CHAPTER TWO: EXPERIMENTAL PROCEDURES

2.1 INTRODUCTION

Cancer is the ultimate genetic pathology; defined and characterised by an accumulation of somatic genomic aberrations, noted since the turn of the twentieth century. The relationship between exposure to DNA damaging agents and the subsequent accrual of mutations over an interval of time has been clearly documented. That a mutagen may attack individual sequence motifs and leave its imprint on a genome which may then be mitigated by the plethora of repair pathways present in the human cell has been acknowledged and will, here, be exploited as a mutational signature. These mutational signatures are inscribed layer upon layer on the cancer genome through the lifetime of the cancer patient.

Utilising the force and scale presented by next-generation sequencing technology, evidence for mutational signatures were explored from the wealth of data proffered by whole genome sequencing strategies in this thesis. However, in order to explore those mutational signatures, a set of high-confidence somatic mutations of all mutation classes for each breast cancer was essential for the nature of the downstream analysis intended. In order to obtain this dataset, a series of wet-bench and informatic procedures were performed and summarised in Figure 2.1. A more detailed description of each of these steps will be provided in the rest of this chapter.

An overview of the overall strategy was as follows:

- **Systematic re-sequencing using high-coverage paired-end next-generation sequencing technology**

DNA was obtained from twenty-one breast cancers and matched normal DNA from women diagnosed with breast cancers and systematic re-sequencing was performed from each of these samples using high-coverage, paired-end second-generation sequencing technology. The samples were obtained from the International Cancer Genome Consortium Breast Cancer Working Group according to local ethical approval. The full spectrum of histopathological subtypes of breast cancers were targeted and compared and contrasted to each other.

- **Employ bespoke bioinformatic algorithms to call all classes of somatic mutation and**

Bioinformatic algorithms were employed to map sequences back to the reference genome, call all variants in tumour and normal with subtraction of normal variation to generate comprehensive catalogues of somatic mutations. Further informatic tools were required to analyse and interpret all classes of somatic variants including substitutions, indels, somatic rearrangements and copy number aberrations. Post-processing filters were developed in order to obtain a dataset with high specificity and sensitivity. An orthogonal method (PCR, capillary sequencing, Roche pyrosequencing) was used to validate subsets of variants as being truly somatic in order to ensure high quality datasets for further analysis.

- **Extract and characterise patterns of somatic mutation and integrated analyses**

Having secured high-quality datasets, patterns of somatic mutation were sought taking types of mutation, sequence context and genomic architecture into consideration, in order to extract understanding regarding processes involved in initiating mutation and insights into DNA repair mechanisms. Excavation of these genomes included analyses of mutational rates and the timing of mutations through the evolution of the cancers. Transcriptomic profiling by expression arrays were performed in order to allow consideration of factors such as expression levels. Integrated analyses of different classes of mutation and transcriptomic data were performed to explore these relationships.

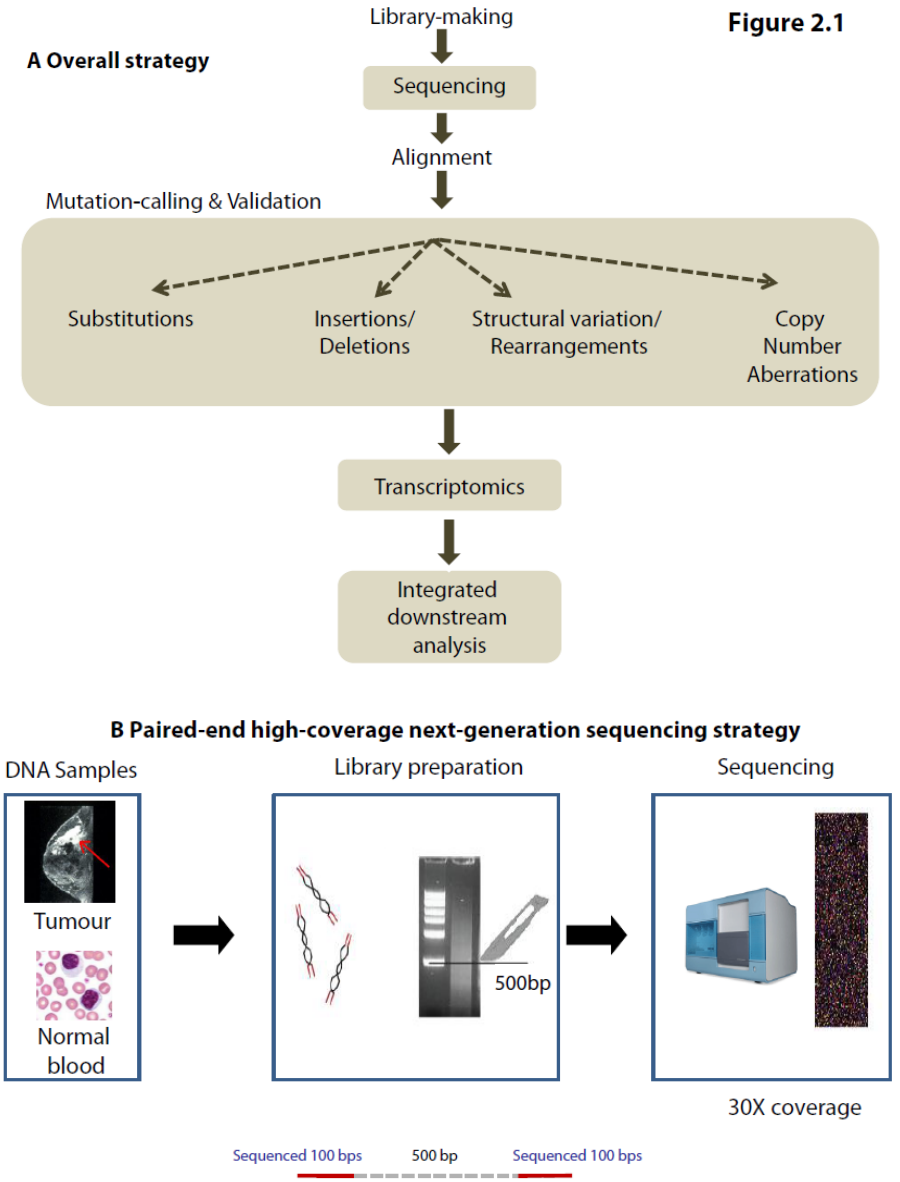

Figure 2.1: A flowchart of the full whole-genome sequencing and analysis strategy for twenty-one breast cancers. DNA obtained from collaborators was used to construct Illumina no-PCR libraries prior to sequencing on Illumina HiSeq 2000 sequencers. Raw sequences of the tumour sample and raw sequences of the normal sample were aligned back to the reference genome build 37 independently. All classes of somatic mutation including substitutions, insertions/deletions, somatic rearrangements and copy number aberrations were sought using a range of bioinformatic tools. Transcriptomics by expression arrays were also performed. A high quality dataset was obtained following post-processing or curation of the datasets, including validation on an orthogonal sequencing platform of a subset of substitutions and all insertions/deletions and rearrangements. The finalised dataset was used for all downstream analyses described in subsequent chapters of this thesis. (B) Paired-end next-generation sequencing strategy. A DNA sample was obtained from the breast cancer and from matched peripheral blood lymphocytes for each patient, fragmented using a Covaris Sonicator separately and following DNA preparation (end-repair, A-tailing and adaptor ligation), gel size-selected 500bp fragments to make a next-generation sequencing library. Each gel slice (library) contained billions of fragments of DNA and was representative for the entire genome of the population of cells in each cancer/matched normal sample. 100bp at both ends of each ~ 500bp fragment was sequenced. Each library was sequenced to generate enough raw sequence to ensure an average coverage of 30-fold per reference base in the genome, hence the term paired-end, high coverage next-generation sequencing strategy.

2.2 THE GENERATION OF ILLUMINA NO-PCR NEXT-GENERATION SEQUENCING LIBRARIES

Breast cancer samples included in this study had previously been subjected to pathology review by two pathologists independently scoring each sample, and only samples with >70% tumour cellularity were accepted for the project. DNA from tumour and matched normal samples were provided by collaborators of the International Cancer Genome Consortium (ICGC) Breast Cancer Working Group. The DNA samples provided were subject to local ethical approval of individual ICGC members. Illumina no-PCR libraries were generated from the DNA samples and a flow diagram of the principles of the library-making process is provided in Figure 2.2.

2.2.1 Starting quantity and fragmentation

Short insert 500bp library construction, flowcell preparation and cluster generation was in accordance with the Illumina no-PCR library protocol (Kozarewa et al., 2009). In brief, 5ug of DNA was brought to 120ul of T0.1E, transferred to a 150ul AFA Covaris vial and sealed. DNA was fragmented using a Covaris AFA DNA Sonicator. Fragment sizes ranging between 300 and 600bp were generated using the following shearing conditions:

- Intensity = 5
- 20% duty cycle
- 200 cycles/burst
- duration 30s
- at 4 ^{o}C.

This was followed by a purification step using a QIAquick protocol to result in 30ul of fragmented DNA. In brief, 600ul of PB buffer was added to the 120ul of fragmented DNA sample and the mixture (720ul) was added to a QIAquick column within a QIAquick tube. Centrifugation at 13,000RPM was performed for 1 minute in a benchtop centrifuge. Flow-through was discarded and the column holding the filter containing the fragmented DNA was replaced into the same tube. 750ul of PE buffer was added and centrifuged for 1 minute at 13,000 RPM. Flow-through was discarded and the column was replaced into the QIAquick tube again. An additional 1 minute of centrifugation was performed in order to remove excess fluid. The QIAquick column was now placed into a fresh tube and 32ul of EB buffer was placed onto the centre of the QIAquick membrane. Following a two minute wait, the

column was centrifuged for a further minute at 13,000RPM. The eluate containing the DNA was retained.

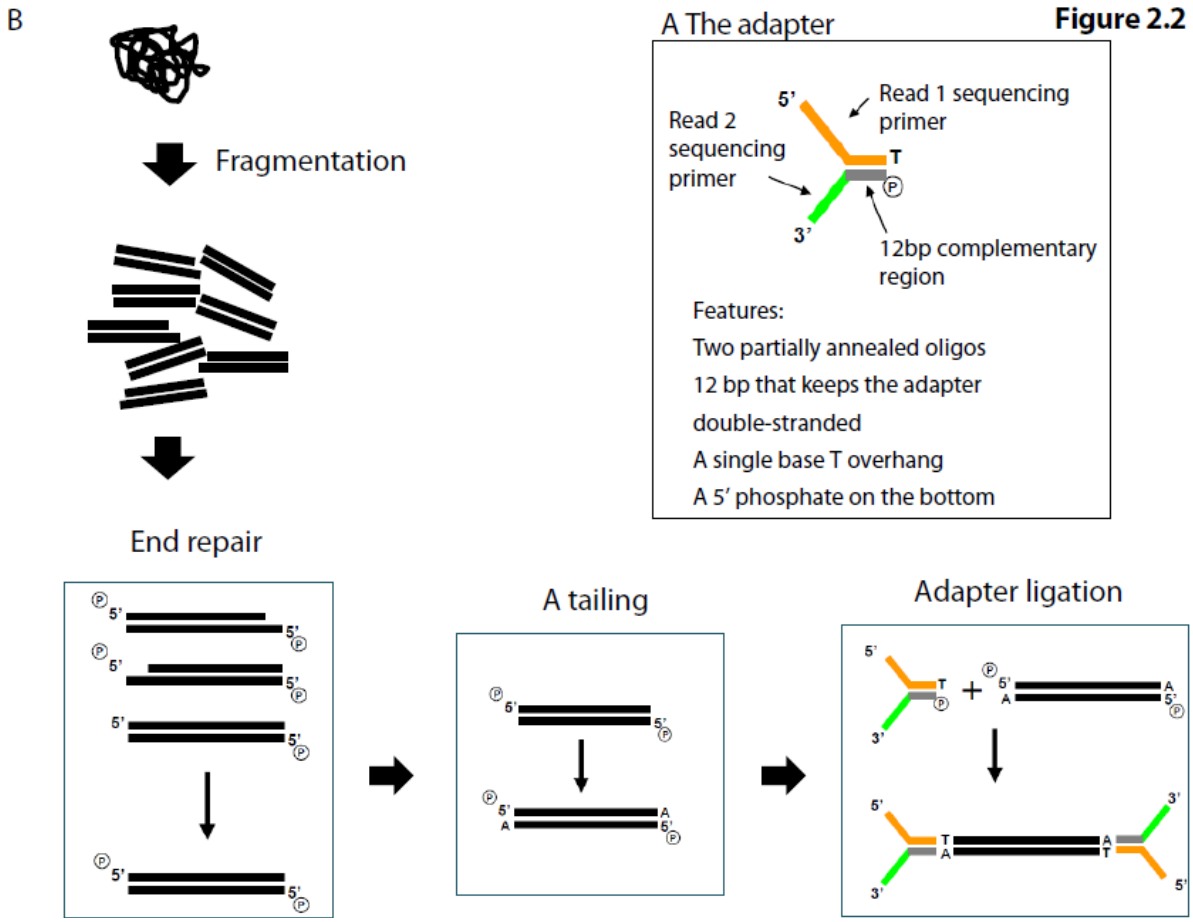

Figure 2.2: Flow diagram of the principles of the DNA preparation process. (A) The Illumina adaptor comprises two partially annealed oligos with a 12bp complementary region which keeps the adapter double-stranded. Sequencing primers for Read 1 and Read 2 of the sequencing process are embedded in the unannealed sections of the adapter. A single T base overhang is present at the 3' end with a 5' phosphate group present on the bottom. (B) DNA preparation of the library-making process. Genomic DNA sonically fragmented and the ragged ends of fragmented DNA are end-repaired to generate blunt ends (end-repair). An A base is added (A-tailing) to increase the efficiency of ligation prior to the adapter ligation step. The Illumina no-PCR library which is finally sent for sequencing therefore comprises billions of ~500bp fragments of DNA with adapters at both ends of each fragment.

2.2.2 End-repair, A-tailing and adaptor ligation

The fragmentation step generated double-stranded fragments which could have ragged edges and overhangs that could reduce the overall efficiency of ligation of the Illumina no-PCR adaptors. Therefore, end-repair and phosphorylation of fragmented DNA was performed using NEB reagents in the following quantities for each library:

No of samples	1
Water (ul)	45
T4 DNA ligase buffer (ul)	10
10mM dNTP Mix (ul)	4
T4 DNA Polymerase (ul)	5
Klenow DNA Polymerase (ul)	1
T4 PNK (ul)	5
Total (ul)	70
DNA (ul)	30
Total volume (ul)	100

QIAquick purification (as described previously) at this stage resulted in 32ul of end-repaired DNA, and was followed by A-tailing, a process which adds an "A" at the 3' end of the double-stranded fragments. This process was developed by Illumina and thought to increase the efficiency of the subsequent ligation step. NEB reagents were used in the following quantities:

No of samples	1
Klenow buffer (ul)	5
1mM dATP (ul)	10
Klenow exo- (ul)	3
Total (ul)	18
DNA(ul)	32
Total volume (ul)	50

A purification step was performed using MinElute columns. In brief, 250ul of Buffer PB was added to the 50ul of sample. The mixture (300ul) was added to a MinElute column, centrifuged at 13,000 RPM for 1 minute and flow-through discarded. 750ul of Buffer PE was added to the MinElute column and centrifuged for 1 min at 13,000 rpm. Flow-through was discarded again and an additional centrifugation was performed to remove excess fluid. Columns were placed into fresh tubes and 12ul of Buffer EB was placed onto the center of the MinElute membrane followed by another centrifugation. Eluate of 10ul of A-tailed DNA was retained

Ligation was carried out with the standard preparation of Illumina no PCR adapter oligo mix using the following quantities:

No of samples	1
Quick ligase buffer (ul)	25
Quick T4 DNA ligase (ul)	5
Total (ul)	30
DNA(ul)	10
Indexed Adaptor (ul)	10
Total volume (ul)	50

Purification was performed using Agencourt Ampure Magnetic beads (SPRI). In brief, 4ul of SPRI beads was added to each 50ul sample. The mixture was vortexed and left to stand for 5 minutes.

Tubes were placed in magnetic racks and left for 3 minutes or until the solution cleared. The clear solution was removed taking care not to displace beads (containing DNA). 200ul of 70% ethanol was added without disturbing beads, allowed to stand for 30 seconds and then gently aspirated and discarded. This was repeated once and the beads were then left to dry on a heated block for 5 minutes at 37^{o}C. 32ul of EB Qiagen solution was added, vortexed and the mixture spun gently. This was left to stand for a further 5 minutes. Tubes were replaced into magnetic racks and solutions left to clear. Clear fluid containing DNA was carefully aspirated and retained in a fresh tube.

Checks were performed by electrophoresis using a DNA 1000 chip on an Agilent Bioanalyser, at each step of the library preparation process to ensure recovery of library and to check overall distribution of fragment sizes obtained (Figure 2.3).

2.2.3 Gel-size selection

50ml of a 2% agarose gel in 1X TAE (1g agarose) suitable for 1 mini-gel, was prepared for each library. Libraries were loaded after mixing with 6X loading dye and run according to the following parameters:

- 60V
- Duration 2 hours
- Chilled at 4 ^{o}C and replaced at 1 hour
- In 1X TAE buffer

For a 500bp library, gels were size-selected at ~700-750bp. Gel slices immediately above and below this bandwidth were also archived in case they were required for the future. Extraction of DNA was performed using Qiagen gel extraction kit. Electrophoresis using a High-Sensitivity chip on an Agilent Bioanalyser was performed to ensure that the library was captured and to ensure that the modal fragment size was in the order of 400-500bp.

Figure 2.3

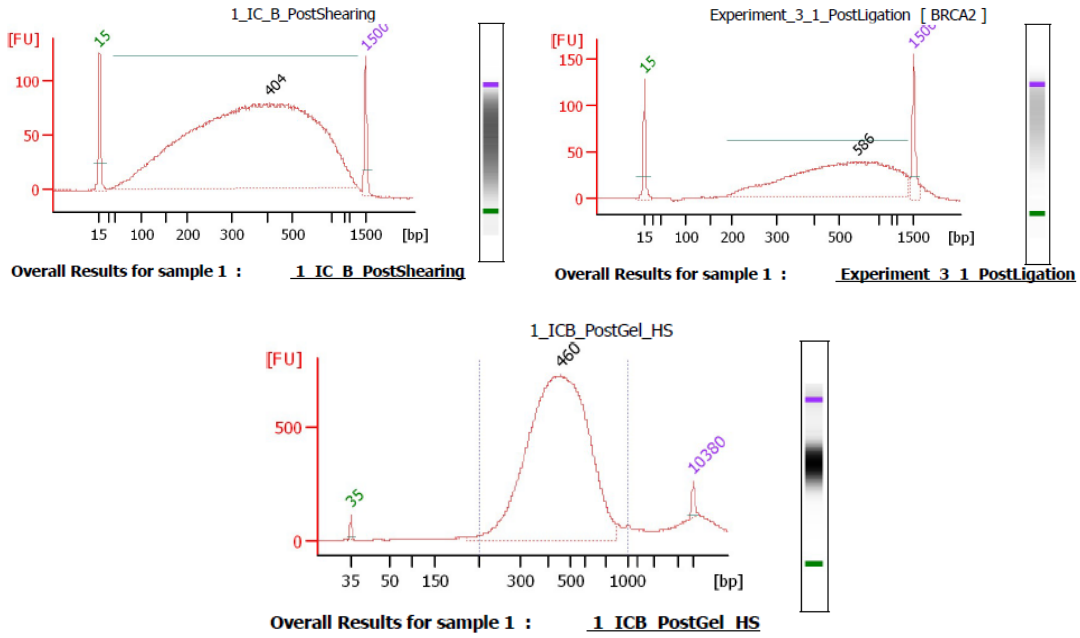

Figure 2.3: Typical Agilent bioanalyser traces (A) post-shearing (B) post end-repair, A-tailing and post adaptor-ligation (C) post gel size selection. Desirable features sought included the correct range of fragment sizes and the absence of a shoulder of remaining adapter or adapter-dimers.

2.2.4 Library quantification using quantitative PCR and sequencing

Illumina library quantification was performed using a real-time PCR assay in order to measure the quantity of fragments which were properly adapter-ligated and the result was used to determine the quantity of library necessary to produce the appropriate quantity of clusters on a single lane of the Illumina GAIIx or Illumina HiSeq. Flow-cell preparation was performed according to the manufacturer's protocol within the sequencing facility. Whole-genome sequencing was performed by the Wellcome Trust Sanger Institute core sequencing facility.

2.3 NEXT-GENERATION SEQUENCING

2.3.1 The principle of Illumina-based next-generation sequencing technology

The principle underlying next-generation sequencing technology (NGS) is not dissimilar to capillary sequencing. The bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is re-synthesised from a DNA template strand. However, in capillary electrophoresis sequencing, one averaged signal is obtained as a representation of a single sequencing reaction from many hundreds of DNA molecules which are mixed in solution. In contrast, NGS obtains many millions of signals across millions of reactions in a massively parallel fashion, with each reaction fixed to a single location on a sequencing chip. This advance enables rapid sequencing entire genomes, with the latest instruments capable of producing hundreds of gigabases of data in a single sequencing run.

2.3.1.1 The modified nucleotide

Standard Sanger capillary sequencing depends on the incorporation of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. The chain-terminating nucleotides lacks a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, and results in termination of DNA strand extension and DNA fragments of varying length. The four types of dideoxynucleotide chain terminators are labeled with fluorescent dyes, each of which emits light at different wavelengths and are thus detectable in an automated fashion following separation by gel electrophoresis.

A development that permits the massively paralleled sequencing approach is the advent of the *reversible* terminator nucleotide. Here, the extension of each DNA molecule occurs base by base. Laser image detection of each fluorescently-labelled nucleotide is followed by cleavage of the fluorescent dye and reversal of the 3’-terminator. Addition of 3’-OH group then allows the extension of the same molecule. This is called sequencing by synthesis and is summarised in Figure 2.4.

Figure 2.4

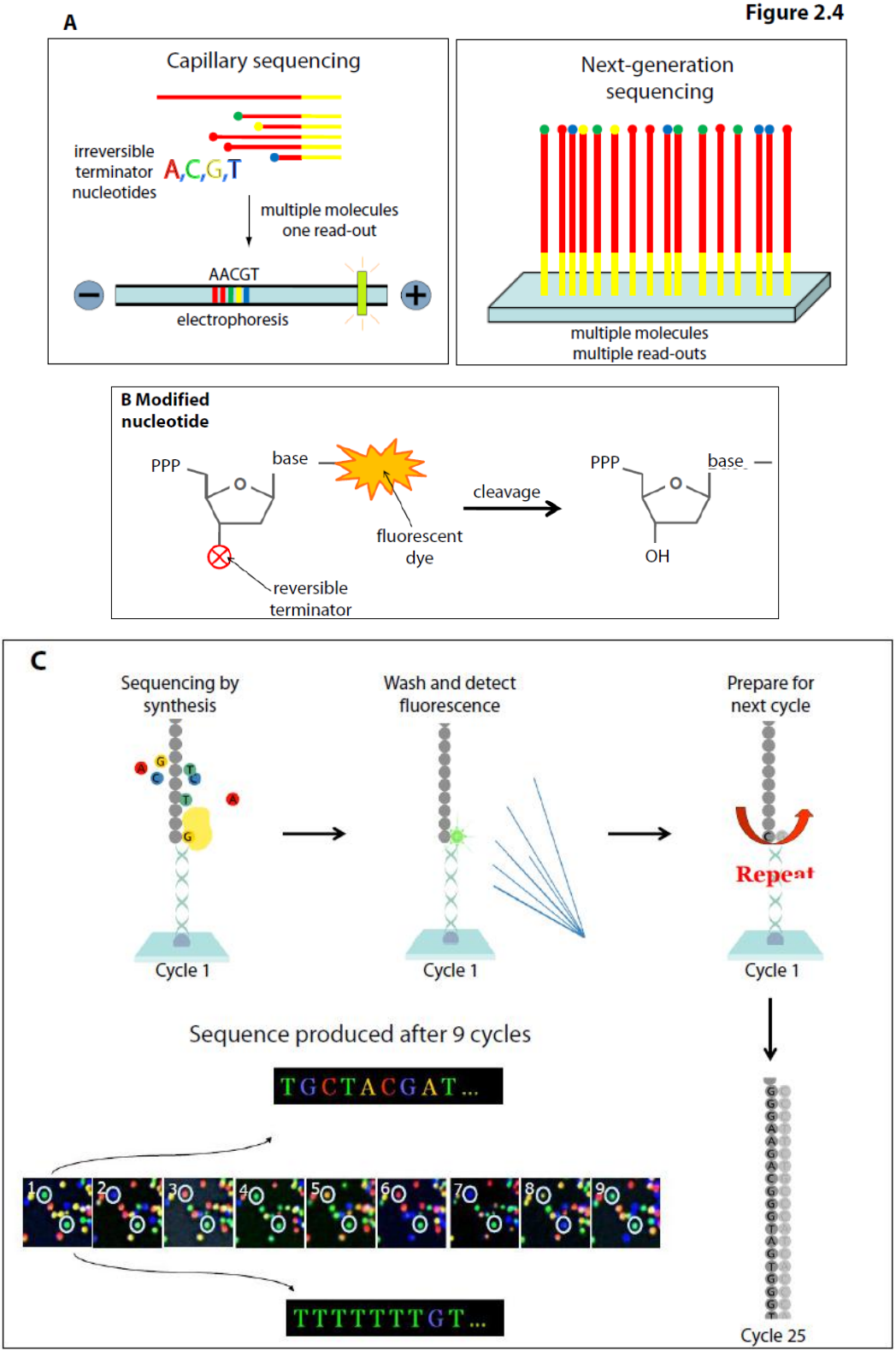

Figure 2.4: The principles of next-generation sequencing (A) In capillary sequencing, many thousands of DNA molecules in solution produce an averaged single read-out after electrophoretic separation. In NGS, many molecules immobilised on a flowcell produce independent read-outs thus increasing the scale of sequencing considerably. (B) The Illumina modified nucleotide contains a reversible terminator at the 3' end and a fluorescent dye which can be cleaved, to allow sequencing-by-synthesis. (C) Sequencing by synthesis: The appropriate and complementary reversible terminator nucleotide attaches at the start of cycle 1. After a wash step, photo-detection of the fluorescence allows identification of the nucleotide which has attached. Cleavage of the reversible terminator and the dye occurs and the cycle is repeated. After multiple cycles, a string of sequence is obtained. (D) Fluorescence colour sequence read-out of clusters of individual fragments on a flowcell. Images adapted from Harold Swerdlow, with thanks.

2.3.1.2 Variations on next-generation sequencing: targeted strategies

The ability to obtain hundreds of gigabases of sequencing data allows re-sequencing of whole genomes in a single experiment. However, variations of this approach can allow defined regions in a genome to be sequenced. This targeted approach involves steps that enrich a library for the regions of interest. The steps involved in making a NGS library for target enrichment are virtually identical to the steps involved in library generation for whole-genome sequencing, with the addition of a target enrichment step where labeled custom-designed oligonucleotide baits (for the desired sequence) is hybridized in solution to fragmented genomic DNA and pulled-down using magnetic beads, capturing the targeted sequence. This approach is commonly used, particularly for sequencing the coding sequences of the human genome and is referred to as exome sequencing. Although, exome-sequencing is not a central part of this thesis, four of the breast cancers in this study were also exome-sequenced and the data were used as a comparison dataset in Chapter 3.

2.3.2 Platforms used in this study

Cluster amplification and 108 or 100 base paired-end sequencing was performed on Illumina GAIIx genome analysers or Illumina Hiseq 2000 analysers respectively, as described in the Illumina Genome Analyser operating manual. Standard quality control metrics including error rates, percentage of purity-filter reads and the total number of bases sequenced were used to characterize process performance prior to alignment. The Core Sequencing pipeline generated data files that contained the sequenced reads and associated qualities (*qseq* files).

2.3.3 Quality control measures

Further quality control metrics for each library and each lane of sequencing were determined within the Cancer Genome Project and were as follows:

- The modal peak of fragment insert sizes for each library was required to be in the region of 400-500bp in size (Figure 2.5a)
- The GC plot (Figure 2.5b) should not show any skewing for or against GC
- The base qualities per cycle plots (Figure 2.5c) should show a high proportion of bases of a good minimum base quality of preferably 25 (see chapter 3 introduction for definition) and above

Figure 2.5

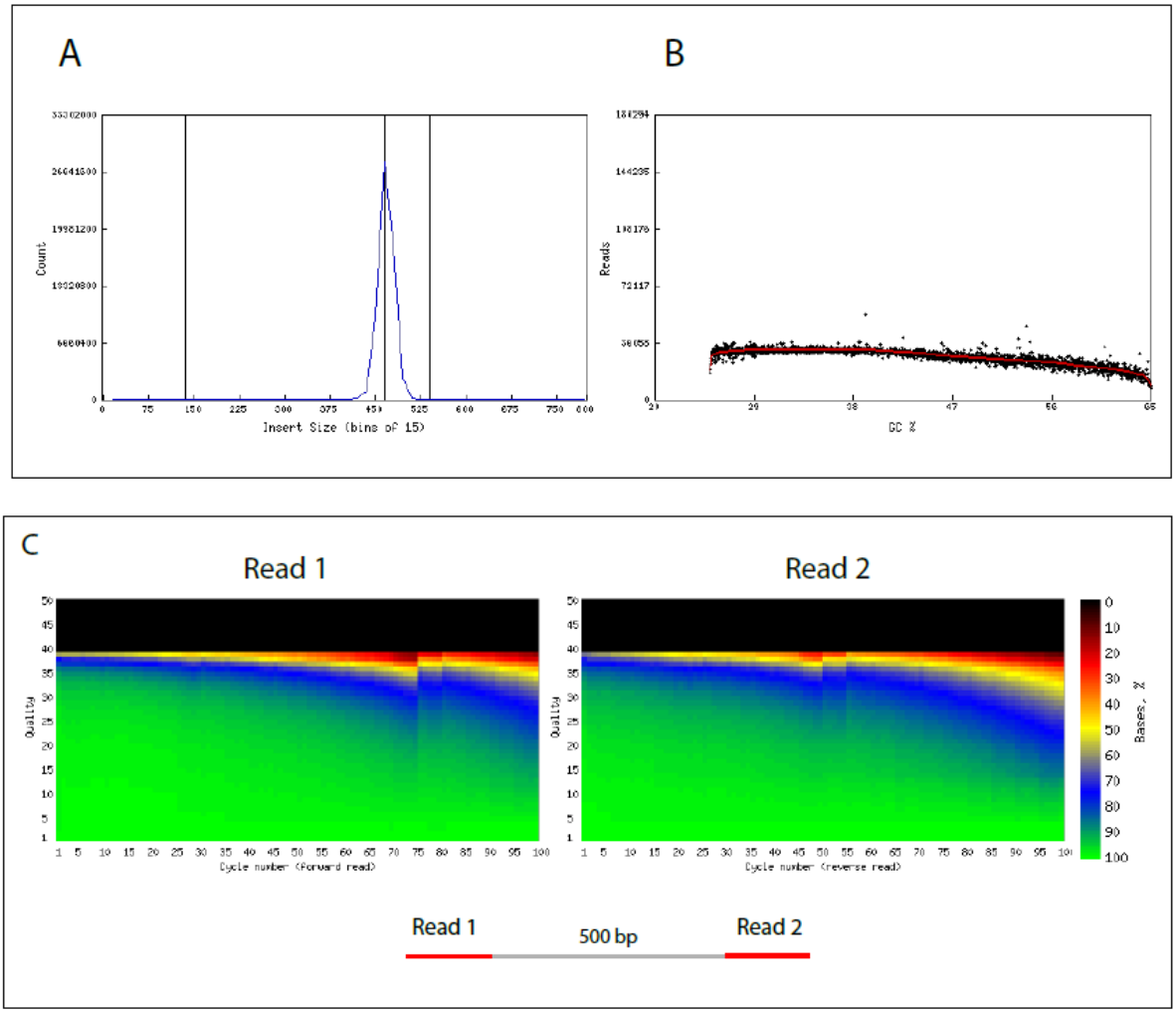

Figure 2.5: Library and sequencing QC metrics used within the Cancer Genome Project. (A) The modal peak of insert sizes should lie between 400-500bp with no evidence of "shoulders" suggestive of adapter contamination or smaller contaminating fragments in the library. (B) The horizontal axis shows the proportion of GC and the vertical axis demonstrates the number of reads. Because the genome differs in its GC content, a good library should have representation from all such regions showing a relatively uniform distribution of reads for all GC fractions. A pro-GC library would show a marked incline of the fitted slope and an anti-GC library would show a steep decline. (C) The base qualities per cycle of sequencing for both reads are demonstrated here. Green corresponds to 100% of bases with a minimum base quality which is on the vertical axis. The vast majority of both plots show green. The general decline towards the ends of reads is a well-known issue with Illumina sequencing. The step-wise change at cycle 75 in Read 1 and cycle 50 in Read 2 represents the laser-detector correction during the sequencing of both reads. This happens at a consistent time with the sequencing of each read.

2.3.4 The alignment of raw sequences to the reference human genome

The genomic DNA from a cancer sample is first fragmented into a library of small segments that can be uniformly and accurately sequenced in millions of parallel reactions. The newly identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold (resequencing), or in the absence of a reference genome (de novo sequencing). The full set of aligned reads reveals the entire sequence of each chromosome in the genomic DNA sample (Figure 2.8).

The raw data produced by the sequencers are contained in a *qseq* file which contains the quality scores, the precise location on the flow cell (lane and tile per lane), the sequencing run and the name of the sequencing machine used for each 100bp or 108bp sequence. Each *qseq* file was converted into a format that was more amenable to downstream manipulation called a *fastq* file. The *fastq* format essentially stored sequence information as concisely as possible but also included quality values for each of the bases sequenced in each read.

Figure 2.6

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure 2.6: The *fastq* format contains four lines per sequence: the first line begins with a '@' character and is a sequence identifier containing details regarding run, lane and position on the flowcell. The second line contains the raw sequence letters. The third line begins with a '+' character and is optionally followed by the same sequence identifier and any other optional description. The fourth line encodes the quality values in ASCII characters (character-encoding format) for the sequence in the second line, and contains the same number of symbols as letters in the sequence.

The sequencing data processing pipeline developed by the Cancer Genome Project starts with the short insert 108bp or 100bp paired-end reads and qualities in *fastq* format for all lanes and libraries generated for a single sample (either tumour or normal) and produces, after alignment to the reference human genome (NCBI37) using standard alignment software called Burrows-Wheeler Aligner (BWA), a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) which represents the sample. The final binary-coded BAM file therefore stores all reads with well-calibrated qualities together with their successful alignments to the genome.

Figure 2.7

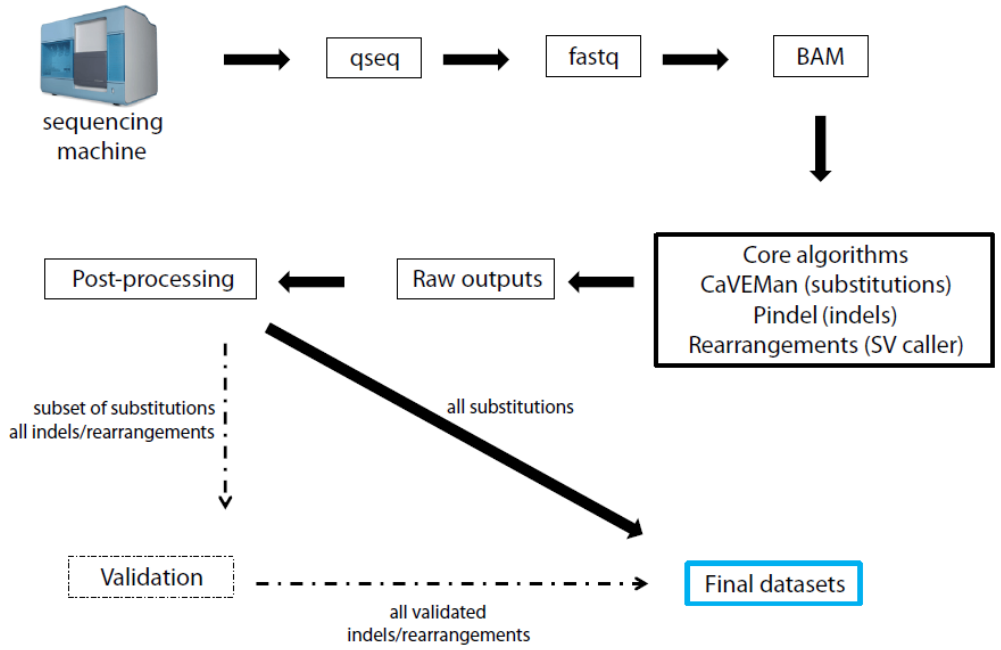

Figure 2.7: The generation of individual sample BAM files. Raw sequence data and quality scores were stored in *qseq* files. These were converted into *fastq* files which contained the sequence reads and qualities stored in a more concise format, amenable to economical storage and efficient computational manipulation. Data that passed QC were aligned back to the reference genome using BWA and a final BAM file was constructed for each library. BAM files were the input file for the subsequent step of calling somatic mutations. SV= structural variation.

2.4 THE PROCESS OF CALLING SOMATIC MUTATIONS

The sequenced reads from a cancer sample were aligned to the reference genome, and the sequenced reads from the matched normal sample were aligned separately to the reference genome. Therefore, two BAM files were generated per patient. The principle of calling somatic mutations involved identifying all variation in the cancer genome and the normal genome independently when compared to the reference genome, subtracting the normal variation (which would include all germline polymorphisms) to generate a final catalogue of somatic variation (Figure 2.8).

Figure 2.8

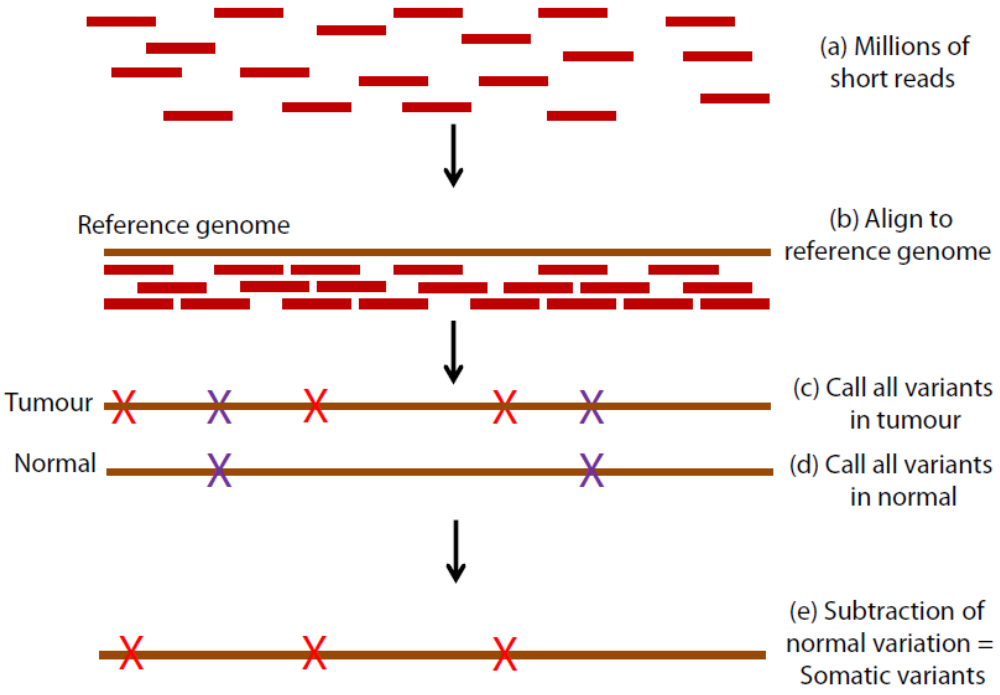

Figure 2.8: The principle of calling somatic mutations in cancer genomes. (a) Millions of short reads generated by sequencers were (b) aligned back to the reference genome separately in tumour and normal genomes. (c) All differences detected when comparing the tumour with the reference genome included somatic (red crosses) and germline variants (purple crosses). (d) All differences in the normal genome relative to reference genome were identified independently (purple crosses). (e) The germline polymorphisms in the normal genome were subtracted from the tumour genome to generate the catalogue of somatic variants for each breast cancer of each patient.

2.4.1 Genomic somatic mutation-calling

For the three mutation classes, substitutions, insertions/deletions and rearrangements, individual calling-algorithms were used. The input files for each of the calling algorithms were BAM files described in the previous section 2.3.4. Each mutation-caller generated a very large set of raw variant calls which comprised true somatic variants as well as a large proportion of false positive calls. In this section of the thesis, a very brief description of the principles of how each mutation-caller works is provided.

2.4.1.1 Substitutions

A bespoke algorithm, CaVEMan (unpublished) was used for calling somatic substitutions. Substitutions were identified, in principle, as alleles called in the tumour genome and not in the germline. Calls were made only from reads that mapped as linked pairs. Post-processing filters were developed to improve the specificity of mutation-calling and will be described in more detail in the next chapter. Copy number status (ploidy) and estimates of normal contamination from SNP6 data processed using ASCAT were used to enhance sensitivity and positive predictive value of substitution detection. CaVEMan will be described in more detail in the following chapter.

2.4.1.2 Indels

Insertions and deletions (indels) in the tumour and normal genomes were called using a modified version of Pindel (https://trac.nbic.nl/pindel/) 0.2.0 on the NCBI37 genome build (Ye et al., 2009). Pindel is a mutation-calling algorithm designed for the detection of small insertions/deletions, the breakpoints of large deletions, medium-sized insertions, inversions, tandem duplications and other structural variants at single-base resolution from next-generation sequencing data. It uses a pattern growth approach to identify the breakpoints of these variants from paired-end short reads. In this thesis, only the ability of Pindel to call small insertions/deletions was exploited, given that an alternative structural-variant caller was available and optimised for calling rearrangements.

During the preparation of a BAM file, all reads were mapped back to the reference genome. A subset of reads, however, mapped with either an indel within a read or could not be mapped although its paired-mate mapped correctly (unmapped singleton). This subset of reads was therefore potentially informative for indels. Pindel identified clusters of these informative reads, and used the mapped paired-mate to determine an anchor point in the reference genome. Having determined the anchor point and using *a priori* knowledge of the fragment insert size, Pindel worked out the orientation and the expected distance from the anchored read where the unmapped reads/reads containing the indel should be mapped. Pindel was able to split these informative reads into two (deletion) or 3 (insertion) smaller fragments, and attempted to align these in independent portions (Figure 2.9). Pindel called variants in tumour and normal separately but did not do a formal comparison. Post-processing filters were put in place to formally assist in the identification of somatic variants.

Somatic indels were required to be present in 5 reads or more in the tumour and not present in the matched normal sample. Variants were also screened against a panel of normal samples and were excluded if present in at least 5% of reads in at least 2 samples from this panel. Despite additional optimisation, the false positive rate in Pindel-called insertions/deletions remained high ~30%.

Therefore, all indels called by Pindel were validated and only validated variants were reported in this study.

2.4.1.3 Copy number

Copy number was determined using the Affymetrix SNP6.0 array for each of the twenty-one breast cancer samples. An informatic tool called "ASCAT" or allele-specific copy number analysis of tumours was used to estimate the fraction of aberrant cells and the tumour ploidy, as well as whole-genome allele-specific copy number profiles. ASCAT is an algorithm (Van Loo et al., 2010) that has considered and modeled the following two properties in cancer; that tumours often deviate from a diploid state (Holland and Cleveland, 2009; Rajagopalan and Lengauer, 2004) and that cancers are likely to comprise multiple populations of both tumour and non-tumour cells (Witz and Levy-Nissenbaum, 2006). ASCAT was therefore able to provide these estimates (Table 7.2) in the twenty-one breast cancers. These estimates were also used to optimise substitution-calling by CaVEMan (see section 3.3). Whole-genome allele-specific copy number profiles allowed regions of gains, losses, amplification and loss of heterozygosity (LOH) to be identified.

Figure 2.9

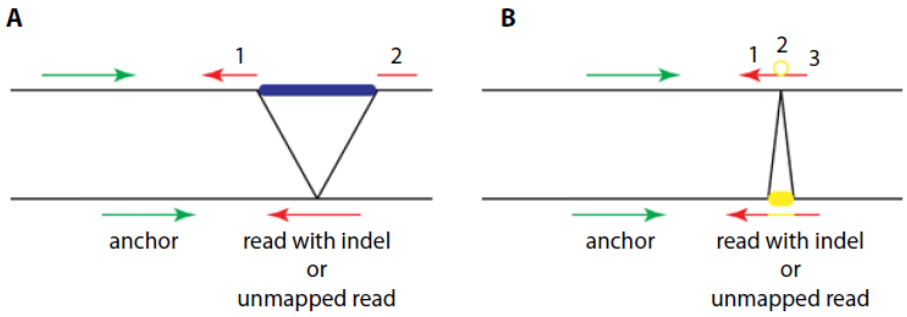

Figure 2.9: The basic principle underlying indel detection. Pindel detects simple deletions (A) and insertions (B) at nucleotide-level resolution (obtained from https://trac.nbic.nl/pindel/). Pindel identifies paired reads that are mapped but contain indels or paired reads with only one end mapped. Pindel uses the mapped read of the pair (green arrows) to determine an anchor point on the reference genome. A sub-region can then be located in the reference genome relative to the anchor read, where Pindel breaks the informative reads into 2 (deletion) or 3 (short insertion) fragments and maps these terminal fragments separately.

2.4.1.4 Rearrangements

Structural variants were called from discordantly mapping paired-end reads from short insert data using MAQ (Mapping and Assembly with Quality) alignments (Campbell et al., 2008; Stephens et al., 2009) (Figure 2.10 for summary). A set of optimisation filters and validation reduced the dataset considerably. Therefore, in order to improve sensitivity of detection, additional candidate structural variants were sought from within the proximity of copy number changes in the following way. All non-telomeric and non-centromeric coordinates of copy number changes were obtained from SNP6 data processed via ASCAT (Van Loo et al., 2010). Rearrangements close to copy number segmentation breakpoints were considered to be somatic if:

- copy number changes were identified for both rearrangement breakpoints and the sum of the distances between the rearrangement breakpoint and the copy number change was below 400kb or if
- a copy number change was identified in conjunction with only one rearrangement, then the distance between the rearrangement breakpoints and the copy number changes was less than 20kb.

If multiple rearrangements were identified for any copy number change, rearrangements closer to a copy number change were preferred over rearrangements further away.

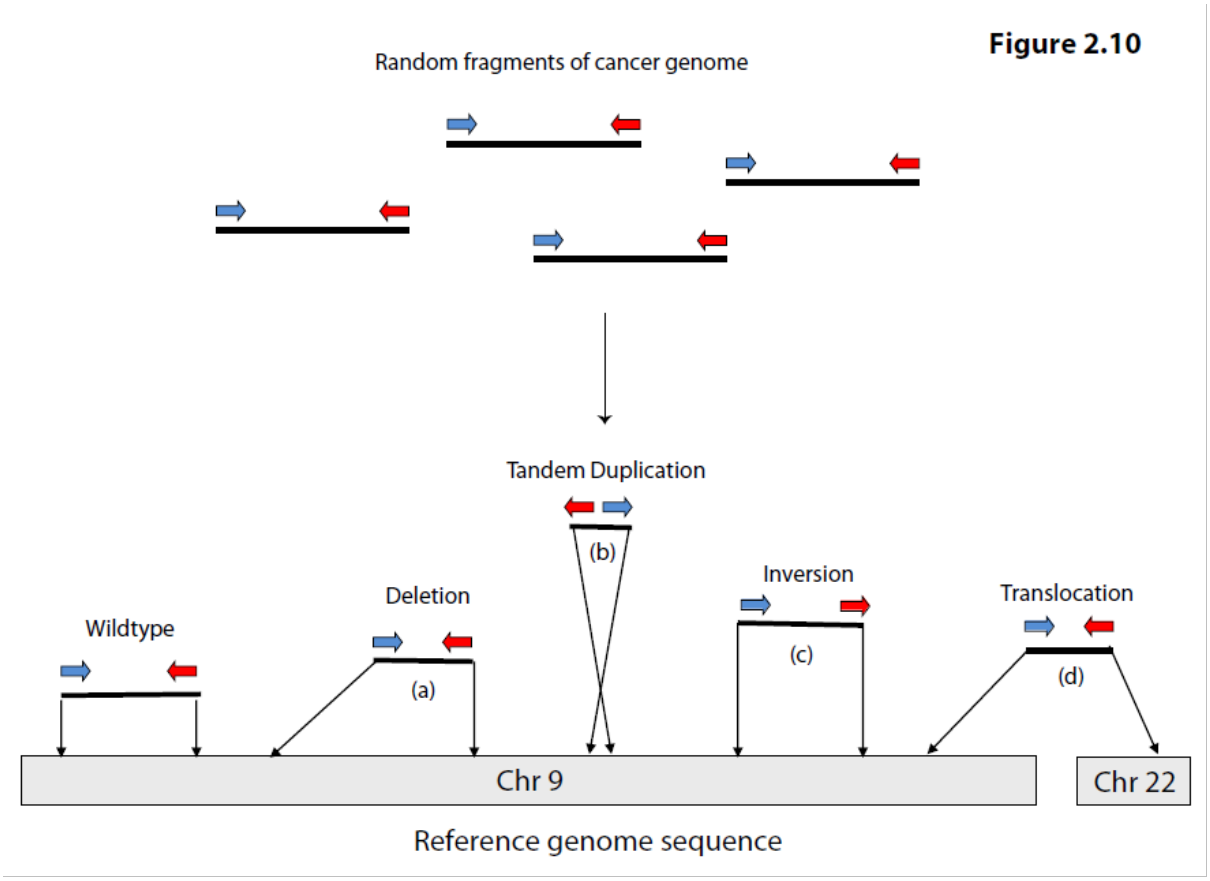

Figure 2.10: The classification of somatic rearrangements. Reads which were mapped at unexpected distances or in the wrong orientation were identified as discordantly mapping reads. Rearrangements were classified according to whether they were a (a) deletion: reads were closer together than expected and there was an associated copy-number change for variants > 200kb (b) tandem duplication: reads were further apart than expected and in the wrong orientation and associated with a copy number change for variants > 200kb (c) inversion: reads were not mapping in the appropriate orientation (d) translocation: reads were mapping to different chromosomes. *Amplicon-associated rearrangements involved reads within a region of high copy number, but are not depicted in this figure.

2.5 VALIDATION

In order to gain insight into the positive predictive value of the mutation-calling algorithms, validation experiments of a subset of putative somatic substitutions and all insertion/deletions and rearrangements were performed. Validation of putative substitutions was performed via Roche pyrosequencing (see section 2.5.2) in 20 tumour-normal pairs and capillary sequencing in 1 tumour-normal pair (PD3890a). All coding substitution variants and a random assortment of intronic and intergenic variants were selected for validation to make up to ~400 PCR products per sample. In addition to the set of variants selected for validation genome-wide, validation was also targeted to several hundred substitutions involved in regions of hypermutation and dinucleotides. The positive predictive value of the calling of substitution variants from the Illumina sequence reads was determined from the proportion of calls confirmed as somatic when sequenced on this orthogonal platform.

2.5.1 Capillary sequencing

Validation of variants was attempted by capillary re-sequencing of the tumour and normal pair. Capillary sequencing failed in ~20% variants. Two attempts at PCR validation for each variant were attempted. Somatic variants were required to be present in the tumour sample and absent in the normal sample. As mentioned previously, a capillary sequencing trace represents an averaged signal obtained from many thousands of DNA molecules in solution. It is believed that a variant has to be present at a sufficient proportion (> 10% of tumour cells) to be detectable by this method. It is therefore acknowledged that variants which are present at a low mutant burden may escape detection by this method of validation and may represent false negative calls.

2.5.2 Roche 454 pyrosequencing

Due to the large number of substitution variants, an alternative large-scale sequencing approach was favoured over the time-consuming, variant-by-variant approach of capillary sequencing. Similar to Illumina Sequencing, 454 sequencing involved a large-scale parallel sequencing approach and was able to generate roughly 400-600Mb of DNA per 10-hour run on a Genome Sequencer FLX using the GS FLX Titanium reagent series. 454 sequencing, also known as pyrosequencing, relies on fixing nebulised and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. The DNA fixed to these beads was then amplified by PCR. Each DNA-bound bead was placed into a ~29µm well together with a mix of sequencing reagents on a 454 PicoTiterPlate, which was essentially a fibre-optic chip. As a validation strategy, 454 pyrosequencing provided an alternative sequencing platform for targeted regions in the genome, allowing sequencing to high coverage. This alternative meant that variants were not subjected to the same systematic biases of Illumina sequencing during the validation step.

2.5.2.1 Library-preparation

Primers were designed to generate PCR amplicons for pyrosequencing of approximately 275-425bp. The PCR reaction was prepared in the following way:

No of samples	1
Whole-genome amplified DNA at 8ng/ul(ul)	4.5
Mixed primers at 4ng/ul (ul)	3
Buffer 10X (ul)	3.58
Taq polymerase(ul)	0.35
dNTPs at 1mM (ul)	3.58
Total volume (ul)	15

The PCR program was as follows:

- 95°C for 15 minutes
- 95°C for 30 seconds
- 60°C for 30 seconds
- 72°C for 30 seconds
- Repeated for 30 cycles
- 72°C for 10 minutes
- 4°C forever

Following the PCR reaction, the enzymes were inactivated using the standard protocol (ExoSAP-IT (Affymetrix)):

No of samples	1
PCR product (ul)	15
Reaction buffer (ul)	3
Dilution buffer (ul)	3.58
Exonuclease 20,000 U/ml(ul)	0.05
Antarctic phosphatase 25,000 U/ml (ul)	0.04
Water (ul)	8.9
Total volume (ul)	26

Using the following programme:

- 37°C for 30 minutes
- 80°C for 15 minutes
- 10°C forever

DNA was purified using Agencourt AMPure magnetic beads (using a similar protocol to that described in Section 2.2.2) and submitted to the 454 sequencing facility for adaptor ligation and sequencing on a Roche 454 Genome Sequencer FLX.

2.5.2.2 Raw data handling

Raw pyrosequencing files for tumour and normal samples were aligned to the reference human genome (NCBI37) using the genome alignment software Burrows-Wheeler Alignment with the addition of Smith-Waterman alignment to allow for longer read lengths (BWA-SW). Similar to Illumina reads, raw 454 pyrosequencing data was converted into pileup files for analysis (Figure 2.11).

Figure 2.11

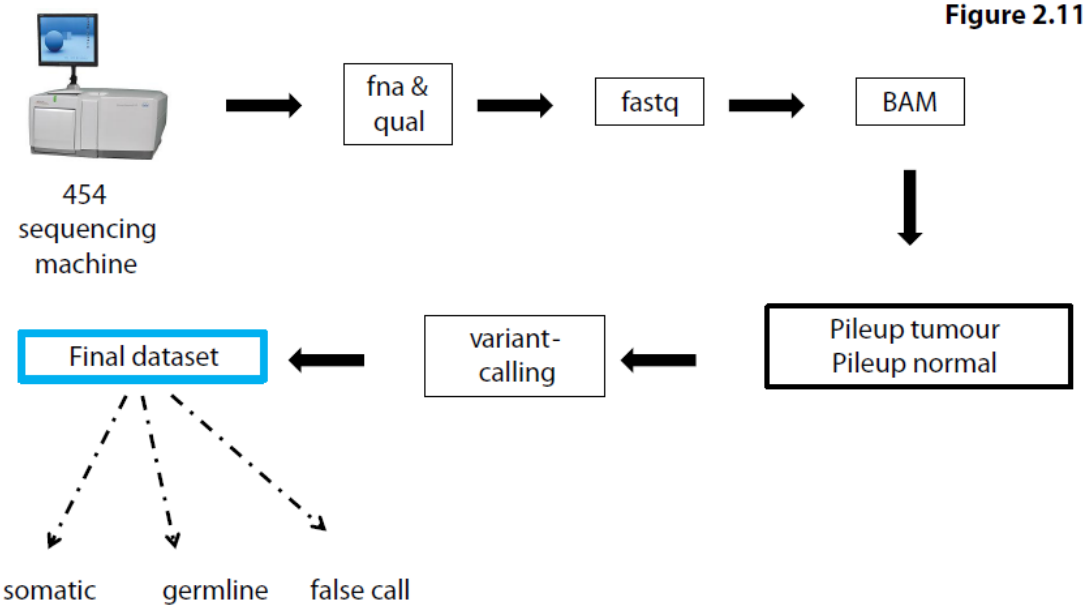

Figure 2.11: Data handling of 454 pyrosequencing output files. Sequence (fna) and quality (qual) files are converted into fastq and then BAM files, similar to the workflow for Illumina sequencing. Two separate pileup files are generated and variant-calling is performed across these two pileup files.

Pileup files were generated for each tumour and matched normal sample separately and variants were called as somatic if they were present in tumour and not in the normal. At least 25 reads of mapping quality of 20 and above and base quality of 25 and above were required to report each variant. To be considered as somatic, variants were required to be present in at least 5% of the reads in the tumour and not in the normal, or if present at a low mutation burden of < 5%, required chi-squared testing to assist in confirmation of somatic status. This imposition of relatively strict criteria could potentially generate false negative calls (true somatic variants called as tumour wild-type) resulting in an underestimation of the specificity of substitution-calling.

For pyrosequencing data, an average coverage of ~657X was achieved for each validated variant. A total of 6334 variants were amplified, of which 5561 met the aforementioned criteria. 4120 variants were found to be somatic, 26 were germline SNPs and 1395 did not show any evidence of the variant in the tumour or normal. The relationship between the variant allele fraction of the 454 experiment mirrored the variant allele fraction of the Illumina experiment in general (Figure 2.4).

Figure 2.12

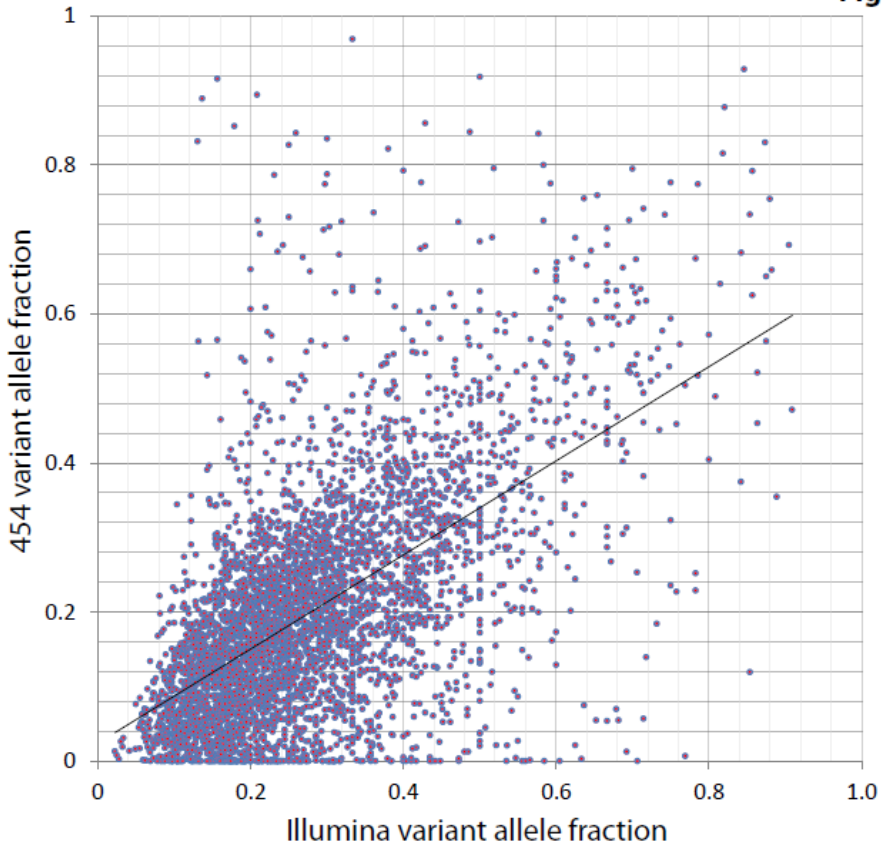

Figure 2.12: The variant allele fractions of the variants validated by 454 pyrosequencing were plotted against the variant allele fractions of the Illumina sequencing experiment to demonstrate that in general the representation of each variant in both experiments were correlated and therefore likely to represent the biological fraction of cells carrying the reported variant in the cancer (r=0.77).

2.5.3 Validation of somatic rearrangement

Structural variants were confirmed by custom-designed PCR across the rearrangement breakpoint (Campbell et al., 2008) or by local reassembly. Structural variants which were PCR-amplified were identified as putative somatic structural variants if a band on gel electrophoresis was seen in the tumour and not in the normal, in duplicate (Figure 2.13). Putative somatic structural variants were then capillary sequenced. Amplicons which were successfully sequenced were aligned back to the reference genome using Blat, in order to identify breakpoints to basepair resolution (http://genome.ucsc.edu/cgi-bin/hgBlat).

Figure 2.13

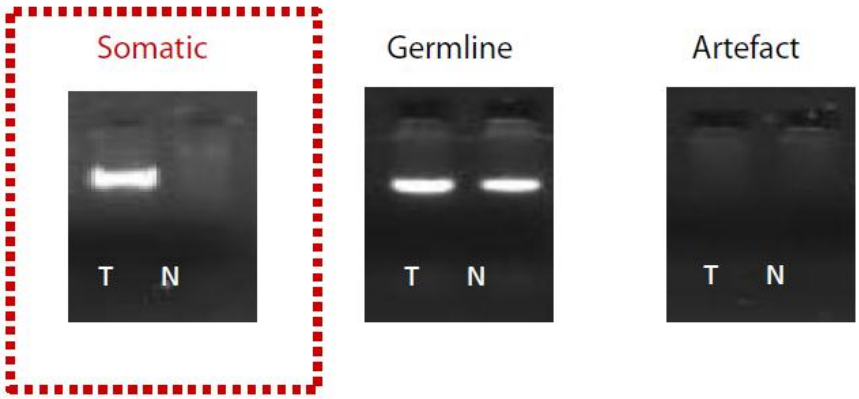

Figure 2.13: The validation step for putative somatic structural variation involved custom-designing primers to putative breakpoints and amplifying both tumour and normal DNA. A band appearing in the tumour (T) lane and not in the normal (N) lane, in duplicate, was taken for capillary sequencing.

For local reassembly, candidate rearrangements in regions of interest had been previously identified as rearrangements in close proximity to copy number changes. Discordantly mapping read pairs that were likely to span breakpoints, as well as a selection of nearby properly-paired reads, were grouped for each region of interest. Using the Velvet de novo assembler (Zerbino and Birney, 2008), reads were locally assembled within each of these regions to produce a contiguous consensus sequence of each region (Figure 2.14). Nearby properly-paired reads were added to increase coverage and to enlarge the resulting contigs. Heterozygous rearrangements, represented by reads from the rearranged derivative as well as the corresponding non-rearranged allele (Figure 2.14D), were instantly recognisable from a particular pattern of five vertices in the de Bruijn graph (a mathematical method used in de novo assembly of (short) read sequences) of component of Velvet (Figure 2.14C). Exact coordinates and features of junction sequence (e.g. microhomology or non-templated sequence) were derived from this. The exact breakpoints were identified by aligning to

the reference genome as though they were split reads. This local reassembly method continues to be under development at the present time.

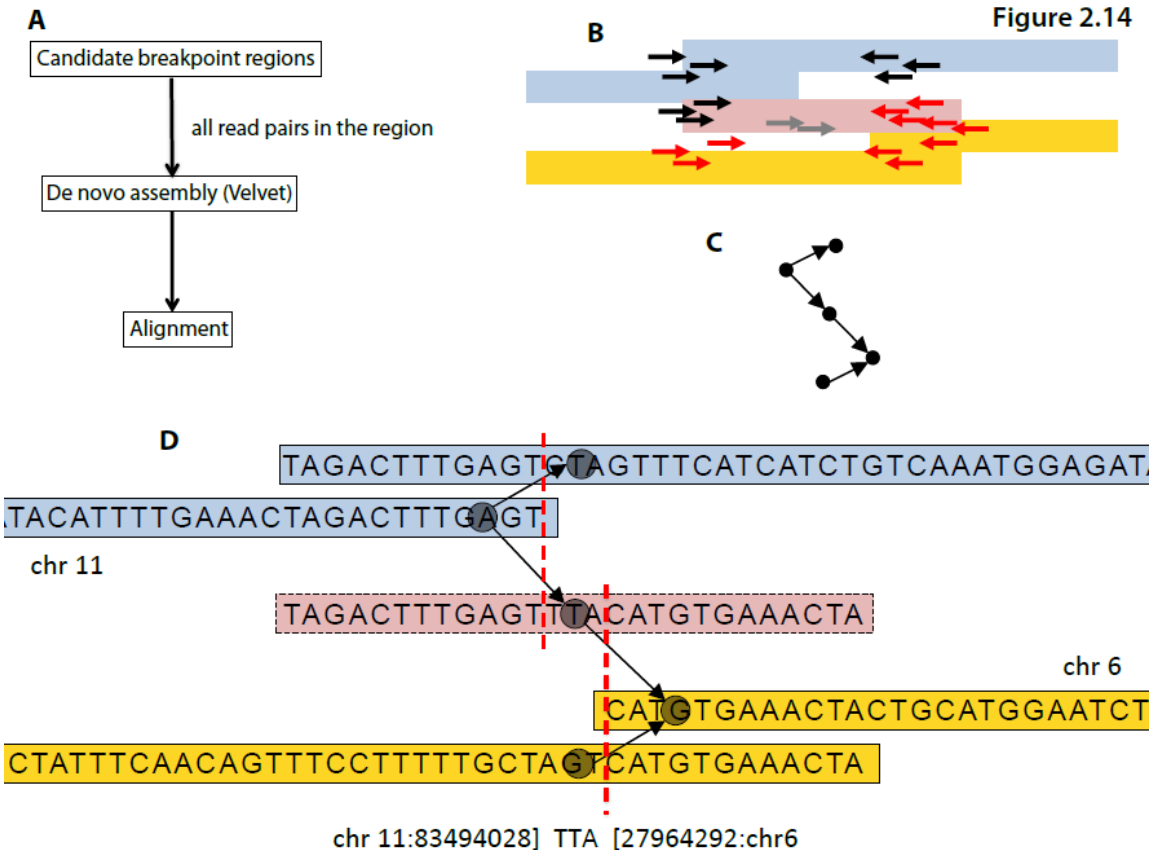

Figure 2.14

Figure 2.14: Validation of somatic rearrangement by local reassembly (Brass Phase II, under development). (A) Workflow of process of validation by de novo assembly. (B) Principle of local reassembly using a somatic interchromosomal translocation as an example. Read pairs are represented by two arrows facing each other on the same horizontal line. Black/red pairs in the centre represent pairs with ends on different chromosomes, i.e. are informative for a rearrangement. Nearby properly-paired reads were included for each region of interest (exclusively black pairs and exclusively red pairs). Reads with one end unmapped (grey arrows) spanning the breakpoints were also included. The coloured rectangles represent the contigs that Velvet was able to decipher: the two blue contigs represent one chromosome (11), and the yellow contigs represent another chromosome (6). The middle pink rectangle is the contig that reports the rearrangement. (C) The pattern of 5 vertices expected from the De Bruijn graph for successfully mapped rearrangement breakpoints. (D) Deciphering the rearrangement. Blue and yellow contigs were reported by properly paired reads. The pink contig reports the rearrangement. A successfully reassembled rearrangement breakpoint shows the pattern of 5 vertices. The breakpoint coordinates can be read from the pink contig. The lateral ends of the pink contig can be mapped back to the reference genome (chromosome 11 on the left and chromosome 6 on the right) until ambiguity is reached towards the middle of the pink contig. In this case, there is a stretch of non-templated sequence (TTA) in the middle of the breakpoint. The other possibility is that the two highlighted contigs meet in the middle and overlap by a few bases, which corresponds to a microhomology at the breakpoint.

2.6 STATISTICAL MEASURES

2.6.1 MONTE CARLO SIMULATION OF SUBSTITUTIONS

In order to assess the likelihood of some of the features or mutational patterns identified in the analyses which will be described in the following chapters, Monte Carlo simulations were performed for each cancer genome. The mutation prevalence of each mutation type (C>A, C>G, C>T, T>A, T>C, T>G) was obtained for each chromosome of each cancer genome. For each genome, 1000 simulations were then performed by generating mutations *in silico,* at the observed mutation rates. For each simulation, a variety of *in silico* parameters could be obtained and compared to observed features in each cancer genome. None of the simulations yielded mutational features according to the observed patterns, hence $p<0.001$ for the observed enrichment of each of those observed phenomena for each cancer genome.

2.6.2 GENERALISED LINEAR MIXED EFFECTS MODEL

Generalised linear models represent a class of fixed effects regression models for different types of dependent variables (including count data). Fixed effects models assume that all observations are independent of each other and are not appropriate for analysis of several types of correlated data structures, in particular, clustered data (where observed subjects are nested within larger units). Random effects can be added into the regression model to account for the variation in the correlation of the data. The resulting model is referred to as a Generalised Linear Mixed Effects Model which includes the usual fixed effects plus the random effects.

In order to examine correlations between mutation prevalence and gene expression as well as to consider transcriptional strand biases, a Generalised Linear Mixed Effects model was used for the analysis. For each mutation-type, the mixed effects comprised

- fixed effects: properties that were always present which here were the transcriptional strands and the expression levels
- random effects: the inter-sample variation.

The overall fitted curve for each mutation-type represents the combined effects across all seventeen cases for which there was expression data. The relationships described in Chapter 6 are based on the model performed here.

2.6.3 KOLMOGOROV-SMIRNOV TEST

The Kolmogorov–Smirnov test (K–S test) is a general nonparametric test which is sensitive to differences in shape of distribution functions between two groups and was therefore used to compare the empirical distribution functions of two groups. The null distribution of this statistic was calculated under the null hypothesis that the groups were drawn from the same distribution. The distributions considered under the null hypothesis were unrestricted, continuous distributions. The K-S test was used to compare differences in distribution functions between observed outcomes and expected outcomes assuming the latter occurred due to random chance (see chapter 7, sections 7.2 and 7.4).