA-MB-175	RPMI 1640, 10%FBS	Dr M.J. O'Hare	(Cailleau et al., 1978)
T47D	RPMI 1640, 10%FBS	ECACC	(Keydar et al., 1979)
ZR-75-1	RPMI 1640, 10%FBS	ECACC	(Engel et al., 1978)
VP229	MCBD201, 10%FBS	G.Lowther	(McCallum and Lowther, 1996)
VP267	MCBD201, 10%FBS	G.Lowther	(McCallum and Lowther, 1996)

Table 2.3. Cell lines, growth conditions and references. ECACC is the European Collection of Cell Cultures, DSMZ is the German Collection of Microorganisms and Cell Cultures. ATCC is American Type Culture Collection. Prof. M. J. O'Hare (LICR/UCL Breast Cancer Laboratory, University College, Middlesex Medical School, London, UK); Dr. C. Tyler-Smith (Department of Pathology, University of Cambridge)

2.3.1. Thawing Splitting and Feeding Cells

Previously, ampoules of cells had been stored freezing medium (10% DMSO, 90% (v/v) culture media) in liquid nitrogen. To begin culture, frozen cells were thawed quickly at 37°C

in a water bath. The cells were quickly re-suspended in their medium and transferred to a 15ml Falcon tube along with 10 ml of pre-warmed culture medium at 37°C. Cells were spun down (500g, 3mins) freezing medium was removed by suction and the cell pellet re-suspended in 5 ml of culture medium and transferred to a T25 flask. Cells were incubated at 37°C, 5% CO_2.

Once adherent cells had reached 80%-90% confluence they could be split and sub-cultured. Culture medium was aspirated and the cells rinsed with 5-10 ml of versene, to remove dead cells and debris. Cells were incubated with 5 ml of versene+ trypsin at 37°C and inspected at 1 min intervals until all the cells were detached from the flask. 10ml of culture medium was added to neutralise the trypsin and the suspension centrifuges (500g, 3 min). The cell pellet was then re-suspended in culture medium and transferred to new flasks. Suspension cell lines were split and sub-cultured as above but without typsinisation.

2.4. Chromosome Preparations

2.4.1. Metaphase Chromosome Preparation for Flow Sorting

Flow sorting of chromosomes was a modification of the procedure previously described (Ng and Carter, 2006; Howarth et al., 2008). Sixteen hours after splitting, cells from ten T150 flasks were blocked in metaphase with 0.1µg/ml colcemid (demecolcecine) and incubated for 20 hours at 37°C. The mitotic cells were separated from adherent interphase cells by banging the flasks 15 times and transferring the supernatant to a new tube. Cell suspensions were centrifuged at 250g for 5 min and the supernatant discarded. For suspension cells, all of the initial sample was centrifuged as above. Cells were then re-suspended in total volume of 25ml polyamine hypotonic solution (75mM KCl, 0.5mM spermidine, 0.2mM spermine, 10mM $MgSO_4$, pH to 8.0, filter sterile) and incubated at room temperature. At 5 min intervals, cell swelling was monitored by mixing 10 µl of the swelling solution with 10µl of Turk's stain (0.01% (w/v) gentian violet, 1% (v/v) acetic acid). Under a phase contrast microscope, swelled cells looked round and chromosomes were visible as speckles inside each cell. Swollen cells were centrifuged at 250g for 5min and the supernatant discarded. The cells were re-suspended in 2ml of polyamine isolation

buffer (0.5mM EGTA, 2mM EDTA, 15mM Tris, 80mM KCl, 50mM NaCl, 0.5mM spermidine, 0.2mM spermine, 30mM DTT, 0.25% (v/v) Triton-X 100. pH to 7.4, filter sterile), incubated on ice for 10 min and gently vortexed for 15s. 10 µl of the cell suspension was placed on a microscope slide and propidium iodide (5 µg/ml) was added. At this stage, chromosomes were visible under a fluorescence microscope. If chromosome clumps remained then the cells were vortexed again. Chromosome preparations were centrifuged at 173g for 1min. The supernatant containing suspended chromosomes was collected and stored at 4°C.

One day prior to flow sorting, chromosomes in suspension were stained with Hoechst 33258 (5µg/ml final concentration), $MgSO_4$ (10mM final) and Chromomycin A3 (40µg/ml final) and incubated at 4°C. Approximately 1 hour before flow sorting, trisodium citrate (100 mM final) and sodium sulphite (250 mM final) were added to improve flow karyotype resolution (van den Engh et al., 1988). The suspension was filtered through a 20µm CellTrics filter under gravity to remove cellular debris and the filtrate kept on ice until sorting. Chromosome preparations were analysed and sorted using a MoFlo flow sorter (Cytomation Bioinstruments) by B.L. Ng (The Wellcome Trust Sanger Institute). The sheath buffer was composed of 10mM Tris-HCl (pH 8.0), 1mM EDTA, 100mM NaCl and 0.5mM sodium azide. For degenerate oligonucleotide primed polymerase chain reaction (DOP-PCR) amplification, aliquots of 300 chromosomes were flow sorted into PCR tubes containing 10µl of sterile PCR water. For GenomiPhi amplification (Amersham Biosciences), 5000 chromosomes were sorted. Figure 2.1 shows the flow karyotype for HCC1187.

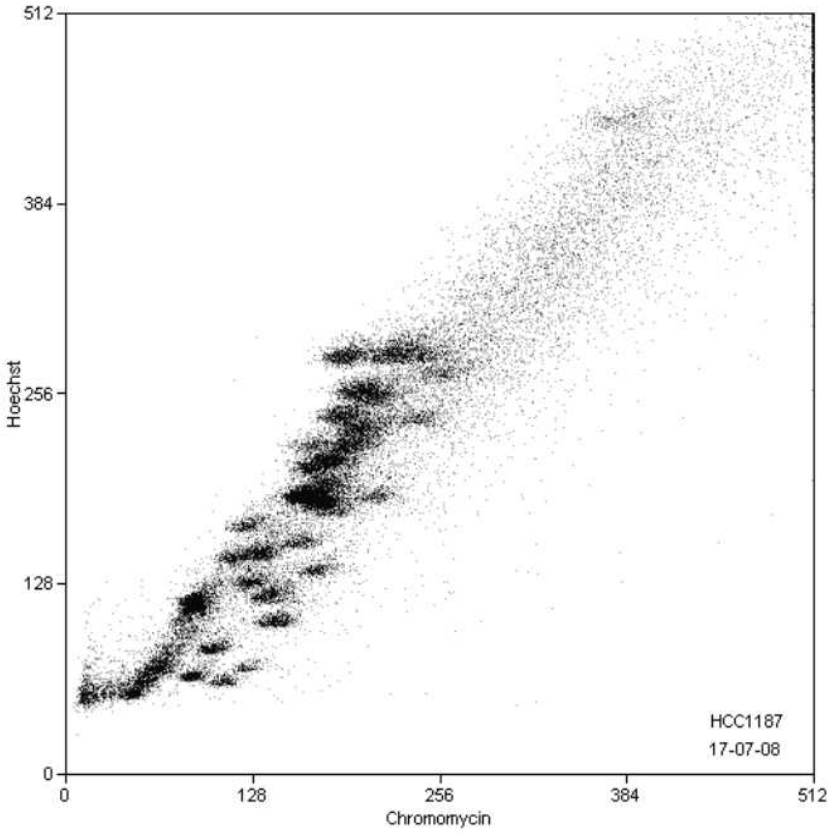


Figure 2.1. Flow Karyotype of HCC1187. Chromomycin fluorescence is on the x-axis, Hoechst fluorescence on the y-axis. Both scales are arbitrary. Flow sorting was performed by Dr. B.L. Ng, Wellcome Sanger Institute.

2.4.2. Metaphase Preparation for FISH

Once adherent cell lines had reached approximately 80% confluence, cultures were split. Twenty hours after splitting, 0.1 µg/ml colcemid was added and cells incubated for 1h30mins. Cells were trypsinised and spun down at 380g/3mins and supernatant removed. 0.075M KCl was slowly added under constant agitation to swell the cells. The swelling mixture was then incubated at 37°C for 15min. Swelling was stopped by addition of 5 drops of ice cold 3:1 methanol/acetic acid fixative and cells were spun down 500g/5mins. The supernatant was removed and the cells fixed by slow addition of cold 3:1 fixative under agitation. The cells were spun down and fixed twice more before a final fixation in 3:2 methanol:acetic acid. Fixed cell suspensions were kept at -20°C until use. 12µl of cell suspension was dropped into a 100µl drop of UP H_2O on a microscope slide

and left to dry. Metaphase spreads were examined using a Nikon phase contrast microscope. Slides with at least 10 visible metaphase spreads were dehydrated in an ethanol gradient (3mins each 75%, 90% and 100%) matured overnight at 37°C prior to FISH hybridizations.

2.4.3. Preparation of DNA Fibres for FISH

Fibre FISH was done as previously described (Mann et al., 1997). Cells fixed as for metaphase FISH were also used to prepare extended DNA fibres. 10 µl of the cell suspension was spread horizontally across a PolyPrep poly-L-lysine coated slide (~1/3 from the top of the slide). The slide was placed in a glass Coplin jar containing 60 ml of lysis buffer (0.5% (w/v) SDS, 50 mM EDTA, 200 mM Tris). The fixed cellular material was approximately 10mm below the surface of the lysis buffer. After 5min, 60 ml of 94% (v/v) ethanol was very slowly dropped on top of the lysis buffer. Chromatin fibres from lysed cells became visible at the point where the ethanol and lysis buffer met after approximately 10min. The slide was then slowly pulled out of the liquid at approximately 75° to the horizontal to extend the fibres along the slide. To fix the fibres in place, the slide was gently placed in a jar of 70% (v/v) ethanol and left for 30 min. Slides were dehydrated through an ethanol series as described above and air dried.

2.5. Fluorescence *in situ* Hybridization (FISH)

2.5.1. Preparation and Labelling of Chromosome Paints

Whole chromosome paints were made from flow sorted chromosomes amplified by degenerate oligonucleotide-primed PCR (DOP-PCR) as described previously (Telenius et al., 1992). Sorted human chromosomes were supplied by Professor M. Ferguson-Smith and Mrs. P. O'Brien, Department of Veterinary Medicine, University of Cambridge. DOP amplification had previously been performed by Dr J.C. Beavis using the below procedure:

Primary amplification of flow sorted chromosomes was performed in a 50 µl reaction containing 1X Buffer D, 200µM dNTPs, 0.05% (v/v) polyoxyethylene ether (W-1), 2µM 6MW primer (5'-CCGACTCGAGNNNNNNATGTGG-3'), 4.5 units Taq polymerase and

~350 flow sorted chromosomes. PCR was performed as in Table 2.5:

Step	Temperature	Time
1	94°C	9 min
2	94°C	1 min 30 s
3	30°C	1 min 30 s
	Repeat steps 2 and 3 for 9 cycles	
4	72°C	3 min
5	94°C	1 min 30 s
6	62°C	1 min 30 s
7	72°C	1 min 30 s
	Repeat steps 4-8 for 29 cycles	
8	72°C	9 min
9	4°C	hold

Table 2.5. Primary DOP PCR programme

Secondary and tertiary amplification of flow sorted products was performed in a 25 µl reaction containing 1X TAPS II buffer, 200µM dNTPs, 2µM 6MW primer, 0.05% (v/v) W-1, 2.25 units Taq polymerase and 5µl of primary or secondary DOP amplification reaction. PCR was as in Table 2.6:

Step	Temperature	Time
1	94°C	9 min
2	94°C	1 min 30 s
3	62°C	1 min 30 s
4	72°C	1 min 30 s
	Repeat steps 2-4 for 29 cycles	
5	72°C	8 min
6	4°C	hold

Table 2.6. Secondary and Tertiary DOP PCR programme

Amplified chromosome DNA was labelled by nick translation with either Biotin-conjugated dUTP or directly labelled Spectrum Orange (Vysis/Abbott), Fluorescein 12-dUTP (Roche). Labelling reactions contained 1X nick translation (NT) buffer (10X NT buffer is 0.5M Tris-HCl, 1mM dithiothreitol (DTT), 0.1M $MgSO_4$), 38µM d(A,C,G)TP, 19µM dTTP, 28µM labelled dUTP, 10 units DNA polymerase I, 0.7-2ng DNAse I and ~0.5µg DNA in a total

volume of 25µl and were incubated for 2 hours at 14°C. The reaction was stopped by addition of 2.5µl 0.5 M EDTA and incubation at 65°C for 10 min. Labelled whole chromosomes paints were stored in the dark at -20°C.

2.5.2. BAC clones and their culture

Appropriate bacterial artificial chromosome (BAC) and fosmid-bearing clones were selected for FISH experiments using the UCSC Genome Browser (http://genome.ucsc.edu). BACs were obtained from BACPAC resources. BAC clones for FISH experiments are listed in Table 2.7.

Gene	Clone Name	Start position (HG18)	End position (HG18)
PUM1	RP11-241O14	chr1: 204979770	chr1:205139387
TRERF1	RP11-7K24	chr6: 42173125	chr6:42228324
CTAGE5	W12-1623H12	chr14:38849330	chr14:38885848
SIP1	W12-1047K24	chr14:38685660	chr14:38728738
MDS1 (3')	RP11-659A23	chr3:170655616	chr3:170810166
MDS1 (5')	RP11-141C22	chr3:170367524	chr3:170542975

Table 2.7 BAC, PAC and fosmid clones used for FISH experiments. Genomic positions are from the HG18 genome build

E.coli bearing BACS or Fosmids were grown overnight on LB agar + chloramphenicol (20mg/ml) for RP11-BAC clones and W12 fosmids or kanamycin (25mg/ml) for RP4 and RP1 PAC clones. Colonies were then picked and then grown in 50ml LB broth culture + chloramphenicol or kanamycin for 16h.

2.5.3. Probe DNA Extraction and Labelling

BAC/fosmid DNA was extracted using Qiagen Spin Miniprep kit (Qiagen) as per manufacturer's instructions. 1000ng of probe DNA was labelled by nick translation with Spectrum Orange (Vysis/Abbott) or Fluorescein 12-dUTP (Roche) using an Abbot Molecular nick translation kit according to the manufacturer's instructions. Fibre FISH probes were labelled indirectly for greater sensitivity. These fosmids were labelled with biotinylated 16-dUTP (Bio-dUTP), or deoxygenin (Dig-dUTP) using a similar procedure to

labelling of chromosome paints except 150ng of DNA was labelled in a total volume of 25ul.

2.5.4. Probe Precipitation

Per FISH hybridization, 3µg human C0t-1 DNA (Roche), 200–500ng whole chromosome paint, 50–100ng of each BAC DNA were co-precipitated with 20 µg glycogen in 100% ethanol for 2 hours at -80°C. Fibre-FISH probe mixtures contained 150-200ng of each labelled fosmid, 2 µg human C0t-1 DNA. Probe mixes were spun down at 13g for 30mins at 4°C. Ethanol was pipetted off, the DNA pellet was then dried in the dark at 37°C for 30mins. The dry pellet was dissolved over 30mins at 37°C in 20µL hybridization buffer (50% v/v deionized formamide, 10% (w/v) dextran sulphate, 2XSSC, 1XDenhardt's solution and 40mmol/l sodium phosphate solution).

2.5.5. FISH Hybridization

FISH was performed as described previously (Alsop et al., 2006; Pole et al., 2006). Hybridisations were usually performed on metaphase spreads from cell lines along with an M62 karyotypically normal control. Probe mixtures were denatured at 70°C for 10min, and incubated at 37°C for one hour prior to hybridisation. Cell DNA in the form of metaphase preparations on microscope slides was denatured at 70°C in 70% deionized formamide–2X SSC for 1min 20s, quenched in ice-cold 70% ethanol for 2 minutes and dehydrated as above. Hybridizations took place in a humid chamber at 37°C for 14-20 hours. Formamide was deionized by stirring 1g mixed bed resin beads per 100 ml formamide for 2 hours. Beads were removed by filtration and 50 ml aliquots were stored at -35°C.

2.5.6. Post Hybridization Washing and Detection

Unbound probe DNA was removed by washing in 2× SSC for 5 minutes at room temperature, twice for 5 minutes at 42°C in 50% formamide–0.5× SSC, twice for 5 minutes at 42°C in 0.5× SSC. For directly labelled probes, coverslips were applied to slides using Vectashield plus 4′,6-diamidino-2-phenylindole (DAPI) mounting medium (Vector

Laboratories).

For indirectly labelled probes, biotin was detected with avidin conjugated to Cy5 (1 mg/ml, stock concentration), diluted 1:200 in 1% (w/v) BSA/4X SST. Prior to use, antibodies were vortexed briefly and centrifuged (16000g, 10 min) to remove self-conjugated aggregates. Following stringent washes, drained slides were blocked with 100 µl 3% (w/v) BSA/4X SST under a plastic cover slip in a dark humid chamber for 45mins. Slides were drained and 100 µl of antibody solution added and incubated for 30-60 min at 37°C. Slides were washed 4 times 5mins in 0.5% (w/v) BSA/4X SST at room temperature, drained briefly, and mounted with Vectashield as above.

2.5.7. Fibre FISH hybridizations and Washes

The probes were prepared and denatured as above. Slides with extended chromatin fibres were denatured for 3mins in 0.5 M NaOH, 1.5 M NaCl and then neutralised in 0.5 M Tris-HCl, 3 M NaCl, pH 7.2. Denatured slides were washed in 2X SSC and dehydrated through an ethanol series, 70% (v/v) ethanol, 90% (v/v) ethanol, and 100% (v/v) ethanol, 3mins each and air dried. Denatured FISH probe mixes were applied and coverslips sealed on as above. Probes were left to hybridise overnight at 37°C in a dark humid chamber. Following hybridisation, the coverslip was removed and slides were washed briefly in 2X SSC at room temperature. To remove unbound probe, slides were washed twice in 50% (v/v) deionised formamide/1X SSC (5 min, 40°C), and twice in 1X SSC (5 min, 40°C).

2.5.8. Fibre FISH Detection of indirectly labelled probes

Following FISH washes, endogenous epitopes were blocked using 100 µl 3% (w/v) BSA in 4XSST under a parafilm cover slip in a dark humid chamber for 1 hour at 37°C. Slides were briefly washed in BSA/4X SST and incubated with 100 µl of antibody solution for 45mins min at 37°C. For more sensitive detection, three antibody hybridisation steps were used. Digoxygenin was detected with a FITC-conjugated mouse anti- digoxygenin (23mg/ml, stock concentration). Biotin was detected with streptavidin conjugated Cy5 (Alexa 488) (1 mg/ml, stock concentration). Antibodies were prepared as in section 2.6.6. Antibody hybridization was with streptavidin and anti-DIG. The first and third hybridizations

were with anti DIG-FITC and streptavidin-Cy5. The second hybridisation was with anti Cy5 streptavidin. Between each antibody hybridization, slides were washed 3 times in 0.5% (w/v) BSA/4X SST (5 min, room temperature).

2.5.9. Image Acquisition and Processing

FISH experiments were visualised with a Nikon E800 microscope mounted with a 100 W mercury lamp light source (Microscope Services and Sales, Ewell, Surrey) and a cooled charge-coupled device camera (Applied Imaging, Newcastle-Upon-Tyne, UK). Slides were illuminated through a 83000 triple band pass filter set with TR, FITC and DAPI excitation filters (Chroma Technology Corp., Rockingham, VT, USA) to visualise FITC, Cy3/Spectrum orange. Cy5 visualisation was through XF93 triple band pass filter set (Omega Optical, Inc., Brattleboro, VT, USA). Composite raw images were pseudocoloured and enhanced using CytoVision software. The Images presented in this thesis were exported from CytoVision in .tiff format and thresholded with Adobe Photoshop CS3 software.

2.6. PCR and Sequencing

2.6.1. Amplification of Sorted Chromosomes for PCR

Five thousand of each flow-sorted chromosome (~5 µl) from the HCC1187 cell line were precipitated overnight at –20°C along with 0.5 µl non-fluorescent Pellet Paint (to make the precipitated DNA visible) and 3.2µl 2.5M NaCl, 35.5µl UP H_2O and 80µl 100% (v/v) ethanol. Chromosome aliquots were centrifuged at 16000g for 20mins at 4°C and the supernatant pipetted off. The pellet washed with 100µl 70% (v/v) ethanol, and then centrifuged as before for 10min. The supernatant was removed with a pipette and the pellet dried for 5min at 37°C and left to slowly re-suspend in 1µl TE. The chromosome DNA was amplified using the GenomiPhi whole genome DNA Amplification Kit (Amersham Biosciences) according to manufacturer's instructions. The amplified DNA was purified by spin column chromatography using MicroSpin G-50 columns packed with Sephadex G-50 (GE Healthcare) as per manufacturer's instructions. DNA was eluted in 50µl TE and the concentration of DNA measured with a Nanodrop instrument.

2.6.2. Genomic DNA preparation

Genomic DNA was extracted from HCC1187, VP229 and VP267 confluent cells using DNAzol reagent (Invitrogen). Cells were scraped from culture plastic in10 ml DNAzol per 10cm^2 of culture flask surface area. The DNAzol/cellular suspension was transferred to a clean tube and DNA was precipitated by adding of 0.5 ml 100% (v/v) ethanol per 1 ml of DNAzol followed by gentle mixing. The DNA precipitate was removed by spooling around a clean pipette tip and washed twice in 1 ml of 95% (v/v) ethanol. The DNA pellet was dried, then re suspended in 500 µl nuclease free water and stored at -20°C.

2.6.3. cDNA Preparation

RNA was extracted from cultured cells using Trizol reagent (Invitrogen) and chloroform.
10µg of total RNA was treated with 1µl of rDNase I and the rDNase to remove genomic DNA. DNAseI was then inactivated with DNAse Removal Reagent. First strand cDNA synthesis was performed using the SuperScript III First-Strand Synthesis Kit (Invitrogen). Oligo-dT priming was used to enrich for mRNA. For reverse transcription, 5µg of DNase-treated RNA, 50ng oligoDT primers and 1µl of 10mM dNTPs were mixed and incubated at 65°C for 5mins, 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 then added and the reaction 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.

2.6.4. PCR of Fusion Transcripts

To search for fused transcripts I used touchdown PCR as it is more sensitive than standard PCR techniques (Korbie and Mattick, 2008). All primers were designed using Primer3 website (http://fokker.wi.mit.edu/primer3/input.htm) (Rozen and Skaletsky, 2000) and were supplied by Eurofins/MWG; all had Tm of 62°C unless otherwise stated. PCR primer sequences are given in Appendix 1.1. PCR reactions were performed using GoGreen PCR master mix (Fermentas). Reactions were set up in a 25µl (25-50ng of template DNA,

12.5µl PCR master mix, 1µl each forward and reverse primer (10mM) and nuclease free water to 25µl). The PCR cycle as in Table 2.6 was run on a DNA Engine Tetrad PTC 225 thermal cycler.

Phase 1	Step	Temperature	Time
1	Denature	95 °C	3 min
2	Denature	95 °C	30 s
3	Anneal	62 +10 °C(a)	45 s
4	Elongate	72 °C	60 s or more(b)
	Repeat steps 2–4 (10–15 times)		
Phase 2	**Step**	**Temperature**	**Time**
5	Denature	95 °C	30 s
6	Anneal	57°C	45 s
7	Elongate	72 °C	2 min
	Repeat steps 5–7 (20–25 times)		
Termination	**Step**	**Temperature**	**Time**
8	Elongate	72 °C	5 min

Table 2.8 Touchdown PCR program as in Korbie and Mattick (2008). (a) Every time steps 2–4 are repeated, the annealing temperature was decreased by 1 °C/cycle, until 62°C was reached.

15µl of each reaction sample was run on a 1.5% (w/v) TBE agarose gel containing ethidium bromide along with Hyperladder I 200 base pair DNA ladder (Bioline). Gels were inspected under UV light Syngene G:BOX Chemi HR16 automated image analyser.

2.6.5. Sanger Sequencing of Fusion Transcripts

PCR of the fusion transcript was performed as above and fusion transcripts were cloned and sequenced. The PCR product was run on a 1.5% agarose TAE gel containing crystal violet, cut out of the gel and purified using S.N.A.P UV-Free Gel Purification Kit (Invitrogen) as per manufacturer's instructions. The PCR product was cloned using pCR-XL-TOPO vector and TOP10 chemically competent cells using TOPO XL PCR cloning kit (Invitrogen) as per manufacturer's instructions. Cells were grown on LB + kanamycin (50mg/ml) and positive transformants were picked and grown overnight. Plasmids were recovered using

Qiagen spin miniprep kit as per instructions. The purified plasmid was precipitated in ethanol, dried, re-suspended in water and sent for sequencing at the University of Cambridge Department of Biochemistry Geneservice.

2.6.6. Sanger Sequencing of Somatically Mutated Regions

Primers for the 85 sequence-level somatic mutations in HCC1187 were previously published (Wood et al., 2007) and listed in Appendix 1.3. PCR was performed on HCC1187 genomic DNA and DNA from flow sorted and amplified chromosomes extracted as described above. PCR reactions were performed using Hotmaster Taq DNA polymerase (5Prime). Reactions had a total volume of 50µl comprising 50ng of template DNA, 5µl of HotMaster PCR Buffer, 2µl 0.1mM dNTPs, 2µl (10pM) each forward and reverse primer and 0.2µl HotMaster Taq polymerase. The PCR cycle used DNA Engine Tetrad PTC 225 thermal cycler as in Table 2.9:

Step	Temperature	Time
1	95ºC	5min
2	95ºC	20s
3	58ºC	20s
4	72ºC	1min
	Repeat steps 2 to 4, 35x	
5	72ºC	5min
6	4ºC	Hold

Table 2.9. PCR amplification program prior to sequencing PCR

10µl of each sample was run on a 1.5% agarose gel to check for a single clear band. As the PCR primers I used had been previously validated, this was the case for all of the loci tested.

PCR products, in 96 well format, were purified using a Nucleofast 96 PCR cleanup kit (Clontech, Mountain View, CA)and vacuum manifold (Qiagen) as per manufacturer's instuctions. DNA was eluted in 20µl of nuclease free water. 3µl of this solution was used as the input for sequencing PCR. The same primers as used for initial amplification were

used for amplification with BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems, Foster City, CA). Per sample, approximately 400ng DNA, BigDye 1.5μlDilution Buffer, 1μl Big Dye v3.1, 1μl Primer (2pmol/ul), water to 10μl were combined for sequencing PCR as in table 2.10.

Step	Temperature	Time
1	96ºC	10 s
2	50ºC	5 s
3	60ºC	4 mins
	Repeat steps 1-3 for 25 cycles	
4	4ºC	hold

Table 2.10. Sequencing PCR program

To precipitate DNA following sequencing PCR, 1μl glycogen and 80μl 70% isopropanol were added to each well. The plate was incubated for 10 minutes at -20ºC and then centrifuged for 30 mins at 3500rpm. The supernatant was then removed by gently turning plate over onto a tissue. The plate was centrifuged, upside down on a tissue at 50g for 1 min to remove any residual isopropanol. The pellet was then washed with 150μl 70% ethanol and spun for 5mins at 3500rpm and the supernatant removed as above. The pellet was left to air dry before being re-suspended in10μl Megabace formamide sequencing buffer. Sequence chromatograms were generated using according an ABI 3700 capillary DNA sequencer according to manufacturer's instructions.

2.6.7. Sanger Sequencing Across Genomic Breakpoints

PCR of genomic structural variant junctions in HCC1187, VP229 and VP267 was performed in the same way. Primers were designed using Pimer3 software (Rozen and Skaletsky, 2000). For VP229 and VP267 DNA sequences flanking structural variant break points were assembled automatically from paired end sequence data (see below).

2.6.8. Pyrosequencing

Assays for SNP quantification were designed using Pyrosequencing Assay Design Software v 1.0 software (Biotage). Primers had Tm of approximately 70ºC to ensure high specificity. Outer primer pairs were tested by standard PCR on genomic DNA prior to pyrosequencing. Pyrosequencing PCR was carried out using PyroMark Gold reagents (Biotage) unless otherwise stated as per manufacturer's instructions. For each reaction 5.0µl Gold Buffer, 4.0µl $MgCl_2$ (25 mM), 2.5µl dNTP (10 mM), 0.3µl enzyme, 34.7µl water, 1.5µl sample genomic DNA (100ng/ul), 1.0µl biotin primer (10nM), 1.0µl non-biotin primer (10mM) in a 50µl total volume was cycled as in Table 2.11:

Step	Temperature	Time
1	95ºC	15min
2	95ºC	20s
3	(Anneling Temp)	20s
4	72ºC	20s
	Repeat steps 2 to 4, 45x	
5	72ºC	5min

Table 2.11. PCR conditions for Pyrosequencing

40µl of each PCR product was added to 3µl steptavidin-sepherose beads (GE Healthcare), 37µl binding buffer and 40µl water and shaken at 1200rpm for 5 mins to bind biotinylated PCR products to streptavidin beads. A pyrosequencing microtitre plate (Biotage) was prepared by adding to each well 1.5µl sequencing primer and 43.5µl annealing buffer. DNA-bound streptavidin beads were washed using a PyroMark Vacuum Prep Workstation (Biotage) for 5s each in 70% ethanol, denaturation solution, 1X washing buffer, water, water. Beads were ejected onto the pre-prepared pyrosequencing plate and their DNA denatured for 3mins at 80ºC. Samples were run on a Pyrosequencer PSQ 96MA (Biotage) using a PyroGold reagent cartridge as per manufacturer's instructions. Results were analysed using Biotage Assay Design software. PCR primers are listed in Appendix 1.5.

2.6.9. Illumina Sequencing

Short insert DNA sequencing libraries were prepared by Dr J.C.Pole and Dr I.Schulte with assistance and supervision from Dr S.F.Chin and Professor C.Caldas (CRUK Cambridge Research Institute). Genomic DNA was extracted as above and libraries constructed from a Paired-End DNA Sample Prep Kit (Illumina) according to the manufacturer's instructions. Sequencing was performed on an Illumina GAIIx sequencer, generating 38 base pair reads, at the Cancer Research UK Cambridge Research Institute.

2.6.10. Quantitative PCR

All Quantitative PCR (qPCR) reactions were performed in triplicate in a 10 µl volume containing 5 µl of SYBR Green PCR master mix (Applied Biosystems), 0.25 µM of primers and 1 µl cDNA (approximately 100ng, but adjusted as described below).To enrich for messenger RNA, cellular RNA was reverse-transcribed using an oligo-dT primer. Wherever practical, I used qPCR best practices as previously described (Bustin, 2005). The PCR cycle for the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) was:

Step	Temperature	Time
1	50ºC	2min
2	95ºC	10min
3	95ºC	15s
4	60ºC	30s
5	72ºC	2min
	Repeat steps 2 to 5, 40x	
6	72ºC	30min

Table 2.12. PCR conditions for quantitative PCR

I used the delta-delta Ct method to quantify the relative level of transcripts in a panel of cell lines (Pfaffl, 2001). The relative expression levels of the **target** gene to **ref**erence gene, in this case the housekeeping gene, *GAPDH*, were based on the difference in Ct (threshold cycle) values between a **control** 'normal' breast cell line, HB4a or HMT, and breast cancer cell lines **sample**s according to the equation:

$$\text{ratio} = \frac{(E_{target})^{\Delta Ct_{target}(\text{control-sample})}}{(E_{ref})^{\Delta Ct_{ref}(\text{control-sample})}}$$

E is the amplification efficiency of each primers pair. In theory a primer pair should be 100% efficient, doubling the amount of DNA present with each round of PCR during the exponential phase so E = 2. In reality primers are not 100% efficient, so E was calculated from the slope of Ct value versus log10 input DNA using the equation:

$$E = 10^{(-1/\text{slope})}$$

For each qPCR primer pair, I calculated E by running qPCR reactions with a serial dilutions of a universal reference cDNA (Clonetech) at100% (approximately 100ng/ul), 50%, 10%, 5%, 1%, 0.5%, 0.1% and 0.01% concentrations.

2.7. Bioinformatics

2.7.1. SNP6 data and Segmentation

SNP6 data for HCC1187 was kindly provided by Dr G. Bignell (Wellcome Trust Sanger Institute). For other cell lines, the segmented SNP6 array CGH data for the Bignell et al. (2010) dataset was downloaded from the Sanger Centre Cancer Genome Project genotype and trace archive http://www.sanger.ac.uk/genetics/CGP/Archive/ under a data access agreement. Copy number segmentation was provided by Dr C.D. Greenman (Wellcome Trust Sanger Institute) and had been processed with the PICNIC algorithm (Greenman et al., 2010). For VP229, data was provided by Dr S.L. Cooke, CRUK Cambridge Research Institute and PICNIC segmentation was performed by Miss C.K. Ng.

2.7.2. Break point regions from segmented SNP6 array CGH data

The Bignell et al. (2010) data was in the form of a list of SNP specific and copy number probes along with their PICNIC-segmented total copy number. To find break point regions

in SNP6 data I used a Perl script that identified copy number transition points. The script output the two SNP or copy number probe positions that flanked the segmented break point along with the break point polarity: copy number gain relative to the preceding segment was a positive break, loss was a negative break. The Perl script in listed in Appendix 2.1.

2.7.3. Genes at SNP6 break points

The list of break point regions from section 2.8.2 was compared with a list of 'gene windows' as described in Chapter 6 using the Perl script in Appendix 2.2. The array CGH data contained many germline copy number variants (CNVs). As normal tissue was not available for all cell lines, I compared my list of copy number steps with list of known CNVs (Redon et al., 2006; Zhang et al., 2009). Any copy number step within 20kb of a known CNV boundary was omitted from the analysis.

2.7.4. Ensembl API scripting to predict gene fusions

After the clustering of similar sequencing reads into structural variant “nodes”, the nodes and their DNA strands were in-putted to a fusion gene prediction Perl script (Batty, 2010).

2.7.5. Ensembl API scripting to retrieve structural variant break point regions

Structural variant nodes and their DNA strands were in-putted to a genome region extraction Perl script (Appendix 2.3). The script found the most conservative estimate of the break point region and used the Ensembl API to extract 1000 base pairs of sequence surrounding the putative structural variant. Repeat sequences in the extracted regions were then masked to avoid non-specific primer annealing.

2.7.6. Circular visualisation of data

Cicle plots ere generated using Circos version 0.48 software downloaded from http://mkweb.bcgsc.ca/circos/ (Krzywinski et al., 2009).

2.8. Statistical Model

2.8.1. Maximum likelihood estimators and confidence intervals

Scripts to generate MLE were written in the R statistical language (R Foundation for Statistical Computing, 2010). Confidence intervals were calculated using a bootstrapping approach (Davison and Hinkley, 1999). The R codes are in Appendix 2.4.

Suppose that *X1,X2, ... ,Xm* is a sample of size m, each taking values in {0; 1}. Of the m observations, an unknown number n take the value 1 whereas the other m - n are independent Bernoulli random variables having:

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, m - n. \qquad (1)$$

The random variable *Y = X1 + ... + Xm* is observed, and the aim is to estimate the parameter *n* and provide some assessment of confidence in that estimate. Note that this is a special case of the problem in which *Xm-n+1, ... ,Xm* are independent Bernoulli random variables having parameter *q*. In this case we can write down the distribution of *Y* , which is the sum of two independent Binomial random variables, *B1 ~ Bin(m – n, p);B2 ~ Bin(n, q)*. Thus:

$$\begin{aligned} \mathbb{P}(Y = y) &= \sum_u \mathbb{P}(B_1 = u)\mathbb{P}(B_2 = y - u) \\ &= \sum_u \binom{m-n}{u} p^u (1-p)^{m-n-u} \binom{n}{y-u} q^{y-u} (1-q)^{n-y+u}. \qquad (2) \end{aligned}$$

2.8.2. Classical approach

In what follows is a sample case where *p = 0:5, q = 1* (3)
In this case (2) reduces to:

$$\mathbb{P}(Y = y) = \binom{m-n}{y-n} \left(\frac{1}{2}\right)^{m-n}, \quad y = n, n+1, \ldots, m. \qquad (4)$$

2.8.3. Finding the MLEs

To find the maximum likelihood estimator of n, we use (4) to write the likelihood as:

$$L(n) = \binom{m-n}{y-n}\left(\frac{1}{2}\right)^{m-n}, n = 0, 1, \ldots, y; \quad = 0, n > y. \qquad (5)$$

This can be maximized numerically. In some cases the MLE is not unique. For example when *m = 16; y = 14*, the values *n = 12; 13* have equal likelihoods.

2.8.4. Confidence intervals

Supposing we have estimated *n* by an MLE, $\hat{n}$. We then simulate B values of the MLE assuming that $\hat{n}$ is the true parameter. This produces simulated values $\hat{\hat{n}}_i$, i = 1, 2, … , B. We suppose that the distribution of $\hat{n} - n$ is well-approximated by that of $\hat{\hat{n}} - \hat{n}$, from which a confidence interval can be approximated. To get a 100(1α) CI, sort the values of $\hat{\hat{n}}_i$, i = 1, 2, … , B into increasing order, and record n_l = Bα =2th value and n_u nu = B(1 α -2)th value. The CI is then has the form:

$$100(1 - a)\%CI = (2\ \hat{n} - n_u\ , l2\ \hat{n} - n_l\ )$$

Chapter 3

The Structure of a Breast Cancer Genome

3.1. Introduction

HCC1187 is a hypotriploid cell line derived from an ER-negative and *ERBB2* non-amplified primary ductal carcinoma (Gazdar et al., 1998; Wistuba et al., 1998). From a genomics perspective, it is one of our most intensively studied models of breast cancer having been investigated by molecular cytogenetics, exome-screening, and massively parallel paired end sequencing (Sjöblom et al., 2006; Wood et al., 2007; Howarth et al., 2008; Stephens et al., 2009). The purpose of this section is to, as fully as possible, define the genomic aberrations of this breast cancer cell line and to use these data to address two questions:

1) How many genes are disrupted by chromosome aberrations in this “typical” breast cancer cell line and how does this compare to the sequence-level mutational burden?
2) Do chromosome aberrations fuse any genes? If so, how many fusion transcripts can be found in HCC1187?

3.2. Previous Data

3.2.1. Spectral Karyotyping (SKY)

SKY was done previously by Dr. M. Grigorova (Grigorova and Edwards, 2004).
The modal chromosome number was 63 and ranged from 61 to 64. By SKY alone, there are 20 structural chromosome abnormalities, including two reciprocal translocations – t(1;8) and t(10;13).

3.2.2. Array Painting

In 2008, HCC1187 was investigated by array painting (Howarth et al., 2008). This allowed a more detailed analysis of the HCC1187 karyotype. Initially, all chromosomes were hybridized to 1Mb BAC arrays and to this resolution, 37 chromosomes were are apparently normal and 29 were structurally abnormal. 24 of these breaks appeared to be balanced with respect to one chromosome and these chromosomes were hybridized to high resolution custom oligonucleotide arrays. This allowed high-resolution identification of

Chromosome Name from Howarth et al. (2008)	Cytogenetic Description	Modal Number of Copies by SKY
A	der(1)(6pter->6p21.1::1p35.2->1q21.3::8p22->8pter)	1
D	der(X)(6pter->6p21.1::1p35->1p21.3::Xp11.22->Xqter)	1
E	der(8)(1q10->1q21.3::8p22->8q22.2::1p31.1->1pter)	1
G	der(20)t(2;20)(q10;q11.21)	2
H	der(8)t(1;8)(p31.1;q22.2)	1
J	der(1)t(1;8)(p13;q22.2)	2
M	del(7)(q36.1)	2
N	der(10)(13qter->13q21::10p12->10q23.1::19q13.41->19qter)	1
O	der(11)t(11;12)(p15.4;p11.22)del(11)(q13.5q21)	1
P	der(19)t(2;19)(p10;p13.3)	1
R	der(11)t(11;16)(p15.3;q22.1)del(11)(q13.5q21)	1
S	der(16)(16pter->16q22.1::11p15.3->11p15.4::12p11.22->12pter)	1
T	der(19)t(2;19)(p16;p13.3)	1
U	i(18)del(18)(q21.2)	1
V	der(2;5) t(2;5)(p10;p10)del(2)(p16p25.1)	1
Y	der(20)t(14;20)({14qter->14q24.3:}{20pter->20qter)	1
b	der(13)t(10;13)(p12;q21.31)	3
c	i(13q)del(13)(q10q31)	1
i	der(19)t(1;19)(p36.22;q13.1)3	1
j	der(?)({20pter->20p13:}{:13q31.1->13qter})3	1
k	trc(1;X;1)(1qter->1p11::Xp21.3->Xq25::1p11->1qter)	1
B	4	2
C	3	2
F	5	2
I	6	2
K	7	2
N	8	1
Q	9,10,11,12	3,2,2,2
W	14	2
X	15	2
Y	16	1
Z	17	2
a	18	2
d	20	1
e	19	1
f/g	21	3
h	22	2
L	X	1

Table 3.1. **Cytogenetic description of HCC1187 karyotype** modified from Howarth et al (2008). Chromosome U was shown to be an isochromosome of 18q by FISH (S.N. not shown)

chromosome breakpoints. All chromosomes are described cytogenetically in Table 3.1, and are named A-Z and a-k, based on their sizes from the flow-sorted karyotype

(Howarth et al., 2008). The positions of chromosome break points defined by array painting are listed in Appendix 3.1.

3.2.3. Massively Parallel Paired End Sequencing

In 2009, Stephens et al. published a survey of 24 breast cancer genomes by massively parallel paired end sequencing (Stephens et al., 2009). HCC1187 was one of the samples used and this provided further data on the structure of the HCC1187 genome. Stephens et al. (2009) reported 11-fold haploid genome coverage for HCC1187 but as the cell line is near triploid this translates to approximately 3.7-fold physical coverage. The authors estimated that they had found 50% of the structural variants in their samples and indeed uncovered substantial somatic variation in the HCC1187 genome. Most rearrangements were small deletions and tandem duplications beyond the resolution of array CGH so could not have been detected by previous methods.

Confirmed somatic structural variations comprised 26 deletions, 9 inversions, 50 insertions (most were probably tandem duplications) and 11 inter-chromosome translocations. The intra-chromosomal rearrangements ranged in size from 3.7kb to 58Mb with a median size of 65.5 kb. All structural variants are listed in table Appendix 3.2.

3.2.4. Exome-wide Mutation Screen and Targeted Resequencing

Wood et al. (2007) reported 79 sequence-level mutations in HCC1187. Six additional mutations are listed in the COSMIC database (Forbes et al., 2010). In total there are 85 sequence-level mutations in the HCC1187 genome comprising 75 base substitutions and 10 indels. All previously reported sequence-level mutations in HCC1187 are listed in Appendix 3.3.

3.3. Analysis Part I. The Genome Structure of HCC1187

3.3.1. Combining Array Painting Data with SNP6 array CGH Data

The breakpoints of the potentially balanced rearrangements in HCC1187 had previously been mapped using custom oligo arrays by array painting (Howarth et al., 2008). The large number of remaining unbalanced breakpoints were only known to within approximately 1Mb. To map these unbalanced rearrangements, high resolution array CGH, provided by Dr G Bignell and analysed by Dr C.D. Greenman of the Wellcome Sanger institute, was used (Bignell et al., 2010; Greenman et al., 2010). Copy number change points had been identified by array CGH segmentation using the PICNIC algorithm (Greenman et al., 2010). All copy number segments from array CGH are listed in Appendix 3.4.

Array painting-derived CGH and SNP6 CGH data data for the HCC1187 genome (Howarth et al., 2008) were compared. There was a very good concordance between the two data sets (Figure 3.1). Where array painting had identified an unbalanced copy number step to 1Mb resolution, the segmented SNP6 break point was always within those bounds (Table 3.2).

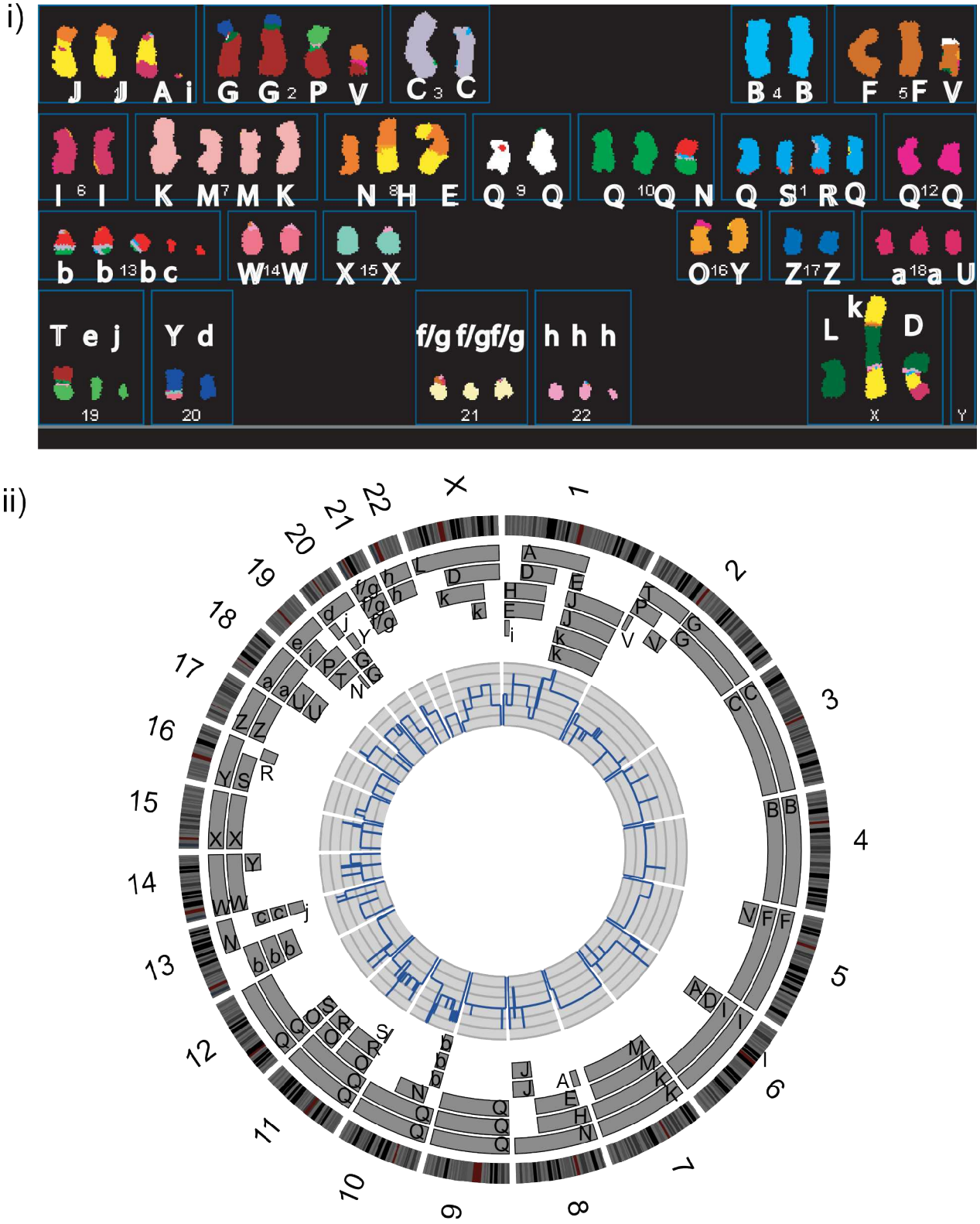


Figure 3.1. The structure of the HCC1187 genome.

Legend overleaf

Figure 3.1. The structure of the HCC1187 genome. i) Spectral Karyotype (Grigorova and Edwards, 2004) with Howarth et al. (2008) chromosome names in white. ii) Circular representation (Krzywinski et al., 2009) of the HCC1187 genome. Chromosome ideograms are arranged clockwise p-terminal to q-terminal around the outside. Moving inward, chromosome segments from array painting (Howarth et al., 2008) are grey rectangles and the derivative chromosome to which they belong are indicated. For example, chromosome D is made from pieces of chromosomes 1, 6 and X. The inner plot is SNP6 array CGH segmented with the PICNIC algorithm (blue line). Segmented copy number is on the y-axis.

Name	Proposed Junction	Chr	Array Painting LHS	Array Painting RHS	PICNIC LHS	PICNIC RHS		Chr	Array Painting LHS	Array Painting RHS	PICNIC LHS	PICNIC RHS
A, D	t(1;6)	1	31160803	31296885	31180697	31183666	;	6	42354900	42358000	42354401	42356874
A	t(1;8)	1	150503923	150563199	150517061	150526980	;	8	14331543	14465908	143981882	143983187
D	t(1;X)	1	96961782	97980642	97790620	97791276	;	X	50231248	54118742	53295047	53299507
E	t(1;8) a	1	84172800	84174600	84233795	84234712	;	8	99290956	101355408	Balanced	Balanced
E	t(1;8) b	1	150503923	150563199	150517061	150526980	;	8	14589000	14591700	Balanced	Balanced
G	t(2;20)	2	89772453	94882962	88841557	88843175	;	20	30605410	30851018	30764109	30768267
H	t(1;8)	1	84135872	85445695	84233795	84234712	;	8	99290956	101355408	Balanced	Balanced
J	t(1;8)	1	115246265	115255974	115260780	115276291	;	8	99290956	101355408	Balanced	Balanced
N	t(10;13)	10	22830000	22875000	22815216	22815579	;	13	58272600	58273300	58272072	58278781
N	t(10;19)	10	84424000	84425500	84425879	84426077	;	19	56139000	56141000	56248520	56250406
O	del(11)q13 .5q21)	11	76153426	78120577	77811711	77814699	;	11	88001258	90989730	90315308	90315329
O	t(11;12)	11	-	-	5535160	5536230	;	12	28228474	28699165	28660139	28664235
P	t(2;19)	2	89772453	94882962	88841557	88843175	;	19	-	-	5745718	5750806
R	t(11;16)	11	10435000	10438000	Balanced	Balanced	;	16	66141900	66142200	Balanced	Balanced
R	del(11)q13 .5q21)	11	76153426	78120577	77811711	77814699	;	11	88001258	90989730	90315308	90315329
S	t(12;16)	12	28228474	28699165	28660139	28664235	;	16	66141900	66142200	Balanced	Balanced
S	t(11;16)	11	-	-	estimate	9302500	;	16	66141900	66142200	Balanced	Balanced
S	t(11;12)	11	-	-	5535160	5536230	;	12	28228474	28699165	28660139	28664235
T	t(2;19)	2	54050000	54053000	54050162	54055907	;	19	5775600	5778000	5745718	5750806
V	t(2;5)	2	89772453	94882962	88841557	88843175	;	5	42897896	50000844	Balanced	Balanced
V	del(2p) (p16p25.1)	2	-	-	105251764	105256524	;	2	54050000	54053000	54050162	54055907
Y	t(14;20)	14	75743298	76653724	72742737	72745332	;	20	>1	<30760000	Balanced	Balanced
b	t(10;13)	10	22830000	22875000	22815216	22815579	;	13	58272600	58273300	58272072	58278781
c	i(13)	13	-	-	88420742	88421559	;	13	-	-	88420742	88421559
i	t(1;19)	1	11042656	11551099	11065194	11065899	;	19	42104981	42928443	42438672	42438948
j	t(13;20)	13	-	-	88420742	88421559	;	20	>1	<30760000	Balanced	Balanced
k	t(1;X)	1	-	-	120264944	120268277	;	X	12228154 3	123854583	122545844	122546391
	t(X;X)	X	26887513	27118094	27115786	27116914	;	X			93199431	93201916
dmin	-	1			117396138	117396585	;				extends as far as 146000000	

Table 3.2. Comparison of array painting CGH with PICNIC-segmented SNP6 array CGH. Mappings are based on the HG18 genome build

All copy number steps present at low resolution were present at high resolution also. This means there had been no recent karyotype evolution in culture and that the HCC1187 karyotype was relatively stable. Figure 3.2 shows a representative unbalanced chromosome break defined by 1Mb array painting CGH and high resolution SNP6 array CGH.

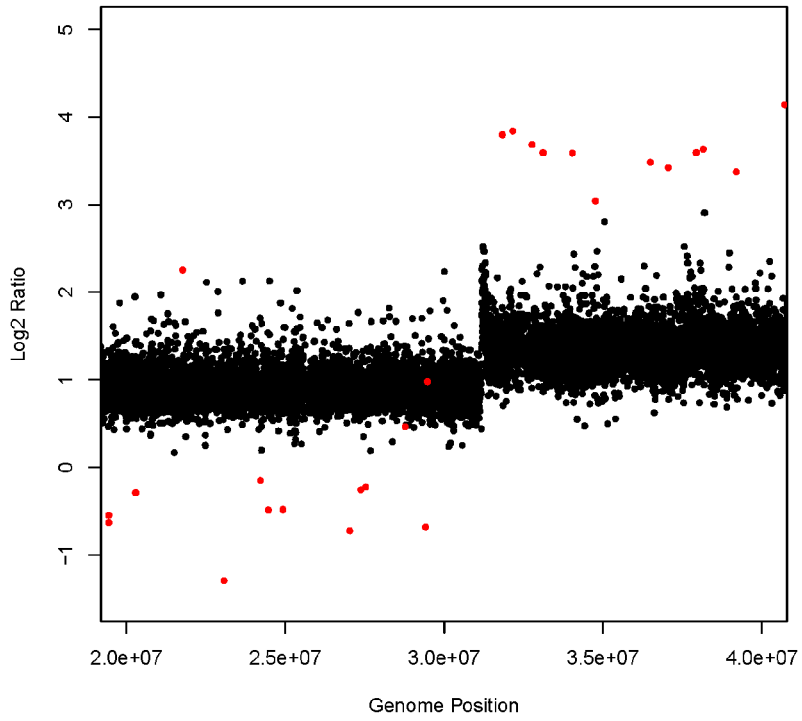


Figure 3.2. Comparison of 1MB BAC array with SNP6 array CGH, chromosome 6p. Red dots are median positions of 1MB BACs form Howarth et al. (2008), black dots are individual probes from Affymetrix SNP6 array from Bignell et al. (2010). The SNP6 array gives a much higher resolution estimate of the breakpoint position. In this case, the midpoint of the BACS flanking the break were 31160802 and 31296885, the SNPs flanking the break were at 31180697 and 31183666.

3.3.2. Incorporating Massively Parallel Paired End Sequence data

I next compared my provisional genome map with the structural variant data reported by Stephens et al. (2009). Again, there was a high concordance between the array based definitions of chromosome break points and the sequencing-derived mappings. Of the 28

cytogenetically visible chromosome junctions, 12 had been sampled by Stephens et al. (2009). These junctions had been mapped to base pair resolution, so were the best mapping available (Table 3.3). Two inter chromosome genomic junctions "StephensDIF2" chr1:31334253+ joined to chr6:41406061- and "StephensDIF6" chr7:148873715+ joined to chr8:143979884- were not anticipated by array painting. StephensDIF2 is likely to represent some complexity at the t(1;6) chromosomes A and D translocation break point and is discussed below in section 3.4.1. StephensDIF6 may represent an unbalanced translocation between the q termini of chromosomes 7 and 8. Chromosome M, listed as del(7)(q36.1) probably has a small piece of chromosome 8, 143.9Mb>qter attached. This small piece of chromosome 8 was below the resolution of the 1 Mb array painting method.

One large intrachromsome inversion could not have been anticipated by array painting: “StephensINV1” and “StephensINV2” indicate an inversion between chr1:157359724 and 215686770. This indicates an inversion of a large segment of 1q, and is presumably present on one or several of chromosomes J or k.

3.3.3. Genes at Chromosome Break Points

After assembling the best mappings for each chromosomal breakpoint, I next asked if any genes had been broken and whether if the 5' portion of one gene had ever been juxtaposed to the 3' end of another, possibly forming a fusion transcript (Table 3.3). When predicting gene fusions, I assumed that chromosome segments containing a telomere formed the termini of derivative chromosomes. Similarly, centromeric breaks were assumed to form centromeres in derivative chromosomes. For segments without telomeres or centromeres, orientation was unclear as genomic breakpoints in many cases had not been cloned. This is true of chromosomes E,N,O and R. The alternative configurations were, however, also considered when investigating possible gene fusions.

Chr	Junction	Stephens Structural Varaint?	Chr	Best Mapping	Strand	Chr	Best Mapping	Strand	Gene A	Gene B	Possible Fusion
A	t(1;8)		1	150522021	+	8	143982535	-	*HRNR (3')*	*LY6E (3')*	
A, D	t(1;6)	DIF1	1	31183855	-	6	42356186	+	*PUM1(5')*	*TRERF1 (3')*	***PUM1-TRERF1***
b	t(10;13)	DIF8	10	22832193	-	13	58272472	-			
c	i(13)		13	88421151	+	13	88421151	+			
D	t(1;X)	DIF4	1	97784960	+	X	53294900	+	*DPYD (3')*	*IQSEC2 (5')*	***IQSEC2-DPYD***
E	t(1;8) a	DIF3	1	84234093	+	8	101058886	+	*TTLL7 (3')*	*RGS22 (3')*	
E	t(1;8) b		1	150522021	+	8	14590350	+	*HRNR (3')*	*SGCZ (5')*	***SGCZ-HRNR***
G	t(2;20)		2	88842366	-	20	30766188	+		*COMMD7 (5')*	
H	t(1;8)		1	84234254	+	8	100323182	-	*TTLL7 (3')*	*RGS22 (3')*	
i	t(1;19)		1	11065541	+	19	42438810	-	*EXOSC10 (3')*	*ZNF585B (3')*	
J	t(1;8)	DIF5	1	115272352	-	8	101059496	-	*SYCP1 (3') / SIKE1 (5')*	*RGS22 (5')*	***RGS22-SYCP1***
j	t(13;20)		13	88421151	-	20	?	+		*?*	
k	t(X;X)		X	27116350	-	X	93200674	-			
k	t(1;X)		1	120266611	-	X	122546118	+	*NOTCH2 (5')*		
M	t(7;8)	DIF6	7	148873715	+	8	143979884	-		*CYP11B2 (5')*	
N	t(10;13)	DIF7	10	22831512	+	13	58272177	+			
N	t(10;19)		10	84425978	+	19	56249463	+	*NRG3 (5')*	*SIGLEC9 (5')*	
O	del(11)q13.5 q21)	DEL17	11	77809792	+	11	90309424	-			
O	t(11;12)	DIF9	11	5536414		12	28661343		*OR52B6 (3') / HBG2 (5')*		
P	t(2;19)		2	88842366	+	19	5748262	-		*NRTN (3')*	
R	del(11)q13.5 q21)	DEL17	11	77809792	+	11	90309424	-			
R	t(11;16)	DIF10	11	10436500		16	66142050		*AMPD3*		

S	t(11;12)	DIF9	11	5536414	-	12	28661343	+	*OR52B6 (3') / HBG2 (5')*		
S	t(11;16)		11	9108545*	+	16	66198168	+	*SCUBE2 (3')*	*CTCF (5')*	*runthrough* ***CTCF-SCUBE2***
T	t(2;19)		2	54053031	+	19	5748262	-	*PSME4 (3')*	*NRTN (3')*	
V	del(2p) (p16p25.1)	DEL2	2	11397776	+	2	55159176	-	*ROCK2 (3')*	*RPS27A (3')*	
V	t(2;5)		2	88842366	+	5	46449370	-	*FLJ40330 (5')*		
Y	t(14;20)	DIF11	14	72736528	-	20	6147141	+	*PSEN1 (5')*		

Table 3.3. Genes at array painting chromosome break points. Break point co-ordinates are defined by the best available mapping from array painting, SNP6 or Stephens et al. (2009).*The mapping of this chromosome translocation was from a cloned genomic junction from Dr. K.D. Howarth.

3.3.4. Sub-Microscopic Aberrations From SNP6 and Massively Parallel paired End Sequencing Data

SNP6 array CGH showed additional rearrangements that were below the resolution of 1Mb array painting. The segmentation algorithm predicted 13 small deletions ranging from 0.26 kb to 2.3Mb with a median size of 257kb. There were also 27 small duplications ranging from 11.7kb to 2.8Mb, median size 320kb. All of these duplications and deletions were absent in the matched normal lymphoblastoid cell line, HCC1187BL. Many of these features were likely to be small interstitial deletions or “head to tail” tandem duplications. Indeed, five of the 13 deletions and 17 of the 27 duplications had associated structural variants from Stephens et al. (2009) that confirmed this inference. I assembled a list of genes at break points and predicted if any gene fusions may have resulted. I assumed each feature was either an interstitial deletion or tandem duplication (Table 3.4).

Massively parallel paired end sequencing uncovered further structural variation that was below the resolution of SNP6 array CGH segmentation. There were 18 apparent deletions ranging from 3.8kb – 2.01Mb with a median size of 16kb, 35 apparent duplications from 4.2kb – 583kb median size 64kb. Additionally, there were 7 inversions ranging from 5.5kb to 17.8kb with a 14.8kb median size. These small balanced rearrangements could not have been detected by array CGH approaches. As for the array CGH segmented deletions and duplications I identified broken genes and possible fusions (Table 3.5).

Chr	Type	Previous Segment End	Size of gained or lost region (kb)	Next Segment Start	Previous Segment CN	Gained or lost region copy number	Next Segment CN	Stephens et al. (2009) SV?	LHS Gene	RHS Gene	Potential Gene Fusion
2	Del	33032718	56.04	33089758	2	0	2	DEL3	*LTBP1(5')*	*LTBP1(3')*	*No – intronic*
2	Del	54050162	1101.61	55159490	2	1	2		*PSME4 (3')*	*RPS27A(3')*	
2	Del	66859298	257.47	67119098	2	1	2	DEL4			
6	Del	1661709	4019.19	5681742	4	3	5	DEL6	*GMDS (3')*	*FARS2(3')*	
6	Del	101044085	205.62	101259352	2	1	2		*SIM1 (3')*	*ASCC3(5')*	
6	Del	134367799	208.05	134580872	2	0	2		*SLC2A12 (3')*	*SGK1(5')*	***SGK1-SLC2A12***
8	Del	124739282	2785.93	127533246	3	1	4		*KLHL38 (3')*		
11	Del	20684394	23364.99	44051626	4	3	4		*NELL1(5')*	*ACCS(3')*	
12	Del	33417710	0.26	33420555	2	4	4			*SYT10(3')*	
12	Del	127475298	646.23	128122073	2	1	2		*TMEM132C (5')*		
14	Del	62771053	630.99	63410248	2	0	2		*KCNH5(3') / RHOJ(5')*	*SYNE2(3')*	***RHOJ-SYNE2***
15	Del	69964951	249.53	70233770	2	0	2		*MYO9A(3')*		
17	Del	34666635	85.62	34766404	2	0	2	DEL22		*FBXL20(5') / CRKRS(3')*	
1	Dup	31180697	149.29	31336568	2	5	4	DIF1/DIF2	*PUM1(5') /PRO0611(3')*	*SNORD85 (3')*	
2	Dup	104926149	319.83	105256524	2	3	2	INS4	*MRPS9(3')*	*TGFBRAP1(3')*	
2	Dup	222210422	651.77	222869850	2	3	2	INS5		*PAX3(3')*	
3	Dup	9483392	1113.06	10597941	2	3	2		*SETD5(3')*	*MIR885(3')*	
3	Dup	57148414	95.13	57255295	2	4	2	INS9	*IL17RD(5') / APPL1(3')*	*APPL1(5') / HESX1(3')*	*Runthrough IL17RD-HESX1*
3	Dup	130771904	131.5	130909903	2	4	2	INS10	*PLXND1(5')*	*TMCC1(3')*	***PLXND1-***

											TMCC1
4	Dup	146293755	428.13	146731470	2	4	2	INS17	*OTUD4(5')*		
4	Dup	199380516	2800.24	2807561	2	3	2			*TNIP2(3') / SH3BP2(5')*	
6	Dup	41394582	959.43	42356874	4	6	2	DIF1	*NCR2(3')*	*TRERF1(3')*	
6	Dup	149546259	497.51	150052602	2	3	2	INS23	*MAP3K7IP2 (3')*	*LATS1(3')*	
8	Dup	102400777	151.95	102558211	3	5	3	INS28	*NACAP1 (3')*		
8	Dup	127530787	11.7	127547507	1	4	3				
8	Dup	127966102	889.5	128857819	3	4	3				
9	Dup	12874003	169.69	13047115	3	5	3	INS29			
9	Dup	113916140	215.02	114139966	3	4	3	INS31	*SUSD1(5')*	*ROD1(3')*	***SUSD1-ROD1***
10	Dup	30165639	272.47	30440477	3	4	3			*KIAA1462 (3')*	
10	Dup	46363383	874.32	47417401	3	4	3		*GPRIN2(3')*		
12	Dup	11769713	58.38	11829511	3	5	3	INS36	*ETV6(3')*	*ETV6(5')*	*No – intronic*
12	Dup	33420367	334.77	33756464	4	4	2		*SYT10(3')*		
12	Dup	34043707	424.29	34490595	2	5	2		*ALG10(3')*		
13	Dup	101878311	365.51	102251373	4	5	4	INS38	*TPP2(3')*	*BIVM(5')*	
14	Dup	38673786	196.33	38881982	2	4	2	INS39	*SIP1(3')*	*CTAGE5 (5') / TRAPPC6B (3')*	***CTAGE5-SIP1***
14	Dup	63658710	251.89	63912211	2	4	2	INS40	*SYNE2(3')*	*ESR2(3')*	
14	Dup	67838563	317.88	68157617	2	4	2	INS41	*RAD51L1 (3')*		
15	Dup	83230545	600	83835487	2	3	2		*SLC28A1 (3')*	*AKAP13(5')*	***AKAP13-SLC28A1***
15	Dup	88492748	146.17	88640915	2	4	2		*SEMA4B (3')*	*CIB1(3')*	
18	Dup	8972450	1692.78	10666165	4	5	4	INS47	*NDUFV2 (3')*		

Table 3.4. Small deletions and Duplications from segmented array CGH. Del=deletion, Dup=duplication, CN=copy number from PICNIC array CGH segmentation. Some of the deletions and duplications had an associated structural variant listed in Stephens et al. (2009) and are named as in

Appendix 3.2. Genes at deletion or duplication break points are listed and the end of the gene retained in a possible fusion transcript is also noted (3' or 5').

Name	SV Type	Chr A	Position A	Strand A	Chr B	Position B	Strand B	Size (kb)	Gene A	Gene B	Potential Effect
DEL10	Del	8	41576622	+	8	41588498	-	11.88	*AGPAT6 (5')*	*AGPAT6 (3')*	Deleted Exons
DEL11	Del	8	70272683	+	8	70277843	-	5.16			
DEL12	Del	10	14616540	+	10	14649227	-	32.69	*FAM107B (3')*	*FAM107B (5')*	within intron
DEL13	Del	10	77414245	+	10	79429479	-	2015.23	*C10orf11 (5')*	*POLR3A (5')*	
DEL15	Del	11	34072921	+	11	34082648	-	9.73	*CAPRIN1 (5')*		
DEL16	Del	11	55503503	+	11	55507819	-	4.32	*OR7E5P (3')*	*OR7E5P (5')*	within intron
DEL18	Del	12	114935765	+	12	114961462	-	25.7	*MED13L (3')*	*MED13L (5')*	Exons Deleted
DEL19	Del	13	47846990	+	13	47850855	-	3.87	*RB1 (5')*	*RB1 (3')*	Exons Deleted
DEL20	Del	14	79857506	+	14	79867739	-	10.23	*DIO2 (3')*	*DIO2 (5')*	within intron
DEL21	Del	14	98850865	+	14	98856028	-	5.16			
DEL23	Del	20	9081436	+	20	9131009	-	49.57	*PLCB4 (5')*	*PLCB4 (3')*	within intron
DEL24	Del	20	52273532	+	20	52281706	-	8.17			
DEL25	Del	20	52887189	+	20	52919407	-	32.22			
DEL26	Del	X	153858946	+	X	153879982	-	21.04	*F8 (3')*	*F8 (5')*	Exons Deleted
DEL5	Del	4	138011731	+	4	138039024	-	27.29			
DEL8	Del	7	110845099	+	7	110867805	-	22.71	*IMMP2L (3')*	*IMMP2L (5')*	within intron
DEL9	Del	7	132786464	+	7	132871038	-	84.57	*EXOC4 (5')*	*EXOC4 (3')*	Exons Deleted
INS1	Ins	1	37593659	-	1	38176917	+	583.26	*ZC3H12A (3')*	*INPP5B (3')*	
INS11	Ins	4	13044591	-	4	13478102	+	433.51	*RAB28 (5')*		
INS12	Ins	4	40464439	-	4	40515180	+	50.74	*NSUN7 (3')*	*APBB2 (3')*	
INS13	Ins	4	79203712	-	4	79238418	+	34.71	*FRAS1 (3')*	*FRAS1 (5')*	within intron
INS14	Ins	4	82495561	-	4	82601352	+	105.79		*RASGEF1B (3')*	

INS15	Ins	4	92041059	-	4	92076272	+	35.21	*FAM190A (3')*	*FAM190A (5')*	Exons Duplicated
INS16	Ins	4	117597380	-	4	117609422	+	12.04			
INS18	Ins	5	37083312	-	5	37175248	+	91.94	*NIPBL (3')*	*C5orf42 (3')*	
INS19	Ins	5	174407766	-	5	174468477	+	60.71			
INS2	Ins	1	190587754	-	1	190724320	+	136.57	*RGS21 (3')*		
INS20	Ins	6	2046552	-	6	2063989	+	17.44	*GMDS (5')*	*GMDS (3')*	
INS21	Ins	6	11391453	-	6	11573487	+	182.03	*NEDD9 (5')*	*TMEM170B (3')*	
INS22	Ins	6	41476568	-	6	41569960	+	93.39			
INS24	Ins	7	104330914	-	7	104516312	+	185.4	*LHFPL3 (3') / LOC723809 (5')*	*MLL5 (5')*	***MLL5-LHFPL3***
INS25	Ins	8	5911073	-	8	5986829	+	75.76			
INS26	Ins	8	6480859	-	8	6586862	+	106	*MCPH1 (3')*	*AGPAT5 (5')*	***AGPAT5-MCPH1***
INS27	Ins	8	79925025	-	8	79950438	+	25.41			
INS3	Ins	2	10155274	-	2	10239261	+	83.99	*RRM2 (3')*	*C2orf48 (5')*	***Runthrough Csorf48-RRM2***
INS30	Ins	9	102203807	-	9	102265942	+	62.14	*C9orf30 (3')*	*TMEFF1 (5')*	***Runthrough TMEF1-C9orf30***
INS32	Ins	11	16826873	-	11	16879687	+	52.81	*PLEKHA7 (5')*	*PLEKHA7 (3')*	Exons Duplicated
INS33	Ins	11	57072794	-	11	57295287	+	222.49	*SMTNL1 (3')*	*CTNND1 (5')*	***CTNND1-SMNTNL1***
INS34	Ins	11	77389514	-	11	77443887	+	54.37			
INS35	Ins	11	111292861	-	11	111354277	+	61.42	*C11orf52 (3')*	*DIXDC1 (5')*	***DIXDC1-C11orf52***
INS37	Ins	12	27532204	-	12	27595557	+	63.35		*PPFIBP1 (5')*	
INS42	Ins	16	65695550	-	16	65828537	+	132.99	*C16orf70 (3')*	*FHOD1 (3')*	
INS43	Ins	16	79615251	-	16	79679344	+	64.09	*CENPN (3')*	*GCSH (3')*	
INS44	Ins	17	46134084	-	17	46194120	+	60.04	*ANKRD40 (5') / LUC7L3 (3')*	*C17orf73 (3')*	
INS45	Ins	17	72865900	-	17	72905624	+	39.72	*SEPT9 (3')*	*SEPT9 (5')*	Exons Duplicated
INS46	Ins	18	606353	-	18	631641	+	25.29	*CLUL1 (3')*	*CLUL1 (3')*	Exons Duplicated
INS48	Ins	19	10398770	-	19	10509924	+	111.15	*PDE4A (3')*		
INS49	Ins	19	16828242	-	19	16940309	+	112.07	*SIN3B (3')*	*CPAMD8 (3')*	

INS50	Ins	20	19112222	-	20	19210849	+	98.63	*SLC24A3 (3')*		
INS6	Ins	2	240060032	-	2	240105411	+	45.38			
INS7	Ins	3	49410887	-	3	49559868	+	148.98	*RHOA (5')*		
INS8	Ins	3	56444738	-	3	56449656	+	4.92	*ERC2 (5')*	*ERC2 (3')*	within intron
INV7	Inv	18	9767694	+	18	9773271	+	5.58	*RAB31*	*RAB31*	within intron
INV8	Inv	18	51504211	+	18	51510656	+	6.45			
INV9	Inv	20	9771873	+	20	9780917	+	9.04			
INV3	Inv	2	42519193	-	2	42533786	-	14.59		*KCNG3 (5')*	
INV5	Inv	13	76416077	+	13	76432425	+	16.35			
INV4	Inv	13	76399932	-	13	76416550	-	16.62			
INV6	Inv	14	73245679	-	14	73263451	-	17.77		*C14orf43(5')*	

Table 3.5. Small deletions and Duplications and inversions not segmented array CGH. Del=deletion, Ins=insertion – likely to be a tandem duplication duplication. Genes at deletion or duplication break points are listed and the end of the gene retained in a possible fusion transcript is also noted (3' or 5').

3.3.5. Broken and Predicted Fusion Genes in HCC1187

To produce a fusion transcript, the 5' end of one gene must be juxtaposed to the 3' end of another. When genes at breakpoints from the HCC1187 structural variants were combined several potential gene fusions were predicted.

I had carried out the bulk of this analysis before Stephens et al. (2009 using data from array painting and array CGH. Cytogenetically visible chromosome aberrations potentially produced five fusion genes: *PUM1-TRERF1, IQSEC2-DPYD, SGCZ-HRNR, RGS22-SYCP1, CTCF-SCUBE2*. Twenty-three additional genes broken but not predicted to be fused. Sub-microscopic duplications and deletions segmented by PICNIC produced eleven possible fusion genes: *SGK1-SLC2A12, RHOJ-SYNE2, IL17RD-HESX, PLXND1-TMCC1, SUSD1-ROD1, CTAGE5-SIP1.* Data from paired end sequencing indicated five further potential fusions: *AKAP13-SLC28A1, MLL5-LHFPL3, AGPAT5-MCPH1, Csorf48-RRM2, TMEF1-C9orf30.* Nine rearrangements were entirely within genes and traversed exons and potentially disrupted gene function: *AGPAT6, MED13L, RB1, F8, EXOC4, FAM190A, PLEKHA7, SEPT9, CLUL1.* Sub-microscopic rearrangements from array CGH and paired end sequencing data broke as many as 77 other genes

All the potential gene-fusions were investigated by RT-PCR and five were shown to be expressed: *PUM1-TRERF1, CTCF-SCUBE2, RHOJ-SYNE2, CTAGE5-SIP1 and ROD1-SUSD1.* Several of these genes may be relevant to breast cancer.

3.4. Expressed Fusion Genes in HCC1187

In the following section I describe how the expressed fusion transcripts were identified. Further investigations recurrence in other samples and expression levels are described in Chapter 6.

3.4.1. *PUM1-TRERF1*

*Pumilio homolog 1 (PUM1) i*s fused to *transcriptional regulating factor 1 (TRERF1)* by to a t(1;6) translocation junction present on chromosomes A and D. At the cDNA level, exon20 of *PUM1* is fused to exon5 of *TRERF1*. I did not observe any different splice isoforms. The fusion transcript is predicted to cause a frame shift in *TRERF1* (Figure 3.4). The *PUM1-TRERF1* breakpoint regions are interesting as SNP6 array CGH showed a gain in copy number approximately 140kb in length immediately distal to the *PUM1* breakpoint and 1Mb copy number gain immediately proximal to the *TRERF1* breakpoint (Figure 3.3). A FISH experiment showed the *PUM1-TRERF1* fusion was present only on the two derivative chromosomes (chromosomes A and D) at the t(1;6) translocation breakpoint. Interphase cells showed two or three untranslocated *PUM1* signals (chromosomes E and H), and one or two untranslocated*TRERF1* signals (chromosome I). An extra *PUM1* signal was occasionally seen in metaphase chromosomes and may have been due to duplication of peak E or H in culture. The modal number of fused signals was four, so it is likely the *PUM1-TRERF1* fusion region and surrounding loci had been duplicated during the evolution of the tumour or cell line. As derivative chromosomes A and D share the same translocation breakpoint, it is probable a single chromosome originally carried the *PUM1-TRERF1* fusion and subsequently duplicated. The 1Mb gain proximal to the *TRERF1* break contains several other genes including cyclin D3 (*CCND3*) and forkhead-box P4 (*FOXP4*), and it is also possible one of these genes is driving the duplication of the region.

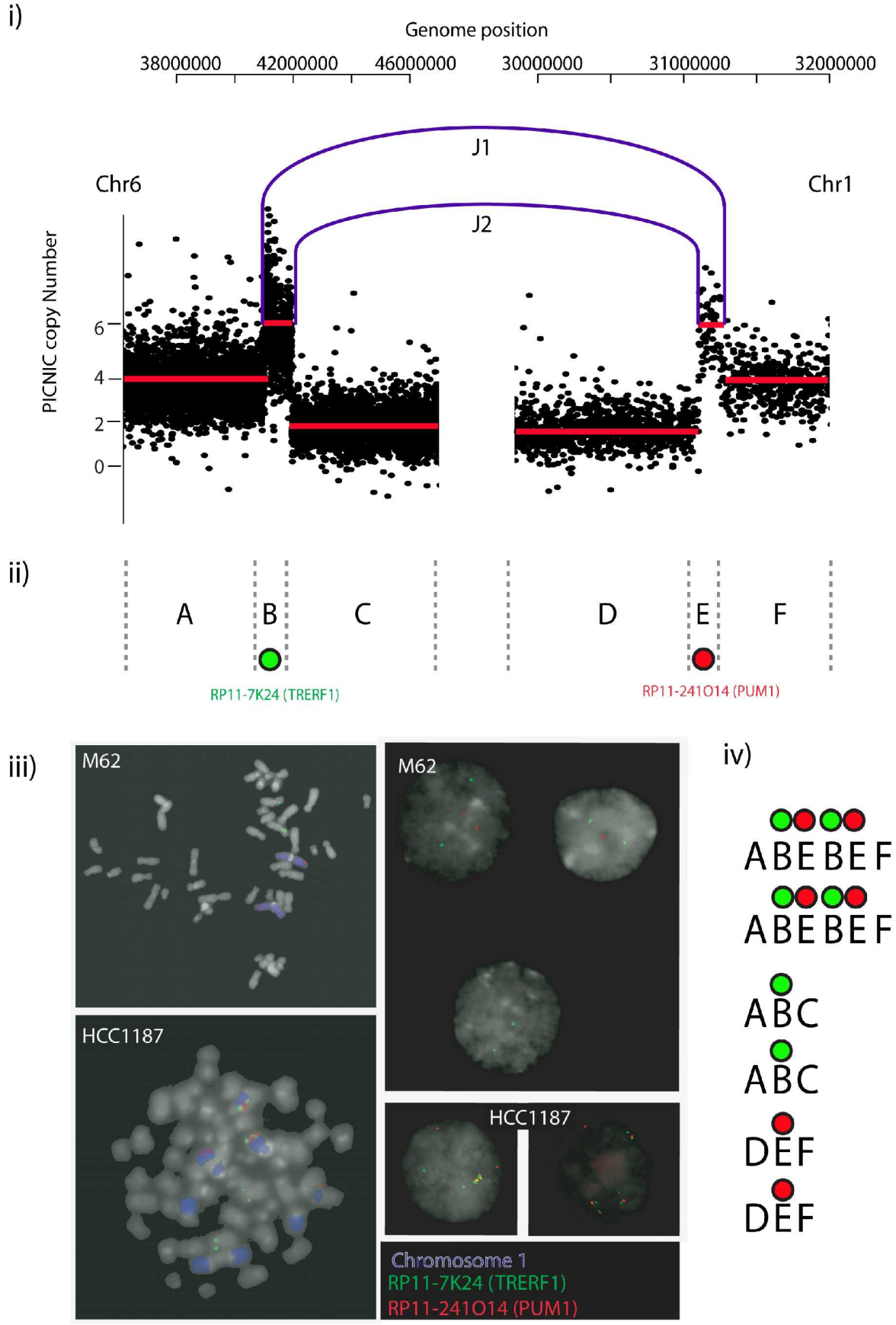


Figure 3.3. Translocation t(1;6) that caused the PUM1-TRERF1 fusion.

Figure 3.3. Translocation t(1;6) that caused the *PUM1-TRERF1* fusion. i) SNP6 array CGH segmented with PICNIC. Black dots are individual probe loci, red lines are segmented copy number. Purple lines are genomic junctions from Stephens et al. (2009). J1=StephensDIF1; chr1:31183855 - ;chr6:42356186 + and J2; StephensDIF2 chr1:31334253 + ;chr6:41406061 -. ii) Regions of chromosomes 1 and 6 with the same segmented copy number named A-C and D-F. Positions of BACs used for FISH are indicated with green and red circles. iii) FISH of the t(1;6) translocation. M62 normal lymphoblastoid cell line metaphase spread shows two normal copies of chromosome 1 (blue), each with a single red FISH signal on 1p (*PUM1*). Two C-group chromosomes have a single green signal on their p-arm (*TRERF1*). Interphase nuclei from M62 show red,red,green,green signals. HCC1187 metaphase spread shows numerous segments of chromosome 1. There are two unpaired green signals and two unpaired red signals and two fused red-green signals, presumably at the t(1;6) translocation junction. HCC1187 interphase nuclei showed a modal pattern of two unpaired reds, two unpaired greens and two red-green-red-green signals. iv) The most probable explanation for the patterns observed is a tandem duplication of the fused region at the translocation break point.

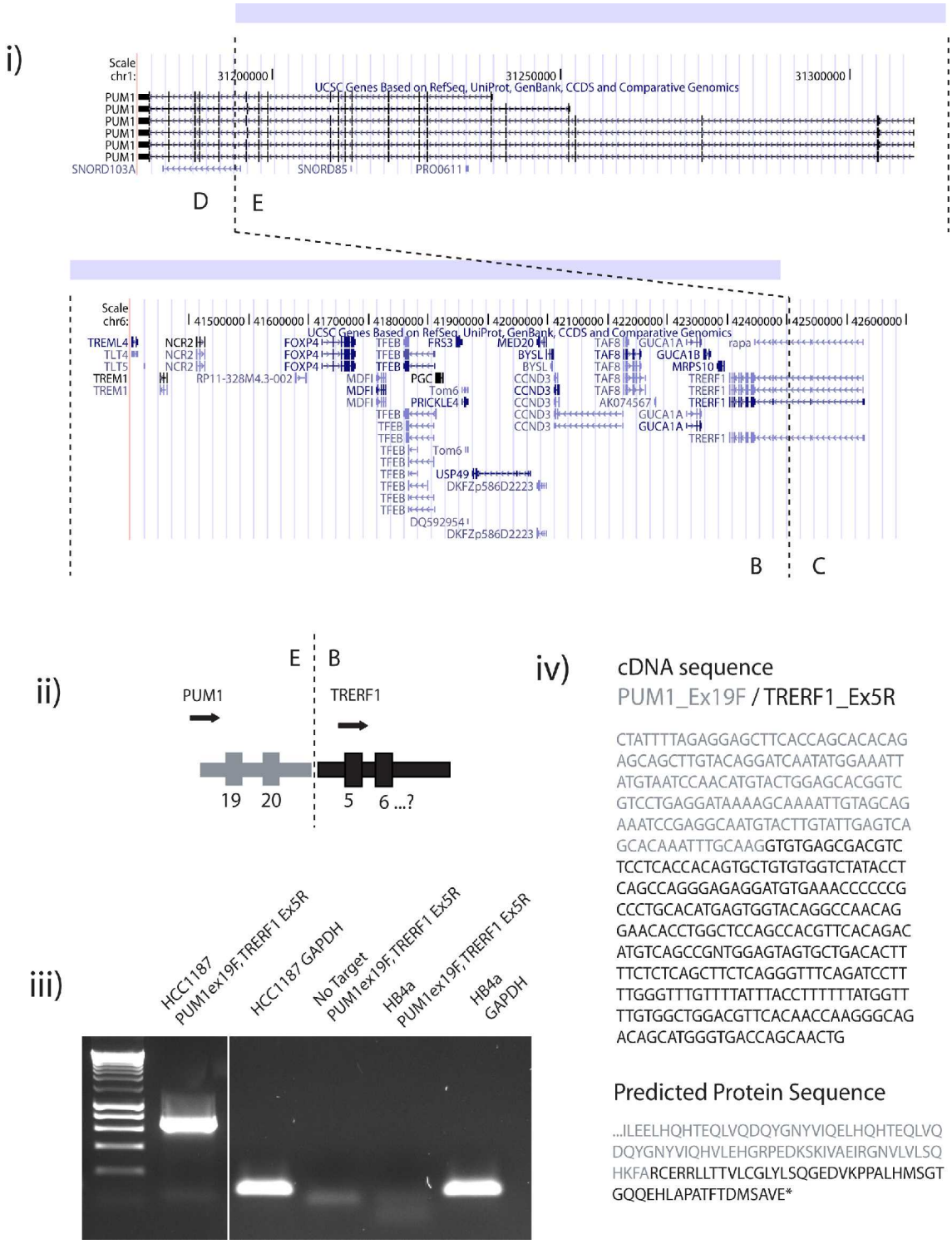


Figure 3.4. RT PCR of the *PUM1-TRERF1* fusion transcript. i) The genomic loci of *PUM1* and *TRERF1*. Dotted lines indicate chromosome breaks. ii) Pictorial representation of fusion transcript. Uncharacterised exon is labelled 'x'. iii) RT-PCR across the fusion transcript junction. The fusion junction was cloned and sequenced giving the sequence shown in iv) *PUM1* exon 20 is fused with *TRERF1* exon 5.Exons

are names as for *PUM1*-001 (ENST00000257075) and *TRERF1*-001 (ENST00000372922). This and all subsequent PCRs used a 200 base-pair ladder.

PUM1 is a member of the evolutionarily-conserved PUF family of RNA binding proteins. The PUF family are characterised by a highly conserved C-terminal RNA-binding domain, composed of eight tandem repeats. PUF-family proteins bind to motifs in the 3'UTR of specific target mRNAs and repress their translation (Spassov and Jurecic, 2003). The mammalian PUF family members are poorly characterized and their mRNA targets are largely unknown although an interesting finding is that *PUM2* may modulate translation of members of the MAPK signalling pathway (Lee et al., 2007). Some have suggested family members interact with the miRNA regulatory system (Galgano et al., 2008). *PUM1* binding, in particular, is required to down-regulate p27 for cell cycle entry and does this by recruiting micro RNAs, miR-221 and miR-222 (Kedde et al., 2010). Conversely, it has been suggested that *PUM1* is expressed at a remarkably constant level over large data sets. This expression profile makes it suitable as a housekeeping gene for investigating differential gene expression in cancers (Gur-Dedeoglu et al., 2009).

The potential role of *TRERF1* in breast cancer is not clear. *TRERF1* encodes a zinc-finger transcriptional regulating protein that interacts with *CBP/p300*. Insertional mutagenesis of the oestrogen-dependent cell line, ZR-75-1, revealed several candidate BCAR (breast cancer anti-estrogen resistance) genes that were thought to underpin oestrogen independence, one of which was *TRERF1/BCAR2* (Dorssers and Veldscholte, 1997; van Agthoven et al., 2009). In this case, retroviral insertion caused increased *TRERF1* expression and the authors postulated that *TRERF1* exerted a dominant growth control and acted as an oncogene by driving oestrogen-independent growth. But conversely, *TRERF1* (TreP-123) acts with steroidogenic factor 1 (*SF-1*) and progesterone receptor to induce expression of G(1) cyclin-dependent kinase inhibitors *p21(WAF1) (p21)* and *p27(KIP1) (p27)*. Knockdown of *TRERF1* in T-47D and MDA-MB-231 cell lines enhanced cell proliferation and lowered p21 and p27 mRNA levels (Gizard et al., 2005, 2006).

There are full length *PUM1* and *TRERF1* transcripts present in HCC1187 (not shown), and as the fusion transcript appears to be out of frame, it is unlikely that the fusion is functional.

3.4.2. *CTCF-SCUBE2*

A near-balanced translocation caused *CTCF* to fuse with *SCUBE2*. The t(11;16) translocation is not exactly reciprocal as it contains a genomic shard – a 1.3kb piece of chromosome 11 sandwiched between the chromosome 11 and 16 translocation break points (Figure 3.5).

The second (untranslated) exon of *CTCF* is fused with second exon of *SCUBE2*. An uncharacterised exon upstream of the annotated *SCUBE2* open reading frame is also present in one of the fusion transcripts (Figure 3.6). The transcriptional start site for all the protein-coding isoforms of *SCUBE2* is in the first exon. As the fusion transcript does not contain the first exon of *SCUBE2* (which contains the translational start site) and the *CTCF* portion of the transcript is untranslated it is not clear where translation would start. I used a bioinformatic algorithm to predict the best Kozak consensus site (Liu et al., 2005) This appeared to be the first AUG in *SCUBE2* exon 2. If we assume this is the true start of translation for the fusion transcripts, the codon 52 leads to a premature stop codon.

CTCF protein is involved in numerous gene-regulation functions including activation, repression, silencing and chromatin insulation (Ohlsson et al., 2001). In the fusion transcript only the first two untranslated exons are present. There are two presumably unbroken copies of the locus on chromosomes Y and R so this translocation is unlikely to have caused loss of *CTCF*.

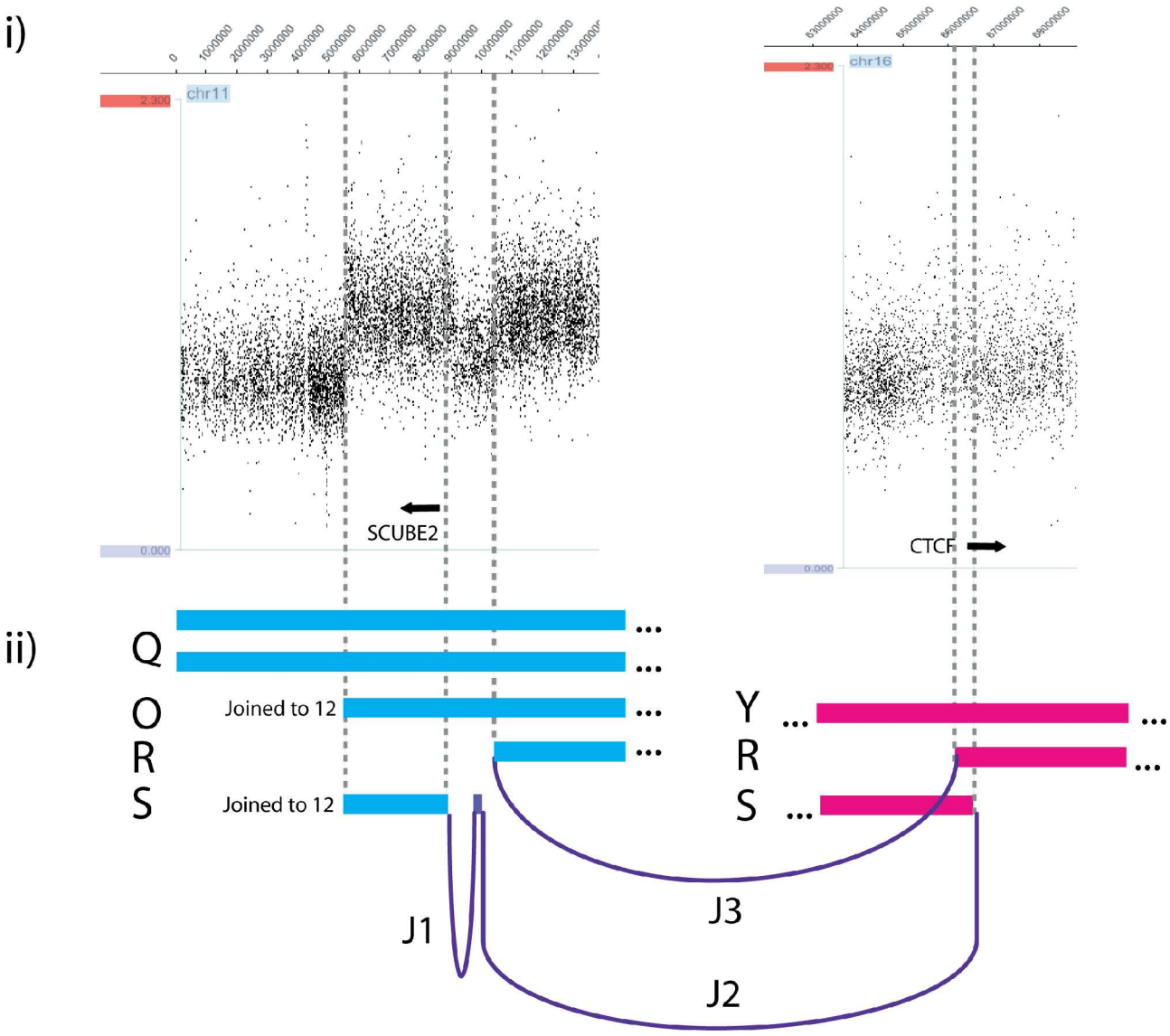
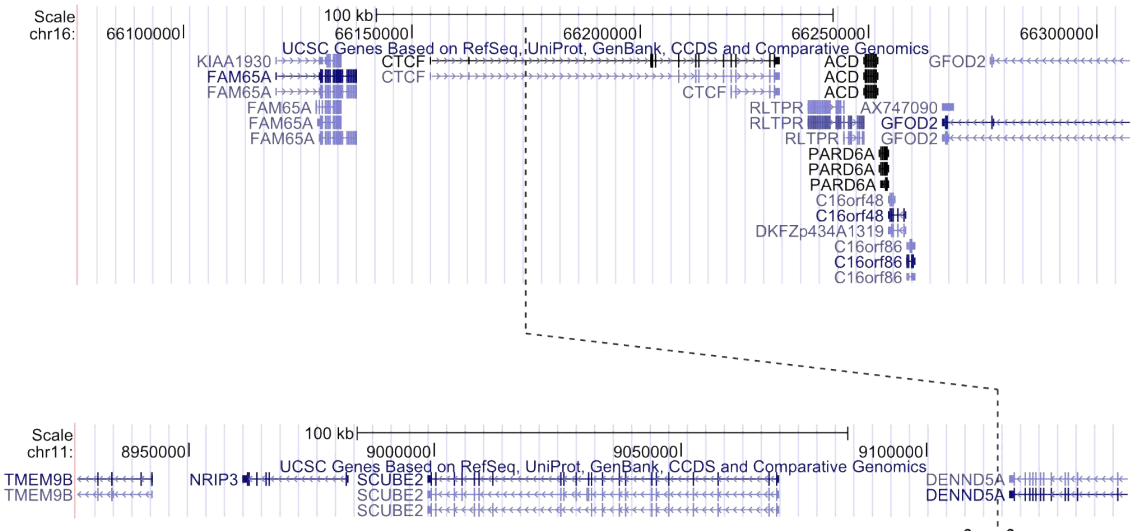
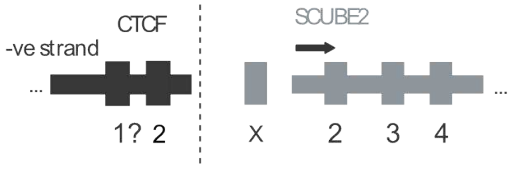


Figure 3.5. Genomic structure of the *CTCF-SCUBE2* regions. Upper scatter plots are SNP6 array CGH from chromosomes 11 and 16 (segmented copy number is not shown as PICNIC missed the copy number step at chr11:9.1Mb). Purple boxes are array painting segments from chromosomes Q,O,R and S; pink boxes are segments of chromosome 16 from chromosomes R, S and Y. Purple lines are genomic junctions: J1 = chr11:9108544 +ve ; chr11:10519531 +ve, J2 = chr11:10520915 +ve ; chr16:66198160 -ve , J3 = chr16:66143420 -ve ; chr11:10438674 +ve. The genomic shard is shown between junctions J1 and J2. Junction sequences were cloned by Dr. K.D. Howarth.

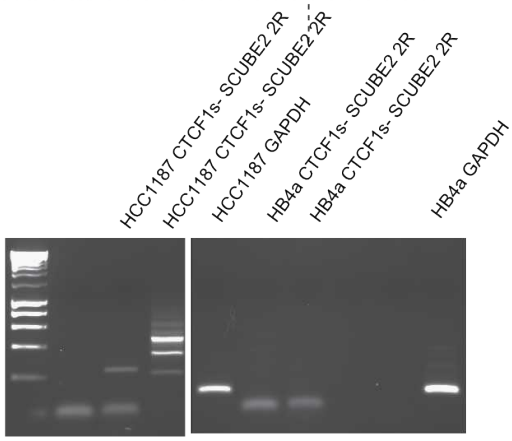
i)



ii)



iii)



iv) cDNA Sequence

GCTGGGTTCTATTTTCCCTCCTCAAACTGACTTT
GCAGCCACGGAGAG

CTCF untranslated exon 2

GATCTGGCGTTTGCTGCCCACGCTGGAGATGTGG
TGGTGTGATCACAGCCTCACTGCAACCTCTGCTCT
GCGTCCCAGGTTCGACTGATCCTCCTGCTTCAGCC
TCCTGAGTAGCTGGGACTACA

SCUBE2 uncharacterised exon
chr11:9100896-9101017

GATGTAGATGAGTGTGCCCAAGGGCTAGATGACTG
CCATGCCGACGCCCTGTGTCAGAACACACCCACCT
CCTACAAGTGCTCCTGCAAGCCTGGCTACCAAGGG
GAAGGCAGGCAGTGTGAGGACATCGATGAATGTG
GAAATGAGCTCAATGGAGGCTGTGTCCATGACTGT
TTGAATATTCCAGGCAATTATCGTTGCACTTGTTTT
GATGGCTTCATGTTGGCTCATGACGGTCATAATTGT
CTTGATGTGGACGAGTGCCTGGAGAACAATGGCG
GCTGCCAGCATACCTGTGTCAACGTCATGGGGAGC
TATGAGTG

Best Kozac consensus start site

SCUBE2 exons 2,3,4 ...

Translation from the best Kozak consensus
start site results in stop at codon 52

MTAMPTPCVRTHPPPTSAPASLATKGKAGSVRTSMNV
EMSSMEAVSMTV-

Figure 3.6. *CTCF-SCUBE2* fusion transcript.

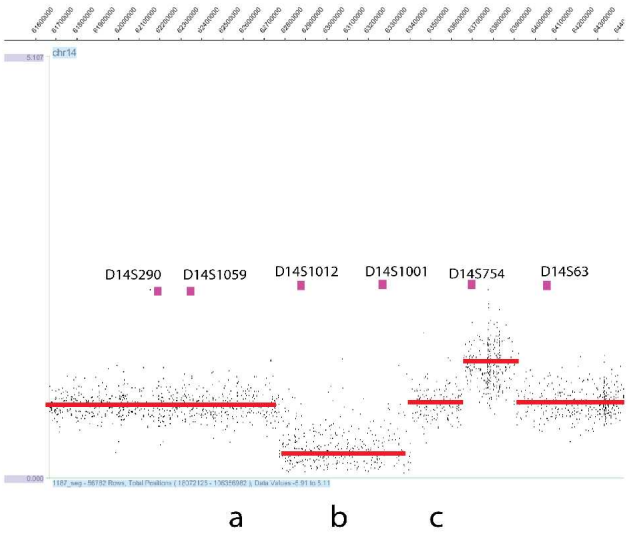
Figure 3.6. *CTCF-SCUBE2* fusion transcript. i) The genomic loci of *CTCF* and *SCUBE2*. Dotted lines indicate chromosome breaks. ii) Pictorial representation of fusion transcript. Uncharacterised exon is labelled 'x'. iii) RT-PCR across the fusion transcript junction. Multiple bands indicate other isoforms may be present. The upper two bands were cloned and sequenced giving the sequence shown in iv) *CTCF* exon 2 is fused with *SCUBE2* exons 2, 3 and 4. The larger fusion transcript includes an uncharacterised *SCUBE2* exon. Both are predicted to form a truncated peptide. Exons are named as in *CTCF*-001 (ENST00000264010) and *SCUBE2*-001 (ENST00000520467).

SCUBE2 (Signal peptide-CUB-epidermal growth factor–like domain-containing protein 2) is a poorly characterised member of the evolutionarily conserved *SCUBE* family. All members contain an N-terminal signal peptide sequence followed by nine EGF-like repeats, and a CUB domain at the C-terminus. CUB proteins are involved in various cellular processes including complement activation, developmental patterning, tissue repair, angiogenesis, cell signalling, fertilisation, haemostasis, inflammation, neurotransmission, receptor-mediated endocytosis, and tumour suppression (Bork and Beckmann, 1993). *SCUBE2* is a secreted surface-anchored glycoprotein that can interact with *SHH* (Sonic Hedgehog) and its receptor *PTCH1* (Patched-1) (Tsai et al., 2009). It has also been suggested that the carboxy terminal of *SCUBE2* can bind and antagonize bone morphogenetic protein activity (Cheng et al., 2009). In the normal breast, *SCUBE2* is expressed in vascular endothelial and mammary ductal epithelial cells and *SCUBE2* is also expressed in a high proportion of primary breast tumours (Cheng et al., 2009). Cheng et al. observed that over-expression in the MCF7 cell line suppressed cell proliferation.

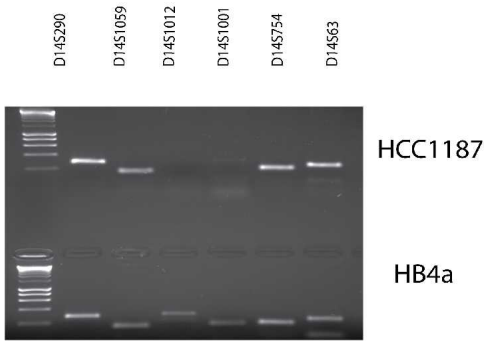
3.4.3. *RHOJ-SYNE2*

This fusion was formed by a homozygous deletion. Array CGH indicated the deletion was homozygous and PCR using STS markers confirmed a 600kb deleted region was entirely absent from HCC1187. The bounds of the homozygous deletion were mapped using PCR markers at 1kb intervals in the regions indicated by the SNP array (not shown). This allowed me to design PCR primers across the genomic junction (Figure 3.7).

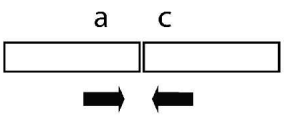
i)



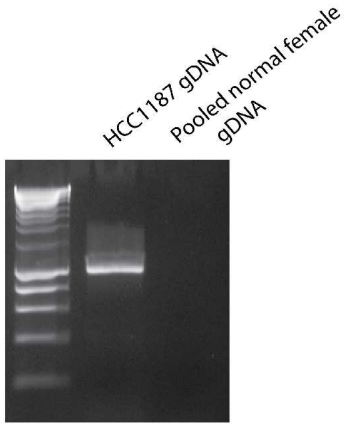
ii)



iii)



iv)



v)

TTTCTTA ATTATTTCAa TGTTTTaAAA GCAGCAATCA TGAACAGGTTTTAT
TGTTTT AGATTTTTCC ACATATAACA AACCTTCCTT TCTACATCCTCTCTT
CTTAC TCTGACCCTA CAGAATTCTC TTTTACGGCC AACAAGTTCCAAGA
GAGGCT ACAGGCTCTG TTTCATGCAT CACCTGACAT AGTACCTTTTCTT
GGAACAA GGTCTCTTCA GTAAAGCCTG CTGTGTAAGC CAGCTATGGTG
GTAGAATAC CCAGCTGAAT GCCTTctgga atatttttt caaggtctgtcaTATAT
GAT TCATTTCTGA CTGTCCCCTA AACACCTGAT ACCTAGCCCCTTGTGT
ACTG CACCCTTGAA CTGAATGTAA TAACTAGCAT TTGTTACCATGTGTT
CTATA GGTACTATGC TTTCTGTTAC CTGTATTATC TCAGTAATATAAGTAC
GTGG GATTTTAGTT TCACTTTTAC AGATGAGGAA ACTCTACTCAGAGA
GGTTAA GTAATTTGCC CAAAATAACA CCTCTAGTAT ATGGCAAAGCTAG
AATTTGA ACCAGGGGGTC AGACTGATTT CAGAACCCAA GCTTTTAACC
ACTACACTAT ACCACAGACA TTTCTAAAAT TCACTTGTTC CTTTAGAATG
TATGCACACG GAGTATTACT CAAGATCAAC ACCCCAGGAA ATCATATTTA
ACCCTATCAG CACATGGAGG TCCTTTTGGT TCTGCTCTAG TCCAGGGGC
ATAAACAGGCA GACTCATGAT ATCCCAGATT CATTTATTTT CCAGCTGTT
TTTTTCTCATC TACTGTCATT TTACTTTTTT AAATTTAAAT TTAAATTTTTA
ATTTTACTT TTTTGTCTTT AACCAGTTAG AATGCTAAGA TTTGATGGTAA
ATAAGACC AAGTATCAAA TATTTAGTAC GAACTCTCCA TTTATCAATGCG
GTTGCACA GAAATGGATG GTt

chr14:62772766 joined to chr14:63406983

Figure 3.7. The RHOJ-SYNE2 genomic locus. i) Segmented SNP6 array CGH. Purple squares are STS markers. ii) PCR of STS markers confirms the deletion in homozygous. iii) schematic of genomic junction formed by a homozygous deletion. iv)

PCR across the genomic junction. v) sequence across the genomic junction. Blue region is from region 'a', red from region 'c'. The black region is non-templated sequence at the genomic junction.

The homozygous deletion removed exons 2-5 of *RHOJ* (Ras homolog gene family, member J) and the first exon of *SYNE2* (spectrin repeat nuclear envelope 2). The RT-PCR product from the HCC1187 *RHOJ-SYNE2* showed that exon1 of *RHOJ* was fused to exon 2 of *SYNE2.* This fusion is predicted to cause a frame shift in the *SYNE2* transcript (Figure 3.8). The resulting peptide has a stop as the 7th codon of the *SYNE2* portion of the fusion transcript. I performed further RT-PCR using the *RHOJ* exon 1 forward primer with a number of reverse primers in *SYNE2* extending as far as exon 25. PCR products were obtained up to *SYNE2* exon 7, but no further. No other splice isoforms were apparent.

RHOJ belongs to the Rho family of small GTP-binding proteins, but not much is known about its function. The fusion causes a homozygous loss of *RHOJ*. The *SYNE-2* gene encodes numerous isoforms spectrin repeat family proteins by way of tissue-specific alternative splicing and transcription initiation. Approximately 20 isoforms ranging in size from several kDa to 1.10 MDa have been described and more probably remain to be characterized (Wheeler et al., 2007). The nesprin genes (*SYNE1, SYNE2* and *SYNE3*) encode a family of ubiquitously-expressed multi-isomeric intracellular proteins. Nesprins interact with emerin and lamin A/C to form a network that links the nucleoskeleton to the inner nuclear membrane, outer nuclear membrane, membraneous organelles, the sarcomere and actin cytoskeleton (Zhang et al., 2007). Recently, Warren et al (2010) showed that isoforms of this structural protein also may have a play a role in signalling as it can tether *ERK1/2* to *PML* nuclear bodies. Knockdown of one of these *SYNE2* isoform resulted sustained *ERK1/2* signal which increased proliferation (Warren et al., 2010).

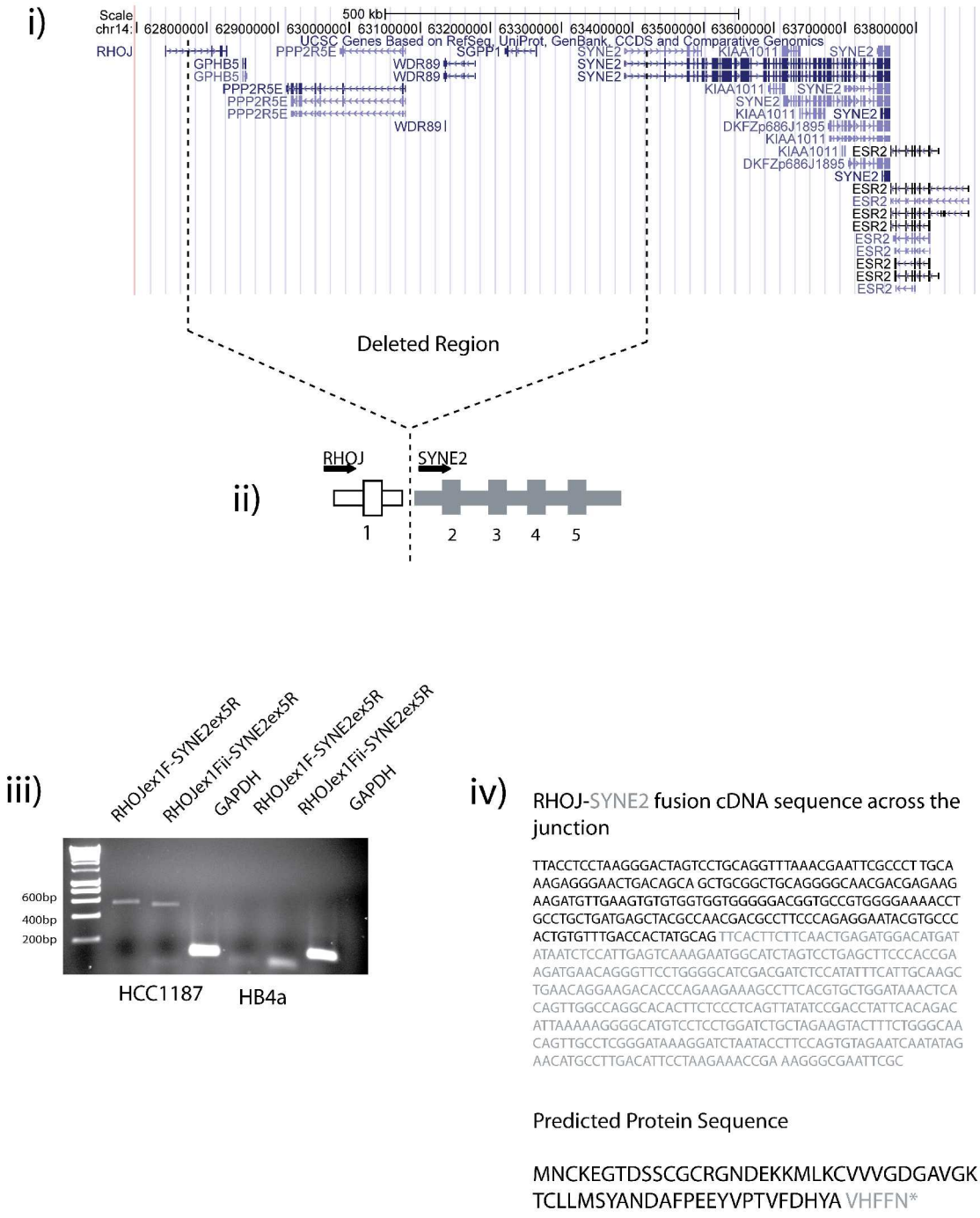


Figure 3.8. *RHOJ-SYNE2* fusion transcript. i) Genomic region encompassing the homozygous deletion (dotted lines). ii) schematic representation of the fusion transcript. iii) RT-PCR shows the *RHOJ-SYNE2* fusion transcript is expressed. iv) Sequence of the fusion transcript and predicted protein sequence. Exons are named for *RHOJ*-001 (ENST00000316754) and *SYNE2*-205 (ENST00000357395).

The deletion also caused loss four other genes: *GPHB5* (glycoprotein hormone beta 5), *PPP2RSE* (Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit epsilon isoform), *WDR89* (WD repeat domain 89) and *SGPP1* (sphingosine-1-phosphate phosphatase 1), and loss of one of these may also play a role in carcinogenesis.

3.4.4. *CTAGE5-SIP1*

A small tandem duplication fused *CTAGE5* exon 20 with *SIP1* exon 10 (Figures 3.9 and 3.10). Initial inspection indicated that the fusion transcript is out of frame. However, exon 10 is the final exon of *SIP1*, so although the frame shift causes a premature stop codon, it is probable that the fusion transcript is translated. The *SIP1* portion of the fusion is predicted to contribute only seven amino acids and its 3' UTR to the hypothetical protein.

CTAGE5 (cutaneous T-cell lymphoma-associated antigen 5) is also called *MGEA6* (meningioma expressed antigen 6) as it was first identified as a menigioma cell-surface antigen (Heckel et al., 1997). *CTAGE5* is overexpressed in meningioma and glioma relative to normal brain (Comtesse et al., 2002) but little is known about its function or its potential role in breast cancer.

Switching of a 3'-UTR by gene fusion is observed in several soft tissue cancers. In these neoplasms, the final exon of the oncogene, *HMGA2,* can be replaced with last few exons and the 3'-UTR of a number of different genes including *RAD51L1, FHIT* and *LPP*. The 3' UTR of wild type *HMGA2* has a let-7 micro RNA binding site which targets the mRNA for degradation. The fusion transcript lacks the micro RNA binding site so allows *HMGA2* mRNA to persist within the cell eventually leading to higher levels of *HMGA2* protein (Mayr et al., 2007). I investigated the analogous possibility with *CTAGE5* and *SIP1*. Quantitative real time PCR showed, however, that *CTAGE5* was expressed at a level similar to control cell lines HMT-3552 and HB4a (not shown).

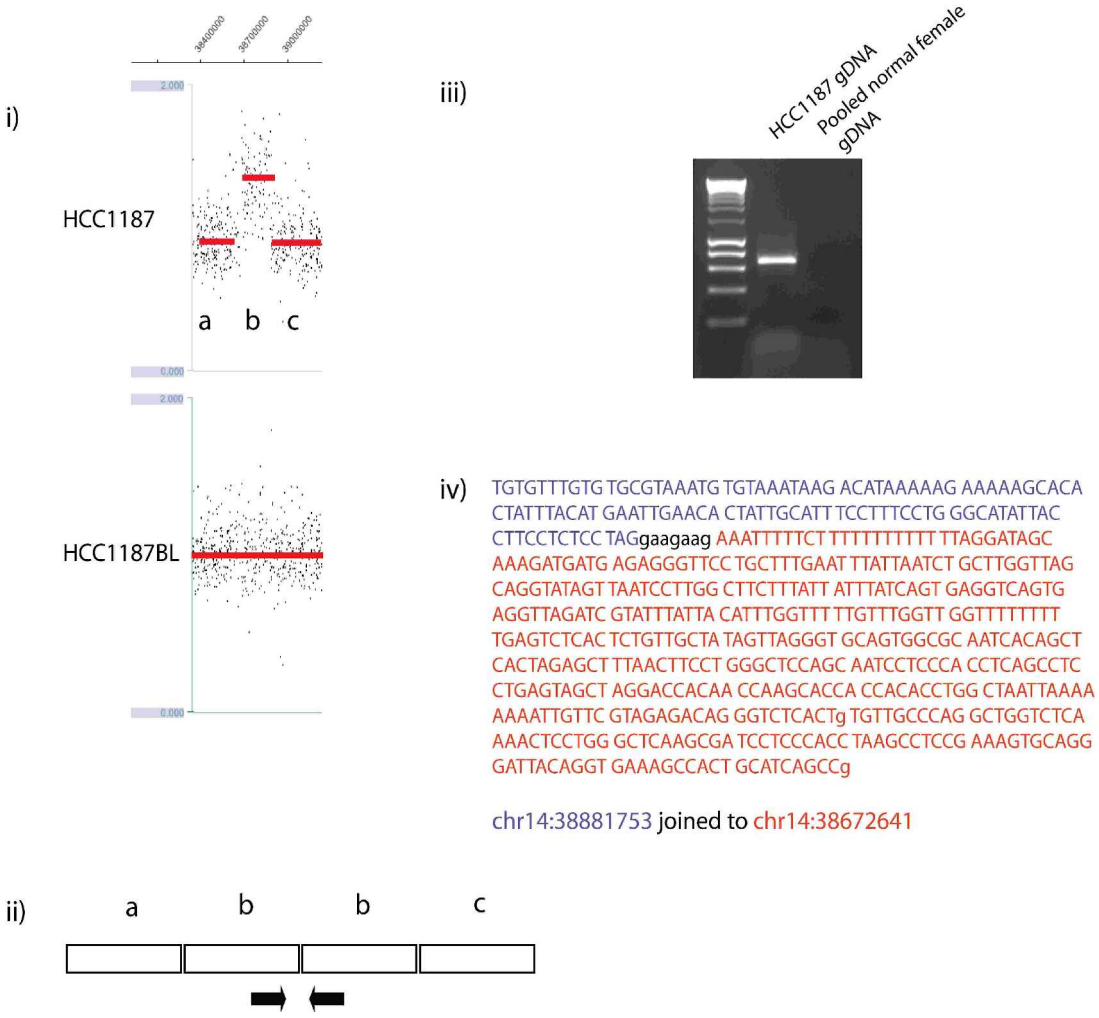


Figure 3.9. The *CTAGE5-SIP1* genomic locus. i) Segmented SNP6 array CGH shows a tandem duplication that was absent from the matched normal lymphoblastoid line HCC1187BL. ii) schematic of genomic junction formed by the tandem duplication. Black arrows indicate position of PCR primers for iii) PCR across the genomic junction. iv) sequence across the genomic junction.

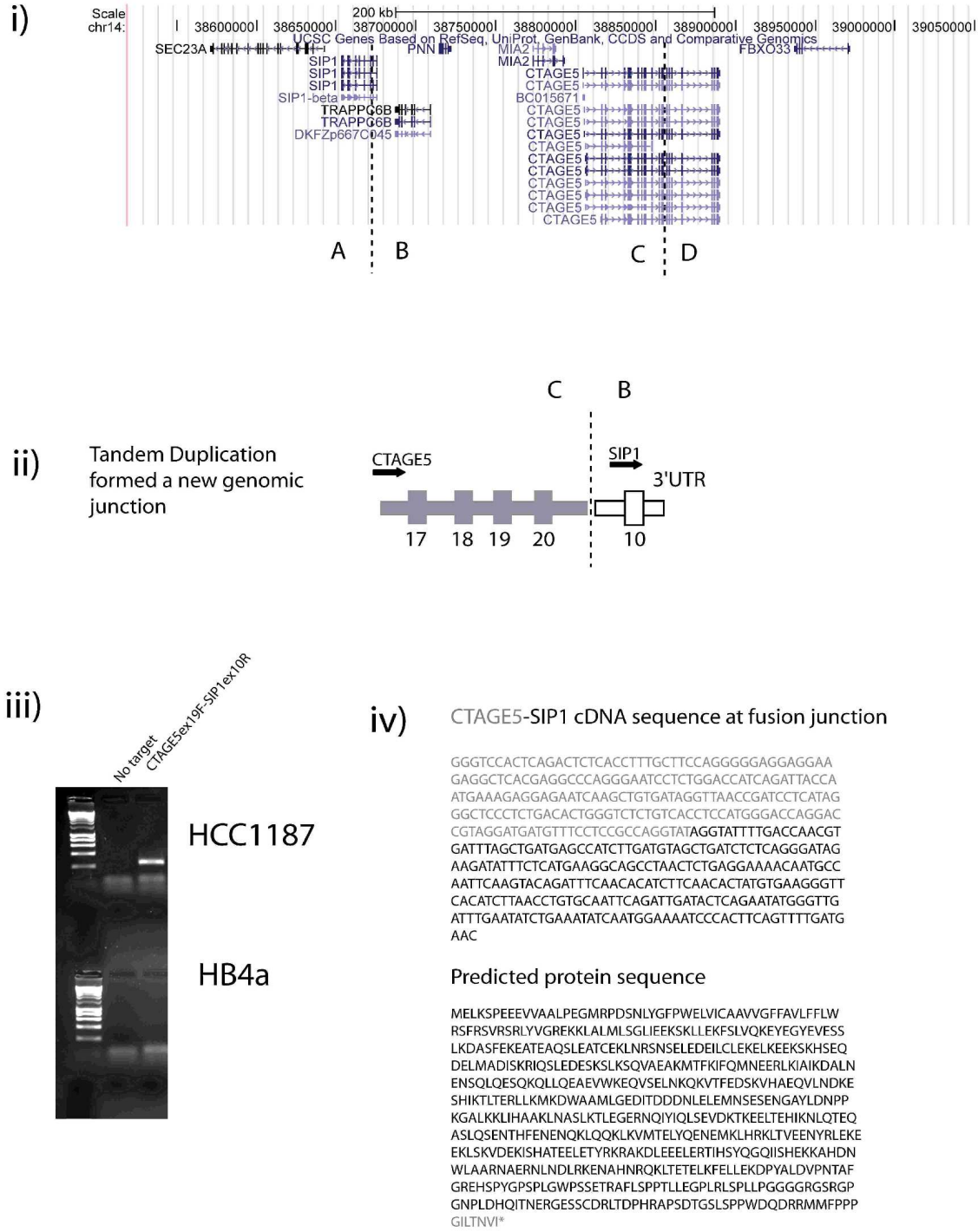


Figure 3.10. *CTAGE5-SIP1* fusion transcript. i) Genomic region encompassing the tandem duplication (dotted lines). ii) schematic representation of the fusion transcript. Iii) RT-PCR shows the *CTAGE5-SIP1* fusion transcript is expressed. iv) Sequence of the fusion transcript and predicted protein sequence. Exons are as in *CTAGE5*-001 (ENST00000341749) and *SIP1* (ENST00000308317).

3.4.5. *SUSD1-ROD1*

A small tandem duplication caused a *SUSD1-ROD1* fusion. *SUSD1* exon 6 is fused to *ROD1* exon2 at the cDNA-level. The *ROD1* portion of the transcript is predicted to undergo a frame shift (figures 3.11 and 3.12).

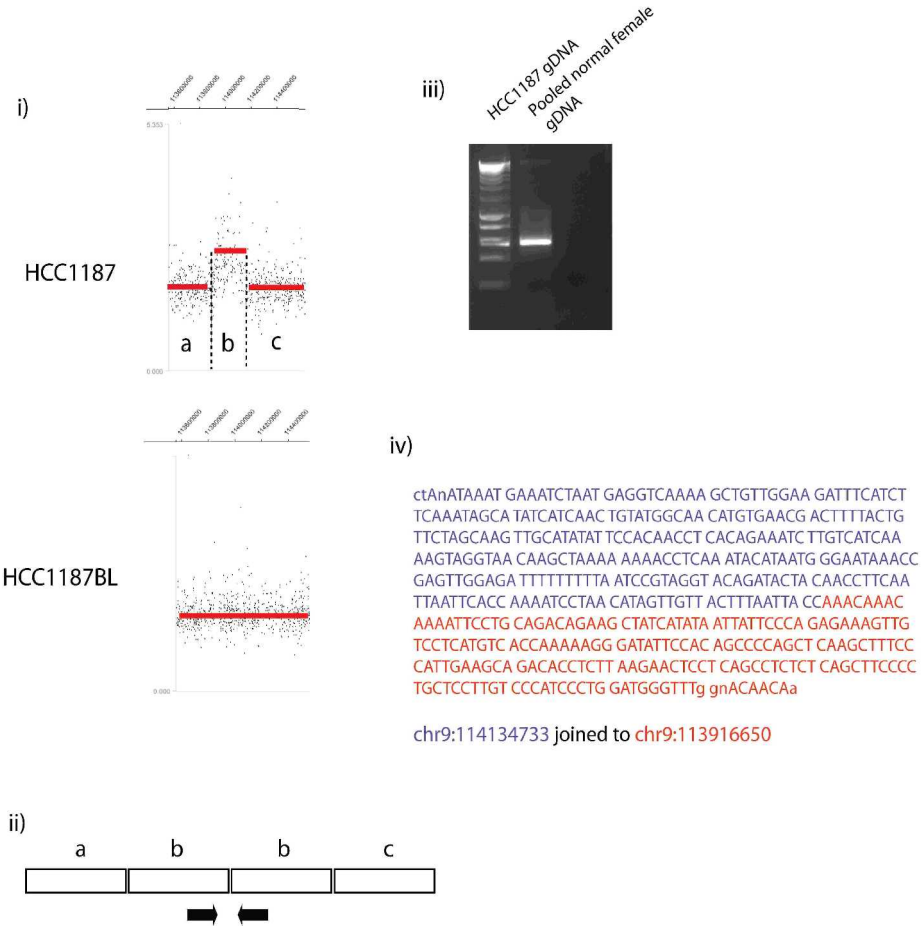


Figure 3.11. The *SUSD1-ROD1* genomic locus. i) Segmented SNP6 array CGH shows a tandem duplication that was absent from the matched normal lymphoblastoid line HCC1187BL. ii) schematic of genomic junction formed by the tandem duplication. Black arrows indicate position of PCR primers for iii) PCR across the genomic junction. iv) sequence across the genomic junction.

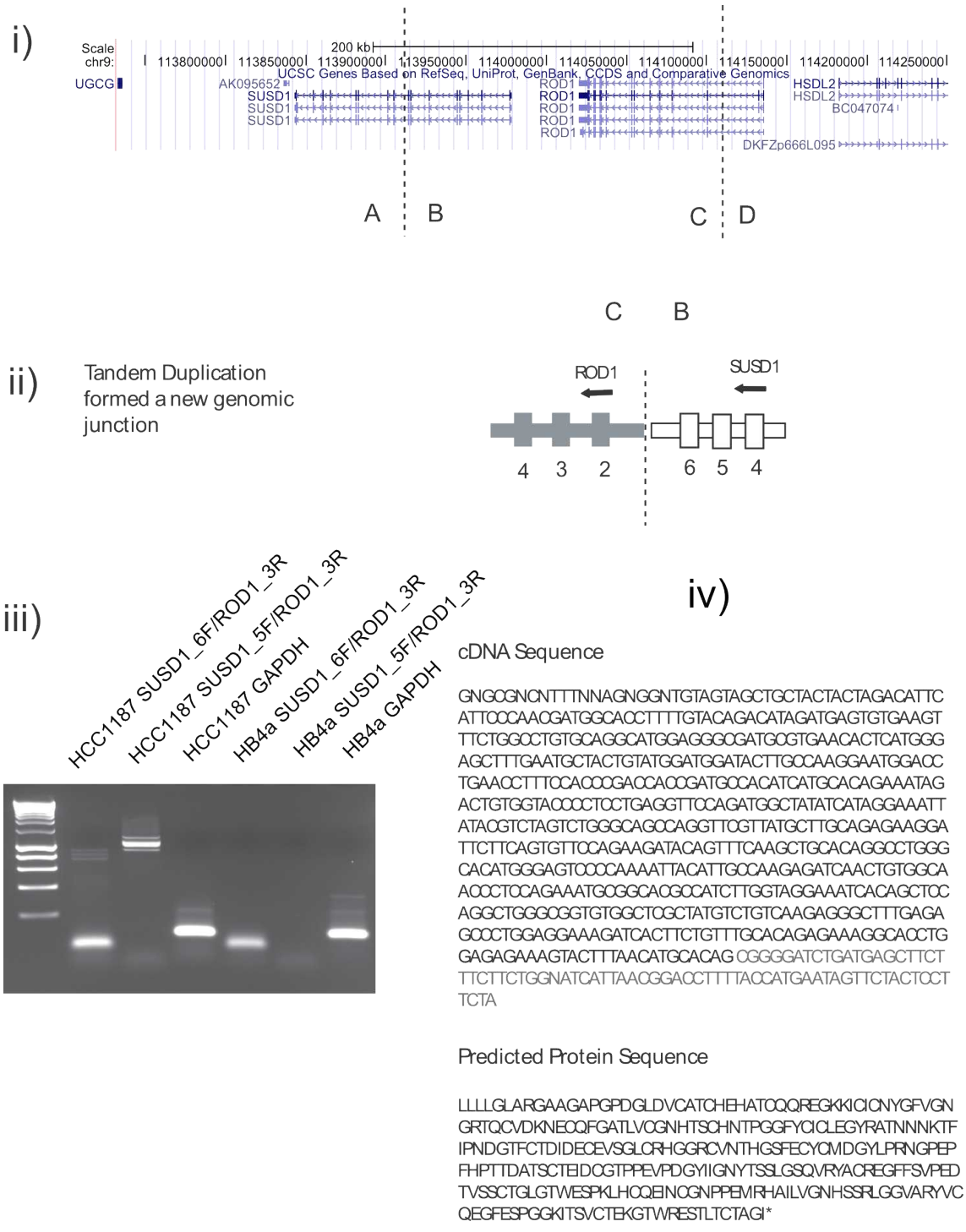


Figure 3.12. SUSD1-ROD1 fusion transcript. i) Genomic region encompassing the tandem duplication (dotted lines). ii) schematic representation of the fusion transcript. Iii) RT-PCR shows the CTAGE5-SIP1 fusion transcript is expressed and that there may be several isoforms present (needs more labels). iv) Sequence of the fusion transcript and predicted truncated protein sequence.

Nothing is known about *SUSD1* (sushi domain containing 1). *ROD* (regulator of differentiation 1) is a functional homologue of *S.pombe nrd1*, an RNA-binding protein that suppresses differentiation (Sadvakassova et al., 2009). Little is known about its function in humans.

As the fusion was formed by a tandem duplication, normal copies of each gene are retained. Unless some type of dominant-negative mechanism is operative it is difficult to postulate on effect for *SUSD1-ROD1.*

3.4.6. *PLXND1-TMCC1*

Another tandem duplication juxtaposed *PLXND1* to *TMCC1*.The fusion transcript is predicted to be in frame (Figure 3.13 and 3.14). *PLXND1* exon 13 is fused with *TMCC1* exon 4. *PLXND1* (*Plexin D1*) a single-pass transmembrane receptor and along with its ligand, semaphorin 3E, it plays a role in the growth of blood vessels (Sakurai et al., 2010). *PLXND1* is expressed on tumour vessels and tumour cells in a number of different tumours (Roodink et al., 2005). *PLXND1* expression is strongly correlated with both invasive behaviour and metastasis in melanoma (Casazza et al., 2010; Roodink et al., 2009) and *PLXND1* is generally expressed in solid tumour musculature but not normal musculature (Roodink et al., 2009). Nothing is known about the function of *TMCC1* (transmembrane and coiled-coil domain family 1). The chimeric protein (if translated) appears to lack the cytoplasmic Plexin domain which is involved in downstream signalling pathways, by interaction with proteins such as *Rac1*, *RhoD, Rnd1* and other Plexin family members (Letunic et al., 2009). It is, therefore, not clear how the chimera may act as an oncoprotein.

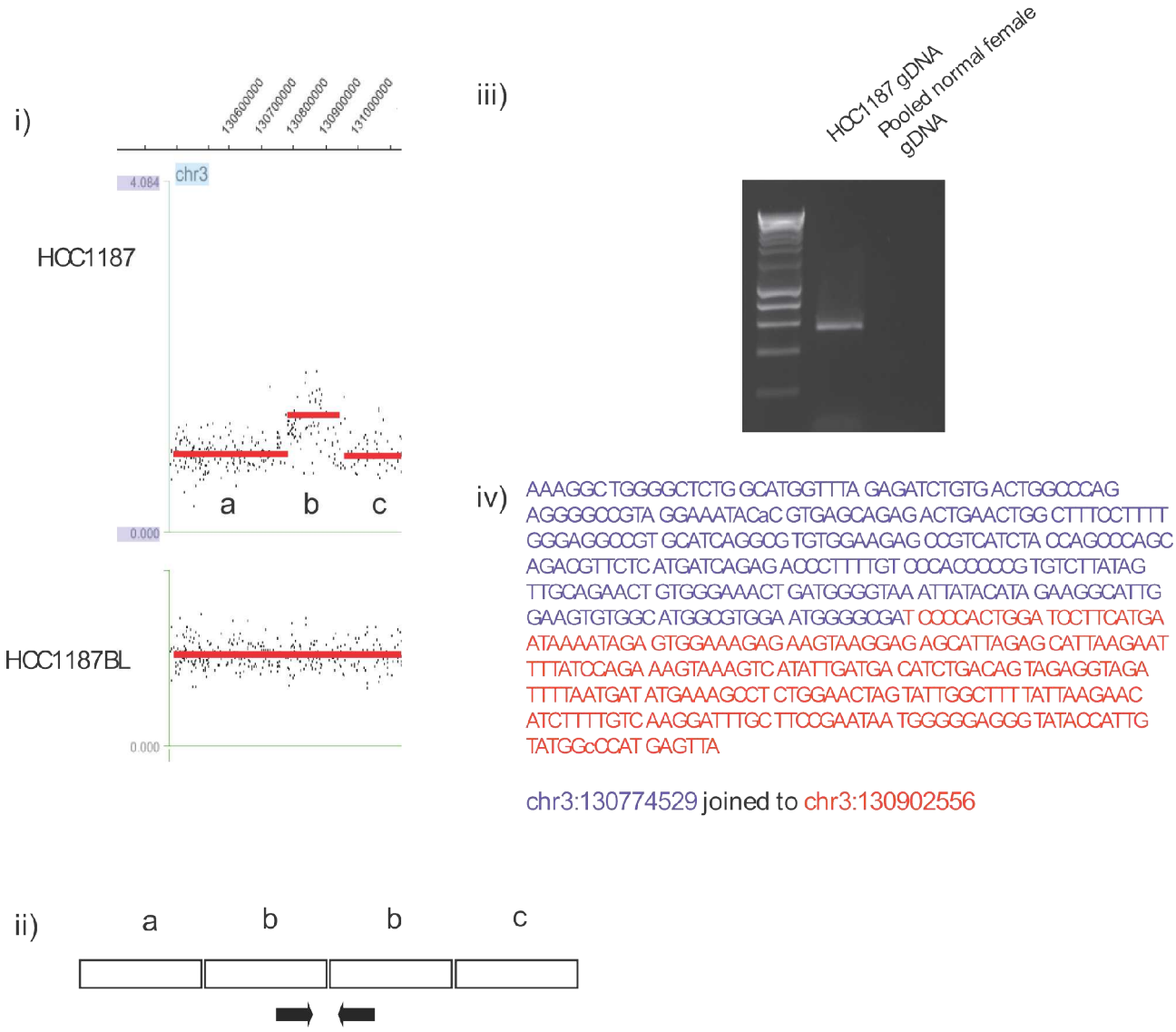


Figure 3.13. *PLXND1-TMCC1* genomic junction. i) Segmented SNP6 array CGH shows a tandem duplication that was absent from the matched normal lymphoblastoid line HCC1187BL. ii) schematic of genomic junction formed by the tandem duplication. Black arrows indicate position of PCR primers for iii) PCR across the genomic junction. iv) sequence across the genomic junction.

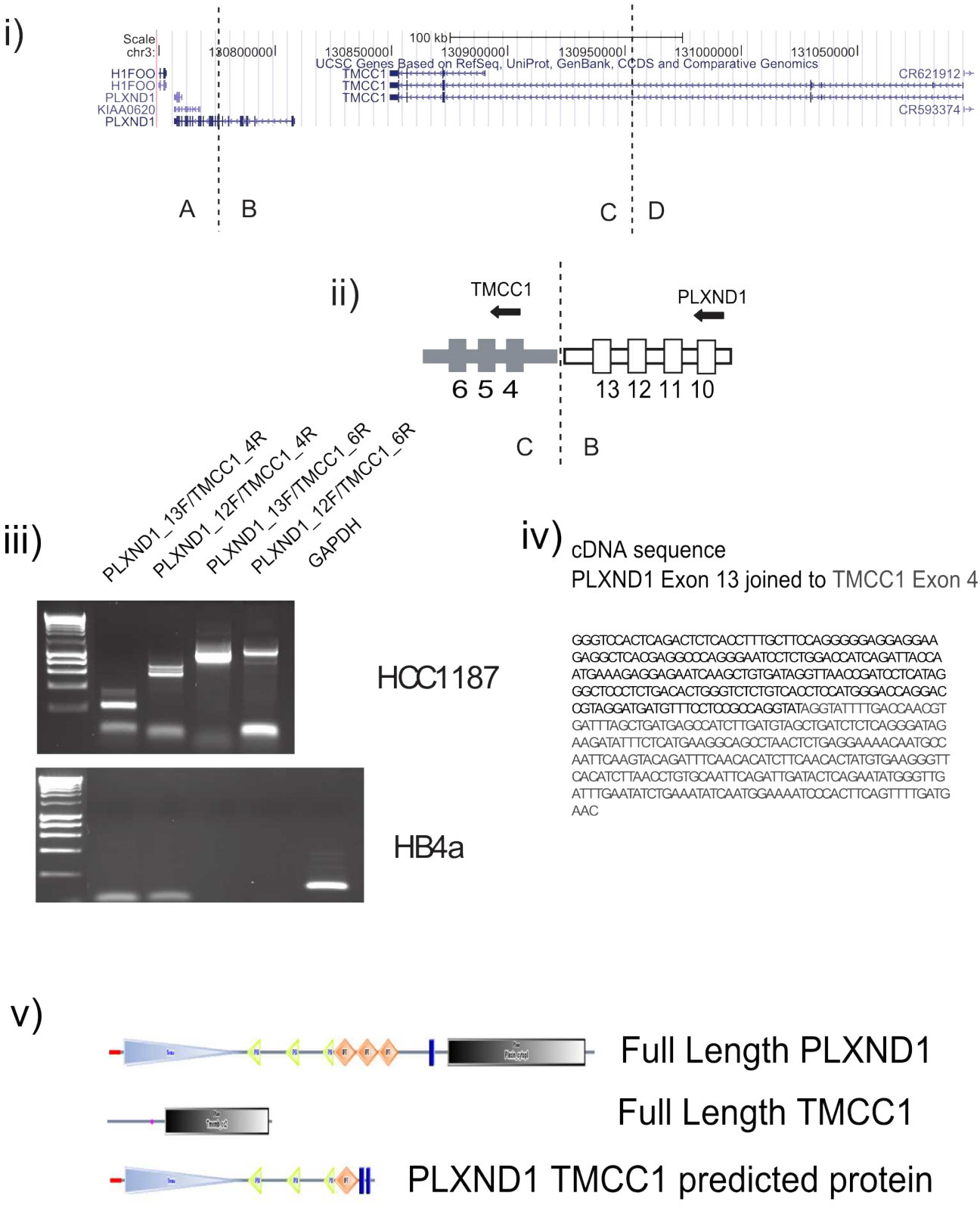


Figure 3.14. ***PLXND1-TMCC1*** fusion transcript. i) Genomic region encompassing the tandem duplication (dotted lines). ii) schematic representation of the fusion transcript. Iii) RT-PCR shows the *PLXND1-TMCC1* fusion transcript is expressed. iv) Sequence

of the fusion transcript generated from the third PCR band and v) predicted protein domains from SMART (Letunic et al., 2009). Grey triangle is a Sema domain, yellow triangle is a PSI domain, orange diamond is an IP domain, blue rectangle is a transmembrane domain, upper black rectangle is a Plexin-homology domain, lower black rectangles for *TMCC1* is a transmembrane domain.

Other reported gene-fusions in HCC1187

Stephens et al. (2009) confirmed expression of *PLXND1-TMCC1*, *CTAGE5-SIP1* and *ROD1-SUSD1* in their study. They did not report expression of *PUM1-TRERF1*, *RHOJ-SYNE2* or *CTCF-SCUBE2*. The authors did, however, find three expressed gene fusions that I did not: *RGS22-SYCP1*, *SGK1-SLC2A12*, *AGPAT5-MCPH1*. The molecular cytogenetic approach showed that *RGS22-SYCP1* and *SGK1-SLC2A12* were candidate fusions but I could not show expression even after extensive RT-PCR. I confirmed expression of the *AGPAT5-MCPH1* fusion transcript by RT-PCR (not shown).

One explanation for this is Stephens et al. used a more sensitive PCR assay. I do not think this is likely as previous real time PCR experiments by Dr KD Howarth and Dr SL Cooke showed *SYCP1* was not detectably expressed in HCC1187 (Cooke, 2007).

3.5. Analysis Part II. Sequence-Level Mutations in HCC1187

Eighty five sequence-level mutations have been identified in the HCC1187 genome comprising, 75 base substitutions and 10 indels. All known sequence-level mutations in HCC1187 are listed in table 3.6 and presented fully in Appendix 3.3. The genetic consequences of these mutations were predicted to be 10 indels, 66 missense, 5 nonsense, 4 synonymous. In two cases, two mutations appear to be in adjacent bases these were: *FLJ20422* (NM_017814.1) 254A>T and 253G>T and *GOLPH4* (NM_014498.2) 935C>T, 934G>C (Wood et al., 2007; Forbes et al., 2010).

3.5.1. Placing Sequence-Level mutations on the Genomic Map

To complete the genomic map of HCC1187, I next placed all of the known sequence-level mutations on it. To find the position of known point mutations, the individual chromosomes of HCC1187 were separated by flow-sorting and exons bearing known point mutations were amplified by PCR and the products directly sequenced. This established the presence or absence of the mutations on each chromosome segment. For example, the mutation V158L in *HSD17B8* is caused by a G>T mutation at chr6:33281286 (HG18). The mutation is listed as heterozygous and could therefore be on any of the four copies of this region in HCC1187, either copy of chromosome I, chromosome A or chromosome D. After flow sorting and sequencing genomic loci from chromosomes I, A and D, the mutation was found on both copies of chromosome I and not found on chromosomes A or D (Figure 3.15)

Of the 85 previously described sequence-level mutations, I was able to confirm 83.Two reported mutations, in *ZNF674* and *HUWE1*, were not found: they presumably occurred in other stocks of the HCC1187 cell line. It was possible to place 75 of these mutations on the genome map by sequencing mutant exons from flow sorted chromosome DNA (Figure 3.17). It was not possible to place eight mutations because they were found in regions where the genome structure was not known precisely (see discussion) or the

chromosomes on which the mutation might have resided were too small to flow sort.

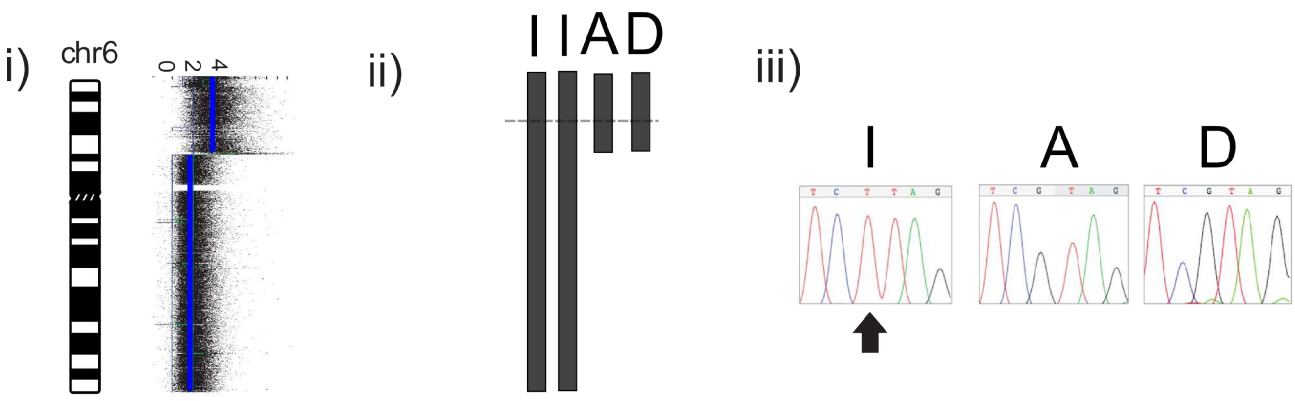


Figure 3.15. Placing sequence-level mutations on the genomic map. i) PICNIC-segmented array CGH. Total copy number is the blue line. Chromosome 6 ideogram is to the left. ii) Segments of chromosome 6 found in the HCC1187 genome, named I, A and D from array painting. Dotted line indicates the position of *HSD17B8*. Iii) Sequence traces show the G>T mutation is present on chromosome I but on chromosome A or D. As the sequence trace from chromosome I shows a homozygous mutation, we can conclude both copies of chromosome I carry the mutation.

3.5.2. Confirmation by pyrosequencing

To confirm that the proportion of mutant and non-mutant copies of each gene in the isolated chromosome preparation was accurate, pyrosequencing was used on a subset of mutations. Seven of the point mutations, in *CD2*, *FLJ20422*, *GPNMB*, *HSD17B8*, *ITIH5L*, *KIAA0427* and *MLL4*, were investigated. In each case, the ratio of normal to mutant alleles found by Sanger sequencing of flow sorted chromosomes was confirmed by pyrosequencing of whole genomic DNA. For example the mutant form of the *HSD17B8* gene was shown by Sanger sequencing to be present on both normal copies of chromosome 6 (chromosome I) and absent from both translocated chromosome 6 copies (chromosomes A and D). Pyrosequencing showed a 50:50 normal to mutant ratio in whole genomic DNA, as expected. Furthermore, pyrosequencing of flow sorted chromosomes confirmed the 0% mutant in the translocated chromosome 6 copies and 100% mutant in chromosome I. Thus non-specific genomic contamination of flow sorted chromosomes, if present at all, was at a level too low to detect by either Sanger sequencing or

pyrosequencing (Figure 316).

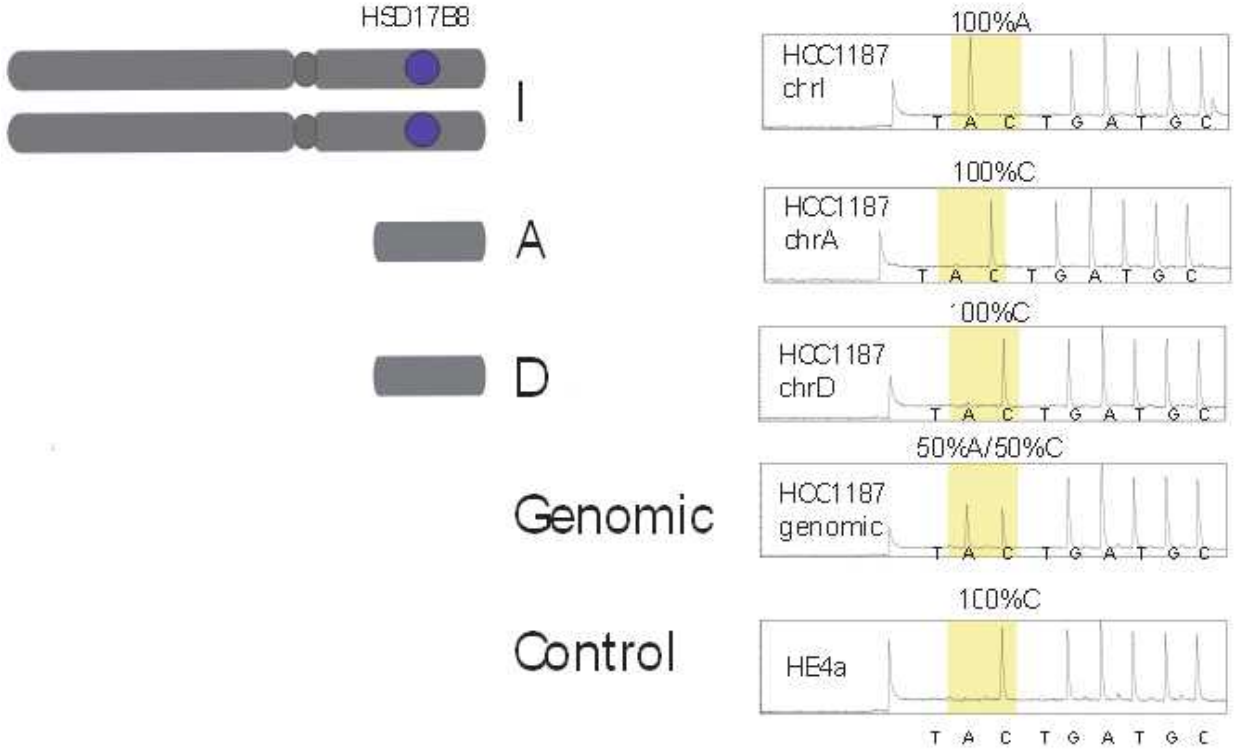


Figure 3.16. Pyrosequencing confirmation of the *HSD17B8* mutation. Left images are chromosome 6 segments I,A and D. The pyrosequencing assay used the reverse strand so the *HSD17B8* G>T mutation would appear to be a C>A. Right hand boxes are quantitative "pyrograms." Control genomic DNA showed 0% mutant bases whereas HCC1187 showed 50% mutant:wild type. Chromosome I showed 100% mutant alleles. As there are two copies of chromosome I in HCC1187 we can conclude that the mutation is *homozygous* with respect to chromosome I. The mutation was not found on chromosomes A and D. Pyrosequencing confirmed the pattern of mutations observed by Sanger sequencing (Figure 3.15).

Gene	Sjoblom et al. (2006) / Wood et al (2007)	Sanger Cosmic Forbes et al. (2010)	Genomic mutation as reported (2004 build)	Amino acid	Mutation Type	Possible Location	Mutation Found	Mutation not found
AMPD2	X		chr1:109885704G>A (homozygous)	R762H	Miss	A	A	
CD2	X		chr1:117019184G>A	C217Y	Miss	A,J,k	A	J, k
SPTA1	X		chr1:155422865A>T	Q1581H	Miss	J,k	J(hetero)	k
SPEN	X	X	chr1:16002504G>T	R1488I	Miss	E,H	E	H
GLT25D2	X		chr1:180641553G>A	V475I	Miss	J,k	k(htereo)	J
PLA2G4A	X		chr1:183651507C>G	H442Q	Miss	J,k	J(hetero)	k
RBAF600	X		chr1:19237486G>A	R1394H	Miss	E,H	E	H
CYP4A22	X		chr1:47323585G>A	G417D	Miss	A,D,E,H	D	A, E, H
PRKAA2	X	X	chr1:56881987C>A	P371T	Miss	A,D,E,H	A	D,E,H
CAMTA1	X		chr1:7730843A>G	L1080L	S	E,H,i	i,H	E
DDX18	X	X	chr2:118291285G>A	G41R	Miss	G	G(hetero)	
ARHGEF4	X	X	chr2:131632752C>G (homozygous)	T441R	Miss	G	G(homo)	
SCN3A	X		chr2:165812042A>G	E946G	Miss	G	G(hetero)	
ZNF142	X		chr2:219333250G>A (homozygous)	R1002H	Miss	G	G(homo)	
SLC4A3	X		chr2:220323514_220323523delGACAAGGACA (homozygous)	fs	INDEL	G	G	
UGT1A9	X		chr2:234462937G>T (homozygous)	S442I	Miss	G	G(homo)	
FLJ21839	X		chr2:27191236G>C (homozygous)	R395P	Miss	G	Genomic(homo)	
SULT6B1	X		chr2:37318343A>C	A108A	S	P,T	P(homo), Genomic(hetero)	T(implied)
LHCGR	X		chr2:48826897G>A	D564N	Miss	P,T	T(homo)	P
GOLPH4	X	X	chr3:169233251C>T	A312V	Miss	C	C(hetero)	
GOLPH4	X	X	chr3:169233252G>C	A312P	Miss	C	C(hetero)	C
RTP1	X		chr3:188400138C>A (homozygous)	R88S	Miss	C	C	
RNU3IP2	X		chr3:51950902C>G (homozygous)	R8G	Miss	C	Genomic(homo)	

BAP1	X		chr3:52415311C>T (homozygous)	Q261X	N	C	C(homo)	
C4orf14	X		chr4:57673742A>G (homozygous)	Q579R	Miss	B	B(homo)	
CTNNA1	X	X	chr5:138294082C>T (homozygous)	Q678X	N	F	F	
PCDHB15	X		chr5:140607486C>T (homozygous)	A719V	Miss	F	F(homo)	
CENTD3	X		chr5:141014054A>C (homozygous)	T1428P	Miss	F	F	
GMCL1L	X		chr5:177546166delA (homozygous)	fs	INDEL	F	F(homo)	
PDCD6	X		chr5:359875G>T	G123C	Miss	F,V	F(hetero)	V
FLJ32363	X		chr5:43541741C>G	S266R	Miss	F,V	F(hetero)	V
C6orf21	X		chr6:31783819C>G	P192R	Miss	I,A,D	I (heterozygous)	A,D
SKIV2L	X		chr6:32036799C>G	L183V	Miss	I,A,D	A,D	I
HSD17B8	X		chr6:33281286G>T	V158L	Miss	I,A,D	I (homo)	A, D
B3GALT4	X		chr6:33353713T>C	V180A	Miss	I,A,D	A, D	I
NCB5OR	X		chr6:84706496G>T	D337Y	Miss	I	I(hetero)	
FLNC	X		chr7:128071176G>T	D185Y	Miss	M,K	M	K
TBXAS1	X		chr7:139064224C>T	R86W	Miss	M,K	M (homo)	K
ABCB8	X		chr7:150179945C>G	A673G	Miss	K	K(hetero)	
PAXIP1		X	chr7:154198087T>G	F457C	Miss	K	K(homo)	
GPNMB	X		chr7:23086956G>T	S519I	Miss	M,K	M (heterozygous)	K
PEBP4	X		chr8:22638372G>C	R149P	Miss	N,E,H	H(homo)	N,E
ADRA1A	X		chr8:26778286G>T	G40W	Miss	N,E,H	N	E, H
FRMPD1	X		chr9:37730240G>A	G572D	Miss	Q	Q(hetero)	
SORCS1	X		chr10:108579379A>C	K223N	Miss	Q	Q(hetero)	
KIAA0934	X		chr10:363080G>A	V1264M	Miss	Q,b	b(hetero)	Q
AVPI1	X		chr10:99429559C>T (homozygous)	Q32X	N	Q	Q(homo)	
ZCSL3	X		chr11:31404466_31404470delTCTTG	fs	INDEL	Q	Q(hetero), O(hetero), R(hetero)	
NUP98	X	X	chr11:3657478G>T (homozygous)	G1652V	Miss	Q	Q(homo)	
OR1S1	X		chr11:57739474T>A	F228I	Miss	Q,O,R	Q(het)O,R (het)	

ZNHIT2	X		chr11:64641527G>C	A59P	Miss	R		R
IPO7	X		chr11:9418649G>T	A923S	Miss	Q,O,R	O	Q,R
TAS2R13	X		chr12:10952719A>G	N149S	Miss	Q,O,S	Q(homo)	Q,O,S
GPR81	X		chr12:121739602_121739601insA (homozygous)	fs	INDEL	Q	Q(homo)	
PPHLN1	X		chr12:41065014G>A (homozygous)	V173M	Miss	Q	Q	
INHBE	X		chr12:56135771G>C	R62T	Miss	Q	Q(het)	
PPP1R12A	X	X	chr12:78693190G>C (homozygous)	Q767H	Miss	Q	Q	
ITR	X		chr13:94052277C>A	T32N	Miss			
NFKBIA	X	X	chr14:34942227_34942226insC (homozygous)	fs	INDEL	W	W(homo)	
WARS	X		chr14:99871016G>C	E455D	Miss	W,Y	Y	W
SMG1		X	chr16:18730825A>C	K3579Q	Miss	Y,S	Y	S
PDPR	X		chr16:68734948A>T	Y546F	Miss		Genomic(het)	
LLGL1	X		chr17:18080933C>G	L522L	S		Genomic(hetero)	
NOS2A	X		chr17:23118990G>T (homozygous)	A679S	Miss	Z	Genomic(prob hetero)	Z
RASL10B	X	X	chr17:31086470G>A (homozygous)	V52M	Miss		Genomic(homo)	
TRIM47		X	chr17:71382450_71382450	fs	INDEL			
TP53	X	X	chr17:7520090_7520088delGGT (homozygous)	G108del	INDEL		Genomic(homo)	
STATIP1	X	X	chr18:31994943_31994944delCT	fs	INDEL	U,a	U(homo)	a
FHOD3	X		chr18:32527271C>T	S533L	Miss	U,a	U(homo)	a
KIAA0427	X		chr18:44541852G>C	V389L	Miss	U,a	U(homo)	a
FLJ20422	X		chr19:19104498A>T	E85V	Miss	e,P,T,i	e	P,T,i
FLJ20422	X		chr19:19104499G>T	E85X	N	e,P,T,i	e	P,T,i
MLL4	X		chr19:40904380C>T	P764L	Miss	P,e,i	P	e, i
APOC4	X		chr19:50140242C>A	P75Q	Miss	e,P,T,i	e	P, T
MYBPC2	X		chr19:55650351C>T	P730L	Miss	e,P,T,i	P	e,T,i
PLCB1		X	chr20: 8667928C>T	A743A	S	G,Y,d	d	G,Y

ITGB2	X	X	chr21:45146037_45146013delTGAACACGCACCCTGATAAGCTGCG	fs	INDEL	F	f(hetero)	
MYH9	X	X	chr22:35012676_35012674delGCA (homozygous)	indel	INDEL	h	h(homo)	
CYP2D6	X		chr22:40851168G>A	G42R	Miss	F	f(hetero)	
PLS3	X		chrX:114703778A>C	D485A	Miss	L,D,k	L	D,k
ZNF674	X		chrX:46144024G>A	E85K	Miss	L,D,k		L, k, Genomic
HUWE1		X	chrX:53537429G>A	R481K	Miss	L,D,k		L,D,k
ITIH5L		X	chrX:54706426C>T	P76L	Miss	L,D,k	k	L,D
SATL1	X		chrX:84168729C>G	S277X	N	L,D,k	k(het), L(homo)	D

Table 3.6. Sequence-level mutations in HCC1187

3.6. Discussion

I generated a complete genomic map of HCC1187 using all known data on the cell line. Figure 3.17 clearly shows that, although the sequence-level mutational burden is quite high, the number of genes disrupted by chromosome aberrations is – at the very least – comparable in scale.

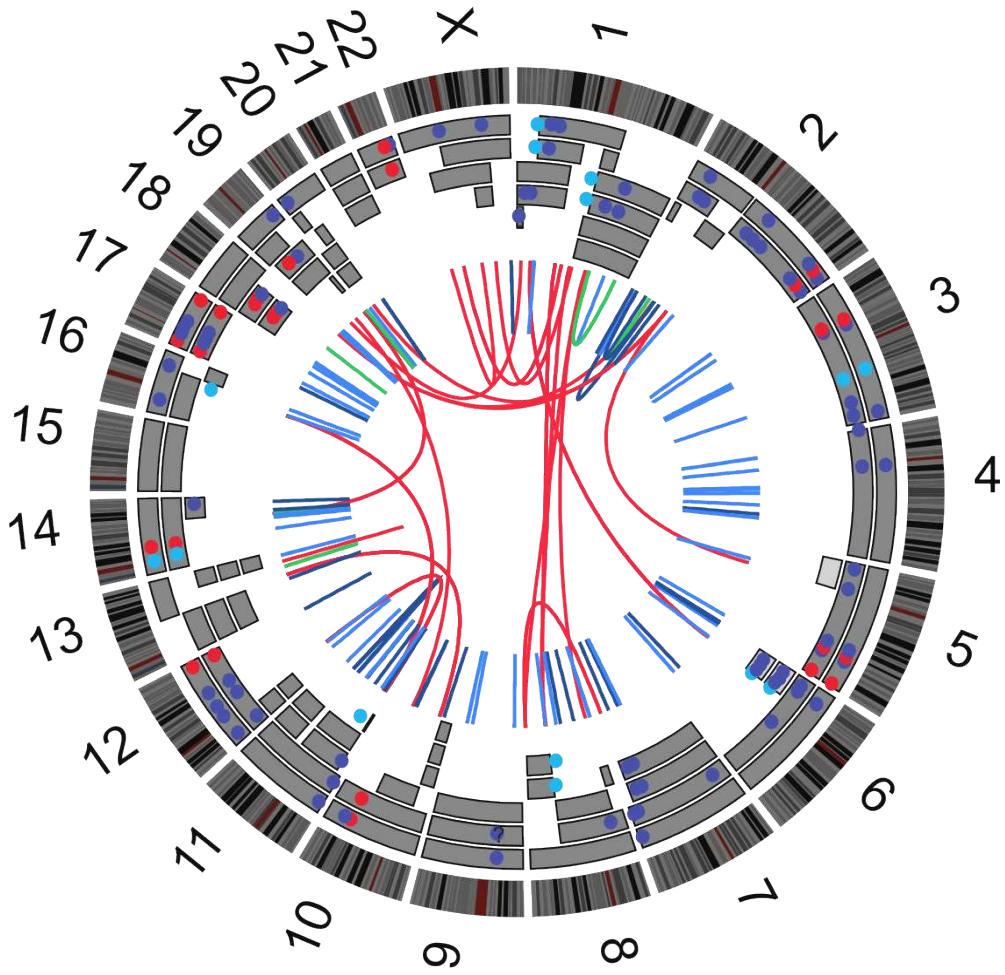


Figure 3.17. The complete genome map of HCC1187. Chromosome ideograms and array painting chromosome segments are arrayed around the outside as in figure 3.1. Blue and red circles are missense and nonsense/ frameshift sequence-level mutations respectively. Light blue circles are expressed fusion genes. They are placed on the chromosome segment identified by resequencing mutated exons. Inner links combine data from Howarth et al. (2008), Stephens et al, (2009), and the present study. Red links are translocations, light blue are duplications, dark blue are deletions and green are inversions.

Approximately 150 genes were at chromosome breakpoints in HCC1187. Three genomic junctions form in-frame and expressed fusion transcripts that may have oncogenic potential. There were four homozygous deletions of genes that could constitute a tumour suppressor loss. Many of the other breaks potentially represent one of the two hits required to inactivate tumour suppressors.

3.6.1. How complete was this analysis?

We can be reasonably certain that most (if not all) of the cytogenetically visible chromosome translocations are accounted for in this analysis. Similarly, small gains and losses that could be identified by segmentation algorithms have probably all been identified. However, some of these were not apparent by array painting. These were regions of chromosome 1q21, 10p, 10q,11 and 12p11 (Figure 3.18). It was not possible to discern the structures of these regions by FISH (not shown).

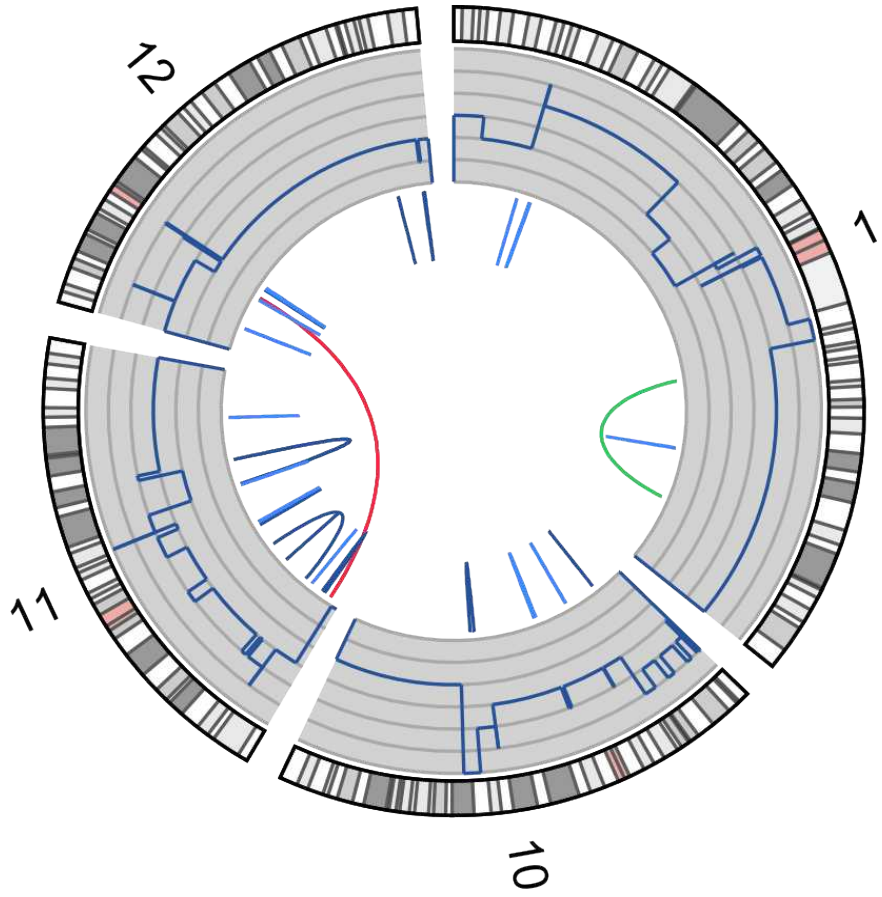


Figure 3.18. Complex regions on HCC1187 chromosomes 1,10, 11 and 12.

Figure 3.18. Complex regions on HCC1187 chromosomes 1,10, 11 and 12. Chromosome ideograms are around the outside. Segmented copy number is the blue line. Intra-chromosomal structural variants are shown as inner links. Light blue are duplications, dark blue are deletions and green are inversions from Stephens et al. (2009).

Stephens at al (2009) identified a hundred or so rearrangements below the resolution of array CGH as well as some balanced inversions (Appendix 3.2). As the authors estimated that their screen only detected 50 percent of rearrangements there is the possibility of one hundred or so further small rearrangements, and therefore more fusion genes.

3.6.2. The rearrangements that fused genes

As Edwards (2010) pointed out, the majority of structural variations in epithelial cancer genomes are likely to be intrachromosomal and sub-microsopic. It is not surprising, then, that 6/9 gene fusion in HCC1187 were formed by tandem duplications and interstitial deletions (Edwards, 2010). We should also consider that tandem duplications, deletions and inversions within genes can result in new and aberrant isoforms being expressed. In HCC1187 this is the case for *RB1* among others. The possibility of a recurrent submicroscopic breast cancer fusion gene is quite possible.

3.6.3. Conclusions

At least 150 genes were disrupted by chromosome rearrangement in HCC1187. As many as nine and possibly more genes are fused and expressed in this cell line, three of which are in frame. But it is still not clear whether we should regard structural rearrangements as equally important as sequence-level mutations. In order to decide this we need to know about the relative timing of these mutational mechanisms. And if sequence-level and structural rearrangement occur at similar times during tumour evolution, there is no reason why we should not regard both mechanisms important in tumour evolution.

Chapter 4

The Evolution of a Breast Cancer Genome

4.1. Introduction

Inferring the evolutionary history of individual tumours has not, to the best of my knowledge, been attempted because of the perception is that it is impossible to look at a single sample at a single time point and infer the order in which its mutations occurred. But given that each individual cancer genome has been described as an 'archaeological record' of a tumour's history or even a 'palimpsest of mutational forces' (Stratton et al. 2009, Greenman et al. 2010 *submitted*) there is a clear utility in finding the relative order in which mutations happened in individual tumours.

There are at least nine fused transcripts in HCC1187 (Chapter 3) and three of these transcripts are predicted to preserve the reading frame of the 3' gene. In addition, we know of 85 sequence-level mutations in HCC1187 from genome-wide screens and targeted resequencing (Sjöblom et al., 2006; Wood et al., 2007; Forbes et al., 2010). Most of these mutations must be passenger events, and, as discussed in Chapter 1, finding driving events amongst an excess of passengers is a considerable challenge.

This task is further complicated by a major unknown in cancer biology: the relative importance and timing of genome rearrangements compared to sequence-level mutation. Some suggest chromosome instability might arise early and be essential to tumour suppressor loss (Nowak et al., 2002; Rajagopalan et al., 2003; Rajagopalan and Lengauer, 2004) while others think that CIN is a late event or contributes little to cancer development (Johansson et al., 1996; Sieber et al., 2002).

In this chapter, I address the above question by considering the evolution of the highly-rearranged karyotype of HCC1187. This allowed me to infer the relative timing and importance of different groups of mutations and retrace the evolutionary history of this tumour.

4.1.1. A common route of evolution for breast cancer genomes.

I was able to infer how the karyotype of HCC1187 had evolved by applying a model first proposed by Muleris et al. (1988) and proved experimentally in breast cancer by Dutrillaux et al. (1991). This study used R-banded karyotype analysis on a large series of breast tumours with a view to elucidating the sequence through which complex karyotypes evolve (Muleris et al., 1988; Dutrillaux et al., 1991).

Dutrillaux et al. (1991) looked at the total number of chromosomes, and the percentage of apparently abnormal chromosomes in 113 breast carcinomas. 65 were near diploid and 48 hyperploid (>50 chromosomes). Chromosome numbers of near diploid tumours had a bimodal distribution, centred around 45-46 and 37-38. Tumours with the fewest chromosomes had hyperploid sidelines in significantly more cases that those with chromosome contents nearer to diploid. This implies a selective pressure for chromosome number to increase once a certain proportion of the diploid genome, around 35%, had been lost. The mechanism of this ploidy increase was endoreduplication – duplication of the entire genome – as there was a large disparity between modal chromosome number and clonal sidelines rather than a series of intermediates.

The authors went on to describe a generalised model for the evolution of breast cancer karyotypes (summarised in Figure 4.1):

1. Occurrence of unbalanced rearrangements decreasing chromosome number and DNA content.
2. Correlatively to the rate of chromosome rearrangements, formation of endoreduplications leading to hyperploid sidelines.
3. Persistence of the near diploid cells and decrease of chromosome number to about 35 and of DNA index to 0.85 or more frequently, elimination of the near diploid cells and complete passage to hyperploidy.
4. further losses of chromosomes in the hyperploid tumours, whose karyotypes can decrease to about 55 chromosomes and a DNA index of 1.35.
5. Eventually, occurrence of a second endoreduplication, leading to an apparent near tetraploidy. (Dutrillaux et al., 1991, p.245)

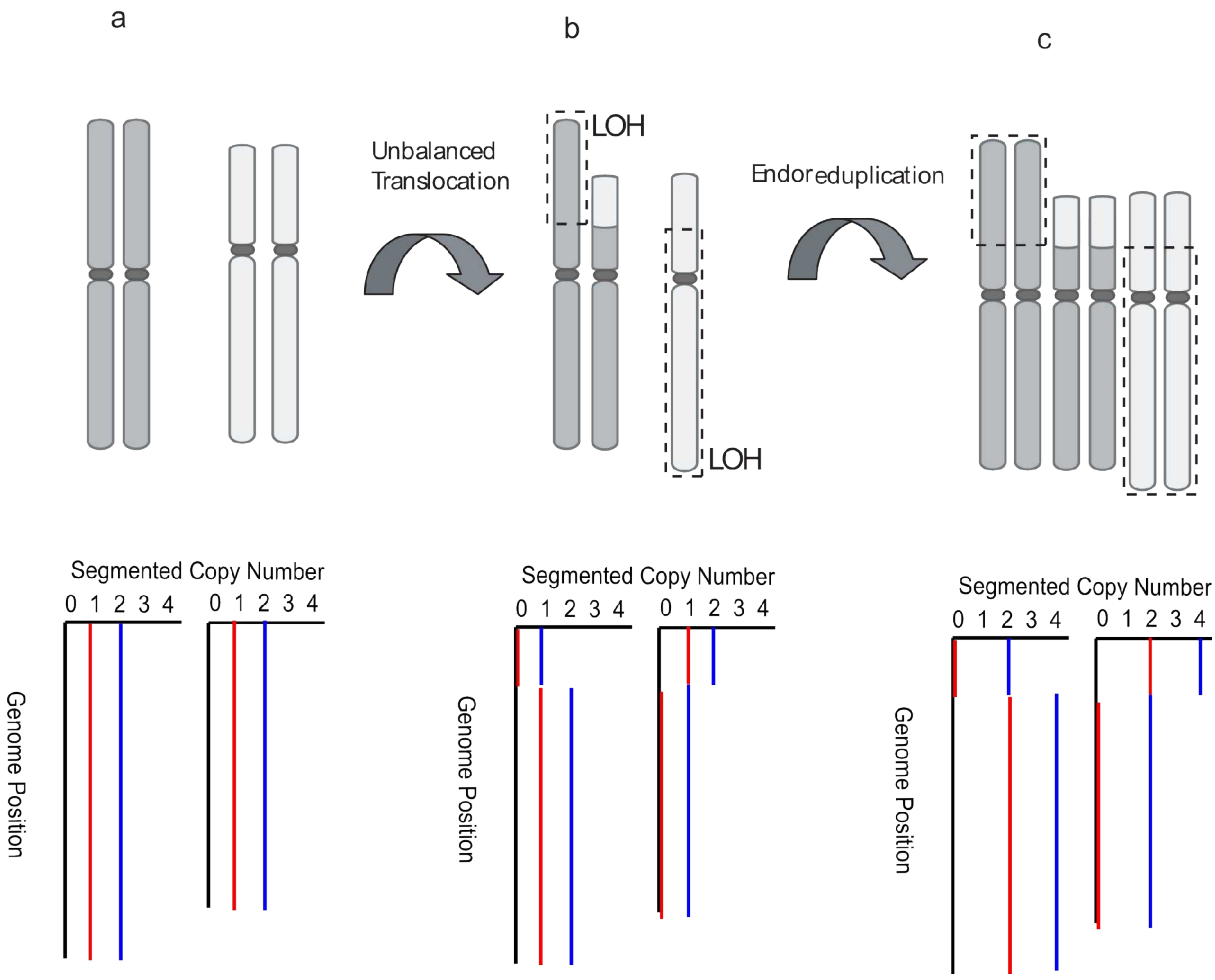


Figure 4.1. The pattern of karyotype evolution followed by most breast tumours, known as 'monosomic' evolution, including an endoreduplication. Upper panels are chromosome ideograms, lower panels are simulated array CGH plots with the total copy number (blue) and minor allele copy number (red) as from PICNIC segmentation. **a) and b),** An unbalanced translocation reduces the chromosome number by one, and leaves regions of loss of heterozygosity (LOH) (dashed boxes). **c)** Often, at some point, endoreduplication occurs, i.e. the whole chromosome complement doubles, to give a duplicated translocation and pairs of chromosome segments showing regions of loss of heterozygosity. The process may then continue with more unbalanced translocations.

Dutrillaux et al. (1991) also note a perceived difficulty in distinguishing between tumours which have gained a few chromosomes and tumours which have lost many before and after endoreduplication. This is true when one only considers chromosome numbers, but when information on chromosome rearrangements are also considered a 'complete

discontinuity' between endoreduplicated and non-endoreduplicted tumours is apparent. The authors often observed two populations of cells, one with approximately twice the number of rearranged chromosomes of the other. In these side lines most of the chromosome translocations appeared in two copies. The authors never observed a series of sidelines with intermediate chromosome numbers meaning that the most probable explanation for the doubled sideline was a simultaneous doubling of the ancestral genome – endoreduplication (Dutrillaux et al., 1991).

4.2. The Evolution of HCC1187

4.2.1. HCC1187 is endoreduplicated

It is very likely that HCC1187 endoreduplicated at some point in its history as many of its chromosome translocation break points appeared in two copies in the modal karyotype. These must have duplicated at some point in their history (Figure 4.3).

Although breast cancer genomes contain hundreds of structural variations, a chromosome rearrangement at any given point in the genome is a rare event. If we see two copies of the same derivative chromosome in a karyotype it is likely to be due to duplication of one original copy. From this fact we can infer the order of events: translocation followed by duplication. For example, HCC1187 chromosomes A and D both have a t(1;6) translocation junction. To SNP6 array resolution both A and D appear to share the same genomic break point. Chromosome A has undergone a further translocation with chromosome 8, and chromosome D a further translocation with X. As neither translocation t(1;8) or t (1;X) was duplicated we can infer each happened after the duplication of the ancestral derivative chromosome der(1)t(1;6).

4.2.2. SNP Allele ratios confirm the HCC1187 endoreduplication

I used data from PICNIC (Predicting Integral Copy Numbers In Cancer) (Greenman et al., 2010) array segmentation algorithm to assign an arbitrary ‘Parent A’ or ‘Parent B’ origin to virtually all regions of the HCC1187 genome. PICNIC is a segmentation algorithm written specifically for rearranged cancer genomes. A useful feature of this algorithm is that it can produce two copy number profiles: one for the total copy number and one for the copy number of the “minor allele.” I combined chromosome segments from array painting, their shared breakpoints from SNP6 arrays and PICNIC zygosity data, and used the fact that shared breakpoints or shared alleles between two loci imply they had a common ancestor (Figures 4.2 and 4.3).

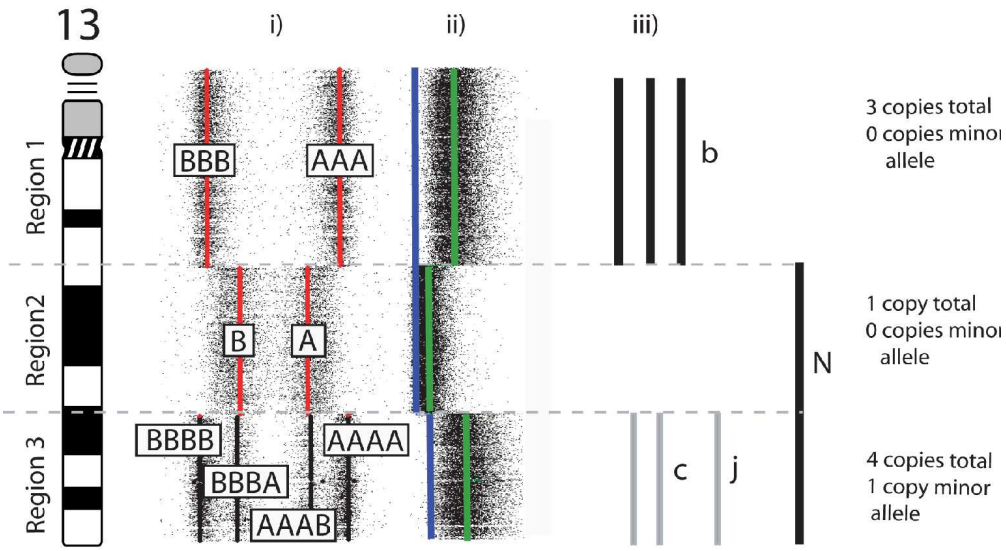


Figure 4.2. Segmentation by PICNIC algorithm reveals 'Parent A' and 'Parent B' origin of segments of chromosome 13. i) Segmentation by PICNIC algorithm (Greenman et al. 2010). 'PICNIC plot' gives zygosity, i.e. parental origins, by plotting SNP calls AA, AB, ABB etc. Equal representation of both SNP alleles, i.e. AB, AABB, etc, would be plotted on the centreline. Pure A calls are to the right, with AA and AAA further out, and pure B to the left. Mixed calls of AAB, AAAB, etc fall between corresponding pure A calls and the centreline. Red lines indicates regions of homozygosity. ii) Total segmented copy number (green), equivalent to CGH, plotted left to right and minor allele copy number (blue line). iii) Segments of chromosome identified by array painting as chromosome b, c, j and N with their inferred parentA/parentB origin (A=black, B=grey). In **Region 1**, Total copy number is three, and region is homozygous, since there are only two combinations of alleles, pure A and pure B. Therefore the three copies of chromosome b share the same arbitrary parental origin. In **Region 2**, there is one copy, homozygous. The Chromosome 13 segments in chromosomes b and N are likely to be products of an ancestral translocation, since their breakpoints are the same to 6kb resolution (unpublished) and they add up to a complete chromosome 13. This means that originally the b and N segments were joined so they must have belonged to the same chromosome. **Region 3**: Four copies with four allele combinations indicating that three copies are from the same parent. Peaks c and j share breakpoints, so are derived from the same chromosome, and must be of different parental origin from peak N to account for the allele combinations. In Conclusion, chromosomes b and N are from one parent, while chromosomes c and j are from the other parent.

The Parent A/Parent B segmentation showed clearly that HCC1187 was endoreduplicated, as almost all chromosome segments that had been identified by array painting had duplicated at some point in their history and were found in the observed karyotype precisely twice (Figure 4.3). It follows that translocation junctions that had been duplicated probably occurred before endoreduplication and those that were not probably occurred after. In HCC1187, 9 chromosome translocations (41%) visible by SKY and chromosome painting occurred before endoreduplication and 13 after. This makes endoreduplication an approximate midpoint in the evolution of this karyotype with respect to chromosome translocations.

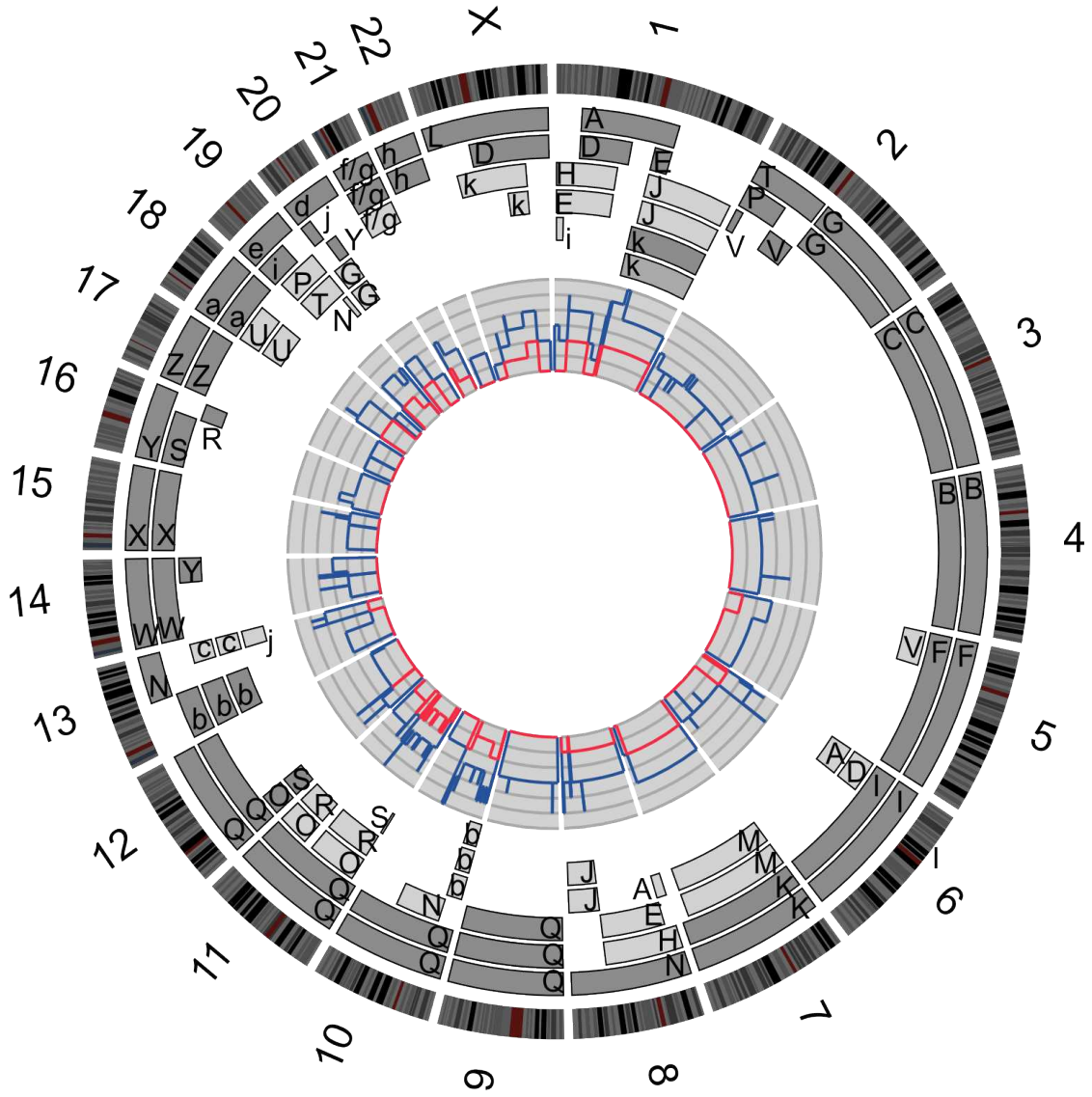


Figure 4.3. Circos plot of the HCC1187 genome: Chromosome ideograms around the outside, oriented clockwise pter to qter. Grey boxes are chromosome segments observed by array painting. Their parent of origin (light grey and dark grey) was deduced as in Figure 4.2 from the number of allelotypes given by PICNIC segmentation. Note that assignment of parents A and B does not transfer between chromosomes. Inner line plots are PICNIC plots: dark blue line, total copy number, equivalent to array CGH. Red line, copy number of the minor allele; where this is zero the genome is homozygous. Chromosome segments that share a translocation breakpoint were assumed to have the same parental origin.

The second signature of endoreduplication was that large portions of the genome were present in two copies but had lost heterozygosity. In these regions it is likely that one parental chromosome was lost and the remaining chromosome duplicated. As this was the case for several whole chromosomes, the simplest explanation is that chromosomes were lost one by one and the remaining copies duplicated simultaneously at endoreduplication. This general scheme of whole chromosome loss and unbalanced translocation followed by an endoreduplication is, in fact, the commonly observed evolutionary route for breast and colorectal cancer genomes described above (Muleris et al., 1988; Dutrillaux et al., 1991). For the three regions of the genome that were triplicated, for example chromosome 9, I assumed one duplication had occurred at endoreduplication and another had occurred later.

4.2.3. Evolution of an endoreduplicated genome

This complex hypotriploid karyotype of HCC1187 is likely to have evolved by successive loss of chromosomes and endoreduplication. Figure 4.4 shows the most probable sequence of evolution. In making this scheme I only had to assume the simplest explanation was most likely. The pre-endoreduplication events consisted mainly of whole chromosome losses and unbalanced translocations, as expected if early karyotype evolution followed what was termed the 'monosomic' route (Muleris et al., 1988). Importantly, using this scheme it was possible to infer the most likely state of the genome immediately before endoreduplication.

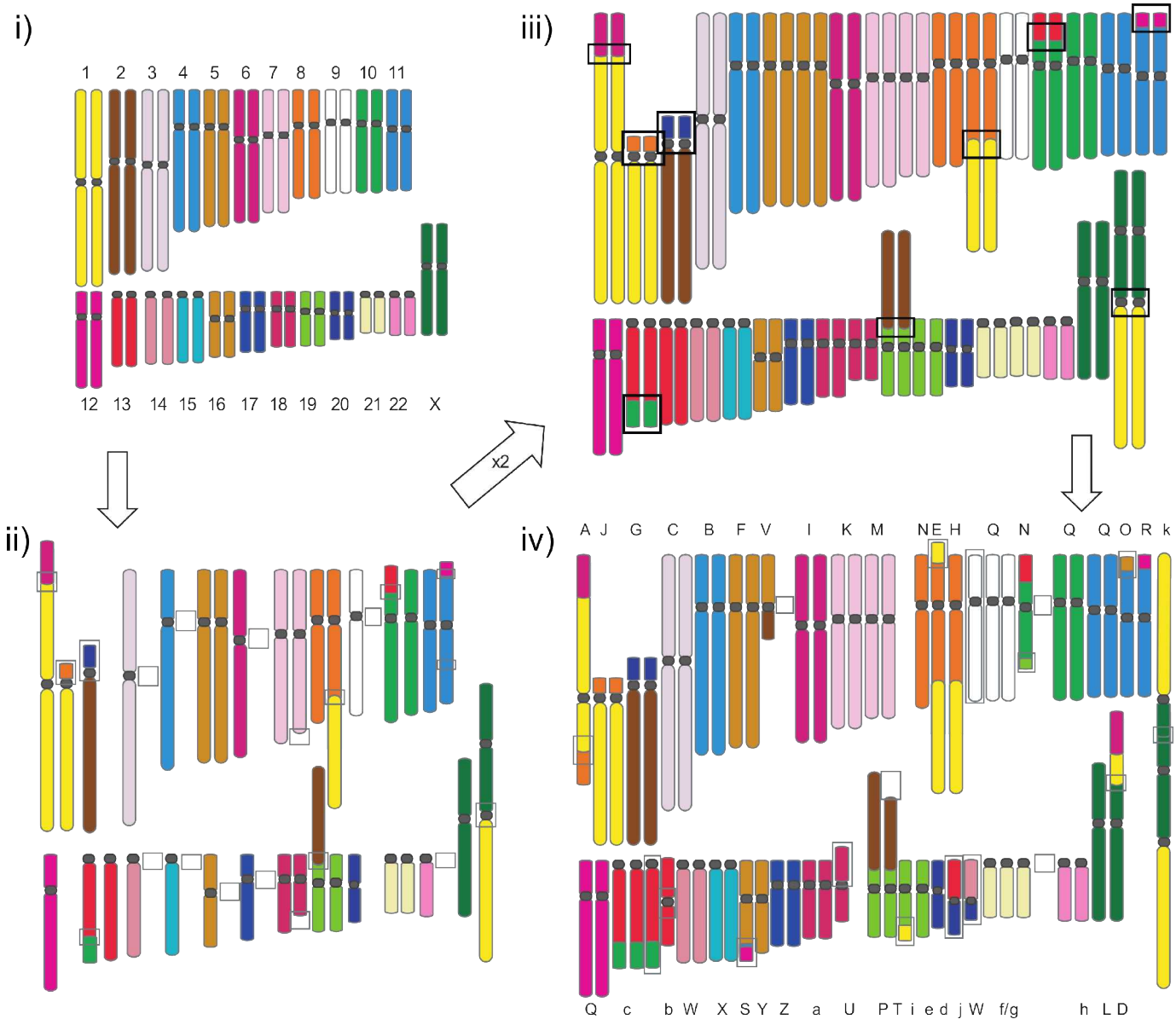


Figure 4.4. Evolution of the HCC1187 karyotype. i) Initial karyotype: chromosomes are shown is SKY pseudo-colours. ii) The first, monosomic, phase of evolution was dominated by whole chromosome losses and unbalanced translocations grey boxes indicate events that probably happened at this stage. iii) At some point, the remaining chromosome complement is doubled by an endoreduplication. Endoreduplication duplicated all of the translocation break points that preceded it. Black boxes show the nine translocations visible by SKY that were doubled. iv) Evolution then continued with further loss and unbalanced translocation. Thirteen of these translocations are seen as

single copies in the final, observed, karyotype. Chromosome names are displayed below each chromosome.

4.3. Duplication of Mutations at Endoreduplication

As it appeared that endoreduplication formed an approximate mid-point in the structural evolution of the HCC1187 genome, I wondered if one could also use endoreduplication to investigate the timing of other mutations. This would be an interesting exercise for three reasons:

1) When working with cell lines one can never by sure if a fusion gene, for example, was formed in culture. Pre-endoredulpication fusion genes must have happened earlier and they would, therefore, be more likely to be *in vivo* events.
2) If chromosome instability started late in the evolution of this line, then most point mutations must precede it. If this was the case, then most point mutations must have happened earlier than endoreduplication. Endoreduplication could help us understand the relative timing of chromosome changes and sequence-level mutations.
3) If a certain class of mutations was concentrated before or after endoreduplication, then this may indicate a *requirement* for them to happen at a given time during tumour evolution. This fact implies selection may have contributed to earlier or later clustering of mutations. Endoreduplication may help us find driving events in the evolution of this genome.

4.3.1. Fusion genes

There are nine expressed fusion genes in HCC1187: *AGPAT5-MCPH1*, *SGK1-SLC2A12*, *CTAGE5-SIP1*, *PLXND1-TMCC1*, *SUSD1-ROD1*, *RGS22-SYCP1*, *RHOJ-SYNE2*, *PUM1-TRERF1* and *CTCF-SCUBE2* (by combining my structural data from Chapter 3 with that of Stephens et al. (2009)). Just as for chromosome translocations, fusion genes present in two copies were likely to have occurred before endoreduplication, while if a

rearrangement occurred after chromosome duplication, it would generally only be present in a single copy.

Duplication of fusion genes was assessed by FISH and array CGH. Of the nine fusion genes, six had clearly been duplicated, two had not been duplicated and one was undetermined. The early, so duplicated, gene fusions were: *CTAGE5-SIP1*, *PLXND1-TMCC1*, *RGS22-SYCP1*, *RHOJ-SYNE2* and *PUM1-TRERF1* and *SGK1-SLC2A12.* The late non-duplicated fusions were *CTCF-SCUBE2*, *SUSD1-ROD1* while *AGPAT5-MCPH1* was undetermined.

4.3.1.1. Fusion genes at chromosome translocation break points

The *PUM1-TRERF1* fusion was an earlier event since derivative chromosomes A and D shared the same translocation breakpoint, a single chromosome originally must have carried the *PUM1-TRERF1* fusion and subsequently duplicated. This is also the case for *RGS22-SYCP1* as it was formed by a translocation t(1;8). This translocation is present on both copies of chromosome J. Chromosome J was observed in two copies so was probably formed before endoreduplication too.

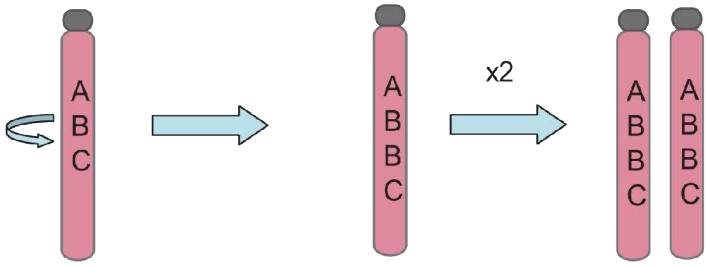
In contrast, fusion of *CTCF* and *SCUBE2* was a later event as it was only found in a single copy and furthermore it had resulted from a balanced translocation of a chromosome that had already been duplicated. The 12;16 junction was in a derivative chromosome made of pieces of 11,12 and 16 (chromosome S). At some time before endoreduplication there was a t(11,16) translocation. At endoreduplication the der(11)t(11;16) duplicated. One copy of this remained (chromosome R) and the chromosome 16 portion of the other took part in a near balanced translocation with chromosome 12 to form the *CTCF-SCUBE2* fusion on the derivative der(16) (chromosome S). Therefore chromosome S had evolved from one of two ancestral copies of chromosome R so must have occurred after endoreduplication.

4.3.1.2. Fusion Genes formed through tandem duplications

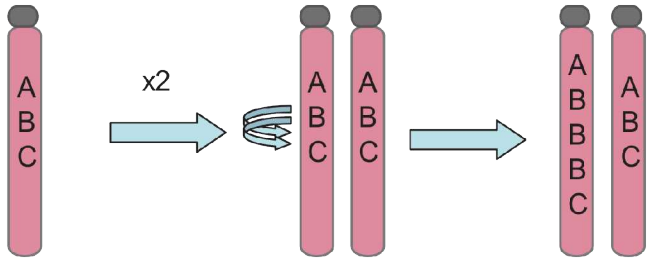
Four of the gene fusions were formed by small tandem duplications. The duplications that formed the *CTAGE5-SIP1* and *PLXND1-TMCC1* fusion probably occurred before endoreduplication, so were earlier events. The duplicated regions were found in 4 copies according to PICNIC segmentation. I first confirmed that the extra copies of this segment were most likely arranged in tandem. Both metaphase and interphase FISH using fosmids showed single fluorescence signals in only two places in the genome, on both copies of chromosome 3 and 14, indicating that the duplications must be very close together, either as 2 + 2 or 3 + 1 copies. FISH on extended chromatin fibres resolved pairs of signals, and individual chromatin fibres always showed two signals, indicating that the fusion was present as a tandem duplication on both homozygous copies of chromosome 3 or 14. The tandem duplications probably happened before endoreduplication, as they was present on both chromosome 3s and 14s in two copies. If the duplications happened after endoreduplication the duplicated region would be present in a 3:1 ratio which I did not observe by fibre FISH (Figure 4.5).

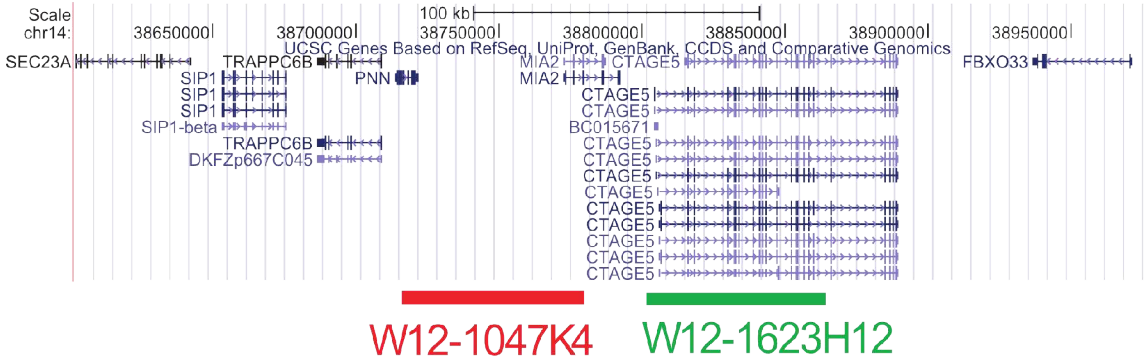
In contrast to the above “doubled” tandem duplications, the events that caused fusion of *SUSD1* to *ROD1* was only found in single copy e.g. the segmented copy number only increased by one. This implies these tandem duplications were not duplicated at endoreduplication so probably happened later. The tandem duplication that fused *AGPAT5* to *MCPH1* was too small to be segmented by PICNIC so its copy number and, therefore, its timing was undetermined.

1. Early tandem dupliction followed by duplication



2. Duplication followed by 2 tandem duplications





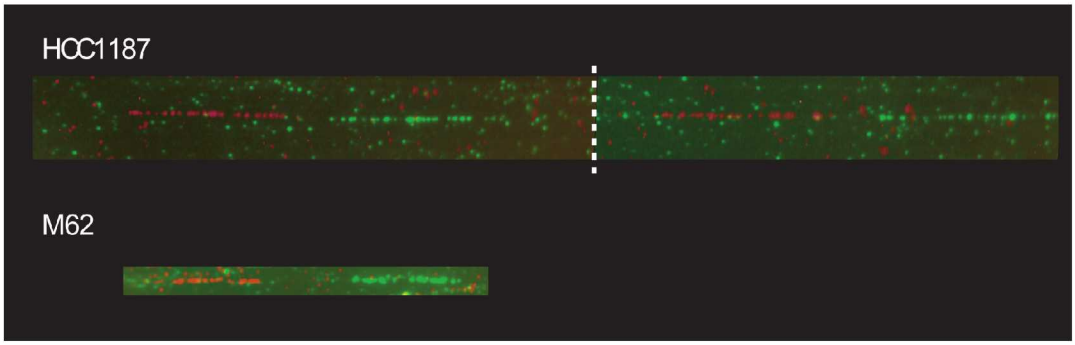


Figure 4.5. Fibre FISH and Evolution of the *CTAGE5-SIP1* fusion. Legend overleaf

Figure 4.5. Fibre FISH and Evolution of the *CTAGE5-SIP1* fusion. There are two explanations for the evolution of the *CTAGE5-SIP1* tandem duplication region on chromosome 14: 1) Early tandem duplication followed by endoreduplication. 2) Endoreduplication followed by two tandem duplications on the same chromosome. Fibre FISH using fosmids, W12-1047K4 and W12-1623, confirmed the first explanation was correct. In control cells FISH on extended chromatin fibres confirmed the probes hybridized close to one another in a red-green configuration. In HCC1187, 100% of DNA fibres showed a red-green-red-green pattern (the figure shows two 60X microscope fields joined together). This means that i) the duplicated region was arrayed in head to tail orientatation, ii) There was only one extra copy of the locus per chromosome. If evolution plan (2) was correct I would have expected to see a 50:50 ratio of single red-green signals to triplicated red-green signals.

4.3.1.3. Fusion genes formed by deletions

The interstitial deletion that formed the *RHOJ-SYNE2* fusion was previously confirmed as homozygous. SNP6 data showed that chromosome 14 was homozygous over its entire length but present in two copies at the *RHOJ-SYNE2* locus. The best explanation is that one copy of chromosome 14 was lost and the remaining copy duplicated during the endoreduplication. As the interstitial deletion that formed the *RHOJ-SYNE2* fusion was homozygous it must have pre-dated the endoreduplication. An identical case for *SGK-SLC2A12* on chromosome six was observed.

4.3.1.4. Duplication of other small deletions and duplications

Most small gains and losses did not fuse genes, but it was still possible to place many of these before or after endoreduplication. As fibre FISH on the *CTAGE5-SIP1* fusion showed, any earlier tandem duplication must have been doubled at endoreduplication. In array CGH, the segmented copy number must increase by two (or be divisible by two) for early events, and only increase by one for later events (Table 4.1).

It was possible to place 19 small duplications before or after endoreduplication, fusions of *CTAGE5-SIP1, PLXND1-TMCC1* and *SUSD1-ROD1* were caused by three of these. Eight duplications (42%) were placed earlier and eleven later. Earlier deletions, typified by the *RHOJ-SYNE2* deletion, would also be duplicated at endoreduplication but the copy number step would decrease by two (or be divisible by two). The copy number of later deletions would only decrease by one. It was possible to place 9 deletions (4 homozygous) of less than 5Mb unambiguously before or after endoreduplication. The fusion of *RHOJ* to *SYNE2* and *SGK1-SLC2A12* were caused by two of these deletions. Five deletions were placed earlier (56%) and four later.

Chr	Type	Preceding Segment End	Size of gained or lost region (kb)	Succeeding Segment Start	Preceding Segment CN	Gained or lost region copy number	Succeeding Segment CN	Doubled at Endoreduplication	Earlier or Later?
2	Deletion	33032718	56.04	33089758	2	0	2	y	E
2	Deletion	54050162	1101.61	55159490	2	1	2	n	L
2	Deletion	66859298	257.47	67119098	2	1	2	n	L
6	Deletion	1661709	4019.19	5681742	4	3	4	n	L
6	Deletion	101044085	205.62	101259352	2	1	2	n	L
6	Deletion	134367799	208.05	134580872	2	0	2	y	E
14	Deletion	62771053	630.99	63410248	2	0	2	y	E
15	Deletion	69964951	249.53	70233770	2	0	2	y	E
17	Deletion	34666635	85.62	34766404	2	0	2	y	E
2	Duplication	104926149	319.83	105256524	2	3	2	n	L
2	Duplication	222210422	651.77	222869850	2	3	2	n	L
3	Duplication	9483392	1113.06	10597941	2	3	2	n	L
3	Duplication	57148414	95.13	57255295	2	4	2	y	E
3	Duplication	130771904	131.5	130909903	2	4	2	y	E
4	Duplication	146293755	428.13	146731470	2	4	2	y	E
4	Duplication	199380516	2800.24	2807561	2	3	2	n	L
6	Duplication	149546259	497.51	150052602	2	3	2	n	L
8	Duplication	127966102	889.5	128857819	3	4	3	n	L
9	Duplication	113916140	215.02	114139966	3	4	3	n	L
10	Duplication	30165639	272.47	30440477	3	4	3	n	L
10	Duplication	46363383	874.32	47417401	3	4	3	n	L
13	Duplication	101878311	365.51	102251373	4	5	4	n	L
14	Duplication	38673786	196.33	38881982	2	4	2	y	E
14	Duplication	63658710	251.89	63912211	2	4	2	y	E
14	Duplication	67838563	317.88	68157617	2	4	2	y	E
15	Duplication	83230545	600	83835487	2	3	2	n	E
15	Duplication	88492748	146.17	88640915	2	4	2	y	E
18	Duplication	8972450	1692.78	10666165	4	5	4	n	L

Table 4.1. Small deletions and tandem duplications placed before or after endoreduplication. Segment copy numbers were generated by PICNIC. Events that had been doubled at endoreduplication were placed earlier. Events not doubled were placed later. Ambiguous cases, for example a copy number change of three, or a surrounding copy number of three, were omitted from this analysis.

4.3.2. Duplication of sequence-level mutations at endoreduplication

In Chapter 3, I placed all known sequence-level mutations on the HCC1187 genome map. To find the position of each mutation, the individual chromosomes of HCC1187 were separated by flow-sorting and exons bearing the mutation were amplified by PCR and the products directly sequenced.

After adding PICNIC minor allele data to my genomic map (Figure 4.3), it was evident that most loci in this genome had duplicated precisely once. I could, therefore, infer whether the common ancestor—before the locus duplicated at endoreduplication—bore a sequence-level mutation or not. If the mutation occurred before duplication it must be present on both copies after the duplication. If a mutation was present on only one of the two loci in the observed karyotype I could infer it happened after endoreduplication (Figure 4.6). The only errors in this analysis would be if a mutation was duplicated or reverted by gene conversion (see Discussion).

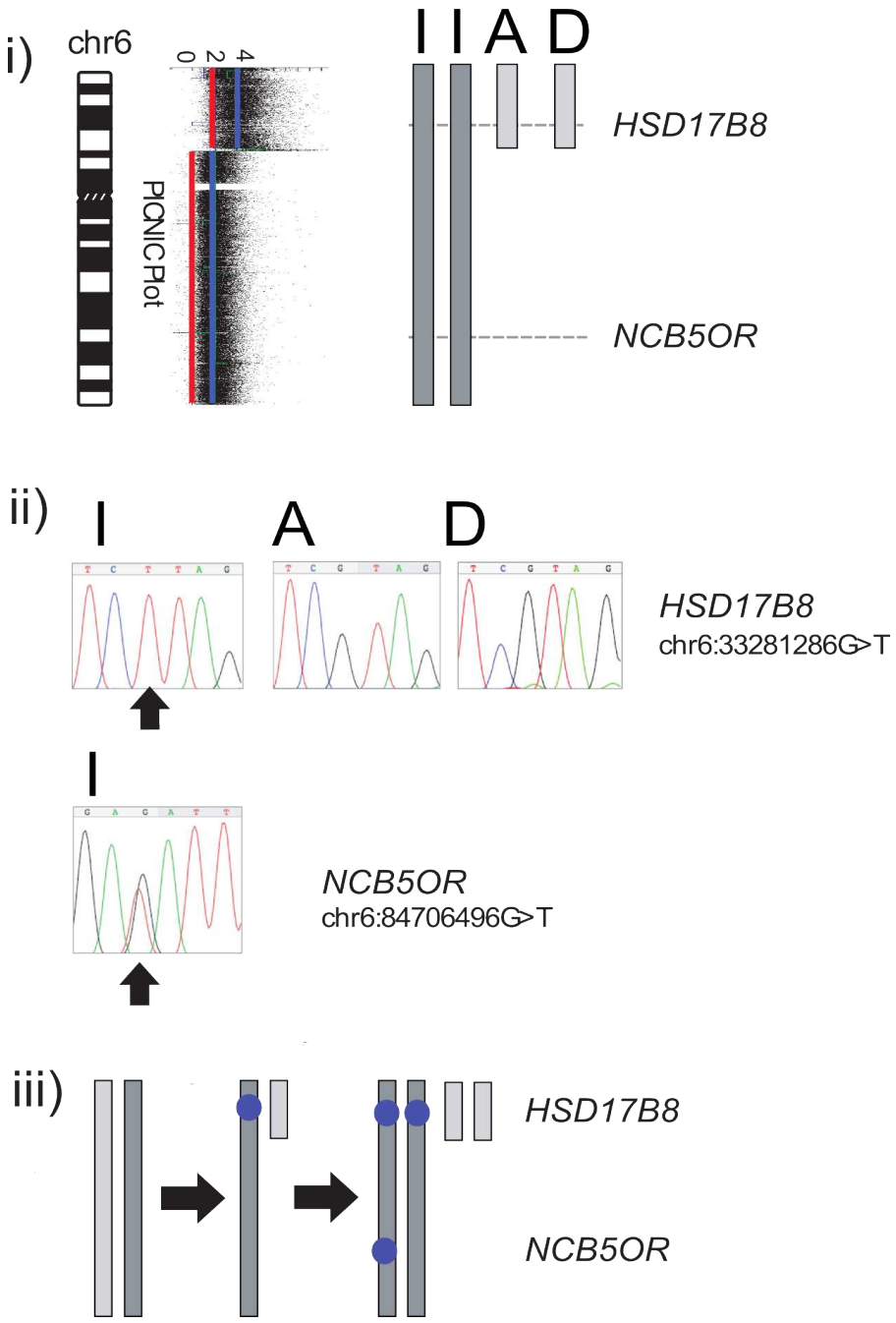


Figure 4.6 The location of point mutations on copies of chromosome 6, and deducing whether they preceded or followed endoreduplication. i) Deducing the parental origin of chromosome 6 segments: the simplest explanation for the allele combinations (blue and red lines on the aCGH plot) in terms of parental origin. Both copies of chromosome 6I originate from parent A and the chromosome 6 segments of A and D originate from parent B. Several small copy number steps are omitted for

clarity. ii**)** Sequence traces show whether mutations are on each isolated chromosome. *HSD17B8*: Chromosome 6I (2 copies) homozygous G>T mutation (black arrow); chromosome 6A and 6D, no mutation. *NCB5OR*: Chromosome 6I, heterozygous mutant (black arrow). iii**)** The likely evolution of the segments of chromosome 6: unbalanced translocation of one copy of chromosome 6 forming der(1)(6pter-6p21.1::1p35->1qter) was followed by duplication of both chromosomes during endoreduplication. *HSD17B8* was mutated on each copy of chromosome 6Į but not on 6A or 6D, while *NCB5OR* was mutated on only one copy of chromosome 6I. The pre-endoreduplication state was likely to be one normal copy of chromosome the other having a mutation in HSD17B8 and suffering unbalanced translocation. The *NCB5OR* mutation occurred after endoreduplication.

Of the 83 previously described sequence-level mutations that I confirmed in Chapter 3, 34 were classed as earlier and 39 later, with only 10 undetermined. Of these 10, 2 were on a chromosome that was too small to be resolved in flow sorting, and 8 were not possible to score, either because they were found on single-copy genome segments, or they were found in a region where parent of origin could not be determined (Figure 4.7, Table 4.2).

All mutations fitted around the scheme of karyotype evolution as expected; if a given mutation had a ‘Parent A’ origin it was only found on Parent A-derived chromosomes. This reinforced the validity of the karyotype evolution scheme and implied that there was no significant incidence in this cell line of other mechanisms that could give duplication of a gene, or removal of one copy, such as gene conversion.

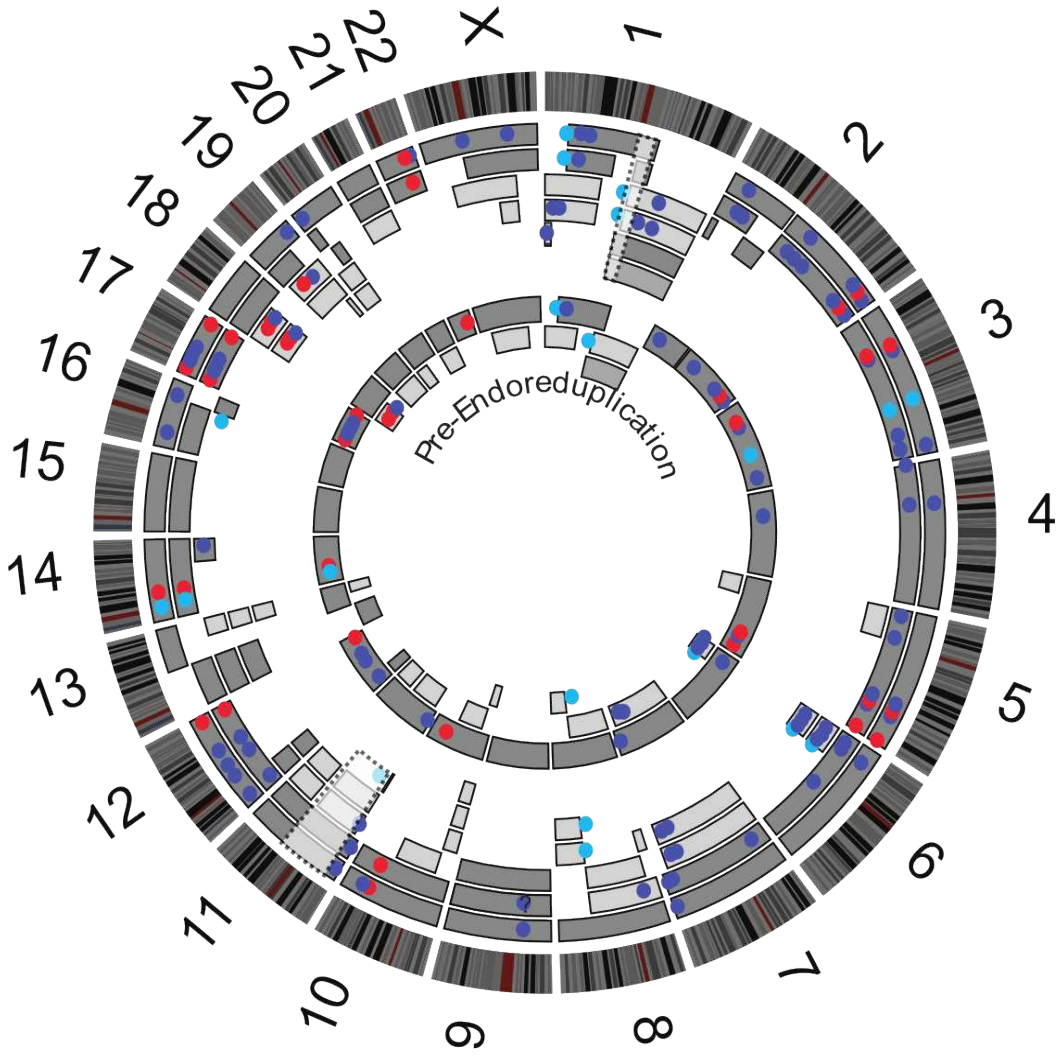


Figure 4.7. Sequence-level mutations and fusion genes before and after endoreduplication. Outer rings are chromosome ideograms and, array painting segments as in Figure 4.3. Inner rings are chromosome segments that must have been present before endoreduplication(equivalent to the state portrayed in Figure 4.4(ii). Coloured dots are different types of mutations, on the outer chromosome segment on which they were observed: truncating (red), non-synonymous (blue), expressed gene fusion (light blue). Mutations that were on two copies of a chromosome segment probably occurred before endoreduplication and are also shown on the inner, pre-endoreduplication genome. Dashed grey boxes on chromosome 1 and 11 indicate regions where parental origin was undetermined, because PICNIC segmentation suggested additional rearrangements had taken place.

Gene	Genomic mutation as reported (2004 build)	cDNA Mutation	Amino acid	Mutation Type	SIFT Score	LogR.E Value	LS-SNP Score	Wood et al (2007) CAN Gene	Early /Late
ARHGEF4	chr2:131632752C>G (homozygous)	1322C>G	T441R	Miss	0	2.31			E
AVPI1	chr10:99429559C>T (homozygous)	94C>T	Q32X	N					E
B3GALT4	chr6:33353713T>C	539T>C	V180A	Miss	0	1.27			E
BAP1	chr3:52415311C>T (homozygous)	781C>T	Q261X	N					E
C4orf14	chr4:57673742A>G (homozygous)	1736A>G	Q579R	Miss	0.64				E
CENTD3	chr5:141014054A>C (homozygous)	4282A>C	T1428P	Miss					E
CTNNA1	chr5:138294082C>T (homozygous)	2032C>T	Q678X	N					E
FHOD3	chr18:32527271C>T	1598C>T	S533L	Miss		0.19			E
FLJ21839	chr2:27191236G>C (homozygous)	1184G>C	R395P	Miss			-1.1		E
FLNC	chr7:128071176G>T	553G>T	D185Y	Miss				X	E
GMCL1L	chr5:177546166delA (homozygous)	741delA	fs	INDEL					E
GPR81	chr12:121739602_121739601insA (homozygous)	165_166insA	fs	INDEL					E
HSD17B8	chr6:33281286G>T	472G>T	V158L	Miss	0.01	0.45	0.11		E
INHBE	chr12:56135771G>C	185G>C	R62T	Miss	0.16	0.07		X	E
KIAA0427	chr18:44541852G>C	1165G>C	V389L	Miss	0.33	1.05		X	E
MYH9	chr22:35012676_35012674delGCA (homozygous)	4200_4202delGCA	indel	INDEL				X	E
NFKBIA	chr14:34942227_34942226insC (homozygous)	427_428insC	fs	INDEL					E
NUP98	chr11:3657478G>T (homozygous)	4955G>T	G1652V	Miss	0.03				E
PAXIP1	chr7:154198087T>G	1370T>G	F457C	Miss	0.18	0			E
PCDHB15	chr5:140607486C>T (homozygous)	2156C>T	A719V	Miss	0.06			X	E
PPHLN1	chr12:41065014G>A (homozygous)	517G>A	V173M	Miss					E
PPP1R12A	chr12:78693190G>C (homozygous)	2301G>C	Q767H	Miss			0.34		E
RASL10B	chr17:31086470G>A (homozygous)	154G>A	V52M	Miss	0.08	0.3	1.27		E
RNU3IP2	chr3:51950902C>G (homozygous)	22C>G	R8G	Miss	0.09				E
RTP1	chr3:188400138C>A (homozygous)	262C>A	R88S	Miss					E
SKIV2L	chr6:32036799C>G	547C>G	L183V	Miss	0.01				E
SLC4A3	chr2:220323514_220323523delGACAA	1291_1300	fs	INDEL					E

	GGACA (homozygous)	delGACAA GGACA 1739_1740							
STATIP1	chr18:31994943_31994944delCT	delCT	fs	INDEL					E
TAS2R13	chr12:10952719A>G	446A>G	N149S	Miss	0.09	0.54			E
TBXAS1	chr7:139064224C>T	256C>T	R86W	Miss	0.01	1.66	-1.04		E
TP53	chr17:7520090_7520088delGGT (homozygous)	322_324delGGT	G108del	INDEL				X	E
TRIM47	chr17:71382450_71382450	1680_1681 insC	fs	INDEL					E
UGT1A9	chr2:234462937G>T (homozygous)	1325G>T	S442I	Miss	0.06	0.17	-1.19		E
ZNF142	chr2:219333250G>A (homozygous)	3005G>A	R1002H	Miss	0.01				E
ABCB8	chr7:150179945C>G	2018C>G	A673G	Miss	0.33			X	L
ADRA1A	chr8:26778286G>T	118G>T	G40W	Miss	0.01	0.4	-0.36		L
C6orf21	chr6:31783819C>G	575C>G	P192R	Miss					L
CAMTA1	chr1:7730843A>G	3240A>G	L1080L	S	1				L
CYP2D6	chr22:40851168G>A	124G>A	G42R	Miss	0.02	-0.97	-0.98		L
CYP4A22	chr1:47323585G>A	1250G>A	G417D	Miss	0	1.41	-1.2		L
DDX18	chr2:118291285G>A	121G>A	G41R	Miss					L
FLJ20422	chr19:19104498A>T	254A>T	E85V	Miss	0.08				L
FLJ20422	chr19:19104499G>T	253G>T	E85X	N					L
FLJ32363	chr5:43541741C>G	798C>G	S266R	Miss					L
FRMPD1	chr9:37730240G>A	1715G>A	G572D	Miss					L
GLT25D2	chr1:180641553G>A	1423G>A	V475I	Miss	0.38	1.26			L
GOLPH4	chr3:169233251C>T	935C>T	A312V	Miss	0.26		-0.66		L
GOLPH4	chr3:169233252G>C	934G>C	A312P	Miss	0.26		-0.66		L
GPNMB	chr7:23086956G>T	1556G>T	S519I	Miss					L
HUWE1	chrX:53537429G>A	1442G>A	R481K	Miss	1	0.27			L
IPO7	chr11:9418649G>T	2767G>T	A923S	Miss	0.3				L
KIAA0934	chr10:363080G>A	3790G>A	V1264M	Miss	0.17	-0.09	-1.46	X	L
LHCGR	chr2:48826897G>A	1690G>A	D564N	Miss	0.01	0.32	0.29		L
LLGL1	chr17:18080933C>G	1566C>G	L522L	S	1				L
MLL4	chr19:40904380C>T	2291C>T	P764L	Miss	0.01				L
MYBPC2	chr19:55650351C>T	2189C>T	P730L	Miss	0.01				L
NCB5OR	chr6:84706496G>T	1009G>T	D337Y	Miss			-1.58	X	L
NOS2A	chr17:23118990G>T (homozygous)	2035G>T	A679S	Miss	0.16		-0.76		L
PDCD6	chr5:359875G>T	367G>T	G123C	Miss	0		-1.27		L
PDPR	chr16:68734948A>T	1637A>T	Y546F	Miss	0.54	1.85			L

PEBP4	chr8:22638372G>C	446G>C	R149P	Miss	0	2.7		L
PLA2G4A	chr1:183651507C>G	1326C>G	H442Q	Miss	0.7	-1.11		L
PLCB1	chr20: 8667928C>T	2229C>T	A743A	S	1	0		L
PLS3	chrX:114703778A>C	1454A>C	D485A	Miss	0	1.75	-0.35	L
PRKAA2	chr1:56881987C>A	1111C>A	P371T	Miss	0.18	0.12	0.87	L
RBAF600	chr1:19237486G>A	4181G>A	R1394H	Miss				L
SATL1	chrX:84168729C>G	830C>G	S277X	N				L
SCN3A	chr2:165812042A>G	2837A>G	E946G	Miss	0	0.31		L
SMG1	chr16:18730825A>C	10735A>C	K3579Q	Miss	0.02			L
SORCS1	chr10:108579379A>C	669A>C	K223N	Miss	0.12			L
SPEN	chr1:16002504G>T	4463G>T	R1488I	Miss				L
SPTA1	chr1:155422865A>T	4743A>T	Q1581H	Miss	0.01	0.34		L
SULT6B1	chr2:37318343A>C	324A>C	A108A	S	1			L
WARS	chr14:99871016G>C	1365G>C	E455D	Miss	0.29		0.59	L
ZNF674	chrX:46144024G>A	253G>A	E85K	Miss	0.28			L
AMPD2	chr1:109885704G>A (homozygous)	2285G>A	R762H	Miss	0			d/k
APOC4	chr19:50140242C>A	224C>A	P75Q	Miss				d/k
C6orf31	chr6:32226401G>A	280G>A	A94T	Miss				d/k
CD2	chr1:117019184G>A	650G>A	C217Y	Miss	0.01			d/k
ITGB2	chr21:45146037_45146013delTGAACA CGCACCCTGATAAGCTGCG	539_563delTGAACACGCACCCTGATAAGC TGCG	fs	INDEL				d/k
ITIH5L	chrX:54706426C>T	227C>T	P76L	Miss	0	3.16		d/k
ITR	chr13:94052277C>A	95C>A	T32N	Miss				d/k
OR1S1	chr11:57739474T>A	682T>A	F228I	Miss	0.47	0.1		d/k
ZCSL3	chr11:31404466_31404470delTCTTG	304_308delTCTTG	fs	INDEL				d/k
ZNHIT2	chr11:64641527G>C	175G>C	A59P	Miss	0.32			d/k

Table 4.2. Sequence-level mutations classed as earlier or later than endoreduplication. Mutations that were duplicated at endoreduplication were classed as earlier and single copy mutations were classed as later. Statistical estimates of a mutation's functionality by SIFT, LogR.E and LS-SNP methods are included (see below) as are Candidate Cancer (CAN) genes from Wood et al. (2007).

4.4. The relative timing of mutations in HCC1187

Overall, the proportions of earlier and later mutations were remarkably similar: 22/50 (44%) of structural changes (translocations, deletions and duplications) and 34/75 (45%) sequence-level changes were classed as earlier (Figure 4.8). Base substitutions predicted to be non-functional by the SIFT (sorting intolerant from tolerant) method(Ng and Henikoff, 2003) were split 9:14 (40% early). Taken together these figures indicate that endoreduplication happened about 40% of the way through the cell line's mutational history (Figure 4.8).

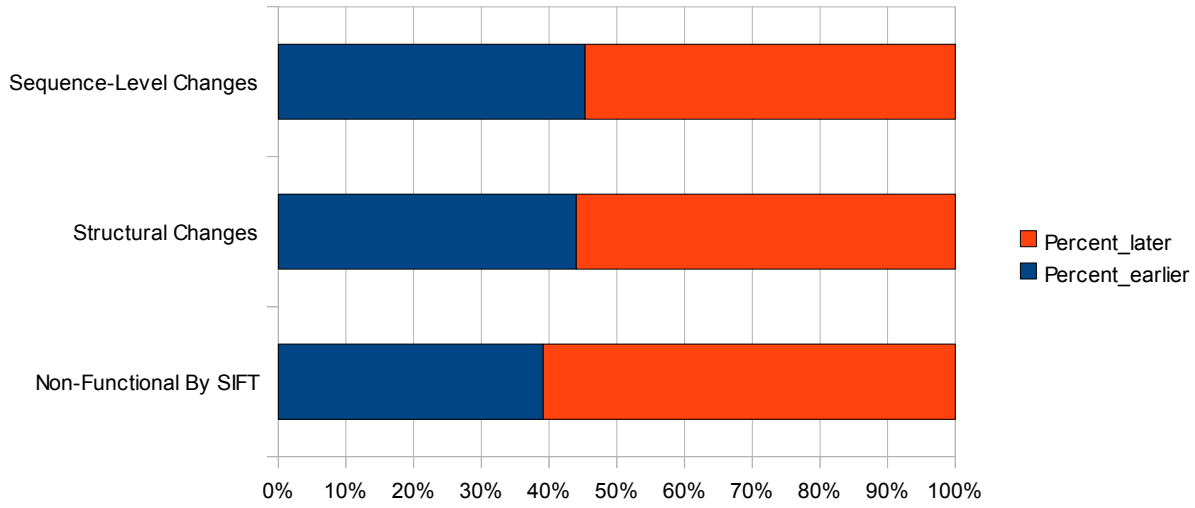


Figure 4.8. The proportions of structural and sequence-level mutations earlier and later than endoreduplication.

The similar proportions of point mutations and structural rearrangements earlier and later implies that the two kinds of mutation occurred broadly in parallel. If genome rearrangement had started substantially later than point mutation, then a higher proportion of sequence-level mutations should have been classified as earlier. Either chromosome instability started before most of the point mutations occurred, or chromosome instability was accompanied by an increased point-mutation rate.

I next investigated where particular subsets of mutations tended to fall earlier or later relative to endoreduplication (Figure 4.9).

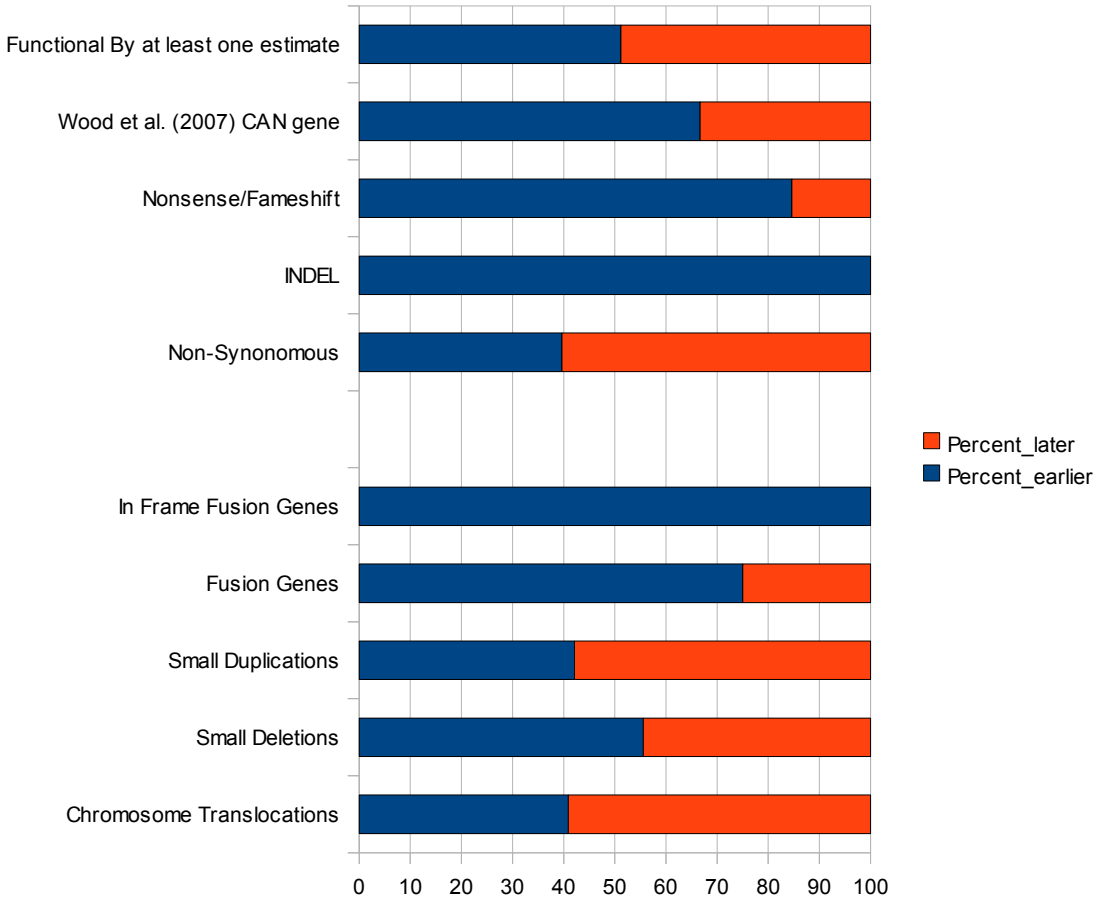


Figure 4.9. Earlier and later classifications of subsets of mutations.

For structural mutations, the earlier group included 9:22 (41%) chromosome translocations. Small duplications were split 8/11 (42%) earlier and small deletions were split 5:4 (55%) earlier. Expressed fusion genes were split 6:2 and all three in-frame fusions fell early.

For sequence mutations, 23:35 (40%) non-synonomous mutations fell early. Mutations found by Sjoblom et al. (2006) and Wood et al. (2007) had previously been investigated by the algorithms SIFT, logR.E. and LS-SNP (Ng and Henikoff, 2003; Clifford et al., 2004;

Karchin et al., 2005; Wood et al., 2007). I repeated these analyses on mutations reported only in the COSMIC database using CANpredict software online (Kaminker et al., 2007). Some mutations could not be analysed by the above methods, for example, *FLNC*. Classifiable mutations that were predicted to be functional by at least one of the estimates were split 21:20 (51%).

Wood et al (2007) identified genes likely to be drivers as 'candidate cancer genes' (CAN) based on their observed versus expected mutation rate and several bioinformatic estimates of functionality. CAN genes showed a bias towards the earlier category that may have been statistically significant (see below). Six of the nine CAN genes found in HCC1187 were found in the earlier category (*INHBE, KIAA0427, MYH9, PCDHB15, RNU3IP2, TP53*) and three in the later category (*ABCB8, KIAA0934, NCB5OR*).

Strikingly different from the overall distribution of mutations in HCC1187 was the proportion of sequence-level truncation mutations in earlier rather than later categories: All eight INDEL mutations happened earlier, and combining this figure with nonsense mutations showed 11:13 (85%) truncation mutations happened earlier.

4.5. Statistical Estimates of the number of non-randomly distributed mutations

Certain classes of mutation appeared to deviate from the expected 40:60 split. Most notably, the mutations predicted to truncate proteins nearly all happened earlier. I next used a statistical model to estimate the number of mutations that showed non-random timing earlier or later. The mathematical model and R scripts to run it was made by Professor S Tavaré, CRUK Cambridge Research Institute. Its application to these data and interpretation is my own work. The model is described in Chapter 2.9 and R scripts to run it are in Appendix 2.4.

The model assumed that any given class of mutations is a mixture of events that have to happen early or late and events that can fall at random. The algorithm finds the most likely number of non-randomly timed mutations (the maximum likelihood estimator, or MLE) and the 95th percentile confidence intervals we can have in that number given an estimate, p, of the timing of endoreduplication. The calculations below make no assumptions about the mutation rate, only that the relative proportions of different classes of mutations before and after endoreduplication were similar in the absence of selection. E.g. if the rate of missense mutation changed after endoreduplication, so did the rate of indel mutation. The degree of non-randomness of the earlier and later classification can be calculated for a range of possible scenarios, but in all cases the implications are that a substantial number of mutations were non-randomly timed.

4.5.1. Maximum Likelihood estimators.

For these calculations, I use the estimated p-value of 0.4 as described above. Consider, for example, the proportion of truncating mutations earlier and later: The observed 11:2 distribution seems improbable. A proportion of these mutations must, presumably, have occur earlier. The MLE in this case is 10 mutations that show non-random timing, i.e. had to happen before endoreduplication. The 95% confidence intervals using a bootstrap approach are 7 and 12, If we then ask the reciprocal question: how many mutations had to happen late, p=0.6, late mutations=2, the MLE for late mutations is 0 (Figure 4.10). I next applied the model to all of the subsets of mutations mentioned above.

Of the eight fusion genes, it was possible to place earlier or later, the most likely number that had to occur early was five with 95th percentile confidence intervals of between two and seven. For small deletions the MLE was three with 95th percentile confidence intervals of between zero and six. It is, therefore, possible that this distribution occurred by chance. A similar case was observed for Wood et al. (2007) CAN genes (MLE=4 or 5 confidence interval 0 to 7) and predicted functional mutations (MLE=8, confidence interval 0 to 16)(Figure 4.10).

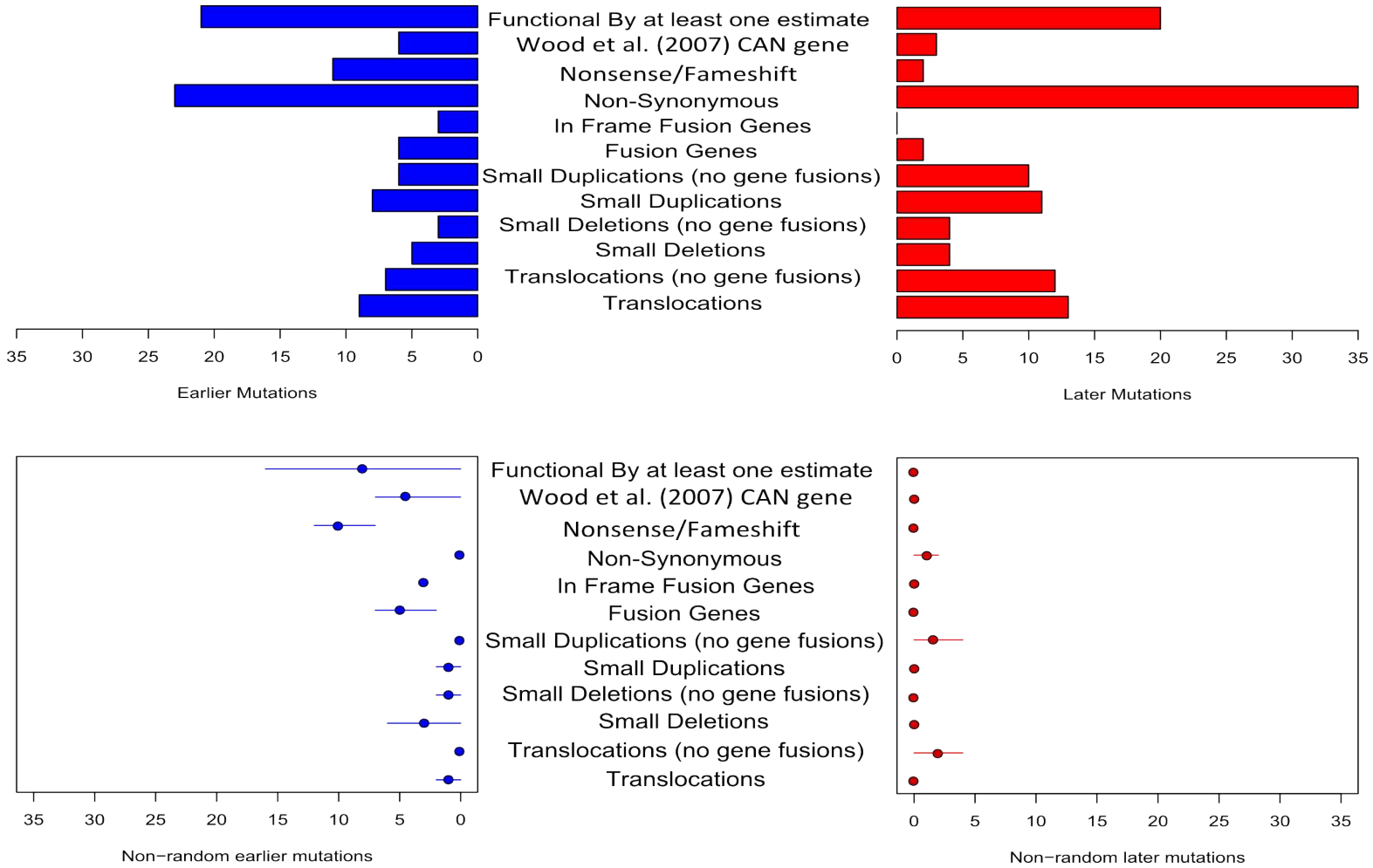


Figure 4.10. Non-randomly timed mutations in HCC1187. Upper plots are the total number of mutations placed earlier (blue)or later (red). Lower plots are MLEs (red and blue dots) and 95th percentile confidence intervals (horizontal red and blue lines).

4.5.2. Did specific mutation rates change over time?

Central to the statistical model was the assumption that the relative proportions of different classes of mutations before and after endoreduplication were similar in the absence of selection. The HCC1187 tumour received undisclosed chemotherapy treatment prior to the derivation of the cell line (Gazdar et al., 1998), so there is a possibility that a certain type of mutation would be artificially concentrated at one stage of tumour evolution. To address this concern I used my statistical model to compare the rates of the different types of mutation earlier and later (Figure 4.11).

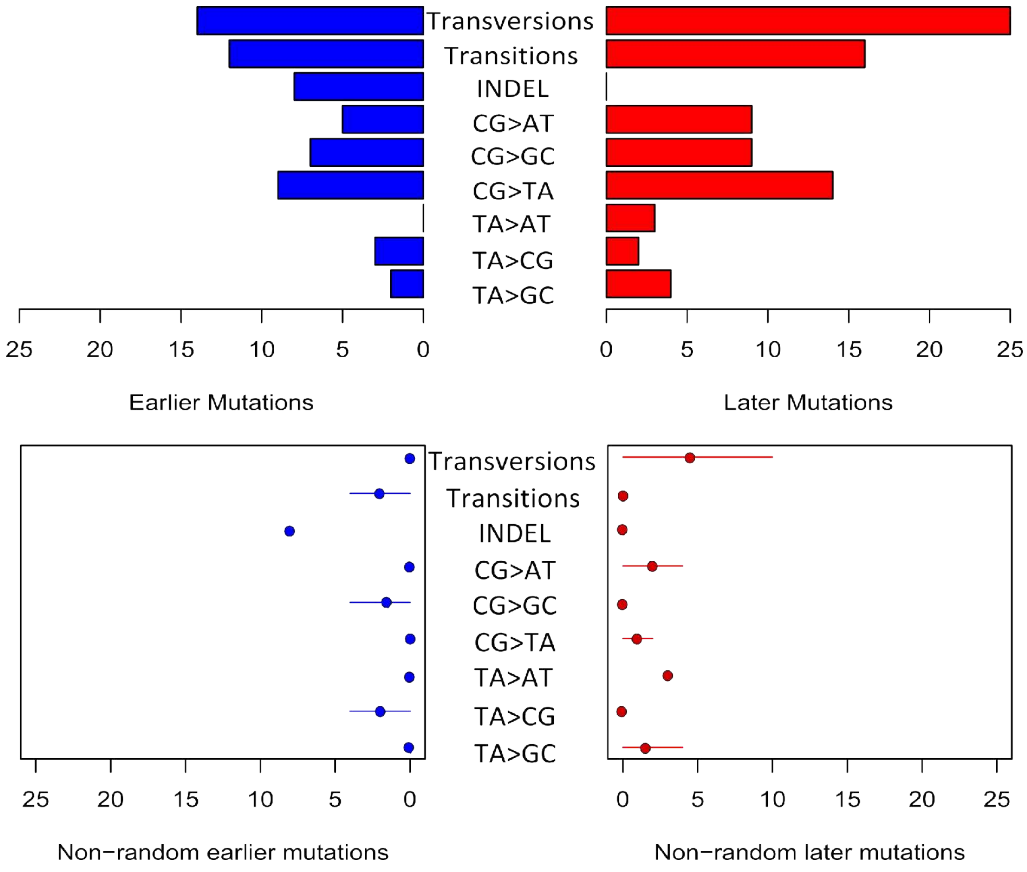


Figure 4.11. The proportion of different types of mutation earlier and later. Upper plots are the total number of mutations placed earlier (blue)or later (red). Lower plots are MLEs (red and blue dots) and 95th percentile confidence intervals (horizontal red and blue lines).

Most types of mutation seemed to accumulate in the expected manner given the p value of 0.4. There did appear to be some small fluctuations earlier and later, but these may well have been due to chance and the small number of mutations sampled. Indel mutations were all concentrated earlier and this is discussed below. There was a possible excess of transversion mutations in the later category (MLE=4 or 5, confidence interval 0-10) due to CG>AT and TA>AT mutations. The G>T transversions could be explained by oxidation in culture (Kino and Sugiyama, 2001), but again, this distribution could be due to random chance. But even if we base our estimate of p only on transitions, we still see a ratio of 43 percent to 57 percent. If chemotherapy or oxidation in culture did generate some mutations in HCC1187 they were not sufficient to have biased my estimate of the timing of endoreduplication, but nevertheless, I discuss this possibility below.

4.5.3. What if endoreduplication were a late event?

Now consider what happens if we allow p to vary (Figure 4.12), i.e. suppose that some of the non-synonymous missense mutations were selected to occur at a specific time. For p values above 0.4 we see that the MLE for early-selected truncating mutations decreases and above p=0.8 the MLE becomes zero. This corresponds to endoreduplication being a relatively late event so the truncation mutations would be mostly earlier due to random chance.

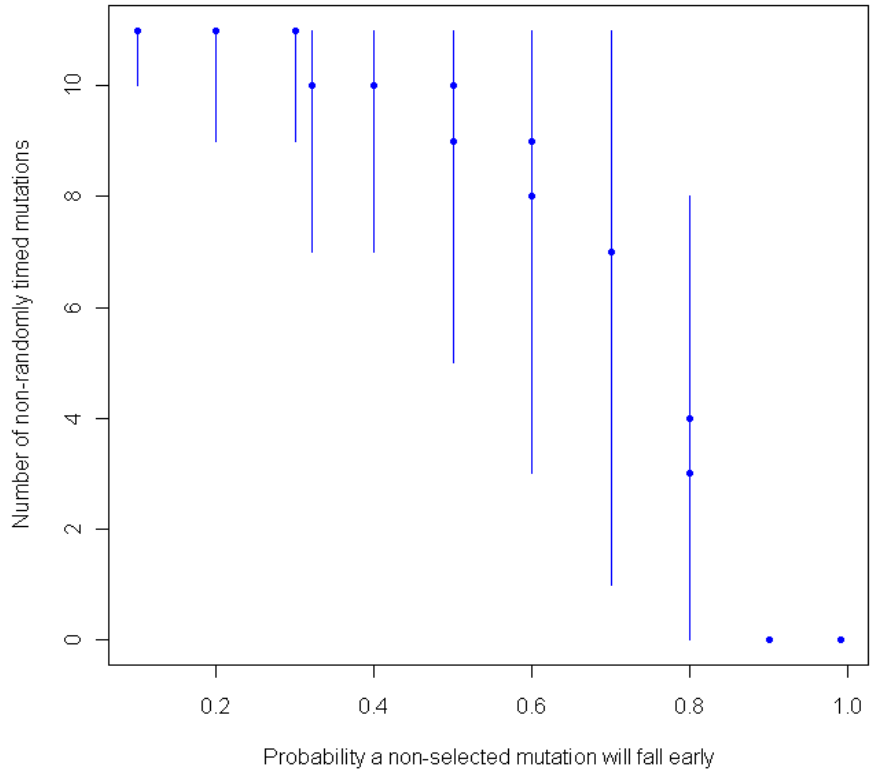


Figure 4.12. Estimates of the number of truncating mutations selected to be early, for various values of p, the probability of a non-selected mutation falling early. MLEs are represented by blue dots, The 95th percentile confidence intervals generated by bootstrapping are vertical bars. Note that for some values of p there are two, equally likely MLEs.

However, if endoreduplication was late, there is a large excess of non-synonymous missense mutations in the late category and we have to conclude that many of the late non-synonymous mutations were had to happen late. For example, if p=0.85, MLE for non-randomly timed *late* non-synonymous mutations is 35 with 95% confidence intervals 30, 38. If we plot this estimate for varying p we see that, the later we suppose endoreduplication to have happened, the more late mutations there must have been (supplementary figure 4). When endoreduplication happened early we see that the MLE for late non-random events drops to zero. This just means that a significant proportion of

mutations did not show non-random timing. It is still likely that there were some non-randomly-timed non-synonymous mutations they just do not cluster earlier or later.

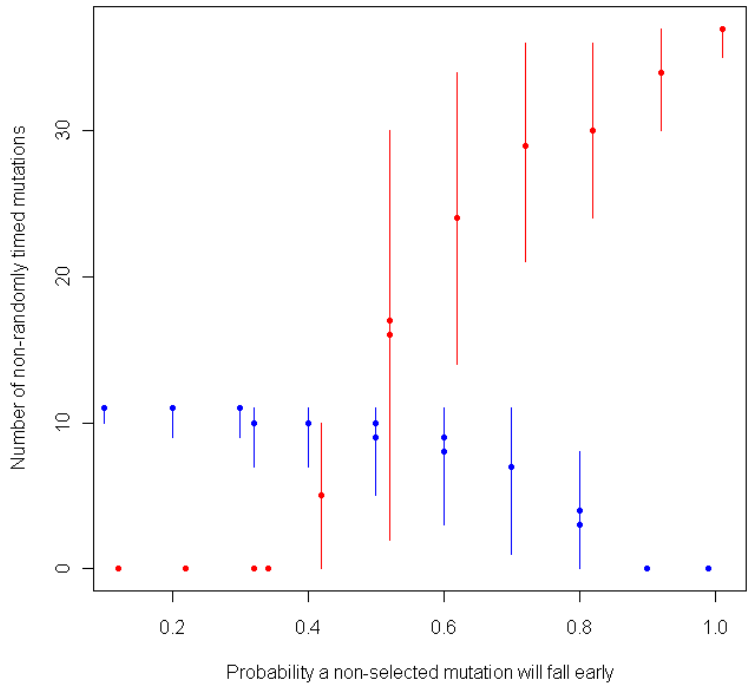


Figure 4.13 Combining maximum likelihood estimates of truncating mutations selected to be early (blue) and non-synonymous missense mutations selected to be late (red).

4.6. Discussion

Endoreduplication in HCC1187 proved to be a useful milestone, because numbers of structural changes and point mutations were roughly equally distributed between the earlier and later categories, so endoreduplication occurred about 40 percent of the way through the evolution of this genome. The earlier versus later classification may help us to understand a variety of issues including the timing and origins of chromosome instability (CIN) and the drivers versus passengers problem.

4.6.1. The timing of CIN

The distribution of mutations allows us to speculate on the timing of CIN in this cell line. There has been much discussion of when CIN occurs, for example some have suggested it as a key facilitator of early tumourgenesis, notably causing loss of heterozygosity of APC in colorectal cancers (Rajagopalan et al., 2003; Rajagopalan and Lengauer, 2004). In contrast, Johannsson et al. (1996) suggested that the extensive rearrangements of carcinoma karyotypes might be late progression events. Others favour a critical role of 'crisis', a transient period of chromosome instability caused by telomere loss (DePinho and Polyak, 2004; Stephens et al., 2011). These latter views suggest that the relative rates of different kinds of mutation would change during the evolution of the tumour.

There is no evidence from these data that different kinds of mutation, e.g. point mutation versus translocation or whole chromosome loss, occurred at radically different times in the development of the tumour. If chromosome instability appeared after a significant number of point mutations had accumulated, we would have seen the majority of point mutations before endoreduplication, which I did not observe. It follows that either CIN started before most of the point mutations, or the onset of CIN was accompanied by an increased point-mutation rate. An important value of the classification is that mutations that may cause chromosome instability must generally be in the 'earlier' group as, by definition, they must pre-date almost all chromosome changes, which in most cases are quite numerous before endoreduplication.

4.6.2. Early Tumour Suppressor Loss

The high proportion of earlier truncating mutations, especially indels (8 earlier vs 0 later), could be explained in two ways: i) the rate of indel mutations was high before endoreduplication and low after, relative to most other types of mutation ii) passenger indels accumulated in the same way as other passenger mutations but more indels accumulated early because they were selected.

I consider (ii) most likely because for 9/11 truncating mutations a chromosome loss before endoreduplication caused loss of the second wild type allele. This is consistent with chromosome instability facilitating early tumour suppressor loss, as has been suggested previously. Indeed, the earlier truncation mutations include known and candidate tumour suppressor genes *TP53, BAP1* (*BRCA1*-associated protein), *CTNNA1* (CateninA1) and *NFKBIA* (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor) (Jensen et al., 1998; Jensen and Rauscher, 1999; Ventii et al., 2008; Liu et al., 2007, 2010; Ding et al., 2010; Osborne et al., 2005); others were *AVPI1*, *GMCL1L*, *GPR81*, *MYH9* , *SLC4A3*, *ELP2* and *TRIM47*. These data, therefore, support the view that early tumour suppressor loss is consistent with tumour evolving monosomically and that driver mutations that cause gene inactivation will be concentrated pre-endoreduplication. An explanation for this phenomenon is that loss or inactivation of two alleles pre-endoreduplication is more likely than loss/inactivation of four alleles post-endoreduplication (Muleris and Dutrillaux, 1996).

4.6.3. Non-random timing of predicted functional substitutions

Gain of function mutations are not under the same numerical constraints as tumour suppressors. Where two hits are required to impair tumour suppressor gene function, only a single mutation is required for oncogenic gains of function and we may, therefore, see these mutations either side of endoreduplication. Some, however, suggest that all useful mutations in a tumour must pre-date the invasive stage (Bernards and Weinberg, 2002; Edwards, 2002). In this case, we might expect to see functional mutations clustering early and there is a suggestion of this in the data. Eight mutations predicted to be functional by at least one of the three bioinformatic estimates were most likely to show non-random timing as were four or five of the CAN genes. Although the 95th percentile confidence intervals included zero, these data are still suggestive of an early bias for functional mutations.

4.6.4. Non-random timing of gene fusions

We can be 95 percent certain that at least two, but as many as eight, fusion genes had to happen early. The most probable estimate is six. This is a surprisingly high proportion given the proposition by Stephens et al. (2009) that most gene fusions were passenger events. Three of the gene-fusions appeared to be in-frame and these all fell early. This makes them likely candidates for selected events. But as Hampton et al. (2008) noted, out of frame gene-fusions may also be selected events as they potentially inactivate one or both of the genes involved. As two of the gene fusions were caused by homozygous deletion, either loss of function of the fused genes themselves or genes deleted in the intervening segment could also be selected events.

4.6.4. The Timing of Endoreduplication

Interpretations of the earlier and later classes depend on when HCC1187 endoreduplicated as endoreduplication can be observed both *in vitro* and *in vivo* (Schwarzacher and Schnedl, 1965; Dutrillaux et al., 1991). There is some evidence that endoreduplication occurred *in vivo* in this case. The original ploidy of HCC1187 was not reported, only that shortly after its derivation, HCC1187 had multiple ploidy indices by flow cytometry (Gazdar et al., 1998; Wistuba et al., 1998). However, around 60 percent of mutations occurred after endoreduplication. It would be surprising if so many happened in culture, given that cell lines largely recapitulate the genomic aberrations observed in primary tumours (Neve et al., 2006; Chin et al., 2007). If endoreduplication happened *in vitro*, only ‘earlier’ mutations happened *in vivo,* and all driver mutations will be in the ‘earlier’ set; whereas if endoreduplication happened *in vivo*, some driving mutations will be present in the ‘later’ group. In either case our estimate of the number of earlier driving mutations remains the same.

4.6.5. How accurate was this analysis?

The above conclusions depend on the accuracy of my earlier and later classification of mutations. I was confident that the tumour had undergone endoreduplication as it showed two characteristic signatures of this phenomenon: multiple duplicated rearrangements and multiple duplicated homozygous regions. Given that there had been an endoreduplication, I reconstructed the main steps of HCC1187 karyotype evolution (Figure 4.4) by assuming that the simplest possible sequence of events had happened. Implicit was the assumption that, as far as possible, all duplications had occurred at endoreduplication. The deduced sequence of chromosome changes was almost exactly consistent with monosomic evolution (allowing some whole-chromosome losses that had occurred without translocation, and two translocations where no loss occurred).

4.6.6. More complex evolutionary routes?

Three duplications could not be explained by endoreduplication: these were three chromosome segments of the same parental origin that were present in three copies, chromosome 9 from parent A, chromosome b (der(13)t(10;13)) and the chromosome 13 portion of chromosome j. The simplest route to these triplications was via endoreduplication followed by an additional single-chromosome duplication.

It is possible a small number of steps in the evolution were more complex than I deduced, but this would not have altered the earlier versus later classification very often. Specifically, if all three triplicated chromosomes had taken the more complex evolutionary route (perhaps duplication followed by endoreduplication, followed by loss), the classification of no more than three point mutations could be affected, moving them from the later category to the undetermined' class. For example, for chromosome 9 from parent A, we assumed that independent duplication followed endoreduplication, so the mutation of *FRMPD1* occurred later; but if the duplication had preceded endoreduplication and a copy was later lost, the mutation would be 'unclassifiable'.

Some mutations were omitted from analysis. These were from the complex regions of 10p and 11q where the parent of origin could not be accurately determined. The omitted mutations comprised eight non-synonymous missense and two truncating mutations. Even if we consider the most unfavourable case, that the two truncating mutations were later, the MLE for non-random early events becomes 9 with 95% confidence interval between 4 to 12.

4.6.7. Gene Conversions?

If a gene conversion had occurred, then a later mutation would appear to have occurred earlier. No gene conversions had to be invoked as all mutations were confined to one parent of origin, implying that gene conversion was rare or absent in this cell line. This is consistent with previous studies which showed that unbalanced translocations and whole chromosome loss account for the bulk of loss of heterozygosity in epithelial cancers (Thiagalingam et al., 2001, 2002; Ogiwara et al., 2008).

4.6.8. A lower estimate of the number of driving mutations in HCC1187

Clustering of a certain type of mutation early could imply one of two things: i) A particular mutational mechanism was more active earlier than later or ii) that non-randomly timed mutations had to happen early, so were selected. If we assume that the second option is correct then a lower estimate for the total number of earlier selected events is nine - two gene-fusions and seven nonsense/frameshift mutations. These are the lower confidence intervals for two of the classes of mutation where this limit was greater than zero. If one, however, adds the MLEs of different types of mutations, the number of non-randomly timed selected events may be in excess of twenty.

Chapter 5

Preliminary analysis of two related cell lines by massively parallel paired-end sequencing

5.1. Introduction

With the advent of massively parallel paired end sequencing it is now possible to rapidly define genome rearrangements in cell lines and primary tumours. However, surveys of primary tumour genomes will probably require high physical coverage given their heterogeneity and contamination with stromal cells. We can reasonably expect ten-fold physical coverage from massively parallel paired end sequencing experiments (Stephens et al., 2009). But if one imagines a triploid tumour genome with thirty percent stromal cell contamination – as is commonly observed – the nominal ten-fold physical coverage of a haploid genome is only about two-fold in real terms. This translates to around 60 percent of clonal, single copy rearrangements being sampled twice or more. For non-clonal rearrangements, this figure would decrease. Thus the results will be biased towards the dominant clone of the particular region of tumour the sequencing library was made from.

It is likely that a higher proportion rearrangements can be sampled in cell lines as they are more clonal and do not contain any stromal cells. But as many cell lines are from late-stage disease and have existed in culture for years one can never be sure that any rearrangement was an *in vivo* event.

A possible solution to these problems is comparative lesion sequencing (Jones et al., 2008; Shah et al., 2009). In this approach, samples from the same patient at two different time points are sequenced and compared. Rearrangements common to both samples probably occurred earlier and *in vivo*, and, if using cell lines, a higher proportion of these rearrangements can be sampled. One such model currently exists for breast cancer, the VP229 and VP267 cell lines. In this chapter, I describe their genome structures using data from massively parallel paired end sequencing.

5.1.1. VP229 and VP267 Cell Lines

The two cell lines, VP229 and VP267, are from the same breast, VP229 being derived from a local excision specimen and VP267 from a mastectomy specimen following Tamoxifen treatment 12 months later (Table 5.1) (McCallum and Lowther, 1996). Early

chromosome analysis of the two lines showed considerable complexity, but several features were present in both cell lines. This implies that VP267 most likely arose by "clonal evolution *in vivo* from cells similar to those which were present in the biopsy used to establish VP229." (McCallum & Lowther 1996 p.258)

Cell Line	Patient Age	Previous Treatment	Histology	ER status (DCC)*	Survival after Op
VP229	47	None	Ductal Grade III	0/0	2 years
VP267	48	Tamoxifen	Ductal Grade III	0/38	1 year

Table 5.1. VP229 and VP267 information. *DCC is a dextran-coated charcoal assay (McGuire and DeLaGarza, 1973).

The karyotypes of both cell lines were among the most complex that Davidson et al. (2000) observed, meaning that techniques such as array CGH and array painting would be very difficult to interpret. This made them a good candidate for investigation by massively parallel paired end sequencing. Studying the genomic structures of these two cell lines is interesting for several reasons:

1) Massively parallel paired-end sequencing would allow me to rapidly define chromosome aberration break points and predict gene-fusions in these two complex genomes
2) Cell lines are known to evolve in culture so one can never be sure if a given rearrangement happened *in vivo* or *in vitro* (Roschke et al., 2002, 2003). As VP229 and VP267 are separate isolates from the same patient, any rearrangement found in both lines was probably present in the original tumour *in vivo*.
3) The cell lines were originally scored as ER-negative (McCallum and Lowther, 1996), but according to the RT-PCR, immunohistochemisrty and gene-induction assays of Ghayad et al. 2009, the VP cell lines express a functionally active ERα. VP229 was sensitive to both commonly used ER agonists, Tamoxifen and Fulvestrant but strikingly, VP267 cells were resistant to both drugs (Ghayad et al., 2009). VP267 is a relapse following Tamoxifen treatment, so the genetic lesion responsible (if retained) would only be found in VP267.

5.2. Bioinformatic Processing of Sequence Data

5.2.1. Library Preparation, Bioinformatic Processing and Physical Coverage Calculations

Sequencing libraries for VP229 and VP267 were constructed by Dr JC Pole and Dr I Schulte. Bioinformatic processing was via a bioinformatic pipeline constructed by Dr K. Howe, Dr S.L. Cooke and Miss C.K. Ng (Wellcome Sanger Institute and CRUK Cambridge Research Institute). Briefly, image analysis and base calling modules FIRECREST and BUSTARD (Illumina) were used to produce raw sequence data. Sequences were aligned to the HG37 reference genome using the Burrows Wheeler alignment (BWA)(Li and Durbin, 2009) as it is faster and possibly more accurate than MAQ (Meyerson et al., 2010). Reads that would not align with BWA were passed on to Novoalign alignment which is very accurate but slow (Hercus, 2009). Run statistics for the two libraries are summarised in table 5.2 and estimated physical genome coverage is shown in table 5.3.

Cell Line	Unique Normal Pairs	Total Bases Covered
VP229	47056862	20661043936
VP267	51024575	21451529751
VP229 / VP267 combined	98081437	42112573687

T**able 5.2 Run statistics for VP229 and VP267 sequencing libraries**

	Haploid genome coverage	Diploid genome coverage	Triploid genome coverage
VP229	6.9	3.4	2.3
VP267	7.2	3.6	2.4
VP229/VP267 combined	14	7	4.7

Table 5.3. Estimated physical genome coverage

Library preparation has a ligation step, so it is possible spurious structural variants would be generated. To counter this, each structural variant had to be supported by at least two

unique paired reads to be considered further. Estimated physical genome coverage was calculated as above and using the Poisson distribution function, the proportion of events hit two times or more was also calculated (Table 5.4).

Cell Line	Haploid genome coverage	Diploid genome coverage	Triploid genome coverage
VP229	0.99	0.8	0.59
VP267	0.99	0.91	0.59
VP229/VP267 combined	0.99	0.99	0.96

Table 5.4. Estimates of the proportion of events hit twice or more using the Poisson function given the coverage estimates from table 5.3.

VP229 and VP267 have modal chromosome numbers of 62 and 59 respectively but it is not possible to state the ploidy of genome as a whole as many loci deviate from the median copy number. Instead, I estimated the total DNA content of VP229 by adding up the total copy number of PICNIC segments from array CGH. Using this approach, there were 8.2 billion bases in VP229 and this translates to 2.7 times the haploid genome. For VP267, array segmentation was not available.

5.2.2. Copy Number Estimation

Generating copy number data from massively parallel sequencing data was done by Miss. E.M. Batty and Dr. K. Howe (Batty, 2010). A description of the work is included here as I later rely on these data for structural variant validation.

The majority of read pairs mapped normally to the reference genome. This was defined as read pairs mapping to the same chromosome, in the correct orientation and separated by a distance of no more than three standard deviations from the median fragment size (Figure 5.1).

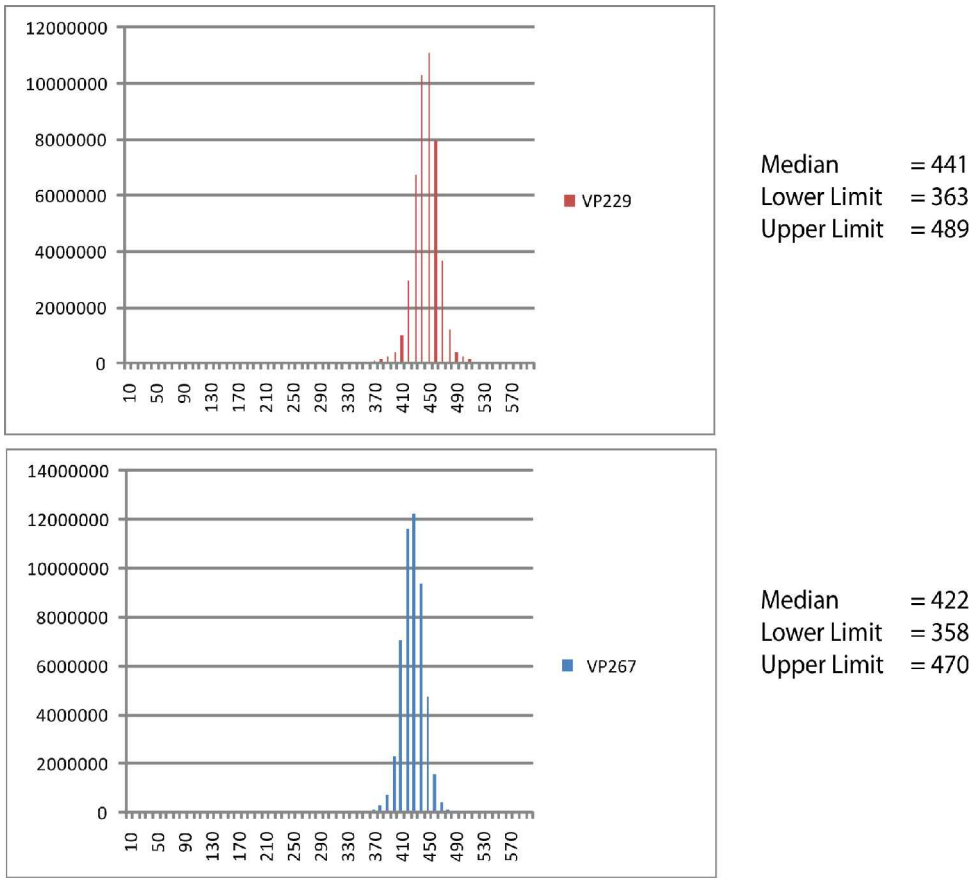


Figure 5.1. **Frequency distribution of sequencing reads**. Any read pair that mapped closer together or further apart than the upper and lower limits (defined as three standard deviations from the median) was considered abnormal.

The number of reads from any given region of the genome should be proportional to its copy number, so gained regions should contain more normally mapping read and lost regions fewer. To generate copy number data comparable to array CGH, some biases in the data had to be removed. Sequencing reads must align uniquely to the genome or they are discarded, so a region rich in repeats would have a disproportionately low number of mapping reads and appear to have decreased in copy number. The level of sequence uniqueness of the reference genome for any given region is recorded as a 'mapability' score. Using mapabilty, bins of varying size across the genome were calculated that should have the same number of normal reads map back to them if the genome being sequenced were normal and diploid (K Howe et al. unpublished).

A second bias was evident in copy number plots and related to GC content. It seemed that GC-rich regions gave a disproportionately high number of reads even with mapability binning and the experimental modifications suggested by Quail et al. (2008). This phenomenon is well documented in array CGH experiments and has been termed a 'GC wave' (Marioni et al., 2007; Leprêtre et al., 2010). To correct the GC wave without a DNA from a matched normal sample a similar procedure to that of Marioni et al. (2007) was used. The number of sequencing reads in each bin were plotted against GC content. The deviation of each point from the normal range was plotted as a loess line. The equation of this line was then used to correct the CG wave, smoothing the data (Batty, 2010). The corrected data was then segmented using circular binary segmentation in a similar way to array CGH (Venkatraman and Olshen, 2007).

This method gave CGH results comparable to that of SNP6 array CGH; an example is shown in Figure 5.2. It is also likely that this method gives a more accurate estimate of the copy number of amplified regions than array CGH. Array CGH approaches are based on hybridization of labelled DNA to the array. Highly amplified regions are found in hundreds of copies so can saturate the array (Chiang et al., 2009). This fact is reflected by a maximum copy number estimate of only 15 by PICNIC (Greenman et al., 2010).

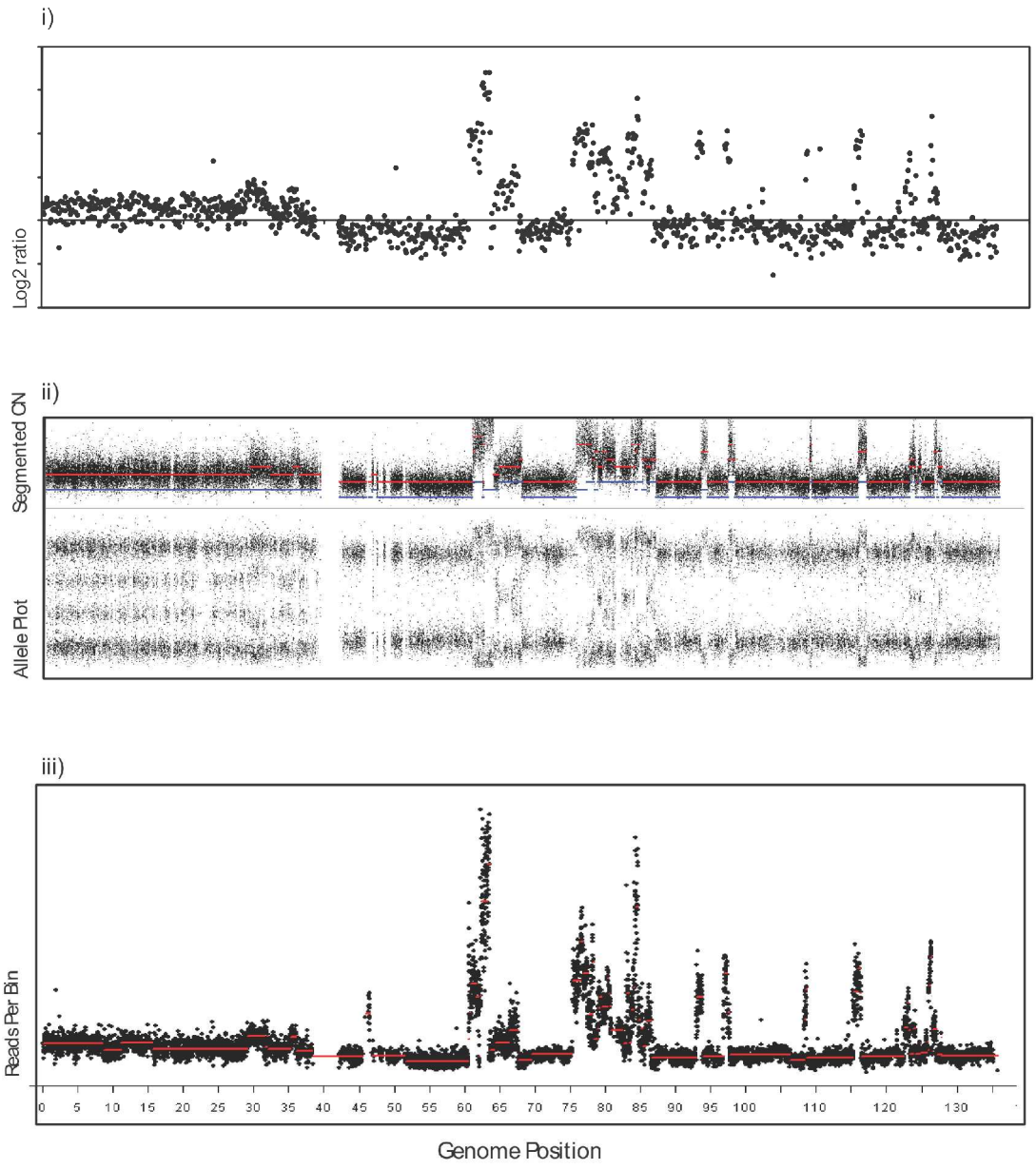


Figure 5.2. Copy number of VP229 chromosome 10 assessed by three methods.
i) Tiling path BAC array (Blood, 2006). ii) SNP6 array CGH. iii) Loess-corrected copy number based on normally-mapping paired reads (Batty, 2010).

5.2.3. Predicted Structural Variants

Read pairs that aligned too far apart, to different chromosomes or in the wrong orientation indicated possible structural variants (SV). The different types of structural variants are

named after their likely effect given the simplest possible interpretation (Figure 5.3).

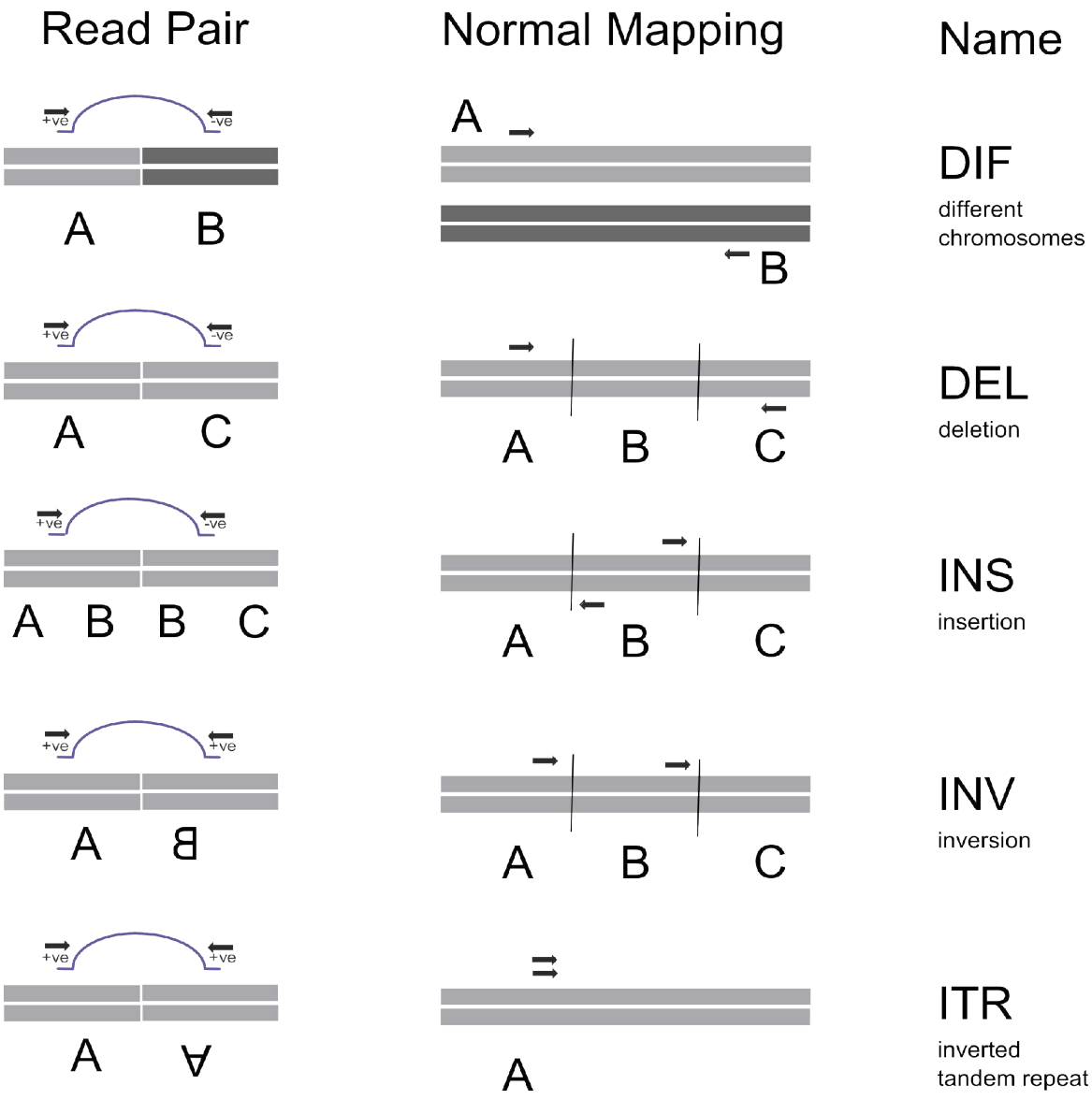


Figure 5.3. Interpretations of aberrantly-mapping read pairs for short insert libraries. Light grey boxes are DNA strands from one chromosome, the upper one is the positive strand. Dark grey boxes are DNA strands from another chromosome. The examples here are all from the positive strand of derivative chromosomes but read pairs originating from the negative strand are also generated in such sequencing experiments.

The numbers of aberrantly mapping read pairs for VP229 and VP267 are below (Table 5.5).

SV Type	VP229	VP267
Total	213056	120540
DIF	36416	88412
INV	74486	10995
DEL / INS	102154	21133

Table 5.5. Summary of aberrantly mapping read pairs. DEL=deletion, DIF=translocation, INS=insertion, INV=inversion, ITR=inverted tandem repeat.

5.2.3. Clustering of Structural Variants

To find read pairs that traversed structural variants, similar reads were clustered together using a custom Perl script (K Howe et al. Unpublished). Each structural variant had to be supported by at least two unique read pairs to be considered further. These clusters of paired reads provided a minimum region for the chromosome break point (Figure 5.4 and table 5.7). If structural variants from the two different cell lines could clustered together they were likely to have been in the common ancestor of VP229 and VP267, named hereafter as "VP-Ancestor."

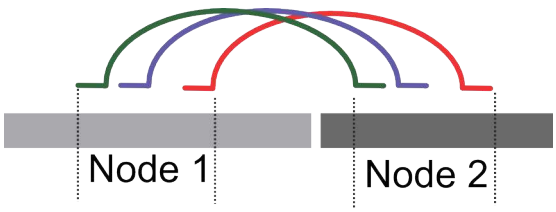


Figure 5.4. Aberrantly mapping reads clustering strategy. Three read-pairs (green,purple and red lines) span an inter-chromosome translocation (or insertion) breakpoint, chromosome A is light grey, chromosome B, dark grey. Similar reads are clustered into 'nodes'. The best estimate for the position of this break point is between the right hand side of node 1 and the left hand side of node 2.

After clustering, deletions, inversions and insertions of less than 10kb were frequent in these data. Most are hypothesised to be germ line variants so were not considered further, as in previous studies (Campbell et al., 2008; Stephens et al., 2009). If the structural variant less than 10kb but was predicted to fuse two genes I first validated the genomic breakpoints by PCR as below. The remaining structural variants are listed in Table 5.6.

SV Type	In both VP229 and VP267 (VP-Ancestor)	Only in VP229	Only in VP267	All
DEL	61	19	30	110
DIF	153	32	73	258
INS	41	14	28	83
INV	81	18	23	122
ITR	6	99	2	107
Total	342	182	156	680

Table 5.6. Predicted structural variants with reads >10kb apart or on different chromosomes

5.3. Validation of Structural Variants

5.3.1.Validation by PCR

To confirm that the bioinformatically-predicted structural variants were really present in each cell line, I performed PCR across genomic breaks as in previous studies (Hampton et al., 2008; Stephens, 2009). There were a large number of structural variants so I could not validate all of them by this method because of time and cost. Instead, I attempted to validate a subset of rearrangements, representing all different types and sizes of rearrangement that were predicted to fuse genes either directly or by runthrough – 103 in total. It was likely that some of the predicted rearrangements were germ line variants and, as there is no matched normal sample available for VP229 and VP267, I used a pool of genomic DNA from ten normal females to check for common germ line variants. Validation of break points by PCR is summarised in Table 5.7 and Figure 5.5.

	Attempted	Present in normal female pool	Present in VP229 or VP267 but not in normal female pool	Could not validate
DEL	25	9	15	1
DIF	42	8	27	7
INV	22	7	13	2
INS	14	6	7	1
Total	103	30	62	11

Table 5.7. Summary of PCR validation of structural variants in VP229 and VP267

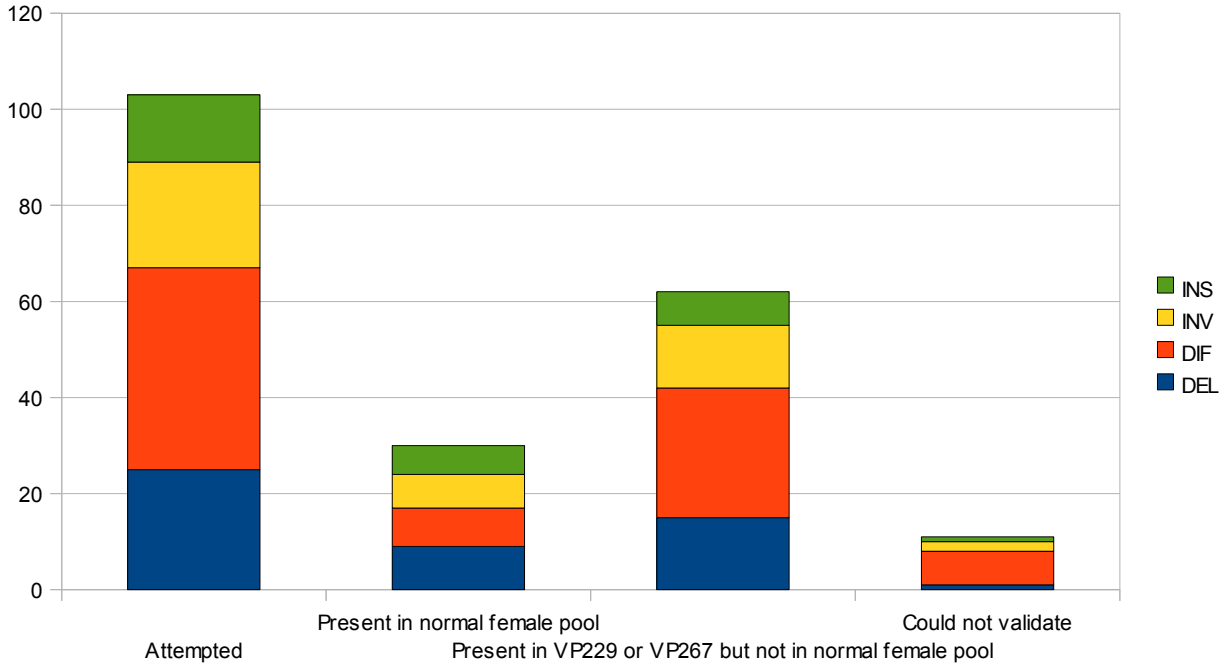


Figure 5.5. Summary of PCR validation of structural variants in VP229 and VP267

Of the 103 attempted validations, 92 gave a PCR product and 30 of these 92 were found in the pool of normal female DNA. This subset of “normal” variants were probably a combination of germline copy number variants and mismappings of the BWA-aligned sequencing reads. There appeared to be no bias towards any specific type or size of intrachromsomal rearrangement. These variants were not investigated any further in this thesis, but the reasons for their mismapping may be useful information that may help improve future alignment strategies. Eleven structural variants could not be validated by PCR. A small subset of these are likely to be due to PCR failure and the remainder, again, due to mismappings or sequencing errors. Interestingly, 9/11 failed PCRs were only predicated in a single library. This suggests spurious structural variants are generated in library preparation but at quite a low rate. This also suggests preparing two independent libraries from the same sample may be a way to cut out spurious structural variation.

Sixty-two predicted structural variants (60%) PCR-validated in VP229 or VP267 and were not present in the normal female pool. This figure means that the majority of predicted structural variants were real but also stresses the importance of validation in such

experiments. When a structural variant was predicted to be in both samples they were found in both samples but when they were predicted to be in a single sample, they were often in both (Table 5.8). This points to the importance of validation of apparently private mutations in both samples, especially when the coverage is low.

	Predicted in VP 229 and VP267	Present in VP229 and VP267	Present in VP229 only	Present in VP267 only
Total	37	35	0	0
DEL	7	7	0	0
DIF	15	14	0	0
INS	4	4	0	0
INV	11	10	0	0

	Predicted in VP229 only	Present in VP229 and VP267	Present in VP229 only	Present in VP267 only
Total	14	4	5	0
DEL	6	2	4	0
DIF	5	1	1	0
INS	1	0	0	0
INV	2	1	0	0

	Predicted in VP267 only	Present in VP229 and VP267	Present in VP229 only	Present in VP267 only
Total	22	12	0	6
DEL	3	1	0	1
DIF	14	8	0	3
INS	3	1	0	2
INV	2	2	0	0

Table 5.8. PCR validation of structural variants in depth. PCR validated junctions are only those that were absent from the normal female pool.

Approximately 30% of structural variants were found in pooled normal DNA from ten donors. There appeared to be no bias towards any specific type or size of intrachromsomal rearrangement. Some of these rearrangements were probably germ line structural variation but it is also possible that some resulted from misalignment of sequencing reads due to SNPs or small indels. The relative proportions of germline variants and spurious alignments is not known.

A second method of structural variant was described previously based around copy number changes (Stephens et al., 2009). If a structural variant is found at an unbalanced copy number change point by array CGH, it is thought to be more likely to be a true structural variant. Copy number plots had been generated from the millions of normally aligning read pairs and was shown to comparable to SNP6 array CGH in resolution (Batty, 2010). The loess-corrected copy number data was segmented using the circular binary segmentation R package *DNA copy* (Venkatraman and Olshen, 2007). This algorithm outputs segments of genome with the same predicted copy number similar to PICNIC segmentation discussed in previous chapters. I compared PCR-validated structural variants from above with copy number steps (Table 5.9).

	Present in VP229 or VP267 but not in normal female pool	**Validated by PCR and within 20kb of a copy number step**	**Validated by PCR but not within 20kb of a copy number step**
DEL	15	7	8
DIF	27	13	14
INV	13	9	4
INS	7	4	3
Total	62	33	29

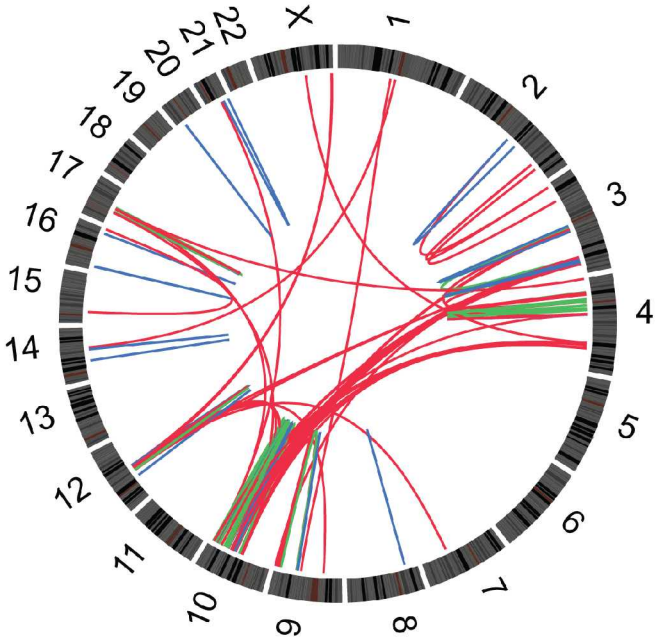
Table 5.9. PCR-validated structural variants at copy number change points.

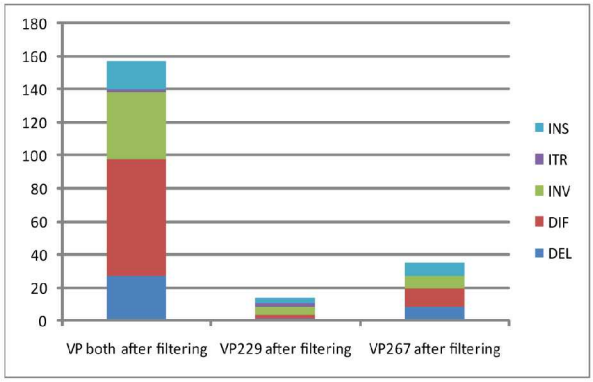
About half of the validated structural variants were within 20kb of a copy number change point. This implies that proximity to a copy number step may be a reasonable method to validate a subset of structural variants. However, around half of the structural variants were not associated with a copy number step and may be copy number neutral, and sub array CGH resolution. In addition, many chromosome breaks are known to be locally complex containing small genomic “shards” (Bignell et al., 2007). Their small size would also make them difficult to detect by copy number change.

5.3. Comparing the genomes of VP229 and VP267

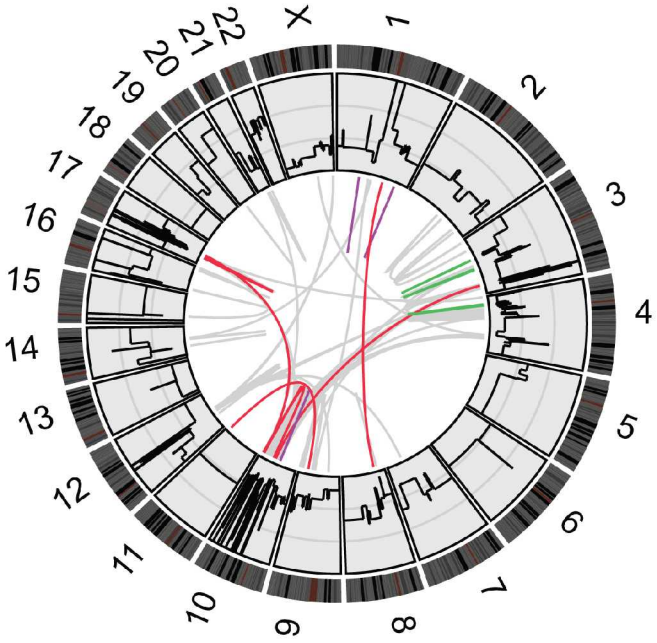
Although the data set is incomplete, it is still possible to make some observations about the genome structures of VP229 and VP267 and the implied common ancestor. Figure 5.6 shows a circular representation of these genomes based around copy number

-validated structural variant junctions.

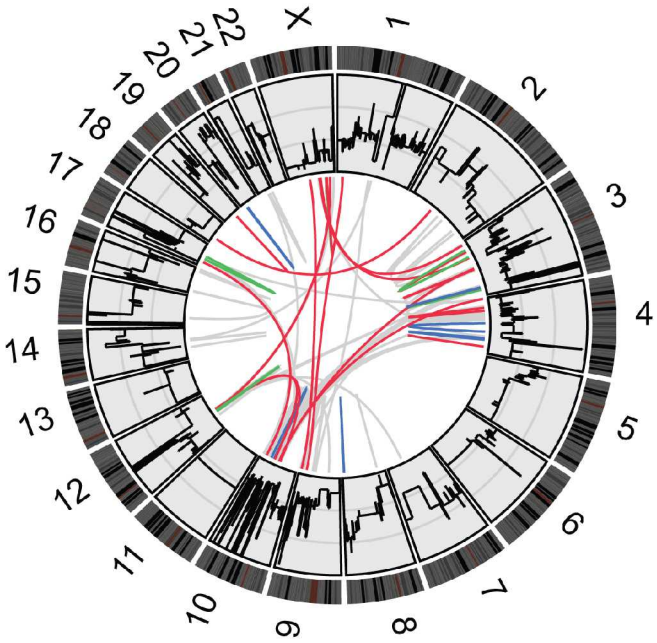




Structural Variants in VP229 and VP267



VP229 only



VP267 only

Figure 5.6. Validated structural variants in VP229 and VP267.

Figure 5.6. Copy number validated structural variants in VP229 and VP267. Upper Circos plot: Structural Variants found in both VP229 and VP267, therefore probably present in the common ancestor. Light blue links are insertions, purple are inverted tandem repeats, green are insertions, red are interchromosomal translocations and dark blue are deletions. Lower Circos plots are structural variants found only in one cell line. Grey histograms are copy number plots generated from loess-corrected binned sequencing reads as described above. Grey links are structural variants found in the common ancestor. Coloured links are structural variants found only in VP229 and VP267. Histogram represents the relative proportions of structural variants in each sample.

5.3.1. The VP-ancestoral genome was highly rearranged

Previous studies have invariably shown that primary tumours contain fewer structural variants than cell lines (Stephens et al., 2009; Varela et al., 2010; Campbell et al., 2010b) and it is tempting to conclude that cell lines accumulate large numbers of structural variants in culture. While some chromosome aberrations probably occur *in vitro* it is likely that the majority are true *in vivo* events. This is clearly illustrated by the genomic complexity of the common ancestor of VP229 and VP267 as it contained in excess of 150 copy number-validated structural variants.

5.3.1. The VP229 and VP267 genomes diverged away from the common ancestor

A smaller proportion of structural variants were found in VP267 but not in VP229 and vice versa (Figure 5.6). This implies that the two cell lines had a common ancestor at some point in time and that they evolved separately to the observed configuration. This evolutionary split likely happened quite late in the evolution of the tumour as the proportion of private VP229 or VP267 mutations is much lower than the shared category.

5.4. Fusion Genes in VP229 and VP267

5.4.1. Bioinformatic Prediction of genes broken and fused

I used a custom Perl script to predict genes broken or fused in VP229 and VP267. This script found genes at or near to chromosome break points and predicted if they were potentially fused to another gene either directly or by runthrough. The script also identified genes where an internal rearrangement may have resulted in alternative isoforms being expressed (Batty, 2010). Potentially fused genes in VP229 and VP267are listed in Table 5.10. A large number of run-through fusions were also predicted which I did not investigate.

5.4.2. Expressed fusion genes in VP229 and VP267

All potential fusions were investigated by RT-PCR. I was assisted by PhD student, S.Flach, and her contribution is noted in table 5.7. Three fusion genes were expressed in both VP229 and VP267: *MDS1-KCNMA1*, *FAM125B-SPTLC1* and *PDLIM1-ZBBX.* On further fusion transcript was found in VP267 only: *TRAPPC9-KCNK9*.

Type	Supp.	Node 1 chr	Node 1 start	Node 1 end	dir	Node 2 chr	Node 2 start	Node 2 end	dir	Possible Fusion	VP 229 gDNA PCR	VP 229 gDNA PCR	RT-PCR by
Del	2	7	1173342	1173753	+	7	1192615	1192690	-1	*ZFAND2A-C7orf50*	y	y	SN
Del	6	8	140704651	140704941	+	8	141348436	141348751	-1	***TRAPPC9-KCNK9***	n	y	SN
Del	6	9	130313996	130314354	+	9	132757339	132757510	-1	*FNBP1-FAM129B*	y	y	SN
Del	8	11	5784201	5784559	+	11	5809301	5809685	-1	*OR52N1-TRIM5*	y	y	SF
Del	3	20	17506460	17506612	+	20	17598409	17598578	-1	*RRBP1-BFSP1*	y	y	SN
Del	4	X	142600166	142600247	+	X	142799082	142799186	-1	*SPANXN2-SPANXN3*	y	y	SN
Ins	7	10	76898989	76899271	-1	10	84730469	84730979	+	*NRG3-SAMD8*	y	y	SF
Ins	2	19	11614744	11614845	-1	19	11638741	11638818	+	*ZNF653-ECSIT*	n	y	SF
Ins	2	22	39425844	39425899	-1	22	39445528	39445564	+	*APOBEC3G-APOBEC3D*	y	y	SN
Ins	22	10	78198382	78198678	-1	10	83671218	83671877	+	*NRG3-C10orf11*	y	y	SF
Diff	2	1	49783251	49783335	+	14	31139376	31139481	+	*SCFD1-AGBL4*	n	y	SF
Diff	9	3	167030872	167031047	+	10	97034481	97034738	-1	***PDLIM1-ZBBX***	y	y	SF
Diff	85	3	169496733	169497135	+	10	84276976	84277214	-1	*MYNN-NRG3*	y	y	SN
Diff	9	3	171158535	171158864	+	10	97032906	97033283	-1	*PDLIM1-TNIK*	y	y	SF
Diff	2	7	102818133	102818319	+	12	63957260	63957303	-	*DPY19L2-DPY19L2P2*	y	y	SN
Diff	7	10	76297188	76297545	+	3	178114514	178114843	-	*ADK-KCNMB2*	y	y	SF
Diff	15	10	77069722	77070093	+	17	26873584	26873946	-	*AC010997.5-UNC119*	y	y	SN
Diff	3	10	77393012	77393230	-	17	27124203	27124413	-	*C17orf63-C10orf11*	y	y	SF
Diff	3	10	77414962	77415055	+	17	27117214	27117289	+	*C10orf11-C17orf63*	y	y	SF
Diff	43	10	78679650	78680299	+	3	169181685	169182111	-	***MDS1-KCNMA1***	y	y	SN
Diff	102	10	79656167	79656503	-	3	178107424	178107859	-	*DLG5-KCNMB2*	y	y	SF
Diff	10	10	84303032	84303414	1	12	66816857	66817212	+	*NRG3-GRIP1*	y	y	SF
Diff	7	10	125636500	125636725	+	21	31094557	31094776	-	*GRIK1-CPXM2*	y	y	SF
Diff	3	13	21750547	21750661	+	11	108585829	108586244	-	*MRP63-DDX10*	y	y	SN
Diff	23	17	35849590	35849900	-	4	782909	-783243	-	*CPLX1-DUSP14*	y	y	SN
Diff	2	19	18953632	18953670	-	8	68218123	68218163	-	*ARFGEF1-UPF1*	y	y	SF

Diff	2	20	51857762	51857845	+	22	29065501	29065807	+	*TSHZ2-TTC28*	n	y	SF
Inv	4	3	108136284	108136419	+	3	110987235	110987406	+	*PVRL3-MYH15*	y	y	SN
Inv	13	9	94647680	94648017	+	9	128534661	128534996	+	*PBX3-ROR2*	y	y	SN
Inv	11	9	94648426	94648733	-	9	127055261	127055508	**-**	*ROR2-NEK6*	y	y	SN
Inv	5	9	94861458	94861760	+	9	129243323	129243573	**+**	***FAM125B-SPTLC1***	y	y	SF
Inv	16	9	127073730	127074038	-	9	130287513	130287774	-	*FAM129B-NEK6*	y	y	SN
Inv	14	10	97789793	97790092	+	10	108715824	108716128	+	*AL356155.1-SORCS1*	y	y	SN
Inv	11	10	98803429	98803655	+	10	99294104	99294473	+	*UBTD1-SLIT1*	y	y	SN
Inv	8	10	124809773	124809955	+	10	127790628	127790985	+	*ACADSB-ADAM12*	y	y	SF
Inv	13	17	35326354	35326611	+	17	36222373	36222607	+	*AATF-AC113211.2*	y	y	SN

Table 5.10. Predicted Fusion Genes in VP229 and VP267. Type: Del = Deletion, Ins = Insertion, Diff = Translocation, Inv = Inversion. Supp.=Number of read pairs that span the genomic junction; a minimum of two were required. Amplified regions typically had more reads crossing junctions. Dir = node direction: (+), the read cluster was in the positive orientation and (–) for the negative strand. All genomic DNA junctions were confirmed by PCR of VP229 and VP267 genomic DNA. None of the above junctions were found in the normal DNA pool. The fusion transcripts whose expression was demonstrable by RT-PCR are in bold.

5.4.2.1. *PDLIM1-ZBBX*

An inter-chromosome rearrangement juxtaposed *PDLIM1* and *ZBBX*. The genomic junction between chromosomes 3 and 10 were found in both VP229 and VP267 (Figure 5.7). *PDLIM* exon 1 is fused with *ZBBX* exon 16 causing an out of frame fusion transcript (Figure 5.9). This is unlikely to produce a homozygous loss of function for either gene as non-rearranged copied of *PDLIM1* and *ZBBX* probably remain on chromosomes 3 and 10.

PDZ and LIM domain protein 1 (PDLIM1), also referred to as carboxyl terminal LIM domain protein 1 (*CLIM1*) is a transcriptional co-regulator that can bind to the LIM domains of nuclear LIM proteins including LIM-homeodomain (LIM-HD) transcription factors. *PDLIM1* is an *ERα* cofactor and it is likely that it plays a role in the regulation of *ERα* target genes in breast cancer cells and primary tumours (Johnsen et al., 2009). Nothing is known of zinc finger, B-box domain containing (*ZBBX*).

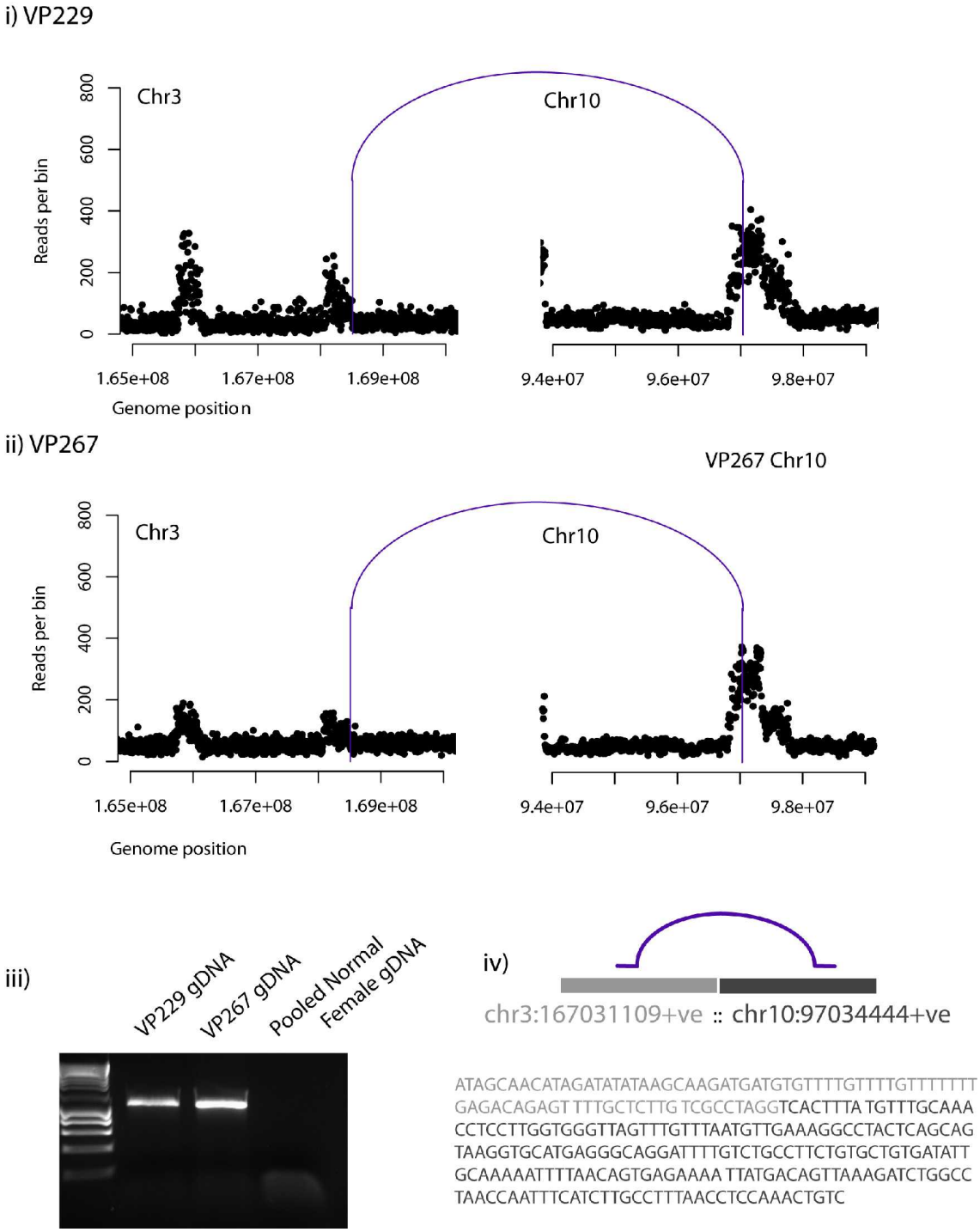


Figure 5.7. *PDIM1-ZBBX* genomic junction. i) VP229 loci of chromosomes 3 and 10. Scatter plots are loess-corrected copy number data, equivalent to array CGH. The purple line is an inter chromosome junction. ii) Equivalent plot from VP267. iii) PCR validation of the genomic junction. iv) schematic representation of the genomic junction and sequence across it.

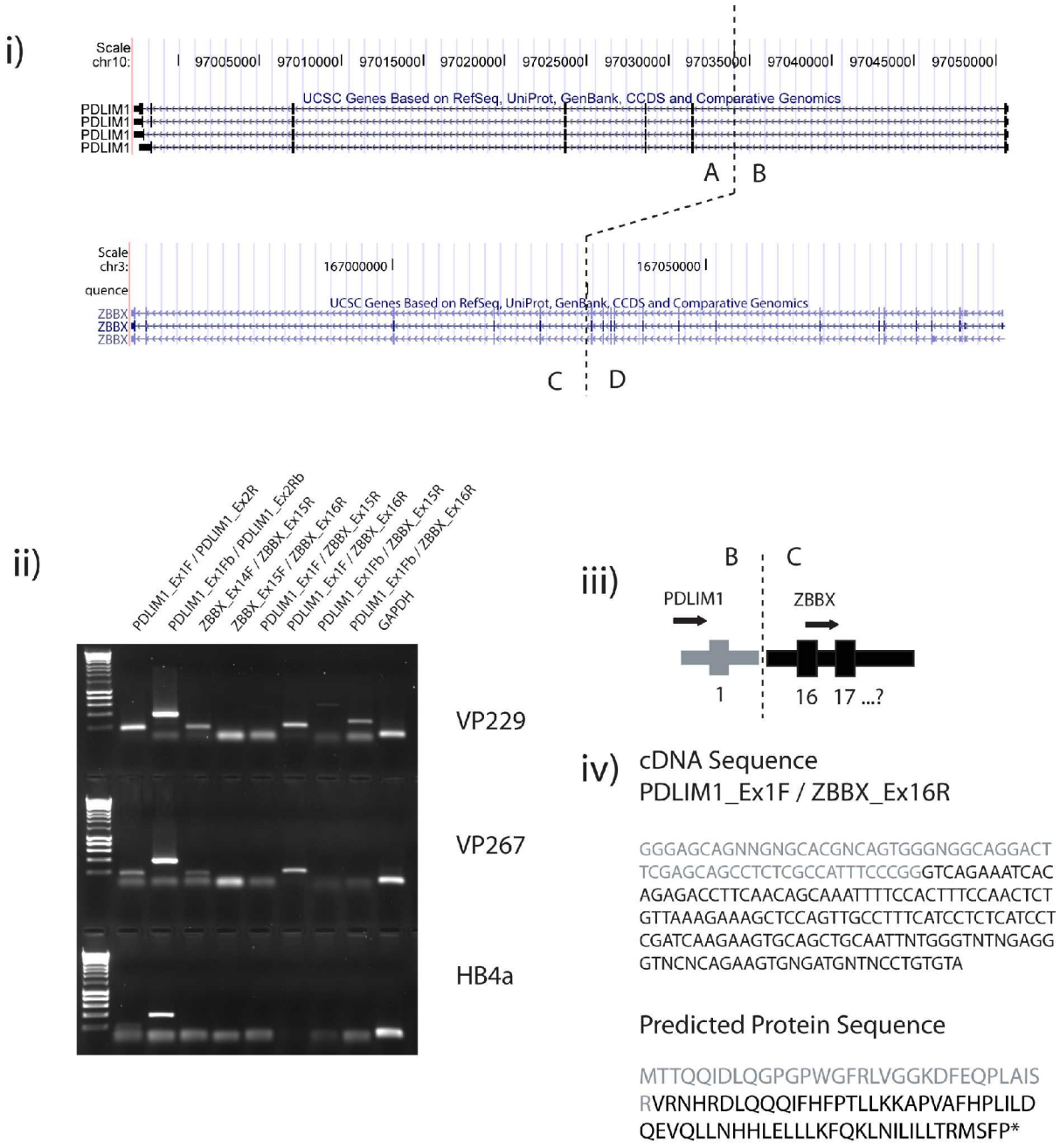


Figure 5.8. RT-PCR of the *PDLIM1-ZBBX* fusion transcript. i) The *PDLIM1* and *ZBBX* genomic loci; dotted lines indicate chromosome break points. ii) RT-PCR of the fusion transcript. iii) Schematic of the fusion transcript exons are named according to *PDLIM1*-001 (ENST00000329399) and *ZBBX*-001 (ENST00000392766).iv) cDNA sequence across the fusion junction is predicted to cause a frame-shift in the ZBBX portion of the transcript.

5.4.2.2. *FAM125B-SPTLC1*

An intra-chromosome rearrangement, nominally an inversion, caused fusion of *FAM125B* and *SPTLC1* (Figure 5.9). RT-PCR showed that *FAM125B* exon 6 was fused with *SPTLC1* exon 4. (Figure 5.10) Only one copy of each locus remains according to PICNIC segmentation of SNP6 VP229 array CGH, so it is possible one or both of these genes are lost by a two-hit mechanism.

Family with sequence similarity 125, member B (*FAM125B)* is a component of the ESCRT-I complex (endosomal sorting complex required for transport I), a regulator of vesicular trafficking process (Tsunematsu et al., 2010). Serine palmitoyltransferase, long chain base subunit 1 (*SPTLC1*) is part of the serine palmitoyltransferase complex, the initial enzyme in sphingolipid biosynthesis (Weiss and Stoffel, 1997). Sphingolipids are a component of cell membranes and are involved in a large number of cellular processes including mitosis, apoptosis, migration, stemness of cancer stem cells and cellular resistance to therapies so may have relevance to cancer (Patwardhan and Liu, 2010).

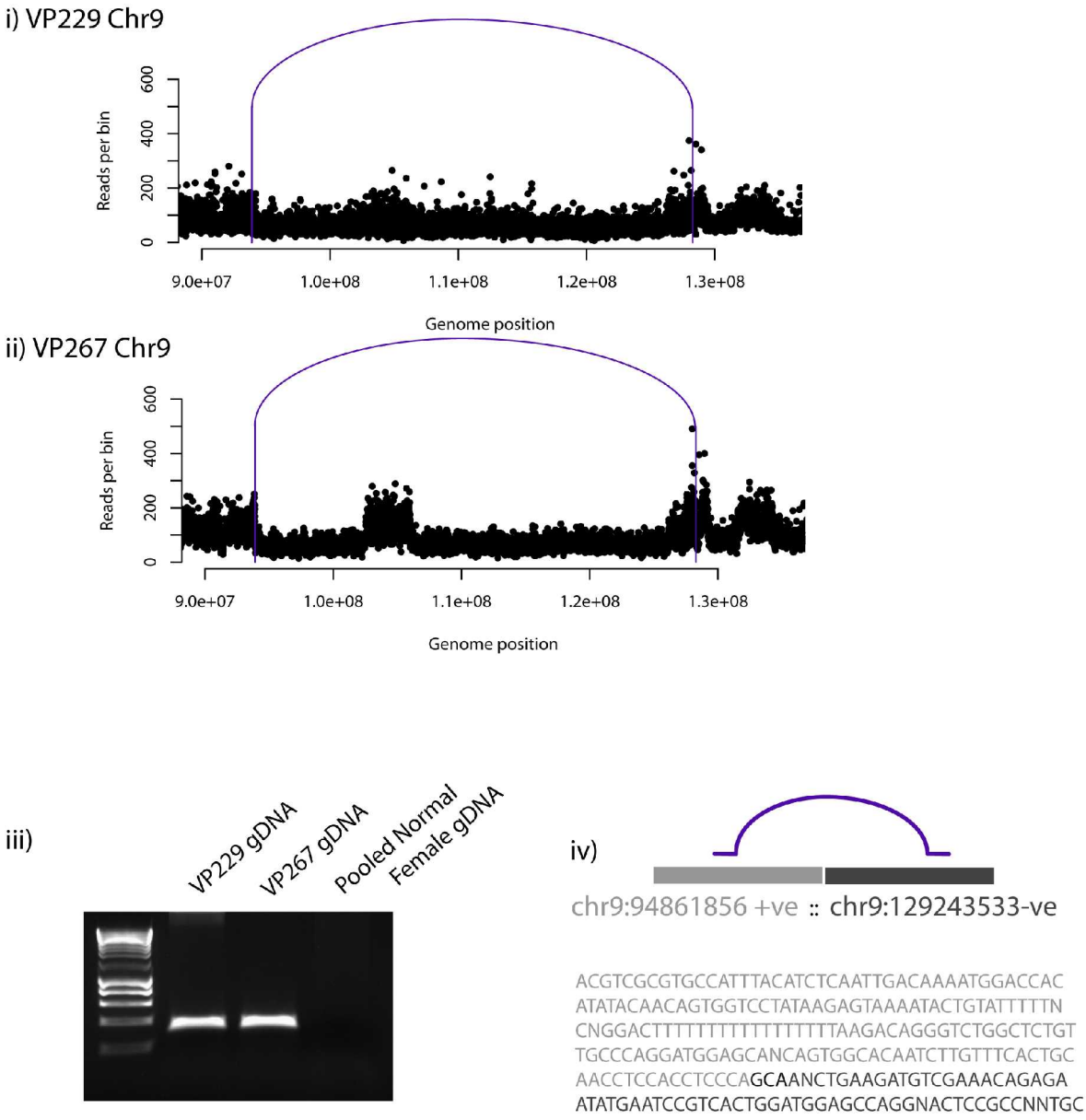


Figure 5.9. *FAM125B-SPTLC1* genomic junction. i) VP229 loci of chromosomes 9. Scatter plots are loess-corrected copy number data, equivalent to array CGH. The purple line is an intra-chromosome junction, classed as an inversion. ii) Equivalent plot from VP267. iii) PCR validation of the genomic junction. iv) schematic representation of the genomic junction and sequence across it.

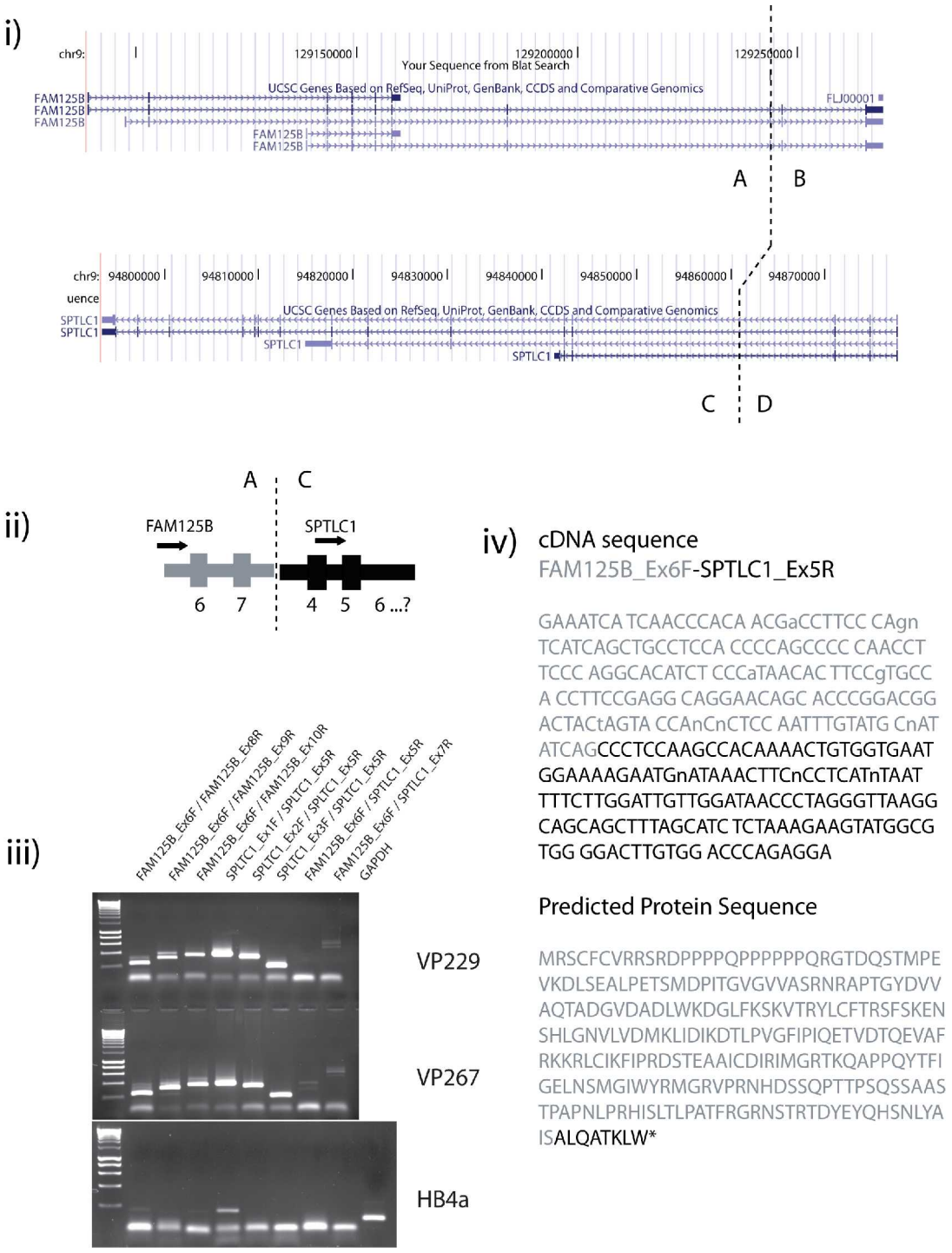


Figure 5.10. RT-PCR of the *FAM125B-SPTLC1* fusion transcript. i) The *FAM125B* and *SPTLC1* genomic loci; dotted lines indicate chromosome break points. ii) RT-PCR of the fusion transcript. iii) Schematic of the fusion transcript exons are named according to *FAM125B-001* (ENST00000361171) and *SPTLC1-001* (ENST00000262554). iv) cDNA sequence across the fusion junction is predicted to cause a frame-shift in the *SPTLC1* portion of the transcript

5.4.2.3. *MDS1-KCNMA1*

An inter-chromosome junction fused *MDS1* to *KCNMA1* (Figure 5.11). The fusion junction is likely to be within the complex co-amplification of chromosomes 3 and 10. The fusion transcript is predicted to be in frame (Figure 5.12).

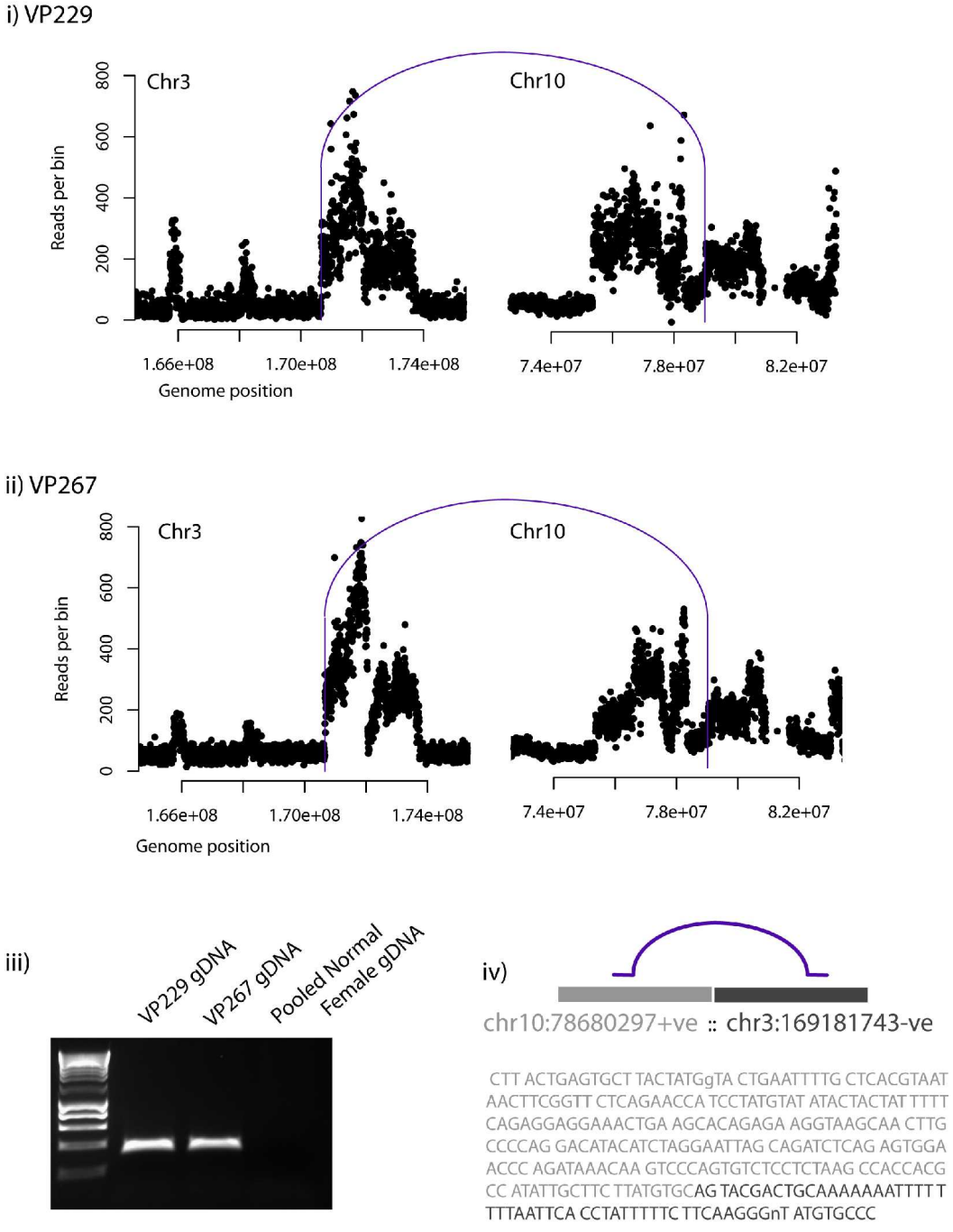


Figure 5.11. *MDS1-KCNMA1* genomic junction. i) VP229 loci of chromosomes 3 and

10. Scatter plots are loess-corrected copy number data, equivalent to array CGH. The purple line is an inter-chromosome junction. ii) Equivalent plot from VP267. iii) PCR validation of the genomic junction. iv) schematic representation of the genomic junction and sequence across it.

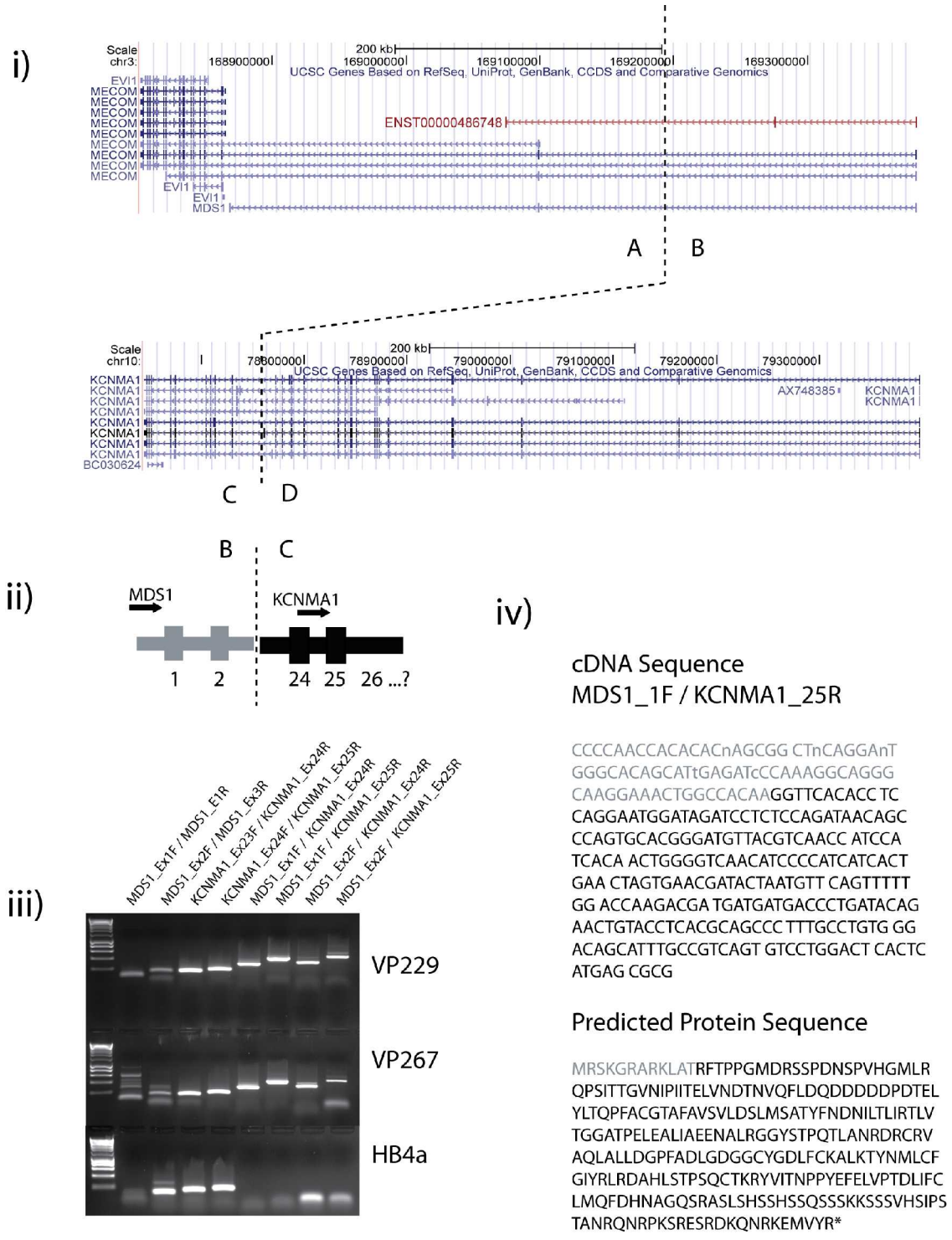


Figure 5.12. RT-PCR of the *MDS1-KCNMA1* fusion transcript. i) The *MDS1* and

KCNMA1 genomic loci; dotted lines indicate chromosome break points. ii) RT-PCR of the fusion transcript. iii) Schematic of the fusion transcript exons are named according to *MECOM*-005 (ENST00000486748) and *KCNMA1*-005 (ENST00000372408) .iv) cDNA sequence across the fusion junction is predicted to cause an in-frame gene fusion

The *MDS1* (Myelodysplasia syndrome-associated protein 1) locus is known to take part in translocations. The human *MDS1* gene was first identified as a component of the *AML1(RUNX1)-MDS1-EVI1* fusion transcript formed through a t(3;21) translocation in some spontaneous myeloid leukaemias (Fears et al., 1996). *MDS1* is 3 kb telomeric to the first exon of *EVI1*, another known target of translocations in leukaemia. Transcription of the *MDS1-EVI1 (MECOM)* locus is complex. Transcripts can contain only *EVI1* exons, only *MDS1* exons, or fusion transcript containing the first two exons of MDS1 (as in the MDS1-KCNMA1 fusion) and exon 2 onwards of *EVI1* (Métais and Dunbar, 2008). Although the fusion transcript is expressed in both normal and leukaemic tissues (Fears et al., 1996) there is some evidence to suggest that *MDS1* drives increased expression of *EVI1* in AML and that these patients have a poorer prognosis (Barjesteh van Waalwijk van Doorn-Khosrovani et al., 2003).

KCNMA1 is the pore forming α-subunit of the large-conductance calcium- and voltage-activated potassium channel, BK_{Ca}. (also called hSlo form *Drosophila* "*slowpoke*" homologue and Maxi-K). BK_{Ca} channels consist of a pore-forming alpha subunit and a regulatory beta subunit (*KCNMB1-4*) which confer the channel with a higher calcium sensitivity. The intracellular C-terminal region of *KCNMA1* consists of a pair of RCK domains each of which contains two primary binding sites for Ca^{2+}, termed 'calcium bowls.' Interestingly, functionally important mutations cluster near the calcium bowls suggesting that this region plays a role in modulating the channel's sensitivity to calcium (Yuan et al., 2010).

Sequencing of the full length *MDS1-KCNMA1* fusion transcript confirmed that *MDS1* exon 2 was fused to *KCNMA1* exon 24. Almost nothing is known about the *MDS1* protein except that the additional amino acids in the *MDS1-EVI1* fusion protein encodes a so-called "PR" domain (*PRD1-BF1/BLIMPI-RIZ* homology), which defines a sub-class of zinc finger genes

(Métais and Dunbar, 2008). For the *KCNMA1* portion of the transcript, most of the N-terminal domains, including the pore domain, is excluded from the fusion transcript and only the C-terminal intra-cellular domains are retained. This region encodes two intracellular RCK domains and a "calcium bowl" which can bind calcium ions. These domains are thought to play a role in the calcium-sensitive opening of the channel (Pico, 2003; Yuan et al., 2010).

5.4.2.4. *TRAPPC9-KCNK9*

A small intra-chromosomal deletion fused *TRAPPC9* and *KCNK9* in VP267 but not in VP229 (Figure 5.13). As a result, we cannot be certain if this was an *in vivo* event. Exon 9 of *TRAPPC9* is fused with exon 2 of *KCNK9*. The fusion is predicted to be in frame (Figure 5.14).

TRAPPC9 (trafficking protein particle complex 9, also known as *NIBP*). Not much is known about this protein except that it is implicated in *NF-kappaB* activation and possibly in intracellular protein trafficking (Mochida et al., 2009). Potassium channel subfamily K member 9 is a protein that in humans is encoded by the *KCNK9* gene and is also referred to as *TASK3* ((*TWIK)-related acid-sensitive-3*). It is one of the members of the superfamily of potassium channel proteins that contain two pore-forming P domains. The predicted fusion protein comprises of several low-complexity regions from the *TRAPPC9* portion of the fusion and the transmembrane potassium pumping domains of *KCNK9*.

Two pore potassium channels are regulated by mechanisms including oxygen tension, pH, mechanical stretch, and G-protein signalling. Some studies indicated that over-expression of *KCNK9* may contribute to the development of cancers such as colorectal (Kim et al., 2004). Overexpression of *KCNK9* in cell lines can confer resistance to hypoxia and may promote tumour formation in nude mice (Mu et al., 2003).

The 8q24 region in which *KCNK9* resides is amplified in a number of cancers, probably due to *MYC* so its amplification in some cancers may well be a passenger event secondary to *MYC* amplification. Bearing this in mind, Mu et al. (2003) showed that "*KCNK9* is the sole over-expressed gene within the amplification epicentre." The *KCNK9*

gene is amplified and over-expressed in approximately 10% breast tumours. It is possible that over expression of the pumping domain occurs due to the gene fusion and this is discussed further in the next chapter.

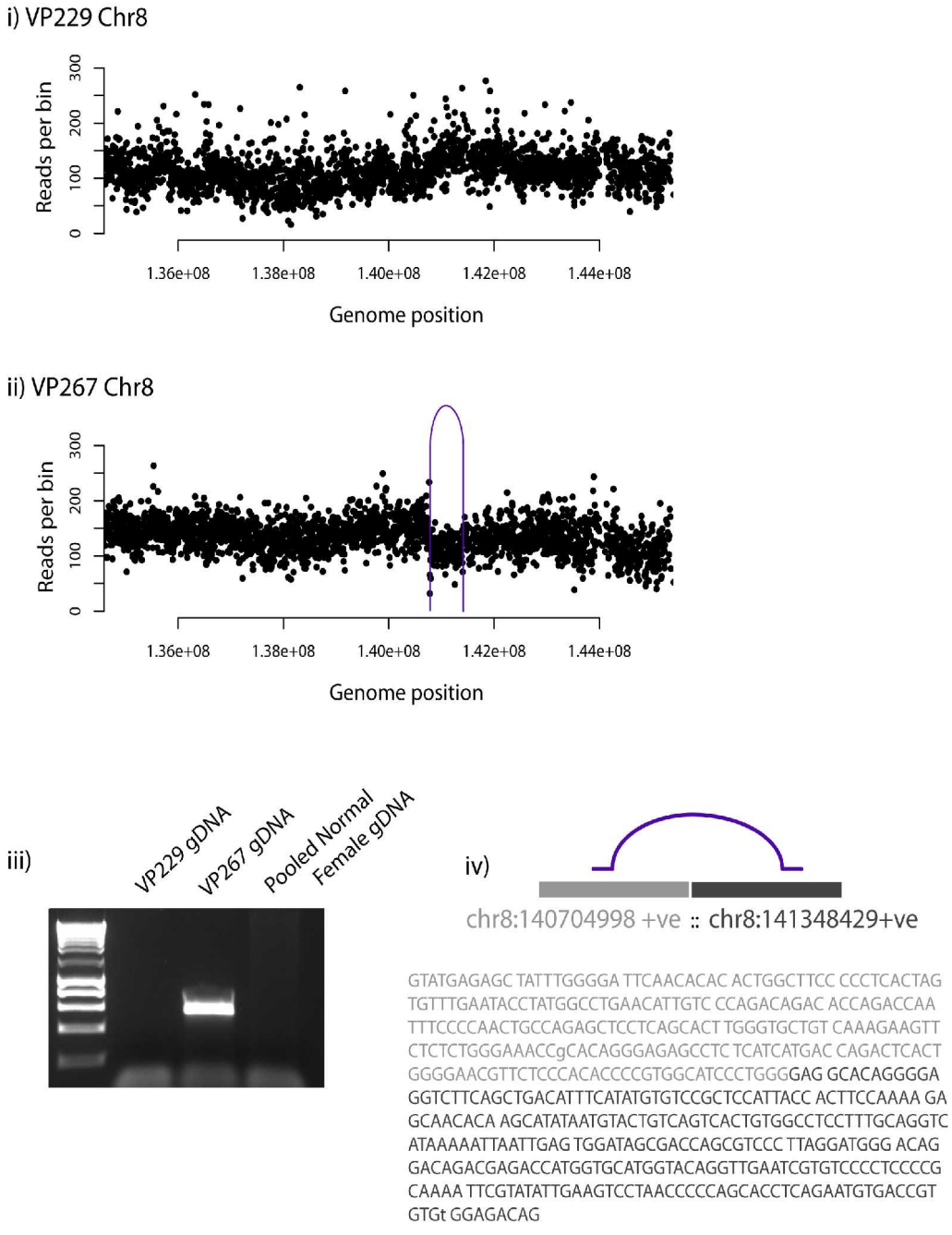


Figure 5.13. *TRAPPC9-KCNK9* genomic junction. i) VP229 chromosome 8. Scatter plots are loess-corrected copy number data, equivalent to array CGH. ii) Equivalent plot from VP267. The purple line is an intra-chromosome junction, called as a deletion. The plot shows a clear copy number step at one of the junctions that is absent from VP229.

iii) PCR showed the genomic junction was found in VP267 only. iv) schematic representation of the genomic junction and sequence across it.

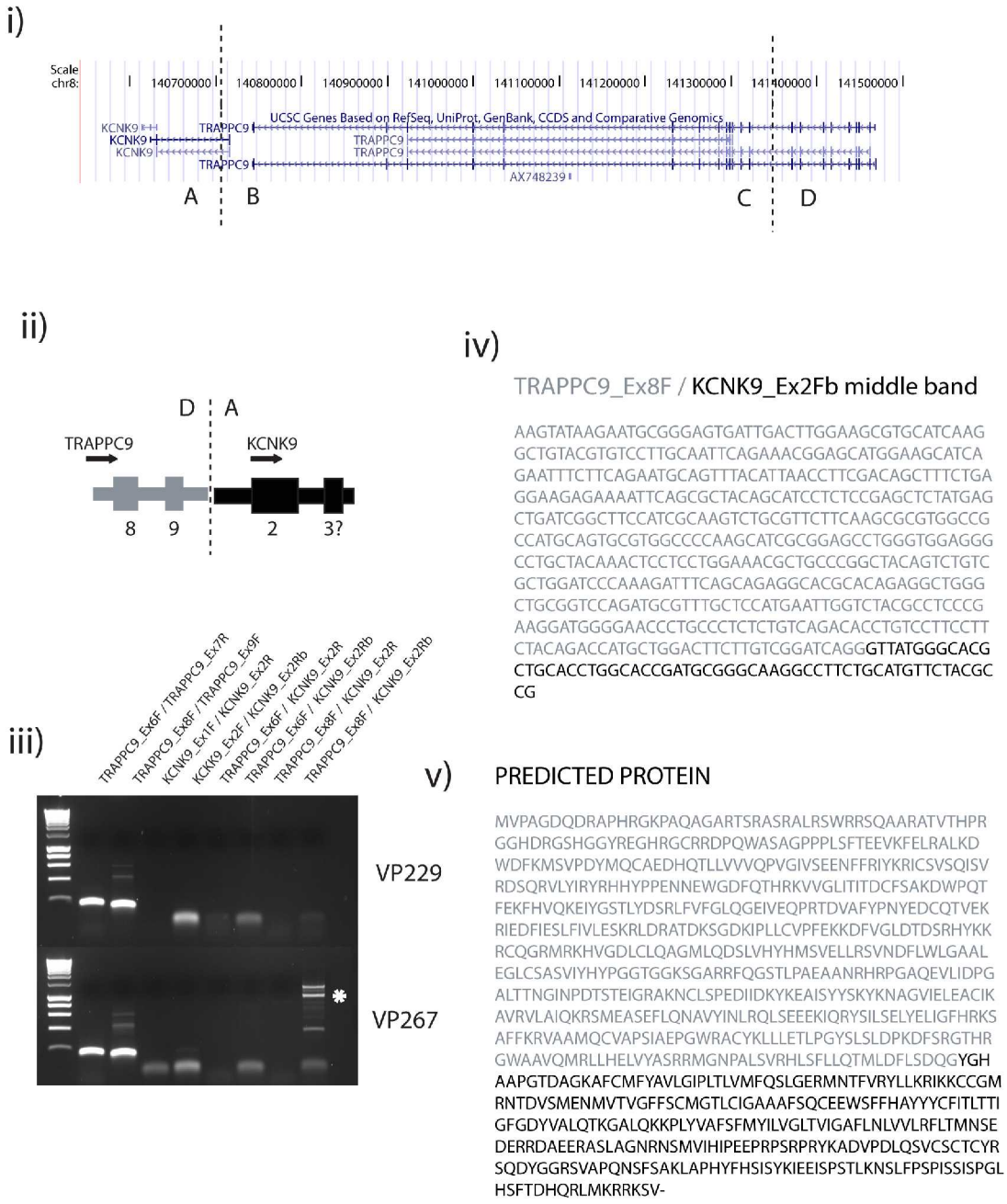


Figure 5.14. RT-PCR of the *TRAPPC9-KCNK9* fusion transcript. i) The *TRAPPC9* and *KCNK9* genomic loci; dotted lines indicate chromosome break points. ii) RT-PCR of the fusion transcript. iii) Schematic of the fusion transcript exons are named according to *TRAPPC9*-001 (ENST00000389328) and *KCNK9*-201 (ENST00000303015). iv) cDNA sequence across the fusion junction is predicted to cause an in-frame gene fusion. This is sequence from the starred band.

5.5. Discussion Part I

5.5.1. How complete are contemporary massively parallel paired end sequencing studies?

Mathematical models of paired end sequencing studies predict that a high proportion of rearrangements are sampled in each experiment. But currently, there is has been little experimental investigation to support theoretical estimates. In 2009, Stephens et al. carried out a survey of 24 breast cancer genomes by massively parallel paired end sequencing. I compared the structural variant data from this study with the chromosome aberrations predicted from array CGH data.

Copy number steps from array CGH (Bignell et al., 2010) represent most types of genomic aberration occurring such as tandem duplication, deletion, amplification and unbalanced translocation. Although the resolution is limited and balanced rearrangements cannot be detected, these data provide a reasonable method to assess the accuracy of massively parallel sequencing experiments as unbalanced rearrangements, evident as segmented copy number steps, must be joined to something else in the genome – most likely another unbalanced copy number step.

I took all PICNIC-segmented copy number steps from the HCC breast cancer cell lines (Gazdar et al., 1998) and VP229 and asked whether there was any associated structural variant (Stephens et al., 2009) within 20kb 5' or 3' of the breakpoint region (Table 5.11).

Cell Line	Estimated Genome Size (Gb)	Physical coverage (haploid genomes)	Corrected Coverage	Copy Number Steps	Number of Copy number steps with an associated SV	Percentage Sampled
HCC1143	10.04	7.6	2.28	353	71	20.11
HCC1187	7.81	11	4.23	157	61	38.85
HCC1395	7.97	6.4	2.41	345	65	18.84
HCC1599	8.79	4.9	1.67	357	20	5.6
HCC1954	12.95	6.7	1.55	448	91	20.31
HCC2157	7.97	4.4	1.66	256	46	17.97
HCC2218	11.71	5.9	1.51	114	30	26.32
HCC38	10.04	9.1	2.72	395	139	35.19
VP229	8.2	6.9	2.52	452	65	14.38

Table 5.11. Physical coverage versus sequence sampling: Data for HCC cell lines is from Stephens et al. (2009). PICNIC copy number steps are from Bignell et al. (2010). VP229 data SNP6 data was provided by Dr SL Cooke (Cambridge CRI). Genome size was estimated from PICNIC total copy number and a "corrected coverage" value calculated e.g. The HCC1187 genome is estimated to be 7.81 gigabases. Given 11 fold coverage of a nominal haploid genome, this translates to 4.23-fold coverage of HCC1187's near triploid genome.

There did seem to be a correlation between physical coverage and the proportion of rearrangements sampled (Figure 5.15). However, these figures are much lower than the Poisson distribution would predict (90 percent of rearrangements for HCC1187 versus 39% seen here) and the reasons why are not entirely clear. A likely factor, however, is the high proportion of repeat elements in the human genome. Currently, each sequence read must map uniquely to be considered in these experiments and those mapping to repeat elements are discarded. It is possible that a translocation break flanked by repeats would not be found by current sequencing methodologies and is discussed further in Chapter 7.

Stephens et al. (2009) did attempt to address this issue by producing a sequencing library with 3kb fragments for HCC1187. These 'mate pairs' were expected to traverse the majority of human repeat elements. The authors stated that "[a]lthough additional rearrangements were detected, a distinct class of repeat-mediated rearrangement was not found." (Stephens et al. 2009. p1008).

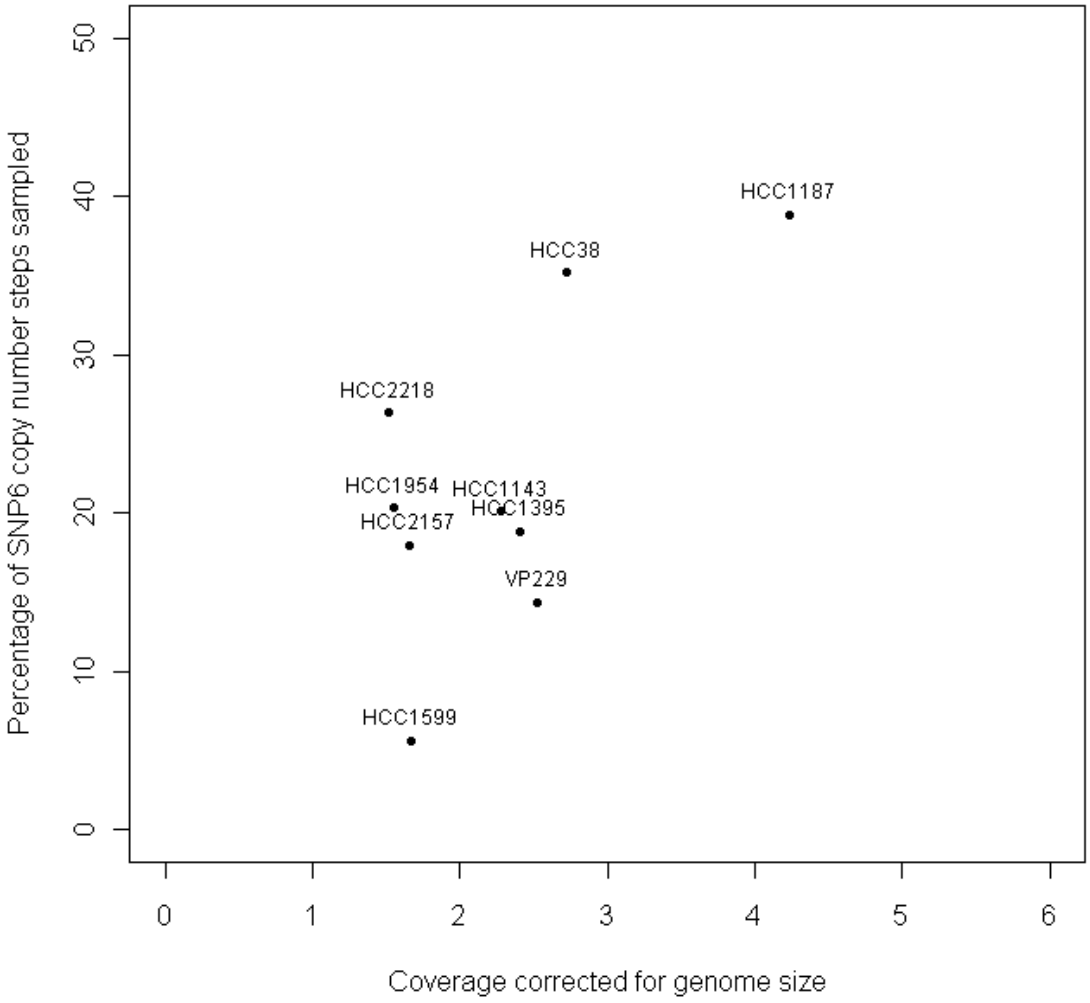


Figure 5.15 Physical coverage versus the proportion of unbalanced rearrangements detected by array CGH.

It is unlikely that increasing the sequence coverage will solve this problem. Pleasance et al. (2009) generated 39-fold sequence coverage for the NCI-H209 cell line. This translates to approximately 410-fold physical coverage of the haploid genome, but it appears that only 32 percent of the unbalanced breaks by CGH were sampled. At this stage, we can only regard genome analysis by paired end sequencing as an incomplete survey of genome rearrangements (Pleasance et al., 2010).

5.5.2. Complexity at Chromosome Breaks

In HCC1187, about half of the predicted fusion genes were expressed but in VP229 and VP267 this was not the case. A possible explanation for this inconsistency is that many chromosome breaks in VP229 and VP267 were within complex amplicons. The genomic junctions of amplicons often contain 'genomic shards' - pieces of DNA, often from unrelated genomic regions, that range from tens of bases to tens of kilobases in size (Bignell et al., 2007) The presence of one or more genomic shards at a junction makes its interpretation much more difficult. The t(11;16) translocation from HCC1187 contains a genomic shard and is shown below to illustrate this difficulty in assembling structural variant data from such regions (Figure 5.16).

In this example, sequencing across genomic break points showed a 1.4kb shard at the junction between the t(11;16) translocation (Chapter 3, chromosome S). In addition, the reciprocal product, (chromosome R) is slightly unbalanced with respect to chromosome 16. Array painting and FISH showed that assembly 1) was probably true and this lead to prediction and RT-PCR verification of the *CTCF-SCUBE2* fusion transcript.

In paired-end sequencing experiments the only source of information is structural variant genomic junctions. If we consider each structural variant in isolation, we cannot predict the *CTCF-SCUBE2* fusion gene (Table 5.12).

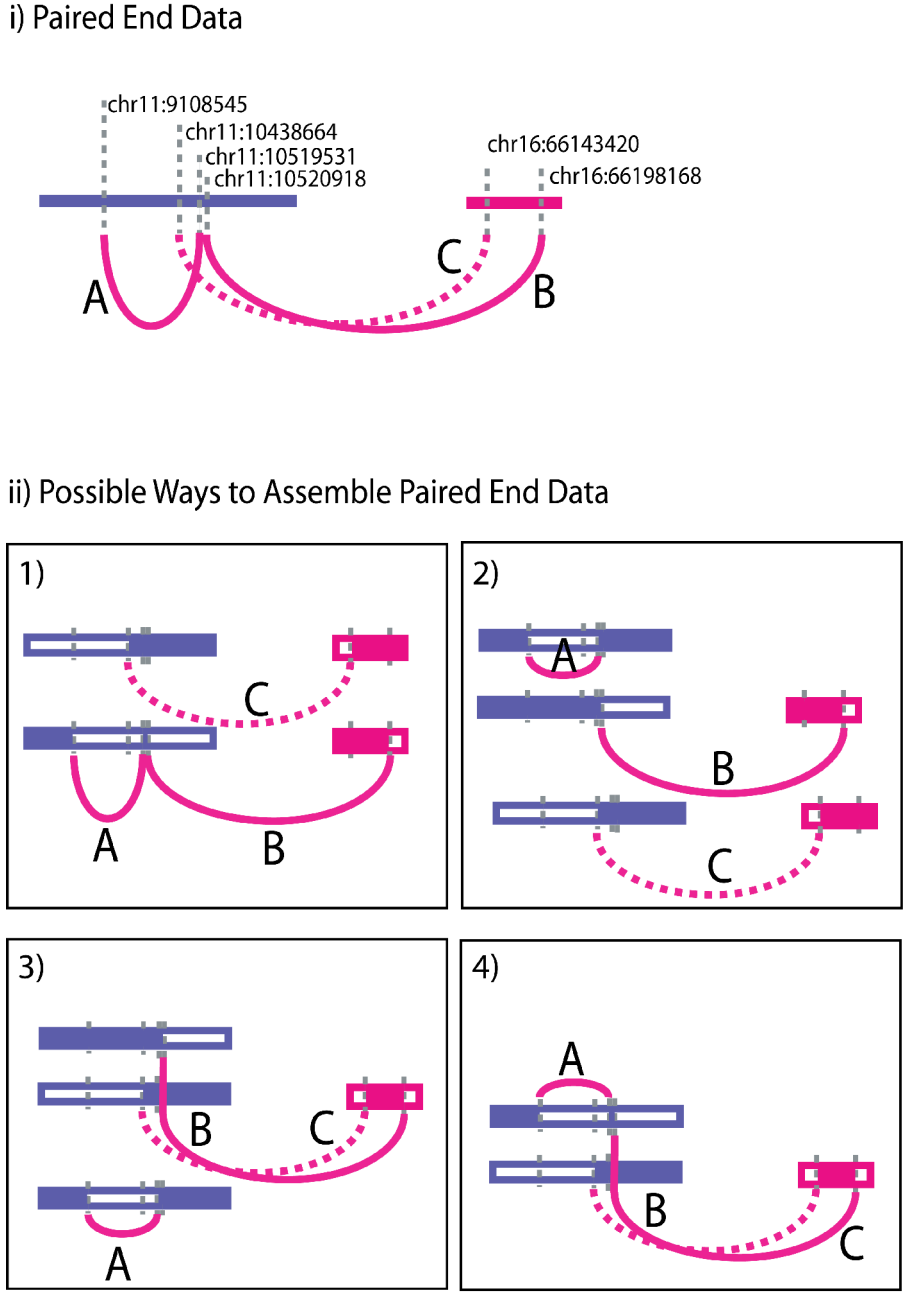


Figure 5.16 Possible Assemblies of the HCC1187 t(11;16) translocation from paired-end sequence data. i) Paired-end sequence data. Purple rectangle is chromosome 11, pink rectangle is chromosome 16. Curved lines are structural variants from Stephens et al. (2009) according to HG18 named A,B. The dotted line is a simulated read as this junction was not detected by Stephens et al. and is named junction C ii) There are four ways to assemble these reads. For example, solution 1) Junction C forms a der(11)t(11;16), Junctions A and B form a near reciprocal product, a der(11)t(16,11), with a small genomic shard at the 11;16 junction.

	Node1				Node2			
Junction	**Chr**	**Position**	**Dir**	**Gene**	**Chr**	**Position**	**Dir**	**Gene**
A	11	9108545	+	***SCUBE2***	11	10519531	+	None
B	11	10520918	+	None	16	66198168	+	***CTCF***
C	11	10438664	-	None	16	66143420	-	None

Table 5.12. Genomic Junctions for the HCC1187 t(11;16) translocation.

We have to assemble the reads into a local genome structure to predict the *CTCF-SCUBE2* fusion. But if the only source of information is structural variant genomic junctions, there are four equally plausible ways to assemble the genomic region and only assemblies 1) and 4) point to the existence of the *CTCF-SCUBE2*.

One could argue that the close proximity of junctions A and B indicate the presence of a shard and this would make assembles 1) and 4) equally plausible. It is likely that in this example the shard could be jumped by a large insert mate-pair sequencing strategy, but as shards of up to 30kb have been reported (Bignell et al., 2007), mate pair strategies will not be able to solve all examples.

5.5.3. How complete was the analysis of VP229 and VP267?

Given the above data on physical genome coverage and break point complexity, it is clear that this analysis is only preliminary and structural variant junctions that contained genomic shards were probably frequent in VP229 and VP267. The above example only contained three genomic junctions, but as multiple shards can be observed at a single junction, other examples are likely to be much more complex. It is also likely that some genomic junctions were not sampled in these experiments making the data incomplete. No attempt was made to assemble complex junctions in VP229 and VP267. It is therefore likely that many more fusion genes remain to be found in these genomes.

5.6. Discussion Part II

5.6.1. Did VP229 and VP267 really evolve from a common ancestor?

As cell line cross contamination is a problem in the field (Chatterjee, 2007) it is reasonable to ask if VP229 and VP267 really were derived from the same patient at different stages of disease (McCallum and Lowther, 1996) and were not, for example, cross contaminants. There are two lines of evidence to support the fact that VP267 was not derived directly from VP229 or vice versa.

Firstly, VP229, is sensitive to Tamoxifen and Fulvestrant but VP267 cells are resistant to both drugs (Ghayad et al., 2009). This fits the story of patient relapse after Tamoxifen treatment (McCallum and Lowther, 1996). An alternative explanation is that VP229 was re-sensitised to the drugs in culture or VP267 evolved drug resistance in culture. Secondly, the private structural variants in each cell line imply that they evolved separately from a common ancestor. Perhaps the different structural variant profiles could be explained by loss of rearranged regions or ongoing rearrangement in culture. It would, however, be surprising if so many rearrangements accumulated in culture, given what we know about the relatively slow rate of *in vitro* evolution of other cell lines (Roschke et al., 2002, 2003; Cooke et al., 2010).

An analagous *in vivo* experiment has recently been reported: Campbell et al., (2010) used massively parallel paired end sequencing to compare genome rearrangement profiles in pancreatic cancer metastases. The authors demonstrated two classes of rearrangement: those found in all metastatic lesions reflecting rearrangements in the common ancestor and those “private” to each metastatic lesion. Although frequencies varied over their ten sample sets, an average of approximately 75 percent of structural variants were found in all the metastases from a single patient (Campbell et al., 2010a). This proportion is strikingly similar to the number of rearrangements observed in VP229 and VP267 relative to the implied common ancestor.

5.6.2. The fusion genes in VP229 and VP267

There were at least three predicted fusion transcripts in VP229 and VP267 – these were probably present in the common ancestor of the two cell lines. VP267 had one additional fusion transcript, *TRAPPC9-KCNK9*, which may also have formed *in vitro*. Both of the out of frame fusion transcripts are likely to be passenger events, as non-rearranged genomic regions and normal transcripts are present in the cell lines. The in frame fusion of *MDS1* to *KCNMA1* may, however, produce a functional protein and was probably formed at a time before the relapse-capable clone had evolved within the primary tumour. Given that a proportion of earlier fusion transcripts in HCC1187 may be selected events it is, therefore plausible that the *MDS1-KCNMA1* gene fusion was a driving event.

Potassium channels couple intracellular chemical signalling to electric signalling and stand at the crossroads of several tumour-associated processes such as cell proliferation, survival, secretion and migration (Kunzelmann, 2005). There is evidence that cell membrane potential is depolarised in early G1 phase and that hyperpolarization accompanies progression to S phase. The likely physiological mechanism for hyperpolarization is the opening of a number of K^+ channels (including *KCNMA1*). Blockade of channel activity leads to membrane depolarization and an arrest in early G1 and several studies have indicated membrane hyperpolarization is essential for transition from G0/G1 and G1/S (Ouadid-Ahidouch et al., 2004; Ouadid-Ahidouch and Ahidouch, 2008)

In cancer, the general opinion seems to be that increased expression of voltage-sensitive ion channels is associated with increasing levels of malignancy but there is currently no clear understanding of why this may be (Kunzelmann, 2005; Fiske et al., 2006). For example, In glioma cell lines, up-regulation and constitutive activation of the *KCNMA1* are correlated with increased malignancy (Liu et al., 2002). Similarly, in prostate cancer, amplification of 10q22 causes over-expression of *KCNMA1* and may lead to increased cell proliferation (Bloch et al., 2007). In osteosarcoma, however, *KCNMA1* is down-regulated and may have a tumour-suppressive function in this cancer type (Cambien et al., 2008).

The only evidence of *KCNMA1* involvement in breast cancer is that channels are up-regulated in MDA-MB-361 breast cancer cells supposedly derived from a brain metastases (Khaitan et al., 2009). A second, possibly relevant, finding was that 17-beta-estradiol binds the regulatory subunit of the channel allowing for activation of channel activity in smooth muscle (Valverde et al., 1999). It is tempting to wonder if an oestrogen-independent late-stage breast cancer somehow needs to circumvent this mechanism.

Chapter 6

Recurrent disruption of genes fused in HCC1187 and VP229/VP267

6.1. Introduction

The structure of breast cancer genomes can be complex even among the epithelial cancers (Heim and Mitelman, 2009). Although some of cytogenetically classifiable chromosome aberrations are clearly non-random, for example loss of 17p (*TP53*) (Baker et al., 1989, 1990), loss of 8p (likely to be driven by loss of *NRG1*) (Adélaïde et al., 2003; Huang et al., 2004; Chua et al., 2009) and amplification of *ERBB2* (Slamon et al., 1989), we know little about the large remainder of cytogenetically unclassifiable aberrations. In the previous chapters, I have described several fusion transcripts. In this section, I attempted to establish if any of the fused genes were disrupted recurrently in a panel of breast cancer cell lines.

6.2. Finding broken genes by array CGH

Genes at unbalanced chromosome breaks in SNP6 array CGH can often be identified. For example, the average breakpoint region for 41 breast cancer cell lines could be predicted to within approximately 3kb (Bignell et al., 2010). As shown in Chapter 3, the computationally -predicted breakpoints nearly always coincide exactly with, or are only a few kilobases away, from the true, experimentally proven, breaks.

Recently, Bignell et al. (2010) released SNP6 array CGH data for 755 cancer cell lines, including 41 from breast cancer. This provided me with a large data set in which to look for recurrently broken genes. I first extracted the break point regions using a custom Perl script. The data was in list form with each SNP and copy number probe having an associated PICNIC-segmented copy number (Bignell et al., 2010). The script found change points in PICNIC-segmented copy number and output the breakpoint regions and their “polarity” (Appendix 2.1). Breakpoints at copy number gains p-terminal to q-terminal were scored as positive. Negative breaks were at copy number losses p-terminal to q-terminal (Figure 6.1).

I next established whether any of the break point regions coincided with genes.

Breakpoints outside of genes can also disrupt gene function, for example by forming runthrough gene fusions. In a runthrough fusion, a broken gene splices into the second (or a subsequent) exon of an unbroken gene near to the translocation breakpoint as is the case for *TAXBP1-AHCY* (Howarth et al., 2008). Because of this, I also included the region upstream of the gene.

A list of gene boundaries and gene "windows" was assembled from the Refseq database by Dr. P.A. Edwards. Gene windows extended up to 200kb before a gene or until another gene on the same strand was reached. Breaks within this window potentially caused runthrough fusion with another gene. The lists of breakpoint regions and gene windows became the input for a second Perl script that matched breakpoints with CCDS genes where the breakpoint region was found entirely within the gene or gene window (Appendix 2.3).[1]

Array CGH only identifies unbalanced chromosome break points and balanced reciprocal translocations or inversions would be missed by this approach. However, it is worth noting that rearrangements historically considered as balanced often lose material at chromosome break points, making them visible by array CGH. For example the genic breakpoints associated with *BCR-ABL* fusion often show copy number changes (Figure 6.3) (Sinclair et al., 2000; De Gregori et al., 2007). In addition, as breast cancers often gain or lose whole chromosomes, an initially balanced rearrangement can become visible by array CGH if one of the chromosomes bearing the reciprocal rearrangement is gained or lost (Howarth et al., 2008).

1 A more efficient SQL-based system is currently under development in the Edwards laboratory

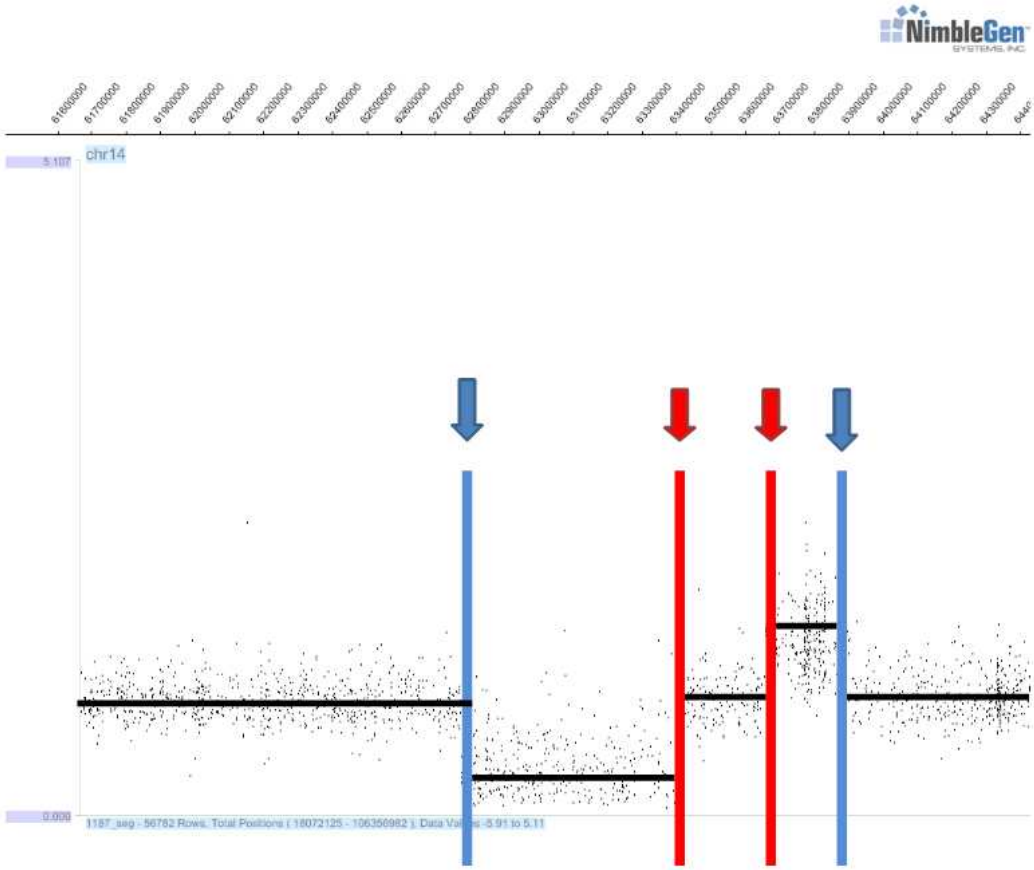


Figure 6.1. Identifying break point regions from PICNIC-segmented SNP6 array CGH data. Figure shows a region of chromosome 14 from the HCC1187 cell line. The data points on the scatter plot are individual probes from the SNP6 array. Copy number segments, defined by PICNIC are the black horizontal lines. Blue lines are 'negative' breakpoint regions as the copy number decreases p-terminal to q-terminal, red lines are positive breaks. The break point regions were defined by the SNP probes flanking the copy number change points.

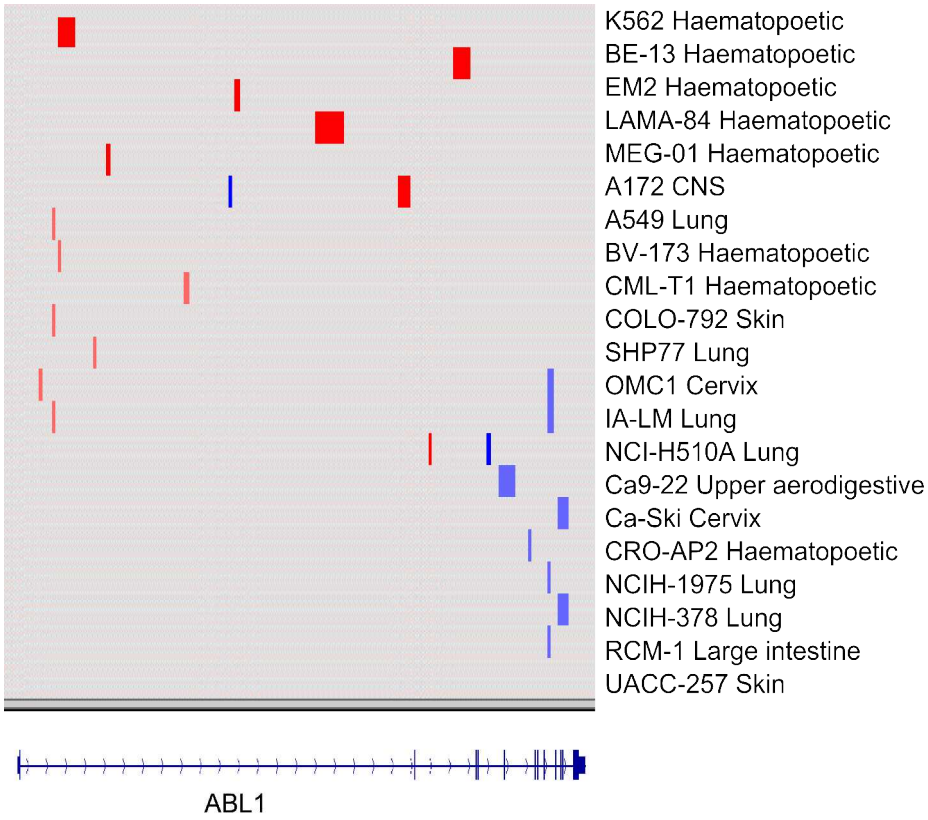


Figure 6.2. Breaks in the *ABL1* gene. Individual samples are listed on the right. Several haematological cancer cell lines show positive breaks in the large first intron of *ABL1* (break point regions are blotted as red bars). This corresponds to the 3' portion of *ABL1* being retained in a possible fusion transcript. Two lung lines also have positive breaks in *ABL1*. Several other lines including haematological, lung, cervix and skin have negative breaks in *ABL1* (blue bars). For these the 5' end of the gene would be retained in a possible fusion transcript.

6.3. Recurrent breaks by array CGH

I used the above methods to look for recurrent breakage in genes found fused in HCC1187 and VP229/VP267 (Table 6.1). In this search for recurrence, I also considered data from massively parallel paired end sequencing from Stephens et al. (2009) and unpublished data from other lab members EM. Batty, JC. Pole and I. Schulte on cell lines MDA-MB-134 and ZR-75-30.

Gene	Total breaks in all cell lines	Total breaks in 41 breast cancer cell lines	Breast cancer breaks retaining 3' end	Breast cancer breaks retaining 5' end
PUM1	7	3	0	3
TRERF1	10	2	1	1
ROD1	3	2	2	0
SUSD1	5	2	0	2
RHOJ	12	1	0	1
SYNE2	11	2	2	1
CTCF	3	1	0	1
SCUBE2	3	1	1	0
CTAGE5	4	1	0	1
SIP1	1	1	0	0
AGPAT5	3	1	0	0
MCPH1	11	1	1	0
SGK1	3	1	0	1
SLC2A12	4	2	1	1
PLXND1	4	1	0	1
TMCC1	123	2	1	1
RGS22	7	1	0	0
SYCP1	5	2	1	2
KCNMA1	30	2	2	0
MECOM	37	6	2	4
FAM125B	7	2	1	1
SPTLC1	3	1	0	0
PDLIM1	1	1	0	1
ZBBX	6	1	1	0
KCNK9	14	2	1	1
TRAPPC9	36	2	0	2

Table 6.1 Breaks in HCC1187 and VP229/VP267 expressed fusion genes by array-CGH. The breaks in HCC1187 and VP229 are included in these figures but not VP267 as SNP6 data was not available.

6.3.1. Breaks in *PUM1* in breast cancer cell lines

Two cell lines in addition to HCC1187 have breaks in *PUM1*: EVSA-T and UACC-812 (Figure 6.3).To investigate a possible gene-fusion in UACC-812, I attempted to amplify the putative 3' fusion partner by RACE. However, this approach only detected the normal copy of the *PUM1* transcript (not shown). The EVSA-T cell line was not available.

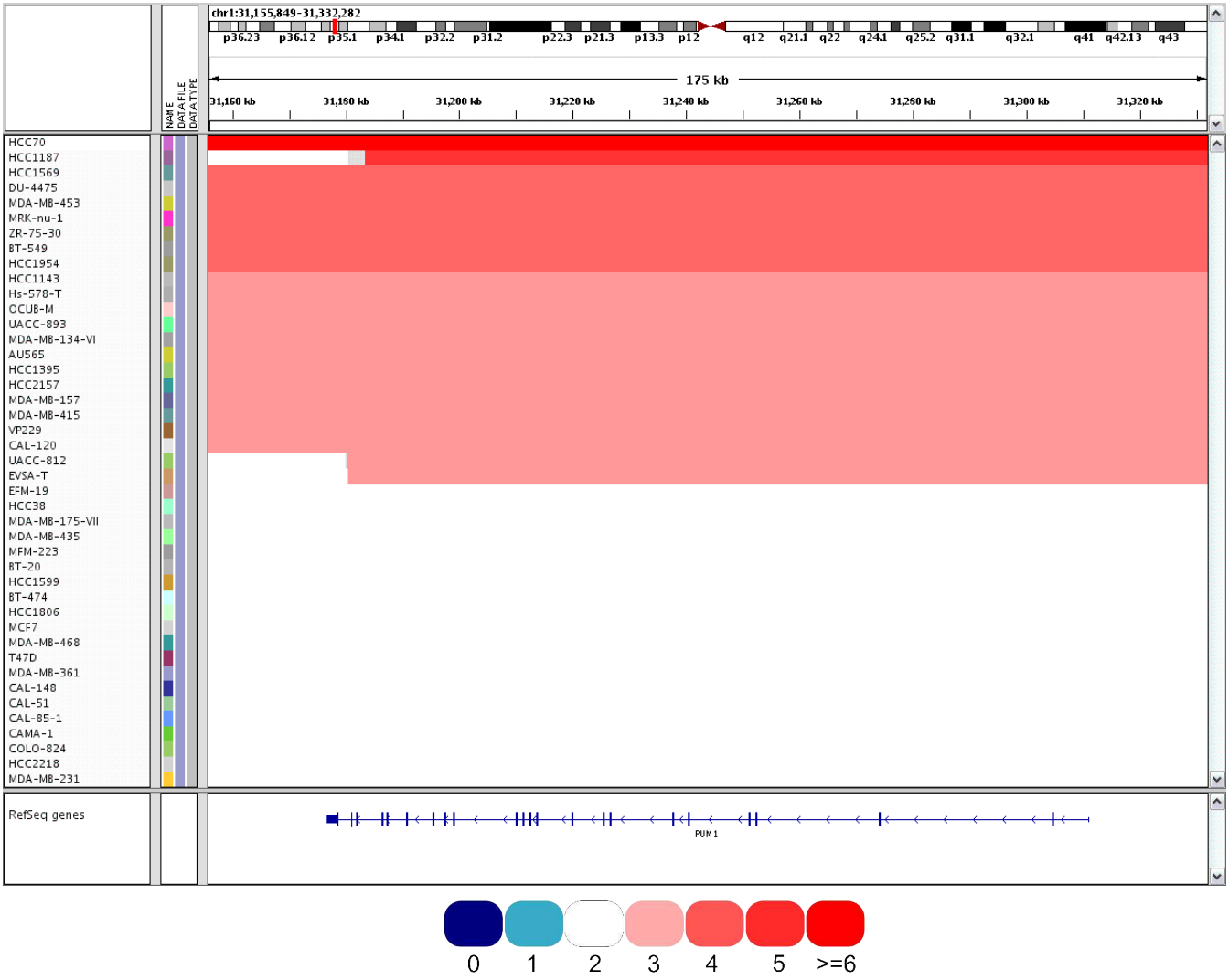


Figure 6.3. Array CGH breaks in *PUM1*. Array CGH copy number is shown as in the above colour scale as. The HCC1187 chromosome break is visible at the top.

6.3.2. Breaks in *TRAPPC9 and KCNK9* in breast cancer cell lines

There are several breaks in the *TRAPPC9* and *KCNK9* regions. AU565, COLO824 and MDA-MB-157 appear to have deletions within the *TRAPPC9* gene. These features are of unknown significance. AU565, MDA-MB-175 and ZR-75-30 all have breaks just up stream of the start of *KCNK9*. These breaks are in the correct configuration to cause runthrough fusion (Figure 6.4).

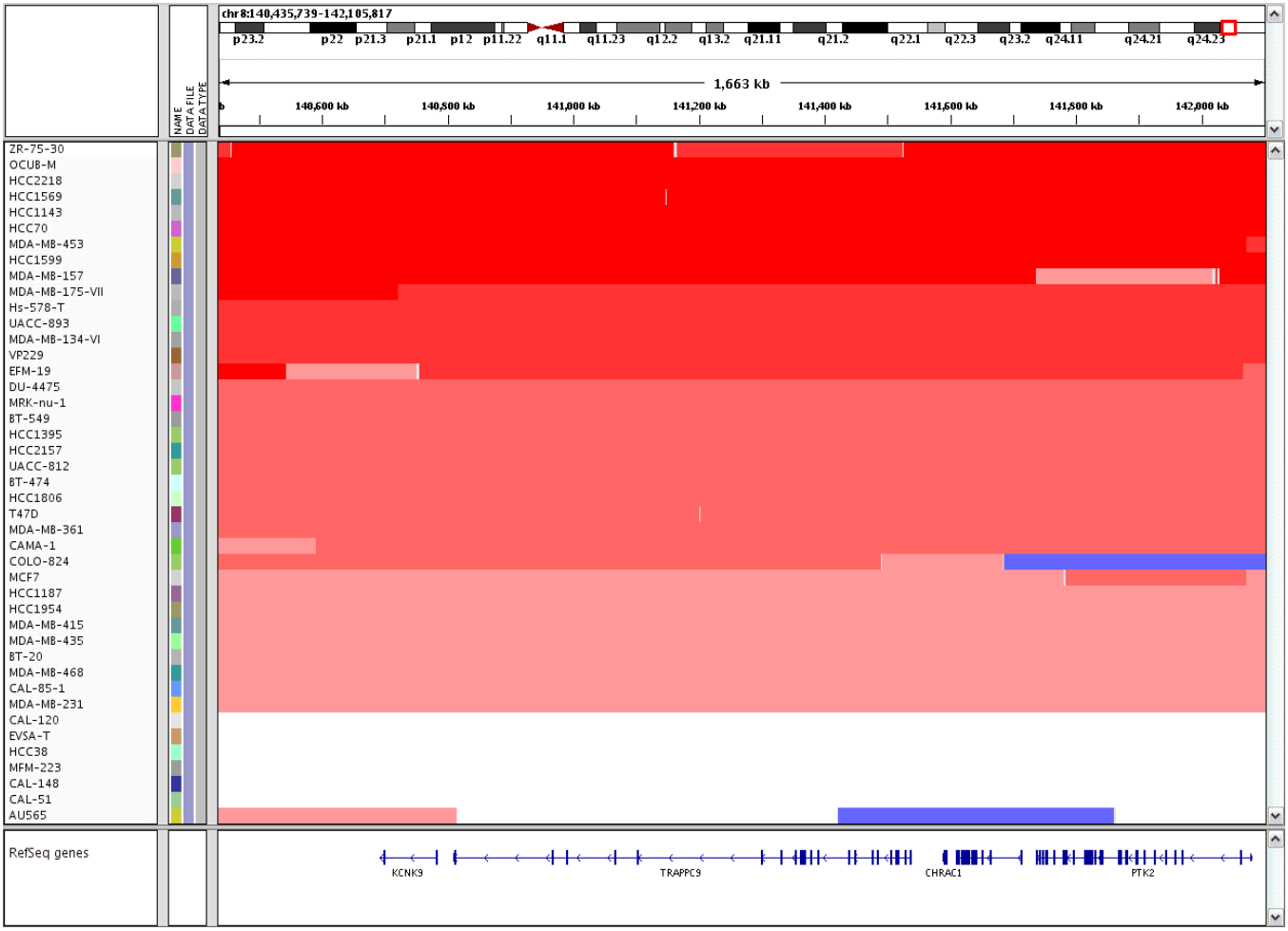
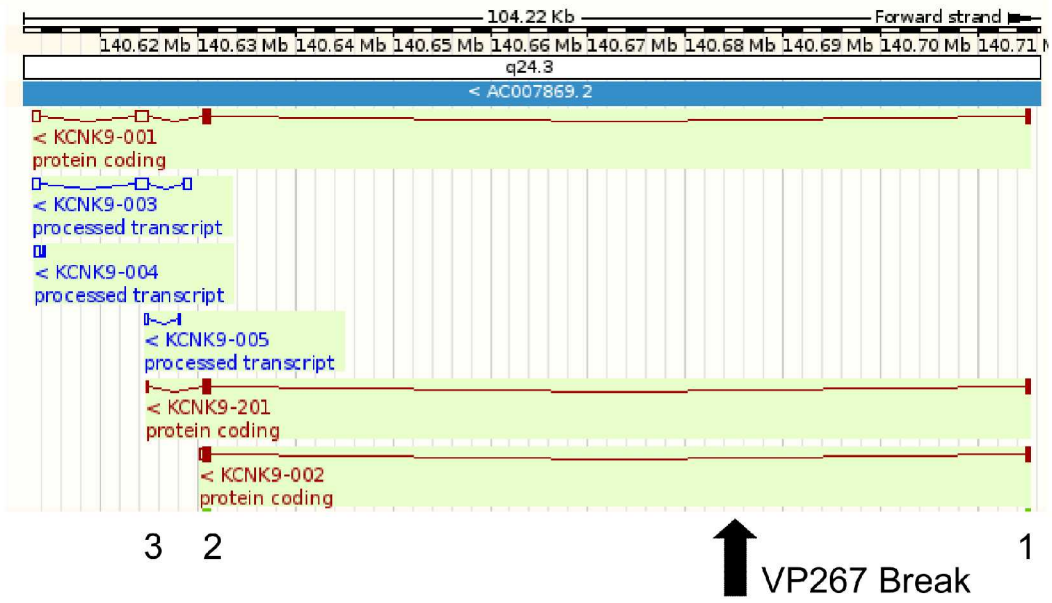


Figure 6.4. Array CGH breaks in *TRAPPC9* and *KCNK9* regions.

KCNK9 is over expressed in breast cancers due to amplification of the genomic locus (Mu et al., 2003). Since the *KCNK9* region is not amplified in VP229 or VP267, I investigated

whether the fusion transcript caused over-expression of the transmembrane domains of *KCNK9* in VP267. Quantitative PCR on cDNA showed a 25-fold increase in the 3' portion of *KCNK9* transcript relative to VP229 (Figure 6.5). The over-expression was probably due to the gene fusion as my control primer pair spanned the genomic break point.

i)



ii)

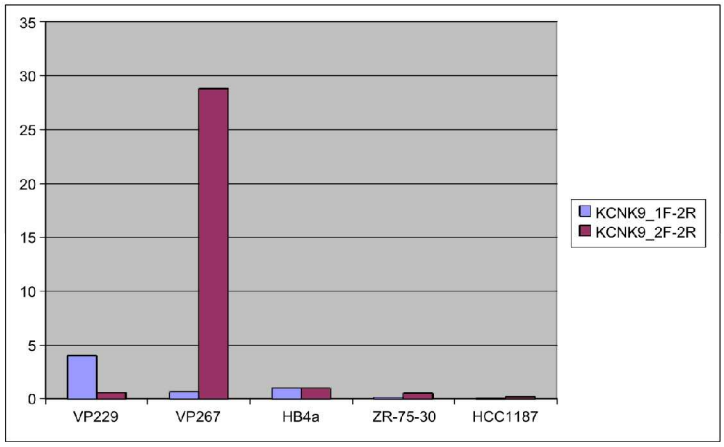


Figure 6.5. Expression levels of *KCNK9* in breast cancer cell lines. i) *KCNK9* genomic locus. VP267 breakpoint is shown. ii) Quantitative RT-PCR for exons 1 and 2 and exon 2 only of *KCNK9*. Y-axis shows expression as a proportion housekeeping gene, *GAPDH*, and scaled to HB4a.

6.3.3. Breaks in *MDS1* (*MECOM*) in breast cancer cell lines

There were several breaks in the *MECOM* locus by array CGH, three of which potentially retained the 5' end of the transcript as in the VP229 and VP267 fusion with *KCNMA1.* These breaks were in HCC1395, AU565 and T47D (Figure 6.6).

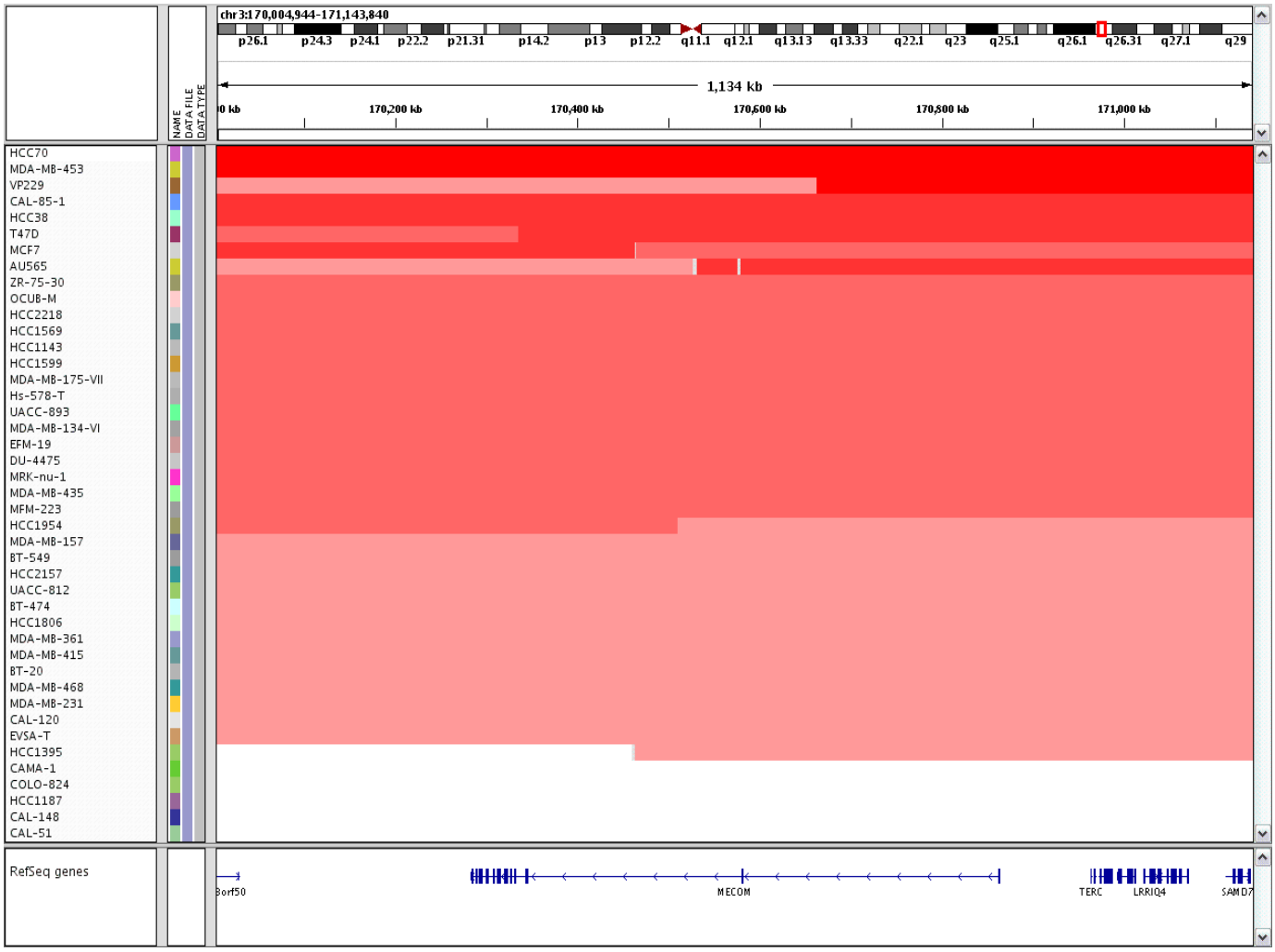


Figure 6.6. Array CGH breaks in MDS1 (MECOM). Array CGH copy number is shown as in the above colour scale as in figure 6.3.

I confirmed these breaks by FISH using two probes, one 5' and one 3' of the *MECOM* locus (Figure 6.7). The AU565 cell line was not available and I instead used SKBr3, another cell line derived from the same patient (Bacus et al., 1990). Just as in VP229 and VP267, a common rearrangement in two cell lines from the same patient probably indicates an *in vivo* event. FISH confirmed that the *MECOM* locus was broken in HCC1395

and SKBr3. In T47D, there were no split signals. It is likely the array CGH segmentation was slightly inaccurate and the true breakpoint was just proximal to the gene.

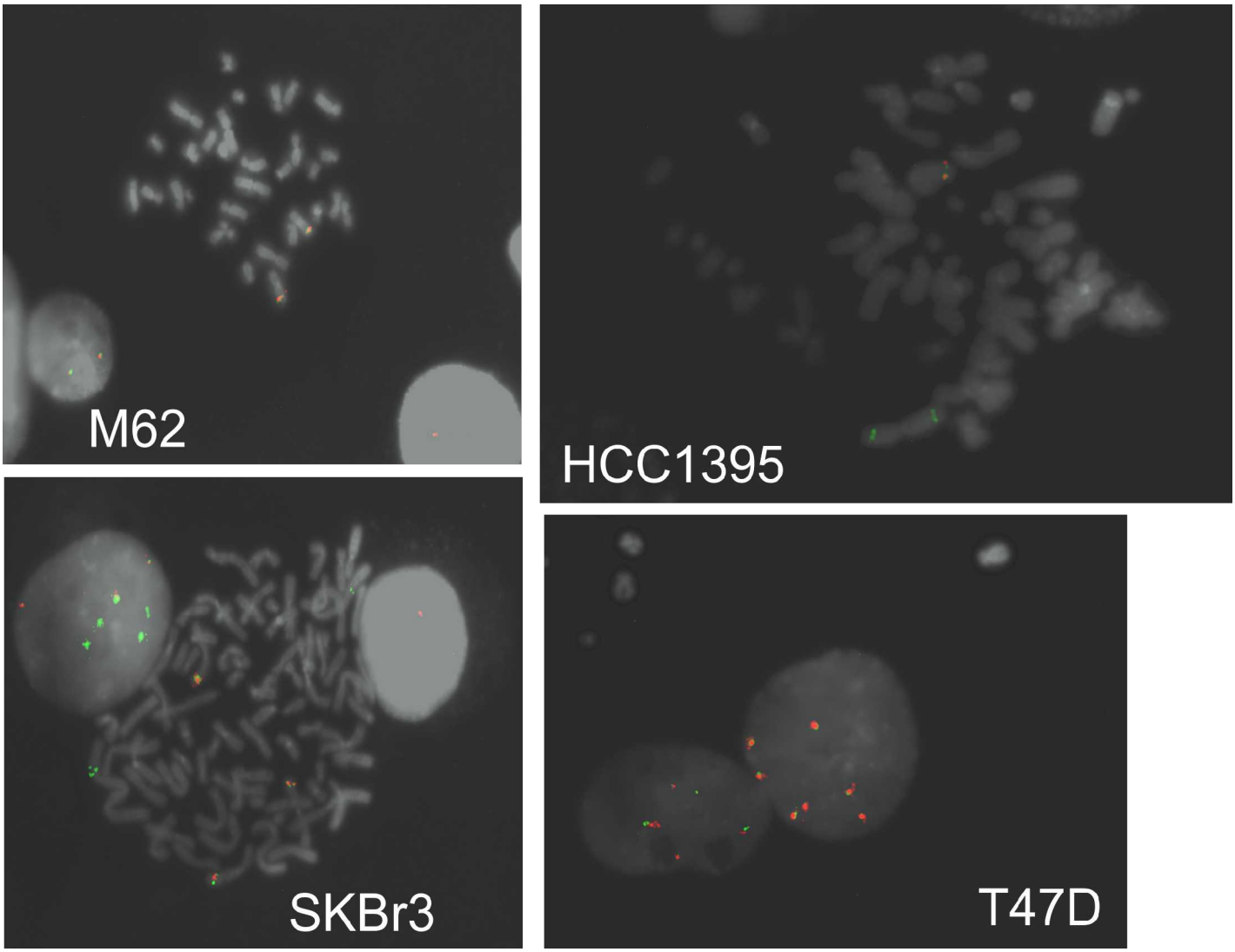


Figure 6.7. FISH to confirm *MDS1* breaks. The 5' end of the gene is in green (BAC RP11-659A23 HG18chr3:170655616-170810166) and the 3' end of the gene is in red (RP11-141C22 HG18chr3:170367524-170542975). Both probes localise to the q-arm of an A group chromosome, likely to be chromosome 3. Unpaired green signals are visible in HCC1395, probably as part of an isochromosome.

In the Stephens et al. (2009) paired end sequencing data there was a predicted fusion of *MDS1* in HCC1395 from an inter-chromosome translocation: chr3:168981393 joined to chr6:84925947 (HG18), predicted to fuse the 5' of *MDS1* into 3' of *KIAA1009*. I could not

show expression of the fusion transcript by RT-PCR. It is possible that this fusion gene is not expressed in HCC1395 or that the paired end sequence data was incomplete. For example, there could be an undefined genomic shard at the translocation junction.

6.3.4. An Internal Rearrangement of *KCNMA1* in BT20

Only one other cell line, besides VP229 and VP267 showed breaks in *KCNMA1*. This was BT20 which had an internal deletion of *KCNMA1*. The breaks could be interpreted as a small interstitial deletion and, if this was the case, an aberrant transcript may result from the locus bearing the deletion. I investigated this possibility by RT-PCR and found a short isoform: *KCNMA1* exon 3 spliced into exon 23. This new isoform was predicted to be out of frame so not likely to produce a functional protein. A full-length *KCNMA1* transcript was also found in BT20, so there was not a homozygous loss of *KCNMA1* (Figure 6.8).

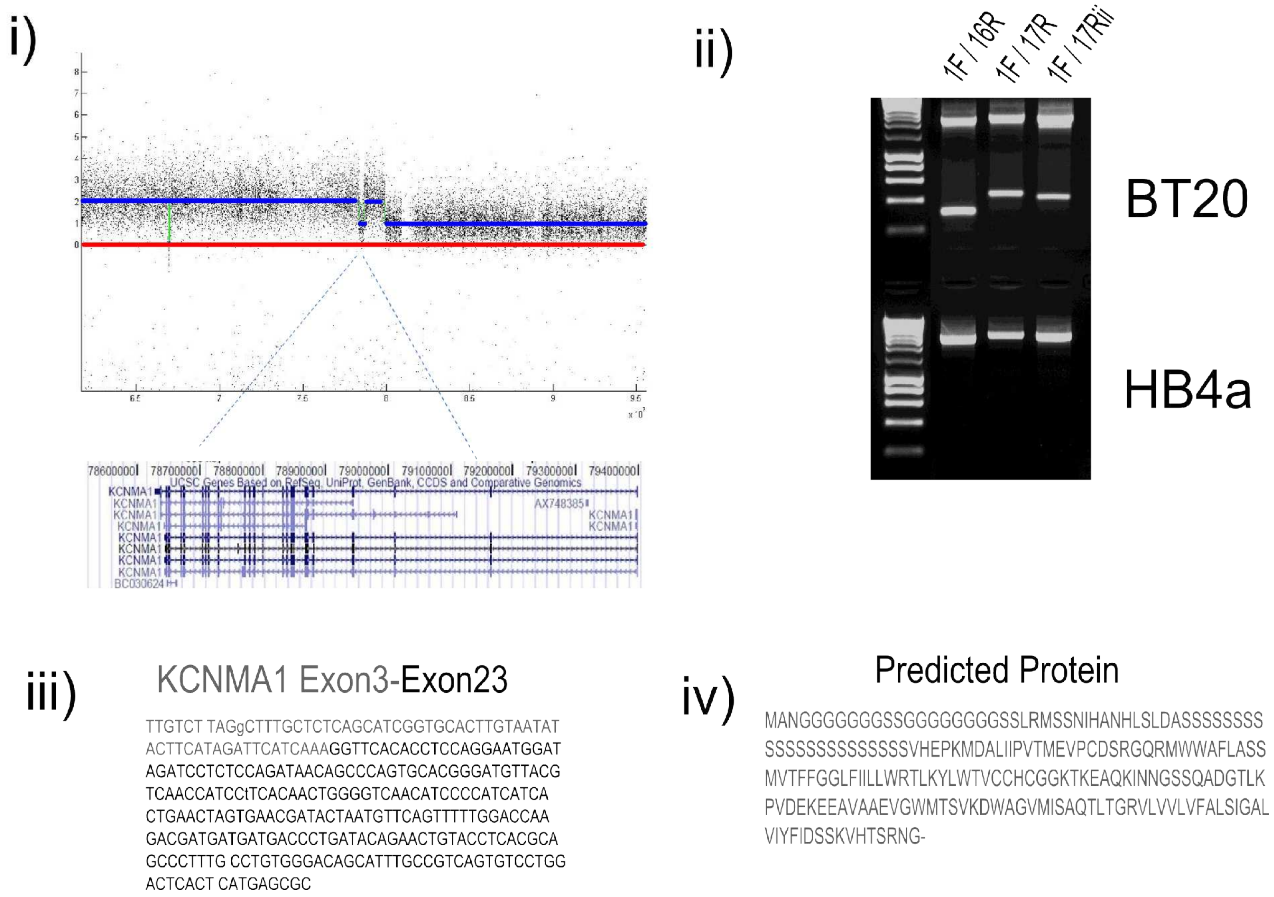


Figure 6.8. Internal deletion of *KCNMA1* in the BT20 cell line. i) PICNIC-segmented SNP6 array CGH. Blue line is the total copy number, red line is minor allele copy number. A small deletion is present at the *KCNMA1* locus. Dotted lines indicate the extent of the deletion. ii) RT-PCR shows a short isoform of KCNMA1, not present in normal breast (HB4a cell line). iii) Sequence across the cDNA junction shows fusion of *KCNMA1* exon 3 with *KCNMA1* exon 23. iv) The protein encoded by the fusion transcript is predicted to be an out of frame. Exons are named as for *KCNMA1*-001 (ENST00000286627)

6.4. Discussion

6.4.1. Methods of analysis

My search for recurrently broken and fused genes was centred around unbalanced break points in array CGH data. The obvious drawbacks of this approach are that it cannot detect balanced rearrangements or rearrangements below approximately 50kb in size. An alternative approach is to search for breaks in primary tumours by tissue microarray FISH. I anticipated some difficulty using this method, as several of my gene-fusions were formed though small tandem duplications and deletions which are difficult to identify by such methods. As the Bignell et al. (2010) array CGH data provided a reasonably large set of samples and all of the gene fusions I described were at unbalanced break points, I thought this was a reasonable way to look for other breaks in these genes.

6.4.2. Recurrent breaks in breast cancer cell lines

I looked in a panel of breast cancer cell lines for recurrent breakage of the expressed fusion genes from HCC1187 and VP229/VP267. Most of these fusion genes were not recurrently broken across many sample but four genes, *PUM1*, *KCNMA1*, *KCNK9* and *MDS1* were broken in other cell lines. It is therefore possible that these genes are fused, probably with partners other than those described here, in other cell lines.

Chapter 7

Discussion

7.1. The Structure of Breast Cancer Genomes

The first aim of this thesis was to define structural rearrangements in breast cancer genomes. I placed an emphasis on finding fusion genes because of their potential clinical utility. I used a combination of molecular cytogenetics and massively parallel paired end sequencing data to achieve this. For HCC1187, the combination of approaches has produced probably one of the most detailed maps of a breast cancer genome to date.

The genome surveys of HCC1187, VP229 and VP267 add to the emerging picture that hundreds of different genes may be disrupted by chromosome aberrations in breast cancer (Hampton et al., 2008; Stephens et al., 2009). Secondly, the average breast cancer expresses several fusion genes. For example, nine in HCC1187, three in VP229, four in VP267, five in MCF7 (Hampton et al. 2008) and seven in ZR-75-30 (Dr I.Schulte, unpublished).

It is difficult to tell which fusion genes are functional in breast tumours as we do not currently know which breast cancer fusion genes are recurrent. Out of frame gene fusions may cause loss of function of either or both genes involved. In frame fusions may be truly oncogeneic gains of functions or could act through dominant-negative mechanisms (Hampton et al. 2008). However, the evolutionary model in chapter four argues that a subset of gene fusions have to be formed at a certain time so were likely to be selected events.

7.1.1. Cell lines as models of breast cancer

As breast cancer cell lines are extensively used as models, the applicability of findings to primary breast tumours are often questioned. This is understandable as many cell lines have survived in culture for decades and often derived from advanced stage tumours (Vargo-Gogola and Rosen, 2007). There is, however, good evidence that breast cancer cell lines, broadly speaking, recapitulate the genomic features of primary tumours.

Neve et al. (2006) compared early and late passage breast cancer cell lines and concluded they had not accumulated substantial new aberrations during culture. The authors went on to showed that, broadly speaking, a panel of 51 breast cancer cell lines showed genomic rearrangements (using 1Mb CGH arrays) and transcriptional profiles similar to those found in a panel of breast tumours. The cell lines also displayed considerable inter-line heterogeneity as observed in primary tumours. Inevitably, some differences were observed between cell lines and primary tumours. For example, cell lines could only be clustered into a single luminal subset, rather than two for primary tumours and the basal-like subset of cell lines can be split into A and B when tumours could not. As cell lines do not contain stromal or normal epithelial cell contamination differences may have been resolved more clearly (Sørlie et al., 2001; Neve et al., 2006; Fridlyand et al., 2006).

Neve et al. (2006) concluded that:

> … the cell line collection mirrors most of the important genomic and resulting transcriptional abnormalities found in primary breast tumors and that analysis of the functions of these genes in the ensemble of cell lines will accurately reflect how they contribute to breast cancer pathophysiologies. (Neve et al., 2006, p.520)

In short, no single cell line can adequately model breast cancer or even a single subtype of it. In stead a panel of cell lines should be consulted, as I attempted to do in chapter six.

7.1.2. The Heterogeneity of Breast Cancer

It is possible that breast cancer is a collection of cancers of the mammary gland, each with a separate set of genetic lesions responsible for its development (Bertucci and Birnbaum, 2008). If this is the case, one would expect to see subtype-specific gene fusions as we do in leukaemias, for example the *SIL-TAL1* fusion gene is only found in T-cell acute lymphoblastic leukaemia (Mansur et al., 2009). If we considered leukaemia to be a single disease, then even the most common subtype-specific fusion genes such as *SIL-TAL1* would be quite rare.

HCC1187 is ER-negative, PR-negative and *ERBB2*-non-amplified so may have originated

from a 'triple negative' breast cancer. As it is likely that triple negative breast cancers have a distinct cell of origin (Foulkes et al., 2010) then a logical next step is to look exclusively within the triple negative category for recurrence of gene fusions such as *PLXND1-TMCC1*. The heterogeneity of breast cancer, its cell of origin and subtypes have proven to be a difficult issue to resolve. In the future, very large studies such as the METABRIC consortium, the Cancer Genome Atlas Research Network and the International Cancer Genome Consortium will provide enough data for definitive classification of breast cancers and allow subtype-specific searches for recurrence.

7.1.3. Is there a better way to find fusion genes in complex genomes?

The purpose of cytogenetics, and the derivatives discussed in this thesis, is to find genes whose expression is altered in cancer. Through the study of chromosome structure, we can, for example, identify genes at chromosome breakpoints. And indeed structural studies that have identified the majority of fusion genes to date (Hampton et al., 2008; Howarth et al., 2008; Stephens et al., 2009).

The two famous examples of fusion genes in common epithelial cancers were found by other, essentially one off, methods. Soda et al. (2007) used a transformation assay to find the *EML4-ALK* fusion gene in non-small-cell lung cancer (Soda et al., 2007). The authors generated a cDNA library from a lung adenocarcinoma and used a retrovirus to insert cDNAs into mouse 3T3 fibroblasts. Tomlins et al. (2005) used a bioinformatic approach to find the *TMPRSS2-ERG* and *TMPRSS2-ETV1* fusion genes. The authors hypothesized that when gene fusion results in the marked over-expression of the 3' gene, this profile should be visible in microarray data. Their cancer outlier profile analysis (COPA) found genes that were highly over-expressed in a subset of prostate tumours (Tomlins et al., 2005). This approach is, however, limited to genes that are highly over-expressed and several famous fusion genes cannot be detected by this method. For example, the outlier profile of *ABL1* and *ALK* in leukaemia and lung cancer datasets respectively are not particularly striking (Rhodes et al., 2004). I did, however, use the online tool *Oncomine* to look for over-expression of the fusion transcripts I had found in breast cancer cell lines (Rhodes et al., 2004). Unfortunately, none had a good outlier profile.

I was reliant on RT-PCR to detect expression of fusion transcripts. The touch down method (Korbie and Mattick, 2008) that I employed was proved to be the most sensitive and specific, but I also considered various PCR derivatives such as splinkerette, vectorette and inverse PCR that amplify from a known sequence to an unknown one (Arnold and Hodgson, 1991; Ochman et al., 1988; Horn et al., 2007). But ultimately the touch-down PCR strategy represented a compromise between sensitivity, specificity, cost and speed.

Transcriptome sequencing approaches can also identify fusion transcripts and are evolving rapidly (Volik et al., 2006; Chinnaiyan et al., 2009; Maher et al., 2009; Zhao et al., 2009; Berger et al., 2010; Sboner et al., 2010). This type of analysis could, in theory, also identify 'bicistonic mRNAs', where there is a fusion between two open reading frames in the absence of chromosomal rearrangement (Guerra et al., 2008). It is, therefore, tempting to conclude that structural studies will give way to transcriptome-based surveys. While this may be justifiable for clinical laboratories, the data presented in this thesis shows that the genomic context in which a fusion gene is found is valuable information for discovery screens.

7.1.4. The mechanisms of fusion gene formation

In order to ask if there are recurrent gene-fusions in breast cancer, it is useful to investigate the mechanisms which may form them. In leukaemias and lymphomas there may be a tendency for genes that are found close together in the interphase nucleus to become fused (Roix et al., 2003). It is also likely that fusion-prone genes often co-localise to the same transcription factory (Osborne et al., 2007), for example, *IGH* and *MYC* in B-cell precursors. But it has also been suggested that these same gene fusions result from the mis-targeting of recombinases *RAG1* and *RAG2*, which usually facilitate somatic rearrangement of immunoglobulin and T-cell receptor loci, to cryptic recombination signal sequences that precede certain genes such as *BCL2*, *LMO2*, *TAL2*, and *TAL1* (Marculescu et al., 2002; Raghavan et al., 2004). Taken together, these observations imply that certain fusions may occur more frequently because of the relative (or combined) input of specific mutational mechanisms.

Breast cancers probably have the most complex genomes of the common cancers. It is likely, as Fridlyand et al. (2006) suggest, that the diverse genomic landscapes of individual breast tumours reflect defects in specific cellular mechanisms. If such defective mechanisms rely on specific DNA motifs such as recombination signal sequences or transcription factor binding sites, then it is possible that specific pairs of genes will be brought into proximity more frequently. An interesting observation is that in prostate cancer, androgen-receptor mediated transcription may play a role in forming gene-fusions (Edwards, 2010). For example, *TMPRSS2* and *ERG* are brought into proximity and cleaved by topisomerase 2B upon androgen-induced transcription (Haffner et al., 2010). Whether an analogous case for oestrogen-regulated genes in breast cancer exists is not yet clear.

If, however, a large proportion of the rearrangements we observe in breast cancer result from some type of non-specific failure of DNA repair – perhaps loss of the homologous recombination repair pathway (Graeser et al., 2010) – we might envisage another possibility: that highly recurrent gene fusions such as *BCR-ABL*, *IGH-MYC*, *TMPRSS2-ERG* are rare, but certain oncogenes fuse somewhat promiscuously as, for example we observe with *MLL* and *RET*. Indeed, the two genes have over thirty fusion partners between them (Mitelman et al., 2010).

7.1.5. Are there recurrent fusion genes in breast cancer?

If genes are brought into proximity more-or-less at random and the probability of many highly-recurrent gene fusions is low, then there is a second possibility: A number of non-recurrent or rare gene fusions and point mutations may all result in the same phenotype.

There is already some data to suggest that this may be the case in breast cancer. For example, Stephens et al. (2009) found a fusion of *ETV6*, part of a known fusion gene in secretory breast carcinoma, and *ITPR2*. The authors could not demonstrate recurrent fusion of *ETV6* with *ITPR2*, but did observe breaks in *ETV6* by FISH in several other tumours. It is therefore possible that *ETV6* is fused with multiple different partners. A

similar possibility presents itself with *MDS1* from the present study. The *MDS1-KCNMA1* fusion was probably an *in vivo* event as it was found in the common ancestor of VP229 and VP267. The locus is a known target of rearrangement in other cancers and it is broken, retaining the 5' end in two other breast cancer cell lines. I did not attempt 3' RACE for *MDS1* so the possibility remains that *MDS1* is fused in multiple cell lines.

7.1.6. Multiple methods of gene disruption and a phenotype-centred view of gene fusions

Gene-fusion is only one method by which gene function can be altered. For example, activating point mutations in *RET* cause multiple endocrine neoplasias and fusions of *RET* contribute to thyroid cancer (Alberti et al., 2003). Perhaps we can see an analogous case for genes such as *KCNK9*. This gene is amplified and over-expressed in breast cancers and it appears as though the *TRAPPC9-KCNK9* fusion in VP267 results in over-expression of the pore domain of *KCNK9*. Thus, it is possible that over-expression of *KCNK9* was achieved by gene-fusion in this case. Other cell lines have breaks slightly upstream of *KCNK9*, so it is possible that over-expression could be achieved by runthrough fusion in these cell lines. The genomic complexity of these regions, however, made it impractical to peruse this hypothesis further.

If we view gene fusion as just one of many ways in which a pathway member can be altered we may begin to see the phenotypic consequences of rare or non-recurrent fusion genes. As well as an internal rearrangement of *KCNMA1* that resulted in expression of a novel isoform, Stephens et al. (2009) also observed an internal rearrangement and novel isoform of *KCNMB2* in another sample. The *KCNMB2* gene encodes an auxiliary beta subunit which influences the calcium sensitivity of the major potassium channel (encoded by *KCNMA1*). Stephens et al. (2009) also reported fusion of *KCNQ5*, a voltage gated potassium channel to *RIMS1*. And from my data, *KCNK9,* another voltage gated channel, is fused. Point mutations in potassium channels are being observed with increasing frequency. From the Wood et al. (2007) mutation screen of eleven breast cancers, there were several point mutations in potassium channel genes including *KCNA5, KCNC2, KCNJ15, KCNQ3*. Interestingly, there were two mutations in *KCNQ5* and *KCNT1* in

colorectal cancer as well as single samples with mutations in *KCNA10, KCNB2, KCNC4, KCND3* and *KCNH4* (Wood et al., 2007). In a recent screen of pancreatic cancer coding exons, there were mutations in *KCNC3, KCNA3, KCNMA1* and *KCNT1* (Yachida et al., 2010) and *KCNH8* were found in three lung carcinomas (Kan et al., 2010). One begins to wonder whether these various mutations all result in aberrant polarisation of cells allowing their transition into G1 phase.

7.2. The Evolution of Breast Cancer Genomes

The breast cancer genomes described in this thesis were complex. The challenge we now face is how to tell which, of the many, mutations are important in cancer evolution. One of the best methods to do this is by looking for recurrence over many samples. For breast cancer, a complex and heterogeneous disease, this would require detailed investigation of very large data sets.

Instead, I took an alternative approach to find mutations that may have been important in tumour development: I interpreted these genomes from an evolutionary perspective. For HCC1187, I classed mutations according to their timing relative to endoreduplication. For VP229 and VP267, I used a comparative lesion sequencing approach to find genome rearrangements in the common ancestor of the two cell lines.

The relative timing of genome rearrangements before or after endoreduplication proved to be very informative as it was an approximate midpoint in the mutational history of the cell line. As certain classes of mutations such as nonsense and indels and gene-fusions clustered early, I was able to estimate the number of selected events in the evolution of this tumour. The comparative lesion sequencing strategy showed that substantial rearrangement had occurred before the clone capable of tamoxifen-resistant relapse had arisen at the primary site. Taken together, these findings indicate that chromosome instability is not a late and irrelevant event but occurs at around the same time as other mutational mechanisms.

7.2.1. Endoreduplication as a cancer genomics tool

HCC1187 is not an isolated example of monosomic evolution followed by endoreduplication. This is, in fact, a common evolutionary route in breast and colon carcinomas (Dutrillaux et al., 1991; Dutrillaux, 1995). There are several cell lines such as T47D, DU4475 and MDA-MB-468 that have clearly evolved through monosomy and then endoreduplicated (Davidson et al., 2000).

I suggest that endoreduplication be used as a tool to investigate tumour evolution. For example, the relative timing of adenoma-carcinoma sequence mutations could be investigated relative to endoreduplication. Over a large dataset, one would always expect to see homozygous (early) mutations of *APC* and two or more late heterozygous mutations of *TP53* if the currently accepted scheme for colorectal cancer evolution is correct (Cho and Vogelstein, 1992).

7.2.2. Comparative lesion sequencing

Comparative lesion sequencing can also tell us about the relative timing of cancer mutations. The bulk of cancer-related mortality is due to metastasis. Some assume that the ability to metastasise is an acquired attribute (Yokota, 2000). Some, however, argue that once a tumour becomes invasive it is inherently able to metastasise (Weiss et al., 1983; Edwards, 2002).

The answer to this debate may come from comparative lesion sequencing. If a mutation within a single cell of a primary tumour confers the ability to metastasise, then all metastatic lesions will contain that mutation but the bulk of the primary tumour will not. The private mutations of the metastatic lesions will display an unrelated variety of passenger mutations.

An opposite case may be argued for drug resistance, where it often appears that a minor population of cells within a tumour can grow out after their drug-sensitive competitors have been killed (Engelman et al., 2007; Turke et al., 2010; Cooke et al., 2010). In this case, only mutations apparently private to the relapse (but probably found in the primary tumour

in low frequency) would be informative about the molecular mechanism behind drug resistance.

7.3. Future Directions

7.3.1. The challenges of massively parallel sequencing

The high-resolution map of the HCC1187 genome allowed me to assess the sensitivity of the Stephens et al. (2009) massively parallel paired end sequencing approach. There was a disparity between the mathematically-predicted sensitivity and the experimentally-observed values. There may be several reasons for the lower-than-expected yield of structural variants from massively parallel paired end sequencing.

At the moment, we are particularly poor at identifying rearrangements involving centromeres, telomeres, sub telomeres and other repetitive sequences. Using algorithms such as MAQ or BWA, sequencing reads must align uniquely in the genome or they are discarded. As reads from repeat regions have multiple mappings, these algorithms cannot identify structural variants within such regions. Furthermore, we know from constitutional cytogenetics that chromosome translocations, deletions and duplications are often mediated by recombination between segmental duplications (Rudd et al., 2009). If this mechanism is operative in cancer too (Darai-Ramqvist et al., 2008), then the resulting chromosome aberrations would be difficult to define with existing approaches.

Secondly, the sequence at chromosome aberration break points is often complex, containing non-templated sequences, small indels and inversions (Bignell et al., 2004; Stephens et al., 2009) such sequences are also difficult to align to the reference genome. The BWA-based alignments in this thesis, for example, only allowed for two mismatches within each short sequencing read. A three base pair indel would, therefore, have been discarded.

Bioinformatics analysis of massively parallel sequencing data is evolving quickly. At the same time, the cost of sequencing is decreasing. In the near future, it is likely that deep

sequence coverage reads could be assembled *de novo* for each new tumour genome (Zerbino and Birney, 2008; Li et al., 2010).This would, in theory, find a higher proportion of genome rearrangements and circumvent the problems of complex genomic shards at chromosome break points.

As somatic rearrangement of tumour DNA probably occurs in most epithelial cancers (Lengauer et al., 1998; Bignell et al., 2007; Beroukhim et al., 2010), two recent studies have shown the potential for this phenomenon to produce personalised tumour biomarkers (Leary et al., 2010; McBride et al., 2010). After structural variant junctions have been identified by paired end sequencing it is relatively easy to design PCR primers flanking the rearrangement junction. This PCR-based assay provides a sensitive and specific test for circulating tumour DNA in patient plasma. Leary et al. (2010) claim to detect tumour DNA molecules at levels lower than 0.001% in patient plasma samples. Importantly, it doesn't matter if the rearrangement junctions are driver or passenger mutations, only that they are found in the primary tumour and are not lost in the relapse or progression clone.

7.3.2. Using structure and sequence together to investigate cancer genome evolution

Recent whole tumour genome sequencing studies have used longer paired end reads to generate high physical and sequence coverage (Pleasance et al., 2010a, 2010b). Such studies define structural and sequence-level mutations in one experiment. These studies have uncovered thousands of sequence-level somatic mutations in individual cancer genomes. Although a huge majority are probably passenger mutations (Stratton et al., 2009), they can still tell us something about the evolution of the cancer genome.

If a genome region bearing a somatic mutation is duplicated, then so is the mutation. For example, if we have two disparate loci, both found in two copies but having lost heterozygosity we can speculate on the relative timing of each duplication by comparing the relative proportions of homozygous (pre-duplication) and heterozygous (post duplication) mutations from each locus. For example, chromosomes thirteen and seventeen of NCI-H209 are two such regions. The simplest explanation for the 2-copy LOH state is chromosomal loss followed by duplication of the remaining copy. For

chromosome thirteen, 59 percent of the mutations are homozygous. For chromosome seventeen, 41 percent of mutations are homozygous. The lower proportion of duplicated mutations on chromosome seventeen implies that it duplicated before chromosome thirteen (assuming the background rate of mutation was the same for both chromosomes). Greenman et al. (2010) used a mathematical model of the accumulation of mutations to estimate that duplication of chromosome thirteen and seventeen happened approximately 78 and 68 percent of the way through the evolutionary history of that cell line respectively. There are homozygous mutations of *RB1* and *TP53* on chromosomes thirteen and seventeen in this cell line. As homozygous mutations probably occurred before the chromosome duplications, we now have an estimate for the latest possible time that the cell line had functional *RB1* and *TP53* proteins (Greenman et al., 2010; Greenman, 2010).

A similar approach to my own was recently reported from the whole genome sequencing of a melanoma cell line, COLO-829 by Pleasance et al. (2010a).

> By combining information on chromosome copy number change with base substitutions the relative order of some mutations on this mitotic lineage can be established. Several genomic regions in COLO-829 show evidence of loss of one parental chromosome, leading to LOH, followed by re-duplication of the remaining copy. In these regions, mutations which occurred before the re-duplication event will be homozygous, whereas those arising after re-duplication will be heterozygous. In most such regions, a small fraction of mutations are heterozygous, indicating relatively late re-duplication. However, in a region of LOH on chromosome 1q, there are more heterozygote substitutions than homozygote, suggesting earlier re-duplication. (Pleasance et al., 2010a, p.5)

Interestingly, Pleasance et al (2010) were able to compare the 'earlier' homozygous mutation spectrum with the later heterozygous spectrum. The earlier category showed signs of increased UV type damage suggesting ultraviolet light exposure was a major early mutational mechanism in this skin cancer. Thus, we can use a complex karyotype as a tool to understand tumour evolution as a whole.

7.4. Conclusions

We should no longer consider genome rearrangement and point mutation as separate from one another. Rather, when sequence-level mutations are viewed in their chromosomal, and therefore evolutionary, context we can find important classes of mutation more easily.

References

Adélaïde, J., Huang, H., Murati, A., Alsop, A. E., Orsetti, B., Mozziconacci, M., Popovici, C., Ginestier, C., Letessier, A., Basset, C., et al. (2003). A recurrent chromosome translocation breakpoint in breast and pancreatic cancer cell lines targets the neuregulin/NRG1 gene. Genes Chromosomes Cancer *37*, 333-45.

Adeyinka, A., Mertens, F., Idvall, I., Bondeson, L., Ingvar, C., Heim, S., Mitelman, F., and Pandis, N. (1998). Cytogenetic findings in invasive breast carcinomas with prognostically favourable histology: a less complex karyotypic pattern? Int. J. Cancer *79*, 361-364.

Agarwal, R., and Kaye, S. B. (2003). Ovarian cancer: strategies for overcoming resistance to chemotherapy. Nat. Rev. Cancer *3*, 502-516.

van Agthoven, T., Veldscholte, J., Smid, M., van Agthoven, T. L. A., Vreede, L., Broertjes, M., de Vries, I., de Jong, D., Sarwari, R., and Dorssers, L. C. J. (2009). Functional identification of genes causing estrogen independence of human breast cancer cells. Breast Cancer Res. Treat *114*, 23-30.

Alsop, A. E., Teschendorff, A. E., and Edwards, P. A. W. (2006). Distribution of breakpoints on chromosome 18 in breast, colorectal, and pancreatic carcinoma cell lines. Cancer Genet. Cytogenet *164*, 97-109.

Arkesteijn, G., Jumelet, E., Hagenbeek, A., Smit, E., Slater, R., and Martens, A. (1999). Reverse chromosome painting for the identification of marker chromosomes and complex translocations in leukemia. Cytometry *35*, 117-124.

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

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*, 161-9.

Arnold, C., and Hodgson, I. J. (1991). Vectorette PCR: a novel approach to genomic walking. PCR Methods Appl *1*, 39-42.

Attard, G., Jameson, C., Moreira, J., Flohr, P., Parker, C., Dearnaley, D., Cooper, C. S., and de Bono, J. S. (2009). Hormone-sensitive prostate cancer: a case of ETS gene fusion heterogeneity. J. Clin. Pathol *62*, 373-376.

Bacus, S. S., Kiguchi, K., Chin, D., King, C. R., and Huberman, E. (1990). Differentiation of cultured human breast cancer cells (AU-565 and MCF-7) associated with loss of cell surface HER-2/neu antigen. Mol. Carcinog *3*, 350-362.

Baker, S. J., Fearon, E. R., Nigro, J. M., Hamilton, S. R., Preisinger, A. C., Jessup, J. M., vanTuinen, P., Ledbetter, D. H., Barker, D. F., Nakamura, Y., et al. (1989). Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science *244*, 217-221.

Baker, S. J., Preisinger, A. C., Jessup, J. M., Paraskeva, C., Markowitz, S., Willson, J. K., Hamilton, S., and Vogelstein, B. (1990). p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. Cancer Res *50*, 7717-7722.

Barjesteh van Waalwijk van Doorn-Khosrovani, S., Erpelinck, C., van Putten, W. L. J., Valk, P. J. M., van der Poel-van de Luytgaarde, S., Hack, R., Slater, R., Smit, E. M. E., Beverloo, H. B., Verhoef, G., et al. (2003). High EVI1 expression predicts poor survival in acute myeloid

leukemia: a study of 319 de novo AML patients. Blood *101*, 837-845.

Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B. J. (2008). Evaluation of Paired-End Sequencing Strategies for Detection of Genome Rearrangements in Cancer. PLoS Computational Biology *4*, e1000051.

Batty, E. (2010). Fusion Genes in Breast Cancer. PhD Thesis. (University of Cambridge).

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.

Bell, D. W. (2010). Our changing view of the genomic landscape of cancer. J. Pathol *220*, 231-243.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

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.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

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*, 287-95.

Bignell, G. R., Santarius, T., Pole, J. C. M., 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*, 1296-303.

Bloch, M., Ousingsawat, J., Simon, R., Schraml, P., Gasser, T. C., Mihatsch, M. J., Kunzelmann, K., and Bubendorf, L. (2007). KCNMA1 gene amplification promotes tumor cell proliferation in human prostate cancer. Oncogene *26*, 2525-2534.

Blood, K. A. (2006). Chromosome Rearrangements in Breast Carcinomas PhD Thesis. (University of Cambridge).

Bock, W. J. (1959). Preadaptation and Multiple Evolutionary Pathways. Evolution *13*, 194-211.

Bodmer, W. F. (2006). Cancer genetics: colorectal cancer as a model. J Hum Genet *51*, 391-6.

Bomme, L., Bardi, G., Pandis, N., Fenger, C., Kronborg, O., and Heim, S. (1998). Cytogenetic

analysis of colorectal adenomas: karyotypic comparisons of synchronous tumors. Cancer Genet. Cytogenet *106*, 66-71.

Bork, P., and Beckmann, G. (1993). The CUB domain. A widespread module in developmentally regulated proteins. J. Mol. Biol *231*, 539-545.

Bos, J. L. (1989). ras oncogenes in human cancer: a review. Cancer Res *49*, 4682-4689.

Bustin, S. A. (2005). Real-time, fluorescence-based quantitative PCR: a snapshot of current procedures and preferences. Expert Rev. Mol. Diagn *5*, 493-498.

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

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

Cambien, B., Rezzonico, R., Vitale, S., Rouzaire-Dubois, B., Dubois, J., Barthel, R., Karimdjee, B. S., Soilihi, B. K., Mograbi, B., Schmid-Alliana, A., et al. (2008). Silencing of hSlo potassium channels in human osteosarcoma cells promotes tumorigenesis. Int. J. Cancer *123*, 365-371.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., and Stratton, M. R. (2008a). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A *105*, 13081.

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. (2008b). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet *40*, 722-729.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109-1113.

Cerveira, N., Ribeiro, F. R., Peixoto, A., Costa, V., Henrique, R., Jerónimo, C., and Teixeira, M. R. (2006). TMPRSS2-ERG gene fusion causing ERG overexpression precedes chromosome copy number changes in prostate carcinomas and paired HGPIN lesions. Neoplasia *8*, 826-832.

Chatterjee, R. (2007). Cell biology. Cases of mistaken identity. Science *315*, 928-931.

Chen, J. J., Silver, D., Cantor, S., Livingston, D. M., and Scully, R. (1999). BRCA1, BRCA2, and Rad51 operate in a common DNA damage response pathway. Cancer Res *59*, 1752s-1756s.

Cheng, C., Lin, Y., Tsai, M., Chen, C., Hsieh, M., Chen, C., and Yang, R. (2009). SCUBE2 Suppresses Breast Tumor Cell Proliferation and Confers a Favorable Prognosis in Invasive Breast Cancer. Cancer Res *69*, 3634 -3641.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., 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.

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

Cho, K. R., and Vogelstein, B. (1992). Genetic alterations in the adenoma--carcinoma sequence. Cancer *70*, 1727-1731.

Chua, Y. L., Ito, Y., Pole, J. C. M., Newman, S., Chin, S., Stein, R. C., Ellis, I. O., Caldas, C., O'Hare, M. J., Murrell, A., et al. (2009). The NRG1 gene is frequently silenced by methylation in breast cancers and is a strong candidate for the 8p tumour suppressor gene. Oncogene *28*, 4041-4052.

Comtesse, N., Niedermayer, I., Glass, B., Heckel, D., Maldener, E., Nastainczyk, W., Feiden, W., and Meese, E. (2002). MGEA6 is tumor-specific overexpressed and frequently recognized by patient-serum antibodies. Oncogene *21*, 239-247.

Cooke, S. L., Temple, J., Macarthur, S., Zahra, M. A., Tan, L. T., Crawford, R. A. F., Ng, C. K. Y., Jimenez-Linan, M., Sala, E., and Brenton, J. D. (2010a). Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer. Br. J. Cancer. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21063398 [Accessed December 29, 2010].

Cooke, S., Ng, C. K. Y., Melnyk, N., Garcia, M. J., Hardcastle, T., Temple, J., Langdon, S., Huntsman, D., and Brenton, J. D. (2010b). Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. Oncogene. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20581869 [Accessed July 23, 2010].

Cooke, S., Pole, J. C. M., Chin, S., Ellis, I. O., Caldas, C., and Edwards, P. A. W. (2008). High-resolution array CGH clarifies events occurring on 8p in carcinogenesis. BMC Cancer *8*, 288.

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

Davison, A. C., and Hinkley, D. V. (1999). Bootstrap methods and their applications (Cambridge: Cambridge University Press).

De Gregori, M., Ciccone, R., Magini, P., Pramparo, T., Gimelli, S., Messa, J., Novara, F., Vetro, A., Rossi, E., Maraschio, P., et al. (2007). Cryptic deletions are a common finding in "balanced" reciprocal and complex chromosome rearrangements: a study of 59 patients. J. Med Genet *44*, 750-762.

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.

Dorssers, L. C., and Veldscholte, J. (1997). Identification of a novel breast-cancer-anti-estrogen-resistance (BCAR2) locus by cell-fusion-mediated gene transfer in human breast-cancer cells. Int. J. Cancer *72*, 700-705.

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., et al. (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N. Engl. J. Med *344*, 1031-1037.

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*, 245-55.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., 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.

Edwards, P. (2010). Fusion genes and chromosome translocations in the common epithelial cancers. J. Pathol *220*, 244-254.

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.

van den Engh, G., Trask, B., Lansdorp, P., and Gray, J. (1988). Improved resolution of flow cytometric measurements of Hoechst- and chromomycin-A3-stained human chromosomes after addition of citrate and sulfite. Cytometry *9*, 266-270.

Erikson, J., Nishikura, K., ar-Rushdi, A., Finan, J., Emanuel, B., Lenoir, G., Nowell, P. C., and Croce, C. M. (1983). Translocation of an immunoglobulin kappa locus to a region 3' of an unrearranged c-myc oncogene enhances c-myc transcription. Proc. Natl. Acad. Sci. U.S.A *80*, 7581-7585.

Esteller, M., Corn, P. G., Baylin, S. B., and Herman, J. G. (2001). A gene hypermethylation profile of human cancer. Cancer Res *61*, 3225-3229.

Falls, D. L. (2003). Neuregulins: functions, forms, and signaling strategies. Exp. Cell Res *284*, 14-30.

Fears, S., Mathieu, C., Zeleznik-Le, N., Huang, S., Rowley, J. D., and Nucifora, G. (1996). Intergenic splicing of MDS1 and EVI1 occurs in normal tissues as well as in myeloid leukemia and produces a new member of the PR domain family. Proc. Natl. Acad. Sci. U.S.A *93*, 1642-1647.

Fiche, M., Avet-Loiseau, H., Maugard, C. M., Sagan, C., Heymann, M. F., Leblanc, M., Classe, J. M., Fumoleau, P., Dravet, F., Mahé, M., et al. (2000). Gene amplifications detected by fluorescence in situ hybridization in pure intraductal breast carcinomas: relation to morphology, cell proliferation and expression of breast cancer-related genes. Int J Cancer *89*, 403-10.

Fiegler, H., Gribble, S. M., Burford, D. C., Carr, P., Prigmore, E., Porter, K. M., Clegg, S., Crolla, J.

A., Dennis, N. R., Jacobs, P., et al. (2003). Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays. BMJ *40*, 664.

Fiske, J. L., Fomin, V. P., Brown, M. L., Duncan, R. L., and Sikes, R. A. (2006). Voltage-sensitive ion channels and cancer. Cancer Metastasis Rev *25*, 493-500.

Forbes, S. A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C. Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res *38*, D652-657.

Ford, A. M., Bennett, C. A., Price, C. M., Bruin, M. C., Van Wering, E. R., and Greaves, M. (1998). Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. Proc. Natl. Acad. Sci. U.S.A *95*, 4584-4588.

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.

Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177.

Galgano, A., Forrer, M., Jaskiewicz, L., Kanitz, A., Zavolan, M., and Gerber, A. P. (2008). Comparative Analysis of mRNA Targets for Human PUF-Family Proteins Suggests Extensive Interaction with the miRNA Regulatory System. PLoS ONE *3*, e3164.

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 1 Mb 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.

Ghayad, S. E., Vendrell, J. A., Bieche, I., Spyratos, F., Dumontet, C., Treilleux, I., Lidereau, R., and Cohen, P. A. (2009). Identification of TACC1, NOV, and PTTG1 as new candidate genes associated with endocrine therapy resistance in breast cancer. J. Mol. Endocrinol *42*, 87-103.

Gizard, F., Robillard, R., Barbier, O., Quatannens, B., Faucompré, A., Révillion, F., Peyrat, J., Staels, B., and Hum, D. W. (2005). TReP-132 controls cell proliferation by regulating the expression of the cyclin-dependent kinase inhibitors p21WAF1/Cip1 and p27Kip1. Mol. Cell. Biol *25*, 4335-4348.

Gizard, F., Robillard, R., Gross, B., Barbier, O., Révillion, F., Peyrat, J., Torpier, G., Hum, D. W., and Staels, B. (2006). TReP-132 is a novel progesterone receptor coactivator required for the inhibition of breast cancer cell growth and enhancement of differentiation by progesterone. Mol. Cell. Biol *26*, 7632-7644.

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. Biostatistics *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.

Grigorova, M., Staines, J. M., Ozdag, H., Caldas, C., and Edwards, P. A. W. (2004). Possible causes of chromosome instability: comparison of chromosomal abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. Cytogenet. Genome Res *104*, 333-340.

Grigorova, M., and Edwards, P. A. (2004). Breast Carcinoma Cell Lines - HCC1187. Available at: http://www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC1187.html [Accessed November 16, 2010].

Guerra, E., Trerotola, M., Dell' Arciprete, R., Bonasera, V., Palombo, B., El-Sewedy, T., Ciccimarra, T., Crescenzi, C., Lorenzini, F., Rossi, C., et al. (2008). A bicistronic CYCLIN D1-TROP2 mRNA chimera demonstrates a novel oncogenic mechanism in human cancer. Cancer Res *68*, 8113-21.

Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S., and Yulug, I. G. (2009). Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. Oncol. Res *17*, 353-365.

Hahn, Y., Bera, T. K., Gehlhaus, K., Kirsch, I. R., Pastan, I. H., and Lee, B. (2004). Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. Proc Natl Acad Sci U S A *101*, 13257.

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. (2008). 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 Research *19*, 167-177.

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

Heckel, D., Brass, N., Fischer, U., Blin, N., Steudel, I., Türeci, O., Fackler, O., Zang, K. D., and Meese, E. (1997). cDNA cloning and chromosomal mapping of a predicted coiled-coil proline-rich protein immunogenic in meningioma patients. Hum. Mol. Genet *6*, 2031-2041.

Heim, S., and Mitelman, F. (2009). Cancer Cytogenetics 3rd ed. (WileyBlackwell).

Heisterkamp, N., Stam, K., Groffen, J., de Klein, A., and Grosveld, G. (1985). Structural organization of the bcr gene and its role in the Ph' translocation. Nature *315*, 758-761.

Hercus, C. (2009). Novoalign. Available at: www.novocraft.com.

Hingorani, S. R., Wang, L., Multani, A. S., Combs, C., Deramaudt, T. B., Hruban, R. H., Rustgi, A. K., Chang, S., and Tuveson, D. A. (2005). Trp53R172H and KrasG12D cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice. Cancer Cell *7*, 469-483.

Hofmann, W., Komor, M., Wassmann, B., Jones, L. C., Gschaidmeier, H., Hoelzer, D., Koeffler, H. P., and Ottmann, O. G. (2003). Presence of the BCR-ABL mutation Glu255Lys prior to STI571 (imatinib) treatment in patients with Ph+ acute lymphoblastic leukemia. Blood *102*,

659-661.

Horn, C., Hansen, J., Schnutgen, F., Seisenberger, C., Floss, T., Irgang, M., De-Zolt, S., Wurst, W., von Melchner, H., and Noppinger, P. R. (2007). Splinkerette PCR for more efficient characterization of gene trap events. Nat Genet *39*, 933-934.

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.

Huang, H., Chin, S., Ginestier, C., Bardou, V., Adélaïde, J., Iyer, N. G., Garcia, M. J., Pole, J. C., Callagy, G. M., Hewitt, S. M., et al. (2004). A recurrent chromosome breakpoint in breast cancer at the NRG1/neuregulin 1/heregulin gene. Cancer Res *64*, 6840-6844.

Huang, Y., and Rane, S. G. (1994). Potassium channel induction by the Ras/Raf signal transduction cascade. J. Biol. Chem *269*, 31183-31189.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., et al. (2010). International network of cancer genome projects. Nature *464*, 993-998.

Hughes-Davies, L., Huntsman, D., Ruas, M., Fuks, F., Bye, J., Chin, S., Milner, J., Brown, L. A., Hsu, F., Gilks, B., et al. (2003). EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell *115*, 523-535.

Jensen, D. E., Proctor, M., Marquis, S. T., Gardner, H. P., Ha, S. I., Chodosh, L. A., Ishov, A. M., Tommerup, N., Vissing, H., Sekido, Y., et al. (1998). BAP1: a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression. Oncogene *16*, 1097-112.

Jensen, D. E., and Rauscher, F. J. (1999). BAP1, a candidate tumor suppressor protein that interacts with BRCA1. Ann. N. Y. Acad. Sci *886*, 191-194.

Johansson, B., Mertens, F., and Mitelman, F. (1996). Primary vs. secondary neoplasia-associated chromosomal abnormalities - balanced rearrangements vs. genomic imbalances? Genes Chromosomes Cancer *16*, 155-163.

Johnsen, S. A., Güngör, C., Prenzel, T., Riethdorf, S., Riethdorf, L., Taniguchi-Ishigaki, N., Rau, T., Tursun, B., Furlow, J. D., Sauter, G., et al. (2009). Regulation of Estrogen-Dependent Transcription by the LIM Cofactors CLIM and RLIM in Breast Cancer. Cancer Res *69*, 128-136.

Jones, D. T. W., Kocialkowski, S., Liu, L., Pearson, D. M., Ichimura, K., and Collins, V. P. (2009). Oncogenic RAF1 rearrangement and a novel BRAF mutation as alternatives to KIAA1549:BRAF fusion in activating the MAPK pathway in pilocytic astrocytoma. Oncogene *28*, 2119-2123.

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

Jones, S., Chen, W., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M.

A., Siegel, C., Velculescu, V. E., et al. (2008b). Comparative lesion sequencing provides insights into tumor evolution. Proc Natl Acad Sci U S A *105*, 4283.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., et al. (2008c). Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. Science *321*, 1801-1806.

Jones, S., Chen, W., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M. A., Siegel, C., Velculescu, V. E., et al. (2008d). Comparative lesion sequencing provides insights into tumor evolution. Proc. Natl. Acad. Sci. U.S.A *105*, 4283-4288.

Jonkers, J., and Berns, A. (2004). Oncogene addiction Sometimes a temporary slavery. Cancer Cell *6*, 535–538.

Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J. A. F., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. Nat Cell Biol. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20818387 [Accessed September 19, 2010].

Keydar, I., Chen, L., Karby, S., Weiss, F. R., Delarea, J., Radu, M., Chaitcik, S., and Brenner, H. J. (1979). Establishment and characterization of a cell line of human breast carcinoma origin. Eur J Cancer *15*, 659-670.

Khaitan, D., Sankpal, U. T., Weksler, B., Meister, E. A., Romero, I. A., Couraud, P., and Ningaraj, N. S. (2009). Role of KCNMA1 gene in breast cancer invasion and metastasis to brain. BMC Cancer *9*, 258.

Kim, C. J., Cho, Y. G., Jeong, S. W., Kim, Y. S., Kim, S. Y., Nam, S. W., Lee, S. H., Yoo, N. J., Lee, J. Y., and Park, W. S. (2004). Altered expression of KCNK9 in colorectal cancers. APMIS *112*, 588-594.

Kino, K., and Sugiyama, H. (2001). Possible cause of G-C-->C-G transversion mutation by guanine oxidation product, imidazolone. Chem. Biol *8*, 369-378.

Kinzler, K. W., Nilbert, M. C., Su, L. K., Vogelstein, B., Bryan, T. M., Levy, D. B., Smith, K. J., Preisinger, A. C., Hedge, P., and McKechnie, D. (1991). Identification of FAP locus genes from chromosome 5q21. Science *253*, 661-665.

Klaus, A., and Birchmeier, W. (2008). Wnt signalling and its impact on development and cancer. Nat Rev Cancer *8*, 387-398.

Knudson, A. G. (1993). Antioncogenes and human cancer. Proc. Natl. Acad. Sci. U.S.A *90*, 10914-10921.

Knudson, A. G., Meadows, A. T., Nichols, W. W., and Hill, R. (1976). Chromosomal deletion and retinoblastoma. N. Engl. J. Med *295*, 1120-1123.

Komarova, N. L., Lengauer, C., Vogelstein, B., and Nowak, M. A. (2002). Dynamics of genetic instability in sporadic and familial colorectal cancer. Cancer Biol. Ther *1*, 685-692.

Korbie, D. J., and Mattick, J. S. (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc *3*, 1452-1456.

Krivtsov, A. V., and Armstrong, S. A. (2007). MLL translocations, histone modifications and leukaemia stem-cell development. Nat. Rev. Cancer *7*, 823-833.

Krivtsov, A. V., Twomey, D., Feng, Z., Stubbs, M. C., Wang, Y., Faber, J., Levine, J. E., Wang, J., Hahn, W. C., Gilliland, D. G., et al. (2006). Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. Nature *442*, 818-822.

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.

Kunzelmann, K. (2005). Ion channels and cancer. J. Membr. Biol *205*, 159-173.

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.

Kytölä, S., Rummukainen, J., Nordgren, A., Karhu, R., Farnebo, F., Isola, J., and Larsson, C. (2000). Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. Genes Chromosomes Cancer *28*, 308-317.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lasfargues, E. Y., Coutinho, W. G., and Redfield, E. S. (1978). Isolation of two human tumor epithelial cell lines from solid breast carcinomas. J. Natl. Cancer Inst *61*, 967-978.

Leary, R. J., Kinde, I., Diehl, F., Schmidt, K., Clouser, C., Duncan, C., Antipova, A., Lee, C., McKernan, K., De La Vega, F. M., et al. (2010). Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. Science Trans Med *2*, 20ra14-20ra14.

Lee, M., Hook, B., Pan, G., Kershner, A. M., Merritt, C., Seydoux, G., Thomson, J. A., Wickens, M., and Kimble, J. (2007). Conserved regulation of MAP kinase expression by PUF RNA-binding proteins. PLoS Genet *3*, e233.

Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K. P., Bhatt, D., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473-477.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature *396*, 643-649.

Leprêtre, F., Villenet, C., Quief, S., Nibourel, O., Jacquemin, C., Troussard, X., Jardin, F., Gibson, F., Kerckaert, J. P., Roumier, C., et al. (2010). Waved aCGH: to smooth or not to smooth. Nucleic Acids Res *38*, e94.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. Nucleic Acids Res *37*, D229-232.

Levitt, N. C., and Hickson, I. D. (2002). Caretaker tumour suppressor genes that defend genome integrity. Trends Mol Med *8*, 179-186.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66-72.

Li, F. P., and Fraumeni, J. F. (1969). Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? Ann. Intern. Med *71*, 747-752.

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

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.

Liaw, D., Marsh, D. J., Li, J., Dahia, P. L. M., Wang, S. I., Zheng, Z., Bose, S., Call, K. M., Tsou, H. C., Peacoke, M., et al. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet *16*, 64-67.

Liu, H., Han, H., Li, J., and Wong, L. (2005). DNAFSMiner: a web-based software toolbox to recognize two types of functional sites in DNA sequences. Bioinformatics *21*, 671 -673.

Liu, T. X., Becker, M. W., Jelinek, J., Wu, W., Deng, M., Mikhalkevich, N., Hsu, K., Bloomfield, C. D., Stone, R. M., DeAngelo, D. J., et al. (2007). Chromosome 5q deletion and epigenetic suppression of the gene encoding [alpha]-catenin (CTNNA1) in myeloid cell transformation. Nat Med *13*, 78-83.

Liu, X., Yu, H., Yang, W., Zhou, X., Lu, H., and Shi, D. (2010). Mutations of NFKBIA in biopsy specimens from Hodgkin lymphoma. Cancer Genet. Cytogenet *197*, 152-157.

Liu, X., Chang, Y., Reinhart, P. H., Sontheimer, H., and Chang, Y. (2002). Cloning and characterization of glioma BK, a novel BK channel isoform highly expressed in human glioma cells. J. Neurosci *22*, 1840-1849.

Lundin, C., and Mertens, F. (1998). Cytogenetics of benign breast lesions. Breast Cancer Res. Treat *51*, 1-15.

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.

Mann, S. M., Burkin, D. J., Grin, D. K., and Ferguson-Smith, M. A. (1997). A fast, novel approach for DNA fibre-fluorescence in situ hybridization analysis. Chromosome Res *5*, 145-147.

Mardis, E. R. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. *361*, 1058-1066.

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med *361*, 1058-1066.

Marioni, J. C., Thorne, N. P., and Tavaré, S. (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics *22*, 1144-1146.

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.

Mathias, N., Bayés, M., and Tyler-Smith, C. (1994). Highly informative compound haplotypes for the human Y chromosome. Hum. Mol. Genet *3*, 115-123.

Mavaddat, N., Antoniou, A. C., Easton, D. F., and Garcia-Closas, M. (2010). Genetic susceptibility to breast cancer. Mol Oncol *4*, 174-191.

McBride, D. J., Orpana, A. K., Sotiriou, C., Joensuu, H., Stephens, P. J., Mudie, L. J., Hämäläinen, E., Stebbings, L. A., Andersson, L. C., Flanagan, A. M., et al. (2010). Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes Chromosomes Cancer. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20725990 [Accessed September 4, 2010].

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.

McGuire, W. L., and DeLaGarza, M. (1973). Improved sensitivity in the measurement of estrogen receptor in human breast cancer. J Clin Endocrinol Metab *37*, 986-989.

METABRIC consortium (2010). METABRIC - Molecular Taxonomy of Breast Cancer International Consortium Generation of a robust molecular taxonomy of clinical annotated breast cancer. Available at: http://science.cancerresearchuk.org/research/who-and-what-we-fund/browse-by-location/cambridge/cambridge-research-institute/grants/carlos-caldas-7199-metabric---molecular-taxonomy-of ; http://molonc.bccrc.ca/ [Accessed December 27, 2010].

Métais, J., and Dunbar, C. E. (2008). The MDS1-EVI1 gene complex as a retrovirus integration site: impact on behavior of hematopoietic cells and implications for gene therapy. Mol. Ther *16*, 439-449.

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet *11*, 685-696.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66-71.

Mitelman, F. (2000). Recurrent chromosome aberrations in cancer. Mutat. Res *462*, 247-253.

Mitelman, F., Johansson, B., Mandahl, N., and Mertens, F. (1997). Clinical significance of cytogenetic findings in solid tumors. Cancer Genet. Cytogenet *95*, 1-8.

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

Mochida, G. H., Mahajnah, M., Hill, A. D., Basel-Vanagaite, L., Gleason, D., Hill, R. S., Bodell, A., Crosier, M., Straussberg, R., and Walsh, C. A. (2009). A truncating mutation of TRAPPC9 is associated with autosomal-recessive intellectual disability and postnatal microcephaly. Am. J. Hum. Genet *85*, 897-902.

Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B., and Kinzler, K. W. (1997). Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science *275*, 1787-1790.

Mu, D., Chen, L., Zhang, X., See, L., Koch, C. M., Yen, C., Tong, J. J., Spiegel, L., Nguyen, K. C. Q., Servoss, A., et al. (2003). Genomic amplification and oncogenic properties of the KCNK9 potassium channel gene. Cancer Cell *3*, 297-302.

Muleris, M., and Dutrillaux, B. (1996). The accumulation and occurrence of clonal and unstable rearrangements are independent in colorectal cancer cells. Cancer Genet. Cytogenet *92*, 11-13.

Muleris, M., Salmon, R. J., and Dutrillaux, B. (1988). Existence of two distinct processes of chromosomal evolution in near-diploid colorectal tumors. Cancer Genet Cytogenet *32*, 43-50.

Muleris, M., Chalastanis, A., Meyer, N., Lae, M., Dutrillaux, B., Sastre-Garau, X., Hamelin, R., Fléjou, J., and Duval, A. (2008). Chromosomal Instability in Near-Diploid Colorectal Cancer: A Link between Numbers and Structure. PLoS ONE *3*, e1632.

Murphree, A. L., and Benedict, W. F. (1984). Retinoblastoma: clues to human oncogenesis. Science *223*, 1028-1033.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al. (2010). Inferring tumor progression from genomic heterogeneity. Genome Res *20*, 68-80.

Negrini, S., Gorgoulis, V. G., and Halazonetis, T. D. (2010). Genomic instability--an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol *11*, 220-228.

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*, 515–527.

Newman, S., and Edwards, P. (2010). High-throughput analysis of chromosome translocations and other genome rearrangements in epithelial cancers. Genome Med *2*, 19.

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

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., Koyama, K., Utsunomiya, J., Baba, S., and Hedge, P. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. Science *253*, 665-669.

Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I., Vogelstein, B., and Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. Proc. Natl. Acad. Sci. U.S.A *99*, 16226-16231.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23-28.

Nowell, P. C. (1962). The minute chromosome (Phl) in chronic granulocytic leukemia. Blut *8*, 65-66.

Nowell, P. C., and Hungerford, D. A. (1960). Chromosome studies on normal and leukemic human leukocytes. J. Natl. Cancer Inst *25*, 85-109.

Nusse, R., and Varmus, H. E. (1982). Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. Cell *31*, 99-109.

Ochman, H., Gerber, A. S., and Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. Genetics *120*, 621-623.

Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet *17*, 520-527.

Osborne, C. S., Chakalova, L., Mitchell, J. A., Horton, A., Wood, A. L., Bolland, D. J., Corcoran, A. E., and Fraser, P. (2007). Myc Dynamically and Preferentially Relocates to a Transcription Factory Occupied by Igh. Plos Biol *5*, e192.

Osborne, J., Lake, A., Alexander, F. E., Taylor, G. M., and Jarrett, R. F. (2005). Germline mutations and polymorphisms in the NFKBIA gene in Hodgkin lymphoma. Int. J. Cancer *116*, 646-651.

Ottesen, G. L., Christensen, I. J., Larsen, J. K., Larsen, J., Baldetorp, B., Linden, T., Hansen, B., and Andersen, J. (2000). Carcinoma in situ of the breast: correlation of histopathology to immunohistochemical markers and DNA ploidy. Breast Cancer Res Treat *60*, 219–226.

Ouadid-Ahidouch, H., and Ahidouch, A. (2008). K+ channel expression in human breast cancer cells: involvement in cell cycle regulation and carcinogenesis. J. Membr. Biol *221*, 1-6.

Ouadid-Ahidouch, H., Roudbaraki, M., Delcourt, P., Ahidouch, A., Joury, N., and Prevarskaya, N. (2004). Functional and molecular identification of intermediate-conductance Ca(2+)-activated K(+) channels in breast cancer cells: association with cell cycle progression. Am. J. Physiol., Cell Physiol *287*, C125-134.

Papatheodorou, I., Crichton, C., Morris, L., Maccallum, P., Davies, J., Brenton, J. D., and Caldas, C. (2009). A metadata approach for clinical data management in translational genomics studies in breast cancer. BMC Med Genomics *2*, 66.

Parada, L. F., Tabin, C. J., Shih, C., and Weinberg, R. A. (1982). Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. Nature *297*, 474-478.

Paterson, A. L., Pole, J. C. M., Blood, K. A., Garcia, M. J., Cooke, S. L., Teschendorff, A. E., Wang, Y., Chin, S., 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 Chromosom. Cancer *46*, 427-439.

Patwardhan, G. A., and Liu, Y. (2010). Sphingolipids and expression regulation of genes in cancer. Prog Lipid Res. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20970453 [Accessed October 31, 2010].

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res *29*, e45.

Pico, A. R. (2003). The RCK domain model of calcium activation in BK channels PhD Thesis. (The Rockefeller University, New York).

Pierotti, M. A., Santoro, M., Jenkins, R. B., Sozzi, G., Bongarzone, I., Grieco, M., Monzini, N., Miozzo, M., Herrmann, M. A., and Fusco, A. (1992). Characterization of an inversion on the long arm of chromosome 10 juxtaposing D10S170 and RET and creating the oncogenic sequence RET/PTC. Proc. Natl. Acad. Sci. U.S.A *89*, 1616-1620.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M., Ordóñez, 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.

Polakis, P. (1997). The adenomatous polyposis coli (APC) tumor suppressor. Biochim. Biophys. Acta *1332*, F127-147.

Polakis, P. (2000). Wnt signaling and cancer. Genes Dev *14*, 1837-1851.

Pole, J., 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.

Pole, J. C. M., Chin, S., Ellis, I. O., Caldas, C., and Edwards, P. A. W. (2008). High-resolution array CGH clarifies events occurring on 8p in carcinogenesis. BMC Cancer *8*, 288.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc. Natl. Acad. Sci. U.S.A *99*, 12963-12968.

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 Meth *5*, 1005-10.

R Foundation for Statistical Computing (2010). R: A language and environment for statistical computing. Available at: http://cran.r-project.org/ [Accessed October 21, 2010].

Rajagopalan, H., and Lengauer, C. (2004). CIN-ful cancers. Cancer Chemother. Pharmacol *54 Suppl 1*, S65-68.

Rajagopalan, H., Nowak, M. A., Vogelstein, B., and Lengauer, C. (2003). The significance of unstable chromosomes in colorectal cancer. Nat. Rev. Cancer *3*, 695-701.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and

integrated data-mining platform. Neoplasia *6*, 1-6.

Ripperger, T., Gadzicki, D., Meindl, A., and Schlegelberger, B. (2009). Breast cancer susceptibility: current knowledge and implications for genetic counselling. Eur. J. Hum. Genet *17*, 722-731.

Roche-Lestienne, C., Soenen-Cornu, V., Grardel-Duflos, N., Laï, J., Philippe, N., Facon, T., Fenaux, P., and Preudhomme, C. (2002). Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. Blood *100*, 1014-1018.

Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A., and Misteli, T. (2003). Spatial proximity of translocation-prone gene loci in human lymphomas. Nat. Genet *34*, 287-291.

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.

Roschke, A. V., Tonon, G., Gehlhaus, K. S., McTyre, N., Bussey, K. J., Lababidi, S., Scudiero, D. A., Weinstein, J. N., and Kirsch, I. R. (2003). Karyotypic Complexity of the NCI-60 Drug-Screening Panel. Cancer Res *63*, 8634-8647.

Rowley, J. D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290-293.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol *132*, 365-386.

Sadvakassova, G., Dobocan, M. C., Difalco, M. R., and Congote, L. F. (2009). Regulator of differentiation 1 (ROD1) binds to the amphipathic C-terminal peptide of thrombospondin-4 and is involved in its mitogenic activity. J. Cell. Physiol *220*, 672-679.

Satoh, S., Daigo, Y., Furukawa, Y., Kato, T., Miwa, N., Nishiwaki, T., Kawasoe, T., Ishiguro, H., Fujita, M., Tokino, T., et al. (2000). AXIN1 mutations in hepatocellular carcinomas, and growth suppression in cancer cells by virus-mediated transfer of AXIN1. Nat. Genet *24*, 245-250.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., et al. (2010). FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol *11*, R104.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS *95*, 5857 -5864.

Schwarzacher, H. G., and Schnedl, W. (1965). Endoreduplication in Human Fibroblast Cultures. Cytogenetics *87*, 1-18.

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.

Sharma, R. P., and Chopra, V. L. (1976). Effect of the Wingless (wg1) mutation on wing and haltere

development in Drosophila melanogaster. Dev. Biol *48*, 461-465.

Shay, J. W., and Roninson, I. B. (2004). Hallmarks of senescence in carcinogenesis and cancer therapy. Oncogene *23*, 2919-2933.

Sieber, O. M., Heinimann, K., and Tomlinson, I. P. (2002). Genomic instability—the engine of tumorigenesis? J. Gastroenterol *37*, 153–163.

Simon, W. E., Hänsel, M., Dietel, M., Matthiesen, L., Albrecht, M., and Hölzel, F. (1984). Alteration of steroid hormone sensitivity during the cultivation of human mammary carcinoma cells. In Vitro *20*, 157-166.

Simpson, J. F., Quan, D. E., Ho, J. P., and Slovak, M. L. (1996). Genetic heterogeneity of primary and metastatic breast carcinoma defined by fluorescence in situ hybridization. Am. J. Pathol *149*, 751-758.

Sinclair, P. B., Nacheva, E. P., Leversha, M., Telford, N., Chang, J., Reid, A., Bench, A., Champion, K., Huntly, B., and Green, A. R. (2000). Large deletions at the t(9;22) breakpoint are common and may identify a poor-prognosis subgroup of patients with chronic myeloid leukemia. Blood *95*, 738-743.

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.

Skoulidis, F., Cassidy, L. D., Pisupati, V., Jonasson, J. G., Bjarnason, H., Eyfjord, J. E., Karreth, F. A., Lim, M., Barber, L. M., Clatworthy, S. A., et al. (2010). Germline Brca2 Heterozygosity Promotes Kras(G12D) -Driven Carcinogenesis in a Murine Model of Familial Pancreatic Cancer. Cancer Cell. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21056012 [Accessed November 12, 2010].

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.

Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., Levin, W. J., Stuart, S. G., Udove, J., and Ullrich, A. (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science *244*, 707-712.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N. Engl. J. Med *344*, 783-792.

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.

Sotiriou, C., Neo, S., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc. Natl. Acad. Sci. U.S.A *100*, 10393-10398.

Spassov, D. S., and Jurecic, R. (2003). The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? IUBMB Life *55*, 359-366.

Speicher, M. R., Gwyn Ballard, S., and Ward, D. C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat. Genet *12*, 368-375.

Speicher, M. R., and Carter, N. P. (2005). The new cytogenetics: blurring the boundaries with molecular biology. Nat. Rev. Genet *6*, 782-792.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C., et al. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet *37*, 590-592.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., et al. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell *144*, 27-40.

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. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005-1010.

Stingl, J. (2009). Detection and analysis of mammary gland stem cells. J. Pathol *217*, 229-241.

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

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. Nature *458*, 719-724.

Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. R., and Chang, E. H. (1982). Mechanism of activation of a human oncogene. Nature *300*, 143-149.

Taylor, A. M. (2001). Chromosome instability syndromes. Best Pract Res Clin Haematol *14*, 631-644.

Teixeira, M. R., Pandis, N., Bardi, G., Andersen, J. A., Mitelman, F., and Heim, S. (1995). Clonal heterogeneity in breast cancer: karyotypic comparisons of multiple intra- and extra-tumorous samples from 3 patients. Int. J. Cancer *63*, 63-68.

Teixeira, M. R. (2006). Recurrent fusion oncogenes in carcinomas. Crit Rev Oncog *12*, 257-271.

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

Telenius, H., Pelmear, A. H., Tunnacliffe, A., Carter, N. P., Behmel, A., Ferguson-Smith, M. A., Nordenskjöld, M., Pfragner, R., and Ponder, B. A. (1992). Cytogenetic analysis by

chromosome painting using DOP-PCR amplified flow-sorted chromosomes. Genes Chromosomes Cancer *4*, 257-263.

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

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K., et al. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat. Genet *41*, 579-584.

Tognon, C., Garnett, M., Kenward, E., Kay, R., Morrison, K., and Sorensen, P. H. (2001). The chimeric protein tyrosine kinase ETV6-NTRK3 requires both Ras-Erk1/2 and PI3-kinase-Akt signaling for fibroblast transformation. Cancer Res *61*, 8909-8916.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X., 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.

Tomlinson, I. P., Novelli, M. R., and Bodmer, W. F. (1996). The mutation rate and cancer. Proc. Natl. Acad. Sci. U.S.A *93*, 14800-14803.

Tomlinson, I., Sasieni, P., and Bodmer, W. (2002). How many mutations in a cancer? Am. J. Pathol *160*, 755-758.

Tsai, M., Cheng, C., Lin, Y., Chen, C., Wu, A., Wu, M., Hsu, C., and Yang, R. (2009). Isolation and characterization of a secreted, cell-surface glycoprotein SCUBE2 from humans. Biochem. J *422*, 119-128.

Tsunematsu, T., Yamauchi, E., Shibata, H., Maki, M., Ohta, T., and Konishi, H. (2010). Distinct functions of human MVB12A and MVB12B in the ESCRT-I dependent on their posttranslational modifications. Biochem. Biophys. Res. Commun *399*, 232-237.

Turke, A. B., Zejnullahu, K., Wu, Y., Song, Y., Dias-Santagata, D., Lifshits, E., Toschi, L., Rogers, A., Mok, T., Sequist, L., et al. (2010). Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. Cancer Cell *17*, 77-88.

Valverde, M. A., Rojas, P., Amigo, J., Cosmelli, D., Orio, P., Bahamonde, M. I., Mann, G. E., Vergara, C., and Latorre, R. (1999). Acute activation of Maxi-K channels (hSlo) by estradiol binding to the beta subunit. Science *285*, 1929-1931.

Varela, I., Klijn, C., Stephens, P., Mudie, L. J., Stebbings, L., Galappaththige, D., van Gulden, H., Schut, E., Klarenbeek, S., Campbell, P. J., et al. (2010). Somatic structural rearrangements in genetically engineered mouse mammary tumors. Genome Biol *11*, R100.

Vargo-Gogola, T., and Rosen, J. M. (2007). Modelling breast cancer: one size does not fit all. Nat. Rev. Cancer *7*, 659-672.

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

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

Venkitaraman, A. R. (2007). Chromosomal instability in cancer: causality and interdependence. Cell Cycle *6*, 2341-2343.

Ventii, K. H., Devi, N. S., Friedrich, K. L., Chernova, T. A., Tighiouart, M., Van Meir, E. G., and Wilkinson, K. D. (2008). BRCA1-Associated Protein-1 Is a Tumor Suppressor that Requires Deubiquitinating Activity and Nuclear Localization. Cancer Res *68*, 6953-6962.

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nat. Med *10*, 789-799.

Volik, S., Raphael, B. J., Huang, G., Stratton, M. R., Bignel, G., Murnane, J., Brebner, J. H., Bajsarowicz, K., Paris, P. L., Tao, Q., et al. (2006). Decoding the fine-scale structure of a breast cancer genome and transcriptome. Genome Res *16*, 394-404.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W., et al. (2003). End-sequence profiling: sequence-based analysis of aberrant genomes. Proc. Natl. Acad. Sci. U.S.A *100*, 7696-7701.

Wang, K., Li, J., Li, S., Bolund, L., and Wiuf, C. (2009). Estimation of tumor heterogeneity using CGH array data. BMC Bioinformatics *10*, 12.

Wang, X. Z., Jolicoeur, E. M., Conte, N., Chaffanet, M., Zhang, Y., Mozziconacci, M. J., Feiner, H., Birnbaum, D., Pébusque, M. J., and Ron, D. (1999). gamma-heregulin is the product of a chromosomal translocation fusing the DOC4 and HGL/NRG1 genes in the MDA-MB-175 breast cancer cell line. Oncogene *18*, 5718-5721.

Warren, D. T., Tajsic, T., Mellad, J. A., Searles, R., Zhang, Q., and Shanahan, C. M. (2010). Novel nuclear nesprin-2 variants tether active extracellular signal-regulated MAPK1 and MAPK2 at promyelocytic leukemia protein nuclear bodies and act to regulate smooth muscle cell proliferation. J. Biol. Chem *285*, 1311-1320.

Watson, M., Barrett, A., Spence, R. A. J., and Twelves, C. (2006). Oncology 2nd ed. (OUP Oxford).

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

Weigelt, B., Geyer, F. C., and Reis-Filho, J. S. (2010b). Histological types of breast cancer: how special are they? Mol Oncol *4*, 192-208.

Weiss, B., and Stoffel, W. (1997). Human and murine serine-palmitoyl-CoA transferase--cloning, expression and characterization of the key enzyme in sphingolipid synthesis. Eur. J. Biochem *249*, 239-247.

Weiss, L., Holmes, J. C., and Ward, P. M. (1983). Do metastases arise from pre-existing subpopulations of cancer cells? Br. J. Cancer *47*, 81-89.

Wheeler, M. A., Davies, J. D., Zhang, Q., Emerson, L. J., Hunt, J., Shanahan, C. M., and Ellis, J. A. (2007). Distinct functional domains in nesprin-1 alpha and nesprin-2beta bind directly to emerin and both interactions are disrupted in X-linked Emery–Dreifuss muscular dystrophy. Exp Cell Res *313*, 2845–2857.

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.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789-792.

Wright, J. B., Brown, S. J., and Cole, M. D. (2010). Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol. Cell. Biol *30*, 1411-1420.

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.

Yuan, P., Leonetti, M. D., Pico, A. R., Hsiung, Y., and MacKinnon, R. (2010). Structure of the human BK channel Ca2+-activation apparatus at 3.0 A resolution. Science *329*, 182-186.

Zhang, Q., Bethmann, C., Worth, N. F., Davies, J. D., Wasner, C., Feuer, A., Ragnauth, C. D., Yi, Q., Mellad, J. A., Warren, D. T., et al. (2007). Nesprin-1 and-2 are involved in the pathogenesis of Emery Dreifuss muscular dystrophy and are critical for nuclear envelope integrity. Hum Mol Genet *16*, 2816.

Zhao, Q., Caballero, O. L., Levy, S., Stevenson, B. J., Iseli, C., de Souza, S. J., Galante, P. A., Busam, D., Leversha, M. A., Chadalavada, K., et al. (2009). Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc. Natl. Acad. Sci. U.S.A *106*, 1886-1891.

Appendix 1. PCR Primers

A.1.1. PCR Primers to confirm fusion gene expression in HCC1187

SGK1_008_1F	CCTTTGCCTCCTGACATGAT
SGK1_008_2F	CCCACCTCAACTACCAGCAT
SGK1_001_1F	AGAAATGCTCAGCCTTCCAA
SGK1_001_2F	TCAGAGTCCCAGCCTGAAGT
SLC2A12_001_3R	ACCAGGAAAGATCTCGCTGA
SLC2A12_001_4R	ACAACAAAAAGCAGGGATGC
SLC2A12_001_5R	GCTCCTGGGGTTTTCTTTTT
RGS22_210_19F	AACAGTTTGCAGCACGTCAG
RGS22_210_20F	AAAAGAAAATCGGGGTCTGG
RGS22_210_21F	TCGCAAAGCATTATTGAATCC
RGS22_210_22F	CCGGAAGGAGTTAGGACCAT
SYCP1_001_24R	CCAAATTTCTGTTTGGCACTTT
SYCP1_001_25R	CCATAAGTGCTACCATTTCAGC
SYCP1_001_26R	TGCTCTCAGTGATGACTGTTCTT
SYCP1_001_27R	CAAAAGTTCAGCTTTGAGATTGG
AGPAT5_201_2F	TTTAGCAAATCATCAAAGCACA
AGPAT5_201_3F	GCTGCCATTGTATGGGTGTT
AGPAT5_201_4F	GATGCGAAACAAGTTGCAGA
AGPAT5_201_5F	TCAGCTAGTCAGGCATTTGCT
MCPH1_201_14Ri	ACGCCAGTTCCTTCTCTTCA
MCPH1_201_14Rii	CGCCAGTTCCTTCTCTTCAC
MCPH1_201_14Riii	CCACAACACATGGCAAACAT
SUSD1_201_3F	AACCACACATCTTGCCACAA
SUSD1_201_4F	TGTGAAGTTTCTGGCCTGTG
SUSD1_201_5F	TGGGAGTCCCCAAAATTACA
SUSD1_201_6F	GTGGCTCGCTATGTCTGTCA
ROD1_004_2R	GGTCCGTTAATGATGCCAGA
ROD1_004_3R	AGGCGAACAGGGAGGTCTAT
ROD1_004_4R	AGTAACGGCAGCTTCCTCAG
ROD1_004_5R	CATTGGAAGGACCTCCAGAA
PLXND1_201_10F	TGCAGCCCTGGAGTGTAGTTT
PLXND1_201_11F	GATTCCTGGACAGCCCTGAG
PLXND1_201_12F	CACCTGTGCATGTGGAGTGA
PLXND1_201_13F	TGAGCCACTGCCTGACAGAT
TMCC1_001_4R	GTTTGGGCAAGGCTGCTTAC
TMCC1_001_4Rii	TGCACTGCCAAATTTGTTCC
TMCC1_001_5R	ATGCTTGCCAGTTCCTGCTT
TMCC1_001_6R	AGCCCTTCTAGCTGCACCAC
RHOJ-1F	GCAAAGAGGGAACTGACAGCA
SYNE2-2R	GTTCATCTTCGGTGGGAAGC
SYNE2-5R	TCGGTTTCTTAGGAATGTCAAGG
PUM1-19F	CTTATGGCTGCCGAGTGATT
TRERF1-7R	GATGGAGCATGTCAGCTTGTT
CTCF-1S	CTGAGCCTGTGGAGCGATT
CTCF-2S	ACCTGAAGCCAAAGAACAAGA
SCUBE2-4R	CACTCATAGGCCCCATGAC

A.1.2. PCR Primers to confirm fusion gene expression in VP229 and VP267

TRAPPC9_203_Ex6F	TCGGACGTGCTAAGAACTGC
TRAPPC9_203_Ex7R	ATGCTTCCATGCTCCGTTTC
TRAPPC9_203_Ex8F	GTGGAGGGCCTGCTACAAAC
TRAPPC9_203_Ex9R	GGGAGGCGTAGACCAATTCA

KCNK9_201_Ex1F	CCGGATCAAGGGGAAGTACA
KCNK9_201_Ex2R	CGGCGTAGAACATGCAGAAG
KCNK9_201_Ex2F	GCTTTACCGACCACCAGAGG
KCNK9_201_Ex3R	GCACAGCAGAGGTCCACTTG
PDLIM1_001_1F	CCATGACCACCCAGCAGATA
PDLIM1_001_2R	TTTTCCCCATCAATGGCTGT
PDLIM1_002_1F	GGTGCCCTGCAAGCTGTT
PDLIM1_002_2R	GGATGACGCTTCCCTTCCT
ZBBX_202_14F	GGCCAAGCTTTGAAGAATCA
ZBBX_202_15R	TGGAATTGTTGCATTCTAAACGA
ZBBX_202_15F	TCGTTTAGAATGCAACAATTCCA
ZBBX_202_16R	GCAGCTGCACTTCTTGATCG
FAM128B_003_2F	GACCAGTATGGCGTGGCTCT
FAM128B_003_3R	TTCTTGCTGTCCTCCTGGTG
FAM128B_003_4F	CTCACGTCCGTGGACCAATA
FAM128B_003_5R	TGCTCGGCTTCTGTCATCAT
MDS1_Ex1F	GGCACAGCATGAGATCCAAA
MDS1_Ex2R	GGGAGGCCTCAGGAAACTT
MDS1_Ex2F	AGGTCCTTGTTTCCCCTTCG
MDS1_Ex3R	AGGGAGGGAGTGCTGGCTAC
KCNMA1_Ex23F	GGGTCAACATCCCCATCATC
KCNMA1_Ex24R	GCGCTCATGAGTGAGTCCAG
KCNMA1_Ex24F	ACAGCATTTGCCGTCAGTGT
KCNMA1_Ex25R	GGTCCCTATTGGCCAGTGTC

A.1.3. PCR Primers for resequencing of HCC1187 somatic mutations

Primers designs were taken from Sjoblom et al. (2006) and Wood et al. (2007)

APOC4 F	CAGAGAGGAGCGGATAAATGG
10ITIH5L /r	GCCAGGATTCAGACTCAAGAAG
10ITIH5L/f	GTTCCCAAAGACTAGCCCATC
11PLS3 /r	GAAGCATCTCCCACTTAACATCC
11PLS3/f	CTTGACGCTGAGCTTCTTGAG
12SATL1 /r	TACCTGATTGGTTTGTGCCTG
12SATL1/f	ACTGAGCCAACCAGTCCTGAG
13ITR /r	AGGTCCAGGTGAGGCTGG
13ITR/f	AATGTCTTCGGCTTATGGCAG
14PEBP4 /r	TCCTTCCAGTTCACTCCCAAG
14PEBP4/f	AGATCCCGCCAAATACAAATC
15OR1S1 /r	CTCAGGGCACCTTTCATATCC
15OR1S1/f	CATTGACAATTTGCTCTTGGG
16IPO7 /r	CTTTAAAGGGCAGGCAGAAAC
16IPO7/f	TTTAAACACCAATTCCTGGAGC
17ZCSL3 /r	TCCAGACAGAAACACTTACCACAT
17ZCSL3/f	TCCATGTTGCATGATTGTGAA
18ZNHIT2 /r	TGGGACAATAGCTATCCCTCAG
18ZNHIT2/f	GACTTCTGTGCCACACTGCTC
19TAS2R13 /r	AGGCATTTGTATGGACCTTGG
19TAS2R13/f	CAATCTCCAGAATTGGGCTG
1PRKAA2 /r	GCCATAATGTCATACGGTTTGC
1PRKAA2/f	CTAACGTCATTGATGATGAGGC
20NUP98 /r	AAGATGCCTGCTTTGACAGTG
20NUP98/f	TTCCTTTCTGTCTTCCTGCC
21KIAA0427 /r	AGAACTCGCAGCCTTCTCG

21KIAA0427/f	AGTAGGCCTGGTCCTGCTTTC
22STATIP1 /r	ATTACCAAACTTGCAGGAGCC
22STATIP1/f	CCCATCTCCCTGTCTTTCTCTC
23FLJ21839 /r	TTACAAGCATGCACCACCAC
23FLJ21839/f	CAAACTGAGAATTCAGGCAGC
24ADRA1A /r	GCTAATCCTTCCTCTTCCATCC
24ADRA1A/f	GTCATGTACTGCCGCGTCTAC
25RBAF600 /r	CCAAAGTGCAGACACCTAACC
25RBAF600/f	ATTGCCCTCAAGGTCAAACAG
26SPEN /r	CTCCCTCAGATGTCTGCTTCC
26SPEN/f	TGAACACAAATCCCACTCACC
27PDCD6 /r	GTATCCATGTCGAGGATCGC
27PDCD6/f	CTGGTTAGTGGTCAGCAGTGG
28FLJ32363 /r	TTTCTGACAACACTGCAGATGAG
28FLJ32363/f	AGGCTGAGATCTGGCCCTC
29PLA2G4A /r	TGCAACATGCAATCCTCTCTC
29PLA2G4A/f	GAGCTAGTAGATCTCACTCTGTGGTTT
2MLL4 /r	TGAATAGCTGCACTTTGGCTC
2MLL4/f	ACGAATCGGTGCTTACTCCTC
30SPTA1 /r	AAGCAATAAAGCTGCCAGGTG
30SPTA1/f	CTGCCTTCACAGCCTTCCTAC
31ABCB8 /r	GCTTTATTGTGAGCAGGAGCAG
31ABCB8/f	TCCTTTCAGATTGGGCATTG
32TBXAS1 /r	CAGAGGCCTGATTGCATCC
32TBXAS1/f	GCTGGTGGAAAGAAATTTGATG
33ZNF674 /r	AATGCCAAGCAAAGAACACAG
33ZNF674/f	CCCTTGCTTGCTATGTGAATG
34LHCGR /r	TGCATACAGAAATGGATTGGC
34LHCGR/f	CATAGACTGGCAGACAGGGAG
35SULT6B1 /r	CCCAGATCTACCTAAATCTTCTTCC
35SULT6B1/f	GCCACCTCCTGTTCCTCAG
36KIAA0934 /r	GGAGCCAGTCTGTGCTGTACC
36KIAA0934/f	TCTGTGGGAAGGAGTCTCTGG
37WARS /r	AATTACCCACAATGCTTTGCC
37WARS/f	GTGGTGTGGGCTGAGCTG
38PDPR /r	GCCTGAGAGCCTCTGCTACA
38PDPR/f	TCAATAGCTAGCGTTCCCTGG
39AMPD2 /r	GAATGGAGACATGCAGAGACC
39AMPD2/f	CTCGCACAAGGTACTACAGCG
3MYBPC2 /r	AAGATGGGCCAGAGGTGG
3MYBPC2/f	CTAAGATTCATGGCCTCGGAC
40C4orf14 /r	GGCTTTCTTAACAGAGCTATGGG
40C4orf14/f	GCACTGCTTCAGTATTGTGCC
41BAP1 /r	ATCAGAGACAAATGCTGTGGG
41BAP1/f	GTATGGCTAGTCGCTGCCTG
42RNU3IP2 /r	AGAGAGGGACGAAGCTGACTC
42RNU3IP2/f	GCGGTGGAAGCCTATCAATAC
43RTP1 /r	CAACCACGGAATCTTATACGG
43RTP1/f	CATTCCTAGGGCTGTTTCCAC
44GOLPH4 /r	TCATCTTTCTGAGGTATGCGAA
44GOLPH4/f	GACATGGAAATCAAGAAATTACATC
46CENTD3 /r	CTGCAGTGGTCTCTCCTTTCC
46CENTD3/f	GCACAGGGTTGTTCCAGAATC
47CTNNA1 /r	AGAAGATTGTCCACAAAGCCC
47CTNNA1/f	TATGCCACAGATTGCCTGTTG
48GMCL1L /r	TGGTTCAGTTTCAAGAAAGGC
48GMCL1L/f	ATGTAGACGCACTGCAGGTTG
49PCDHB15 /r	AACAGACGGTCGGAAAGCTAC

49PCDHB15/f	GTCGTACCAGCTGCTCAAGG
4PLCB1 /r	AACCCACAGACCTATGAGCAC
4PLCB1/f	CTGAGACAGGAGAATGGCTTG
50ARHGEF4 /r	TGCTAATGCTCAAACTGGCTG
50ARHGEF4/f	GAAATGTAGAGGCTCTTGCCC
51UGT1A9 /r	GGGCTGATTAATTTATGCAAAG
51UGT1A9/f	AGCAGCCATGAGCATAAAGAG
52ZNF142 /r	CAACGGCTCTGTTTACAGGTG
52ZNF142/f	TAAGAAGCACCACCTTGACCC
53DDX18 /r	CCAGGCTAATCTCTAACATACTGGTG
53DDX18/f	CAGGAAGCCATGGAAGTTCTC
54SCN3A /r	AACAATAAGGCACATGGTTTGG
54SCN3A/f	ACATGGGCATATCTTTGGATG
55SLC4A3 /r	CCCAACCCACTCAGTGAAGTC
55SLC4A3/f	CCTGACCCTCCACTACTCACC
56ITGB2 /r	CTGGAGTCCCTCAGACGACC
56ITGB2/f	CATCTTGCTTTCTCCACCTCC
57MYH9 /r	AGCCCAGGCTTTCTCTGATG
57MYH9/f	TTTCATAACTGGGCAGATCCC
58CYP2D6 /r	TGCCATGTATAAATGCCCTTC
58CYP2D6/f	TTTATAAGGGAAGGGTCACGC
59NCB5OR /r	GAAATGATGAATGAGGAAATGG
59NCB5OR/f	TCAGACTTCAGATGGTTTGGC
5CAMTA1 /r	ATGGGTAGATTTCTTCCACGG
5CAMTA1/f	AATCGTAAAGCATTTGTTTCCC
60PAXIP1 /r	AGGTGGAACCTGATGCTGC
60PAXIP1/f	GCAGTGCTGTTTAGCCAAGTG
61AVPI1 /r	GGCCAAGTAACTAGCTCCAGG
61AVPI1/f	CCTAGGATTAGCCAGGACCC
62GPR81 /r	GTGATGAACACAATTGCCACC
62GPR81/f	CTGTGTGGTTTCTGCTTCCAC
63FRMPD1 /r	CCAAAGTGGAAGAAGGTGGAC
63FRMPD1/f	ATAGGACGGTCTTAGGCCAGC
64SORCS1 /r	AGCATCCACCCAACAAGACTC
64SORCS1/f	GAAACCTCCTGAGAGCCATTC
65PPHLN1 /r	AATGGCTAGCCCAGATACCAG
65PPHLN1/f	AGGAAAGAGTTTAGGGCCACC
66PPP1R12A /r	GGTTCGATGAATCACAAGTTAGG
66PPP1R12A/f	GGAAATTTGAATTACTTTGGCG
67INHBE /r	CCTCCTTCTCCTGCCACAC
67INHBE/f	CCCAGCAATCAGACTCAACAG
68NFKBIA /r	TGCCTGGACTCCTTAAGTTGG
68NFKBIA/f	CCTGTCTAGGAGGAGCAGCAC
69SMG1 /r	AGGAGTTTCTCTCTTGCACCG
69SMG1/f	ACCACTACCACCTATCCCGTG
6C6orf31 /r	AGTCTCCTGCAGGTAAGGGTC
6C6orf31/f	GCAGATGGCCTAGATACAGCC
70NOS2A /r	TATGCCCTAACAGGCTCTTGC
70NOS2A/f	GTTACACGAAACACACGGCTC
71RASL10B /r	TACAGCAGTTTGGAATCCAGC
71RASL10B/f	GCACTAAGCCCACCTCTTGTC
72TP53 /r	GAGGAATCCCAAAGTTCCAAAC
72TP53/f	ACGTTCTGGTAAGGACAAGGG
73LLGL1 /r	ACCAGACCCTCCAGCTCATC
73LLGL1/f	ACTCGGGTAGCCCTGACATC
74TRIM47 /r	CTCTTCAGCACGGATATGCAG
74TRIM47/f	GAACCAAAGGTGTCAAGAGGG
7B3GALT4 /r	TCAGCAGGAATTTCCCATAGC

7B3GALT4/f	AACTGGGCTGAGAAACACTGC
8SKIV2L /r	TAGTCCAGGAGCTGGAGTTGG
8SKIV2L/f	CCTTTACTGTCATCTGCTGGG
9HUWE1 /r	TCTGCTTTACCTGCCATCTCTAC
9HUWE1/f	CAACTGTGTATCTGCTTGCAGTG
ADRA1A f2	CCTTCATGTGGCCTTCTGAG
ADRA1A r2	GTCGATGGAGATGATGCAGAG
APOC4 R	CAAGAGATCTCGCTGTGTTGC
BAP1 f2	CCACTCCAAGTCCCACCTTT
BAP1 r2	CTGCCAGGATATCTGCCTCA
C6orf21 F	GTGTACGACGTCTTGGTGCTC
C6orf21 F	TTATCCAGGCATGGTGGC
C6orf21pyro F	CTCCCCTACAGGATCCCAGTT
C6orf21pyro S	TGCAATGTCCTCCTGT
C6orf31 f2	CCACTACCATCAGTCTGGCAC
C6orf31 r2	CCTGCCCTTGTTCCTCTATCC
CD2 F	TCTGTGAGCCTGGGAGTTATG
CD2 pyro F	AGTAATGGGCTCTCTGCCTGGA
CD2 pyro S	TATCTCATCATTGGCATA
CD2 R	TGCAGATTCAAGGTGTCATCC
CYP4A22 F	TCCAATGACCCTTGGAGAATA
CYP4A22 R	AGCACCAGAGCCAGGATAGTT
FHOD3 F	TGCTGTGTGCATCAGGAAAC
FHOD3 R	TGAAATCATCACTACTGCCCTG
FLJ20422 F	GAGCCAAGGTTGTGGAAAGAG
FLJ20422 pyro F	GTGCTCACCGTCCCTCTTG
FLJ20422 pyro S	ATGCCCCGTTCCAGC
FLJ20422 R	TCCACAAACCACTGGTACTCC
FLJ32363 f2	CATTGCCAGATCATGAAATCC
FLJ32363 r2	CATAAGCAACACATTTGCTAGGG
FLNC F	CTCTGAGGGTGTTGGGTGAAC
FLNC R	GCGATGGAAAGGAGTGATGTC
GLT25D2 F	AACCAAAGCTGTGCTTCATCC
GLT25D2 R	CAGGACACTCACCATCTCTGC
GPNMB F	CACCAGTGTCTTGCAAACTGTC
GPNMB pyro F	ACACAAGGAATACAACCCAATAGA
GPNMB pyro S	GAATACAACCCAATAGAAAA
GPNMB R	TGCCTGCAGTATAATCCCTCTC
HSD17B8 F	CAAGGTGGCGATCTCTGAAC
HSD17B8 pyro R	GGGTTCTCCTCTCTATACCACTTG
HSD17B8 pyro S	CAACCTGACCTTTCCTA
HSD17B8 R	CCTTGGATGCTGCATAGTTTG
HUWE1f2	TGGAATTTATGAGGAAGAATGAAAA
HUWE1r2	GAAATCAACTGTGTATCTGCTTGC
IPO7 pyro F	TATTTGGAGATTCTGGCTAAGCA
IPO7 pyro S	GAGATTCTGGCTAAGCA
ITIH5l pyro F	ACCTAAAACACCTCCTCCTGTCT
ITIH5l pyro S	TTGACCTGGATCTGC
ITR f2	GTCGGTGCAGACCTGGAG
ITR r2	AGGTGAGGCTGGGGAAGTC
KIAA0427 pyro F	AACAGCATGCGGAACAACAG
KIAA0427 pyro S	AACAACAGCAGCGAC
MLL4 f2	GGTCACCACTCCTGTTAAGGC
MLL4 pyro R	ATCCGGGCTTTTTCCAGG
MLL4 pyro S	GCATCTGCTGTGGTGA
MLL4 r2	GGATCACAGAAAGGCAGGTTC
PEB4 pyro R	TGGAAGGATGGGGAGGTCTTA
PEB4 pyro S	GATAGACAAAGAACTGGTAG

PLCB1 f2	CCAGGTGTGTCCTTAATGTCC
PLCB1 r2	TGTTACATAACAAAATTACAAAGCAGA
PRKAA2 pyro R	TTTGGGCTTAGTCGTATTCAGTG
PRKAA2 pyro S	GCTGTCTGCTATAAGAGGTG
SATL1 f2	ACAAGTAGGCACCAGCCAATC
SATL1 r2	TGCCTCCCTTACTCTTTCAGC
TRIM47 f2	GATTTGGACAGCGACACAGC
TRIM47 r2	AGCATAGAAGGCCAAGGCAC
WARS f2	TGACTGGGCTGGGATTATTG
WARS r2	AGGAAGGAGCCACTCAGGAC
ZNHIT2 f2	ACCTGCCCTCGCTGTAATG
ZNHIT2 r2	GCATTATCCAGCTCCTCCAG

A.1.4. Real time PCR primers

GAPDH_RT_F	GCAAATTCCATGGCACCGT
GAPDH_RT_R	TCGCCCCACTTGATTTTGG
KCNK9_201_Ex1F	CCGGATCAAGGGGAAGTACA
KCNK9_201_Ex2R	CGGCGTAGAACATGCAGAAG
KCNK9_Ex2int_F	CGTTGACTACCATTGGGTTCG
KCNK9_Ex2int_R	CAGCCCCACCAGGATATACA

A.1.5. PCR Primers for Pyrosequencing

Primer Name	5' Biotin	Sequence (5' to 3')	Tm
C6orf21pyro F		CTCCCCTACAGGATCCCAGTT	70.5
C6orf21pyro R	Biotin	TCCTGCCAGGTCACAGAGTC	70.5
C6orf21pyro S		TGCAATGTCCTCCTGT	52.9
CD2 pyro F		AGTAATGGGCTCTCTGCCTGGA	73.4
CD2 pyro R	Biotin	AGAAAACGAGCAGTGCCACAAAG	73.8
CD2 pyro S		TATCTCATCATTGGCATA	49.4
FLJ20422 pyro F		GTGCTCACCGTCCCTCTTG	70.9
FLJ20422 pyro R	Biotin	GGCTATTACCCAGGGCATCC	71.6
FLJ20422 pyro S		ATGCCCCGTTCCAGC	61.2
GPNMB pyro F		ACACAAGGAATACAACCCAATAGA	67.7
GPNMB pyro R	Biotin	ACTCAGGCCTTTGCTTCTGAC	69.8
GPNMB pyro S		GAATACAACCCAATAGAAAA	51.3
HSD17B8 pyro F	Biotin	TCGTGGTTCCATCATCAACAT	70.1
HSD17B8 pyro R		GGGTTCTCCTCTCTATACCACTTG	68.4
HSD17B8 pyro S		CAACCTGACCTTTCCTA	50.7
IPO7 pyro F		TATTTGGAGATTCTGGCTAAGCA	69
IPO7 pyro R	Biotin	TTTAAAGGGCAGGCAGAAACTA	69
IPO7 pyro S		GAGATTCTGGCTAAGCA	51.2
ITIH5l pyro F		ACCTAAAACACCTCCTCCTGTCT	68.4
ITIH5l pyro R	Biotin	GAAATTGGAGATAAAGGCAAGATG	69.1
ITIH5l pyro S		TTGACCTGGATCTGC	50
KIAA0427 pyro F		AACAGCATGCGGAACAACAG	71.2
KIAA0427 pyro R	Biotin	TCGGACACAGCCTTCTGGTA	70.8
KIAA0427 pyro S		AACAACAGCAGCGAC	50.6
MLL4 pyro F	Biotin	GAGCAACGGGCCACAGAC	71.9
MLL4 pyro R		ATCCGGGCTTTTTCCAGG	71.2
MLL4 pyro S		GCATCTGCTGTGGTGA	56.3
PEB4 pyro F	Biotin	ACCGGCACACAGTGGCTT	71.8

PEB4 pyro R		TGGAAGGATGGGGAGGTCTTA	71.8
PEB4 pyro S		GATAGACAAAGAACTGGTAG	48.1
PRKAA2 pyro F	Biotin	TGATAGTGCCATGCATATTCCC	70.8
PRKAA2 pyro R		TTTGGGCTTAGTCGTATTCAGTG	69.3
PRKAA2 pyro S		GCTGTCTGCTATAAGAGGTG	54.6

A.1.6. PCR Primers to confirm somatic structural variants in VP229 and VP267

i) Predicted structural variants

Predicted in	Type of SV	Node 1 chr	Node 1 start	Node 1 end	Node 1 direction	Node 2 chr	Node 2 start	Node 2 end	Node 2 direction	Primer name
VP267	DIF	1	19501578	19501771	1	9	115804584	115804726	-1	A1
Vpboth	DIF	3	167030872	167031047	1	10	97034481	97034738	-1	A10
VPBoth	DIF	3	170442107	170442495	1	4	144355821	144356190	-1	A12
Vpboth	DIF	3	171158535	171158864	1	10	97032906	97033283	-1	A13
VP229	DEL	3	171409858	171409978	1	3	171410591	171410850	-1	A14
VPBoth	INS	5	749192	749453	-1	5	840835	840876	1	A15
VP229	DIF	5	10296160	10296260	1	7	146469598	146469634	1	A16
VPBoth	INV	6	2903166	2903329	1	6	3064030	3064149	1	A17
VPBoth	DEL	6	29856913	29856955	1	6	29896035	29896072	-1	A18
Vpboth	DIF	1	110132251	110132512	1	9	84978114	84978363	-1	A2
Vpboth	DIF	7	102818133	102818319	1	12	63957260	63957303	-1	A21
VPBoth	INV	9	84673335	84673384	-1	9	95650438	95650487	-1	A22
Vpboth	INV	9	94861458	94861760	1	9	129243323	129243573	1	A23
VP267	DIF	9	115804619	115804742	1	1	19501845	19502151	-1	A24
VPBoth	INV	9	127073730	127074038	-1	9	130287513	130287774	-1	A25
VPBoth	DEL	9	130313996	130314354	1	9	132757339	132757510	-1	A26
VP229	DEL	10	14704339	14704762	1	10	14705031	14705468	-1	A27
VP267	DIF	10	76297188	76297545	1	3	178114514	178114843	-1	A28
VPBoth	INS	1	144837271	144837307	-1	1	146474210	146474251	1	A3
VPBoth	DIF	10	77063134	77063482	-1	3	100659737	100660078	-1	A30
Vpboth	DIF	10	77069722	77070093	1	17	26873584	26873946	1	A31
VP267	DIF	10	77393012	77393230	-1	17	27124203	27124413	-1	A32
VP267	DIF	10	77414962	77415055	1	17	27117214	27117289	1	A33
VPBoth	DIF	10	78224422	78224758	1	3	178426171	178426491	1	A34
Vpboth	INV	10	78229401	78229776	-1	10	122537421	122537463	-1	A35
Vpboth	DIF	10	78679650	78680299	1	3	169181685	169182111	-1	A36
VP229	DIF	10	78679650	78680299	1	3	169181872	169182064	1	A36
VP267	DIF	10	78679650	78680299	1	3	169181872	169182064	1	A36
Vpboth	INV	10	79656068	79656493	1	10	84276990	84277205	1	A37
VPBoth	DIF	10	79656167	79656503	-1	3	178107424	178107859	-1	A38
VP229	DEL	10	80127208	80127257	1	10	80127854	80128006	-1	A39
Vpboth	DEL	2	242852936	242853243	1	2	243035841	243036157	-1	A4
VPBoth	DIF	10	84303032	84303414	1	12	66816857	66817212	1	A40
Vpboth	INV	10	93897014	93897410	1	10	102245634	102246026	1	A41
VP267	DIF	10	109724338	109724379	1	22	29065779	29065914	-1	A42

Vpboth	INV		10	124809773	124809955	1		10	127790628	127790985	1	A43
VP267	INS		11	308533	308595	-1		11	314064	314148	1	A44
VP229	DEL		11	47289854	47290156	1		11	47290578	47290834	-1	A45
VP229	DIF		11	77656124	77656163	1		7	64623948	64623999	1	A46
VP229	DEL		12	69079758	69079913	1		12	69080496	69080619	-1	A47
VP267	DIF		13	21750547	21750661	1		11	108585829	108586244	-1	A48
VP229	DIF		13	61034148	61034183	-1		14	74200278	74200314	-1	A49
VP267	DIF		3	9461581	9461652	1	X		102375851	102375932	1	A5
VPBoth	INS		14	106349889	106349928	-1		14	106359128	106359223	1	A50
VPBoth	DIF		14	106483843	106484213	1		15	22479869	22479909	-1	A51
VP229	DEL		15	52175561	52175956	1		15	52176296	52176440	-1	A52
VPBoth	INV		15	78282315	78282356	1		15	79054483	79054549	1	A53
VP267	DEL		16	611910	612133	1		16	675407	675637	-1	A54
VP267	INV		16	70152159	70152198	-1		16	74397297	74397478	-1	A55
VPBoth	INV		17	35326354	35326611	1		17	36222373	36222607	1	A56
VP267	DIF		17	37230708	37231060	-1		3	171648030	171648408	-1	A57
VP229	DEL		17	78163335	78163371	1		17	78362163	78362206	-1	A58
VP267	INS		19	11614744	11614845	-1		19	11638741	11638818	1	A59
VP267	INS		3	35662797	35662951	-1		3	56242409	56242583	1	A6
VP267	DIF		19	18953632	18953670	-1		8	68218123	68218163	-1	A60
VP229	DEL		19	37346031	37346070	1		19	41295756	41295849	-1	A61
VP229	DEL		19	39434948	39435080	1		19	39435655	39435730	-1	A62
VPBoth	INS		19	43244102	43244259	-1		19	43359087	43359402	1	A63
VP229	INS		19	43356453	43356512	-1		19	43418232	43418272	1	A64
VP229	DIF		19	52596051	52596107	1		6	157823962	157824122	1	A65
VP229	INV		19	54804916	54804957	-1		19	55105319	55105362	-1	A66
VP267	DEL		20	17506460	17506612	1		20	17598409	17598578	-1	A67
VP267	DIF		20	51857762	51857845	1		22	29065501	29065807	1	A68
VPBoth	INV		22	20329738	20329796	-1		22	20653465	20653516	-1	A69
VPBoth	DIF		3	107806125	107806388	1		4	73975049	73975377	-1	A7
VP267	DIF		22	29065304	29065537	-1		3	24817529	24817579	-1	A70
VPBoth	INV		22	36626954	36626996	1		22	36657767	36657818	1	A71
VPBoth	INS		22	39425844	39425899	-1		22	39445528	39445564	1	A72
VP229	INS	X		29329361	29329472	-1	X		31278768	31278858	1	A73
VP229	DIF	X		100133833	100133870	1		12	7240008	7240314	-1	A74
VP229	DEL	X		142600166	142600247	1	X		142799082	142799186	-1	A75
VP229	INV		3	108136284	108136419	1		3	110987235	110987406	1	A8
VP267	INV		3	108136284	108136419	1		3	110987235	110987406	1	A8
VPBoth	DIF		3	149679452	149679489	-1		7	83100972	83101067	-1	A9
VPBoth	DEL		10	116257973	116258272	1		10	116364481	116364798	-1	Del20
VP229	DEL		8	41675207	41675366	1		8	41704343	41704453	-1	Del21

VPBoth	DEL	16	27336191	27336599	1	16	27351346	27351623	-1	Del34
VPBoth	DIF	17	35329033	35329210	1	4	778121	778512	1	Fus1
VP229	INS	3	10115221	10115261	-1	3	11932819	11932869	1	Fus10
Vpboth	DIF	10	125636500	125636725	1	21	31094557	31094776	-1	Fus14
VP229	DEL	4	3156552	3156602	1	4	3437871	3437927	-1	Fus15
Vpboth	DIF	3	169496733	169497135	1	10	84276976	84277214	-1	Fus16
Vpboth	INS	10	78198382	78198678	-1	10	83671218	83671877	1	Fus17
VPBoth	INS	10	76898989	76899271	-1	10	84730469	84730979	1	Fus19
VP267	INS	10	76898989	76899271	-1	10	84730993	84731135	1	Fus19
VPBoth	DEL	11	5784201	5784559	1	11	5809301	5809685	-1	Fus20
Vpboth	INV	9	94648426	94648733	-1	9	127055261	127055508	-1	Fus21
VP267	DIF	10	62684277	62684470	-1	12	69045557	69045632	-1	Fus22
VPBoth	INV	9	94647680	94648017	1	9	128534661	128534996	1	Fus23
VP267	DIF	1	49783251	49783335	1	14	31139376	31139481	1	Fus24
VP229	INV	12	48446975	48447325	-1	12	52404710	52404905	-1	Fus25
VP267	DEL	8	140704651	140704941	1	8	141348436	141348751	-1	Fus27
VPBoth	INV	10	98803429	98803655	1	10	99294104	99294473	1	Fus28
Vpboth	DEL	10	102235324	102235661	1	10	114568184	114568344	-1	Fus29
VPBoth	DEL	7	1173820	1173899	1	7	1192727	1192916	-1	Fus30
VPBoth	DEL	7	1173342	1173753	1	7	1192615	1192690	-1	Fus30
VP229	DEL	7	1173342	1173753	1	7	1192727	1192916	-1	Fus30
Vpboth	INV	10	97789793	97790092	1	10	108715824	108716128	1	Fus5
Vpboth	DIF	12	69990188	69990524	1	9	37924987	37925228	1	Fus6
VPBoth	DIF	17	35849590	35849900	-1	4	782909	783243	-1	Fus7
Vpboth	INV	10	96875800	96876182	1	10	114571574	114571949	1	Runthru18
Vpboth	DIF	17	35670072	35670392	-1	4	778819	779114	-1	Runthru2
VPBoth	DIF	10	124148806	124149214	1	3	171407291	171407709	-1	Ruthru15

i) VP229/VP267 Structural Variant Primers

Primer name	VP229PCR	VP267PCR	NormalfemalePCR	F primer	R primer
A1	y	y	y	TGGTTCCATTCTTAACTCTATCTGA	AAATGAGTGTGGCCATCTGT
A10	y	y	n	AACCACACATATGAACACAGCA	TGAGTGGTCATGCTTCAGGA
A12	y	y	n	CACAATATCCCAGCATGAGG	AATTTGGGACAGGCTTGATG
A13	y	y	n	CAGGAATAGCTCTCCCAGTCC	ATTCCCATCCAGGAAGGTCT
A14	y	y	y	GATCTGCTTACCCCCATCAA	CGCACTTATGTGTCACTTCCTT
A15	y	y	n	GTGGACGTCACAGACCAGAG	GAGCTGTGGCTTTCCTGTTC
A16	y	n	n	AGCTTTGTTCAGGCACGAGT	GAGCAGATCACGAGGTCAGG

A17	y	y	y	CCCAGAGCAGAGAAGAGCAG	GCTCGTCAAGTGTGGGAAA
A18	y	y	n	AGGCCTTGTTCTCTGCTTCA	CTTACCCCATCTCAGGGTGA
A2	y	y	n	AGCAAGATTTGCAAGCAGGT	TGTAAATTGTCAGATGAAGGCAGT
A21	y	y	n	CACGCCTGGCTTTTTGTATT	TCATCGCGCATCATAGTCTC
A22	y	y	y	GATGGGCAAGCATCTGTTCT	GATTTACCTTCGAGGCATGG
A23	y	y	n	AGCCAGTCTCAAATCGCCTA	GTGGGACCTGTCCTCTTTCA
A24	y	y	n	GAAACACAGGAACTGTCAGCA	GTTTGCACTTACCCTGAGGA
A25	y	y	n	TGCAGTGAGCTCTCAGGAAA	GGGTCTAGGTTGCTGTGGAT
A26	y	y	n	AGAGGTCCCCTCCCATGA	GTACTCCAGGCGGTTTTCAA
A27	y	y	y	TGCATGCCCTTCTGATGATA	TCTATGTTCCAAGCCTCCTCA
A28	y	y	n	TGTCATTTTGCAAATGGTTCTT	TGCGCTATTAAATTTGGAGACC
A3	y	y	y	TTGCCTACCATTTCCTCTGAA	GGTGACCAAAGATTGCAAAAA
A30	y	y	y	CCACTCTGCTTGCCTTATCC	GATGCCATCTGTCCACCTCT
A31	y	y	n	GACGTTTGGGACCTGAGAAA	CAGTTGGTCCAGCTCCTACC
A32	y	y	n	TTCCCAGGTTTCTAAGTGCAG	TGACCTTACAAAAACCACCTTTT
A33	y	y	n	GCCTATTCATTTTATCACCATACTTC	GCTACCCAAACAGAATGAGAAGA
A34	y	y	y	ACATGTGGGCACACAGAAGA	TGTGGCCTTAGAGTGGGACT
A35	y	y	n	AACTTGCTTGCCTCTGGTGT	GAGCACAGCTGGGTATTAGCA
A36	y	y	n	AGAGGGGGAAGGCTGAACTA	GGGCACATATCCCTTGAAGA
A36	y	y	n	AGAGGGGGAAGGCTGAACTA	GGGCACATATCCCTTGAAGA
A36	y	y	n	AGAGGGGGAAGGCTGAACTA	GGGCACATATCCCTTGAAGA
A37	n	n	n	CGTGTTCTAACAAAGATCTGTCAA	CACTGTCCAAAGGATGTTGC
A38	y	y	n	TTTTAATGATACTTTCTTTCTCTGACA	GCTTGTAACTTTCAAAAATAGTTGAGG
A39	y	y	y	CTCGCTGCAATTTGTCTCTG	TCTTTGCAGGCAATGTGTGT
A4	y	y	y	GTCTGTTTGCTCTGCCTCCT	TCCCATTAACTTGATTTCTGCTC
A40	y	y	n	TGGAGTCTGTTACCTGAGAGTTAGAA	GCCCAGTCATGGTTTGTGTA
A41	y	y	n	AGGGGGTCCACCAATTCTAC	GGGAAAGGACTTTGCTAGGG
A42	n	n	n	GATGGAACCCTGGTTAGGTG	AAAAATTCAAGCATATGGAAAAACTTA
A43	y	y	n	GGCATGAGAAACATTTCAGTTT	GGGAGAAAGCCCTCATCTTC
A44	y	y	y	AACTGAAACGACAGGGGAAA	CTGGAGCCTCCTCCTAGACC
A45	y	y	y	TTTTCCTCCTTCCCACACAC	GGCCATACCACAGAGTGACA
A46	y	y	y	TTGTCGATTTGCTGAACAGG	GCTGCCAGCTGTCTATTTGA
A47	y	y	y	GCTGAGGATTGTGGAACCAT	GCTTTGCCTGACCAAAGTCT
A48	y	y	n	ACGCACCGCTTTCCTCTC	GCCAAAGATTGTAGTGATTTCCTT
A49	n	n	n	TCATACCCATAAATCCCACTTC	GGTCAAGAAGAGTTGCTAGATATTAAA
A5	n	y	n	AATGTGCACAACCCACTTCA	AACAGAGCCTCAGAGCCAAA
A50	y	y	y	CTGCCCACCCTATCTTAGCC	GCCACATAGGAGCTCACCAG
A51	y	y	y	GGCTTTATTCATCCCGGTTT	TCAGCTTCATCTCCCATTGA
A52	y	y	y	TCCATCGGTCCCTCAAATAA	CTGTTCATCCCTGCATCCTT
A53	y	y	y	GTCCTCCAGGAAGCACTGAG	GACAGCTGCCTATGGGATGT
A54	n	n	n	GAGCTTAGGTGGCACAGAGG	CAGCGGTTGGTGATGTCATA

A55	y	y	n	TGCAACCATCTCTTCCTTCC	TGGCTTGGTCAGAGTGTGAG
A56	y	y	n	CTTGGGGATCTAGGCATTCA	GTGGTGGCACCTGTACTCCT
A57	n	n	n	GGGTCTCATTTCCATGTTTTG	GAAGCCAGTTTACTGCTGCT
A58	y	n	n	TTCTCCCTCGCTGTGAAACT	CCGTCATTAACCCACCATCT
A59	n	y	n	CAGGCTTGACGAACAAGTGA	GCCCTCGTTTTTCTTTTTGA
A6	n	y	n	AGCAAACACAAGGCCAAGAT	CCAAAGTCAAAAGGCAGTGG
A60	y	y	n	CTCAAAAGGCAAAGCAGCTC	TGTCCTGTGCCTTATAAACAGTG
A61	y	n	n	TGGGTTTGCCATTCTCCT	TCCCTATCTCCCTTCCTTCA
A62	y	y	y	GGAGCATTACAAGCAGTGAGAC	GGAAGTGGTCAAGTTCTCAGC
A63	y	y	y	CCCTGTCCCTCTCTGGTGTA	GGAGGAAGCAGAGTGACTGG
A64	y	y	y	GCAATTACAAGGGTGGATGG	GTCCTTCAGAGGCTGACACC
A65	n	n	n	TAGGCCAGGCATAGCAGTTC	GCCTGGCGATATACCTTCTG
A66	y	y	y	TGTTCCCAAAACGTTGAACA	GACGGGATGTAGCAGCAAAT
A67	y	y	n	TAAATGACCCCTCCCCTTGT	TGCAGCAGCCAGTGAGTC
A68	n	y	n	GCCCTAACCTAAGTCGAAGG	AGATGTTTTCCACATACATGCTT
A69	y	y	y	GGAGACCCAGCTATGACACG	TATCTGTGTTTCCCCCAGGA
A7	y	y	y	ATTAAAAAGGGCAGGGCAGT	GGTAAACTTTTAGGGAGCTAGGTAAT
A70	y	y	n	TTTTGGCCCTAACTGGTCAC	CATTTGCAAACCAAATCACA
A71	y	y	y	CAGAAGGGTCCTGCTGTGTT	GTTACCTCCATTGGGCACTC
A72	y	y	n	CCTACGCAAAGCCTATGGTC	CCCAGGAAAGTTAGGGAGGA
A73	n	n	n	ATTTGTGGTTTGGCAAAAGG	AAAGCAGGTCATTGCTTTTCA
A74	n	n	n	TCGGGGACATGAGTTTATCT	CCTTTTCCAGGGCTAACTCC
A75	y	y	n	TTAATCCAGCCGTGCTTAGG	TAGGAGCAAGGGGAAGTTCA
A8	y	y	n	TGTGGCTTTGTACACTTCTGTCT	AAAAGGAAAGGGGACTTGGA
A8	y	y	n	TGTGGCTTTGTACACTTCTGTCT	AAAAGGAAAGGGGACTTGGA
A9	y	y	y	TGCAGCATTTTCTTTTTGCT	TGGTGCTGTAACTCAAACATCA
Del20	y	y	n	GAACCCATCCTTGGACAGAA	CAGCATGACTGCCTTGCTTA
Del21	y	n	n	AGCCTGAATGTCAAGGTGCT	GGTAGGGTGGACTGTGTGCT
Del34	y	y	y	GCCTATTCGCTGGGTCTTC	AGGCACCTGCAGAGAGAGAG
Fus1	y	y	n	TGATACATGCAAAAACGTGGA	GTCCTAGGTGGGAGGGAGAG
Fus10	y	y	y	TGGCATCAGTAATTGGAGCA	CAATGTACTGCTGGGGTACAAA
Fus14	y	y	n	GGCTTAAAGCTTGGGACACA	CTCCAACCAGGCTGTTGTTT
Fus15	y	n	n	GGAGAGGCCTCCTGATTTTC	AGCACCCCCAGAACCTTAGT
Fus16	y	y	n	ACGGAAACGTGGAAAATCAC	CCAGCAATGACTCCAGTGAA
Fus17	y	y	n	AAAGTGCAGGCAAGGAGAAA	CCCTGGGATTCACAAATATCA
Fus19	y	y	n	TCCATTATCCAAAGAGTTCATTCA	AGCGTGGTCCACCTTAAAAA
Fus19	y	y	n	TCCATTATCCAAAGAGTTCATTCA	AGCGTGGTCCACCTTAAAAA
Fus20	y	y	n	TCAATCCCTGTCTCCTTCCA	ATAGTTGCCCTGCTGATTGG
Fus21??	y	y	n	ATGGCCTCCTCTCCTGCT	AGCTGCGGTCTCACTCTAGG
Fus22	n	n	n	CTGCAACTGTTGGATGAAATG	TTCAAAGACCCCCAAATTGT
Fus23??	y	y	n	TGCGTTTAAATCAAATCAACGA	AACAAGATGTGTTGATATTTGGAGAG

Fus24	n	y	n	GCCTTTCTGGACTTCTGTTCC	TGGAAACATTAGAAAGGGCAAC
Fus25	n	n	n	CCAGAGCACTTGCATTTTGA	AGCTTCTCCCACCTGGATCT
Fus27	n	y	n	TTCAACACACACTGGCTTCC	TGCTTGTGTTGCTCTTTTGG
Fus28	y	y	n	TCTCTCTCTCCCTGCCACTG	CTGGGGTAGAAAAGGTGGTG
Fus29	y	y	n	GCCTTTCTGAGTGGGAACAA	AGTAATCCTCAGCCCCATCG
Fus30	y	y	n	TGGGAGAAAGATCATTTGCTAT	ACAGACACCCTTTGGACCAC
Fus30	y	y	n	CACCCAGGTGCTACTTTACGA	GAAGTGGAGCCCCATTGAG
Fus30	y	y	n	CACCCAGGTGCTACTTTACGA	GAAGTGGAGCCCCATTGAG
Fus5	y	y	n	AGGGTGCTCCTTTCCTTCTC	GCAGACCTAGAGGCTGTGCT
Fus6	y	y	n	TGTCTTGTGCTGGGAATTGT	CTCTGGGCACACATACATGC
Fus7	y	y	n	GCGAGAAAGCAAATCCGATA	CACTCCCATCCTCAGGTGTT
Runthru18	y	y	y	GGTCATTCGGGAGCATATTG	GGGCCCACATTCCTTTTATT
Runthru2	y	y	y	TTTCTGAGTATTCTTTCTCCCAAGA	GGTGTCCTCTGTCCGTCTGT
Ruthru15	n	n	n	TGATGGCATGTGCCTGTAAT	TTCAAGGCTGCAGTGAGCTA

iii) Inverted Tandem Repeat and Small Deletion Predicted Structural Variants

Predicted in	Type of SV	Node 1	Node 1 start	Node 1 end	Node 1 direction	Node 2 chr	Node 2 start	Node 2 end	Node 2 direction	Primer Name
Vpboth	ITR	4	75042125	75042539	1	4	75042125	75042539	-1	ITR5
VP229	DEL	15	70479894	70480192	1	15	70480652	70480863	-1	VP229_smallDEL5
VP229	DEL	12	19554147	19554458	1	12	19554657	19555116	-1	VP229_smallDEL1
VP229	DEL	19	19535993	19536032	1	19	19536772	19536925	-1	VP229_smallDEL8
VP229	DEL	1	156622751	156623199	1	1	156623465	156623925	-1	VP229_smallDEL2
VP229	DEL	9	17612391	17612676	1	9	17613036	17613351	-1	VP229_smallDEL4
Vpboth	ITR	16	48084954	48085365	1	16	48084954	48085365	-1	ITR4
VP229	DEL	11	71712233	71712644	1	11	71712910	71713324	-1	VP229_smallDEL3
Vpboth	ITR	11	66277922	66278273	1	11	66277922	66278273	-1	ITR2
Vpboth	ITR	4	188170325	188170496	1	4	188170325	188170496	-1	ITR6
Vpboth	ITR	12	92734889	92735301	1	12	92734889	92735301	-1	ITR3
Vpboth	ITR	10	85628541	85628978	1	10	85628541	85628978	-1	ITR1
VP229	DEL	5	31860066	31860119	1	5	31860779	31860830	-1	VP229_smallDEL7
VP229	ITR	11	79634637	79634965	1	11	79634637	79634965	-1	ITR32
VP229	DEL	18	54691525	54691641	1	18	54692203	54692319	-1	VP229_smallDEL6

iv) Inverted Tandem Repeat and Small Deletion Primer Sequences

Primer Name	F primer sequence
ITR5	AAGTTCCTCTGGCTGCGTAA
VP229_smallDEL5	ACAGCCATCTTGGGAAACAC
VP229_smallDEL1	AGGAGACTGCCACCATGC
VP229_smallDEL8	CAGTGTGGGGAGGCTATTTG
VP229_smallDEL2	CCCACAGAGCCTTAAGCAAC
VP229_smallDEL4	CTGGGAGGATGCATTTCACT
ITR4	GGCTTGGTAACTGGTGGAAA
VP229_smallDEL3	GTGGACCCTGAACAGGTTGT
ITR2	GTGGCTGACCCAGAGATTGT
ITR6	TCATAGCTTGTGCCGAACAG
ITR3	TGCAACAACTCCTGCAAGTC
ITR1	TGGGTTCTAAGGGTGTCCTG
VP229_smallDEL7	TGGTGGTTTCATTGGAGGAT
ITR32	TGTCGAACTTCCAGAAATAAAAT
VP229_smallDEL6	TTGCATTCATTCAAGGCTCA

A.1.7. PCR Primers to confirm somatic structural variants in HCC1187

HCC1187Fus1F	TGGATTGGTTTCTCTTTCTCC
HCC1187Fus1R	CCTTCTACCGCCTCCTCAC
HCC1187Fus2F	GGTCAGCCATCATCTGTGTC
HCC1187Fus2R	TGGAGACAATAAGTTGGAGCAA
HCC1187Fus3F	GCATGGTGGCTTACACCTG
HCC1187Fus3R	GGGCAAAGGTTTTATGGCTA
HCC1187Fus4F	AACTCATGGCCCATACAATG
HCC1187Fus4R	TTCTTCCACCTAAGCCTTGC
HCC1187Fus5F	TCCTGGATATCACCCTTGAGA
HCC1187Fus5R	CTGAAAATGAACGCAGGACA
HCC1187Fus6F	TGCCCAACGTGGTAAGTAAA
HCC1187Fus6R	TTACGTGCTCAGGGGAGCTA
HCC1187Fus7F	AGCAGTGTGCAATCTGCATT
HCC1187Fus7R	TGTTGTCAAAACCCATCCAG
HCC1954Fus1F	GAGAGGGTGGCAATGTGAGT
HCC1954Fus1R	CAGTGGTGGTATCCTGTTTATCA
RHOJgDNA_F	GCACATGGAAACACATGGAA
SYNE2gDNAR	GCAGTACACAAGGGGCTAGG

Appendix 2. Perl and R Scripts

Appendix 2.1. Perl script to extract break point regions from segmented SNP6 data

```
#!/usr/bin/perl -w
#Finds the breakpoint data from PICNIC segmented .csv files in a directory
use warnings;
use strict;

my$output = "BreastLinesPICNICbreaks.txt";
open(OUTPUT, ">$output");
#Loops through all the files in the directory
@files = <*>;
foreach $file (@files) {
#Prints current file name
print $file . "\n";
my $infile  = $file;
my @data;
print "loading $infile ...\n";
open( INFILE, "<$infile")or die( "Couldn't open file $infile: $!\n" );
@data=<INFILE>;
close( INFILE );
print "$infile loaded\n";
my $linecounter=0;
foreach(@data){
            #Ignores file header
            #The input file is comma delimited: chromosome, position, Affymetrix
            #probe ID, Alternative ID, nucleotide 1, nucleotide 2, probe
            #intensity, segmented copy number, genotype 1, genotype 2.
            next if ($_=~m/#/);
            $linecounter++;
            chomp$_;
            my ( $chr , $pos , $affyID ,  $RSID , $nuc1 , $nuc2 , $copy ,
            $genotype_intensity , $segmentCN , $genotype1 , $genotype2 ) =
            split(',', $_);
            my ( $ch , $chrNo ) = split('r', $chr);

            my ( $chrB , $posB ,$affyIDB ,  $RSIDB , $nuc1B , $nuc2B , $copyB ,
            $genotype_intensityB , $segmentCNB , $genotype1B , $genotype2B ) =
            split(',', $data[$linecounter]);
            my ( $chB , $chrNoB ) = split('r', $chrB);

            if ($chrNo=~m/X/)
            {$chrNo=23}
            if ($chrNoB=~m/X/)
            {$chrNoB=23}
            if ($chrNo=~m/Y/)
            {$chrNo=24}
            if ($chrNoB=~m/Y/)
            {$chrNoB=24}

            #Finds chage points in copy number between one line and the next
            if(($chrNo==$chrNoB)&&($segmentCN!=$segmentCNB)){
                  print OUTPUT "$infile   $chr:$pos-$posB   $segmentCN
                  $segmentCNB\n";
                  }
            }
      }
close(OUTPUT);
```

Appendix 2.2. Perl Script to find genes at SNP6 break points

```
#Finds the breakpoint regions from the above script that coincide with genes or
#gene windows
use warnings;
use strict;

#Load all PICNIC breakpoint regions, all positive and negative strand genes and
#gene windows
my $infile1  = "Cell_LinesPICNICbreaksAll.txt";
my $infile2 = "tableOfPlusGenesGenome.txt";
my $infile3 = "tableOfMinusGenesNegGenome.txt";
my $infile4 = "geneWindowsPLusWDes.txt";
my $infile5 = "geneWindowsMinusWDes.txt";
my $infile6  = "genes_at_all_PICNIC_breaks.txt";
my$output = "genes_at_all_PICNIC_breaks.txt";

#Set up arrays to hold each data file
my @PICNIC_breaks;
my @pos_genes;
my @neg_genes;
my @pos_windows;
my @neg_windows;

#Open files
open( INFILE1, "<$infile1")or die( "Couldn't open file $infile1: $!\n" );
open( INFILE2, "<$infile2")or die( "Couldn't open file $infile2: $!\n" );
open( INFILE3, "<$infile3")or die( "Couldn't open file $infile3: $!\n" );
open( INFILE4, "<$infile4")or die( "Couldn't open file $infile2: $!\n" );
open( INFILE5, "<$infile5")or die( "Couldn't open file $infile3: $!\n" );
open(OUTPUT, ">$output");

#Assign files to arrays
@PICNIC_breaks=<INFILE1>;
@pos_genes=<INFILE2>;
@neg_genes=<INFILE3>;
@pos_windows=<INFILE4>;
@neg_windows=<INFILE5>;
print "$infile1 loaded\n";
print "$infile2 loaded\n";
print "$infile2 loaded\n";

#Close input files
close( INFILE1 );
close( INFILE2 );
close( INFILE3 );

print "Processing ...\n";
foreach(@PICNIC_breaks)
      {
      #print "$_";
      chomp$_;
      my ( $cellline , $break_chr , $break_start , $break_end , $prevCN ,
      $afterCN ) = split("    " , $_);
      my$break_polarity=$afterCN-$prevCN;

if($break_polarity > 0)
            {
      #Positive breaks positive genes
```

```
      #Any break that retains the 5' end of the gene or any break within the
      #gene window for a runthrough
      #if loop pulls out positive breakpoint polarities
            foreach(@pos_windows)
                  {
                  chomp$_;
                  my ( $genechr1 , $window_start , $window_end , $gene_name ) =
                  split(/\t/, $_);
                  my ( $genech1 , $genechr ) = split(/chr/, $genechr1);


                  #The PICNIC breakpoint region has to be entirely within the
                  #gene window. For this to happen
                  # the window start has to be less than the PICNIC start and
                  #the window end has to be greater
                  # than the PICNIC end.


if(($break_chr==$genechr)&&($window_start<$break_start)&&($window_end>$break_end
))
                        {
                        #Output line
                        print OUTPUT "$cellline $break_chr  $break_start
                        $break_end  $gene_name  3' end retained\n";
                        }
                  }
                  #Positive breaks negative genes
                  #These retain the 3' end of the gene. The break has to be
                  #within the gene and exclude the 5' window.
                  #in this case the gene window end is adjusted by subtracting
                  #the window       #size defined above.
                  foreach(@neg_genes)
                  {
                  chomp$_;
                  my ( $genechr1 , $window_start , $window_end , $gene_name ) =
                  split(/\t/, $_);
                  my ( $genech1 , $genechr ) = split(/chr/, $genechr1);


if(($break_chr==$genechr)&&($window_start<$break_start)&&($window_end>$break_end
))
                        {
                        #Output line
                        print OUTPUT "$cellline $break_chr  $break_start
                        $break_end  $gene_name  5' end retained\n";
                        }
                  }
            }

                  #negative breaks negative genes
                  #These retain the 5' end of the gene or form possible
                  #runthroughs so consider the whole gene window
                  #this else loop is left over from the if loop above. All the
                  #-ve breaks are considered here
      else
            {
            foreach(@neg_windows)
                  {
                  chomp$_;
```

```
                my ( $genechr1 , $window_start , $window_end , $gene_name ) =
                split(/\t/, $_);
                my ( $genech1 , $genechr ) = split(/chr/, $genechr1);


if(($break_chr==$genechr)&&($window_start<$break_start)&&($window_end>$break_end
))
                        {
                        #Output line
                        print OUTPUT "$cellline $break_chr  $break_start
                        $break_end  $gene_name  3' end retained\n";
            }
                  }

      #negative breaks positive genes
      #These can only retain the 3' end of the gene so discount the window
            foreach(@pos_genes)
                  {
                  chomp$_;
                  my ( $genechr1 , $window_start , $window_end , $gene_name ) =
                  split(/\t/, $_);
                  my ( $genech1 , $genechr ) = split(/chr/, $genechr1);



if(($break_chr==$genechr)&&($window_start<$break_start)&&($window_end>$break_end
))
                        {
                        #Output line
                        print OUTPUT "$cellline $break_chr  $break_start
                        $break_end  $gene_name  5' end retained\n";

                        }
                  }
            }
      }

close OUTPUT;

#Load the file you just created in order to get rid of duplicate entries
print "Removing duplicates ...\n";
open( infile6, "<$infile6")or die( "Couldn't open file $infile6: $!\n" );
my@PICNIC_genes=<infile6>;
open(OUTPUT, ">$output");

#Make a hash of all the lines of the input file and returns all unique keys
#then prints to the output file

my %hash   = map { $_, 1 } @PICNIC_genes;
my @unique = keys %hash;
foreach(@unique)
      {
      print OUTPUT "$_";
      }
close OUTPUT;
print "finished\n";
```

Appendix 2.3. Perl script to extract break point regions for PCR verification

```
#!/usr/bin/perl -w
#Pull out genome slices and construct a sequence to design primers for PCR
#verification
#This script will output "Slice1[]Slice2" ready for primer3 and pre-mask all
#repeats

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

#make a connection to the Ensembl database

my $registry = 'Bio::EnsEMBL::Registry';
$registry->load_registry_from_db(
      -host => 'ensembldb.ensembl.org',
      -user => 'anonymous'
      );

#Open input file should be as below $cellline, $breakID , $SVtype , $SVsupport ,
#$node1chr , $node1start , $node1end , $node1dir , $node2chr , $node2start ,
#$node2end , $node2dir
#eg.
#VP229      Fus10 INS   2     3     10115221    10115261    -1    3     11932819
#     11932869    1

my $infile  = "VP229breaks.txt";
print "loading all breaks...\n";
open( INFILE, "<$infile")or die( "Couldn't open file $infile: $!\n" );
my @allbreaks;
@allbreaks=<INFILE>;
close( INFILE );
print "$infile loaded\n";

#Create an output file
my $outfile = "VP229breakpointregions.txt";
open( OUTFILE, ">$outfile");

#set the size of the genome slice you want to get back
my $slice_size=1000;
foreach (@allbreaks)
      {
      chomp$_;
      my( $cellline, $breakID , $SVtype , $SVsupport , $node1chr , $node1start ,
$node1end , $node1dir , $node2chr , $node2start , $node2end , $node2dir ) =
split(/\s/, $_);

            #Positive read, negative read = positive strand joined to positive
            #strand: Simplest case.
            #Pull out sequence to LHS of node1 end and to RHS of node2 start

            if(($node1dir =~ m/^1/)&&($node2dir =~ m/^-1/))
                  {
                  print "positive and negative : $node1dir , $node2dir\n";
                  my $first_slice_start=$node1end-$slice_size;
```

```
        my $first_slice_end=$node1end;
        my $second_slice_start=$node2start;
        my $second_slice_end=$node2start+$slice_size;
        print "will fetch $node1chr : $first_slice_start -
        $first_slice_end [] $node2chr : $second_slice_start -
        $second_slice_end \n";

        #make slice adapter1 by linking it to the ensembl database
        my $slice_adaptor1 = $registry->get_adaptor( 'Human', 'Core',
        'Slice' );
        #fetch the slice defined by chr node start and node end
        my $chromslice1 = $slice_adaptor1-
        >fetch_by_region('chromosome', $node1chr, $first_slice_start,
        $first_slice_end);
        #Fetch the same region repeatmasked
        my $repeat_slice1 = $chromslice1->get_repeatmasked_seq();
        #compare the two slices and output repeat-masked sequence
        my $sequence1 = $repeat_slice1->seq();
        #repeat for slice2
        my $slice_adaptor2 = $registry->get_adaptor( 'Human', 'Core',
        'Slice' );
        my $chromslice2 = $slice_adaptor2-
        >fetch_by_region('chromosome', $node2chr, $second_slice_start,
        $second_slice_end);
        my $repeat_slice2 = $chromslice2->get_repeatmasked_seq();
        my $sequence2 = $repeat_slice2->seq();
        print "$sequence1 [] $sequence2 \n";
        print OUTFILE "$breakID $sequence1 [] $sequence2\n";
        }
#Negative read, positive read = positive strand joined to positive
#strand, but the read is from negative strand.
#Like the above, just flipped over.
if(($node1dir =~ m/^-1/)&&($node2dir =~ m/^1/))
        {
        print "negative and positive : $node1dir , $node2dir\n";
        my $first_slice_start=$node2end-$slice_size;
        my $first_slice_end=$node2end;
        my $second_slice_start=$node1start;
        my $second_slice_end=$node1start+$slice_size;
        print "will fetch $node1chr : $first_slice_start -
        $first_slice_end [] $node2chr : $second_slice_start -
        $second_slice_end \n";
        my $slice_adaptor1 = $registry->get_adaptor( 'Human', 'Core',
        'Slice' );
        my $chromslice1 = $slice_adaptor1-
        >fetch_by_region('chromosome', $node1chr, $first_slice_start,
        $first_slice_end);
        my $repeat_slice1 = $chromslice1->get_repeatmasked_seq();
        my $sequence1 = $repeat_slice1->seq();
        my $slice_adaptor2 = $registry->get_adaptor( 'Human', 'Core',
        'Slice' );
        my $chromslice2 = $slice_adaptor2-
        >fetch_by_region('chromosome', $node2chr, $second_slice_start,
        $second_slice_end);
        my $repeat_slice2 = $chromslice2->get_repeatmasked_seq();
        my $sequence2 = $repeat_slice2->seq();
        print "$sequence1 [] $sequence2 \n";
        print OUTFILE "$breakID $sequence1 [] $sequence2\n";
        }
```

```
#Positive read, positive read = Positive strand, node1 joined to
#negative strand node2.
#pull out positive node and join it to the reverse complement of the
#negative strand
if(($node1dir =~ m/^1/)&&($node2dir =~ m/^1/))
      {
      print "positive and positive : $node1dir , $node2dir\n";
      #As for +ve read joined to -ve read
      my $first_slice_start=$node1end-$slice_size;
      my $first_slice_end=$node1end;
      my $second_slice_start=$node2end-$slice_size;
      my $second_slice_end=$node2end;
      print "will fetch $node1chr : $first_slice_start -
      $first_slice_end [] $node2chr : $second_slice_start -
      $second_slice_end \n";
      my $slice_adaptor1 = $registry->get_adaptor( 'Human', 'Core',
      'Slice' );
      my $chromslice1 = $slice_adaptor1-
      >fetch_by_region('chromosome', $node1chr, $first_slice_start,
      $first_slice_end);
      my $repeat_slice1 = $chromslice1->get_repeatmasked_seq();
      my $sequence1 = $repeat_slice1->seq();
      my $slice_adaptor2 = $registry->get_adaptor( 'Human', 'Core',
      'Slice' );
      my $chromslice2 = $slice_adaptor2-
      >fetch_by_region('chromosome', $node2chr, $second_slice_start,
      $second_slice_end);
      my $repeat_slice2 = $chromslice2->get_repeatmasked_seq();
      my $sequence2 = $repeat_slice2->seq();

      #make reverese complement of slice2
      my$revcomp = reverse $sequence2;
      # The Perl translate/transliterate command does reverse
      #compliment and ignores repeats masked to "N".
      $revcomp =~ tr/ACGTacgt/TGCAtgca/;
      print "$sequence1 [] $revcomp \n";
      print OUTFILE "$breakID $sequence1 [] $revcomp\n";
      }

if(($node1dir =~ m/^-1/)&&($node2dir =~ m/^-1/))
      {
      print "negative and negative : $node1dir , $node2dir\n";
      my $first_slice_start=$node1start;
      my $first_slice_end=$node1start+$slice_size;
      my $second_slice_start=$node2start;
      my $second_slice_end=$node2start+$slice_size;
      print "will fetch $node1chr : $first_slice_start -
      $first_slice_end [] $node2chr : $second_slice_start -
      $second_slice_end \n";
      my $slice_adaptor1 = $registry->get_adaptor( 'Human', 'Core',
      'Slice' );
      my $chromslice1 = $slice_adaptor1-
      >fetch_by_region('chromosome', $node1chr, $first_slice_start,
      $first_slice_end);
      my $repeat_slice1 = $chromslice1->get_repeatmasked_seq();
      my $sequence1 = $repeat_slice1->seq();
      my $slice_adaptor2 = $registry->get_adaptor( 'Human', 'Core',
      'Slice' );
      my $chromslice2 = $slice_adaptor2-
```

```
                    >fetch_by_region('chromosome', $node2chr, $second_slice_start,
                    $second_slice_end);
                    my $repeat_slice2 = $chromslice2->get_repeatmasked_seq();
                    my $sequence2 = $repeat_slice2->seq();

                    #make reverese complement of slice1
                    my$revcomp = reverse $sequence1;
                    $revcomp =~ tr/ACGTacgt/TGCAtgca/;
                    print "$revcomp [] $sequence2 \n";
                    print OUTFILE "$breakID $revcomp [] $sequence2\n";
                    }
      }

print "Finished!\n";
close ( OUTFILE );
```

Appendix 2.4. R script to generate maximum likelihood estimators and 95th percentile confidence intervals

```
#This R script was written by Professor S. Tavaree, Cambridge CRI and is based
#on a statistical model also written by Professor Tavaree.

#Load the nnet package
library(nnet)

#Total number of any giver class of mutations
m <- 75

#Number of mutations that fell early
y <- 34

#Proportion of mutations that fell early by chance
p=0.4

lik <- numeric(y+1)
for(n in 1:(y+1)){lik[n]<-dbinom(y-n+1,m-n+1,p)}
# clean up rounding
lik <- signif(lik,digits=6)
mle <- which(lik == max(lik)) - 1

confint <- function(m,nhat,B,alpha,p){
nhathat <- numeric(B)
# Simulate B values of y = nhat + Bin(m-nhat,p)
for(i in 1:B){ y <- nhat + rbinom(1, m-nhat, p)
lik <- numeric(y+1)
for(n in 0:y){lik[n+1]<-dbinom(y-n,m-n,p)}
lik <- signif(lik,digits=6) # clean up rounding!
nhathat[i] <- which.is.max(lik) - 1 #chooses MLE at random if ties
}
snhathat <- sort(nhathat,decreasing=FALSE)
nl <- snhathat[floor(B * alpha/2)]
nu <- snhathat[ceiling(B * (1 - alpha/2))]
cll <- 2 * nhat - nu
clu <- 2 * nhat - nl
print(cll)
print(clu)
}

#Returns the maximum likelihood estimator (MLE)

mle

#Returns the 95th perrcentile confidence intervals. This example is for a sample
#of 34 early mutations, the mle calculated above (7 in this case), and run for
#10000 iterations, with p of 0.4
#confint(34,5,10000,0.05,0.321)

confint(m,mle,10000,0.05,p)
```

Appendix 3. HCC1187 Mutations

Name	Proposed Junction	Chr	Break LHS	Break RHS		Chr	Break LHS	Break RHS
A, D	t(1;6)	1	31160802.5	31296885	;	6	42354900	42358000
A	t(1;8)	1	150503922.5	150563199	;	8	14331542.5	14465907.5
D	t(1;X)	1	96961781.5	97980642	;	X	50231247.5	54118741.5
E	t(1;8) a	1	84172800	84174600	;	8	99290955.5	101355408
E	t(1;8) b	1	150503922.5	150563199	;	8	14589000	14591700
G	t(2;20)	2	89772452.5	94882962	;	20	30605409.5	30851017.5
H	t(1;8)	1	84135871.5	85445694.5	;	8	99290955.5	101355408
J	t(1;8)	1	115246265	115255974	;	8	99290955.5	101355408
N	t(10;13)	10	22830000	22875000	;	13	58272600	58273300
N	t(10;19)	10	84424000	84425500	;	19	56139000	56141000
O	del(11)q13.5q21)	11	76153425.5	78120576.5	;	11	88001257.5	90989730
O	t(11;12)	11	-	-	;	12	28228473.5	28699164.5
P	t(2;19)	2	89772452.5	94882962	;	19	-	-
R	t(11;16)	11	10435000	10438000	;	16	66141900	66142200
R	del(11)q13.5q21)	11	76153425.5	78120576.5	;	11	88001257.5	90989730
S	t(12;16)	12	28228473.5	28699164.5	;	16	66141900	66142200
S	t(11;16)	11	-	-	;	16	66141900	66142200
S	t(11;12)	11	-	-	;	12	28228473.5	28699164.5
T	t(2;19)	2	54050000	54053000	;	19	5775600	5778000
V	t(2;5)	2	89772452.5	94882962	;	5	42897896	50000844
V	del(2p)(p16p25.1)	2	-	-	;	2	54050000	54053000
Y	t(14;20)	14	75743298	76653723.5	;	20	>1	<30760000
b	t(10;13)	10	22830000	22875000	;	13	58272600	58273300
c	i(13)	13	-	-	;	13	-	-
i	t(1;19)	1	11042655.5	11551099	;	19	42104980.5	42928442.5
j	t(13;20)	13	-	-	;	20	>1	<30760000
k	t(1;X)	1	-	-	;	X	122281542.5	123854583
L	t(X;X)	X	26887513	27118094	;	X	-	-

Table A3.1. Chromosome break points defined by array painting. Break point regions were defined by BACs or oligo probes flanking copy number shifts. Genome co-ordinates are as in the HG18 genome build.

Name	SV Type	Chr A	Position A	Strand A	Chr B	Position B	Strand B	Size	Support
StephensDEL1	DEL	2	996170	+	2	1007188	-	11018	6
StephensDEL2	DEL	2	11397776	+	2	55159176	-	43761400	1
StephensDEL3	DEL	2	33031637	+	2	33097132	-	65532	16
StephensDEL4	DEL	2	66861264	+	2	67119301	-	258074	5
StephensDEL5	DEL	4	138011731	+	4	138039024	-	27293	8
StephensDEL6	DEL	6	1664850	+	6	5686186	-	4021373	6
StephensDEL7	DEL	6	134369018	+	6	134579664	-	210646	6
StephensDEL8	DEL	7	110845099	+	7	110867805	-	22706	8
StephensDEL9	DEL	7	132786464	+	7	132871038	-	84574	8
StephensDEL10	DEL	8	41576622	+	8	41588498	-	11876	7
StephensDEL11	DEL	8	70272683	+	8	70277843	-	5160	6
StephensDEL12	DEL	10	14616540	+	10	14649227	-	32687	11
StephensDEL13	DEL	10	77414245	+	10	79429479	-	2015234	8
StephensDEL14	DEL	11	9108490	+	11	10519862	-	1411372	7
StephensDEL15	DEL	11	34072921	+	11	34082648	-	9764	8
StephensDEL16	DEL	11	55503503	+	11	55507819	-	4353	4
StephensDEL17	DEL	11	77809792	+	11	90309424	-	12499632	12
StephensDEL18	DEL	12	114935765	+	12	114961462	-	25697	8
StephensDEL19	DEL	13	47846990	+	13	47850855	-	3902	1
StephensDEL20	DEL	14	79857506	+	14	79867739	-	10233	17
StephensDEL21	DEL	14	98850865	+	14	98856028	-	5163	11
StephensDEL22	DEL	17	34671312	+	17	34759972	-	88697	3
StephensDEL23	DEL	20	9081436	+	20	9131009	-	49573	11
StephensDEL24	DEL	20	52273532	+	20	52281706	-	8211	7
StephensDEL25	DEL	20	52887189	+	20	52919407	-	32255	4
StephensDEL26	DEL	X	153858946	+	X	153879982	-	21036	17
StephensDIF1	DIF	1	31183855	-	6	42356186	+	N/A	19
StephensDIF2	DIF	1	31334253	+	6	41406061	-	N/A	8
StephensDIF3	DIF	1	84234093	+	8	101058886	+	N/A	8
StephensDIF4	DIF	1	97784960	+	X	53294900	+	N/A	6
StephensDIF5	DIF	1	115272352	-	8	101059496	-	N/A	9
StephensDIF6	DIF	7	148873715	+	8	143979884	-	N/A	5
StephensDIF7	DIF	10	22831512	+	13	58272177	+	N/A	5
StephensDIF8	DIF	10	22832193	-	13	58272472	-	N/A	3
StephensDIF9	DIF	11	5536414	-	12	28661343	+	N/A	2

StephensDIF10	DIF	11	10520685	+	16	66197968	+	N/A	3
StephensDIF11	DIF	14	72736528	-	20	6147141	+	N/A	4
StephensINS1	INS	1	37593659	-	1	38176917	+	583223	5
StephensINS2	INS	1	190587754	-	1	190724320	+	136566	6
StephensINS3	INS	2	10155274	-	2	10239261	+	83987	7
StephensINS4	INS	2	104933183	-	2	105257264	+	324081	5
StephensINS5	INS	2	222206770	-	2	222863211	+	656441	3
StephensINS6	INS	2	240060032	-	2	240105411	+	45379	3
StephensINS7	INS	3	49410887	-	3	49559868	+	148981	4
StephensINS8	INS	3	56444738	-	3	56449656	+	4918	12
StephensINS9	INS	3	57150092	-	3	57255624	+	105532	8
StephensINS10	INS	3	130774767	-	3	130902421	+	127654	7
StephensINS11	INS	4	13044591	-	4	13478102	+	433476	4
StephensINS12	INS	4	40464439	-	4	40515180	+	50706	5
StephensINS13	INS	4	79203712	-	4	79238418	+	34671	6
StephensINS14	INS	4	82495561	-	4	82601352	+	105791	5
StephensINS15	INS	4	92041059	-	4	92076272	+	35213	7
StephensINS16	INS	4	117597380	-	4	117609422	+	12042	1
StephensINS17	INS	4	146298490	-	4	146725491	+	427001	16
StephensINS18	INS	5	37083312	-	5	37175248	+	91936	6
StephensINS19	INS	5	174407766	-	5	174468477	+	60711	4
StephensINS20	INS	6	2046552	-	6	2063989	+	17402	2
StephensINS21	INS	6	11391453	-	6	11573487	+	182034	4
StephensINS22	INS	6	41476568	-	6	41569960	+	93357	11
StephensINS23	INS	6	149549556	-	6	150018566	+	468975	3
StephensINS24	INS	7	104330914	-	7	104516312	+	185363	5
StephensINS25	INS	8	5911073	-	8	5986829	+	75721	5
StephensINS26	INS	8	6480859	-	8	6586862	+	106003	8
StephensINS27	INS	8	79925025	-	8	79950438	+	25413	8
StephensINS28	INS	8	102399808	-	8	102559056	+	159248	19
StephensINS29	INS	9	12879468	-	9	13045861	+	166393	14
StephensINS30	INS	9	102203807	-	9	102265942	+	62100	3
StephensINS31	INS	9	113916716	-	9	114134450	+	217699	9
StephensINS32	INS	11	16826873	-	11	16879687	+	52779	2
StephensINS33	INS	11	57072794	-	11	57295287	+	222493	5
StephensINS34	INS	11	77389514	-	11	77443887	+	54338	2
StephensINS35	INS	11	111292861	-	11	111354277	+	61416	6
StephensINS36	INS	12	11774592	-	12	11804932	+	30340	17
StephensINS37	INS	12	27532204	-	12	27595557	+	63318	3
StephensINS38	INS	13	101879040	-	13	102251960	+	372920	6

StephensINS39	INS	14	38672867	-	14	38881647	+	208780	5
StephensINS40	INS	14	63654523	-	14	63910709	+	256151	7
StephensINS41	INS	14	67838965	-	14	68157538	+	318573	11
StephensINS42	INS	16	65695550	-	16	65828537	+	132952	2
StephensINS43	INS	16	79615251	-	16	79679344	+	64093	2
StephensINS44	INS	17	46134084	-	17	46194120	+	60036	3
StephensINS45	INS	17	72865900	-	17	72905624	+	39689	2
StephensINS46	INS	18	606353	-	18	631641	+	25253	4
StephensINS47	INS	18	8929214	-	18	10803919	+	1874670	3
StephensINS48	INS	19	10398770	-	19	10509924	+	111119	4
StephensINS49	INS	19	16828242	-	19	16940309	+	112067	6
StephensINS50	INS	20	19112222	-	20	19210849	+	98592	7
StephensINV1	INV	1	157359299	+	1	215686367	+	58327068	5
StephensINV2	INV	1	157359724	-	1	215686770	-	58327047	2
StephensINV3	INV	2	42519193	-	2	42533786	-	14593	4
StephensINV4	INV	13	76399932	-	13	76416550	-	16618	4
StephensINV5	INV	13	76416077	+	13	76432425	+	16348	5
StephensINV6	INV	14	73245679	-	14	73263451	-	17772	13
StephensINV7	INV	18	9767694	+	18	9773271	+	5577	2
StephensINV8	INV	18	51504211	+	18	51510656	+	6446	4
StephensINV9	INV	20	9771873	+	20	9780917	+	9045	5

Table A3.2. Structural variants reported by Stephens et al. (2009). DEL=deletion, INV=Inversion, INS=Insertion/tandem duplication, DIF=Interchomosome translocation. Support refers to the number of paired sequencing reads that crossed the genomic junction.

Sjoblom (2006), Wood (2007)	COSMIC (2010)	Genomic mutation as reported (2004 build)	Genomic Mutation (2009 build)	cDNA Mutation	Amino acid	Mutation Type
X	X	chr2:131632752C>G (homozygous)	chr2:131799020-131799020	1322C>G	T441R	Miss
X		chr10:99429559C>T (homozygous)	chr10:99439569-99439569	94C>T	Q32X	N
X		chr6:33353713T>C	chr6:33245735-33245735	539T>C	V180A	Miss
X		chr3:52415311C>T (homozygous)	chr3:52440271-52440271	781C>T	Q261X	N
X		chr4:57673742A>G (homozygous)	chr4:57832814-57832814	1736A>G	Q579R	Miss
X		chr5:141014054A>C (homozygous)	chr5:141033870-141033870	4282A>C	T1428P	Miss
X	X	chr5:138294082C>T (homozygous)	chr5:138266183-138266183	2032C>T	Q678X	N
X		chr18:32527271C>T	chr18:34273273-34273273	1598C>T	S533L	Miss
X		chr2:27191236G>C (homozygous)	chr2:27279585-27279585	1184G>C	R395P	Miss
X		chr7:128071176G>T	chr7:128477225-128477225	553G>T	D185Y	Miss
X		chr5:177546166delA (homozygous)	chr5:177613560-177613560	741delA	fs	INDEL
X		chr12:121739602_121739601insA (homozygous)	chr12:123214721-123214722	165_166insA	fs	INDEL
X		chr6:33281286G>T	chr6:33173308-33173308	472G>T	V158L	Miss
X		chr12:56135771G>C	chr12:57849504-57849504	185G>C	R62T	Miss
X		chr18:44541852G>C	chr18:46287854-46287854	1165G>C	V389L	Miss
X	X	chr22:35012676_35012674delGCA (homozygous)	chr22:36688174-36688176	4200_4202delGCA	indel	INDEL
X	X	chr14:34942227_34942226insC (homozygous)	chr14:35872475-35872476	427_428insC	fs	INDEL
X	X	chr11:3657478G>T (homozygous)	chr11:3700902-3700902	4955G>T	G1652V	Miss
	X	chr7:154198087T>G	chr7:154567154-154567154	1370T>G	F457C	Miss
X		chr5:140607486C>T (homozygous)	chr5:140627302-140627302	2156C>T	A719V	Miss
X		chr12:41065014G>A (homozygous)	chr12:42778747-42778747	517G>A	V173M	Miss
X	X	chr12:78693190G>C (homozygous)	chr12:80190722-80190722	2301G>C	Q767H	Miss
X	X	chr17:31086470G>A (homozygous)	chr17:34062357-34062357	154G>A	V52M	Miss
X		chr3:51950902C>G (homozygous)	chr3:51975862-51975862	22C>G	R8G	Miss
X		chr3:188400138C>A (homozygous)	chr3:186917436-186917436	262C>A	R88S	Miss
X		chr6:32036799C>G	chr6:31928820-31928820	547C>G	L183V	Miss
X		chr2:220323514_220323523delGACAAGGACA (homozygous)	chr2:220498009-220498018	1291_1300delGACAAGGACA	fs	INDEL
X	X	chr18:31994943_31994944delCT	chr18:33740945-33740946	1739_1740delCT	fs	INDEL
X		chr12:10952719A>G	chr12:11061452-11061452	446A>G	N149S	Miss

X		chr7:139064224C>T	chr7:139611040-139611040	256C>T	R86W	Miss
X	X	chr17:7520090_7520088delGGT (homozygous)	chr17:7579363-7579365	322_324delGGT	G108del	INDEL
	X	chr17:71382450_71382450	chr17:73870855-73870855	1680_1681insC	fs	INDEL
X		chr2:234462937G>T (homozygous)	chr2:234680937-234680937	1325G>T	S442I	Miss
X		chr2:219333250G>A (homozygous)	chr2:219507745-219507745	3005G>A	R1002H	Miss
X		chr7:150179945C>G	chr7:150742297-150742297	2018C>G	A673G	Miss
X		chr8:26778286G>T	chr8:26722369-26722369	118G>T	G40W	Miss
X		chr6:31783819C>G	chr6:31675840-31675840	575C>G	P192R	Miss
X		chr1:7730843A>G	chr1:7796577-7796577	3240A>G	L1080L	S
X		chr22:40851168G>A	chr22:42526670-42526670	124G>A	G42R	Miss
X		chr1:47323585G>A	chr1:47611565-47611565	1250G>A	G417D	Miss
X	X	chr2:118291285G>A	chr2:118575055-118575055	121G>A	G41R	Miss
X		chr19:19104498A>T	chr19:19243498-19243498	254A>T	E85V	Miss
X		chr19:19104499G>T	chr19:19243499-19243499	253G>T	E85X	N
X		chr5:43541741C>G	chr5:43505984-43505984	798C>G	S266R	Miss
X		chr9:37730240G>A	chr9:37740240-37740240	1715G>A	G572D	Miss
X		chr1:180641553G>A	chr1:183909896-183909896	1423G>A	V475I	Miss
X	X	chr3:169233251C>T	chr3:167750549-167750549	935C>T	A312V	Miss
X	X	chr3:169233252G>C	chr3:167750550-167750550	934G>C	A312P	Miss
X		chr7:23086956G>T	chr7:23313716-23313716	1556G>T	S519I	Miss
	X	chrX:53537429G>A	chrX:53520704-53520704	1442G>A	R481K	Miss
X		chr11:9418649G>T	chr11:9462073-9462073	2767G>T	A923S	Miss
X		chr10:363080G>A	chr10:373080-373080	3790G>A	V1264M	Miss
X		chr2:48826897G>A	chr2:48915246-48915246	1690G>A	D564N	Miss
X		chr17:18080933C>G	chr17:18140208-18140208	1566C>G	L522L	S
X		chr19:40904380C>T	chr19:36212540-36212540	2291C>T	P764L	Miss
X		chr19:55650351C>T	chr19:50958539-50958539	2189C>T	P730L	Miss
X		chr6:84706496G>T	chr6:84649777-84649777	1009G>T	D337Y	Miss
X		chr17:23118990G>T (homozygous)	chr17:26094863-26094863	2035G>T	A679S	Miss
X		chr5:359875G>T	chr5:306875-306875	367G>T	G123C	Miss
X		chr16:68734948A>T	chr16:70177447-70177447	1637A>T	Y546F	Miss
X		chr8:22638372G>C	chr8:22582427-22582427	446G>C	R149P	Miss
X		chr1:183651507C>G	chr1:186919850-186919850	1326C>G	H442Q	Miss
	X	chr20: 8667928C>T	chr20:8719928-8719928	2229C>T	A743A	S
X		chrX:114703778A>C	chrX:114880798-114880798	1454A>C	D485A	Miss
X	X	chr1:56881987C>A	chr1:57169966-57169966	1111C>A	P371T	Miss
X		chr1:19237486G>A	chr1:19492180-19492180	4181G>A	R1394H	Miss
X		chrX:84168729C>G	chrX:84362584-84362584	830C>G	S277X	N

X		chr2:165812042A>G	chr2:165986535-165986535	2837A>G	E946G	Miss
	X	chr16:18730825A>C	chr16:18823324-18823324	10735A>C	K3579Q	Miss
X		chr10:108579379A>C	chr10:108589389-108589389	669A>C	K223N	Miss
X	X	chr1:16002504G>T	chr1:16257198-16257198	4463G>T	R1488I	Miss
X		chr1:155422865A>T	chr1:158609792-158609792	4743A>T	Q1581H	Miss
X		chr2:37318343A>C	chr2:37406692-37406692	324A>C	A108A	S
X		chr14:99871016G>C	chr14:100801263-100801263	1365G>C	E455D	Miss
X		chrX:46144024G>A	chrX:46387770-46387770	253G>A	E85K	Miss
X		chr1:109885704G>A (homozygous)	chr1:110173662-110173662	2285G>A	R762H	Miss
X		chr19:50140242C>A	chr19:45448402-45448402	224C>A	P75Q	Miss
X		chr6:32226401G>A	chr6:32118423-32118423	280G>A	A94T	Miss
X		chr1:117019184G>A	chr1:117307142-117307142	650G>A	C217Y	Miss
X	X	chr21:45146037_45146013delTGAACACGCACCCTGATAAGCTGCG	chr21:46321585-46321609	539_563delTGAACACGCACCCTGATAAGCTGCG	fs	INDEL
	X	chrX:54706426C>T	chrX:54689701-54689701	227C>T	P76L	Miss
X		chr13:94052277C>A	chr13:95254276-95254276	95C>A	T32N	Miss
X		chr11:57739474T>A	chr11:57982898-57982898	682T>A	F228I	Miss
X		chr11:31404466_31404470delTCTTG	chr11:31447890-31447894	304_308delTCTTG	fs	INDEL
X		chr11:64641527G>C	chr11:64884951-64884951	175G>C	A59P	Miss

Table A3.3. Coding sequence mutations in HCC1187 as reported by Wood et al (2007) and the COSMIC database