UNIVERSITY OF MILANO-BICOCCA

Development and Implementation of Novel Applications of Massively-parallel Sequencing in Precision Medicine

by

Robert Sebastian Steinfelder

Department of Biotechnology & Biosciences
Translational and Molecular Medicine (DIMET)

Tutor: **Dr Antonella Ronchi**
Examiner: **Dr Nigel Mongan**

Cycle XXVII, 2014-2015

"The most valuable investment is that in the Human"

Jean-Jacques Rousseau (★1712 - †1778) philosopher, scientist and author

Contents

List of Figures

List of Tables

Abbreviations

bp	**b**ase **p**air(s)
kb	**K**ilo**b**ases
Mb	**M**ega**b**ases
NGS	**N**ext-**g**eneration **s**equencing
RNA	**r**ibo**n**ucleic **a**cid
DNA	**D**eoxyribo**n**ucleic **a**cid
SNV	**S**ingle **n**ucleotide **v**ariant
CNA	**C**opy **n**umber **a**lteration
InDel	**In**sertion or **Del**etion
TP53	**T**umour **p**rotein p**53**
KRAS	**K**irsten **ra**t **s**arcoma viral oncogene homologue
PTEN	**P**hosphatase and **ten**sin homologue
IHC	**I**mmuno**h**isto**c**hemistry
FISH	**F**luorescence ***i**n **s**itu* **h**ybridisation
qPCR	**Q**uantitative **p**olymerase **c**hain **r**eaction
RT-PCR	**R**eal-**t**ime **p**olymerase **c**hain **r**eaction

SNP	**S**ingle-**n**ucleotide **p**olymorphism
HRP	**H**orse**r**adish **p**eroxidase
ddNTP	**d**i**d**eoxy**n**ucleotide triphosphate(s)
Gb	**G**iga**b**ases
FFPE	**F**ormalin-**f**ixed and **p**araffin **e**mbedded
DIN	**D**NA **i**ntegrity **n**umber
BCL	**B**ase **c**a**l**ling format
BWA	**B**urrows- **W**heeler **A**ligner
OCT	**O**ptimal **C**utting **T**emperature compound
SAM	**S**equencing **a**lign**m**ent format
BAM	**B**inary **a**lign**m**ent format
IGV	**I**ntegrative **g**enomics **v**iewer
VCF	**V**ariant **c**all **f**ormat
SPRI	**S**olid **P**hase **R**eversible **I**mmobilisation
CRC	**C**olo**r**ectal **C**ancer
NSCLC	**N**on-**s**mall **C**ell **L**ung **C**ancer
BAQ	**B**ase **a**lignment **q**uality
emPCR	**e**mulsion **p**olymerase **c**hain **r**eaction
US	**U**niversal **s**equence oligonucleotide

Chapter 1

Introduction

Sequencing parts of the genome of cells showing abnormal growth behaviour, known as cancer cells, helped in understanding the key causes and mechanisms of how cells are transformed and evade growth and differentiation regulation. With help of early pioneers, such as Fredreick Sanger, technologies were developed, which are used reading genetic molecules. Nowadays, sequencing DNA fragments of up to a few hundred to a thousand nucleotides in size with less than one error in 10,000 bases is possible. Since this method, called Sanger sequencing, was published in 1977, technology has remarkably improved and replaced later by methods offering a faster throughput at a lower cost per base. Massively parallel sequencing became commercially available just over 10 years ago producing hundreds of millions of reads of under 300 bases in length, reaching out for the psychological target of

$1,000 as a cost for sequencing an entire human genome within a few days. Kits or panels to prepare *libraries*, a collection of DNA prepared for sequencing, are commercially available helping with automation, improve reproducibility and give the option to target specific regions of the genome to be sequenced. A few nanograms of extracted DNA from tumour tissue promises enough potential for a comprehensive prognostic and diagnostic prediction. Cataloguing of mutations and integration with association studies reveals someone's personal tumour evolution and pathologists can predict cancer sensitivities or resistances of commonly prescribed drugs. Kits for library preparation and sequencing are just about to be used in a clinical environment, meaning limitations and reliability need to be assessed, or validated, to prove its benefits for regular use in precision medicine. Utilising massively parallel sequencing involves a careful design of regions to choose being cost-effective but comprehensive at the same time. Further, sources of noise that can cause a bias in decision making, are important to be revealed and reduced to a minimum, where possible. Protocols for sequencing, data analysis and reporting are key factors that need to be defined to help clinical pathologists with diagnosis and support them in making decisions about individual cancer treatment.

After a brief overview of molecular cancer diagnostics and methods of cancer profiling, a custom designed cancer panel is introduced, described and discussed. The panel targets a selection of commonly reported cancer genes, which was assessed by applying it on 280 samples,

which were partially tested by a separate method named pyrosequencing. Subsequently a smaller custom design panel is described and assessed separately, as it is based on a different technology providing a shorter turnaround time and was designed to replace current assays for genetic marker testing in diagnostic medicine. Chapter four describes a protocol for introducing unique cellular barcodes by direct emulsion PCR providing a new approach for sequencing many single cells in parallel at a low cost. Finally the work is summarised and future perspectives are given.

1.1 Cancer in Humans

Cancer in all its different varieties and manifestations is one of the most common causes of human deaths worldwide. Motivation of understanding causes, development and researching new treatments is apparent. Over the past decades, the scientific community has revealed a tremendous amount of detail about the causes and development of cancer at different stages. Starting from fundamental basis that genetic material can act as a potential cause for cancer, by looking at malignant cells under a light microscope [40], which was later confirmed by transforming normal NIH3T3 cells into cancer cells with abnormal growth behaviour [57, 111]. Since sequencing technology became readily available, cancer formation can be traced back to its initial set of mutations that caused genetic instability and abnormal growth. Understanding

the evolution and why single mutations manifest among different types of cancer has been one of the biggest challenges in molecular medicine in recent years. In this section genomic evolution of malignant cells is explained based on random mutation events and selection.

1.1.1 Genomic Evolution of Cancer

During a life-cycle of an organism cells are exposed to a large number of random events of genetic change, called *somatic mutations*, such as single nucleotide variation (SNV), insertion or deletion (InDel) of bases, chromosomal rearrangement or copy number alteration (CNA). They differ from *germline mutations* shared by all cells in an organism, as they are inherited from the parental generation. Although most somatic changes are repaired, it may happen that a change can manifest and is passed on to the next generation after cell division. Neutral mutations have little if not any effect on cell physiology, while mutations resulting in a major disadvantage undergo negative selection and are eradicated sooner or later. The event of accumulating a combination of mutations, however resulting in a growth advantage, overcome the host's defence mechanisms and grow uncontrolled and invade surrounding tissue by metastasis is very rare [117]. A single mutation event occurs randomly across the entire genome at a fairly low rate. Current estimates vary between 1.1×10^{-8} and 2.5×10^{-8} per base-pair per generation, resulting in about $30 - 80$ mutations per cell [27, 20]. In order for a cell

to become malignant several independent mutations in certain genes are required, while only around 1% of the human genome encodes for genes [39]. Hence, accumulating a combination of mutations transforming a cell into cancer would be very rare, which disagrees with actual occurrence among the human population. This is explained by the fact that very early mutations in cancer pathogenesis increase mutation rate within a cell drastically, e.g. by disrupting a *tumour suppressor gene*, such as tumour protein p53 (TP53), which plays an important role in regulating cell division and DNA repair [63, 33]. Tumour suppressors prevent cells from turning into cancer by definition, when they are functional. They often play an important role in preventing permanent DNA damage, apoptosis or cell division. Indeed, cancer cells show a clearly higher mutation rate than normal cells, while tumour suppressors have been damaged by a mutation at the same time [39, 38]. It explains why certain cancer types often show a similar set of mutations, because a certain number of disrupted genes must have arisen at a very early stage causing an initial increase in the mutation rate. As cancer progresses, individual cells still collect further mutations independently supporting cell-to-cell heterogeneity within a tumour. Selection pressure causes cells carrying mutations providing growth advantages to out-compete others and potentially invade surrounding tissue eventually. Cell heterogeneity grows and population originated from one initial malignant cell changes over time. When treating a tumour with chemotherapy, a minor population of cell phenotypes can

show a resistance to a given drug. Whilst they have been suppressed by a majority of tumour cells being very sensitive to that drug, they receive a significant growth advantage, when the dominating population was eradicated. Evolution of tumours are not fully understood, due to its enormous complexity, although a great progress was made providing an insight into clonal evolution and mutation acquisition in cancer over the past years [87]. Two fundamental principles have been described so far: linear and branched cancer evolution [10]. In the model of linear cancer evolution, cancer progresses sequentially, as shown in Figure 1.1. Over time a single cell in a tumour acquires a beneficial mutation, followed by selection or genetic drift until the new population outcompetes an ancestral clone [9]. In the branched model of cancer evolution, however, subclones may grow independently whilst clonal displacement may be incomplete, see Figure 1.2. The branched model can be considered as a generalisation of the linear model of clonal evolution in cancer. Both models describe clonal heterogeneity increasing over time as more and more clonal subpopulations arise, while chances of total eradication of a subclone decrease.

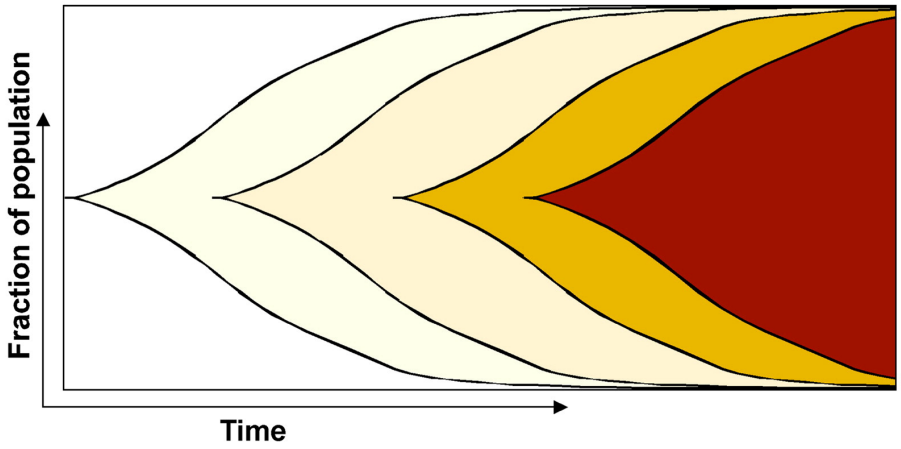


FIGURE 1.1: Model of linear evolution shows a sequential growth. Each newly arisen subpopulation shares mutations from its predecessor. Image from Marusyk et al. [79]

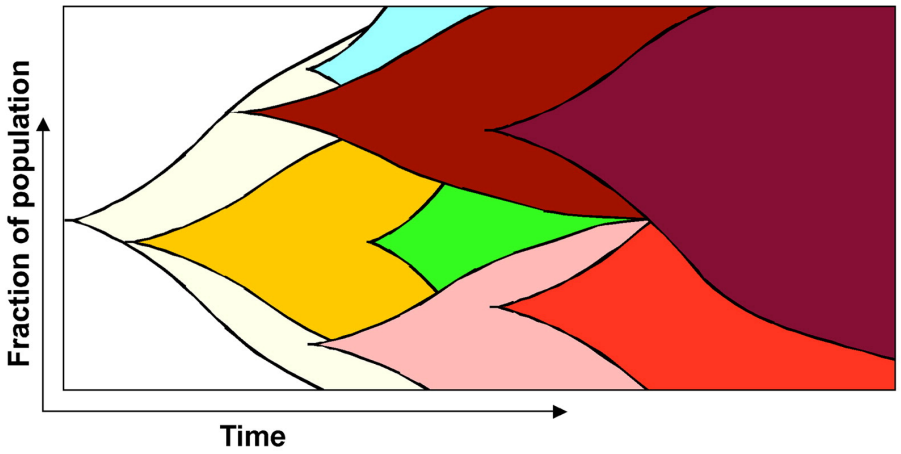


FIGURE 1.2: In the model of branched evolution subclones co-exist within a tumour. Image from Marusyk et al. [79]

1.1.2 Impact of Somatic Mutations in Cancer

Whilst all cancers carry somatic mutations some of them are crucial providing any growth advantage, called *driver mutations*. Other mutations that have been acquired as a side effect of genetic instability, but do not play any role in further development of the cancer are named *passenger mutations*. As outlined in Figure 1.3, driver mutations are always positively selected at any given stage of the cancer evolution and are usually, but not always, retained throughout all cancer stages. Passenger mutations show no or only very little functional impact and never contribute to cancer development by definition. Determination if a given mutation is a driver or a passenger often remains challenging [127].

Screening a patient for driver mutations is in most cases limited to regions where known driver mutations are often present. These regions are described as *mutation hot spots*. By limiting the search space, by looking for driver mutations that are commonly present and show clinical relevance is often sufficient, since the same cancer type in two patients often shows a related pathogenesis. A very popular region that is tested for driver mutations is the oncogene *Kirsten rat sarcoma viral oncogene homologue* (*KRAS*). Efficiency of established drug therapies highly depends on whether mutations are found in this gene or not. If functionality of KRAS is changed by a mutation, it is likely to be a driver mutation, resulting in resistance to drugs that inhibit *epidermal*

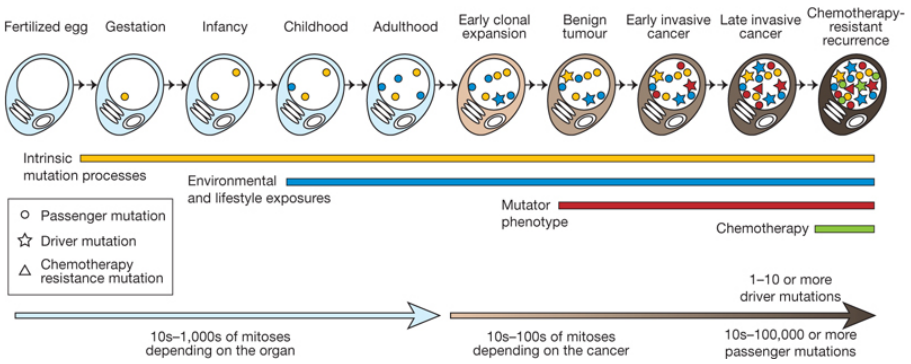


FIGURE 1.3: Schematic view of mutation acquisition in cancer from the early stage of a host to a late stage cancer (germline mutations are not shown). During life somatic mutations accumulate. During cancer development driver mutations are added over time causing clonal expansion. Tumour subpopulations may share a large set of these driver mutations. Image from Stratton et al. [117]

growth factor receptor (*EGFR*) signalling pathway, such as Cetuximab [71], as a permanently transcribed KRAS bypasses signal inhibition [107]. Hence, only patients without an activating mutation in KRAS should receive this drug as part of their treatment.

Zygosity can play an important role in predicting the impact of a mutation. Many tumour suppressor genes need all alleles to be affected before an effect can manifest, called the two-hit hypothesis, whilst mutations in oncogenes may already be effective with one allele changed. Due to an increased cell division and replication rate of malignant cells, they contain on average an increased number of copies of the genome compared to normal cells. It makes predicting of the functional impact of mutations more difficult if zygosity or ploidy has changed. Present

compound heterozygosity, where the same gene is mutated independently on two sites and in different copies, disrupting the function of the protein makes impact prediction even more challenging, as the expected effect is equivalent to a homozygous mutation, but not seen as such. There are cases known where zygosity of a mutation is of clinical relevance. One example is therapeutic targeting of *Phosphatase and tensin homologue* (*PTEN*) in a large variety of different cancer types. If PTEN is fully disrupted, an activation of the PI3K/AKT/mTOR pathway is the consequence, which can be targeted by specific drugs, such as BKM120 [24].

Due to positive selection of mutations supporting cell survival and resistance against apoptosis, individual tumour types can show a very similar set of driver mutations at a given stage. Hence, for drug sensitivity or resistance screening of a given tumour or tissue type it is very often possible to name a finite set of regions allowing sufficient prediction power for a comprehensive clinical diagnosis [77]. At present, described regions are often limited to genes, exons or even just a few codons of genes that are frequently mutated and are associated with drug sensitivities or resistances. Mutations in exons can have a direct impact on the functionality of a protein, as summarised in Table 1.1. Prediction of the severity of a mutation is often difficult and so is subsequent *in vivo* validation.

Mutation	**Effect**	**Example**	**Impact**[1]
Silent	Synonymous substitution	GCC (A) → GCA (A)	None
Missense	Non-synonymous substitution with amino acid of similar chemical properties	TTA (L) → ATA (I)	Low/ Moderate
	Non-synonymous substitution with amino acid of different chemical properties	CGT (R) → CTT (L)	Moderate/ High
Nonsense	Substitution to stop codon	TAT (Y) → TAA (STOP)	High
Frame shift	Nucleotide insertion/deletion causing frame-shift	TTA(L) → TATA (L?)	High
Inframe	Insertion/deletion of multiple bases keeping reading frame	TTATTA (LL) → TTA (L)	Moderate

Stop lost	Stop codon is changed to an amino acid	TAA (STOP) → TAT (Y)	High
Splice site	Non-silent mutation at a splice site	AGG (donor) → TGG	High

TABLE 1.1: A summary of mutations and their impact on the structure of a protein (e.g. amino acid changes). Functional effect prediction is very limited due to chemical complexity of proteins.

Finally, it should be pointed out that not only exonic mutations play a role in cancer progression and pathogenesis. Recent studies have shown that aberrations in regulatory elements of the genome, including non-coding RNA and epigenetic changes are important factors and will eventually lead to new approaches in cancer therapies in the future [129, 21, 93], although they have not been regularly used in clinical diagnostics, yet.

1.2 Diagnostic Biomarker Detection

Cancer diagnostics are essential for indentifying and locating a tumour as well as measuring progression stage and extent. For this, pathologists test for presence or absence of *biomarkers* [3], indicators for a certain state of a biological system -in this case- a tumour [91]. Biomarkers

[1] Potential impact on the protein

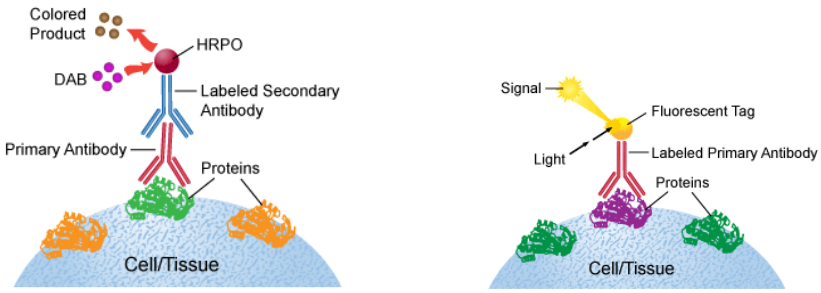
that allow effective and sensitive testing are eligible to be used for clinical diagnostics and prognostics. Testing for biomarkers aid in early detection, providing prognostic and predictive information for treatment selection and therapy guidance [53]. In most cases, it involves testing for absence or presence of a protein, RNA or for DNA, named *genetic marker*.

In this section an overview of popular methods for biomarker testing is given: immunohistochemistry (IHC), uses labelled antibodies to detect antigens or proteins that are produced exclusively by the tumour and is commonly used in diagnostics. A similar technique is *fluorescence in situ hybridisation* (FISH), where a short, highly specific DNA or RNA probe marks complementary genetic molecules to test for presence or absence of a DNA sequences or chromosomes of interest. *Quantitative polymerase chain reaction* (qPCR) or *real-time PCR* (RT-PCR) is a modification of the conventional PCR to detect presence or absence of DNA or RNA that are typical for cancer. Microarrays commercially rose in the 1990s and test for multiple DNA or RNA molecules in parallel. Sequencing technology reads out target regions, exomes or the entire genome or transcriptomes from a tissue section directly.

1.2.1 IHC & FISH

Despite the fact that IHC was introduced almost 80 years ago, it is still considered the gold standard for *in situ* detection of many biomarkers

in cancer. It has proven to be a reliable method to check for presence or absence of proteins that are exclusive for cancer cells [97]. A manufactured antibody is combined with a colour-producing enzyme or fluorophore and subsequently viewed under a microscope, as outlined in Figure 1.4. The efficiency of this method is mainly based on the binding affinity of the chosen antibody and availability of an antigen. Preferably only antibodies are selected that show a high binding specificity, to make biomarker testing robust.



(A) Indirect labelling of proteins with a pair of primary and secondary antibodies. Oxidising horseradish peroxidase enzyme (HRPO) catalyses inactive 3,3'-Diaminobenzidine (DAB) to emit coloured light.

(B) Direct labelling of proteins with a single antibody with attached tag, which can be detected under a fluorescence microscope.

FIGURE 1.4: Illustration of IHC with either a pair of primary and secondary antibodies for higher sensitivity or using a light emitting fluorophore. Image from Leinco technologies™ [118]

It is possible that due to sample preparation methods antigens can be masked and binding reaction is, therefore, insufficient. Another common issue is poor quality of the light emitter that can lead to false

results. By testing positive and negative controls in the same experiment confidence can be increased [15]. IHC for biomarker detection is described as a precise, reproducible, method that helps answering key clinical questions, whilst being often more cost-effective than other approaches [119]. Although, it is technically feasible for automated screening of treated tissue, high-throughput screening of many markers is less feasible.

A related technique is FISH, which was introduced in 1982 [61]. The fundamental idea is based on tagging a fluorescent dye to a short DNA or RNA probe. If a designed probe binds to the complementary sequence, the fluorophore can be spotted under a fluorescence microscope, as sketched in Figure 1.5. Using multiple dyes of different spectra, several loci can be tested for individually. FISH is commonly used for screening chromosomal rearrangements, relocations or deletions. It is further possible to use RNA probes for testing expression [59]. It is important that probes bind uniquely and strongly to the locus, otherwise the result may lead to false conclusions. Resolution of relocations is down to about one *megabase* (Mb), but largely depend on the quality of probe and fluorophore. As with IHC, FISH is suitable for testing clinically relevant aberrances due to high sensitivity and robustness at low cost [102]. For an increasing number of markers automation is necessary [74].

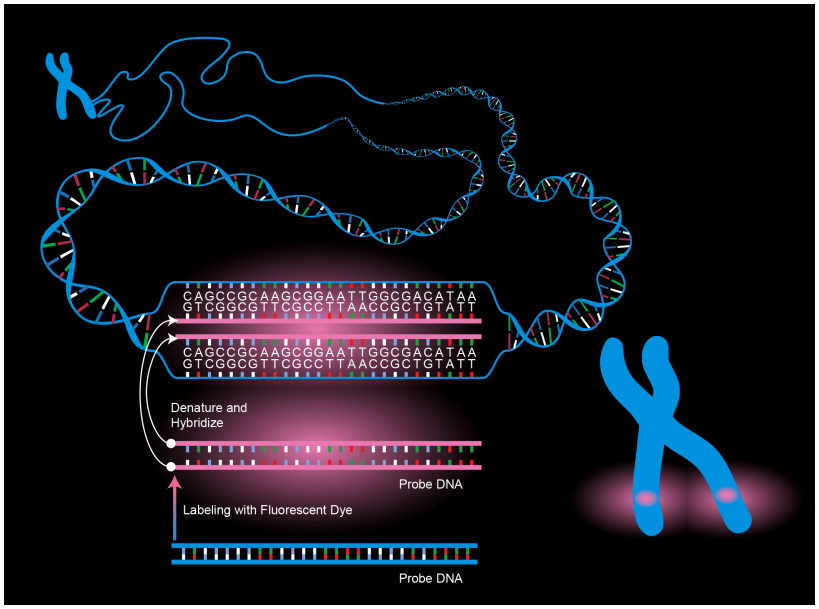


FIGURE 1.5: FISH is based on tagging a probe with a fluorescent dye. The probe hybridises to a complementary sequence, which is then viewed under a microscope. The technique is great for detecting chromosomal re-arrangements, copy number alterations or deletions. Image from NIH public gallery [62].

1.2.2 Quantitative PCR

Quantitative (qPCR) or real-time PCR (RT-PCR) allows DNA or RNA fragments to be amplified and quantified in one step [43]. Conventional PCR amplifies double- or single-stranded DNA or RNA templates, by using a DNA polymerase and a pair of primers, flanking the region of interest. A sequential number of usually 20-30 heat controlled reaction cycles of *denaturation*, *annealing* and *extension* doubling template after every cycle. A PCR can be modified to obtain quantitative information with help of fluorescent dyes transmitting a light signal relative to the quantity of template [58]. The fundamental principle of quantitative PCR is based on a fluorescent dye bound to a probe of oligonucleotides that shows no, or very little fluorescence in the inactive state. If a probe hybridises to a complementary DNA or RNA sequence, the dye transforms into its active state, causing light emitting at a certain wavelength that can be detected. As outlined in Figure 1.6, light intensities initially show exponential growth, followed by a plateau phase. After every discrete time point within the exponential growth phase fluorescence can be measured between two samples. Ratio at a given cycle is constant and is interpreted as the relative difference between the quantity of input template. Further improvements have been made over the years, such as TaqMan® assays, detecting *small nucleotide polymorphisms* (SNPs) or somatic mutations and small InDels by designing probes that only amplify specific genotypes if present.

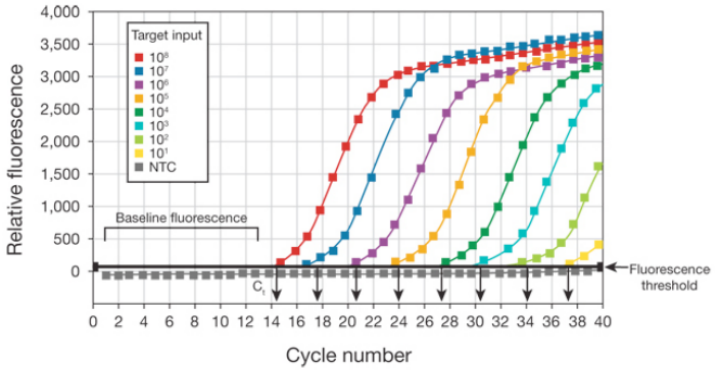


FIGURE 1.6: Quantitative PCR for multiple samples with different amount of starting material. Curves show measured fluorescence signal at discrete time points. Each curve shows a baseline (no detectable signal), an exponential phase (fluorescence doubles after every cycle) and a plateau phase (no increase of fluorescence after a cycle). During the exponential phase at cycle C_t the signal intensity ratio between two samples $\frac{S_i[C_t]}{S_j[C_t]}$ is defined as the relative difference of input material between samples S_i and S_j. Image from Thermo Fischer Scientific [98].

Performance of qPCR has been addressed in several studies on DNA and RNA [90, 6]. A low cost per sample, high degree of automation and simultaneous testing for several markers, called multiplexing is available. Further, technical replication has been established, which leads to sensitive results with a low error-rate. It is suitable for quantification of low quantity of sequences of DNA to a high degree of confidence, which makes it suitable for detection of free circulating tumour cells in blood [29]. Absolute quantification of the initial template is also possible by a sequential number of qPCR reactions on the same sample [112].

1.2.3 Microarray

Whilst testing multiple genetic markers with qPCR is limited by the number of parallel amplification reactions that can be performed, microarrays overcome this limitation and screen thousands of markers at once. Over the past decades a wide range of different microarray technologies have been developed for the analysis of DNA or RNA samples. Microarrays are based on hybridisation to complementary oligo probes anchored to a solid surface. As sketched in Figure 1.7, DNA or RNA fragments are labelled with fluorophores that emit a light signal. After hybridisation to probes and washing off unbound material, light emitting of fluorophores is excited by a laser and subsequently detected. The more of a fragment is present in a sample, the larger and brighter a cluster. Adding multiple fluorophores for labelling different samples allows relative transcript abundance estimation between samples [106, 124]. Another popular application for microarrays is somatic genotyping, where oligo probes complementary to target sequences carrying a SNV or InDel are designed. After shearing DNA into fragments, only complementary fragments hybridise to probes. There are other microarray technologies available as well, e.g. for detecting CNAs in conjunction with a suitable control [51].

The advantage of microarrays is the possibility to test for many genetic biomarkers simultaneously [42]. Commercial applications offer a wide range of pre-designed probe collections for many different applications.

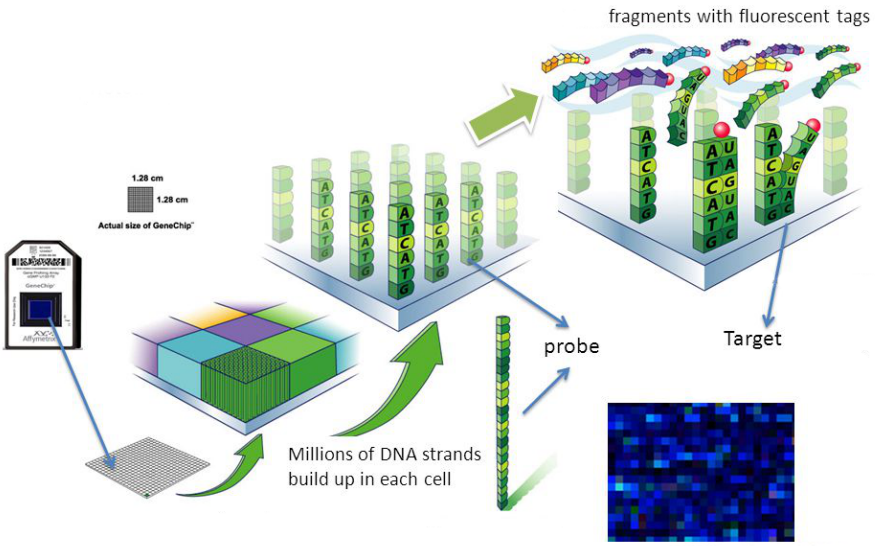


FIGURE 1.7: General principle of microarrays. Each chip carries up to 500,000 clusters with millions of DNA probes each. Fragments are prepared by adding a fluorescent dye to every fragment. Matching fragments hybridise to probes. After washing off unbound fragments, fluorescence intensities are measured for every cluster. Image adapted from Affymetrix [36].

Microarrays are still widely used, not only because of relatively low cost per sample, but also because their limitations are well known and common technical biases are understood and can be accounted for by statistical error correction methods and replication. They are a valuable tool for screening many genetic markers in cancer diagnostics today [82, 16, 28, 101, 37].

1.2.4 Sequencing

Reading out DNA or RNA molecules, known as sequencing, became very popular in the mid-seventies, when sequencing was pioneered by Frederick Sanger [105]. Since then, Sanger sequencing was improved constantly in terms of quality, throughput and length of fragment sequenced [52, 26], providing a read length of up to 1,400 bp in a few hours and at a high quality per base, i.e. less than 1 error in 10,000 bases sequenced [76]. High consistency between technical replicates and a low turnaround time, make it suitable for testing of genetic markers in a limited number of genetic regions. A very common example is routine mutation hot-spot KRAS testing of codons 12 and 13 of exon 2 and codon 61 of exon 3. Sanger sequencing in precision medicine provides a reliable mutation detection rate to a limit of 5% – 10% *variant allele frequency* (VAF) present in a sample [78, 121]. As outline in Figure 1.8, a detector captures a fluorescence signal for base-calling. In cases of low allelic frequencies of less than 10%, the signal can be covered by a much stronger signal.

Sanger sequencing was the first popular method for sequencing of DNA or, less common, RNA fragments. Although still popular, high complexity of genomics and transcriptomics make large-scale screenings very cost-intensive and time-consuming. Over the years, a number of different strategies have evolved to overcome throughput and sensitivity limitations, of which two are further explained. The first method,

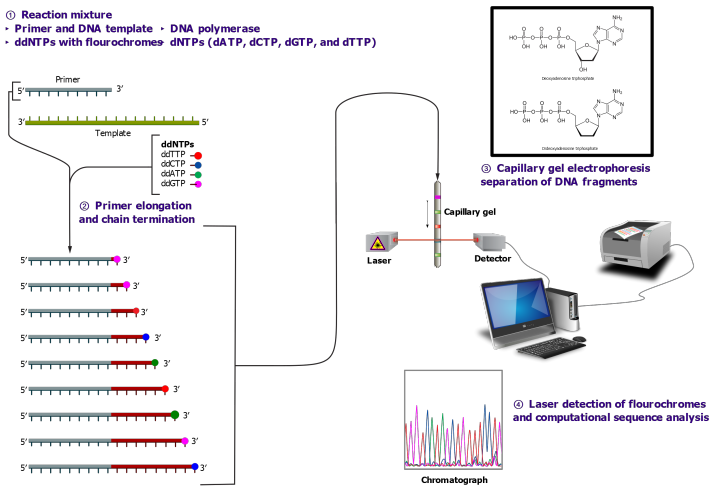
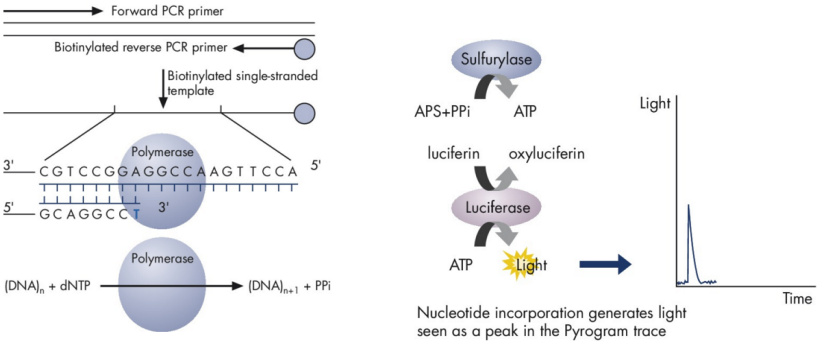


FIGURE 1.8: Principles of capilliary sequencing. DNA fragments (templates) are duplicated by adding a complementary primer. Primer elongation is terminated by random insertion of a labeled *dideoxynucleotide triphosphates* (ddNTPs). Products are seperated by size through gel. Fluorescence signals are automatically detected and translated into one of the for bases respectively. Two overlapping signals can indicate a SNP or mutation in the final Sanger trace. Image from Estevezj [30].

named pyrosequencing, was firstly described in 1996 and is based on the detection of pyrophosphates, which are released when a nucleotide is incorporated by a polymerase, as outlined in Figure 1.9 [103]. Hence it is called sequencing by synthesis. Pyrosequencing can be utilised as a benchtop sequencer, producing single sequencing reads of up to a few hundred bases in length or as a high-throughput method for targeted, exome or whole-genome sequencing producing up to a million reads per run. Studies have shown that pyrosequencing has an increased sensitivity over Sanger sequencing [121, 92], between 2% and 10%.



(A) After amplification and biotinylation of a strand, sequencing primers are hybridised. A pyrophosphate (PPi) is released when a nucleotide is incorporated by a polymerase during amplification.

(B) Sulfurylase converts pyrophosphate and adenosine 5' phosphosulfate (APS) to ATP, which luciferase needs to convert luciferin to oxyluciferin causing light emission. Detected signal results in an intensity peak over time in a pyrogram.

FIGURE 1.9: Sequencing by synthesis pyrosequencing approach: Nucleotide incorporation emits light, which can be detected. Intensity is then translated back into the base that was incorporated. Images from Qiagen [95].

Benchtop versions are still widely used for testing a limited number of genetic markers in clinical samples due to its robustness and low-cost per sample. Minimum input requirement for pyrosequencing is $10ng$ of genomic DNA according to manufacturers' protocol [94]. Pyrosequencing, however, can struggle in repetitive regions and double mutations in neighbouring loci.

Another sequencing by synthesis technology was commercially introduced by Solexa and later by Illumina®. The principle is based on

Instrument	**MiSeq**	**NextSeq 500**	**HiSeq 2500**
Output (Gb)	15	120	1,500
Sequencing reads (millions)	25	400	5,000
Max. read length	2×300bp	2×150bp	2×150bp
Run time	< 3 days	< 2 days	< 6 days
Key applications	Small genomes, exomes, transcriptomes, gene panels	Genomes, exomes, transcriptomes, gene panels	Large-scale genomes, exomes, transcriptomes, targeted gene panels

TABLE 1.2: Benchmark information of a selection Illumina® instruments. All specifications are declared by Illumina® [48].

reversible dye-terminator nucleotide incorporation at one base per cycle [12]. In a first step, fixed adapter sequences are ligated to 5' and 3' ends of DNA fragments, called *library preparation*. Complementary adapter sequences bound to a glass slide or *flow-cell* allow DNA fragments to partially hybridise. Molecules fixated to the flow-cell by hybridisation are amplified to clusters and fluorescence signals are detected base after base for each cluster, as outlined in Figure 1.10 - Figure 1.12. The first instruments were able to sequence up to one billion prepared fragments simultaneously of 20-40 bases in size [32].

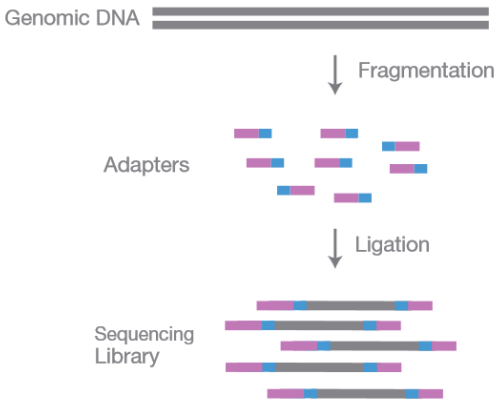


FIGURE 1.10: DNA is sheared or digested into smaller fragments (grey) and forward and reverse adapters (purple, pink) are ligated to both ends. Image from Illumina® [2].

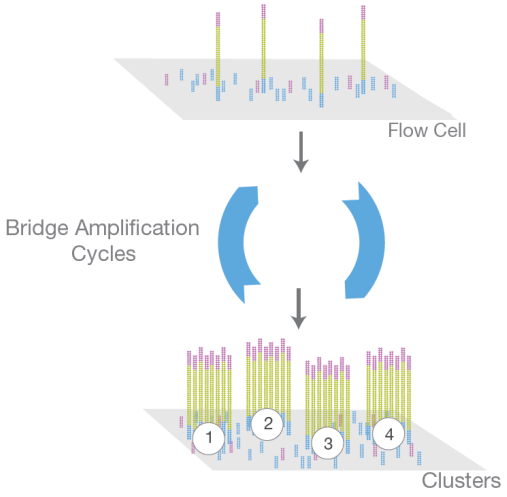


FIGURE 1.11: Mix of fragments with ligated adapters are loaded onto a flow-cell. It carries a surplus of complementary forward and reverse adapter sequences. Fragments are bridge-amplified to clusters. Image from Illumina® [2].

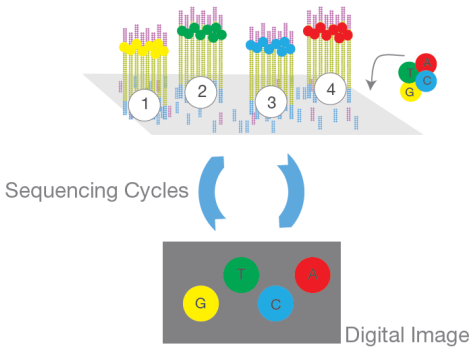


FIGURE 1.12: Labelled reversible terminator stopping strand extension, primers and DNA polymerase are added. Each of the four nucleotide is labelled with a fluorophore emitting a different colour. Nucleotides are incorporated at each cluster by the enzyme. Fluorescence is triggered by a laser and flow cell is imaged. Subsequently the terminator is cleaved off and a new base can be incorporated. Image from Illumina® [2].

Massively parallel sequencing technologies had to make sample preparation quicker and easier to handle for a high number of samples, due to the drastically increased throughput. Over the past years technology was further improved by increasing the number length of sequencing reads. Also a range of benchtop instruments were introduced decreasing sequencing time and read length by sacrificing yield. A summary of three commonly used instruments are given in Table 1.2, comparing benchtop versions of different sizes with instruments of medium or high throughput. Basecalling from Illumina® sequencing data is on average less confident than from Sanger traces. Studies report the error rate to lie between 0.1% and 1%, i.e. 1 to 10 in 1,000 can be expected to

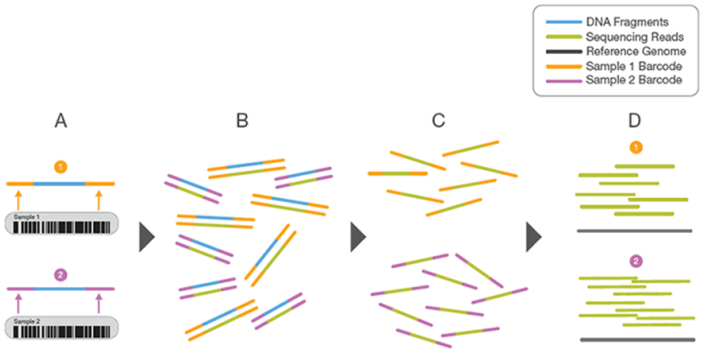


FIGURE 1.13: Concept of barcoding or *multiplexing* of two samples (orange, purple). A) Two specific barcode sequences are added to separate collections of fragments (libraries). B) Libraries are pooled and sequenced (fragments and barcodes). C) Samples are separated according to their identified barcodes (*de-multiplexing*). D) Each set of reads is processed differently. Image from Illumina® [2].

be miscalled, ten times higher on average than current Sanger sequencing [96]. Typically confidence drops towards the end of reads beyond around 100bp. Due to the very low cost per base, compared to other sequencers, every genetic region is covered multiple times by different reads increasing confidence. Necessary coverage or confidence depends on the application. Although bases are sequenced multiple times, the available output per sequencing run is often still higher than needed for a single sample. Illumina®, therefore, has introduced a method of sharing a run by attaching sample-specific barcodes to every fragment and pooling those, as outlined in Figure 1.13. Due to a unique barcode attached to each fragment, its origin can be determined afterwards [113].

Illumina® sequencing platforms are an efficient method to sequence arbitrary sets of DNA or RNA, hence the technology is used more and more in molecular biology and for clinical applications. It is important to say, however, that due to the high amount of sequencing data generated from every run, design, processing, and interpretation of the results is more challenging. Especially in cases where entire exons or genes are sequenced, the testing becomes more and more hypothesis-free in the sense that irrespectively of the known clinical impact of the reportes results are [75].

Technologies, such as Oxford Nanopore™ or Pacific Biosciences® improve constantly their throughput for sequencing single molecules in the near future. Currently error rates, throughput and cost per single base do not meet all criteria necessary for diagnostic applications. These technologies are, however, very promising and may replace current sequencing technologies eventually. As sketched in Figure 1.14, a broad range of sequencing technologies have risen over the past decades, each with their own strengths and weaknesses. Current sequencers, such as the Illumina® instruments show a great decrease in the cost per base, by sacrificing the length of each fragment sequenced, while other instruments, based on SMRT sequencing provide an increased read length targeting single molecules [99].

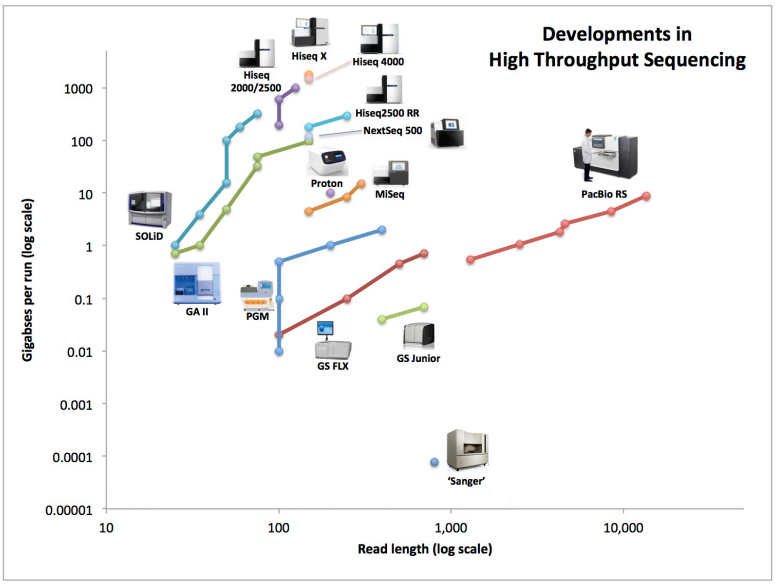


FIGURE 1.14: Developments of sequencing in terms of read length and throughput. Image from Nederbragt [88].

1.3 Sample Preservation Methods

Most collected clinical samples are immediately preseverd and stored for a later use. Quite a few different methods to preserve tissue exist, such as storing them in stabilisation reagents or embedding in *Optimal Cutting Temperature Compound* (OCT) [13]. The most two common protocols are *formalin-fixation and paraffin embedding* (FFPE) or deep freezing of freshly collected tissue. Both methods have strengths and weaknesses which are further discussed, as their impact on DNA quality can be high. This section gives a brief overview of freezing and paraffin embedding and their influence on cell morphology and genetic material.

1.3.1 Measuring DNA Integrity

Applications for genetic marker testing require a minimum concentration of extracted DNA of a sufficient quality to produce good results that satisfy requirements for a diagnostic interpretation. Typically, DNA extracted from FFPE tissue shows a broad range of degrees of fragmentation and concentration. A drop in quality or quantity may have an effect, especially on sequencing [14]. The concentration of extracted DNA is routinely quantified by measuring light or fluorescence absorption [4]. Instruments such as Qubit (fluorometer) or NanoDrop (spectrophotometer) are commercially available to quantify DNA within seconds at relatively low cost. Measuring the degree of fragmentation, called

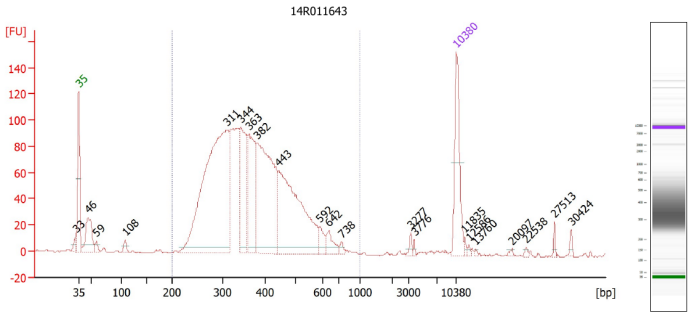


FIGURE 1.15: An Agilent Bioanalyzer trace and gel-like image of an Illumina® TruSeq® library. It was prepared by adding 63bp to the 5' (57bp adapter and 6bp barcode) and 57bp-65bp to the 3' end (57bp adapter), resulting in 120bp that needs substracting to receive the initial fragment size distribution.

integrity, in a sample is somewhat more complex, as loading enough genetic material onto a gel should be avoided to save material of precious samples, while high sensitivity chips use up less genetic material, but can measure fragment sizes only up to a few thousand bases in size, often insufficient to see level of degradation, as shown in Figure 1.15. Another problem is that measured concentrations vary depending on the instrument used. For example, NanoDrop tends to overestimate concentrations from DNA extracted from FFPE tissue, while Qubit is sensitive to low salt concentrations [85].

Another possible method is to measure grade of DNA integrity with help of an Agilent TapeStation 2200 instrument, which generates a DNA integrity number (DIN), based on concentration and fragmentation of extracted DNA [54]. The system is able to analyse and quantify

genomic DNA fragments from 200bp to over 60kbp. The software generates an electropherogram and a gel-like image, as shown in Figure 1.17 and Figure 1.16. An improved shearing protocol, based on DNA integrity of each sample, reduces negative effects on sequencing outcome by improving target enrichment and library preparation steps. A DIN number is a measure of degradation and concentration of genomic DNA prior to shearing [50]. DIN scores are useful to obtain a reasonable estimate of DNA integrity without running an argarose gel, for quality assessment prior library preparation for sequencing. About 100ng of raw extracted DNA as input is recommended for a run, but a reliable estimation of the integrity can be achieved from 10ng. Agilent recommends a minimum DIN score of 3 for achieving consistent results of library preparation sufficient for sequencing and subsequent data analysis, although this number may differ for different methods [35]. Results tend to be very noisy or inconsistent if minimum criteria are not met, i.e. DIN score falls below a certain value or the amount of available genetic material is low. Further the exact method of how the DNA integrity score is determined is not publicly available making it difficult to understand relationship between concentration, integration and fragment size of DNA.

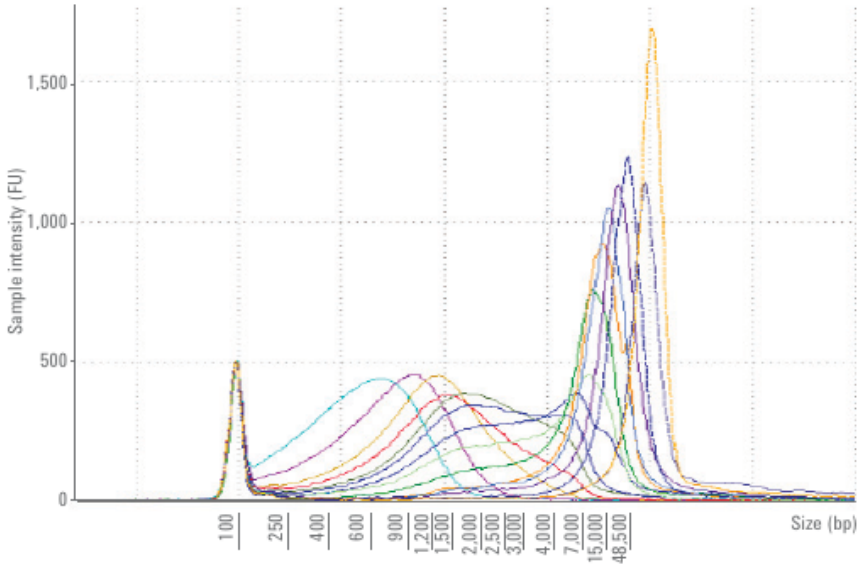


FIGURE 1.16: Electropherogram from Agilent TapeStation 2200. Peaks at the higher end indicate concentrated and intact DNA. Broad and low peaks at the lower end indicate degenerated or fragmented DNA.

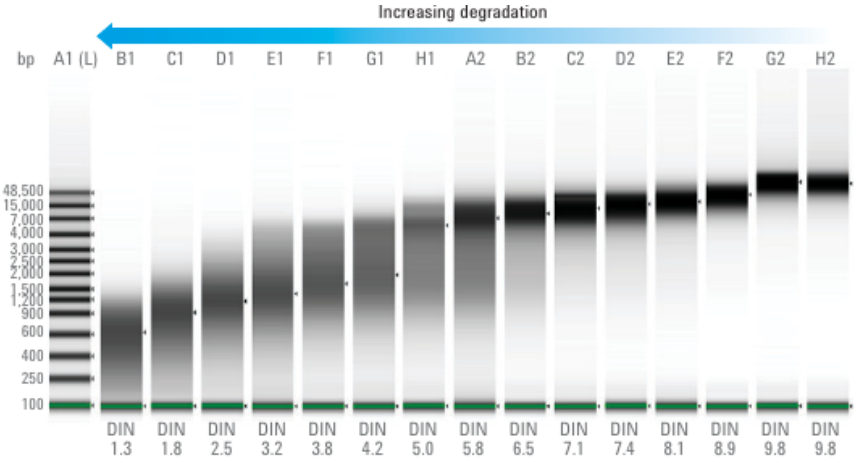


FIGURE 1.17: Gel-like image from Tape Station 2200. Below each lane the computed DIN. Bright smears cause lower DINs, while dark and sharp bands towards larger sizes result in higher DINs.

1.3.2 Fresh-frozen Tissue

Sections that are snap-frozen immediately after surgery are considered the gold standard for many molecular assays. The protocol involves embedding sections in a preservation media, such as OCT and then storing below −70 °C. For sample use such as staining for IHC or extraction of genetic material, tissue is fixated with acetone, ethanol or formalin [81]. The protocol is simple, but requires long-term storage in a dedicated freezer space. Shipping frozen samples carries a risk of being thawed, meaning samples could be irretrievably damaged or even lost. Snap-freezing can keep morphology intact in most cases, but some proteins may be damaged causing freezing artefacts [110]. For some tissue sections keeping cell morphology intact is difficult, especially if the samples are (partially) thawed or were previously not embedded properly. Genetic content such as DNA and RNA is kept relatively intact as freezing prevents enzymes, such as DNAses or RNAses degradation [41]. Hence, freezing keeps DNA integrity generally at a high level. Fresh tissue freezing remains the method of choice for many applications due to its fast protocol and great cell preservation. Snap-freezing is, however, prone to artefacts due to improper freezing and can lead a loss of cell morphology information.

1.3.3 Formalin-fixing and Paraffin Embedding

Tissue preservation by formalin-fixing and paraffin embedding (FFPE) is the most common method used by pathologists. A tissue slice of about 3-4mm in thickness is harvested, put into a cassette and immediately fixated in 10% formalin by incubating tissue for up to 48 hours at room temperature. Different protocols decrease fixation time to reduce the negative effects of formalin on genetic material [123]. The sample tissue is then rinsed and kept in *phosphate-buffered saline* (PBS) or 70% ethanol until it is incubated in 70%, 95% and 100% ethanol for dehydration. Finally, two xylene washes are performed before the sample is embedded into paraffin. Dehydration is necessary to allow the embedding of fixated samples safely into wax. After the block has cooled and hardened it can be stored at room temperature. The block can later be sectioned and stained, such as shown in Figure 1.18 or de-paraffinised and processed for DNA extraction. The described protocol is based on the work of Canene-Adams [13]. There is, however no standardised FFPE protocol that is used by every lab and it is unlikely that protocols will be fully standardised in the future.

Formaldehyde stops most enzymatic activity, which inhibits digestion of the cells and degradation of a tissue sample. Cross-linking between proteins in tissue preserves cell morphology. There are many factors throughout the protocol that can have an effect on the quality of the

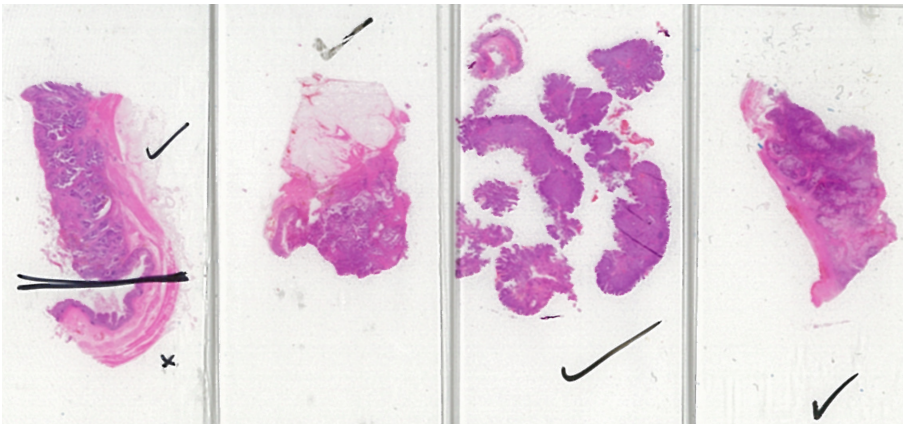


FIGURE 1.18: Sectioned and hematoxylin and eosin (H&E) stained samples that were fixed and embedded. Hematoxylin stains the genetic material, such as DNA and RNA, blue. Eosin stains eosinophilic structures, such as proteins, in a bright red or pink. A checkmark indicates tumour burden of at least 70%, a cross means less than 70% and the black line indicates where to cut to reach 70%. Determined by a pathologist.

genetic material and morphology, such as tissue type, fixation time, formalin concentration, processing temperature, pH or storage conditions [11, 86]. Whilst cell morphology is well preserved of formalin-fixed and paraffin embedded tissue, integrity of extracted DNA shows a broad range of quality, from relatively intact to complete degradation.

High fragmentation and chemical modifications, e.g. cytosine deamination can introduce artefacts when DNA from formalin treated tissue is used for genetic marker detection, such as qPCR, microarray or sequencing [130]. Protocols for genetic marker detection may need to be

adjusted, to achieve good results. These constraints are sometimes difficult to meet, e.g. only a very limited amount of poorly preserved tissue is available. Recommendations exist to overcome some of these obstacles. One option is an increase of number of PCR cycles or using special reagents or newly engineered enzymes to improve the DNA yield and quality [128, 56, 108, 115]. Integrity and quantity still can vary from sample to sample, depending on many factors, some of which were mentioned above. Although some studies exist, no universal guidelines exist to preserve genetic material using FFPE protocols best [122, 85]. FFPE tissue should be treated cautiously to maintain DNA integrity and reduce loss of material from poor extraction methods. Low frequency mutation detection on DNA extracted from FFPE tissue may not be possible below a certain threshold in some cases, due to modification of genetic material from formalin treatment.

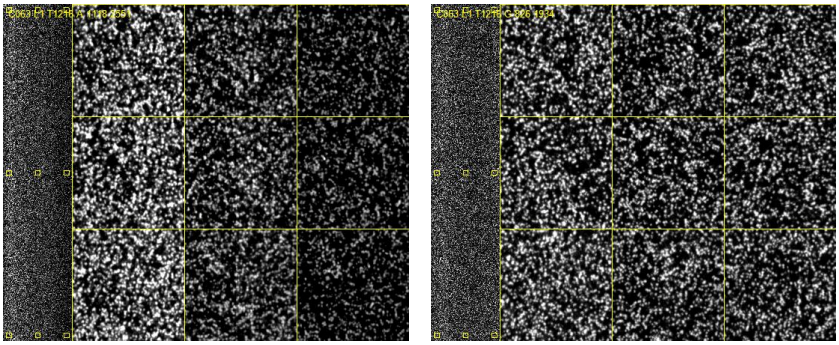
1.4 Low-frequency Mutation Detection in Humans

Massively parallel sequencing instruments produce a vast amount of data within a few days. Analysis of sequencing data, therefore, needs to be automated. Bioinformatics, as an interdisciplinary field between mathematics, computer science, biology and statistics provides solutions for the analysis of bulk data generated by massively parallel sequencing. Images taken from clusters are translated into bases in a step named *base-calling* that are subsequently formatted into sequencing reads containing strings of called bases coupled with a confidence estimation. The read sets are cleaned from contamination and poor quality bases, followed by alignment or mapping against a reference based on similarity. Resulting alignment files are used to find differences in the read data from that reference, named *variant calling*. After filtering for likely artefacts, remaining variants are put into context by comparing them to a public databases, such as dbSNP [109], COSMIC [34] or ClinVar [60] and by predicting the potential effects on transcribed proteins. In this section the fundamental principles of the analysis of sequencing data for mutation detection will be explained and an introduction of the standardised formats that are commonly used will be given.

1.4.1 Base-calling and Conversion into FASTQ Format

Base-calling on Sanger sequencing reads is performed by translating a fluorescence signal to one of the four nucleotides adenine (A), cytosine (C), guanine (G), thymine (T) or the unknown base (N). Illumina® sequencers generate high-resolution image files from discrete positions on the flow-cell and all detectable clusters per sequencing cycle are captured, see Figure 1.19. Storing entire runs with billions of image files would allocate many terabytes of disk space from one sequencing run. As this is impractical, images are processed to identify locations of clusters and directly translated into bases. Only a small subset of images from fixed coordinates on the flow-cell is stored to check for cluster density by-eye. The clusters are identified from the taken images along with meta-information, such as instrument, unique run id, coordinates and intensities. The raw image file is discarded afterwards. There are several algorithms for base-calling published, each trying to provide reliable and robust interpretation of the image data [80].

Basecalling files (BCL files) are compressed binary files, which are usually not used as input for subsequent sequencing analysis directly. Instead, they are converted into a human-readable file format, named FASTQ. Introduced by the Wellcome Trust Sanger Institute in FASTQ file nucleotide sequences are bundled together with estimated quality



(A) Fluorescent channel for base A (B) Fluorescent channel for base G

FIGURE 1.19: Two images of identified clusters on an Illumina's HiSeq 2000 flow-cell. Both images show a tile, a section of a sequencing lane (1216), at cycle 63.

scores. A quality score $\mathcal{Q}$ is defined as [19]:

$$\mathcal{Q} = -10 \log_{10} p \tag{1.1}$$

where p is the probability that a corresponding base-call is incorrect [31]. It is also known as Phred-Score. Since Illumina® pipeline version 1.8 the $\mathcal{Q}$-score is encoded as ASCII 33 to 126 characters (ASCII+33), although Illumina® describes symbols only up to ASCII 73 [131]. Sequence bases and qualities derived from a cluster are bundled together with collected metainformation. A single read in FASTQ format from a NextSeq 500 looks like this:

```
@NS500781:4:HCKH5BGXX:1:11101:1162:1050 1:N:0:13 Comment
CTGAGNAGCTGGGCTCCCGCTCTGGTGGGACACGCTGCCATCATTACTTTGATTAC
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEE
```

The header line starts always with an '@' and contains the metainformation, further described in Table 1.3. The second line is the actual read sequence. The third line starts with a +, but is not used in most cases, hence the line is usually empty besides the plus sign. The fourth line contains the determined Q–scores in ASCII+33 encoding.

The Illumina® demultiplexer and converter *bcl2fastq* generates one (for single-end runs) or two (for paired-end runs) files per sample [131]. The software is able to separate the samples according to their barcode that was attached, from a provided sample sheet. As shown

Field	Description
NS500781	sequencer ID
4	run id
HCKH5BGXX	flow-cell id
1	lane id
11101	tile within lane
1162	'x'-coordinate of cluster on tile
1050	'y'-coordinate of cluster on tile
1	member in a pair (forward - 1, reverse - 2)
N	read filter (filtered - Y, not filtered - N)
0	control number (no control bits - 0, otherwise - even number
13	index or index id

Table 1.3: FASTQ read header description for a read sequenced with a NextSeq 500 instrument.

in Figure 1.20, a sample sheet is a *comma-separated text* file (CSV) that maps the index or barcode sequences uniquely to samples and provides further information about the experiment. Created FASTQ files are sorted according to their flow-cell coordinates. For paired-end runs it is ensured that both files have the same number of reads and the same order. If the other member in a pair, called *mate*, cannot be identified, a dummy read would be created, which is a read consisting only of Ns and Q–scores of 0. Most downstream applications, such as sequencing aligners expect the same order of read pairs and use the pairing information. Unlike the BCL format, the FASTQ file format is standardised and accepted by virtually all bioinformatics software as input files, hence it will be called *raw data* from now on.

1.4.2 Data Trimming and Read Alignment

Illumina® sequencing is prone to a number of different errors that were characterised so far [83]. Reads can be charged with base-calling uncertainties or errors for various reasons, such as insufficient fluorophore cleavage or weak intensity signals. As shown in Figure 1.21, there is a typical quality drop towards the 3'-end of each read, due to deterioration of involved sequencing reagents over time, i.e. change of pH. Another problem is caused by read extension by the polymerase into the flow-cell adapter sequence for short fragments, which does not contain any genetic information, as shown in Figure 1.22.

[Header]							
IEMFileVersion	4						
Investigator Name	LH						
Experiment Name	VAL269_Run2						
Date	10/03/2016						
Workflow	GenerateFASTQ						
Application	FASTQ Only						
Assay	TruSeq LT						
Description	VAL269_Run1						
Chemistry	Default						
[Reads]							
151							
151							
[Settings]							
ReverseComplement	0						
[Data]							
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	Sample_Project	Description
15R8492	15R8492				GCCAAGAC	VAL269_Run2	VAL269_Run2
15R8494	15R8494				CGAACTTA	VAL269_Run2	VAL269_Run2
15R8495	15R8495				ACCTCCAA	VAL269_Run2	VAL269_Run2
15R8445	15R8445				CTGTAGCC	VAL269_Run2	VAL269_Run2
15R8455	15R8455				CGCTGATC	VAL269_Run2	VAL269_Run2
15R8472	15R8472				ATTGAGGA	VAL269_Run2	VAL269_Run2

FIGURE 1.20: Example of a sample sheet [47]. A sample sheet consists of four parts: [header], [reads], [settings] and [data]. The header contains information about the experiment, the read section describes the read length and run settings. The settings part is used to give further instructions, such as the use of custom applications. The data part describes samples, assigns a unique sample ID to an index and further information for individual samples can be given.

As every base in the genome is usually sequenced multiple times, removing low confidence base-calls or adapter bases should not produce any gaps, but increases subsequent alignment quality. Data trimmer always try to find an ideal trade-off between efficiency and accuracy. A very pedantic trimming may result in a minor to mediocre boost in quality, but the advantage is defeated if trimming takes too long to finish. Incomplete or "over-"trimming contradicts the purpose of trimming, as the trimmed data should be less biased and not more. A recently published trimming software called "Skewer" uses bit-masked

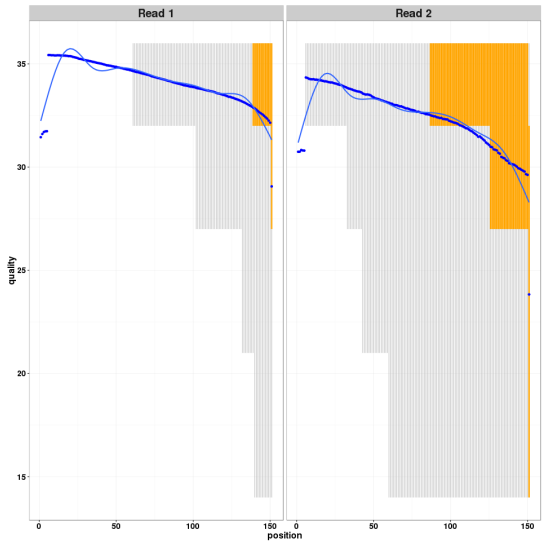


FIGURE 1.21: Quality of sequencing data containing 4.5M sequenced fragments along the read position. Grey lines indicate 10% and 90% quantiles, orange the lower and upper quartiles, blue dots show the median Q-score, light blue line a Lowess curve, from a local regression [8]. Towards the end of the reads the average quality drops. The reverse read shows a lower average quality.

k-difference matching, a dynamic programming algorithm that runs in runtime class $\mathcal{O}(kn)$, where k is the maximum number of differences between the adapter sequence and the read, while n denotes the read length [49]. The core algorithm is based on calculation of an alignment with the minimum Levenshtein, or edit distance. An alignment reports sequence similarity under a defined metric. The edit distance between two sequences $a = a_1,...a_i,...a_n$ and $b = b_1,...,b_j,...,b_m$ with lengths n and m, respectively, is defined as:

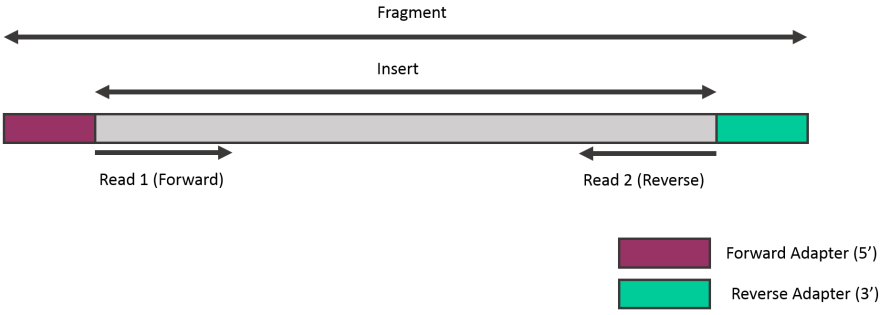


FIGURE 1.22: A schematic diagram showing origin of forward and reverse reads. If the insert size is shorter than the desired read length, a read will contain parts of the adapter sequence.

$$E_{a,b}(i,j) = \; min \begin{cases} E_{a,b}(i-1,j-1) + \delta_{a_i,b_j} \\ E_{a,b}(i-1,j) + 1 \\ E_{a,b}(i,j-1) + 1 \end{cases} \tag{1.2}$$

where $\delta_{a_i,b_j} = 0$ if $a_i = b_j$ and otherwise $\delta_{a_i,b_j} = 1$. Besides is $E_{a,b}(0,0) = 0$ by definition. The algorithm solves this mathematical problem by spanning a dynamic programming matrix and backtracks the minimum alignment distance. The software is extended to support $\mathcal{Q}$–scores and utilises paired-end read information. The asymptotic memory requirement of $\mathcal{O}(m)$, where m is the length of the adapter sequence. The algorithm has proven fast and efficient in day-to-day work.

Targeted re-sequencing for mutation detection relies on comparing sequenced reads against a reference genome. Automated comparison of

two sequences against each other is useful even before high-throughput methods were introduced. First algorithms were based on dynamic programming methods, such as Needleman-Wunsch algorithm [89] or Smith-Waterman algorithm [114], of which Skewer uses a modification. They are both very sensitive methods that deliver optimal results. Unfortunately both algorithms are very memory and time consuming, making them infeasible for comparing a large number of sequences. Later heuristics were introduced such as seeding-based algorithms used by FASTA [72] or BLAST [1], which work very well for searching a query sequence in a large database of reference sequences. Aligning data from massively parallel sequencing to a reference, however, requires finding hits of a many short query sequences in one large genome. Seeding-based algorithms are impractical for such problems, as they work better on longer input queries to deliver sufficient results. Further, time and memory requirements are lower than other dynamic programming approaches, but insufficient when aligning a large amount of short sequencing read data to a large reference genome.

One of the first short read mapper was introduced in 2008 named MAQ [69]. Similar to ELAND (2008) [5], the reference sequence was pre-processed and several indexed hash tables were generated to find accurate gap-free hits with up to two mismatches from reads of a finite length. Later the Burrows-Wheeler Transform was utilised for string matching, allowing gaps and arbitrary mismatches in the alignment in complexity class $\mathcal{O}(n)$, i.e. the computation time is based on the

length of the query sequence n, possible by previous indexing of the genome. Further improvements were made to efficiently align reads of an increased length beyond 150bp [68, 67]. The Burrows-Wheeler-Aligner (BWA) is able to trim data automatically towards the end of the reads if too many mismatches would cause alignment score dropping too much. Although this *soft-clipping* of bases can be used for an on-the-fly trimming of adapter sequences, it should be mentioned that BWA does not utilise calculated $\mathcal{Q}$–scores from the raw data. There are some cases possible, where BWA may introduce unwanted alignment artefacts from untrimmed data, e.g. from base calling errors inside the adapter.

The standard output for short read alignment is given in *sequence alignment/map* (SAM) format. It is a TAB-delimited text format consisting of a header and an alignment section [70]. Each line in a SAM header starts with an '@' character, followed by a two-letter record explaining the field and a `TAG:value` tuple. The purpose of a SAM header is to assign meta-information for aligned data, such as the sorting order of the alignment, the reference used or sample origin. A typical header looks like this:

```
@HD     VN:1.5  SO:coordinate
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
@SQ     SN:chr4 LN:191154276
```

```
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@RG     ID:14R010275    LB:Source_TSB_v2        PL:INextSeq\
        SM:14R010275    PU:-    CN:SBS
@PG     ID:bwa  VN:0.7.12-r1039 CL:bwa mem -M /hg19/genome.fa\
        14R010275-R1.fq.gz 14R010275-R2.fq.gz PN:bwa
```

Lines starting with HD contain format version and sorting order of the alignment, SQ tags indicate information about used genome reference and chromosome length. RG stores a unique sample ID, kit and instrument used. It can be further customised according to local conventions. PG lines track programs used for file modifications with all set parameters acting as a history to recreate the file from raw data if necessary. More fields can be added if desired; the full standard is documented by the SAM file specifications [66]. A header protects an alignment file against unwanted corruptions, such as a file name change. The alignment section of a SAM file has at least 11 tab-separated columns and arbitrary number of rows. Every row stores information about one single read alignment. The individual fields are explained in Table 1.4. The alignment sections of a SAM file can be extended by adding optional columns after the 11 mandatory fields. These must follow a TAG:TYPE:VALUE format and should follow predefined specifications

to avoid misinterpretation by other users or software tools. An example alignment row with additional optional fields looks like this (\ continues line):

```
M01706:46:000000000-ALDMT:1:1109:13015:25512   147    chr6   54635473\
60      67M84S  =       54635369        -171 \
ATATGATCCAACAATAGAGAATTCCTACAGGAAGCAAGTAGTAATTTATAGAGAACCGTGTCTCTTGAA\
AATTCCTGAGAGAACAGGTCTAAAATAGTGCAACGGAATAAAAGCAAGTACATGAGGACTGGCGGGGAG
AGCTATCCTTGTG  EFFGGBFCAEFF2G2DDF1FFFFGGBFGD1DBFHHFGBG@1@22DB2@222B/B\
B//00F0B1B22D221@0B0211D2 B21HGG1HFB2A2D11111//1B0222121DB11122D221D1\
1DB1B0000111GGGF3FB11>>1A>>11        NM:i:4  MD:Z:46G2G5A1C9 AS:i:47\
XS:i:27 BC:Z:CACTGTCCGCATACA
```

It contains more than the 11 mandatory fields, which is permitted as long as the formatting is correct. In this case the BWA aligner added further information about the alignment. `NM:i:4` indicates that the Levenshtein distance to the reference is 4. The tag `MD:Z:46G2G5A1C9` explains the sequential matching of the positions in read coordinates, from left to right (46 matches, then a G mismatch, then 2 matches and so on), similar to the *CIGAR* string, the collapsed per-base alignment operations. `AS:i:47` reflects the best alignment score, while `XS:i:27` is the score of the second best alignment. `BC:Z:CACTGTCCGCATACA` indicates a barcode for that read. The SAM format has proven to be very flexible and efficient. There is a binary option available, called *binary SAM format* (BAM), which uses file compression based on BGZF, an extension of the Gzip file standard [23].

SAM file manipulation and conversion/compression can be performed with Samtools, a suite of software tools written in C [70]. Sequencing alignments can be visualised with a genome browser, such as the *Integrative Genomics Viewer* (IGV) [100]. It requires an additional SAM index file, for efficient data access, which can be created with Samtools. As shown in Figure 1.23, individual read alignments can be inspected across the genome. The options of IGV are very versatile and provide a good method of visualisation and comparison of aligned sequencing data.

Column	Description
1	Read header from FASTQ file
2	Bitwise flag encoding read alignment properties
3	Reference sequence name
4	Leftmost mapping position of read (1-based)
5	Mapping quality (Phred scale)
6	CIGAR string
7	Read header of mate
8	Mapping position of mate
9	Length of alignment
10	Sequence that has been aligned
11	ASCII+33 scores of sequence

Table 1.4: All mandatory fields in a valid SAM alignment file [66].

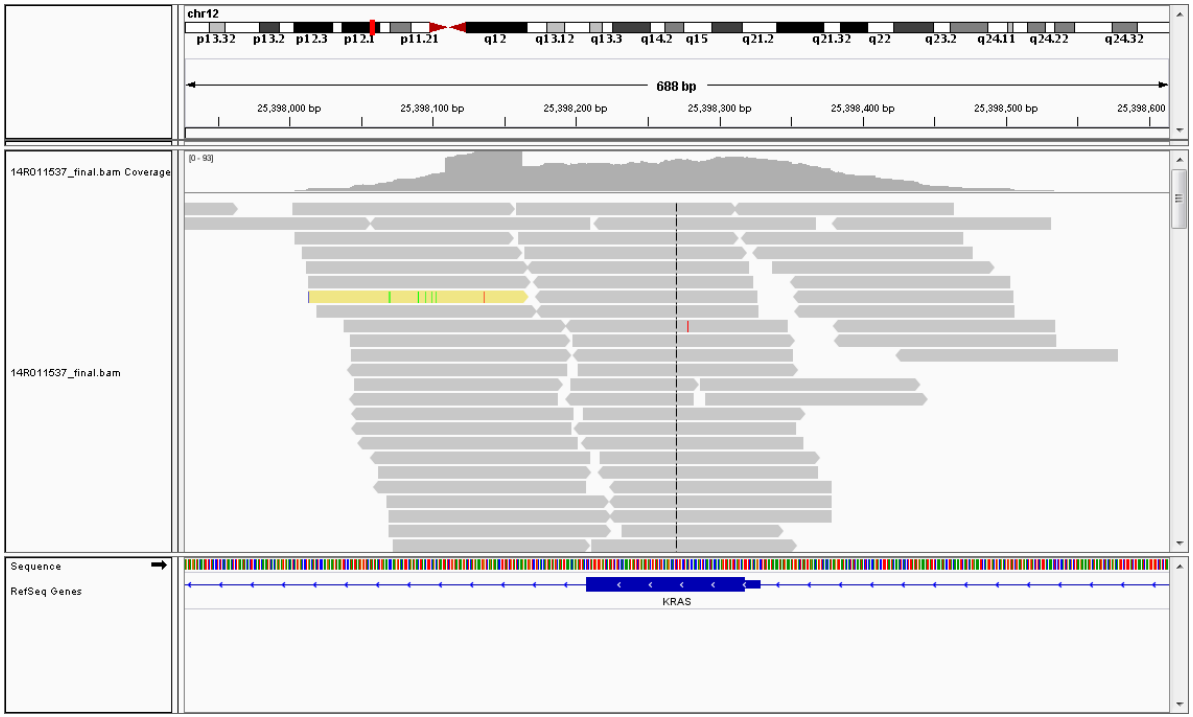


FIGURE 1.23: Alignment visualisation with Integrative Genomics Viewer of reads mapping to KRAS exon 2. Reads are shown as grey arrows, reflecting read orientation. Colours indicate different base or read properties depending on user defined settings. At the top a per-base coverage is plotted as a histogram along the genome. The bottom shows the base sequence and annotation of genomic features. Additional tracks can be loaded for extended analysis.

1.4.3 Alignment Quality Improvement Strategies

Alignments can be very noisy, as every read is exposed to a relatively high error rate. A high coverage per reference base allows many artefacts to be identified and subsequently corrected. Common sources of error identified are duplicated reads, sequencing errors and misalignments. Errors from Illumina® sequencing consist mainly of miscalled bases of poor quality [104], hence errors can be corrected by sequencing each base multiple times and ignoring bases called below a certain quality threshold. It becomes difficult if artefacts are amplified by PCR, for example single base substitutions during library amplification. A read duplicate is defined as a read thwat was derived more than once from the same initial DNA fragment. Thus, duplicates do not add any more valuable information, which is why they count as noise[25]. There are two possible classes of duplicates described here: PCR duplicates and optical duplicates. A PCR duplicate is generated when the sequencing library is amplified by PCR and the same fragments are then sequenced multiple times. The amount of PCR duplicates is usually around 10%, but in cases of low complexity libraries, PCR duplicates can make up over 90% of the data. An optical duplicate is generated if a cluster on the flow-cell is read out twice during the base-calling. This happens especially if a cluster is very big, because the fragment was over-amplified during bridge amplification or if too many fragments were loaded onto the flow-cell. Depending on the library preparation

protocol used, there are different ways of excluding duplicates from a dataset. In cases of randomly sheared DNA, a read is flagged as a PCR duplicate if start position and CIGAR string are identical, i.e. the read information is completely identical. Optical duplicates are determined by investigating the flow-cell coordinates. If two reads are very close to each other, they are treated as they were derived from the same cluster. To flag a read as duplicate the bitwise flag in the SAM file is set to 1024 [7]. Other tools will ignore this read or read pair automatically.

Sequencing alignment on massively parallel sequencing data can be only performed in a reasonable time by applying heuristic methods receiving good results in most cases. Sometimes, however, may not find the true origin of a sequenced read pair or maybe the true position is returned, but not with the best alignment, as artificial mismatches were introduced. A very common problem is the alignment quality around homopolymers, repetitive regions or insertions/deletions of bases [22]. Established strategies are based on identification and relocation of misaligned reads either by realignment with a more sensitive algorithm directly, or by local or global reassembly of the reads into a consensus sequence first, followed by realignment [65, 44]. For longer InDels a local re-assembly often leads to better results [120]. Sequencing data from FFPE samples shows an increase in the amount of mismatches, which many assembly algorithms struggle with. The ABRA realigner has an node pruning option since version v0.95 that improves efficiency in noisy data [84]. ABRA spans a De-Bruijn graph of k–mers for each

region of interest from the aligned reads, traverses the graph to build de-novo *contigs*, a set of potentially overlapping DNA fragments that build a consensus. These contigs are then aligned back to the reference. Due to increased length of the contigs over the short reads, insertions and deletions are less punished by BWA. The result is a locally alternated reference, which is then used to align reads against. If reads align better to the alternative than the initial reference their alignment entry is updated in the SAM file. It is crucial to only work with data of very high quality, as keeping poorly aligned reads in the data set causes error prone results.

1.4.4 Variant Calling and Filtering

Calling variants from massively parallel sequencing data means comparing base distribution of sequenced bases to a reference base at every position of interest in the genome. Although a SAM file contains all information needed, the format is not ideal to efficiently look at called bases, as the data is stored read-wise. Hence, data is formatted into a pileup format prior analysis. It streamlines the aligned bases along the reference , by channelling the reads at every position. It results in a data matrix, where each row is one genomic position, with reference base, read coverage and the base information collected from reads covering that position, bundled together with mapping quality extracted from the SAM file by using Samtools mpileup:

```
chr1 11346862  G 19 .,.-1A,-1a,-1a,-1a,,.-1A.-1A. @@GGOHH3GGG/EGHEHGG
chr7 140453136 A 25 ,,t,..TT,,,TT..tt....T... EEDC>ABHHHHCBEHHHHGHHHAHH
```

A pileup is a human readable file of six columns: sequence identifier or chromosome (1), 1-based position in sequence (2), reference base (3), depth of coverage (4), a representation of the bases at position (5) and the base quality in ASCII+33 (6). Column five is encoded with a string of characters described in Table 1.5. A graphical visualisation of such a pileup is shown in Figure 1.24. From a pileup file the per base frequency and quality distribution is extracted. Fisher's Exact Test can be used to derive the likelihood of genotypes from base distribution at each position [64]. It is important that a variant caller uses suitable assumptions to give meaningful results even at a lower coverage. For example, calling somatic variants from a heterogeneous cell population involves a different design than for germline variants in a cell line. In addition, several strategies exist for the calling itself. One popular method is matched tumour-normal pair sequencing. A pair of samples from a patient is collected, one from a tumour and a non-tumour acting as a control. Both are sequenced and aligned, subsequently the variant caller calls genotypes in each sample and distincts germline variants from the tumour-free sample and reports tumour-exclusive mutations only. This is a great method to receive the set of somatic mutations present in a tumour even at a low frequency. This approach is, however, difficult to be used for diagnostics, as most hospitals cannot provide a sample pair for various reasons. Many variant callers do not

officially support a tumour exclusive analysis, as the paired sample approach is usually desired. Other variant callers do not work very well with heterogeneous populations, as they are designed for calling variants on diploid organisms, hence they fail in calling low-frequency somatic mutations. VarScan2's genotyping module supports calling low-frequency mutations from only one sample and also has proven to deliver very consistent results for SNVs and small InDels [55, 116]. Other types of genetic alteration, such as copy number alterations or loss of heterozygosity events are difficult to be detected by this approach due to the missing normal sample, hence they are not further discussed.

Character	**Description**
.	Match to reference base on the forward strand
,	Match to reference base on the reverse strand
A,C,G,T,N	A mismatch to reference base on forward strand
a,c,g,t,n	A mismatch to reference base on reverse strand
+[0-9][ACGTNacgtn]	Insertion of one or more bases
-[0-9][ACGTNacgtn]	Deletion of one or more bases
^	Start of a read segment, followed by mapping quality
$	End of read segment

TABLE 1.5: Symbols used in a pileup file to describe the aligned bases at a given position. Read segments are defined to reconstruct read sequences from pileup.

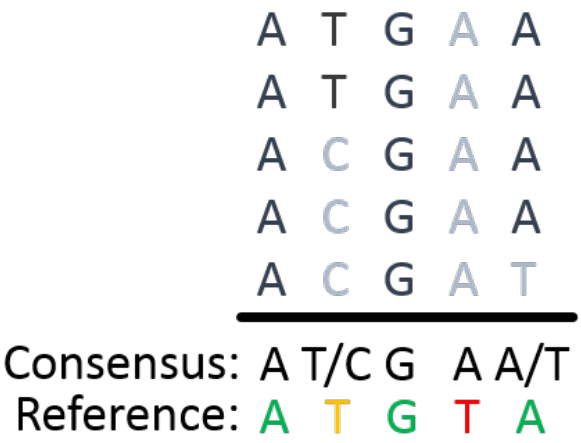


FIGURE 1.24: Pileup format represent the read alignment per base in the aligned reference. In a pileup every aligned base per position is lined up. Indicating a match (light grey) or mismatch (dark grey) to the reference base helps building a consensus of present genotypes. A variant caller uses statistical testing to classify a consensus as a variant base position (heterozygous - yellow, homozygous - red) or in concordance to a reference base (green).

	Reference allele count	**Alternative allele count**	**Coverage**
Observed	a_R	a_S	$A = a_R + a_S$
Expected	$e_R = A - e_S$	$e_S = \lfloor \frac{A}{1000} \rfloor$	$A = e_R + e_S$

TABLE 1.6: Contingency table used by VarScan2 to estimate significance of observed results.

VarScan2 core algorithm stores for every base two observations: a_R for number of reads supporting the reference allele, a_S for the number of reads supporting an alternative allele. Under the assumption of a baseline error of 0.1%, the algorithm tests with Fisher's exact test the likelihood that a_R and a_S would be observed if they were exclusively derived from the reference sequence. This is denoted as e_R for estimated counts supporting the reference allele and e_S for the alternative allele respectively. To be more specific, VarScan2 spans a 2×2 contingency table at each base position, as seen in Table 1.6. Under the null hypothesis H_0 that no mutation is present, VarScan2 computes the probability p under the hypergeometric distribution to observe a_R and a_S with respect to e_R and e_S utilising Fisher's exact test:

$$p_{e_R,e_S,a_R,a_S} = \frac{\binom{e_R+e_S}{e_R}\binom{a_R+a_S}{a_R}}{\binom{e_R+e_S+a_R+a_S}{e_R+a_R}} = \frac{(e_R+e_S)!\,(a_R+a_S)!\,(e_R+a_R)!\,(e_S+a_S)!}{e_R!e_S!a_R!a_S!(e_R+e_S+a_R+a_S)!} \tag{1.3}$$

which can be read as the probability to observe Table 1.6 under the assumption that there is no variant allele present. To obtain the p-value, the probability observing the derived contingency table or an even more extreme event under the null hypothesis, VarScan2 computes

$$\text{p-value} = \sum_{i=0}^{m} p^{i}_{\hat{e}_R, \hat{e}_S, \hat{a}_R, \hat{a}_S} \tag{1.4}$$

where $m = \min(e_S, a_R)$ and $\hat{e}_S = e_S - i$, $\hat{a}_R = a_R - i$, but $\hat{e}_R = e_R + i$, $\hat{a}_S = a_S + i$ to keep coverage A constant.

Theoretically this works for all mutation frequencies, but it is trivial to see that for lower mutation frequencies, it becomes more difficult to distinct between sequencing errors and a mutation by not adjusting coverage A. Practical studies have shown that VarScan2 works best from 500x coverage per base or higher [116]. Further, the software allows to specify a number of different filter criteria to account for artefacts in the data, such as low quality bases or poor coverage after calling variants, as Fisher's exact test can be quite sensitive to noise that is not related to sequencing errors, e.g. errors from PCR amplification or misalignment. It is, therefore, recommended to not perform a test if ratio $\frac{a_S}{A}$ is below a certain threshold that can be defined by the user beforehand. Determining further filter criteria can be a considerable challenge, as there is always the risk of over-filtering and not reporting a true variant, as there are no universal filter criteria known that are ideal in all cases. Since release 2.3.1, however, VarScan2 has an integrated false-positive filter to remove reports that were most likely called from artefacts.

VarScan2 reports variants in the variant call format 4.0 (VCF), introduced by the 1000 genomes project [20]. Similar to a SAM file, it is a human-readable text file with a header part starting with a pound sign ('#'):

```
##fileformat=VCFv4.1
##source=VarScan2
##INFO=<ID=ADP,Number=1,Type=Integer,Description="Average per-sample\
depth of bases with Phred score >= 15">
##INFO=<ID=WT,Number=1,Type=Integer,Description="Number of samples\
called reference (wild-type)">
##INFO=<ID=HET,Number=1,Type=Integer,Description="Number of samples\
called heterozygous-variant">
##INFO=<ID=HOM,Number=1,Type=Integer,Description="Number of samples\
called homozygous-variant">
##INFO=<ID=NC,Number=1,Type=Integer,Description="Number of samples\
not called">
##FILTER=<ID=str10,Description="Less than 10% or more than 90% of\
variant supporting reads on one strand">
##FILTER=<ID=indelError,Description="Likely artifact due to indel\
reads at this position">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=SDP,Number=1,Type=Integer,Description="Raw Read Depth as\
reported by SAMtools">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Quality Read Depth\
of bases with Phred score >= 15">
##FORMAT=<ID=RD,Number=1,Type=Integer,Description="Depth of reference-\
supporting bases (reads1)">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Depth of variant-\
supporting bases (reads2)">
##FORMAT=<ID=FREQ,Number=1,Type=String,Description="Variant allele\
frequency">
```

```
##FORMAT=<ID=PVAL,Number=1,Type=String,Description="P-value from\
Fisher's Exact Test">
##FORMAT=<ID=RBQ,Number=1,Type=Integer,Description="Average quality of\
reference-supporting bases (qual1)">
##FORMAT=<ID=ABQ,Number=1,Type=Integer,Description="Average quality of\
variant-supporting bases (qual2)">
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  Sample1
```

After a double-pound sign ('##') fields being present in INFO and FORMAT columns are described, as they can be arbitrarily customised. The meta-information makes sure that the data can be automatically parsed and analysed independently what a variant caller reported. The mandatory line starting with a single pound ('#') provides a description of the present columns. The body contains of at least eight mandatory columns with one row per variant, e.g. :

```
chr7 36858643 rs10255208 A G . PASS ADP=633;WT=0;HET=1;HOM=0;NC=0;\
GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ\
0/1:255:635:633:327:306:48.34%:2.5945E-114:51:53
```

The ID field contains a dbSNP database entry, if there is one. `FORMAT` lists all present subfields, while Sample1 is a generic placeholder for the sample ID. It should be renamed, to allow efficient merging or processing, even if file names change. In addition, the field lists the obtained values for that sample from the variant caller. If multiple samples have been analysed simultaneously, they would be added as extra columns, each with a unique ID assigned. The format allows adding

further fields subsequently by updating the header and attaching the values to the INFO field. In a similar way, filtering can be performed and written into the FILTER column, by adding filter criteria to the header accordingly.

1.4.5 Genetic Variant Annotation and Effect Prediction

Every human being carries millions of SNVs, InDels and other structural variants compared to the official reference sequence, such as hg19 or hg38, without showing any abnormal phenotypes or diseases [20]. The logical consequence is that a variant on its own is not sufficient as a diagnostic marker, unless it is integrated with further information from other sources, such as association studies, structural, phenotypic or pathway information. A number of software tools for such an analysis exist, such as ANNOVAR [126], VAAST [45] or DAVID [46]. Within the scope of this thesis, variant annotation and effect prediction is performed with a combination of SnpEff and SnpSift [17, 18]. Both tools are written in Java and work very well together, are reasonably fast, support natively the latest VCF file format and can quickly be integrated into existing pipelines. SnpEff collects information from many different resources and provides them bundled in a database. This is used for functional annotations of variants, which are then written back to the VCF file, by adding the ANN tag into the INFO section. Every annotation contains 16 mandatory fields separated by a vertical bar,

which are listed in Table 1.7. For each variant there can be multiple effect annotations, which are ordered decreasingly by the predicted impact.

Field	Description
1. Allele/ALT	Alternative base (repeat from VCF file column 5)
2. Annotation	Effect annotation(s) using sequence ontology from table 1.1
3. Impact	Estimation of the putative impact on the protein, analogous to table 1.1
4. Gene Name	Gene name of nearest gene
5. Gene ID	Gene ID of nearest gene
6. Feature type	Type of nearest annotated feature
7. Feature ID	ID of nearest annotated feature
8. Transcript	Coding, Non-coding or ENSEMBL biotype
9. Rank	Exon/intron rank / total number of exons/introns
10. HGVS.c	Variant in DNA HGVS annotation
11. HGVS.p	Variant in Protein HGVS annotation, if available
12. cDNA	Position in cDNA / cDNA length

13. CDS	Coding sequence position / length of coding sequence
14. Protein	Position and protein of affected amino acid / length of protein
15. Distance	Distance to nearest feature, if useful
16. Errors	Any errors from the search, if any

TABLE 1.7: The 16 mandatory fields of the SnpEff ANN sub-field [125]. The annotation allows further downstream filtering based on added information, such as gene name, HGVS or impact.

SnpSift is a collection of tools for arbitrary manipulations of VCF files. This includes filtering, further annotation with any additional variant databases, such as COSMIC [34], dbSNP [109], dbNSFP [73] or ClinVar [60]. While SnpEff focusses on the structural annotation, SnpSift has its strengths in combining functional and phenotyping information about the called variants and filter them by a given criteria. As SnpEff and SnpSift are maintained by the same group, they both benefit from a similar syntax and respectively accept the output of one tool as their own input. If additional databases are provided for further annotation, SnpSift checks the first two fields in a row in the list of variants and compares it to the database, if it has been found, it adds the entire line from the database to the INFO field. Annotated variants allow better predictions of the functional impact, especially if the mutation

is reported and the predicted outcome has been supported by clinical studies.

1.5 Scope of Thesis

Massively parallel sequencing has reached a point where robustness, throughput, cost and turnaround time can compete with other assays that have been briefly described in this chapter. Many target capture methods have been released and allow limiting sequencing to regions of interest. Although this is a crucial step in diagnostic methods to decrease cost and time, while focussing on genes that show clinical relevance, such as genetic markers, it is prone to introducing artefacts, which could lead to false results. Hence, only very few target enrichment panels are validated and, therefore, suitable for diagnostic procedures, most of which are designed for a very specific disease making them potentially unattractive for other applications or cancer types.

Within the scope of this thesis two target capture methods are investigated and applied on a large scale as a base for clinical validation. Further, a new approach is described to maintain information from which cell sequencing reads were derived, by introducing a degenerative barcode into each amplicon during library preparation followed by sequencing on a MiSeq instrument.

In Chapter 1 the fundamental principle, design process and enrichment by hybridisation and target capture based on the Agilent SureSelect XT technology is described. The designed panel was tested on two Horizon reference standards: samples with annotated mutations with known frequencies in a number of genes. The results were used to boost poorly performing regions by adjusting capture probe concentration. The revised panel was tested on 278 clinical samples from various primary tumour tissues that were previously pyrosequenced for a profound sensitivity estimation in a selection of mutation hotspots. The genes that were targeted with this enrichment panel were selected based on scientific and clinical impact described in the literature for a broad range of different cancer types. The sequencing results were analysed to define input quality criteria and panel limitations. Further, best practices for analysis, variant calling and filtering were determined.

In Chapter 2 a custom designed target enrichment panel based on the Agilent HaloPlex HS system is introduced. The panel targets 9 genes and was designed to replace current assays in clinical diagnostics based on molecular inversion probes for target enrichment and amplification. A panel based on molecular inversion probes was chosen due to a simplified protocol and faster preparation process compared to conventional enrichment via target capture. Further, it overcomes various issues that can arise from other enrichment methods, such as amplicon sequencing. Based on 48 clinical samples that were enriched and sequenced the panel was assessed and general difficulties with molecular

inversion probes enriching regions from DNA extracted from FFPE tissue are explained.

In Chapter 3 a new method of high-throughput cellular barcoding of target regions by encapsulation of cells into droplets via emulsion and performing a direct PCR using uniquely barcoded primer libraries attached to small microparticles, called *beads* is described. The first section briefly explains the protocol of how beads are loaded with barcoded primers, while the second part introduces the cellular barcoding method. Subsequently the results from a pilot experiment from two mixed NIH3T3 and K562 cell cultures will be presented. Both cultures were sequenced with pyrosequencing in KRAS exon 2 and 3 to test for any present mutations. NIH3T3 carried a heterozygous SNV in KRAS exon 2, whilst no SNVs or InDels could be found in K562. Cells from both cultures were mixed in an 4:1 ratio followed by preparation, sequencing and analysis. The heterozygous mutation in NIH3T3 cells could be confirmed, but also 13 other low-frequency mutations were found that would have been missed with conventional amplicon sequencing. Clustering analysis of a selection of identified cells, revealed a cellular evolution of both cell cultures. Moreover, present artefacts and potential error sources arising from high-throughput cellular barcoding based on direct emulsion PCR are described and possible solutions are given. The thesis closes with a brief summary of the discussed experiments and gives future perspectives that can build on the results of this research.

References

[1] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[2] *An Introduction to Next-Generation Sequencing Technology*. http://www.illumina.com/technology/next-generation-sequencing.html. Illumina. May 2016.

[3] JK Aronson. "Biomarkers and surrogate endpoints". In: *British journal of clinical pharmacology* 59.5 (2005), pp. 491–494.

[4] Carlos F Barbas et al. "Quantitation of DNA and RNA". In: *Cold Spring Harbor Protocols* 2007.11 (2007), pdb–ip47.

[5] David R Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *nature* 456.7218 (2008), pp. 53–59.

[6] J Brabender et al. "Epidermal growth factor receptor and HER2-neu mRNA expression in non-small cell lung cancer is correlated with survival". In: *Clinical Cancer Research* 7.7 (2001), pp. 1850–1855.

[7] Broadinstitue. *Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*. http://broadinstitute.github.io/picard. Nov. 2015.

[8] Vince Buffalo. *qrqc: Quick Read Quality Control*. R package version 1.24.0. 2012. URL: http://github.com/vsbuffalo/qrqc.

[9] Rebecca A Burrell and Charles Swanton. "The evolution of the unstable cancer genome". In: *Current opinion in genetics & development* 24 (2014), pp. 61–67.

[10] Rebecca A Burrell and Charles Swanton. "Tumour heterogeneity and the evolution of polyclonal drug resistance". In: *Molecular oncology* 8.6 (2014), pp. 1095–1111.

[11] Gianni Bussolati et al. "Formalin fixation at low temperature better preserves nucleic acid integrity". In: *PLoS One* 6.6 (2011), e21043.

[12] Bruno Canard and Robert S Sarfati. "DNA polymerase fluorescent substrates with reversible 3-tags". In: *Gene* 148.1 (1994), pp. 1–6.

[13] Kirstie Canene-Adams. "Preparation of formalin-fixed paraffin-embedded tissue for immunohistochemistry". In: *Methods in enzymology* 533 (2012), pp. 225–233.

[14] Danielle Mercatante Carrick et al. "Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue". In: *PloS one* 10.7 (2015), e0127353.

[15] Daniel E Carvajal-Hausdorf et al. "Quantitative measurement of cancer tissue biomarkers in the lab and in the clinic". In: *Laboratory Investigation* 95.4 (2015), pp. 385–396.

[16] Gary A Churchill. "Fundamentals of experimental design for cDNA microarrays". In: *Nature genetics* 32 (2002), pp. 490–495.

[17] P. Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". In: *Fly* 6.2 (2012), pp. 80–92.

[18] P. Cingolani et al. "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift". In: *Frontiers in Genetics* 3 (2012).

[19] Peter JA Cock et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic acids research* 38.6 (2010), pp. 1767–1771.

[20] 1000 Genomes Project Consortium et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–1073.

[21] Mark A Dawson and Tony Kouzarides. "Cancer epigenetics: from mechanism to therapy". In: *Cell* 150.1 (2012), pp. 12–27.

[22] Mark A DePristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature genetics* 43.5 (2011), pp. 491–498.

[23] L Peter Deutsch. "GZIP file format specification version 4.3". In: (1996).

[24] Lloye M Dillon and Todd W Miller. "Therapeutic targeting of cancers with loss of PTEN function". In: *Current drug targets* 15.1 (2014), p. 65.

[25] Hui Dong et al. "Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System". In: *Acta biochimica et biophysica Sinica* 43.6 (2011), pp. 496–500.

[26] N J Dovichi and Jianzhong Zhang. "How capillary electrophoresis sequenced the human genome". In: *Angewandte Chemie International Edition* 39.24 (2000), pp. 4463–4468.

[27] John W Drake et al. "Rates of spontaneous mutation". In: *Genetics* 148.4 (1998), pp. 1667–1686.

[28] Aron C Eklund and Zoltan Szallasi. "Correction of technical bias in clinical microarray data improves concordance with known biological information". In: *Genome Biol* 9.2 (2008), R26.

[29] Mark G Erlander et al. "Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification". In: *The Journal of Molecular Diagnostics* 13.5 (2011), pp. 493–503.

[30] Estevezj. *Sanger sequencing*. https://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg. Dec. 2012.

[31] Brent Ewing and Phil Green. "Base-calling of automated sequencer traces using phred. II. Error probabilities". In: *Genome research* 8.3 (1998), pp. 186–194.

[32] Milan Fedurco et al. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies". In: *Nucleic acids research* 34.3 (2006), e22–e22.

[33] Cathy A Finlay, Philip W Hinds, and Arnold J Levine. "The p53 proto-oncogene can act as a suppressor of transformation". In: *Cell* 57.7 (1989), pp. 1083–1093.

[34] SA1 Forbes et al. "The catalogue of somatic mutations in cancer (COSMIC)". In: *Current protocols in human genetics* (2008), pp. 10–11.

[35] Marcus Gassmann and Barry McHoull. *DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System & Genomic DNA ScreenTape*. Tech. rep. G5991-5258EN. Application note. Agilent Technologies, 2014.

[36] GeneChip. *Microarray*. Computer Desktop Encyclopedia. Curtesy of Affymetrix. June 2015.

[37] Annuska M Glas et al. "Converting a breast cancer microarray signature into a high-throughput diagnostic test". In: *BMC genomics* 7.1 (2006), p. 1.

[38] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[39] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer". In: *cell* 100.1 (2000), pp. 57–70.

[40] David Hansemann. "Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung". In: *Virchows Archiv* 119.2 (1890), pp. 299–326.

[41] Jakob Hedegaard et al. "Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue". In: *PloS one* 9.5 (2014), e98187.

[42] Michael J Heller. "DNA microarray technology: devices, systems, and applications". In: *Annual review of biomedical engineering* 4.1 (2002), pp. 129–153.

[43] Russell Higuchi et al. "Simultaneous amplification and detection of specific DNA sequences". In: *Bio/technology* 10.4 (1992), pp. 413–417.

[44] Nils Homer and Stanley F Nelson. "Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA". In: *Genome Biol* 11.10 (2010), R99.

[45] Hao Hu et al. "VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix". In: *Genetic epidemiology* 37.6 (2013), pp. 622–634.

[46] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic acids research* 37.1 (2009), pp. 1–13.

[47] *Illumina Experiment Manager User Guide*. 1.9 (15031335 Rev. J). ILLUMINA PROPRIETARY. Illumina. Mar. 2015.

[48] Illumina. *Sequencing power for every scale*. Website. http://www.illumina.com/systems/sequencing.html. May 2016.

[49] Hongshan Jiang et al. "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads". In: *BMC bioinformatics* 15.1 (2014), p. 1.

[50] Hyunju Jung et al. *The DNA Integrity Number (DIN) Provided by the Genomic DNA ScreenTape Assay Allows for Streamlining of NGS on FFPE Tissue Samples*. Tech. rep. 5991-5360EN. Application note. Agilent Technologies, 2014.

[51] Evangelia Karampetsou, Deborah Morrogh, and Lyn Chitty. "Microarray Technology for the Diagnosis of Fetal Chromosomal Aberrations: Which Platform Should We Use?" In: *Journal of clinical medicine* 3.2 (2014), pp. 663–678.

[52] Graham Kemp. "Capillary Electrophoresis". In: *Biotechnology and applied biochemistry* 27.1 (1998), pp. 9–17.

[53] Scott E Kern. "Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures". In: *Cancer research* 72.23 (2012), pp. 6097–6101.

[54] Christoph Kirsch and Eva Schmidt. *The DNA Integrity Number (DIN) Provided by the Agilent 2200 TapeStation System is an Ideal Tool to Optimize FFPE Extraction*. Tech. rep. 5991-5246EN. Application note. Agilent Technologies, 2015.

[55] Daniel C Koboldt et al. "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome research* 22.3 (2012), pp. 568–576.

[56] Lasse S Kristensen et al. "Quality assessment of DNA derived from up to 30 years old formalin fixed paraffin embedded (FFPE) tissue for PCR-based methylation analysis using SMART-MSP and MS-HRM". In: *BMC cancer* 9.1 (2009), p. 1.

[57] Theodore G Krontiris and Geoffrey M Cooper. "Transforming activity of human tumor DNAs". In: *Proceedings of the National Academy of Sciences* 78.2 (1981), pp. 1181–1184.

[58] Mikael Kubista et al. "The real-time polymerase chain reaction". In: *Molecular Aspects of Medicine* 27.2-3 (2006). Real-time Polymerase Chain Reaction, pp. 95 –125. ISSN: 0098-2997.

[59] Sunjong Kwon. "Single-molecule fluorescence in situ hybridization: quantitative imaging of single RNA molecules". In: *BMB reports* 46.2 (2013), pp. 65–72.

[60] Melissa J Landrum et al. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic acids research* 42.D1 (2014), pp. D980–D985.

[61] Pennina R Langer-Safer, Michael Levine, and David C Ward. "Immunological method for mapping genes on Drosophila polytene chromosomes". In: *Proceedings of the National Academy of Sciences* 79.14 (1982), pp. 4381–4385.

[62] Darryl Leja. *Fluorescence In Situ Hybridization (FISH)*. 2010. URL: https://www.genome.gov.

[63] Arnold J Levine. "p53, the cellular gatekeeper for growth and division". In: *cell* 88.3 (1997), pp. 323–331.

[64] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.

[65] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997* (2013).

[66] Heng Li. *Sequence Alignment/Map Format Specification*. Tech. rep. The SAM/BAM Format Specification Working Group, Nov. 2015.

[67] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5 (2010), pp. 589–595.

[68] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[69] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome research* 18.11 (2008), pp. 1851–1858.

[70] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[71] Astrid Lievre et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer". In: *Cancer research* 66.8 (2006), pp. 3992–3995.

[72] David J Lipman and William R Pearson. "Rapid and sensitive protein similarity searches". In: *Science* 227.4693 (1985), pp. 1435–1441.

[73] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions". In: *Human mutation* 32.8 (2011), pp. 894–899.

[74] Elise MJ van der Logt et al. "Fully Automated Fluorescent in situ Hybridization (FISH) Staining and Digital Analysis of HER2 in Breast Cancer: A Validation Study". In: *PloS one* 10.4 (2015), e0123201.

[75] Katja Lohmann and Christine Klein. "Next generation sequencing and the future of genetic diagnosis". In: *Neurotherapeutics* 11.4 (2014), pp. 699–707.

[76] Dianne I Lou et al. "High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing". In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19872–19877.

[77] Fotios Loupakis et al. "PTEN expression and KRAS mutations on primary tumors and metastases in the prediction of benefit from cetuximab plus irinotecan for patients with metastatic colorectal cancer". In: *Journal of Clinical Oncology* 27.16 (2009), pp. 2622–2629.

[78] Umberto Malapelle et al. "Sanger sequencing in routine KRAS testing: a review of 1720 cases from a pathologist's perspective". In: *Journal of clinical pathology* (2012), jclinpath–2012.

[79] A Marusyk and K Polyak. "Tumor heterogeneity: causes and consequences". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 105–117.

[80] Tim Massingham and Nick Goldman. "All Your Base: a fast and accurate probabilistic approach to base calling". In: *Genome Biol* 13.2 (2012), R13.

[81] Patrick Micke et al. "Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens". In: *Laboratory investigation* 86.2 (2006), pp. 202–211.

[82] Lance D Miller et al. "Optimal gene expression analysis by microarrays". In: *Cancer cell* 2.5 (2002), pp. 353–361.

[83] André E Minoche, Juliane C Dohm, Heinz Himmelbauer, et al. "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems". In: *Genome Biol* 12.11 (2011), R112.

[84] Lisle E Mose et al. "ABRA: improved coding indel detection via assembly-based realignment". In: *Bioinformatics* 30.19 (2014), pp. 2813–2815.

[85] Yuki Nakayama et al. "Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions". In: *PloS one* 11.3 (2016), e0150528.

[86] Soo Kyung Nam et al. "Effects of fixation and storage of human tissue samples on nucleic Acid preservation". In: *Korean journal of pathology* 48.1 (2014), p. 36.

[87] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond". In: *Genome research* 25.10 (2015), 1499–1507.

[88] Lex Nederbragt. "Developments in NGS". In: (2015).

[89] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[90] K Page et al. "Detection of HER2 amplification in circulating free DNA in patients with breast cancer". In: *British journal of cancer* 104.8 (2011), pp. 1342–1348.

[91] Joseph F. Paone et al. "Serum UDP-galactosyl transferase as a potential biomarker for breast carcinoma". In: *Journal of Surgical Oncology* 15.1 (1980), pp. 59–66. ISSN: 1096-9098. DOI: 10.1002/jso.2930150110.

[92] Deniz Pekin et al. "Quantitative and sensitive detection of rare mutations using droplet-based microfluidics". In: *Lab on a Chip* 11.13 (2011), pp. 2156–2166.

[93] SW Piraino and SJ Furney. "Beyond the exome: the role of non-coding somatic mutations in cancer". In: *Annals of Oncology* 27.2 (2016), pp. 240–248.

[94] *PyroMark PCR Handbook*. Qiagen. May 2009.

[95] *PyroMark Q24 MDx User Manual*. Qiagen. Jan. 2016.

[96] Michael A Quail et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC genomics* 13.1 (2012), p. 1.

[97] JA Ramos-Vara and MA Miller. "When Tissue Antigens and Antibodies Get Along Revisiting the Technical Aspects of ImmunohistochemistryThe Red, Brown, and Blue Technique". In: *Veterinary Pathology Online* 51.1 (2014), pp. 42–87.

[98] *Real-time PCR handbook*. CO32085 0812. With curtesy of Thermo Fischer Scientific. Thermo Fischer Scientific. Aug. 2012.

[99] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. "The advantages of SMRT sequencing". In: *Genome Biol* 14.6 (2013), p. 405.

[100] James T Robinson et al. "Integrative genomics viewer". In: *Nature biotechnology* 29.1 (2011), pp. 24–26.

[101] Mark D Robinson, Alicia Oshlack, et al. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biol* 11.3 (2010), R25.

[102] Scott J Rodig et al. "Unique clinicopathologic features characterize ALK-rearranged lung adenocarcinoma in the western population". In: *Clinical Cancer Research* 15.16 (2009), pp. 5216–5223.

[103] Mostafa Ronaghi et al. "Real-time DNA sequencing using detection of pyrophosphate release". In: *Analytical biochemistry* 242.1 (1996), pp. 84–89.

[104] Michael G Ross et al. "Characterizing and measuring bias in sequence data". In: *Genome Biol* 14.5 (2013), R51.

[105] Fred Sanger and Alan R Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of molecular biology* 94.3 (1975), pp. 441–448.

[106] Mark Schena et al. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235 (1995), pp. 467–470.

[107] Suzanne Schubbert, Kevin Shannon, and Gideon Bollag. "Hyperactive Ras in developmental disorders and cancer". In: *Nature Reviews Cancer* 7.4 (2007), pp. 295–308.

[108] Michal R Schweiger et al. "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number-and mutation-analysis". In: *PloS one* 4.5 (2009), e5548.

[109] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[110] Shan-Rong Shi et al. "Evaluation of the value of frozen tissue section used as gold standard for immunohistochemistry". In: *American journal of clinical pathology* 129.3 (2008), pp. 358–366.

[111] Chiaho Shih et al. "Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts". In: (1981).

[112] Mano Sivaganesan et al. "Improved strategies and optimization of calibration models for real-time PCR absolute quantification". In: *Water research* 44.16 (2010), pp. 4726–4735.

[113] Andrew M Smith et al. "Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples". In: *Nucleic acids research* (2010), gkq368.

[114] Temple F Smith and Michael S Waterman. "Identification of common molecular subsequences". In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[115] Jérome Solassol et al. "KRAS mutation detection in paired frozen and formalin-fixed paraffin-embedded (FFPE) colorectal cancer tissues". In: *International journal of molecular sciences* 12.5 (2011), pp. 3191–3204.

[116] Lucy F Stead et al. "Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution". In: *Human mutation* 34.10 (2013), pp. 1432–1438.

[117] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (2009), pp. 719–724.

[118] Leinco Technologies. *Immunohistochemistry Protocol for Frozen Sections*. Website. Curtesy of Leinco Technologies. 2015.

[119] Steven M Teutsch et al. "The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP working group". In: *Genetics in Medicine* 11.1 (2009), pp. 3–14.

[120] K Thaker, R Shah, and M Berger. "The IMPACT of INDEL realignment: Detecting insertions and deletions longer than 30 base pairs with ABRA". Poster. Nov. 2014.

[121] Athanasios C Tsiatis et al. "Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications". In: *The Journal of Molecular Diagnostics* 12.4 (2010), pp. 425–432.

[122] Eliezer M Van Allen et al. "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine". In: *Nature medicine* 20.6 (2014), pp. 682–688.

[123] EH Van Beers et al. "A multiplex PCR predictor for aCGH success of FFPE samples". In: *British journal of cancer* 94.2 (2006), pp. 333–337.

[124] Laura J Van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *nature* 415.6871 (2002), pp. 530–536.

[125] *Variant annotations in VCF format*. http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf. Cingolani, Pablo et al. Jan. 2015.

[126] Kai Wang, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". In: *Nucleic acids research* 38.16 (2010), e164–e164.

[127] Kai Wang et al. "Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer". In: *Nature genetics* 46.6 (2014), pp. 573–582.

[128] Stephen Q Wong et al. "UV-Associated Mutations Underlie the Etiology of MCV-Negative Merkel Cell Carcinomas". In: *Cancer research* 75.24 (2015), pp. 5228–5234.

[129] Hongping Xia and Kam M. Hui. "Mechanism of Cancer Drug Resistance and the Involvement of Noncoding RNAs". In: *Current Medicinal Chemistry* 21.26 (2014), pp. 3029–3041. ISSN: 0929-8673/1875-533X. DOI: 10.2174/09298673216661404141019 39.

[130] Shawn E Yost et al. "Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast

cancer specimens". In: *Nucleic acids research* 40.14 (2012), e107–e107.

[131] *bcl2fastq2 Conversion Software Guide*. 2.17 (15051736 Rev. G). ILLUMINA PROPRIETARY. Illumina. July 2015.

Chapter 2

Design of a Comprehensive Cancer Panel for Precision Medicine

Testing genetic markers for their prognostic and diagnostic potential requires focusing to a limited number of regions to be targeted, such as the exome, genes, exons or even just a few commonly mutated codons, often referred as mutation hotspots, which are supported by clinical studies. A comprehensive sequencing panel should ideally have full diagnostic and prognostic potential, i.e. identification of the causes of cancer and predicting a likely outcome. In this chapter the design of the

panel is described and results of an initial pilot experiment for panel improvement and subsequent test on 278 FFPE samples are discussed.

2.1 Introduction

In precision medicine standardised and validated protocols need to be in place to allow fast processing and a reliable, clear result. Hence, off-the-shelf target sequencing panels or kits have been released to ensure a stable environment that is robust and suitable for automation. Customisable target enrichment kits are based on different technologies like the off-the-shelf solutions, but benefit from relatively arbitrary regions than can be individually defined. Two main methods for target enrichment are commonly used. The first method is based on multiplexed PCRs to amplify regions of interest, called *amplicon-seq*. The second method captures targets with designed baits or probes by hybridisation and washing off DNA fragments of unwanted regions. Multiplexed PCR approaches benefit from a short preparation time, but the number and length of regions that can be targeted is limited. In addition, design and optimisation of multiplexed PCR primers amplifying every region equally well, disregarding limiting factors -such as GC content- and without biasing allele frequencies, is a considerable challenge often causing unsatisfactory results [146]. In cases where it is desired to preserve the allele frequencies throughout the enrichment and sequencing, artefacts raising from PCR amplification should be avoided. Mutation

frequency of TP53, for example, is used as an indicator of tumour progression [187, 193]. Hence, a method preserving information from the input material is preferred. Target capture methods, like the Agilent Custom SureSelect XT, are preferred, as their enrichment principle is not based on PCR. Extracted DNA is sheared into smaller fragments, which are then prepared for sequencing and subsequently target regions are captured and amplified [208]. Hence, an entire library is amplified rather than individual regions. Another important part of designing a custom designed cancer panel is to select the genes to be tested for. Current gene selections from cancer panels that are commercially available show a broad selection of different combination of genes. They are either associated with specific cancer types or covering a large selection of cancer genes. Some panels offer entire exons, others limit their focus on mutation hotspots connected with a known clinical significance. Depending on the application, both attempts are useful. Targeting mutation hotspots imply a certain hypothesis about samples that are tested, which leads to a clear outcome. Sequencing entire exons or even genes can, in contrast, be seen as a hypothesis-free approach, where all findings are reported leaving room for interpretation of the results [172]. Hypothesis-free approaches allow researchers and doctors to explore the cancer genome for scientific purposes, such as genetic marker detection. Moreover, if a known cancer gene is reported to be clinically relevant re-testing can be evaded if the gene was already sequenced. Cancer tissue samples are precious and re-testing would

require another potentially invasive operation or are even impossible in cases where a patient deceased or the tumour was defeated. Most cancer panels are not clinically validated to be used for diagnostic and prognostic testing and modifications to a kit would require at least a partial re-validation. Every additional gene picked, however, increases the amount of sequencing required. It means that target sequencing for a broad range of applications requires a carefully balanced trade-off between the amount of sequencing and the number of genes covered. The cancer panel designed in this context consists of 69 genes with known diagnostic and prognostic potential. They are selected tumour suppressor and oncogenes playing a major role in many common cancer types, such as colon, colorectal cancer, melanoma, lung and breast cancer. Current trends in literature were considered to pick important or versatile genes playing either a key role in a specific type or are often reported to be important in many different cancer types. The resulting gene list is shown in Table 2.1.

Gene symbol	**Often referred cancer types**	**Clinical relevance**	**Source**
ABL1	Leukaemia	Tyrosine kinase inhibitor resistance	[207]
AKT1/2/3	Non-small lung, ovarian, breast, colorectal, pancreatic	AKT inhibition sensitivity	[163, 133]

ALK	Non-small lung, neuroblastoma, anaplastic large-cell lymphoma	ALK inhibition sensitivity	[194]
APC	Colon, pancreatic, liver, colorectal, stomach, desmoid, hepatoblastoma, glioma	Tumour suppressor	[137]
ATM	Lymphoma, leukaemia, glioma, melanoma, prostate, breast, stomach, bladder, lung, ovarian	Tumour suppressor	[171]

BRAF	Melanoma, non-small lung, colorectal, skin, leukaemia	Proteasome inhibitor resistance, oncogene	[151, 166]
CBL	Leukaemia, myelodysplastic syndrome	Tumour suppressor	[140]
CDH1	Gastric, breast, prostate	Tumour suppressor	[158]
CDK4	Melanoma	Cyclin-dependant kinase inibition sensitivity	[138, 196]
CDKN2A	Melanoma, pancreatic, glioma, ovarian, lung, skin, leukaemia	Tumour suppressor	[154]
CEBPA	Leukaemia	Tumour suppressor	[180]
CRLF2	Leukaemia	Oncogene	[147]

CSF1R	Intestinal, skin, stomach, leukaemia	Oncogene, colony-stimulating factor-1 receptor inhibitor sensitivity	[188]
CTNNB1	Colorectal, ovarian, desmoid, melanoma, neoplasm	Beta-catenin inhibition sensitivity, oncogene	[135]
EGFR	Non-small lung, glioma, pancreatic, neoplasm, brain, colorectal, prostate, colon	Oncogene, EGFR inhibition sensititvity, monoclonal antibody inhibition sensitivity	[212]
ERBB2/4	Breast, gastic, stomach, uterine, salivary duct, glioma, non-small lung, ovarian, neuroblastoma	Oncogene, monoclonal antibody inhibition sensitivity	[210, 162, 195]

EZH2	Prostate, breast, bladder, uterine, renal, melanoma, lymphoma	Oncogene, EZH2 inhibition sensitivity	[145, 202]
FBXW7	Breast, colorectal, leukaemia, lung, skin	Tumour suppressor	[206]
FGFR1/2/3	Lymphoma,breast, non-small lung, gastric, bladder, myeloma, breast	Oncogene, FGFR inhibition sensitivity	[169]
FLT3	Leukaemia, colorectal	Oncogene, FLT3 inihibition sensitivity	[157]
FOXL2	Ovarian, testicular	Tumour suppressor	[141]
GATA1/2	Leukaemia, non-small lung, colorectal, skin	Oncogene, Proteasome inhibitor sensitivity	[197, 174]

KDR	Colorectal, non-small lung, intestinal, skin	Oncogene	[198]
KIT	Melanoma, leukaemia, gastric, lymphoma, breast, neoplasm, colorectal, bladder, liver	Oncogene, tyrosine-kinase inhibitor sensitivity, tyrosine-kinase inhibitor resistance	[178]
KRAS	Lung, colorectal, pancreatic, leukaemia, neuroblastoma, colon, skin, breast	Oncogene, EGFR inhibition resistance	[204]
MAP2K1	Non-small lung, melanoma, colorectal	Tumour suppressor	[144]

MET	Gastric, colorectal, glioma, ovarian, small lung, breast	Oncogene, EGFR inhibition resistance, MET inhibition sensitivity	[155, 139]
MLH1	Stomach, non-small lung, colorectal, ovarian, leukaemia, colon, endometrial, intestinal, skin	Tumour suppressor	[143]
MPL	Colorectal, leukaemia, lung, skin	Oncogene	[192]
NF1/2	Breast, leukaemia, melanoma, colorectal, glioma, lung, ovarian	Tumour suppressor	[134]

NOTCH1/2	Oesophageal, leukaemia, stomach, intestinal, skin, breast	Oncogene, tumour suppressor, NOTCH inhibition sensitivity	[205, 181, 165]
NPM1	Leukaemia, skin	Oncogene	[159]
NRAS	Melanoma, leukaemia, myeloma, colorectal	Oncogene, EGFR inhibition resistance	[183]
PDGFRA	Gastric, leukaemia	Oncogene, PDGF inhibition sensitivity	[164]
PIK3CA	Colon, glioma, gastric, breast, enome-trial,neoplasm, lung	Oncogene, phosphoinositide 3-kinases inhibition sensitivity	[182]
PIK3R1/5	Breast, endometrial, renal, intestinal cancer, skin, brain	Oncogene, phosphoinositide 3-kinases inhibition sensitivity	[200, 149]

PTCH1	Basal cell, bone, skin, vulvar, gastric, colorectal	Tumour suppressor	[148]
PTEN	Prostate, endometrial, glioblastoma,	Tumour suppressor, PI3K/AKT/mTor inhibition sensititvity, PTEN inhibition sensitivity,	[179, 24]
PTPN11	Leukaemia, neuroblastoma, melanoma, breast, lung, colorectal, stomach, endometrial, intestinal	Oncogene, protein tyrosine phosphatase inhibition sensitivity	[213, 142]
RB1	Small lung, breast, sarcoma, eye, fallopian tube, intestinal	Tumour suppressor	[156]

RET	Lung, sarcoma, breast, eye, medullary thyroid, breast	Oncogene, RET kinase inhibition sensitivity	[185]
RUNX1	Leukaemia, neoplasm, ovarian, breast	Tumour suppressor	[167]
SMAD4	Colorectal, pancreatic, gastric	Tumour suppressor	[132]
SMARCB1	Bone, intestinal, rhabdoid, central nervous system	Tumour suppressor	[176]
SMO	Basal cell, glioblastoma, medulloblastoma, skin	Oncogene, SMO inhibition sensitivity	[191]
SRC	Colon, breast, prostate, skin	Oncogene, SRC/ABL inhibition sensitivity	[186]

STK11	Non-small lung, ovarian, pancreatic, testicular, gastric, cervical, skin, breast	Tumour suppressor	[175]
TET2	Prostate, leukaemia, neoplasm	Tumour suppressor	[199]
TP53	Lung, ovarian, colon, oesophageal, neoplasm, skin, coloretal, breast, glioma and many others	Tumour suppressor	[184]
TSHR	Intestinal, skin, thyroid	Oncogene	[209]

VHL	Renal, kidney, paratesticular, intestinal	Tumour suppressor, VEGF inhibition sensitivity, mTOR inhibition sensitivity, monoclonal anitbody inhibition sensitivity	[153]
WT1	Leukaemia, kidney, lung, skin	Oncogene, tumour suppressor, monoclonal antibody sensitivity	[150, 211]

TABLE 2.1: The selected 69 genes for the comprehensive bait-based sequencing panel. Tumour suppressor and oncogenes indicate prognostic capabilities, while sensitivties or resistances indicate a diagnostic potential.

The panel is beneficial for clinicians and researches who are in need of a broad view across typically mutated and druggable targets for the next few years. For example, in cases where a patient suffers from a disease and a screening might reveal a drug sensitivity in a different tumour type. As the list of references indicates, most findings have been published over the past two to three years indicating that many of these genes are under further investigation and may be utilised as genetic markers in the future.

2.2 Probe Design and Evaluation on 2 Pilot Samples

From the gene list introduced above, a list of hg19 genome coordinates was derived spanning all exons of the target genes including 10 bases up and downstream at both ends. Over 20,000 different molecular probes were designed spanning all targeted regions, called flanking regions. These molecular probes are single-stranded, biotinylated RNA fragments of 120bp in length. These probes hybridise to complementary sequences present in the target regions. Due to the length of the probes, they are robust against minor variation in the region, so that they still hybridise strongly enough in cases of SNVs or short InDels being present. A brief overview of the selection process is shown in Figure 2.1.

The probe cocktail was used for a first pilot experiment to assess enrichment and sequencing performance of the kit. Extracted genomic DNA from two pooled FFPE cell cultures were prepared, enriched and sequenced on a MiSeq instrument. Sample BRAF20 (Horizon Catalogue ID HD232) was a mix of two cell lines of which one had a heterozygous mutation in BRAF (p.V600E), while the other one was reported WT at that position. Both cell lines were mixed by Horizon to reach a final allelic frequency of 20%. Sample QUANTREF (Horizon Catalogue ID HD200) was a mix of multiple cell lines carrying 11 known mutations in various frequencies listed in Table A.1. After comparison of both

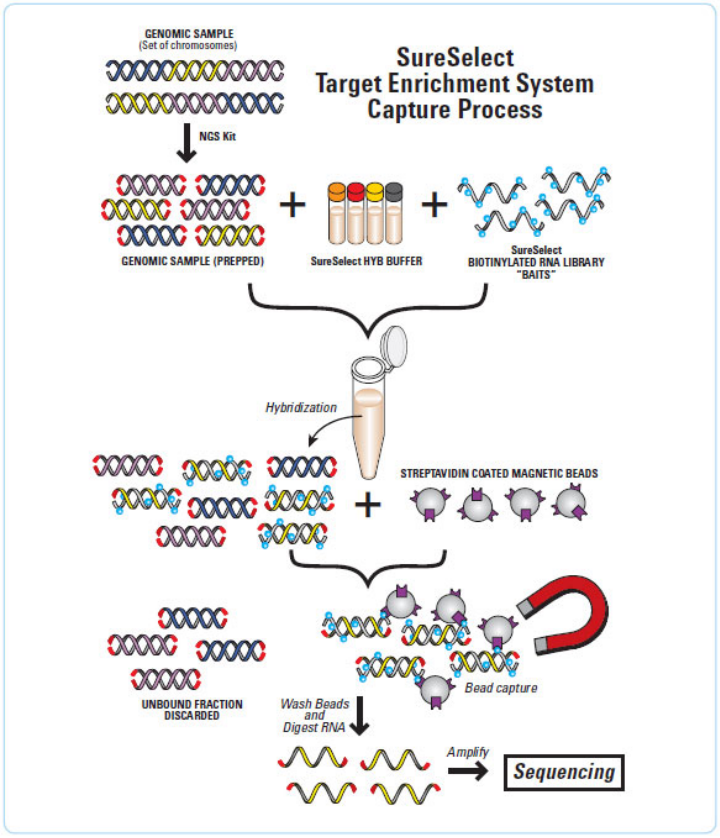


FIGURE 2.1: DNA is sheared into small fragments and sequencing adapters are ligated. Biotinylated RNA probes, hybridise to complementary fragments. Streptavidin coated iron beads are added and capture probes with hybridsed fragments. Captured DNA fragments are amplified and sequenced. Image from Agilent [201].

kits, 278 clinical samples of different primary tumour types were prepared, sequenced and analysed using the revised kit (version 2) and an Illumina NextSeq 500 sequencing instrument. In this section the protocol is briefly described and the results of the two kit versions are compared.

2.2.1 DNA Extraction from FFPE tissue and Quantification

Purification of DNA from FFPE sections required removal of wax, cell fragments and cross-linked proteins. DNA had to be freed from contaminants collecting as much genetic material as possible. For this study the following protocol was used using the QIAamp DNA FFPE Tissue Kit [190]:

1. Scrape up to 8 sections of $5-10\mu$m tissue from an FFPE block and transfer to 1.5ml microcentrifuge tube
2. Add 1ml xylene to tube and vortex vigorously for 5 seconds for de-paraffinisation
3. Spin at $> 13,000$ rpm for 2 minutes
4. Remove and discard supernatant, leaving the pellet intact
5. Add 1ml of 96% – 100% ethanol on to the pellet to remove xylene residues

6. Vortex vigorously for 5 seconds
7. Spin at > 13,000 rpm for 2 minutes
8. Remove and discard supernatant

Optional. For large amounts of tissue, steps 5.-8. can be repeated to ensure all xylene has been removed from sample

9. Dry pellet by opening lid of the tube and wait for at least 10 minutes until all ethanol has evaporated
10. Resuspend dried pellet in 180μl buffer ATL provided by Qiagen.
11. Add 20μl of Proteinase K to lyse tissue
12. Briefly vortex and incubate for 14-18 hours or overnight at 56 °C on a heat block on a shaker
13. Increase temperature to 90 °C for 1 hour to inactivate Proteinase K
14. Spin briefly until all liquid is collected at the bottom of the tube
15. Add 200μl of Buffer AL provided by Qiagen and vortex briefly
16. Add 200μl of 96% – 100% ethanol straight away and vortex briefly
17. Spin briefly until all liquid is collected at the bottom of the tube
18. Transfer entire lysate to a QIAmp MinElute column in a 2ml collection tube without wetting rim of spin column

19. Close lid and spin at 8,000 rpm for 2 minutes
20. Ensure that all lysate has passed the column. If necessary repeat step 19. with greater speed. Replace collection tube and discard the flow through
21. Open lid of spin column and add 500μl of buffer AW1 provided by Qiagen without wetting the rim
22. Close lid and spin at 8,000 rpm for 2 minutes
23. Ensure that all buffer has passed the column. If necessary repeat step 22. with greater speed. Replace collection tube and discard the flow through
24. Open lid of spin column and add 500μl of buffer AW2 provided by Qiagen without wetting the rim
25. Spin at > 13,000 rpm for 3 minutes to completely dry the membrane of column
26. Place spin column in a clean 1.5ml microcentrifuge tube and discard collection tube with the flow through
27. Open the lid of the column and add 60μl of buffer ATE provided by Qiagen on to the membrane
28. Close lid and incubate at room temperature for 5 minutes
29. Spin at > 16,000 rpm for 1 minute to collect extracted and cleaned DNA

Note: All buffers are used at room temperature, i.e. 15 °C – 25 °C prior use.

2.2.2 Library preparation and Target Enrichment

Target enrichment for both samples was carried out by following the Agilent SureSelect XT Target Enrichment System for Illumina Paired-End Sequencing Library protocol [201] using the custom designed panel version 1. For an economic sample use, only 200ng of input material has been used for each sample. The protocol offered also a preparation guide for 3,000ng, which was not followed as the amount of genetic material per sample was limited. Target enrichment and library preparation steps are performed according to suppliers protocol, the workflow is shown sketched in Figure 2.2, the detailed protocol for library preparation and target enrichment is listed in Section A.3.

From both samples (BRAF20, QUANTREF) three Bioanalyzer 2100 electropherograms were generated to measure fragment size distribution and concentration after shearing, enrichment and pooling, shown in Figure A.5. Libraries created had a fragment size distribution between 250bp-700bp, i.e. the insert size was between 130bp and 580bp, as 120bp were subtracted for adapters and barcode sequences that were ligated prior target capture. From prepared libraries selected regions were captured by hybridising designed probes followed by a minimum

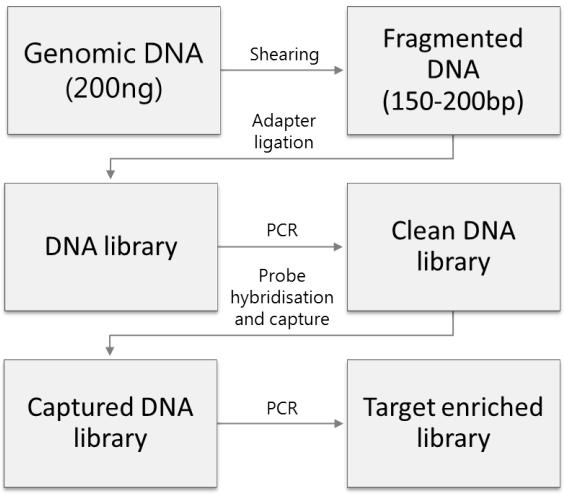


FIGURE 2.2: Workflow of the Agilent SureSelect XT library preparation and target enrichment.

number of PCR cycles; the full protocol is outlined in Section A.4. Bioanalyzer traces of pooled samples showed a size distribution between 250bp and 550bp, so the majority of captured inserts were between 130bp and 430bp in size meaning a minor amount of adapter contamination in reads derived from small fragments was expected as read length was set to 150bp paired-end. In summary, library preparation and target enrichment finished with anticipated results. Both samples were indexed to be sequenced on the same flow-cell, barcodes used are listed in Table 2.2.

Sample ID	Index	Index ID
BRAF20	GCCAAT	A006
QUANTREF	CTTGTA	A012

TABLE 2.2: Indexes for samples BRAF20 and QUANTREF from the SureSelect XT target enrichment and library preparation version 1.

2.2.3 Sequencing and Alignment

Samples BRAF20 and QUANTREF were sequenced on a MiSeq instrument with a paired-end 150bp read length to account for the insert size range of about 150bp-450bp, as shown in Figure A.5. A longer read length would have caused more read pairs to overlap or show adapter contamination, in conjunction with an increased run time and higher sequencing cost negating the benefit of a higher yield. BCL sequencing data was converted and demultiplexed utilising the Illumina® bcl2fastq2 converter [131], commands used are listed in section A.5. Raw data was trimmed for adapter contaminants and low quality bases, parameters set for trimming are specified in Table 2.3. Alignment was performed with BWA mem version 0.7.12 and realigned with ABRA version 0.96 inside of target regions, duplicates were marked with Picardtools version 2.0.1 and data was written to a binary SAM, sorted and indexed, full commands are listed in Section A.6.

Setting	Value
5' adapter (forward)	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
3' adapter (reverse)	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Max. error rate	0.1
Max. InDel rate	0.03
Min. mean Q-Score	25
End-read Q-Score threshold	30
Min. read length	18bp

TABLE 2.3: Trimming parameters used to remove adapter sequence contamination and poor quality bases. Mean quality of bases needs to be at least of a Q-Score of 25, bases are trimmed off the end until threshold is reached. Last base in a read needs to be at least of Q-Score 30 or higher. A read and its pair is deteleted, if read length drops below 18bp.

2.2.4 Results of Panel Version 1

Sequencing data generated for BRAF20 and QUANTREF were trimmed and aligned, a summary of the results is provided in Table 2.4. As shown in Figure 2.8, over 50bp were added for designing hybridisation probes on either side of the target regions increasing the area spanned by probes. This was necessary to increase on-target coverage near the ends of target regions. It caused, however, a decrease in design efficiency. The percentage of bases aligned off-target was between 40% and 52% meaning they did not align to any of the regions that were selected. There were two main factors that caused reads to align outside of target regions. Probes were designed such that at least one base needed to be overlap with a target region. During the target capture and enrichment, probes did not have to fully bind to fragments, meaning they captured a fragment that partially or entirely covered an off-target region, as illustrated in Figures 2.3 - 2.7. Adding probes that partially covered flanking regions was a necessary trade-off to span the whole sequence within target regions at a sufficient high coverage. Overall percentage of bases that aligned on-target, however, decreased. Some reads started and ended much further up- or downstream off the exons, as shown in Figure 2.9, which were not near any flanking regions. These fragments were possibly captured by probes that were hybridising to fragments that were of a similar sequence as target regions, such as pseudogenes. Other regions that were difficult to be captured are highly repetitive

elements or a regions with a problematic GC content. Another potential source of reads aligning off-target were library fragments that were not removed by the capture process, due to incomplete washes. Even though this seemed to be negligible fraction of reads looking at a small section of the genome, it added up to hundreds of thousands of reads per sample.

The duplication rates were within the expected limitations in both versions of the kit. A high number of duplicates in the sequencing data would have indicated a poor library complexity that was "over-sequenced", i.e. substantially more reads would have been sequenced than unique fragments were present in the library. A low-complexity library can be caused by a low adapter ligation efficiency due to a problematic GC content, degenerated DNA or sources of contamination. Another cause for a low-complexity library could have been an insufficient enrichment process, e.g. due to poor probe hybridisation.

Furthermore, 1.33% of the target bases showed no coverage at all. As the number was nearly the same for both samples, it indicated a problem with some of the probes designed. It was unlikely that an increase in sequencing would have changed this number significantly meaning probes needed adjustment to cover more bases and, therefore, increase panel efficiency. Using coverage results from sequencing, the probe cocktail was adjusted to increase coverage in all poorly captured regions, while reducing concentration of probes in regions that performed above

average. Figure 2.10, shows the probe coverage per target gene from the first enrichment. Whilst some regions were well covered, probes targeting other regions performed poorly. The probe concentrations were adjusted to balanced coverage across all regions. In doing so, the new probe mix was used for re-capture and sequencing samples again followed by a direct comparison of both versions.

	Version 1	
Metric	**BRAF20**	**QUANTREF**
Size of target region (in bp)	218,694	218,694
Size of region spanned by Agilent Probes (in bp)	325,696	325,696
Design efficiency (in %)	67.15	67.15
Raw reads	11,196,860	8,209,828
Aligned and pass-filter reads	11,178,292	8,195,378
Duplicates (in %)	22.99	31.47
Insert size	212.55 ± 69.55	220.97 ± 75.38

Mismatch rate (in %)	0.2776	0.23
On-target bases (in %)	47.02	58.50
Off-target bases (in %)	52.98	41.50
Targeted bases covered ≥ 100x (in %)	98.68	98.69
Unique bases on-target (in %)	19.23	20.88
Targeted bases not covered (in %)	1.33	1.33

TABLE 2.4: Sequencing results from pilot sequencing of kit version 1. The metrics were generated with a combination of Picardtools [7] and Samtools [70].

FIGURE 2.3: Illustration of two probes (blue, purple) hybridising to a region of interest (green). The regions directly up- and downstream are *flanking regions* (orange). Regions further apart are off-target regions that are not flanking (grey). Probe 1 (purple) fully overlaps with target region. Probe 2 (blue) overlaps only partially with target region.

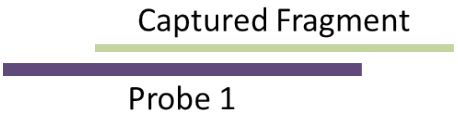


FIGURE 2.4: Probe 1 can hybridise to fragments that fully overlap target region.

FIGURE 2.5: Probe 1 can hybridise to fragments that partially overlap the target region.

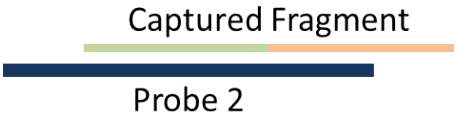


FIGURE 2.6: Probe 2 can hybridise to fragments that partially overlap the target region.

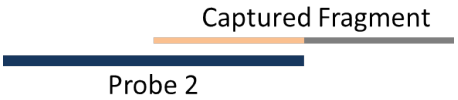


FIGURE 2.7: Probe 2 can hybridise to fragments that do not overlap the targeted region at all.

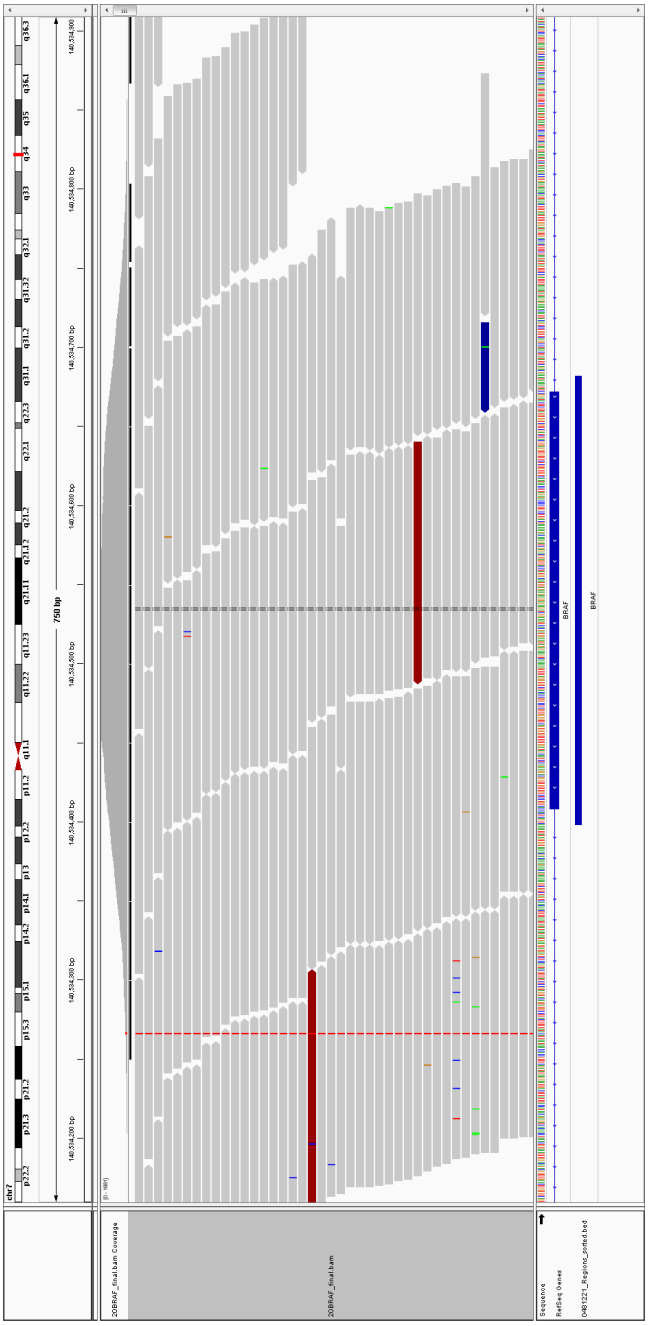


FIGURE 2.8: Probe design for BRAF exon 3. 60bp are added on both ends of the exon to increase region length for probe design. Reads align further downstream due to partial binding of probes to fragments.

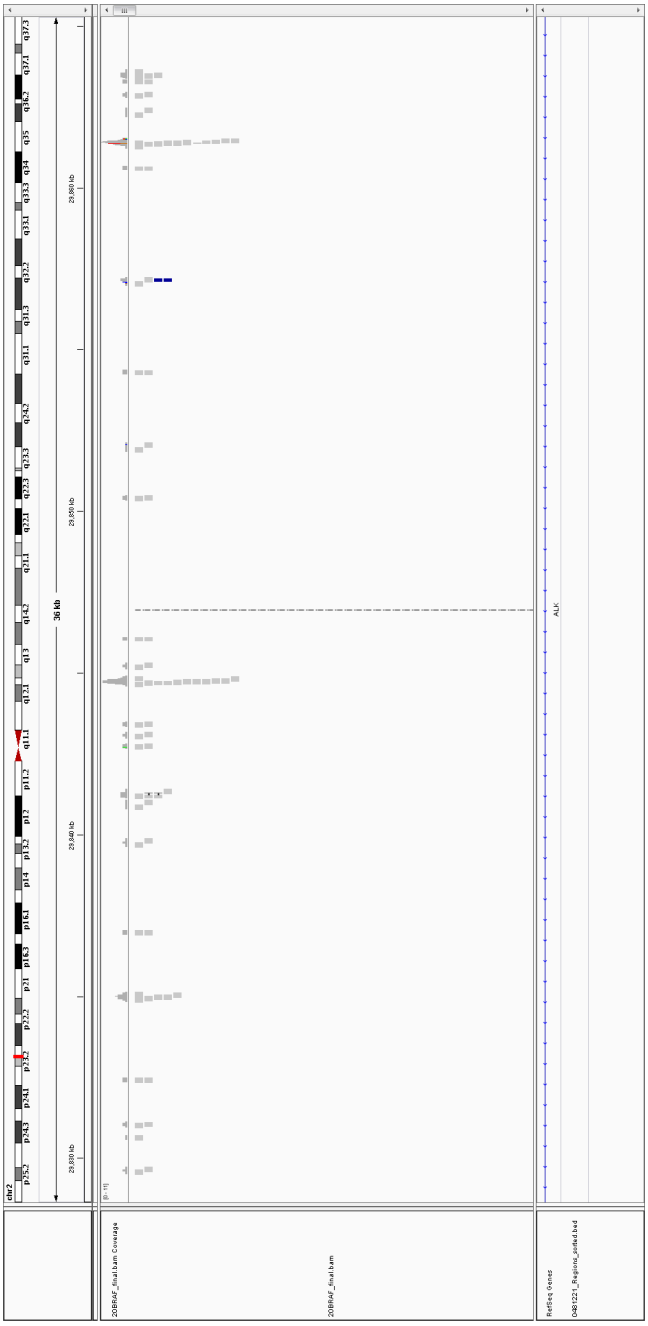


FIGURE 2.9: Example of reads aligning off-target due to loose binding stringency of some probes and incomplete washes during the target capture.

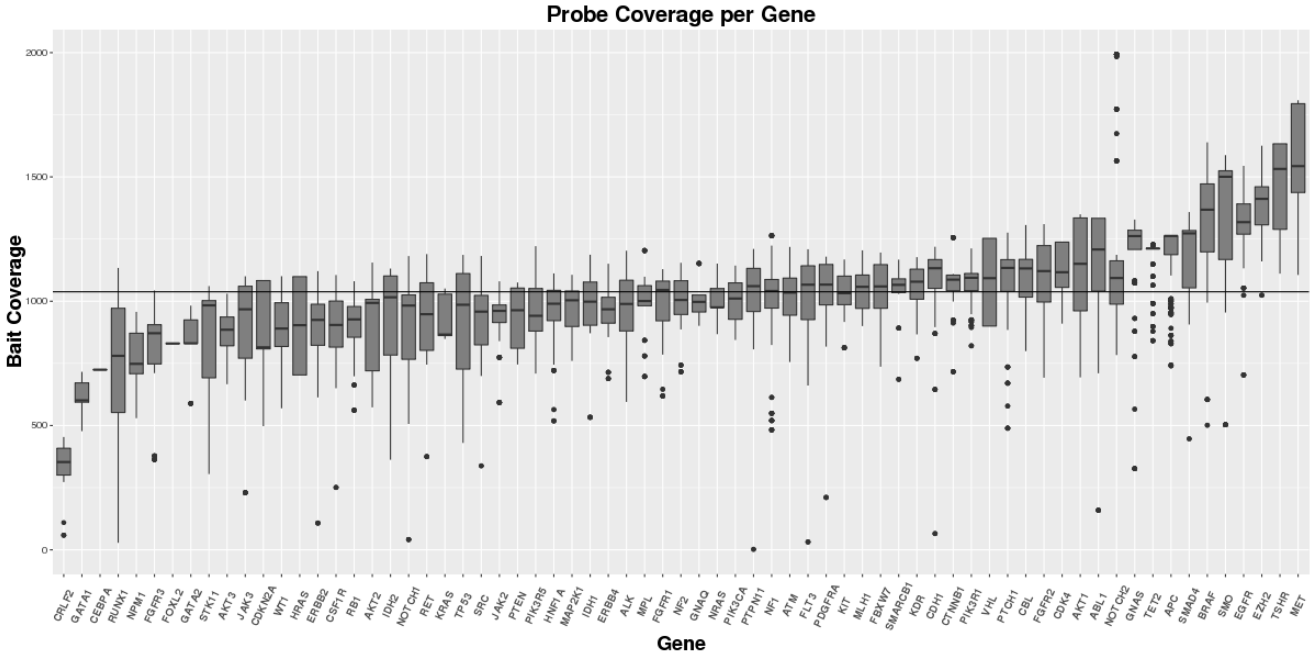


FIGURE 2.10: Probe coverage plotted per gene as a boxplot. Every gene is covered by a number of probes. Reads were tracked from which probe they were derived from. Mean probe coverage across all genes was 1038, indicated as a horizontal line. Probe concentrations below mean coverage were increased in the probe cocktail for a revised version.

2.2.5 Comparison of Probe Coverages

After the probe concentrations were adjusted to balance coverage along target regions, the same two samples were prepared and enriched with the adjusted probe cocktail (version 2). Subsequently they were sequenced again on the MiSeq as a 150bp read paired-end run. Both samples were barcoded again, indexes are listed in Table 2.5. Sequencing and alignment results are provided in Table 2.6 comparing alignment statistics with those obtained from the initial panel (version 1). Since only concentrations of probes were adjusted design efficiency was not changed. Due to the lower number of reads per sample in version 2 duplication rate was little lower, but also the estimated library complexity was higher from samples sequenced with the adjusted probe mix. The reason was a lower binding specificity of some of adjusted probes resulting in a higher number of reads aligning off-target. Enriched fragments from off-target regions added to the overall number of fragments resulting in a higher library complexity. The mismatch rate, defined as number of mismatches among raw reads divided by the number of bases sequenced, was almost twice as high in the second version of the kit, explained by the fact that off-target reads were not removed, hence they increased the rate of mismatches being present.

The fact that more reads aligned off-target was explained by the fact that probes were added partially overlapping with flanking regions, as shown in Figure 2.11. Other probes added were designed from

Sample ID	Index	Index ID
BRAF20	AAACAT	A013
QUANTREF	CTTGTA	A012

TABLE 2.5: Indexes for samples BRAF20 and QUANTREF from the SureSelect XT target enrichment and library preparation version 2.

homologous regions, or from repetitive elements and regions with a problematic GC content, shown in Figure 2.12 and Figure 2.13. It was decided that a higher off-target rate was a fair sacrifice for a panel with a better coverage in some regions. Off target reads were ignored in any further data analysis steps, as they could cause artefacts, due to potentially low coverage and poor alignment quality.

To assess coverage performance of both kit versions, it was important to understand how the probe adjustment influenced the on-target coverage. This was achieved by investigating cumulative coverage of all target regions, shown in Figure 2.14 and Figure 2.15. Regions that had been previously missed were covered causing a drop of uncovered bases from 1.33% in version 1 to 1.24% in version 2, even with less reads sequenced. Another important aspect is that due to a better probe balancing less regions were covered very highly. The minimum amount of sequencing is ruled by the poorest covered base of interest, as it needs enough reads aligned that a subsequent variant calling is possible down to a certain allelic threshold. Coverage beyond a certain factor may improve variant calling in some cases, but there would be less to no benefit, due to given library complexity and the lowest variant allele

frequency to be detected. In summary, the version 2 probe cocktail appeared to cover slightly more bases inside target regions, at the cost of a higher number of reads aligning off-target. From this perspective, the amount of sequencing necessary to make sure all target regions had a sufficient coverage for subsequent data analysis had to be further investigated.

	Version 1		**Version 2**	
Metric	**BRAF20**	**QUANTR.**	**BRAF20**	**QUANTR.**
Size of target region (in bp)	218,694	218,694	218,694	218,694
Size of region spanned by Agilent Probes (in bp)	325,696	325,696	325,696	325,696
Design efficiency (in %)	67.15	67.15	67.15	67.15

Raw reads	11,196,860	8,209,828	8,006,822	6,788,788
Aligned and pass-filter reads	11,178,292	8,195,378	7,925,468	6,722,604
Duplicates (in %)	22.99	31.47	20.36	28.55
Insert size	212.55 ± 69.55	220.97 ± 75.38	283.13 ± 88.07	290.65 ± 89.83
Mismatch rate (in %)	0.2776	0.23	0.62	0.6046
On-target bases (in %)	47.02	58.50	30.78	31.17

Off-target bases (in %)	52.98	41.50	69.22	68.83
Targeted bases covered ≥ 100x (in %)	98.68	98.69	98.85	98.82
Unique bases on-target (in %)	19.23	20.88	11.81	10.48
Targeted bases not covered (in %)	1.33	1.33	1.24	1.24

Table 2.6: Sequencing results from pilot sequencing of both versions.

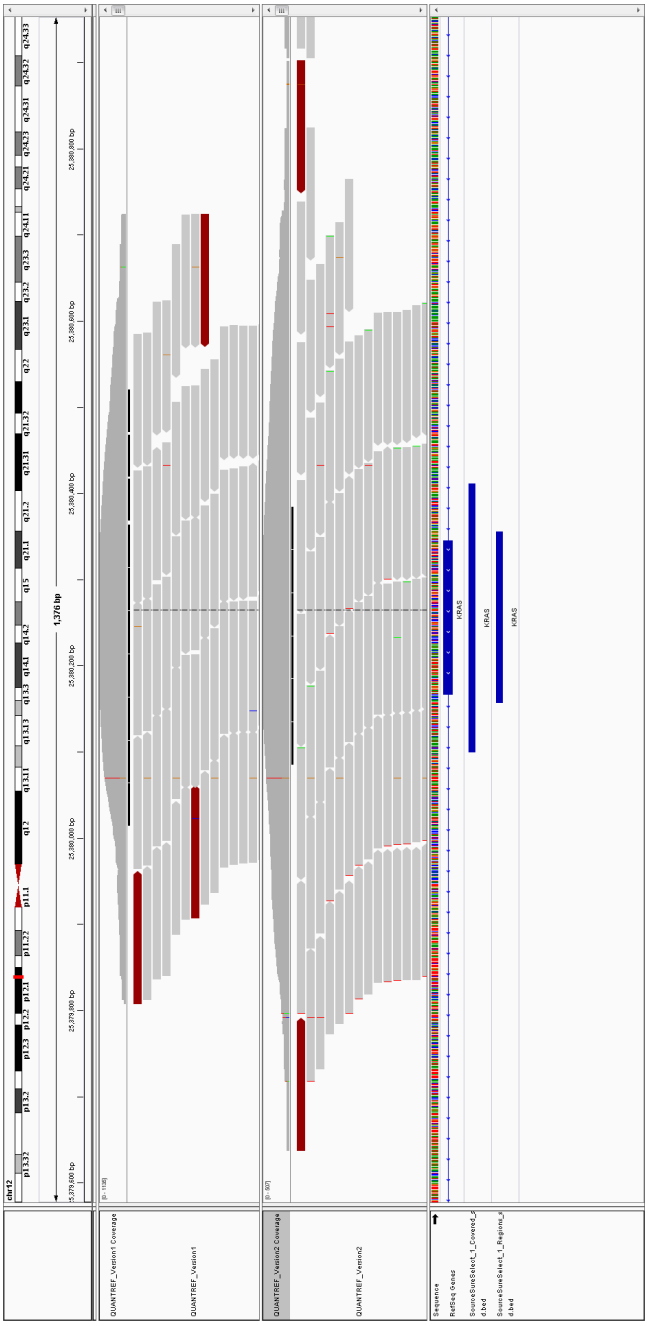


FIGURE 2.11: Regions covered by probes and target regions plotted next to each other with sequencing data for KRAS exon 3. On top of the 10bp added to target regions by the design software, probes were designed over 50bp up- and downstream of the exon. A small amount of reads align even further apart, especially in kit version 2.

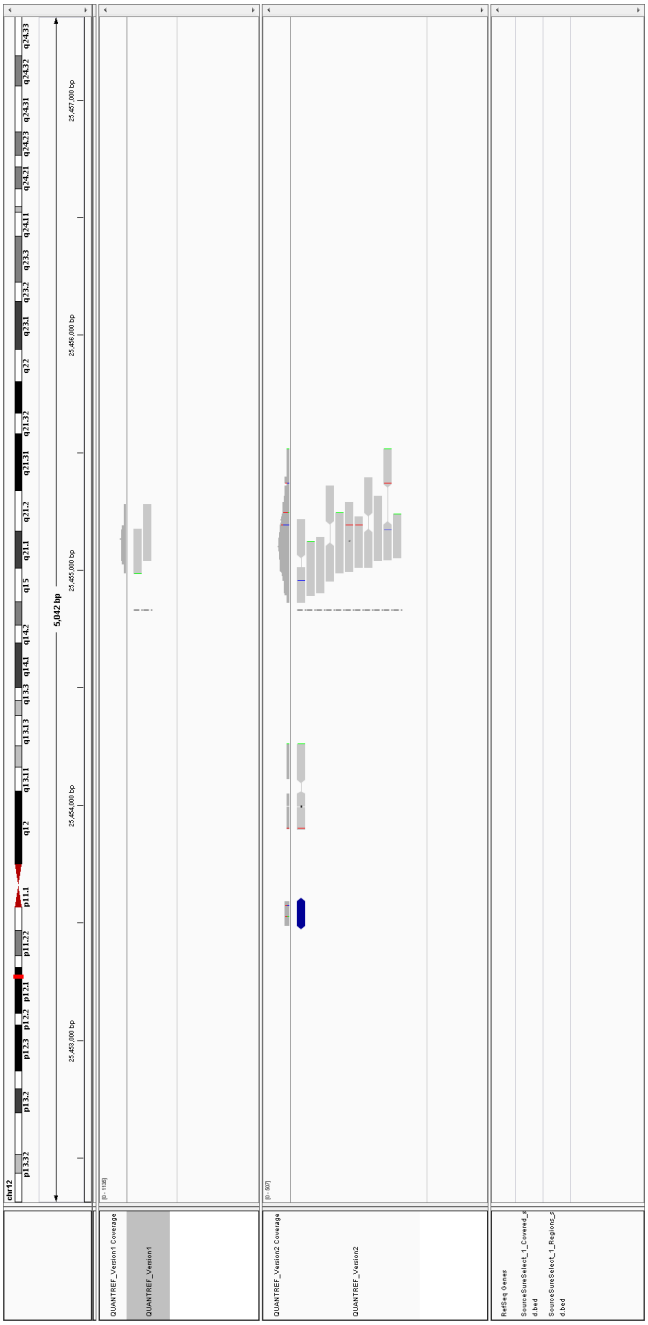


FIGURE 2.12: Example of probes hybridising off-target. Version 2 (bottom) shows an increase in off-target bases covered.

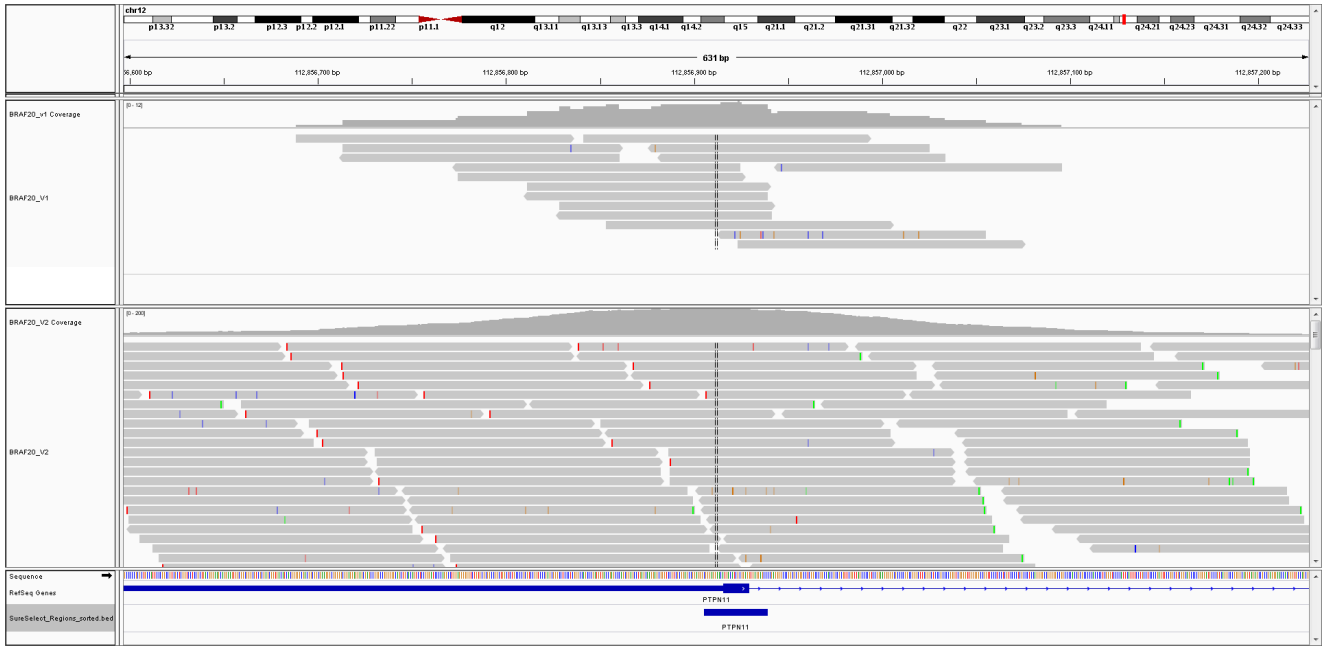


Figure 2.13: Coverage plot and alignment visualisation of the first exon of PTPN11. The first version (top) shows insufficient coverage, version 2 (bottom) provided a much higher coverage across the entire exon. PTPN11 exon 1 shows an increased GC content (65%), which affected binding properties of designed probes.

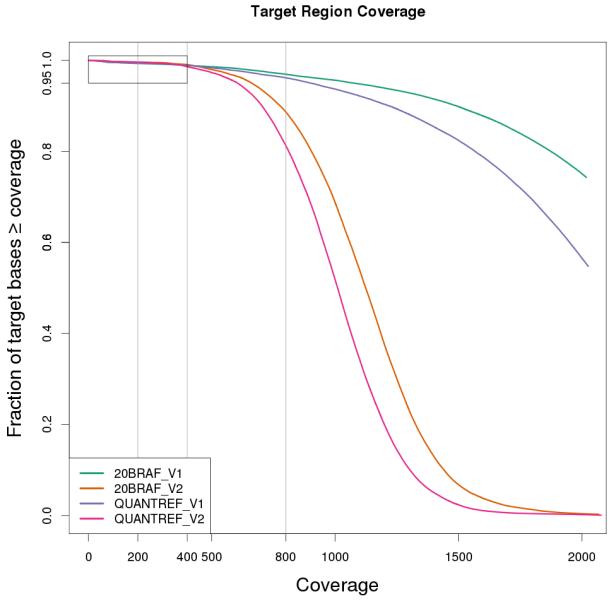


FIGURE 2.14: Cumulative coverage of target regions. In version 1 over 80% of regions sequenced 1500x or higher. In version 2, coverage drops quicker after 800x, indicating a more balanced distribution of coverage.

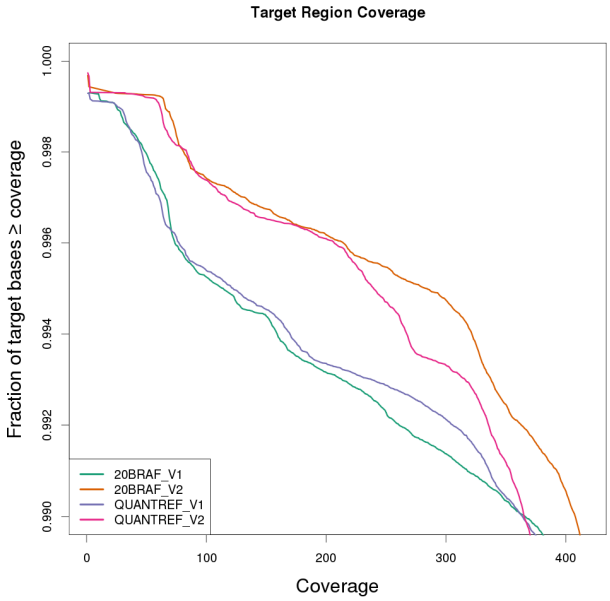


FIGURE 2.15: Cumulative coverage first 400 and top 1% (rectangular in Figure 2.14.

2.2.6 Variant Calling, Filtering and Annotation

After checking kit performance, variant calling was performed on all four samples with Samtools and VarScan2 using parameter settings listed in Table 2.7. The alignment data was translated into a pileup format and variant calling was performed with VarScan2 to call variants based on Fisher's exact test. The resulting variant lists were saved as VCF files and compared, i.e. samples BRAF20 prepared with version 1 and version 2 were compared to each other and sample QUANTREF prepared with version 1 and version 2. An overview of the results is given in Table 2.8.

Parameter	Value
Samtools mpileup	
Recalculation of base alignment quality	False
Maximum depth/coverage	10,000
Adjusting mapping quality	$\sqrt{\frac{50-m}{50}} \cdot 50$, where m is the initial mapping quality
VarScan2	
Minimum variant frequency	1%
Minimum p-value	0.05
Strand-filter	90%

TABLE 2.7: Variant calling parameters used. The full commands are listed in Section A.6. The base alignment quality (BAQ) was not recalculated, as VarScan2 is not expecting BAQ scores. Maximum depth was increased, as Samtools would cap the coverage at 250x. Mapping quality was adjusted by factor 50, to reduce the impact of reads with excessive mismatches.

Variants were filtered with VarScan2 fpfilter module to flag calls that were most likely false-positives. Filtering criteria were based on quality and structural criteria, listed in Table A.3. It was decided to keep potential false positives in the data and flag them rather than removing.

For sample BRAF20 only one variant was validated: a BRAF V600E mutation that was identified in pilot runs. Variant calling results for sample QUANTREF were also consistent for both samples, especially for short InDels. As a first step, reported variants in sample QUANTREF were compared to a list of 11 validated variants provided by Horizon, listed in Table A.1. Triple Venn diagrams summarising the results are shown in Figure 2.18 and Figure 2.19. 9 out of the 11 variants were confirmed, but two mutations could not be detected. They were present in the sample at a low frequency. According to Horizon, the InDel delE746-A750 was present at 2% allele frequency and SNV T790M at 1% allele frequency. The deletion turned out to be present at a slightly lower frequency in the alignment data, i.e. 0.88% in the first version of the kit and 0.2% in version 2. As the variant frequency threshold was set to 1% this variant dropped out in both runs. A low frequency caused also SNV T790M to be missed. The variant was present in the sequencing data obtained from kit version 1, but was not picked up by VarScan2 as it was present at 0.92% allele frequency. Even though base-coverage was only one third in version 2 compared to version 1, SNV T790M was present slightly above the detection threshold, i.e. 1.05% when looking at the alignment data in IGV. In the pileup file, however,

it was revealed that many of the reads supporting the variant position were quality filtered by Samtools. An increase of the allele frequency threshold would have caused VarScan2 to report this mutation. This would increase, however, the number of potential false positives, caused by formalin fixation, amplification, sequencing or alignment artefacts.

It was decided that introducing a minimum frequency threshold of 5% would be suitable for clinical tests, as all present variants and InDels above of a higher frequency were correctly identified. Venn diagrams, shown in Figure 2.16 and Figure 2.17 support this decision, as only mutations below a frequency of 5% caused a disagreement between both versions.

	BRAF20		QUANTREF	
Variants in target regions	**Version 1**	**Version 2**	**Version 1**	**Version 2**
SNVs (total)	196	182	354	338
SNVs filtered	54	40	52	34
Insertions (total)	7	7	9	8
Insertions (filtered)	1	1	0	0
Deletions (total)	24	24	44	45
Deletions (filtered)	4	4	3	4
Validated (all)	1 (100%)	1 (100%)	9 (81.2%)	9 (81.2%)

TABLE 2.8: SNVs and small InDels called by VarScan2. One variant was known in sample BRAF20, while 11 variants were validated in QUANTREF, both by Horizon. Two known variants were not identified.

After initial variant calling and filtering, all pass-filter variants were further annotated by public databases, i.e. Clinvar [60] and dbSNP [109]. Further effects of variants were added, as it provided further filter criteria reducing the number of variants and InDels to focus on. In doing so, variants of low impact could be excluded from the dataset, such as known SNPs, synonymous or intronic variants in flanking regions. Altogether, over 45% of variants were further excluded, in both data sets, as shown in Figure 2.20. Variants that were predicted to be of low impact were safely ignored, as they do not have any effect on the protein by definition.

Theoretically, VarScan2 requires a minimum amount of coverage per base to apply Fisher's exact test and provide reliable results. To estimate robustness of VarScan2 given the described settings, alignment files were downsampled and resulting alignments were used as input for VarScan2 again. In doing so, it was assessed how on-target coverage was affecting the variant calling. Downsampling was performed by randomly choosing a lower number of read-pairs from an alignment file. In Figure 2.21 the number of aligned reads from sample QUANTREF was plotted against the number of variants that were identified by VarScan2, with all variants and only those passing the false-positive filter. It turned out that with version 1 a maximum number variants were reported between 4M and 5M reads sequenced. An increase in the number of reads beyond that did not cause the number of called variants to rise, but rather to drop slightly. VarScan2 was able to distinct

more reliably between low frequency mutations and noise in the data at a higher coverage. For kit version 2 less variants were called from the same number of reads, probably caused by the rebalancing and lower coverage in some regions. An optimal amount of sequencing depends on library complexity meaning it is sample-dependant to certain extend. In this case, between 6M and 8M reads was considered ideal. The coverage requirements for the revised panel were a bit higher, due to the fact that more regions could be covered sufficiently at the cost of a higher off-target rate.

In summary, both kits performed well to make massively parallel sequencing available for genetic testing of multiple prognostic and diagnostic markers in parallel. A sufficient coverage and a removal of off-target reads, PCR duplicates and reads of low quality prior to variant calling were identified to be necessary.

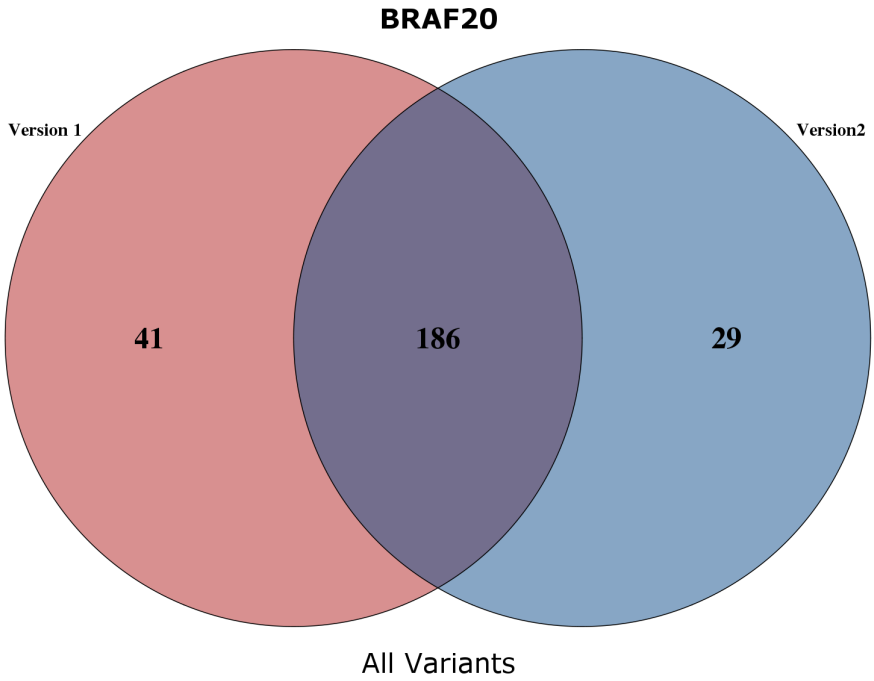


Figure 2.16: Venn diagram variants (unfiltered) of both versions of sample BRAF20.

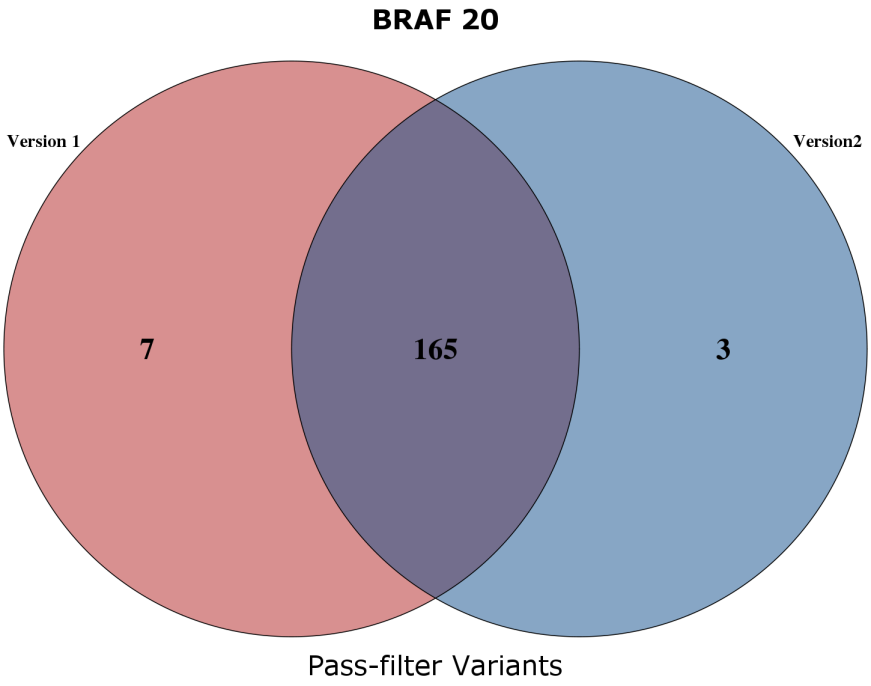


FIGURE 2.17: Venn diagram of BRAF20 from variants (pass-filter).

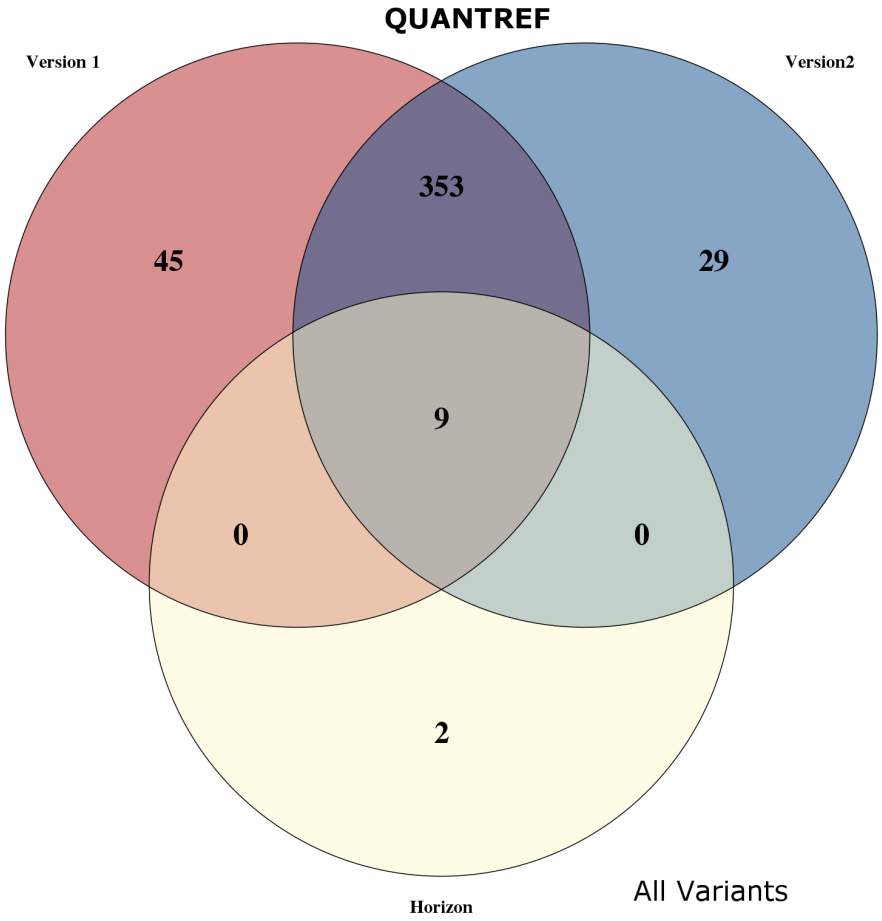


FIGURE 2.18: Triple Venn diagram of variants (unfiltered) in QUANTREF from both versions and Horizon reference.

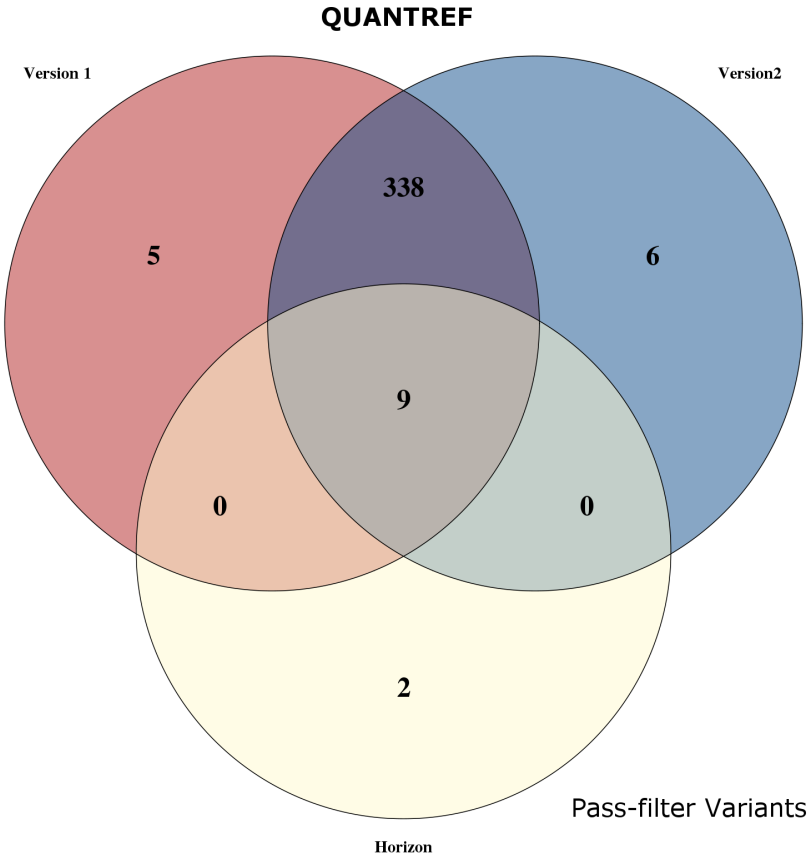


FIGURE 2.19: Triple Venn diagram of variants (pass-filter) in QUANTREF from both versions and Horizon reference.

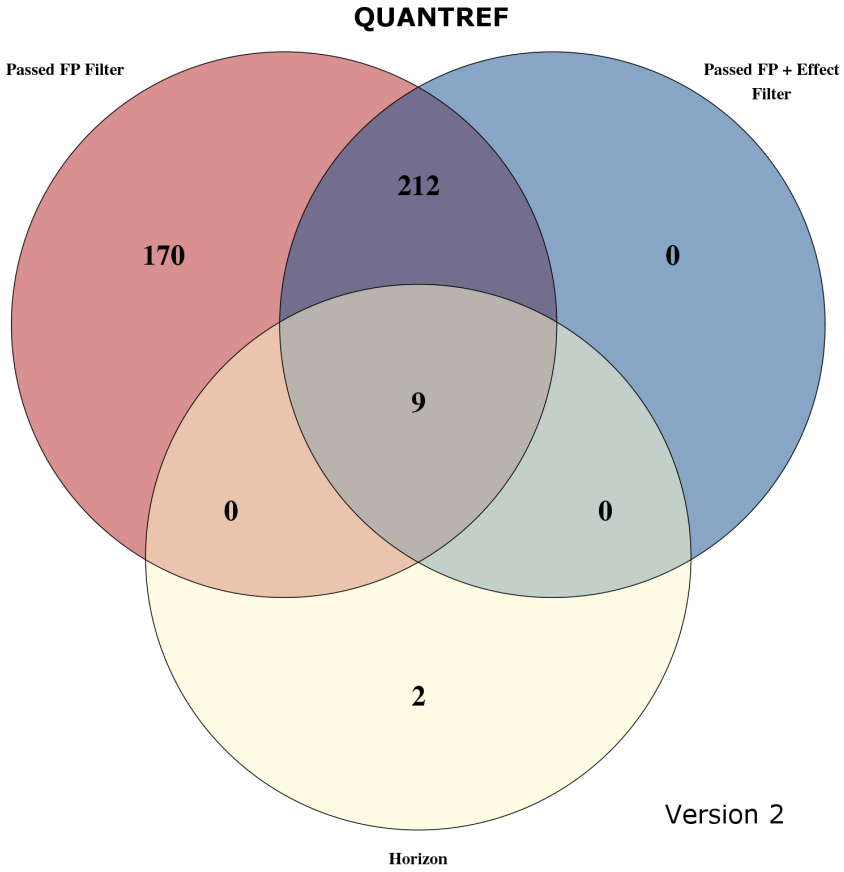


Figure 2.20: Triple Venn diagram of variants called from QUANTREF version 2 sequencing data. Removal of mutations predicted as low impact reduced the amount of variants by over 45%.

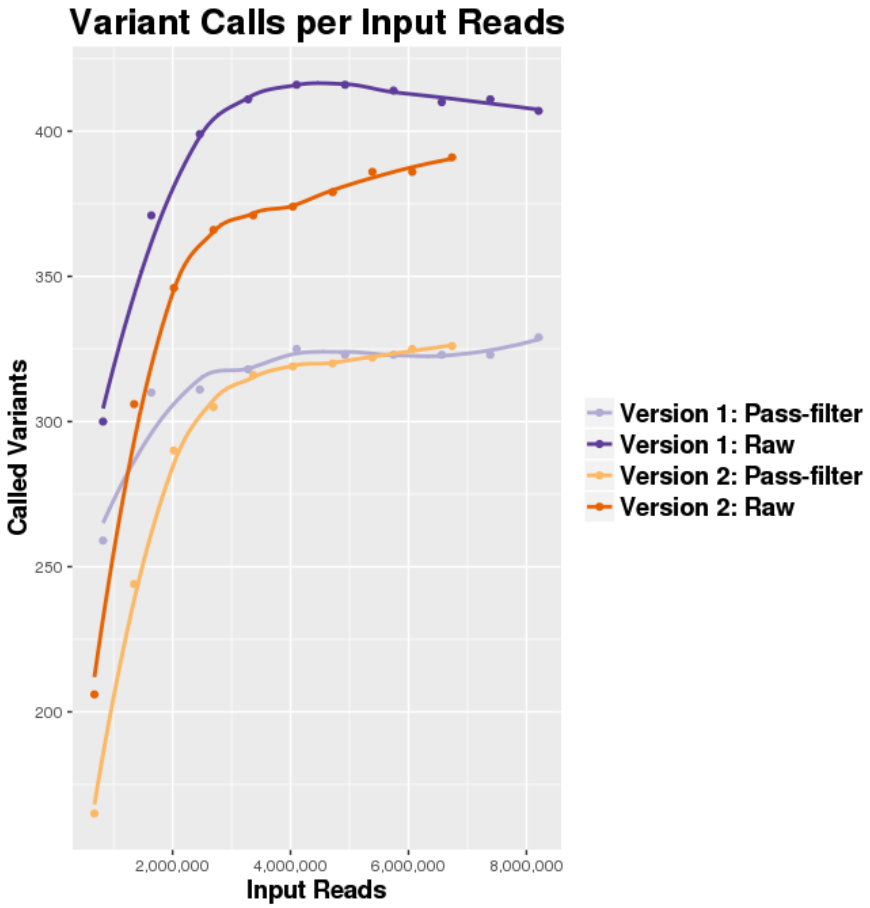


FIGURE 2.21: Variant calling and filtering from QUANTREF alignment downsampled from 10% – 100%.

2.3 Performance Evaluation on 278 Samples

After both versions of the kits were assessed, 278 clinical samples were prepared with kit version 2 and subsequently sequenced, aligned and variants were called. This large-scale experiment was performed to get a deeper understanding kit-performance under real conditions by using clinical FFPE samples.

2.3.1 DNA extraction and Quality Check

DNA was extracted from FFPE samples following the DNA extraction protocol described in Section 2.2.1. DNA was quantified using the Qubit instrument, DNA integrity of input material was measured with an Agilent TapeStation 2200. Enriched libraries were quantified with a high-sensitivity chip and a Agilent Bioanalyzer 2100 instrument. Concentrations and DNA integrity scores are listed in Table B.1. Extracted DNA from samples were prepared, enriched, pooled and sequenced, even if concentrations or integrity scores failed to meet input requirements specified by the manufacturer. They were, however, flagged as low-input or low/failed DIN samples throughout the experiment. Concentrations and DNA integrities of extracted DNA showed a broad range, based on numerous factors such as the amount of tissue used for extraction, embedding protocol and storage conditions. Results were compared by institute and tissue type that was received, shown

in Figure 2.22 and Figure 2.23. Extraction of higher quantities of DNA was more difficult for some cancer types than for others, due to the amount of tissue that was present. Tissue samples that were lung or breast cancers were received from fine needle aspirations, as it less invasive and generally safer than surgery, but resulting in less tissue embedded from which DNA can be extracted. Boxplots from measured DIN scores by tissue type are shown in Figure 2.24 and Figure 2.25. The amount of genetic material that was collect by the extraction kit was very consistent for any of the three tissue providers. The DNA integrity scores were considerably lower from Birmingham Queen Elizabeth Hospital (BI), showing an average DIN score of around 2, while samples provided by United Kingdom National External Quality Assessment Service (UK NEQAS) showed an assigned DNA integrity on average of 3.2 and extracted DNA from Source BioScience had an average DIN score of 3.9. The interval of DIN scores defined by Agilent was difficult to handle in practice sometimes, as samples performed poorly got assigned the value "NA", which meant they were excluded by any data analysis software. As a consequence, the average quality was overestimated. In order to penalise poorly performing samples, the DIN was manually adjusted to 0 for poorly performing samples and only keep the value "NA" for samples not measured. This step is not ideal, as it causes a definition gap between 0 and 1, which made it impossible to interpret a DIN of 0 any further. Hence, this approach was only used for the visualisation purposes. For all other analysis steps, the value

was set back to "NA" again. To see if DIN score and the input material showed any trend both were visualised in a pairwise scatter plot, as shown in Figure 2.26. To see a mathematical correlation between two datasets $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_n\}$, each of size n, the Pearson correlation coefficient r was defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.1}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and analogously $\bar{y}$ are defined as the mean of set x and y respectively. The Pearson correlation coefficient is defined on the interval $[-1, 1]$, where 0 means no correlation, 1 a perfect positive correlation and -1 a perfect negative correlation.

It could be excluded that there was a strong relationship between concentration and DIN score of extracted DNA, as the calculated Pearson correlation was 0.056 indicating that integrity does not strongly rely on the provided concentration. From this it was concluded that measuring DNA integrity really provided additional information, although it was not possible to deeply understand how the DIN score was exactly determined, as it was calculated by a proprietary algorithm.

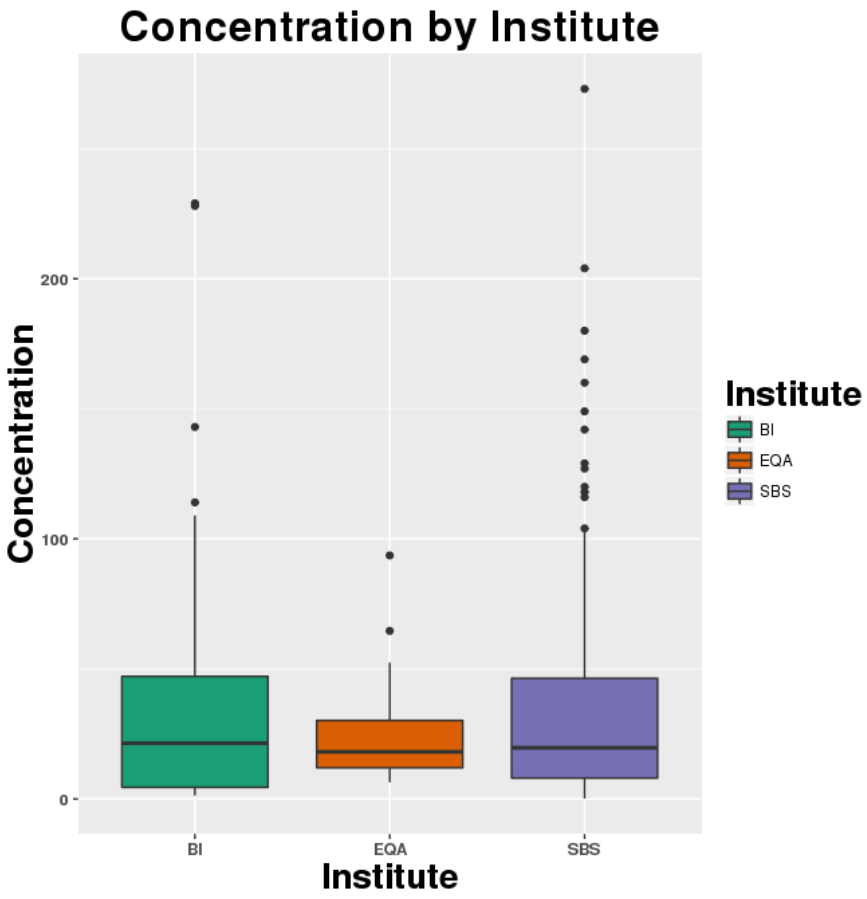


Figure 2.22: Boxplot of extracted DNA concentrations from FFPE tissue plotted by institute.

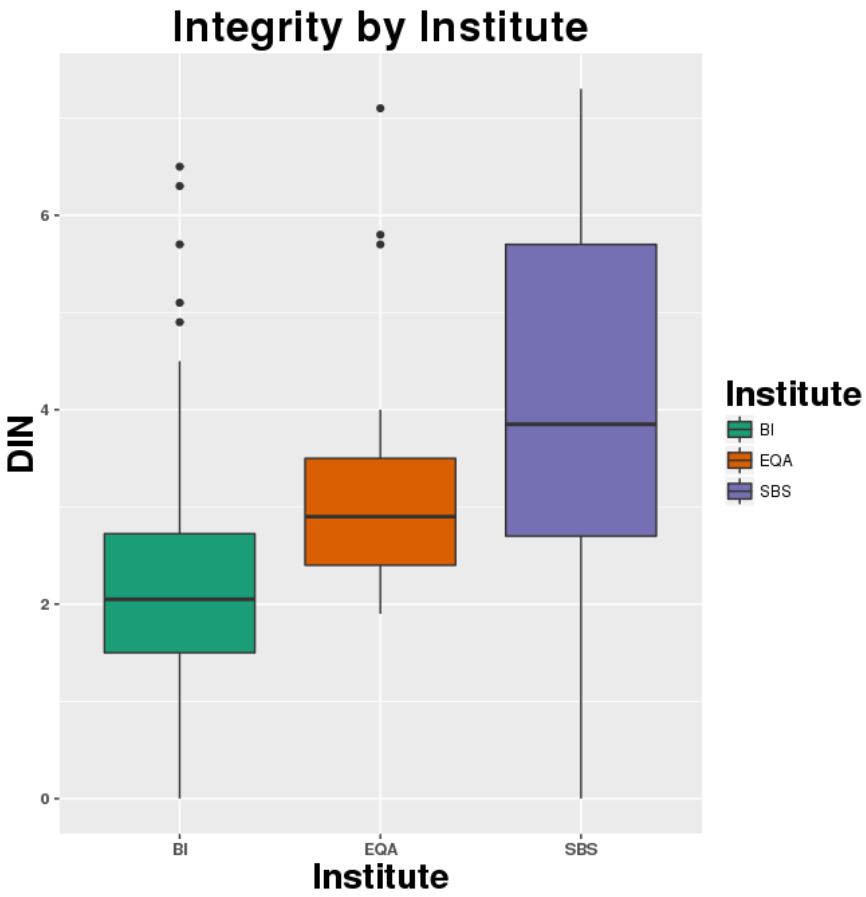


FIGURE 2.23: Boxplot of DIN scores of DNA from extracted FFPE tissue plotted by institute.

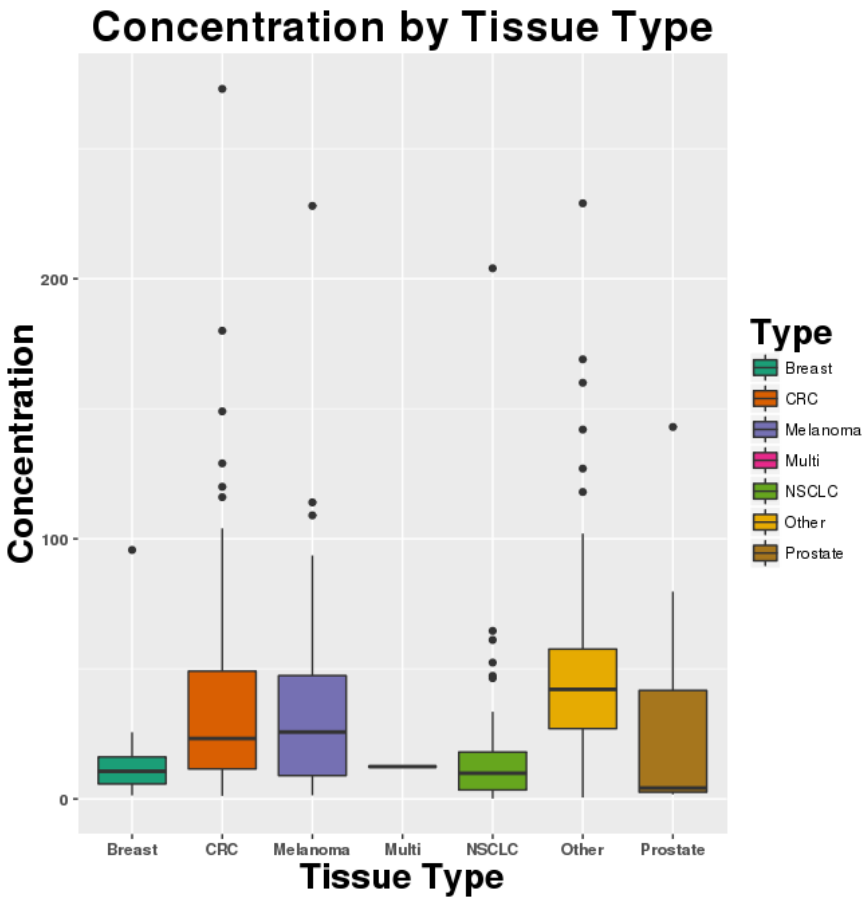


FIGURE 2.24: Boxplot of concentrations of DNA from extracted FFPE samples. Plotted by the tissue DNA was extracted from.

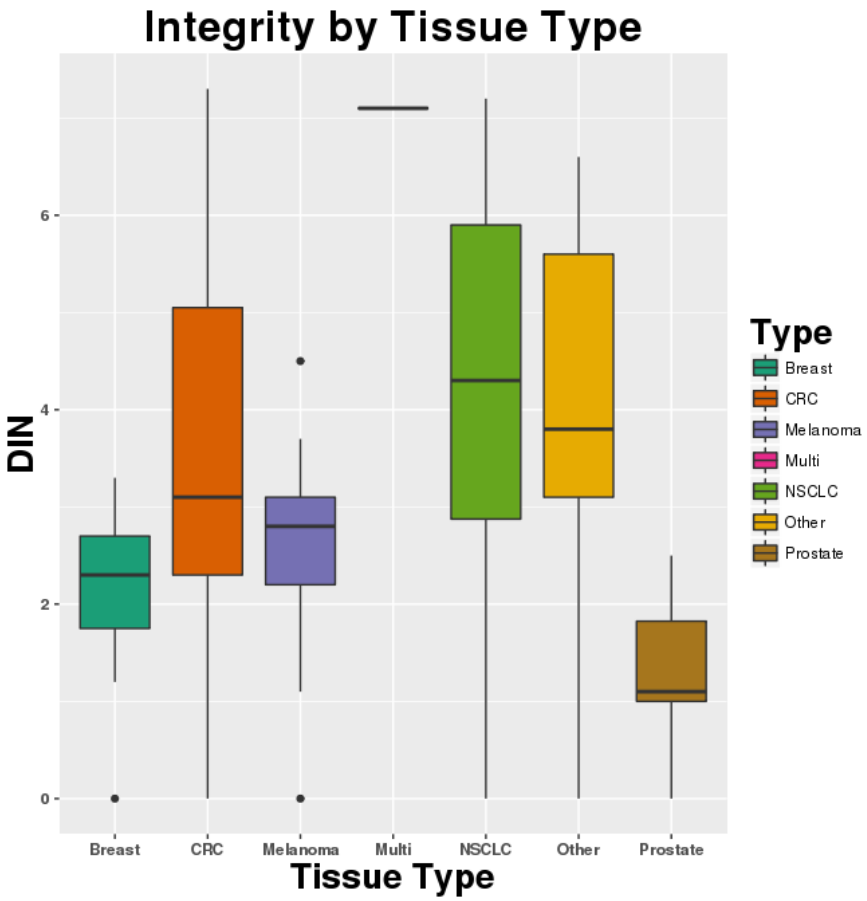


FIGURE 2.25: Boxplot of DIN scores of extracted DNA from FFPE samples. Plotted by the tissue DNA was extracted from.

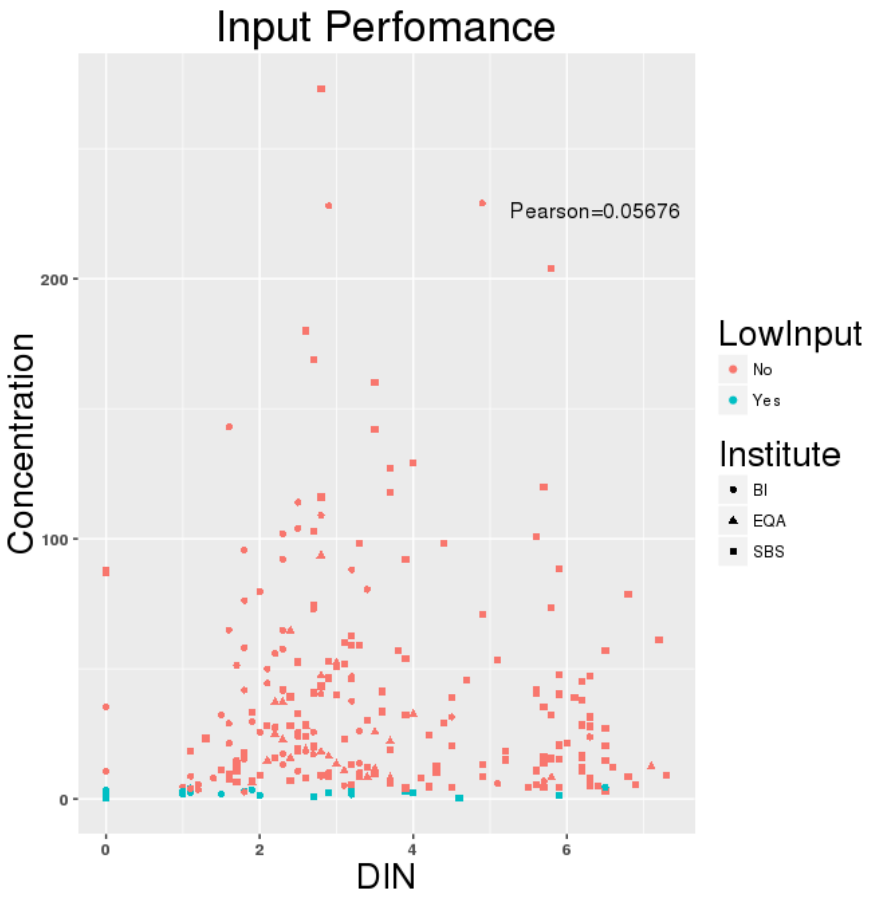


FIGURE 2.26: Scatter plot of DIN scores against concentrations of extracted material.

2.3.2 Sequencing and Alignment

After an initial quality check of the extracted DNA and the generated libraries, all 278 prepared clinical samples were merged into six separate pools, two of which were a mix of 43 indexed samples, while the other four were made from 48 samples each. Every pool was sequenced separately on a NextSeq 500 instrument. Generated BCL files were translated and demultiplexed to FASTQ files. The raw data was adapter and quality trimmed and aligned. Subsequently duplicates were flagged, and remaining reads were re-aligned within target regions. Reads that were not removed by Skewer are plotted by pool, shown in Figure 2.28. Five out of four sequencing runs were within the expected range of sequencing yield, but unfortunately the clustering failed on pool 4. The yield was below 10% of the average, reaching not nearly the necessary amount of sequencing required for subsequent analysis.

The raw data was not further quality checked with any software suite, such as FastQC [136], which collects standard quality and statistical metrics, to assess raw sequencing data. The main reason was the lack of universal validity in the sense that a lot of calculated metrics reported a problem with a same. Samples tested with FastQC, the software reported a problem with the GC distribution present in each read as FastQC was expecting an average of 45%, as shown in Figure 2.27. The true GC content from the target regions, however, was much higher, due

to the fact that exons generally show a higher GC content [173]. A sample with a high number of off-target reads, therefore, could potentially pass FastQC filter criteria, although this sample would be considered poorly performing from a different perspective. In addition, calculating a number of quality metrics on each sample would have been time consuming while strict pass or fail criteria prior alignment were generally hard to define including all eventualities. Moreover, there would be no countermeasures known in case a sample would have failed a quality check. FASTQ files were pre-processed by trimming anyway and other methods to compensate for an unexpected bias were neither known, nor would they be safe, as their effects on the subsequent data analysis could not be predicted. Hence, it was decided to check sequencing, trimming and alignment reports and to evaluate a sample from the aligned data instead without additional quality control of the raw data.

The next step was to check the distribution of sequencing yield per sample. From the pilot it was presumed that the optimal number of reads was about 6 million to 8 million reads excluding duplicates. If excluding pool 4, about 90% of the samples showed more than 6 million reads, as shown in Figure 2.29. Failed samples were still further analysed to observe their variant calling potential. It remained to see what factors influence on-target coverage and library complexity.

Table 2.9 lists a number of metrics that were collected from Picard-tools. Although a wide range of metrics were available provided by

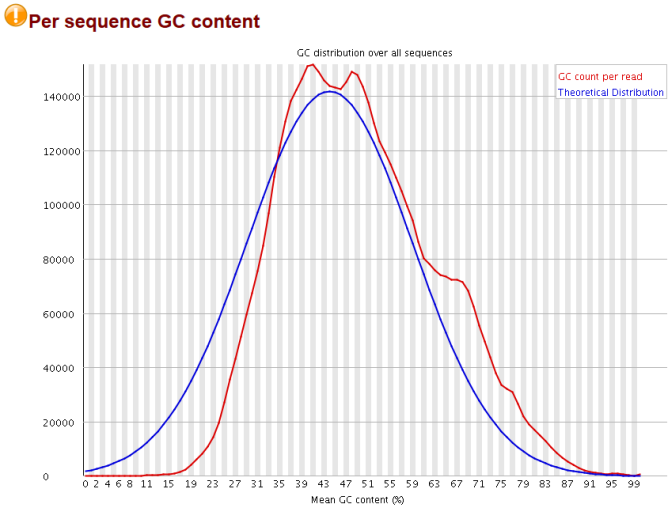


FIGURE 2.27: Reported GC content per sequence in sample 14R011759. FastQC expects a theoretical distribution. The enriched regions, however, follow a different distribution causing FastQC to print a warning.

Picardtools and Samtools only a handful of metrics were picked, as they seemed most helpful to determine how much sequencing was necessary to reach optimal on-target coverage, given the broad variety of quality and quantity of extracted DNA received from different institutes. The coverage per base was defined to be optimal exactly when VarScan2 could reliably distinct between a mutation with a frequency of 5% or higher and noise. When coverage was sufficient, no more variants were reported, even if coverage was further increased, seen as a plateau effect that was previously observed in the pilot experiment. Hence, metrics collecting information about the on-target coverage were selected. It

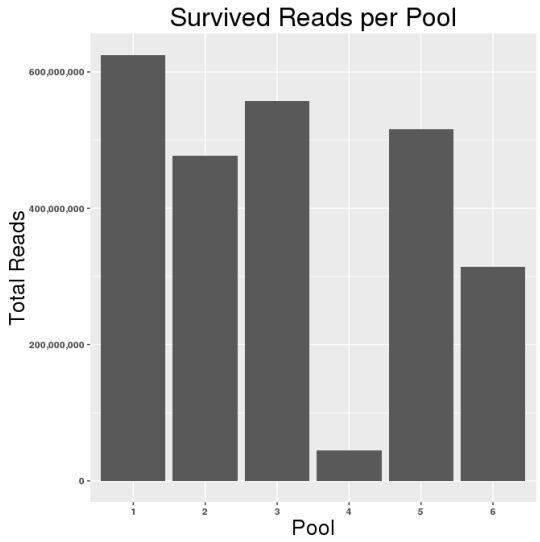


FIGURE 2.28: Histogram of all reads survived split by pool. Pool 4 underperformed significantly, leaving only about 10% of reads available than expected.

was desired to find metrics that correlate with any of the initial quality measures, such as quantity or integrity. A correlation to input quality or quantity measures would allow to define strict input criteria prior shearing reducing the overall amount of samples failed to be sufficiently sequenced. Table 2.10 lists a matrix of pairwise Pearson correlations from various metrics and input measures to detect possible dependencies between certain criteria. The first observation made was that the DNA integrity correlated to a fair extend with mean on-target coverage per sample, as shown in Figure 2.30. Moreover, no correlation between the total number of trimmed reads or the duplication rate was found, as shown in Figure 2.31 and Figure 2.32.

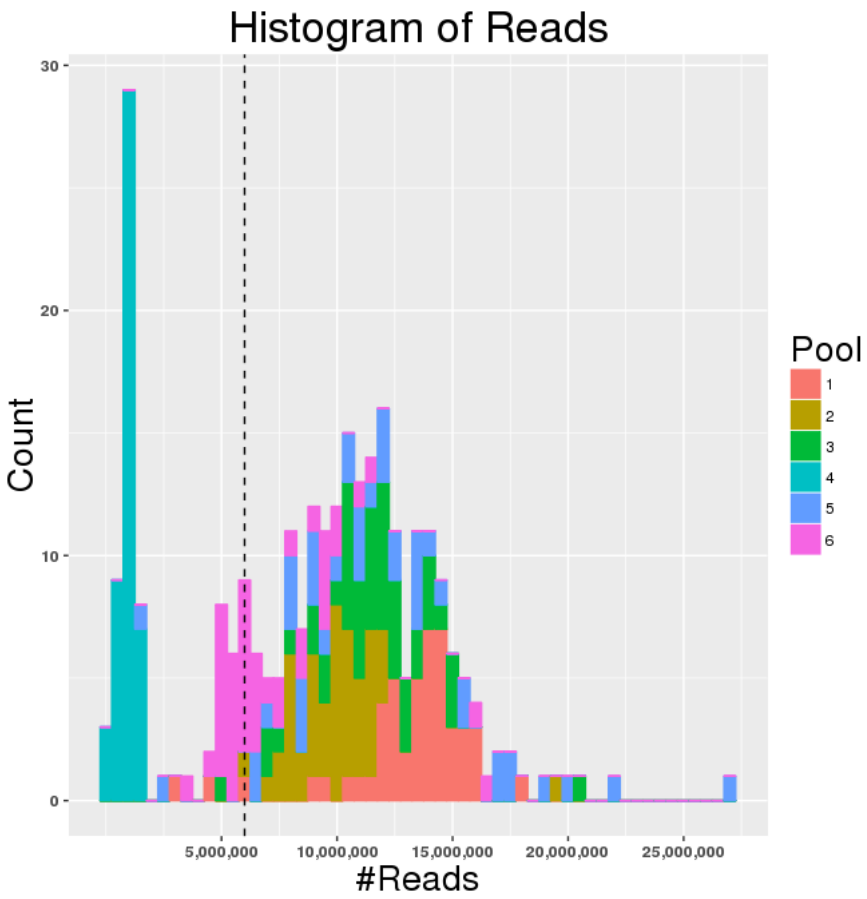


FIGURE 2.29: Histogram of reads survived trimming. The dashed line indicates a threshold of 6 million reads. If pool 4 is excluded, over 90% of the samples had contained 6 million reads sequenced or more.

Metric	**Description**
Duplication rate	Rate of reads marked as duplicates (Dupl.)
Bases on-target	Number of aligned bases on-target (Ot)
Rate on-target covered ≥ 100X	Rate of bases in target region that are covered at least 100X (≥ 100X)
Rate of bases on-target and useful	De-duplicated, on-target bases aligned (Useful)
Mean on-target coverage	The average coverage of bases on-target (Mot)

Table 2.9: A selection of metrics used to evaluate kit performance

	DIN	Conc.	Raw	Dupl.	Ot	≥ 100X	Useful	Mot
DIN		-0.012	0.030	0.009	0.473	0.278	0.461	0.474
Conc.	-0.012		0.079	0.071	0.075	0.175	-0.009	0.074
Raw	0.031	0.079		0.340	0.530	0.552	-0.387	0.527
Dupl.	0.009	0.071	0.340		-0.084	-0.001	-0.575	-0.092
Ot	0.474	0.075	0.530	-0.0839		0.559	0.408	0.995
≥ 100X	0.278	0.175	0.552	-0.002	0.559		0.085	0.574
Useful	0.461	-0.01	-0.387	-0.575	0.408	0.085		0.408
Mot	0.474	0.075	0.526	-0.092	0.995	0.575	0.408	

TABLE 2.10: Matrix of pairwise Pearson correlation coefficients rounded to three digits. Table headings are described in Table 2.9. Pearson correlation ≥ 0.25 are highlighted in green, as they indicate a certain degree of dependency, larger than noise. Correlation between trimmed reads and useful bases on-target is caused by samples from pool 4, hence it was ignored.

By demonstrating that the DINs correlated noticeably with the on-target coverage, but not with duplication rate or the number of trimmed reads, DIN score could be utilised as an indicator of how much sequencing is needed for a sample to reach sufficient on-target coverage. In cases where library complexity was high enough a sample could have been sequenced more to compensate for a higher off-target rate. It may prevent that samples fail to meet the minimum criteria needed for variant calling for the future. In cases of low complexity, however, this method would not result in any improvement.

A possible reason for this relationship could be based on the fact that intact DNA takes longer to hybridise against the 120bp probes. Strongly degraded DNA fragments can hybridise much quicker and therefore occupy a probe increasing the number of off-target reads. By sequencing more the absolute number of fragments sequenced would be higher meaning the number of fragments enriched from a target region would be higher as well. Hence, more sequencing could compensate for partially degraded DNA to a certain extend.

2.3.3 Variant Calling

Sequencing data was used for variant calling and filtering to assess sensitivity of the panel given a minimum mutation frequency of 5%. The hypothesis-free approach of massively parallel sequencing made it difficult to determine performance at every single base, i.e. the designed

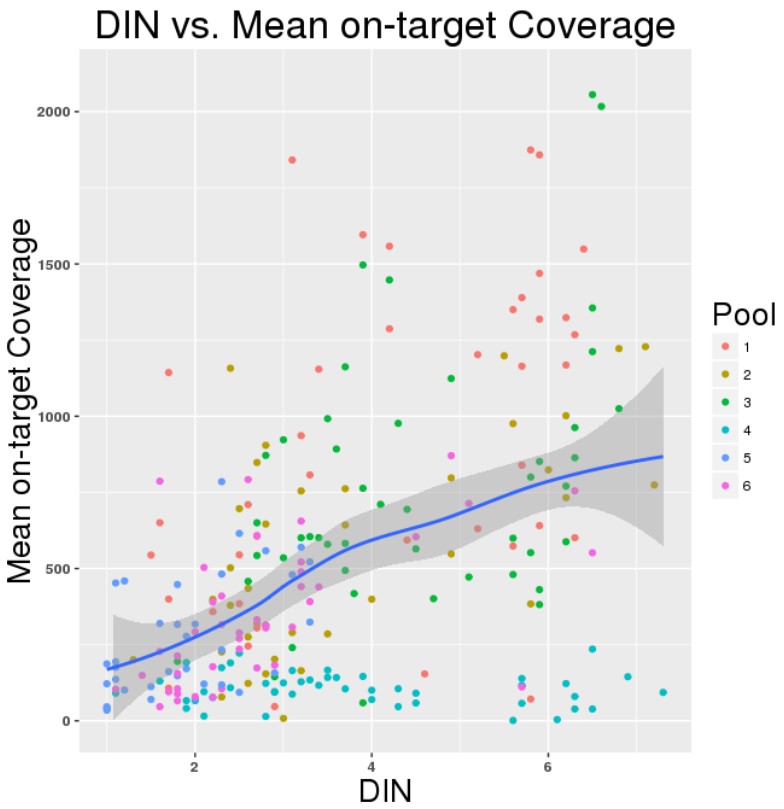


FIGURE 2.30: DIN score plotted against the mean on-target coverage. If pool 4 was excluded from the data, Pearson correlation coefficient raised to 0.626 (not shown).

cancer panel tests covers 219kbp. Hence, only a small subset of these loci were practically tested with by another assay. In order to get a performance estimate, extracted DNA from samples were tested with pyrosequencing for known mutation hotspots. Depending on their primary tumour type these sites differed: while colorectal cancer was tested for mutations in KRAS codons 12,13,61; NRAS codons 12,13,

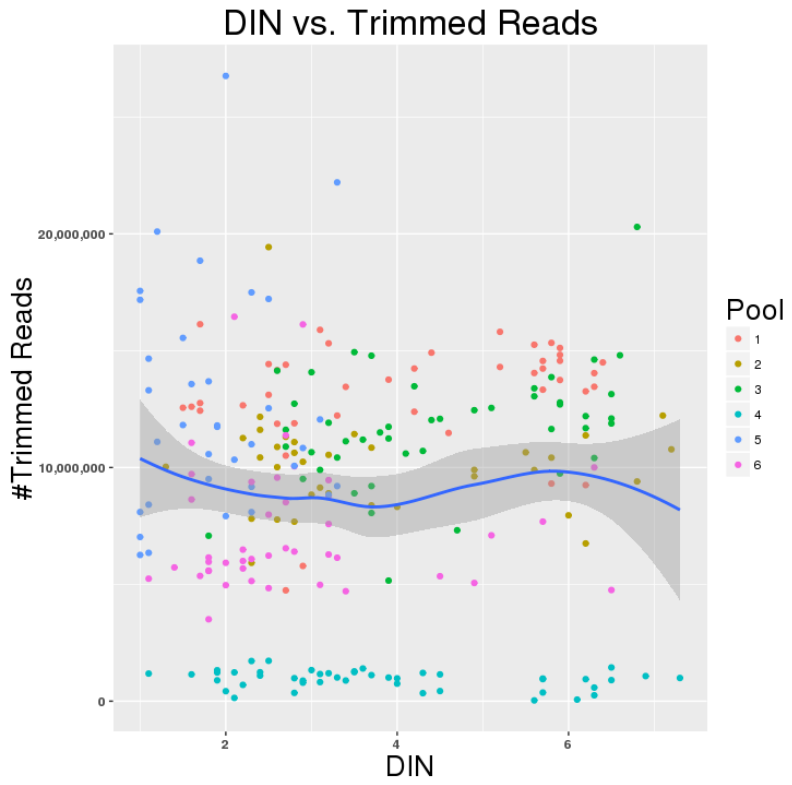


FIGURE 2.31: Scatter plot of DIN score against number of trimmed reads.

61 and BRAF codon 600, non-small cell lung cancer was tested for EGFR codons 719,858-861 and deletions in exon 19. Melanoma was screened for BRAF codon 600 only, while other samples were not tested. It was desired to understand how reliable mutations could be called based on obtained sequencing data on the selected loci and how much coverage would be optimal. Further the data was searched for any type of artefacts that had an impact on the variant calling and filtering.

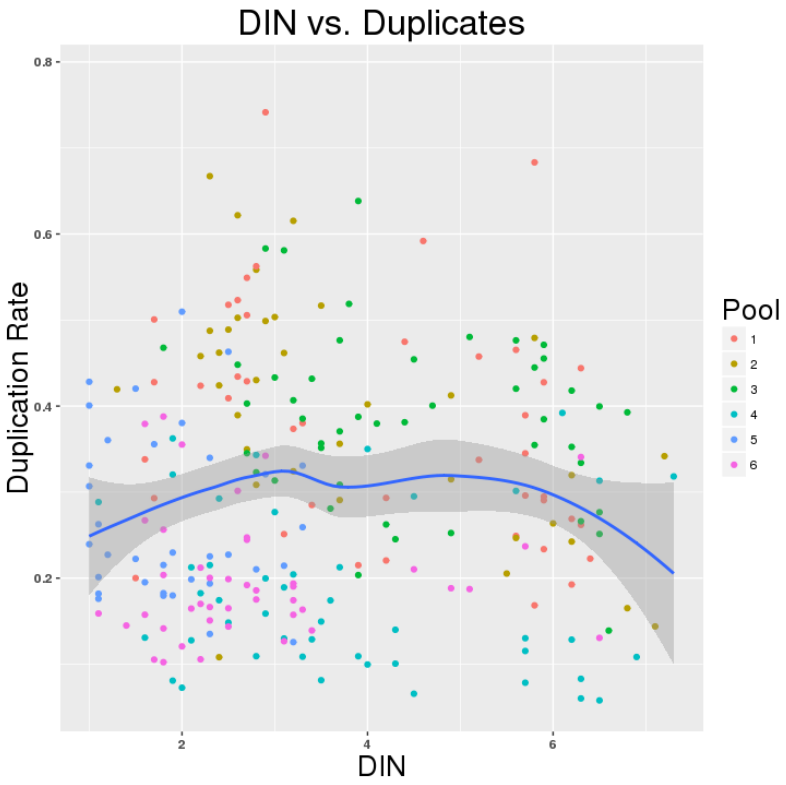


FIGURE 2.32: Scatter plot of DIN score against duplication rate

Variants were called, annotated and filtered for all samples excluding pool 4. SNVs and small InDels in genes KRAS, NRAS, BRAF and EGFR were reported. Known SNPs or synonymous variants were ignored. The results were compared to the calls from pyrosequencing, which are listed in Table 2.11. In eight cases a mutation was reported, which could not be confirmed with pyrosequencing, hence the mutation results *clashed*. All clashing sites had a mutation frequency of less than 10% reported by VarScan2. Estimated tumour burden for these samples were

between 5% and 25%, which was significantly lower than recommended. Clinical testing with pyrosequencing usually comes with a caveat for tissue samples with a tumour burden of below 70% meaning that there is an increased chance that a mutation is missed if tumour burden drops below that. Massively parallel sequencing with target capture and enrichment may work better in cases of lower tumour burden than current pyrosequencing assays used. In two cases mutations were missed by VarScan2 that were reported by pyrosequencing.

Site	Mutated sites	Confirmed (tested)	Unconfirmed (not tested)	Clash	Missed
KRAS codon 12	48	35	12	1	1
KRAS codon 13	11	10	1	0	0
KRAS codon 61	8	3	2	3	0
BRAF codon 600	14	10	4	0	0
NRAS codon 12	0	0	0	0	0
NRAS codon 13	0	0	0	0	0
NRAS codon 61	7	4	4	0	0
EGFR codon 719	3	1	1	1	1
EGFR deletions	3	1	0	2	0
EGFR codons 858-861	3	1	1	1	0

Table 2.11: A list of confirmed non-synonymous mutations identified in genes KRAS, NRAS, BRAF and EGFR. Data from pool 4 was not included. First column contains the total number SNVs/InDels, the second how many could be confirmed. Column three shows the number of samples not tested with pyrosequencing. Fourth column shows how many results clashed, i.e. pyrosequencing could not confirm a mutation. The last column shows the number of samples with missed mutations.

One sample (14R012024) had a relatively low sequencing yield of less than 6 million reads and a DIN score of 3.9. The duplication rate was reported at 64% indicating a low library complexity. Low DNA integrity caused a relatively high number of off-target reads leaving only 1.77% of usable reads on-target, i.e. 5.4% of target bases were sufficiently covered. As shown in the alignment visualisation in Figure 2.33, the base of interest was covered at 40X leaving it insufficiently covered for variant calling, as the mutation was reported to be present at a low frequency as well and had to be repeated with pyrosequencing once due to obscure results, because the estimated tumour burden was reported at 10%. As expected, a poor on-target coverage and a high duplication rate caused the test to fail, especially in cases of low mutation frequencies. The case of the other sample (14R011773) was different. It had a high DIN score of 5.8 and over 15 million reads sequenced, while only 17% were duplicates. Nonetheless, VarScan2 reported a false-negative result. Mutation frequency and tumour burden were both above 50% meaning the mutation should be present in a sufficient number of reads. By looking at the alignment, shown in Figure 2.34, the base shows a high coverage, but the mutation was almost not present in the reads. This was confirmed by looking at the pileup data for that base. The mutated allele was present only twice in over 2000 reads, which is within the noise threshold. It seemed that the reference allele was preferred in the capture or sequencing process causing a false-negative result.

Figure 2.33: Sample 14R012024 alignment around locus chr12:25398284.

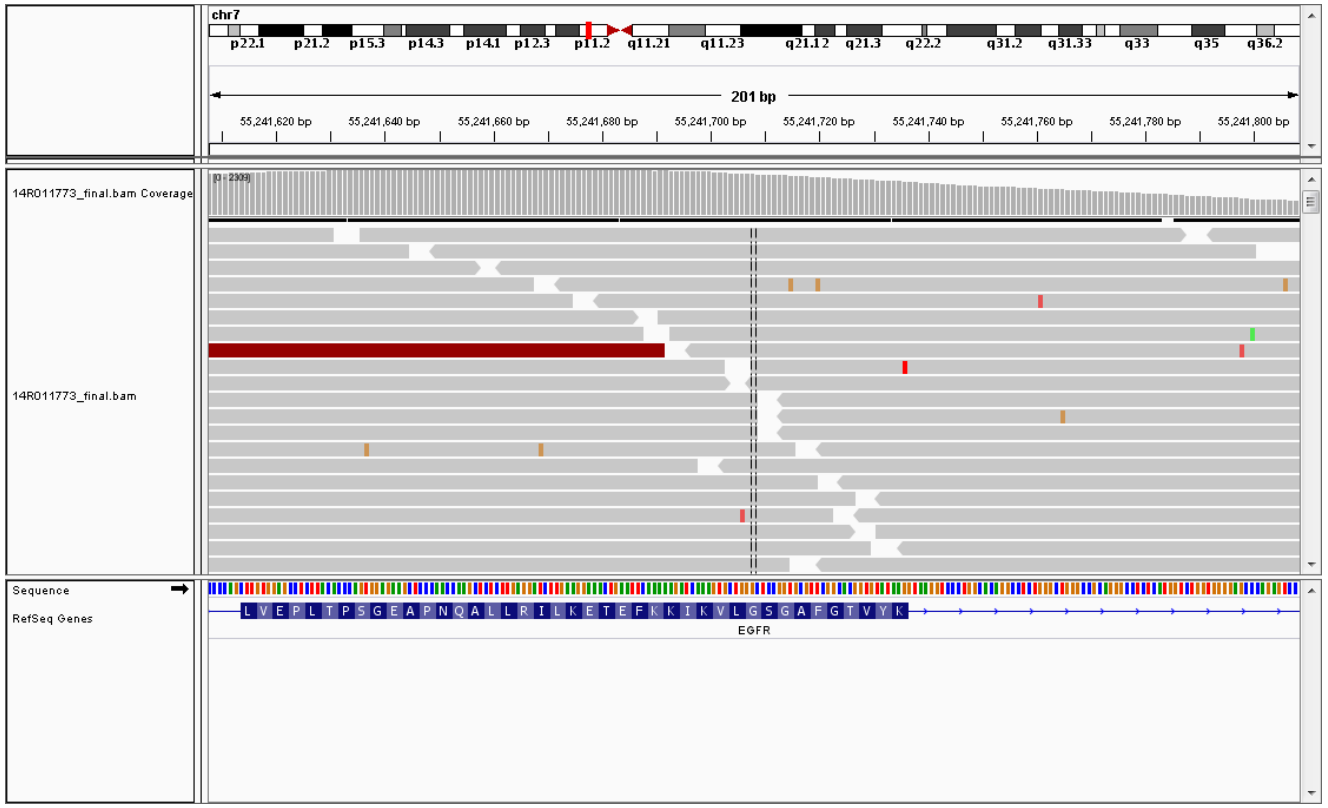


Figure 2.34: Sample 14R011773 alignment around locus chr7:55241708.

From the collected results the sensitivity

$$\widehat{Sensitvity} = \frac{TP}{TP + FN} \tag{2.2}$$

of the enrichment panel can be estimated, where the number of observed true positives TP and false negatives FN were derived from Table 2.11 by counting clashing sites as true low frequency mutations that were missed by pyrosequencing. The estimated sensitivity for the sequencing panel version 2 for all tested loci was calculated as

$$\widehat{Sensitvity}^{V2} = \frac{67}{67 + 2} = 0.971 = 97.10\%, \tag{2.3}$$

if poorly covered sample 14R012024 had been included. Under real conditions this sample would have been excluded due to insufficient on-target coverage, resulting in a sensitivity rate of 0.9853 = 98.53%, i.e. one missed mutation out of 68 tested.

Another interesting case is sample 14R012187, where pyrosequencing reported a p.Gln61Arg mutation in KRAS, while VarScan2 and snpEff has reported p.Gln61Lys. The reason for the discrepancy was a double mutation meaning two SNVs were adjacent to each other in the genome. The mutation can be interpreted as "DelInsAA": a deletion of two bases followed by an insertion of two adenines. The software reading the

pyrogram may have struggled with the 25% double mutation as it was only detected after manually checking the peaks after the variant calling results were analysed.

Utilising massively parallel sequencing and analysing results from all 69 genes that were selected revealed over ten thousand different SNVs and InDels, dozens to hundreds in each sample, even if common SNPs, synonymous and intronic variants were filtered. The vast majority of them turned out to be artefacts from sequencing and formalin treatment as they were low frequency mutations with no report in dbSNP, COSMIC or ClinVar. Only 274 mutations were annotated by ClinVar and 95 of these are either classified as probably pathogenic, pathogenic or as other -such as risk factors- in 25 different genes. On average this was around 1 mutation with clinical relevance per sample. The vast majority of these mutations were, however, rarely present, as there was only one sample for them reported to carry the SNV or InDel. Some mutations were associated with certain cancer types, such as the previously mentioned codon 12 and 13 mutations in KRAS that were commonly reported. Studies showed that about 42% of all colorectal cancers carry either of these mutations [203], which fitted to what was observed, as around 40% of all CRC samples carried a mutation in one or both codons. A number of other cancer types were positive for KRAS codon 12 and 13 mutations too, supporting that they are important driver mutations in other cancer types as well. Altogether around 60% of all samples carried a mutation in KRAS codon 12 or 13 respectively. Figure

2.35 shows a histogram of moderate and high impact predicted variants and small InDels per sample. The mean number of calls was 18 variants per sample, with a standard deviation of 15.1. One sample had to be excluded (14R011779), as over 1,300 medium or high impact mutations were reported for that sample, which was believed to be an artefact. Sample 14R011779 did not show any drops in input quality, quantity, complexity or on-target coverage. Over 90% of these mutations were uniquely present in that sample and the majority were not reported in dbSNP or COSMIC indicating that the majority of them were probably not real. Variants reported by VarScan2 had passed initial filter criteria and showed a sufficient coverage. Furthermore, they did not seem to appear randomly across the genome, as they carried common SNPs present among the British population. Subsequently most calls were filtered by excluding variants below 5% frequency. Either these were artefacts from the embedding that caused a high amount of low frequency mutations that were not tumour related or the sample was collected from a primary tumour at a very late stage and drastically mutated due to very high genetic instability, i.e. showed a high number of subclonal heterogeneity. Four more samples (H12-0020486, H12-0021728, H12-0023107, H12-0024623) showed a higher presence of low-frequency mutations with predicted moderate and high impact as well. Between 59 and 169 medium or high impact SNVs or InDels per sample, hence they were believed to be mainly artefacts as well.

In order to see if there was a general trend, mean coverage was plotted

against the number of filtered variants, shown in Figure 2.36. The presumption from the pilot experiment that on average a minimum on-target coverage leads to a plateau after which VarScan2 does not report any additional variants could be confirmed. The effect became even stronger if low frequency variants and InDels below 5% were excluded, as shown in Figure 2.37. Higher coverages supported VarScan2, however, to further reduce the number of potential false-positives, shown as a slight reduced number of variants called from samples with a high coverage.

As mentioned before many mutations were shared among different cancer types, as they are typical driver mutations that were present in several tumours. Hence, it was considered that samples could be further analysed by clustering them according to their mutation profile. A pragmatic approach was a principal component analysis (PCA), which transforms high dimensional variables into a reduced space by eliminating linear correlations. The more dimensions could be reduced, the stronger the correlations in the data [189]. It is a very popular tool for spotting linear correlations between variables in high dimensional datasets. In this case, each sample was interpreted as a multidimensional vector storing information about presence or absence of mutations across all samples. By plotting the first two dimensions of samples in the transformed space caused correlating samples to cluster, while very unrelated samples had a higher distance from each other. Random mutation patterns indicating an underlying problem with samples,

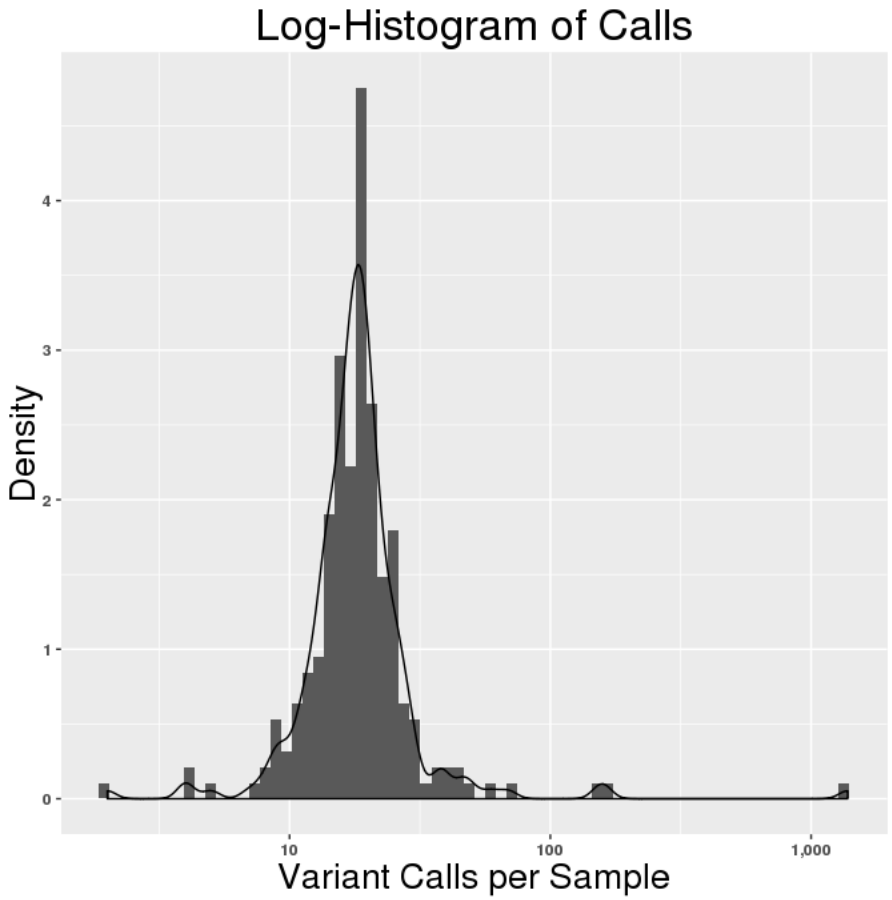


FIGURE 2.35: Histogram of moderate and high impact mutations.

would result in a random dispersion across the reduced space. Figure 2.38 shows a plot of the samples along the two highest principal components [214]. In addition every sample was colour-coded according to their originating primary tissue. Most samples aggregated around two clusters independently from their primary tissue type. Many of the samples with a higher distance to both clusters showed often a lower

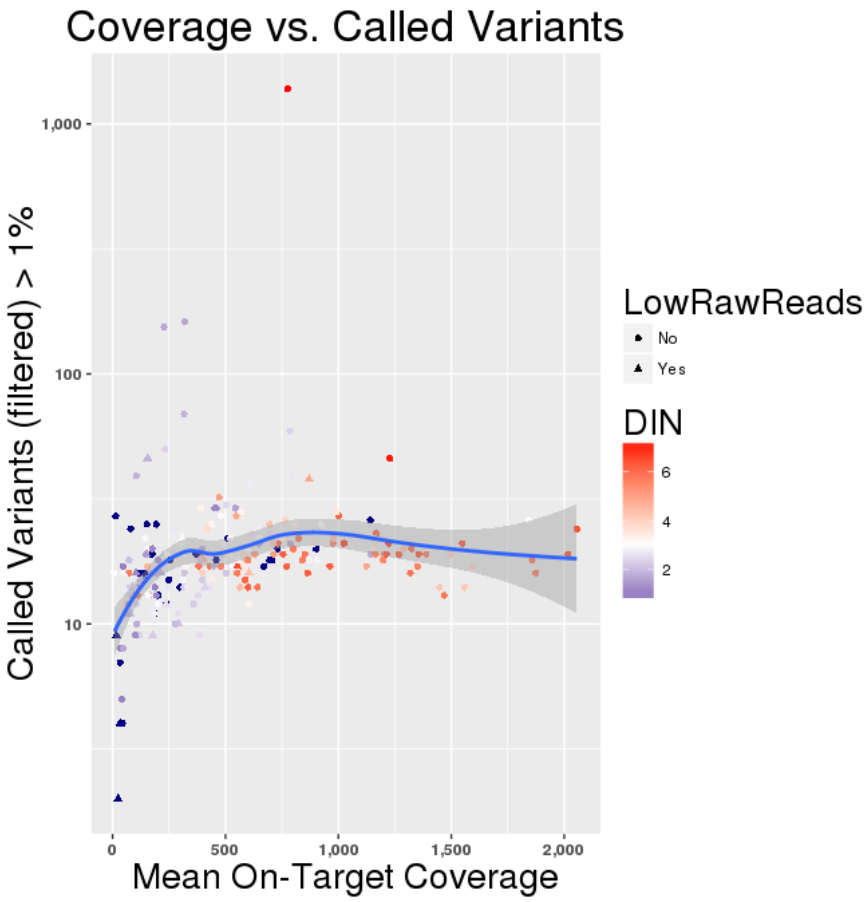


FIGURE 2.36: Mean on-target coverage and pass-filter variants of moderate and high impact, including low-frequency variants. Low number of calls is related to low sequencing yield or low DIN (< 3).

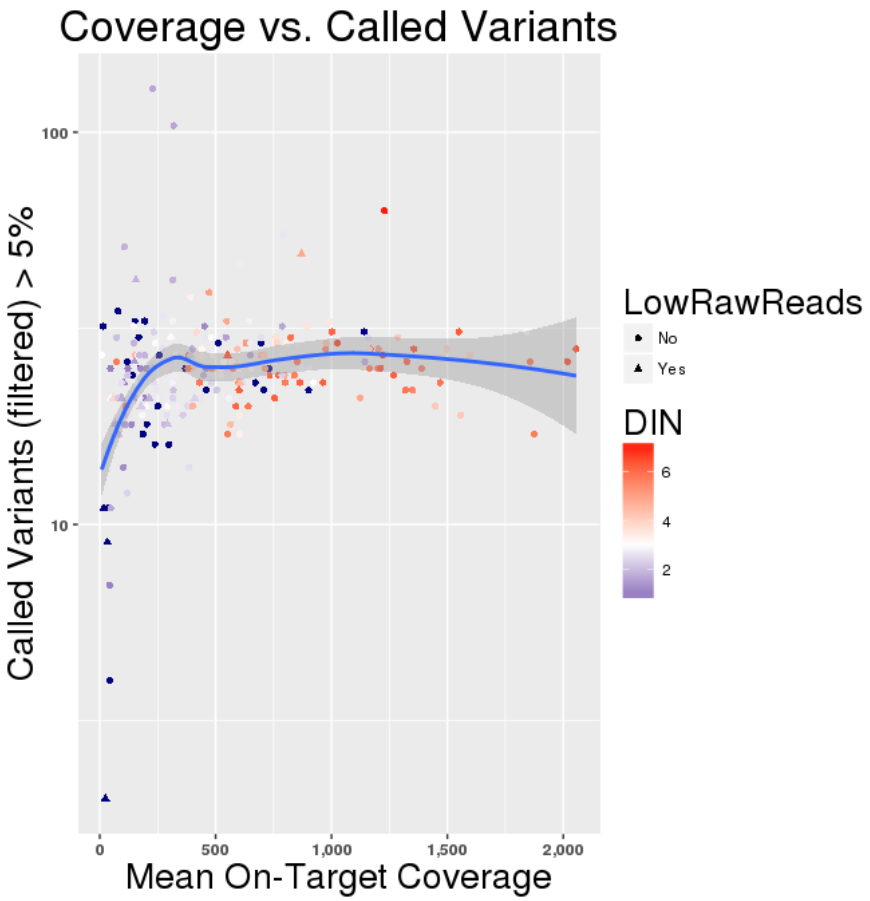


FIGURE 2.37: Mean on-target against number of pass-filter variants with moderate and high impact mutations.

mean on-target coverage. Although this was not true for all samples it gave an indication that they could be exposed to a higher technical noise than samples located in either of the two clusters. The results promise potentially new approaches for further quality and confidence estimation of reported variants.

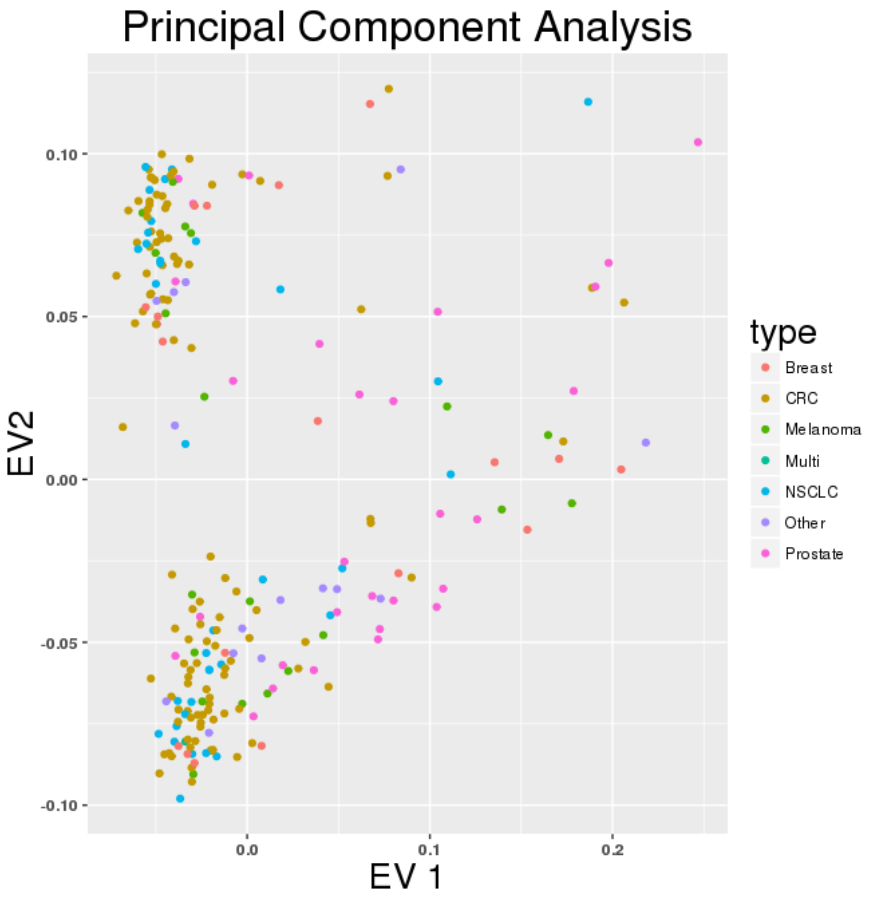


FIGURE 2.38: Plot of samples along the two highest principal components. Only pass-filter variants with a mutation frequency of > 5% were used.

2.4 Discussion

Utilising massively parallel sequencing for precision medicine promises a better throughput and a high number of genetic markers to be tested for by keeping the amount of precious tissue constant. The target capture and enrichment method based on 120bp RNA hybridisation from Agilent SureSelect XT Custom was successfully tested on 230 clinical samples on a NextSeq 500 instrument. The panel showed good performance even at lower input concentrations or insufficient overall amount of genetic material available. It could be shown target enrichment on samples with low DNA integrity showed a lower success rate. Subsequently, a profound correlation analysis gave indication that poor DNA integrity may be compensated by more sequencing in some cases, where complexity was sufficient, but without a profound understanding of the DIN values, only a general trend could be observed. On average 6 million to 8 million reads showed a saturation of the number of variants and small InDels above an allelic threshold of 5% or higher that were identified in most samples. Mutations of a lower frequency showed a significantly higher false-positive rate while their pathological interpretation was very limited in most cases. Limitations of the designed kit were assessed by comparison of results in known mutation hotspots in a selection of genes with reported results from pyrosequencing assays. Library preparation and sequencing problems, alignment artefacts, sample contaminations and a low tumour burden

are potential caveats as with other technologies currently in use. One mutation was missed even if all input and coverage criteria appeared to be within defined requirements resulting in a false negative result in 1 out of 68 cases tested. Massively parallel sequencing showed an overall better detection rate in cases of double mutations and InDels or when DNA was extracted from tissue showing a low tumour burden. Coverage, library complexity and DNA integrity criteria however need to be met. Some improvements can be considered targeting various key aspects of the process. Probes of 120bp in length have some disadvantages, as they can increase off-target coverage if flanking regions are partially covered. In addition, some of the probes designed had a low hybridisation stringency causing them to capture off-target regions. In cases where the amount of sequencing is critical, such as for high-throughput testing, low-stringency probes might be reduced or even removed in some cases. Moreover, one false negative result was observed potentially caused by a biased allele amplification. The reasons for the false negative could not be determined. The coverage in that position was sufficient and both strands were equally amplified and sequenced, i.e. common artefacts such as strand bias could be excluded [160, 161]. A homologous was not found by BLAT [170] excluding a false enrichment and probes showed a high stringency for that region. As no chemistry is perfect it may a rare coincidence of several factors combined. Potentially a higher amount of shorter probes of higher specificity may reduce risk of false-negatives and increase

sensitivity of the panel [177]. The library preparation process could be improved to preserve more fragments throughout the adapter ligation process, i.e. increasing library complexity [168]. As library preparation consists of a number of processing steps, each involving a number of enzymes all having their own weaknesses decreasing effectiveness, such as a high or low GC content, poor DNA concentration or a high level of degradation [104]. Over the past few years a range of improved enzymes have become available to face these problems and to improve yield and prepare libraries of higher complexity to capture sufficient fragments even in cases of low DNA integrity.

Furthermore, data analysis and quality of reported variants may be improved by adjusting filter criteria. More and more possible error sources and artefacts are revealed that can occur in data from massively parallel sequencing, if these problems can be accounted for, it improves results by reducing the number of false positive mutations that are reported. Just recently, results of the CGA-ICGC DREAM-3 SNV Challenge were published [152]. A number of new filter criteria were described that may even prove a new gold standard for filtering sequencing data from massively parallel sequencing data.

Last but not least, it should be mentioned that going from testing single loci, such as mutation hotspots, to entire coding genes, is a big step, as it is difficult to get profound estimates of sensitivity and specificity across the entire genome. Cost and time constrains make it difficult to prove

every mutation right, but showing a base in a sample is not mutated is even harder, which would be necessary for a profound estimation of specificity. Hence, showing that no mutation are missed from real samples in all cases is impossible at the moment leaving hypothesis-free approaches difficult to interpret outside of known mutation hotspots.

References

[132] Hafid Alazzouzi et al. "SMAD4 as a prognostic marker in colorectal cancer". In: *Clinical Cancer Research* 11.7 (2005), 2606–2611.

[133] Walter Alexander. "Inhibiting the akt pathway in cancer treatment: three leading candidates". In: *Pharmacy and Therapeutics* 36.4 (2011), p. 225.

[134] Deborah A Altomare and Joseph R Testa. "Perturbations of the AKT signaling pathway in human cancer". In: *Oncogene* 24.50 (2005), pp. 7455–7464.

[135] Jamie N Anastas and Randall T Moon. "WNT signalling pathways as therapeutic targets in cancer". In: *Nature Reviews Cancer* 13.1 (2013), pp. 11–26.

[136] Simon Andrews. *FastQC: A quality control tool for high throughput sequence data*. Website. http://bioinformatics.babraham.ac.uk/projects/fastqc/. Oct. 2015.

[137] Koji Aoki and Makoto M Taketo. "Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene". In: *Journal of cell science* 120.19 (2007), pp. 3327–3335.

[138] Uzma Asghar et al. "The history and future of targeting cyclin-dependent kinases in cancer therapy". In: *Nature reviews Drug discovery* 14.2 (2015), pp. 130–146.

[139] Alberto Bardelli et al. "Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer". In: *Cancer discovery* 3.6 (2013), pp. 658–673.

[140] Heiko Becker et al. "Tracing the development of acute myeloid leukemia in CBL syndrome". In: *Blood* 123.12 (2014), pp. 1883–1886.

[141] B A Benayoun et al. "The forkhead factor FOXL2: A novel tumor suppressor?" In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 1–5.

[142] Mohamed Bentires-Alj et al. "Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia". In: *Cancer research* 64.24 (2004), pp. 8816–8820.

[143] Valérie Bonadona et al. "Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome". In: *Jama* 305.22 (2011), pp. 2304–2310.

[7] Broadinstitue. *Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*. http://broadinstitute.github.io/picard. Nov. 2015.

[144] Jennifer L Bromberg-White, Nicholas J Andersen, and Nicholas S Duesbery. "MEK genomics in development and disease". In: *Briefings in functional genomics* 11.4 (2012), pp. 300–310.

[145] CJ Chang and MC Hung. "The role of EZH2 in tumour progression". In: *British journal of cancer* 106.2 (2012), pp. 243–247.

[146] F Chang and Marilyn M Li. "Clinical application of amplicon-based next-generation sequencing in cancer". In: *Cancer genetics* 206.12 (2013), pp. 413–419.

[147] I-Ming Chen et al. "Outcome modeling with CRLF2, IKZF1, JAK, and minimal residual disease in pediatric acute lymphoblastic leukemia: a Children's Oncology Group study". In: *Blood* 119.15 (2012), pp. 3512–3522.

[148] Jon H Chung and Fred Bunz. "A loss-of-function mutation in PTCH1 suggests a role for autocrine hedgehog signaling in colorectal tumorigenesis". In: *Oncotarget* 4.12 (2013), p. 2208.

[149] Magdalena Cizkova et al. "PIK3R1 underexpression is an independent prognostic marker in breast cancer". In: *BMC cancer* 13.1 (2013), p. 545.

[150] Tao Dao et al. "Targeting the intracellular WT1 oncogene product with a therapeutic human antibody". In: *Science translational medicine* 5.176 (2013), 176ra33–176ra33.

[24] Lloye M Dillon and Todd W Miller. "Therapeutic targeting of cancers with loss of PTEN function". In: *Current drug targets* 15.1 (2014), p. 65.

[151] Mark Ewalt et al. "Real-time PCR-based analysis of BRAF V600E mutation in low and intermediate grade lymphomas confirms frequent occurrence in hairy cell leukaemia". In: *Hematological oncology* 30.4 (2012), pp. 190–193.

[152] Adam D Ewing et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection". In: *Nature methods* 12.7 (2015), pp. 623–630.

[153] Monica Hoyos Flight. "Anticancer drugs: A sweet blow for cancer cells". In: *Nature Reviews Drug Discovery* 10.10 (2011), pp. 734–734.

[154] William D Foulkes et al. "The CDKN2A (p16) gene and human cancer." In: *Molecular Medicine* 3.1 (1997), p. 5.

[155] Ermanno Gherardi et al. "Targeting MET in cancer: rationale and progress". In: *Nature Reviews Cancer* 12.2 (2012), pp. 89–103.

[156] David W Goodrich. "The retinoblastoma tumor-suppressor gene, the exception that proves the rule". In: *Oncogene* 25.38 (2006), pp. 5233–5243.

[157] Tiziana Grafone et al. "An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment". In: *Oncology reviews* 6.1 (2012).

[158] F Graziano, B Humar, and P Guilford. "The role of the E-cadherin gene (CDH1) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice". In: *Annals of oncology* 14.12 (2003), pp. 1705–1713.

[159] Silvia Grisendi et al. "Nucleophosmin and cancer". In: *Nature Reviews Cancer* 6.7 (2006), pp. 493–505.

[160] Yan Guo et al. "Exome sequencing generates high quality data in non-target regions". In: *BMC genomics* 13.1 (2012), p. 1.

[161] Yan Guo et al. "The effect of strand bias in Illumina short-read sequencing data". In: *BMC genomics* 13.1 (2012), p. 1.

[162] Carolina Gutierrez and Rachel Schiff. "HER2: biology, detection, and clinical implications". In: *Archives of pathology & laboratory medicine* 135.1 (2011), pp. 55–62.

[163] Z Han et al. "Reversal of multidrug resistance of gastric cancer cells by downregulation of Akt1 with Akt1 siRNA." In: *Journal of experimental & clinical cancer research: CR* 25.4 (2006), pp. 601–606.

[164] Carl-Henrik Heldin. "Targeting the PDGF signaling pathway in tumor treatment". In: *Cell Communication and Signaling* 11.1 (2013), p. 1.

[165] Daniel Herranz et al. "Metabolic reprogramming induces resistance to anti-NOTCH1 therapies in T cell acute lymphoblastic leukemia". In: *Nature medicine* (2015).

[166] Matthew Holderfield et al. "Targeting RAF kinases for cancer therapy: BRAF mutated melanoma and beyond". In: *Nature reviews. Cancer* 14.7 (2014), p. 455.

[167] Yoshiaki Ito, Suk-Chul Bae, and Linda Shyue Huey Chuang. "The RUNX family: developmental regulators in cancer". In: *Nature Reviews Cancer* 15.2 (2015), pp. 81–95.

[168] Marcus B Jones et al. "Library preparation methodology can influence genomic and functional predictions in human microbiome research". In: *Proceedings of the National Academy of Sciences* 112.45 (2015), pp. 14024–14029.

[169] Masaru Katoh and Hitoshi Nakagama. "FGF receptors: cancer biology and therapeutics". In: *Medicinal research reviews* 34.2 (2014), pp. 280–300.

[170] W James Kent. "BLAT-the BLAST-like alignment tool". In: *Genome research* 12.4 (2002), pp. 656–664.

[171] Kum Kum Khanna. "Cancer risk and the ATM gene: a continuing debate". In: *Journal of the National Cancer Institute* 92.10 (2000), pp. 795–802.

[172] Dmitriy Khodakov, Chunyan Wang, and David Yu Zhang. "Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches". In: *Advanced drug delivery reviews* (2016).

[173] Piotr Kozlowski, Mateusz de Mezer, and Wlodzimierz J Krzyzosiak. "Trinucleotide repeats in human genome and exome". In: *Nucleic acids research* (2010), gkq127.

[174] Madhu S Kumar et al. "The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer". In: *Cell* 149.3 (2012), pp. 642–655.

[60] Melissa J Landrum et al. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic acids research* 42.D1 (2014), pp. D980–D985.

[175] Virpi Launonen. "Mutations in the human LKB1/STK11 gene". In: *Human mutation* 26.4 (2005), pp. 291–297.

[176] Ryan S Lee et al. "A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers". In: *The Journal of clinical investigation* 122.8 (2012), pp. 2983–2988.

[177] Stefan H Lelieveld et al. "Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions". In: *Human mutation* 36.8 (2015), pp. 815–822.

[178] Johan Lennartsson and L Ronnstrand. "The stem cell factor receptor/ c-Kit as a drug target in cancer". In: *Current cancer drug targets* 6.1 (2006), pp. 65–75.

[70] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[179] Hui Jun Lim, Philip Crowe, and Jia-Lin Yang. "Current clinical regulation of PI3K/PTEN/Akt/mTOR signalling in treatment of human cancer". In: *Journal of cancer research and clinical oncology* 141.4 (2015), pp. 671–689.

[180] TC Lin et al. "CEBPA methylation as a prognostic biomarker in patients with de novo acute myeloid leukemia". In: *Leukemia* 25.1 (2011), pp. 32–40.

[181] Camille Lobry, Philmo Oh, and Iannis Aifantis. "Oncogenic and tumor suppressor functions of Notch in cancer: its NOTCH what you think". In: *The Journal of experimental medicine* 208.10 (2011), pp. 1931–1935.

[182] R Marone et al. "Targeting phosphoinositide 3-kinasemoving towards therapy". In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1784.1 (2008), pp. 159–185.

[183] Fausto Meriggi et al. "The emerging role of NRAS mutations in colorectal cancer patients selected for anti-EGFR therapies". In: *Reviews on recent clinical trials* 9.1 (2014), pp. 8–12.

[184] Patricia AJ Muller and Karen H Vousden. "Mutant p53 in cancer: new functions and therapeutic opportunities". In: *Cancer cell* 25.3 (2014), pp. 304–317.

[185] Lois M Mulligan. "RET revisited: expanding the oncogenic portfolio". In: *Nature Reviews Cancer* 14.3 (2014), pp. 173–186.

[186] F Musumeci et al. "An update on dual Src/Abl inhibitors". In: *Future medicinal chemistry* 4.6 (2012), pp. 799–822.

[187] Magali Olivier, Monica Hollstein, and Pierre Hainaut. "TP53 mutations in human cancers: origins, consequences, and clinical use". In: *Cold Spring Harbor perspectives in biology* 2.1 (2010), a001008.

[188] Sharmila Patel and Mark R Player. "Colony-stimulating factor-1 receptor inhibitors for the treatment of cancer and inflammatory disease". In: *Current topics in medicinal chemistry* 9.7 (2009), pp. 599–610.

[189] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[190] *QIAamp DNA FFPE Tissue Handbook*. 06/2012. Qiagen. June 2012.

[104] Michael G Ross et al. "Characterizing and measuring bias in sequence data". In: *Genome Biol* 14.5 (2013), R51.

[191] Charles M Rudin. "Vismodegib". In: *Clinical Cancer Research* 18.12 (2012), pp. 3218–3222.

[192] Elisa Rumi et al. "Clinical effect of driver mutations of JAK2, CALR, or MPL in primary myelofibrosis". In: *Blood* 124.7 (2014), pp. 1062–1069.

[193] Antonio Russo et al. "The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment". In: *Journal of clinical oncology* 23.30 (2005), pp. 7518–7528.

[194] Alice T Shaw et al. "Ceritinib in ALK-rearranged non–small-cell lung cancer". In: *New England Journal of Medicine* 370.13 (2014), pp. 1189–1197.

[195] Q Sheng and J Liu. "The therapeutic potential of targeting the EGFR family in epithelial ovarian cancer". In: *British journal of cancer* 104.8 (2011), pp. 1241–1245.

[196] Karen E Sheppard and Grant A McArthur. "The cell-cycle regulator CDK4: an emerging therapeutic target in melanoma". In: *Clinical Cancer Research* 19.19 (2013), pp. 5320–5328.

[109] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[197] Ritsuko Shimizu, J D Engel, and M Yamamoto. "GATA1-related leukaemias". In: *Nature Reviews Cancer* 8.4 (2008), pp. 279–287.

[198] Martha L Slattery, Abbie Lundgreen, and Roger K Wolff. " VEGFA, FLT1, KDR and colorectal cancer: assessment of disease risk, tumor molecular phenotype, and survival". In: *Molecular carcinogenesis* 53.S1 (2014).

[199] E Solary et al. "The Ten-Eleven Translocation-2 (TET2) gene in hematopoiesis and hematopoietic diseases". In: *Leukemia* 28.3 (2014), pp. 485–496.

[200] Len Stephens, Roger Williams, and Phillip Hawkins. "Phosphoinositide 3-kinases as drug targets in cancer". In: *Current opinion in pharmacology* 5.4 (2005), pp. 357–365.

[201] *SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library Protocol*. Version B2. Agilent Technologies. Apr. 2015.

[202] J Tan et al. "EZH2: biology, disease, and structure-based drug discovery". In: *Acta Pharmacologica Sinica* 35.2 (2014), pp. 161–174.

[203] Cecily P Vaughn et al. "Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer". In: *Genes, Chromosomes and Cancer* 50.5 (2011), pp. 307–312.

[204] Yuanxiang Wang et al. "Targeting mutant KRAS for anticancer therapeutics: a review of novel small molecule modulators". In: *Journal of medicinal chemistry* 56.13 (2013), pp. 5219–5230.

[205] Zhiwei Wang et al. "Targeting Notch signaling pathway to overcome drug resistance for cancer therapy". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1806.2 (2010), pp. 258–267.

[206] Zhiwei Wang et al. "Tumor suppressor functions of FBW7 in cancer development and progression". In: *FEBS letters* 586.10 (2012), pp. 1409–1418.

[207] E Weisberg and J D Griffin. "Mechanism of resistance to the ABL tyrosine kinase inhibitor STI571 in BCR/ABL–transformed hematopoietic cell lines". In: *Blood* 95.11 (2000), pp. 3498–3505.

[208] Stephen Q Wong et al. "Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours". In: *Scientific reports* 3 (2013).

[209] Mingzhao Xing. "Molecular pathogenesis and mechanisms of thyroid cancer". In: *Nature Reviews Cancer* 13.3 (2013), pp. 184–199.

[210] Min Yan et al. "HER2 expression status in diverse cancers: review of results from 37,992 patients". In: *Cancer and Metastasis Reviews* 34.1 (2015), pp. 157–164.

[211] L1 Yang et al. "A tumor suppressor and oncogene: the WT1 story". In: *Leukemia* 21.5 (2007), pp. 868–876.

[212] Chetan Yewale et al. "Epidermal growth factor receptor targeting in cancer: a review of trends and strategies". In: *Biomaterials* 34.34 (2013), pp. 8690–8707.

[213] Bing Yu et al. "Targeting protein tyrosine phosphatase SHP2 for the treatment of PTPN11-associated malignancies". In: *Molecular cancer therapeutics* 12.9 (2013), pp. 1738–1748.

[214] Xiuwen Zheng et al. "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24 (2012), pp. 3326–3328.

[131] *bcl2fastq2 Conversion Software Guide*. 2.17 (15051736 Rev. G). ILLUMINA PROPRIETARY. Illumina. July 2015.

Chapter 3

A Modified Amplicon Cancer Sequencing Panel

Testing genetic markers by massively parallel sequencing benefits from robustness and performance even from a low amount of starting material, even if more than just a few genes are tested. In this chapter a new cancer panel is introduced with a low number of target genes. A speed up of sample processing and library preparation time is intended by using a modified amplicon-seq panel. Conventional amplicon sequencing, an enrichment method where regions are amplified with a multiplexed PCR, is prone to a range of technical biases that can distort genotyping results and cause a biased enrichment of regions. This panel uses two major improvements: target capture by adding biotinylated probes

ligating target regions into rings prior amplification, called *molecular inversion probes* and tagging amplicons with a molecular barcode [220, 222]. The unique molecular barcodes enable PCR duplicate removal and allow correction for sequencing errors after the data was aligned. This can help overcome described limitations to make it suitable to be used in clinical diagnostics [218]. After a brief introduction and an overview of potential benefits from the designed panel, modified amplicon sequencing method based on the Agilent HaloPlex HS technology is described. The panel covers nine genes that were selected to replace current methods of clinical testing in precision medicine. It was tested on 48 clinical samples that were previously sequenced with pyrosequencing at known mutation hotspots. Results are compared and discussed to address strengths, limitations and caveats of this method in context of currently established methods.

3.1 Motivation

Many cancer panels have been released based on different enrichment strategies and genes or regions targeted. They divide into two main principles. The first enrichment method is based on-target capture, where DNA is sheared into fragments, sequencing adapters are ligated and fragments of interest are captured with biotinylated, single-stranded

DNA or RNA probes hybridising to a complementary sequence. Captured fragments are pulled-down with magnetic beads, washed, amplified followed by sequencing. The second method enriches regions by amplification, where primer pairs are designed for each target region for a multiplexed PCR. Depending on the number and sizes of these regions a few dozen to hundreds of such PCR primers are designed to sufficiently amplify these regions. Subsequently, the amplicons undergo a clean-up and adapter ligation process to build a library that is sequenced [215].

Target capture-based methods usually preserve mutation frequencies better, as the number artefacts from amplification is reduced. Random shearing of DNA causes random double-strand breaks allowing subsequent removal of PCR duplicates, as every fragment in a sequencing library is believed to be unique causing reads to all have different starting positions and alignments. One important aspect for clinical use is a relatively high amount of input DNA needed. The Agilent SureSelect XT protocol, for example, recommends 200ng - 3000ng of input for preparing an enriched library. There is very little control over quantity and quality of FFPE samples that are tested, hence input recommendations may not always be met. Another important consideration is the time it takes to prepare a library from extracted DNA. Target capture protocols take between 36 and 96 hours due to an extended probe hybridisation step, as described in Section A.4. Clinicians and pathologists, however, need a result within two to three days in

most cases. In addition with time needed for sequencing, there is very little time left for data analysis or a repeat in case of a failed process.

Amplicon sequencing benefits from a much simpler preparation protocol, requiring less input DNA and time. Qiagen's GeneRead™ V2 panels, for example, can prepare a library in under three hours [217]. Enrichment protocols based on multiplexed PCR, however, are limited by the number of PCR primers that can be used in a reaction. Overlapping primer pairs need a careful design to prevent some regions to be amplified higher than others. A strong imbalance in amplification leaves target regions covered very unevenly, which subsequently needs to be compensated by an increase of sequencing possibly negating cost advantage from a simplified sample preparation protocol. Further, amplicon sequencing is neither a safe method of reliably keeping mutation frequencies throughout the process, nor can duplicates be distinct from uniquely sequenced molecules afterwards, as all amplicons were derived by the same primer pair sharing the same start and end position. Related to this, data analysis pipelines are significantly harder to configure for all eventualities, especially if the number of primer pairs is very high.

Both methods show drawbacks in different aspects of their technologies and make high-throughput diagnostic testing in personalised medicine a challenge. Barcoded molecular inversion probes could be a potential solution as they benefit from a faster probe hybridisation, but preserve

allelic frequencies throughout the enrichment process. A commercial enrichment panel based on barcoded molecular inversion probes is the Agilent HaloPlex HS system. It combines the strengths of a target capture process with those from amplicon enrichment. Turnaround time to build a sequencing library from extracted DNA lies between 8 and 24 hours, depending on the number and size of target regions.

3.2 Agilent HaloPlex HS Target Enrichment

Agilent HaloPlex HS can be seen as a hybrid between target capture and amplification that can enrich regions between 1kb and 5Mb in size and was designed to work from 50ng of input material, although for DNA extracted from FFPE tissue the number should be increased to 200ng where possible, as the system is sensitive to degraded DNA. Extracted DNA is digested, then biotinylated probes bind to complementary motifs, which are then captured, barcoded and amplified. Although it sounds similar to other target capture methods, the fundamental difference lies in the probes library. Instead of capturing fragments directly by hybridisation, the probes consist of two parts. The first part consists of two binding sites that hybridise to the target. After the second probe was added, the target region is ligated into a uniquely barcoded, biotinylated ring that is captured. Molecular inversion probes speed up the enrichment process compared to conventional probes as library preparation and amplification steps are combined. Further,

binding sites of the probes are shorter reducing time necessary for hybridisation. Unlike with conventional probes, where DNA can be randomly sheared into small fragments, the molecular inversion probes need directed fragments to ensure target region are are not disrupted, as otherwise circularisation would fail. DNA is, therefore, digested in 30 minutes with a mix of 16 different restriction enzymes instead of random shearing. It ensures that DNA fragments all have known start and end sites, as shown in Figure 3.1. Extracted DNA is aliquotted into 8 wells mixed with two different restriction enzymes. Unlike random shearing, a fixed number of fragment sizes can be expected, as shown in Figure 3.2. Due to the different principle of probe hybridisation there is no adapter ligation step necessary. Instead, digested DNA is directly mixed with the probes for hybridisation. As described before, each probe consists of two parts: a flexible oligo that carries two binding sites matching the target and a fixed part, which is used for target capture, ligation, barcoding and amplification, as shown in Figure 3.4 and Figure 3.3. After probe hybridisation the entire region between probe binding sites is circularised and ligated to a ring. Hybridisation of the probes takes about 2 hours to complete and ligation is performed in another 15 minutes. Due to the flexible probe, rings are biotinylated, hence they can be captured by streptavidin coated, magnetic beads, sketched in Figure 3.6. It separates target regions from unwanted DNA fragments. The captured rings are not eligible for sequencing, yet. They need to be amplified by PCR using the bridge primer sequence present

in the probe to generate a set of linearised amplicons, as shown in Figure 3.7. The pooled libraries are now ready to be sequenced at the desired run length. It requires dual-indexing to read both barcodes. The sequencing instrument performs two indexing runs in addition to the forward and reverse sequencing. The first index is the usual sample barcode, while the second index read is the moleculare barcode, hence, it is not used for demultiplexing. Instead, determined molecular barcode sequences are written into a dictionary mapping the unique read id to a molecular barcode sequence, which can be used later for de-duplication and error correction, e.g.

```
@M00762:267:000000000-AMM19:1:2105:14516:1761 2:N:0:ATTGAGGA+TACAATATAC
TACAATATAC
+
11>11B1BBD
```

The dictionary is a FASTQ file that contains the unique read ID in conjunction with the molecular barcode sequence (`TACAATATAC`). In summary, every molecule carries two barcodes, the first is sample specific, while the second is probe specific, hence, the origin of every amplicon can be identified by the two barcodes. If two reads carry the same combination of molecular and sample barcode and alignes to the same locus it is a PCR duplicate. In other words, introduced molecular barcodes do not have to be entirely random, as they need to be unique for a probe only. With an additional analysis step barcodes from the dictionary can be used to collapse read duplicates and to

correct for sequencing errors, as outlined in Figure 3.8. Due to the target capture step, the system does not rely on efficiency of region-specific PCR primers and potential artefacts that are related to it. In addition, target capture benefits from a fast hybridisation step and focusses on one amplification step towards the end. Subsequent correction of sequencing errors and duplication removal are even not available for other capture-based enrichment methods, as duplicates carrying a sequencing error are not identified as such. There are multiple HaloPlex HS probes spanning the region of interest, as shown in Figure 3.5 to increase coverage, confidence and reduce risk of low coverage of a region due to poor enrichment, e.g. because a probe was unable to bind due to a variant or InDel in either of the binding sites. As for other target enrichment methods, some regions are difficult to design probes for, e.g. due to homologous or repetitive regions or a problematic GC content. Unlike other target capture methods that use long oligos for hybridisation, the binding sites of molecular inversion probes are much shorter and need to be present as a pair with a certain distance towards each other, just like primers, but on the same strand. It makes probe design more difficult in some cases.

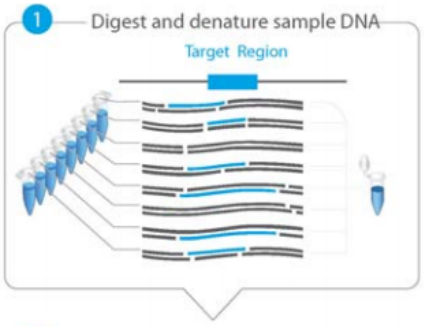


FIGURE 3.1: Agilent HaloPlex HS target enrichment step 1: DNA digest. A combination of enzymes digests DNA at known sites. Image from Agilent [219].

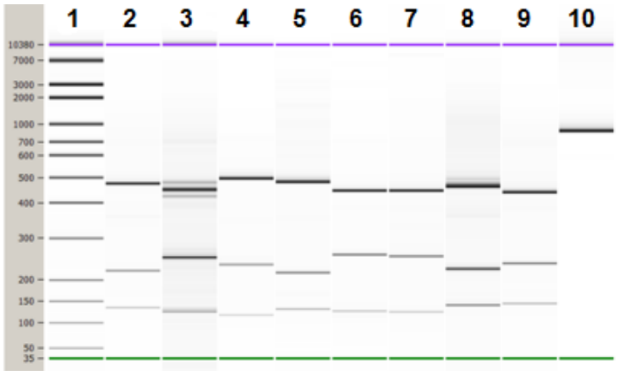


FIGURE 3.2: Bioanalzyer electropherogram from DNA Digest. A fixed number multiple sharp bands are obtained from digestion. Image from Agilent [219].

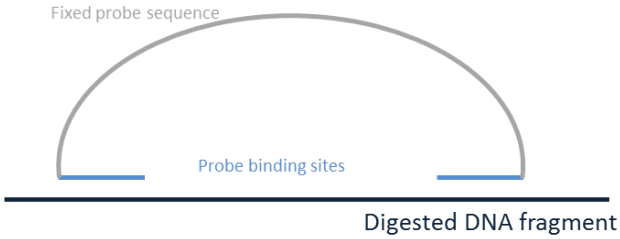


FIGURE 3.3: A simplified picture of the "flexible" probe that carries two binding sites that are used for hybridisation (light blue). Both sites hybridise to DNA fragment (dark blue). In between there is a fixed sequence (grey). Probe is shown as a half-circle.

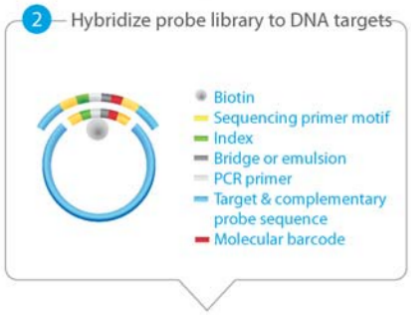


FIGURE 3.4: Full picture of both parts: "flexible" and "fixed". DNA target is shown as a half-circle. The flexible probe hybridises to the DNA and in addition contains complementary sequences for Illumina adapter, Illumina barcode, PCR primers (bridge, PCR) and a molecular barcode unique for the probe. The fixed probe is biotinylated and is built of Illumina adapter, amplification primers (bridge, PCR) and the unique molecular barcode. Image from Agilent [219]

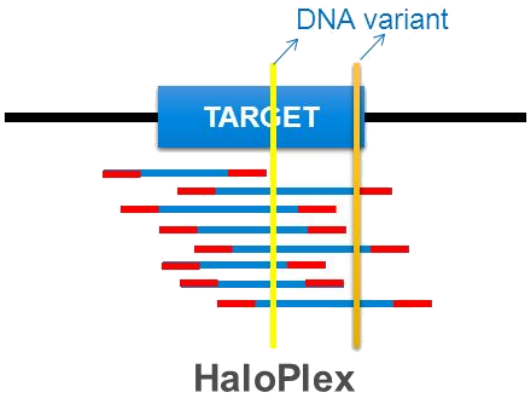


FIGURE 3.5: Multiple probes are designed for a target with different start and stop sites to increase library complexity and to compensate for if a probe cannot bind due to mutations in either of the binding sites of a probe. Image from Agilent [219].

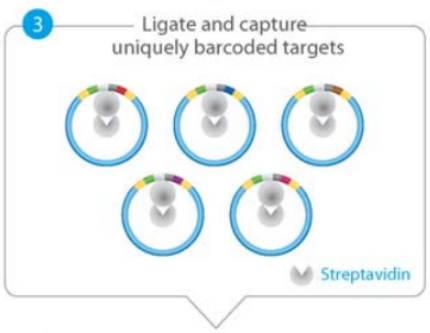


FIGURE 3.6: Agilent HaloPlex HS target enrichment step 2: target capture. After ligation, biotinylated rings are captured with streptavidin coated beads. Image from Agilent [219].

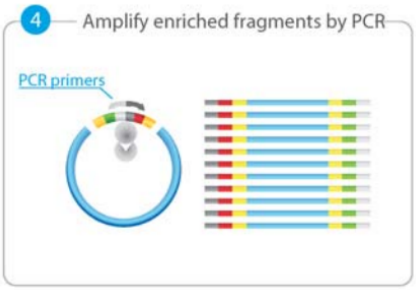


FIGURE 3.7: Agilent HaloPlex HS target enrichment step 3: amplification. Enriched rings are amplified by PCR with complementary PCR and bridge primers. Amplicon library is purified and can be pooled with with libraries from other samples. Image from Agilent [219].

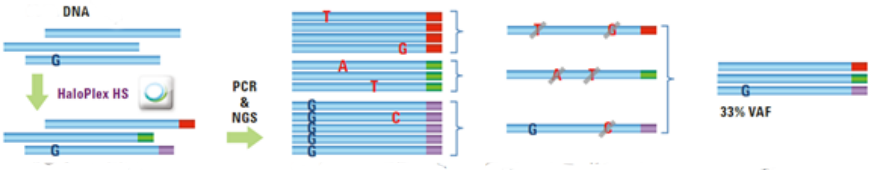


FIGURE 3.8: Summary of the HaloPlex HS system and subsequent deduplication and read error correction. Due to the molecular barcodes, reads (blue) are tagged by their derived fragment (indicated by coloured ends). The molecular barcode can be used to collapse duplicated amplicons and correct sequencing errors. This is supposed to preserve variant allele frequencies throughout the target enrichment. Image from Agilent [219].

3.3 The Panel Design

As motivated before, the panel was designed to replace current hotspot testing in KRAS, NRAS, BRAF, EGFR, PIK3CA, KIT and PDGFRA. Their diagnostic potential is summarised in Table 2.1. In addition exons of genes TP53 and HRAS were added. TP53 is commonly mutated in many different cancer types and due its key role in many different pathways including regulation of cell apoptosis and DNA repair, it is of interest for many clinicians [230]. HRAS is another member of the RAS-gene family, playing an important role in cell division and growth. It is an oncogene playing a central role in several types of cancer, such as bladder cancer [216, 233]. Recent studies have shown that HRAS is another important participant in the EGFR downstream signalling pathway, hence it may prove a diagnostic marker in the near future, just like the other two members of the RAS family [227]. The final panel targeted 23,228 bases including 10 bases up- and downstream of each exon, of which over 99% of the bases were covered with nearly 2,000 probes. The insert size per fragment, i.e. the distance between both probe binding sites, was adjusted to reach a desired read length of 150bp paired-end. The panel captured most coding exons of the target genes, but a few bases in some genes were missed, listed in Table 3.1. For these regions no probes could be designed, due to structural constrains. Although no coverage could expected from these regions, the panel design was accepted. There was one known mutation site in

EGFR reported in COSMIC (COSM1550027) that could not be tested for. Hence, the base had to be excluded. Further, a 55bp in TP53 was reported not to be covered, although the reported region was not exonic according to RefSeq [228]. The HaloPlex HS design software, SureDesign, used additional annotation databases to RefSeq, i.e. Ensembl [223], CCDS [229], Gencode [221], VEGA [231], dbSNP [109] and UCSC cytoband track [224]. As shown in Figure 3.9 potentially interesting features were automatically added by SureDesign as they could be of interest. These features were not utilised for any genetic marker testing, hence these regions were ignored. Although they may decrease cost-efficiency bit, they were kept in case they can be utilised in the future.

Like with other target capture methods probes could bind at different efficiencies. Designed HaloPlex HS probes were classified into three main groups: maximum specificity, maximum coverage and balanced. As the name implies, maximum specificity probes uniquely mapped to one locus in the genome with no mismatch. Such probes were preferred, as they captured only target regions and had a low chance of hybridising elsewhere in the genome. Balanced probes were considered by the design algorithm on regions that could not be covered sufficiently with maximum specificity probes. They were limited to 2 matches elsewhere in the genome. On average this meant 1 out of 2 or 1 out of 3 amplicons sequenced were derived from the fragment of interest. It increased the number of off-target reads sequenced. Maximum coverage

probes were the least stringent option and were only considered in cases where target bases were not or insufficiently covered by probes with a higher specificity. There were up to ten different matches elsewhere in the genome allowed for probes from that category. In other words, up to 90% of these amplicons could subsequently align off-target to the reference genome. These probes increased off-target rate a lot and were only considered as a last resource, as they needed a much higher concentration in the final mix than probes of higher specificity to achieve the same number of on-target amplicons. As shown in Figure 3.11, most binding sites of probes bound uniquely to target regions. Some bases were exclusively covered by maximum coverage probes meaning enrichment efficiency in those region could potentially drop, if probes were not sufficiently balanced. SureDesign added probes in flanking regions of some exons to partially improve coverage. Due to enzymatic digestion of the DNA and design of the probes resulting amplicons always covered the desired part of a target region, as shown in Figure 3.10. In some cases they were needed to provide sufficient number of probes covering target regions for a robust amplification.

As described before size distribution of sequencing libraries generated by HaloPlexHS are not random around an average fragment size. Whilst Bioanalyzer traces from randomly sheared libraries show a smooth curve, Agilent HaloPlex HS traces show usually a jagged curve of fragment sizes scattered between 200bp and 500bp in the electropherogram, as shown in Figure 3.12A. During library validation of the custom

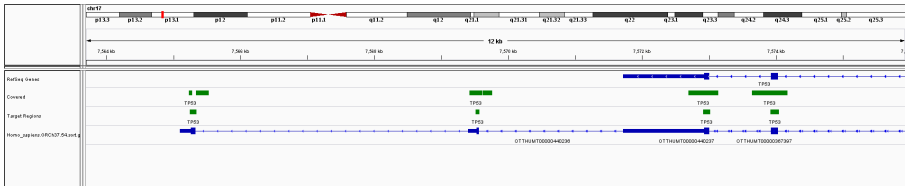


FIGURE 3.9: Section of TP53 gene annotation databases used by the probe design algorithm. In addition to the RefSeq gene annotation (top track), VEGA (bottom track) annotates two additional coding exons. The algorithm considered these regions as well and designed probes spanning that region (green tracks).

designed probe mix, Bioanalyzer traces showed a very unusual size distribution, as shown in Figure 3.12B. An unusually large concentration of fragments around 280bp was observed, represented as a spike in the electropherogram. Although it cannot be said for sure what the exact reasons for that spike were, it was assumed to be a problem with a subset of probes generating all fragments of the same length, potentially off-target regions. Even after the panel was re-manufactured that signal intensity spike reoccurred at the same size. This could have had an impact on the sequencing results, such as an increased duplication rate or an imbalanced enrichment, sequencing yield per sample was increased during evaluation step, as a precaution. Agilent recommended a minimum sequencing of 14.675Mbp per sample, i.e. about 50,000 paired-end reads of 150bp in length, which translates into a theoretical coverage of 630x per sample prior collapsing read duplicates. For this experiment, this number was increased by factor 20, i.e. about 1 million reads were sequenced per sample.

Gene (Exon)	**Locus**	**Length**	**Amino acid**
PIK3CA (12)	chr3:178937506-178937519	13bp	632-636
PIK3CA (13)	chr3:178937833-178937838	5bp	670-672
EGFR (24)	chr7:55266407-55266410	3bp	901
TP53 (NA)	chr17:7565280-7565335	55bp	UTR/Intronic

TABLE 3.1: Target bases not covered with designed HaloPlex HS probes. Only one reported mutation in COSMIC was found: a mutation in amino acid 901 of EGFR (COSM1550027). The base was excluded from testing.

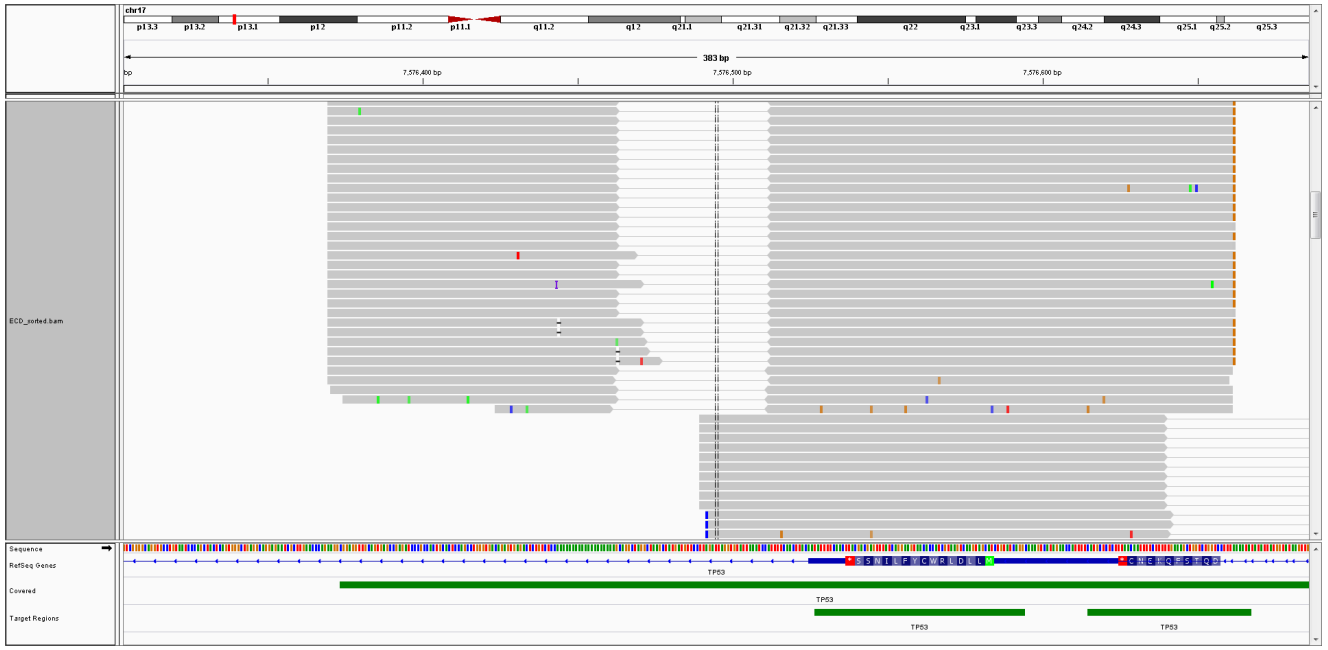


FIGURE 3.10: An example of an off-target probe that was designed. Shown is an SAM file prior de-duplication or error correction. Read pairs are connected with a line. Amplicons derived from probes with the same binding sites show the same start and end position. The forward reads align off-target, reverse reads span the target region.

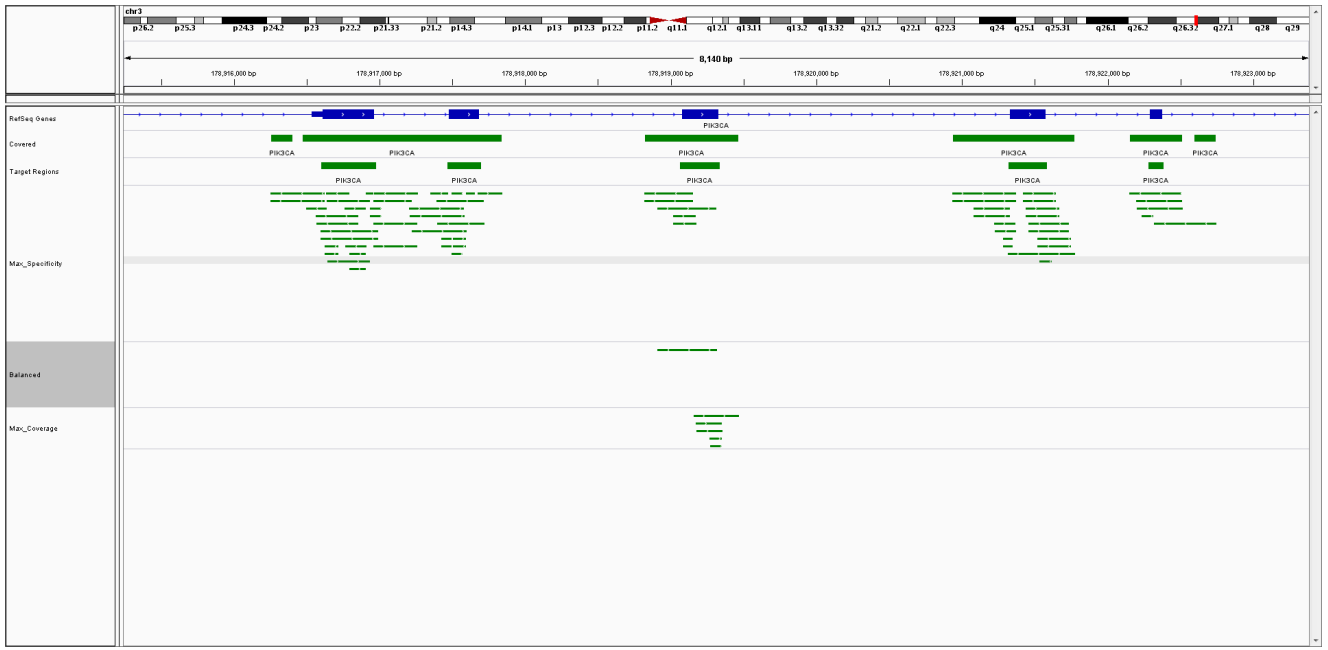
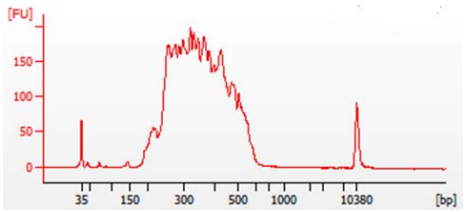
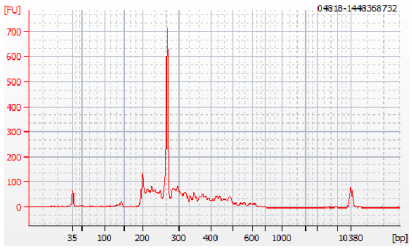


FIGURE 3.11: Designed HaloPlex HS probes spanning five exons of PIK3CA (tracks 1-3). Most bases are covered by probes with maximum specificity (track 4), a small fraction of additional probes designed are balanced (track 5) or maximum coverage (track 6).



(A) Excpected Bioanalyzer traces from an example Agilent HaloPlex HS library. Signal intensities (FU) are jagged, as size distribution of amplicons is determined by probe design.



(B) Bioanalyzer traces of a library enriched with custom designed probes. Target enrichment was performed on enrichment control DNA (ECD) for test purposes. A dominating peak around 280bp may indicate a problem with enrichment and/or amplification.

FIGURE 3.12: Bioanalyzer trace of an example HaloPlex HS library that is expected and a trace from a library produced by the custom HaloPlex HS custom design panel.

3.4 Performance Evaluation on 48 samples

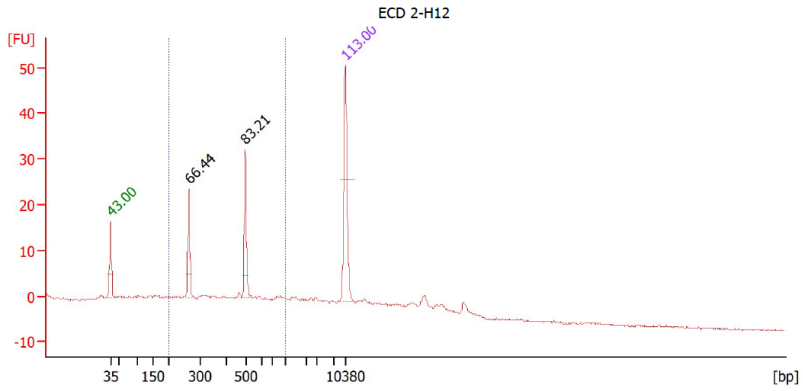
Panel efficiency and robustness against low input DNA was assessed by enriching and sequencing 48 clinical FFPE samples from various primary tumour tissue types, such as colorectal cancer or lung. In this section results from enrichment and library preparation, sequencing and subsequent data analysis will be presented and described. Analysis results are then compared to current assays estimating sensitivity and specificity of the panel for known mutation hotspots. In doing so, the samples were initially tested by pyrosequencing for known mutations in KRAS codons 12, 13 and 61; NRAS codons 12, 13 and 61; BRAF codon 600; EGFR codon 719, 858-861 and deletions in EGFR. To demonstrate results obtained from the panel are reproducible, some samples were processed and sequenced twice. Replicates follow an "_A", "_B" or "_1" "_2" notation at the end of every sample identifier.

In the last section results are discussed and potential improvements are described that may be considered to reach minimum standards required for clinical diagnostics.

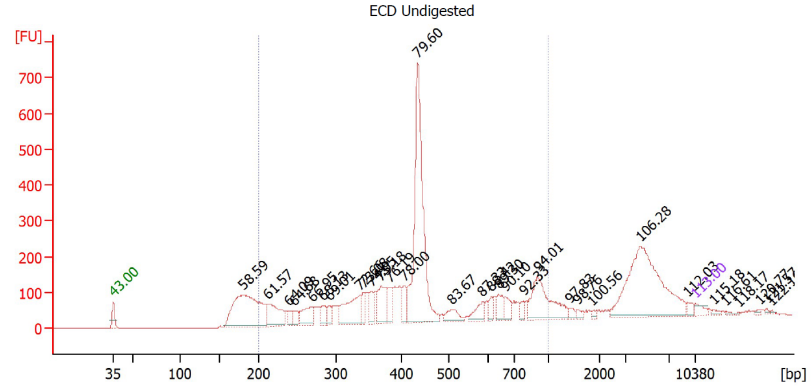
3.4.1 Target enrichment with Customised HaloPlex HS Panel

As mentioned before, there is very little control over input quality or quantity of clinical FFPE samples that are to be tested for genetic markers. Hence, in some cases, input quality and quantity criteria

recommended by manufacturer were barely met or even missed. Like with other target enrichment methods robustness against low input is, therefore, a necessity. If a test would fail on a majority of samples because of too strict input criteria, that diagnostic test is not of much use. Comprehensive DNA quality testing, as with an Agilent TapeStation 2200 instrument measuring DNA integrity, ideally, would require a sacrifice of an additional 100ng of DNA input, which may be better used by increasing input amount for the enrichment. Hence, instead of assessing DNA quality and subsequently risk a subsequent failure, extracted DNA was directly used for target enrichment, even at a higher concentration of up to 250ng, where possible. As previously described, DNA obtained from FFPE samples could be fragmented and carry artificial SNVs, due to treatment with formalin or suboptimal storage conditions after embedding. An increase of genetic material as input could help to compensate for some of these artefacts, seen as a high number of low-frequency mutation artefacts, a decrease in library complexity and a low on-target coverage. After extraction of DNA from tissue with Qiagen's QIAamp DNA FFPE Tissue Kit [190], concentrations of captured material were quantified by Qubit instrument for every sample, summarised in Table B.2. Figure 3.13 shows a representative Bioanalyzer trace of ECD DNA to check if enzymatic digestion of DNA was successful in order to avoid library DNA from being sacrificed from every sample. As expected, three peaks between 100bp and 500bp were identified indicating that the digestion worked



(A) Bioanalyzer trace from enzymatic digestion of control DNA (ECD).



(B) Raw input of ECD DNA, prior digestion.

FIGURE 3.13: Two representative bioanalyzer traces to control enzymatic digestion has worked. To save precious sample DNA, only ECD has been used.

as expected.

Probe hybridisation, target capture and enrichment was carried out as described above. Subsequently 25 rounds of PCR amplified the captured targets ligated to rings, as specified by manufacturer. The

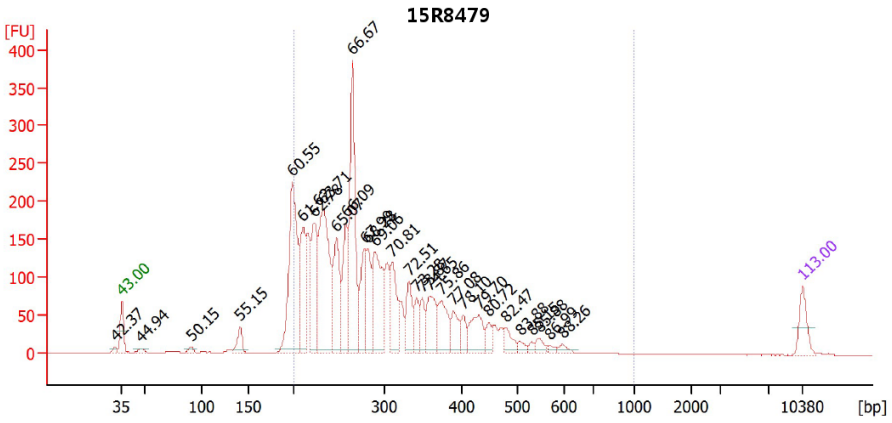


FIGURE 3.14: A Bioanalyzer trace from an captured and amplified library (15R8479_2) enriched with custom designed HaloPlex HS panel. Library shows an expected spike and a jagged curve of size distributions between 200bp and 500bp.

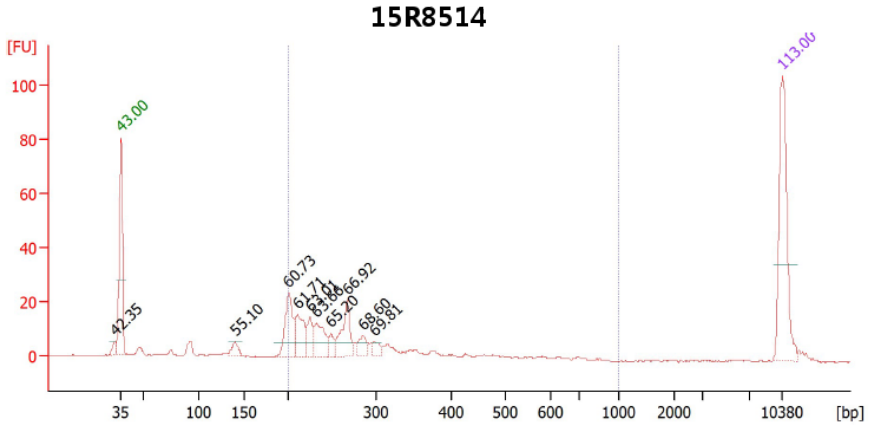


FIGURE 3.15: Representative example of a Bioanalyzer trace enriched from low concentrated DNA (15R8514). Signal intensities are lower and some expected peaks are not seen at all.

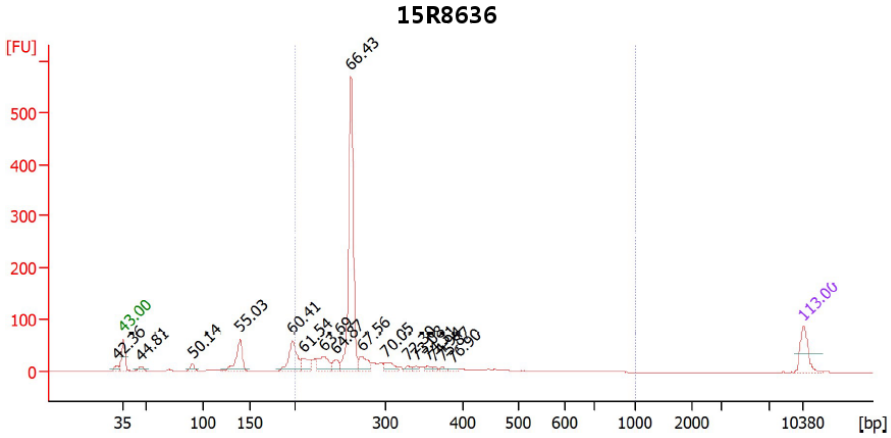


FIGURE 3.16: Bioanalyzer trace showing almost exclusively the typical spike at 280bp (15R8636), this may indicate a problem with the enrichment.

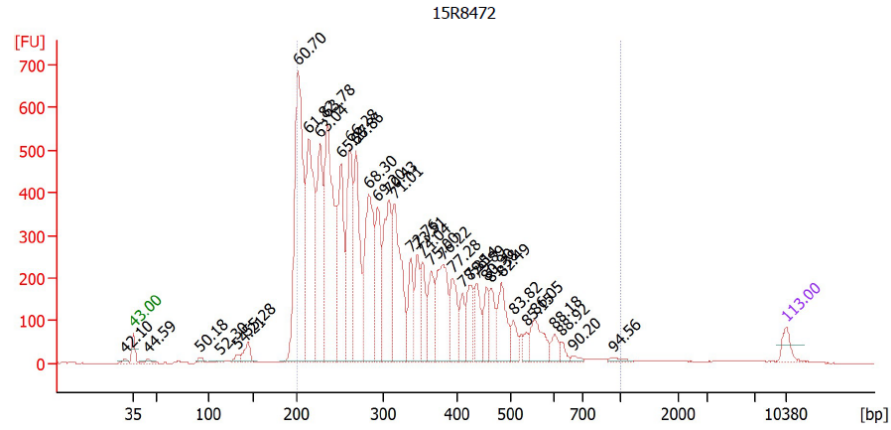
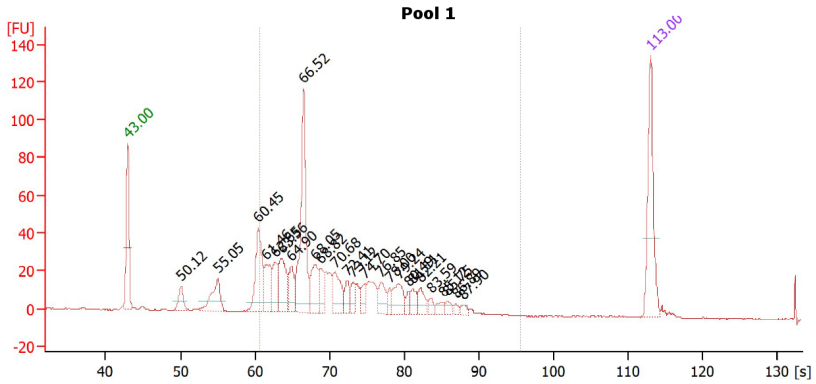
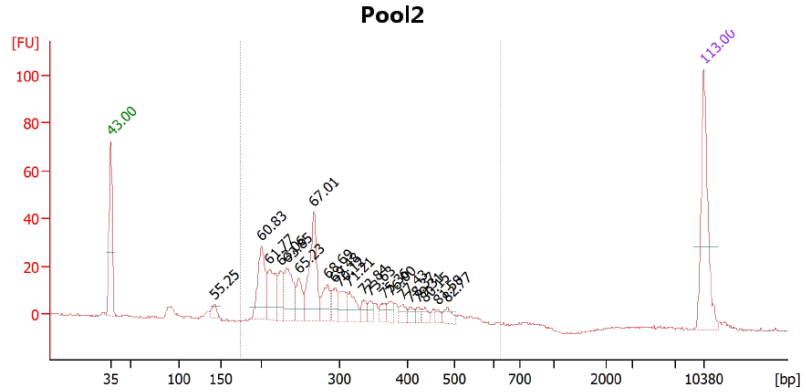


FIGURE 3.17: Bioanalyzer trace of a sample (15R8472) showing the typical jagged signal intensity curve, but no dominating spike at 280bp. The signal intensities are very high in general. It is believed that the spike is probably lost in the strong signals.



(A) Pooled libraries from the first 24 samples. Agilent Bioanalyzer software was unable to identify the upper marker in another sample of the chip, hence fragment size was not converted in bp.



(B) Pooled libraries from the second 24 samples.

FIGURE 3.18: Bioanalyzer traces of the pooled libraries. Expected jagged curve and spike can be seen by eye.

number of PCR cycles was dependant on the number of probes designed, as a larger number of probes required fewer PCR cycles to avoid unwanted amplification artefacts. Enriched and amplified libraries were loaded onto a Bioanalyzer instrument. A selection of representative electropherograms is shown in Figures 3.14 - 3.17. Libraries were pooled in an equimolar concententration for most cases. Libraries that were insufficiently concentrated were nonetheless sequenced, but were flagged throughout the analysis. Figure 3.18 shows Bioanalyzer traces of both pooled libraries prior sequencing of 24 prepared samples each. Sequencing was performed by two MiSeq runs, both configured as a 150bp paired-end sequencing with dual-indexing. Probes were designed to have a desired distance of at least 300bp for sequencing with a 150bp paired-end run. As the bioanalyzer traces imply, some generated fragments were shorter resulting in reads to overlap. The run length was a trade-off between high yield and challenges in the probe design. First index was configured as an eight base-pair multiplexing index, shown in Table B.2. The second index (i5) was the unique 10bp molecular barcode for each enriched fragment, which is not used for demultiplexing. Instead the instrument was configured to sequence the two indexes, but only use the first index for demultiplexing, while the second index was written to a dictionary FASTQ file. The method of generating the dictionary files deviated slightly from manufacturer's protocol, because the bcl2fastq algorithm was unable to generate a sample-specific dictionary file. Instead bcl2fastq always tried to use

both indexes (i7,i5) for demultiplexing, which obviosuly failed. Hence, a run-specific dictionary file was created; the commands are listed in Section B.3. Due to the fact that sequencing reads always have a unique ID, no information was lost or changed during this step.

The number of reads survived after trimming was plotted for every sample, as shown in Figure 3.19. All samples showed a sufficient amount of reads sequenced, even one sample (15R8476) showing a decreased number of reads. Sequencing yield of the first pool was lower than of pool 2. In addition, some samples were probably not mixed equimolarly. An uneven pooling was mainly caused by technical bias. Quantification with a Qubit instrument could be skewed [85] and pipette inaccuracies could cause a noticeable bias as well. In addition, during bridge amplification shorter fragments may hybridised faster to the adapter sequences on the flow-cell than longer fragments giving them an advantage for an improved amplification. As composition of fragments differed from library to library, sequencing yield varied. Due to the increase in sequencing,however, a re-sequencing of individual samples could be evaded.

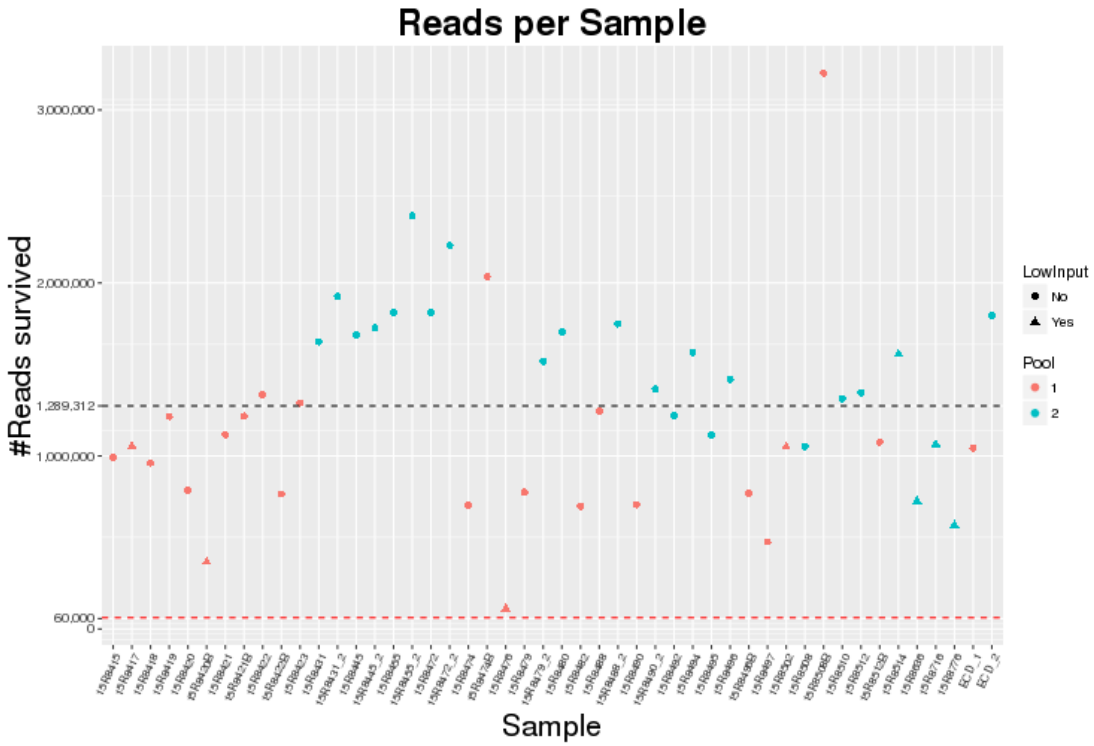


FIGURE 3.19: Sequencing yield of samples in a logarithmic plot. Only reads that survived trimming were taken into account. Dashed black line indicates average coverage per sample, dashed red line indicates minimum number of reads per sample recommended by manufacturer.

3.4.2 Bioinformatics Analysis

Data from all 48 samples was demultiplexed and converted into two raw FASTQ files per sample, one for the forward and one for the reverse read, commands listed in Section B.3. Data was trimmed and aligned with Skewer using parameters listed in Table 2.3. In addition first base (cytosine) of the reverse read (read 2) was trimmed, as it was not derived from genomic sequence, but an artefact from the ligation. Keeping this base would have introduced a false-positive mismatch at that position, as shown in Figure 3.20.

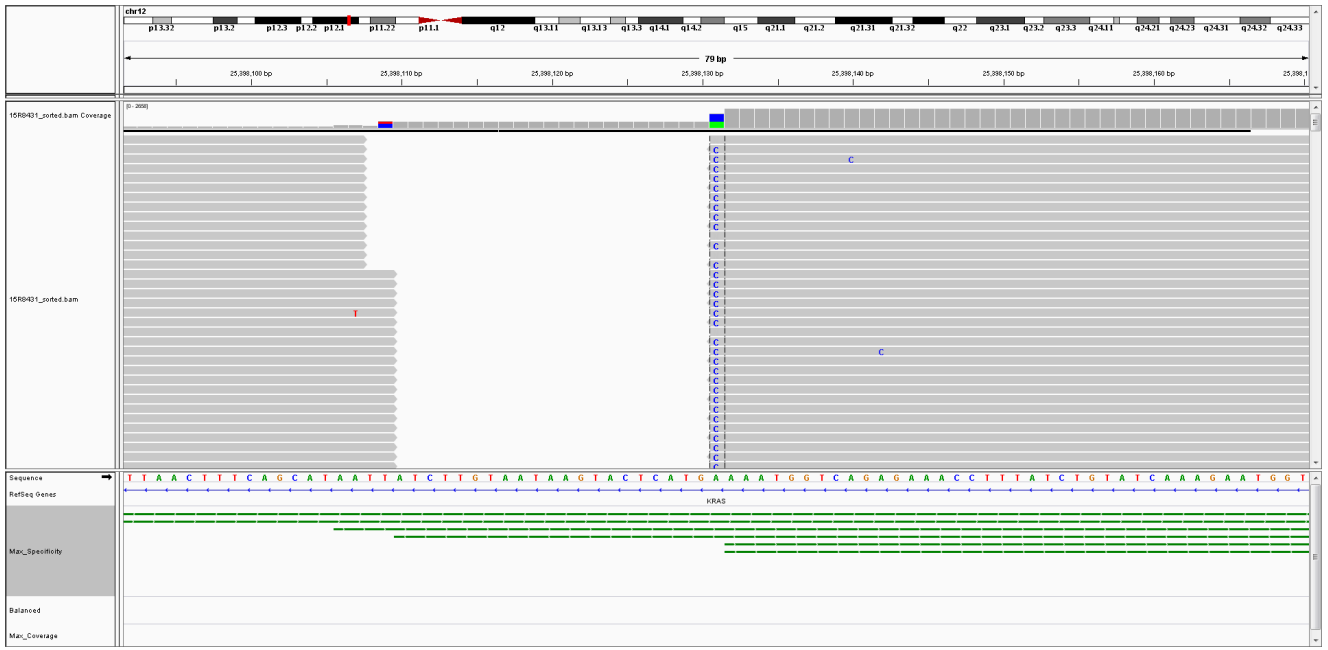


FIGURE 3.20: Agilent HaloPlex HS bases carry a cytosine residue from the ligation step in the first base of the reverse read. If this base is not trimmed, reads can still be aligned, but a false mismatch is introduced potentially causing a false-positive SNV returned by the variant caller.

Data was aligned with BWA using parameters given in Table A.6. Library de-duplication was performed by utilising the introduced molecular barcode, sequenced as i5 index. AgilentMBCDedup, a proprietary software provided by Agilent was used, command is listed in Section B.4 [225]. It collapsed duplicates and corrected sequencing errors from a provided SAM and dictionary files. In addition it removed reads from the SAM file that aligned off-target. Unfortunately the software removed reads from the alignment file entirely rather than flagging them. Hence, it would have been cumbersome to distinct between an overall duplication rate and the number of reads that aligned off-target, because reads aligning off-target were still riddled with duplicates. Instead, the following metric was used:

$$R = S - E, \tag{3.1}$$

where R indicates the number of reads remained after collapsing and off-target removal, S the number of reads survived the trimming and E all reads that were removed, because they were either duplicates or aligned off-target. Analogously to the percentage of duplicates, the percentage of remaining reads is defined as

$$R_{\text{Rate}} = \frac{R}{S} \tag{3.2}$$

Figure 3.21 shows the number of raw reads that survived trimming against the number of remaining reads per sample. As with enrichment methods based on random shearing, the unique on-target rate varied from sample to sample. On average around 32% of the reads remained, but in two cases (15R8476, 15R8514) almost 98% of the reads were removed by AgilentMBCDedup. The reasons for a high number of duplicates and off-target mappings were believed to be a combination of poor DNA integrity and low input quantity. Samples with low input showed a high rate of reads remaining in some cases, but on the other hand reads with sufficient input lost a relatively high number of reads during read collapsing -sometimes up to 85%. Like with other enrichment methods a high quality of input material can compensate for a low input quantity. Due to the very limited amount of genetic material that was available, DNA integrity could not be measured for any further analysis.

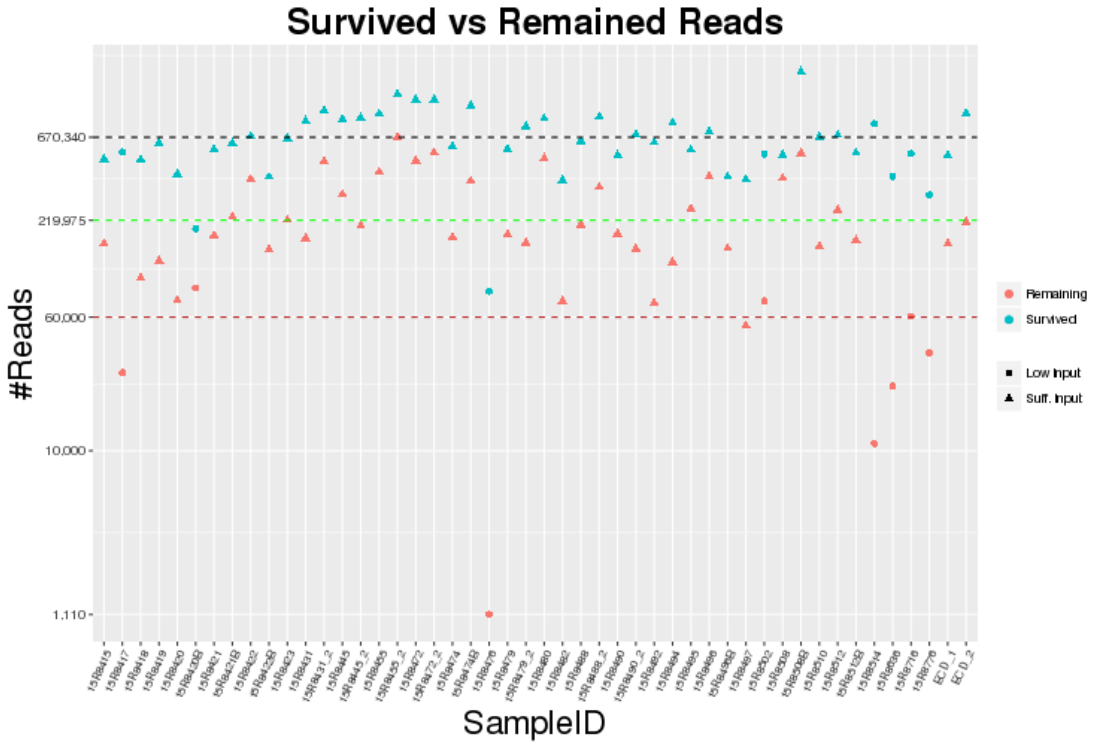


FIGURE 3.21: Number of survived reads against the remaining reads per sample on a logarithmic scale. Dashed black line indicates mean number of reads survived trimming, dashed green line mean number of reads remained after AgilentMBCDedup was applied, dashed red line indicates minimum sequencing recommended.

Although most samples showed a sufficient number of remaining reads consulting the minimum number of reads defined by Agilent, it had to be evaluated of coverage was sufficient for subsequent data analysis. Just like with other methods, probes may work at different efficiencies depending on numerous factors. Hence, probe performance was further investigated along all target genes, as shown in Figure 3.22 and Section C. For every probe a boxplot was drawn. If all probes had performed equally well, no increase or decrease of mean coverage would be observed. This was not the case, as probes performed at different enrichment efficiencies. One factor would be the mentioned probe specificity, but also structural properties of the probes may change their hybridisation efficiency. In addition, some regions showed a drop in coverage, because binding sites of some of the probes fell into regions of a SNP or mutation drastically decreasing binding capabilities of the probes. Hence, it was important to check the per-probe coverage in every gene prior variant calling to see if any regions could not be tested. A failure of a limited number of probes was not considered to be critical as long as regions showed sufficient coverage in general.

Variants were called with VarScan2 on the collapsed reads on every sample individually using parameters given in Section A.6. No sample was excluded, to assess limitations of the panel. Table 3.2 lists mutations that were identified with pyrosequencing from 48 samples that were tested, including technical replicates and compares it to the sequencing results. In 8 out 48 samples there was a disagreement in the results

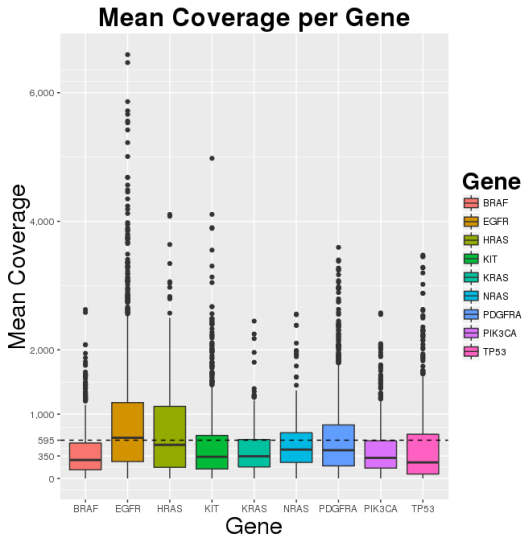


FIGURE 3.22: Coverage boxplot per gene. The dashed black line indicates the average coverage across all genes and probes.

between pyrosequencing and the data obtained from the HaloPlex HS panel. All missed mutations were in KRAS codon 12, which indicated an underlying problem for that region. By looking at samples that were enriched and sequenced twice, it was observed that the mutation was also missed in the replicate indicating the problem was real not due to random noise. Moreover, all tumour burden was above 50% in all cases and variant allele frequencies were above 10%, hence they should have been identified given sufficient coverage in that region.

Figure 3.23 shows the number of remaining reads per sample with those highlighted that missed a mutation. Although a few samples with a low number of input reads missed a mutation in KRAS, the majority of

samples showed a sufficient yield, even if the library was enriched with less extracted DNA than recommended. Hence, the cause of problem was not explained by insufficient sequencing. This was also supported by the fact that coverage was not dropping unusually in the target regions, where the mutation was missed compared to a sample where the mutation was correctly identified, as shown in Figure 3.24, where coverage is plotted along KRAS exon 2 of samples missing a mutation together with a sample where the p.(Gly12Asp) amino acid change in KRAS was correctly identified. Coverage dropped in that region in all samples, due to probe design, but coverage was always sufficient for variant calling of more than 10% mutation allele frequency.

As a next step, pileup files were compared to see if the variant was potentially missed by VarScan2, a selection of pileups is listed in Table 3.3. It was observed that the mutated allele was not present in the sequencing data, or only at a very low frequency ($< 1\%$). In conjunction with coverage information, it was finally excluded that this is caused by a misaligned or due to a false negative by VarScan2. Further, from the pileup it was noticed that all reads in that region were obtained from the same strand, which was caused by probe design in conjunction with a setting in Samtools, as shown in Figure 3.25. Half of the probes were designed such that the insert size was shorter than one read length, causing reads to overlap. By default Samtools ignores reads derived from fragments with an insert size shorter than twice the read length. Anomalous read pairs are usually not considered, because they can bias

allele frequencies, as every captured fragment is counted twice where reads overlap. But even when switching off the anomalous read filter, shown in Section A.6, the mutation was not or barely present in any of the reads. Although a biased presentation of the alleles was reported before, known as strand bias [160, 161, 177], it was unusual that always the same region in KRAS was affected, but not in all samples.

One possible explanation would be a heterozygous mutation or InDel being present in the binding sites of the probes. This would have caused a probe to preferably bind to the non-mutated allele and support a biased enrichment. In conjunction with a low number of probes designed, only the reference allele would have been amplified in these cases. Unfortunately, there was not enough genetic material left, to test for this hypothesis directly, as there were a number of known SNPs and mutations falling into the first 20 base-pairs of both amplicon sites, reported by dbSNP and COSMIC. The chances that all probes were affected in more than a few samples, however, was too low. In addition SNVs would have been reported in cases where the binding sites were covered by other probes. Since that was not the case, this hypothesis was dismissed.

Another reason could be degenerated DNA. Unlike long capture probes that potentially bind even to short fragments of DNA, molecular inversion probes have two binding sites and are enrich DNA fragments shorter than the distance between both sites, as shown in Figures 3.26 -

3.29. These types of artefacts are typical for amplicon sequencing and cannot be overcome by HaloPlex HS [232]. Table 3.4 summarises the amplicons produced from probes spanning the locus chr12:25,398,284. It was shown before that mutation detection with target resequencing based on HaloPlex systems can cause mutations to be missed in cases where binding sites of probes were too far apart from each other and DNA was partially degenerated, like DNA extracted from FFPE tissue [226]. In the cases of low DNA integrity in samples where a p.(Gly12Asp) mutation in KRAS was missed, only one probe with a short distance between both binding sites (134bp) was left for enrichment. In combination with the chosen read length, half of derived reads were produced from the same fragment, the number of unique fragments was further reduced by 50%, causing a biased enrichment. Possible solutions described suggest a redesign of the panel to reduce insert size below 250bp and limit sequnecing read length to 100bp paired-end. There was, however, no guarantee that a new design would allow unbiased target enrichment in all cases. Hence, target enrichment based on barcoded molecular inversion probes does not seem to be generally eligible for clinical diagnostics on FFPE samples.

Sample	KRAS		NRAS		BRAF		EGFR	
	Pyro	Halo-Plex	Pyro	Halo-Plex	Pyro	Halo-Plex	Pyro	Halo-Plex
15R8415	WT	WT	Gln61 Lys	Gln61 Lys	NA	WT	NA	WT
15R8417	WT	WT	WT	WT	NA	WT	NA	WT
15R8418	WT	WT	Gly12 Cys	Gly12 Cys	NA	WT	NA	WT
15R8419	WT	WT	WT	WT	NA	Val600 Glu	NA	WT
15R8420	Gly12 Val	WT	NA	WT	NA	WT	NA	WT
15R8420B	Gly12 Val	WT	NA	WT	NA	WT	NA	WT

15R8421	WT	WT	WT	WT	NA	WT	NA	WT
15R8421B	WT	WT	WT	WT	NA	WT	NA	WT
15R8422	WT	WT	WT	WT	NA	WT	NA	WT
15R8422B	WT	WT	WT	WT	NA	WT	NA	WT
15R8423	Gly12 Asp	Gly12 Asp	NA	WT	NA	WT	WT	WT
15R8474	NA	Gly12 Asp	NA	WT	NA	WT	WT	WT
15R8474B	NA	Gly12 Asp	NA	WT	NA	WT	WT	WT
15R8476	NA	WT	NA	WT	NA	WT	WT	WT
15R8479	Gly12 Ala	Gly12 Ala	NA	WT	WT	WT	NA	WT

15R8482	NA	WT	NA	WT	NA	WT	Exon 19 Del	Exon 19 Del
15R8488	Gly12 Asp	Gly12 Asp	NA	WT	NA	WT	NA	WT
15R8490	Gly12 Val	WT	NA	WT	WT	WT	NA	WT
15R8496B	Gly12 Ser	Gly12 Ser	NA	WT	WT	WT	NA	WT
15R8497	NA	WT	NA	WT	NA	WT	WT	WT
15R8502	NA	WT	NA	WT	NA	WT	WT	WT
15R8508B	NA	WT	NA	WT	NA	WT	WT	WT
15R8512B	Gly12 Val	Gly12 Val	NA	WT	NA	WT	NA	WT

ECD_1	NA	WT	NA	WT	NA	WT	NA	WT
15R8431	Gly12 Val	WT	NA	WT	WT	WT	NA	WT
15R8431_2	Gly12 Val	WT	NA	WT	WT	WT	NA	WT
15R8445	WT	WT	WT	WT	NA	WT	NA	WT
15R8445_2	WT	WT	WT	WT	NA	WT	NA	WT
15R8455	NA	WT	NA	WT	NA	WT	WT	WT
15R8455_2	NA	WT	NA	WT	NA	WT	WT	WT
15R8472	NA	WT	NA	WT	NA	WT	WT	WT
15R8472_2	NA	WT	NA	WT	NA	WT	WT	WT
15R8479_2	Gly12 Ala	Gly12 Ala	NA	WT	WT	WT	NA	WT

15R8480	NA	WT	NA	WT	NA	WT	WT	WT
15R8488_2	Gly12 Asp	Gly12 Asp	NA	WT	NA	WT	NA	WT
15R8490_2	Gly12 Val	WT	NA	WT	WT	WT	NA	WT
15R8492	Gly12 Val	Gly12 Val	NA	WT	WT	WT	NA	WT
15R8494	WT	WT	WT	WT	Val600 Glu	Val600 Glu	NA	WT
15R8495	Gly12 Val	WT	NA	WT	WT	WT	NA	WT
15R8496	Gly12 Ser	Gly12 Ser	NA	WT	WT	WT	NA	WT
15R8508	NA	WT	NA	WT	NA	WT	WT	WT

15R8510	WT	WT	WT	WT	NA	WT	NA	WT
15R8512	Gly12 Val	Gly12 Val	NA	WT	NA	WT	NA	WT
15R8514	Gly12 Val	WT	NA	WT	NA	WT	NA	WT
15R8636	NA	WT	NA	WT	NA	WT	WT	WT
15R8716	NA	p.Gly12 Ala	NA	WT	NA	WT	WT	WT
15R8776	Gly12 Val	WT	NA	WT	NA	WT	NA	WT
ECD_2	NA	WT	NA	WT	NA	WT	NA	WT

Table 3.2: Comparison of HaloPlex HS (HaloPlex) results and Pyrosequencing (Pyro) for KRAS codons 12, 13, 61; NRAS codons 12, 13, 61; BRAF codon 600, deletions (Del) in EGFR or EGFR codons 719, 858-861. Result is either not tested (NA), no findings (WT) or shows the mutation in HGVS.p notation. A red cell indicates a disagreement/missed mutation.

Sample	Bur-den	Cov.	Pileup
15R8420	60%	116	....
158420B	60%	87	....
15R8431	60%	73	....
15R8431_2 (15R8431B)	60%	168	....
15R8423 (Control)	40%	90	T.....TT.....T.....T..T.....TT.....T ....T..T..T.T....TT..T.T.....TT....T..T.....

TABLE 3.3: Pileup files of locus chr12:25,398,284 (ref.: cytosine) and estimated tumour burden of samples where the mutation was missed. All bases that passed the internal Samtools filter were derived only from the forward strand, as reads with anomalous insert sizes were excluded, hence a all reads showed a reduced coverage, as one of the amplicons have an insert size smaller than twice the read length, shown in Figure 3.25.

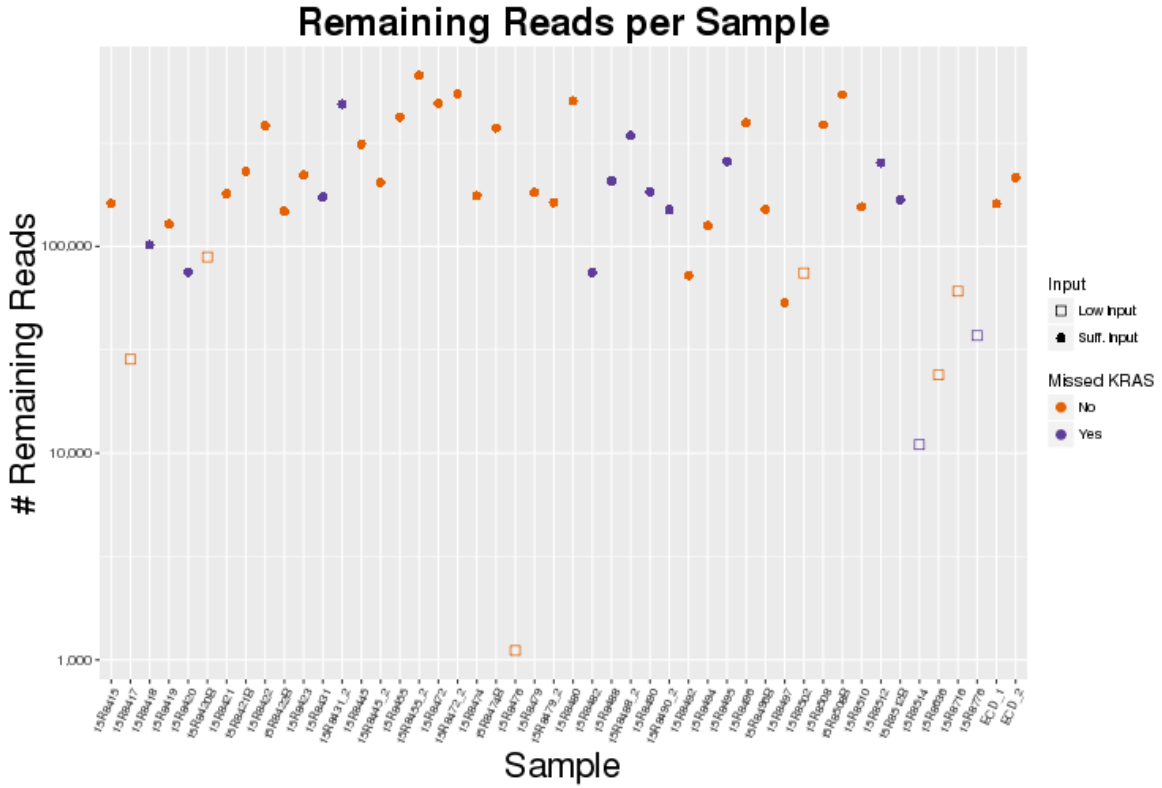


Figure 3.23: Scatterplot of remaining reads per sample on a logarithmic scale. Samples flagged as low input are plotted as squares. Samples with missed mutation in KRAS are drawn purple.

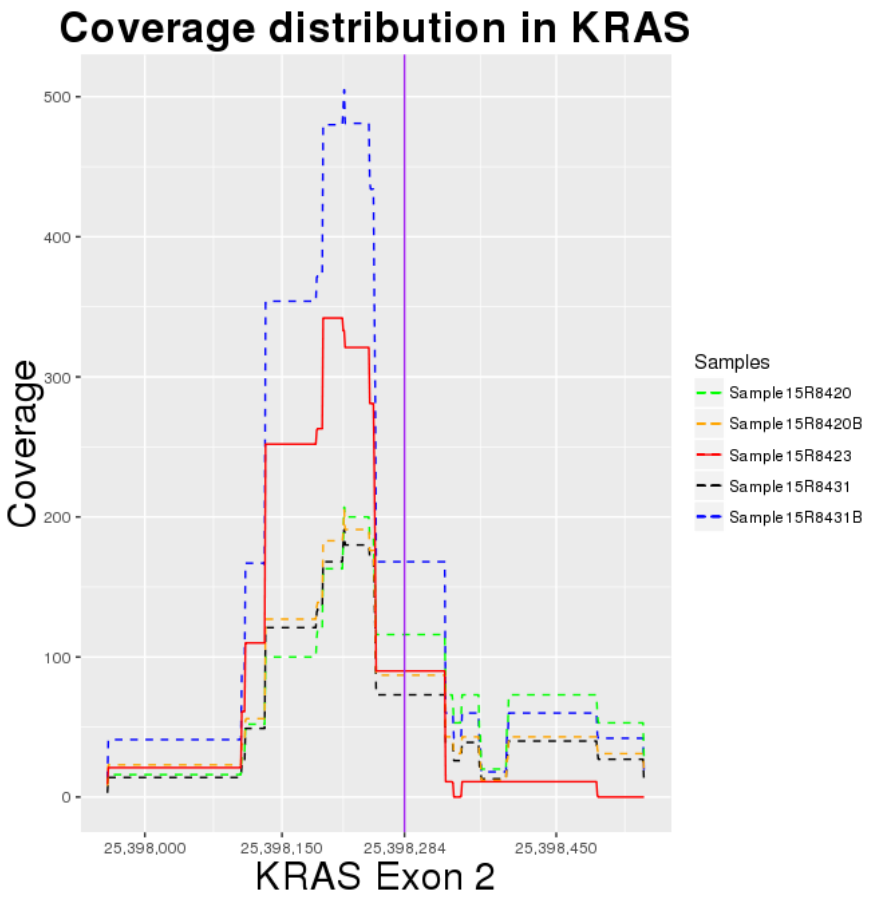


FIGURE 3.24: coverage distribution of KRAS. Coverage distribution along the exon follows a shape of a "Manhatten Skyline", caused by probe design. Although coverage generally drops around the mutation hotspot (purple line), a mutation was detected in the red control sample (15R8423).

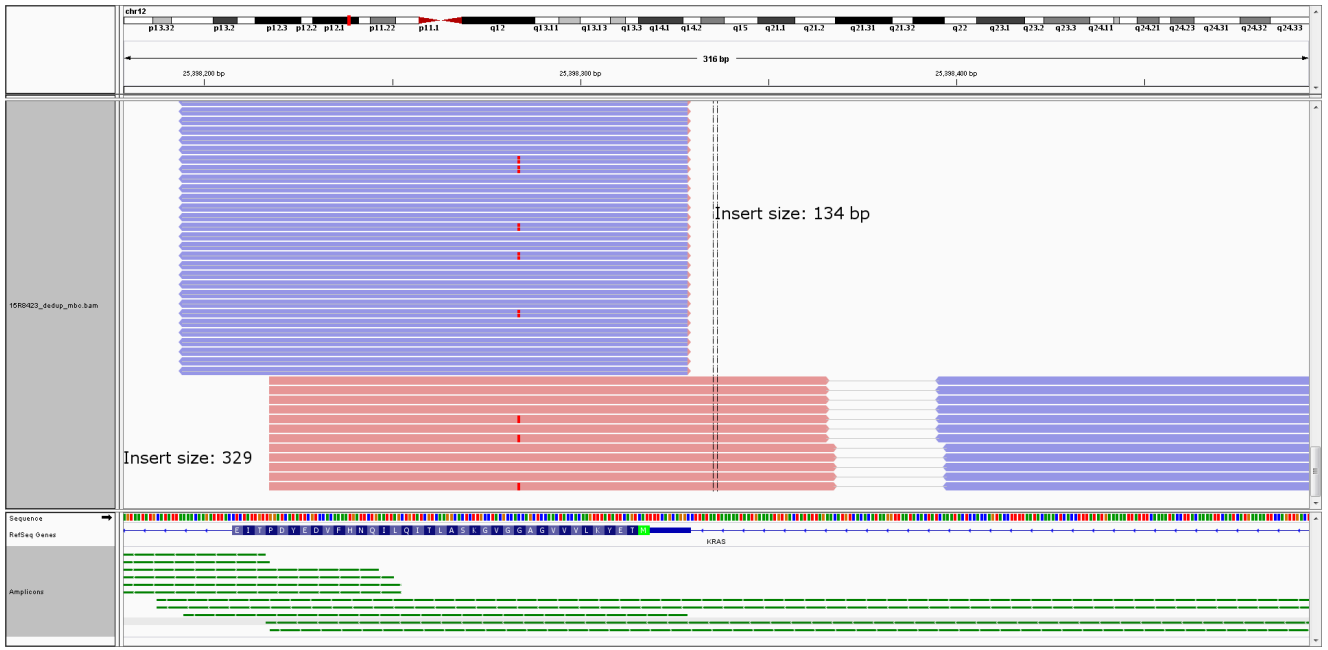


FIGURE 3.25: A probe in KRAS generated amplicons that were shorter than read length. Samtools ignores such read pairs by default, as it can cause a bias in allele frequencies leaving only reads derived from the forward strand (red).

FIGURE 3.26: Intact DNA (blue) with a heterozygous SNV (red asterisk).

FIGURE 3.27: After enzymatic digestion, probes (grey-blue) target regions of interest for capture and enrichment.

FIGURE 3.28: Highly fragmented or degenerated DNA (blue) with the same heterozygous SNV (red asterisk).

FIGURE 3.29: After enzymatic digestion probes (grey-blue) cannot bind to fragments smaller than their distance between two binding sites, causing an amplification bias.

Probe ID	**Covering region**	**Dist.**
764770_001714	chr12:25398187-25398495 (+)	308
764770_001715	chr12:25398187-25398495 (-)	308
764770_001716	chr12:25398194-25398328 (-)	134
764770_001717	chr12:25398216-25398545 (-)	329
764770_001718	chr12:25398217-25398546 (+)	329

TABLE 3.4: Listed probes spanning locus chr12:25,398,284. 4 out of 5 probes are above 300bp. Although the resulting amplicons are desired for a 150bp read length, they fail binding to DNA fragments below that size.

3.5 Discussion

Multi-gene testing for many sites parallel of genes with diagnostic and prognostic potential in various tumour types is not a trivial method, since sequencing entire genes is an hypothesis-free approach, i.e. every detection is reported leaving clinical impact unclear in most cases [172]. Even by reducing the number of genes to be tested still tens of thousands of bases are tested for. Validation of every single site is, therefore, currently technically impossible. Within the scope of this work, validation was limited to a few codons in a few genes by using a customised HaloPlex HS target enrichment panel based on barcoded molecular inversion probes on 48 clinical samples that were FFPE tissues from various primary tumour types that were previously tested in sites to be of clinical relevance by a number of validated pyrosequencing assays. It turned out that the panel missed mutations in codon 12 in KRAS in, probably caused by highly fragmented DNA obtained from poorly preserved FFPE samples, causing the validation to fail. Although a redesign of the probes was considered to reduce distance between both binding sites of the probes to make it more robust against high degeneration, it was decided that molecular inversion probes are not suitable for clinical diagnostics, as the restrictions of sample qualities were too strong, as the risk of a wrong test result is too high. Even after a complete redesign, probes could potentially cause

a biased enrichment, which is hard to exclude for all sample qualities and regions.

References

[215] Carol Beadling et al. "Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping". In: *The Journal of Molecular Diagnostics* 15.2 (2013), pp. 171–176.

[216] PD Da Forno et al. "BRAF, NRAS and HRAS mutations in spitzoid tumours and their possible pathogenetic significance". In: *British Journal of Dermatology* 161.2 (2009), pp. 364–372.

[217] *GeneRead™DNAseq Targeted Panels V2 Handbook*. V2. For targeted enrichment prior to next-generation sequencing. QIAGEN. June 2015.

[218] Phillip N Gray, Charles LM Dunlop, and Aaron M Elliott. "Not All Next Generation Sequencing Diagnostics are Created Equal: Understanding the Nuances of Solid Tumor Assay Design for Somatic Mutation Detection". In: *Cancers* 7.3 (2015), pp. 1313–1332.

[160] Yan Guo et al. "Exome sequencing generates high quality data in non-target regions". In: *BMC genomics* 13.1 (2012), p. 1.

[161] Yan Guo et al. "The effect of strand bias in Illumina short-read sequencing data". In: *BMC genomics* 13.1 (2012), p. 1.

[219] *HaloPlex HS Target Enrichment System For Illumina Sequencing Protocol*. B0. Copyright by Agilent Technologies. Agilent. June 2015.

[220] Paul Hardenbol et al. "Multiplexed genotyping with sequence-tagged molecular inversion probes". In: *Nature biotechnology* 21.6 (2003), pp. 673–678.

[221] Jennifer Harrow et al. "GENCODE: the reference human genome annotation for The ENCODE Project". In: *Genome research* 22.9 (2012), pp. 1760–1774.

[222] Joseph B Hiatt et al. "Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation". In: *Genome research* 23.5 (2013), pp. 843–854.

[223] T Hubbard et al. "The Ensembl genome database project". In: *Nucleic acids research* 30.1 (2002), pp. 38–41.

[224] Donna Karolchik et al. "The UCSC genome browser database". In: *Nucleic acids research* 31.1 (2003), pp. 51–54.

[172] Dmitriy Khodakov, Chunyan Wang, and David Yu Zhang. "Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches". In: *Advanced drug delivery reviews* (2016).

[177] Stefan H Lelieveld et al. "Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions". In: *Human mutation* 36.8 (2015), pp. 815–822.

[225] *MBCDeduplication Read Me File*. Version 1.0. Copyright by Agilent Technologies Inc. Agilent. 2015.

[226] Lotte NJ Moens et al. "HaloPlex Targeted Resequencing for Mutation Detection in Clinical Formalin-Fixed, Paraffin-Embedded Tumor Samples". In: *The Journal of Molecular Diagnostics* 17.6 (2015), pp. 729–739.

[85] Yuki Nakayama et al. "Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions". In: *PloS one* 11.3 (2016), e0150528.

[227] Hans Prenen, Sabine Tejpar, and Eric Van Cutsem. "New strategies for treatment of KRAS mutant metastatic colorectal cancer". In: *Clinical Cancer Research* 16.11 (2010), pp. 2921–2926.

[228] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 35.suppl 1 (2007), pp. D61–D65.

[229] Kim D Pruitt et al. "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". In: *Genome research* 19.7 (2009), pp. 1316–1323.

[190] *QIAamp DNA FFPE Tissue Handbook*. 06/2012. Qiagen. June 2012.

[230] Ana I Robles and Curtis C Harris. "Clinical outcomes and correlates of TP53 mutations and cancer". In: *Cold Spring Harbor perspectives in biology* 2.3 (2010), a001016.

[109] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[231] Laurens G Wilming et al. "The vertebrate genome annotation (Vega) database". In: *Nucleic acids research* 36.suppl 1 (2008), pp. D753–D760.

[232] Stephen Q Wong et al. "Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing". In: *BMC medical genomics* 7.1 (2014), p. 1.

[233] J Xiao et al. "Association between urothelial carcinoma after kidney transplantation and aristolochic acid exposure: the potential role of aristolochic acid in HRas and TP53 gene mutations". In: *Transplantation proceedings*. Vol. 43. 10. Elsevier. 2011, pp. 3751–3754.

Chapter 4

Cellular Barcoding Followed by Massively Parallel Sequencing

Mutation detection on heterogeneous cell populations, such as cancer, with massively parallel sequencing is achieved by measuring signal intensities from millions of independent short fragments. Other systems, such as Oxford Nanopore or PacBio have increased this read length to tens of thousands of bases to allow sequencing up to entire molecules [99, 234]. These long reads allow prediction of linkage between SNPs or mutations in a cell per sample depending on read or molecule length. Single-cell sequencing approaches go one step further by sequencing

the genome, exome or transcriptome from a single cell [236]. Recently published methods are either based on separation of individual cells, from which DNA or RNA is extracted, amplified and sequenced or technologies utilising microfluidics, where a cell is encapsulated in a droplet and subsequently processed [259, 238].

In this chapter a new method will be introduced to allow hundreds to thousands of cells to be sequenced within a few hours of preparation time by *direct emulsion PCR* (demPCR). Emulsion PCR is commonly used for pyrosequencing and for amplification of complex gene libraries [258, 240], whilst direct PCR or colony PCR is often used for amplification of genetic material without prior DNA extraction [248, 249].

After a brief motivation, the protocol of cellular barcoding is described and the results of the sequencing analysis of a mixed cell population of K562 and NIH3T3 cells are presented and discussed.

4.1 Introduction

The understanding of evolution and the heterogeneity in cancer is of great interest to understand the development and progress of a disease over time. Genetic complexity of a tumour has been targeted for many years, but since the introduction of massively parallel sequencing technologies, single cell sequencing has become very popular among researchers [87]. Although impressive instruments for these tasks are

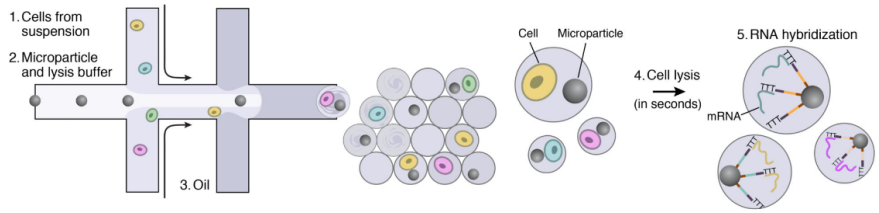
available, the number of cells that can be analysed is still limited to a few thousand cells at a time, due to cell separation and processing time needed per cell [235]. Further single cell RNA sequencing is often reduced to sequencing 5'- or 3'-ends of every transcript in a cell, to reduce the amount of sequencing per cell necessary [257]. Barcoding single cells was previously performed for RNA sequencing, which uses the facts that mRNAs are relatively short and provided with a poly-A tail allowing entire transcriptomes to be captured and subsequently barcoded and amplified. Further the amount of RNA present in a single cell is much higher than DNA [246]. Recently, transcripts of thousands of cells were barcoded and sequenced by capturing each droplet with a bead of poly-T oligonucleotide primers, shown in Figure 4.1 [244]. Analysing and comparing over 44,00 single transcriptomes of mouse retinal cells revealed numerous potentially new cell types with different gene expression patterns, indicating a higher cell heterogeneity in healthy organisms than previously expected. Due to genetic instability in cancer it can be expected that cell diversity is even larger, as subclonal driver mutations move a tumour constantly forward, especially during treatment. Subclonal changes and complexity in cancer were targeted with ultra-deep sequencing of exomes [243] or by sequencing bulk tissue from several regions from a tumour in a single individual [237]. It revealed not only clonal evolution, but also large tumour heterogeneity. By tracing mutations in single cancer stem cells, progression of resistance evolution of a tumour to therapies was documented [250].

Due to the vast potential that is provided by single cell sequencing, it would be a powerful tool to be utilised for personalised medicine.

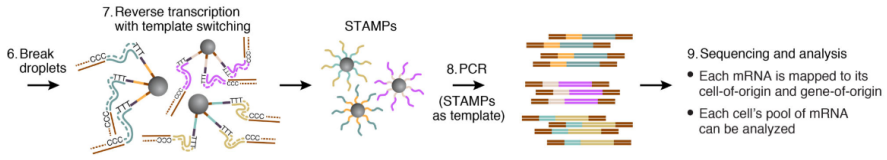
The method described here, tries to overcome current limitations in single cell sequencing, such as cost, turnaround time and DNA amplification directly from single cancer cells in microdroplets based on emPCR. The concept has proven very useful by generating billions of little droplets being available as microreactors containing only a few nanoliters of reaction volume that is encapsulated. Microfludics approaches keep full control over ever single droplet meaning cells are transferred into droplets, lysed, amplified one by one and finally sequenced [251]. Hence, after a cell was lysed, neutralisation buffers could be added. This option is not available when giving up control over every reaction, as emulsion PCR randomly encapsulates cells and once the emulsion was formed, no further modifications can be made. DNA from eukaryotes is compacted and bound to histones and cell compartments provide a poor environment and conditions for a polymerases meaning a direct amplification has proven difficult [252]. Within the scope of this work a novel polymerase provided by Clontech was used. It allowed direct amplification of cells without prior lysis or mechanical treatment even when isolated in droplets. Uniquely barcoded primers for target amplification, barcoding and library preparation were anchored on a microparticle, named *bead*. The protocol consists of the following major steps: first, to obtain a large number of beads, each bead had to be filled with a collection of oligonucleotides carrying a

bead-specific barcode, which differed to all other beads, th sequencing adapter and a target-specific primer. Second, the emulsification of bead and cell in the same microreactor. Third, cell lysis within a droplet. Fourth, barcoding of cellular DNA, library preparation and amplification. Fifth, breaking emulsing and capturing DNA libraries barcoded individually for each cell, sixth, amplification of the library and sequencing. Finally, the molecular barcode had to be removed from each sequencing read and written as a tag into the meta informatiom, to trace back from which cell the read pair was derived from. Data analysis revealed subclonal mutations in some cells that would have been missed with conventional amplicon sequencing. Phylogenetic trees were generated from a selection of sequenced cells to visualise and trace evolution of cell cultures.

Within the scope of this experiment the general protocol was designed and it was shown to give meaningful results, based on two mixed cell cultures, i.e. NIH3T3 with a human KRAS gene transfected carrying a heterozyous mutation in codon 12 and K562 being wild-type in KRAS codon 12 [255]. Both cells populations were mixed, emulsfied, i.e. transferred into aqueous droplets with a barcoded primer selection, amplified and prepared and subsequently sequenced. Results showed that introducing cellular barcodes with demPCR is a fast and cheap method that will proof useful in personalised medicine in the future.



(A) Cells are merged with beads carrying a barcoded collection of poly-T oligonucleotide, hence every bead is unique. After cell was lysed RNA is hybridised to barcoded primer sequences.



(B) After hybridisation is complete, droplets are broken and beads are captured. After reversed transcription of the RNA, barcoded cDNAs are amplified and sequenced

FIGURE 4.1: Brief overview of single cell mRNA library preparation using Drop-Seq. Images from Macosko et al. [247]

4.2 Material and Methods

The experiment was split into the following major steps: 1) Obtaining a large number of uniquely barcoded beads carrying a bead-specific barcode and a selection of primers for target enrichment. 2) Emulsification of beads and cells. 3) cell lysis and barcoding of cellular DNA with library preparation and amplification. 4) Breaking emulsion and capturing DNA libraries barcoded individually for each cell. 5) Amplification of the library and sequencing. 6) Split the barcode from reads

and adding it to SAM file to trace back from which cell a read pair was derived from.

4.2.1 Generating Uniquely Barcoded Beads

To generate a large number of uniquely barcoded beads, every bead had to be loaded with one single fragment carrying a degenerative barcode carrying the sequencing adapter, a unique barcode and a *universal sequence oligonucleotide* (US), as shown in Figure 4.4. This was achieved by adding a very low concentration of oligonucleotides to a surplus of beads to ensure the number of beads carrying an oligo was very low, as otherwise beads could carry multiple barcodes, shown in Figure 4.2. Under the assumption that the number of beads was high and every oligo was able to bind independently from other oligos the anchoring process followed the Poisson distribution:

$$P(X = k) = P_\lambda(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \tag{4.1}$$

With an event rate λ, which was defined as

$$\lambda = \frac{N_o}{n_B}, \tag{4.2}$$

where N_o was the number of oligos and n_B the number of beads respectively. The ratio of beads and oligos define an upper bound of how many beads would carry no ($k = 0$), one ($k = 1$) or more than one ($k > 1$) of the unique oligos. Concentration ratio between oligos and beads was adjusted so less than 0.001% of beads were carrying more than one oligo. Beads were then saturated with a surplus of a biotinylated reverse forward primer and US, shown in Figure 4.3. Beads that were successfully loaded with a barcoded primer sequence were enriched, using magnetic enrichment beads carrying a complementary oligo to the sequencing adapter. Beads were pulled down by a magnet and any beads without a unique oligo were washed off, sketched in Figure 4.5. Emulsification of beads in a large volume of oil and PCR mix, ensured that beads were encapsulated in a single droplet, shown in Figure 4.6. The oil recipe was tested to be stable throughout the PCR to prevent droplets from merging. Several rounds of PCR and a mixture of primers sharing the sequencing adapter amplified the full nucleotide with a number of gene primers anchored onto the bead, outlined in Figures 4.7 - 4.12. After PCR was completed the emulsion was broken and beads were captured, washed and stored at 4°C until further use. The full protocol is listed in Section D.2.

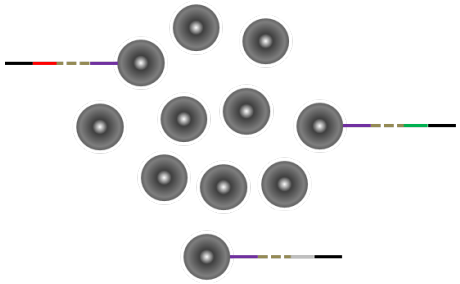


FIGURE 4.2: Biotinylated oligos at a low concentration bind to non-magnetic streptavidin coated silica beads.

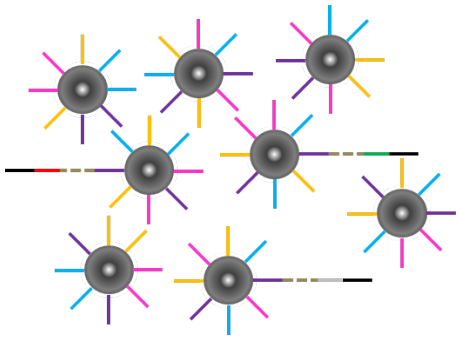


FIGURE 4.3: Beads are saturated with a high concentration of biotinylated primer and the US (Anchor Primers).

FIGURE 4.4: Unqiue oligos consist of three parts: primer, the US, a degenerative barcode and sequencing adapter

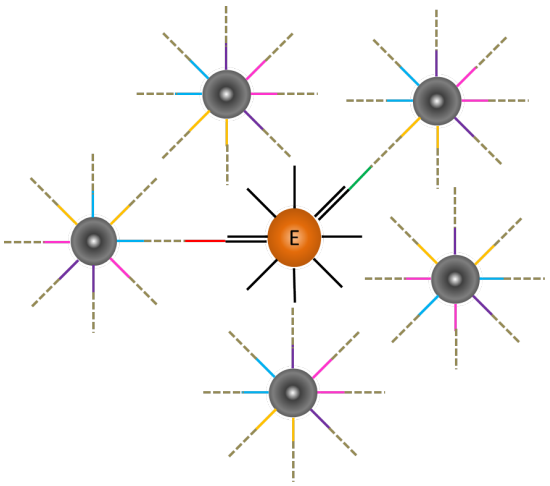


FIGURE 4.5: Beads carrying a unique oligo are enriched by adding magnetic beads carrying a reverse complement to sequencing adapter (Enrichment Sequence).

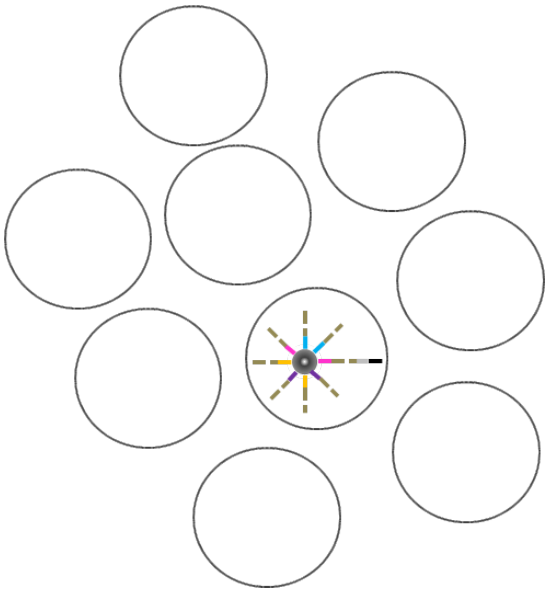


FIGURE 4.6: Beads with unique oligos are emulsified in a large volume of PCR mastermix (aqueous) and oils.

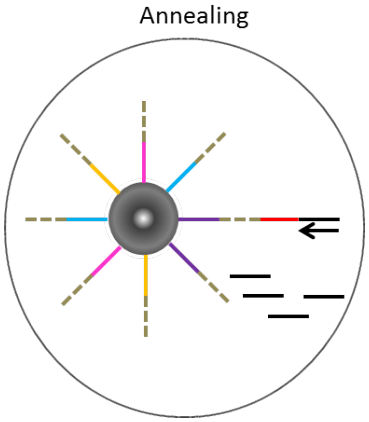


FIGURE 4.7: After the first annealing step, a Bead-loading Primer anneals to the unique oligo. It carries a 3' primer overhang.

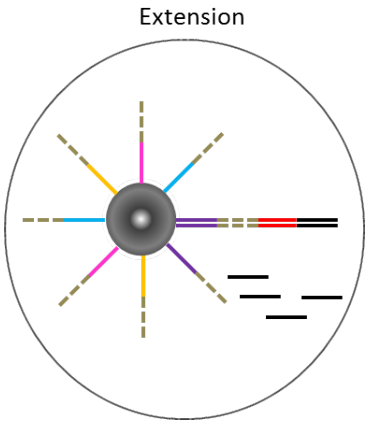


FIGURE 4.8: Bead loading primer is extended to a full copy of the Unique Oligo.

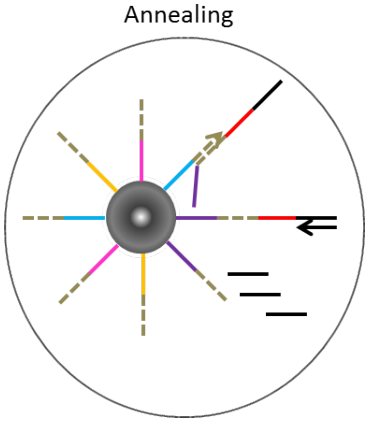


FIGURE 4.9: In the second cycle the copy hybridises to a US, because the complementary sequence is occupied by a new Bead-loading Primer.

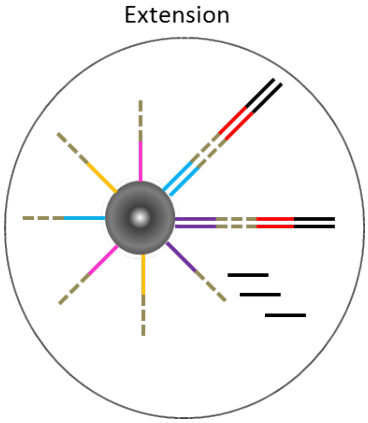


FIGURE 4.10: Complementary sequences hybridise.

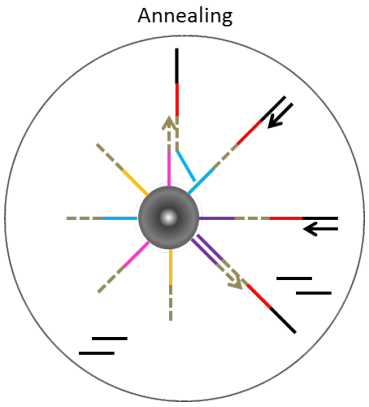


FIGURE 4.11: Copies of unique oligo double after every cycle. After a limited number of cycles the entire bead is saturated with uniquely barcoded primer sequences.

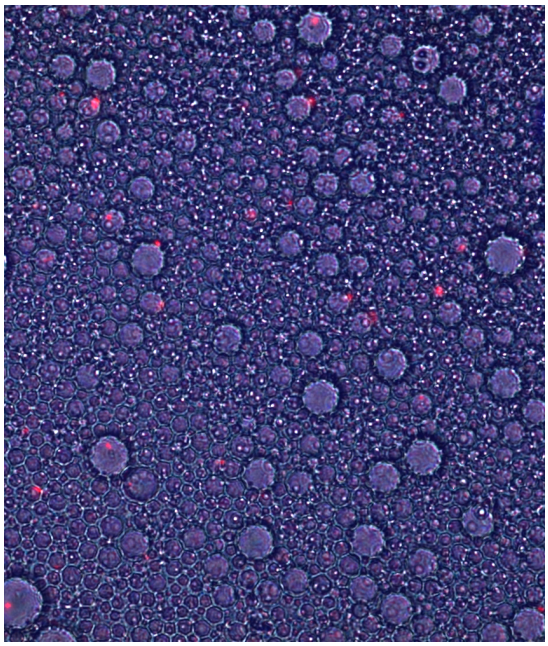


FIGURE 4.12: Emulsified beads loaded with biotinylated oligos labelled with Texas Red.

4.2.2 Single Cell Direct Emulsion PCR

As a proof-of-principle of the protocol, two cell lines, K562 and NIH3T3 were pooled in an 4:1 ratio. Exons two and three of human KRAS were amplified with demPCR and sequenced with introduced cellular barcodes on a MiSeq instrument; the entire protocol is listed in Section D.3. Uniquely barcoded beads and cells were emulsified, i.e. encapsulated in a large quantity of microreactors [241]. Concentrations of beads and cells were adjusted such that as many cells as possible can be barcoded and amplified. Hence, concentration of beads was chosen to tolerate multiple beads in one droplet, resulting in multiple barcodes could be derived from the same cell. Target regions were amplified and carried a 15 nucleotide degenerative barcode and sequencing adapters attached when at least one bead was present, shown in Figures 4.13 - 4.20. After 20 cycles of PCR barcoded libraries were recovered and the library was amplified in a subsequent PCR. Cleaned amplicons were quantified and loaded onto a MiSeq flow-cell, where the library was sequenced as spike in on a 150bp paired-end sequencing run.

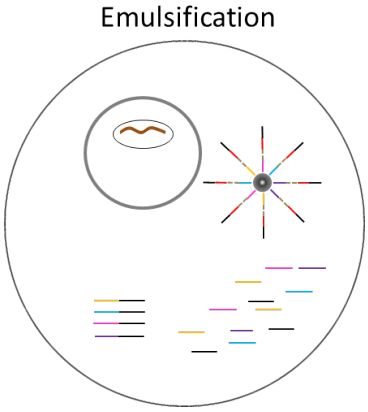


FIGURE 4.13: Cells and beads are emulsified in direct PCR mix containing Amplification Primer, Bead-loading Primer and Reverse Primer.

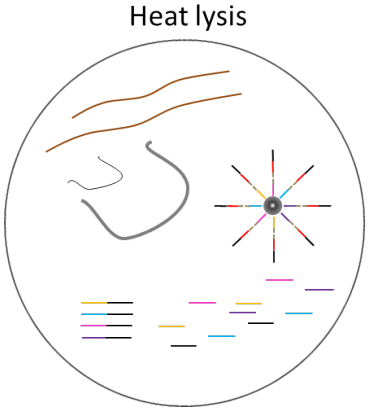


FIGURE 4.14: Cell is lysed during the first heat cycle

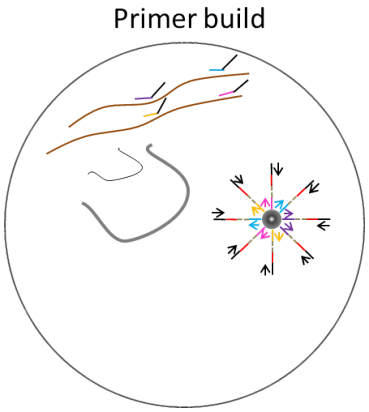


FIGURE 4.15: In the first annealing step primers amplify only barcoded primer library from the bead.

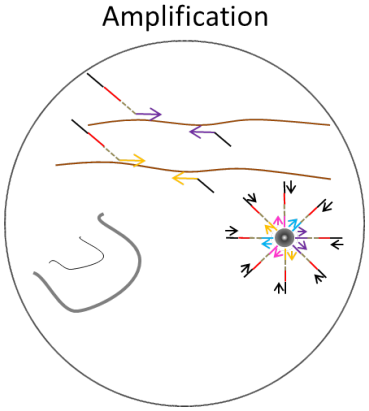


FIGURE 4.16: With barcoded primers amplified from the bead target region of cellular DNA is amplified with barcode and adapter overhang.

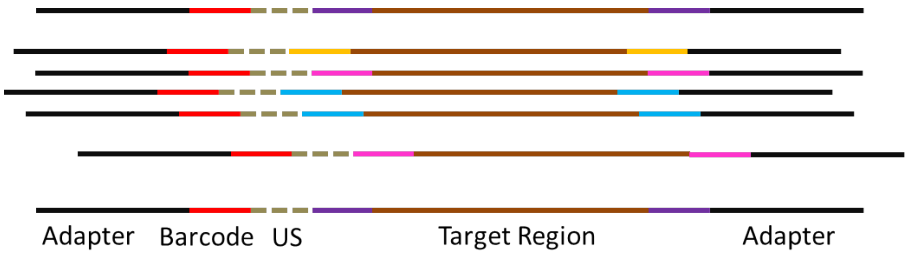


FIGURE 4.17: Resulting libraries are purified after emulsion is broken.

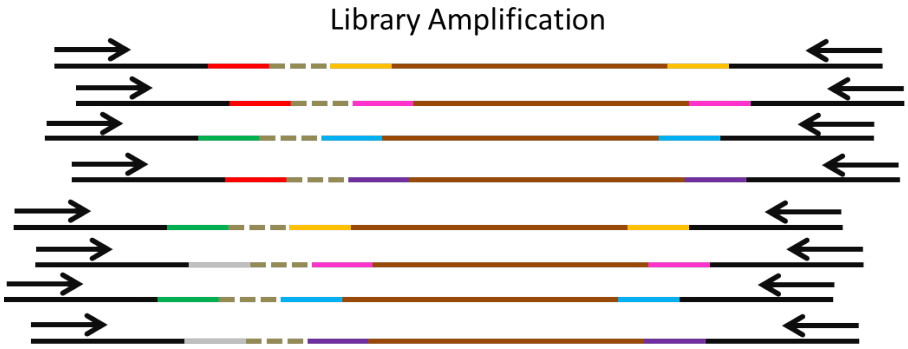


FIGURE 4.18: Libraries from all cells are collected and amplified.

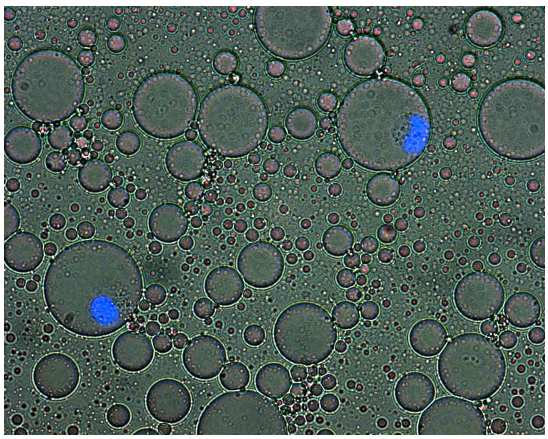


FIGURE 4.19: Cells captured in droplets. Nucleus stained with DAPI (blue).

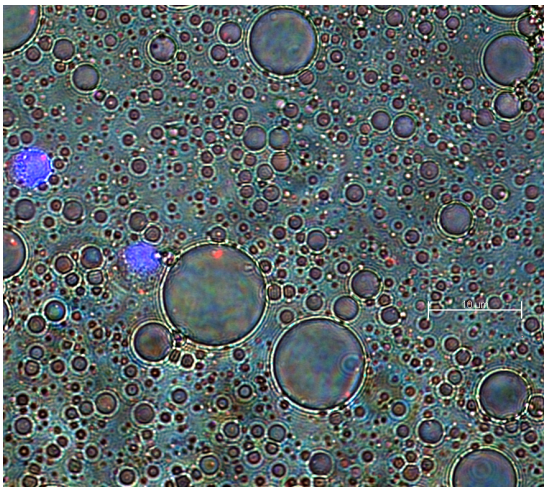


FIGURE 4.20: Cells and beads captured in droplets. White ruler indicates 10μm

4.2.3 Data Analysis

Sequencing data was translated into FASTQ data with bcl2fastq [131]. A script scanned the forward read automatically for the US and 15 bp barcode sequence that were located in the first 35 nucleotides of the forward read. The script was written in GNU R and is listed in Section D.4 [242]. When the script found a molecular barcode in front of the US in the read, it cut of the entire 35 bases and wrote the barcode into the FASTQ header:

```
@M01706:46:000000000-ALDMT:1:2114:13559:25487 1:N:0:0 BC:Z:ACACACAGTCGCGGT
```

The mapper BWA maintained the comment section and wrote the barcode into the SAM file as a custom tag [68]. This tag allowed filtering, sorting and grouping of alignments based on that flag; the script is listed in Section D.5. The resulting SAM file was visualised with IGV [100], which provided a feature to group alignments by custom tags. An example is shown in Figure 4.21. The read data shows a clear heterozygous mutation in reads with the same barcode, as expected. Mutations in NIH3T3 was previously confirmed via pyrosequencing, pyrograms are shown in Figures 4.22 - 4.25.

FIGURE 4.21: Example of the barcoded forward reads. It can be seen that reads were obtained from the same cell (labelled ACACACAGTCGCGGT) and the read data shows a heterozygous mutation indicating it was a NIH3T3 cell sequenced. Short reads indicate that only the primer was sequenced. This is caused by a missing size selection of the library after amplification.

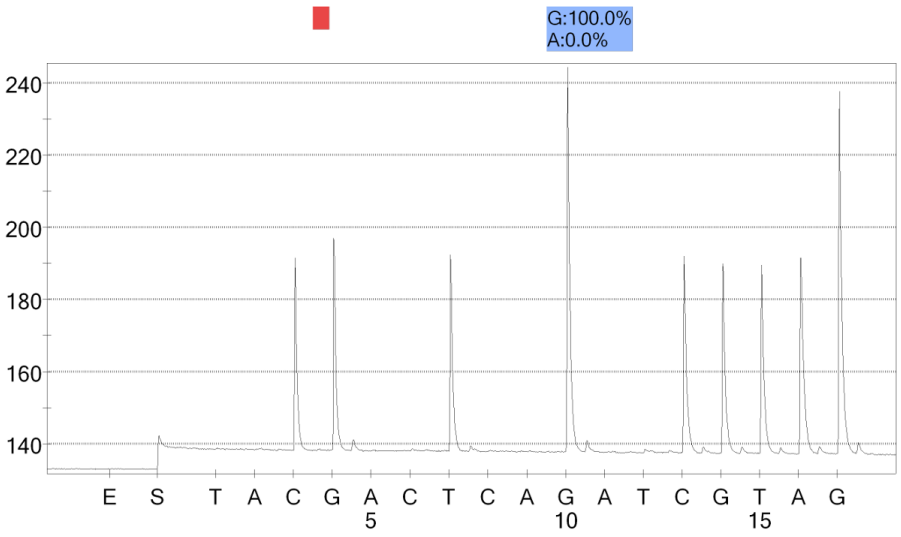


FIGURE 4.22: Pyrogram of sequenced KRAS codons 12 and 13 from DNA extracted from NIH3T3 cells. The red square indicates the expected $C \rightarrow G$ change.

After filtering of poor-quality reads and removal of adapters with Skewer [49], reads without a barcode were deleted. The remaining 374,000 reads represented over 1,700 different barcodes. Unfortunately, a large portion (2/3) of reads were just the forward primers, as the library was not cleaned by a size selection step. But still hundreds of cells were successfully sequenced giving a deep insight into both cell cultures.

VarScan2 reported two variants across KRAS exon 2 and 3, one in locus chr12:25,398,285 ($C \rightarrow G$), the KRAS p.(Gly12Arg) mutation detected by pyrosequencing and a second in locus 12:25,398,311 ($T \rightarrow C$). Both variants were reported to be present below 25% variant allele frequency.

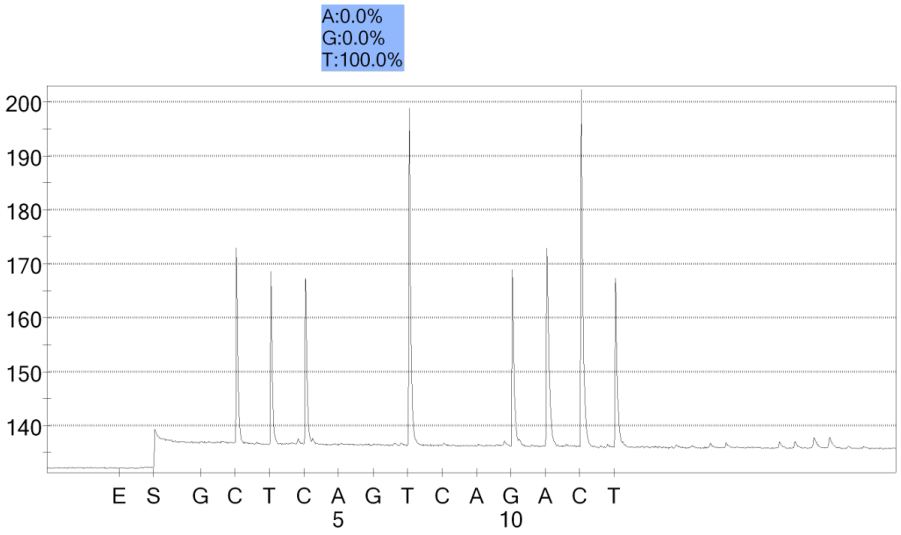


FIGURE 4.23: Pyrogram from sequenced NIH3T3 cells from pyrosequencing. The pyrogram is a diagram plotting light intensity over time.

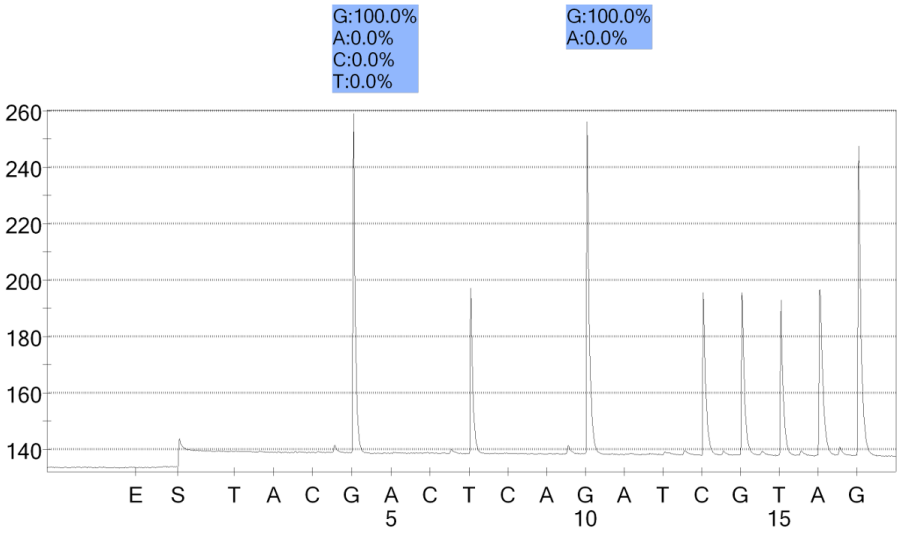


FIGURE 4.24: Pyrograms of sequenced KRAS codons 12 and 13 from DNA extracted from K562 cells.

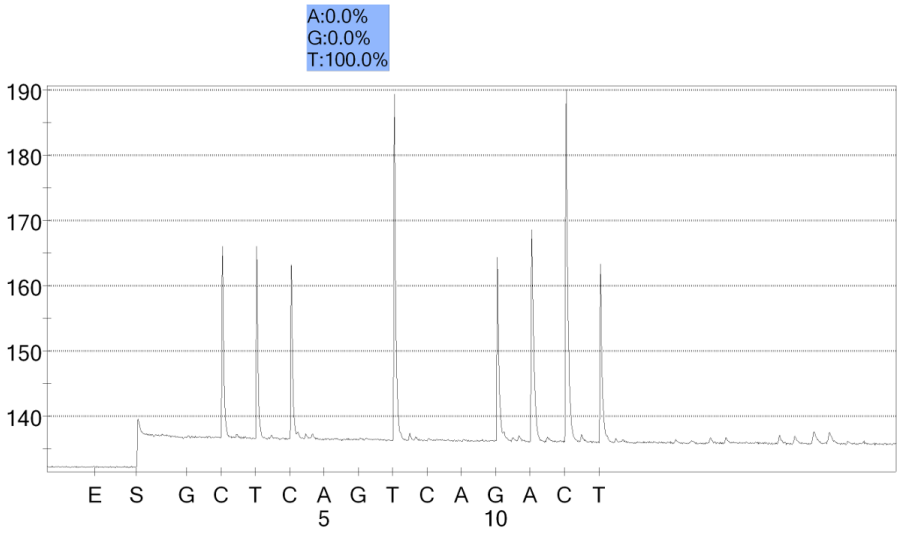


FIGURE 4.25: Pyrograms from sequenced K562 cells from pyrosequencing. The pyrogram is a diagram plotting light intensity over time.

By looking at the reads from single cells, however, it was revealed that there were more SNVs present than VarScan2 initially reported, e.g. a homozygous mutation in locus chr12:25,380,351 being present in some cells, but not all, as shown in Figure 4.26. By analysing over 100 cells individually, a few other interesting observations were made. Some of the variants detected in single cells were believed to be noise from multiple cells captured in the same droplet, e.g. cell CTCCCCAGCCCGACC showed a 17% $C \rightarrow G$ change at locus chr12:25,398,285. This can either be explained by PCR artefacts, a change in ploidy or multiple cells captured in the same microreactor, shown in Figure 4.27. At locus chr12:25,380,351 it appeared that the cell was heterozygous, but

carried also some mutations that were linked to a neighbouring SNV in locus chr12:25,380,344 which were showing an $A \rightarrow G$ change. This was very unlikely to be caused by a combination of sequencing and/or amplification errors. Hence, multiple cells got captured in the same droplet and had the same barcode assigned. With a limited number of cells being barcoded, it was almost always possible to identify these cases by eye. Unfortunately, there was no known SNV in K562 that could have been used for a further investigation how many barcodes had a mixed signal.

Theoretically only a limited number of reads per cell were necessary to find hetero- or homozygous mutations, as NIH3T3 was assumed to be diploid [254]. K562 was, however, expected to be polyploid [256]. For estimating ploidy from single-cell massively parallel sequencing data a sufficient number of variant sites, e.g. SNPs or mutations, would have been necessary, which was not the case, as only two short regions were targeted [253]. Nevertheless, unusual variant allele frequencies in the data could be explained by noise and ploidy change in K562. For example, the KRAS p.(Gly12Arg) mutation was expected to be present at 40%, as NIH3T3 cells were spiked in at 80% carrying a heterozygous mutation, but it was present in 25% of the aligned reads.

Overall, 14 different cell-specific mutations were found, some of them were present in many cells, while others were rare. Although cells were

not exposed to a strong selective pressure, it was interesting to see that some cells accumulated additional mutations.

To see clonal evolution of both cell cultures and if K562 and NIH3T3 were amplified separately, cellular barcodes were split into individual SAM files, variants were called on 25 cells and clustered to a phylogenetic tree [214]. Variant calling was performed with an allelic threshold of 20%, to account for polyploidy and amplification bias, but to decrease the effect of mixed signals, where two cells were trapped in the same droplet, the command is listed in Section D.6.

Some cells showed insufficient coverage in that region for variants to be called, shown in Figure 4.29. These were subsequently removed from the data set. Subsequently a phylogenetic tree of a subset of 25 cells was drawn in Figure 4.30. Cells were selected as they showed sufficient coverage in both exons and gave a nice representation of the data that was obtained. The tree was built from 6 SNVs present in the 25 cells including the known mutation in NIH3T3 cells in KRAS exon 2. One cell seems to have a very different genotype, hence it was believed to be contamination, as it showed three homozygous mutations uniquely present in the dataset. The other cells were mainly agglomerated by cell type, except four cells that showed the p.(Gly12Arg) mutation, but were clustered to K562 cells.It was believed that this was caused by droplets with multiple cells that were evenly amplified, as shown in Figure 4.28.

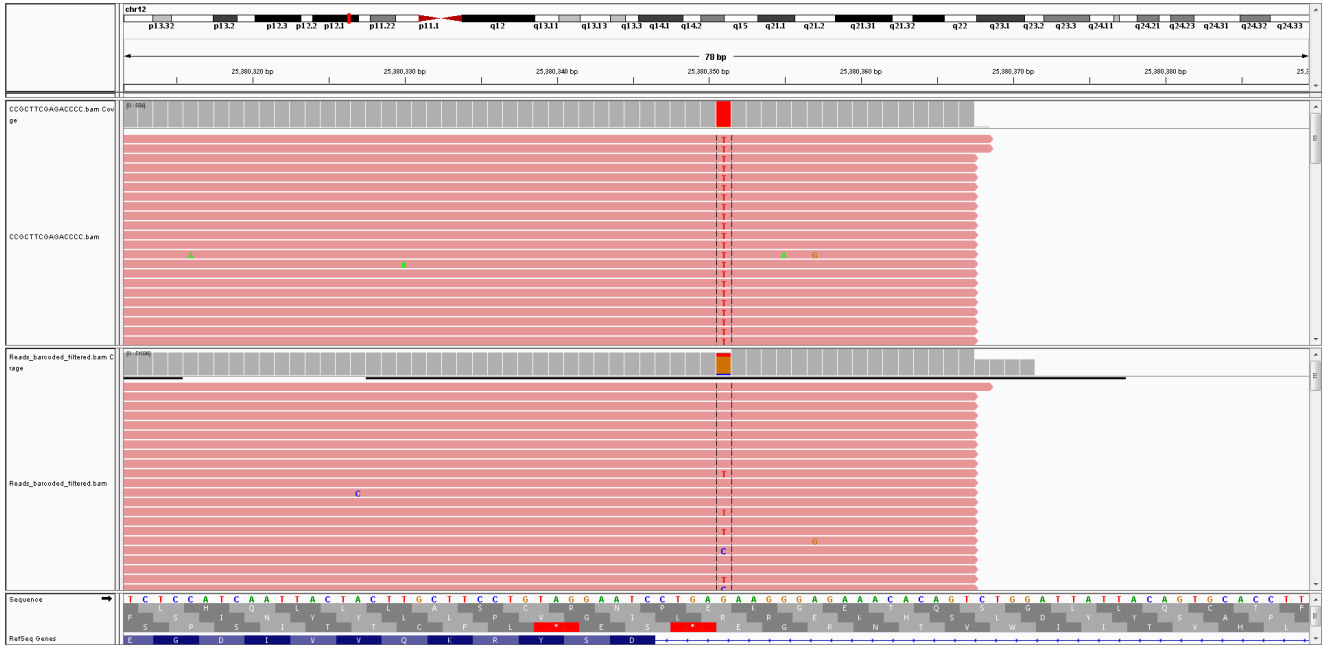


Figure 4.26: Top - Homozygous SNV detected from read data of a single barcode (CCGCTTCGAGACCCC). Bottom - Combined reads (barcodes merged).

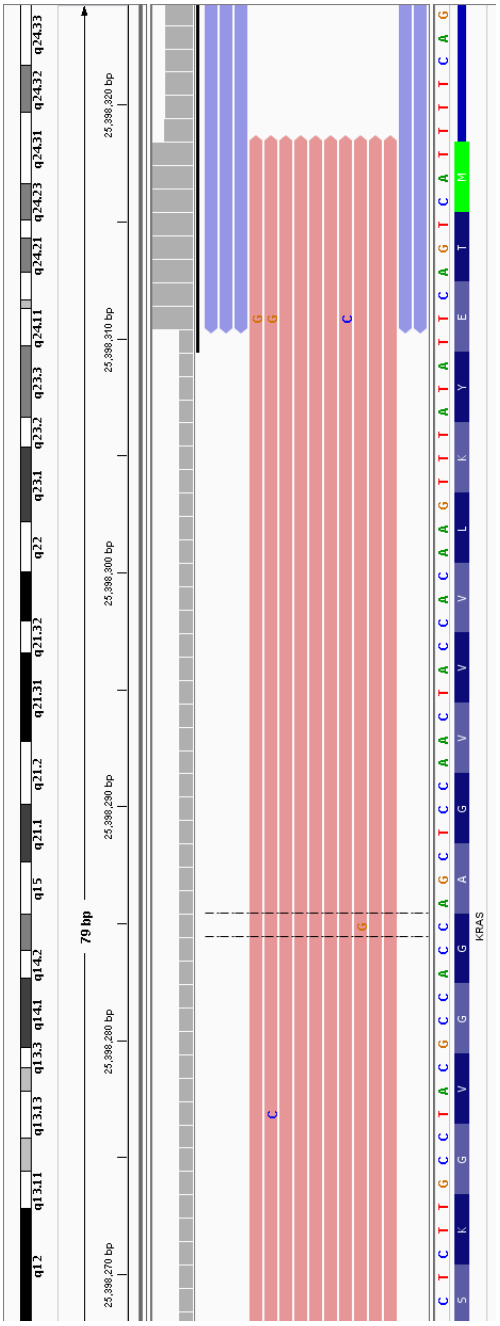


FIGURE 4.27: $C \rightarrow G$ change in locus chr12:25,398,285 is present at below 20% in read data assigned with barcode CTCCCCAGCCCGACC.

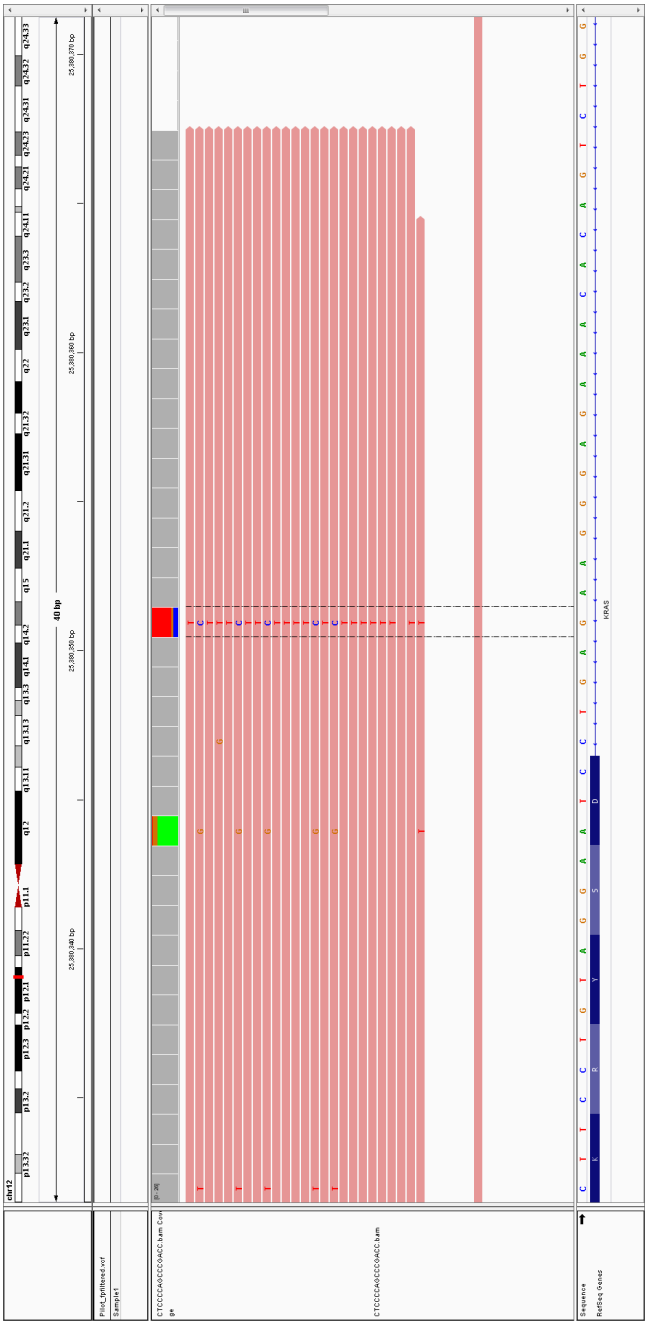


FIGURE 4.28: The same cell (CTCCCCAGCCCGACC) shows three SNVs near each other. This is likely to be derived from two cells showing exclusive genotypes.

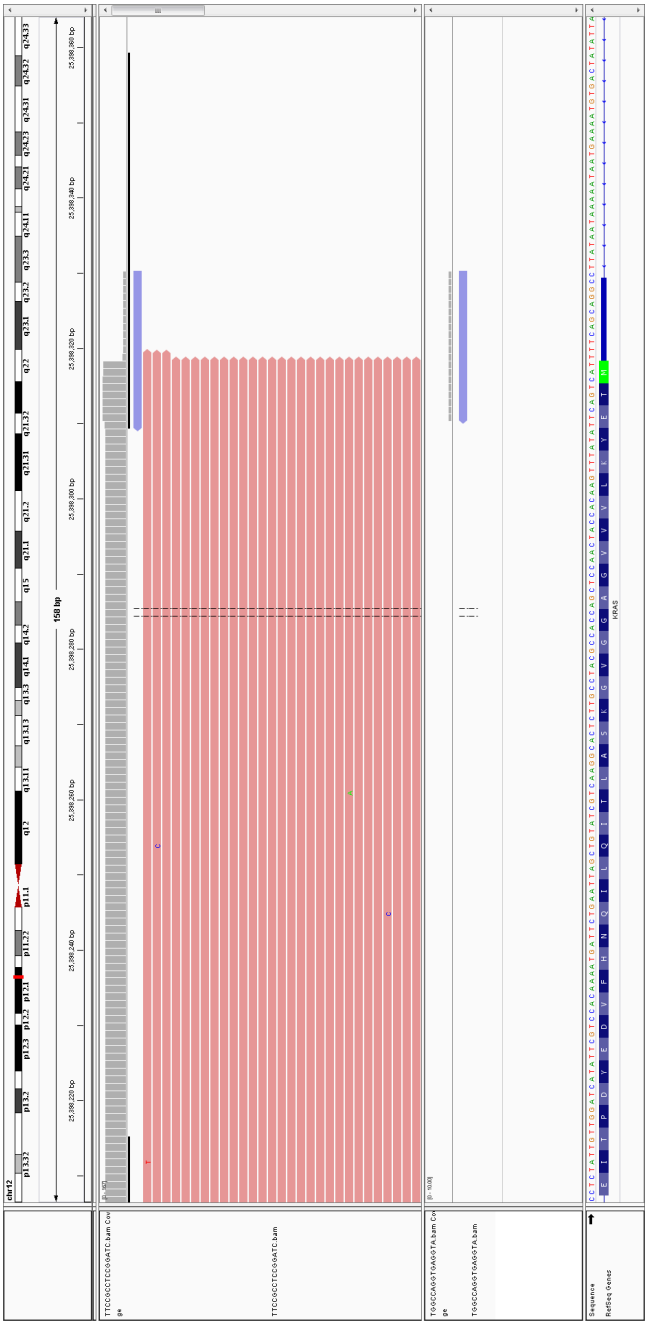


FIGURE 4.29: Alignments of two barcodes (top-TTCCGCCTCCGGATC, bottom-TGGCCAGGTGAGGTA). The top track shows sufficient coverage for a variant calling, the bottom track has insufficient reads for variant calling.

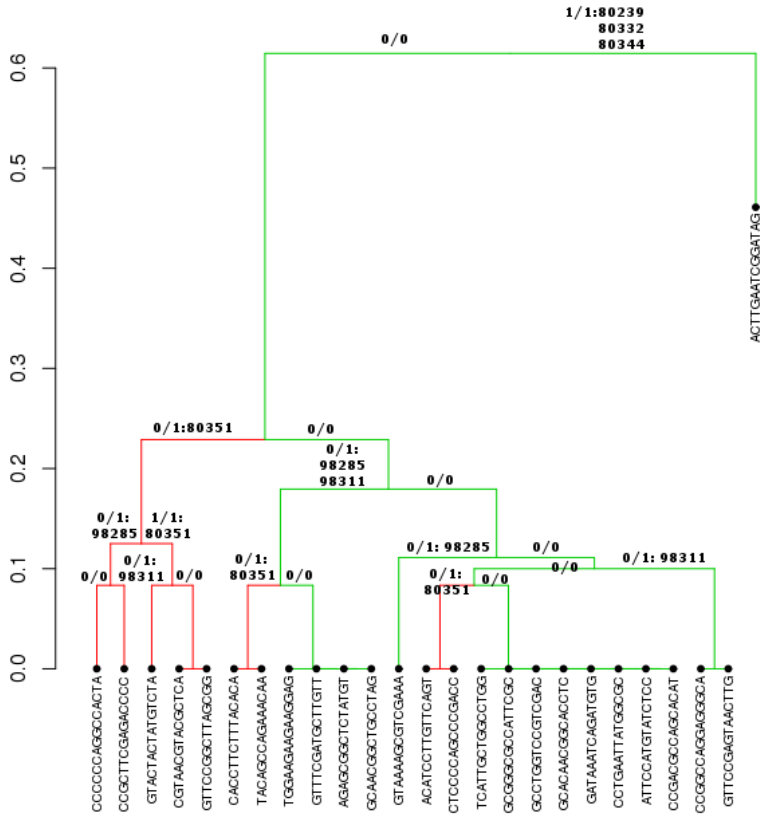


FIGURE 4.30: A phylogenetic tree of 25 selected cells. Every edge is labelled with the mutation(s) gained. Red edges indicate NIH3T3 cells.

4.3 Discussion

It was shown that the introduced protocol amplifies target regions, barcodes single cells and prepares them for sequencing on a large scale. The method was demonstrated to be quick and more cost-effective than other methods, such as microfludics, as it requires only two PCRs. High-throughput single cell sequencing with emulsion PCR revealed a very detailed picture of genetic evolution of K562 and NIH3T3 cell cultures. Some parameters need some fine adjustments, as droplets sometimes captured multiple cells or free DNA from damaged cell and the protocol was updated by adding a size selection step with magnetic SPRI beads [239]. Like with other amplicon based sequencing approaches, the design of multiplexed primers is crucial to avoid an unbalanced amplification to not leave regions poorly covered. Some artefacts, however, were compensated for by applying stringent filtering mechanisms, e.g. during the variant calling.

The method can be further adjusted for improved results and higher throughput, e.g. by applying a random barcode sequence on every primer and bead as well, to use potential duplicates for removal of sequencing or amplification errors. This is possible, but it would make manufacturing of uniquely barcoded beads even less efficient, as it is right now. Instead, other technologies for manufacturing beads promise a significantly better yield, lower cost and a higher speed. One alternative approach was described by Macosko et al. as the "split-pool"

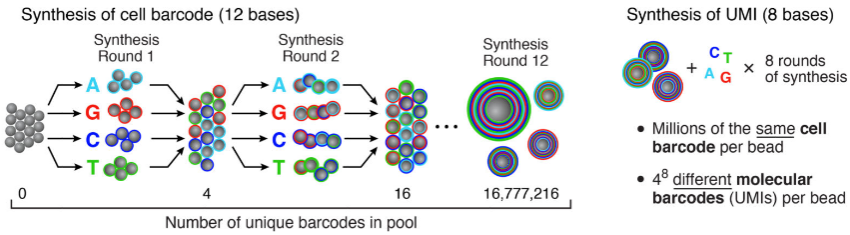


FIGURE 4.31: Split-pool approach for on-bead nucleotide synthesis. Beads are split into four pools, each adds one of the four nucleotides (A,C,G,T) directly onto the bead. After every step, beads are pooled and split again until the barcode is completely synthesised. For receiving a degenerative barcode per primer (UMI) all beads are pooled and bases are added randomly in each cycle. The same principle can be used for loading primer sequences and adapters. Image from Macosko et al. [247].

approach, shown in figure 4.31. It does not require an emulsion for bead separation, hence preparation of large volumes of emulsions for a relatively small fraction of beads would not be necessary anymore. By switching to the split-pool approach oligos could be redesigned so that, similar to barcoded molecular inversion probes the degenerative barcode sequence could be placed in the i5 index region of the adapter [222]. This would make the US useless, hence it could be removed. This would increase sequencing yield and simplify subsequent data analysis. Further, this method could be applied to microfluidics as well since single cell direct PCR in droplets is possible. This could decrease noise ratio to a minimum, as the risk of capturing multiple cells in a droplet would be reduced.

The potential of sequencing large quantities of single cells from a

population is large, for example in cancer diagnostics, where mutations can be present even in a tiny amount of cells in the entire population that need to be screened for. Even in this small pilot, which was serving as a proof of concept, it was shown that homozygous mutations were present in some cells, which were not spotted by performing analysis on traditional amplicon-seq data. Latest massively parallel sequencing instruments, such as the Next500 or HiSeq 4000 produce a vast amount of sequencing read at a reasonable length to sequence target regions of hundreds of thousands of single cells at a sufficient coverage in a single run. Current data formats and algorithms, such as the SAM and VCF standards have proven versatile and useful for processing of large amounts of sequencing data from many samples. Focussing on single cell data will, however, reveal new types of artefacts and require new concepts in quality control, normalisation and data analysis including and beyond variant calling. Of course, diagnostic are not yet feasible, as tissue samples are usually preserved with FFPE protocols that introduce artificial mutations making single cell genotyping a challenge. Further, formalin causes cross-linkages between proteins and DNA making cell dissociation from FFPE tissue cumbersome. Although some approaches exist using harsh method such as pepsin and heat treatment [245], these only dissolve out a minor fraction of cells, with an unknown amount of free DNA from lysed cells.

References

[234] Daniel Branton et al. "The potential and challenges of nanopore sequencing". In: *Nature biotechnology* 26.10 (2008), pp. 1146–1153.

[235] H Christina Fan, Glenn K Fu, and Stephen PA Fodor. "Combinatorial labeling of single cells for gene expression cytometry". In: *Science* 347.6222 (2015), p. 1258367.

[236] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". In: *Nature Reviews Genetics* 17.3 (2016), pp. 175–188.

[237] Marco Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England Journal of Medicine* 366.10 (2012), pp. 883–892.

[238] Mira T Guo et al. "Droplet microfluidics for high-throughput biological assays". In: *Lab on a Chip* 12.12 (2012), pp. 2146–2155.

[239] Trevor L Hawkins et al. "DNA purification and isolation using a solid-phase." In: *Nucleic Acids Research* 22.21 (1994), p. 4543.

[222] Joseph B Hiatt et al. "Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation". In: *Genome research* 23.5 (2013), pp. 843–854.

[240] Christian Hoffmann et al. "DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations". In: *Nucleic acids research* 35.13 (2007), e91.

[241] Machiko Hori, Hajime Fukano, and Yosuke Suzuki. "Uniform amplification of multiple DNAs by emulsion PCR". In: *Biochemical and biophysical research communications* 352.2 (2007), pp. 323–328.

[242] Kurt Hornik. *R FAQ*. 2016. URL: https://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

[49] Hongshan Jiang et al. "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads". In: *BMC bioinformatics* 15.1 (2014), p. 1.

[243] Scott R Kennedy et al. "Detecting ultralow-frequency mutations by Duplex Sequencing". In: *Nature protocols* 9.11 (2014), pp. 2586–2606.

[244] Allon M Klein et al. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". In: *Cell* 161.5 (2015), pp. 1187–1201.

[245] MPG Leers et al. "Heat pretreatment increases resolution in DNA flow cytometry of paraffin-embedded tumor tissue". In: *Cytometry* 35.3 (1999), pp. 260–266.

[68] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[246] FJ Livesey. "Strategies for microarray analysis of limiting amounts of RNA". In: *Briefings in functional genomics & proteomics* 2.1 (2003), pp. 31–36.

[247] Evan Z Macosko et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161.5 (2015), pp. 1202–1214.

[248] B Mercier et al. "Direct PCR from whole blood, without DNA extraction." In: *Nucleic acids research* 18.19 (1990), p. 5908.

[249] H Mirhendi et al. "Colony-PCR is a rapid and sensitive method for DNA amplification in yeasts". In: *Iranian J Publ Health* 36.1 (2007), pp. 40–44.

[87] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond". In: *Genome research* 25.10 (2015), 1499–1507.

[250] Nicholas E Navin and James Hicks. "Tracing the tumor lineage". In: *Molecular oncology* 4.3 (2010), pp. 267–283.

[251] Richard Novak et al. "Single-Cell Multiplex Gene Detection and Sequencing with Microfluidically Generated Agarose Emulsions". In: *Angewandte Chemie International Edition* 50.2 (2011), pp. 390–395.

[252] Sharad Pathak et al. "Counting mycobacteria in infected human cells and mouse tissue: a comparison between qPCR and CFU". In: *PLoS One* 7.4 (2012), e34931.

[253] Davide Prandi et al. "Unraveling the clonal hierarchy of somatic genomic aberrations". In: *Genome Biol* 15.8 (2014), p. 439.

[254] S Pulciani et al. "ras gene Amplification and malignant transformation." In: *Molecular and cellular biology* 5.10 (1985), pp. 2836–2841.

[255] Simonetta Pulciani et al. "Transforming genes in human tumors". In: *Journal of cellular biochemistry* 20.1 (1982), pp. 51–61.

[99] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. "The advantages of SMRT sequencing". In: *Genome Biol* 14.6 (2013), p. 405.

[100] James T Robinson et al. "Integrative genomics viewer". In: *Nature biotechnology* 29.1 (2011), pp. 24–26.

[256] Camelia Iancu Rubin, Deborah L French, and George F Atweh. "Stathmin expression and megakaryocyte differentiation: a potential role in polyploidy". In: *Experimental hematology* 31.5 (2003), pp. 389–397.

[257] Antoine-Emmanuel Saliba et al. "Single-cell RNA-seq: advances and future challenges". In: *Nucleic acids research* 42.14 (2014), pp. 8845–8860.

[258] Richard Williams et al. "Amplification of complex gene libraries by emulsion PCR". In: *Nature methods* 3.7 (2006), pp. 545–550.

[259] Yong Zeng et al. "High-performance single cell genetic analysis using microfluidic emulsion generator arrays". In: *Analytical chemistry* 82.8 (2010), pp. 3183–3190.

[214] Xiuwen Zheng et al. "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24 (2012), pp. 3326–3328.

[131] *bcl2fastq2 Conversion Software Guide*. 2.17 (15051736 Rev. G). ILLUMINA PROPRIETARY. Illumina. July 2015.

Chapter 5

Final discussion and Future Perspectives

Within the scope of this thesis three projects in massively parallel sequencing in precision medicine were examined to understand the principles of target enrichment panels to limit sequencing to relevant parts of the genome, i.e. important cancer genes of interest for pathologists. Selection of 69 genes were carefully decided with recent publications of scientific community and personal interaction with scientists and doctors to ensure an optimal choice by not selecting too many regions to increase cost and time for subsequent bioinformatics' analysis.

A first part of the work was to design a custom enrichment Agilent

SureSelect XT panel based on 120bp RNA capture probes. To ensure detection of all potential driver mutations in genes, hybridisation probes were designed for all exons [266]. To investigate capture potential the panel was tested on two samples carrying a selection of validated mutations in a number of genes. As probes show different binding potential, capture efficiencies differed leaving some genetic regions poorly covered, while others were overrepresented. To achieve a more balanced capture, probe concentrations were adjusted according to their observed enrichment capabilities. Although this increased the number of miscaptured DNA fragments, the new probe cocktail was more balanced and showed better coverage of previously underperforming target regions. The adjusted panel was tested on a large number of clinical samples to assess performance, sensitivity and limitations of the technology and how much sequencing was required to identify SNVs or short InDels of DNA extracted from FFPE tissue showing a large variety of quality and quantity characteristics. By testing a large number of samples it was possible to make claims about quantity and quality of necessary input material and the effect on variants that could be identified. Moreover precise filter criteria based on structural and functional properties could be defined to increase robustness of the assay in case samples were poorly preserved. Finally filtered variants from samples were clustered by dimension reduction. It was observed that samples tend to cluster independently from their primary tumour type. It was speculated that this was caused by SNPs present among the

human population, but also because mutations in some cancer genes are commonly present among many different cancer types, which may be used to identify potential outliers in the future.

Potential more work is possible in this area. In particular to get a better understanding of how integrity and quality of extracted DNA influences the target enrichment. There were signs in the data that an increase in sequencing can increase results from substandard material in some cases. A further investigation of that presumption would decrease the number of drop-outs and be of great value for target enrichment methods used in precision medicine. Another improvement to be considered is to improve library preparation, by increasing the end-repair and adapter ligation process. Efficiency of the current enzyme mix that is used are sensitive to fluctuations in local GC content [104], which causes an enrichment bias, which is compensated by adjusting probe concentrations, which is clearly not ideal, as it was shown that probes are affected by problematic GC contents as well. Instead modified enzymes could improve general performance of library preparation, resulting in a better DNA yield prior to enrichment resulting in less input material needed and results in better target capture even on poor quality DNA [267, 261]. Further, a decrease in probe length could improve binding specificity and decrease captures of fragments from flanking regions of target exons. Although this would increase the overall number of probes resulting in a more complex probe concentration adjustment, the higher binding specificity could help with

capture of difficult regions and simultaneously decrease time needed for hybridisation. The sequencing data of the analysed samples can be further used for redefining additional filter criteria, test different variant callers and use them in association studies to identify new potential driver mutations in various cancer types. The panel will be of benefit for clinicians and pathologists who require broad view across typically mutated and druggable targets, giving new insights and help with further development in cancer treatment and personalised medicine.

The second part covered the design of a different target enrichment method based on barcoded molecular inversion probes for target capture and amplification using a molecular barcode for detection of PCR duplicates and sequencing error correction promising an advantage over conventional amplicon sequencing methods. Instead of providing a comprehensive cancer panel for exploring known cancer genes in potentially new or unusual tumours or cancer phenotypes, this panel was designed to replace and extend current assays of genetic marker testing in commonly mutated genes that have an effect on pathways targeted by commonly prescribed drugs. The panel covered translated exons according to RefSeq [228] and VEGA [231]. Other databases have been utilised, but did not add additional regions. Due to the fact that the panel was significantly smaller than other panels, up to 24 samples could be loaded onto an Illumina® MiSeq instrument to be cost-efficient and reduce turnaround time to obtain the results as fast as possible. As probes were designed such that the resulting amplicon size was above

300 bp, the MiSeq was configured for a 150bp paired-end run. Another hurdle to take was the large variety of quality and quantity of the input material. As there is very little control over clinical samples that are to be tested, the panel must provide reliable and consistent results even in cases of poorly preserved tissue or low amount of available input material. To assess limitations of this enrichment method, 48 clinical FFPE samples, including technical replicates, were prepared and tested, some of which showed very low DNA concentration. As the amount of available DNA material was very limited, extensive DNA quality control steps had to be skipped, to evade risk of a sample failing to be tested due to a miss of minimum input concentration requirements. Instead samples were tested with pyrosequencing for known mutation hotspots in four genes. It was shown that there was a problem with the probe design causing the panel to lack detection power for eight samples. There were strong indications that the structure of probes can cause problems on DNA with a lower integrity, although this could not be entirely proved due to lack of genetic material, which made a procedure of exclusion necessary. The effect was previously observed with other amplicon-based methods and HaloPlex systems before [232, 226]. Due to the uncertainty of methods requiring an intact fragment between two binding sites of a primer or probe, it was concluded that sensitivities of amplicon sequencing-based approaches do not reach a rate to be acceptable for diagnostic use, even if the technology benefits from a lower turnaround time than capture probe-based enrichment

methods. The panel could be adjusted by reducing the amplicon size, i.e. reducing distance between both binding sites of a probe and reducing the sequencing read length, but there would be no guarantee that this would be sufficient in all cases. Further studies are required to better understand the level of degradation of FFPE samples provided every day for diagnostic use. Although a number of studies exist based on utilising PCR [262, 265, 260], impact on target region enrichment and amplification are not well understood. The designed HaloPlex HS did not show an unusual drop in coverage in that region indicating a problem with the enrichment per se. Instead alleles were unequally amplified causing a false call in a high proportion of samples tested. Although this can happen with other enrichment methods as well, as no chemistry is perfect, a test that misses a mutation in half of the samples cannot be accepted in diagnostic testing.

In the third part a new method was introduced where degenerative barcodes were introduced into every amplicon from single cell DNA. Unlike other methods based on cell separation and massive amplification of DNA or methods using microfluidics as used for single cell RNA sequencing, this method is based on a direct multiplex PCR from beads carrying uniquely barcoded primer sequences. It was shown that DNA from single cells can be directly amplified without prior DNA extraction. Primer sequences were anchored on silica beads and introduced into droplets formed by emulsification of the PCR mix containing the

cells and oil. Resulting barcoded amplicons were cleaned with a conventional cleanup kit and directly sequenced on a MiSeq instrument. K562 and NIH3T3 cell populations were sequenced by pyrosequencing in KRAS codons 12, 13 and 61 to determine genotypes in both cell cultures. Subsequently both cell populations were mixed in a defined ratio, emulsified and KRAS exon 2 and 3 were amplified and barcoded from each cell. The barcode was extracted from every read and aligned to the reference. The SAM file was split by the determined barcodes and a selection of these barcodes were investigated individually. We found additional SNVs in NIH3T3 cells that were used for generating a phylogenetic tree. Besides that both cell types clustered individually, it was noticed that some barcodes returned mixed signals in some sites, potentially caused by loose DNA released by damaged cells or multiple cells being trapped in the same droplet. Although this was accounted for by adjusting variant calling filter criteria, a few cases remained where alleles were equally well amplified. It was revealed that fine tuning of individual parameters is required, such as the number of cells per reaction volume. Further improvements can be considered, i.e. a more efficient method of manufacturing uniquely barcoded beads with the split-pool approach, where nucleotides are attached directly onto beads base after base with intermediate pooling and splitting to introduce random barcodes. In cases where noise level needs to be reduced to a minimum, microfluidics can be applied reducing the risk of multiple cells trapped in the same droplet. Although the experiment

was designed as a proof-of-principle, single cell analysis revealed clonal evolution of both cell cultures that had been missed by conventional amplicon sequencing, due to their low frequency in the overall sequencing data. The potential benefit is large as current technologies are cost and time consuming, whilst yield tends to be low. This methods does not require any expensive instruments and a sequencing library of barcoded single cells was achieved with essentially two PCRs. This would be very useful for applications in diagnostic and prognostic testing, which would benefit tremendously from higher accuracy. Further, understanding of clonal evolution in specific cancer types and genetic heterogeneity based on clonal evolution has just begun [263, 237, 87]. Further fields beyond human cancer research are considerable, such as metagenomics and microbial applications [268] or detection of rare genetic diseases, such chromosome mosaicisms [264] benefit from novel single cell sequencing technologies. The method described within the scope of this thesis promises an easy to handle, cost-effective and fast method for sequencing single cells.

References

[260] Parviz Ahmad-Nejad et al. "Assessing quality and functionality of DNA isolated from FFPE tissues through external quality assessment in tissue banks". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 53.12 (2015), pp. 1927–1934.

[261] Maryke Appel et al. *KAPA HyperPlus/SeqCap EZ workflow: Improving Data Quality and Turnaround Times for Targeted Next-Generation Sequencing of FFPE DNA*. Tech. rep. KAPA Biosystems, 2016.

[262] Krzysztof Bielawski et al. "The suitability of DNA extracted from formalin-fixed, paraffin-embedded tissues for double differential polymerase chain reaction analysis". In: *International journal of molecular medicine* 8.5 (2001), pp. 573–578.

[263] Li Ding et al. "Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing". In: *Nature* 481.7382 (2012), pp. 506–510.

[237] Marco Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England Journal of Medicine* 366.10 (2012), pp. 883–892.

[264] Kurt Hirschhorn, Wayne H Decker, and Herbert L Cooper. "Human intersex with chromosome mosaicism of type XY/XO: Report of a case". In: *New England Journal of Medicine* 263.21 (1960), pp. 1044–1048.

[265] Steve Michalik and Christopher Williams. "Qualitative multiplex PCR assay for assessing DNA quality from FFPE tissues and other sources of damaged DNA". In: *Life Science* (2008), p. 23.

[226] Lotte NJ Moens et al. "HaloPlex Targeted Resequencing for Mutation Detection in Clinical Formalin-Fixed, Paraffin-Embedded Tumor Samples". In: *The Journal of Molecular Diagnostics* 17.6 (2015), pp. 729–739.

[87] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond". In: *Genome research* 25.10 (2015), 1499–1507.

[228] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 35.suppl 1 (2007), pp. D61–D65.

[266] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. "The promise of whole-exome sequencing in medical genetics". In: *Journal of human genetics* 59.1 (2014), pp. 5–15.

[104] Michael G Ross et al. "Characterizing and measuring bias in sequence data". In: *Genome Biol* 14.5 (2013), R51.

[267] G Rush et al. "Novel Improvements to the Illumina TruSeq Indexed Library Cconstruction, Amplification and Quantification Protocols for Optmized Multiplexed Sequencing". Poster. Feb. 2011.

[268] Susannah Green Tringe et al. "Comparative metagenomics of microbial communities". In: *Science* 308.5721 (2005), pp. 554–557.

[231] Laurens G Wilming et al. "The vertebrate genome annotation (Vega) database". In: *Nucleic acids research* 36.suppl 1 (2008), pp. D753–D760.

[232] Stephen Q Wong et al. "Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing". In: *BMC medical genomics* 7.1 (2014), p. 1.

Acknowledgements

My sincere gratitude to all hospitals, institutes and doctors who provided me with samples and cells. I wish all the best to the patients and their families that were part of this study. I want to say thank you to Dr Tom Burr, who supported me in my work, for his hints, insights and perspectives. Thank you to Dr Antonella Ronchi for all the advice and tips and for hosting me for the time in Milan in the Department of Biotechnology & Biosciences. Thank you for my colleagues and friends, Isaura, Sudhi, Gloria, Jake, Priscilla, Mirabela, Leanne, Kevin, Sarah, Dao and Depesh for their help and support.

This work was enabled by the Marie Curie fellowship program and the HemID consortium for providing training and guidance. Thank you for this opportunity and my European Union providing a border-free and safe environment for science and partnerships.

Partially this work was further financially supported by the Technology Strategy Board (Innovate UK) of the United Kingdom.

Thank you to my family and friends in Germany, Italy, Spain, Poland, Canada, USA, France, Netherlands, Greece, England, Austria. A particular thanks to my father Hans-Jürgen and Marlies for mental support and care packages with the finest treats. In memory of my mother Ilona. She died from cancer at the age of 50.

Appendix A

Protocols and Description of Pilot Samples

A.1 Genotypes of Samples Used

HGVS.p	Gene	Locus	Freq.	Sample
V600E	BRAF	chr7: 140453136	20%, 10.5%	BRAF20, QUANTR.
D816V	KIT	chr4: 55599321	10.0%	QUANTR.
delE746 -A750	EGFR	chr7: 55242463	2.0%	QUANTR.

L858R	EGFR	chr7: 55259515	3.0%	QUANTR.
T790M	EGFR	chr7: 55249071	1.0%	QUANTR.
G719S	EGFR	chr7: 55241707	24.5%	QUANTR.
G13D	KRAS	chr12: 25398281	15.0%	QUANTR.
G12D	KRAS	chr12: 25398284	6.0%	QUANTR.
Q61K	NRAS	chr1: 115256530	12.5%	QUANTR.
H1047R	PIK3CA	chr3: 178952085	17.5%	QUANTR.
E545K	PIK3CA	chr3: 178936091	9.0%	QUANTR.

Table A.1: Sample BRAF20 that were purchased from Horizon with their annotated mutation rate used for a first evaluation of the performance of the comprehensive cancer panel. Mutations are verified by Sanger sequencing, quality was tested with agarose gel electrophoresis and qPCR, quantification was performed with Quantifluor™, all performed by Horizon.

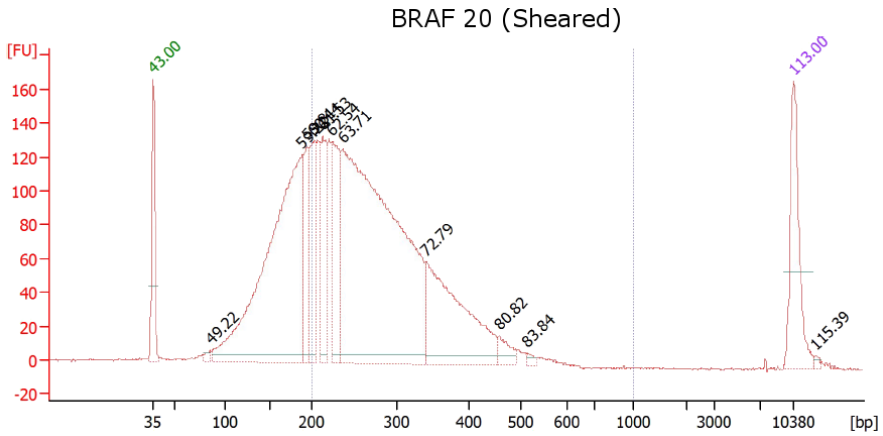


Figure A.1: BRAF20 after shearing.

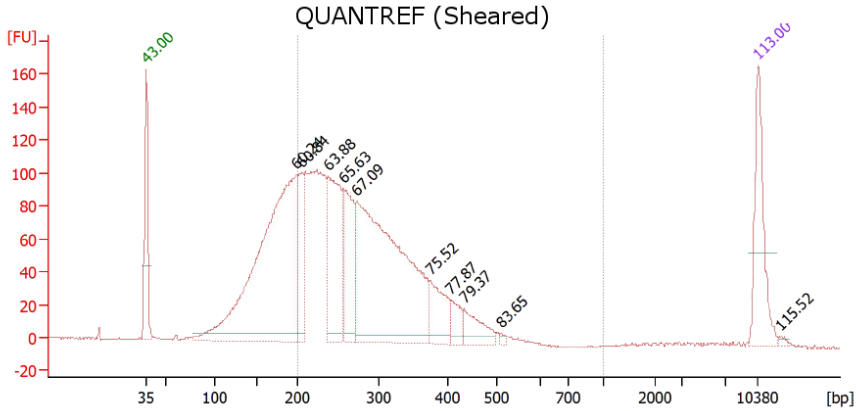


Figure A.2: QUANTREF after shearing.

A.2 Pilot Bioanalyzer Traces

Peak	Concentration (in pg/μl)
BRAF20 (Shearing)	

85	8.24
188	763.59
194	92.94
201	122.25
212	127.39
223	153.96
236	1,055.10
340	264.60
458	20.60
513	4.60
QUANTREF (Shearing)	
199	650.66
205	93.71
238	201.82
257	130.07
273	556.78
374	64.01
406	26.32
432	39.20

509	3.23
BRAF20 (Library)	
45	134.52
284	12,274.09
417	4,118.55
424	7,221.48
1,031	43.88
1,331	31.83
1,499	25.70
1,646	21.72
1,777	46.52
2,183	50.46
3,102	39.82
5,353	14.25
8,078	4.22
QUANTREF (Library)	
50	16.11
284	2,487.50
331	4,117.19

344	8,007.86
454	3,778.56
877	35.98
1,027	23.68
1,198	16.40
1,629	16.54
1,875	14.22
2,062	27.25
2,683	14.04
3,210	20.99
5,425	7.52
6,626	3.18
7,633	4.22
8,759	2.22
Version 1 (Pool)	
300	84.29
312	17.34
327	26.06
331	16.69

348	27.05
360	21.11
370	17.59
385	22.36
404	22.31
477	7.69
491	8.24
528	10.32

Table A.2: Concentration of Bioanalyzer Peaks from kit version 1

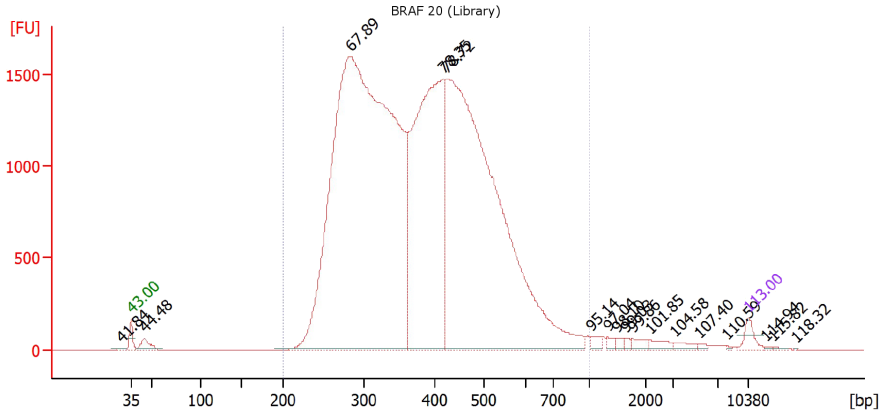


FIGURE A.3: BRAF 20 after library amplification

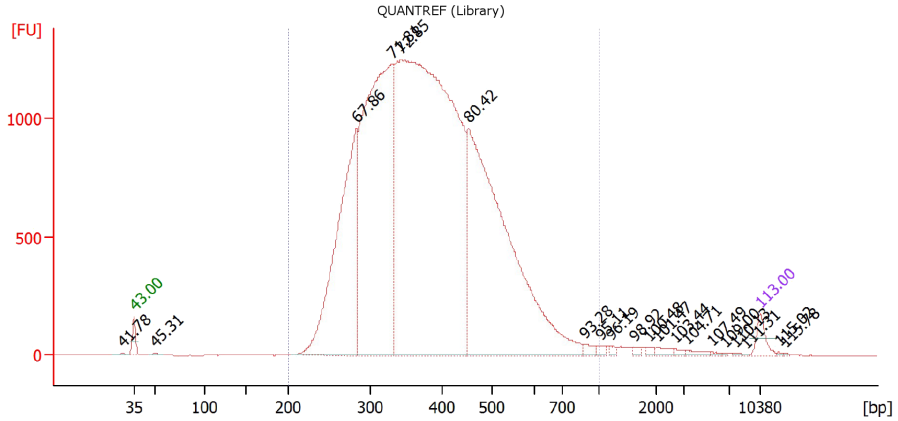


FIGURE A.4: QUANTREF after library amplification

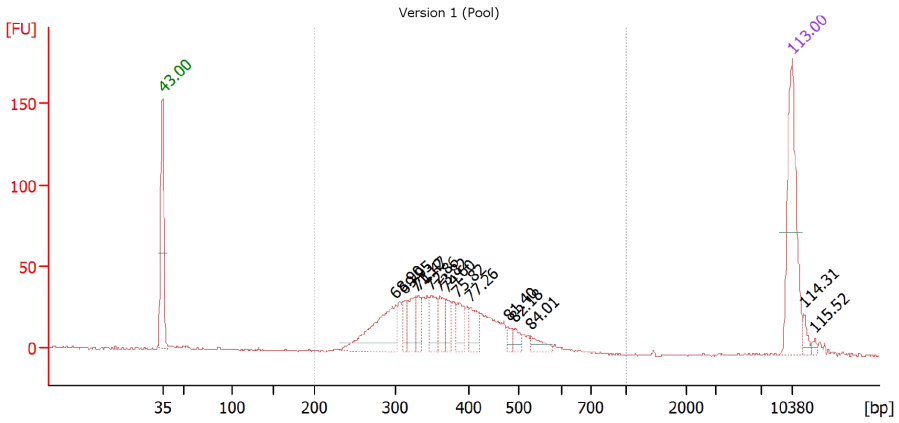


FIGURE A.5: Bioanalyzer traces of samples treated with kit version 1. Concentrations are given in table A.2

A.3 Agilent Sure Select XT Custom Library Preparation Protcol

1. Random DNA shearing by sonication to generate DNA fragments of 150bp-200bp in size using the following settings on the Covaris instrument:

Setting	Value
Duty factor	10%
Peak incident power	175
Cycles per burst	200
Treatment time	360 seconds
Bath temperature	4 °C – 8 °C

2. assessment of the fragment size distribution and quantification using the Agilent Bioanalyzer 2100 instrument

3. Repair the fragment ends, for subsequent adapter ligation

   (a) Prepare the end repair master mix for each sample according to the Agilent protocol:

Reagent	Volume per reaction
Nuclease-free water	35.2μl
10x end repair buffer	10μl
dNTP mix	1.6μl
T4 DNA polymerase	1μl
Klenow DNA polymerase	2μl
T4 Polynucleotide Kinase	2.2μl
Total	52μl

(b) Add 52μl to each sample of about 50μl.

(c) Mixing by pipetting up and down several times.

(d) Incubate mix at 20 °C for 30 minutes

4. DNA purification using magnetic SPRI beads, such as Ampure XP beads from Beckman Coulter

   (a) Add 180μl bead suspension to each sample and pipette up and down 10 times for mixing

   (b) Incubate samples for 5 minutes at room temperature to let DNA bind to the carboxyl molecules that coat the beads.

   (c) Put the sample tube or plate into a magnetic separation device to pull down beads with bound DNA and wait 3-5 minutes until solution is completely clear

(d) Carefully remove the supernatant which contains enzymes, dNTPs, buffer and DNA fragments < $100bp$ due to their low electrostatic properties [239]

(e) While keeping the beads in the magnetic separation device wash beads twice for one minute in 200μl of 70% ethanol per sample

(f) Seal the tube or plate, spin briefly to collect residual ethanol and put back to the magnetic separation device and remove the ethanol

(g) Dry samples by removing the seal and keep samples at 37 °C for 3 – 5 minutes until all ethanol has evaporated

(h) Resuspend beads in 32μl nuclease-free water to each sample

(i) Vortex well and spin briefly until all liquid has been collected at the bottom of the tube

(j) Incubate for 2 minutes at room temperature, then put on to the magnetic separation device for 2-3 minutes until solution is clear. The DNA has separated from the beads due to the low salt concentration.

(k) Transfer supernatant to a fresh tube or PCR plate. The beads are discarded

5. 3' adenylation of the fragments for downstream for subsequent adapter ligation

(a) Prepare the adenylation master mix according to the Agilent protocol:

Reagent	Volume per reaction
Nuclease-free water	11μl
10x Klenow polymerase buffer	5μl
dATP	1μl
Klenow DNA polymerase	3μl
Total	20μl

(b) Add 20μl to approximately 30μl of end-repaired and purified DNA sample and mix well py pipetting up and down several times

(c) Incubate for 30 minutes at 37°C to let the Klenow polymerase add a 3' adenosine to each fragment

6. DNA purification using magnetic SPRI beads, such as Ampure XP beads from Beckman Coulter

(a) Add 90μl bead suspension to each 50μl sample and pipette up and down 10 times for mixing

(b) Incubate samples for 5 minutes at room temperature to let DNA bind to the carboxyl molecules that coat the beads.

(c) Put the sample tube or plate into a magnetic separation device to pull down beads with bound DNA and wait 3-5 minutes until solution is completely clear

(d) Carefully remove the supernatant which contains enzymes, dNTPs, buffer and DNA fragments $< 100bp$ due to their low electrostatic properties

(e) While keeping the beads in the magnetic separation device wash beads twice for one minute in 200μl of 70% ethanol per sample

(f) Seal the tube or plate, spin briefly to collect residual ethanol and put back to the magnetic separation device and remove the ethanol

(g) Dry samples by removing the seal and keep samples at 37 °C for 1 – 2 minutes until all ethanol has evaporated

(h) Resuspend beads in 15μl nuclease-free water to each sample

(i) Vortex well and spin briefly until all liquid has been collected at the bottom of the tube

(j) Incubate for 2 minutes at room temperature, then put on to the magnetic separation device for 2-3 minutes until solution is clear. The DNA has separated from the beads due to the low salt concentration.

(k) Transfer 13μl supernatant to a fresh tube or PCR plate. The beads are discarded

7. Ligation of Illumina sequencing adapters

   (a) Prepare the ligation master mix accoding to Agilent protocol:

Reagent	Volume per reaction
Nuclease-free water	15.5μl
5x T4 DNA ligase buffer	10μl
Diluted SureSelect adapter oligo mix	10μl
T4 DNA ligase	1.5μl
Total	37μl

   (b) Add 37μl to each 13μl sample and mix well py pipetting up and down

   (c) Incubate the mix for 15 minutes at 20 °C

8. DNA purification using magnetic SPRI beads, such as Ampure XP beads from Beckman Coulter

   (a) Add 90μl bead suspension to each 50μl sample and pipette up and down 10 times for mixing

   (b) Incubate samples for 5 minutes at room temperature to let DNA bind to the carboxyl molecules that coat the beads.

(c) Put the sample tube or plate into a magnetic separation device to pull down beads with bound DNA and wait 3-5 minutes until solution is completely clear

(d) Carefully remove the supernatant which contains enzymes, dNTPs, buffer and DNA fragments $< 100bp$ due to their low electrostatic properties

(e) While keeping the beads in the magnetic separation device wash beads twice for one minute in 200μl of 70% ethanol per sample

(f) Seal the tube or plate, spin briefly to collect residual ethanol and put back to the magnetic separation device and remove the ethanol

(g) Dry samples by removing the seal and keep samples at 37 °C for 1 – 2 minutes until all ethanol has evaporated

(h) Resuspend beads in 32μl nuclease-free water to each sample

(i) Vortex well and spin briefly until all liquid has been collected at the bottom of the tube

(j) Incubate for 2 minutes at room temperature, then put on to the magnetic separation device for 2-3 minutes until solution is clear. The DNA has separated from the beads due to the low salt concentration.

(k) Transfer at least 30μl supernatant to a fresh tube or PCR plate. The beads are discarded

9. Library amplification

   (a) Prepare the pre-capture PCR master mix according to the Agilent protocol:

Reagent	Volume per reaction
Nuclease-free water	6μl
SureSelect primer	1.25μl
SureSelect ILM indexing pre-capture PCR reverse primer	1.25μl
5x Herculase II reaction buffer	10μl
100mM dNTP mix	0.5μl
Herculase II Fusion DNA polymerase	1μl
Total	20μl

   (b) Add 20μl to each 30μl sample and mix by pipetting up and down. Run the PCR on a thermal cycler:

      i. 98 °C for 2 minutes
      ii. 98 °C for 30 seconds
      iii. 65 °C for 30 seconds
      iv. 72 °C for 1 minute

v. Repeat steps ii. - iv. 9 times

vi. 72 °C for 10 minutes

vii. 4 °C for ∞

10. DNA purification using magnetic SPRI beads, such as Ampure XP beads from Beckman Coulter

    (a) Add 90μl bead suspension to each 50μl sample and pipette up and down 10 times for mixing

    (b) Incubate samples for 5 minutes at room temperature to let DNA bind to the carboxyl molecules that coat the beads.

    (c) Put the sample tube or plate into a magnetic separation device to pull down beads with bound DNA and wait 3-5 minutes until solution is completely clear

    (d) Carefully remove the supernatant which contains enzymes, dNTPs, buffer and DNA fragments < 100bp due to their low electrostatic properties

    (e) While keeping the beads in the magnetic separation device wash beads twice for one minute in 200μl of 70% ethanol per sample

    (f) Seal the tube or plate, spin briefly to collect residual ethanol and put back to the magnetic separation device and remove the ethanol

    (g) Dry samples by removing the seal and keep samples at 37 °C for 1 – 2 minutes until all ethanol has evaporated

(h) Resuspend beads in 30μl nuclease-free water to each sample

(i) Vortex well and spin briefly until all liquid has been collected at the bottom of the tube

(j) Incubate for 2 minutes at room temperature, then put on to the magnetic separation device for 2-3 minutes until solution is clear. The DNA has separated from the beads due to the low salt concentration.

(k) Transfer at least 30μl supernatant to a fresh tube or PCR plate. The beads are discarded

11. Assessment of the fragment size distribution and quantification using the Agilent Bioanalyzer 2100 instrument

A.4 Agilent SureSelect XT Hybridisation and Capture

1. Hybridisation of library DNA to capture probes

   (a) 750ng are needed in a volume of 3.4μl. Either dilute with nuclease-free water or use a vacuum concentrator

      i. Add the entire 30μl DNA library to a 1.5ml microcentrifuge tube. Break off cap and cover with parafilm and poke a hole in the parafilm

      ii. Dehydrate with vacuum concentrator on $\leq 45\,^{\circ}C$

iii. Refill to a final concentration of $221\frac{\text{ng}}{\mu\text{l}}$ and mix well by pipetting, vortexing and spinning.

iv. Transfer library at desired concentration to new tube

Prepare the hybridisation buffer according to Agilent protocol:

Reagent	Volume per reaction
Surelect Hyb 1	6.63μl
Surelect Hyb 2	0.27μl
Surelect Hyb 3	2.65μl
Surelect Hyb 4	3.45μl
Total	13μl

(b) Prepare the SureSelect block mix according to Agilent protocol:

Reagent	Volume per reaction
SureSelect indexing block 1	2.5μl
Surelect block 2	2.5μl
Surelect ILM indexing block 3	0.6μl
Total	5.6μl

(c) Add to each library 5.6μl of the SureSelect block mix and mix well by pipetting up and down

(d) Seal the tube/plate well and incubate 5 minutes at 95 °C, followed by at least 5 minutes at 65 °C

(e) Prepare 5μl of a 10% dilution of SureSelect RNAse block and keep on ice

(f) Prepare capture library hybridisation mix for < 3Mb targets according to Agilent protocol:

Reagent	Volume per reaction
Hybridisation buffer mixture from step	13μl
10% RNAse block solution from step	5μl
Capture library	2μl
Total	20μl

(g) Add 20μl capture library hybridisation mix to library and SureSelect block mix while still at 65 °C and mix well by pipetting up and down 10 times

(h) Incubate entire hybridisation mix for 16 hours at 65 °C with a heated lid at 105 °C to let probes bind to complementary ssDNA

2. Capture hybridised DNA using Dynabeads MyOne Streptavidin T1 magnetic beads

   (a) Transfer the entire hybridisation mix to a tube containing 200μl Dynabeads MyOne Streptavidin T1 magnetic beads and mix well by pipetting up and down

   (b) Cap wells to avoid evaporation and incubate at room temperature for 30 minutes to let the probes bind to the streptavidin

   (c) Briefly spin the tube or plate in a centrifuge until all liquid is collected at the bottom

   (d) Place next to a magnetic separation device for 3 minutes until solution is clear

   (e) Remove and discard supernatant containing DNA library fragments that were not targeted

   (f) Resuspend beads in 200μl SureSelect wash buffer 1 and mix by pipetting up and down

   (g) Incubate for 15 minutes at room temperature and repeat steps (d) and (e)

   (h) Resuspend beads with 65 °C SureSelect wash buffer 2 by pipetting up and down

   (i) Cap the well or tube and incubate for 10 minutes at 65 °C

   (j) Place next to a magnetic separation device for 3 minutes until solution is clear

(k) Remove and discard supernatant containing DNA library fragments that were not targeted

(l) Repeat steps (h) to (k) twice

(m) Resuspend beads in 30μl nuclease-free water

3. Amplification and 6bp indexing/barcoding of captured DNA library

(a) Prepare post-capture PCR mix according to Agilent protocol:

Reagent	Volume per reaction
Nuclease-free water	22.5μl
5x Herculase II reaction buffer	10μl
Herculase II Fusion DNA polymerase	1μl
100mM dNTP mix	0.5μl
SureSelect ILM indexing post-capture forward PCR primer	1μl
Total	35μl

(b) Prepare 35μl post-capture PCR mix per sample in a fresh well of a PCR plate or a fresh tube

(c) Add 1μ of appropriate SureSelect PCR primer index 1-16 to each sample, so each sample is uniquely barcoded

(d) Add 14μl bead-bound target-enriched DNA to the mix and mix well by pipetting up and down, the rest can be kept at −20 °C

(e) Transfer plate or tube to thermal cycler and run the following PCR programme:

i. 98 °C for 2 minutes

ii. 98 °C for 30 seconds

iii. 57 °C for 30 seconds

iv. 72 °C for 1 minute

v. Repeat steps ii. - iv. 16 times

vi. 72 °C for 10 minutes

vii. 4 °C for ∞

4. DNA purification using magnetic SPRI beads, such as Ampure XP beads from Beckman Coulter

(a) Add 90μl bead suspension to each 50μl sample and pipette up and down 10 times for mixing

(b) Incubate samples for 5 minutes at room temperature to let DNA bind to the carboxyl molecules that coat the beads.

(c) Put the sample tube or plate into a magnetic separation device to pull down beads with bound DNA and wait 3-5 minutes until solution is completely clear

(d) Carefully remove the supernatant which contains enzymes, dNTPs, buffer and DNA fragments < $100bp$ due to their low electrostatic properties

(e) While keeping the beads in the magnetic separation device wash beads twice for one minute in 200μl of 70% ethanol per sample

(f) Seal the tube or plate, spin briefly to collect residual ethanol and put back to the magnetic separation device and remove the ethanol

(g) Dry samples by removing the seal and keep samples at 37 °C for 1 – 2 minutes until all ethanol has evaporated

(h) Resuspend beads in 30μl nuclease-free water to each sample

(i) Vortex well and spin briefly until all liquid has been collected at the bottom of the tube

(j) Incubate for 2 minutes at room temperature, then put on to the magnetic separation device for 2-3 minutes until solution is clear. The DNA has separated from the beads due to the low salt concentration.

(k) Transfer at least 30μl supernatant to a fresh tube or PCR plate. The beads are discarded

5. Assessment of the fragment size distribution and quantification using the Agilent Bioanalyzer 2100 instrument

6. Libraries can finally be pooled according to determined concentration

A.5 Demultiplexing for Agilent SureSelect XT Libraries

```
bcl2fastq -o Fastq_Raw --no-lane-splitting --use-bases-mask y150n,i6,y150n \
--sample-sheet run_SureSelect.csv > demultiplexing.log
```

A.6 Bioinformatics Commands

```
#Alignment with BWA mem 0.7.12 and sorting
bwa mem -M hg19.fa [sample]_r1.fastq [sample]_r2.fastq | \
samtools view -bT  hg19.fa - | \
samtools sort - [sample]_sorted.bam

#Mark Duplicates with Picardtools
java -jar picard.jar MarkDuplicates I=[sample]_sorted.bam\
R=hg19.fa METRICS_FILE=[sample]_metrics.txt O=[sample]_dedup.bam\
CREATE_INDEX=true

#Re-alignment with Abra-0.96
#Note: target_regions must be sorted and comments removed
java -jar abra-0.96.jar --in [sample]_dedup.bam --out\
[sample]_realigned.bam --ref hg19.fa --working abra_tmp \
--targets target_regions.bed --mad 1000 --mnf 5 --mbq 150\
--ib --maxn 50000
```

```
#Variant calling
samtools mpileup -l target_regions_sorted.bed -f hg19.fa -B\
-d 10000 -C50 [sample]_final.bam | java -jar VarScan.v2.3.9.jar\
mpileup2cns --min-var-freq 0.01 --output-vcf 1  --variants 1\
--p-value 0.05 > [sample]_raw.vcf

# Revised Variant calling from HaloPlex HS data
 samtools mpileup -l target_regions_sorted.bed -f hg19.fa -A -B\
-d 10000 -C50 [sample]_final.bam | java -jar VarScan.v2.3.9.jar\
mpileup2cns --min-var-freq 0.01 --output-vcf 1  --variants 1\
--p-value 0.05 > [sample]_raw.vcf

# Generating a readcount file for the fpfilter in two steps
#1) Write the vcf file to a bed file
vcf2bed < [sample]_raw.vcf > [sample]_raw.bed && bedtools slop -i\
[sample]_raw.bed -g human.hg19.genome -b 50  > [sample]_padded.bed
2) Generate a readcount
bam-readcount -f hg19.fa [sample]_final.bam -l [sample]_padded.bed \
> [sample]_readcount.metrics
#Variant filtering
java -jar VarScan.v2.3.9.jar fpfilter [sample]_raw.vcf \
[sample]_readcount.metrics --output-file [sample]_fpfiltered.vcf \
--keep-failures --min-var-freq 0.05 --min-var-count 8 \
--min-ref-basequal 28 --min-var-basequal 30
#Note: Fixing bug in the readcount/fpfilter module.
#Deletions are always filtered
sed -i \
"s/\([[:space:]]\)\(NoReadCounts\)\([[:space:]]\)/\1PASS\3/"\
[sample]_fpfiltered.vcf
#Note: Fixing a bug in fpfilter module. Homozygous
#variants are always filtered
sed -i \
"s/\([[:space:]]\)\(RefBaseQual\)\([[:space:]]\)
```

```
\(.*HOM=1\)/\1PASS\3\4/"\  [sample]_fpfiltered.vcf
#Annotation with snpEFF and snpSIFT
java -jar SnpSift.jar annotate -c snpEff.config -dbsnp\
[sample]_fpfiltered.vcf > [sample]_dbSNP.vcf
java -jar SnpSift.jar annotate -c snpEff.config cosmic.vcf\
[sample]_fpfiltered.vcf > [sample]_cosmic.vcf
java -jar SnpSift.jar annotate -c snpEff.config -clinvar\
[sample]_fpfiltered.vcf > [sample]_clinVar.vcf
java -jar SnpSift.jar dbnsfp -c snpEff.config -collapse
[sample]_fpfiltered.vcf > [sample]_dbNSFP.vcf
java -jar snpEff.jar -s [sample]_snpeff_summary.html -c \
snpEff.config -q hg19 [sample]_fpfiltered.vcf > [sample]_snpEff.vcf
```

Commands for variant calling, filtering and annotation. Bam-readcount file is similar to a pileup, but contains more information about the per-base coverage, needed by VarScan2 [269]. SnpSift is used for database annotations, such as dbSNP [109], clinvar [60] and dbNSFP [270] or COSMIC [34] or any possible combination. snpEff predicts the structural change, adds the gene ID and many more information. Unfortunately, the false-positive filter contains two bugs in version 2.3.9. Firstly, it is not testing if a variant is a deletion, which means no reads support the variant and therefore is filtered. Secondly, homozygous variants are always flagged as false variant, because the base quality of reference supporting reads is too low, as there are none.

Parameter	**Value**	**Description/Reasons**
Minimum variant-supporting reads	8	This ensures a high confidence in variant calls, given sufficient coverage
Minimum allele frequency	0.05	If coverage is high enough, low-frequency variants can be called reliably
Minimum average read position of variant-supporting reads	0.1	Avoid structural read bias
Minimum average relative distance to 3' end	0.1	Avoid a variant to be called from reads with low quality towards 3' end.
Minimum variant-supporting strandness	0.01	Ensures that variant is seen on both strands

Minimum allele coverage to perform strandness test	5	For lower numbers statistical testing is not feasible and is showing poor support
Minimum average base quality for reference allele	28	Reference alleles are likely to be seen, hence quality can be a little lower
Minimum average base quality for variant allele	30	Variant allels need to be of high cofidence
Maximum average relative read-length difference	0.25	If reads are massively soft-clipped, the reads cannot be trusted.
Maximum mismatch quality sum of variant-supporting reads	100	If reads contain too many mismatches, the variant position cannot be trusted

Minimum average mapping quality for reference allele	30	Similar to the base quality, although mapping quality is not standardised
Minimum average mapping quality for variant allele	30	Similar to the base quality, although mapping quality is not standardised
Maximum average mapping quality difference	50	If quality differs too much between reference- and variant-supporting reads, it indicates a sequencing or alignment problem

Table A.3: Criteria for VarScan2's false-positive filter module. Most values have been set according to best practice guidelines predefined by VarScan2. Some have been adjusted specifically for calling variants from pooled samples without a tumour-normal pair sequencing, such as minimum allele frequency, minimum number of variant-supporting reads and minimum average base quality for reference allele.

Appendix B

Clinical Samples and Read Collapsing

B.1 Samples for Agilent SureSelect XT Panel Validation

Sample ID	Source	Type	DIN	Conc.[2]
		Pool 1		
14R011537	SBS	NSCLC	5.8	204
14R011557	SBS	CRC	6.3	4.77
14R011558	SBS	CRC	N/A (0)	87.9

14R011569	SBS	CRC	4.2	24.5
14R011570	SBS	CRC	4.4	98.1
14R011571	SBS	CRC	N/A (0)	1.13
14R011572	SBS	CRC	3.4	12.3
14R011573	SBS	NSCLC	2.7	0.706
14R011610	SBS	CRC	4.2	5.13
14R011611	SBS	CRC	5.7	120
14R011612	SBS	CRC	1.7	10.6
14R011613	SBS	CRC	2.8	273
14R011614	SBS	CRC	5.9	47.7
14R011615	SBS	CRC	5.9	15.1
14R011616	SBS	CRC	3.3	98.1
14R011617	SBS	CRC	5.6	101
14R011618	SBS	CRC	2.2	15.7
14R011643	SBS	CRC	6.2	15.9
14R011644	SBS	CRC	1.7	6.38
14R011646	SBS	CRC	6.4	4.98

14R011647	SBS	CRC	2.5	19.1
14R011648	SBS	CRC	5.6	11
14R011680	SBS	CRC	3.2	5.33
14R011681	SBS	NSCLC	2.9	2.47
14R011687	SBS	NSCLC	3.9	3.09
14R011689	SBS	NSCLC	6.3	8.1
14R011691	SBS	NSCLC	5.9	20.7
14R011693	SBS	NSCLC	6.2	28.4
14R011695	SBS	CRC	1.6	7.61
14R011696	SBS	CRC	2.6	24.0
14R011697	SBS	CRC	1.5	11.3
14R011755	SBS	CRC	3.1	60.1
14R011756	SBS	CRC	2.7	103
14R011757	SBS	CRC	5.7	13.6
14R011758	SBS	CRC	5.7	16.1
14R011759	SBS	CRC	2.7	40.8
14R011760	SBS	CRC	5.9	88.4
14R011761	SBS	CRC	2.6	7.84

14R011762	SBS	CRC	2.5	52.6
14R011763	SBS	CRC	5.2	18.5
14R011764	SBS	CRC	1.7	13.1
14R011767	SBS	NSCLC	5.2	15.0
14R011769	SBS	NSCLC	N/A (0)	0.42
14R011771	SBS	NSCLC	N/A (0)	0.103
14R011773	SBS	NSCLC	5.8	15.5
14R011775	SBS	NSCLC	N/A (0)	0.478
14R011777	SBS	NSCLC	N/A (0)	0.62
14R011822	SBS	NSCLC	4.6	0.47
Pool 2				
14R001551	SBS	Melanoma	N/A (0)	1.35
14R003348	SBS	CRC	1.3	23.2
14R006030	EQA	NSCLC	2.4	64.6
14R006031	EQA	NSCLC	5.8	8.29

14R006032	EQA	NSCLC	3.0	52.4
14R006034	EQA	NSCLC	2.8	18.1
14R006035	EQA	Melanoma	3.5	25.7
14R006036	EQA	Melanoma	2.6	27.7
14R006037	EQA	Melanoma	2.4	15.7
14R006038	EQA	Melanoma	3.1	10.8
14R006039	EQA	Melanoma	2.8	93.6
14R006040	EQA	CRC	2.2	37.2
14R006041	EQA	CRC	2.6	18.9
14R006042	EQA	CRC	3.7	22.1
14R006043	EQA	CRC	2.3	32.2
14R006044	EQA	CRC	2.3	22.7
14R010290	EQA	Multi	7.1	12.4
14R011645	SBS	CRC	3.7	6.27
14R011693B	SBS	NSCLC	6.2	12.3
14R011779	SBS	NSCLC	7.2	61.1
14R011821	SBS	CRC	5.6	5.59
14R011838	SBS	NSCLC	2.4	7.2

14R011860	SBS	CRC	2.6	180.0
14R011861	SBS	CRC	4.0	2.48
14R011864	SBS	Melanoma	3.2	2.87
14R011866	SBS	CRC	4.9	8.56
14R011867	SBS	CRC	5.5	4.22
14R011910	SBS	CRC	2.7	20.2
14R011911	SBS	CRC	6.2	16.6
14R011912	SBS	CRC	2.9	8.79
14R011949	SBS	CRC	3.2	62.7
14R011950	SBS	CRC	4.9	13.0
14R011952	SBS	CRC	2.8	116.0
14R011954	SBS	CRC	N/A	149
14R011955	SBS	CRC	N/A	42
14R011956	SBS	CRC	N/A	49.1
14R011957	SBS	CRC		12.9
14R011958B	SBS	NSCLC	N/A	26.9
14R011958A	SBS	CRC	2.5	18.5
14R011960	SBS	NSCLC	N/A	3.1

14R011962	SBS	NSCLC	N/A	1.01
14R012020	SBS	CRC	6.8	78.8
14R012026	SBS	CRC	N/A	4.66
14R012036	SBS	NSCLC	N/A	6.94
14R012039	SBS	NSCLC	N/A	16.1
14R012049-1A	SBS	CRC	N/A	104.0
14R012056	SBS	CRC	6.0	21.5
14R012060	SBS	NSCLC	N/A	0.821
Pool 3				
14R012024	SBS	CRC	3.9	4.14
14R012034	SBS	CRC	4.4	29.0
14R012049-2A	SBS	CRC	N/A (0)	87.2
14R012059	SBS	CRC	4.9	71.0
14R012061	SBS	CRC	4.7	45.7
14R012118	SBS	Melanoma	3.9	54.0
14R012119	SBS	CRC	3.8	57.0

14R012120	SBS	CRC	6.5	57.0
14R012121	SBS	CRC	3.1	52.0
14R012122	SBS	CRC	2.7	74.3
14R012123	SBS	CRC	3.6	41.3
14R012124	SBS	CRC	6.5	27.0
14R012125	SBS	CRC	5.1	53.3
14R012126	SBS	CRC	3.5	160.0
14R012127	SBS	CRC	2.7	169.0
14R012128	SBS	CRC	5.6	42.1
14R012129	SBS	CRC	3.0	51.0
14R012130	SBS	NSCLC	6.3	47.3
14R012131	SBS	NSCLC	1.8	17.7
14R012133	SBS	NSCLC	3.7	19.0
14R012134	SBS	NSCLC	6.8	8.53
14R012138	SBS	NSCLC	4.1	7.94
14R012181	SBS	CRC	6.5	20.5
14R012182	SBS	CRC	5.9	40.3
14R012183	SBS	CRC	4.5	39.1

14R012184	SBS	CRC	N/A (0)	0.533
14R012185	SBS	CRC	5.9	1.3
14R012186	SBS	CRC	3.7	118.0
14R012187	SBS	CRC	3.5	142.0
14R012188	SBS	CRC	6.2	37.8
14R012189	SBS	CRC	4.2	4.52
14R012190	SBS	CRC	6.6	12.2
14R012193	SBS	NSCLC	5.9	4.39
14R012195	SBS	NSCLC	4.3	10.1
14R012233	SBS	CRC	3.4	30.3
14R012234	SBS	CRC	3.2	59.0
14R012235	SBS	CRC	2.6	28.5
14R012236	SBS	CRC	3.0	40.1
14R012237	SBS	CRC	3.3	59.0
14R012238	SBS	CRC	5.6	40.8
14R012239	SBS	NSCLC	6.3	5.3
14R012242	SBS	CRC	5.8	32.4

14R012276	SBS	CRC	3.7	127.0
14R012290	SBS	CRC	6.2	45.2
14R012291	SBS	CRC	5.8	73.5
14R012292	SBS	CRC	3.9	92.0
14R012293	SBS	CRC	2.9	53.0
14R012294	SBS	CRC	2.8	43.4
Pool 4				
14R010267	EQA	CRC	3.1	10.7
14R010268	EQA	CRC	4.0	32.6
14R010270	EQA	CRC	1.9	6.32
14R010271	EQA	CRC	3.5	11.5
14R010272	EQA	NSCLC	3.4	8.45
14R010273	EQA	NSCLC	2.9	16.3
14R010274	EQA	NSCLC	5.7	14.9
14R010275	EQA	NSCLC	3.0	13.4
14R010276	EQA	Melanoma	2.2	25.0
14R010277	EQA	Melanoma	2.8	47.4
14R010278	EQA	Melanoma	2.1	14.7

14R010279	EQA	Melanoma	3.7	8.06
14R012043	SBS	NSCLC	4.5	20.5
14R012132	SBS	NSCLC	2.9	46.4
14R012191	SBS	NSCLC	6.3	27.9
14R012240	SBS	NSCLC	5.6	10.9
14R012244	SBS	NSCLC	N/A (0)	0.728
14R012321	SBS	CRC	1.9	6.95
14R012326	SBS	CRC	2.4	28.1
14R012328	SBS	CRC	3.9	32.4
14R012329	SBS	CRC	2.5	32.7
14R012330	SBS	NSCLC	6.9	5.45
14R012333	SBS	CRC	6.2	10.8
14R012334	SBS	CRC	2.1	28.1
14R012335	SBS	NSCLC	3.5	9.41
14R012337	SBS	NSCLC	5.7	4.39
14R012339A	SBS	NSCLC	4.3	12.5
14R012339B	SBS	NSCLC	3.6	33.5

14R012374	SBS	CRC	6.3	31.5
14R012375	SBS	CRC	4.0	129.0
14R012377	SBS	CRC	1.9	33.4
14R012380	SBS	CRC	7.3	8.94
14R012382	SBS	CRC	2.3	41.9
14R012402	SBS	CRC	5.7	35.3
14R012403	SBS	NSCLC	N/A (0)	2.25
14R012436	SBS	NSCLC	1.6	9.62
14R012454	SBS	CRC	3.1	22.9
14R012455	SBS	Melanoma	4.5	4.33
14R012460	SBS	CRC	2.4	39.2
14R012461	SBS	CRC	2.0	9.18
14R012341	SBS	CRC	3.3	8.66
14R012383	SBS	CRC	1.1	18.3
14R012384	SBS	CRC	6.1	38.8
14R012439	SBS	NSCLC	6.5	3.1
14R012448	SBS	NSCLC	4.3	12.2

14R012452	SBS	Melanoma	2.8	8.94
14R012458	SBS	CRC	3.2	13.2
14R012496	SBS	NSCLC	6.5	14.6
Pool 5				
H04-0020636	BI	Prostate	N/A (0)	2.68
H07-0016101	BI	Prostate	N/A (0)	3.26
H08-0000803	BI	Prostate	1.1	4.17
H08-0009827	BI	Prostate	N/A (0)	2.96
H09-0003157	BI	Prostate	1.0	1.75
H09-0015981	BI	Prostate	1.1	2.42
H09-0022497	BI	Prostate	1.1	2.58
H09-0024365	BI	Prostate	N/A (0)	1.73
H10-0000806	BI	Prostate	N/A (0)	1.76
H10-0003176	BI	Prostate	1.2	5.6
H10-0007699	BI	Prostate	1.0	4.56

H11-0000751	BI	Prostate	1.0	2.52
H12-0005377	BI	Prostate	1.0	3.03
H12-0010556	BI	Prostate	1.0	1.99
H12-0011977	BI	Breast	N/A (0)	10.6
H12-0018944	BI	Prostate	1.9	3.51
H12-0019484	BI	Prostate	1.8	41.8
H12-0020486	BI	Breast	1.8	95.7
H12-0020515	BI	Prostate	1.5	32.4
H12-0020586	BI	Prostate	N/A (0)	35.3
H12-0020607	BI	Breast	3.2	1.6
H12-0021258	BI	Breast	3.3	10.2
H12-0021538	BI	Prostate	2.3	41.7
H12-0021574	BI	Breast	2.3	13.2
H12-0021728	BI	Prostate	2.3	64.8
H12-0021736	BI	Prostate	1.1	4.02
H12-0022064	BI	Breast	1.5	1.7

H12-0022066	BI	Breast	2.0	1.32
H12-0022738	BI	Breast	3.3	13.7
H12-0023107	BI	Prostate	1.6	143.0
H12-0023109	BI	Prostate	1.9	29.7
H12-0023396	BI	Breast	1.7	14.9
H12-0023436	BI	Prostate	N/A	1.76
H12-0023657	BI	Prostate	2.1	44.4
H12-0023881	BI	Breast	2.5	25.6
H12-0026932	BI	Breast	2.9	10.3
H13-0000392	BI	Breast	2.5	10.7
H13-0002251	BI	Breast	2.3	17.3
H13-23147	BI	CRC	3.1	23.2
H14-0007483	BI	Breast	2.0	25.6
H14-0008076	BI	Breast	2.8	8.69
H14-0009506	BI	Breast	1.2	3.59
H14-0009690	BI	Breast	1.8	2.63
Pool 6				
246812-A12	SBS	Other	3.2	46.2

H07-18185	BI	Melanoma	1.1	8.75
H09-2347-B1	BI	Other	6.5	4.44
H10-16659-1	BI	Other	2.3	57.6
H12-0015672	BI	CRC	1.8	76.3
H12-0024078	BI	Prostate	1.8	58.1
H12-0024623	BI	Prostate	1.6	64.9
H12-0024858	BI	Prostate	2.5	24.0
H12-0025258	BI	Prostate	2.0	79.7
H12-03931	BI	Melanoma	2.9	228.0
H12-13980-A5	BI	Other	3.4	80.5
H12-14836-10	BI	Other	2.2	55.8
H12-22598	BI	Melanoma	1.6	21.4
H12-25779	BI	Melanoma	2.7	25.7
H12-3680	BI	Other	1.8	3.0
H12-8147	BI	CRC	2.8	40.3
H13-00527A	BI	Melanoma	2.7	72.9
H13-00527B	BI	Melanoma	2.5	114.0

H13-02640	BI	Melanoma	3.1	5.13
H13-12973	BI	CRC	1.8	15.4
H13-17391	BI	Melanoma	3.2	88.1
H13-17420	BI	Other	3.2	47.1
H13-18218	BI	CRC	2.1	49.9
H13-18868	BI	CRC	2.5	104.0
H13-18879	BI	CRC	2.3	92.2
H13-1909-6	BI	Other	4.5	31.5
H13-20944	BI	CRC	2.7	17.1
H13-21073	BI	CRC	3.2	37.6
H13-22140	BI	CRC	1.8	16.4
H13-22212	BI	CRC	2.2	27.2
H13-25256	BI	Melanoma	3.3	26.0
H13-25706	BI	Melanoma	2.8	109.0
H13-26294	BI	Melanoma	2.2	27.8
H13-8045-4	BI	Other	2.0	25.7
H14-00544	BI	Melanoma	1.6	29.0
H14-7316-18	BI	Other	2.3	102.0

H14-8067	BI	Breast	2.6	18.3
H14-9300	BI	Breast	1.4	7.9
H15-1177	BI	Other	6.3	23.8
H15-1420-A2	BI	Other	4.9	229.0
H15-2165-4	BI	Other	1.7	51.3
H15-360	BI	Other	5.1	5.83
H15-901	BI	Other	5.7	6.9

TABLE B.1: Overview of all sequenced and analysed clinical samples. Tissue was provided by Source BioScience (SBS), Birmingham Queen Elizabeth Hospital (BI) and UK NEQAS (EQA). The samples were either colorectal cancer (CRC), non-small cell lung cancer (NSCLC), melanoma (Melanoma), prostate (Prostate), breast (Breast) or not specified (Other). DIN Scores are generated with an Agilent TapeStation 2200. 1 is the lowest possible score (highly degenerated, lowly concentrated) 10 the highest (highly intact, highly concentrated). N/A means either DNA input concentration was too low to quantify or sample was not measured. DNA was quantified with Qubit instrument, red cells indicate that DNA amount available was below 200ng, i.e. not enough volume present or DNA had to be vacuum concentrated. Other samples were diluted to provide the exact amount of 200ng as input.

B.2 Samples for Agilent HaloPlex HS Panel Validation

[2]Concentration quantified by Qubit, in ng/μl

Sample ID	**Illumina index (i7)**	**Conc.** [3]	**Type**[4]
		Pool 1	
15R8415	ATGCCTAA	5.48	CRC
15R8417	AGCAGGAA	0.83	
15R8418	ATCATTCC	3.05	CRC
15R8419	AACTCACC	3.57	CRC
15R8420	AACGCTTA	2.71	CRC
15R8421	AGCCATGC	5.43	CRC
15R8422	GAATCTGA	11.00	CRC
15R8423	GAGCTGAA	4.90	CRC
15R8474	GCCACATA	4.93	NSCLC
15R8476	GCTAACGA	0.93	NSCLC
15R8479	GGAGAACA	5.20	CRC
15R8488	GTACGCAA	4.98	CRC
15R8490	AACGTGAT	5.61	CRC
15R8420_B	AAACATCG	1.35	CRC
15R8421_B	ACCACTGT	6.05	CRC

15R8422_B	CAGATCTG	4.00	CRC
15R8474_B	CATCAAGT	9.82	NSCLC
15R8496_B	AGTACAAG	4.88	CRC
15R8508_B	CACTTCGA	10.20	CRC
15R8512_B	GAGTTAGC	4.30	CRC
15R8502	CTGGCATA	2.26	NSCLC
15R8482	ATCCTGTA	3.26	NSCLC
15R8497	AAGGTACA	2.50	NSCLC
ECD_1	ACATTGGC	5.11	Control
Pool 2			
15R8492	GCCAAGAC	2.51	CRC
15R8494	CGAACTTA	3.79	CRC
15R8495	ACCTCCAA	2.3	CRC
15R8445	CTGTAGCC	9.46	CRC
15R8455	CGCTGATC	11.70	NSCLC
15R8472	ATTGAGGA	15.10	NSCLC
15R8431	GACTAGTA	4.61	CRC
15R8480	GATAGACA	16.70	NSCLC

15R8510	GCGAGTAA	4.24	CRC
15R8496	GCTCGGTA	10.50	CRC
15R8508	GGTGCGAA	12.80	NSCLC
15R8512	GTCGTAGA	7.74	CRC
15R8514	ATTGGCTC	0.157	CRC
15R8445_2	AAGGACAC	4.83	CRC
15R8455_2	ACTATGCA	17.10	CRC
15R8472_2	ACACGACC	14.90	NSCLC
15R8431_2	CCTAATCC	11.40	CRC
15R8479_2	AGAGTCAA	4.86	CRC
15R8488_2	GATGAATC	8.74	CRC
15R8490_2	GACAGTGC	4.23	CRC
15R8636	CGGATTGC	1.52	NSCLC
15R8716	AGTCACTA	2.09	NSCLC
15R8776	CTGAGCCA	0.692	CRC
ECD_2	CCGACAAC	7.68	Control

TABLE B.2: Samples used for evaluation of the custom Agilent HaloPlex HS panel. Red cells indicate insufficient DNA concentration according to manufacturer's guidelines. ECD is a control provided by Agilent for restriction digest validation. It contains genomic DNA mixed with an 800bp PCR product with restriction sites for all enzymes used.

B.3 Demultiplexing of HaloPlex HS Libraries

```
# Demultiplexing using Illumina i7 index
bcl2fastq -o Fastq_Raw --no-lane-splitting --use-bases-mask y150n,i8,n10,y150n \
--sample-sheet runHaloPlexHS.csv > demultiplexing.log
# Creating Barcode Dictionary
bcl2fastq -o Barcode_Dictionary --no-lane-splitting --use-bases-mask y150n,i6,i10,y150n \
--sample-sheet runHaloPlex.csv > demultiplexing_dict.log
#Note: Dictionary FASTQ is written to file ``Undetermined_S0_I2_001.fastq.gz''
```

B.4 Read De-duplication and Error Correction

```
# Deduplication, removal of reads off-target and sequencing error correction.
 java -jar AgilentMBCDedup.jar -X mbc_tmp -IB -b Amplicons.bed -o [sample]_dedup_mbc.bam \
 [sample]_sorted.bam Undetermined_S0_I2_001.fastq.gz
```

[3] Concentration quantified by Qubit, in ng/μl

[4] CRC - (metastatic) colorectal cancer; NSCLC - non-small cell lung cancer

Appendix C

HaloPlex HS Coverage Plots

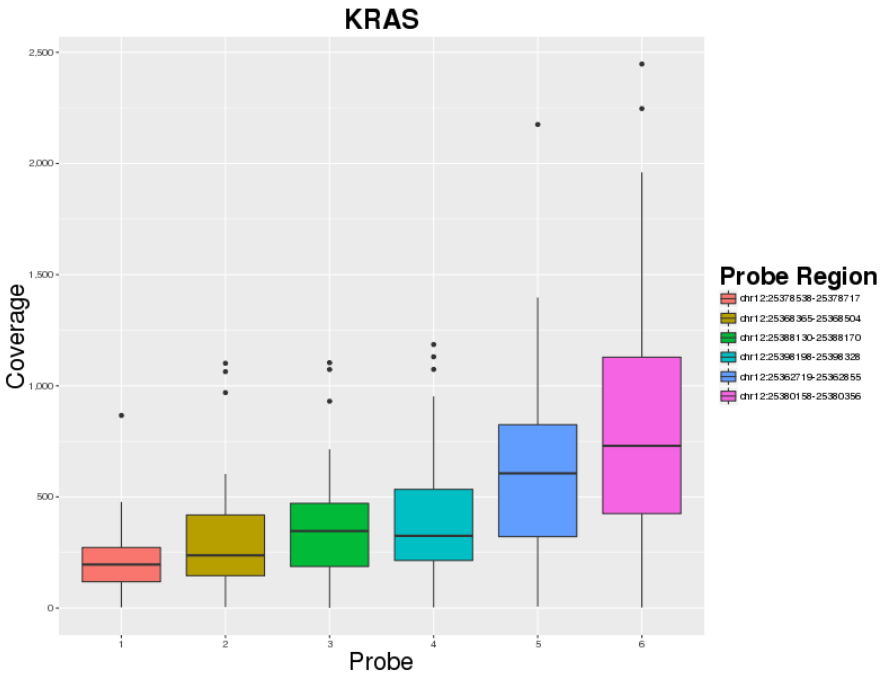


FIGURE C.1: KRAS coverage distribution by probe.

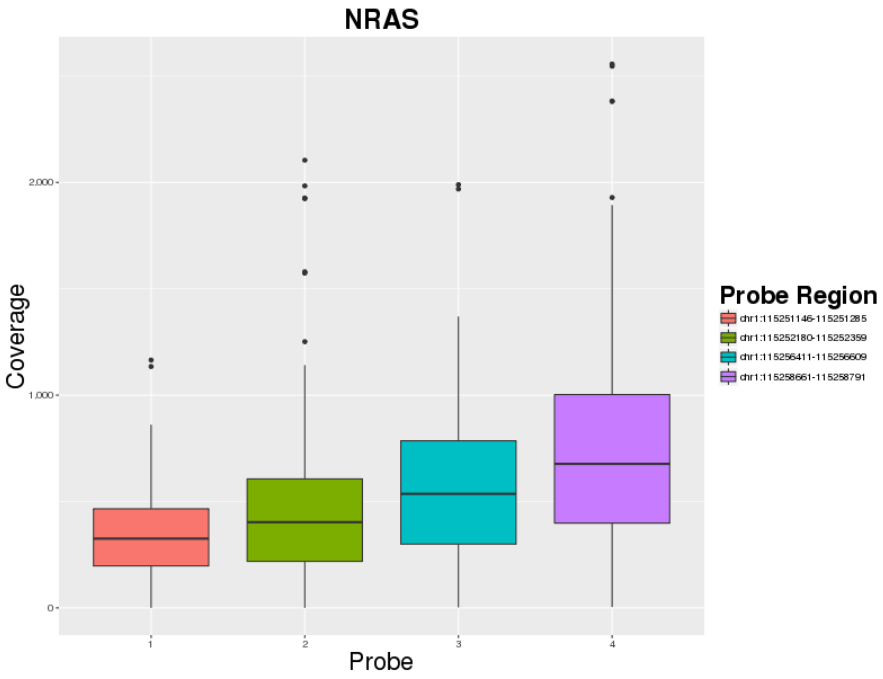


FIGURE C.2: NRAS coverage distribution by probe.

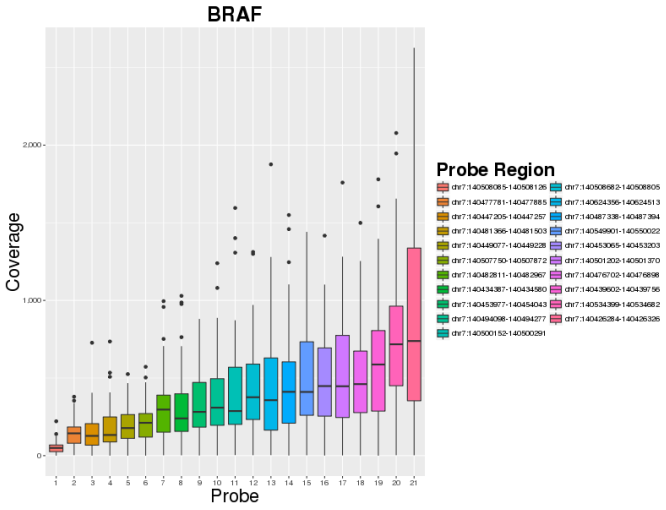


FIGURE C.3: BRAF coverage distribution by probe.

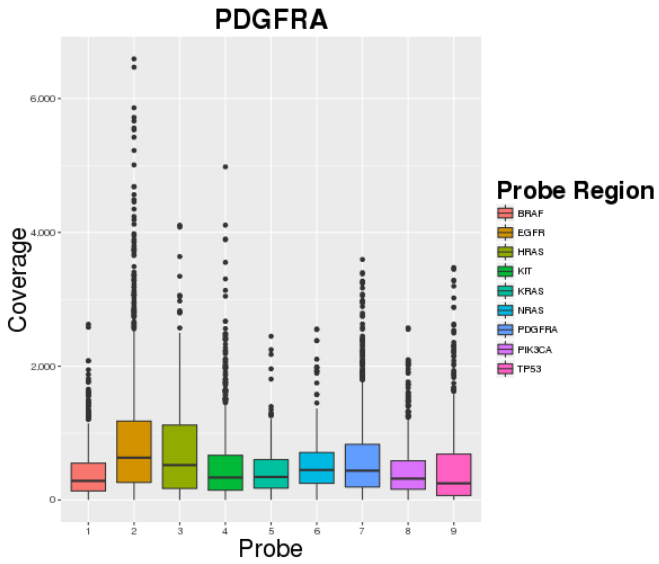


FIGURE C.4: EGFR coverage distribution by probe.

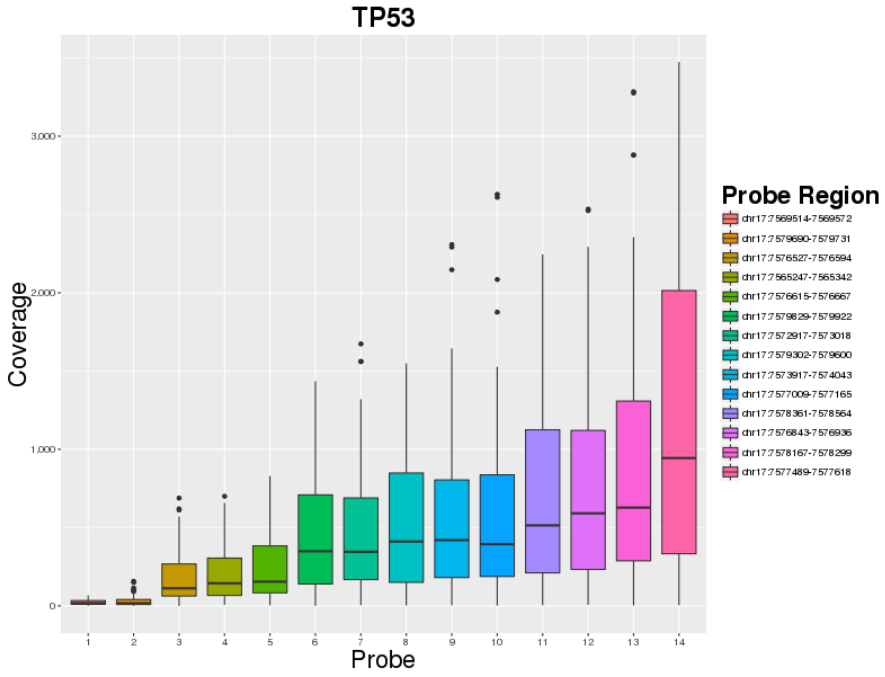


FIGURE C.5: TP53 coverage distribution by probe.

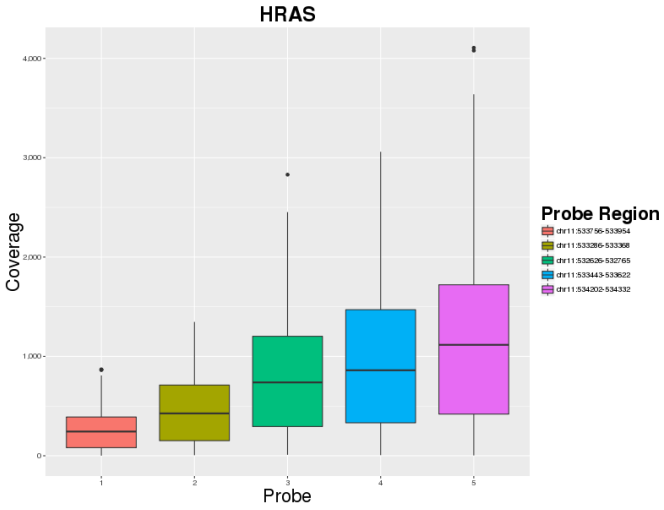


FIGURE C.6: HRAS coverage distribution by probe.

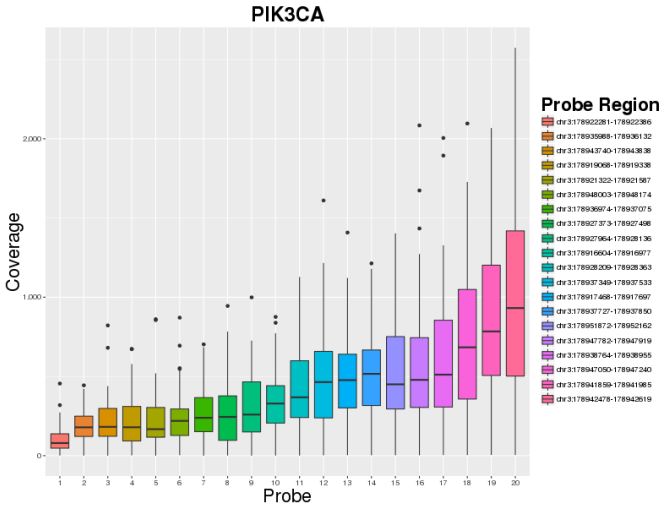


FIGURE C.7: PIK3CA coverage distribution by probe.

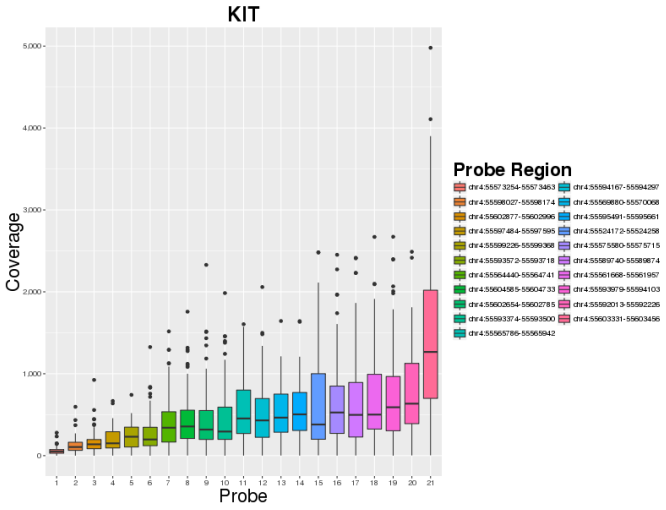


FIGURE C.8: KIT coverage distribution by probe.

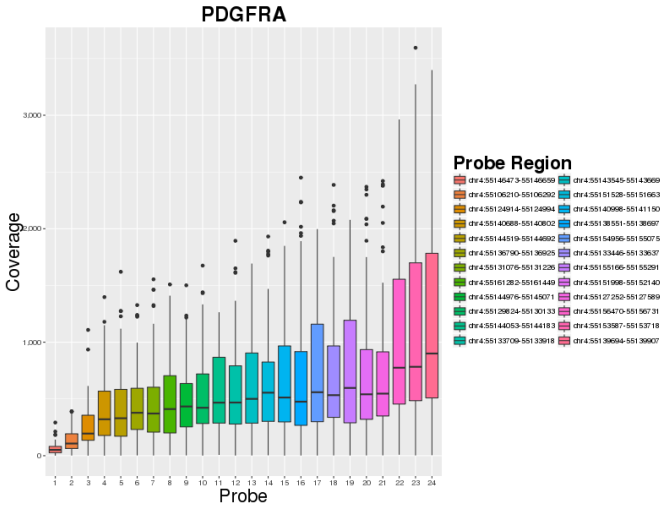


FIGURE C.9: PDGFRA coverage distribution by probe.

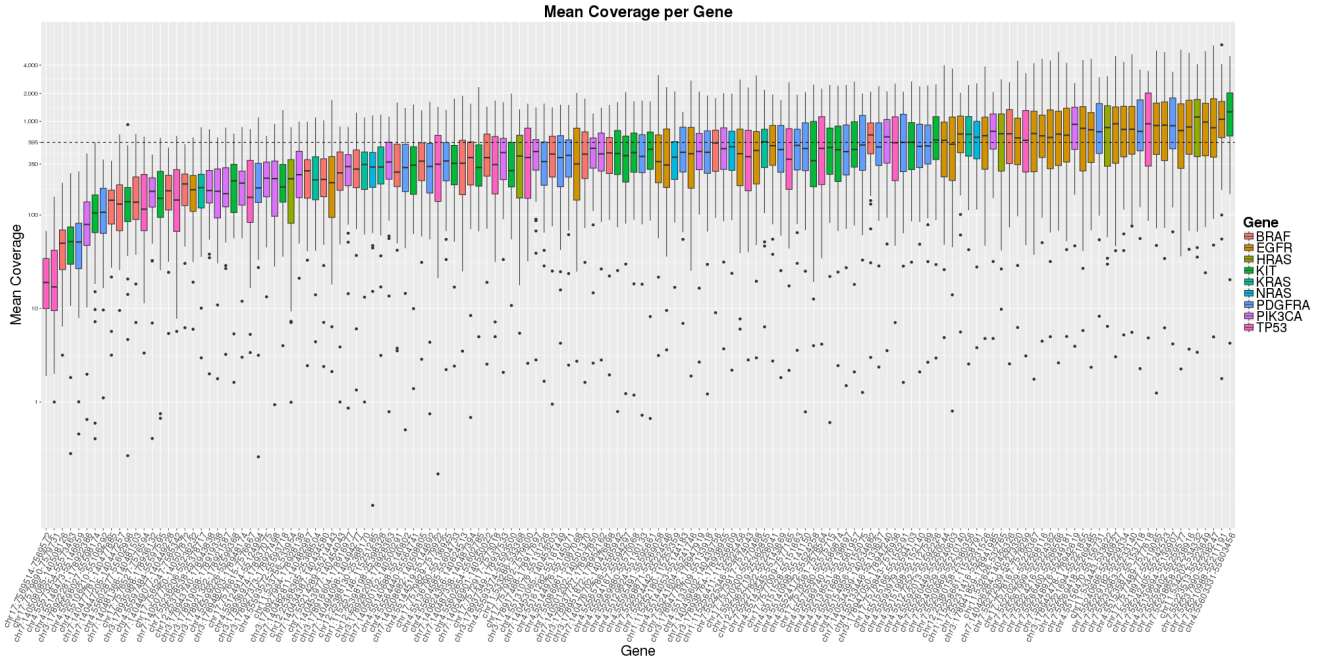


FIGURE C.10: Logarithmic coverage distribution of all genes by probe.

Appendix D

Cellular Barcoding Protocols

D.1 Table of Oligonucleotides and Reagents

List of Reagents

Reagent	Supplier/Recipe	Description
Silica Microspheres, Streptavidin 1μm (Cat. 24760-2)	Polysciences, Inc	Non-magnetic silica beads to carry barcoded primer collection
2x B/W buffer	10 mM Tris-HCL (pH 7.5), 1 mM EDTA, 2 M NaCl	Used for washes (1x) and coupling nucleic acids (2x)

Denaturation Solution	0.2M NaOH	Used for nucleotide dissociation
Annealing buffer	10mM Tris (pH 7.5-8), 50 mM NaCl, 1mM EDTA	For annealing oligonucleotides
TE buffer	10mM Tris-Hcl (pH 75-8), 1mM EDTA (pH 8)	For storing DNA or beads
Dynabeads® MyOne™ Streptavidin C1 (Cat. 65001)	Thermo Fischer Scientific	Enrichment beads
Water (Cat. W4502 Sigma)	Sigma-Aldrich®	Nuclease free water
Takara Ex Taq HS DNA Polymerase (Cat. RR006A)	Clontech	Hot start Ex Taq DNA Polymerase and buffer
dNTP Mix (Cat. R0191)	Thermo Fischer Scientific	For PCR reactions
ABIL WE09	Evonik	For emulsion
Mineral Oil (Cat. M5904)	Sigma-Aldrich®	For emulsion

Tegosoft DEC	Evonik	For emulsion
Isobutanol (Cat. 82059)	Sigma-Aldrich®	For breaking emulsions
Diethyl ether (Cat. 346136)	Sigma-Aldrich®	For breaking emulsions
Ethanol (Cat. E7023)	Sigma-Aldrich®	Alcohol is always of use
Triton® X-100 (Cat. T8787)	Sigma-Aldrich®	Maintain bead suspension
Trypsin-EDTA (0.5%) (Cat. 15400054)	Thermo Fischer Scientific	For cell dissociation
DBPS (Cat. 14190169)	Thermo Fischer Scientific	Washing and preparing cells
Monarch® PCR & DNA Cleanup Kit (Cat.T1030S)	New England BioLabs®	Clean up kit for amplicon recovery
Agencourt AMPure XP	Beckman Coulter (Cat. A63880)	Optional PCR product clean up and size selction

Terra™ PCR Direct Polymerase Mix	Clontech	Enzyme for genomic DNA amplification from cells in emulsions [272]

Table D.1: List of reagents used.

Oligo	Sequence (listed in 5′ → 3′ direction)
Unique oligo	5′-[Btn]- CAGTCATTTTCAGCAGGCC (Genomic primer) AGGACGTCAACGGAATGCTC (Universal sequence) NNNNNNNNNNNNNNN (Barcode sequence) AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT-3′ (Illumina adapter)
KRAS Anchor Primer 1	5′-[Btn]-CAGTCATTTTCAGCAGGCCAGGACGTCAACGGAATGCTC-3′
KRAS Anchor Primer 2	5′-[Btn]-AAGGGAGAAACACAGTCTGGAGGACGTCAACGGAATGCTC-3′
Bead-loading Primer	5′-AATGATACGGCGACCACCGA-3′
Enrichment Sequence	5′-[Btn] AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
KRAS Reverse Primer 1	5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT (Illumina rev. adapter) GGTCCTGCACCAGTAATATG-3′ (Reverse genomic primer)
KRAS Reverse Primer 2	5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT (Illumina rev. adapter) CACAAAGAAAGCCCTCCCCA-3′ (Reverse genomic primer)
KRAS Ampl. Primer 1	5′-CAGTCATTTTCAGCAGGCCAGGACGTCAACGGAATGCTC-3′

KRAS Ampl. Primer 2	5′-AAGGGAGAAACACAGTCTGGAGGACGTCAACGGAATGCTC-3′
KRAS FW Primer1	GGCCTGCTGAAAATGACTG
KRAS REV Primer1	GGTCCTGCACCAGTAATATG
KRAS FW Primer2	CCAGACTGTGTTTCTCCCTT
KRAS REV Primer2	CACAAAGAAAGCCCTCCCCA
Library Ampl. Primer 1	5′-AATGATACGGCGACCACCGA-3′
Library Ampl. Primer 2	5′-CAAGCAGAAGACGGCATACG-3′

Table D.2: Oligonucleotides used for Cellular Barcoding. KRAS multiplex primers were designed by Lurkin et al. [271]. [Btn.] means biotin.

D.2 Loading Unique Oligo on Beads

1. Prepare Enrichment beads:
   (a) Transfer 200μl Dynabeads® MyOne™ Streptavidin C1 into a tube
   (b) Wash[4] beads in 1ml 2x B/W buffer twice
   (c) Resuspend in 200μl 2x B/W buffer
   (d) Add 20μl Enrichment Sequence oligos
   (e) Incubate for 15 minutes at room temperature[5] while gently shaking
   (f) Wash in 1ml 1x B/W buffer twice
   (g) Wash in 500μl 1x TE buffer twice
   (h) Resuspend in 200μl 1x TE-buffer and keep at 4°C until use
2. Load Unique oligos and enrichment:
   (a) Wash in 2ml 2x B/W buffer twice
   (b) Resuspend in 2ml 1x B/W buffer
   (c) Add 100μl of Unique oligo in concentration of 100fM while gently shaking
   (d) Incubate for 15 minutes at room temperature while gently shaking

[4] Suspend, spin 3.5 minutes at > 16,000g and remove supernatant
[5] 18 °C – 25 °C

(e) Wash in 2ml Denaturation solution

(f) Wash in 2ml Annealing buffer twice

(g) Add 40μl Enrichment beads

(h) Keep for 5 minutes at 95 °C on a heatblock

(i) Keep at room temperature for 15 minutes

(j) Then put on ice for 3 minutes

(k) Gently put next to a tube magnet and separate for 3 minutes at room temperature

(l) Transfer supernatant into new tube or discard. Do not touch the pellet! The supernatant contains beads without unique oligo

(m) Resuspend beads with 250μl Annealing buffer

(n) Repeat steps 2h - 2l twice

(o) Resuspend brown pellet with 200μl Denaturation solution and vortex for 30 seconds

(p) Place next to a tube magnet for 3 minutes

(q) Transfer supernatant to fresh tube. The supernatant contains beads carrying the unique oligo

Optional. Resuspend brown pellet with denaturation solution, vortex for 30 seconds, place next to a tube magnet for 3 minutes and add to the enriched beads

Optional. Place Enriched beads next to a tube magnet for three minutes and transfer supernatant to fresh tube, to remove potential residues of enrichment beads

(r) Wash in 1ml 1x B/W twice. Carefully watch the white pellet

(s) Resuspend beads in 100μl 1x B/W buffer

(t) Add 100μl of Anchor Primer mix (1+2)

(u) Incubate 15 minutes at room temperature while gently shaking

(v) Wash in 200μ 1x B/W buffer

(w) Wash in 100μl 1x TE-buffer twice

(x) Resuspend in 20μl and keep at 4°C until further use

3. Emulsification:

Comment. Beads are emulsified for filling entire bead collection with uniquely barcoded primers. In doing so, loaded beads are split into separate reactions each using 10μl of beads.

(a) Aqueous Recipe:

Reagent	Volume per reaction
Nuclease-free water	32.5μl
10x Ex Taq HS buffer	5μl
dNTP mix 2.5mM	1μl
Bead-loading Primer (10μM)	1μl
Loaded beads	10μl
Ex Taq HS	0.5μl
Total	50μl

4. Oil Recipe (for 50μl aqueous mix):

Reagent	Volume per reaction
ABIL WE 09	219μl
Mineral Oil	60μl
Tegosoft DEC	21μl
Total	300μl

Note. Keep oils on ice prior use!

Note. Cut the front of the tip to accurately pipette the necessary volume!

5. Vortex oil mix at maximum speed then add aqueous mix dropwise over 1 minute into the tube of oil

6. Vortex at maximum speed for 2 more minutes at room temperature. Altogether the mix needs 3 minutes to be vortexed.

Optional. Check 1μl of the emulsion under a microscope, an example is given in figure D.1

7. Immediately transfer creamy white emulsion to a thermal cycler for a PCR:
   (a) 98 °C for 10 seconds
   (b) 56 °C for 30 seconds
   (c) 72 °C for 1 minute
   (d) Repeat steps (b) - (c) 29 times
   (e) 4 °C for ∞
8. Break emulsions by adding 1ml Isobutanol for 350μl of emulsion
9. Vortex for 10 seconds at maximum speed
10. Spin 2 minutes at > 16,000g
11. Remove supernatant and add 1ml fresh Isobutanol and 250μl Diethyl ether
12. Vortex for 10 seconds at maximum speed
13. Spin 2 minutes at > 16,000g
14. Carefully remove supernatant and resuspend white pellet with 250μl 70% Ethanol

15. Spin 2 minutes at > 16,000g
16. Resuspend white pellet in 1x B/W buffer and with 1% Triton® X-100
17. Incubate 90 minutes - 120 minutes at room temperature while gently shaking until beads are fully resuspended
18. Wash beads in 1x TE-buffer
19. Resuspend in 20μl 1x TE-buffer and keep at 4°C until further use

Beads can be tested by PCR:

Reagent	Volume per reaction
Nuclease-free water	*
10X ExTaq HS buffer	2.5μl
dNTPs 2.5mM	1μl
KRAS Rev. primer mix (10μM)	0.5μl
KRAS Ampl. Primer mix (10μM)	0.5μl
Bead-loading Primer (10μM)	0.25μl
Fully loaded Beads	0.25μl
Ex Taq HS	0.125μl
DNA template (< 500ng)	*
Total	25μl

PCR programme:

1. 98 °C for 10 seconds
2. 56 °C for 30 seconds
3. 72 °C for 1 minute
4. Repeat steps 2. - 3. 29 times
5. 4 °C for ∞

Expected are two products, 274bp for KRAS exon 3 and 282 bp for KRAS exon 2. As a control a PCR with KRAS FW+REV primer mixes is advised. The amplicons are, however, 120 bp shorter. Note: the primer mix amplifies the uniquely barcoded primer mixes on the bead, the resulting forward primers then amplify the targeted regions. The resulting PCR products could already be sequenced, due to the Illumina adapter overhangs on both ends, but the amplicons are not barcoded on per-cell cell level, of course.

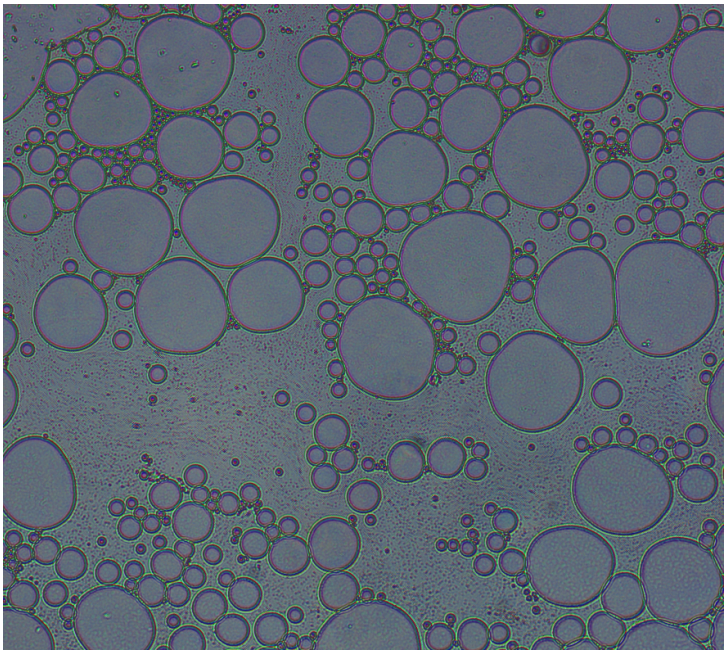


FIGURE D.1: Example of a stable emulsion under a microscope 60X zoom.

D.3 Direct Emulsion PCR Library Amplification

1. Cell preparation:
   (a) NIH3T3 cells were trypsinised, washed in DPBS and counted by using a hemocytometer
   (b) K562 cells were washed in DPB Sand counted by using a hemocytometer
   (c) NIH3T3 and K562 were mixed in an 80 (NIH3T3):20 (K562) ratio
   (d) 100,000 cells were transferred into a fresh tube, spinned at 800g for 3 minutes to avoid any cells to break
   (e) Cell mix was resuspended in 5μl 1x Direct PCR buffer
2. Prepare mix for emulsion using a direct polymerase mix
   (a) Add the following reagents directly on to the cell suspension:

Reagent	Volume per reaction
1. Nuclease-free water	7.25μl
2. 2x Direct PCR buffer	25μl
3. dNTPs 10mM	1μl
4. KRAS Rev. primer mix (100μM)	0.25μl
5. KRAS Ampl. Primer mix (100μM)	0.25μl
6. Bead-loading Primer (100μM)	0.25μl
7. Fully loaded Beads	10μl
8. 1x BSA	5μl
9. Direct Taq	1μl
Total	50μl

(b) Oil Recipe (for 50μl aqueous mix):

Reagent	Volume per reaction
ABIL WE 09	219μl
Mineral Oil	60μl
Tegosoft DEC	21μl
Total	300μl

(c) Aqueous mix containing cells is carefully pipetted up and down to mix the mix, but not to break any cells

Note. Keep oils on ice prior use!

Note. Cut the front of the tip to accurately pipette the necessary volume!

(d) Vortex oil mix at maximum speed then add aqueous mix drop-wise over 1 minute into the tube of oil

(e) Vortex at maximum speed for 2 more minutes at room temperature. Altogether the mix needs 3 minutes to be vortexed.

Optional. Check 1μl of the emulsion under a microscope.

(f) Directly transfer the emulsion to a thermal cycler for PCR using the following programme:

i. 98 °C for 2 minutes

ii. 98 °C for 10 seconds

iii. 58 °C for 15 seconds

iv. 68 °C for 1 minute

v. Repeat steps ii. - iv. 20 times

vi. 4°C for ∞

Note. Due to the increase in temperature and longer cycles, emulsion can become unstable after more than 20 cycles

(g) After PCR has completed, add 1ml Isobutanol to break the emulsion

(h) Vortex for 10 seconds or until white emulsion has completely dissolved

(i) Spin for 2 minutes at > 16,000g

(j) The Isobutanol and the aqueous phase have been separated into two layers

(k) Remove the top layer carefully to not remove any of the aqueous phase

(l) Resuspend aqueous phase in 800μl Isobutanol and 200μl Diethyl ether

(m) Vortex for 10 seconds

(n) Spin for 2 minutes at > 16,000g

(o) The Isobutanol and the aqueous phase have been separated into two layers

(p) Remove carefully the top layer with all remaining oils

(q) Resuspend the aqueous phase with 250μl 70% Ethanol

(r) Mix by pipetting up and down a few times

(s) Spin for 4 minutes at $> 16,000$g

(t) Remove the to layer

(u) Clean up the PCR product with a cleanup kit to wash away primers, beads or remaining cell fragments

Note. Do not run a gel on the library to see if the PCR has worked after this step, as library needs amplification first

(v) Keep at 4°C until further use

3. Library amplification and sequencing

(a) Amplifcation of successfully barcoded amplicons by PCR:

Reagent	Volume per reaction
Nuclease-free water	39.75μl
10x Ex Taq HS buffer	5μl
dNTPs (2.5mM)	2μl
Library Ampl. primer mix (10μM)	1μl
emPCR product	2μl
Ex Taq HS	0.25μl
Total	50μl

(b) Transfer reaction mixes to a thermal cycler for a PCR:

i. 98 °C for 10 seconds

ii. 56 °C for 30 seconds

iii. 72 °C for 1 minute

iv. Repeat steps ii. - iv. 29 times

v. 4 °C for ∞

(c) Resulting library has sufficient concentration to test run on a gel, see figure D.2

(d) Clean up PCR product with a cleanup kit

Optional. Clean up product with magnetic SPRI beads, such as AMPure XP beads, adjust concentration to remove DNA fragments below 200bp

(e) Sequence library with a Miseq/NextSeq 500 instrument. Recommended run length is 100bp paired-end

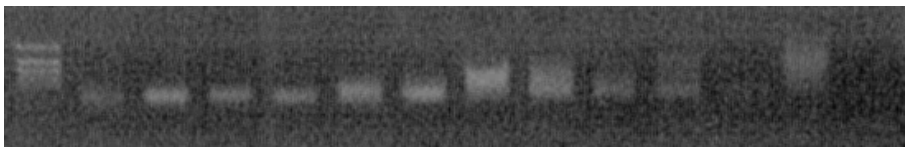


FIGURE D.2: Gel image of single cell direct emulsion PCR. Lane 1: ladder, lane2 : single cell direct emulsion PCR (unique barcodes), lane 3: purified DNA emulsion PCR from beads (unique barcodes), lane 4: direct emulsion PCR (non-unique barcodes), lane 5: purified DNA emulsion PCR from beads (non -unique barcodes), lane 6: direct PCR from beads (unique barcodes), lane 7: direct PCR from beads (non-unique barcodes), lane 8: purified DNA PCR from beads (unique barcodes), lane 9: purified DNA PCR from beads (non-unique barcodes), lane 10: direct emulsion PCR (primers), lane 11: direct PCR (primers), lane 12: purified DNA emulsion PCR (primers), lane 13: purified DNA PCR (primers), lane 13: negative control (water)

D.4 Extract Barcode from Read Data

```
##Extract_cellbarcode _and_write_to_FASTQ_header.R
#Load libraries
library(ShortRead)
library(parallel)

#US
useq <- "GAGCATTCCGTTGACGTCCT"
#FASTQ Read 1
path <- "./Pilot_R1.fastq.gz"

writeBarcodeToHeader <- function(i,reads, bc.len=15, useq=useq){
  useq.rc <- as.character(reverseComplement(DNAString(useq)))
  #If universal#sequence has been found
  if(grepl(useq,sread(reads)[i]) == TRUE){
    bc <- sub(paste0(useq,".*"),"", sread(reads)[i])
    if(nchar(bc)==bc.len){ Barcode found
      reads[i] <- narrow(reads[i], start=(bc.len+nchar(useq)+1))
      reads[i] <- renew(reads[i], id=BStringSet(x=paste0(id(reads[i]),
      " BC:Z:",bc)))
      reads[i] # Read with BC in header and trimmed FASTQ
    }else{
      reads[i] <- narrow(reads[i],start=(nchar(bc)+nchar(useq)+1))
      reads[i] <- renew(reads[i], id=BStringSet(x=paste0(id(reads[i]),
      " BC:Z:", paste0(rep("N", length=bc.len), collapse=""))))
      reads[i] # Undetermined Barcode
    }
    # Reverse complement of US found
  } else if(grepl(useq.rc,sread(reads)[i]) == TRUE) {
    bc <- sub(paste0(useq.rc,".*"),"",sread(reads)[i])
    if(nchar(bc)==bc.len){ #Barcode found
      reads[i] <- narrow(reads[i],start=(bc.len+nchar(useq)+1))
      reads[i] <- renew(reads[i],
```

```
      id=BStringSet(paste0(id(reads[i])," BC:Z:",bc)))
      reads[i] # Read with barcode in header and trimmed FASTQ
    } else {
      reads[i] <- narrow(reads[i],start=(nchar(bc)+nchar(useq)+1))
      reads[i] <- renew(reads[i], id=BStringSet(paste0(id(reads[i]),
      " BC:Z:", paste0(rep("N", length=bc.len), collapse=""))))
      reads[i] # Undertemined Barcode
    }
  } else { # Read does not contain US/barcode
    warning(paste("Barcode of", id(reads)[i],
    "could not be determined. Leaving it untouched!"))
    reads[i]
  }
}
#Call the function
reads.new <- mclapply(1:length(reads.raw),
                      writeBarcodeToHeader,
                      reads=reads.raw,
                      bc.len=15,
                      useq=useq,
                      mc.cores=32) # Adjust to machine spec
for(r in 1:length(reads.new)){ #write to FASTQ file
  writeFastq(reads.new[[r]], "Reads_barcoded_R1.fastq.gz", mode='a')
}
```

D.5 Carry Barcode to SAM

```
bwa mem -t 2 -M -C hg19.fa Reads_barcoded_R1.fastq.gz\
Pilot_R2.fastq.gz > pilot.sam
```

D.6 Variant Calling on Single-Cell Data

```
samtools mpileup -f genome.fa -AB -d 10000 -C50 [cell].bam |\
java -jar ./VarScan.v2.3.9.jar mpileup2cns --strand-filter 0\
--min-var-freq 0.20 --variants 1 --output-vcf 1 > [cell].vcf
```

Bibliography

[260] Parviz Ahmad-Nejad et al. "Assessing quality and functionality of DNA isolated from FFPE tissues through external quality assessment in tissue banks". In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 53.12 (2015), pp. 1927–1934.

[132] Hafid Alazzouzi et al. "SMAD4 as a prognostic marker in colorectal cancer". In: *Clinical Cancer Research* 11.7 (2005), 2606–2611.

[133] Walter Alexander. "Inhibiting the akt pathway in cancer treatment: three leading candidates". In: *Pharmacy and Therapeutics* 36.4 (2011), p. 225.

[134] Deborah A Altomare and Joseph R Testa. "Perturbations of the AKT signaling pathway in human cancer". In: *Oncogene* 24.50 (2005), pp. 7455–7464.

[1] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[2] *An Introduction to Next-Generation Sequencing Technology*. http://www.illumina.com/technology/next-generation-sequencing.html. Illumina. May 2016.

[135] Jamie N Anastas and Randall T Moon. "WNT signalling pathways as therapeutic targets in cancer". In: *Nature Reviews Cancer* 13.1 (2013), pp. 11–26.

[136] Simon Andrews. *FastQC: A quality control tool for high throughput sequence data*. Website. http://bioinformatics.babraham.ac.uk/projects/fastqc/. Oct. 2015.

[137] Koji Aoki and Makoto M Taketo. "Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene". In: *Journal of cell science* 120.19 (2007), pp. 3327–3335.

[261] Maryke Appel et al. *KAPA HyperPlus/SeqCap EZ workflow: Improving Data Quality and Turnaround Times for Targeted Next-Generation Sequencing of FFPE DNA*. Tech. rep. KAPA Biosystems, 2016.

[3] JK Aronson. "Biomarkers and surrogate endpoints". In: *British journal of clinical pharmacology* 59.5 (2005), pp. 491–494.

[138] Uzma Asghar et al. "The history and future of targeting cyclin-dependent kinases in cancer therapy". In: *Nature reviews Drug discovery* 14.2 (2015), pp. 130–146.

[4] Carlos F Barbas et al. "Quantitation of DNA and RNA". In: *Cold Spring Harbor Protocols* 2007.11 (2007), pdb–ip47.

[139] Alberto Bardelli et al. "Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer". In: *Cancer discovery* 3.6 (2013), pp. 658–673.

[215] Carol Beadling et al. "Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping". In: *The Journal of Molecular Diagnostics* 15.2 (2013), pp. 171–176.

[140] Heiko Becker et al. "Tracing the development of acute myeloid leukemia in CBL syndrome". In: *Blood* 123.12 (2014), pp. 1883–1886.

[141] B A Benayoun et al. "The forkhead factor FOXL2: A novel tumor suppressor?" In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 1–5.

[142] Mohamed Bentires-Alj et al. "Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia". In: *Cancer research* 64.24 (2004), pp. 8816–8820.

[5] David R Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry". In: *nature* 456.7218 (2008), pp. 53–59.

[262] Krzysztof Bielawski et al. "The suitability of DNA extracted from formalin-fixed, paraffin-embedded tissues for double differential polymerase chain reaction analysis". In: *International journal of molecular medicine* 8.5 (2001), pp. 573–578.

[143] Valérie Bonadona et al. “Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome”. In: *Jama* 305.22 (2011), pp. 2304–2310.

[6] J Brabender et al. “Epidermal growth factor receptor and HER2-neu mRNA expression in non-small cell lung cancer is correlated with survival”. In: *Clinical Cancer Research* 7.7 (2001), pp. 1850–1855.

[234] Daniel Branton et al. “The potential and challenges of nanopore sequencing”. In: *Nature biotechnology* 26.10 (2008), pp. 1146–1153.

[7] Broadinstitue. *Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*. http://broadinstitute.github.io/picard. Nov. 2015.

[144] Jennifer L Bromberg-White, Nicholas J Andersen, and Nicholas S Duesbery. “MEK genomics in development and disease”. In: *Briefings in functional genomics* 11.4 (2012), pp. 300–310.

[8] Vince Buffalo. *qrqc: Quick Read Quality Control*. R package version 1.24.0. 2012. URL: http://github.com/vsbuffalo/qrqc.

[9] Rebecca A Burrell and Charles Swanton. “The evolution of the unstable cancer genome”. In: *Current opinion in genetics & development* 24 (2014), pp. 61–67.

[10] Rebecca A Burrell and Charles Swanton. "Tumour heterogeneity and the evolution of polyclonal drug resistance". In: *Molecular oncology* 8.6 (2014), pp. 1095–1111.

[11] Gianni Bussolati et al. "Formalin fixation at low temperature better preserves nucleic acid integrity". In: *PLoS One* 6.6 (2011), e21043.

[12] Bruno Canard and Robert S Sarfati. "DNA polymerase fluorescent substrates with reversible 3-tags". In: *Gene* 148.1 (1994), pp. 1–6.

[13] Kirstie Canene-Adams. "Preparation of formalin-fixed paraffin-embedded tissue for immunohistochemistry". In: *Methods in enzymology* 533 (2012), pp. 225–233.

[14] Danielle Mercatante Carrick et al. "Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue". In: *PloS one* 10.7 (2015), e0127353.

[15] Daniel E Carvajal-Hausdorf et al. "Quantitative measurement of cancer tissue biomarkers in the lab and in the clinic". In: *Laboratory Investigation* 95.4 (2015), pp. 385–396.

[145] CJ Chang and MC Hung. "The role of EZH2 in tumour progression". In: *British journal of cancer* 106.2 (2012), pp. 243–247.

[146] F Chang and Marilyn M Li. "Clinical application of amplicon-based next-generation sequencing in cancer". In: *Cancer genetics* 206.12 (2013), pp. 413–419.

[147] I-Ming Chen et al. "Outcome modeling with CRLF2, IKZF1, JAK, and minimal residual disease in pediatric acute lymphoblastic leukemia: a Children's Oncology Group study". In: *Blood* 119.15 (2012), pp. 3512–3522.

[148] Jon H Chung and Fred Bunz. "A loss-of-function mutation in PTCH1 suggests a role for autocrine hedgehog signaling in colorectal tumorigenesis". In: *Oncotarget* 4.12 (2013), p. 2208.

[16] Gary A Churchill. "Fundamentals of experimental design for cDNA microarrays". In: *Nature genetics* 32 (2002), pp. 490–495.

[17] P. Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". In: *Fly* 6.2 (2012), pp. 80–92.

[18] P. Cingolani et al. "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift". In: *Frontiers in Genetics* 3 (2012).

[149] Magdalena Cizkova et al. "PIK3R1 underexpression is an independent prognostic marker in breast cancer". In: *BMC cancer* 13.1 (2013), p. 545.

[19] Peter JA Cock et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic acids research* 38.6 (2010), pp. 1767–1771.

[20] 1000 Genomes Project Consortium et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–1073.

[216] PD Da Forno et al. "BRAF, NRAS and HRAS mutations in spitzoid tumours and their possible pathogenetic significance". In: *British Journal of Dermatology* 161.2 (2009), pp. 364–372.

[150] Tao Dao et al. "Targeting the intracellular WT1 oncogene product with a therapeutic human antibody". In: *Science translational medicine* 5.176 (2013), 176ra33–176ra33.

[21] Mark A Dawson and Tony Kouzarides. "Cancer epigenetics: from mechanism to therapy". In: *Cell* 150.1 (2012), pp. 12–27.

[22] Mark A DePristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature genetics* 43.5 (2011), pp. 491–498.

[23] L Peter Deutsch. "GZIP file format specification version 4.3". In: (1996).

[24] Lloye M Dillon and Todd W Miller. "Therapeutic targeting of cancers with loss of PTEN function". In: *Current drug targets* 15.1 (2014), p. 65.

[263] Li Ding et al. "Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing". In: *Nature* 481.7382 (2012), pp. 506–510.

[25] Hui Dong et al. "Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System". In: *Acta biochimica et biophysica Sinica* 43.6 (2011), pp. 496–500.

[26] N J Dovichi and Jianzhong Zhang. "How capillary electrophoresis sequenced the human genome". In: *Angewandte Chemie International Edition* 39.24 (2000), pp. 4463–4468.

[27] John W Drake et al. "Rates of spontaneous mutation". In: *Genetics* 148.4 (1998), pp. 1667–1686.

[28] Aron C Eklund and Zoltan Szallasi. "Correction of technical bias in clinical microarray data improves concordance with known biological information". In: *Genome Biol* 9.2 (2008), R26.

[29] Mark G Erlander et al. "Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification". In: *The Journal of Molecular Diagnostics* 13.5 (2011), pp. 493–503.

[30] Estevezj. *Sanger sequencing*. https://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg. Dec. 2012.

[151] Mark Ewalt et al. "Real-time PCR-based analysis of BRAF V600E mutation in low and intermediate grade lymphomas confirms frequent occurrence in hairy cell leukaemia". In: *Hematological oncology* 30.4 (2012), pp. 190–193.

[152] Adam D Ewing et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection". In: *Nature methods* 12.7 (2015), pp. 623–630.

[31] Brent Ewing and Phil Green. "Base-calling of automated sequencer traces using phred. II. Error probabilities". In: *Genome research* 8.3 (1998), pp. 186–194.

[235] H Christina Fan, Glenn K Fu, and Stephen PA Fodor. "Combinatorial labeling of single cells for gene expression cytometry". In: *Science* 347.6222 (2015), p. 1258367.

[32] Milan Fedurco et al. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies". In: *Nucleic acids research* 34.3 (2006), e22–e22.

[33] Cathy A Finlay, Philip W Hinds, and Arnold J Levine. "The p53 proto-oncogene can act as a suppressor of transformation". In: *Cell* 57.7 (1989), pp. 1083–1093.

[153] Monica Hoyos Flight. "Anticancer drugs: A sweet blow for cancer cells". In: *Nature Reviews Drug Discovery* 10.10 (2011), pp. 734–734.

[34] SA1 Forbes et al. "The catalogue of somatic mutations in cancer (COSMIC)". In: *Current protocols in human genetics* (2008), pp. 10–11.

[154] William D Foulkes et al. "The CDKN2A (p16) gene and human cancer." In: *Molecular Medicine* 3.1 (1997), p. 5.

[35] Marcus Gassmann and Barry McHoull. *DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System & Genomic DNA ScreenTape*. Tech. rep. G5991-5258EN. Application note. Agilent Technologies, 2014.

[236] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". In: *Nature Reviews Genetics* 17.3 (2016), pp. 175–188.

[36] GeneChip. *Microarray*. Computer Desktop Encyclopedia. Curtesy of Affymetrix. June 2015.

[217] *GeneRead™DNAseq Targeted Panels V2 Handbook*. V2. For targeted enrichment prior to next-generation sequencing. QIAGEN. June 2015.

[237] Marco Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England Journal of Medicine* 366.10 (2012), pp. 883–892.

[155] Ermanno Gherardi et al. "Targeting MET in cancer: rationale and progress". In: *Nature Reviews Cancer* 12.2 (2012), pp. 89–103.

[37] Annuska M Glas et al. "Converting a breast cancer microarray signature into a high-throughput diagnostic test". In: *BMC genomics* 7.1 (2006), p. 1.

[156] David W Goodrich. "The retinoblastoma tumor-suppressor gene, the exception that proves the rule". In: *Oncogene* 25.38 (2006), pp. 5233–5243.

[157] Tiziana Grafone et al. "An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment". In: *Oncology reviews* 6.1 (2012).

[218] Phillip N Gray, Charles LM Dunlop, and Aaron M Elliott. "Not All Next Generation Sequencing Diagnostics are Created Equal: Understanding the Nuances of Solid Tumor Assay Design for Somatic Mutation Detection". In: *Cancers* 7.3 (2015), pp. 1313–1332.

[158] F Graziano, B Humar, and P Guilford. "The role of the E-cadherin gene (CDH1) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice". In: *Annals of oncology* 14.12 (2003), pp. 1705–1713.

[159] Silvia Grisendi et al. "Nucleophosmin and cancer". In: *Nature Reviews Cancer* 6.7 (2006), pp. 493–505.

[238] Mira T Guo et al. "Droplet microfluidics for high-throughput biological assays". In: *Lab on a Chip* 12.12 (2012), pp. 2146–2155.

[160] Yan Guo et al. "Exome sequencing generates high quality data in non-target regions". In: *BMC genomics* 13.1 (2012), p. 1.

[161] Yan Guo et al. "The effect of strand bias in Illumina short-read sequencing data". In: *BMC genomics* 13.1 (2012), p. 1.

[162] Carolina Gutierrez and Rachel Schiff. "HER2: biology, detection, and clinical implications". In: *Archives of pathology & laboratory medicine* 135.1 (2011), pp. 55–62.

[219] *HaloPlex HS Target Enrichment System For Illumina Sequencing Protocol*. B0. Copyright by Agilent Technologies. Agilent. June 2015.

[163] Z Han et al. "Reversal of multidrug resistance of gastric cancer cells by downregulation of Akt1 with Akt1 siRNA." In: *Journal of experimental & clinical cancer research: CR* 25.4 (2006), pp. 601–606.

[38] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[39] Douglas Hanahan and Robert A Weinberg. "The hallmarks of cancer". In: *cell* 100.1 (2000), pp. 57–70.

[40] David Hansemann. "Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung". In: *Virchows Archiv* 119.2 (1890), pp. 299–326.

[220] Paul Hardenbol et al. "Multiplexed genotyping with sequence-tagged molecular inversion probes". In: *Nature biotechnology* 21.6 (2003), pp. 673–678.

[221] Jennifer Harrow et al. "GENCODE: the reference human genome annotation for The ENCODE Project". In: *Genome research* 22.9 (2012), pp. 1760–1774.

[239] Trevor L Hawkins et al. "DNA purification and isolation using a solid-phase." In: *Nucleic Acids Research* 22.21 (1994), p. 4543.

[41] Jakob Hedegaard et al. "Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue". In: *PloS one* 9.5 (2014), e98187.

[164] Carl-Henrik Heldin. "Targeting the PDGF signaling pathway in tumor treatment". In: *Cell Communication and Signaling* 11.1 (2013), p. 1.

[42] Michael J Heller. "DNA microarray technology: devices, systems, and applications". In: *Annual review of biomedical engineering* 4.1 (2002), pp. 129–153.

[165] Daniel Herranz et al. "Metabolic reprogramming induces resistance to anti-NOTCH1 therapies in T cell acute lymphoblastic leukemia". In: *Nature medicine* (2015).

[222] Joseph B Hiatt et al. "Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation". In: *Genome research* 23.5 (2013), pp. 843–854.

[43] Russell Higuchi et al. "Simultaneous amplification and detection of specific DNA sequences". In: *Bio/technology* 10.4 (1992), pp. 413–417.

[264] Kurt Hirschhorn, Wayne H Decker, and Herbert L Cooper. "Human intersex with chromosome mosaicism of type XY/XO: Report of a case". In: *New England Journal of Medicine* 263.21 (1960), pp. 1044–1048.

[240] Christian Hoffmann et al. "DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations". In: *Nucleic acids research* 35.13 (2007), e91.

[166] Matthew Holderfield et al. "Targeting RAF kinases for cancer therapy: BRAF mutated melanoma and beyond". In: *Nature reviews. Cancer* 14.7 (2014), p. 455.

[44] Nils Homer and Stanley F Nelson. "Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA". In: *Genome Biol* 11.10 (2010), R99.

[241] Machiko Hori, Hajime Fukano, and Yosuke Suzuki. "Uniform amplification of multiple DNAs by emulsion PCR". In: *Biochemical and biophysical research communications* 352.2 (2007), pp. 323–328.

[242] Kurt Hornik. *R FAQ*. 2016. URL: https://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

[45] Hao Hu et al. "VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix". In: *Genetic epidemiology* 37.6 (2013), pp. 622–634.

[46] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic acids research* 37.1 (2009), pp. 1–13.

[223] T Hubbard et al. "The Ensembl genome database project". In: *Nucleic acids research* 30.1 (2002), pp. 38–41.

[47] *Illumina Experiment Manager User Guide*. 1.9 (15031335 Rev. J). ILLUMINA PROPRIETARY. Illumina. Mar. 2015.

[48] Illumina. *Sequencing power for every scale*. Website. http://www.illumina.com/systems/sequencing.html. May 2016.

[167] Yoshiaki Ito, Suk-Chul Bae, and Linda Shyue Huey Chuang. "The RUNX family: developmental regulators in cancer". In: *Nature Reviews Cancer* 15.2 (2015), pp. 81–95.

[49] Hongshan Jiang et al. "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads". In: *BMC bioinformatics* 15.1 (2014), p. 1.

[168] Marcus B Jones et al. "Library preparation methodology can influence genomic and functional predictions in human microbiome research". In: *Proceedings of the National Academy of Sciences* 112.45 (2015), pp. 14024–14029.

[50] Hyunju Jung et al. *The DNA Integrity Number (DIN) Provided by the Genomic DNA ScreenTape Assay Allows for Streamlining of NGS on FFPE Tissue Samples*. Tech. rep. 5991-5360EN. Application note. Agilent Technologies, 2014.

[51] Evangelia Karampetsou, Deborah Morrogh, and Lyn Chitty. "Microarray Technology for the Diagnosis of Fetal Chromosomal Aberrations: Which Platform Should We Use?" In: *Journal of clinical medicine* 3.2 (2014), pp. 663–678.

[224] Donna Karolchik et al. "The UCSC genome browser database". In: *Nucleic acids research* 31.1 (2003), pp. 51–54.

[169] Masaru Katoh and Hitoshi Nakagama. "FGF receptors: cancer biology and therapeutics". In: *Medicinal research reviews* 34.2 (2014), pp. 280–300.

[52] Graham Kemp. "Capillary Electrophoresis". In: *Biotechnology and applied biochemistry* 27.1 (1998), pp. 9–17.

[243] Scott R Kennedy et al. "Detecting ultralow-frequency mutations by Duplex Sequencing". In: *Nature protocols* 9.11 (2014), pp. 2586–2606.

[170] W James Kent. "BLAT-the BLAST-like alignment tool". In: *Genome research* 12.4 (2002), pp. 656–664.

[53] Scott E Kern. "Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures". In: *Cancer research* 72.23 (2012), pp. 6097–6101.

[171] Kum Kum Khanna. "Cancer risk and the ATM gene: a continuing debate". In: *Journal of the National Cancer Institute* 92.10 (2000), pp. 795–802.

[172] Dmitriy Khodakov, Chunyan Wang, and David Yu Zhang. "Diagnostics based on nucleic acid sequence variant profiling: PCR, hybridization, and NGS approaches". In: *Advanced drug delivery reviews* (2016).

[54] Christoph Kirsch and Eva Schmidt. *The DNA Integrity Number (DIN) Provided by the Agilent 2200 TapeStation System is an Ideal Tool to Optimize FFPE Extraction*. Tech. rep. 5991-5246EN. Application note. Agilent Technologies, 2015.

[244] Allon M Klein et al. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". In: *Cell* 161.5 (2015), pp. 1187–1201.

[55] Daniel C Koboldt et al. "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome research* 22.3 (2012), pp. 568–576.

[173] Piotr Kozlowski, Mateusz de Mezer, and Wlodzimierz J Krzyzosiak. "Trinucleotide repeats in human genome and exome". In: *Nucleic acids research* (2010), gkq127.

[56] Lasse S Kristensen et al. "Quality assessment of DNA derived from up to 30 years old formalin fixed paraffin embedded (FFPE) tissue for PCR-based methylation analysis using SMART-MSP and MS-HRM". In: *BMC cancer* 9.1 (2009), p. 1.

[57] Theodore G Krontiris and Geoffrey M Cooper. "Transforming activity of human tumor DNAs". In: *Proceedings of the National Academy of Sciences* 78.2 (1981), pp. 1181–1184.

[58] Mikael Kubista et al. "The real-time polymerase chain reaction". In: *Molecular Aspects of Medicine* 27.2-3 (2006). Real-time Polymerase Chain Reaction, pp. 95 –125. ISSN: 0098-2997.

[174] Madhu S Kumar et al. "The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer". In: *Cell* 149.3 (2012), pp. 642–655.

[59] Sunjong Kwon. "Single-molecule fluorescence in situ hybridization: quantitative imaging of single RNA molecules". In: *BMB reports* 46.2 (2013), pp. 65–72.

[60] Melissa J Landrum et al. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic acids research* 42.D1 (2014), pp. D980–D985.

[61] Pennina R Langer-Safer, Michael Levine, and David C Ward. "Immunological method for mapping genes on Drosophila polytene chromosomes". In: *Proceedings of the National Academy of Sciences* 79.14 (1982), pp. 4381–4385.

[269] David Larson and Travis Abbott. *Count DNA sequence reads in BAM files*. Website. https://github.com/genome/bam-readcount. Apr. 2016.

[175] Virpi Launonen. "Mutations in the human LKB1/STK11 gene". In: *Human mutation* 26.4 (2005), pp. 291–297.

[176] Ryan S Lee et al. "A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers". In: *The Journal of clinical investigation* 122.8 (2012), pp. 2983–2988.

[245] MPG Leers et al. "Heat pretreatment increases resolution in DNA flow cytometry of paraffin-embedded tumor tissue". In: *Cytometry* 35.3 (1999), pp. 260–266.

[62] Darryl Leja. *Fluorescence In Situ Hybridization (FISH)*. 2010. URL: https://www.genome.gov.

[177] Stefan H Lelieveld et al. "Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions". In: *Human mutation* 36.8 (2015), pp. 815–822.

[178] Johan Lennartsson and L Ronnstrand. "The stem cell factor receptor/ c-Kit as a drug target in cancer". In: *Current cancer drug targets* 6.1 (2006), pp. 65–75.

[63] Arnold J Levine. "p53, the cellular gatekeeper for growth and division". In: *cell* 88.3 (1997), pp. 323–331.

[64] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.

[65] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997* (2013).

[66] Heng Li. *Sequence Alignment/Map Format Specification*. Tech. rep. The SAM/BAM Format Specification Working Group, Nov. 2015.

[67] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5 (2010), pp. 589–595.

[68] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[69] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome research* 18.11 (2008), pp. 1851–1858.

[70] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[71] Astrid Lievre et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer". In: *Cancer research* 66.8 (2006), pp. 3992–3995.

[179] Hui Jun Lim, Philip Crowe, and Jia-Lin Yang. "Current clinical regulation of PI3K/PTEN/Akt/mTOR signalling in treatment of human cancer". In: *Journal of cancer research and clinical oncology* 141.4 (2015), pp. 671–689.

[180] TC Lin et al. "CEBPA methylation as a prognostic biomarker in patients with de novo acute myeloid leukemia". In: *Leukemia* 25.1 (2011), pp. 32–40.

[72] David J Lipman and William R Pearson. "Rapid and sensitive protein similarity searches". In: *Science* 227.4693 (1985), pp. 1435–1441.

[73] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions". In: *Human mutation* 32.8 (2011), pp. 894–899.

[270] Xiaoming Liu et al. "dbNSFP v3. 0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs". In: *Human mutation* (2015).

[246] FJ Livesey. "Strategies for microarray analysis of limiting amounts of RNA". In: *Briefings in functional genomics & proteomics* 2.1 (2003), pp. 31–36.

[181] Camille Lobry, Philmo Oh, and Iannis Aifantis. "Oncogenic and tumor suppressor functions of Notch in cancer: its NOTCH what you think". In: *The Journal of experimental medicine* 208.10 (2011), pp. 1931–1935.

[74] Elise MJ van der Logt et al. "Fully Automated Fluorescent in situ Hybridization (FISH) Staining and Digital Analysis of HER2 in Breast Cancer: A Validation Study". In: *PloS one* 10.4 (2015), e0123201.

[75] Katja Lohmann and Christine Klein. "Next generation sequencing and the future of genetic diagnosis". In: *Neurotherapeutics* 11.4 (2014), pp. 699–707.

[76] Dianne I Lou et al. "High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing". In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19872–19877.

[77] Fotios Loupakis et al. "PTEN expression and KRAS mutations on primary tumors and metastases in the prediction of benefit from cetuximab plus irinotecan for patients with metastatic colorectal cancer". In: *Journal of Clinical Oncology* 27.16 (2009), pp. 2622–2629.

[271] Irene Lurkin et al. "Two multiplex assays that simultaneously identify 22 possible mutation sites in the KRAS, BRAF, NRAS and PIK3CA genes". In: *PLoS One* 5.1 (2010), e8802.

[225] *MBCDeduplication Read Me File*. Version 1.0. Copyright by Agilent Technologies Inc. Agilent. 2015.

[247] Evan Z Macosko et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets". In: *Cell* 161.5 (2015), pp. 1202–1214.

[78] Umberto Malapelle et al. "Sanger sequencing in routine KRAS testing: a review of 1720 cases from a pathologist's perspective". In: *Journal of clinical pathology* (2012), jclinpath–2012.

[182] R Marone et al. "Targeting phosphoinositide 3-kinasemoving towards therapy". In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1784.1 (2008), pp. 159–185.

[79] A Marusyk and K Polyak. "Tumor heterogeneity: causes and consequences". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805.1 (2010), pp. 105–117.

[80] Tim Massingham and Nick Goldman. "All Your Base: a fast and accurate probabilistic approach to base calling". In: *Genome Biol* 13.2 (2012), R13.

[248] B Mercier et al. "Direct PCR from whole blood, without DNA extraction." In: *Nucleic acids research* 18.19 (1990), p. 5908.

[183] Fausto Meriggi et al. "The emerging role of NRAS mutations in colorectal cancer patients selected for anti-EGFR therapies". In: *Reviews on recent clinical trials* 9.1 (2014), pp. 8–12.

[265] Steve Michalik and Christopher Williams. "Qualitative multiplex PCR assay for assessing DNA quality from FFPE tissues and other sources of damaged DNA". In: *Life Science* (2008), p. 23.

[81] Patrick Micke et al. "Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens". In: *Laboratory investigation* 86.2 (2006), pp. 202–211.

[82] Lance D Miller et al. "Optimal gene expression analysis by microarrays". In: *Cancer cell* 2.5 (2002), pp. 353–361.

[83] André E Minoche, Juliane C Dohm, Heinz Himmelbauer, et al. "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems". In: *Genome Biol* 12.11 (2011), R112.

[249] H Mirhendi et al. "Colony-PCR is a rapid and sensitive method for DNA amplification in yeasts". In: *Iranian J Publ Health* 36.1 (2007), pp. 40–44.

[226] Lotte NJ Moens et al. "HaloPlex Targeted Resequencing for Mutation Detection in Clinical Formalin-Fixed, Paraffin-Embedded Tumor Samples". In: *The Journal of Molecular Diagnostics* 17.6 (2015), pp. 729–739.

[84] Lisle E Mose et al. "ABRA: improved coding indel detection via assembly-based realignment". In: *Bioinformatics* 30.19 (2014), pp. 2813–2815.

[184] Patricia AJ Muller and Karen H Vousden. "Mutant p53 in cancer: new functions and therapeutic opportunities". In: *Cancer cell* 25.3 (2014), pp. 304–317.

[185] Lois M Mulligan. "RET revisited: expanding the oncogenic portfolio". In: *Nature Reviews Cancer* 14.3 (2014), pp. 173–186.

[186] F Musumeci et al. "An update on dual Src/Abl inhibitors". In: *Future medicinal chemistry* 4.6 (2012), pp. 799–822.

[85] Yuki Nakayama et al. "Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions". In: *PloS one* 11.3 (2016), e0150528.

[86] Soo Kyung Nam et al. "Effects of fixation and storage of human tissue samples on nucleic Acid preservation". In: *Korean journal of pathology* 48.1 (2014), p. 36.

[87] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond". In: *Genome research* 25.10 (2015), 1499–1507.

[250] Nicholas E Navin and James Hicks. "Tracing the tumor lineage". In: *Molecular oncology* 4.3 (2010), pp. 267–283.

[88] Lex Nederbragt. "Developments in NGS". In: (2015).

[89] Saul B Needleman and Christian D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[251] Richard Novak et al. "Single-Cell Multiplex Gene Detection and Sequencing with Microfluidically Generated Agarose Emulsions". In: *Angewandte Chemie International Edition* 50.2 (2011), pp. 390–395.

[187] Magali Olivier, Monica Hollstein, and Pierre Hainaut. "TP53 mutations in human cancers: origins, consequences, and clinical use". In: *Cold Spring Harbor perspectives in biology* 2.1 (2010), a001008.

[90] K Page et al. "Detection of HER2 amplification in circulating free DNA in patients with breast cancer". In: *British journal of cancer* 104.8 (2011), pp. 1342–1348.

[91] Joseph F. Paone et al. "Serum UDP-galactosyl transferase as a potential biomarker for breast carcinoma". In: *Journal of Surgical Oncology* 15.1 (1980), pp. 59–66. ISSN: 1096-9098. DOI: 10.1002/jso.2930150110.

[188] Sharmila Patel and Mark R Player. "Colony-stimulating factor-1 receptor inhibitors for the treatment of cancer and inflammatory disease". In: *Current topics in medicinal chemistry* 9.7 (2009), pp. 599–610.

[252] Sharad Pathak et al. "Counting mycobacteria in infected human cells and mouse tissue: a comparison between qPCR and CFU". In: *PLoS One* 7.4 (2012), e34931.

[189] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[92] Deniz Pekin et al. "Quantitative and sensitive detection of rare mutations using droplet-based microfluidics". In: *Lab on a Chip* 11.13 (2011), pp. 2156–2166.

[93] SW Piraino and SJ Furney. "Beyond the exome: the role of non-coding somatic mutations in cancer". In: *Annals of Oncology* 27.2 (2016), pp. 240–248.

[253] Davide Prandi et al. "Unraveling the clonal hierarchy of somatic genomic aberrations". In: *Genome Biol* 15.8 (2014), p. 439.

[227] Hans Prenen, Sabine Tejpar, and Eric Van Cutsem. "New strategies for treatment of KRAS mutant metastatic colorectal cancer". In: *Clinical Cancer Research* 16.11 (2010), pp. 2921–2926.

[228] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 35.suppl 1 (2007), pp. D61–D65.

[229] Kim D Pruitt et al. "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". In: *Genome research* 19.7 (2009), pp. 1316–1323.

[254] S Pulciani et al. "ras gene Amplification and malignant transformation." In: *Molecular and cellular biology* 5.10 (1985), pp. 2836–2841.

[255] Simonetta Pulciani et al. "Transforming genes in human tumors". In: *Journal of cellular biochemistry* 20.1 (1982), pp. 51–61.

[94] *PyroMark PCR Handbook*. Qiagen. May 2009.

[95] *PyroMark Q24 MDx User Manual*. Qiagen. Jan. 2016.

[190] *QIAamp DNA FFPE Tissue Handbook*. 06/2012. Qiagen. June 2012.

[96] Michael A Quail et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC genomics* 13.1 (2012), p. 1.

[266] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. "The promise of whole-exome sequencing in medical genetics". In: *Journal of human genetics* 59.1 (2014), pp. 5–15.

[97] JA Ramos-Vara and MA Miller. "When Tissue Antigens and Antibodies Get Along Revisiting the Technical Aspects of ImmunohistochemistryThe Red, Brown, and Blue Technique". In: *Veterinary Pathology Online* 51.1 (2014), pp. 42–87.

[98] *Real-time PCR handbook*. CO32085 0812. With curtesy of Thermo Fischer Scientific. Thermo Fischer Scientific. Aug. 2012.

[99] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. "The advantages of SMRT sequencing". In: *Genome Biol* 14.6 (2013), p. 405.

[100] James T Robinson et al. "Integrative genomics viewer". In: *Nature biotechnology* 29.1 (2011), pp. 24–26.

[101] Mark D Robinson, Alicia Oshlack, et al. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biol* 11.3 (2010), R25.

[230] Ana I Robles and Curtis C Harris. "Clinical outcomes and correlates of TP53 mutations and cancer". In: *Cold Spring Harbor perspectives in biology* 2.3 (2010), a001016.

[102] Scott J Rodig et al. "Unique clinicopathologic features characterize ALK-rearranged lung adenocarcinoma in the western population". In: *Clinical Cancer Research* 15.16 (2009), pp. 5216–5223.

[103] Mostafa Ronaghi et al. "Real-time DNA sequencing using detection of pyrophosphate release". In: *Analytical biochemistry* 242.1 (1996), pp. 84–89.

[104] Michael G Ross et al. "Characterizing and measuring bias in sequence data". In: *Genome Biol* 14.5 (2013), R51.

[256] Camelia Iancu Rubin, Deborah L French, and George F Atweh. "Stathmin expression and megakaryocyte differentiation: a potential role in polyploidy". In: *Experimental hematology* 31.5 (2003), pp. 389–397.

[191] Charles M Rudin. "Vismodegib". In: *Clinical Cancer Research* 18.12 (2012), pp. 3218–3222.

[192] Elisa Rumi et al. "Clinical effect of driver mutations of JAK2, CALR, or MPL in primary myelofibrosis". In: *Blood* 124.7 (2014), pp. 1062–1069.

[267] G Rush et al. "Novel Improvements to the Illumina TruSeq Indexed Library Cconstruction, Amplification and Quantification Protocols for Optmized Multiplexed Sequencing". Poster. Feb. 2011.

[193] Antonio Russo et al. "The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment". In: *Journal of clinical oncology* 23.30 (2005), pp. 7518–7528.

[257] Antoine-Emmanuel Saliba et al. "Single-cell RNA-seq: advances and future challenges". In: *Nucleic acids research* 42.14 (2014), pp. 8845–8860.

[105] Fred Sanger and Alan R Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of molecular biology* 94.3 (1975), pp. 441–448.

[106] Mark Schena et al. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235 (1995), pp. 467–470.

[107] Suzanne Schubbert, Kevin Shannon, and Gideon Bollag. "Hyperactive Ras in developmental disorders and cancer". In: *Nature Reviews Cancer* 7.4 (2007), pp. 295–308.

[108] Michal R Schweiger et al. "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number-and mutation-analysis". In: *PloS one* 4.5 (2009), e5548.

[194] Alice T Shaw et al. "Ceritinib in ALK-rearranged non–small-cell lung cancer". In: *New England Journal of Medicine* 370.13 (2014), pp. 1189–1197.

[195] Q Sheng and J Liu. "The therapeutic potential of targeting the EGFR family in epithelial ovarian cancer". In: *British journal of cancer* 104.8 (2011), pp. 1241–1245.

[196] Karen E Sheppard and Grant A McArthur. "The cell-cycle regulator CDK4: an emerging therapeutic target in melanoma". In: *Clinical Cancer Research* 19.19 (2013), pp. 5320–5328.

[109] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[110] Shan-Rong Shi et al. "Evaluation of the value of frozen tissue section used as gold standard for immunohistochemistry". In: *American journal of clinical pathology* 129.3 (2008), pp. 358–366.

[111] Chiaho Shih et al. "Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts". In: (1981).

[197] Ritsuko Shimizu, J D Engel, and M Yamamoto. "GATA1-related leukaemias". In: *Nature Reviews Cancer* 8.4 (2008), pp. 279–287.

[112] Mano Sivaganesan et al. "Improved strategies and optimization of calibration models for real-time PCR absolute quantification". In: *Water research* 44.16 (2010), pp. 4726–4735.

[198] Martha L Slattery, Abbie Lundgreen, and Roger K Wolff. " VEGFA, FLT1, KDR and colorectal cancer: assessment of disease risk, tumor molecular phenotype, and survival". In: *Molecular carcinogenesis* 53.S1 (2014).

[113] Andrew M Smith et al. "Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples". In: *Nucleic acids research* (2010), gkq368.

[114] Temple F Smith and Michael S Waterman. "Identification of common molecular subsequences". In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[199] E Solary et al. "The Ten-Eleven Translocation-2 (TET2) gene in hematopoiesis and hematopoietic diseases". In: *Leukemia* 28.3 (2014), pp. 485–496.

[115] Jérome Solassol et al. "KRAS mutation detection in paired frozen and formalin-fixed paraffin-embedded (FFPE) colorectal cancer tissues". In: *International journal of molecular sciences* 12.5 (2011), pp. 3191–3204.

[116] Lucy F Stead et al. "Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing:

Applications in Tumor Subclone Resolution". In: *Human mutation* 34.10 (2013), pp. 1432–1438.

[200] Len Stephens, Roger Williams, and Phillip Hawkins. "Phosphoinositide 3-kinases as drug targets in cancer". In: *Current opinion in pharmacology* 5.4 (2005), pp. 357–365.

[117] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (2009), pp. 719–724.

[201] *SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library Protocol*. Version B2. Agilent Technologies. Apr. 2015.

[202] J Tan et al. "EZH2: biology, disease, and structure-based drug discovery". In: *Acta Pharmacologica Sinica* 35.2 (2014), pp. 161–174.

[118] Leinco Technologies. *Immunohistochemistry Protocol for Frozen Sections*. Website. Curtesy of Leinco Technologies. 2015.

[272] *Terra™ PCR Direct Polymerase Mix User Manual*. 31416th ed. Cat. Nos. 639269, 639270, 639271. Clontech Laboratories Inc. 2015.

[119] Steven M Teutsch et al. "The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP working group". In: *Genetics in Medicine* 11.1 (2009), pp. 3–14.

[120] K Thaker, R Shah, and M Berger. "The IMPACT of INDEL realignment: Detecting insertions and deletions longer than 30 base pairs with ABRA". Poster. Nov. 2014.

[268] Susannah Green Tringe et al. "Comparative metagenomics of microbial communities". In: *Science* 308.5721 (2005), pp. 554–557.

[121] Athanasios C Tsiatis et al. "Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications". In: *The Journal of Molecular Diagnostics* 12.4 (2010), pp. 425–432.

[122] Eliezer M Van Allen et al. "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine". In: *Nature medicine* 20.6 (2014), pp. 682–688.

[123] EH Van Beers et al. "A multiplex PCR predictor for aCGH success of FFPE samples". In: *British journal of cancer* 94.2 (2006), pp. 333–337.

[124] Laura J Van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *nature* 415.6871 (2002), pp. 530–536.

[125] *Variant annotations in VCF format*. http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf. Cingolani, Pablo et al. Jan. 2015.

[203] Cecily P Vaughn et al. "Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer". In: *Genes, Chromosomes and Cancer* 50.5 (2011), pp. 307–312.

[126] Kai Wang, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". In: *Nucleic acids research* 38.16 (2010), e164–e164.

[127] Kai Wang et al. "Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer". In: *Nature genetics* 46.6 (2014), pp. 573–582.

[204] Yuanxiang Wang et al. "Targeting mutant KRAS for anticancer therapeutics: a review of novel small molecule modulators". In: *Journal of medicinal chemistry* 56.13 (2013), pp. 5219–5230.

[205] Zhiwei Wang et al. "Targeting Notch signaling pathway to overcome drug resistance for cancer therapy". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1806.2 (2010), pp. 258–267.

[206] Zhiwei Wang et al. "Tumor suppressor functions of FBW7 in cancer development and progression". In: *FEBS letters* 586.10 (2012), pp. 1409–1418.

[207] E Weisberg and J D Griffin. "Mechanism of resistance to the ABL tyrosine kinase inhibitor STI571 in BCR/ABL–transformed hematopoietic cell lines". In: *Blood* 95.11 (2000), pp. 3498–3505.

[258] Richard Williams et al. "Amplification of complex gene libraries by emulsion PCR". In: *Nature methods* 3.7 (2006), pp. 545–550.

[231] Laurens G Wilming et al. "The vertebrate genome annotation (Vega) database". In: *Nucleic acids research* 36.suppl 1 (2008), pp. D753–D760.

[232] Stephen Q Wong et al. "Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing". In: *BMC medical genomics* 7.1 (2014), p. 1.

[208] Stephen Q Wong et al. "Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours". In: *Scientific reports* 3 (2013).

[128] Stephen Q Wong et al. "UV-Associated Mutations Underlie the Etiology of MCV-Negative Merkel Cell Carcinomas". In: *Cancer research* 75.24 (2015), pp. 5228–5234.

[129] Hongping Xia and Kam M. Hui. "Mechanism of Cancer Drug Resistance and the Involvement of Noncoding RNAs". In: *Current Medicinal Chemistry* 21.26 (2014), pp. 3029–3041. ISSN: 0929-8673/1875-533X. DOI: 10.2174/09298673216661404141019 39.

[233] J Xiao et al. "Association between urothelial carcinoma after kidney transplantation and aristolochic acid exposure: the potential role of aristolochic acid in HRas and TP53 gene mutations". In:

Transplantation proceedings. Vol. 43. 10. Elsevier. 2011, pp. 3751–3754.

[209] Mingzhao Xing. "Molecular pathogenesis and mechanisms of thyroid cancer". In: *Nature Reviews Cancer* 13.3 (2013), pp. 184–199.

[210] Min Yan et al. "HER2 expression status in diverse cancers: review of results from 37,992 patients". In: *Cancer and Metastasis Reviews* 34.1 (2015), pp. 157–164.

[211] L1 Yang et al. "A tumor suppressor and oncogene: the WT1 story". In: *Leukemia* 21.5 (2007), pp. 868–876.

[212] Chetan Yewale et al. "Epidermal growth factor receptor targeting in cancer: a review of trends and strategies". In: *Biomaterials* 34.34 (2013), pp. 8690–8707.

[130] Shawn E Yost et al. "Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens". In: *Nucleic acids research* 40.14 (2012), e107–e107.

[213] Bing Yu et al. "Targeting protein tyrosine phosphatase SHP2 for the treatment of PTPN11-associated malignancies". In: *Molecular cancer therapeutics* 12.9 (2013), pp. 1738–1748.

[259] Yong Zeng et al. "High-performance single cell genetic analysis using microfluidic emulsion generator arrays". In: *Analytical chemistry* 82.8 (2010), pp. 3183–3190.

[214] Xiuwen Zheng et al. "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24 (2012), pp. 3326–3328.

[131] *bcl2fastq2 Conversion Software Guide*. 2.17 (15051736 Rev. G). ILLUMINA PROPRIETARY. Illumina. July 2015.