Analysis of genomic alterations in morphologically normal tissue in prostate cancer patients reveals a potential role in tumour development.

Claudia Buhigas

This dissertation is submitted for the degree of *Doctor of Philosophy*.

University of East Anglia

Norwich Medical School

September 2020

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Claudia Buhigas Novoa

September 2020

AKNOWLEDGEMENTS

I would like to thank Dr Daniel Brewer, for his dedicated supervision and support throughout my PhD that greatly contributed to developing my data analysis skills. I would also like to thank Professor Colin Cooper for his support, guidance and insights into prostate cancer research. I would like to extend my gratitude to Prostate Cancer UK for funding this project.

I am indebted to Dr David Wedge, for his support and help with data analysis and interpretation of the results. A big thank you to Dr Rachel Hurst, for her patience and instruction in the lab. I am grateful to the cancer genetics team, especially Dr Jeremy Clark and Dr Ghanasyam Rallapalli, for always giving useful feedback on my research. To Dr Emma Manners and especially Dr Gemma Kay, for her assistance and supervision of the sequencing experiments. I am also thankful to Iñigo Martincorena and Dr Federico Abascal for their essential advice about experiment design.

A special thank you to Dr Anne Warren, for her patience and assistance with the sample collection, and Dr Kate Manley, whose insights were very useful for the preparation of this thesis.

I would also like to express my gratitude to the ICGC prostate group, for their invaluable advice and insights into my research.

Finally, I am grateful to my friends and family, that have contributed to these years in Norwich being a wonderful experience.

ABSTRACT

Up to 80 % of cases of prostate cancer present with multifocal tumour lesions leading to the hypothesis of a field effect present in an apparently normal prostate that predisposes it to cancer development. In this thesis we explore the development of the field effect in the prostate by analysing normal tissues.

We first applied Whole Genome DNA Sequencing (WGS) to morphologically normal tissue and benign prostatic hyperplasia (BPH) samples (n = 44) from men with and without prostate cancer. Substitutions (P =7.1x10^{-03}, Wilcoxon rank sum test) and indels (P = 9.5x10^{-04}) were significantly higher in morphologically normal samples, including BPH, from men with prostate cancer (median = 436) compared to those without (median = 141). Subclonal expansions under selective pressure were significantly associated with prostate cancer presence (P = 3.5x10^{-02}, Fisher exact test). Phylogenies reveal lineages were sometimes shared between BPH and normal tissues but were completely distinct from tumour clones.

Secondly, we gathered 95 samples from previously analysed normal tissue of a prostate cancer patient and performed deep targeted sequencing (> 500X) on a panel of 98 prostate cancer associated genes. We identified hundreds of mutations and validated the majority of the mutations previously found for this patient. Some genes showed repeated mutations in specific areas of the prostate whereas others were spread across the prostate. Apart from gene *MUC3A*, we did not find evidence of positive selection.

Our results show that field characterisation of the human prostate is associated with selected clonal expansions in morphologically normal tissue/BPH that expand under selective pressure by mechanisms that are distinct from those occurring in adjacent cancer, but that are allied to the presence of the cancer. Expansions are characterised by lack of recurrent driver mutations, by almost complete absence of structure variants/copy number alterations and by distinct mutational processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

aCGH	Array comparative genomic hybridization
AJCC	American Joint Committee on Cancer
ALK	Anaplastic lymphoma kinase
AR	Androgen receptor
AUC	area under the curve
BAF	B-allele frequency
BAM	Binary Alignment Map
BCL	Binary Base Call
BPH	Benign prostatic hyperplasia
BRASS	Breakpoints via Assembly
BWA	Burrows-Wheeler Alignment
BWA-MEM	Burrows- Wheeler Alignment – maximal exact match
BWA-SW	Burrows-Wheeler Alignment - Smith-Waterman
BWT	Burrows-Wheeler Transform
CAFs	cancer associated fibroblasts
CCF	cancer cell fraction
CNA	copy number alteration
COSMIC	catalogue of somatic mutations in cancer
CRUK	cancer research UK
CSCC	cutaneous squamous cell carcinoma
DBS	deoxyribonucleotide triphosphates
dNTPs	double strand break
DP	dirichlet process
DRE	digital rectal examination
EM	expectation-maximization
ESCC	Esophageal squamous cell carcinoma
FACS	fluorescence-activated cell sorting
FFPE	formalin-fixated paraffin-embedded
HCC	hepatocellular carcinoma
H&E	hematoxylin and eosin
HNSCC	head and neck squamous cell carcinoma

HR	homologous recombination
IMD	inter-mutational distance
INDEL	insertion and deletion
LCRs	low copy repeats
LUTS	lower urinary track symptoms
MMBIR	microhomology-mediated-break-induced replication
mpMRI	multiparametric magnetic resonance imaging
NAHR	non-allelic homologous recombination
NHEJ	non-homologous end joining
NGS	next-generation sequencing
NICE	National Institute for health and Care Excellence
NMF	non-negative matrix factorization
PC	principal component
PCA	principal component analysis
PCAWG	Pan Cancer Analysis of Whole Genomes
PCR	polymerase chain reaction
PIA	proliferative inflammatory atrophy
PIN	prostatic intra-epithelial neoplasia
PI-RADS	prostate imaging-reporting and data system
PRISM	pair-read informed split-read mapping
PSA	prostatic specific antigen
PSAD	prostatic specific antigen density
QP	quadratic programming
qPCR	quantitative PCR
RTA	real time analysis
SAM	Sequence Alignment Map
SBS	sequencing by synthesis
scRNA-seq	single cell RNA sequencing
scSeq	single cell sequencing
scWGS	single cell whole genome sequencing
SNP	single nucleotide polymorphism
SNV	single stranded DNA
ssDNA	single nucleotide variant

TCGA	The Cancer Genome Atlas
TNM	tumour node metastasis
TRUS	transrectal ultrasound
UICC	Union for International Cancer Control
UNC	universal neutrality curve
UGS	urogenital sinus
VAF	variant allele frequency
WES	whole exome sequencing
WGS	whole genome sequencing

CHAPTER 1 : INTRODUCTION

1.1 CANCER

Cancer is a group of diseases which are all characterized by an increased proliferation of cells and the subsequent invasion of other organs in the body (metastasis). In the hallmarks of cancer[1], Hanahan & Weinberg describe a series of characteristics that have to be present in cells that are essential for cancer development: sustainability of proliferative signaling, evasion of growth suppressors, resistance to programmed cell death (apoptosis), limitless replication, increased angiogenesis (blood vessels growth that enable the progression of the tumour), and invasion of other organs in the body (metastasis) (Figure 1.1). Recently, two other important characteristics have been proposed: alterations in the energy metabolism of the cell and immune system evasion by the cancer cells[2]. Genome instability and inflammation are known factors that contribute to the acquisition of the hallmarks of cancer.

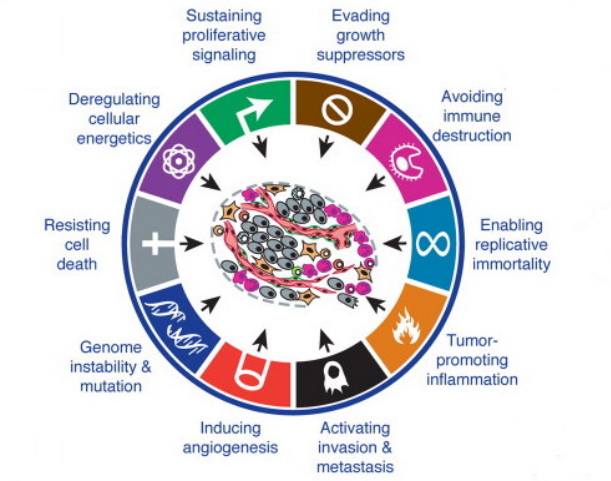


Figure 1.1: Ten hallmark traits. Adapted from Hanahan *et al.*[1].

These characteristics are generally acquired by cells through genetic alterations (mutations) and can lead to an increase rate of alterations in the cancer genome. If mutations start to accumulate in oncogenes, tumour suppressor genes or DNA repair genes, an abnormal cell can escape growth and regulatory control mechanisms, which in turn leads to the development of a tumour. Therefore, cancer is a disease of heritable changes to the genome.

It is widely assumed that cancer progression is driven by natural selection[3]. According to the clonal evolution model[4], cells carrying an advantageous mutation will grow into a bigger population called a clone. Clonal expansions will then diversify and suffer selection pressures in a highly dynamic microenvironment[5]. During this process there may be mutations that are fully clonal (present in 100% of cells) and subclonal (present only in a subset of cells).

Eventually, clones can gain the ability to invade surrounding tissues and metastasise, which can lead to the patient's death.

The evolution of clones is regulated by the occurrence of advantageous mutations or "drivers" that encourage cancer development[6]. A study by Martínez-Jiménez *et al.*[7] implemented the Integrative OncoGenomics (IntOGen) pipeline on more than 28,000 tumours from 66 cancer and identified 568 driver genes, which included previously reported genes and 152 potential novel driver genes. Some well-known driver genes reported in this study are tumour suppressors *TP53, PTEN, ARID1A,* epigenetic modifiers such as *KMT2C* and *KMT2D* and oncogenes *PIK3CA* and *KRAS.* Potential novel driver genes include *RASA1* and *FOXA2.*

There are also neutral mutations that do not influence cancer evolution and these are referred to as "passenger mutations". However, there is evidence that mutations classified as "passenger" could have an impact in cancer development under specific circumstances[5]. An example would be a mutation that normally would be identified as a passenger having an effect only when there is another mutation in another gene.

There are multiple models of cancer evolution. The classical model involves the consecutive acquisition of mutations and subsequent growth of subclones under selection pressure[4]. In Figure 1.2 we can see an example of a branched model of evolution, in which there is a high clonal heterogeneity since the early stages of tumour development. However, there is increasing evidence that cancer does not always follow a continuous steady development[8], but a model of punctuated evolution occurs, in which rapid changes occur at the beginning[8] (Figure 1.3). In these circumstances, events that promote tumour progression can occur simultaneously in short periods of time. Chromothripsis, first reported in lymphocytic leukemia by Stephens *et al.*[9], is an example one of these events. This mutational phenomenon is characterised by a massive genome shattering and reassembly that leads to hundreds of genomic rearrangements localised in a couple of chromosomes[10]. Point mutations can also arise in very short periods, which results in specific patterns of genomic alterations called *kataegis,* first observed in breast cancer[11]. It is primarily characterized by clusters of 10-20 base substitutions concentrated in 1-2 kilobases near genomic rearrangement sites[11]. Mutations tend to be C>T and C>G, usually appearing in a TpC mutation context, and located near large genomic structural variants

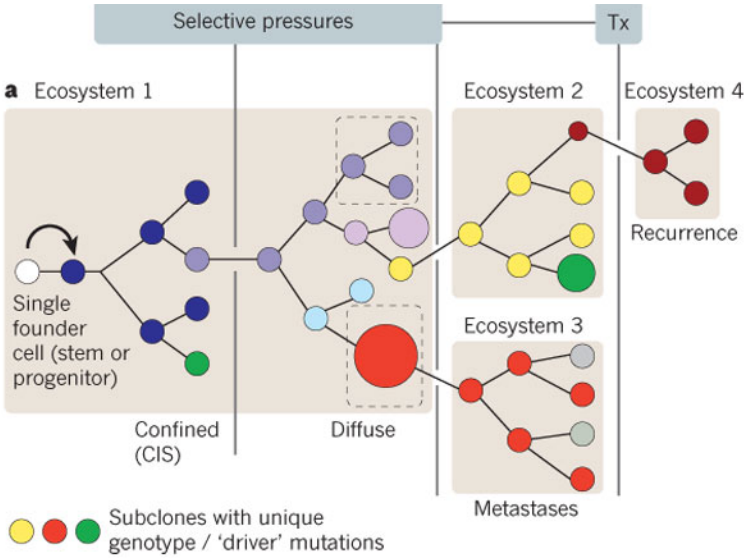


Figure 1.2: Cancer clones under selective pressures. Some subclones become dormant while others expand. Each coloured circle represents different subclones. Vertical lines indicate the presence of selective pressures. Adapted from Greaves *et al.*[5].

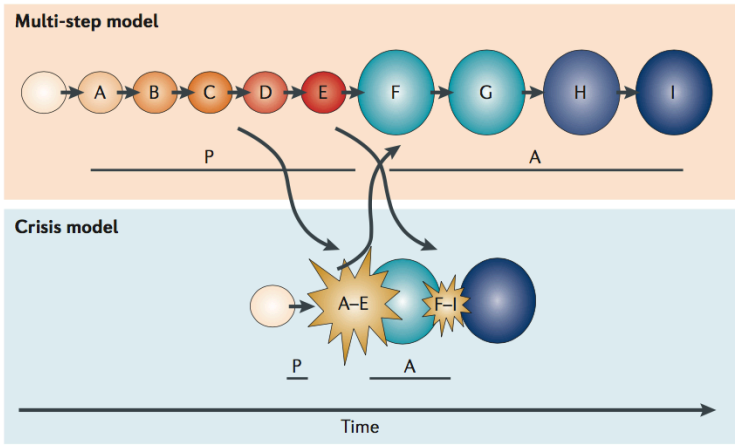


Figure 1.3: Stepwise model of evolution versus punctuated mutation events. Mutations A-E are needed for clonal expansion initiation and cancer development, and occur in the premalignant phase (P). Mutations F-H represent ongoing evolution leading towards the acquisition of more aggressive characteristics (Phase A). In the crisis model (punctuated evolution), the premalignant phase is almost non-existent[8].

1.1.1 CANCER HETEROGENEITY

Cancer heterogeneity is characterised by the detection of different cellular clones or subclones in the same tumour or between the primary tumour and metastases. They show different patterns of gene expression, histology and metastatic potential. Genetic heterogeneity has been reported in colon[12], lung[13], breast[14], ovary[15], kidney[16] and prostate[17,18], among other cancers. This is exemplified in a study by Gerlinger *et al.*[19], where branching tumour evolution was observed during the development of kidney cancer. In one patient, a small combination of mutations was observed in all specimens from the same tumour, but there was a subset of driver mutations that where found only in specific regions, indicating a highly intratumour heterogeneity. Phenotypic convergent evolution, in which similar adaptive strategies occur through different mechanisms, was detected in driver gene *SETD2*, where deleterious mutations were found in spatially separated samples.

Cancer heterogeneity is a hallmark of the dynamic evolution of the disease but creates a problem in determining the risk of progression and managing treatment[16,20]. As a single biopsy only accounts for a small amount of tissue, this means that a key biopsy may be missed, leading to a false picture of the state of the disease[16,21]. Even with a good characterization of the different branching clones, it is difficult to predict which ones are more aggressive.

1.1.1.1 IMPORTANT ROLE OF SUBCLONES IN CANCERS WITH HIGH INTRATUMOUR HETEROGENEITY

High genetic heterogeneity commonly leads to the forming of multiple and diverse subclones that could engage in both competition and cooperation. A study by Inda *et al.*[22] in a glioblastoma multiforme mouse xenotransplant model reported an example of cooperativity, where a cell subpopulation carrying a *EGFR* mutation promoted growth of all tumour cells. Sometimes cooperation between subclones has been critical for specific subpopulations that initially lack metastatic potential to invade other tissues, aided by processes driven by another, more invasive subpolulation of cells. This scenario is described by Chapman *et al.*[23] after observing that protease activity of cells around the primary tumour was a key step for other cells to metastasise as well. The ongoing branching process favoured by high genetic heterogeneity may also lead towards the development of treatment resistance and evasion of the immune system. This event occurs when a subclone is not affected by treatment and consequently becomes the dominant clone[19]. However, higher clonal diversity does not

necessarily predict a worse patient outcome. A study by Andor *et al.*[24] examining 12 different cancer types observed that survival decreased when more than two clones coexist in the same tumour but increased when more than 4 clones were present.

1.1.2 GENETIC AND EPIGENETIC ALTERATIONS IN CANCER

Cancer is characterized by heritable changes in the genome. These alterations or mutations are caused by exposure to mutagenic agents or errors in DNA replication and repair. They can be classified as somatic (occurring in a cell during the lifetime of the patient) or germline (inheritable from the cell of a parent). Germline mutations are present in every cell of the body from birth. Both germline and somatic have been associated with risk of cancer development. However, in this thesis we will focus only on somatic mutations.

1.1.2.1 POINT MUTATIONS

Point mutations are substitutions of a single nucleotide for another that can occur anywhere in the genome. When they occur in protein-coding regions of genes they are classified into three types: silent mutations or synonymous (the gene codes for the same aminoacid); missense mutations (the gene codes for a different aminoacid); and nonsense mutations (they code for a stop codon, so the aminoacid is not fully translated). The last two are also referred to as non-synonymous mutations. When they are somatic or single nucleotide variants (SNVs), they occur in a single cell in somatic tissue, some of them can lead to cancer[25]. Genes *TP53, PIK3CA* and *BRAF* have been found to be mutated in more than 10 % of patients in a wide range of cancers[25].

1.1.2.2 INSERTIONS AND DELETIONS

Insertions and deletions (INDELS) occur when one or more nucleotides are added or subtracted from DNA sequence during replication. Indels that occur in coding regions can alter the end protein product. They are classified in two types: frameshift and non-frameshift indels. Non-frameshift indels are characterized by addition or elimination of a multiple of three base pairs, which would introduce or delete one or more aminoacids but would not alter the rest of the protein. On the other hand, frameshift indels introduce a reading frame change that alters the protein sequence from the location where the indel occurred. Although both types are damaging, frameshift indels are more likely to result in the protein losing its function. In some cases, DNA fragments are inserted and deleted simultaneously, producing what it is called a

complex indel. Though complex indels represent a small proportion of the total indels, they have been detected in key cancer-associated genes[26]. Most of these affected genes are tumour suppressors genes (*PIK3R1, TP53, ARID1A, PTEN* and *ATRX*) and oncogenes (*EGFR, ALK, MET*).

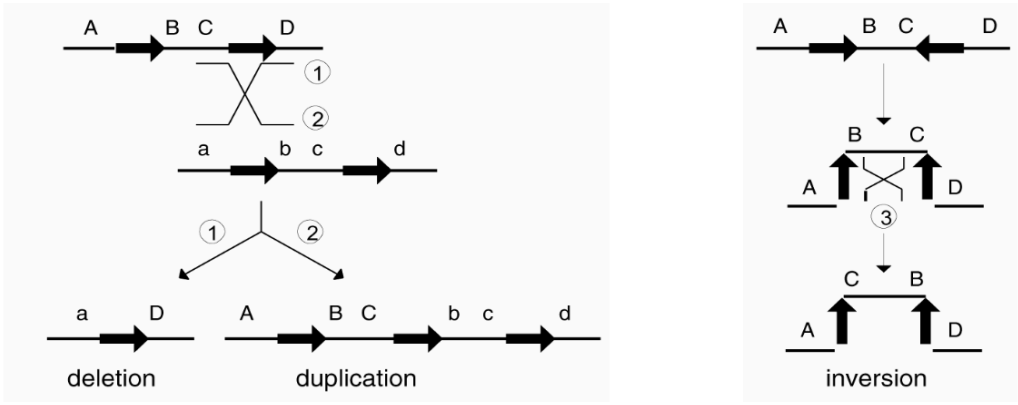
1.1.2.3 GENOMIC REARRANGEMENTS AND COPY NUMBER ALTERATIONS (CNAs)

Genomic rearrangements or breakpoints constitute DNA changes of a size that range from 100 base pairs to various megabases, therefore affecting large chromosomal regions. They can take the form of deletions, insertions, duplications, inversions (genetic material is inverted) and translocations (genetic material is exchanged between chromosomes). Rearrangements can be highly complex when several joining sites and more than two chromosomes are involved[27]. Sometimes small regions of DNA with sequence homology or "microhomology" are observed in the genomic rearrangements. They involve regions of homology of less than 70 bp that occur at the junctions of the rearrangement[28] (Figure 1.4). Genomic rearrangements are described as balanced when the exchange of genetic material does not lead to a gain or loss of the number of a gene. On the other hand, an unbalanced rearrangement could result in gene copy number alterations (CNAs). A copy number alteration is considered when the DNA fragment lost or gained is between 1kb and 3 Mb in size[29]. The effect can be neutral, but in many cases its occurrence is associated with cancer development, progression and metastasis[27]. In fact, many studies[30–32] show that the higher the percentage of the genome with CNAs, the worst the outcome for the patient. It has been observed that there are regions in the genome or "hotspots" where rearrangements arise in tumours but not in healthy tissues. Another study[33] found breakpoints events are observed in both tumour and normal, but the distribution across the genome is different: cancer associated breakpoints are found recurrently in specific regions whereas in normal tissue they are distributed in a uniform fashion.

Chromosomal rearrangements can produce gene fusions, which is an event that results in the combination of unrelated genes. A widely studied gene fusion is the Philadelphia chromosome found in chronic myelogenous leukemia. This translocation joins the viral oncogene *ABL* in chromosome 9 and the *BCR* gene in chromosome 22 in patients suffering chronic myelogenous leukemia[34]. The result is an altered kinase that increases the proliferation of myeloid cells by inhibiting apoptosis. In a similar fashion, the anaplastic lymphoma kinase (ALK) is altered by a gene fusion between the *ALK* gene and the protein EML4 triggered by an inversion. This event is observed in 2-5% of non-small cell lung cancers[27].

Wenli Gu *et al.*[35] described several mechanisms that can drive chromosomal rearrangements that can result in copy number changes[35]. During these processes separate segments of DNA are joined, which can result in fusion of genes, inversions of the DNA segment and translocations. Most cases are caused by Non-allelic homologous recombination (NAHR), by which two regions of DNA of 10-300 kb (called Low-Copy repeats or LCRs) that share a high similarity are aligned during meiosis or mitosis (instead of chromosomal allelic copies). This "crossover" can produce genomic rearrangements (Figure 1.4) that are localised in specific areas of the genome or hotspots, and therefore they tend to generate an increase of recurrent CNAs. Another studied mechanism that could result in copy number changes is non-homologous end joining (NHEJ)[36], a double-strand breaks repair pathway. NHEJ is likely to contribute to non-recurrent copy number changes by rejoining sequences that are not homologous. Finally, CNAs can occur through microhomology-mediated-break-induced replication (MMBIR). This mechanism involves joining sequences with microhomology regions after errors during DNA replication[28,37].

A



B

```
KIAA1549 GTATGCATATGTGTTTGTATGGAGAGAGTGAGGGGGAGAG
    BRAF TTATTTATGAATTAGCACATGGCCTTTGCCCTCAAGGAAG

  Fusion GTATGCATATGTGTTTGTATGGCCTTTGCCCTCAAGGAAG
```

```
TMPRSS2 GAAATGTGTCCATCATGTGG|GCCAGGGGTTCGTCTGCCTT
    ERG AATGGAAAACCCACAAAAGA|ACTGGACCAGGGACCAAAAA

  Fusion GAAATGTGTCCATCATGTGG|ACTGGACCAGGGACCAAAAA
```

Figure 1.4: (A) Genomic rearrangements[35]. (B) Microhomology at a rearrangement junctions. Adapted from Ottaviani *et al.*[28].

1.1.2.4 EPIGENETIC ALTERATIONS

Epigenetics is the study of processes that determine changes in cells of heritable nature but that do not make a change in the DNA sequence. One of the main epigenetic mechanisms is DNA methylation. A base located in the 5' position of a cytosine is methylated and constitutes an epigenetic mark in human DNA. Most of these marks occur in CpG dinucleotides which tend to be concentrated in specific regions of the genome but can also be located independently. In normal cells, methylation of dense CpG regions (also called CpG islands) is unusual, whereas these regions appear to be hypermethylated in cancer cells[38]. Hypermethylation of CpG has been associated with gene repression involved in silencing genes related to cancer suppression.

Other known mechanisms are histone modification and DNA and RNA protein interactions[38]. Histones are DNA-binding proteins that can suffer posttranslational modifications such as acetylation and methylation, among others. Along with methylation, these processes determine if chromatin is active (open) or inactive (closed), which in turn influences the possibility of gene expression/repression. Inactive chromatin has been commonly associated with cancer, as this prevents many critical regulatory proteins binding to DNA[39].

1.2 PROSTATE CANCER AND BENIGN PROSTATIC HYPERPLASIA

1.2.1 OVERVIEW

Prostate cancer is one of the four most common cancers worldwide, with an estimated 1,276,106 new cases diagnosed and 358,989 deaths a year (in 2018), most of them occurring in developed countries[40]. The majority of tumours are characterised by highly proliferating cells that form adenocarcinomas of acinar type. They originate from epithelial tissue and are characterized by the formation of acini and tubules[41]. Up to 75% arise in the peripheral zone, with only 20% and 5% arising from the transition and central zones respectively. Healthy prostate epithelium presents a secretory layer of columnar cells and a basal layer. As cancer progresses, the basal layer is lost, and the cells suffer a series of changes, such as uncontrolled growth, alterations in cell shape and cellular atypia (enlarged nuclei and nucleoli).

In approximately half of the prostate cancers detected by the PSA blood test (prostate specific antigen, section 1.2.2), the disease progresses slowly and poses no risk for the patient[42]. Radical radiotherapy or prostatectomy constitute the most common treatments with curative intent for

intermediate and high risk prostate cancer, both of which can produce lasting side effects that significantly affect the patient's quality of life, such as incontinence and erectile dysfunction[43]. It is important to make the distinction between indolent and aggressive prostate cancers, in order to give the appropriate treatment and avoid unnecessary overtreatment. However, prediction of cancer progression at initial diagnosis is still unfeasible and radical overtreatment may still be performed. So far risks factors for prostate cancer, and specifically, aggressive prostate cancer, are not clearly determined and there is limited understanding about the early stages of prostate cancer development. Advancing age is associated with lower survival, with a higher proportion of men (26%) presenting high-risk disease by age 75[44]. Also, in many cases prostate cancer is asymptomatic and can even be characterised as indolent disease. In an autopsy study by Jahn *et al.*[45] it was revealed that asymptomatic prostate tumours were detected in 36% of Caucasian and 51% of African-American men aged 70-79. Further studies are needed to deepen our understanding of the disease and to identify biomarkers that would allow for more accurate prognostic tests. This would greatly reduce the long-term life altering side effects that patients suffer from unnecessary overtreatment and the economic impact on health care systems.

1.2.2 THE PROSTATE

The prostate is a gland approximately the size of a walnut (11-16g) and forms part of the male reproductive system[46] (Figure 1.5B). It secretes an alkaline fluid that constitutes 30% of semen (prostate secretions), along with sperm and fluid from the seminal vesicles. The prostate secretions are mainly composed of calcium, zinc, citric acid, acid phosphatase, albumin, and prostatic specific antigen (PSA)[47,48].

Histologically, the prostate gland is composed of a secretory epithelium comprised of columnar cells and basal cells. It contains 30-50 tubuloalveolar glands, that excrete into 15-25 separate excretory ducts that open to the urethra. It is contained in a fibromuscular stroma composed of smooth muscle and connective tissue. There are different zones: the peripheral zone (contains large glands, the ducts open into the urethra), central zone (submucosal glands) and transition zone (mucosal glands) and anterior fibromuscular zone or stroma (Figure 1.5A).

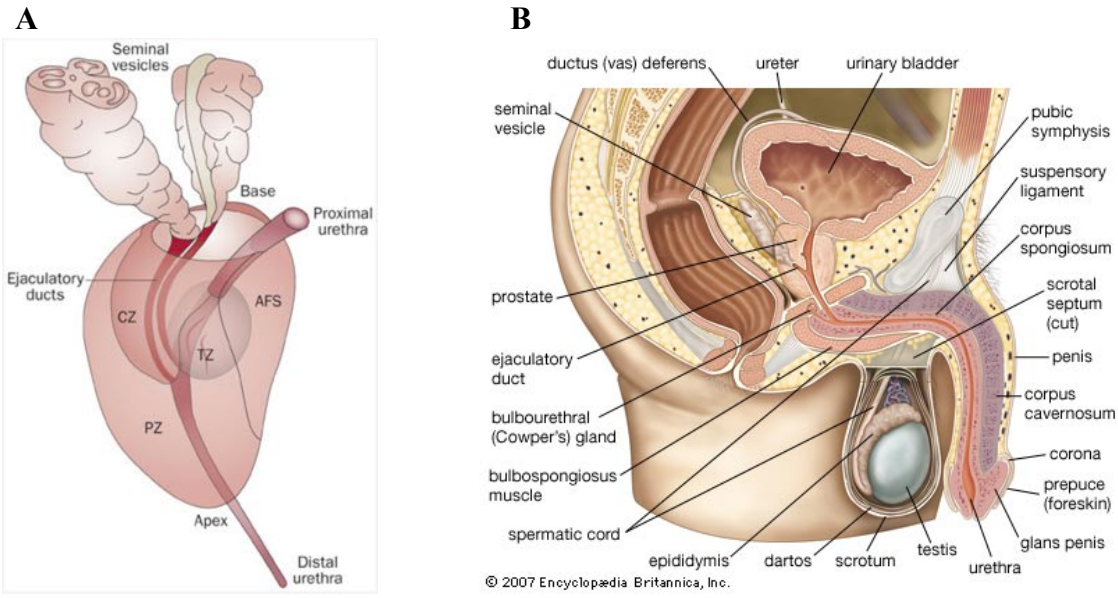


Figure 1.5: (A) Three zones of the prostate: the central zone (CZ), the transition zone (TZ) and the peripheral zone (PZ). The anterior fibromuscular stroma is also depicted. Based on a figure from Wadhera *et al.*[274]. (B) Male reproductive system. Taken from Encyclopaedia Britannica.

1.2.3 PROSTATE DEVELOPMENT

In the embryo, the male gonad starts to differentiate under the influence of the Y chromosome. Around the 10th gestational week, the production of fetal androgens by the gonad will stimulate a region of the urogenital sinus (UGS) mesenchyme to interact with the underlying epithelium. As a result, epithelial outgrowths, called prostatic buds, form and continue to interact with the mesenchyme during prostate development. This interaction will mediate duct formation, regulate epithelial proliferation and lead to the expression of epithelial androgen receptors (ARs) and specific prostate proteins[49].

Some aspects of prostatic development such as epithelial morphogenesis, proliferation and androgen receptor expression are mediated by androgen receptors present in the UGS mesenchyme[49], but the expression of androgen dependent secretory proteins is mediated by the epithelial AR. Similarly, the UGS mesenchyme will grow and differentiate into prostatic smooth muscle that will organise around acini, and later will constitute the stroma of the prostate[50].

In summary, three stages of development have been identified. During the first 20-30 weeks (bud stage) the ducts of the prostate are simple, columnar cells are observed in the basal layer and no lumen can be seen. Around 31-36 weeks (bud-tubule stage) cellular acini are present in both the peripheral and transition zones of the prostate. In the last weeks before birth, cluster of acinotubular glands form. Secretory protein PSA can be detected from 32 weeks onwards. The prostate suffers morphogenesis and growth during this prenatal stage, and then a second phase of growth occurs during puberty. During these periods of growth, cell proliferation is higher than cell death, but the normal adult prostate is characterized by a very low cell turnover rate[49]. Additional growth is considered to be pathological, which can occur during neoplastic processes but also with other conditions such as benign prostatic hyperplasia (BPH).

1.2.4 RISK FACTORS OF PROSTATE CANCER

Prostate cancer risk factors are not clearly defined, but it is widely known that the main contributing factor is old age. 64% of new diagnoses in the US are in men at least 65 years old[51]. Other factors such as family history and ethnicity are also associated with the disease[52]. Men with a family history of prostate cancer are at greater risk of developing it themselves, with the risk doubling when there is a first degree relative that has had prostate cancer. Interestingly, men with specific heritable genetic alterations increase prostate cancer risk, most prominent of which are mutations in the *BRCA1* and *BRCA2* suppressor genes that are involved in processes of DNA repair. In total, 70 germline variants have been identified to have an effect in prostate cancer risk and account for 30% of the heritable component of prostate cancer[53]. It has also been reported that prostate cancer incidence rate is higher among African-American men (185.4 per 100,000) in comparison to Caucasians (US:107 per 100,000, EU:34-100.1 per 100,000)[54], whereas Chinese men have the lowest rate (1.7 per 100,000)[54]. Although still controversial[55], these differences may be explained by environmental factors such as obesity, radiation exposure, androgenic (anabolic) steroids and red meat consumption.

1.2.5 DETECTION OF PROSTATE CANCER

Clinical assessment of suspected prostate cancer includes measurement of serum PSA and digital rectal examination (DRE) of the prostate. The PSA test measures the amount of PSA in the blood. Elevated levels of PSA have been associated with prostate cancer[48]. DRE is a procedure that involves the insertion of a gloved finger in the anus, to facilitate examination of

the surface of the prostate which lies behind the rectal wall. Abnormalities such as enlargement or changes in consistency are identified which could indicate the presence of a tumour.

PSA is an unreliable test because of its low specificity and sensitivity. A high PSA result can also be caused by BPH (Benign Prostatic Hyperplasia) or other causes such as prostatic infection or inflamation. Nevertheless, higher levels are expected in the presence of prostate cancer, and as an independent variable, it is still a better predictor than other strategies alone, such as DRE and transrectal ultrasound (TRUS)[55]. A level around 4 ng/ml indicates presence of disease 25% of the time[56] and the risk increases with higher levels of PSA. Patients with levels >100 ng/ml are almost certain to have metastatic disease and are informed about the probability of requiring hormonal therapy or chemotherapy[57]. The exact cut-off level for what is considered the normal PSA level has not been established, but levels of 2-3 ng/ml are often considered normal in young men[56]. Many institutions employ age and race-specific reference ranges for 'normal' PSA in clinical practice.

After assessment of PSA, DRE and a thorough clinical history to identify relevant risk factors (such as ethnicity, age and family history) patients are frequently offered multiparametric magnetic resonance imaging (mpMRI) as a first line screening tool. mpMRI is generally only offered to those patients who would be suitable candidates for radical intervention to treat their prostate cancer. Results from mpMRI are evaluated using a five-point Likert scoring system[58] and the Prostate Imaging-Reporting and Data System (PI-RADS). Higher scores (4-5) denote greater suspicion of underlying malignancy, whereas lower scores (1-2) indicate a low suspicion. Depending on clinical risk factors, patients with PI-RADS scores greater than 3 may be offered a biopsy to confirm[57]. PSA density (PSAD) (serum PSA (ng/ml) divided by prostate volume (mL)) of greater than 0.10 ng/ml^2 may also identify those patients for which biopsy is indicated. For scores of 1 or 2, serum PSA is usually repeated after 3-6 months to ensure a return to baseline levels. The detailed information regarding tumour location obtained from the mpMRI is used to determine optimum placement of the biopsy needle to increase the chances of detecting cancer.

Historically, the most common type of prostate biopsy was TRUS. The biopsy, guided by ultrasound allows the sample collection from the prostate for further analysis under the microscope by pathologists. A thin needle is inserted through the rectum wall into the prostate. A small "core" of prostate tissue is removed, and this is normally done 12 times. In spite of taking multiple "cores", it is still possible to miss tumour tissue i.e. sampling error. More

recently, the transperineal approach has been utilized more frequently in combination with the results of the mpMRI to provide greater accuracy in detection of prostate cancer. In this approach, the biopsy needles are inserted through the perineum through a metal grid, enabling systematic sampling of prostatic tissue and improved access to the anterior prostate.

As prostate cancer is multifocal in nature in 80% of cases[17], complete histologic and genomic representation of the cancer is often impossible, as foci may harbour different aberrations. Consequently, sometimes upwards of 30 cores are collected. Based on images of the biopsies a score called "Gleason score" is assigned by a histopathologist for classification purposes (section 1.2.6.1). If cancer is diagnosed, other tests such as isotope bone scans are performed to rule out the presence of metastatic disease before the patient can progress to radical treatment.

1.2.6 RISK STRATIFICATION AND TREATMENT

1.2.6.1 GLEASON SCORE

The Gleason score is the most common grading system in prostate cancer using biopsy samples. Depending on the level of differentiation and abnormal glandular growth patterns observed in the biopsy, a grade between 1 and 5 is given (Figure 1.6). The score is the sum of the most and second-most dominant types of glandular growth patterns in the tumour (e.g 3+4= 7). These values are also reported in addition to the Gleason score. This is relevant, as having a higher score in the most dominant pattern indicates a more advanced disease. If only one type is found, that grade is doubled e.g 3+3. As a rule, a grade must comprise at least 5% of the tumour to be considered[55]. A Gleason score below 6 is difficult to interpret, as there is uncertainty about the presence of cancer.

The Gleason score is currently one of the best prognostic factors for predicting clinical behavior and response to treatment[55] and has been modified throughout the years[59]. In a study by L Egevad *et al*.[60], the Gleason score had a very strong prognostic value as a predictor of death from cancer. Only 23% of patients with a Gleason score of 6 died from prostate cancer, in contrast to 70% of patients with a 8-10 Gleason score[60].

If the patient has had their entire prostate removed (radical prostatectomy), no cancer foci can be missed, a more accurate evaluation can be performed and the resulting Gleason score is referred to as pathological Gleason score. The extent to which cancer has spread is also better

assessed. For this reason, two categories in term of classification can be made: clinical stage (assessment done without radical prostatectomy) and pathological stage (assessment following surgery).

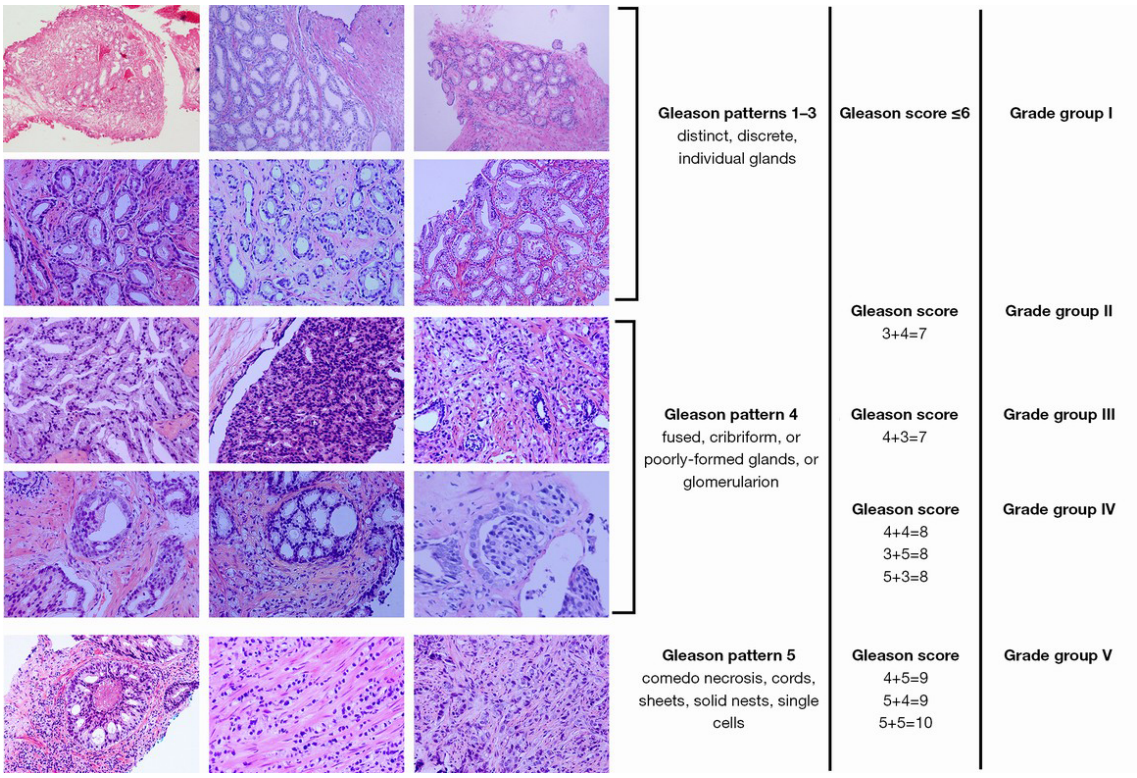


Figure 1.6: Prostate cancer grading system, adapted from Chen Ni *et al.*[59] using Gleason scores. Column 1-3, from left to right: most and second-most dominant types of glandular growth patterns for each group of H&E microscope slides from prostate samples; sum of each grade given to each pattern; grade prognostic groups according to Gleason score that range from I (most favorable) to V (least favorable).

1.2.6.2 TNM STAGE

The American Joint Committee on Cancer (AJCC) and UICC (International Union for Cancer Control) are the most common staging system for prostate cancer, called the TNM (Tumour node metastasis) system (Table 1.1)[55]. This system is used to evaluate treatment options and prognosis. It considers 3 characteristics that were first described by Brierley *et al.*[61]: description of the primary tumour site (T category), the possibility of spreading to the lymph nodes (N category), and the presence of metastasis (M category) (see Table 1.1).

T – Primary Tumour
TX Primary tumour cannot be assessed T0 No evidence of primary tumour T1 Clinically inapparent tumour that is not palpable T1a Tumour incidental histological finding in 5% or less of tissue resected T1b Tumour incidental histological finding in more than 5% of tissue resected T1c Tumour identified by needle biopsy (e.g., because of elevated PSA) T2 Tumour that is palpable and confined within prostate T2a Tumour involves one half of one lobe or less T2b Tumour involves more than half of one lobe, but not both lobes T2c Tumour involves both lobes T3 Tumour extends through the prostatic capsule [1] T3a Extracapsular extension (unilateral or bilateral) including microscopic bladder neck involvement T3b Tumour invades seminal vesicle(s) T4 Tumour is fixed or invades adjacent structures other than seminal vesicles: external sphincter, rectum, levator muscles, and/or pelvic wall.
N - Regional Lymph Nodes
NX Regional lymph nodes cannot be assessed N0 No regional lymph node metastasis N1 Regional lymph node metastasis
M – Distant Metastasis [2]
M0 No distant metastasis M1 Distant metastasis M1a Non-regional lymph node(s) M1b Bone(s) M1c Other site(s)
[1] Invasion into the prostatic apex or into (but not beyond) the prostatic capsule is not classified as T3, but as T2. [2] When more than one site of metastasis is present, the most advanced category is used. (p)M1c is the most advanced category

Table 1.1: TNM staging system. Adapted from Brierley *et al*.[61].

1.2.6.3 D'AMICO/ NICE CATEGORIES

In conjunction to the TNM system, disease management is determined by the assessment of PSA and Gleason score. A classification system using these two parameters that predicts the risk of progression in localized prostate cancer was developed by D'Amico and described in the NICE (National Institute for Health and Care Excellence) guidelines[57,62] (Table 1.2).

Level of risk	**PSA**		**Gleason score**		**Clinical stage**
Low risk	<10 ng/ml	and	≤6	and	T1 to T2a
Intermediate risk	10–20 ng/ml	or	7	or	T2b
High risk[1]	>20 ng/ml	or	8-10	or	≥T2c

Abbreviation: PSA, prostate-specific antigen.

[1] High-risk localised prostate cancer is also included in the definition of locally advanced prostate cancer.

Table 1.2: D'Amico categories. Adapted from NICE (2019)[57].

1.2.6.4 PRIMARY TREATMENT

For patients that are considered low risk according to the D'Amico classification system "Active Surveillance" is favoured over other forms of more aggressive treatment such as radical prostatectomy or radiotherapy. This consists of close monitoring of PSA, regular DRE examinations and mpMRI. For example, PSA readings occur every 3-4 months during the first year and every six months after the second year of diagnosis; DRE is performed annually and mpMRI is performed every 12-18 months. In case of suspicious clinical changes another biopsy is taken. This approach improves the patient's quality of life and reduces overtreatment. If cancer is locally advanced or the cancer is considered of intermediate risk, the most common treatment offered is radical prostatectomy (removal of the prostate), which has been well documented to increase prostate cancer specific survival[55]. The surgery includes the resection of the seminal vesicles and enough surrounding tissue in an attempt to ensure a negative margin. This procedure has common adverse side effects such as urinary incontinence and impotence, which can decrease quality of life substantially[43]. Other treatments for high risk patients include radiotherapy with or without hormonal therapy[55].

1.2.7 PRE-NEOPLASTIC LESIONS

Prostatic intra-epithelial neoplasia (PIN) and proliferative inflammatory atrophy (PIA) are two histologically classified precancerous lesions that are multifocal and heterogeneous[63,64] (Figure 1.7). PIN is characterized by non-invasive hyper-proliferative cells that have a malignant morphology, but still present an intact basal layer, whereas PIA shows inflammatory infiltrates, generally associated with atrophic tissue. It is widely accepted that both PIN and PIA act as a precursor to prostate cancer[63,65,66]. When present simultaneously with prostate cancer it has been observed that both chromosomal and genetic alterations correlate with primary tumours, indicating their involvement in prostate cancer development[67]. Specifically, PIA presents an increased expression of genes *Ki67, bcl-2, GSTP-1* and *COX-2*. However, the opposite has also been reported: PIN foci can be distinct to those observed in the main tumour mass[68].

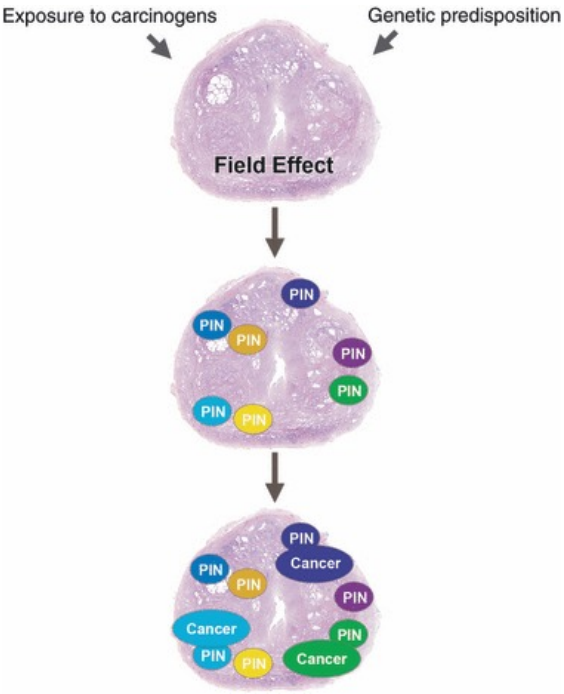


Figure 1.7: Field effect and progression from prostatic intraepithelial neoplasia (PIN) to multifocal prostate carcinoma[275].

1.2.8 BENIGN PROSTATIC HYPERPLASIA

Benign prostatic hyperplasia is characterized by an enlargement of the prostate, produced by increased proliferation of both epithelial and stromal cells. Nodules form in the transition zone of the prostate (around the urethra), and continue growing while pressing on the urethra[69]. This increase in pressure can result in bladder outflow obstruction and can lead to the development

of LUTS (Lower Urinary Tract Symptoms), among other disorders. The main risk factor for the development of this condition is age (as is the case in prostate cancer), with more than 70% men being affected by the age of 70 years[69]. The mechanisms that lead to this enlargement are believed to be mediated by androgens, estrogens, growth factors, inflammatory and immune mediators[70].

The two main factors that influence prostate development are androgens and stromal tissue. However, in the adult, prostate growth is quiescent. Therefore, there may be genetic, physiologic or environmental aspects favouring new growth. Because of this reason, BPH has been called a "reawakening" of processes that are involved in the organogenesis of the prostate[71].

The role of estrogen has also been investigated in BPH. Estrogens could have an "imprinting role" in the fetal prostate in utero[72]. A physiological increase of 50% of estrogen during gestation in mice had the consequence of an enlarged prostate in adulthood, showing a 6.6 fold increase in androgen receptor levels in comparison to controls. The number of glands also increased, especially in the areas BPH originates in the adult. This outcome suggests that estrogens can alter the sensitivity of the prostate to androgens and that is a long-lasting effect that could lead to dysfunction in growth regulatory patterns later in life.

Inflammation is also thought to be involved in the development of BPH[73]. A study in stromal fibroblasts revealed the presence of upregulated genes related to the secretion of inflammatory mediators[73]. These mediators (secreted by the stroma) produced an increase in the proliferation of both epithelial and stromal cells[74]. The presence of heterogeneous bacterial and viral strains has been reported in patients with BPH, which could explain the increased production of inflammatory mediators observed in these patients[73].

It is widely thought that BPH does not lead to prostate cancer development, although in many cases both conditions coexist[70]. Some studies[70,75] have linked BPH to prostate cancer, but the nature of this association is unclear and it is a topic of controversy within the urological community. Histologically, they are very different: most prostate cancers originate from epithelial cells in the peripheral zone of the prostate, whereas BPH originates in the transition zone[70]. However, BPH and Prostate cancer do have many aspects in common. As observed previously, both conditions are dependent on androgens for growth, respond to antiandrogen

therapy, show a high grade of inflammation and share genetic and epigenetic alterations[75].Therefore, it would be logically to consider BPH as a possible risk factor for prostate cancer. Furthermore, observations from historical autopsy data show that 83% of prostate cancer arises in the presence of BPH.

In a 100 single-nucleotide polymorphism (SNP) study by Saaristo *et al.*[76] BPH patients carrying the mutation in *HBOX13* gene (previously associated with an increased risk of hereditary prostate cancer[77]) were observed to be 4.6 times more likely to develop prostate cancer. Similarly, epigenetic alterations in tumour suppressor genes (*14-3-3σ, RASSF1A*), heavy metal binding genes (*MT1G*) and genes that code for proteins that are ABC transporters (*MDR1*) have been found in both BPH and prostate cancer have been found in both BPH and prostate cancer[75].

A recent study by Liu *et al.*[78] that examined the mutational, methylation and transcriptional landscape of BPH samples detected somatic substitutions and nine cancer mutational signatures (see section 1.3.4) in all samples. The most prevalent one was signature 1, which has been associated with age, the main contributor to BPH onset. In addition, analysis of the methylation profile in BPH revealed hypermethylation in CpGs from gene promoter regions and identified two different BPH subgroups: the first was referred to as "the stromal signature", defined by the differential expression of specific genes that had been previously associated with stromal samples by Tomlins *et al.*[79]; the second was associated with obesity and hypertension in the patients that was also confirmed by transcription changes in genes related to fatty acid and protein metabolism. However, clear evidence of BPH arising as a result of a cancerous process was not established because of the lack of driver mutations.

1.3 NEXT GENERATION SEQUENCING

Next-generation sequencing (NGS or high-throughput sequencing) refers to a group of technologies that allow researchers to rapidly sequence DNA and RNA by establishing the order of the bases A, T/U, C and G at low cost. Some examples are Illumina/Solexa sequencing (that use sequencing by synthesis, described in section 1.3.1), Roche 454 sequencing (pyrosequencing), Ion torrent: Proton/PGM sequencing (sequencing by synthesis, based on hydrogen ions detection) and SOLiD sequencing (sequencing by ligation). Whole genome sequencing (WGS) is characterised by sequencing all the regions in the genome, whereas whole

exome sequencing (WES) involves the selection of protein coding regions (exons and splice sites). It is also possible to perform targeted sequencing (described in section 1.3.2) and select only regions of interest, a method that allows scientists which reduces the cost considerably. Sequencing of the transcriptome can quantify gene expression levels, editing of RNA and alternative splicing[80], and has widely replaced microarray technologies. Other sequencing approaches like Bisulfite-Seq and ChIP-seq are used for the detection of changes in DNA methylation and histone modifications.

1.3.1 ILLUMINA SEQUENCING (SEQUENCING BY SYNTHESIS)

Illumina sequencing was developed by Shankar Balasubramanian and David Klenerman in 1998[81] and it is the platform used for sequencing experiments in this thesis. This method[82] is one of the most frequently used and it is currently generating more than 90% of the sequencing data in the world. The DNA has to go through a multistep process called library preparation before sequencing. Library preparation methods involve DNA fragmentation to short pieces (100-250bp) and adapter ligation to both 5' and 3' ends of the piece. An individual index (short sequence of 6-8 nucleotides) can also be added to identify each read during data analysis. Library preparation methods differ in many aspects such as the DNA fragment size and indexes used. The steps described above are a representative example.

Once library preparation is accomplished, the sequencing process is performed inside of a glass flow cell. Attached to the bottom of the cell are short nucleotide sequences called oligonucleotides. When the DNA enters the flow cell, the adapters will match the complementary oligonucleotides while sequencing takes place. First, thousands of DNA are replicated through a process called bridge amplification, where polymerase enzymes create a complementary strand (forward or reverse) and the original one is washed away. Secondly, the new strand bends and the adapter at the end attaches to oligonucleotides again for the same process to be repeated. A complementary strand (equal to the original) is created. Thirdly, the double strand DNA is denatured and both strands attach to new oligonucleotides to undergo bridge amplification again. This clonal amplification aids researchers to control the quality of the sequences and identify sequencing artefacts by allowing to compare forward and reverse strands.

After amplification a polymerase uses modified deoxyribonucleotide triphosphates (dNTPS) containing a reversible terminator with a fluorescent label that blocks any further polymerization (Figure 1.8). Each base is added one by one and will be detected by a camera (each base has a unique fluorescent emission). This sequencing reaction is carried on millions of template molecules at the same time, and it is done separately for each one of the four bases. When the camera records the images, the terminators are removed and the cycle is repeated. Base calls are recorded by the Real Time Analysis (RTA) software on the Illumina platform for every cycle and stored on BCL files as sequencing progresses. This is a binary file is the raw data output of the sequencing run and contains each base call and a quality score assigned to it. After the run is completed the BCL files are converted to FASTQ files, a text-based file format that stored the raw sequence data and the quality scores. For each flow cell lane two FASTQ files are created per sample during a paired-end run.

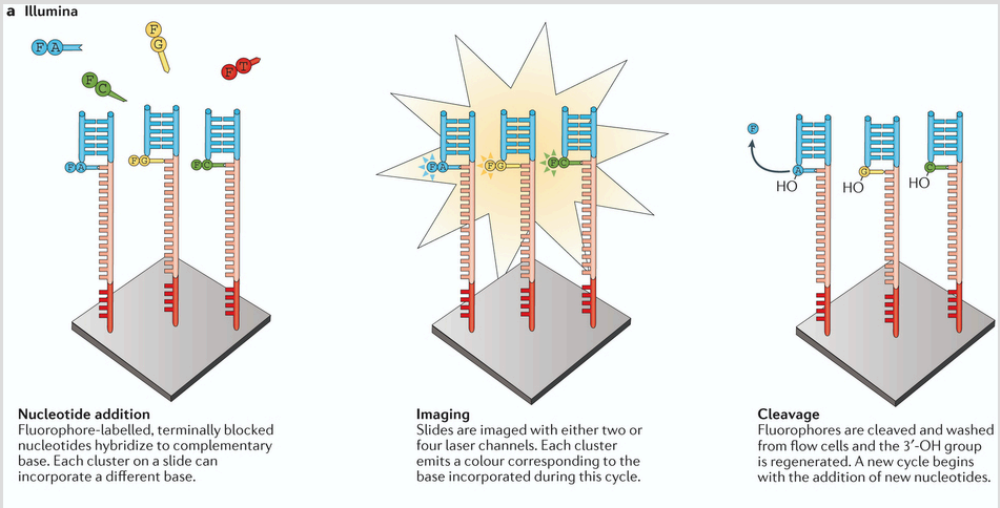


Figure 1.8: Sequencing by synthesis.

1.3.2 TARGETED SEQUENCING

Targeted sequencing involves the isolation of the DNA of only a subset of genes or regions and sequenced for the study of specific areas of interest, reducing costs and time. Common approaches to select the areas of interest are the amplification by PCR (polymerase chain reaction) of target regions by using oligonucleotide probes and hybridization capture method. In this thesis we used the hybridization capture method. This approach consists of adding biotinylated DNA or RNA bait molecules to the DNA of interest that represent the targeted DNA regions that will hybridize to the previously fragmented DNA. Then, using a magnet, the bait molecules are captured along with the regions of interest, obtaining a sequencing library

enriched for those regions (Figure 1.9). This method is of interest for studies that require sequencing with a very high sensitivity, such as the detection of subclones in heterogeneous tumours, validation of mutations with a great deal of certainty or mutations that occur at a very low frequency.

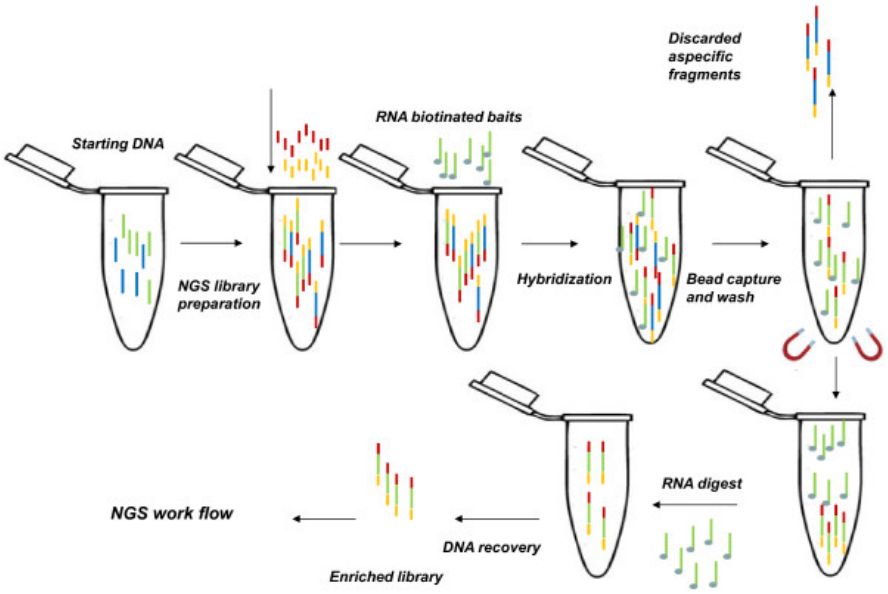


Figure 1.9: Library preparation for targeted sequencing. Adapted from Rizzi *et al*.[276].

1.3.3 SOMATIC VARIANT CALLING

1.3.3.1 POINT MUTATIONS

Many algorithms have been developed in order to detect single nucleotide variants (SNVs) from aligned sequencing data. Variation in the genome could be detected by measuring the number of occurrences of the variant alleles for each nucleotide in comparison to the reference allele. However, all the mutation calling methods have to account for sequencing errors, alignment errors and a variable proportion of cells with the mutation. For these reasons, different computational approaches have been developed to perform SNV calling with confidence and distinguish true variants from sequencing artifacts. Different approaches are used depending on the samples analysed: matched tumour-normal or single sample. In the first scenario, commonly used methods are probabilistic models and heuristic algorithms. Probabilistic models aim to estimate the probability for the joint genotypes $(G_T|G_N)$ by using the Bayes's rule[83]. This estimation assumes diploidy in tumour and normal samples, and therefore the variant allele frequency (VAF) of true variants is expected to be around 0.5-1.0.

A few examples are SAMtools[84], SNVSniffer[85] and CaVEMan[86]. CaVEMan achieves a high level of specificity and sensitivity, by applying post-processing filters[11]. This ensures that the false positives are removed and the true somatic variants are kept. Compared to other mutation callers it has been found to be amongst the top performers in terms of sensitivity and specificity[87].

However, the assumption of diploidy can pose a limitation when detecting variants present in subclones with different ploidy (as is frequently the case) or when the aim is to detect very low frequency variants from high coverage sequencing data. For these cases it is recommended to choose other callers such as Strelka[88], MuTect[89], LoFreq[90], MuSe[91] and deepSNV[92] that do not assume diploidy and jointly model the allele frequencies ($f_T|f_N$).

Another approach is to use heuristic algorithms (applied in tools like VarScan2[93] and Shimmer[94]) that take into account different factors from the sequencing reads such as allele count, read quality and read depth and apply thresholds to select potential SNVs.

1.3.3.2 DETECTION OF STRUCTURAL VARIATION: INDELS AND REARRANGEMENTS

Some strategies to identify structural variation are de-novo assembly of unaligned reads (SOAPindel[95], BRASS (https://github.com/cancerit/BRASS)), splitting of reads (Pindel[96], PRISM[97]) and depth coverage analysis (CVNator[98]). Assembly can be performed *de novo* by using de Bruijn graphs[99], which requires a large amount of computational resources. The splitting of reads works by breaking the sequencing reads and mapping them separately to a reference genome. The exact location, size and type of indel (deletion, insertion or complex) is given by the mapping location in the reference and its orientation. It is very accurate for short (1-20 bp) and medium sized indels (1-50-1000 bp) but often miss larger indels because of mapping limitations[100]. Read-depth of coverage approaches frequently miss small indels but are more effective for identification of changes above 1 kb and copy number variants. They are based on the detection of regions with significantly higher or lower coverage. Read depth is expected to be proportional across the genome, so a change produced by duplicated or deleted regions results in read depth variations.

1.3.3.3 DETECTION OF COPY NUMBER ALTERATIONS

In the past, common approaches for CNAs identification were developed for array based technologies, such as whole genome array comparative genomic hybridization (aCGH)[101] and single nucleotide polymorphism (SNP) arrays[102]. aCGH examines the relative frequency of probe DNA segments between two genomes, while SNPs arrays compares the probe intensities at known loci with another genome to detect allele changes. With the development of new high throughout sequencing, strategies based on whole genome sequencing data have been developed. In a previous study the battenberg algorithm was used for detecting clonal and subclonal CNAs[103] . Copy number alterations may lead to changes in allelic frequency in tumour samples. Allele frequencies come from binomial distributions, as their value can be estimated from the proportion of reads present of each allele. When these frequencies are very different, estimates of the B-allele frequency (BAF) can be obtained. This approach is described in detail in Methods, section 2.9.5.

1.3.4 MUTATIONAL SPECTRA IN CANCER

Cancer genomes carry somatic mutations that result from mutational processes that occur from the time the egg is fertilized[104]. In the majority of cases these mutations are repaired, but some are not and lead to a fingerprint of that process on the genome. The causes that drive the mutational processes can be exogenous or endogenous. Tobacco smoke and ultraviolet light are the most common exogenous causes. In lung cancer, it has been observed that tobacco smokers had on average a 10-fold increase of somatic mutations in their genome, in comparison to non-smokers[105,106]. For example, in the case of tobacco smoke, the mutational process is characterized by an increase of the number of C>A transversions[107].

Different mathematical approaches can be used to decipher the mutational signatures that contribute to each sample from the SNV called[104,108,109]. Perhaps the most common is the application of Non-Negative Matrix Factorization (NMF) to trinucleotides. The six substitution types are considered (C>A, C>G, C>T, T>A, T>C, T>G), as well as the bases present before (5') and after (3') the mutated base. This provides 96 possible scenarios of mutation types (6 substitution types * 4 types of 5' base * 4 types of 3' base). Mutational signatures are defined as different weightings of these 96 types. For each sample, the relative proportion of mutations caused by each signature are reported.

This model was used to analyse 4,938,362 mutations from 7,042 cancers and was able to decipher more than 20 different mutational signatures[109]. These signatures are used regularly in the community and are referred to as the COSMIC (Catalogue Of Somatic Mutations In Cancer) mutational signatures. Signature 1 is thought to result from an endogenous mutational process started by the deamination of 5-methylcytosine and is also characterized by small numbers of small insertions and deletions. Along with signature 5 they have been associated with age. Other signatures have also been linked to specific biological processes, such as signature 3, 4 and 7. Signature 3 has been associated with a very high number of large insertions and deletions produced by a failure of DNA double-strand break-repair by homologous recombination and BRCA mutations. Signatures 4 and 7 have been associated to tobacco smoking and ultraviolet light exposure, respectively.

A more recent study by Alexandrov *et al.*[110] has updated these signatures by analysing 23,829 samples from a wide range of cancer types, including samples from patients treated with chemotherapy. This dataset was comprised of 2,780 highly curated PCAWG (Pan-Cancer Analysis of Whole Genomes) whole genomes, 19,184 exomes and 1,865 other whole genomes. All previous signatures except signature 25 were confirmed in this analysis and a total of thirteen new signatures were detected (Figure 1.10). Two of them (31 and 35) were associated with previous chemotherapy treatment. In addition, some signatures were found to be more complex than previously thought and were comprised of multiple previously undetected mutational processes that result in different substitution patterns. Consequently, these signatures (7,10 and 17) were split into two or more signatures.

However, the NMF method requires a very high number of samples (>200). A solution to this problem is applying quadratic programming methods. This approach involves using the previously known signatures discovered by Alexandrov *et al*.[110] to estimate the contribution of each signature for a given sample (Methods, section 2.12).

1.3.5 DETECTION OF CLONAL EXPANSIONS

A clone is a group of cells that carry the same genetic alterations. A clonal expansion is characterized by the increasing number of a group of cells. The number of clones that are present in a tumour sample can be detected by analysis of the variant allele frequency (VAF) of SNVs detected by NGS technologies. Mutations that are early or clonal, will be found in the

majority of cancer cells in a sample and so are expected to have a high VAF, whereas late or subclonal mutations would be expected to occur in fewer cells and so have a low VAF. Other aspects that are necessary to estimate are tumour purity and copy number alterations (CNAs)[21]. Combining the VAF with this information gives a Cancer Cell Fraction (CCF), so CNAs and tumour purity information is important for a correct estimation of the CCF. Substitutions with similar CCFs will be clustered together into clones or subclones.

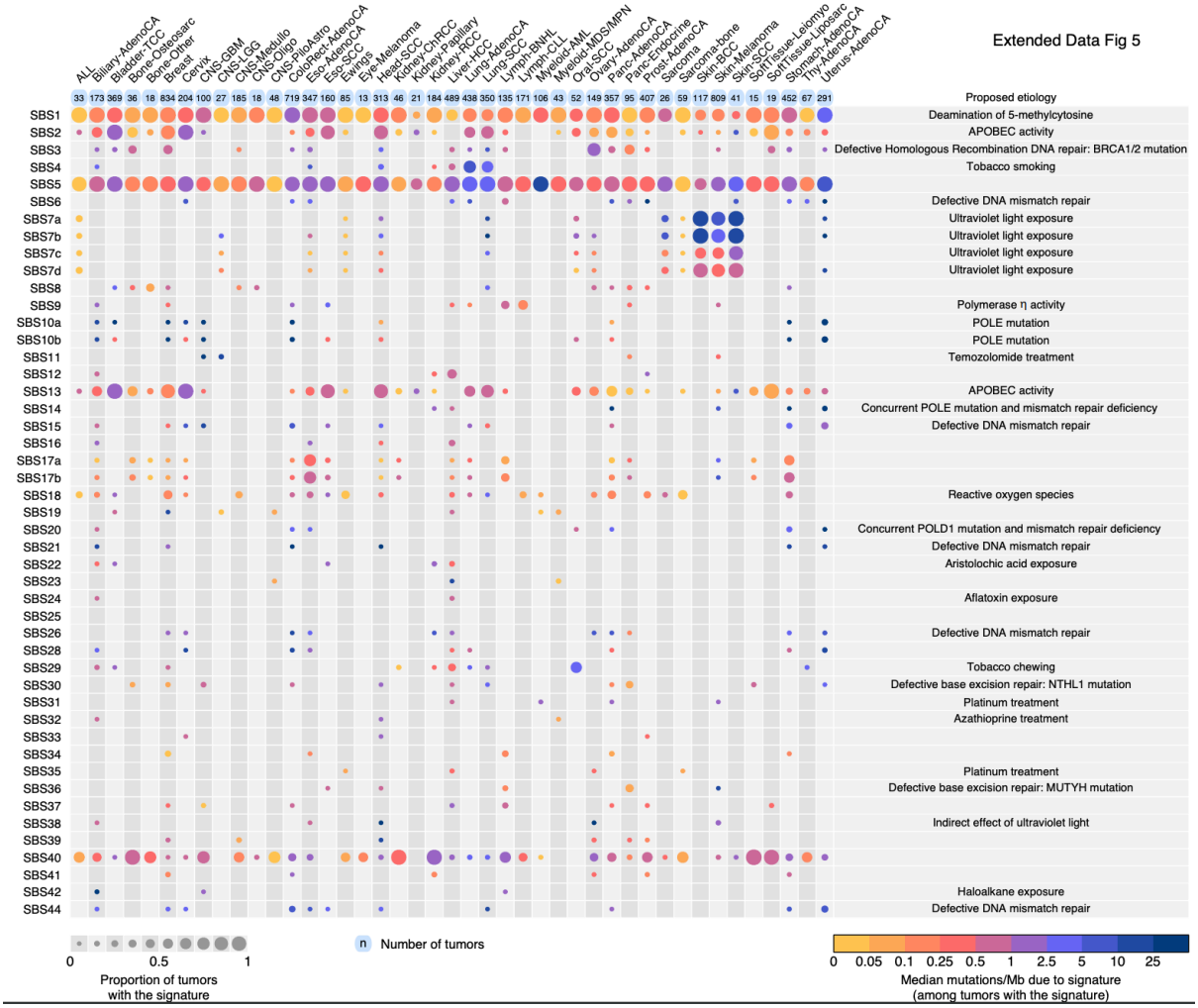


Figure 1.10: Mutational signatures across human cancer[110]

In recent years, computational methods have been developed to analyse the presence of clonal and subclonal expansions in tumours. Most of them are based on Bayesian analysis, a method that allows to create a model of clonal structure with many uncertainties, such as the number of subclones and tumour cell fractions (which are unknown)[21]. DPClust, a Bayesian Dirichlet process[111] has been used in many studies[17,103,112] for this purpose (Methods, section 2.11.1). This method models clusters of clonal and subclonal somatic mutations, estimates the number of clones, the fraction of cells carrying that clone and the number of mutations in each clone.

Other approaches, like CloneHD[113], build the subclonal architecture using CNAs and somatic variants by producing a model of relationships applying Hidden Markov Models.

1.3.6 METHODS FOR DETECTING DRIVER GENES

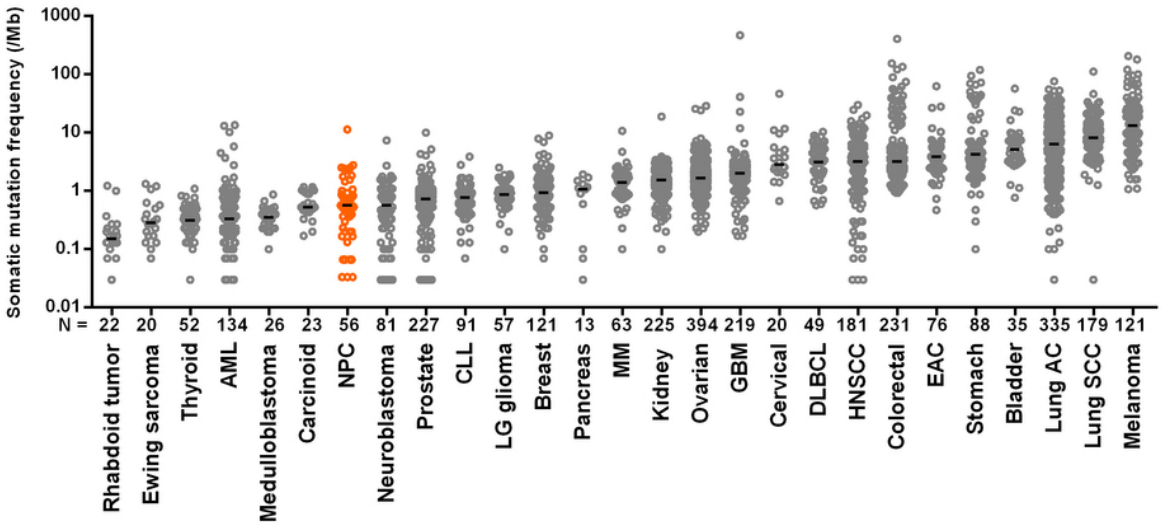
Driver genes can be detected by analysing substitutions patterns in multiple samples. Somatic mutations that are silent (synonymous) are considered to be neutral or passenger mutations, whereas non-synonymous mutations are considered to have the potential to be drivers, or positive selected. The most common method for analysing this is the calculation of the non-synonymous to synonymous ratio in each gene/region of interest. Selection is considered to be present when the dN/dS ratio deviates from what it would be expected by chance (dN/dS ~1). However, one of the main problems of these methods is the inability to determine whether a mutation is pathogenic or not, as non-synonymous substitutions can be passenger mutations as well[114]. For this reason, many methods identify those genes that are mutated more frequently than it would be expected by chance in comparison to a simulated background mutation rate, such as MutSig2CV[115], OncodriveCLUST[116] and in some cases combine this information with a predicted functional impact for each mutation (OncodriveFML[117]). Other methods such as dNdScv[118] (described in Methods 2.15.2) include the correction for gene length, and calculates the mutation rate for each type of mutation separately (missense, nonsense and splice variants).

1.4 PROSTATE CANCER GENOME

1.4.1 GENETIC ALTERATIONS

Primary prostate cancer has a low coding mutation rate, similar to those of acute myeloid leukemia and breast cancer; 7 to 15-fold lower than melanoma or lung cancer (Figure 1.11). Up to 80% of cases of prostate cancer constitute multifocal disease[17,20]. Multifocal prostate cancer is characterized by the presence of multiple tumour foci in the prostate. In recent studies it has been observed that the different foci are of independent origin[17,119]. The genomic lesions in prostate cancer show a very high variability not only between foci (intratumour heterogeneity), but also between patients. Intratumour heterogeneity is characterized by the detection of different cellular clones or subclones in the same tumour, which show different patterns of gene expression, histology and metastatic potential.

Many studies have reported multiple genomic alterations in prostate cancer but it is best characterized by the presence of rearrangements[120]. They can cause the dysregulation of multiple genes related to prostate development, androgen signaling, PI3K and WNT pathways



and cell cycle regulation[121]. PI3K pathway in involved in cell proliferation, cell survival and angiogenesis whereas WNT pathway plays an important role in development, cell adhesion and polarity, among others. The most common genetic alterations are discussed below.

1.4.1.1 SNPs

An evaluation of SNPs associated with prostate cancer risk in the large PRACTICAL (Prostate Cancer Association Group To Investigate Cancer Associated Alterations in the Genome) consortium [122] identified 63 novel variants, of which 38 SNPs were discovered in gene-rich regions (intronic, missense and UTR regions), affecting genes strongly involved with cell cycle and DNA repair pathways such as *ATM, CDKN1B, INCENP* and *HAUS6,* among others. These variants, in combination with 85 previously identified loci associated a higher susceptibility of prostate cancer development explain up to 28.4 % familial relative risk for the disease.

1.4.1.2 SNVs

Figure 1.11: Somatic coding mutations rates of human cancers[277].

Unlike other cancers like renal cell carcinoma[123], no common initiating mutation has been detected in prostate cancer[124] and the most recurrent genes are at relatively low level. In a study by Zhao *et al.*[125], 333 genes were classified as drivers in prostate cancer, of which *SPOP, TP53,*

SPTA1, AHNAK, HMCN1, ATM, FOXA1 CSMD3, LRP1B and *FREM2* are among the most recurrently mutated genes. Other studies also report highly recurrent mutated genes such as *COL5A1, MED12*[126] and *ARID1A, CASZ1, CNOT3, PIK3R1, TBX3* and *ZMYM3*[127]. A recent study by Armenia *et al*[128]. examined whole exome data from 1013 samples from primary and metastatic prostate tumours. Metastatic samples showed an enrichment of mutations in genes *TP53, AR, PTEN, RB1, FOX1, APC*, *BRCA2* and epigenetic modifiers *KMT2C* and *KMT2D* in comparison to primary tumours, whereas the *SPOP* gene was significantly enriched in primary tumours. In addition, they detected 97 novel mutated genes at very low frequencies, of which 70 had been previously reported in cancer and 9 were specific to prostate cancer. A few examples are gene *PIK3R2* (involved in the PI3K pathway), *SPEN*, (involved in the androgen signaling pathway), DNA repair genes *MRE11A* and *PALB2*. Ubiquitin protease genes *USP28, USP7* and *CUL3* were also significantly mutated.

1.4.1.3 INSERTIONS AND DELETIONS

Chromosomal deletions of tumour suppressor gene *PTEN* at chromosome 10q23.3 are present in 40% of localized prostate cancers and in 60% of metastatic cancer[129], and has been associated with a higher risk of metastasis. Other known prostate cancer genes that present recurrent indel events are *TP53, AR, KMT2C, KMT2D, RB1, APC, BRCA2, CDK12, ZFHX3* and *PIK3CB* and were significantly enriched in metastatic samples in comparison to primary tumours[121,127,130].

1.4.1.4 GENOMIC REARRANGEMENTS

The gene fusion of the *ERG* gene from the ETS family of transcription factors (involved in transcription regulation) and androgen-responsive promoter *TMPRSS2* (both located in chromosome 21) have been reported in 50% to 75% of cases[131]. It has been associated with a worse outcome than *TMPRSS2-ERG* negative cancers in some studies[132], but this association is still controversial[133]. Other genes from the ETS family such as *ETV1, ETV4* and *ETV5* are also frequently involved in fusions with other partners other than *TMPRSS2*[134]. Overall, these rearrangements can lead to gene activation or repression, which in turn can affect oncogenic signaling processes[131]. Because they tend to be distributed homogeneously within a discrete tumour lesion it is thought to be an early event in the disease[132]. As *TMPRSS2* is androgen-responsive, the *TMPRSS2-ERG* fusion and androgen receptor (AR) relationship has been investigated. Several studies report an association between structural rearrangements and AR activity in early onset prostate cancer[135]. The AR achieves this by inducing a spatial proximity

between the genes present in the rearrangements[131]. Interestingly, an abundance of androgen driven rearrangements have been reported in early onset prostate cancer, whereas more elderly patients showed non-androgen associated rearrangements[136].

Sequencing of the transcriptome has revealed that other rearrangements between several *RAF* kinases such as *SLC45A3-BRAF* and *ESRP1-CRAF1* are present in 1-2 % of prostate cnacer patients. A higher Gleason score and more advanced disease was a common trend among these patients[137]. Similarly, the analysis of RNA-Seq data from 14 primary prostate cancers and matched normal tissue revealed 37 novel fusions in tumour tissue and 3 of them were recurrent: *SDK1-AMACR, RAD50-PDLIM4* and *CTAGE5-KHDRBS3*[138].

1.4.1.5 COPY NUMBER ALTERATIONS (CNAs)

As mentioned earlier, chromosomal rearrangements can lead to a higher frequency of copy number alterations (CNAs). The most frequently observed copy number losses include deletion of chromosome 8p (affecting tumour suppressor gene *NKX3*), loss of chromosomal region 13q13.1-q31.1 (which surrounds gene *RB1*) and deletions at the 16q region. In contrast, frequent gains have been reported in chromosome 8q, chromosome 7 and chromosome 16p[139]. The percentage of the genome affected by CNAs in prostate cancer has been found to be correlated with tumour grade, biochemical recurrence and metastasis[17,120]. For example, it is known that the androgen receptor (AR) gene undergoes amplification (copy number gains) after hormonal therapy and it is a negative prognostic factor for overall survival[140,141]. Similarly, a study by Camacho *et al.*[30] reported nine specific copy number changes (two deletions and seven gains) associated with relapse.

Specific copy number changes have also been associated to different evolutionary patterns in prostate cancer regarding ETS rearrangements[127]. Deletions between *TMPRSS2* and *ERG* genes in ETS positive cancers occur at early stages of development and usually occurs simultaneously with deletions in tumour suppressor *PTEN*, deletions at 17q21.31 and the amplification at 16p13.3[30]. On the other hand, ETS negative cancers are characterised by copy number losses in genes *CHD1, RGMB, BRCA12, RB1* and *FOXO1*[127] and amplifications in genes *EGFR* and *MYC*[30]. *CHD1* has been specifically associated with the initiation of ETS negative cancers, as this event prevents the TMPRSS2-ERG rearrangement. Other copy number losses have been observed in both ETS positive and negative cancers, which seem to be triggered by whole genome duplication.

1.4.2 MUTATIONAL SPECTRA IN PROSTATE CANCER

Initially, prostate cancer was associated with signatures 1 and 6[109]. Later, a study by Cooper *et al.*[17], signatures 1A and 5 were identified in tumour tissue from a group of three prostates. Signature 8 was identified in some samples of one prostate in tumour tissue only, whereas signatures 1A and 5 were present in both tumour and normal tissue from two of the prostates. These results show that abnormal processes are also at work in morphologically normal tissue, giving further evidence that a field effect is at work (Figure 1.12).

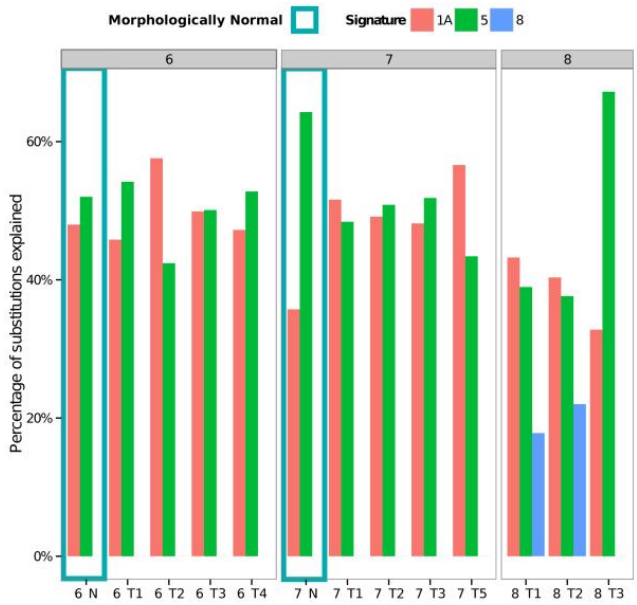


Figure 1.12: Relative contributions of mutational signatures for each sample for a group of three prostates[17].

More recently[142], the analysis of 23,829 samples revealed an association between signatures 1, 2, 3, 5, 6, 8, 12, 13, 18, 33, 37, 39, 40, and 41 and prostate cancer (Figure 1.10). Signature 1 and 5 have been associated to ageing and have been found in many cancers but also normal somatic cells[143]. Therefore, they have been referred to as clock-like signatures, as mutations are produced at a continuous rate. Cell proliferation is a critical contributor to signature 1, but little is known about the biological processes driving signature 5[143]. Signature 2 and signature 13 tend to occur in the same samples and its presence is associated with the activity of the AID/APOBEC family of cytidine deaminases. Similarly, signature 6 is related to defective DNA mismatch repair[144]. Signatures 8, 12, 33, 37, 39, 40 and 41 have unknown aetiology.

1.4.3 EPIGENETIC ALTERATIONS

It has been observed that epigenetic alterations patterns are less variable between tumours than other genetic alterations[145]. For example it is widely known that gene *GTSP1* (Glutathione S-Transferase pi 1) shows hypermethylation in more than 90% prostate cancer tissues[146,147]. Similarly, tumour suppressor gene *RASSF1A* and caretaker gene *APC* are also hypermethylated in malignant prostate tissue[148,149] and associated with aggressive prostate cancer[148,150,151]. A group of 8 candidate genes have been proposed as methylation prostate cancer biomarkers (*AOX1, CCDC281, GABRE, GAS6, HAPLN3, KLF8, MOB3B* and *SLC18A2*)[152–154]. In addition, hypomethylation of some markers seems to be another mechanism associated with prostate cancer, especially as a late event, when metastasis is occurring[38]. Epigenetic modifications have been detected even in the neoplastic lesions such as PIN (see section 1.2.7)[150,155].

1.5 FIELD CANCERIZATION

Field cancerization or the field effect, is a phenomenon characterized by the presence of molecular, genetic and epigenetic alterations in morphologically normal tissue that can lead to the development of cancer. These alterations are usually not apparent under histological examination[156]. This idea was first introduced by Slaughter *et al.*[157] after observing the presence of multiple tumours in 11% of patients with oral squamous cell carcinomas. It was proposed that the otherwise morphologically normal tissue comprised a "field" that led to tissue carcinogenesis. These lesions appeared as multiple tumours in different areas of the alimentary tract mucosa, suggesting multifocal disease.

The presence of somatic substitutions, mitochondrial mutations and methylation changes in the preneoplastic field from patients with cancer has been investigated in different types of cancer. It has been observed that these mutations occur in groups of cells in the morphologically normal tissue, indicating the presence of clonal expansions. These clones can harbour mutations in cancer associated genes that are also found in the primary tumour. For example, *TP53* mutations have been observed in the morphologically normal mucosa of HNSCC (head and neck squamous cell carcinoma) patients with cancer[158]. Similarly, Park *et al.*[159] reported abnormal methylation levels in colon cancer associated genes *SFRP2, TFPI2, NDRG4* and *BMP3* both in adjacent and non-adjacent normal tissue of patients. Abnormal molecular

changes are not always shared between tumour and morphologically normal tissues. In liver, multiple clones were supported by somatic substitutions and indels that were also detected in cirrhotic liver but not in hepatocellular carcinoma (HCC) samples. Although no shared driver mutations were found between normal/cirrhotic liver and distant HCC at that stage, it is speculated that the continuous competition of evolving clones in normal liver tissue could eventually result in neoplastic transformation[160]. Clonal proliferation was associated with mutation burden, although it was observed that clonal events do not necessarily lead to cancer development.

RNA sequencing data collected from morphologically normal tissue from patients with a wide range of cancers also confirms the findings from whole genome sequencing: somatic mutations and clonal expansions are a common feature[161]. In some cases, clonal expansions were associated with the presence of cancer or neoplastic lesion. Tissues that had a direct exposure to environmental carcinogenic factors (ultraviolet radiation, smoking and nutritional habits), or had a very high proliferation rate, such as skin, lung and esophagus, had the highest mutation burden[161]. This finding is consistent with previous observations regarding tobacco smoke contribution to a field effect in oral squamous cell carcinoma[156]. However, the association between environmental factors underlying field cancerization is lacking in many types of cancers. Other etiologic factors are considered, such as age, diet, inflammation and hormones. Therefore, the conjunction of heritable genetic and epigenetic factors, exogenous carcinogen exposure and lifestyle factors could determine the propensity to tissue somatic alterations by affecting tissue microenvironment. Because of its potential relevance in cancer initiation and development it is one of the main focus of this thesis.

1.5.1 GENETIC ALTERATIONS IN MORPHOLOGICALLY NORMAL TISSUES

Although the mutational landscape in morphologically normal tissues from donors without cancer has been less extensively studied, there are records that tissue with a histologically morphologically normal appearance can harbour a significant amount of mutations, early clonal expansions, distinct expression profiles and methylation changes that could potentially lead to tumour development in many tissues.

For example, somatic mutations have been detected in the exomes of sun-exposed morphologically normal skin[118] and esophagus[162], where they found 3,760 mutations from 4

donors and 8,919 mutations from 9 donors, respectively. Both studies revealed that a group of positively selected genes were driving clonal expansions. Specifically, genes *NOTCH1, NOTCH2, NOTCH3, FAT1* and *TP53* were found to be mutated both in normal skin and esophagus, all known drivers of esophageal squamous cell carcinomas (ESCC) and cutaneous squamous cell carcinomas (CSCC). Comparable findings have been reported in blood, where the detection of clonal expansions in healthy patients over 65 has been associated with a significant increase in the risk of leukemia[163–166].

A recent study by Colom *et al.*[167] further investigated the clonal evolutionary dynamics in normal tissue of the esophagus in mouse (control) and in esophageal normal tissue treated with a mutagen. Clonal expansions were detected in both tissues and eight positively selected genes were found in mutagen-treated tissue. Mutations in these genes were also present in the control tissue but apart from *NOTCH1*, there was no evidence of selection. In both cases the number of clones decreased with time followed by expansion of existing clones, but the rate of clonal loss and subsequent expansion was much higher in the mutagen-treated tissue due to a stronger clonal competition. Clonal growth was observed to be controlled by the nature of the neighbouring cells: an advantageous clone surrounded by wild type cells will expand rapidly whereas the presence of competing clones nearby diminishes overall growth of all clones. Interestingly, in normal esophageal tissue (control) a higher number of clones was observed, which is explained by the slower rate of clonal loss. These results indicate that competition between specific clones in normal tissues is determinant to preserve homeostasis.

Overall, there is strong evidence that morphologically normal tissues harbour mutations and that in some instances there is strong positive selection of cancer associated genes. This indicates that in many cases normal tissues are comprised of a mixture of evolving clones that eventually could increase the risk for cancer development[168]. However, the mechanisms that lead to these changes and their role in cancer initiation are not understood.

1.5.2 FIELD EFFECT IN PROSTATE CANCER

The possible presence of a field effect in prostate cancer is supported by two main observations: the multifocal nature of up to 80% of prostate cancers and molecular evidence[67]. Multifocality coupled with the occurrence of multiple and heterogeneous genetic alterations of a distinct origin, it is a clear indication of a potential field affecting the prostate as a whole.

In a recent study by Cooper *et al.*[17], mutations and clonal expansions were found at high levels in morphologically normal tissue (under the microscope it looks normal), in many cases distant from the tumour, suggesting evidence for the presence of fields effects and their involvement in prostate cancer evolution. Chromosomal rearrangements were also reported by Shancheng Ren *et al.*[138] in RNA-seq data from morphologically normal tissue adjacent from the tumour. In a study by Risk *et al*.[169] biomarkers were identified to play a role in the cancerization of the prostate. They found an overexpression of cancer-associated genes (*ERG*, *HOXC4, HOXC5* and *FOLH1*) in benign tissue from men with prostate cancer compared to men without prostate cancer. In addition, genes *COX-2* and *PCA3* showed an increase in expression in normal tissue areas that are adjacent to the tumour in patients that suffered recurrence[170].

Abnormal alterations in morphologically normal tissue are also detected at epigenetic level, a finding noted by Chai *et al.*[156]. Several studies have highlighted the importance of analysed the methylation status of key genes in terms of predicting prostate cancer development after an initial negative biopsy. Hypermethylation in genes *APC, GTSP1* and *RASSF1* has been observed in morphologically normal tissue in many studies[155,171–173] and it has proven to be a better predictor of cancer development than histopathological examination alone[171–173]. Hypermethylation in genes *APC* and *GTSP1* was reported in 95% and 43% respectively in patients with an initial negative biopsy that later developed prostate cancer[171]. In a study by Møller *et al.*[174] they explored the possible presence of a field effect of epigenetic nature by assessing a group of 9 genes in morphologically normal tissue from men with and without prostate cancer. These genes had been previously reported as hypermethylated in prostate tumour tissue[146,147,152–154]. They identified a four gene methylation signature (*AOX1, GSTP1, HAPLN3* and *SLC18A2*) that was specific to morphologically normal samples from men with prostate cancer only. This four gene methylation signature had a higher positive predictive value of developing prostate cancer than the PSA test.

The presence of a field effect is of clinical relevance, as the effect of the field needs to be removed, not just the tumour foci, if recurrent cancer is to be prevented[156]. The early detection of a field effect in the prostate could also help to risk stratify disease at diagnosis and requirements for closer surveillance or repeated biopsies[171–173]. It may also help identify new drug therapies or mechanisms that could be targeted by treatment.

1.6 THESIS AIMS

The objective of this thesis is to gain insights about the mechanisms that drive multifocal prostate cancer development in the early stages of the disease. This could lead to the development of better diagnostic tools that would allow to identify aggressive cancers from indolent ones thus avoid unnecessary treatment. We explore the possibility of the presence of a field effect in the morphologically normal tissue of the prostate by carrying two different approaches. First, we analysed whole genome sequencing data from a group of morphologically normal tissue samples from men with and without prostate cancer to determine if the changes observed could be driving the development of multifocal cancer in the prostate. We used different methods to identify the genomic alterations, detect clonal expansions, mutational signatures and driver genes. We report a high number of genetic alterations in normal tissues and an association between mutation burden and clonal expansion presence and prostate cancer.

Secondly, we performed targeted sequencing with high coverage on a 100 gene panel in 96 samples from normal tissue of one prostate. A detailed representation of the mutational landscape of the prostate in these genes is provided. We observed a group of mutated genes in multiple samples, suggesting the presence of clonal patches.

1.7 CHAPTER SUMMARIES

In Chapter 1 we describe the biological knowledge necessary to understand the key aspects of cancer biology, prostate cancer. We also introduce different approaches for the analysis of next generation sequencing technologies required for the characterisation of abnormal genetic alterations.

In Chapter 2 we describe the sequencing and computational methods applied in this thesis.

In Chapter 3 we examine the mutational landscape for morphologically normal tissues men with and without cancer, matched tumours and cell cultured fibroblasts.

In Chapter 4 the results of the subclonal architecture reconstruction of morphologically normal samples (including samples with BPH and cell cultured fibroblasts) and tumour from patients with and without cancer are presented. We describe the evolutionary relationship between the

tumour, normal and BPH samples and observe a clear association between clonal expansions and the presence of cancer.

In Chapter 5 we present the results obtained from the deep targeted sequencing of a panel of 98 genes that are relevant in prostate cancer of 96 morphologically normal and tumour samples from a prostate cancer patient. Detection of substitutions is performed and assessed, in combination with the spatial location of each sample.

In Chapter 6 we summarise our key findings, address their implications and examine different options for further research in the field.

CHAPTER 2 : METHODS

2.1 SUMMARY

In this chapter the laboratory techniques, computational and statistical methods implemented in this thesis are described. A detailed description of all the steps followed from the sample collection to the analysis of the genomic data obtained after sequencing. We outline sample collection and library preparation techniques, NGS technologies, NGS data processing approaches and tools. We then describe machine learning methods that allowed us to detect clonal expansions and mutational signatures, pathway analysis and methods to assess a mutation's functional impact and statistical analyses used to find associations between clinical variables and genomic data.

2.2 PREPARATION OF THE PROSTATE

Here we describe the steps taken to collect samples from frozen tissue for Whole Genome Sequencing (WGS) and from formalin-fixated paraffin-embedded (FFPE) tissue for targeted sequencing as outlined by Warren *et al*[175]. Specimens from prostatectomy and cystoprostatectomy performed at Addenbrooke's hospital in Cambridge were processed as follows: they were weighed, inked (to identify the correct orientation later on) and measured in three dimensions. The prostates were cut transversely into 3 slices: one of them was selected for DNA extraction and subsequent WGS and processed further (see section 2.3) whereas the rest (apex and basal part of the prostate) were processed and eventually formalin-fixed and paraffin-embedded for future analyses such as the patchwork experiment (see section 2.4).

2.3 PROCESSING SELECTED TISSUE FOR WGS

4-6 mm cores were taken from the selected slice, inked at one end (for correct identification of anatomical features and orientation) and frozen. The slice from which the cores were taken was then pinned on a cork to prevent it from shrinking and formalin-fixated as described by Egevad[176] (Figure 2.1). This slice represents a "map" that indicates the location where each core was retrieved from. A section immediately adjacent was hematoxylin and eosin (H&E) stained, allowing the identification of normal and tumour tissue and subsequent selection of

samples. Each core could be mapped to the original formalin-fixated slice using the inked marks that were visible both in the H&E and formalin-fixated slice (Figures 2.1D and 2.2A).

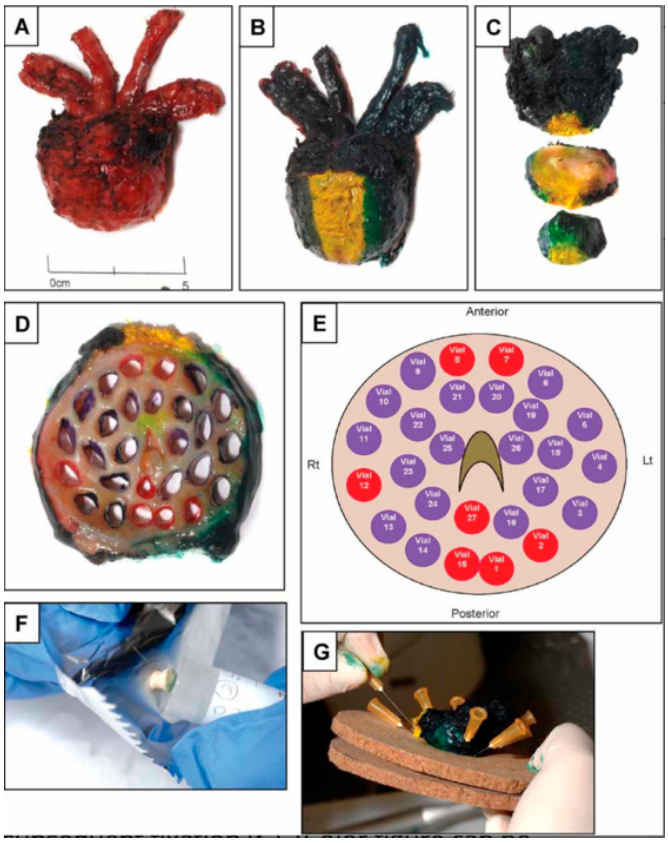


Figure 2.1: Sample tissue sampling from fresh radical prostatectomy specimens. Adapted from Warren *et al.*[175]. **(A)** Prostate was removed prostatectomy with the seminal vesicles intact, **(B)** inked for the correct identification of the anatomical regions, **(C)** cut transversely into 5 mm slices **(D)** and multiple punch biopsies were taken from each slice. **(E)** Each core's location was recorded on a map diagram, and then each core was frozen at 80 C **(F)**. The remaining fresh slices were pinned to a cork and fixated **(G)**.

2.3.1 PROCESSING OF THE TISSUE CORES

A longitudinal section was taken from each frozen core and tumour cellularity was estimated (Figure 2.2B). Then, a transverse single 5 µm section was taken from the frozen cores and H&E stained. After this, six adjacent 6×50 µm sections were cut. This process was repeated multiple times. The six 6×50 µm were used for DNA extraction, whereas the 5 µm was used to assess the presence or absence of cancer in central pathology review by three histopathologists (Figure 2.2 C). All these steps were performed in Cambridge by Anne Warren. Then, I measured the

distance (in mm) (using the H&E stained slice adjacent to the one where the cores were retrieved from, Figure 2.2A) between all the morphologically normal samples and their respective tumours, where present.

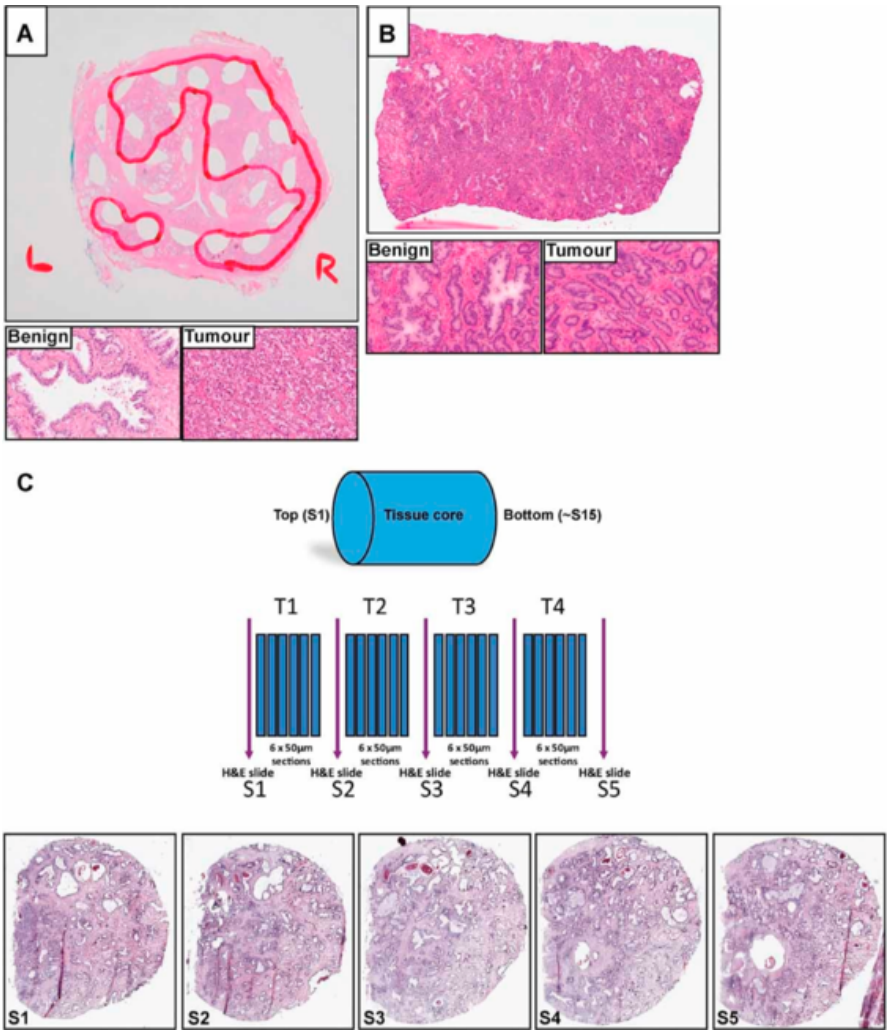


Figure 2.2: Processing frozen tissue from prostate, adapted from Warren *et al.*[175]. **(A)** An adjacent section was taken from the slice where the cores were taken from and H&E stained. Matched normal and tumour samples were selected by comparing the inked holes on the FFPE H&E stained section with the map. **(B)** A frozen section was taken longitudinally from each tissue core and tumour cellularity was estimated for the whole core. **(C)** Single transverse frozen sections of 5 µm were taken (S1-S5) and H&E stained and the presence of tumour was assessed. In between each pair of 5 µm sections, six 6×50 µm sections were taken for DNA extraction (T1-T4).

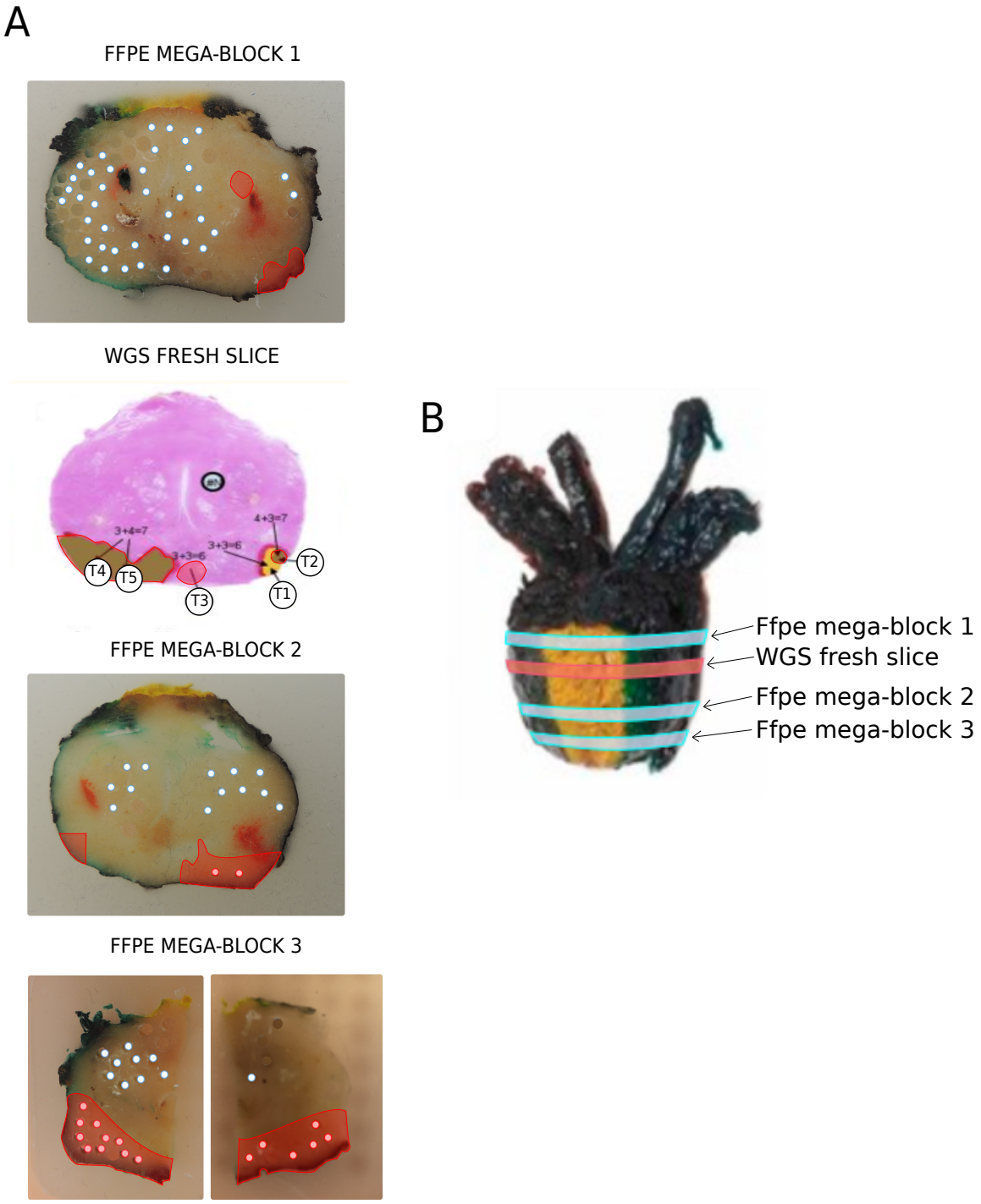


Figure 2.3: (A) FFPE mega-blocks used for the patchwork experiment and fresh slice used for WGS. 1 mm^3 punches were taken: 77 from normal tissue and 18 from tumour (tumour area shown in red). In 15 cases two 1 mm^3 samples were taken from the same punch. **(B)** Mega-block 1 and 2 were taken from above and below the fresh WGS slice, mega-block 3 was from the bottom of the prostate. Fresh slice was adapted from Cooper *et al.*[17]

2.4 SAMPLE COLLECTION FROM FFPE TISSUE FOR TARGETED SEQUENCING

Basal and apical slices described in section 2.2 were also pinned to a cork and formalin fixated. All the formalin fixated parts of the prostate (the apical slice, the slice where the cores were taken from and the basal slice) were cut into 5 mm slices or mega-blocks (Figure 2.3) and stored. Three of these FFPE slices or mega-blocks were selected from one prostate that had been previously processed for whole genome sequencing (WGS), as described in section 2.3. An adjacent section from the slice used for WGS (Figure 2.2A) had previously been taken, H&E stained and reviewed by 3 histopathologists to confirm the presence or absence of cancer, so the FFPE mega-blocks could be mapped to this section and normal and tumour tissue identified. Two of the FFPE blocks were from above and below the frozen tissue slice (section 2.3) and the third was from the bottom of the prostate. A total of 95 punches were taken using a 1 mm^3 punch: 77 punches from normal (39 from mega-block 1, 8 from mega-block 2 and 10 from mega-block 3) and 18 from tumour tissue (2 from mega-block 2 and 16 from mega-block 3) as illustrated in Figure 2.3.

2.5 DETAILED SPECIFICATION FOR OUR SEQUENCING EXPERIMENTS

2.5.1 WGS EXPERIMENT

DNA from whole blood samples and frozen tissue was extracted and quantified using a ds-DNA assay (UK-Quant-iT PicoGreen® dsDNA Assay Kit for DNA) following manufacturer's instructions with a Fluorescence Microplate Reader (Biotek SynergyHT, Biotek). Acceptable DNA had a concentration of at least 50ng/µl in TE (10mM Tris/1mM EDTA), with an OD 260/280 between 1.8-2.0. This was performed by researchers at CRUK CI.

Paired-end whole genome sequencing (WGS) of the samples was performed at Illumina, Inc. (Illumina Sequencing Facility, San Diego, CA USA). 1 ug of DNA was used to generate paired-end libraries following the End Sample Prep Kit (Catalog # PE-102-1002). A Covaris E220 was used for DNA fragmentation. 300 bp inserts were selected manually by agarose gel electrophoresis and PCR amplified for 10 cycles. Final quality of libraries was assessed with the Agilent Bioanalyzer and quantified by qPCR and picogreen fluorimetry.

100 base paired-end reads were sequenced on the Illumina HiSeq2000 using TruSeq Sequencing by synthesis (SBS) chemistry (described in section 1.3.1) v3 to a target depth of 50× for the tumour samples and 30× for morphologically normal and blood samples. Cell cultured fibroblasts were sequenced on the Illumina HiSeq 2000.

2.5.2 TARGETED SEQUENCING EXPERIMENT

DNA from 1 mm^3 punches from FFPE blocks was extracted using allprepDNA/RNA kit from Qiagen and quantified (Qubit dsDNA HS Assay Kit) with a qubit fluorometer following manufacturer's instructions. Total DNA had a total concentration of at least 64 ng in TE (10mM Tris/1mM EDTA), with an OD 260/280 between 1.8-2.0.

At least 10 ng of DNA was used to generate paired-end libraries following the SureSelectXT Low Input Target Enrichment System for Illumina from Agilent (GB9707B). The Sureselect enzymatic fragmentation kit (p/n 5191-4080) was used to generate 150-200 bp DNA fragments. The DNA ends were repaired, dA-Tailed and a molecular barcoded adaptor was ligated to each sample. Libraries were then amplified and purified with AMPure XP beads. Assessment of the pre-capture library was performed with the Agilent Tapestation instrument (p/n G2991AA). At least 500 ng of pre-capture was used for hybridization to a target-specific probe capture library that was designed by Agilent. The 95 target regions were captured using streptavidin-coated magnetic beads. Then libraries were amplified and purified using AMPure XP beads. Assessment of the libraries post-capture was performed using the Agilent Tapestation instrument. Final libraries were pooled with a final concentration of 4 nM. Paired-end targeted sequencing of the samples was performed at the Quadram Institute (Norwich Research Park, UK).100 base paired-end reads were sequenced on the Illumina NextSeq 500 using NextSeq 500/550 High Output Kit v2.5 (300 Cycles) to a target depth of 500× for all samples the blood control. Two runs of sequencing were performed to achieve a higher coverage. Sequencing data was merged for all the computer based downstream analyses.

Target regions included in the probe capture library comprised a panel of 98 genes (Figure C.2, Appendix C): 62 genes commonly involved in prostate cancer[127] , 11 of which have been reported at the very low frequency[128], 15 genes that were also targeted by Martincorena *et al*. in skin[118], known to be involved in a wide range of cancers and 22 genes detected in morphologically normal tissue with WGS that were reported as having potential functional significance (see Chapter 3, section 3.5.4).

2.6 QUALITY CONTROL

Quality control is used to identify low quality sequence reads and other problems that occur in sequencing experiments such as the presence of contaminants (adaptors or foreign DNA). Sequencers generate a quality control report by default, but it only addresses problems related to the sequencing process. They also produce a quality score per base of sequence. Other checks can be performed after sequencing and necessary to evaluate other metrics. The tool FASTQC was used for this purpose (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), which is the most frequently used to assess the quality of FASTQ files. For both the WGS and the targeted sequencing experiment we looked at the quality values across all bases and sequences, the proportion of each base in every sequence, the GC content present in each sequence, undetermined base calls (designated as N), sequence length distribution, mismatched pairs, duplication levels and overrepresented sequences (such as adapters and other contaminants).

2.7 ALIGNMENT

In order to identify variants from our sequencing data, the reads need to be mapped to the human reference genome. Here we used the Burrows-Wheeler alignment tool (BWA)[177,178]. This algorithm uses Burrows-Wheeler Transform (BWT), which allows gapped alignment and mismatches, informs about mapping quality and outputs multiple possible alignments. It can be used for both single-end and paired-end reads. The output files are FASTQ format and the output is a SAM (Sequence Alignment Map) file format, a text-based format first introduced by Heng Li *et al.*[179] that is used to store sequences aligned to a reference genome.

2.7.1 BURROWS-WHEELER TRANSFORM

We have an alphabet Σ where symbol $ is not part of Σ and always appear at the end of a string $X = a_0 a_1 \dots a_{n-1}$ so that $a_{n-1} = \$$. If we let $X[i] = a_i, i = 0,1, \dots, n-1$ be the *i*th symbol of $X, X[i,j] = a_i \dots a_j$ a substring and $X_i = X[i, n-1]$ a suffix of *X*. The suffix *S* of *X* is comprised of a permutation of integers 0, …, *n -1* and so *S(i)* is the start position of the smallest suffix *i*th. This permutation produces *n* strings that are lexicographically sorted. Therefore, the BWT of *X* can be expressed as *B*[i] =$ when *S*[i]=0 and *B*[i]=*X*[*S*(i)-1] when *S*[i]>0. An example of how the BWT is constructed is illustrated in Figure 2.4.

Let *W* be a substring of *X*, the position of each appearance of *W* in *X* appears in an interval in the suffix array. All the suffixes that have the prefix *W* are sorted together, and so we can define the intervals as:

$$\underline{R}(W) = \min\{k: W\ is\ the\ prefix\ of\ X_{S(k)}\}$$
$$\overline{R}(W) = \max\{k: W\ is\ the\ prefix\ of\ X_{S(k)}\}$$

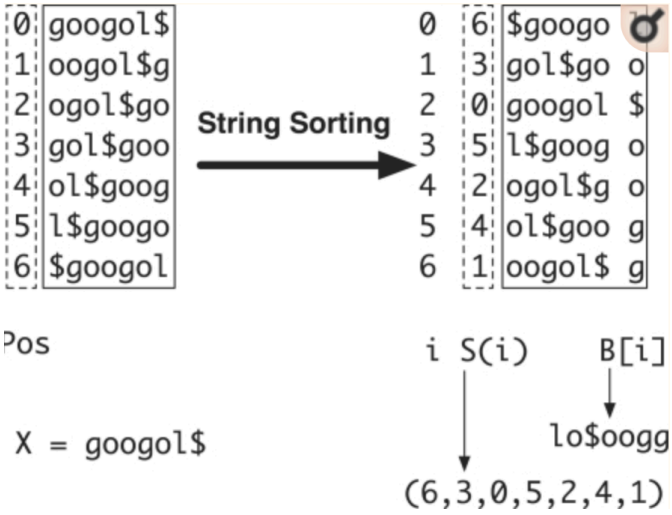


Figure 2.4: Constructing suffix array and BWT string for *X*= googol$. String X is circulated to generate seven strings, which are then lexicographically sorted. After sorting, the positions of the first symbols form the suffix array (6, 3, 0, 5, 2, 4, 1) and the concatenation of the last symbols of the circulated strings gives the BWT string lo$oogg. Adapted from Heng Li *et al.*[180]

The interval $[\underline{R}(W), \overline{R}(W)]$ is called the "SA interval" of *W*. In figure *X*, the string "go" has an "SA interval" [1, 2] and suffix values [3, 0]. The "SA intervals" of a substring *X* inform about the positions of each component of the string. In order to align the short sequences to a reference genome the algorithm searches for the "SA intervals" of substrings *X* that match the query (reference genome). As there are *n* permutations there may be many different matches, some more exact than others. To achieve the most exact match, the intervals $\underline{R}$ and $\overline{R}$ are calculated iteratively from the end of *W* in a process called "backward search"[181].

2.8 POST-ALIGNMENT PROCESSING

After alignment, the SAM files can be converted to a BAM (Binary Alignment Map) format. The BAM format is a compressed binary version of a SAM file and stores the same information. At this point, the information of the alignment is stored randomly (the order of appearance in the FASTQ files) in the BAM file. In order to access the data easily, the BAM file needs to be sorted by alignment coordinates (ordered by chromosome and position) and indexed. Indexing generates a text file containing all the ordered coordinates of our BAM file that allows to extract information of interest from it quickly. Before performing variant calling we need to assess a few aspects: quality of the alignment and duplication levels. This analysis is commonly performed with Picard tools (http://broadinstitute.github.io/picard/), a command line tool specifically designed for the analysis of Illumina NGS data. It is used to produce a wide range of metrics that provide important information (depending on the type of NGS performed) about the quality of a BAM file and quality of the alignment. This analysis provides information about several features such as the total number of reads that aligned to the reference genome, the number of high quality aligned reads, the number of mismatches, the number of reads aligned in pairs (both reverse and forward strand were aligned), among others. Other aspects that can be assessed with Picard tools are duplication levels and coverage metrics. In this thesis we analysed the alignment quality, duplication levels and coverage.

2.8.1 DUPLICATION LEVELS

Duplication is a very important aspect that can affect the analysis of a sequencing experiment. If high duplication levels are observed, two or more reads that originate from the same fragment of DNA can be observed. Duplicates can be classified as PCR duplicates or optical duplicates. PCR duplicates are produced when there is enrichment bias caused by PCR overamplification for some regions of the library. The may not be exactly identical but have a very high sequence similarity. Optical duplicates are observed when during sequencing a single cluster of reads is misread as two different clusters and the sequencing platform reports two different read calls when there is only one. This type of duplicates is completely identical.

2.9 VARIANT CALLING

2.9.1 CA VEMAN

CaVEMan detects single nucleotide variants (SNVs) by calculating a probability score (parameter θ) for likely phenotypes at each genomic position, given prior information regarding reference alleles, copy number or ploidy, the fraction of aberrant tumour cells that are present in each cancer sample and sequencing quality scores. This is done by applying an EM (expectation-maximization) algorithm. CaVEMan achieves a high level of specificity and sensitivity, by applying post-processing filters[86,87]. Some filters include mutation probability threshold, read depth and comparison of the somatic variant to a matched normal (a true variant should not be present in the normal sample). More complex filters can be tailored to specific projects, which further increases reliability of the algorithm. They can be classified in three different categories: filters dependent on intrinsic thresholds of sequencing, filters for elimination of systematic sequencing artefacts produced by next-generation sequencing and filters for specific genomic features that contribute to mapping errors. For example, filters included in the Cancer Genome Project Wellcome Trust Sanger Institute pipeline are: DTH (less than one third of mutant alleles have a base quality equal or higher than 25), RP (coverage is less than 8), MN (more than 0.05 of mutant alleles with a base quality equal or higher than 15 are found in a matched normal), UM (at least 0.05 of mutant alleles with a base quality equal or higher than 15 are found in at least 2 unmatched normal samples), PT (mutant alleles are all on one direction of the read and in the second half of the read, MQ (mean mapping quality of the mutant allele reads was below 21), SR (position falls within a simple repeat), CR (position falls withing a centromeric repeat), PH (mutant reads are on one strand and mean mutant base quality was below 21), HSD (position falls within a high sequencing depth region), GI (position falls within a germline indel), VUM (position has 3 or more mutant allele present in at least 1 % of unmatched normal samples), SE (coverage is equal or higher than 10 on each strand but mutant allele is only present on one strand) and MNP (The proportion between tumour sample mutant allele proportion and normal sample allele proportion is below 0.2).

2.9.1.1 EM ALGORITHM TO CALCULATE PARAMETER θ

The EM algorithm is an iterative approach that calculates maximum-likelihood estimates for model parameters[182]. It is especially useful when the data is incomplete or has latent variables (variables that are hidden). In a scenario with complete data, to apply EM we need some observed data y with a parametric density $p(y|\theta)$ and a description of the desired complete data

X. It is assumed that the data can be modelled as a continuous random vector *X* with density $p(x|\theta)$, where $\theta \in \Omega$ for some set Ω. In summary, we observe the realization *y* of the random vector *Y* that depends on *X*, but *X* cannot be observed directly.

The objective is to make a guess about the complete data *X* and estimate the parameter θ that maximizes the *log-likelihood* of *X*.

$$\theta_{MLE} = \underset{\theta \in \Omega}{\operatorname{argmax}}\, logp(y|\theta).$$

When the data is incomplete, the data *X* is the observed data *Y* + missing data (latent) data *Z*, such that *X = (Y, Z)*.

There are two main steps: the estimation step and the maximization step.

1) We use the conditional probability distribution $p\left(y, z\middle|y, \theta^{(m)}\right)$ in order to calculate the *conditional expected log-likelihood*, also called the *Q*-function. The parameter θ is initialised with *m* =0. The conditional expectation is computed using the $\theta^{(m)}$ from the previous iteration.

$$Q\left(\theta\middle|\theta^{(m)}\right) = \int_X logp(y, z|\theta)p(y, z|y, \theta^{(m)})dx = \int_Z logp(y, z|\theta)p(z|y, \theta^{(m)})dz$$
$$= E_{Z|y,\theta^{(m)}}[logp(y, Z|\theta)].$$

2) A new estimate of the parameter θ that maximizes the *Q*-function is found. The new estimate is $\theta^{(m+1)}$.

$$\theta^{(m+1)} = \underset{\theta \in \Omega}{\operatorname{argmax}}\, Q(\theta|\theta^{(m)}).$$

These two steps are performed until the estimate converges to a stable value after many iterations.

2.9.2 DeepSNV

The deepSNV algorithm was first developed by Gerstung *et al.*[92] is used to compare a sample of interest (test) to a control. For each genomic position, the number of observed nucleotide counts from both strands in the test and the control is modelled with a hierarchical binomial

model. For each base a likelihood ratio test is obtained to address the problem of the expected variation between the two types of samples (test vs control), and account for the increased number of SNVs expected in the test in comparison to the control sample.

2.9.3 PINDEL

Pindel[96] is a popular method used to identify insertions and deletions that has been used in many studies[16,43]. It implements a pattern growth algorithm to identify break points of deletions and medium sized insertions and their respective fragments in comparison to the reference genome from paired–end reads[96]. With the given output Pindel maps all the reads to the reference genome using SSAHA2[183] (Sequence Search and Alignment by Hashing Algorithm), an algorithm that performs searches on big DNA databases. Then, it selects those paired reads that are only mapped in one end and whose other end cannot be mapped anywhere in the genome. The mapped end must have no mismatched bases. The mapped end is used to find the direction of the unmapped read by comparing to the reference genome and determines if it is a deletion or an insertion. Some aspects that lead to missing variants are the presence of repeats in the reference genome, as it would prevent unique mapping. Also, only perfect matching is considered, which means that SNPs or sequencing errors in the regions of indels may be the cause of missing true positives[96]. In order to remove false positive variants, multiple filters are usually applied. For example, filters commonly included by the Cancer Genome Project Wellcome Trust Sanger Institute pipeline are: F004 (medium read depth strand bias check), F005: high read depth strand bias check, F006 (small call excessive repeat check), F010 (variant must not exist within the unmatched normal panel), F012 (germline check), F018 (sufficient depth), F015 (no normal calls), F016 (verify indel condition) and F017 (variant must not overlap with a simple repeat).

2.9.4 BRASS

Brass (Breakpoints via assembly) is an algorithm developed at the Wellcome Trust Sanger institute to find genomic rearrangements in paired-end NGS sequencing data. It consists of two phases. In Brass phase I, discordant reads are detected and used to find regions of interest in the genome (it reports read mapping orientations). These regions are only considered if there are enough reads supporting them, appear in a difficult region in the genome or if they were found in the matched normal sample. In Brass phase II, a de novo assembly using Velvet[184] is performed on the reads mapping around the breakpoint windows determined in Brass phase I.

If this produces the expected pattern then the end result is a mapping of the breakpoint to base pair resolution.

2.9.5 BATTENBERG ALGORITHM

The Battenberg algorithm[103] is used to detect copy number alterations in sequencing data. Copy number alterations produce allelic imbalances. To detect them, the B- allele frequency (BAF) of germline SNPs (and therefore assign the same BAF for positions of interest surrounding the SNP) is calculated as the fraction of the total number of reads from allele B ($r_{B,i}$) and the sum of the total number of reads from alleles A ($r_{A,i}$) and B ($r_{B,i}$). The alleles of all SNPs are paired according to whether they come from the mother or the father, a process that is known by haplotype phasing[103,185]. This technique results in BAF values that are more accurately and one can proceed to classify CNAs as clonal or subclonal (section 2.9.5.1). Here we will express the number of reads from alleles A and B as a function of the number of chromosome copies (integer value) for each allele ($r_{A,i} = n_A$ and $r_{B,i} = n_B$):

$$BAF_i = \frac{n_{B,i}}{n_{A,i} + n_{B,i}}, \quad Eq.1$$

The BAF of a germline heterozygous SNP should fall around 0.5 if there are no allelic imbalances, so lower or higher values indicate the presence of CNAs (loss or gain, respectively). To obtain accurate detection of CNAs it is important to first estimate the fraction of tumour cells present in the sample, also referred to as tumour purity (ρ). Thus, we account for the number of chromosome copies in tumour cells ($n_{A,t}$ and A and $n_{B,t}$) and the number of normal chromosome copies in normal cells ($n_{A,n}$ and $n_{B,n}$) in the sample:

$$BAF_i = \frac{\rho n_{B,t} + (1-\rho) n_{B,n}}{\rho\left(n_{A,t} + n_{B,t}\right) + (1-\rho)(n_{A,n} + n_{B,n})}, \qquad Eq.2$$

BAF values from samples with variable tumour purity are illustrated in Figure 2.5.

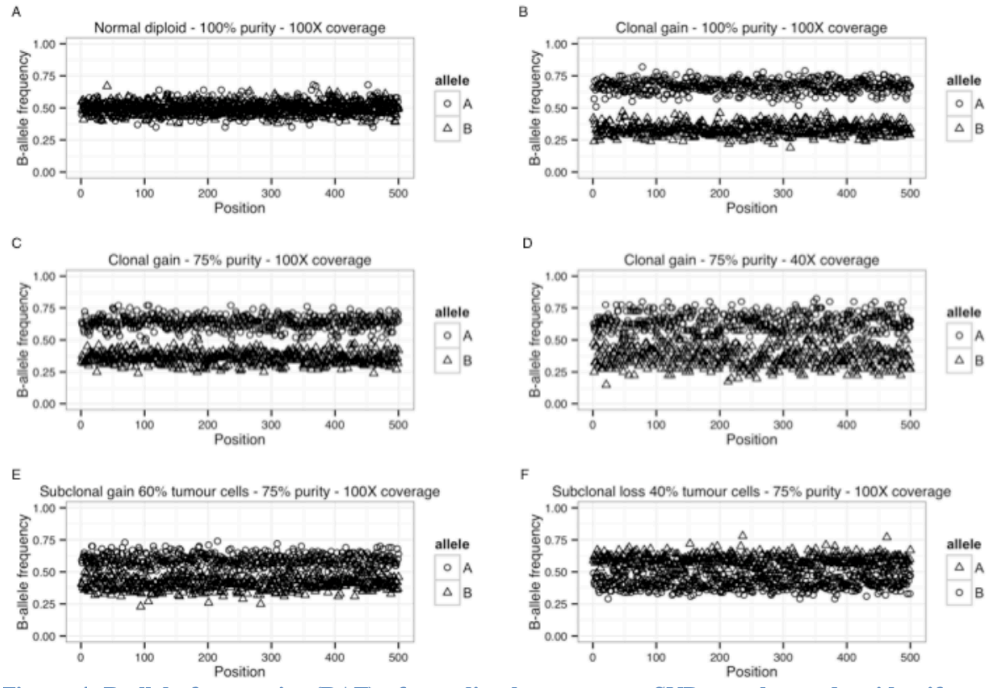


Figure 2.5: B-allele frequencies (BAF) of germline heterozygous SNPs can be used to identify copy number aberrations. A-F show that the BAF is noisy, and that it gets increasingly more difficult to separate the bands as the purity or coverage goes down and when the aberration is subclonal. To reduce the noise, SNPs can be phased to determine which allele is the B-allele. By combining the SNPs over longer stretches of DNA it becomes possible to detect subclonal aberrations. Adapted from Dentro *et al.*[188]

Case A shows a region from a tumour sample with tumour purity of 100% with no CNAs: the values fall around 0.5 and form a single thick band in the middle. In contrast, case B shows a clonal gain that is represented by two separate bands: allele A presents more reads that allele B. Cases C and D demonstrate the importance of sequencing depth and tumour purity on CNAs estimation. In case C, a lower purity (75 %) produces less defined bands. The effect is even more pronounced in case D where a lower coverage (40x) increases the noise and the bands from both alleles are mixed. Finally, cases E and F depict the difficulties in detecting subclonal CNAs. As the changes are occurring only in a percentage of the total cells and there can be more than one subclone, the BAF values are not constant, producing poorly defined bands. In this case, read pairs corresponding to both variant alleles and read pairs corresponding to the reference allele and the variant allele are found simultaneously. This scenario can only occur

if each read pair come from different cell populations (subclones). For this reason, it is much easier to estimate CNAs that are clonal. The challenge of BAF variation due to low coverage has been addressed by performing *haplotype phasing*[103,185]. This technique involves taking CNA information from SNPs in the same region and pairing the alleles of all SNPs according to whether they come from the mother or the father. After phasing, BAF values are more accurately and one can proceed to classify CNAs as clonal or subclonal (section 2.9.5.1).

2.9.5.1 DETECTING SUBCLONAL COPY NUMBER ALTERATIONS

In order to detect subclonal CNAs, we assume that there is clonal change when the BAF deviates from the expected value of 0.5 is made. Taking the copy number values for each allele n_A and n_B for the estimated clonal and subclonal CNAs (integer and non-integer values, respectively), we can calculate the allele frequency $\hat{h}_f$ under the assumption that copy number values are clonal. In order to do this, we have to round up or down the subclonal values (a non-integer value) to the closest integer value above and below the real value. Thus, there are four possible scenarios:

Both A and B alleles are rounded down:

$$\hat{h}_f = \frac{\rho \lfloor n_B \rfloor + 1 - \rho}{\rho(\lfloor n_A \rfloor + \lfloor n_B \rfloor + (1-\rho)2}, \qquad Eq.3$$

Both alleles A and B are rounded up:

$$\hat{h}_f = \frac{\rho \lceil n_B \rceil + 1 - \rho}{\rho(\lceil n_A \rceil + \lceil n_B \rceil + (1-\rho)2}, \qquad Eq.4$$

The A allele is rounded up and the B allele is rounded down:

$$\hat{h}_f = \frac{\rho \lfloor n_B \rfloor + 1 - \rho}{\rho(\lceil n_A \rceil + \lfloor n_B \rfloor + (1-\rho)2}, \qquad Eq.5$$

The A allele is rounded down and the B allele is rounded up:

$$\hat{h}_f = \frac{\rho \lceil n_B \rceil + 1 - \rho}{\rho(\lfloor n_A \rfloor + \lceil n_B \rceil) + (1-\rho)2}, \qquad Eq.6$$

Now we can compare the observed allele frequency h_f to the calculated clonal $\hat{h}_f$ values from all the four scenarios. If the observed h_f is significantly different from $\hat{h}_f$ in all four cases a subclonal CNA is identified.

2.10 MUTATIONAL SIGNATURES DETECTION METHODS

2.10.1 NON-NEGATIVE MATRIX FACTORIZATION

This method is a machine learning technique first introduced by Pateero *et al.*[186]that has been adapted to perform *de novo* signature extraction. Computational frameworks such as SigProfiler[187,188] and SignatureAnalizer[refs] are based on the non-negative matrix factorization method. This approach allows us to find latent features that cannot be directly observed, but the result of these features combined. A key characteristic of this approach is that all the values have to be positive. For a matrix A $(U \times V)$ with nonnegative elements, matrices W $(U \times K)$ and H $(K \times V)$ have to be found so that A is the product of matrices W and H:

$$A = WH$$

$$a_i = [w_{i1} w_{i2} \ldots w_{ik}] \times \begin{bmatrix} h_1 \\ \ldots \\ h_k \end{bmatrix} = \sum_{j=1}^{k} w_{ij} \times h_i$$

In order to achieve this, the dimensions of factor matrices W and H have to be lower than those of the original matrix A. Each row in H is a component and each row in W is a weight of that component.

In the context of signature extraction, we have to consider 96 possible scenarios of mutation types K. These 96 scenarios refer to the 96 possible mutated trinucleotides. Mutational signatures are derived from specific patterns of these 96 trinucleotides, therefore, a mutational signature is described as a discrete probability density function over the domain of mutation types K. Thus, a mutational process's signature P_l is described as a combination of these probabilities and expressed as nonnegative K-tuple $P_1 = [p_1^1, p_1^2, \ldots p_1^k]^T$, where $\sum_{k=1}^{K} p_1^k = 1$

and p_1^k is the probability of the mutation process P_1 to originate a mutation type K of all K types. Therefore, a set of mutational signatures can be represented as a nonnegative mutational signature matrix:

$$P\ (K \times N) = \begin{pmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{pmatrix},$$

where K is the number of mutation types (96 possible trinucleotides) and N is the number of signatures (needs to be determined). Each mutational process is supported by a specific number of mutations in the genome, which can be classified as the exposure of that genome to that mutational process.

The exposure of G genomes to a set of processes N can be expressed as:

$$E\ (N \times G) = \begin{pmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{pmatrix},$$

where N represents the signatures and G is the number of genomes.

Finally, the mutational catalogue of a cancer genome could be represented as:

$$M\ (K \times G) = \begin{pmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{pmatrix},$$

where K represents the number of mutation types (all trinucleotide combinations) and G the number of genomes.

With this NMF model, mutational signatures of a cancer genome can be seen as a superposition of signatures present in that genome and their exposures. For a set of G genomes and N mutational signatures, this can be expressed as:

$$\mathrm{M} \approx \mathrm{P} \times \mathrm{E}$$

However, there are limitations to this method: the number of signatures that can be detected is dependent on the number of genomes available and the number of mutations per sample. When a dataset does not meet those requirements, there is a possibility of refitting previously validated signatures (extracted using NMF or other methods) from large cohorts of cancer genomes to individual samples. Some approaches used for this purpose have been developed such as Non-linear Programming Methods and the SigProfilerSingleSample tool (an updated version of SigProfiler) (described in Methods, section 2.10.2 and section 2.10.3, respectively).

2.10.2 NON-LINEAR PROGRAMMING METHODS

Non-linear programming is an approach used to solve an optimization problem. A set of *n* decision variables $x_1, x_2, \dots, x_n$ are selected in order to optimize (maximize or minimize) an objective function. This function is subject to constraints (called equalities or inequalities), that determine what is called the feasible region the decision variables can be selected from in order to optimize the objective function. A general non-linear program could be described as:

$$\begin{aligned} &Maximize\ f(x_1, x_2, \dots, x_n) \\ &subject\ to:\ g_1(x_1, x_2, \dots, x_n) \leq b_1, \\ &\qquad g_m(x_1, x_2, \dots, x_n) \leq b_m \,„ \qquad Eq.1 \end{aligned}$$

One type of non-linear programming method is Quadratic Programming (QP). Here the objective function is a quadratic function of several decision variables dependent on linear constraints. In a Quadratic Programming problem, we have a *n*-dimensional vector *m*, an $n \times n$ dimensional symmetric positive matrix Q here referred to as $S^T S$ (the superscript T implies transposition), an $m \times n$ dimensional real matrix *A* and an *m*-dimensional real vector *w*.

Applied to signature refitting, the dimensional vector of real values *m* corresponds to the normalized observed vector of mutation scenarios (96 x 1); *S* can be defined as a 96 x *k* matrix (reference signatures), where each column represents a contribution of mutational scenarios to one signature and *k* is the number of known mutational signatures; *w* is a matrix of weights *k*

x 1 to be estimated. The objective is to estimate a n-dimensional vector *w* that will minimize the difference between observed vector *m* and *S.*
The quadratic programming problem can be expressed as follows:

$$minimize\ (m - Sw)^T\ (m - Sw) = m^T m - w^T S^T m - m^T Sw + w^T S^T Sw$$

$$subject\ to\ \sum_j w_j = 1, w_j \geq 0\ ,\ \ Eq.2$$

which is the same as:

$$minimize - m^T Sw + \frac{1}{2} w^T S^T Sw$$

$$subject\ to\ \sum_j w_j = 1,\ \ w_j \geq 0, \qquad Eq.3$$

This problem can be solved using the dual method of Goldfarb and Idnani[189].

2.10.3 SIGNATURE REFITTING USING SIGPROFILER

The updated version of the tool for mutational signature analysis SigProfiler[187,188] SigProfilerSingleSample (https://github.com/AlexandrovLab/SigProfilerSingleSample) uses previously extracted signatures and estimates the number of somatic mutations that are associated with a known set of mutational signatures in a single sample.

The contribution of each signature is estimated using non-linear programming method described by Byrd[190]. However, the tool applies a set of rules that limit the number of signatures that can be attributed to a sample. These rules ensure that not all reference signatures are used for the mutational profile reconstruction by incorporating previous biological knowledge. The aim of this measure is to control for inaccurate signature attributions with a low biological plausibility.

2.11 CLUSTERING METHODS

Clustering methods are used to find structure and groups of similar objects in the data. They reveal useful relationships between objects. The kind of clustering techniques described in this section are classified as unsupervised, which means that they perform the division into groups without any external information or labels of the objects. In this thesis the agglomerative hierarchical clustering and Bayesian clustering analysis were applied to detect related groups of mutational signatures and reconstruct the clonal architecture, respectively.

In bayesian statistics, a prior distribution $p(\theta)$ is assigned to all unknown parameters in the model and a posterior distribution $p(\theta|y)$ for all parameters is inferred given the observed data (Eq.1). The prior distribution represents subjective beliefs about the parameters, and $p(y|\theta)$ is defined as the likelihood function or the probability of obtaining the data given the parameters. According to Bayes's theorem:

$$p(\theta|y) \sim p(\theta) \times p(y|\theta), \qquad Eq.1$$

Here we will focus on nonparametric bayesian methods, where there is at least one infinite-dimensional parameter. This feature allows the creation of complex models, where the model's complexity grows as more data is observed. The number of clusters do not have to be specified in advance, and it is inferred by the observation of the data. A very popular prior distribution in bayesian nonparametric inference is the Dirichlet Process[191] (see section 2.11.1)

2.11.1 DETECTION OF SUBCLONAL POPULATIONS USING A BAYESIAN DIRICHLET PROCESS

In this section we describe the methods used to reconstruct the clonal architecture of tumour and normal samples from NGS data for each patient. The clustering method used to detect the subclonal populations was developed by Dentro *et al*.[192] and has been applied in a vast number of studies[17,103,193,194]. It aims to determine the number of subclones (clusters) within a tumour, the fraction of cancer cells in each cluster and the number of mutations that belongs to each cluster, which are all unknown parameters. Subsequently, the subclonal architecture for each patient can be illustrated as a phylogenetic tree (see section 2.11.1.2).

Reconstruction of the clonal architecture relies on clustering together mutations (in this case, SNVs) with similar cellular cell fraction (CCF) values, under the assumption that each mutation has occurred only once in the lifetime of a tumour. As stated in section 2.11.1.1, CCFs depend on CNAs and VAF values. VAF is a measure of the proportion of reads with the variant, the sequencing depth determines how accurate this can be obtained. Sequencing depth is a changing measure in any sequencing experiment. Therefore, CCF can vary greatly due to changes in sequencing depth. In order to account for this diversity, an error model is used:

$$r_i \sim Bin(r_{tot,i}, p_i),$$

Where the number of variant reads can be compared to the number of successes of *N* independent coin tosses, being *N* the total read depth. The number of successes (variant reads) are modelled through a binomial distribution, r_i being the number of variant reads at location *i,* $r_{tot,i}$ the total depth at location *i,* and p_i the probability of observing a mutant read. The probability p_i is defined by the proportion of expected reads if the mutation is fully clonal (ζ_i) and the true fraction of cells carrying the mutation (π_i). Therefore,

$p_i = \zeta_i \pi_i$.

As we will discuss later, ζ_i can be obtained from the tumour purity and copy number status of the location. The estimation of π_i is performed using a Bayesian Dirichlet process in *n* dimensions, with *n* representing the number of samples. The dimensions represent the number of related samples for each patient (normal, tumour and BPH). A Dirichlet process is a stochastic process used in Bayesian nonparametric models that has two parameters: a base probability distribution (P_0) and dispersion parameter α. It can be seen as a distribution over distributions, as each draw from a Dirichlet process is a distribution in itself sampled from another distribution (the base probability distribution or P_0). The sample space of P_0 is possibly infinite, so each draw is performed from an unknown number of distributions. With this method we can co-estimate all the unknown parameters: the number of subclones (clusters) or contributing distributions within a tumour, the fraction of cancer cells in each cluster and the number of mutations that belongs to each cluster.

The stick-breaking representation of the Dirichlet process[195] is often used for this purpose:

$P = \sum_{h=1}^{\infty} w_h \pi_{\theta_h}, with\ \pi_h \sim P_0$.

π_{θ_h} represents a location in the CCF space and ω_h is the probability weight probability weight of the *h*th mutation cluster. This probability can be defined as:

$\omega_h = V_h \prod_{l>h}(1 - V_l),\ with\ V_h \sim \beta(1, \alpha).$

V_h are parts of a unit length stick that are sequentially broken off from the remaining stick (as depicted in Figure 2.6). Each iteration produces a smaller V_h, as the remaining stick decreases each time a new part is broken off. Each partition represents a fraction of the total cluster weight (number of SNVs) and a CCF is assigned by resampling using the SNVs allocated to that cluster. Using the associated weight and stick location of a SNV, the probability that a SNV is generated by that substick is calculated for each substick and SNV. This process is performed repeatedly until all SNVs are assigned to a cluster. Therefore, clusters (cellular populations), are estimated, determined by the accumulation of the weight and SNVs in specific locations.

2.11.1.1 CCFs ESTIMATION

The CCF is estimated from the VAF and the number of copies of alleles in that position. The VAF is the percentage of sequence variant reads (r_{mut}) divided by the total amount of reads (r_{ref}) at that position. Therefore, the VAF or f_i of a SNV *i* can be calculated as follows:

$$f_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}, \qquad Eq.1$$

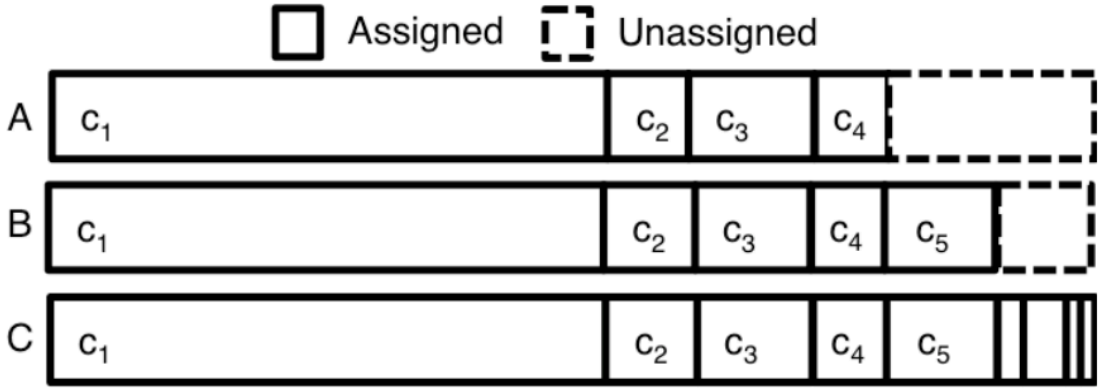


Figure 2.6: Stick-breaking schematic. The stick-breaking property of the Dirichlet process (DP)is used to estimate the number of mutation clusters in the data. For each mutation, a stick of arbitrary length is broken into randomly sized bits that represent a cluster. At point A, breaks have been introduced, corresponding to clusters $c_1 - c_4$. B shows the stick after introducing break 5, whereas C shows the completed stick-breaking procedure. The size of each broken part represents the weight associated with a cluster and influences the mutation assignments, in which a high weight makes it more likely that a mutation is assigned to that cluster. These weights are updated after probabilities for each cluster have been obtained for each mutation, eventually converging on a solution. Adapted from Dentro *et al.*[192].

Allele frequencies are affected by CNAs, if there are any. If there is a subclonal loss or gain, the VAF can increase or decrease depending on which allele (variant or reference) has been affected by the CNA. Because of this reason, it is important to have the CNAs for all locations. An example is illustrated in Figure 2.7.

Therefore, for each SNV there is a copy number state, also called multiplicity or m_i for a given mutation *i*. In order to have an accurate estimation of the number of cells containing a mutation, we must consider the following:

$$u_i = CCF_i m_i, \qquad Eq.2$$

As stated earlier, a clonal SNV is present in 100% of cells so it has a CCF of 1.0. Therefore, the number of chromosome copies or multiplicity m_i has to be an integer and $u_i \geq 1$. A subclonal mutation, however, is present in less than 100% of cells (it has a CCF less than 1.0). The multiplicity value m_i=1, as it is only carried by one chromosome copy (if there are not any CNAs). In this case $u_i < 1$.

In summary, to calculate the value u_i for a given location *i,* we have to consider the fraction of tumour cells of our sample (ρ), the number of chromosome copies in those cells for that specific locus ($\rho n_{tot,n,i}$), and the fraction of normal cells ($1-\rho$) with the number of chromosome copies in normal cells ($n_{tot,n,i}$) at locus *i*:

$$u_i = f_i \frac{1}{\rho}\left[\rho n_{tot,t,i} + (1-\rho) n_{tot,n,i}\right], \qquad Eq.3$$

Normal cells are considered to have a diploid copy number state ($n_{tot,n,i}$= 2), whereas tumour cells copy number state $n_{tot,t,i}$ can be calculated by copy number analysis (discussed in section 2.9.5). For SNVs 3 and 4 in figure 6, the values for u_i would be:

$$\frac{4}{4+6} x \frac{1}{0.8} x[0.8x2 + 0.2x2] = 1.000$$

$$\frac{11}{11+9} x \frac{1}{0.8} x[0.8x3 + 0.2x2] = 1.925$$

2.11.1.2 PHYLOGENY RECONSTRUCTION

Phylogenetic trees that illustrate the evolutionary relationships between subclones can be constructed using the CCF information for each clone/subclone and by applying the *pigeonhole principle*. This principle specifies that if we have m containers or *pigeonholes* and n items or *pigeons* to store in the containers and $n > m$ there has to be a container with more than one item. Following this principle in subclonal reconstruction, we assume that the sum of the CCF of a subclone or subclones has to be smaller than the CCF of their ancestor. Phylogenies can be linear (each cluster is a descendant from an older cluster) or parallel. Figure 2.8 illustrates these two scenarios. In case A we have a linear phylogeny with CCFs of 100%, 80% and 40% for clusters 1, 2, and 3, respectively. As 100% cluster 1 + 80% cluster 2 >100%, cluster 2 with a CCF of 80% has to be a descendant of cluster 1. The same applies to cluster 3: cluster 2 80% + cluster 3 40% >100%, so cluster 3 is a descendant of cluster 2. In the parallel phylogeny, however, cluster 2 has a CFF of 50%. As cluster 2 50% + cluster 3 40% < 100, cluster 3 could be a descendant of cluster 2 or cluster 1.

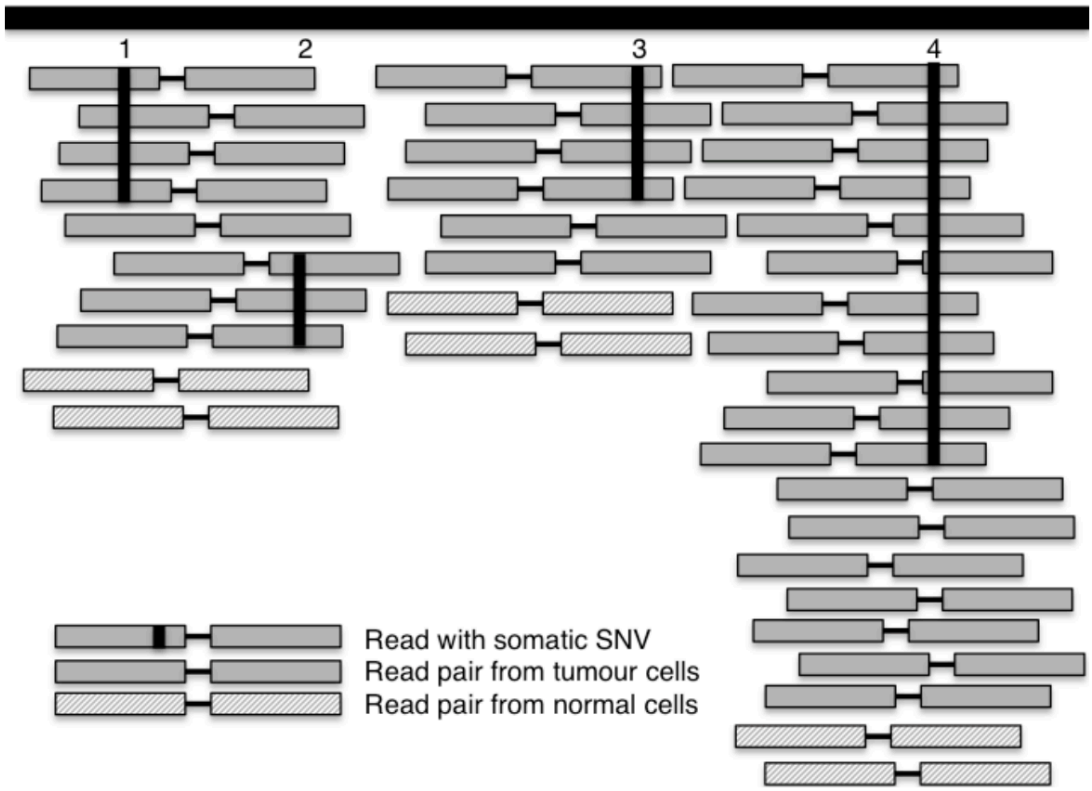


Figure 2.7: Copy number alterations affect variant allele frequencies. Allele frequencies of single nucleotide variants (SNVs) must be transformed to cancer cell fractions (CCFs), accounting for copy nuber changes, before they can be clustered to identify subclonal populations. This illustration shows four SNVs in different (sub)clonal populations and in regions with different copy number states, to illustrate this principle. SNVs 1 and 2 are clonal and subclonal respectively and appear in a nonaberrated copy number state. SNV 3 coincides with a subclonal deletion, with the SNV falling on the retained allele (i.e., the other allele is subclonally deleted). SNV4 has occurred before a gain and is therefore carried by two chromosome copies. Even though SNV 1, 3, and 4 are clonal, their allele frequencies differ because of copy number alterations (CNAs). Adapted from Dentro *et al*.[192]

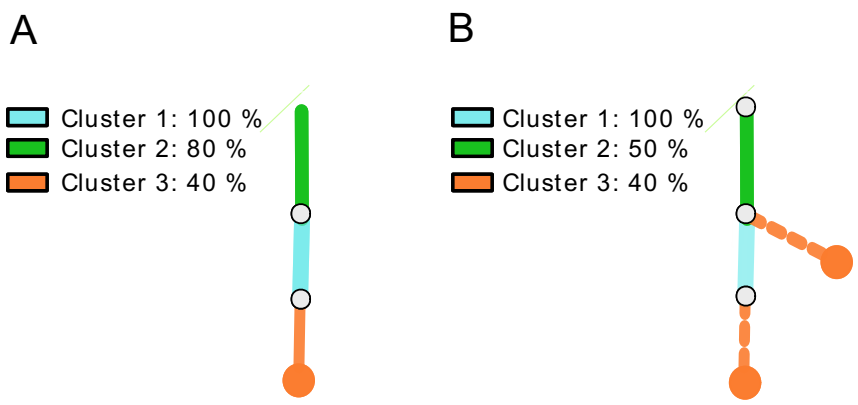


Figure 2.8: Phylogeny reconstruction applying pigeonhole principle. A) A linear phylogeny is represented. B) Parallel phylogeny, where cluster 3 has to possible ancestors.

2.11.2 HIERARCHICAL CLUSTERING

The description by Frigui *et al.*[196] was used for the elaboration of this section. This type of clustering is referred to as hierarchical or nested, as opposed to partitional or unnested clustering. Whereas partitional clustering is just a separation of the data objects into non-overlapping groups, hierarchical clustering allows the existence of subclusters. This means that there can be overlapping groups (nested clusters) that can be organised as a tree. Therefore, a cluster or node can be seen as the combination of its subclusters or subnodes and the root of the tree is the object that contains all the clusters.

There are multiple ways of performing hierarchical clustering, but here we are going to focus on agglomerative hierarchical clustering approach. This technique generates the tree by separating each single point into a cluster and iteratively merging the closest clusters until a single cluster (root of the tree) is obtained. The closest clusters are merged based on their similarity or proximity measure, which is defined using different approaches:

Maximum linkage: the maximum distance between the two points from two different clusters that are farther apart.
Single link: the shortest distance between two points of two different clusters.
Average link: the average distance between each point in one cluster and all the points in the other cluster. This approach was the one used in this thesis.

The distance between any pair of points can be calculated in different ways, which can be quantitative measures or similarity/dissimilarity measures. In this thesis we used quantitative measure called Euclidean distance. It is calculated by measuring the line segment between two points or computing the absolute value of the difference between two points.

2.12 PCA

Principal component analysis (PCA) was first introduced by Karl Pearson[197] and is a widely used technique that is commonly applied for dimensionality reduction and visualisation of data. The main purpose of this analysis is to decrease the number of dimensions of a dataset that presents a high number of attributes (interrelated variables) while retaining as much variation as possible. It also allows to find hidden patterns in the data and identify correlated variables. In order to attain this, the set of observations is transformed to new variables called principal components (PCs). These components are ordered so the first one accounts for as much variability as possible and unlike the original variables, they are linearly uncorrelated.
Let *x* be a vector of *p* random variables. The first step is to generate a covariance matrix $(n \times n)$ that contains the covariances between all possible pairs of *p* variables present on the data set. The covariance between two variables can be calculated using the formula:

$$cov(X,Y) = \frac{1}{n-1}\Sigma_i^n 1\ (X_i - x)(Y_i - y).$$

The variances are represented by the diagonal elements of this matrix, while the rest of the elements correspond to the covariances between variables. Thus, high covariance values (values > 0) indicate redundancy, or correlation between variables. In order to decrease redundancy, the original variables have to be transformed so that the covariance is closer to 0. This is known as *eigen decomposition*, where a square matrix *A* is decomposed into paired *eigenvectors* and *eigenvalues*. An eigenvector $\vec{v}$ is a non-zero vector of a linear transformation of a given matrix *A* that is changed by a scalar factor or eigenvalue λ determined by that linear transformation so that:

$$A\vec{v} = \lambda\vec{v}, \qquad Eq.1$$

The eigenvalues of A can be calculated using the following equation, where *I* corresponds to the identity matrix:

$\vec{v}(A - \lambda I) = 0$, which is equal to:

$$det\begin{pmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2k} \\ \cdots & \cdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} - \lambda \end{pmatrix}$$

After eigenvectors and eigenvalues are calculated and sorted in decreasing order, the eigenvectors with the lowest eigenvalues are removed (reducing dimensionality) to form a matrix *d x k* from the chosen eigenvectors or *k*. The order eigenvectors/eigenvalues are sorted now corresponds to the order of the principal components, the highest pair corresponds to the first principal component.

2.13 BOOTSTRAP

Bootstrapping is a statistical method first introduced by Bradley Efron in 1979[198] used to do a wide range of statistical inferences by measuring the accuracy of sample estimates such as variance, confidence intervals and standard errors. It is also commonly used to assess the stability of results, especially useful when the original sample size is small. Bootstraping is based on resampling with replacement from the original data to create many simulated sample sets of the same size as the original dataset. The original probability distribution J is analogous to the empirical distribution $\hat{J}$ obtained after resampling the original data.

2.14 POSITIVE SELECTION ANALYSIS

2.14.1 IDENTIFYING NEUTRAL EVOLUTION FROM THE VAF DISTRIBUTION

Evolutionary dynamics of tumours are very complex and are essential for the understanding of cancer development and designing successful treatment strategies. As discussed in the Introduction (section 1.1), the evolutionary processes in cancer development can be neutral or driven by selection.

In order to determine the evolutionary dynamics in our samples we used a model developed by Williams *et al*[199]. that is based on the analysis of the variant allele frequency (VAF) distribution of a sample. This model assumes that tumours that evolve neutrally are originated when a single cell gained a significant amount of tumour inducing mutations and therefore all the descendants harbour them. If subclonal mutations occurring later on accumulate at a steady rate and are not under selection pressures we can infer that these are neutral. In this scenario, the VAF of a neutral mutation *f* increases in a constant manner over time, so the age of the tumour is equal to *f*. This can be expressed as follows:

$$\frac{dM}{dt} = \mu\pi\lambda N(t), \qquad Eq.1$$

where *N(t)* is the number of tumour cells at time *t* with a dividing rate of λ for each unit *t,* μ is the mutation rate that results from cell division and π is the average number of chromosome sets in the tumour cell. However, cell division is not always successful, as cells can undergo apoptosis and differentiation. For this reason, instead of μ here we consider the fraction of effective cell divisions or β where both lineages survive. As cells grow exponentially, we can model the mean number as a function of time:

$$N(t) = e^{\lambda\beta t}, \qquad Eq.2$$

To solve this, we integrate over the growth function *N(t)* in a time interval $[t_0, t]$:

$$M(t) = \frac{\mu\pi}{\beta}\left(e^{\lambda\beta t}e^{\lambda\beta t_0}\right), \qquad Eq.3$$

which inform about the total number of subclonal mutations accumulated in a tumour between t_0 and t. The variant allele frequency *f* is defined as the inverse of the number of alleles in a population and therefore can be obtained as follows:

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda\beta t}}, \qquad Eq.4$$

As *f* and *t* are interchangeable, we can redefine equation X like so:

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right), \qquad Eq.5$$

This equation represents the distribution of mutations when under a neutral model of evolution, which is a power law distribution (the cumulative number of mutations at a frequency *f* is proportional to 1/*f*). By using the R package Neutralitytestr[199], we tested this model of neutral evolution on our data. Using linear models the empirical distribution of mutations (VAF) is fit against a 1/*f* power law distribution. The goodness of fit measure R^2 (variance explained by the model) is calculated. Neutral evolution is reported when $R^2 \geq 0.98$. In a more recent paper, Williams *et al.*[200] improved their previous approach by using the curve described by equation 5 and checking whether the empirical data collapsed into this curve after normalization. To compare how well does the empirical data fit against this hypothetical distribution three metrics are used: the area between the empirical data and the theoretical distribution, the Kolmogorov distance between the two distributions and the Euclidean distance of all points between the empirical data and the theoretical distribution. These three metrics showed a much better performance to fit of the empirical data to the neutral 1/*f* power law distribution.
All these metrics are illustrated in Figure 2.9.

2.14.2 FINDING POSITIVE SELECTION AT GENE LEVEL

As stated in section 1.3.6 (See Introduction), there are many methods for analysis of positive selection that consist of the calculation of a dN/dS ratio. When the dN/dS ratio deviates from ~1 positive (or negative) selection is suspected.

In this thesis a dN/dS method[118] adapated from Greenman *et al.*[114] was used to detect the presence of driver genes in morphologically normal tissue. This approach has proven to be more accurate than traditional methods, as it takes into account the background mutation

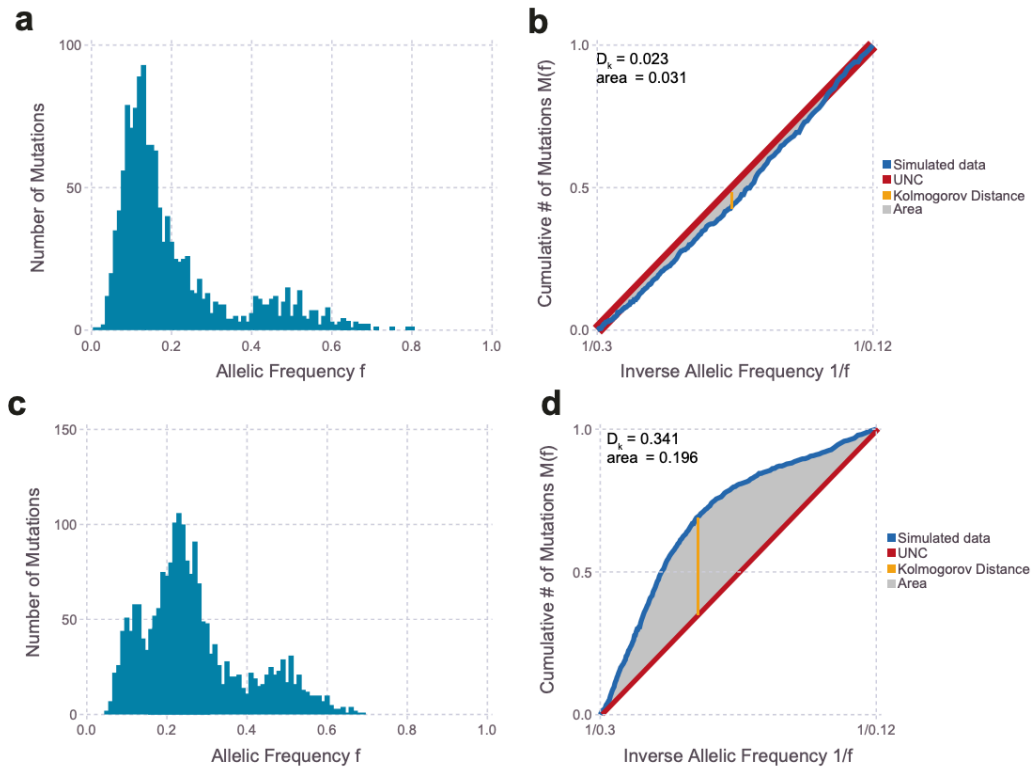


Figure 2.9: Test statistics for neutrality. (A) Example of the VAF distribution of a neutral tumour and (C) a tumour with one selected subclone. To test the model of neutrality three different test statistics were used to compare the empirical data (blue line) to the expected normalised normalised distribution (red line or universal neutrality curve (UNC)). The statistics used were the area between the curves (AUC, grey area), the Kolmogorov distance (orange line) and the Euclidean distance between all points on the two curves. Adapted from Williams *et al.*[197]

spectrum, the length of the sequence of each gene and the variation of the mutation rate across genes. An infinite site dN/dS model is applied, which allows us to have a separate dN/dS ratio for missense, nonsense and essential splice site mutations so we can assess the presence of selection in these mutation types independently. Essential splice site mutations are described as those positions located -2 and -1 upstream of an exon start and +1, +2 and +5 downstream of an exon end.

Each of these mutation types is treated as a random variable modeled through a Poisson distribution. 192 parameters were used to assess the mutation potential of each base and transcriptional strand bias in the coding strand (the nucleotides up and downstream a given base, *e.g.* ACG>ATG). In total, the model has 195 parameters after including all mutation types

(missense, nonsense, splice site and synonymous mutations, so that each type of mutation is modeled as:

$$\lambda_{syn,ACG>ATG} = t * r_{ACG>ATG} * S_{syn,ACG>ATG}$$

$$\lambda_{mis,ACG>ATG} = t * r_{ACG>ATG} * S_{mis,ACG>ATG} * w_{mis}$$

$$\lambda_{non,ACG>ATG} = t * r_{ACG>ATG} * S_{non,ACG>ATG} * w_{non}$$

$$\lambda_{spl,ACG>ATG} = t * r_{ACG>ATG} * S_{spl,ACG>ATG} * w_{spl}$$

$r_{ACG>ATG}$ accounts for the relative rate of ACG>ATG transitions per ACG site in the sequence, *t* accounts for the local mutation rate, $w_{mis|non|ss}$ represent the rate of missense mutations in relation to synonymous mutations, and *S* represents the number of sites that can have a missense mutation in that gene sequence at that location. Therefore, *t* controls for the gene to gene variation in coverage or in the background mutation rate, and *S* controls for the length of the sequence of each gene.

The likelihood of observing a number of missense ACG>ATG mutations in a specific gene ($n_{mis,ACG>ATG}$), with and expected rate ($\lambda_{mis,ACG>ATG}$) is expressed as:

$$L_{mis,ACG>ATG} = Pois\left(\lambda_{mis,ACG>ATG} \middle| n_{mis,ACG>ATG}\right) = \lambda^n * \frac{e^{-\lambda}}{n!}, \qquad Eq.1$$

Using Poisson regression, the maximum-likelihood estimates and confidence intervals for the 195 parameters are calculated. As we have 192 parameters to define the trinucleotide context and 4 different mutation types (missense, nonsense, essential splice sites), we calculate the joint likelihood as the product of each individual likelihood of the 192 trinucleotide rates * 4 types of mutations:

$$L = \prod_{j\in\{1,2,\ldots,192\}} \left[Pois\left(\lambda_{syn,j} \middle| n_{syn,j}\right) * Pois\left(\lambda_{mis,j} \middle| n_{mis,j}\right) * Pois\left(\lambda_{non,j} \middle| n_{non,j}\right) * Pois\left(\lambda_{spl,j} \middle| n_{spl,j}\right)\right], \quad Eq.2$$

The estimates of the 192 parameters are calculated from the total number of mutations and are considered constant across all genes. Then, maximum-likelihood estimates are obtained for *t*, w_{mis}, w_{non}, w_{spl} for each gene. Finally, a likelihood ratio test is used to compare the observed model to a neutral model where $w_{mis} = 1$, $w_{non} = 1$ and w_{spl}=1. False discovery rate is controlled by applying the Benjamini-Hochberg method for multiple testing[201].

2.15 FUNCTIONAL IMPACT ANALYSIS

In order to assess the functional impact of our set of variants we applied the wANNOVAR tool[202]. wANNOVAR annotates the functional impact for each variant by analysing the position (chromosome, location, reference and alternate nucleotides) of each mutation. It applies previously developed tools (SIFT[203], PolyPhen2 HVAR[204], MutationTaster[205], MutationAssessor[206], FATHMM[207], MetaLR[208]) and generates a summary of the functional predictions for those variants.

Although there are slight differences, all these tools are based on analysing the protein sequence conservation patterns. The rationale is that mutations affecting highly variable sites tend to be better tolerated than mutations in highly conserved regions. In order to obtain this information, an alignment of the protein of interest and related proteins with shared function (same family or homologous proteins) is performed to examine the conservation patterns across the amino acid sequence. Based on the alignment results, the normalized probabilities for every possible amino acid change are calculated. Those above or equal to an established cutoff are considered neutral. These tools vary regarding the range of features analysed apart from conservation levels, the databases consulted for the annotations and the methods followed to find homologous sequences. For example, PolyPhen2 HVAR is known for integrating conservation patterns from sequence alignments and three protein structure features that are essential for protein stability and function. MutationTaster, on the other hand, considers other features such as splice sites regions and its effects, protein length and mRNA stability.

2.16 DETECTION OF KATAEGIS EVENTS

For the detection and visualization of clustered mutations in all samples, the R package "ClusteredMutations" was used. The inter-mutational distance (IMD) or distance between each somatic substitution and the substitution immediately prior, was calculated and then represented in "rainfall plots". We considered *kataegis* events when there were more than 10 mutations clustered within 1kb.

CHAPTER 3 : MUTATIONAL LANDSCAPE OF MORPHOLOGICALLY NORMAL TISSUES

3.1 SUMMARY

The presence of genetic changes drives cancer development by altering the characteristics of the cell. Mutations have been identified in genes associated with tumour suppression, DNA repair, cell proliferation and cell adhesion.
In this chapter we characterise the mutational landscape of morphologically normal tissues and tumours from patients with and without prostate cancer. We report and compare the specific genetic alterations (SNVs, indels, structural variants and copy number changes) and mutational signatures present in each group of samples (tumour, morphologically normal tissue from cancer patients and morphologically normal tissue from healthy men). Clear associations could be made between the nature of the sample and the characteristics of the mutational landscape, such as mutation burden and presence of cancer associated genes. Cancer presence was associated with a higher number of SNVs and INDELs in comparison to samples from patients without cancer. CNAs and rearrangements were only found in tumour samples. A wide range of mutational processes were detected in all morphologically normal samples, including previously associated prostate cancer associated processes.

3.2 BACKGROUND

Next generation sequencing technologies have become an essential tool for the characterization of mutational landscape of prostate cancer. As described in the Introduction (section 1.4) many studies[17,18,124,132] have reported that a wide range of genomic changes such as DNA copy number alterations, rearrangements and gene fusions, single point mutations and indels promote cancer development in the prostate. Genomic alterations occur throughout the in lifetime of a cell. Most of them are passenger mutations that do not affect cell behaviour, but sometimes mutations in key genes lead to the development of abnormal cell clones that can lead to cancer. For this reason, the characterisation of morphologically normal tissues is of utmost importance to understand early stages of cancer development.

Overall, there is evidence that morphologically normal tissues harbour mutations and that in some instances there is strong positive selection of cancer associated genes[118]. This indicates that in many cases normal tissues are comprised of a mixture of evolving clones that eventually

could increase the risk for cancer development[168]. However, this has not been explored in prostate.

In prostate, the presence of an a wide range of genetic alterations in morphologically normal tissue form prostate cancer patients[17] suggest that this tissues have the potential of becoming a preconditioned epithelium that acts as a "field" that could lead to cancer, a theory evidenced by the fact that failure to remove margins has been associated with local recurrence[209]. As described in the introduction (Introduction, section 1.5.2), this field could be epigenetic in nature.

3.3 MATERIALS

3.3.1 SAMPLES

We analysed 89 samples (summarized in Table 3.1) from 30 patients with prostate cancer obtained after prostatectomy at Addenbrooke's hospital: 39 samples from morphologically normal tissue and 38 from tumour tissue. An extra 7 morphologically normal tissue samples were collected from men without prostate cancer: 5 samples collected at autopsy at the Tissue and Research Pathology/Pitt Biospecimen Core at the University of Pittsburgh and two from cystoprostatectomy collected at Addenbrooke's hospital. An extra five samples of cell cultured fibroblasts derived from stroma were collected: two at York Teaching Hospital NHS Foundation Trust and three at Castle Hill Hospital in Hull. 10 samples from normal tissue were classified as having benign prostatic hyperplasia (BPH) by the histopathologist Dr. Anne Warren. Multiple morphologically normal samples from the same patient were taken in six cases (Patients 0065, 0073, 0077, 0006, 0007 and 0008) (Table 3.2). All patients had a matched tumour except patient 0240. Matched control from blood and lymphocytes were included for all epithelial and fibroblast samples, respectively.

Samples were collected subject to ICGC standards of ethical consent (https://icgc.org/). Ethical approval for the morphologically normal samples (including BPH) and fibroblasts was obtained from the NHS East of England-Cambridge (REC [03/018]) and from the NHS Hull and East Yorkshire (REC ref 07/H1304/121), respectively. Patients gave informed consent and identities were anonymised at source.

Blood samples were used as normal controls for tumour and morphologically normal tissue. Cell cultured lymphocytes were used as controls for the cell cultured fibroblasts.

	SAMPLES		
PATIENTS	**Normal tissue**		**Tumour tissue**
Cancer (29)	**Normal tissue**	**30 (Prostatectomy)**	**38 (Prostatectomy)**
	BPH	**9 (Prostatectomy)**	
	Fibroblasts	**5 (cell culture)**	
Non-cancer (7)	**Normal tissue**	**6 (2 Cystoprostatectomy and 5 Autopsy)**	
	BPH	**1**	

Table 3.1: Samples collected from morphologically normal, BPH and tumour tissues from patients with and without cancer.

A

Samples	Type	Samples	Type
0006	Tumour 1	0006	Normal 1
0006	Tumour 2	0006	Normal 2
0006	Tumour 3	0006	Normal 3
0006	Tumour 4		
0007	Tumour 1	0007	Normal 1
0007	Tumour 2	0007	Normal 2
0007	Tumour 3	0007	Normal 3
0007	Tumour 4		
0007	Tumour 5		
0008	Tumour 1	0008	Normal 1
0008	Tumour 2	0008	Normal 2
0008	Tumour 3	0008	Normal 3

B

Samples	Type
0065	Normal
0065	BPH
0073	Normal
0073	BPH
0077	Normal
0077	BPH

Table 3.2: A) Three patients with multiple samples from normal and tumour tissue. B) Three patients have an additional sample from BPH.

3.4 METHODS

3.4.1 SAMPLE COLLECTION AND SEQUENCING

A total of 89 samples was collected and reviewed by an histopathologist (see Methods, section 2.3). DNA was extracted and whole genome sequenced (Methods, section 2.5.1). The preprocessing steps were performed by researchers at Cambridge University and the library preparation and sequencing was performed by Illumina Inc. The tool FASTQC (described in Methods, section 2.6) was used to assess the quality of FASTQ files.

3.4.2 PRE-VARIANT CALLING PROCESSING: ALIGNMENT AND DUPLICATE REMOVAL

Sequencing data from each lane was aligned to the GRCh37 reference human genome[178] using the Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW, described in Methods, section 2.7) v0.5.9-r16+rugo using parameters -1 32 -t 6[178]. SAM files can be converted to a BAM format using Samtools[179]. Picard tools (Methods, section 2.8) option was used to assess the quality of the alignment (option *CollectAlignmentSummaryMetrics*), remove duplicated reads (option *MarkDuplicates*) and calculate coverage of the region of interest (option *CollectHsMetrics*). This data has been submitted to the European Genome-Phenome Archive (EGAD00001000689).

3.4.3 VARIANT CALLING

Substitutions, insertions and deletions were detected using the Cancer Genome Project Wellcome Trust Sanger Institute pipeline. Somatic mutations (SNVs), INDELs, structural variants and copy number alterations were called using CaVEMan, Pindel, Brass and the Battenberg algorithm respectively (described in Methods, section 2.9). The filters applied for SNVs and indels are mentioned in Method sections 2.9.1 and 2.9.3 as part of the Cancer Genome Project Wellcome Trust Sanger Institute pipeline.

3.4.4 VISUAL VALIDATION

Visual validation of single nucleotide variants was made for all substitutions of 5 samples using G-browse (Figure 3.3), a genome browser used to view short read sequences. Samples with

differing number of mutations were chosen to determine if there is a relationship between the number of variants obtained and the quality of the calls.

3.4.5 FUNCTIONAL IMPACT OF MUTATIONS

In order to assess the functional impact of our set of variants we applied the wANNOVAR tool[202]. wANNOVAR annotates the functional impact for each variant by analysing the position (chromosome, location, reference and alternate nucleotides) of each mutation. In addition, I compared our variants to the following existing variation databases containing gene annotation datasets: UCSC Genome Browser[210], 1000 Genomes Project[211], dbSNP[212], COSMIC[213] and TCGA[214].

3.4.6 POSITIVE SELECTION

A previously described dN/dS method developed by Martincorena *et al.*[118] (see Methods, section 2.15.2) was used to detect positive selection in coding variants. dN/dS ratios were quantified for missense, nonsense and essential splice mutations using the package R dNdScv (https://github.com/im3sanger/dndscv).

3.4.7 MUTATIONAL SIGNATURE DETECTION

Signature refitting using quadratic programming methods (see Methods, section 2.10.1) was performed using the recently published new mutational catalogue from Alexandrov *et al.*[142] . All the mutational signatures that currently reported in COSMIC were included in the analysis except signature 25 which was not confirmed in the new catalogue[142]. Signatures 27, 43, 45-60 have been classified as possible sequencing artefacts, but their inclusion in the analysis allows the detection of artefacts in our samples.

Using the whole list of reference signatures may lead to overfitting, where the majority of signatures of the list are assigned in most samples (overcalling). A common approach to solve this is the previous selection of signatures to include in the analysis. The limitation of the number of signatures can be performed based on different criteria. In this thesis we explore signature stability as a factor that affects correct signature assignment as described in Results sections 3.5.5.1.2 and 3.5.5.1.3. Alternatively, we used the mutational analysis tool SigProfilerSingleSample[187,188] (described in section 2.10.3) that includes techniques to limit the number of signatures based on previous biological knowledge.

3.5 RESULTS

3.5.1 QUALITY CONTROL

3.5.1.1 FASTQC ANALYSIS

Overall, samples had a good FASTQC report confirming the good quality of the data: per base and sequence quality average was > 30, all bases were proportionately distributed, no overrepresented sequences such as adapters were in the data and duplication levels were low. Regarding GC content we found that a small proportion of the reads differ considerably from the expected normal distribution (the sum of the deviations from the normal distribution represented more than 15 % of the reads). An example of some of these metrics is depicted in Figure 3.1. The proportion of unique reads reported in normal (including BPH and BPH fibroblasts) and tumour ranged from 86% -99% and 88% -98%, respectively (Figure 3.2)

3.5.1.2 SEQUENCING METRICS

A mean coverage of 53X was achieved for normal samples that ranged from 31X to 65X (including BPH and BPH fibroblasts) and of 59X for tumours that ranged from 49X to 92X. The alignment to the reference genome was of good quality, with a percentage of mapped reads that ranged from 85 % - 91 % in normal samples and 88 % - 91 % in tumours. These metrics are represented in Figure 3.2.

3.5.1.3 VISUAL VALIDATION

Visual validation of single nucleotide variants was made for all substitutions for five samples using G-browse (Figure 3.3). A set of rules were followed in order to determine whether the variant was valid or not: presence of the variant in at least one read of the control (from blood), strand bias, very high number of mutations surrounding the variant and multiple base changes at the mutant position all indicated a likely force variant. Strand bias was considered when the variant was only present in the backward or forward read or when only one of the strands had good quality reads. Samples with differing number of mutations were chosen to determine if there is a relationship between the number of variants obtained and the quality of the calls (n = 28 to 1213). This showed that the majority of the variants (86 % in 0074, 84 % in 0077_BPH, 75 % in 0159, 82 % in 0240 and 60 % in 0008_N3) were considered to be either good or tentative quality (Figure 3.4) and that there was no association between the quality of variants

and the total number of variants of each sample (P = 0.38, F-test). Sample 0008_N1 was removed from future analyses as the number of variants was very low (n = 28) and half of them were of poor quality.

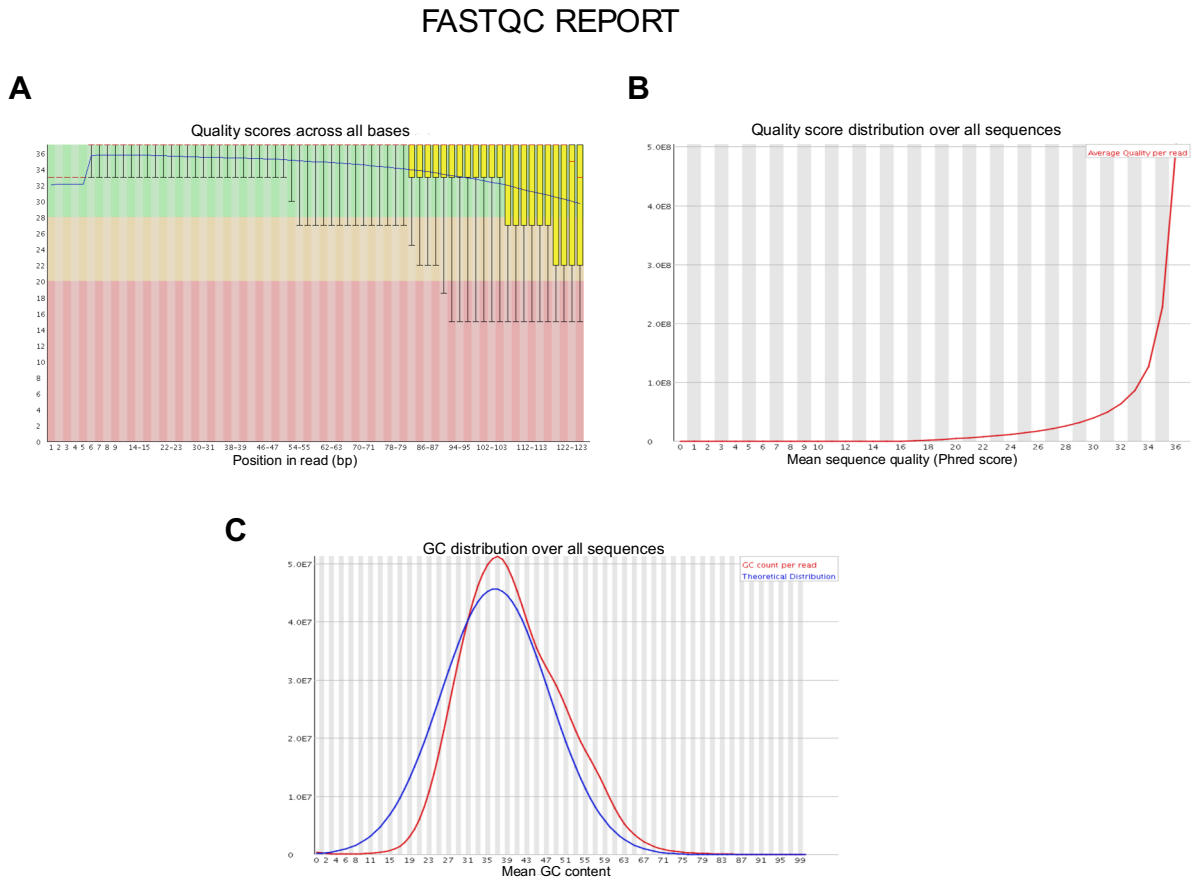


Figure 3.1: An example of quality metrics produced by the FASTQC software. This is for sample 6N1. (A) Quality values across all bases of all sequences. Mean quality, median value and inter-quartile range are represented by the blue line, the central red line and the yellow box, respectively. Points above 90 % and below 10 % are indicated by the lower whiskers. (B) Per base of read sequence quality score. (C) GC content across the length of each sequence in comparison to a modelled normal distribution of GC content. (D) Relative number of sequences with different duplication levels.

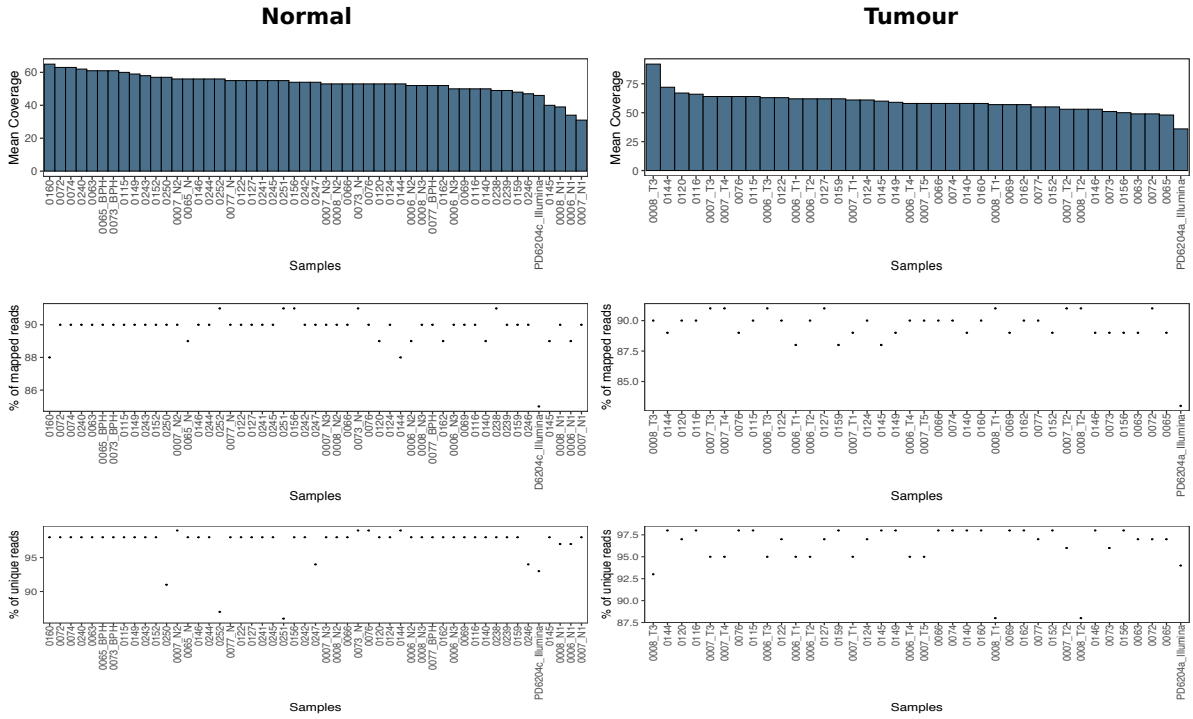


Figure 3.2: Coverage and Alignment metrics: mean coverage is 53X for all normal samples. The % of mapped reads is the proportion of reads that aligned successfully to the reference genome. The % of unique reads plot is the number of reads that remain after removing PCR duplicates and shows very low levels of duplication.

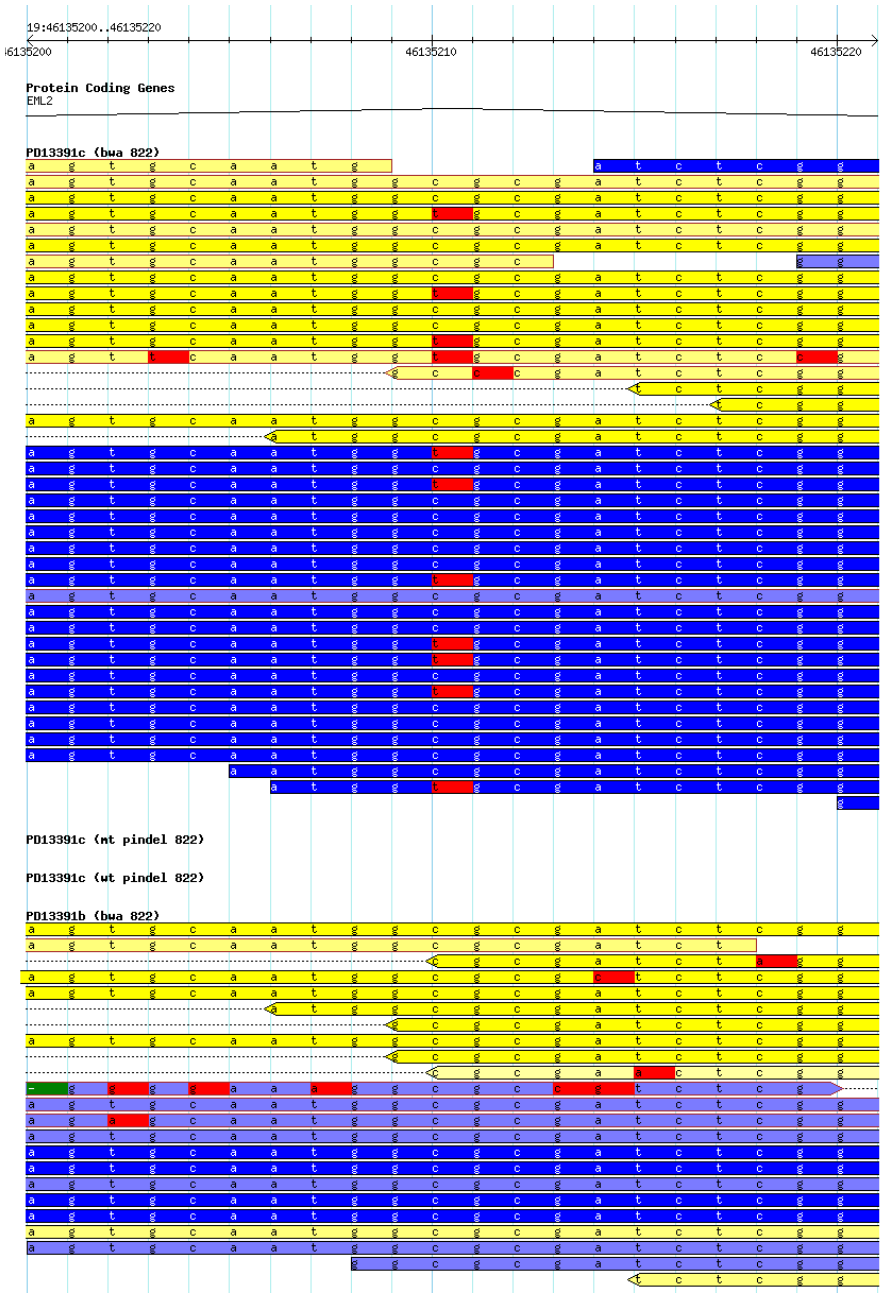


Figure 3.3: Example of variant at 46135211 bp in chromosome 19 for sample 0074. Blue and yellow lines represent the forward and backward reads, respectively. Quality of the read is indicated by the intensity of the colours. This variant occurs both in high quality backward and forward reads and it is absent in the control sample. There is no evidence of indels around the variant.

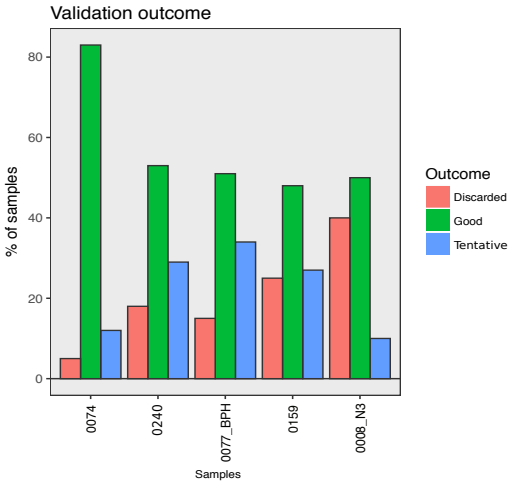


Figure 3.4: Visual validation results for five samples. Variants inspected using G-browse.

3.5.2 MUTATIONAL LANDSCAPE

SNVs, INDELs, breakpoints and copy number changes were called in all 89 samples (Figure 3.5, Table A.1 for normal (including BPH) and fibroblasts samples and Table A.2 for tumour samples in Appendix A). 26,135 SNVs (median of 421 per sample), and 17,370 indels (median of 445) were identified in morphologically normal samples, whereas tumour samples harboured a total of 97,745 SNVs (median of 2,560.5) and 11,087 indels (median of 265). No copy number alterations and only 8 rearrangements (median of 0) (not represented in figure 3.5) were detected across all morphologically normal (including BPH) patients (sample 0063 (n=1), 0127 (n=3), 0073_N (n=1), 0074 (n=1), 0006_N1 (n=1) and sample 0006_N3 (n=1), whereas a median of 22 copy number alterations and a median of 40 rearrangements were found in cancer. Overall, morphologically normal samples had significantly fewer number of substitutions ($P = 6.81\text{x}10^{-12}$), indels ($P = 4.51\text{x}10^{-03}$), copy number alterations (0 in normal) and rearrangements ($P < 2.2\text{x}10^{-16}$). This absence of copy number in normal samples is notable.

In the 166 tumours reported by Wedge *et al.*[127] they found a group of 32 tumour samples with very few CNAs, that we will refer to as "quiet". A comparison between these "quiet" tumours, the high CNAs counterparts and the normal samples analysed in this thesis is presented in section 3.5.7.

Fibroblasts also harboured a high number of SNVs (6,597; median of 1116). The majority of the variants in morphologically normal tissue (including BPH and fibroblasts from BPH) were present in non-coding regions, with only 283 mutations present in exons.

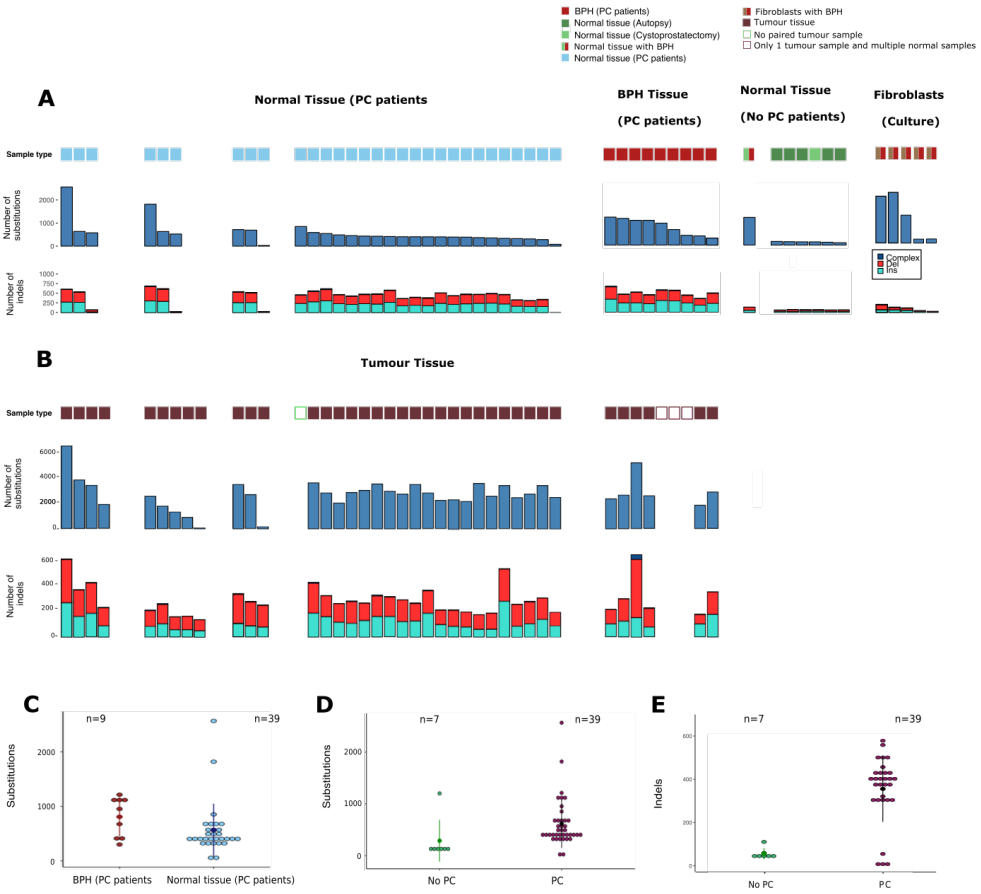


Figure 3.5: Mutational landscape. **(A)** From top to bottom: sample type (morphologically normal tissue in prostate cancer patients, BPH tissue in prostate cancer patients, tissue from non-prostate cancer patients, BPH fibroblast cell culture); number of SNVs detected per sample; number of INDELs (insertions, deletions and complex insertions/deletions) per sample. Each column represents a sample and they are ordered according to sample type and decreasing number of SNVs (see order in column "Samples", Table A.1, Appendix A). 8 rearrangements (not represented in figure) were detected across all patients (sample 0063 (n=1), 0127 (n=3), 0073_N (n=1), 0074 (n=1), 0006_N1 (n=1) and sample 0006_N3 (n=1). No copy number alterations were detected. **(B)** Number of SNVs detected per sample; number of INDELs (insertions, deletions and complex insertions/deletions) per sample. Each column represents a sample and they are ordered according to sample type and decreasing number of SNVs (see order in column "Samples", Table A.2, Appendix A). **(C)** Plot showing the relationships between the number of SNVs between BPH samples and normal samples in prostate cancer patients. **(D)** Plot showing the relationships between the number of SNVs between samples from people with or without prostate cancer. **(E)** the number of INDELs between samples from people with or without prostate cancer.

3.5.3 ASSOCIATION WITH CLINICAL FEATURES AMONG NORMAL SAMPLES

Substitutions and indels were significantly higher in morphologically normal samples from men with prostate cancer compared to those without: SNVs, median 436 for cancer vs 141 non-cancer, $P = 7.1\text{x}10^{-03}$, Wilcoxon rank sum test and; Indels, median for cancer 445 vs 62 non-cancer, $P = 5.5\text{x}10^{-03}$, Wilcoxon rank sum test. Notably, cystoprostatectomy sample 0239 had an exceptionally high number of mutations (1202) in comparison to the other non-cancer patients. There is some evidence that a higher number of mutations is present in BPH samples in patients with prostate cancer compared to other normal tissue (median 952 for BPH compared to 424 for morphologically normal tissue, $P = 0.05$, Wilcoxon rank sum test).

There was no evidence of an association between the number of SNVs and the distance within the prostate between morphologically normal and tumour samples ($\rho = -3.1\text{x}10^{-02}$, $P = 0.85$, Spearman's correlation, Figure 3.6). Similarly, although age is a known contributor to prostate cancer development, only a weak non-significant association was found between age and the number of mutations in morphologically normal samples ($\rho = 0.26$, $P = 0.082$ Spearman's correlation, Figure 3.7). The stromal content of each sample was not linked to whether the sample had BPH ($P = 0.58$, Wilcoxon rank sum test, Figure 3.8, Table A.3 in Appendix A) or the total number of substitutions ($P = 0.09$, Figure 3.9).

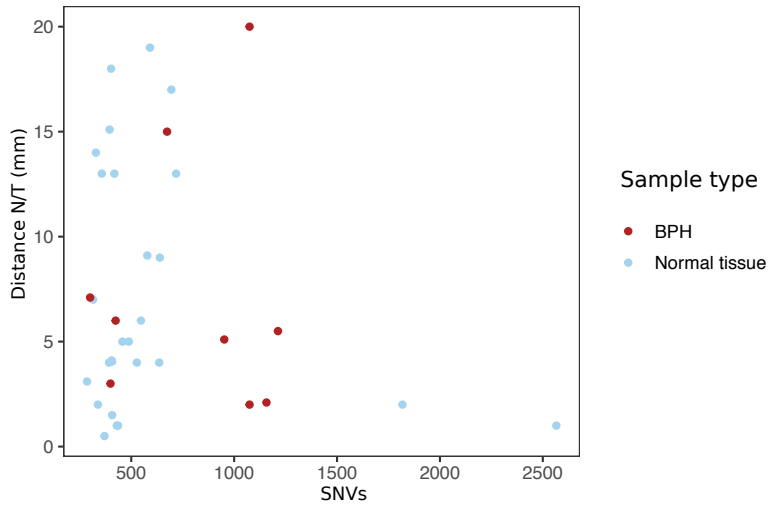


Figure 3.6: N/T distance in relation to total number of SNVs: correlation between the normal tumour distance (in mm) and the total number of SNVs for all samples from prostate cancer patients. There was no association between the variables ($\rho = -3.1\text{x}10^{-02}$, $P = 0.85$, Spearman's correlation).

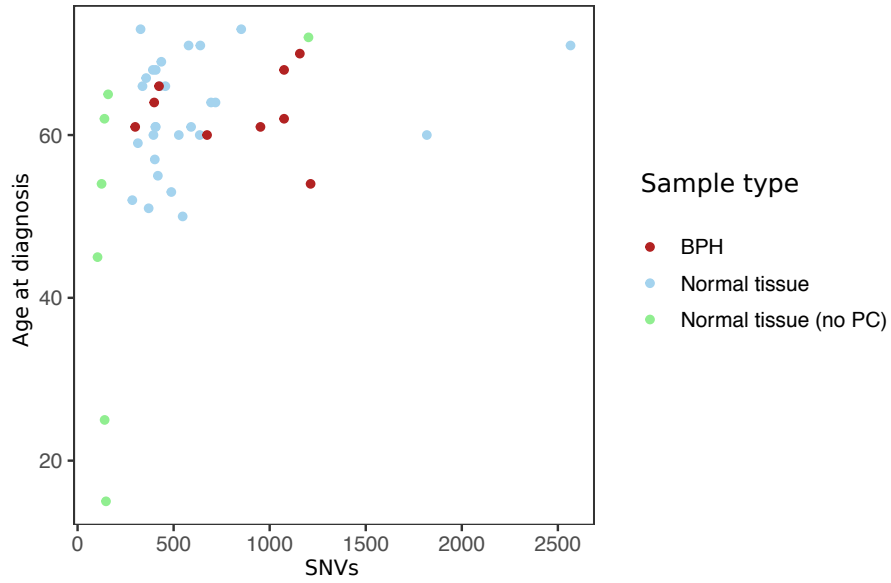


Figure 3.7: Age distribution of patients in relation to total number of SNVs: correlation between the number of SNVs and the age of the patient across all samples. The relationship was not significant ($\rho = 0.26$, $P = 8.2\text{x}10^{-02}$ Spearman's correlation).

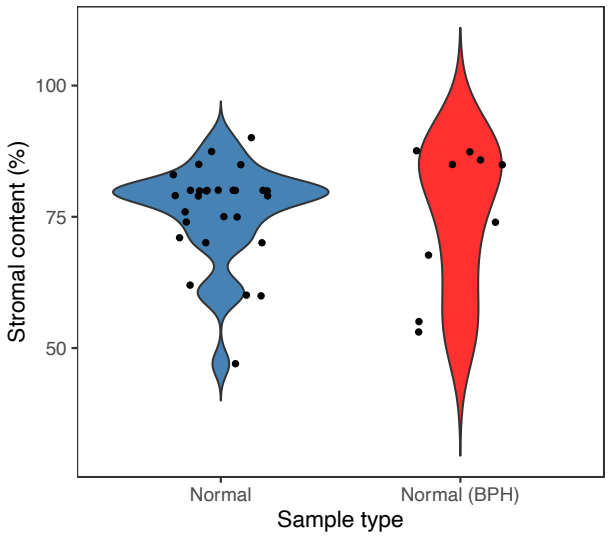


Figure 3.8: Violin plots showing the relationship between stromal content and the presence or absence of BPH.

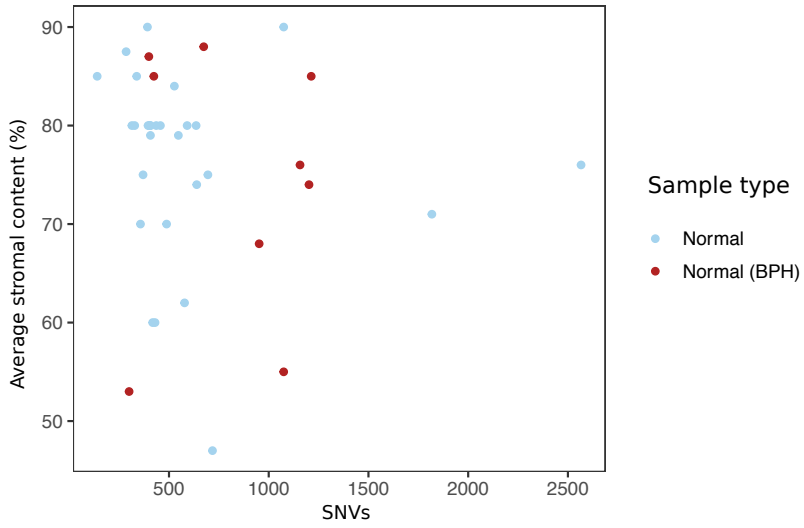


Figure 3.9: Relationship between stromal content and the total number of SNVs in prostate cancer patients. No correlation was found between the two variables (P = 0.09, Spearman's correlation).

3.5.4 GENE MUTATIONS WITH PREDICTED FUNCTIONAL IMPACT IN NORMAL TISSUE

In morphologically normal, fibroblasts and BPH samples a total of 283 substitutions and indel mutations were observed in coding regions of 165 genes. wANNOVAR[202] (described in Methods, section 2.16), was used to predict the functional impact of these group of variants. 113 of the 283 mutations show a potential functional significance and 7 occurred in cancer-related genes (*PPARG, BRCA1, GATA1, WHSC1, POLE, FAT1* and *HOXD11*) reported in the cancer gene census[213] (Table A.4 in Appendix A). Of these, mutations in *GATA1, WHSC1, FAT1* and *POLE* were only observed in samples from a primary prostate fibroblast culture. Mutations with predicted functional impact also included the genes *MIR671, SOBP, CTHRC1, IQGAP1, L1TD1*, *FOXJ3, ATP1A3, PHF12, BCAT1, GMPR2, ADAM28*, DHX32, *DSG3, DDX19A, KIAA1217, PPARG, PTK2B, RPL18* and *XKRX*, which have also been classified by The Cancer Genome Atlas Research Network (TCGA) as a prognostic marker (genes that correlate poorly with the patient survival) for many cancers[214]. All 113 mutations were present in one sample, except for 5 cases: mutations affecting genes *GYPA* and *NACAD* were present in multiple samples from different patients and mutations in genes *BCAT1, FAT2* and *MIR671* were present in two samples from the same patient. Of all 113 in this set, only *BRCA2* and

ADAM28 have been previously classified as recurrently mutated drivers in prostate cancer[127,128]. Dn/dS driver detection (see Methods, section 2.15.2) was performed but no significant hits were found.

Interestingly, from the 113 genes with a predicted functional impact, 13 were also observed to be mutated in tumour samples (Table A.4 in Appendix A). However, mutation in gene *ACOT1* was occurring in a matched tumour from the same patient.

3.5.5 MUTATIONAL SIGNATURES

3.5.5.1 MUTATIONAL SIGNATURE DETECTION USING QUADRATIC PROGRAMMING METHODS

3.5.5.1.1 PRELIMINARY RESULTS

SNVs were assigned to the mutational signatures defined in Alexandrov *et al.*[142] using a quadratic programming method (see Methods, section 2.12) in all normal and tumour samples except for the three samples with less than 100 SNVs (PD2604c_illumina, 0008_N3 and 0007_T4) (Figure 3.10). In these preliminary results all signatures were used. Overall, we detected 25 signatures across morphologically normal (including BPH), cell cultured fibroblasts and tumour samples.

From the 13 signatures detected in tumours there is a group of 8 signatures that have been previously reported as prostate cancer signatures[142] : 1, 3, 5, 8, 18, 37, 39 and 40). Five of these are recurrent (present in at least 5 samples). There are no recurrent signatures among the remaining five signatures discovered in tumour. In contrast, normal samples (not including fibroblasts) show a much more variable pattern: a high number of signatures was detected overall (23), with 10 of them previously reported in prostate cancer of which 8 are recurrent. In comparison to tumours, 5 non-previously associated prostate cancer signatures (9, 19, 30, 32) were detected recurrently. Fibroblasts samples harboured 12 mutational signatures, of which 2 (signature 24 and 16) were unique. These initial results are presented in Figure 3.10.

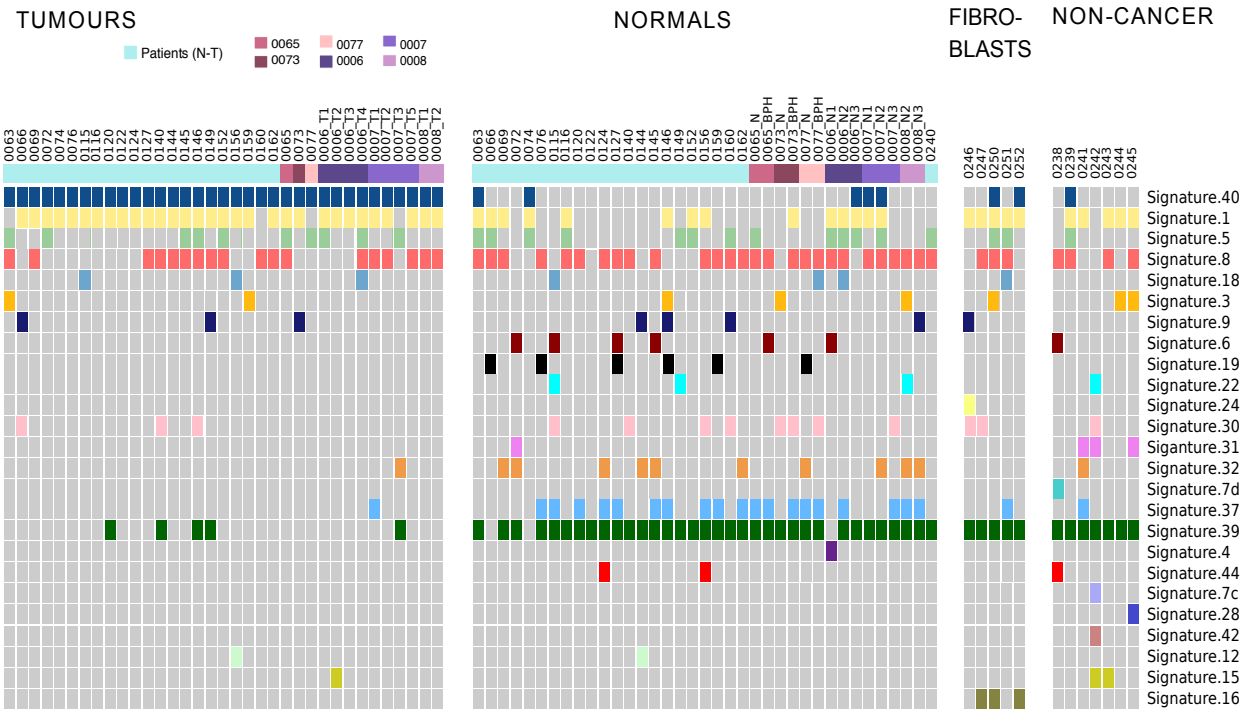


Figure 3.10: Mutational signatures detected in tumour and matched morphologically normal tissue from prostate cancer patients and normal tissue from men without prostate cancer. All signatures were used in this preliminary analysis.

3.5.5.1.2 STABILITY ANALYSIS

To estimate the confidence and stability of the detected signatures, we followed the method described by Huang *et al*[215]. With this approach, the SNVs of each patient was perturbed 1000 times using random resampling with replacement. For each bootstrap, signature contributions were estimated using quadratic programming methods. Subsequently, we compared the signature's contributions obtained from the original data and after bootstrap by calculating the mean squared error (MSE) between the estimated proportion of signatures in each case (Figure 3.11. In agreement with previous results[215], some of the signatures detected showed a remarkable variability after bootstrap. This is the case of signatures 5, 37, 54, 32, 30, 19, 8, 6 and 3, among others.

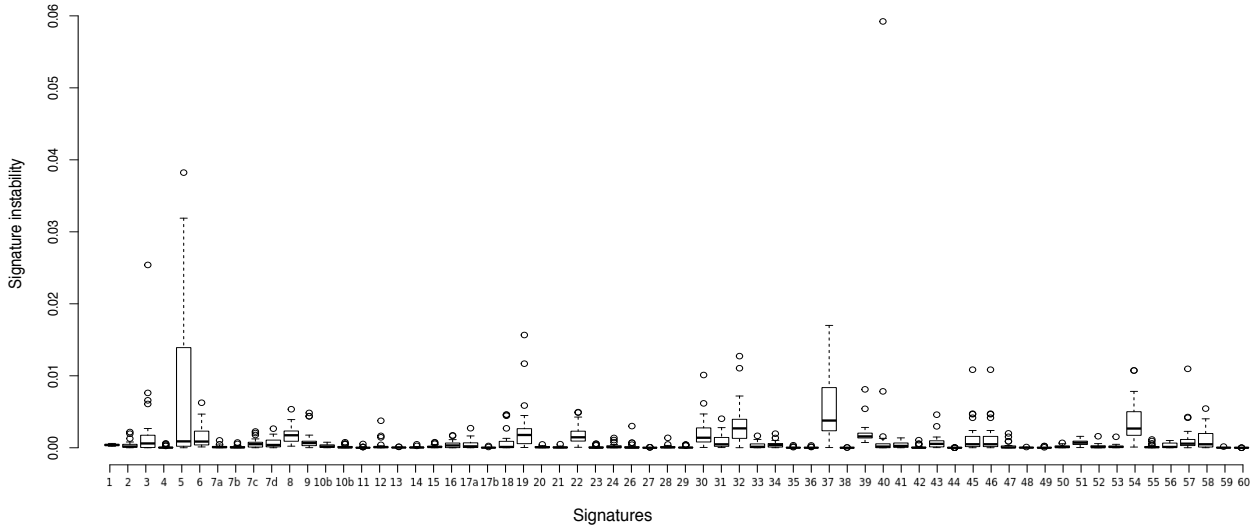


Figure 3.11: Mean squared error (MSE) of all the mutational signatures contributions after bootstrap across all samples.

3.5.5.1.3 INCREASING ROBUSTNESS OF SIGNATURE SELECTION

We decided to use a similar approach in order to increase the robustness of signature detection. We introduced a preliminary step to select which signatures would be included in the analysis and thus, reducing the possibility of overfitting. First, each patient mutational profile was randomly altered using bootstrapping and signature contributions were estimated using quadratic programming methods for each bootstrap. Signatures that were detected in the lowest number of iterations of bootstrap resampling were removed one at a time until all the signatures were present in at least 90% of iterations. Second, quadratic programming methods were applied to the original data using the selected signatures after bootstrapping. Following the recommendation by Rosenthal *et.al*[216] only those signatures that contributed to at least 6 % of SNVs were used in the final assignment.

3.5.5.1.4 MUTATIONAL SIGNATURES AFTER BOOTSTRAP

The analysis was repeated after including only those signatures that showed to be robust (selected signatures after bootstrapping, present in at least 90% of replicates), as described in section 3.5.5.3. Overall, the total number of signatures (not including fibroblasts) detected after bootstrap decreased from 25 to 20 in the final set: 6 in tumours, 20 in morphologically normal and 12 in fibroblasts (with overlaps among all the groups). These results still show a much more diverse set of signatures in morphologically normal (n=20) in comparison to tumour

(n=6). In tumour, all non-recurrent signatures that lacked a previous association with prostate cancer were no longer detected. In morphologically normal tissue the signatures 12, 15, 16, 28 and 42 disappeared completely. Interestingly, they had very low contributions (~10%), accounted only for a small number of mutations (median of 9) and were present in less than 3 samples.

Mutational signatures 1, 3, 5, 8, 18 and 40 were detected both in tumour and morphologically normal tissue/BPH samples (Figure 3.12). All of these signatures have been previously been identified in prostate cancer samples[142]. Signature 1 (associated with ageing), signature 5 and the recently discovered signature 40 were overrepresented in tumour samples ($P = 2.39\text{x}10^{-05}$, $P = 9.62\text{x}10^{-04}$ and $P = 5.85\text{x}10^{-08}$ respectively, Fisher's exact test). The aetiologies of signature 5, signature 8 and signature 40 are unknown[142].

Fourteen signatures (4, 6, 7b, 7c, 9, 19, 22, 26, 30, 31, 32, 37, 39, 44) were unique to morphologically normal tissue/BPH. There were no significant differences between morphologically normal tissue and BPH samples for any of the signatures ($P > 0.05$; Fisher's exact test). Signatures 39, 9, 37, 3, 32, 30 and 6 were present in at least five samples, whereas the rest are only present in single samples. Signatures 39 and 37 have been previously observed in prostate cancer, but their aetiology is unknown[142].

Interestingly, the set of signatures found in cell cultured fibroblasts (1, 3, 5, 9, 40) resembled more the one found in tumours than in morphologically normal tissue. Both groups harboured the same signatures in similar proportions ($P < 0.05$ for all signatures, Fisher's exact test) with the exemption of signature 9 that was only present in fibroblasts ($P = 1.28\text{x}10^{-02,}$ Fisher's exact test).

In order to explore whether samples of the same nature (morphologically normal, tumour, normal-BPH and cell cultured fibroblasts) share similar mutational signatures, we applied unsupervised hierarchical clustering (see Methods, section 2.11.2) and principal component analysis (PCA) (see Methods, section 2.13). For clustering, a dendrogram was constructed using the relative contributions of the different mutational processes of each sample. The proximity between each cluster was defined as the average distance between each point in one cluster and all the points in the other cluster. The Euclidean distance was calculated to measure the similarity between clusters. We also performed multiscale bootstrap resampling using the R package pvclust (http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/).

These analyses suggest that the overall pattern of signatures in normal tissue and tumour are different (Figures 3.13 and 3.14), with morphologically normal/BPH samples showing a much more diverse set of signatures (n=20) compared to tumours (n=6).

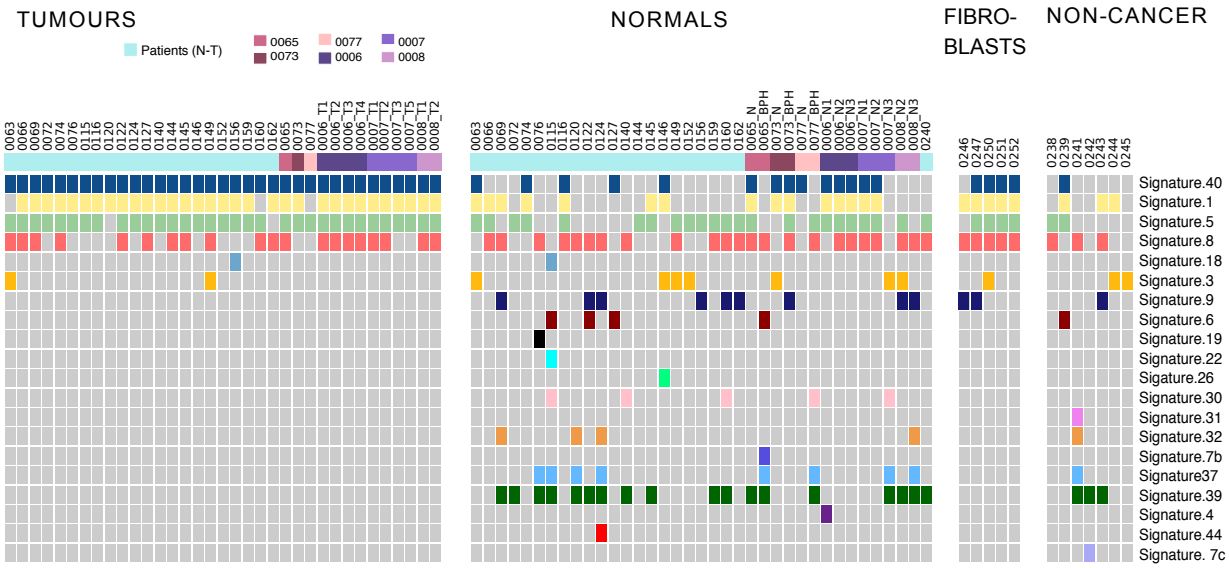


Figure 3.12: Mutational signatures detected in tumour and matched morphologically normal tissue from prostate cancer patients and normal tissue from men without prostate cancer. To estimate the confidence and stability of the detected signatures, bootstrapping was performed in order to perturb each patient's mutational profile. Six patients had more than two samples analysed and one morphologically normal sample did not have a matched tumour.

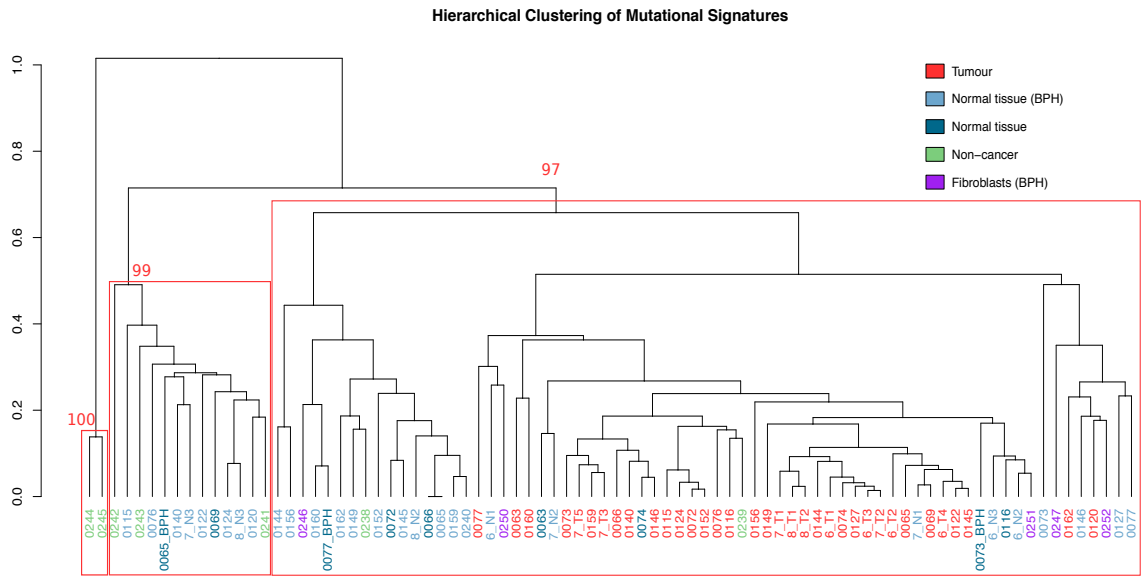


Figure 3.13: Hierarchical clustering of mutational signatures: dendrogram constructed by unsupervised hierarchical clustering using the relative contributions of mutational signatures in each sample. The distance between two clusters was defined as the average distance between each point in one cluster and all the points in the other cluster. Clustering uncertainty was assessed by multiscale bootstrap resampling using the R package pvclust. Red squares represent the major clusters that are significantly supported by the data ($P < 0.05$; number in red is the number of resamplings where the cluster was present).

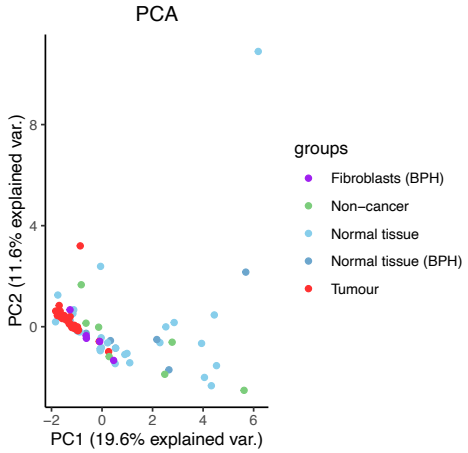


Figure 3.14: Principal component analysis of mutational signatures. PCA was performed using the relative contributions of mutational signatures in each sample. Each sample type (tumour, BPH, normal and non-cancer tissue) is represented by a different colour (red, dark blue, light blue and green).

3.5.5.2 MUTATIONAL SIGNATURE SELECTION USING SIGPROFILER

Mutational signatures were also inferred for each sample using SigProfilerSingleSample[187,188] using the set of signatures defined by Alexandrov *et. al*[24]. Mutational signatures 1, 5, 8, 18 and 40 were detected both in tumour and in morphologically normal tissue/BPH samples (Figure 3.15). All of these signatures have previously been identified in prostate cancer samples[110]. Signature 1 was overrepresented in tumour samples ($P = 4.89\text{x}10^{-03}$, Fisher's exact test). The total number of signatures detected morphologically normal tissue decreased drastically from 20 (obtained using quadratic programming methods and bootstrap and presented in section 3.5.5.1.4) to 7 signatures. The cosine similarity between the reference signatures and the reconstructed profiles was significantly higher in tumour samples in comparison to normal samples (median of 0.97 for tumour vs 0.88 normal), likely the result of a lower number of SNVs in normal tissues. Furthermore, a clear association between the number of SNVs for each set of samples and the cosine similarity was observed ($P = 1.69\text{x}10^{-07}$ $P = 4.26\text{x}10^{-03}$ in normal and tumour, respectively, F statistic).

Three signatures (3, 4, and 28) were unique to morphologically normal tissue. There were no significant differences between non-BPH morphologically normal tissue and BPH. Signatures 4 and 28 were present in only one sample, whereas signature 3 is present in 10 samples.

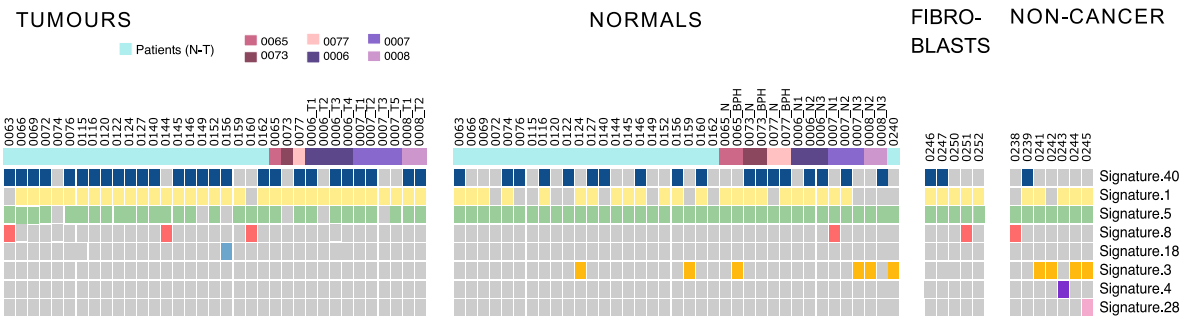


Figure 3.15: Mutational signatures detected using SigProfiler in tumour and matched morphologically normal tissue from prostate cancer patients and normal tissue from men without prostate cancer.

3.5.6 HYPERMUTATION ZONES OR KATAEGIS

19 *kataegis* events (clusters of at least 10 mutations that occurred within 1000kb) were detected in 12 tumour samples that ranged from 10-27 (median of 13) (Table A.5, Appendix A). (Figure 3.16). No *kataegis* events were detected in normal tissues as hypermutation being very low, with the most mutated 1000 kb regions harbouring a maximum of 6 mutations.

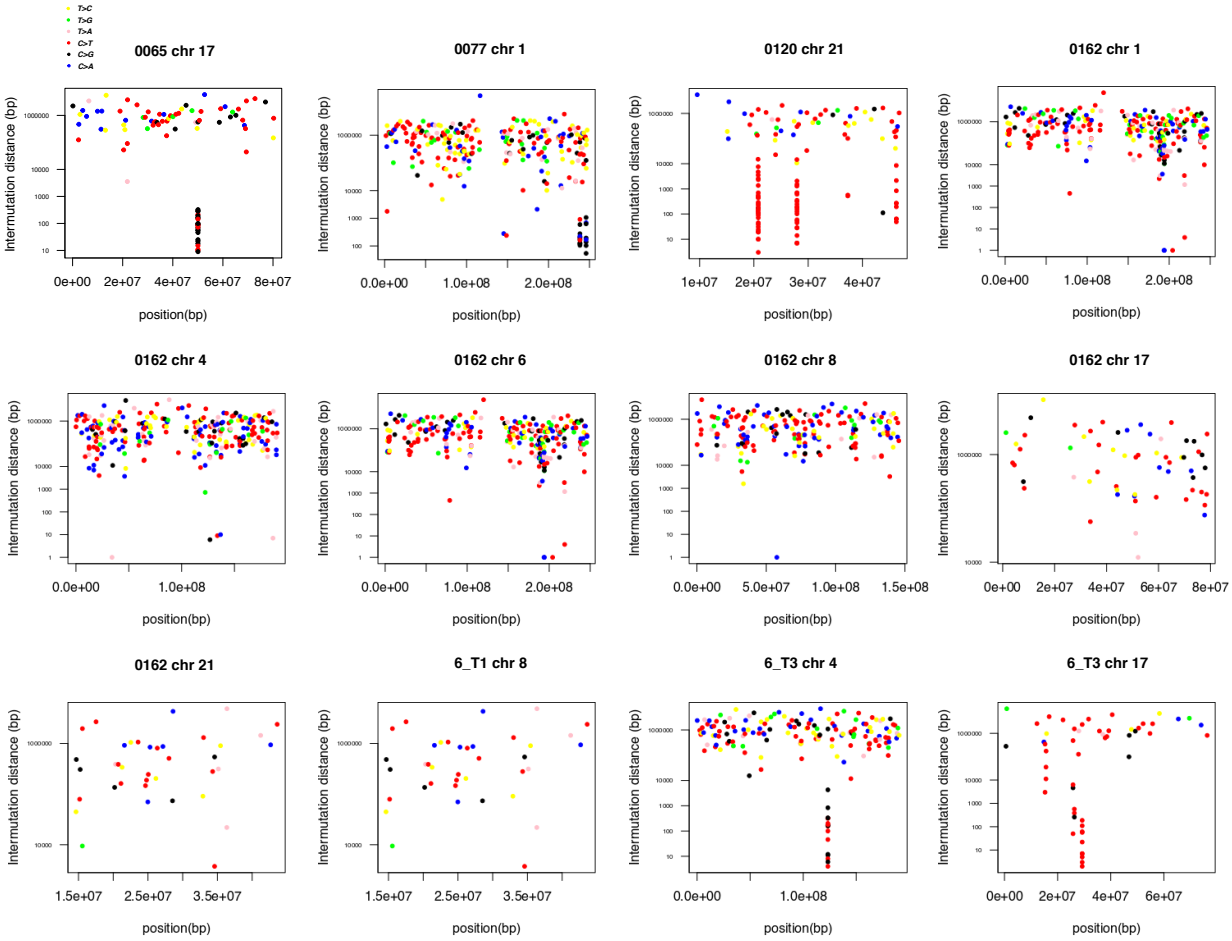


Figure 3.16: *Kataegis* events in tumour samples by chromosome. The plots represent regional clustering of mutations. IMD is represented in the y-axis on a log base 10 scale. Mutations are ordered on the x-axis according to genomic position in the genome. The colour of each dot represents a nucleotide change: T > C (yellow), T > G (green) T > A (pink), C > T (red), C > G (black), C > A (blue).

3.5.7 COMPARISON BETWEEN NORMAL AND "QUIET" TUMOURS

Substitutions and indels were significantly lower in morphologically normal samples in comparison to tumour samples reported in Wedge *et* al.[127] whose total genome affected by CNAs was ≤ 6 % ($P = 1.49x10^{-12}$ for SNVs and $1.34x10^{-13}$ for indels, Wilcoxon rank sum test, Table 3.3). However, these "quiet" tumours had significant fewer alterations overall than the high CNAs counterparts ($P = 5.06x10^{-07}$ for SNVs and $6.77x10^{-07}$ for indels, Wilcoxon rank sum test). Only 7 mutations of which 4 had shown a potential functional significance in this thesis (*ACOT1, MST1, MUC20* and *ARL13B,* section 3.5.4) overlapped between normal samples and these group of tumour samples. All these mutations were present in those tumours with more than 6 % of the genome altered by CNAs.

% of genome affected by CNAs	Number of samples	Average SNVs	Average indels	Samples
< 6	32	2214	2177	Tumour samples from Wedge et al.[128]
> 6	134	2902	2867	Tumour samples from Wedge et al.[128]
0	44	427	670	Normal samples from WGS

Table 3.3: Proportion of tumour samples affected by CNAs examined by Wedge *et al.*[128] The average number of SNVs and indels is shown for each group. The third row shows the average of each type of mutation for the normal samples in this thesis.

Chromosome	Position	Gene	Mutation type	Sample present
1	1637160	CDK11B	silent	Normal/Tumour (non-quiet)
14	74004547	ACOT1	missense	Normal/Tumour (non-quiet)
21	46918360	SLC19A1	missense	Normal/Tumour (non-quiet)
3	49725021	MST1	missense	Normal/Tumour (non-quiet)
3	195452645	MUC20	missense	Normal/Tumour (non-quiet)
16	1291928	TPSAB1	silent	Normal/Tumour (non-quiet)
3	93761891	ARL13B	missense	Normal/Tumour (non-quiet)

Table 3.4: List of genes that were both present in tumour samples analysed by Wedge *et al.*[128] and the normal samples here examined.

3.6 DISCUSSION

These results demonstrate a number of critical features about the mutations present in non-neoplastic tissue taken from cancerous and non-cancerous prostates. A high number of substitutions and indels along with multiple mutational processes were detected in morphologically normal prostate cancer samples, a finding that confirms the findings previously reported by Cooper *et al.*[17]. They also noted prostates 0006 and 0007 lacked telomere attrition, a feature associated with aging cells and observed in the blood from a 115 old woman where mutations were suspected to have arisen due to somatic mosaicism[217]. In contrast, quiescent tissues such as the brain lacked mutations. Given that the prostate is a rather quiescent tissue, the number of mutations reported in this chapter is striking and constitute evidence of a field effect in the prostate.

The number of SNVs and indels was found to be significantly higher in morphologically normal samples from cancer patients compared to non-neoplastic samples. Our results support the idea that the presence of substitutions and indels possibly contributing to subclonal expansions in non-neoplastic tissue is a feature associated with cancer development. However, it is not clear whether these clonal expansions initiate the cancerous process or occur after cancer has already developed. This finding is similar to that previously reported in leukemia[163–166], normal skin and esophagus[118,162].

Morphologically normal tissues were characterised by very few rearrangements and a complete lack of copy number alterations and *kataegis* events. This is in total contrast to the overall spectrum of mutations observed in prostate cancer samples that in general contain an abundance of all classes of genetic alteration and it is specifically characterised by a high number of rearrangements such as *TMPRSS2-ERG*. A previous study by Mehdi *et al.*[218] analysed samples from 2647 subjects and reported that CNAs occur naturally in 4-9 % of the genome in healthy individuals. However, reports regarding the presence of copy number alterations in normal tissue of the prostate are mixed. A study by Yu *et al.*[219] found that normal tissue adjacent to prostate tumours in prostate cancer patients harboured CNAs, some of which were also observed in the tumours. Of note, our samples were collected from morphologically normal areas that were distant (5 mm or more) from the tumour in 20/33 of cases. On the other hand, the examination of whole genome sequencing data from BPH tissue (non-neoplastic)

from prostate samples revealed almost no CNAs in comparison to tumours[78]. Several studies[160,220] that analysed whole genome sequencing data from normal liver and lung in patients showed similar results to the ones reported in this thesis: copy number changes and rearrangements were common in tumours but not in normal samples.

Conversely, the occurrence of CNAs is generally but not always a characteristic of prostate tumours. A study by Wedge *et al.*[127] reported that a significant proportion of prostate tumour samples presented SNVs and indels but minimal CNAs. Although the number of SNVs and indels in these samples was significantly higher than in morphologically normal samples, it was significantly lower than in the high CNAs counterparts. In addition, mutational processes detected in these samples were more similar to those observed in the tumours analysed in this thesis. These findings suggest that the mechanisms that generate CNAs (along with chromosomal rearrangements) are not always operative in neoplastic processes and are not a requirement for prostate cancer development. Alternatively, the lower number of CNAs may reflect an early stage of the disease[30].

Homologous recombination (HR), non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) and microhomology-mediated break-induced replication (MMBIR) are DBS repair mechanisms that could result in CNAs, rearrangements and hypermutation. Specifically, the last three are more mutation prone than the highly accurate homologous recombination (HR). A study by Ponder *et al.*[221] showed that the activation of the stress response has the potential to favour the more error prone MMBIR mechanism over HR. DBS repair mechanisms generate single stranded DNA (ssDNA), which due to its higher vulnerability to damage has been associated to *kataegis* events[222]. The absence of these three types of genetic alterations in normal samples suggest that this type of DNA damage by DBS and errors in the repairing mechanisms (or both) occur at a lower rate in normal samples. In contrast, the clear association of SNVs and indels in normal tissues with prostate cancer presence supports the potential increase of replication errors and non-DBS DNA damage produced by endogenous or exogenous environmental factors.

It has to be noted that the number of indels in morphologically normal tissue is higher than in tumours. It has been observed that the number of indels in human genomes is normally below 10 % of the total number of SNVs[223,224]. Our data shows that in normal tissue the total number of indels is 63 % of the total SNVs, whereas in tumour is 11 % of total SNVs. The high number

of indels in morphologically normal tissue could be an inaccurate result and more advanced filtering criteria should be considered. However, given that indel variants were not used for any of the analyses presented in this thesis and therefore the number of indel variants do not have any implications on the interpretation of results, no further measures were applied.

Prostate cancer associated mutational signatures[142] were present in both morphologically normal and tumour tissue using the different approaches presented in this thesis. Non-linear methods detected a highly diverse set of signatures (20) in morphologically normal tissues, even after applying bootstrap techniques. This result is in contrast to the 6 signatures detected in tumour. The lack of previous association to prostate cancer of the majority of these signatures and the low representation in the normal samples (35 % of the signatures were present in only 10 % of the samples) suggest that the mutational signature profiles in normal tissues include incorrectly assigned signatures. This is further supported by the fact that signature refitting performs less accurately as the number of mutations per sample decreases[225]. Morphologically normal samples in this study had fewer mutations in comparison to tumours (median of 421 vs 2560.5, respectively). Bootstrapping and signature selection increase the robustness of the detection, but there are still false signature assignments. In comparison, mutational profiles obtained using SigProfilerSingleSample showed a small set of 7 signatures of which only two signatures (Signature 4 and 28) have not been previously associated with prostate cancer[110]. A greater number of mutations was significantly associated with the higher similarity of each sample to the reconstructed profile.

The presence of prostate cancer associated signatures in normal tissue of the prostate has been previously reported by Cooper *et al.*[17] and suggests that the same processes driving prostate cancer are at least partly responsible for the mutations in normal tissue. Other studies[226,227] have also shown that mutational processes commonly occur in morphologically normal tissues, including normal tissues from BPH[78]. Of the 9 mutational signatures detected by Deli Liu *et al.*[78] we detected only signature 1 in 9/10 of the BPH samples. Signature 1 has been associated with age, which is one of the main causes of both prostate cancer development and BPH onset.

Higher mutation rates were observed in BPH samples in comparison to normal samples without BPH. BPH disease has been associated with the hyper-proliferation of stromal tissue[228]. Whether the increase of alterations in BPH normal samples is related an early cancerous process or unique to BPH pathogenesis alone is unknown. Although stromal proliferation has

been observed to be much higher than that in epithelial tissue, no association between stromal content and mutation burden was found. A better identification of the stromal/epithelial compartments through laser-capture microdissection could render different results.

A high number of mutations was detected in the cell cultured fibroblasts from BPH samples and four of them occurred in genes reported in the Cancer Gene Census[213]. It has been observed that stromal cells from BPH, unlike stromal cells from normal prostatic tissue, have capability of inducing growth of prostatic epithelia in vivo[228]. However, these growths are non-neoplastic.

Non-cancer sample 0239 had BPH, a high number of substitutions (1202), a pattern that was absent in the other non-cancer patients. This finding suggests that development of BPH alone is a factor that can increase mutation burden in normal tissues, likely the result of a higher proliferation rate observed in BPH cells[229]. It seems plausible that the increased accumulation of somatic mutations could over time increase the risk for cancer development[168], a theory in agreement with previous studies reporting that although a causal link has not been established between BPH and prostate cancer, an association exists[70,75,230].

Apart from isolated cases (*PPARG, BRCA1, GATA1, HOXD11, WHSC1*, FAT1 and *POLE*) we did not find evidence for the presence of mutations in known or novel genetic drivers, which highlights the possible importance of epigenetic alterations in driving expansion. This observation is consistent with previous findings[78] in BPH tissue where they found somatic mutations but no recurrent mutations that would indicate positively selected genes. Overall, these group of genes have been associated with leukemia[231,232],breast[233–236], bladder[233,237] colon[238,239], kidney[240], endometrial[241], head and neck carcinoma[242–244], pancreatic[233] and prostate[245–247] cancers. These genes have been associated with tumour suppression (*BRCA1* and *FAT1*[248,249]), DNA repair (*POLE*[250]), morphogenesis (*HOXD11*), epigenetic regulation (*WHSC1*[251]), lipid metabolism (*PPARG*[252]) and red blood cell development (*GATA1*[253]). Notably, four of the potential driver genes (*GATA1, WHSC1, FAT1* and *POLE*) were only observed in samples from primary prostate fibroblasts from BPH samples.

The finding of a few mutations in a small group of genes is in line with reports from cancers where no initiating event has been identified and supports the suggestion that there might be epigenetic cause for the mutations we observe and the field effect. This hypothesis is supported by several studies that have reported high hypermethylation levels in genes such as *APC*,

GTSP1 and *RASSF1* in morphologically normal tissue in the prostate[155,171–173] and it has proven to be a better predictor of cancer development than histopathological examination alone[171–173]. Hypermethylation in genes *APC* and *GTSP1* was reported in 95% and 43% respectively in patients with an initial negative biopsy that later developed prostate cancer[171].

In summary, these results show potential evidence of a field cancerization in the prostate and provide key insights into genomic evolution of prostate cancer at very early stages of development. The examination of the subclonal architecture presented in the next chapter will help to improve our understanding of evolutionary dynamics in normal tissues from men with and without prostate cancer and support the findings presented in this chapter.

CHAPTER 4 : RECONSTRUCTION OF THE SUBCLONAL ARCHITECTURE

4.1 SUMMARY

In this chapter we study clonal expansions detected in morphologically normal (including BPH samples) and tumour samples using Bayesian Dirichlet clustering process (see Methods, section 2.11.1) using allele frequency of variants and copy number changes. Phylogenies were reconstructed to illustrate how the clones observed evolved and to determine whether normal and tumour samples from the same patient shared a progenitor. Variant allele frequency (VAF) distributions were analysed in all samples to explore the evolution dynamics and report those samples where neutral evolution is suspected.

4.2 MATERIALS

4.2.1 DATASETS

We analysed the same set of samples described in section of chapter 3 (Table A.1, Table A.2, Appendix A). Ethical approval was obtained as described in section 3.3.1 of Chapter 3.

4.3 METHODS

4.3.1 DATA PROCESSING

Two steps were performed before reconstructing the subclonal architecture: the adding of low frequency variants and the removal of variants that have been reported as single nucleotide polymorphisms (SNPs).

4.3.2 LOW FREQUENCY VARIANTS

The Bayesian Dirichlet process uses the substitutions that were detected using the CaVEMAN variant caller. Many variants are discarded as they have insufficient coverage i.e a very low allele frequency (< 5 %). However, for the detection of subclones and the reconstruction the correct phylogenetic tree these low allele frequency variants are crucial. To rectify this, from multiple samples from a patient, we retrieve substitutions with low allele fractions in one sample if they were detected (passing all CaveMAN filters) in another sample from the same patient. In order to do this, we retrieved all the read counts for each base (A, C, G and T) for all the variants detected by CaveMAN from the BAM file using the tool *Bam-readcount*

(https://github.com/genome/bam-readcount). This tool produces the number of reads for each base at each position in a list provided by the user. If a sample had at least one read with a variant reported with certainty in a related sample, the variant was included for that sample in the analyses performed in this chapter. This approach was applied to all the patients where more than one sample was analysed.

4.3.3 CLONAL EXPANSION DETECTION

The subclonal architecture of normal and tumour samples from individual prostate was reconstructed using a Bayesian Dirichlet process adapted to cluster substitutions from whole genome sequencing data in *n* dimensions[112] where *n* is the number of related samples from a patient (see Methods, section 2.11.1). For those cases where there was only one sample such as non-cancer patients (cases 0239, 0240, 0241, 0242, 0243, 0244 and 0245) and fibroblasts (cases 0247, 0250, 0251 and 0252) (Table A.1, Appendix A) a 1-dimensional Dirichlet process was applied. The fraction of cells carrying a particular mutation was estimated from the mutant allele fraction, copy number alterations (CNAs) and cellularity (see Methods, section 2.11.1). Copy number alterations were identified using the Battenberg algorithm (see Methods, section 2.9.5). The prior distribution was defined by $P_0 \sim U(0, 1)$ and $\alpha \sim \Gamma(0.01, 0.01)$.
These priors have been used previously[103]. The posterior distribution of the parameters of interest (the number of clusters, the fraction of cancer cells in each cluster and the number of mutations that belongs to each cluster) is estimated by Gibbs sampling. Only those clones supported by at least 1.5 % of total substitutions for each patient were kept. For cases 0006-0008, clusters supported by less than 1.5 % of total substitutions but previously validated by deep sequencing[17] were used for the phylogeny reconstruction.

To illustrate the relationship among different clones, phylogenetic trees were constructed using the pigeonhole principle (See Methods, section 2.11.1.2). In all cases the allele frequencies of the subclone were significantly different to the estimated background rate ($P < 0.05$). Each tree was annotated with their corresponding genomic alterations: substitutions, indels, rearrangements and copy number changes. The assignment of substitutions to each clone or subclone was performed using the information generated by the Bayesian Dirichlet process clustering. For indels, rearrangements and copy number we assigned to each cluster the number of alterations detected for each sample. We also calculated shared alterations between multiple

samples for the same patient and checked if the pattern of unique and shared mutations were in agreement with the subclonal architecture identified using the Dirichlet process.

4.3.4 NEUTRAL EVOLUTION ANALYSES

Neutrality analyses were performed using the R package Neutralitytestr[199]. This package uses variant allele frequencies from sequencing data and fits a distribution that is predicted from a neutral model of evolution https://github.com/marcjwilliams1/neutralitytestr. In brief, this model proposed by Williams *et al.*[200] predicts that subclonal mutations (with allele frequency < 0.25) follow a 1/f power law distribution (Methods, section 2.15.1). For these analyses, only those mutations with VAF >0.1 were considered, as recommended by the package authors. Subclonal clusters were removed from further analysis when evidence for neutrality was found. Three different metrics (area under the curve, Kolmogorov distance and Euclidean distance) (see Methods, section 2.15.1) were used to fit the allele frequency distribution to a 1/f power law distribution.

4.4 RESULTS

4.4.1 DATA PROCESSING

A mean of 480, 301, 444 (SD of 450, 149 and of 422, respectively) low frequency variants were added to each normal, BPH and tumour sample from all the remaining samples from each patient, respectively (see Figure 4.1, Tables A.1 and A.2, Appendix A). A significantly higher number of low frequency variants were added to patients where three or more normal and tumour samples were taken (P =1.7x10^{-04} in morphologically normal, P = 1.5x10^{-06} in tumour, Wilcoxon rank sum Test).

Variants that were reported to be at the position of SNPs in SNPdb were identified in all samples using tool wANNOVAR (Methods, section 2.16). In order to explore the effects of SNPs on the subclonal reconstruction, the Bayesian Dirichlet process was applied both including and excluding SNPs. For the main results that are outlined in detail in sections 4.4.4 - 4.4.8 we remove the variants at SNPs. An average mean of 266 SNPs per sample in normal and an average of 201 SNPs in tumour tissue were removed (Figure 4.2, Tables A.1 and A.2,

Appendix A). I explore the effect of variants located at SNPs on the phylogenies in section 4.4.9.

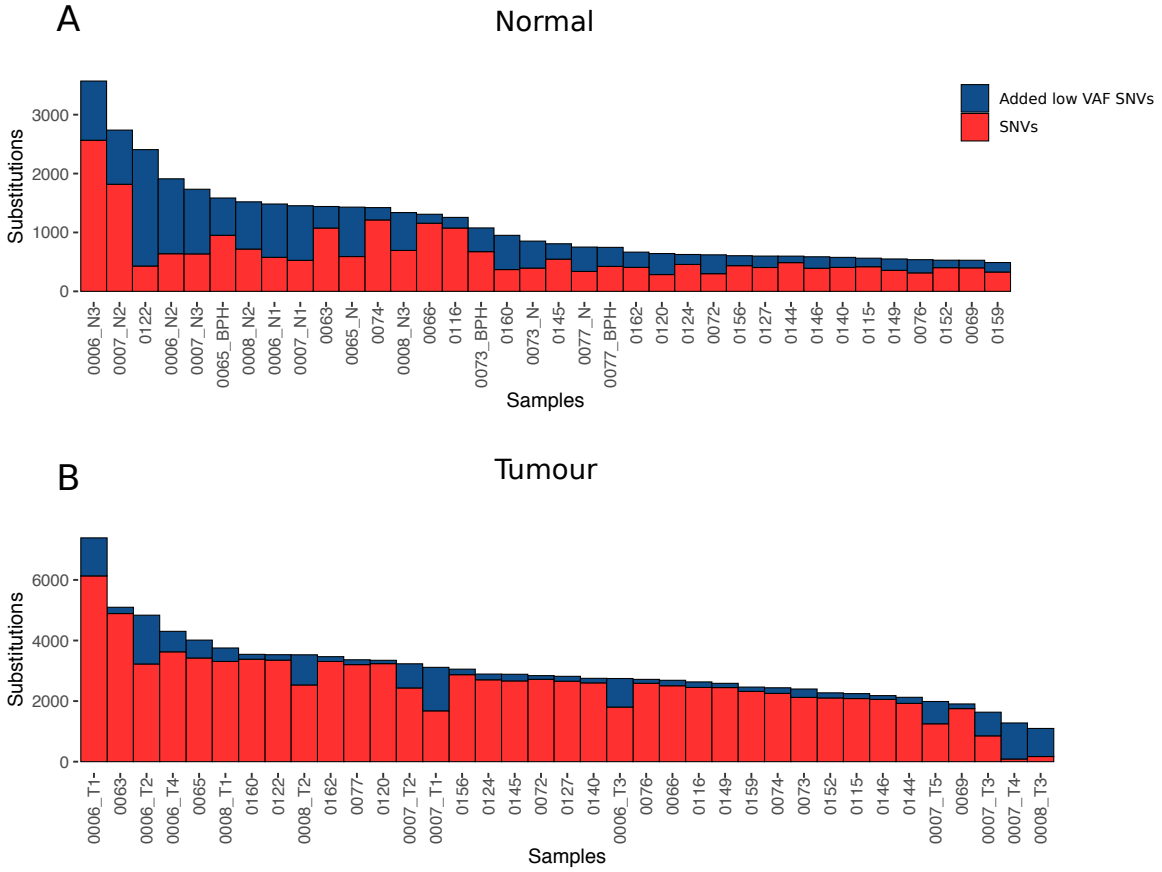


Figure 4.1: Number of SNVs per sample detected by CaVEMan (red) and added low frequency variants found in the paired sample (blue) in normal samples (A) and tumour samples (B).

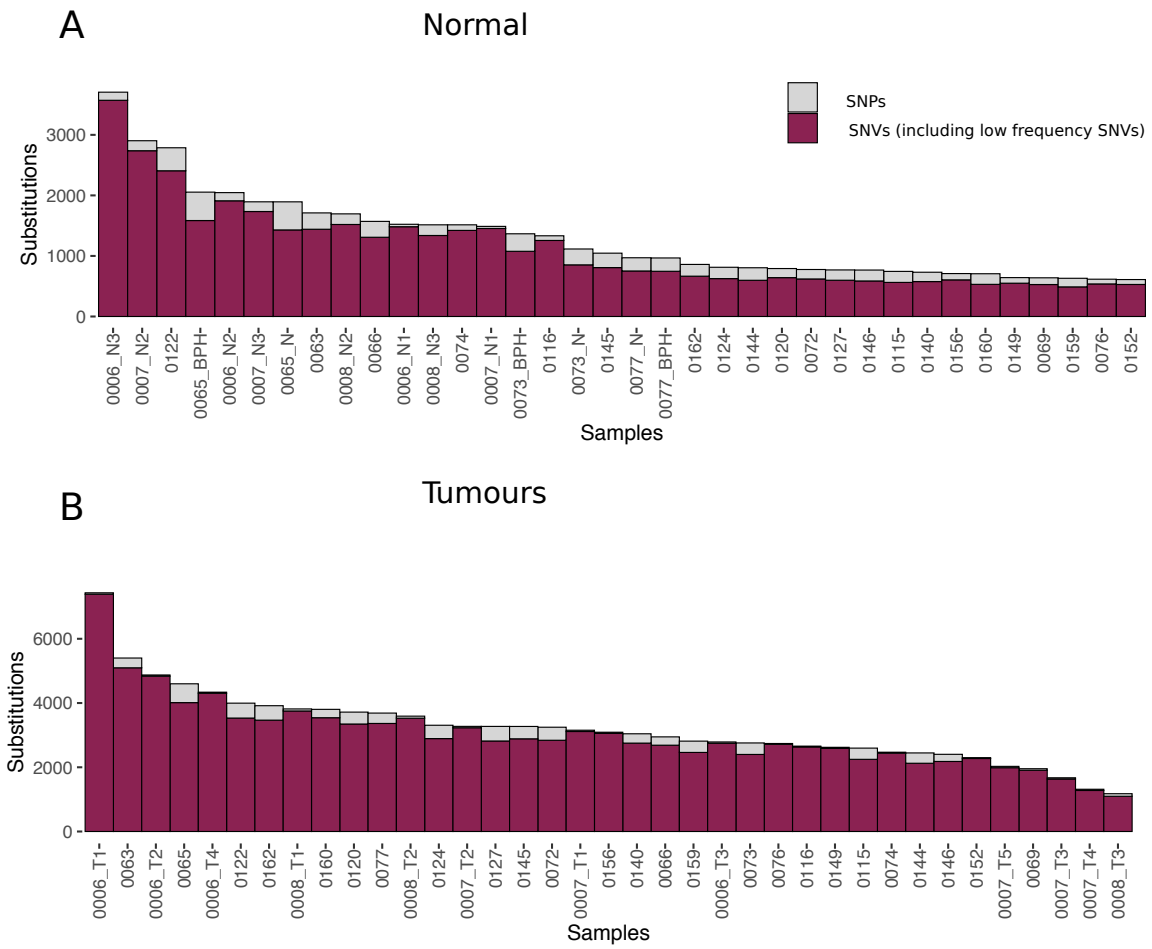


Figure 4.2: Number of removed SNPs per sample (grey) relative to the total number of SNVs (including low frequency SNVs) in normal (A) and tumour samples (B).

4.4.2 NEUTRALITY ANALYSES

A total of 134 clusters (Tables B.1-B.8) associated with clones were identified using the Bayesian Dirichlet process (median of 3 and 95% CI of 2.51, 4.25). Clusters where evidence for neutral evolution was detected were removed: 20 clusters in morphologically normal samples from men with cancer (Tables B.1-B.5, Table B7, Appendix B), 6 clusters from men without prostate cancer (Table B.8, Appendix B) and 4 clusters from fibroblasts samples (B.6, Appendix B). The clone removed was always the one with the lower CCF, as neutral mutations tend to have a VAF around or below 0.25. The VAF distribution and the fitted distribution predicted from a neutral model of evolution is represented for all samples in Figure B.1,

Appendix B. Neutral cases show a distribution that resembles 1/f power law distribution ($P > 0.05$; Area under the curve, Kolmogorov distance and Euclidean distance).

Samples	Evolution	*P* (Area under curve)	*P* (Kolmogorov distance)	*P* (Euclidean distance)	No of clones	Sample type	Gleason	Total SNVs
0063	neutral	0.413	0.359	0.366	0	BPH	3+4	1075
0065_N	non-neutral	0.03	0.06	0.03	3	Normal tissue	3+4	591
0065_BPH	neutral	0.23	0.4	0.21	2	BPH	3+4	952
0066	non-neutral	0.001	0.002	0.001	1	BPH	3+3	1157
0069	neutral	0.249	0.276	0.235	0	BPH	3+4	399
0072	non-neutral	0.008	0.014	0.008	1	BPH	3+3	30
0073_N	non-neutral	0.09	0.18	0.08	3	Normal tissue	3+4	395
0073_BPH	neutral	0.77	0.63	0.68	1	BPH	3+4	674
0074	non-neutral	0.002	0.005	0.002	1	BPH	3+4	1213
0076	non-neutral	0.003	0.007	0.005	2	Normal tissue	3+4	314
0077_N	non-neutral	0.01	0.04	0.01	1	Normal tissue	3+3	338
0077_BPH	non-neutral	0.02	0.01	0.03	1	BPH	3+3	424
0115	non-neutral	0.02	0.02	0.02	1	Normal tissue	3+4	418
0116	neutral	0.16	0.21	0.9	0	BPH	3+4	1075
0120	non-neutral	0.02	0.03	0.03	1	Normal tissue	3+4	285
0122	neutral	0.28	0.37	0.29	0	Normal tissue	3+4	430
0124	neutral	0.09	0.16	0.08	0	Normal tissue	3+4	457
0127	neutral	0.08	0.15	0.07	0	Normal tissue	4+5	405
0140	neutral	0.14	0.17	0.17	0	Normal tissue	4+5	407
0144	neutral	0.25	0.41	0.25	0	Normal tissue	4+3	488
0145	neutral	0.22	0.35	0.24	0	Normal tissue	3+3	547
0146	neutral	0.11	0.19	0.13	1	Normal tissue	3+4	392
0149	non-neutral	0.02	0.02	0.02	1	Normal tissue	3+3	357
0152	neutral	0.05	0.08	0.07	0	Normal tissue	3+3	402
0156	neutral	0.12	0.24	0.09	1	Normal tissue	4+5	436
0159	non-neutral	0.009	0.005	0.01	1	Normal tissue	3+4	328
0160	neutral	0.18	0.12	0.16	0	Normal tissue	4+3	370
0162	non-neutral	0.02	0.02	0.02	1	Normal tissue	3+5	407
0006_N1	non-neutral	0.01	0.02	0.01	1	Normal tissue	3+4	578
0006_N2	neutral	0.16	0.23	0.16	0	Normal tissue	3+4	639
0006_N3	neutral	0.06	0.12	0.05	1	Normal tissue	3+4	2566
0007_N1	neutral	0.11	0.17	0.12	1	Normal tissue	4+3	527
0007_N2	neutral	0.74	0.84	0.91	2	Normal tissue	4+3	1818
0007_N3	neutral	0.51	0.68	0.59	0	Normal tissue	4+3	636
0008_N2	neutral	0.59	0.84	0.65	1	Normal tissue	3+3	718
0008_N3	neutral	0.59	0.72	0.72	0	Normal tissue	3+3	695
0238	neutral	0.32	0.33	0.32	0	Normal tissue (No PC)		140
0239	non-neutral	0.009	0.01	0.008	2	Normal tissue (No PC)		1202
0240	neutral	0.78	0.77	0.9	1	Normal tissue		852
0241	neutral	0.42	0.35	0.44	0	Normal tissue (No PC)		141
0242	neutral	0.18	0.26	0.15	0	Normal tissue (No PC)		125
0243	neutral	0.93	0.51	0.78	0	Normal tissue (No PC)		148
0244	neutral	0.86	0.58	0.7	0	Normal tissue (No PC)		159
0245	neutral	0.24	0.51	0.23	0	Normal tissue (No PC)		104
0246	neutral	0.111	0.076	0.114	0	Cultured Fibroblasts (BPH)		234
0247	non-neutral	0.045	0.043	0.051	1	Cultured Fibroblasts (BPH)		238
0250	neutral	0.134	0.175	0.138	1	Cultured Fibroblasts (BPH)		2431
0251	neutral	0.295	0.505	0.229	3	Cultured Fibroblasts (BPH)		2578
0252	neutral	0.427	0.539	0.469	2	Cultured Fibroblasts (BPH)		1116

Table 4.1: Summary of samples harbouring clonal expansions. Neutrality evolution (green) is suspected when P > 0.05 (Area under the curve; Kolmogorov distance and Euclidean distance).

4.4.3 SUBCLONAL ARCHITECTURE RECONSTRUCTION

After removal of clones with evidence of neutral evolution, a total of 25 clusters under selective pressure were identified in morphologically normal (including BPH) and fibroblasts (Tables B.1-B.6 and Table B.8, Appendix B) and 48 identified in tumour samples (Tables, B.1-B.5, Appendix B). Overall, the total number of clones is 91 (59 in tumour samples and 32 in normal samples), as some clusters were comprised of shared clones between multiple samples for the same patient.

Of the morphologically normal samples (including BPH) harbouring clonal expansions under selective pressure, 79 % had only one subclone, whereas 21 % had two or more (Table 4.1). This is in contrast to tumour samples, where among samples with clonal expansions 12 % of samples harboured one subclone and 88 % harboured 2 or more clones (P = 0.407, Wilcoxon rank sum test, Tables B.1- B.5). 4 out of 5 BPH fibroblasts from normal tissue harboured one clone or more. The total number of clones in normal tissue (including BPH) was not associated with the patient's Gleason score (P = 0.163, Wilcoxon rank sum test, Table 4.1) or the total number of SNVs per sample (P = 0.670, Wilcoxon rank sum test, Table 4.1). Similarly, the number of clones was not significantly different in fibroblasts samples in comparison to normal (including or excluding BPH, P = 0.186 and P = 0.197, Wilcoxon rank sum test, Table 4.1) or between normal and BPH samples (P = 0.828).

The number of samples with subclonal expansions under selective pressure was significantly higher in morphologically normal tissue taken from cancer patients (23/37; 65 %) compared to samples from non-cancer patients (1/7;14 %; $P = 3.5\text{x}10^{-02}$, Fisher exact test, Table 4.1; fibroblasts not included).

The cancer cell fraction (CCF) of the clonal expansions was significantly higher in samples from BPH fibroblasts compared to morphologically normal samples from cancer patients (BPH not included): median of 47 for BPH fibroblasts vs 35 for morphologically normal samples ($P = 4.43\text{x}10^{-02}$, Wilcoxon rank sum test, Figure 4.3). No significant differences were found between the CCFs of morphologically normal and BPH (median of 39) samples (P = 0.252, Wilcoxon rank sum test, Figure 4.3) and BPH fibroblasts and epithelial BPH samples (P =

0.275, Wilcoxon rank sum test, Figure 4.3). However, the CCF was slightly associated with the stromal content (P = 5.82x10^{-02}, F-statistic, Figure 4.4) and the percentage of epithelial content (Figure 4.5) was significantly higher than the CCF for each sample (P = 1.81x10^{-06}, paired T-test).

Among the samples where clonal expansions under selection pressure were detected, no association was observed between CCF and Gleason score (P = 0.151, Wilcoxon rank sum test, Table 4.1). Examples of the posterior distribution of the fraction of cells harbouring a mutation modelled using the Bayesian Dirichlet process are represented in Figure B.2, Appendix B.

Phylogenies from 17 out of 29 men were reconstructed (Figures 4.6, 4.7) and described in detail in the next sections (4.4.4 - 4.4.6). Only men where there was evidence of positive selection in at least one normal sample were reconstructed. Subclones in three patients (0065, 0073, 0077) could have been positioned differently in the phylogeny and equally agree with the data. In the following description I use the lineages where they have been placed in a linear configuration. The alternative lineages are illustrated in Figure B.4, Appendix B. The subclonal architecture identified with the Bayesian Dirichlet process was always supported by the shared indels identified by Pindel.

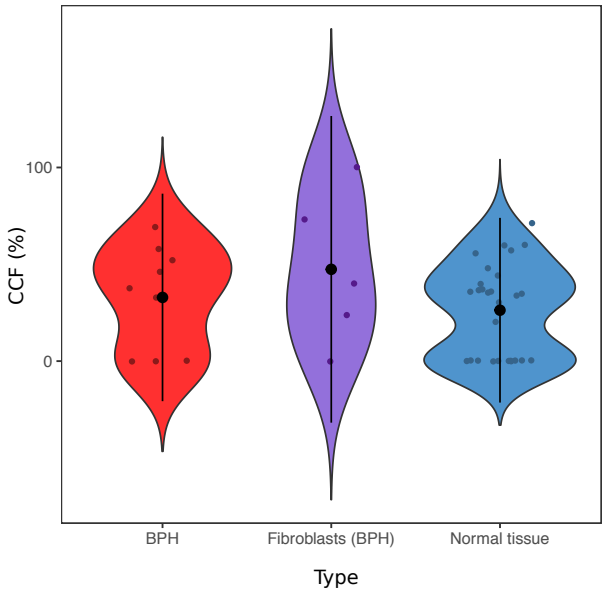


Figure 4.3: Boxplots showing the relationship between the cellular cell fraction (CCF) and the type of normal samples from prostate cancer patients (normal, normal with BPH and BPH fibroblasts).

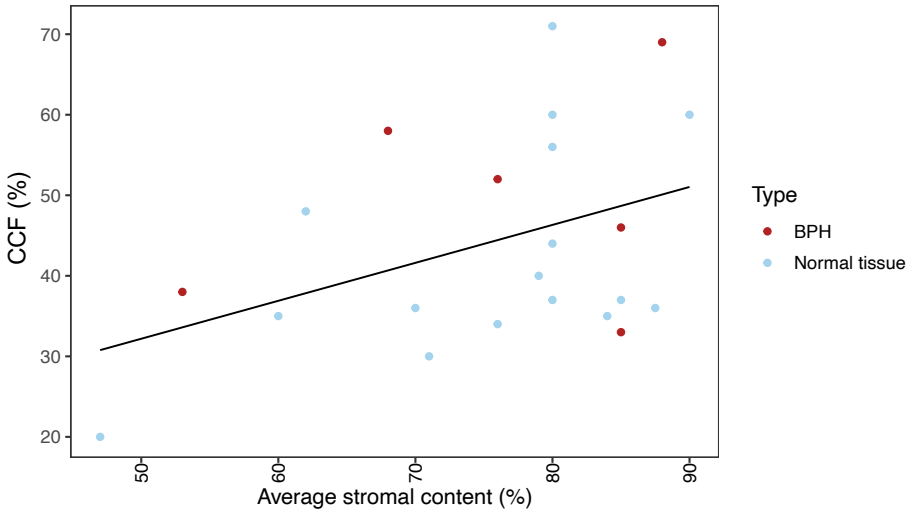


Figure 4.4: Relationship between the average stromal content and the CCF for each morphologically normal sample from men with prostate cancer ($P = 5.82\text{x}10^{-02}$, F-statistic).

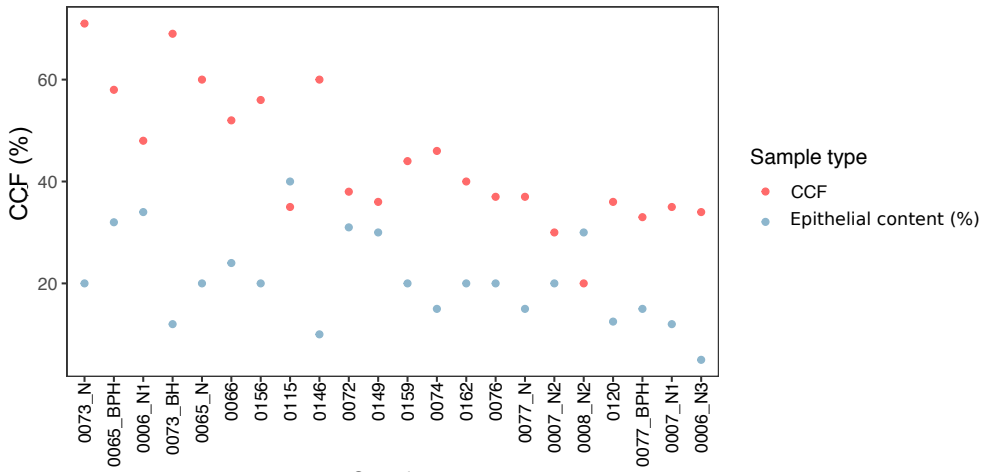


Figure 4.5: Comparison between the CCF and the epithelial content for each morphologically normal sample from men with prostate cancer. The CCF is significantly higher than the epithelial content, $P = 1.81\text{x}10^{-06}$, paired T-test.

4.4.4 COMPLEX MEN

Phylogenies for these patients have been previously reconstructed using fewer normal samples[17]. A total of 64, 62 and 45 substitutions supporting the detected clusters had been previously validated[17] by PCR/deep sequencing in all tumour samples from patients 0006, 0007 and 0008, respectively. In morphologically normal, 1 substitution for patient 0006 and 4 substitutions for patient 0007 supporting the detected clusters had been previously validated[17]. The reconstruction of the subclonal architecture performed for this thesis resembles very accurately the ones depicted by Cooper *et al*[17]. for those samples already analysed (Figure 4.6 vs Figure B.3, Appendix B). The slight differences (described in the next sections) are due to excluding those tumour samples where there was neutral evolution was suspected and a more stringent filtering of clusters was performed: only those clusters that were supported by 1 % of the total number of substitutions for each patient were considered. All three phylogenies (Figure 4.6) are characterised by harbouring one or more subclonal clusters in at least one morphologically normal sample. No shared clones are found between normal and tumour samples.

4.4.4.1 PATIENT 0006

A total of 11 clusters (Table B.1, Appendix B) were detected for this patient and 2 of them were found in morphologically normal samples (Figure 4.6 A). In comparison to the phylogeny presented by Cooper *et al*.[17] (Figure B.3, Appendix B), we note the absence of shared clones T1/T2/T3/T4/N1 and T1/T2/T3, as it was only supported by 1 SNV and one SV. These clones have not been considered in this representation due to a more stringent filtering.

In the normal tissues, one cluster was detected in samples N1 and N2 and two in sample N3. However, there was evidence of neutral evolution in normal sample N2 and N3, so these were not considered in the phylogeny. Mutations with a potential functional significance occur in genes *RYR3* and *MEPE* and were assigned to the subclonal expansion detected in N1.

There are three independent clones from samples T2 (CCF=100 %), T3 (CCF=100 %) and T4 (CCF=100 %), with an extra T4 subclone in 46% of cells. Shared clones are found between T1/T2 and T1/T4 (CCF =100 %). T1 appears to be a mixture of the T2 and T4 lineages. A unique subclone is detected in sample T1, but as there are two lineages found for T1 (T1/T2

and T1/T4), it is not possible to know whether this clone (93% of cells) emerges from T1/T2 or T1/T4, or both.

4.4.4.2 PATIENT 0007

We detect 11 clusters in this patient (Table B.2, Appendix B), with 2 of them present in normal samples (Figure 4.6 A). In this phylogeny shared clones T1/T2/N1 and T3/T4/T5detected by Cooper *et al.*[17] (Figure B.3, Appendix B) are not included because the SNVs supporting those clusters (10 and 3, respectively) do not pass the new filtering criteria. One shared subclone between samples N1 and N2 is reported in 35 % and 33% of cells, respectively. Interestingly, the shared subclone N1/N2 harbours mutations with a potential functional significance (see Chapter 3, section 3.5.4) affecting coding genes *ADAM28, BCAT1* and *FAT2*. Another subclone is observed in 22 % of cells in sample N2. This subclone (represented linearly in Figure 4.6 A) could have also been positioned in a parallel fashion (Figure B.4 A, Appendix B). An independent clone in sample T3 and shared lineages between T1/T2 (100%/56%) and T4/T5 (100%/100%) are reported in this patient. Two extra subclones from T5 (73% and 62%) and T1 (100% and 100%) and T2 (81%) have evolved in a linear fashion from shared lineage T4/T5 and T1/T2, respectively.

4.4.4.3 PATIENT 0008

We identified 9 clusters (Table B.3, Appendix B): 7 were present in tumour samples and 1 was present in normal sample N2 (Figure 4.6 A). In contrast to the phylogeny reconstructed by Cooper *et al*.[17] (Figure B.3, Appendix B), shared clone T1/T2/T3/N1 and the independent clone in N1 have not been included because they failed to meet the new filtering criteria. Specifically, sample N1 was not included in the subclonal reconstruction analysis because of the low number of SNVs for this sample (28). This phylogeny is characterised by an independent subclone in N2 (20%) and a T1/T2/T3 lineage comprised of 2 shared clones (100%/100%/13%) and (98%/96%6%). Two shared clones are also observed between T1/T2 (100%/100% and 93%/55%, respectively) that evolve in two unique subclones in T1 (51% and 43%) and unique subclone in T1 (56%). Cluster 12 (Table B.3, Appendix B) could not be placed in the phylogeny according to the pigeonhole principle. This cluster shows mutations in a small fraction of the tumour samples (6% in T1, 7% in T2 and 22% in T3) that are shared with normal samples N2 and N3 and are supported by 100 SNVs and one indel. According to the pigeonhole principle a shared T1/T2/T3/N2/N3 should be the first common ancestor, which means that any clone evolving from that cluster would have a CCF smaller than the ancestor. As there is a shared

cluster in 100% of cells in T1/T2 (supported by 1526 of SNVs, 123 indels and 25 SVs) it is impossible to include both clusters 12 and cluster 1 in this phylogeny. Given the stronger evidence of overall mutations supporting shared clone T1/T2 and its previous validation by Cooper *et al*[17]., we included cluster 1 and discarded cluser12.

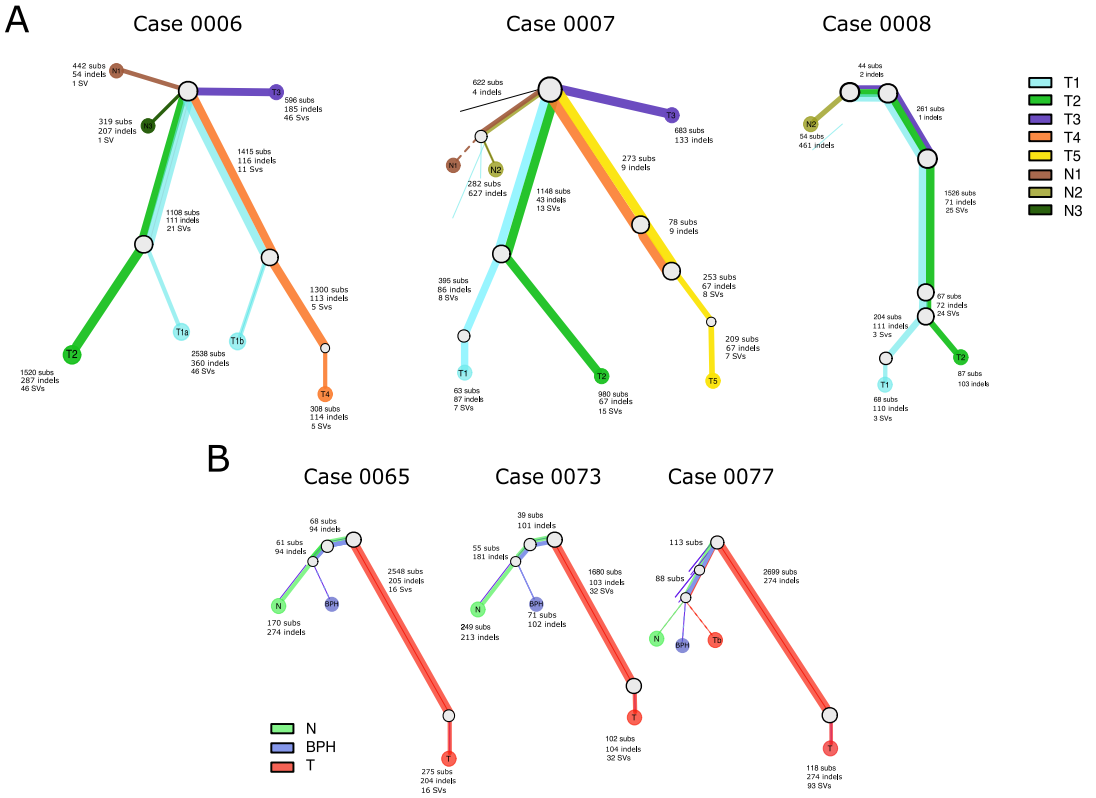


Figure 4.6: Subclonal architecture of patients with multiple samples: (A-B) Phylogenies revealing the relationships between sample clones for each case. Each coloured line represents an independent clone/subclone in a sample. When two or more coloured lines are together they represent one clone that is in all samples represented. The length of the line is proportional to the weighted number of substitutions present in each clone, the thickness shows the cell fraction associated with that clone (Also shown in Tables B.1-B.4, Appendix B). For example, case 0077 contains a shared subclone with 8% N, 33% BPH and 2% T supported by 113 substitutions and 4 indels. Dotted lines are associated with samples that have no evidence of a unique sample clone: it represents a possible clone evolving from the shared clones. The very low fraction tumour subclone (< 4%) shared with normal and BPH tissue in case 0077 and between normal and tumour in case 0072 suggests cancer tissue contained some of the N/BPH cells.

4.4.5 PATIENTS WITH BOTH BPH AND NORMAL SAMPLES

Three cases had samples in normal, BPH and tumour tissue. In the majority of cancers, the morphologically normal tissue and cancer had distinct lineages (as illustrated by the Complex Men and cases 0065 and 0073; Figure 4.6 A and B). 5 and 6 clusters were detected for cases 0065 and 0073 (Table B.4, Appendix B), respectively. An almost identical subclonal architecture for these two patients (Figure 4.6 B) is observed, with an independent clonal cluster in the tumour and a subclonal cluster in the normal sample (N). Notably, BPH and morphologically normal tissue had a shared lineage: two shared subclones in ~ 60 % and ~40 % of cells in case 0065 and two shared subclones in ~70 % and ~50 % of cells in case 0073. In both cases these subclones have been represented linearly (one subclone is the descendant of the other) in Figure 4.6 B but they could have also evolved in parallel, as the sum of their CCF <100. The alternate configuration of these phylogenies is illustrated in Figure B.4 B (Appendix B). A unique BPH subclone with evidence of selection (CCF of 64%) is detected only in sample 0073. Mutations with a potential functional significance occur in gene *ANKRD20A2* in the unique N subclone.

Patient 0077 harboured 4 clusters (Table 4.6) and had a similar configuration to the previous patients (0065 and 0073) with one exception: this phylogeny presents two subclones with 2% contribution in the tumour targeted sample, 8-38% in the morphologically normal sample and 33% in the BPH sample, consistent with a model in which the tumour targeted sample contains some of the N/BPH subclone. In this phylogeny, a shared N/BPH/T cluster should be the first common ancestor. However, according to the pigeonhole technique a cluster that evolves from an ancestor should have a CCF equal or smaller than the ancestor. Here the shared N/BPH/T clusters have mutations from tumour in 2 % of cells, and a fully clonal tumour cluster could not have emerged from this cluster. As the cluster in the tumour sample is fully clonal, there is no possibility of another tumour clone evolving in parallel to the main tumour clone. The only explanation is that the tumour sample harbours a small percentage of normal/BPH cells. Regarding the normal and BPH contributions, they could have emerged in a parallel fashion (alternate configuration, Figure B.4 B, Appendix B) or linearly (Figure 4.6 B) for the same reasons discussed in case 0006.

4.4.6 PATIENTS WITH ONE NORMAL AND ONE TUMOUR

Phylogenies for 11 patients with only morphologically normal and tumour samples were constructed. Clusters detected for these 11 cases ranged between 2 and 3 (see Table B.5 and Figure 4.7). There was a fully clonal cluster in the tumour in all cases. An extra tumour subclone emerging from the main tumour clone was present in 8 cases (all except for cases 0076, 0149, and 0156). At least one unique normal sample with evidence of clonal expansion was present in these 11 patients at levels of 27-60 % of cells (Table B.5, Appendix B). 7 samples (cases 0072, 0076, 0120, 0146, 0156, 0159, 0162, Figure 4.7, Table B.5, Appendix B) showed evidence of the subclonal normal cells being present in the cancer tissue at low cellular fraction (< 13%; median of 3, IQR of 2), a scenario already described in patient 0077 for clusters 3 and 5 of that phylogeny. Coding mutations with a potential functional significance were found in 3 genes (*RDH10, SOBP* and *MAP3K4*) in the normal subclone for patient 0066, 7 genes (*PHF12, FOXJ3, L1TD1, NPFFR1, COL6A1, ZNF687* and *UNC80*) in the normal subclone for patient 0074 and in one gene (*TIE1*) for patient 0162.

4.4.7 NON-CANCER PATIENTS

Subclones were detected in all cases with a CCF that ranged from 24% -32% (median of 31) but were not considered because of evidence of neutral evolution except in case 039 (Table B.8, Appendix B). This case presented two subclones (31 % and 47 % of cells, respectively) (Figure 4.8). Interestingly, this patient is characterised by having BPH. The number of substitutions in non-cancer samples is considered low for clonal expansion detection in all cases (~140 substitutions) except patient 239 (1202 substitutions).

4.4.8 FIBROBLASTS

Clonal expansions were also detected in 4/5 fibroblasts samples (0247, 0250, 0251 and 0252) (Figure 4.8). A single subclonal cluster was detected in sample 247 cluster with a CCF of 25 % supported by 137 substitutions. Sample 0250 harboured one cluster in 40 % of cells and samples 0251 and 0252 presented 3 and 2 clusters each in 54 %, 86% and 100% of cells and 40 % and 77 % of cells, respectively (see Table B.6, Appendix B).

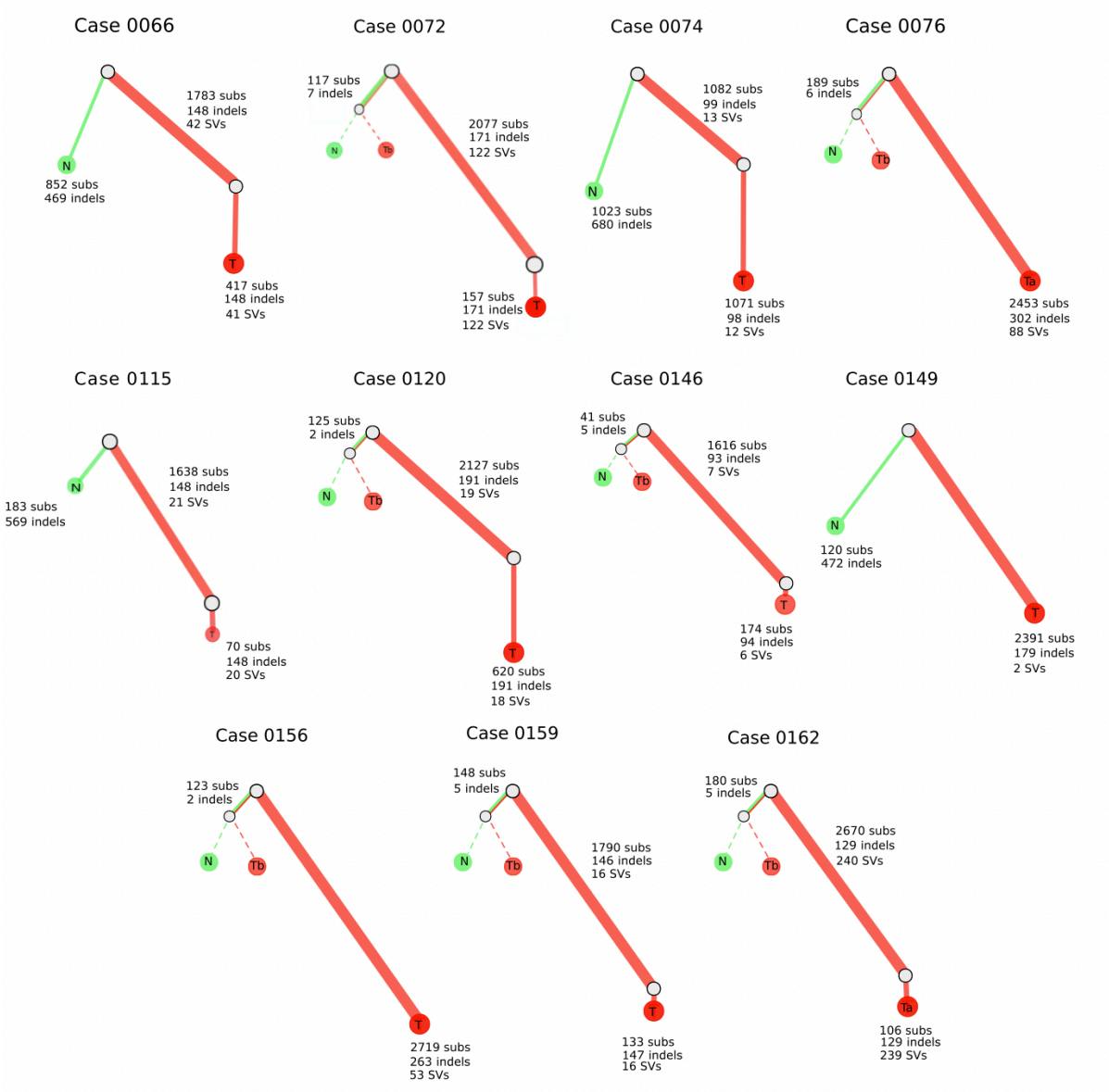


Figure 4.7: Subclonal architecture of patients with morphologically normal and matched tumour (N-T). Phylogenies revealing the relationships between sample clones for each case. Each coloured line represents an independent clone/subclone in a sample. When two or more coloured lines are together they represent one clone that is in all samples represented. The length of the line is proportional to the weighted number of substitutions present in each clone, the thickness shows the cell fraction associated with that clone. For example, case 0120 contains a shared subclone with 36% N and 2% T supported by 125 substitutions and 7 indels. Dotted lines are associated with samples that have no evidence of a unique sample specific clone. The very low fraction tumour subclone (2- 13 %) shared with normal and T tissue in all three case suggests cancer tissue contained some of the N/BPH cells.

We found a median of 9 coding mutations supporting all the fibroblast clusters (see Table B.6, Appendix B). 18 of them had a potential functional significance (see Chapter 3, section 3.5.4). Interestingly, among these we find the cancer related genes *FAT1, POLE,* and *ACR* (reported in the cancer gene census[213]) and genes (*DSG3, RPL18, KIAA1217* and *DHX32)*, which have been classified by The Cancer Genome Atlas Research Network (TCGA) as a prognostic marker for many cancers (see Chapter 3, section 3.5.4). Mutations in *DSG3* and *POLE* occur simultaneously in cluster 4 (CCF of 74 %) from patient 0252, *KIAA1217* and *RPL18* are mutated in cluster 4 (CCF of 100 %) from patient 0251, *DHX32* and *ACR* in cluster 2 (CCF of 40 %) from patient 0250 and mutations in *FAT1* correspond to cluster 2 (CCF of 54 %) in patient 0251.

4.4.9 EFFECT OF SNPs IN THE SUBCLONAL ARCHITECTURE

Clusters were also detected when variants at SNP locations were included (see Tables B.9 to B.14, Appendix B). Overall, all cases were very similar to the ones described in sections 4.4.4 - 4.4.6 (Tables B.1 to B.5 and Table B.7), with the main lineages being the same. However, we observe that SNPs affect those N/T subclones where we detected a small contribution (< 13%) in the tumour sample. There are three instances (patients 0069, 0140 and 0149) where a NT subclone is detected only when variants at SNPs were included. For the remaining NT clusters, we observe a small decrease of the tumour contribution for these subclones, although is not significant ($P = 1.07\text{x}10^{-01}$, Wilcoxon rank sum test, Figure 4.9).

4.4.10 SAMPLE MAPPING

The distance (in mm) between normal and tumour samples, multiple normal and multiple tumour samples was taken (Table A.1, Appendix A) and associations between the distance between the different samples and the subclonal architecture were explored. The minimum distance (mm) between cancer and normal samples for the samples (median of 19) with independent lineages was greater compared to samples with normal infiltration in the tumour sample (median of 7.1) (Figure 4.10), but there was no statistical significance ($P = 1.8\text{x}10^{-01}$, Wilcoxon rank sum test). Similarly, there was no significant association between the proximity of the normal sample to the matched tumour and whether the cancer was multifocal ($P = 0.852$,

Wilcoxon rank sum test, Figure 4.11, Table A.1, Appendix A) or whether the clones were multiple or single ($P = 0.307$, Wilcoxon rank sum test, Figure 4.12).

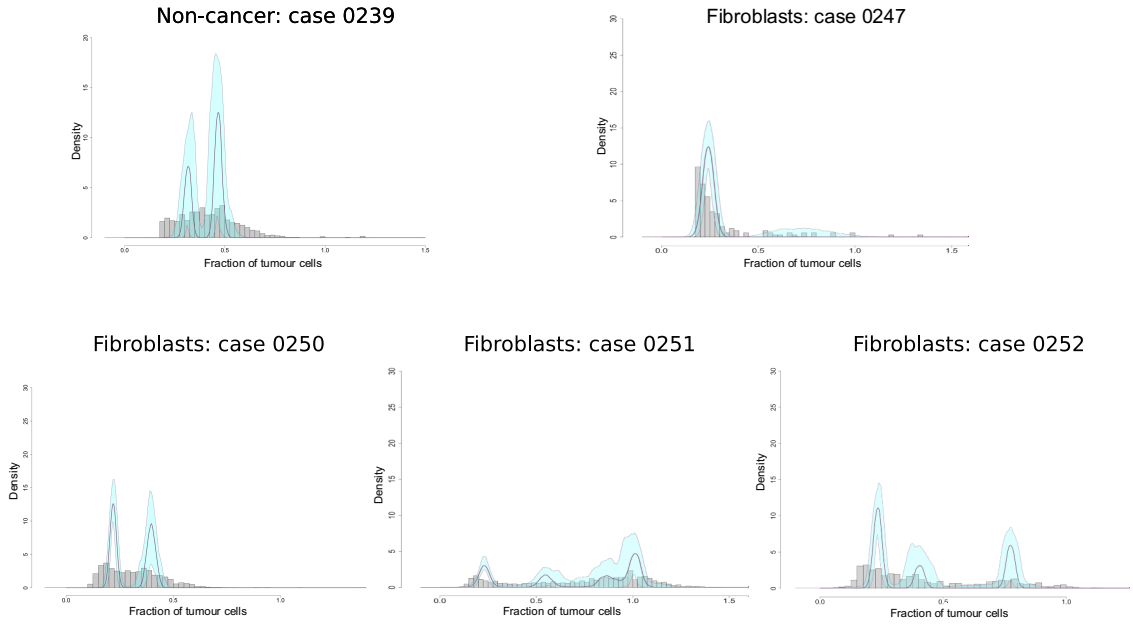


Figure 4.8: Example density plots of cell cultured fibroblasts and morphologically normal samples from patients where phylogenies could not be reconstructed. They show the posterior distribution of the fraction of cells bearing a mutation, modelled by a one-dimensional Bayesian Dirichlet process. The median density is indicated by the purple line and 95% confidence intervals by the blue region. The grey histogram shows the observed frequency density of mutations as a function of the fraction of cells bearing the mutation.

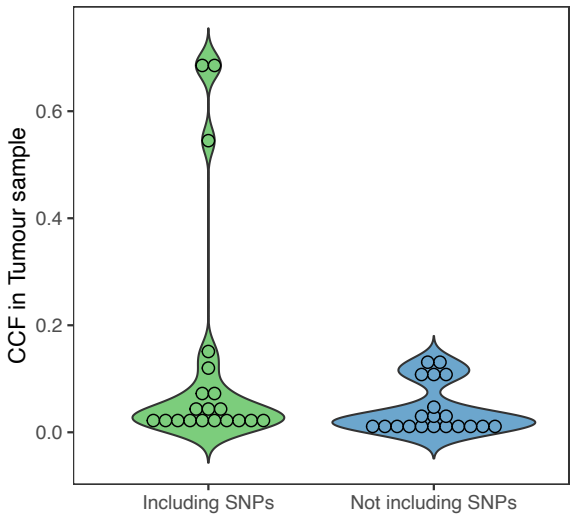


Figure 4.9: Violin plots showing the relationship between tumour infiltration and the inclusion of SNPs for the clonal expansion detection analyses. No association was found (P = 1.07x10^{-01}, Wilcoxon rank sum test).

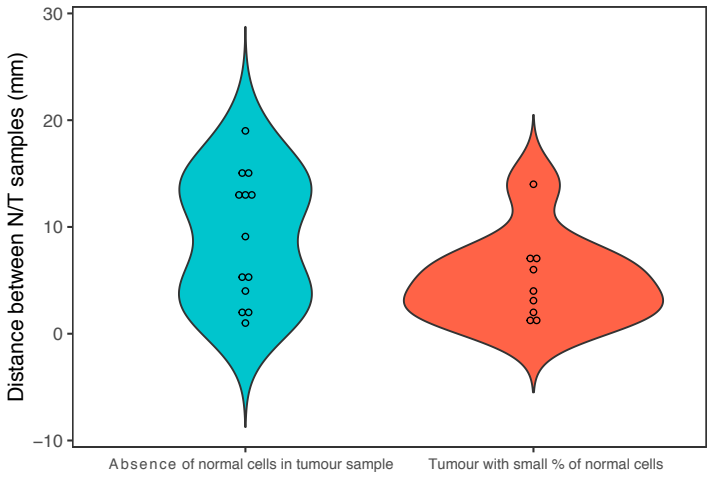


Figure 4.10: Violin plots showing the relationship between normal infiltration of the tumour sample and the distance in mm between normal samples with subclonal expansions and the matched tumour sample. No association was found (P = 1.8x10^{-01}, Wilcoxon rank sum test).

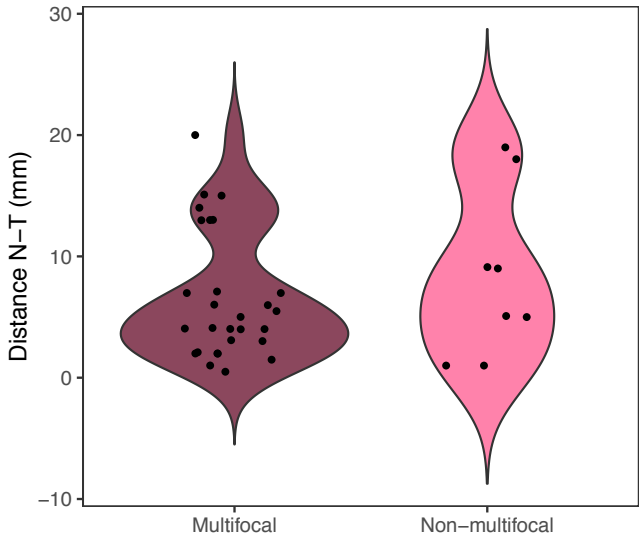


Figure 4.11: Violin plots showing the relationship between multifocality and the distance in mm between normal samples and the matched tumour sample. No association was found (P = 0.852, Wilcoxon rank sum test).

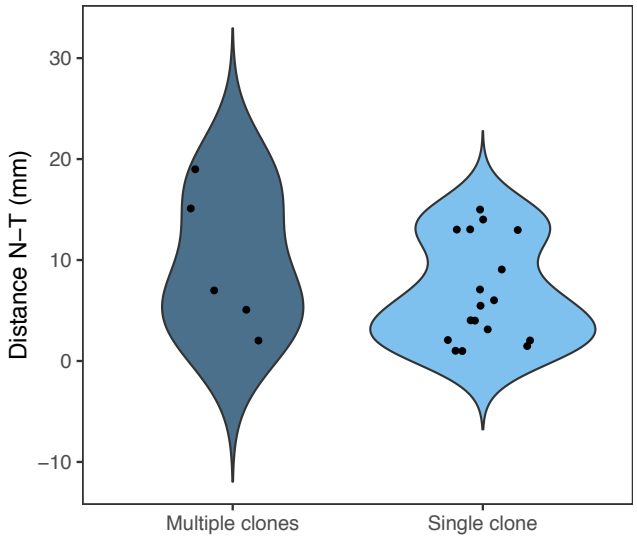


Figure 4.12: Violin plots showing the relationship between the number of clonal expansions and the distance in mm between normal samples and the matched tumour sample. Only samples with at least one clonal expansion were included in the analysis. No association was found (P = 0.307, Wilcoxon rank sum test).

4.5 DISCUSSION

In this chapter we have characterised the subclonal architecture of both tumour and morphologically normal tissue for patients with prostate cancer, morphologically normal tissue in men without cancer and cell cultured fibroblasts from normal tissue with BPH. Our results demonstrate that the majority of patients harbour subclonal expansions under selective pressure in morphologically normal samples (including those from BPH), in contrast to samples from men lacking cancer (1/7). In some cases, multiple subclones were detected (0006, 0065, 0073, 0239) which shows that there may be more than one independent expansion in a single prostate. No link was observed between the number of subclones and clinical parameters. These observations suggest that the presence of the subclonal expansions could be linked to the field effect and the development of cancer. Although the implications of the presence of clonal expansions in normal tissues for cancer development are still being debated, there is evidence that shows the presence of multiple clones undergoing selection that carry driver genes in the normal human eyelid and esophagus[118,162]. Their presence could potentially lead to cancer development following a multistage model of carcinogenesis[254]. This model postulates that the accumulation of a specific number of sequential genetic changes in driver genes precedes cancer initiation. Although a few protein coding mutations have been found supporting clonal expansions, in this study we could not confirm the presence of positively selected genes in normal tissue due to the low number of coding mutations detected. This scenario would be in agreement with the notable absence of clonal expansions under selective pressure and significantly lower number of overall genetic alterations (see chapter 3, section 3.5.3) in morphologically normal tissue from men without cancer. The prostates from men with cancer have undergone notable changes as a whole. These findings are slightly different to those reported by Martincorena *et al.*[118] in skin and esophagus where the clonal expansions were detected in normal tissue from patients without cancer. However, the skin and esophagus are tissues with an incredibly high proliferation rate in comparison to prostate and they are also exposed to known mutagens (ultraviolet light and diet).

The phylogenies also reveal that subclones in morphologically normal samples are distinct from those in the tumour, but morphologically normal and BPH samples from the same patient on the other hand often have a shared lineage (Figure 2b).

It has been already discussed in chapter 3 (section 3.6) that the higher proliferation rate in BPH may be contributing to a heavier mutation burden and occurrence of clonal expansions. This is further supported by the finding of two subclonal expansions under selection in the BPH non-cancer sample 0239. An association between BPH and cancer development has been shown in several studies, although the link remains controversial. Similarly, fibroblasts from BPH samples showed evidence of clonal expansions in four samples. Substitutions were present in at least 25%, 40%, 100% and 77% of cells in cases 0247, 0250, 0251 and 0252, respectively. Interestingly, these features have been also associated with cancer-associated fibroblasts or CAFs[255]. Whether these subclones are driving the BPH disease process alone or contributing to cancer development needs further examination using other approaches such as single-cell sequencing. A previous study that explored the spatial competition of clonal expansions in normal esophagus has revealed that the multiple clones that have similar fitness or are evolving neutrally can coexist in normal tissue while tissue homeostasis is maintained. When a clone is surrounded by other clones with similar fitness, the growth rate of each clone population is diminished. The genotype of neighbouring cells is as important as the acquisition of advantageous mutations. Therefore, the higher mutation burden and clonal expansions observed in normal tissues from prostate cancer patients could increase the risk of cancer development by the acquisition of driver mutations over time, but it is also dependent on the discrepancy in fitness among the surrounding cells.

Overall, these results further support the hypothesis of a field cancerization in the prostate and provide key insights into genomic evolution of prostate cancer at very early stages of development.

CHAPTER 5 : PATCHWORK EXPERIMENT

5.1 SUMMARY

Prostate cancer is a highly heterogeneous disease that is multifocal in nature. This genomic diversity poses a great challenge in order to characterise the mutational landscape in cancer and detect driver genes. In this chapter we describe "The Patchwork experiment", an approach that allows us to explore the presence of a field effect in the prostate by targeted sequencing a large number of samples (n = 95) from morphologically normal and tumour tissue from one single prostate patient. This patient had been previously WGS sequenced before for previous analyses presented in chapters 3 and 4. We detected 21 recurrently mutated genes and clonal expansions affecting 10 genes and one driver gene under positive selection.

5.2 MATERIALS

5.2.1 SAMPLES

We analysed 96 samples (summarized in Figure 5.1) collected from ICGC prostate patient number 0007 (Methods, section 2.4) obtained after prostatectomy at Addenbrooke's hospital: 77 samples from morphologically normal tissue, 18 from tumour tissue and one control from blood. Ethical approval was obtained from the NHS East of England-Cambridge REC [03/018]. Samples were collected subject to ICGC standards of ethical consent (https://icgc.org/).

5.3 METHODS

5.3.1 DATA COLLECTION AND QUALITY CONTROL

95 punches of 1 mm^3 were taken from three different 5 mm FFPE block slices (Methods, section 2.4). A previously collected blood sample for the WGS experiment was used as control. I extracted the DNA, constructed the libraries and performed targeted sequencing using the Illumina Nextseq sequencing system at the Quadram Institute (Methods, sections 2.5.2). The tool FASTQC (Methods, section 2.6) was used to assess the quality of FASTQ files.

5.3.2 PRE-VARIANT CALLING PROCESSING: ALIGNMENT AND DUPLICATE REMOVAL

Reads were mapped to the GRCh37 reference human genome using the Burrows-Wheeler Aligner's maximum exact matches (BWA-MEM)[177] algorithm. This algorithm is similar to BWA-SW but faster and more accurate. SAM files from the two sequencing runs were converted to a BAM format and merged using Samtools[179]. However, the sequencing metrics of the first sequencing run were calculated and are depicted in Figure C.1 (Appendix C). Picard tools (Methods, section 2.8) option was used to assess the quality of the alignment (option *CollectAlignmentSummaryMetrics*), remove duplicated reads (option *MarkDuplicates*) and calculate coverage of the region of interest (option *CollectHsMetrics*).

5.3.3 VARIANT CALLING AND POSITIVE SELECTION ANALYSIS

Substitutions were called for each sample against a deep sequenced matched blood control using a likelihood ratio test from the R package deepSNV (Methods, section 2.9.2).
Positive selection in coding variants was detected using the package R dNdScv (https://github.com/im3sanger/dndscv). dN/dS ratios were quantified for missense, nonsense and essential splice mutations. This is a previously described dN/dS method developed by Martincorena *et al.*[118] (see Methods, section 2.15.2) that was also used in Chapter 3, section 3.5.4.

5.4 RESULTS

5.4.1 DNA YIELD AND LIBRARY METRICS

The mean yield of the total DNA obtained from each punch was significantly higher in tumour (mean of 392 ng) than in morphologically normal (mean of 181 ng) samples ($P = 2.16\text{x}10^{-4,}$ Wilcoxon rank sum test, Figure 5.1). A median of 1730 pg/ul (IQR of 575) DNA post-capture was obtained after library preparation, with a percentage on target (DNA molecules that were successfully captured by the hybridization process to the designed RNA probes) that ranged from 84 % to 97 % (median of 94). The median library size was 328 bp (Figure 5.2).

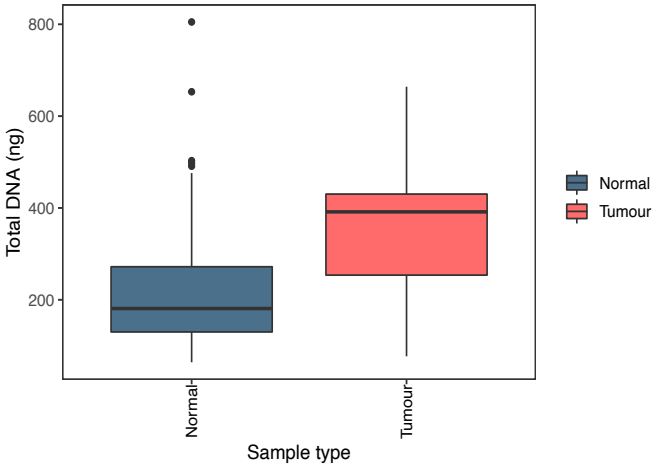


Figure 5.1: Total DNA yield (ng) in morphologically normal and tumour samples.

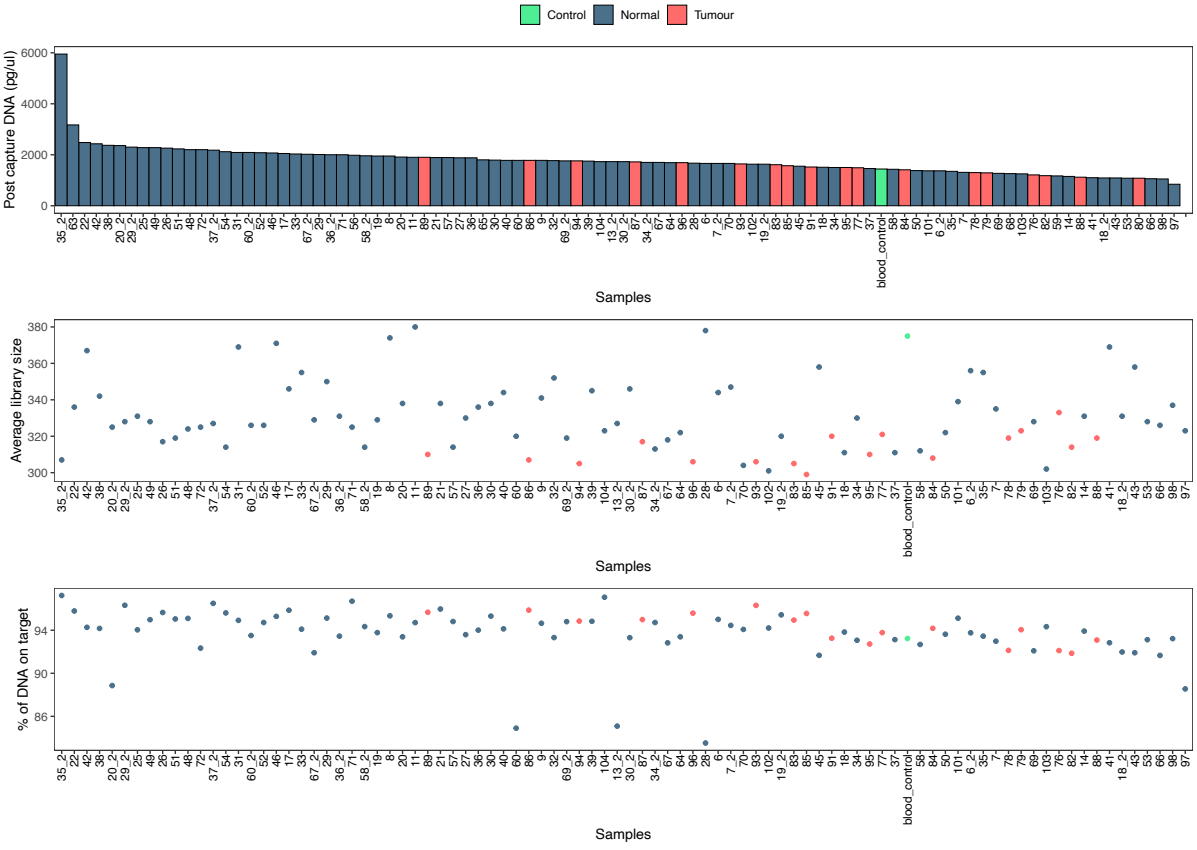


Figure 5.2: Library metrics for each sample. (A) Post-capture DNA (pg/ ul), (B) library size (bp) and (C) percentage of hybridized DNA within the targeted region of interest.

5.4.2 QUALITY CONTROL

5.4.2.1 FASTQC REPORT

The FASTQC report indicates a very good quality of the reads: per base and sequence quality average was > 30, all bases were proportionately distributed. Adapters were found in the first round of sequencing and were removed. Estimated PCR duplication levels were very high: unique reads reported in normal and tumour ranged from 15% - 84% and 22% - 38%, respectively (Figure 5.4). Regarding GC content we found that a proportion of the reads do not follow the expected normal distribution (the sum of the deviations from the normal distribution represented more than 30% of the reads). An example of some of these metrics is depicted in Figure 5.3.

5.4.2.2 SEQUENCING METRICS

A mean on target coverage (after removing duplicates and off target capture reads) of 331X was achieved after two sequencing runs for normal samples (maximum of 648X and minimum of 83X, IQR of 124) and of 398X for tumours (maximum of 484x and minimum of 152X, IQR of 122). The blood control sample had a mean coverage was 878X. The alignment to the reference genome was of good quality, with a percentage of mapped reads that ranged from 91% - 99% in normal samples (median of 98%) and 85% - 99% in tumours (median of 98%). Overall, a median of 38 % of reads were classified as off target capture and were discarded in normal and tumour samples (ranging from 16 – 86 in normal and of 29 to 59 in tumour). All these metrics are represented in Figure 5.4. Per base coverage across the total target region (protein coding regions of 98 genes) is shown in Figure C.2, Appendix C. Coverage for each gene across all samples ranged from a median of 133X in SMARCA1 (the gene with the lowest median coverage) and 1196 in MUC3A (the gene with the highest median coverage). The distribution of mutations detected across genes had a weak correlation with coverage (Spearman's $\rho = 0.39$, $P = 5.6 \times 10^{-5}$) (Figure 5.5).

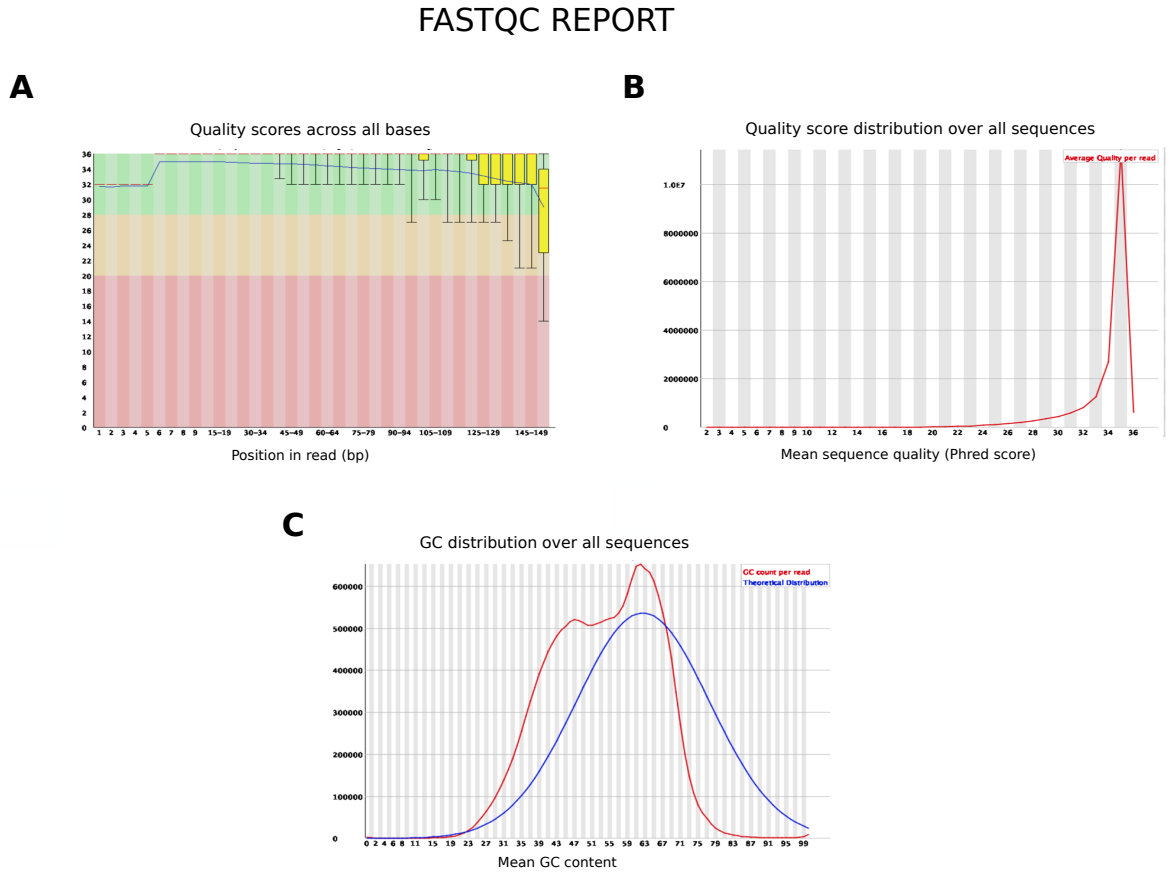


Figure 5.3: Quality metrics for the 96 samples of the Patchwork experiment. (A) Quality values across all bases of all sequences. Mean quality, median value and inter-quartile range are represented by the blue line, the central red line and the yellow box, respectively. Points above 90% and below 10% are indicated by the lower whiskers. (B) Per sequence quality score. The y-axis represents the number of sequences. (C) GC content across the length of each sequence in comparison to a modelled normal distribution of GC content. The y-axis represents the number of sequences.

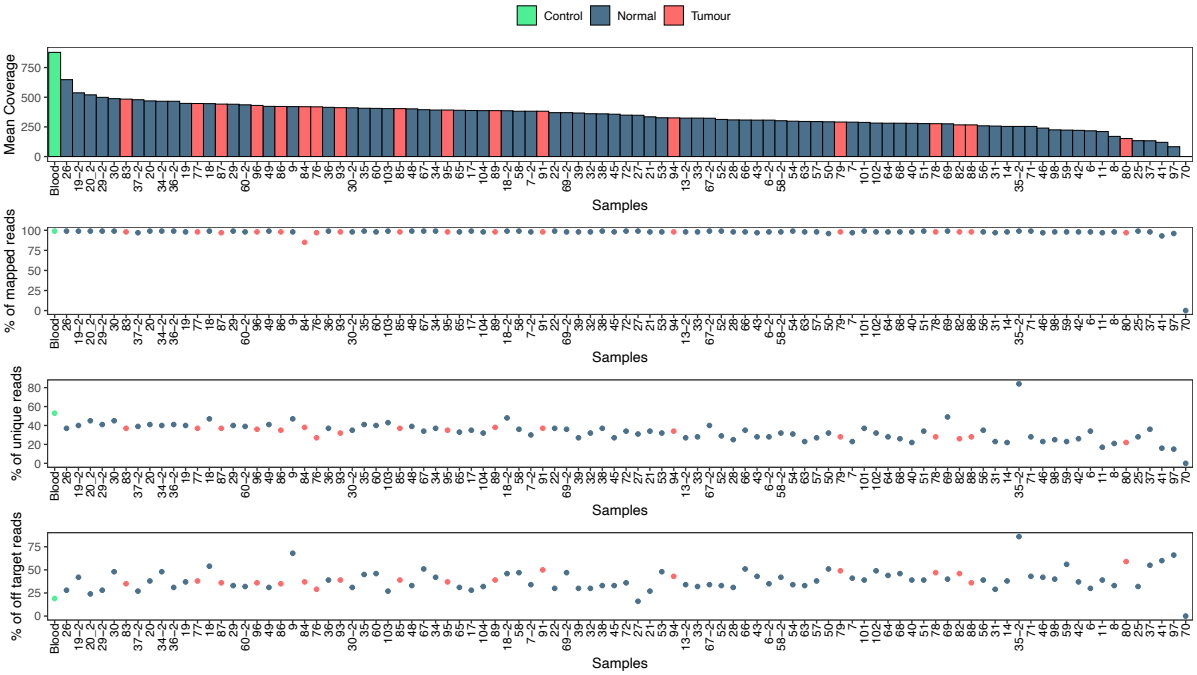


Figure 5.4: Coverage and Alignment metrics: The percentage of mapped reads represents the reads that aligned successfully to the reference genome. The percentage of unique reads show moderate to high levels of duplication.

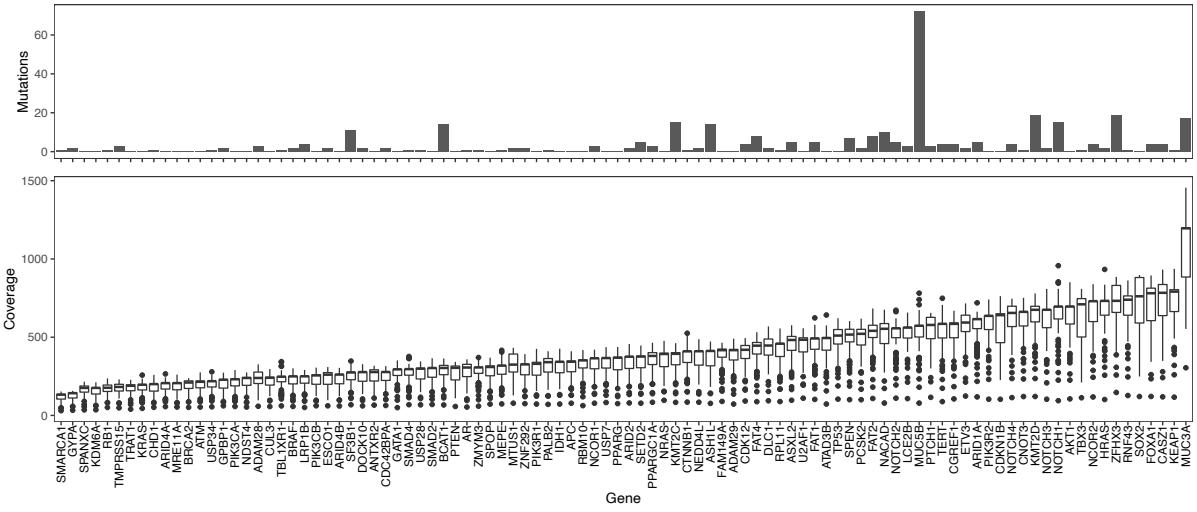


Figure 5.5: Mutations and coverage for each gene and across samples. There was a moderate variability of coverage for each gene. An association was found between the number of mutations detected per gene and their coverage (Spearman's $\rho=3.9\text{x}10^{-1}$, $P=5.6\text{x}10^{-5}$).

5.4.3 DETECTED SNVs

Overall, 354 SNVs were identified with support from at least three reads (Table C.1, Appendix C): mutations were detected in 47 out the 77 normal samples with a total of 169 mutations. Mutations were detected in all 18 tumour samples with a total of 185 (Table C.1, Appendix C). 258 SNVs (73 %) were non-synonymous. The mutation burden (measured as substitutions/megabase) was significantly higher in tumour (P = 5.5x10$^{-3,}$ Wilcoxon rank sum test, Figure 5.6) than in normal samples (mean of 0.06 and 0.03, respectively). The variant allele frequency ranged from ~ 0.6 % to 42 % in tumour (median of 2 %) and ~ 1 % to 36 % (median of 3) in normal samples and was significantly higher in the tumour samples (P = 4.4x10$^{-5,}$ Wilcoxon rank sum test, Figure 5.7).

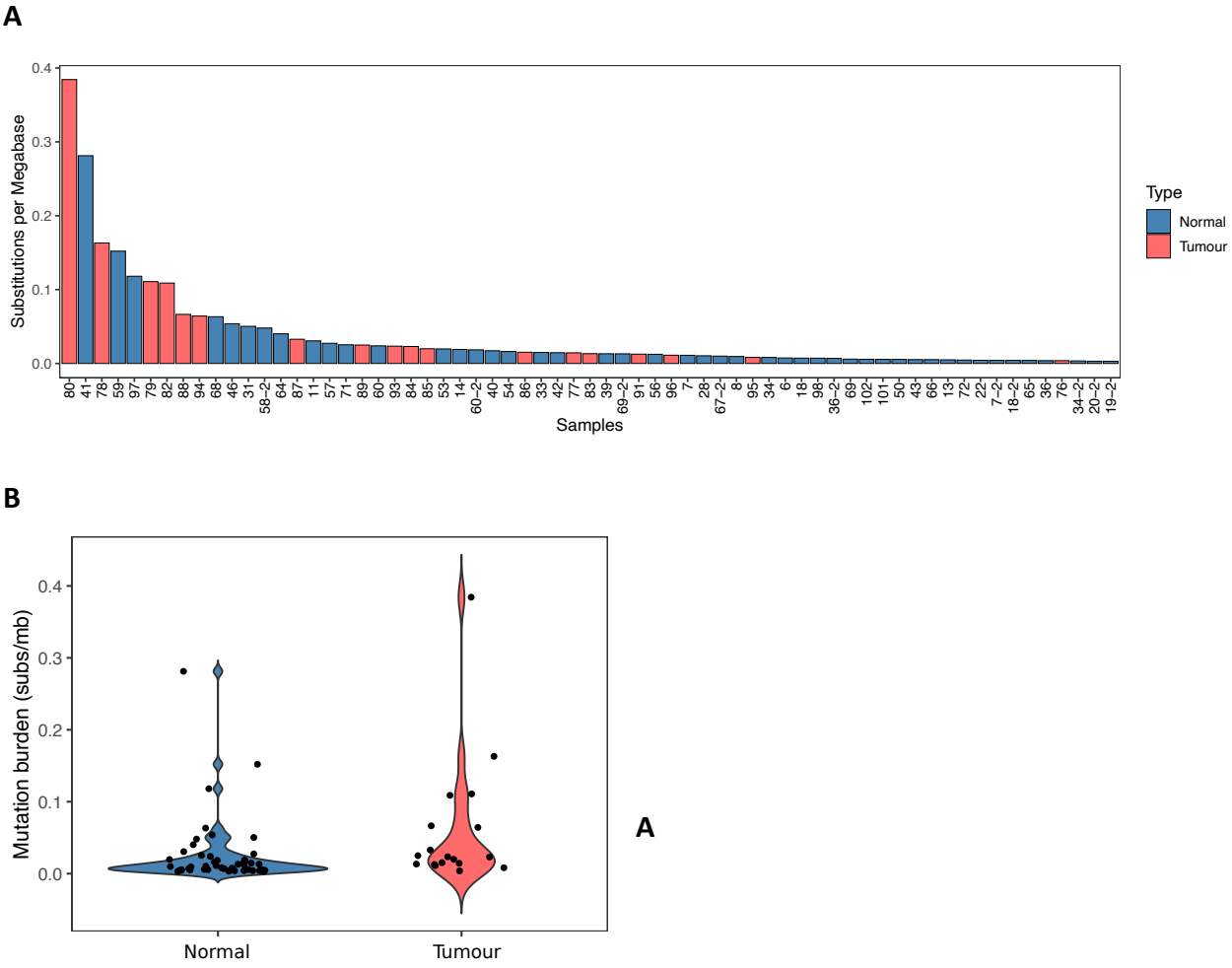
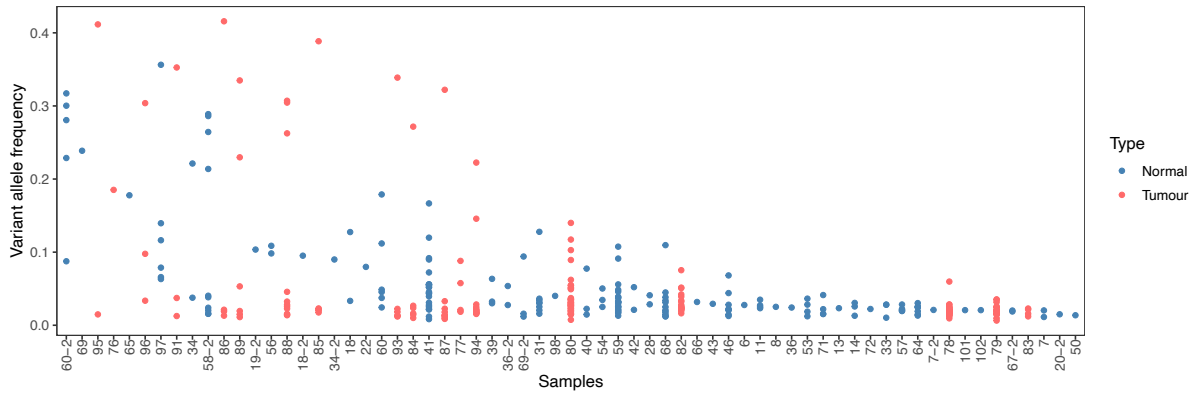


Figure 5.6: (A) Mutation burden across samples. **(B)** Boxplots showing the relationship between mutation burden and sample type (Normal/Tumour).



B

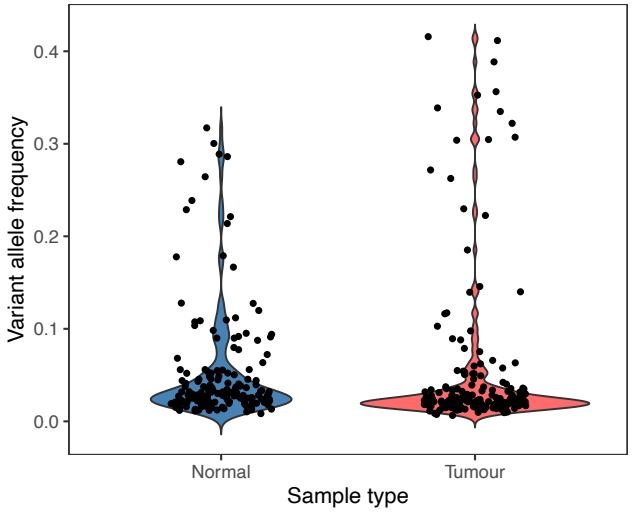


Figure 5.7: (A) Variant allele frequency across all samples. 354 mutations were detected in normal (blue) and tumour samples (red). **(B)** Boxplots showing the relationship between variant allele frequency and sample type (Normal/Tumour).

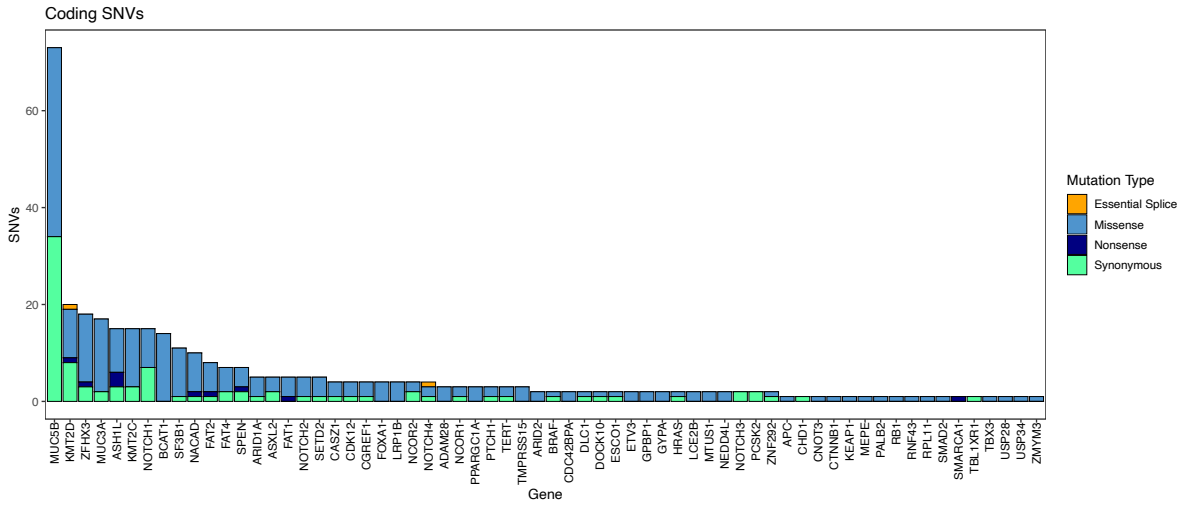


Figure 5.8: Number of mutations for each gene classified by mutation type (Synonymous vs Non-synonymous (Missense, nonsense and essential splice).

Of the 98 genes targeted, we identified at least one mutation in 63 of them (Figure 5.8). Mutated genes across samples are represented in Figure 5.9. A complete list of all the genes (mutated and non-mutated is shown in Table C.2, Appendix C).

For 24 genes (*ARID1A*, *ASHL1, ASXL2, BCAT1, CASZ1, CDK12, CGREF1, FAT1, FOXA1, FAT2, LRP1B, MKT2C, KMT2D, MUC3A, MUC5B, NACAD, NOTCH1, NOTCH2, PPARG1A, SETD2, SF3B1, SPEN, TMPRSS15* and *ZFHX3*) mutations were present in more than 3 samples (Figure 5.9). A spatial representation of all the mutations is shown for the whole prostate for those mutations reflecting clonal expansions or that affected the same gene in at least 3 samples (Figures 5.10 and 5.11).

Mutations detected from WGS for this patient were validated (Chapter 3, Figure A.1, Appendix A) in genes *BCAT, FAT2, ADAM28, GPBP1* and *TMPRSS15* (in the normal samples) and *RFPL11* and *SF3B1* in tumour samples. All mutations in these genes occurred in the same genomic location (in multiple normal/Tumour samples) except for gene *FAT2*, for which we report an extra five mutations in different genomic locations, three in normal samples and two in tumour.

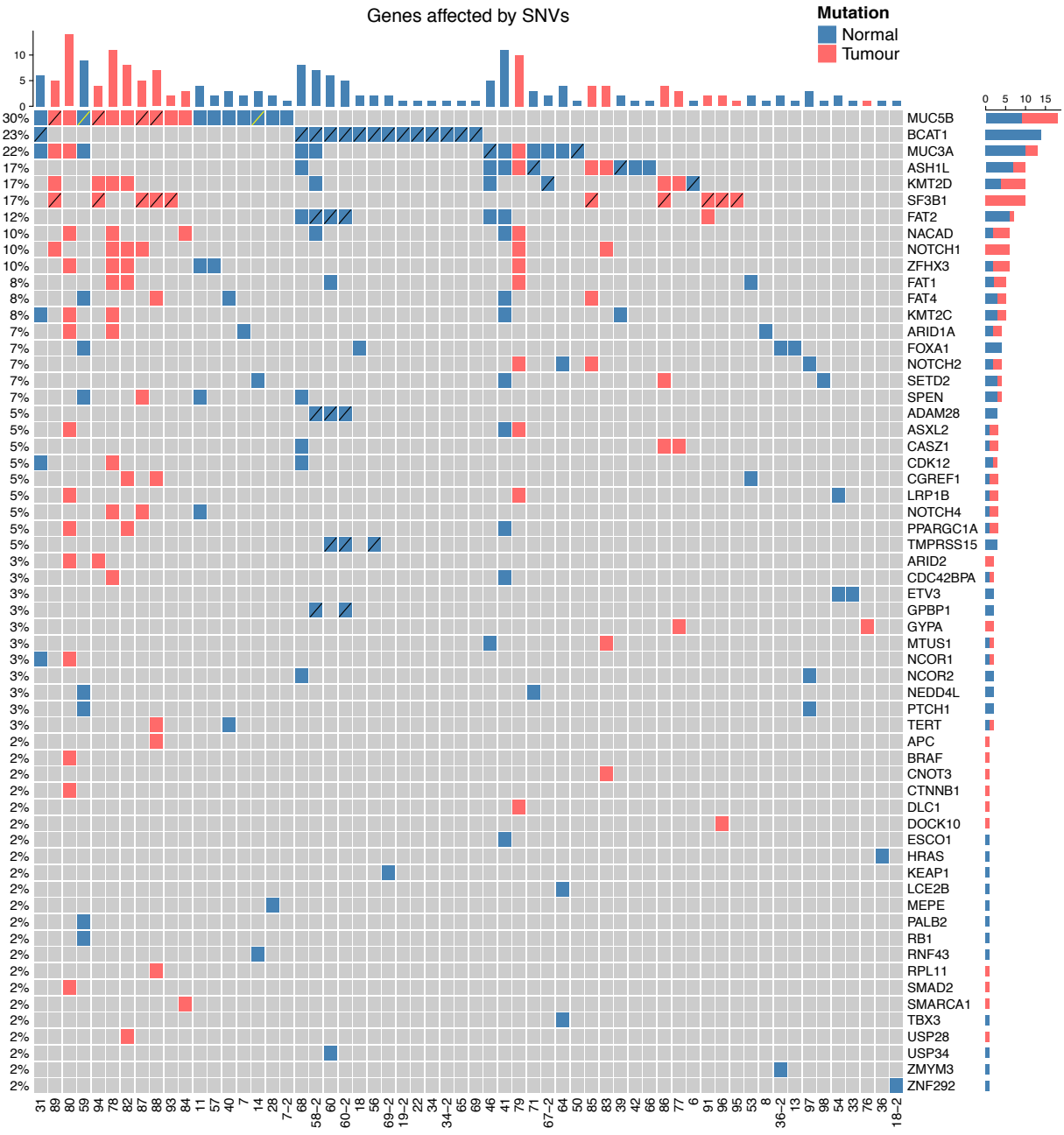


Figure 5.9: Mutated genes across all samples (Tumour =red; Normal = blue). Squares with diagonal lines represent samples where the mutation occurred at the exact same genomic location. Gene *MUC5B* has two different mutations occurring in the same location in multiple samples (represented by black and yellow lines).

5.4.3.1 POSITIVE SELECTION AND RECURRENT MUTATIONS IN DRIVERS

Using the dN/dS method (Methods, section 2.15.2) the gene *MUC3A* was identified to be under positive selection in the normal samples, harbouring 10 mutations that were protein altering mutations (Table 5.1). In addition, genes *MUC5B, ZFHX3, BCAT1, KMT2C, KMT2D, SF3B1* and *ASHL1* were recurrently mutated but the number of non-synonymous mutations did not reach significance.

Gene	Synonymous	Missense	Nonsense	Splice	$qval_{all}$
MUC3A	0	10	0	0	1.63x10-3

Table 5.1: Gene MUC3A is detected to be under significant positive selection after correcting for multiple hypotheses testing (qval < 0.05).

5.4.3.2 CLONAL EXPANSIONS

We identified mutations in the same genomic position overlapping several samples, suggesting seven clonal expansions. Normal tissue revealed four small clones that were supported by mutations in *MUC5B, MUC3A, KMT2C* and *ASHL1* and a larger clone in *BCAT1* that harboured several subclones affecting genes *ADAM28, FAT2, GPBP1* and *TMPRSS15* (Figure 5.11). In tumour, we identified two clones supported by mutations in genes *SF3B1* and *MUC5B*. Mutations in genes *BCAT1*, *ADAM28, FAT2, GPBP1*, *TMPRSS1* and *SF3B1* were validated by the slice used for WGS for the same patient.

5.4.3.2.1 CLONE 1

Mutations affecting gene in *BCAT1* overlapped in 14 samples (18, 22, 31, 34, 56, 60, 65, 68, 69, 19-2, 34-2, 58-2, 60-2, 69-2) and supported a large clone and subclones A, B and C (Figure 5.11 B). Mutations in *BCAT1* were present both above and below the slice used for WGS analysed in Chapter 3, sometimes separated by at least 2.5 cm from top to bottom of the prostate. CCFs in this large clone ranged from 3 % to ~ 60 % (Figure 5.10). Subclone A affecting gene *TMPRSS15* was occurring in samples 56, 60 and 60-2 and could have either originated from the *BCAT1* clone or independently as it has a low cellular fraction (the

pigeonhole principle is explained in Methods, section 2.11.1.2); subclone B (*ADAM28* and *FAT2*) was shared between samples 58-2, 60 and 60-2 and subclone C (gene *GPBP1*) could have either originated from the *BCAT1* gene clone or subclone B. These subclones have been represented linearly in Figure 5.11 B using the pigeonhole principle. All subclones (A, B, and C) were present in the slice immediately below the slice used for WGS, the median distance between the samples supporting each subclone was of 8.25, 5 and 5.5 mm, respectively (Figure 5.11 A).

5.4.3.2.2 CLONE 2

A large clone in tumour tissue was detected affecting gene *SF3B1* across ten samples (85, 86, 87, 88, 89, 91, 93, 94, 95 and 96). Mutations had an average CCF of 81 % and clustered together in the bottom slice of the prostate in an area of 0.7 cm^2 for *SF3B1*, 20 mm below the one used for WGS (Figures 5.10 and 5.12).

5.4.3.2.3 CLONE 3

A smaller subclone in the bottom slice of the prostate was affecting gene *MUC5B* comprised 4 samples (87, 88, 89 and 94). Mutations had an average CCF of 59 % clustered in an area of 0.15 cm^2 (Figures 5.10 and 5.12).

5.4.3.2.4 CLONES 4, 5, 6 AND 7

Mutations were detected in no more than two samples affecting genes *MUC5B, MUC3A, KMT2D* and *ASHL1* (Figure 5.12), suggesting the presence of small clones. CCFs ranged from 3 % to 6 % (Figure 5.10). All mutations were from different slices. Substitutions in *MUC5B* and *MUC3A* occurred in the slices immediately above and below the one used for WGS, whereas substitutions in *KMT2D* and *ASH1L* were further apart, one in the top slice (immediately above the one used for WGS) and the other in the bottom slice (20 mm below the WGS slice).

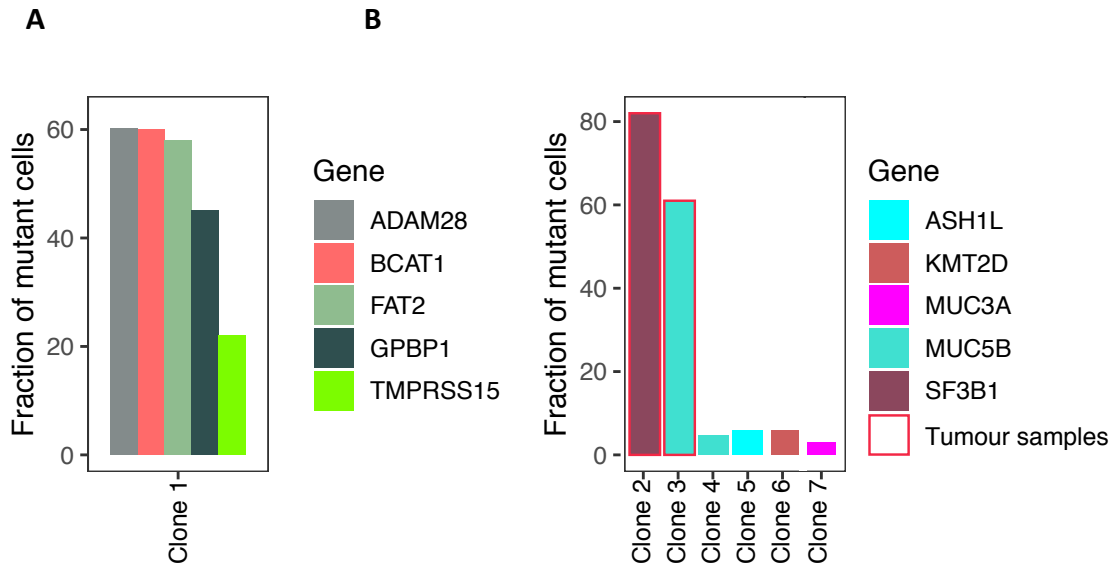


Figure 5.10: (A) CCF of the large clone affecting the gene *BCAT1* and its subclones (genes *ADAM28/FAT2, GPBP1* and *TMPRSS15)* supported by mutations overlapping multiple samples. (B) CCF of independent clones. Tumour samples are highlighted with red lines.

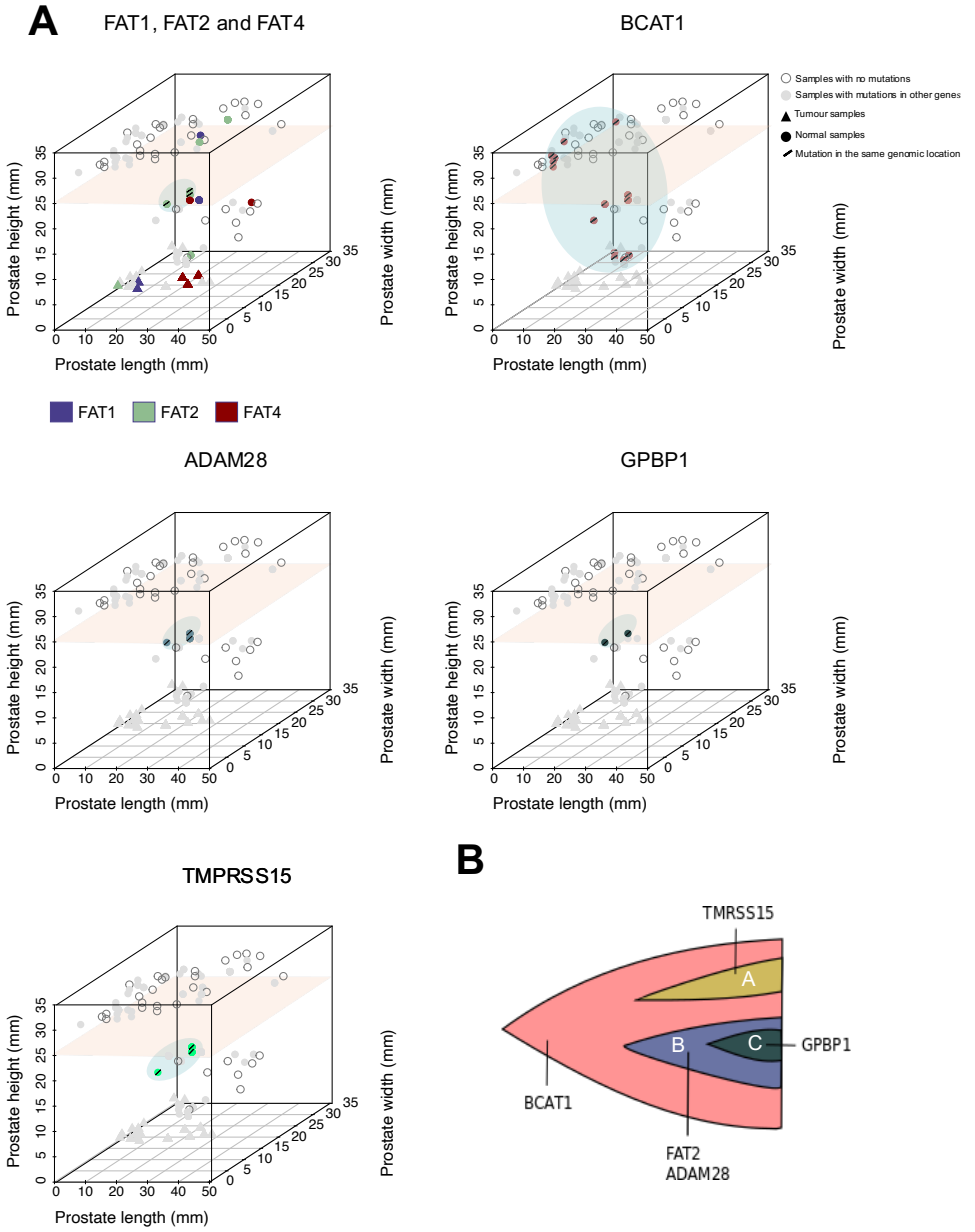
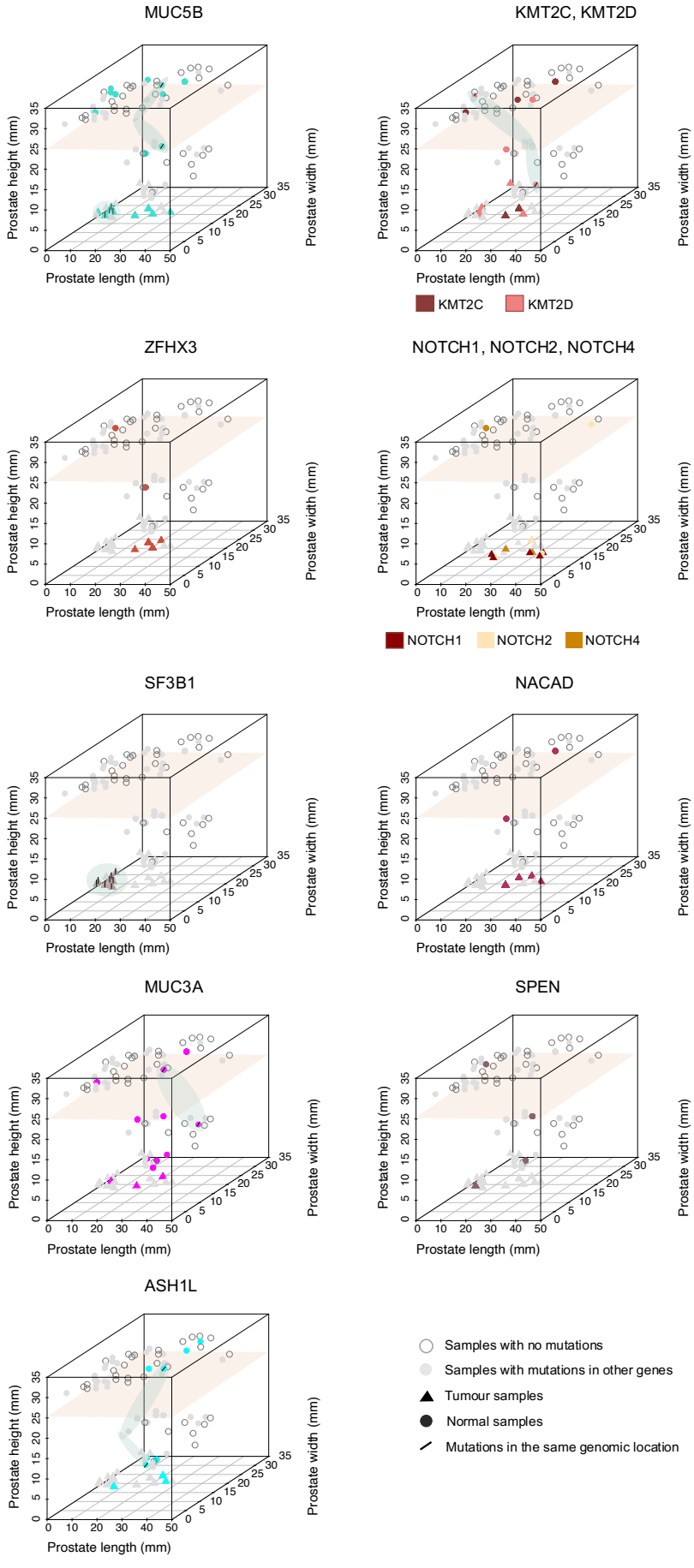


Figure 5.11: (A) Spatial representation of recurrent mutations supporting large clone affecting the BCAT1 gene that overlaps 14 samples and its respective subclones. Each gene/family of genes is represented separately with a different colour. Dots (for normal samples) and triangles (for tumour samples) indicate the location of the sample with the mutation. Grey dots/triangle show samples that did not harbour a mutation in that specific gene and empty dots represent samples with no mutations. Dots with black lines represent mutations occurring in the same genomic location. The horizontal plane marks the area where the sample for the WGS experiment was taken. Clonal expansions are highlighted with round light blue circles. **(B)** Subclonal architecture of the large BCAT1 clone.


MUC5B
KMT2C, KMT2D
KMT2C
KMT2D
ZFHX3
NOTCH1, NOTCH2, NOTCH4
NOTCH1
NOTCH2
NOTCH4
SF3B1
NACAD
MUC3A
SPEN
ASH1L
Prostate height (mm)
Prostate width (mm)
Prostate length (mm)
Samples with no mutations
Samples with mutations in other genes
Tumour samples
Normal samples
Mutations in the same genomic location

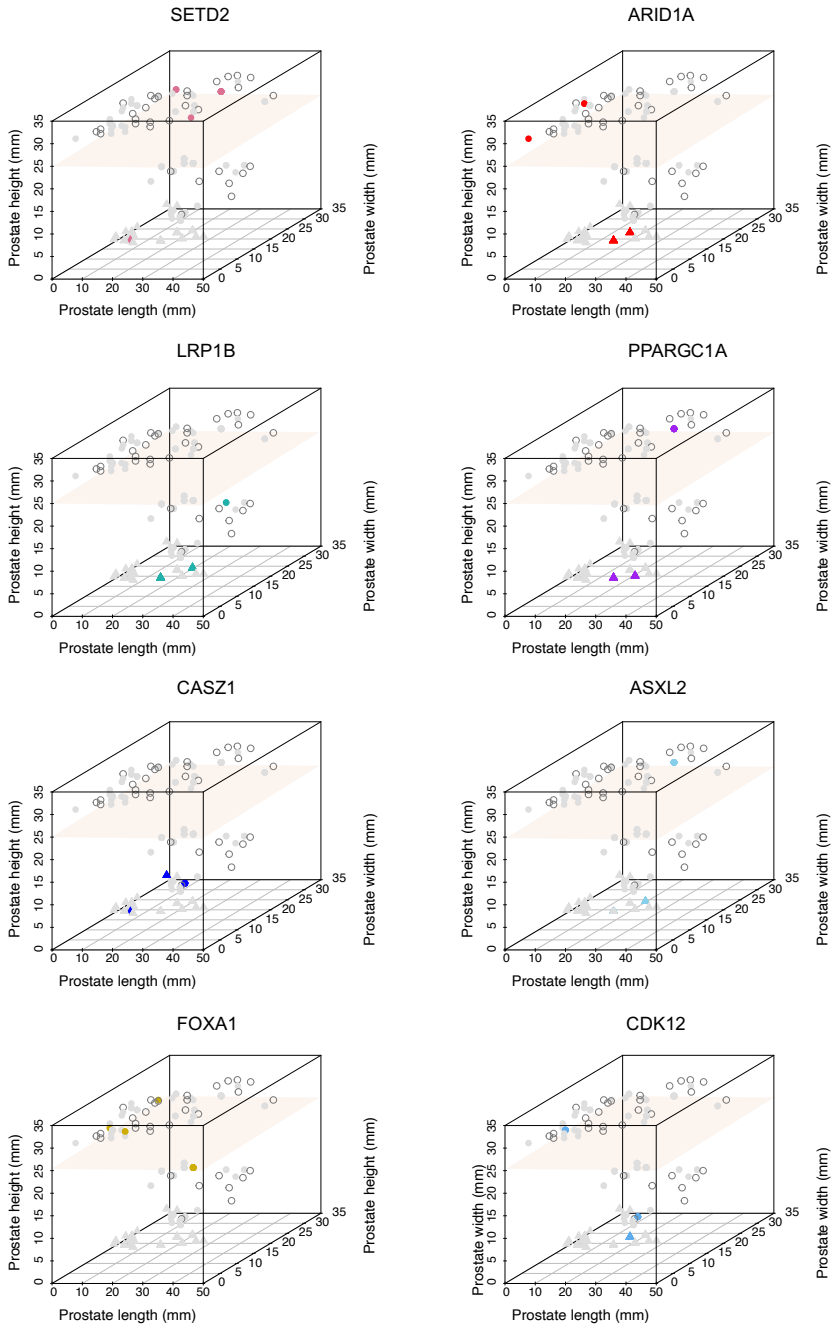


Figure 5.12: Spatial representation of recurrent mutations in the prostate. Each gene/family of genes is represented separately with a different colour. Dots (for normal samples) and triangles (for tumour samples) indicate the location of the sample with the mutation. Grey dots/triangle show samples that did not harbour a mutation in that specific gene and empty dots represent samples with no mutations. Dots with black lines represent mutations occurring in the same genomic location. The horizontal plane marks the area where the sample for the WGS experiment was taken. Clonal expansions are highlighted with round circles.

5.5 DISCUSSION

In this chapter we have targeted-sequenced 98 cancer associated genes in morphologically normal and tumour tissue from one patient with prostate cancer.

We found that the mutation burden was significantly lower in normal tissue in comparison to tumour. The overall mutation burden in prostate tumours has been reported to be higher (0.83-0.93 per Mb) in WES[130,256] and targeted sequencing experiments (0.33 per Mb)[257] than the one observed here, but much lower than in melanoma and lung cancer[109]. A possible explanation of the low mutation burden in tumours is the small sample size (18). Similarly, the mutation burden in normal in prostate is much lower than the observed in other normal tissues such as sun-exposed skin[118]. It is also plausible that there are low mutation rates among the genes targeted. Sequencing a broader panel of genes (or the whole exome) would address this problem.

We identified gene *MUC3A* was recurrently mutated and under positive selection in normal tissue. The *MUC3A* gene encodes for a membrane-associated mucin and it has been found recurrently mutated in gastric[258], ovarian, pancreatic, endometrial and lung squamous cell carcinoma[259], although the majority of these mutations were non-damaging. Altered expression levels of *MUC3A* have also been linked to poor prognosis in breast[260], gastric[261], appendiceal[262], esophagus[263] and clear-cell renal cell carcinoma[264].

Other genes, such as *MUC5B, ZFHX3, BCAT1, KMT2C, KMT2D, SF3B1* and *ASHL1* were recurrently mutated but the number of protein-altering mutations did not reach significance. However, the low detection of drivers could be a result of not having enough mutations in only one patient. Including more samples from other patients could greatly increase the possibility of detecting more genes as significant[118].

We detected seven clonal expansions occurring in the same group of cells simultaneously, however, only one small clone contained the driver mutation *MUC3A*. Mutations supporting clonal expansions were observed to be in nearby samples in genes *TMPRSS15, ADAM28, FAT2* and *GPBP1* (median distance of 8.25, 5, 5 and 5.5 mm, respectively), *SF3B1 (*clustered in an area of 0.7 cm^2*)* and *MUC5B (*clustered in an area of 0.15 cm^2), except for mutations in *BCAT1,* that were distributed across the prostate. All of them except *MUC5B* mutations were validated

by WGS. Interestingly, clones overlapping more distant samples (*ASH1L, MUC3A, KMT2D* and normal clone in *MUC5B*) were located in different prostate slices. This could be due to a lower vertical characterization of the prostate (only 3 slices).

However, clones with mutations in genes *KMT2D* and *ASH1L* had samples with no mutations in those genes in between. Given the low number of samples with detected mutations for these genes (two), the low CCF (< 6 %) and the lack of validation from the WGS experiment it is plausible that these mutations could be artefacts. Another explanation could be the presence of cell seeding, a phenomena that occurs when cells disseminate to other regions[22]. This scenario would suit the observed distribution of mutations affecting gene *BCAT1* that supported the large clone 1. Some of them were present in samples close together and in the same vertical plane, but there were also mutations from the top slice separated by 17 mm and other samples in between without detected mutations in *BCAT1*. This clone has been validated by WGS.

All clonal expansions harbouring the same mutation in different samples were found in one tissue type (normal or tumour): a clone harbouring mutations in *BCAT1* (and its subclones) was unique to normal, whereas *SF3B1* and *MUC5B* clones were unique to tumour tissue. As these genes contain the largest clonal expansions, these observations suggest there are no shared clones between normal and tumour. This is consistent with the findings in chapter 4. However, the majority of the mutated genes are found in both tissues.

At this stage, these results validate previous findings and show that mutations and clonal expansions are occurring in cancer associated genes, sometimes affecting multiple areas in normal tissues of prostate cancer patients. These results are in line with the presence of a field effect in normal tissue, although the frequency of drivers detected in this thesis in the normal prostate is low. Expanding the number of patients could increase the probability of detecting more low frequency mutations in potential drivers. Conversely, this result suggests that the field effect could be governed by epigenetic mechanisms.

CHAPTER 6 : DISCUSSION AND FUTURE WORK

In this thesis we aimed to understand the nature of the field effect in multifocal prostate cancer by analysing next generation sequencing data from morphologically normal tissue from men with and without prostate cancer. It had been previously reported that morphologically normal tissue of the prostate harboured a high number of mutations and clonal expansions. The two projects presented in this thesis are relevant for the understanding of early changes in cancer development that potentially lead to prostate cancer. In this section we summarise our results and suggest future research instructions to expand the knowledge about multifocal disease and the early stages of cancer initiation.

6.1 PROJECT 1: THE CHARACTERISATION OF THE MUTATIONAL LANDSCAPE AND SUBCLONAL ARCHITECTURE RECONSTRUCTION IN MORPHOLOGICALLY NORMAL TISSUES OF THE PROSTATE

For this project we analysed WGS data from morphologically normal tissue epithelium (including BPH samples) and fibroblasts cell cultures from men with and without prostate cancer and compared the normal data to their matched tumour when prostate cancer was present (Chapters 3 and 4). We confirmed high levels of mutations in the normal prostate as previously reported[17]. The contribution of each class of genetic alteration greatly differed between normal and tumour tissue: CNAs, chromosomal rearrangements and occasional *kataegis* events were specific to tumour tissue, whereas SNVs and indels were observed in both tissues. Prostate cancer is highly characterised by a high number of CNAs rearrangements such as *TMPRSS2-ERG* and there is evidence that proportion of the genome altered by these alterations have been associated with the progression of the disease[30].

The number of mutations in normal tissue from men with cancer was significantly higher in comparison to the normal prostates of men lacking cancer. Interestingly, a higher mutation rate was observed in BPH samples in comparison to normal samples without BPH, possibly due to hyper-proliferation of stromal tissue[228] that characterises BPH. Whether this finding is related to cancer development is unknown. In line with these results, the subclonal architecture reconstruction revealed that a high proportion of the normal samples from men with prostate cancer harbour subclonal expansions under selective pressure. Moreover, multiple unrelated subclones have been identified in a single prostate. Shared clones were found between multiple

normal and BPH/normal samples from the same patient but not between normal and tumour samples. In contrast, no subclonal expansions under selective pressure were detected in men without cancer.

No known or novel genetic drivers were identified in normal tissues except isolated instances (*PPARG, BRCA1, GATA1, HOXD11, WHSC1*, FAT1 and *POLE*), which strongly suggests the importance of other mechanisms such as transcriptomic and epigenetic alterations driving the clonal expansions. Interestingly, mutations in *GATA1, WHSC1, FAT1* and *POLE* were only observed in samples from a primary prostate fibroblast culture. Clones from cell cultured fibroblasts also showed a higher cellular cell fraction than morphologically normal clones, which has also been reported in CAFs, a known contributor to carcinogenesis[255].

Mutational processes (described by Alexandrov *et al.*[110]) were also detected in normal, BPH, BPH fibroblasts and tumour tissues. Tumour samples were defined by a group of signatures (1, 5, 8, 18, and 40) that have been previously associated with prostate cancer. This group was also observed in normal, BPH and BPH fibroblasts, in addition to signatures 4 and 28. Among the signatures identified, we highlight signatures 1, 5, 8. The presence of these processes in normal and BPH tissues has been previously reported by Cooper *et al*.[17] and Liu *et al*.[78] (Signature 1). We could confirm the widely described[225] observation that a lower number of SNVs results in a worse performance of the signature refitting process. However, the techniques applied by SigProfilerSingleSample tool in order to limit the number of signatures based on previous biological knowledge greatly improves the outcome of the mutational profile reconstruction.

6.2 THE PATCHWORK EXPERIMENT

For the second project of this thesis targeted sequencing was performed on 98 genes from morphologically normal and tumour tissue samples from one patient with prostate cancer. The genes on the panel were carefully selected according to the following criteria: included were known cancer and prostate cancer genes reported in the literature and genes that were found mutated in our group of normal samples. Mutations in genes *BCAT1, FAT2, ADAM28, GPBP1* and *TMPRSS15* that were previously detected by WGS for this patient were validated with this experiment. The gene *MUC3A* was recurrently mutated and identified as driver by the dN/dS method. Mutations in this gene have been observed (most of them non-damaging) in a variety of cancers: gastric[258], ovarian, pancreatic, endometrial and lung squamous cell carcinoma[259].

Likewise, *MUC3A* has been found to be differentially expressed and associated with poor prognosis in breast[260], gastric[261], appendiceal[262], esophagus[263] and clear-cell renal cell carcinoma[264].

Other recurrently mutated genes were *MUC5B, ZFHX3, BCAT1, KMT2C, KMT2D, SF3B1* and *ASHL1* but were not detected as drivers. Increasing the number of samples/patients could help improve the detection of significant driver genes[118].

We detected a total of 7 clonal expansions supported by mutations that were present in nearby multiple samples, and 5 of them (*BCAT1, FAT2, ADAM28, TMRPSS15* and *GPBP1*) were previously detected by WGS. A small clone contained mutations in gene *MUC3A* that were present in distant samples from different FFPE blocks. This scenario is also observed in clones containing mutations in genes *ASH1L, KMT2D* and *MUC5B*. It is plausible that these clones are affecting the tissue between these FFPE blocks, so further sequencing from this area would be needed in order to validate this. All clonal expansions were unique to each tissue type, indicating no intermixing between morphologically normal and tumour tissue, a finding that is in line with the results reported in chapter 4. Overall, our results show that frequency of drivers in the normal prostate is low, but expanding the number of patients could provide with more information.

6.3 OVERALL DISCUSSION AND LIMITATIONS OF THE RESEARCH

The analysis of the mutational landscape of normal prostate tissue from men with and without prostate cancer and the results obtained from the patchwork experiment reveal that these tissues harbour a high number of mutations, clonal expansions and mutational processes. All these findings are specific to prostate cancer patients. This suggests that the normal tissue of men with prostate cancer has undergone significant changes that could be associated with the development of cancer and the field effect and are in line with other studies that reported high numbers of mutations and clonal expansions in normal samples from esophagus[162], skin[118] and blood. In blood, the presence of clonal expansions has been significantly linked to an increased risk of developing leukemia overtime[163–166]. It is plausible that a higher frequency of mutations increases the possibility of developing cancer following a multistage model of carcinogenesis[254]. Nonetheless, we did not find the driving mechanism responsible of this field effect, as very few protein coding mutations were detected in normal tissues and only gene

MUC3A was detected as a potential driver by the dN/dS method. More studies (discussed in sections 6.4.1.2 and 6.4.2) are needed to validate this result and possibly uncover new potential driver genes. This scenario is consistent with results reported in a recent study by Liu *et al.*[78] in BPH tissue where they found somatic mutations but no recurrent mutations that would indicate positively selected genes. It has to be noted that the coverage reached in this study was no higher than 100X and therefore not enough to reveal very low frequency mutations. Overall, the lack of recurrent mutations in protein coding regions suggests that the development of a field effect could be governed by either epigenetic mechanisms, altered expression levels or both and needs to be further examined.

Likewise, more understanding is needed of the effect of BPH and the development of the field effect. BPH alone was a factor that was associated with a higher mutation burden even in non-cancer patients. Given that there is only one non-cancer patient with BPH, this association has to be further examined by analysis. Lastly, the present study failed to prove a relationship between age and mutation burden due to patients with prostate cancer having a similar age range and a very low number of non-prostate cancer patients (n =7). Similarly, the commonly observed association between age and prostate cancer development reported in many studies[51,265] was lacking in this thesis. These limitations will be addressed as described in section 6.4.1.

At the moment, more studies that explore the nature of the field effect are needed, as most studies focus on the examination of normal tissue adjacent to the tumour and therefore do not address the mechanisms that cause multiple distant lesions. This work will allow us to identify molecular features within normal tissue that could be used as biomarkers of early disease or targets for potential therapeutic intervention with the aim to prevent prostate cancer, slow its progression or even stop it.

6.4 FUTURE WORK

At this stage there a few factors that are crucial to deepen our understanding of the early stages of cancer development: expanding the number of samples for some of the tissue types, integrating data from multiple sequencing platforms, achieving a much more precise characterisation of all normal cell types in the prostate and reconstructing the subclonal architecture at single-cell resolution.

6.4.1 EXPANDING THE NUMBER OF SAMPLES

6.4.1.1 WGS EXPERIMENT

Firstly, increasing the number of samples of morphologically normal tissue from men without prostate cancer would be beneficial, as this group was only comprised of 7 samples. WGS sequencing for all the analyses described in Chapter 3 would be performed on at least 20 more non-cancer samples, of which at least 8 must have BPH and an another 8 must be from fibroblasts (4 from morphologically normal and 4 from BPH). This approach would allow us to achieve three goals. First, obtaining a more robust association between clonal expansions under selective pressure in normal tissues and prostate cancer presence, which would validate the results present in this thesis. Second, we would elucidate whether the changes observed in BPH (both epithelium and stroma) are unique to BPH or are also related to the development of a field effect. Thirdly, establishing the already reported relationship between age, mutation burden and clonal expansions in normal tissues is observed in the prostate. Apart from WGS, further experiments could be performed on all samples proposed here (see section 6.4.2).

6.4.1.2 PATCHWORK EXPERIMENT

As mentioned in section 5.5, the number of mutations detected in only one prostate was not high enough for detection of positively selected genes. As stated by Martincorena *et al.*[118], the inclusion of more samples from other patients could increase the possibility of detecting more genes under positive selection that reach significance[118]. Expanding the number of prostates to five (one from a man without prostate cancer) would help elucidate whether there is a notable proportion of potential driver genes in morphologically normal tissues of the prostate and whether it is affected by the presence of cancer. Future experiments regarding these expanded group of prostates are described in section and 6.4.3.

6.4.2 IDENTIFYING THE INITIATING EVENT OF THE FIELD EFFECT. INTEGRATING DATA FROM MULTIPLE SEQUENCING PLATFORMS

By using WGS we have obtained a good characterisation of normal tissue of the prostate and observed clear differences between morphologically normal tissues from men with and without prostate cancer. However, this technique has its limitations. As described in chapter 1, there are many changes that can alter gene function without altering the DNA sequence. There is evidence that the field effect could be epigenetic in nature. Integrating data from multiple sequencing platforms is essential to achieve this. Bisulphite sequencing could be used to

characterize methylation levels across the expanded group of samples as suggested in section 6.4.1.1, including the new category of BPH samples from men without prostate cancer. It has been previously reported that methylation profiles in normal tissues from men with prostate cancer are different to those from men lacking cancer[266,267]. More importantly, hypermethylation in genes *APC* and *GTSP1* and specific CpG loci[268] in normal tissues from men with a prostate negative biopsy can successfully predict cancer development later on. Likewise, we could carry out experiments to examine gene expression levels across already analysed samples and the new ones proposed in section 6.4.1.1 by using RNA-seq. Previous studies using microarrays reported that adjacent morphologically normal tissue from men with cancer displayed different gene expression patterns in comparison to men without cancer[269,270]. Combining RNA-sequencing, methylation and WGS data would greatly improve our understanding the nature of the field effect and holds promise for the development of an early diagnostic test by defining field effect methylation and expression profiles.

6.4.3 DEEP CHARACTERISATION OF NORMAL CELLS AT SINGLE CELL RESOLUTION AND BULK SAMPLES

The patchwork experiment revealed the presence of mutations in cancer related genes in one prostate. However, there was only one positively selected gene detected. Firstly, expanding the number of prostates to four could increase the possibility of detecting more positively selected genes[118]. Including at least a sample from a man without prostate cancer would also be ideal to compare the differences of mutational landscape in normal tissues from men that have developed prostate cancer and those who have not, and further validate the findings reported in this thesis. Secondly, a broader coverage of coding regions could be accomplished by performing exome sequencing instead of performing targeted sequencing only on a panel of genes. As a result, we would not miss potential driver genes that have not been targeted. Thirdly, single cell experiments could be carried out on all four prostates (including patient 0007 used for the patchwork experiment) in combination to exome sequencing from bulk samples. Single-cell whole genome sequencing (scWGS) and single-cell RNA sequencing (scRNA-Seq) would be performed for the detection and validation of potential driver genes and differential expression analyses, respectively. Single cell sequencing (ScSeq) has proven to be a very useful technique to reconstruct the subclonal architecture and study cancer evolution. Previously detected mutations through WGS and targeted DNA sequencing will be validated for patient 0007. Cells from frozen prostate normal epithelium (including BPH) from men with and without prostate cancer, normal stroma (from BPH and non-BPH samples) and

tumour could be isolated by using fluorescence-activated cell sorting (FACS). The subclonal architecture would be reconstructed at single cell resolution, providing an accurate subclonal composition for each tissue type and revealing subclone/subclone interactions. ScSeq technologies have been applied to normal tissues of the prostate by Gervaise *et al*.[271] and Crowley *et al*.[272] that both performed scRNA to characterise the normal anatomy of the prostate, and Wouter *et al*.[273], who studied the regeneration potential of normal stem cells. However, no scSeq experiments have been performed to address the nature of the field effect in the prostate cancer.

This would greatly enhance our understanding of early cancer development as it would allow a better characterization of genetic heterogeneity in morphologically normal and tumour tissues, possible detection of cancer invasion at an early stage and discovery of rare cell populations, which is a critical step in the understanding of drug resistance development. Stromal-epithelium interactions could be assessed in relation to the development of the field effect and BPH pathogenesis.

Overall, continuing this line of research is critical for the future development of better screening and prognostic tests through the dentification of specific drivers and expression/methylation profiles from prostate tumours and normal tissues from patients with and without prostate cancer. Deepening our understanding about the mechanisms responsible for the early stages of cancer initiation also has the potential to lead to characterisation of aggressive and indolent disease and to development of better treatments. The correct detection and prognostic tools are essential to improve the wellbeing of prostate cancer patients by avoiding the long term and life changing side effects that result from sometimes unnecessary treatments such as radical prostatectomy.

6.5 CONCLUSION

In summary, the findings presented in this thesis show further evidence of a field cancerization in the prostate and provide insights regarding the clonal dynamics of morphologically normal tissue. We detected a high number of mutations, clonal expansions and mutational processes in morphologically normal tissue from men with cancer in comparison to those without, indicating that the normal prostate in cancer patients had undergone notable changes as a whole. However, we can confirm that there is no clone mixing between clonal expansions from

tumour and morphologically normal tissues. Whether these findings in normal tissues are driving the disease process or on the other hand, are a result of tumour development is still unknown. Subclonal architecture reconstruction at single-cell resolution and examination of the epigenetic and expression profiles of normal tissues from men with and without cancer are needed to deepen our understanding or early stages of cancer development and the field effect.

REFERENCES

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
3. Pepper, J. W., Findlay, C. S., Kassen, R., Spencer, S. L. & Maley, C. C. Cancer research meets evolutionary biology. *Evol. Appl.* **2**, 62–70 (2009).
4. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–8 (1976).
5. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
6. Stratton, M., Campbell, P. & Futreal, P. The cancer genome. *Nature* **458**, 719–724 (2009).
7. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
8. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat Rev Genet* **13**, 795–806 (2012).
9. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
10. Rode, A., Maass, K. K., Willmund, K. V., Lichter, P. & Ernst, A. Chromothripsis in cancer cells: An update. *Int. J. Cancer* **138**, 2322–2333 (2016).
11. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
12. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. **29**, 1120–1127.
13. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
14. Park, S. Y., Gönen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. **120**, (2010).
15. Khalique, L. *et al.* Genetic intra-tumour heterogeneity in epithelial ovarian cancer and its implications for molecular diagnosis of. 286–295 (2007) doi:10.1002/path.
16. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

17. Cooper, C. S. *et al.* Analysis of the Genetic Phylogeny of Multifocal Prostate Cancer Identifies Multiple Independent Clonal Expansions in Neoplastic and Morphologically Normal Prostate Tissue. *Nat. Genet.* **47**, 367–372 (2015).
18. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized , multifocal prostate cancer. *Nat. Publ. Gr.* **47**, 736–745 (2015).
19. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
20. Shoag, J. & Barbieri, C. E. Clinical variability and molecular heterogeneity in prostate cancer. *Asian J. Androl.* **0**, 0 (2016).
21. Turajlic, S., McGranahan, N. & Swanton, C. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochim. Biophys. Acta - Rev. Cancer* **1855**, 264–275 (2015).
22. Inda, M.-M. *et al.* Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes Dev.* **24**, 1731–1745 (2010).
23. Chapman, A. *et al.* Heterogeneous tumor subpopulations cooperate to drive invasion. *Cell Rep.* **8**, 688–695 (2014).
24. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
25. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science (80-.).* **349**, 1483 LP – 1489 (2015).
26. Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* **22**, 97–104 (2016).
27. Hasty, P. & Montagna, C. Chromosomal Rearrangements in Cancer: Detection and potential causal mechanisms. *Mol. Cell. Oncol.* **1**, e29904 (2014).
28. Ottaviani, D., LeCain, M. & Sheer, D. The role of microhomology in genomic structural variation. *Trends Genet.* **30**, 85—94 (2014).
29. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
30. Camacho, N. *et al.* Appraising the relevance of DNA copy number loss and gain in prostate cancer using whole genome DNA sequence data. *PLoS Genet.* **13**, e1007001–e1007001 (2017).
31. Mirchia, K. *et al.* Total copy number variation as a prognostic factor in adult astrocytoma subtypes. *Acta Neuropathol. Commun.* **7**, 92 (2019).

32. Roy, D. M. *et al.* Integrated Genomics for Pinpointing Survival Loci within Arm-Level Somatic Copy Number Alterations. *Cancer Cell* **29**, 737–750 (2016).
33. Marczok, S., Bortz, B., Wang, C. & Pospisil, H. Comprehensive Analysis of Genome Rearrangements in Eight Human Malignant Tumor Tissues. *PLoS One* **11**, e0158995 (2016).
34. Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J. Clin. Invest.* **117**, 2033–2035 (2007).
35. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).
36. Weterings, E. & Van Gent, D. C. The mechanism of non-homologous end-joining: A synopsis of synapsis. *DNA Repair (Amst).* **3**, 1425–1435 (2004).
37. Hastings, P. J., Ira, G. & Lupski, J. R. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLOS Genet.* **5**, e1000327 (2009).
38. Park, J. Y. Promoter Hypermethylation in Prostate Cancer. *Cancer Control* **17**, 245–255 (2010).
39. Nebbioso, A., Tambaro, F. P., Dell'Aversana, C. & Altucci, L. Cancer epigenetics: Moving forward. *PLOS Genet.* **14**, e1007362 (2018).
40. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
41. Humphrey, P. A. Histological variants of prostatic carcinoma and their significance. *Histopathology* **60**, 59–74 (2012).
42. Bangma, C. H. & Roobol, M. J. Defining and predicting indolent and low risk prostate cancer. *Crit. Rev. Oncol. Hematol.* **83**, 235–241 (2012).
43. Bul, M. *et al.* Active Surveillance for Low-Risk Prostate Cancer Worldwide : The PRIAS Study. **63**, 597–603 (2013).
44. Bechis, S. K., Carroll, P. R. & Cooperberg, M. R. Impact of age at diagnosis on prostate cancer treatment and survival. *J. Clin. Oncol.* **29**, 235–241 (2011).
45. Jahn, J. L., Giovannucci, E. L. & Stampfer, M. J. The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. *Int. J. cancer* **137**, 2795–2802 (2015).
46. Leissner, K.-H. & Tisell, L.-E. The Weight of the Human Prostate. *Scand. J. Urol. Nephrol.* **13**, 137–142 (1979).

47. Owen, D. H. & Katz, D. F. A review of the physical and chemical properties of human semen and the formulation of a semen simulant. *J. Androl.* **26**, 459–469 (2005).
48. Kuriyama, M. *et al.* Quantitation of Prostate-specific Antigen in Serum by a Sensitive Enzyme Immunoassay. *Cancer Res.* **40**, 4658 LP – 4662 (1980).
49. Marker, P. C., Donjacour, A. A., Dahiya, R. & Cunha, G. R. Hormonal, cellular, and molecular control of prostatic development. *Dev. Biol.* **253**, 165–174 (2003).
50. Cunha, G. R. *et al.* The endocrinology and developmental biology of the prostate. *Endocr. Rev.* **8**, 338–362 (1987).
51. Heinzer, H. & Steuber, T. Prostate cancer in the elderly. *Urol. Oncol. Semin. Orig. Investig.* **27**, 668–672 (2009).
52. Carter, H. B. *et al.* Early Detection of Prostate Cancer: AUA Guideline. *J. Urol.* **190**, 419–426 (2013).
53. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45**, 10.1038/ng.2560 (2013).
54. Haas, G. P., Delongchamps, N., Brawley, O. W., Wang, C. Y. & de la Roza, G. The Worldwide Epidemiology of Prostate Cancer: Perspectives from Autopsy Studies. *Can. J. Urol.* **15**, 3866–3871 (2008).
55. Mottet, N. *et al.* Full-Text. (2015).
56. Heidenreich, A. *et al.* EAU guidelines on prostate cancer. Part 1: Screening, diagnosis, and local treatment with curative intent - Update 2013. *Eur. Urol.* **65**, 124–137 (2014).
57. National Institue for Clinical Excellence. Recommendations | Prostate cancer: diagnosis and management | Guidance | NICE. (2019).
58. Dickinson, L. *et al.* Scoring systems used for the interpretation and reporting of multiparametric MRI for prostate cancer detection, localization, and characterization: could standardization lead to improved utilization of imaging within the diagnostic pathway? *J. Magn. Reson. Imaging* **37**, 48–58 (2013).
59. Chen, N. & Zhou, Q. The evolving Gleason grading system. *Chinese J. Cancer Res.* **28**, 58–64 (2016).
60. Egevad, L., Granfors, T., Karlberg, L., Bergh, A. & Stattin, P. Prognostic value of the Gleason score in prostate cancer. *BJU Int.* **89**, 538–542 (2002).
61. Brierley J, Gospodarowicz MK, Wittekind C (2017). Wiley, C. *TNM classification of malignant tumours.*
62. D'Amico, A. V. *et al.* Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate

cancer. *J. Am. Med. Assoc.* **280**, 969–974 (1998).

63. Montironi, R., Mazzucchelli, R., Lopez-Beltran, A., Cheng, L. & Scarpelli, M. Mechanisms of Disease: high-grade prostatic intraepithelial neoplasia and other proposed preneoplastic lesions in the prostate. *Nat. Clin. Pract. Urol.* **4**, 321–332 (2007).
64. De Marzo, A. M., Marchi, V. L., Epstein, J. I. & Nelson, W. G. Proliferative inflammatory atrophy of the prostate: implications for prostatic carcinogenesis. *Am. J. Pathol.* **155**, 1985–1992 (1999).
65. Qian, J., Wollan, P. & Bostwick, D. G. The extent and multicentricity of high-grade prostatic intraepithelial neoplasia in clinically localized prostatic adenocarcinoma. *Hum. Pathol.* **28**, 143–148 (1997).
66. McNeal, J. E. & Bostwick, D. G. Intraductal dysplasia: A premalignant lesion of the prostate. *Hum. Pathol.* **17**, 64–71 (1986).
67. Nonn, L., Ananthanarayanan, V. & Gann, P. H. Evidence for field cancerization of the prostate. *Prostate* **69**, 1470–1479 (2009).
68. Bostwick, D. *et al. Independent origin of multiple foci of prostatic intraepithelial neoplasia - Comparison with matched foci of prostate carcinoma*. *Cancer* vol. 83 (1998).
69. Wein, A. J., Between, C., Urinary, L. & Symptoms, T. Benign prostatic hyperplasia. *Surgery* **32**, 1010–1011 (2004).
70. Ørsted, D. D. & Bojesen, S. E. The link between benign prostatic hyperplasia and prostate cancer. *Nat. Rev. Urol.* **10**, 49–54 (2013).
71. Cai, Y. Benign prostatic hyperplasia is a reawakened process of persistent Mu È llerian duct mesenchyme. (2001).
72. Biology, S. & Biology, F. Prostate enlargement in mice due to fetal exposure to low doses of estradiol or diethylstilbestrol and opposite effects at high doses. **94**, 2056–2061 (1997).
73. Chughtai, B., Lee, R., Te, A. & Kaplan, S. Role of inflammation in benign prostatic hyperplasia. *Rev. Urol.* **13**, 147–50 (2011).
74. Robert, G. *et al.* Inflammation in benign prostatic hyperplasia: a 282 patients' immunohistochemical analysis. *Prostate* **69**, 1774–1780 (2009).
75. Miah, S. & Catto, J. BPH and prostate cancer risk. *Indian J. Urol.* **30**, 214–218 (2014).
76. Saaristo, L. *et al.* 1039 Genetic testing in identification of BPH patients developing later prostate cancer. *Eur. Urol. Suppl.* **12**, e1039 (2013).

77. Ewing, C. M. *et al.* Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.* **366**, 141–149 (2012).
78. Liu, D. *et al.* Integrative multiplatform molecular profiling of benign prostatic hyperplasia identifies distinct subtypes. *Nat. Commun.* **11**, 1987 (2020).
79. Tomlins, S. A. *et al.* Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.* **39**, 41–51 (2007).
80. Shyr, D. & Liu, Q. Next generation sequencing in cancer research and clinical application. *Biol. Proced. Online* **15**, 4 (2013).
81. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun. (Camb).* **47**, 7281–7286 (2011).
82. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
83. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).
84. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
85. Liu, Y., Loewer, M., Aluru, S. & Schmidt, B. SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. *BMC Syst. Biol.* **10 Suppl 2**, 47 (2016).
86. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1-15.10.18 (2016).
87. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, (2015).
88. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
89. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
90. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
91. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data.

Genome Biol. **17**, 178 (2016).

92. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
93. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
94. Hansen, N. F., Gartner, J. J., Mei, L., Samuels, Y. & Mullikin, J. C. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics* **29**, 1498–1503 (2013).
95. Li, S. *et al.* SOAPindel: efficient identification of indels from short paired reads. *Genome Res.* **23**, 195–200 (2013).
96. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
97. Jiang, Y., Wang, Y. & Brudno, M. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**, 2576–2583 (2012).
98. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
99. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–9 (2008).
100. K, Y. & G, H. Structural Variation Detection from Next Generation Sequencing. *J. Next Gener. Seq. Appl.* **01**, (2015).
101. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207–211 (1998).
102. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166–1174 (2008).
103. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
104. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
105. Imielinski, M. *et al.* NIH Public Access. **150**, 1107–1120 (2013).
106. Govindan, R. *et al.* GENOMIC LANDSCAPE OF NON-SMALL CELL LUNG CANCER IN SMOKERS AND NEVER SMOKERS. *Cell* **150**, 1121–1134 (2012).
107. Rodin, S. N. & Rodin, A. S. Origins and selection of p53 mutations in lung

carcinogenesis. *Semin. Cancer Biol.* **15**, 103–112 (2005).

108. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).
109. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).
110. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
111. Dunson, D. B. Nonparametric Bayes Applications to Biostatistics. 1–47 (1992).
112. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
113. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Rep.* **7**, 1740–1752 (2017).
114. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
115. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
116. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
117. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
118. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880 LP – 886 (2015).
119. Svensson, M. a *et al.* Testing mutual exclusivity of ETS rearranged prostate cancer. *Lab. Invest.* **91**, 404–412 (2011).
120. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
121. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
122. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

123. Cairns, P. Renal Cell Carcinoma. *Cancer Biomarkers* **9**, 461–473 (2011).
124. Mitchell, T. & Neal, D. E. The genomic evolution of human prostate cancer. *Br. J. Cancer* **113**, 193–198 (2015).
125. Zhao, X. *et al.* Integrative analysis of cancer driver genes in prostate adenocarcinoma. *Mol. Med. Rep.* **19**, 2707–2715 (2019).
126. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
127. Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* **50**, 682–692 (2018).
128. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
129. Phin, S., Moore, M. W. & Cotter, P. D. Genomic Rearrangements of PTEN in Prostate Cancer. *Front. Oncol.* **3**, 240 (2013).
130. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
131. White, N. M., Feng, F. Y. & Maher, C. A. Recurrent rearrangements in prostate cancer: causes and therapeutic potential. *Curr. Drug Targets* **14**, 450–9 (2013).
132. Rubin, M. A., Maher, C. A. & Chinnaiyan, A. M. Common gene rearrangements in prostate cancer. *J. Clin. Oncol.* **29**, 3659–3668 (2011).
133. Pettersson, A. *et al.* THE TMPRSS2:ERG REARRANGEMENT, ERG EXPRESSION, AND PROSTATE CANCER OUTCOMES: A COHORT STUDY AND META-ANALYSIS. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1497–1509 (2012).
134. Clark, J. P. & Cooper, C. S. ETS gene fusions in prostate cancer. *Nat. Rev. Urol.* **6**, 429–439 (2009).
135. Wang, L. *et al.* Increased Androgen Receptor Gene Copy Number is Associated With TMPRSS2-ERG Rearrangement in Prostatic Small Cell Carcinoma. **907**, 900–907 (2015).
136. Weischenfeldt, J. *et al.* Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. *Cancer Cell* **23**, 159–170 (2013).
137. Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat. Med.* **16**, 793–798 (2010).
138. Ren, S. *et al.* RNA-seq analysis of prostate cancer in the Chinese population identifies

recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.* **22**, 806–821 (2012).

139. Williams, J. L., Greer, P. A. & Squire, J. A. Recurrent copy number alterations in prostate cancer: An in silico meta-analysis of publicly available genomic data. *Cancer Genet.* **207**, 474–488 (2014).
140. Visakorpi, T. *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat. Genet.* **9**, 401–406 (1995).
141. Lolli, C. *et al.* Testosterone levels and androgen receptor copy number variations in castration-resistant prostate cancer treated with abiraterone or enzalutamide. *Prostate* **79**, 1211–1220 (2019).
142. Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv* 322859 (2018) doi:10.1101/322859.
143. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
144. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2018).
145. Esteller, M., Corn, P. G., Baylin, S. B. & Herman, J. G. A gene hypermethylation profile of human cancer. *Cancer Res.* **61**, 3225–3229 (2001).
146. Nelson, W. G., De Marzo, A. M. & Yegnasubramanian, S. Epigenetic alterations in human prostate cancers. *Endocrinology* **150**, 3991–4002 (2009).
147. Lee, W. H. *et al.* Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11733–11737 (1994).
148. Sathyanarayana, U. G. *et al.* Aberrant Promoter Methylation of Laminin-5-Encoding Genes in Prostate Cancers and Its Relationship to Clinicopathological Features. *Clin. Cancer Res.* **9**, 6395–6400 (2003).
149. Yegnasubramanian, S. *et al.* Hypermethylation of CpG Islands in Primary and Metastatic Human Prostate Cancer. *Cancer Res.* **64**, 1975–1986 (2004).
150. Kang, G. H., Lee, S., Lee, H. J. & Hwang, K. S. Aberrant CpG island hypermethylation of multiple genes in prostate cancer and prostatic intraepithelial neoplasia. *J. Pathol.* **202**, 233–240 (2004).
151. Rosenbaum, E. *et al.* Promoter hypermethylation as an independent prognostic factor for relapse in patients with prostate cancer following radical prostatectomy. *Clin. Cancer Res.* **11**, 8321–8325 (2005).

152. Haldrup, C. *et al.* DNA Methylation Signatures for Prediction of Biochemical Recurrence After Radical Prostatectomy of Clinically Localized Prostate Cancer. *J. Clin. Oncol.* **31**, 3250–3258 (2013).
153. Kristensen, H. *et al.* Hypermethylation of the GABRẼmiR-45Z̃miR-224 promoter in prostate cancer predicts biochemical recurrence after radical prostatectomy. *Clin. Cancer Res.* **20**, 2169–2181 (2014).
154. Sørensen, K. D. *et al.* Genetic and epigenetic SLC18A2 silencing in prostate cancer is an independent adverse predictor of biochemical recurrence after radical prostatectomy. *Clin. Cancer Res.* **15**, 1400–1410 (2009).
155. Henrique, R. *et al.* Epigenetic Heterogeneity of High-Grade Prostatic Intraepithelial Neoplasia: Clues for Clonal Progression in Prostate Carcinogenesis. *Mol. Cancer Res.* **4**, 1 LP – 8 (2006).
156. Chai, H. & Brown, R. E. field effect. *Ann. Clin. Lab. Sci.* **39**, 331–338 (2009).
157. Slaughter, D. P., Southwick, H. W. & Smejkal, W. "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer* **6**, 963–968 (1953).
158. Waridel, F. *et al.* Field cancerisation and polyclonal p53 mutation in the upper aero-digestive tract. *Oncogene* **14**, 163–169 (1997).
159. Park, S.-K. *et al.* Field Cancerization in Sporadic Colon Cancer. *Gut Liver* **10**, 773–780 (2016).
160. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
161. Yizhak, K. *et al.* A comprehensive analysis of RNA sequences reveals macroscopic somatic clonal expansion across normal tissues. (2018).
162. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. **917**, 911–917 (2018).
163. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
164. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
165. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
166. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).

167. Colom, B. *et al.* Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nat. Genet.* **52**, 604–614 (2020).
168. Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer- associated mutations in normal tissues. 1–12 (2018).
169. Risk, M. C. *et al.* Differential gene expression in benign prostate epithelium of men with and without prostate cancer: evidence for a prostate cancer field effect. *Clin. Cancer Res.* **16**, 5414–5423 (2010).
170. Cohen, B. L. *et al.* Cyclooxygenase-2 (cox-2) expression is an independent predictor of prostate cancer recurrence. *Int. J. Cancer* **119**, 1082–1087 (2006).
171. Trock, B. J. *et al.* Evaluation of GSTP1 and APC methylation as indicators for repeat biopsy in a high-risk cohort of men with negative initial prostate biopsies. *BJU Int.* **110**, 56–62 (2012).
172. Partin, A. W. *et al.* Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. *J. Urol.* **192**, 1081–1087 (2014).
173. D., S. G. *et al.* Clinical Utility of an Epigenetic Assay to Detect Occult Prostate Cancer in Histopathologically Negative Biopsies: Results of the MATLOC Study. *J. Urol.* **189**, 1110–1116 (2013).
174. Møller, M. *et al.* Heterogeneous patterns of DNA methylation-based field effects in histologically normal prostate tissue from cancer patients. *Sci. Rep.* **7**, 40636 (2017).
175. Warren, A. Y. *et al.* Method for SamplingTissue for ResearchWhich Preserves Pathological Data in Radical Prostatectomy. **202**, 194–202 (2013).
176. Egevad, L. Handling of radical prostatectomy specimens. *Histopathology* **60**, 118–124 (2012).
177. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* **1303**, (2013).
178. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
179. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
180. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
181. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 390–398 (2000). doi:10.1109/SFCS.2000.892127.

182. Gupta, M. R. & Chen, Y. Theory and use of the em algorithm. *Found. Trends Signal Process.* **4**, 223–296 (2010).
183. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res.* **11**, 1725–1729 (2001).
184. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
185. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
186. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).
187. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
188. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv* (2020) doi:10.1101/2020.12.13.422570.
189. Goldfarb, D. & Idnani, A. A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* **27**, 1–33 (1983).
190. Byrd, R., Hribar, M. E. & Nocedal, J. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM J. Optim.* **9**, 877–900 (1999).
191. Statistics, M. A Bayesian Analysis of Some Nonparametric Problems Author (s): Thomas S . Ferguson Source : The Annals of Statistics , Vol . 1 , No . 2 (Mar ., 1973), pp . 209-230 Published by : Institute of Mathematical Statistics Stable URL : http://www.jstor.org/s. *Statistics (Ber).* **1**, 209–230 (2010).
192. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
193. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
194. Gundem, G. *et al.* Europe PMC Funders Group The Evolutionary History of Lethal Metastatic Prostate Cancer. **520**, 353–357 (2015).
195. Paisley, J. & Jordan, M. I. A constructive definition of the beta process. (2016).
196. Frigui, H. Clustering: Algorithms and applications. *2008 1st Int. Work. Image Process. Theory, Tools Appl. IPTA 2008* (2008) doi:10.1109/IPTA.2008.4743793.
197. Pearson K. Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**, 559–572 (1901).

198. Statistics, M. Bootstrap Methods : Another Look at the Jackknife Author (s): B . Efron Source : The Annals of Statistics , Vol . 7 , No . 1 (Jan ., 1979), pp . 1-26 Published by : Institute of Mathematical Statistics Stable URL : http://www.jstor.org/stable/2958830. *Statistics (Ber).* **7**, 1–26 (2008).
199. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238 (2016).
200. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
201. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. in (1995).
202. Chang, X. & Wang, K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.* **49**, 433–436 (2012).
203. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
204. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
205. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
206. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).
207. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
208. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
209. Kim, J. *et al.* Unfavourable prognosis associated with K-ras gene mutation in pancreatic cancer surgical margins. *Gut* **55**, 1598–1605 (2006).
210. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
211. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
212. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*

29, 308–311 (2001).

213. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
214. Network, C. G. A. R. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
215. Huang, X., Wojtowicz, D. & Przytycka, T. M. Genome analysis Detecting presence of mutational signatures in cancer with confidence. **34**, 330–337 (2018).
216. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
217. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res.* **24**, 733–742 (2014).
218. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
219. Yu, Y. P. *et al.* Genome abnormalities precede prostate cancer and predict clinical relapse. *Am. J. Pathol.* **180**, 2240–2248 (2012).
220. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
221. Ponder, R. G., Fonville, N. C. & Rosenberg, S. M. A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation. *Mol. Cell* **19**, 791–804 (2005).
222. Sakofsky, C. J. *et al.* Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep.* **7**, 1640–1648 (2014).
223. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLoS Genet.* **12**, 1–15 (2016).
224. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
225. Baez-Ortega, A. & Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinform.* **20**, 77–88 (2019).
226. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260 (2016).
227. Ju, Y. S. *et al.* dynamics in the early human embryo. *Nat. Publ. Gr.* **543**, 714–718 (2017).

228. Barclay, W. W., Woodruff, R. D., Hall, M. C. & Cramer, S. D. A system for studying epithelial-stromal interactions reveals distinct inductive abilities of stromal cells from benign prostatic hyperplasia and prostate cancer. *Endocrinology* **146**, 13–18 (2005).
229. Hayes, G. M. *et al.* Regional Cell Proliferation in Microdissected Human Prostate Specimens after Heavy Water Labeling In Vivo: Correlation with Prostate Epithelial Cells Isolated from Seminal Fluid. *Clin. Cancer Res.* **18**, 3250 LP – 3260 (2012).
230. Chokkalingam, A. P. *et al.* Prostate Carcinoma Risk Subsequent to Diagnosis of Benign Prostatic Hyperplasia A Population-Based Cohort Study in Sweden. (2003) doi:10.1002/cncr.11710.
231. Shimizu, R., Engel, J. D. & Yamamoto, M. GATA1-related leukaemias. *Nat. Rev. Cancer* **8**, 279–287 (2008).
232. Swaroop, A. *et al.* An activating mutation of the NSD2 histone methyltransferase drives oncogenic reprogramming in acute lymphocytic leukemia. *Oncogene* **38**, 671–686 (2019).
233. Consortium, A. P. G. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
234. Rosen, E. M., Fan, S., Pestell, R. G. & Goldberg, I. D. BRCA1 gene in breast cancer. *J. Cell. Physiol.* **196**, 19–41 (2003).
235. Cao, X. H. *et al.* FAT1 expression in different breast lesions and its down-regulation in breast cancer development. *Int. J. Clin. Exp. Pathol.* **10**, 7242–7248 (2017).
236. Yu, S. *et al.* Comprehensive analysis of the GATA transcription factor gene family in breast carcinoma using gene microarrays, online databases and integrated bioinformatics. *Sci. Rep.* **9**, 4467 (2019).
237. Cazier, J. B. *et al.* Whole-genome sequencing of bladder cancers reveals somatic CDKN1A mutations and clinicopathological associations with mutation burden. *Nat. Commun.* **5**, (2014).
238. Guerra, J. *et al.* POLE somatic mutations in advanced colorectal cancer. *Cancer Med.* **6**, 2966–2971 (2017).
239. Yu, J., Liu, M., Liu, H. & Zhou, L. GATA1 promotes colorectal cancer cell proliferation, migration and invasion via activating AKT signaling pathway. *Mol. Cell. Biochem.* **457**, 191–199 (2019).
240. Peters, I., Dubrowinskaja, N., Tezval, H. & Kramer, M. W. Decreased mRNA expression of GATA1 and GATA2 is associated with tumor aggressiveness and poor outcome in clear cell renal cell carcinoma. 267–275 (2015) doi:10.1007/s11523-014-

0335-8.

241. Imboden, S. *et al.* Phenotype of POLE-mutated endometrial cancer. *PLoS One* **14**, 1–15 (2019).
242. Saloura, V. *et al.* WHSC1 Promotes Oncogenesis through Regulation of NIMA-Related Kinase-7 in Squamous Cell Carcinoma of the Head and Neck. *Mol. Cancer Res.* **13**, 293 LP – 304 (2015).
243. Lin, S. C. *et al.* FAT1 somatic mutations in head and neck carcinoma are associated with tumor progression and survival. *Carcinogenesis* **39**, 1320–1330 (2018).
244. de Barros e Lima Bueno, R. *et al.* HOX genes: potential candidates for the progression of laryngeal squamous cell carcinoma. *Tumor Biol.* **37**, 15087–15096 (2016).
245. Castro, E. & Eeles, R. The role of BRCA1 and BRCA2 in prostate cancer. *Asian J. Androl.* **14**, 409–414 (2012).
246. Li, N. *et al.* AKT-mediated stabilization of histone methyltransferase WHSC1 promotes prostate cancer metastasis. *J. Clin. Invest.* **127**, 1284–1302 (2017).
247. Elix, C., Pal, S. K. & Jones, J. O. The role of peroxisome proliferator-activated receptor gamma in prostate cancer. *Asian J. Androl.* **20**, 238–243 (2018).
248. Silver, D. P. & Livingston, D. M. Mechanisms of BRCA1 tumor suppression. *Cancer Discov.* **2**, 679–684 (2012).
249. Katoh, M. Function and cancer genomics of FAT family genes (review). *Int. J. Oncol.* **41**, 1913–1918 (2012).
250. Ogi, T. *et al.* Three DNA Polymerases, Recruited by Different Mechanisms, Carry Out NER Repair Synthesis in Human Cells. *Mol. Cell* **37**, 714–727 (2010).
251. Campos-Sanchez, E. *et al.* Wolf-Hirschhorn Syndrome Candidate 1 Is Necessary for Correct Hematopoietic and B Cell Development. *Cell Rep.* **19**, 1586–1601 (2017).
252. Lefterova, M. I. *et al.* PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes Dev.* **22**, 2941–2952 (2008).
253. Ferreira, R., Ohneda, K., Yamamoto, M. & Philipsen, S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.* **25**, 1215–1227 (2005).
254. Armitage, P. Multistage models of carcinogenesis. *Environ. Health Perspect.* **63**, 195–201 (1985).
255. Foster, D. S. *et al.* Abstract 10: Characterizing the Clonal Nature of Cancer Associated Fibroblasts. *Plast. Reconstr. Surg. Glob. Open* **6**, 8 (2018).
256. Huang, F. W. *et al.* Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function ERF Mutations. *Cancer Discov.* **7**, 973–983 (2017).

257. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
258. Cui, J. *et al.* Comprehensive characterization of the genomic alterations in human gastric cancer. *Int. J. cancer* **137**, 86–95 (2015).
259. King, R. J., Yu, F. & Singh, P. K. Genomic alterations in mucins across cancers. *Oncotarget* **8**, 67152–67168 (2017).
260. Rakha, E. A. *et al.* Expression of mucins (MUC1, MUC2, MUC3, MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer. *Mod. Pathol.* **18**, 1295–1304 (2005).
261. Wang, R.-Q. & Fang, D.-C. Alterations of MUC1 and MUC3 expression in gastric carcinoma: relevance to patient clinicopathological features. *J. Clin. Pathol.* **56**, 378–384 (2003).
262. Shibahara, H. *et al.* A comprehensive expression analysis of mucins in appendiceal carcinoma in a multicenter study: MUC3 is a novel prognostic factor. *PLoS One* **9**, e115613–e115613 (2014).
263. Arul, G. S. *et al.* Mucin gene expression in Barrett's oesophagus: an in situ hybridisation and immunohistochemical study. *Gut* **47**, 753–761 (2000).
264. Niu, T. *et al.* Increased expression of MUC3A is associated with poor prognosis in localized clear-cell renal cell carcinoma. *Oncotarget* **7**, 50017–50026 (2016).
265. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA. Cancer J. Clin.* **70**, 7–30 (2020).
266. Luo, J. H. *et al.* Genome-wide methylation analysis of prostate tissues reveals global methylation patterns of prostate cancer. *Am. J. Pathol.* **182**, 2028–2036 (2013).
267. Yang, B. *et al.* Methylation profiling defines an extensive field defect in histologically normal prostate tissues associated with prostate cancer. *Neoplasia (United States)* **15**, 399–408 (2013).
268. Yang, B. *et al.* Validation of an epigenetic field of susceptibility to detect significant prostate cancer from non-tumor biopsies. *Clin. Epigenetics* **11**, 168 (2019).
269. Chandran, U. R. *et al.* Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. *BMC Cancer* **5**, 45 (2005).
270. Yu, Y. P. *et al.* Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy. *J. Clin. Oncol.* **22**, 2790–2799 (2004).

271. Henry, G. H. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep.* **25**, 3530-3542.e5 (2018).
272. Crowley, L. *et al.* A single-cell atlas of the mouse and human prostate reveals heterogeneity and conservation of epithelial progenitors. *bioRxiv* 2020.05.29.123836 (2020) doi:10.1101/2020.05.29.123836.
273. Karthaus, W. *et al.* Regenerative potential of prostate luminal cells revealed by single-cell analysis. *Science (80-.).* **368**, 497–505 (2020).
274. Wadhera, P. An introduction to acinar pressures in BPH and prostate cancer. *Nat Rev Urol* **10**, 358–366 (2013).
275. Montironi, R., Bostwick, D. G. & Lopez-beltran, A. Staging of prostate cancer. (2012) doi:10.1111/j.1365-2559.2011.04025.x.
276. Rizzi, E., Lari, M., Gigli, E., De Bellis, G. & Caramelli, D. Ancient DNA studies: new perspectives on old samples. *Genet. Sel. Evol.* **44**, 21 (2012).
277. Lin, D.-C. *et al.* The genomic landscape of nasopharyngeal carcinoma. *Nat Genet* **46**, 866–871 (2014).

APPENDIX A

SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 3

Cases	Sample type	Snvs	Indels	CNAs	SVs	Age	PSA	Pathological_stage	Gleason	Multifocal	Distance N-T (mm)
0006_N3	Normal tissue	2566	580	0	1	71	7	T3aNxMx	3+4	N	1
0006_N2	Normal tissue	639	482	0	0	71	7	T3aNxMx	3+4	N	9
0006_N1	Normal tissue	578	72	0	1	71	7	T3aNxMx	3+4	N	9.1
0007_N2	Normal tissue	1818	656	0	0	60	10.1	T1cNxMx	4+3	Y	2
0007_N3	Normal tissue	636	520	0	0	60	10.1	T1cNxMx	4+3	Y	4
0007_N1	Normal tissue	527	25	0	0	60	10.1	T1cNxMx	4+3	Y	4
0008_N2	Normal tissue	718	467	0	0	64	6.7	T2aN0Mx	3+3	Y	13
0008_N3	Normal tissue	695	495	0	0	64	6.7	T2aN0Mx	3+3	Y	17
0008_N1	Normal tissue	28	14	0	0	64	6.7	T2aN0Mx	3+3	Y	
0240	Normal tissue	852	559	0	0	73	5.2	unknown	unknown	Y	
0065_N	Normal tissue	591	468	0	0	61	12.8	T1cNxMx	3+4	N	19
0145	Normal tissue	547	634	0	0	50	10.85	T1cNxMx	3+3	Y	6
0144	Normal tissue	488	459	0	0	53	6.4	T1cNxMx	4+3	N	5
0124	Normal tissue	457	428	0	0	66	5.63	T1cNxMx	3+4	Y	5
0156	Normal tissue	436	460	0	0	69	19.03	T1cNxMx	4+5	N	1
0122	Normal tissue	430	461	0	0	66	8.38	T3aNxMx	3+4	Y	1
0115	Normal tissue	418	578	0	0	55	7	T1cNxMx	3+4	Y	13
0140	Normal tissue	407	349	0	0	68	9.34	T2aNxMx	4+5	Y	4.05
0162	Normal tissue	407	369	0	0	61	11.05	T2aNxMx	3+5	Y	1.5
0127	Normal tissue	405	352	0	3	61	7.63	T1cNxMx	4+5	Y	4.1
0152	Normal tissue	402	500	0	0	57	6.5	T1cNxMx	3+3	N	18
0073_N	Normal tissue	395	419	0	0	60	6.3	T1cNxMx	3+4	Y	15.1
0146	Normal tissue	392	451	0	0	68	7.3	T1cNxMx	3+4	Y	4
0160	Normal tissue	370	440	0	0	51	15.4	T1cNxMx	4+3	Y	0.5
0149	Normal tissue	357	480	0	0	67	13.9	T2aNxMx	3+3	Y	13
0077_N	Normal tissue	338	437	0	1	66	7	T1cNxMx	3+3	Y	2
0159	Normal tissue	328	353	0	0	73	6.79	T1cNxMx	3+4	Y	14
0076	Normal tissue	314	346	0	0	59	7.3	T1cNxMx	3+4	Y	7
0120	Normal tissue	285	354	0	0	52	7	T1cN0M0	3+4	Y	3.1
PD6204c	Normal tissue	23	16	0	0	unknown	unknown	unknown	unknown	unknown	
0074	BPH	1213	691	0	0	54	6.07	T1cN0Mx	3+4	Y	5.5
0066	BPH	1157	474	0	0	70	20.74	T1cNxMx	3+3	Y	2.1
0063	BPH	1075	518	0	1	68	10.6	T1cNxMx	3+4	Y	20
0116	BPH	1075	420	0	0	62	11.13	T1cNxMx	3+4	Y	2
0065_BPH	BPH	952	583	0	0	61	12.8	T1cNxMx	3+4	N	5.1
0073_BPH	BPH	674	572	0	0	60	6.3	T1cNxMx	3+4	Y	15
0077_BPH	BPH	424	461	0	1	66	7	T1cNxMx	3+3	Y	6
0069	BPH	399	347	0	0	64	7.57	T1cNxMx	3+4	Y	3
0072	BPH	300	567	0	0	61	5.4	T1cNxMx	3+3	Y	7.1
0239	Normal tissue (no PC)	1202	134	0	0	72	unknown				
0244	Normal tissue (no PC)	159	62	0	0	65	unknown				
0243	Normal tissue (no PC)	148	77	0	0	15	unknown				
0241	Normal tissue (no PC)	141	67	0	0	25	unknown				
0238	Normal tissue (no PC)	140	55	0	0	62	unknown				
0242	Normal tissue (no PC)	125	58	0	0	54	unknown				
0245	Normal tissue (no PC)	104	60	0	0	45	unknown				
0246	Cultured fibroblasts (BPH)	234	21	0	0	unknown	unknown				
0247	Cultured fibroblasts (BPH)	238	51	0	0	unknown	unknown				
0250	Cultured fibroblasts (BPH)	2431	184	0	0	unknown	unknown				

Table A.1: Morphologically normal sample summary. Coloured cells indicate multiple samples. In yellow are those samples not included in the mutational signature analysis. Sample order corresponds to the sample order displayed in Figure 3.5 A.

Cases	Sample_type	Snvs	Indels	CNAs	SVs	Age	PSA	Pathological_stage	Gleason	Multifocal
0006_T1	Tumour	6133	614	33	80	71	7	T3aNxMx	3+4	N
0006_T4	Tumour	3626	375	7	23	71	7	T3aNxMx	3+4	N
0006_T2	Tumour	3225	428	15	31	71	7	T3aNxMx	3+4	N
0006_T3	Tumour	1802	230	8	42	71	7	T3aNxMx	3+4	N
0007_T2	Tumour	2434	199	7	35	60	10.1	T1cNxMx	4+3	Y
0007_T1	Tumour	1678	251	10	9	60	10.1	T1cNxMx	4+3	Y
0007_T5	Tumour	1251	165	9	22	60	10.1	T1cNxMx	4+3	Y
0007_T3	Tumour	853	162	2	1	60	10.1	T1cNxMx	4+3	Y
0007_T4	Tumour	88	138	3	16	60	10.1	T1cNxMx	4+3	Y
0008_T1	Tumour	3309	381	18	55	64	6.7	T2aN0Mx	3+3	Y
0008_T2	Tumour	2533	255	14	49	64	6.7	T2aN0Mx	3+3	Y
0008_T3	Tumour	170	223	6	1	64	6.7	T2aN0Mx	3+3	Y
0065	Tumour	3423	420	27	32	61	12.8	T1cNxMx	3+4	N
0145	Tumour	2666	316	78	83	50	10.85	T1cNxMx	3+3	Y
0144	Tumour	1925	261	83	75	53	6.4	T1cNxMx	4+3	N
0124	Tumour	2703	273	13	31	66	5.63	T1cNxMx	3+4	Y
0156	Tumour	2872	267	46	53	69	19.03	T1cNxMx	4+5	N
0122	Tumour	3352	323	18	39	66	8.38	T3aNxMx	3+4	Y
0115	Tumour	2804	305	24	41	55	7	T1cNxMx	3+4	Y
0140	Tumour	2602	287	53	126	68	9.34	T2aNxMx	4+5	Y
0162	Tumour	3308	263	40	479	61	11.05	T2aNxMx	3+5	Y
0127	Tumour	2657	382	25	32	61	7.63	T1cNxMx	4+5	Y
0152	Tumour	2104	200	43	58	57	6.5	T1cNxMx	3+3	N
0073	Tumour	2123	211	46	64	60	6.3	T1cNxMx	3+4	Y
0146	Tumour	2055	192	23	13	68	7.3	T1cNxMx	3+4	Y
0160	Tumour	3381	173	21	12	51	15.4	T1cNxMx	4+3	Y
0149	Tumour	2446	187	22	2	67	13.9	T2aNxMx	3+3	Y
0077	Tumour	3204	553	74	187	66	7	T1cNxMx	3+3	Y
0159	Tumour	2321	298	36	32	73	6.79	T1cNxMx	3+4	Y
0076	Tumour	2588	308	46	88	59	7.3	T1cNxMx	3+4	Y
0120	Tumour	3236	382	31	37	52	7	T1cN0M0	3+4	Y
PD6204a	Tumour	2300	179	48	72	unknown	unknown	unknown	unknown	unknown
0074	Tumour	2255	208	21	25	54	6.07	T1cN0Mx	3+4	Y
0066	Tumour	2507	301	9	83	70	20.74	T1cNxMx	3+3	Y
0063	Tumour	4890	640	83	41	68	10.6	T1cNxMx	3+4	Y
0116	Tumour	2450	218	13	97	62	11.13	T1cNxMx	3+4	Y
0069	Tumour	1752	170	10	24	64	7.57	T1cNxMx	3+4	Y
0072	Tumour	2719	349	55	144	61	5.4	T1cNxMx	3+3	Y

Table A.2: Tumour samples summary. Coloured cells indicate multiple samples. In yellow are those samples not included in the mutational signature analysis. Sample order corresponds to the sample order displayed in Figure 3.5 B.

Sample	% Cancer	% Dysplasia	% Lymph	Average epithelial (%)	Average stroma (%)	SNVs	CCF	Type
0006_N1	0	4	0	34	62	578	48	Normal
0006_N2	0	0	0	26	74	639	0	Normal
0006_N3	0	19	0	5	76	2566	34	Normal
0007_N1	0	3	1	12	84	527	35	Normal
0007_N2	0	4	5	20	71	1818	30	Normal
0007_N3	0	0	0	20	80	636	0	Normal
0008_N2	0	23	0	30	47	718	20	Normal
0008_N2	0	0	0	25	75	695	0	Normal
0063	0	0	2.5	7.5	90	1075	0	Normal
0065_N	0	0	0	20	80	591	60	Normal
0065_BPH	0	0	0	32	68	952	58	Normal (BPH)
0066	0	0	0	24	76	1157	52	Normal (BPH)
0069	0	0	0	13	87	399	0	Normal (BPH)
0072	0	3	0	31	53	300	38	Normal (BPH)
0073_N	0	0	0	20	80	395	71	Normal
0073_BPH	0	0	0	12	88	674	69	Normal (BPH)
0074	0	0	0	15	85	1213	46	Normal (BPH)
0076	0	0	0	20	80	314	37	Normal
0077_N	0	0	0	15	85	338	37	Normal
0077_BPH	0	0	0	15	85	424	33	Normal (BPH)
0115	0	0	0	40	60	418	35	Normal
0116	0	0	0	45	55	1075	0	Normal (BPH)
0120	0	0	0	12.5	87.5	285	36	Normal
0122	0	0	0	40	60	430	0	Normal
0124	0	0	0	20	80	457	0	Normal
0127	0	0	0	20	80	405	0	Normal
0140	0	0	0	20	80	407	0	Normal
0144	0	0	0	30	70	488	0	Normal
0145	0	0	1	20	79	547	0	Normal
0146	0	0	0	10	90	392	60	Normal
0149	0	0	0	30	70	357	36	Normal
0152	0	0	0	20	80	402	0	Normal
0156	0	0	0	20	80	436	56	Normal
0159	0	0	0	20	80	328	44	Normal
0160	0	0	0	25	75	370	0	Normal
0162	0	0	1	20	79	407	40	Normal

Table A.3: Percentage of epithelial and stromal tissue across all morphologically normal samples from prostate cancer patients.

Chr	Position	Ref	Alt	Type	Gene	Function	Found in dbSNP	COSMIC	TCGA	Present in Tumour
4	171010770	C	G	exonic	AADAT	Nsym	.			
14	74004547	C	T	exonic	ACOT1	Nsym	rs199627073			Yes
22	51177701	G	A	exonic	ACR	Nsym	.	Tier1		
8	24192995	G	A	exonic	ADAM28	Nsym	.		Yes	
4	175897727	C	T	exonic	ADAM29	stopgain	.			Yes
4	71388554	A	G	exonic	AMTN	Nsym	.			
10	61846440	A	T	exonic	ANK3	Nsym	.			
9	42368472	C	A	exonic	ANKRD20A2	Nsym	.			
3	93761891	C	A	exonic	ARL13B	Nsym	rs139997243			
1	1431048	A	G	exonic	ATAD3B	Nsym	rs201429000			
19	42492630	T	C	exonic	ATP1A3	Nsym	.		Yes	
1	154309899	C	T	exonic	ATP8B2	Nsym	rs767034036			
12	24989522	G	T	exonic	BCAT1	Nsym	.		Yes	
17	41245732	G	A	exonic	BRCA1	Nsym	.	Tier1		Yes
6	32261213	C	T	exonic	C6orf10	Nsym	.			
22	37888677	C	T	exonic	CARD10	Nsym	rs377599502			
17	77757713	G	C	exonic	CBX2	Nsym	.			
1	93657591	C	T	exonic	CCDC18	Nsym	.			Yes
1	227441857	C	T	splicing	CDC42BPA	Nsym	.			
7	150934857	G	T	exonic	MIR671	Nsym	.		Yes	
3	130383856	A	C	exonic	COL6A6	Nsym	.			
3	15563062	G	A	exonic	COLQ	Nsym	rs374642884			
8	2836275	C	T	exonic	CSMD1	Nsym	rs200039361			Yes
8	104394781	G	A	exonic	CTHRC1	Nsym	.		Yes	
8	143960482	G	T	exonic	CYP11B1	Nsym	.			
16	70398492	G	A	exonic	DDX19A	Nsym	.		Yes	
12	12974315	A	G	exonic	DDX47	Nsym	.			
10	127527604	A	G	exonic	DHX32	Nsym	.		Yes	
19	49868880	C	G	exonic	DKKL1	Nsym	.			
21	34958343	C	T	exonic	DONSON	Nsym	.		Yes	
18	29049153	C	T	exonic	DSG3	Nsym	rs561461640		Yes	
5	118274932	G	T	exonic	DTWD2	Nsym	.			
4	187077231	C	T	exonic	FAM149A	Nsym	rs368472968			
2	170387991	C	T	exonic	FASTKD1	Nsym	rs752543615			
4	187524513	C	T	exonic	FAT1	Nsym	.	Tier1	Yes	
5	150885254	A	T	exonic	FAT2	Nsym	.			
1	42645391	G	A	exonic	FOXJ3	Nsym	.		Yes	
1	151065765	A	G	exonic	GABPB2	Nsym	.			
X	48651642	C	T	exonic	GATA1	Nsym	.	Tier1		
14	24706276	G	A	splicing	GMPR2	Nsym	.		Yes	
15	74368296	G	A	exonic	GOLGA6A	Nsym	rs542174200			
15	75561228	C	T	exonic	GOLGA6C	Nsym	rs778683471			
5	56526784	C	T	exonic	GPBP1	Nsym	rs775619360			
9	125797606	T	A	exonic	GPR21	Nsym	.			
4	145041707	C	A	exonic	GYPA	Nsym	rs7658293			
11	5269623	G	C	exonic	HBG1	Nsym	rs56205611			
2	176972669	C	G	exonic	HOXD11	Nsym	.	Tier1		
X	114141604	G	T	exonic	HTR2C	Nsym	.			Yes
15	91025262	T	G	exonic	IQGAP1	Nsym	.		Yes	
10	24832217	G	T	exonic	KIAA1217	Nsym	.		Yes	
12	52696929	C	A	exonic	KRT86	Nsym	rs771522388			Yes
21	46117387	C	A	exonic	KRTAP10-12	Nsym	.			
1	62673042	G	C	exonic	L1TD1	Nsym	.			
7	107718670	G	A	exonic	LAMB4	Nsym	.			
X	78011018	A	G	exonic	LPAR4	Nsym	.			
6	161528978	G	A	exonic	MAP3K4	Nsym	.			
6	131931221	T	C	exonic	MED23	Nsym	.			
4	88766379	C	G	exonic	MEPE	stopgain	.			
12	86373908	C	T	exonic	MGAT4C	Nsym	rs770702724			
20	3026740	G	C	exonic	MRPS26	Nsym	.			
3	49725021	C	T	exonic	MST1	Nsym	rs114429531			Yes
4	187455234	A	T	exonic	MTNR1A	Nsym	rs576228243			

7	100646974	G	A	exonic	MUC12	Nsym	rs112087460			
7	100679390	A	G	exonic	MUC17	Nsym	rs150982179			Yes
3	195452645	G	A	exonic	MUC20	Nsym	rs200616967			Yes
11	1264393	A	G	exonic	MUC5B	Nsym	rs753807150			
11	1264418	C	G	exonic	MUC5B	Nsym	rs201556927			
17	8526330	T	C	exonic	MYH10	Nsym	.			
7	45123906	T	G	exonic	NACAD	Nsym	rs61740887			
16	5083441	C	G	exonic	NAGPA	Nsym	.			Yes
10	72025930	A	C	exonic	NPFFR1	Nsym	.			
16	14859247	C	T	exonic	NPIPA2	Nsym	rs753086489			
9	99699418	C	A	exonic	NUTM2G	Nsym	.			
1	159283843	C	A	exonic	OR10J3	Nsym	.			
14	20389481	C	T	exonic	OR4K5	Nsym	rs372302210			
5	140221312	A	C	exonic	PCDHA8	Nsym	.	Tier1		
19	15580517	G	A	exonic	PGLYRP2	Nsym	.			
17	27277948	C	T	exonic	PHF12	Nsym	.		Yes	
10	3190389	C	T	exonic	PITRM1	Nsym	.			
8	110467001	A	T	exonic	PKHD1L1	Nsym	.			
12	133240651	T	C	exonic	POLE	Nsym	rs539312991	Tier1		
7	72412591	G	A	exonic	POM121	Nsym	rs201184041			
2	130832415	C	T	exonic	POTEF	Nsym	rs757139860			
22	16287673	C	G	exonic	POTEH	Nsym	.			
3	12421223	G	A	exonic	PPARG	Nsym	.	Tier 1	Yes	
1	12921443	T	C	exonic	PRAMEF2	Nsym	rs201306561			
8	27293809	C	T	exonic	PTK2B	Nsym	rs201274282		Yes	
12	9317737	G	A	exonic	PZP	Nsym	rs150415784			
8	74233189	G	C	exonic	RDH10	Nsym	.			
22	29837565	C	G	exonic	RFPL1	Nsym	.			
19	49120614	T	A	exonic	RPL18	Nsym	.		Yes	
15	33873844	G	T	exonic	RYR3	Nsym	.			
2	120209639	C	T	exonic	SCTR	Nsym	.			
9	135163738	T	C	exonic	SETX	Nsym	.			Yes
19	52133292	A	G	exonic	SIGLEC5	Nsym	rs1973019			
9	131115799	G	A	exonic	SLC27A4	Nsym	rs148488076			
6	107956183	C	T	exonic	SOBP	Nsym	.		Yes	
15	43892272	T	C	exonic	STRC	Nsym	.			
6	46658843	C	T	exonic	TDRD6	Nsym	rs369734344			
16	4312380	T	A	exonic	TFAP4	Nsym	.			
1	43784988	G	C	exonic	TIE1	Nsym	.			
5	72419763	C	T	exonic	TMEM171	Nsym	rs767193589			
19	55889178	C	T	exonic	TMEM190	Nsym	rs146395464			
21	19685332	A	T	exonic	TMPRSS15	Nsym	.			
13	43180704	A	C	exonic	TNFSF11	Nsym	.			
2	210777327	G	T	exonic	UNC80	Nsym	.			
4	9213373	C	T	exonic	USP17L10	Nsym	rs759473039			
2	61528508	A	C	exonic	USP34	Nsym	.			Yes
4	1957776	G	T	exonic	WHSC1	Nsym	.	Tier1		
X	100177851	C	T	exonic	XKRX	Nsym	.		Yes	
19	38057101	T	C	exonic	ZNF571	Nsym	.			
1	151261719	C	A	exonic	ZNF687	Nsym	.			
19	58385906	G	C	exonic	ZNF814	Nsym	rs545083939			

Table A.4: Mutations in coding regions with functional significance: Functional impact was assessed using wANNOWAR[11].

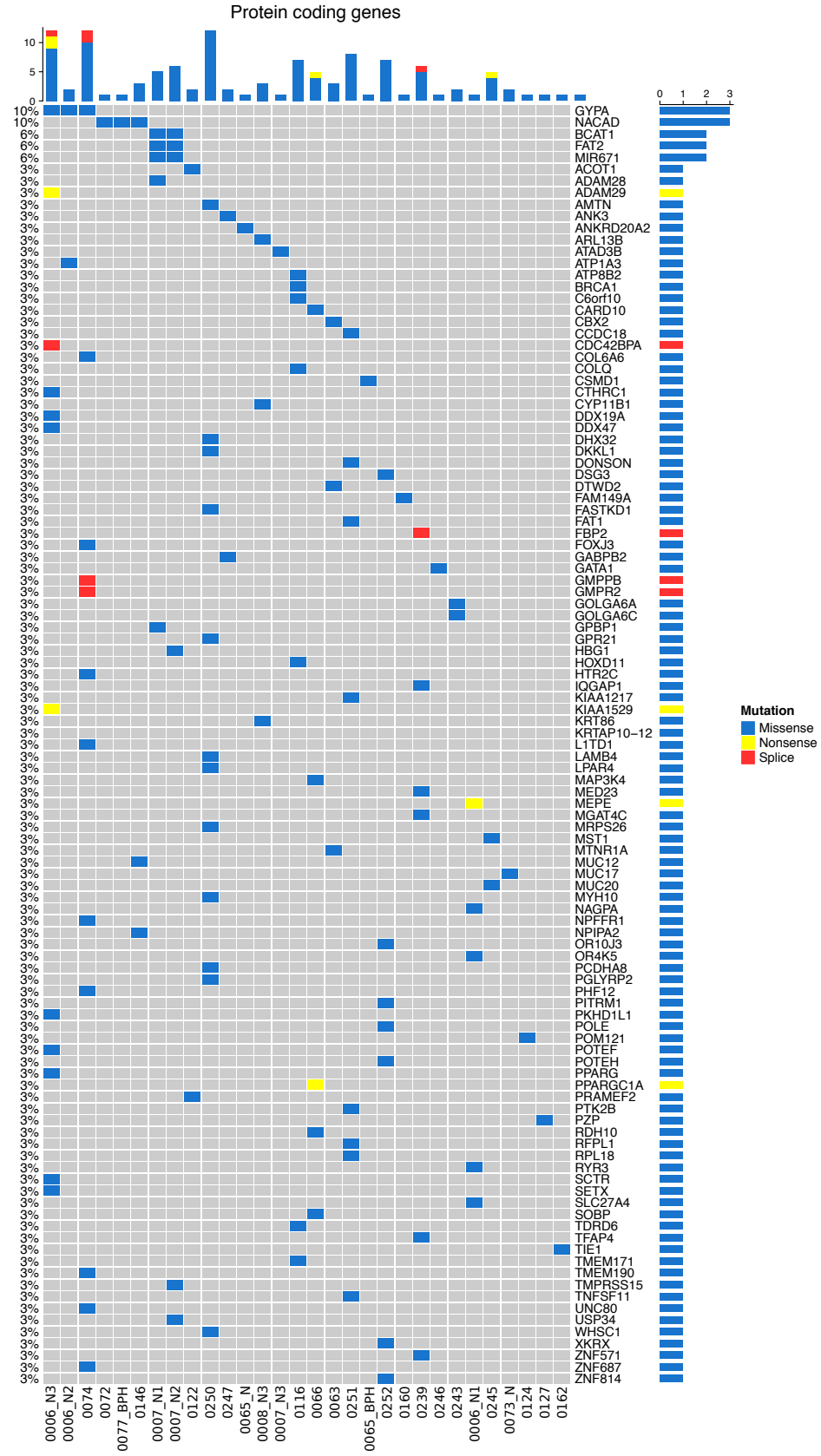


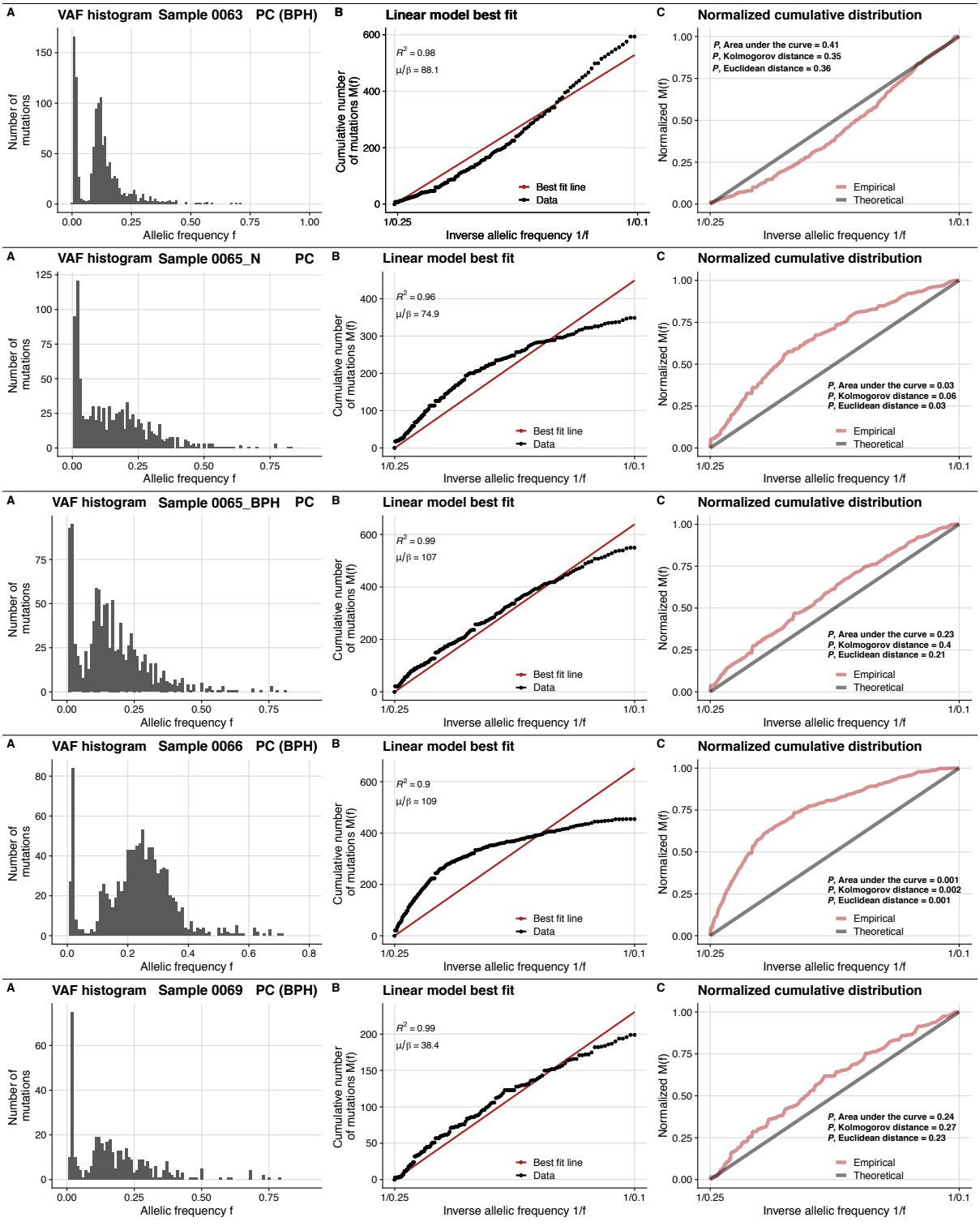
Figure A.1: Protein coding genes across all samples with predicted functional impact.

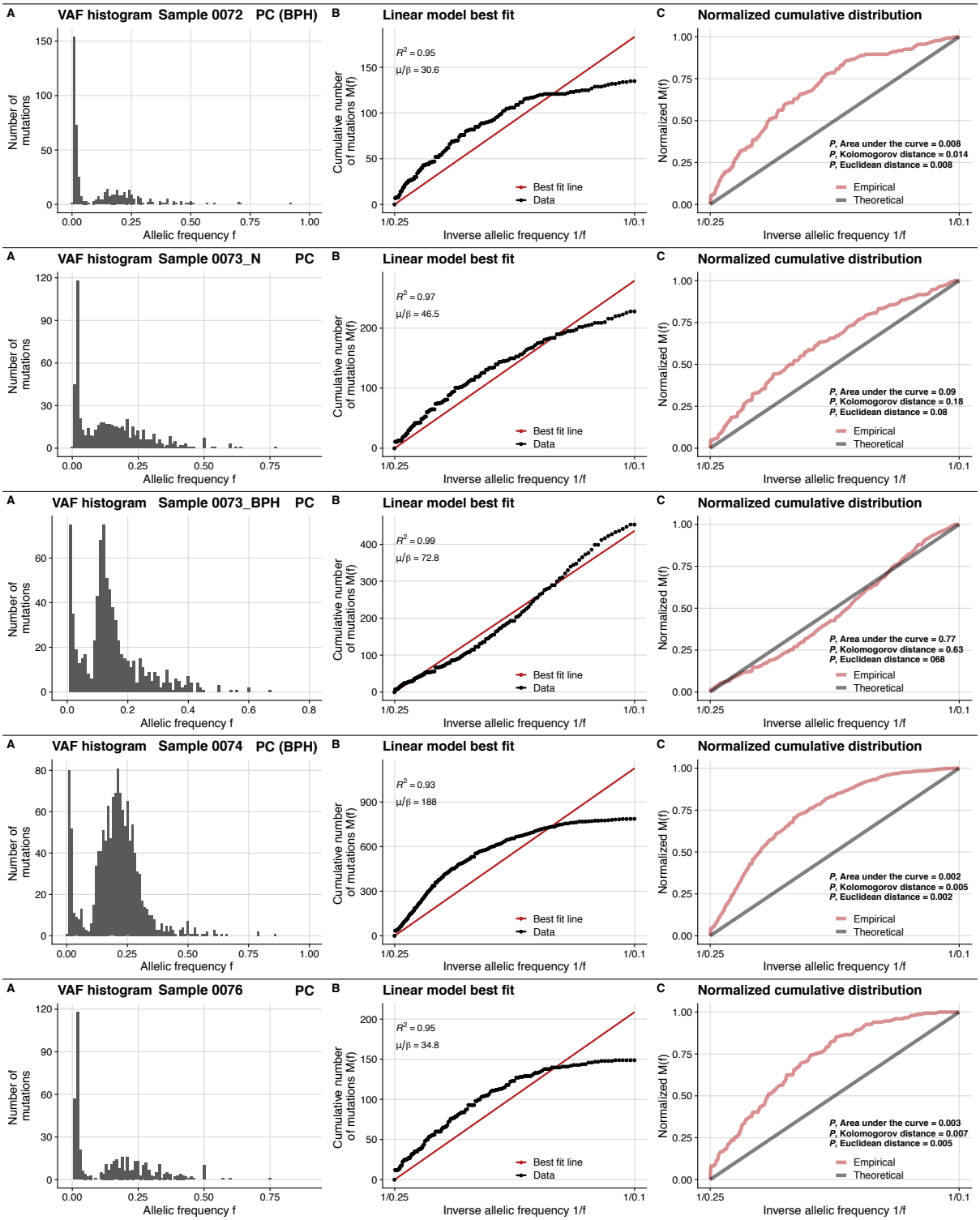
Samples	Chr	Position	Distance	Number of mutations within 1kb
162	17	50057108	2643	27
65	17	50057108	2643	26
120	21	20809988	2728	20
162	8	37614329	1062	20
162	21	20809988	2728	20
162	4	123264467	2866	17
6_T3	4	123264467	2866	17
77	1	238324648	3419	14
162	1	238324648	3419	14
162	1	242076771	490	13
162	6	116067481	938	12
120	21	27943158	1192	11
162	17	29253772	461	11
162	21	27943158	1192	11
162	Y	58909134	490	11
6_T3	17	29253772	461	11
6_T1	8	37614329	1062	11
120	21	27939898	2733	10
162	21	27939898	2733	10

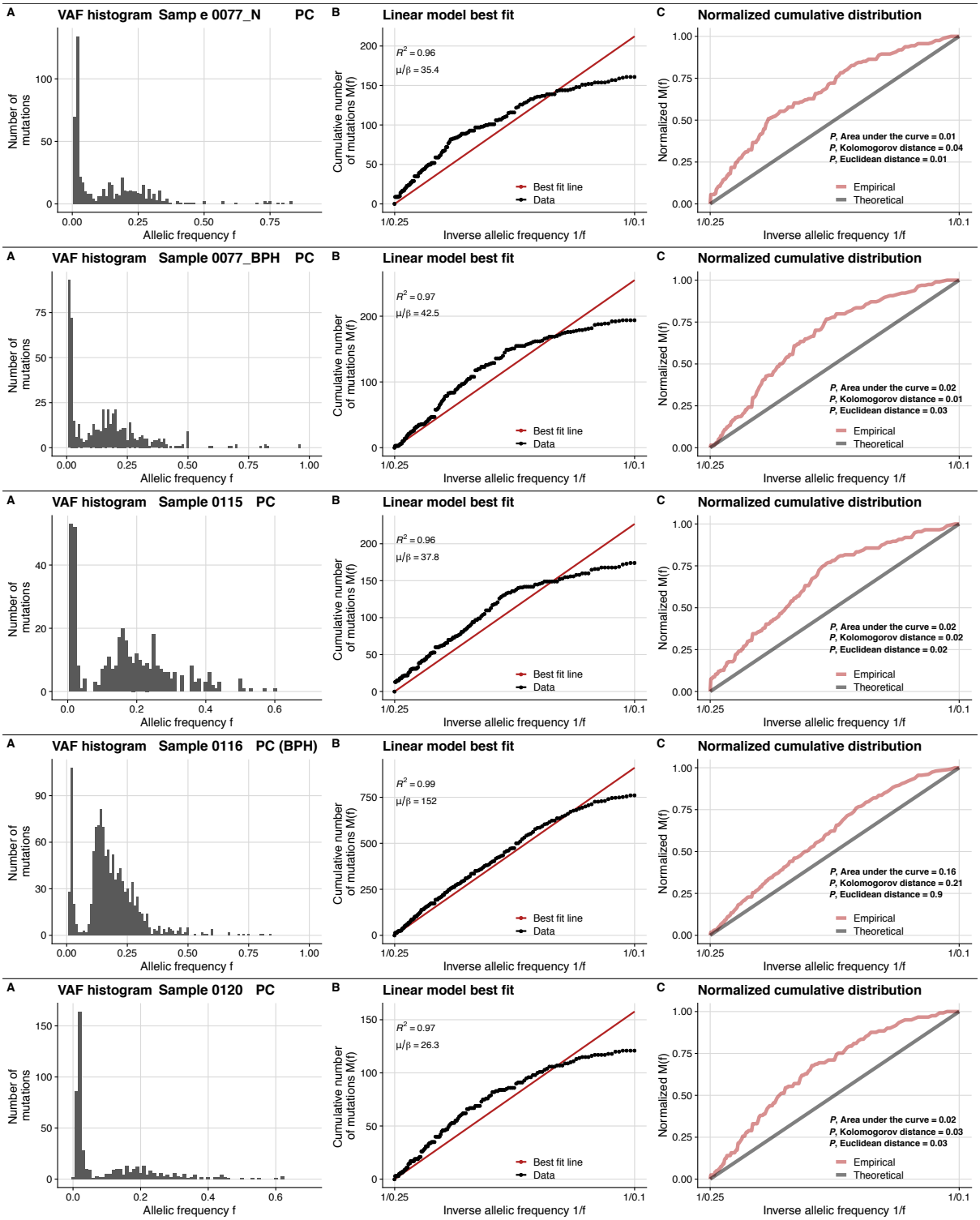
Table A.5: Kataegis events in tumour samples

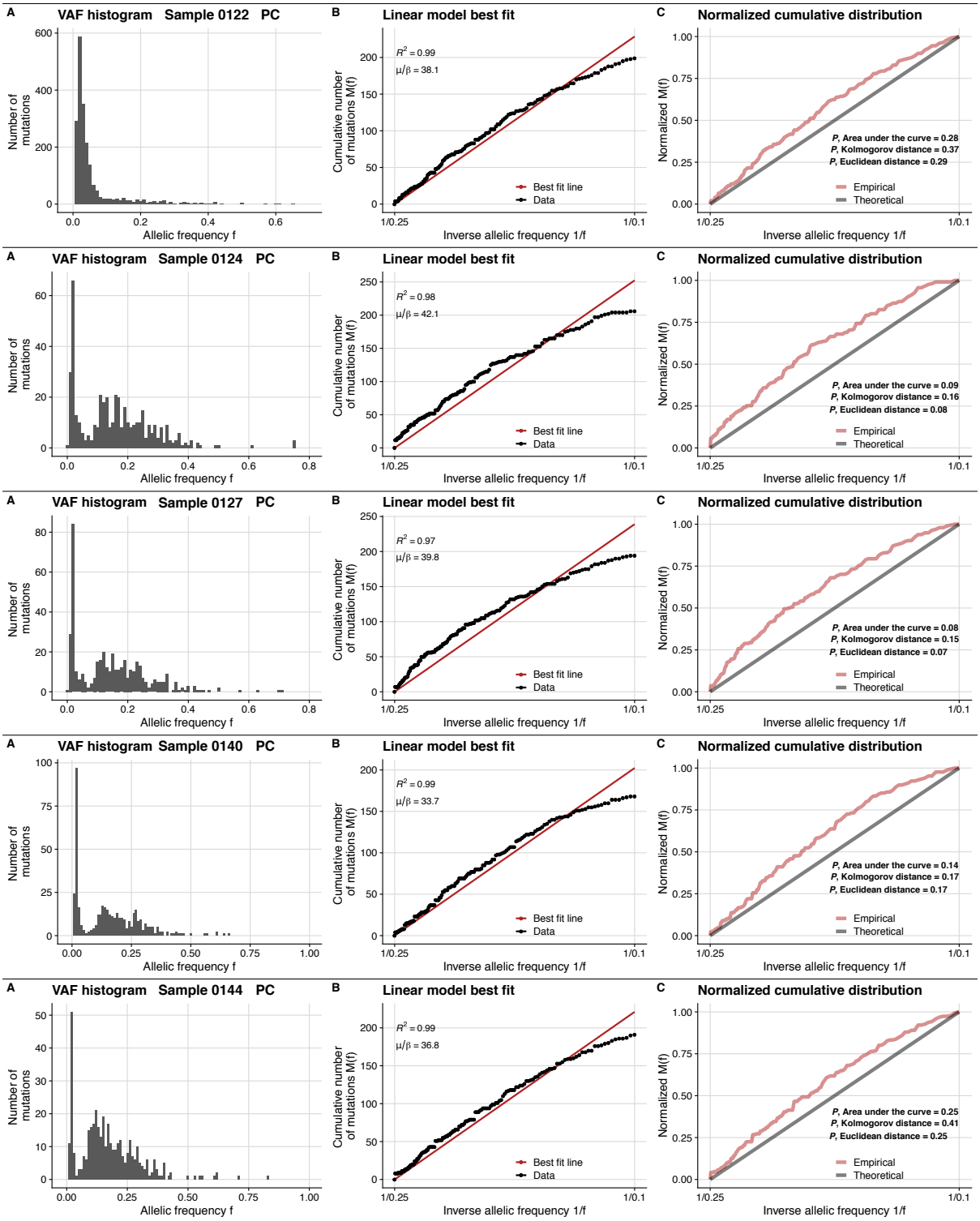
APPENDIX B

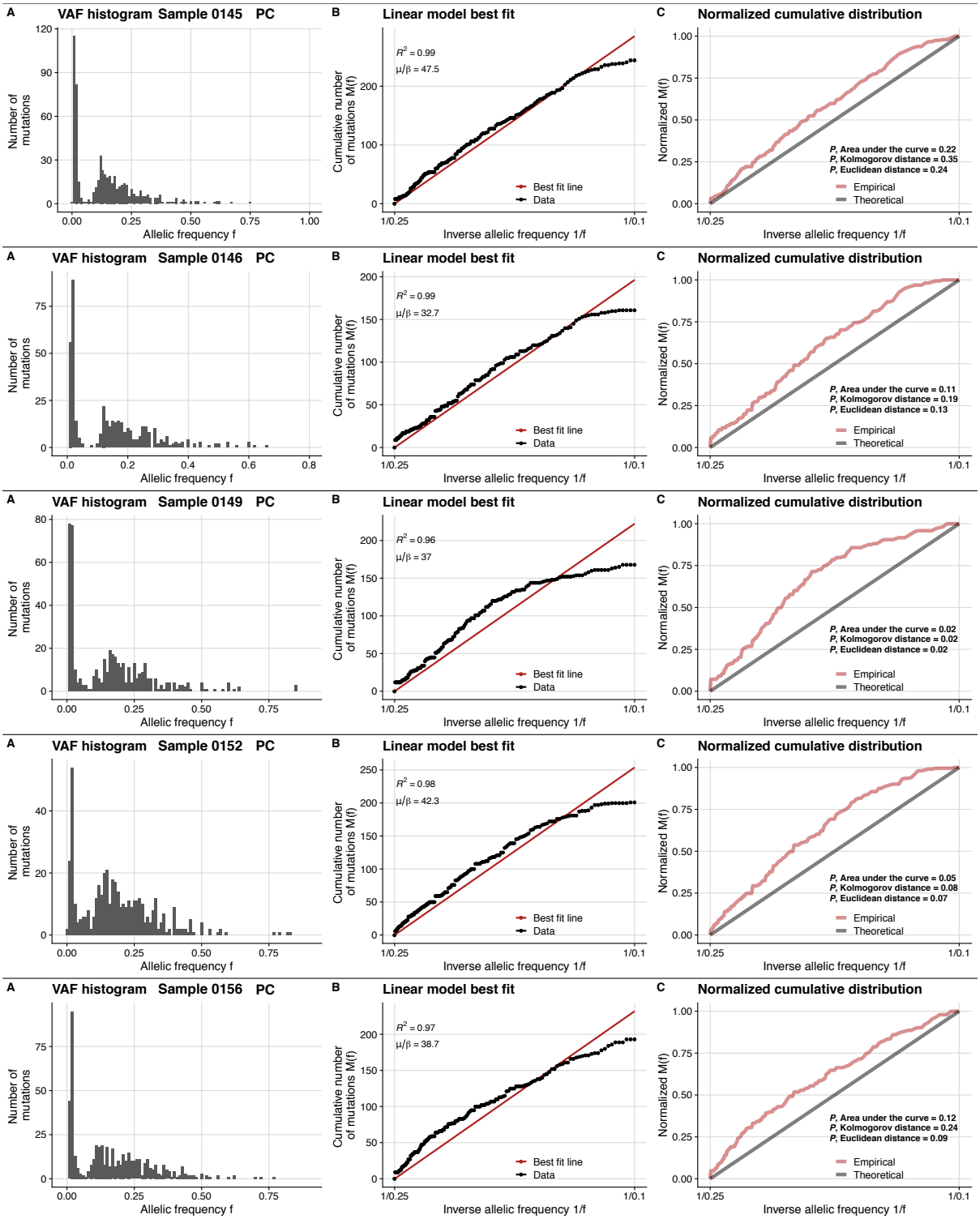
SUPPLEMENTARY TABLES ANF FIGURES FOR CHAPTER 4

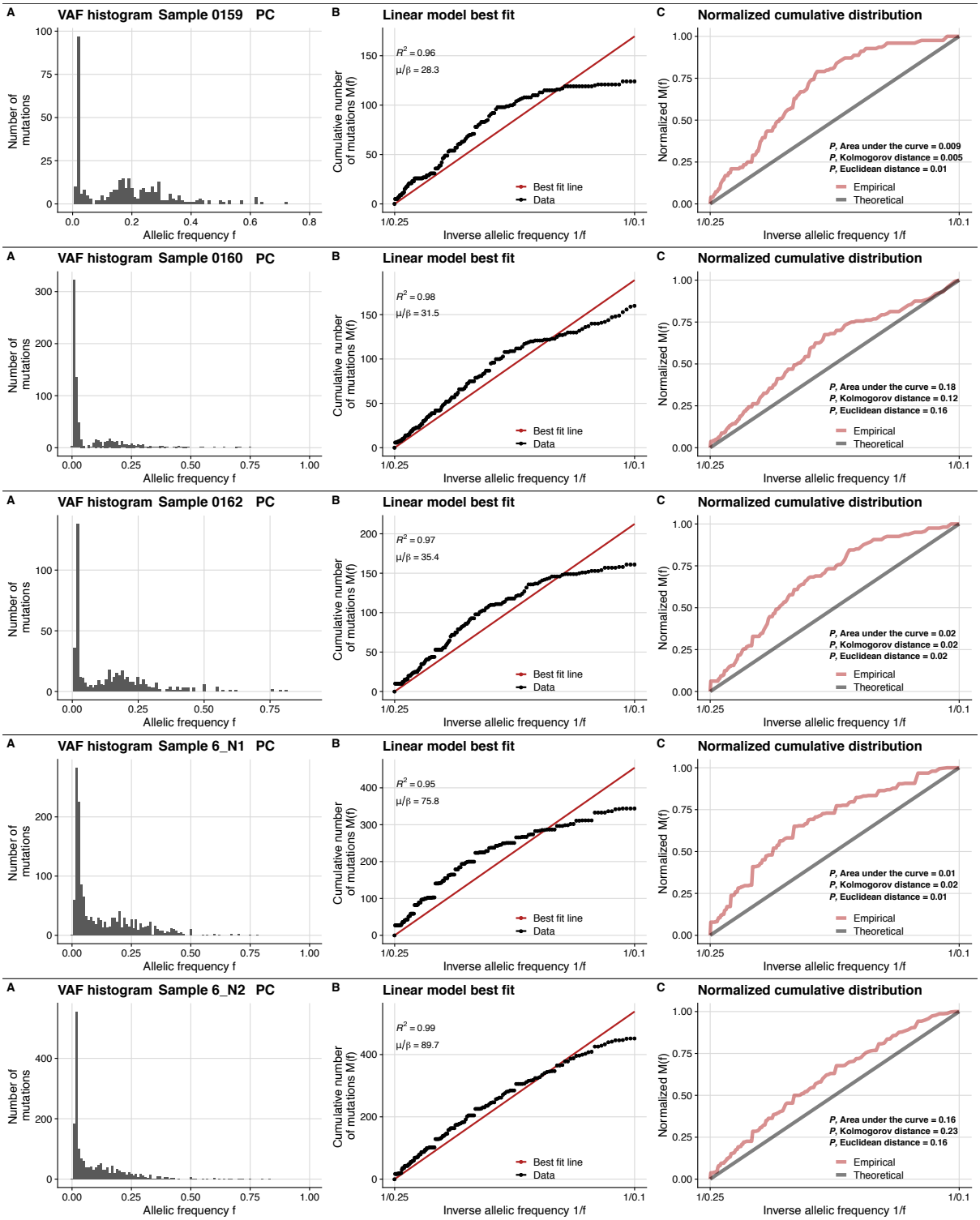


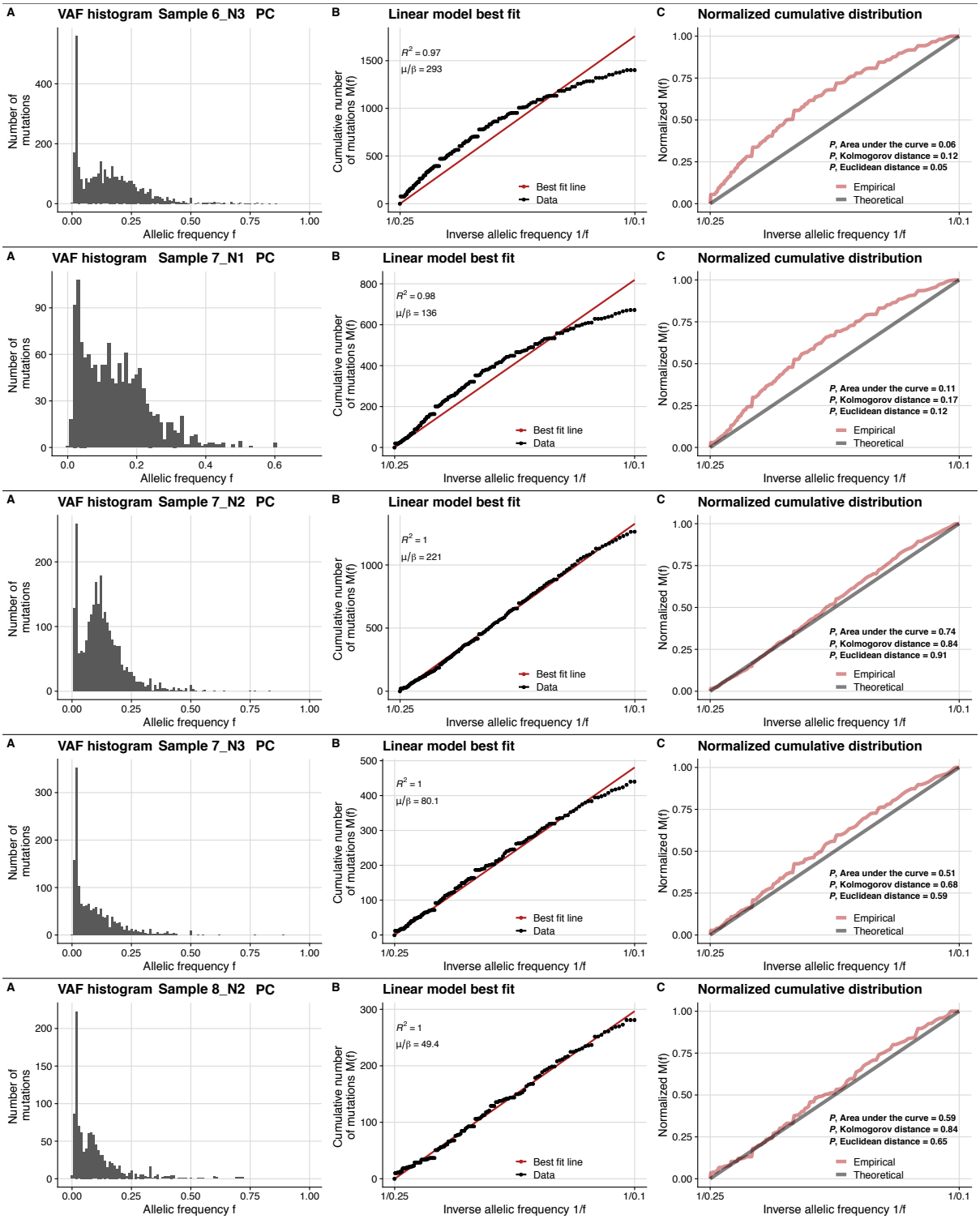

A
VAF histogram Sample 0072 PC (BPH)
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.95
μ/β = 30.6
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.008
P, Kolomogorov distance = 0.014
P, Euclidean distance = 0.008
Empirical
Theoretical
A
VAF histogram Sample 0073_N PC
B
Linear model best fit
R^2 = 0.97
μ/β = 46.5
C
Normalized cumulative distribution
P, Area under the curve = 0.09
P, Kolomogorov distance = 0.18
P, Euclidean distance = 0.08
A
VAF histogram Sample 0073_BPH PC
B
Linear model best fit
R^2 = 0.99
μ/β = 72.8
C
Normalized cumulative distribution
P, Area under the curve = 0.77
P, Kolomogorov distance = 0.63
P, Euclidean distance = 068
A
VAF histogram Sample 0074 PC (BPH)
B
Linear model best fit
R^2 = 0.93
μ/β = 188
C
Normalized cumulative distribution
P, Area under the curve = 0.002
P, Kolomogorov distance = 0.005
P, Euclidean distance = 0.002
A
VAF histogram Sample 0076 PC
B
Linear model best fit
R^2 = 0.95
μ/β = 34.8
C
Normalized cumulative distribution
P, Area under the curve = 0.003
P, Kolomogorov distance = 0.007
P, Euclidean distance = 0.005

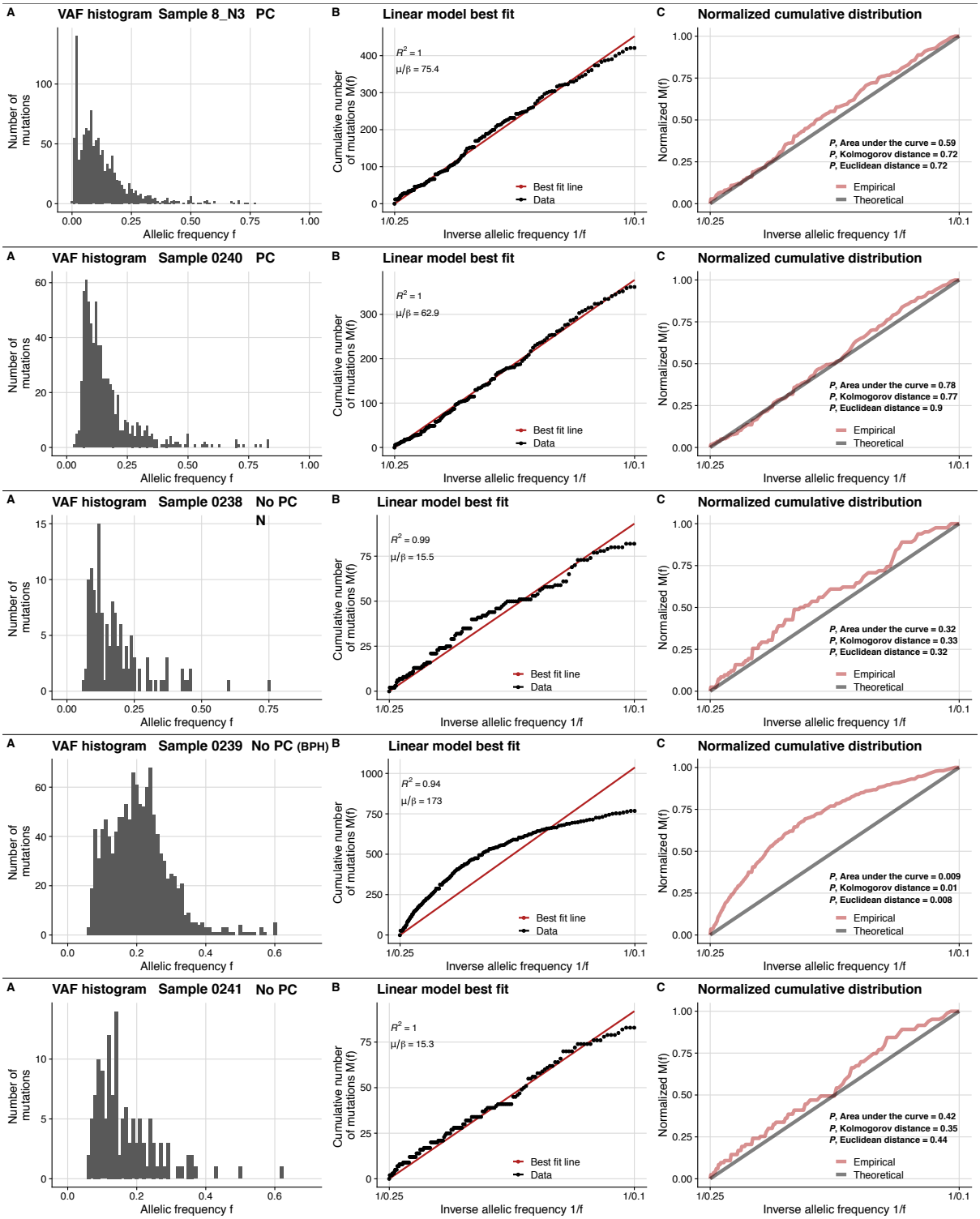

A
VAF histogram Samp e 0077_N PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.96
μ/β = 35.4
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.01
P, Kolomogorov distance = 0.04
P, Euclidean distance = 0.01
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0077_BPH PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.97
μ/β = 42.5
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.02
P, Kolomogorov distance = 0.01
P, Euclidean distance = 0.03
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0115 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.96
μ/β = 37.8
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.02
P, Kolomogorov distance = 0.02
P, Euclidean distance = 0.02
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0116 PC (BPH)
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 152
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.16
P, Kolomogorov distance = 0.21
P, Euclidean distance = 0.9
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0120 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.97
μ/β = 26.3
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.02
P, Kolomogorov distance = 0.03
P, Euclidean distance = 0.03
Empirical
Theoretical
Inverse allelic frequency 1/f

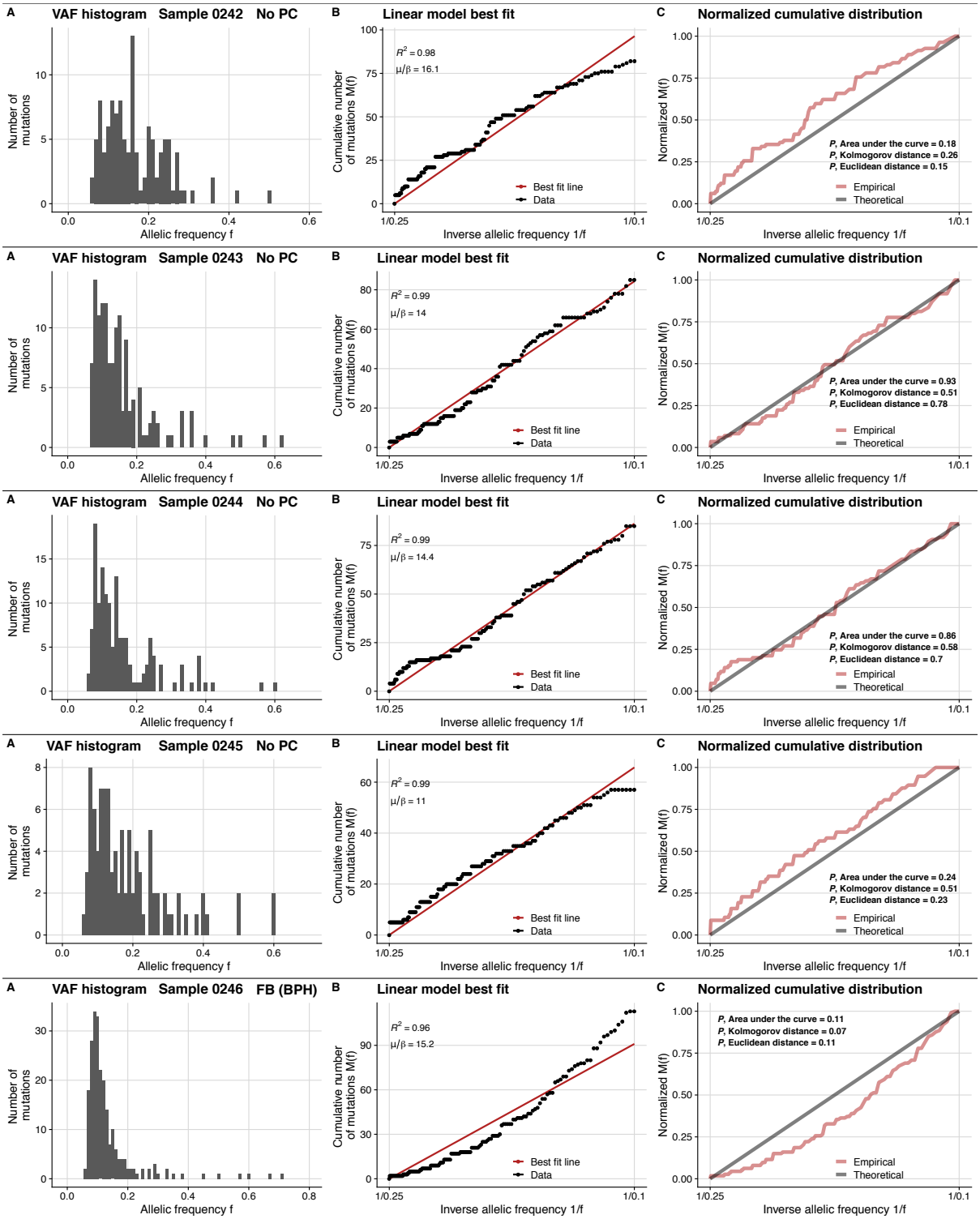

A
VAF histogram Sample 0122 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 38.1
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.28
P, Kolmogorov distance = 0.37
P, Euclidean distance = 0.29
Empirical
Theoretical
A
VAF histogram Sample 0124 PC
B
Linear model best fit
R^2 = 0.98
μ/β = 42.1
C
Normalized cumulative distribution
P, Area under the curve = 0.09
P, Kolmogorov distance = 0.16
P, Euclidean distance = 0.08
A
VAF histogram Sample 0127 PC
B
Linear model best fit
R^2 = 0.97
μ/β = 39.8
C
Normalized cumulative distribution
P, Area under the curve = 0.08
P, Kolmogorov distance = 0.15
P, Euclidean distance = 0.07
A
VAF histogram Sample 0140 PC
B
Linear model best fit
R^2 = 0.99
μ/β = 33.7
C
Normalized cumulative distribution
P, Area under the curve = 0.14
P, Kolmogorov distance = 0.17
P, Euclidean distance = 0.17
A
VAF histogram Sample 0144 PC
B
Linear model best fit
R^2 = 0.99
μ/β = 36.8
C
Normalized cumulative distribution
P, Area under the curve = 0.25
P, Kolmogorov distance = 0.41
P, Euclidean distance = 0.25


A
VAF histogram Sample 0145 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 47.5
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.22
P, Kolmogorov distance = 0.35
P, Euclidean distance = 0.24
Empirical
Theoretical
A
VAF histogram Sample 0146 PC
B
Linear model best fit
R^2 = 0.99
μ/β = 32.7
C
Normalized cumulative distribution
P, Area under the curve = 0.11
P, Kolmogorov distance = 0.19
P, Euclidean distance = 0.13
A
VAF histogram Sample 0149 PC
B
Linear model best fit
R^2 = 0.96
μ/β = 37
C
Normalized cumulative distribution
P, Area under the curve = 0.02
P, Kolmogorov distance = 0.02
P, Euclidean distance = 0.02
A
VAF histogram Sample 0152 PC
B
Linear model best fit
R^2 = 0.98
μ/β = 42.3
C
Normalized cumulative distribution
P, Area under the curve = 0.05
P, Kolmogorov distance = 0.08
P, Euclidean distance = 0.07
A
VAF histogram Sample 0156 PC
B
Linear model best fit
R^2 = 0.97
μ/β = 38.7
C
Normalized cumulative distribution
P, Area under the curve = 0.12
P, Kolmogorov distance = 0.24
P, Euclidean distance = 0.09


A
VAF histogram Sample 0159 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.96
μ/β = 28.3
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.009
P, Kolmogorov distance = 0.005
P, Euclidean distance = 0.01
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0160 PC
B
Linear model best fit
R^2 = 0.98
μ/β = 31.5
C
Normalized cumulative distribution
P, Area under the curve = 0.18
P, Kolmogorov distance = 0.12
P, Euclidean distance = 0.16
A
VAF histogram Sample 0162 PC
B
Linear model best fit
R^2 = 0.97
μ/β = 35.4
C
Normalized cumulative distribution
P, Area under the curve = 0.02
P, Kolmogorov distance = 0.02
P, Euclidean distance = 0.02
A
VAF histogram Sample 6_N1 PC
B
Linear model best fit
R^2 = 0.95
μ/β = 75.8
C
Normalized cumulative distribution
P, Area under the curve = 0.01
P, Kolmogorov distance = 0.02
P, Euclidean distance = 0.01
A
VAF histogram Sample 6_N2 PC
B
Linear model best fit
R^2 = 0.99
μ/β = 89.7
C
Normalized cumulative distribution
P, Area under the curve = 0.16
P, Kolmogorov distance = 0.23
P, Euclidean distance = 0.16


A
VAF histogram Sample 6_N3 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.97
μ/β = 293
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.06
P, Kolmogorov distance = 0.12
P, Euclidean distance = 0.05
Empirical
Theoretical
A
VAF histogram Sample 7_N1 PC
B
Linear model best fit
R^2 = 0.98
μ/β = 136
C
Normalized cumulative distribution
P, Area under the curve = 0.11
P, Kolmogorov distance = 0.17
P, Euclidean distance = 0.12
A
VAF histogram Sample 7_N2 PC
B
Linear model best fit
R^2 = 1
μ/β = 221
C
Normalized cumulative distribution
P, Area under the curve = 0.74
P, Kolmogorov distance = 0.84
P, Euclidean distance = 0.91
A
VAF histogram Sample 7_N3 PC
B
Linear model best fit
R^2 = 1
μ/β = 80.1
C
Normalized cumulative distribution
P, Area under the curve = 0.51
P, Kolmogorov distance = 0.68
P, Euclidean distance = 0.59
A
VAF histogram Sample 8_N2 PC
B
Linear model best fit
R^2 = 1
μ/β = 49.4
C
Normalized cumulative distribution
P, Area under the curve = 0.59
P, Kolmogorov distance = 0.84
P, Euclidean distance = 0.65


A
VAF histogram Sample 8_N3 PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 1
μ/β = 75.4
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.59
P, Kolmogorov distance = 0.72
P, Euclidean distance = 0.72
Empirical
Theoretical
A
VAF histogram Sample 0240 PC
B
Linear model best fit
R^2 = 1
μ/β = 62.9
C
Normalized cumulative distribution
P, Area under the curve = 0.78
P, Kolmogorov distance = 0.77
P, Euclidean distance = 0.9
A
VAF histogram Sample 0238 No PC N
B
Linear model best fit
R^2 = 0.99
μ/β = 15.5
C
Normalized cumulative distribution
P, Area under the curve = 0.32
P, Kolmogorov distance = 0.33
P, Euclidean distance = 0.32
A
VAF histogram Sample 0239 No PC (BPH)
B
Linear model best fit
R^2 = 0.94
μ/β = 173
C
Normalized cumulative distribution
P, Area under the curve = 0.009
P, Kolmogorov distance = 0.01
P, Euclidean distance = 0.008
A
VAF histogram Sample 0241 No PC
B
Linear model best fit
R^2 = 1
μ/β = 15.3
C
Normalized cumulative distribution
P, Area under the curve = 0.42
P, Kolmogorov distance = 0.35
P, Euclidean distance = 0.44


A
VAF histogram Sample 0242 No PC
Number of mutations
Allelic frequency f
B
Linear model best fit
R^2 = 0.98
μ/β = 16.1
Cumulative number of mutations M(f)
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
Normalized M(f)
P, Area under the curve = 0.18
P, Kolmogorov distance = 0.26
P, Euclidean distance = 0.15
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0243 No PC
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 14
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
P, Area under the curve = 0.93
P, Kolmogorov distance = 0.51
P, Euclidean distance = 0.78
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0244 No PC
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 14.4
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
P, Area under the curve = 0.86
P, Kolmogorov distance = 0.58
P, Euclidean distance = 0.7
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0245 No PC
Allelic frequency f
B
Linear model best fit
R^2 = 0.99
μ/β = 11
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
P, Area under the curve = 0.24
P, Kolmogorov distance = 0.51
P, Euclidean distance = 0.23
Empirical
Theoretical
Inverse allelic frequency 1/f
A
VAF histogram Sample 0246 FB (BPH)
Allelic frequency f
B
Linear model best fit
R^2 = 0.96
μ/β = 15.2
Best fit line
Data
Inverse allelic frequency 1/f
C
Normalized cumulative distribution
P, Area under the curve = 0.11
P, Kolmogorov distance = 0.07
P, Euclidean distance = 0.11
Empirical
Theoretical
Inverse allelic frequency 1/f

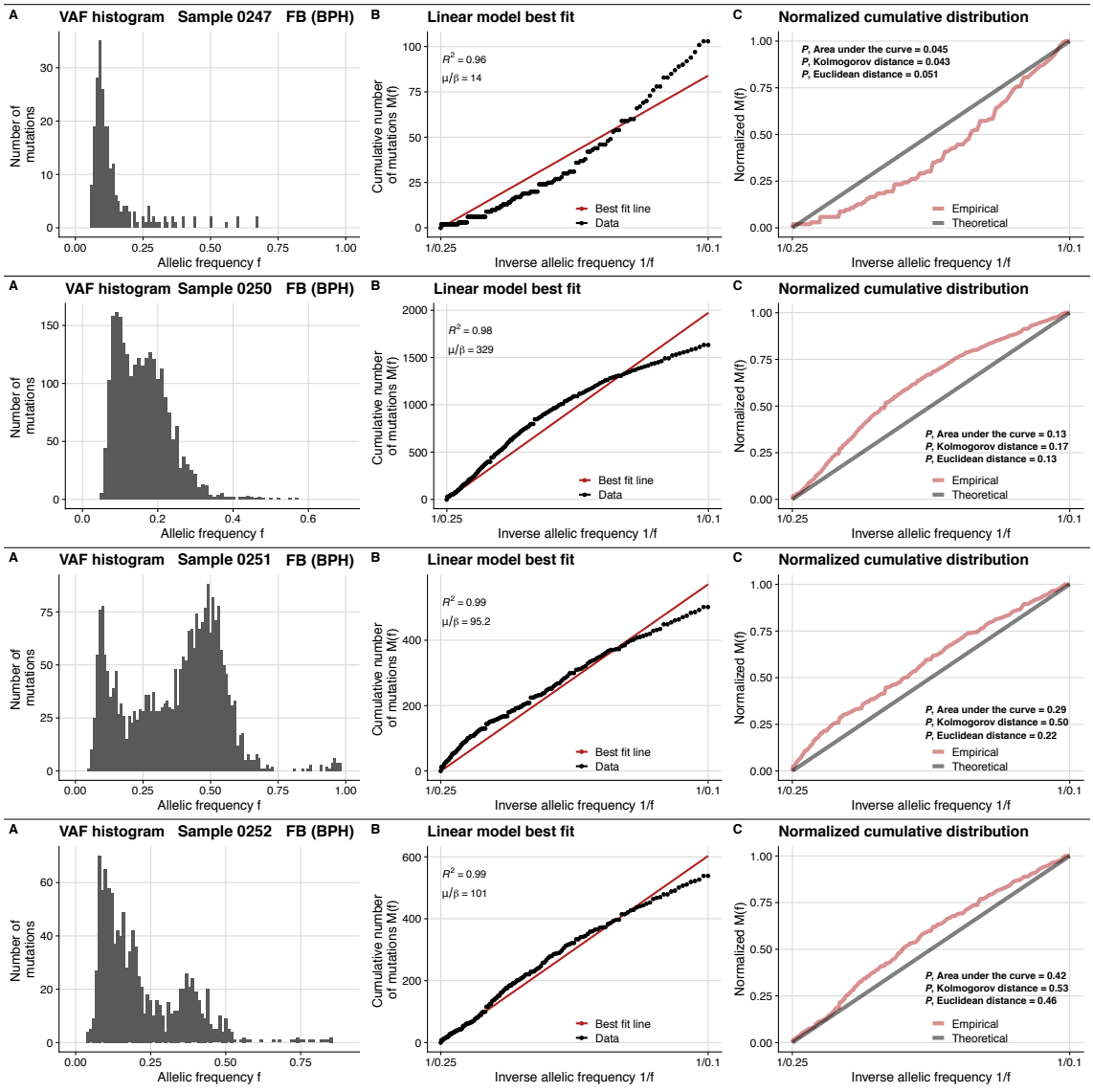


Figure B.1: (A) VAF distribution of all samples. (B) Cumulative distribution and least squares best fit line with R^2 values and estimated mutation rates ($\boldsymbol{\mu/\beta}$). (C) Normalized cumulative distribution and Area under the curve, Kolmogorov distance and Euclidean distance value.

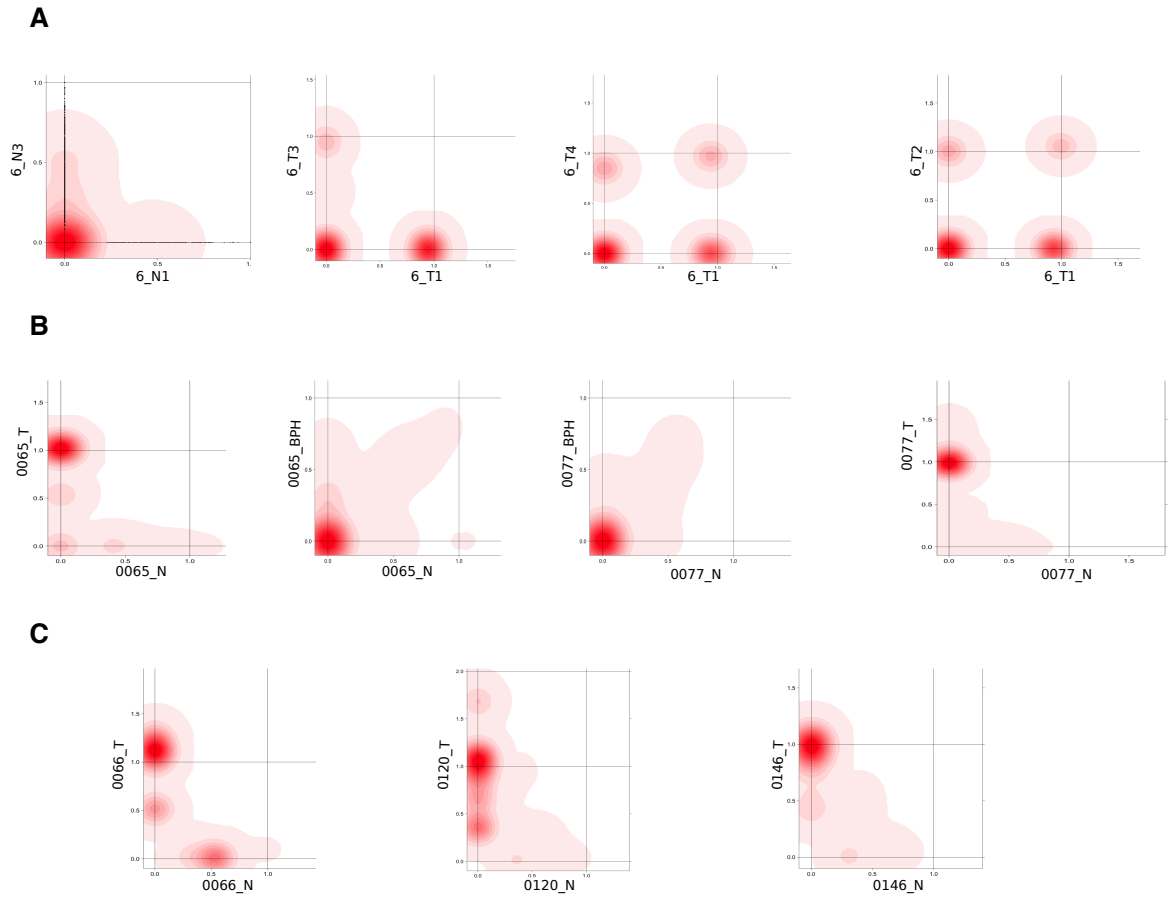


Figure B.2: 2D density plots of the posterior distribution of the fraction of cells (modelled using the Bayesian Dirichlet process) harbouring a mutation for 6 samples. (A) In case 6 we observe shared clonal mutations (6_T1/6_T4; 6_T1_6_2), unique clones in 6_T1, 6_T3) and two unique subclones (6_N1 and 6_N3). (B) BPH cases 0065 and 0077 show shared subclonal mutations in cases 0065 (N/BPH) and 0077 (N/BPH), unrelated subclones in 0065_N, 0077_N and unrelated clones in 0065_T and 0077_T. (C) Single cases 0066, 0120 and 0146 present a tumour clonal cluster with an extra subclonal tumour subclone in ~ 50% of cells and a subclonal cluster in the normal sample.

	cluster	T1	T2	T3	T4	N1	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
6	2	0	0	0	0	0	0	0.346	319	207	1	2.96	34% in N3	
	4	0	0	0	0	0.481	0	0	442	54	1	4.11	48% in N1	*RYR3, MEPE*
	7	0	1.030	0	0	0	0	0	1520	287	8	14.14	100% in T2	
	8	0	0	0	0.463	0	0	0	308	114	5	2.86	46% in T4	
	9	0.931	0	0	0	0	0	0	2538	360	46	23.61	93% in T1	*HIST1H2BJ, NUBPL, SRF1, GNPDA2, SLC13A3, DUOX1, CCR3, DCC, RRP9, CDK2AP2, ACSS3, PTEN, SORBS1, TTC35, SETX, ZBBX, DNPEP, RYR2, C19orf38, AGAP2, ASTN1*
	10	0	0	0	0.955	0	0	0	1300	113	5	12.09	95% in T4	
	19	0.950	0	0	0.983	0	0	0	1415	116	11	13.16	95% in T1, 100% in T4	*CNGA4, MYO1F, HS35T4, PDGFRB, CALCRL, AVPR1B*
	25	1.016	1.091	0	0	0	0	0	1108	111	21	10.3	100% in T1, 100% in T2	*PTPRO, DDX4, ANKRD17, LCA5, SLC22A16, CDC5L*
	26	0	0	1.129	0	0	0	0	554	185	46	5.15	100% in T3	
	3	0	0	0	0	0	0.353	0	332	316	0	2.62	35 % in N2	
	1	0	0	0	0	0	0	0.230	751	208	0	5.93	23 % in N3	

Table B.1: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants for case 0006 (prostate cancer patient with multiple normal and a tumour samples). Two subclone in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	T1	T2	T3	T4	T5	N1	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
7	1	1.184	0.575	0.003	0.008	0.005	0.002	0.003	0.003	1148	43	0	21.71	100% in T1, 56% in T2	*HIC1, APOL3, KIT, TRIM49, KIAA1614, NCOA7*
	2	0.014	0.003	0.002	0.011	0.005	0.355	0.304	0.011	622	4	0	11.76	35% in N1, 30% in N2	*ADAM28, BCAT1, FAT2*
	6	0.010	0.001	0.003	0.001	0.733	0.001	0.003	0.002	253	67	8	4.78	73% in T5	
	10	1.242	0	0.003	0.011	0.003	0.001	0.002	0.003	395	87	8	7.47	100% in T1	*RPL11*
	12	0.020	0.004	0.004	0	0.620	0.001	0.002	0.002	209	67	7	3.95	62% in T5	
	13	0.052	0.812	0.003	0.011	0.004	0.001	0.002	0.002	980	67	15	18.53	81% in T2	
	14	0.010	0.002	0.964	0.009	0.006	0.001	0.003	0.002	683	133	1	12.92	96% in T3	
	20	0.010	0.003	0.002	0.947	0.966	0.001	0.003	0.003	273	9	0	5.16	94% in T4, 96% in T5	
	3	0.010	0.006	0.014	0.027	0.009	0.000	0.180	0.204	237	206	0	3.94	18 % in N2, 20 % in N3	
	9	0.006	0.003	0.002	0.015	0.004	0.015	0.216	0.003	282	627	0	4.69	22% in N2	
	40	0.019	0	0.005	0.024	0.006	0.21	0.3035	0.0108	64	3	0	1	21% in N1, 30% in N2	
	21	1.291	0.004	0.002	0.009	0.004	0	0.004	0.0033	63	86	7	1	100% in T1	*C12orf63*
	17	0.005	0	0.003	1.506	0.974	0	0.0011	0.0004	78	9	0	1.3	1005 in T4, 97% in T5	

Table B.2: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants for case 0007 (prostate cancer patient with multiple normal and a tumour samples). Two subclone in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	T1	T2	T3	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
8	1	0.977	1.000	0	0	0	1526	71	25	41.41112619	100 % in T1, 100 % in T2	*PHRF1, C22orf43, CEACAM1, SIGLEC11, ZCCHC8, PHF10, RASAL2,*
	21	0.518	0.013	0.014	0.005	0.004	204	111	3	5.535956581	51 % in T1	
	35	0.926	0.545	0.004	0	0	67	72	24		93 % in T1, 55 % in T2	*MIA3*
	33	0.003	0.555	0.009	0.001	0.002	87	100	0	2.360922659	56 % in T2	
	24	0.433	0	0	0	0	68	110	3		43% in T1	*AASDH*
	27	0	0	0	0.200	0	54	461	0	1.465400271	20 % in N2	
	40	1.015	0.993	0.138	0	0	44	2	0	1.194029851	100 %in T1, 100 % in T2, 13 % in T3	
	18	0.976	0.959	0.06	0.016	0.004	261	1	0	7	98% in T1, 96% in T2, 6% in T3	
	7	0.029	0.037	0.065	0.184	0.191	206	0	0	5.6	3% in T1, 4% in T2, 6% in T3, 18% in N2, 19% in N3	
	12	0.063	0.068	0.223	0.157	0.254	100	0	0	2.7	6% in T1, 7% in T2, 22% in T3, 16% in N2, 25% in N3	

Table B.3: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants for case 0008 (prostate cancer patient with multiple normal and a tumour samples). One subclone in the normal sample was not considered due to suspected evidence of neutral evolution (green). Cluster 12 (grey) was discarded because it was not possible to include it in the phylogeny according to the pigeonhole principle (see section 4.4.4.3). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	N	BPH	T	SNVs	Indels	Svs	Total SNVs (%)	CCF values	Coding genes
0065	1	0	0	0.534	275	205	16	6.66	53% in Tb	*CHD1, ZNF616, SLAM9*
	2	0	0	1.018	2548	204	16	60.34	100% in Ta	*AGRN, NAP1L4, DNAH9, OR4C5, A1CF, CES1, KIAA0182, CHST10, RANBP2, PLBD2, COL29A1, LRP1B, FBXO5, OBSCN*
	3	0	0.295	0	419	386	0	11.75	29 % in BPH	
	4	0.396	0	0	170	274	0	11.76	39% in N	*ANKRD20A2*
	5	0.396	0.389	0	61	94	0	1.74	40% in N, 39% n BPH	
	7	0.595	0.578	0	68	94	0	1.94	60% in N, 58% in BPH	

	cluster	N	BPH	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0073	1	0	0	0.526	102	104	32	3.69	52% in Tb	
	2	0	0	0.996	1680	103	32	60.91	100% in Ta	*FAN86, SMARCC2, KNCK13, KIAA1409, PLXNC1, TRRAP, SLCO20A1, INPP4B*
	3	0	0.276	0	560	181	0	20.3	28 % in BPH	
	5	0	0.643	0	71	181	0	2.57	64 % in BPH	
	8	0.464	0	0	249	213	0	9.02	46% in N	
	9	0.500	0.440	0	55	102	0	1.99	50% in N, 44% in BPH	
	11	0.717	0.689	0	39	101	0	1.41	71% in N,69% in BPH	

	cluster	N	BPH	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0077	1	0.002	0.001	0.375	118	274	93	3.77	37% in Tb	
	2	0.002	0.002	0.985	2699	274	94	86.34	100% in Ta	*LYPD3, ARID2, SPOP, MAP4, SPATS2, GLIS1, U2AF2, IBTR, C10orf12, SULT1C3, MGMT, TGFBI, MTMR15, APOBEC3F, KIF26B*
	3	0.084	0.333	0.020	113	0	0	3.61	8% in N, 33% in BPH, 2% in T	
	5	0.380	0.335	0.023	88	0	0	2.81	37 % in N, 33 % in BPH, 2 % in T	

Table B.4: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants from prostate cancer patients with a normal, a BPH and a tumour sample. Two subclones in the BPH sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0066	1	0.001	0.512	417	148	41	13.65	51% in Tb	*GLIPR1L2, LPHN2, CCDC18*
	2	0.002	1.105	1783	148	42	58.4	100% in Ta	*C2orf63, USP34, DPYD, CAP29, SLC3A1, ETV3, TIPRL, AP1G2, TIMM50, TGM7, TUBGCP4, NCOR2*
	3	0.528	0	852	469	0	27.9	52% in N	*RDH10, SOBP, MAP3K4*

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0072	1	0.003	0.496	157	171	122	6.6	50% in Tb	
	2	0.004	1.056	2077	171	122	87.41	100% in Ta	*POLRMT, DDX17, RTN2, GDF2, CPT2, PAWR2, RASGRF2, DUSP6, ZNF518, NCKAP5, ARHGEF5*
	3	0.383	0.031	117	7	0	4.92	38% in N,3% in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0074	1	0.002	0.497	1071	98	12	33.46	49% in Tb	*HYAL1, C3orf63, ZNF280D, CDH8, MIR1538, NAA35, IRF5, TRA2B*
	2	0.002	0.953	1082	99	13	33.81	95% in Ta	*DSG2, RYR1, CYP2D6, CRX, MIR1279, CBX8, GOT1, C6orf170, SLC22A5*
	3	0.468	0	1023	680	0	31.96	46% in N	*PHF12, FOXJ3, L1TD1, NPFFR1, COL6A1, ZNF687, UNC80*

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0076	1	0.002	0.880	2453	302	88	92.46	88% in T	KCNV2, FBXO10, EHD4, LRP6, MYOI5A, LRP10, GOLGA8A, CUL7, KLC4, SIGLEC7, BTBD18, EXOC3L, NDUFV1, FAM65A, KCNQ5, PABPC1, TET2, NUP205, DNAI2, NAGPA, GGTLC1, BAIAP2L2
	2	0.370	0.022	189	6	0	7.12	37% in N, 2% in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0115	1	0.005	0.390	70	148	20	3.61	39% in Tb	
	2	0.001	1.051	1638	148	21	84.52	100% in Ta	*MIER3, DOCK7, NOLC1, ALDH2, DYNLT1, TTLL2, TTN, NCK15D*
	3	0.355	0	183	569	0	9.44	35% in N	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	
0120	1	0.003	0.359	620	191	18	21.54	36% in Tb	*APC, CTNNA2, MORC3*
	2	0.002	1.052	2127	191	19	73.9	100% in Ta	*AHSG, PLCL1, KIAA1751, EFCABUB, TAF3, ENO1, NKX3-2, IPO4, CSNK1A1L, CD33, HEATR6*
	3	0.364	0.020	125	2	0	4.34	36% in N, 2% in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0146	1	0.000	0.427	174	94	6	8.73	43% in Tb	*KIAA1683, FSTL5*
	2	0.002	0.987	1616	93	7	81.12	100% in Ta	*IQCE, ELP2, P2RY2, HIP1, RNF219, KIAA1731, TDRD1, EPB41L2, PTTG1, ATP13A5, TIPRL*
	3	0.303	0.009	130	446	0	6.52	30% in N	
	6	0.608	0.047	41	5		2	60 % in N, 4 % in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0149	1	0.003	1.560	2391	179	2	94.54	100% in T	*C3orf63, CSPP1, CHRDL2, ROBO2, IMPA1, PPP1CC, BTNL8, CEP170, TTN, INCENP, MME*
	2	0.359	0	120	472	0	4.74	36% in N	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0156	1	0.002	1.208	2719	263	53	90.75	100% in T	*TRIP13, CYYR1, TRIOBP, KRT12, ITGB2, IL6ST, CCDC85B, KIAA0317, NRG3, C9irf64, RARS2, DDX10, CSMD3, RAD50, TG, PMPCA, LRP1B, ABCB11, RBM44, CD5L, DNAH3, NKX3-1, CBY3*
	2	0.269	0.110	123	2	0	4.1	26 % in N, 11 % in T	
	5	0.558	0.131	123	2	0	4.1	56% in N, 13% in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0159	1	0	0.530	133	147	16	6.42	53% in Tb	*INTS9, PLEKHA8*
	2	0.003	0.985	1790	146	16	86.43	100% in Ta	*PADI3, MYH7, KIAA171, RAG1, MYT3, AMBN, NUMA1, SIRT7, GOLGB1, GPR111*
	3	0.444	0.016	148	5	0	7.14	44% in N, 2% in T	

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0162	1	0	0.580	106	129	239	3.58	58% in Tb	*SDCCAG3*
	2	0.004	0.986	2670	129	240	90.32	100% in Ta	*LRRK2, GUCA2A, WDR6, WNT5A, JUN, CDH8, EHBP1, NOLC1, CELSR2, HSD3B2, ATP6V0A4, TTN, KISS1, ZC3H13, B46ALT4, BPNT1*
	3	0.401	0.037	180	5	0	6.08	40% in N,4% in T	*TIE1*

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values	Coding genes
0240	1	0.277	NA	365	279	0	42.8	28%	
	2	0.573	NA	73	280	0	8.56	57%	

Table B.5: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants from prostate cancer patients with a normal and a tumour sample (except case 0240, where there is no matched tumour). Three subclones in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0246	1	0.25	149	21	63.6	25%	

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0247	1	0.244	134	51	56.3	24%	*GABPB2, ANK3*

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0250	1	0.219	1141	92	46.9	22%	*MYH10, FASTKD1, MRPS26, WHSC1, AMTN, PCDHA8, LAMB4, GPR21*
	2	0.397	1158	92	47.6	40%	*DHX32, PGLYRP2, DKKL1, ACR*

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0251	1	0.23	528	57	20.4	23%	*DONSON, PTK2B*
	2	0.545	428	58	16.6	54%	*CCDC18, FAT1, RFPL1*
	3	0.869	36	58	1.39	87%	
	4	1.009	1150	58	44.6	100%%	*KIAA1217, TNFSF11, RPL18*

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0252	2	0.236	545	56	48.8	24%	
	3	0.403	177	57	15.8	40%	*OR10J3, PITRM1, ZNF814*
	4	0.726	302	57	27	73%	*POLE, DSG3*

Table B.6: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants from BPH fibroblasts from men without prostate cancer. Four subclones were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0063	1	0	0.483	546	256	20	10.53	48 in Tb
	2	0.002	1.013	3937	257	21	76	100 % in Ta
	3	0.250	0	702	635	1	13.5	25 % in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0069	1	0	0.609	256	85	12	24.5	61% in Tb
	2	0.002	1.016	1257	85	12	73.85	100 % in Ta
	3	0.329	0	158	347	0	9.28	33 % in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0116	1	0.003	0.394	2347	214	97		39% in T
	2	0.326	0	900	416	0		32% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0122	1	0.021	0.614	1641	157	19	54.86	2% in N, 61% in Tb
	2	0.050	1.007	1130	158	20	37.78	5% in N, 100% in Ta
	3	0.320	0.018	204	8	0	6.82	32% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0124	1	0	0.427	290	89	10	11.68	43% in Tc
	2	0.002	0.742	524	89	10	21.11	74% in Tb
	3	0	1.055	1446	90	11	58.25	100% in Ta
	4	0.318	0.000	221	423	0	8.9	32% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0127	1	0.002	0.463	466	189	16	19.63	46% in Tb
	2	0.001	1.024	1676	188	16	70.62	100% in Ta
	3	0.340	0.021	192	5	0	8.09	34% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0140	1	0.001	0.983	2338	284	126	91.29	94% in T
	2	0.289	0	186	346	0	7.26	30% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0144	1	0	0.374	132	128	37	7.22	37% in Tb
	2	0.001	1.050	1458	129	38	79.75	100% in Ta
	3	0.271	0.016	238	4	0	13.01	27% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0145	1	0.002	0.440	734	103	27	28.9	44% in Tc
	2	0.003	0.754	380	102	28	14.96	75% in Tb
	3	0.004	1.061	1132	102	28	44.58	100% in Ta
	4	0.300	0.022	288	9	0	11.34	30% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0152	1	0.002	0.562	722	97	29	19.22	56% in Tb
	2	0	0.988	1243	97	29	33.09	100% in Ta
	3	0.350	0.054	221	6	0	5.88	35% in N, 5% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0160	1	0.007	0.998	3022	165	12	94.43	100 % in T
	2	0.345	0.106	163	9	0	5	34 % in N, 10 % in T

Table B.7: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants from prostate cancer patients with a normal and a tumour sample. Subclones in the normal sample were not considered due to suspected evidence of neutral evolution. Because there was only one normal subclonal cluster for these patients the whole phylogeny was not constructed in Figures 4.6 and 4.7. Each row represents a cluster identified by the Bayesian Dirichlet process.

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0238	1	0.282	118	55	84.2	28%	

	cluster.r	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0239	1	0.261	318	67	26.4	26%	*MED23*
	2	0.542	712	67	59.2	54%	*IQGAP1, TFAP4, ZNF571*

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0241	1	0.305	90	67	63.8	30%	

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0242	1	0.249	60	58	48	25%	

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0243	1	0.266	108	77	72.9	29%	

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0244	1	0.247	109	62	68.5	25%	

	cluster	N	SNVs	Indels	Total SNVs (%)	CCF values	Coding genes
0245	1	0.261	64	60	61.5	26%%	

Table B.8: Subclonal hierarchies identified by the Bayesian Dirichlet process not including SNP variants from a morphologically normal sample from men without prostate cancer. None of the subclones except one were considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, indels and genes with predicted functional significance are annotated for each cluster.

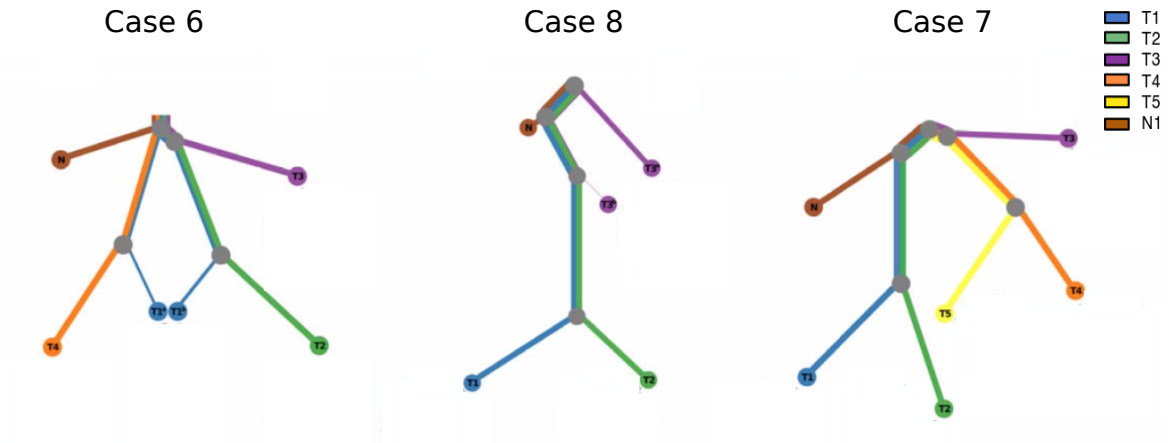


Figure B.3: Phylogenies of three patients with multifocal prostate cancer reconstructed by Cooper *et al.*[17]. Each line represents a clone from a sample. The length of each line is proportional to the weighted quantity of mutations on a logarithmic scale. Adapted from Cooper *et al.*[17].

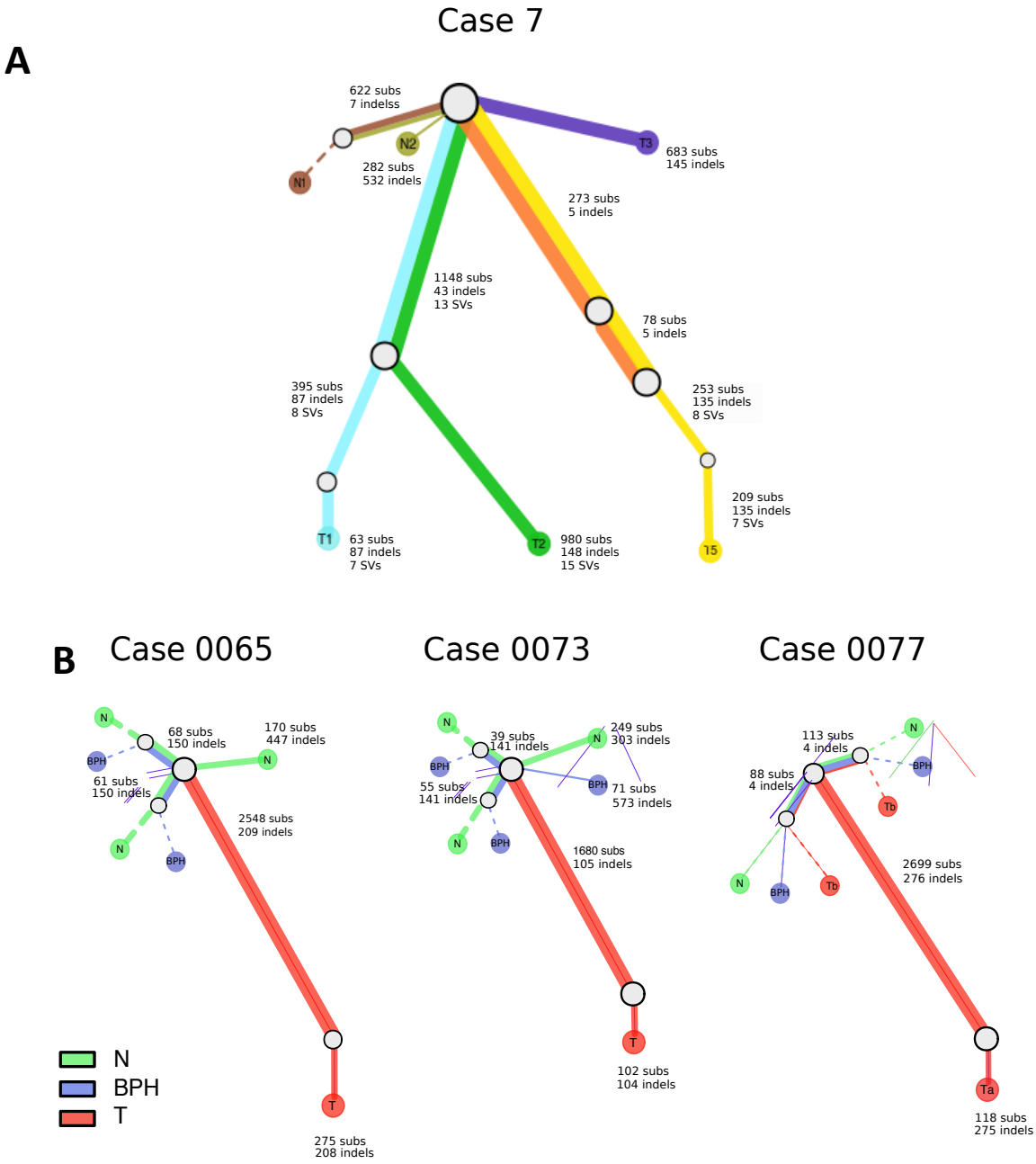


Figure B.4: Alternate configuration representing the subclonal architecture of patient 0007 and patients with normal, BPH and tumour samples. (A) In case 0007, subclone in N2 is positioned in parallel, as opposed to linearly in Figure 4.6 A. (B) All multiple. shared N/BPH subclones and unique normal and BPH subclones have been positioned in parallel (evolving independently), in contrast to having originated linearly (Figure 4.6 B).

	cluster	T1	T2	T3	T4	N1	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values
6	8	0.01	0.00	0.00	0.00	0.00	0.00	0.35	1052	207	1	7.22	27% in N3
	23	0.01	0.00	0.00	0.00	0.48	0.00	0.00	503	54	1	3.45	48% in N1
	12	0.01	1.03	0.00	0.00	0.00	0.00	0.00	1720	287	8	11.8	100% in T2
	13	0.01	0.00	0.00	0.49	0.00	0.00	0.00	407	114	5	2.79	49% in T4
	9	0.80	0.02	0.00	0.00	0.00	0.00	0.00	3418	360	46	23.46	80% in T1
	15	0.01	0.01	0.00	0.98	0.00	0.00	0.00	1308	113	5	8.97	98% in T4
	16	0.93	0.02	0.00	0.98	0.00	0.00	0.00	1746	116	11	11.98	100% in T1, 100% in T4
	26	1.01	1.09	0.00	0.00	0.00	0.00	0.00	1288	111	21	8.84	100% in T1, 100% in T2
	4	0.01	0.00	0.98	0.00	0.00	0.00	0.00	1318	185	46	9.04	100% in T3
	30	0.01	0.01	0.00	0.00	0.00	0.34	0.00	314	316	0	2.15	34% in N2
	45	0.01	0.00	0.00	0.00	0.00	0.00	0.23	179	208	0	1.22	1% in N3

Table B.9: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants for case 0006 (prostate cancer patient with multiple normal and a tumour samples). Two subclones in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, and indels with predicted functional significance are annotated for each cluster.

	cluster	T1	T2	T3	T4	T5	N1	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values
7	1	1.23	0.39	0.00	0.01	0.01	0.00	0.00	0.00	1861	43	0	25.13	100% in T1, 32% in T2
	3	0.01	0.00	0.01	0.02	0.00	0.34	0.34	0.02	740	4	0	9.99	34% in N1, 34% in N2
	12	0.01	0.00	0.00	0.00	0.68	0.00	0.00	0.00	625	67	15	8.44	68% in T5
	9	0.05	0.81	0.00	0.01	0.00	0.00	0.00	0.00	1130	67	15	15.25	80% in T2
	24	0.01	0.00	1.02	0.01	0.01	0.00	0.00	0.00	696	133	1	9.39	100% in T3
	31	0.01	0.00	0.01	1.06	0.93	0.00	0.00	0.00	485	9	0	6.54	100% in T4, 100% in T5
	19	0.01	0.00	0.01	0.01	0.00	0.00	0.23	0.00	304	627	0	4.1	23% in N2
	73	0.01	0.00	0.00	0.00	0.00	0.20	0.17	0.00	83	3	0	1.12	20% in N1, 17% in N2

Table B.10: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants for case 0007 (prostate cancer patient with multiple normal and a tumour samples). Two subclones in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, and indels with predicted functional significance are annotated for each cluster.

	cluster	T1	T2	T3	N2	N3	SNVs	Indels	SVs	Total SNVs (%)	CCF values
8	1	0.877	0.856	0.042	0.016	0.014	2692	2	0	61.1	88% in T1, 86% in T2, 4% in T3
	43	0.002	0.579	0.016	0.003	0.003	96	100	0	2.17	58 % in T2
	39	0	0	0	0.201	0	61	363	0	1.38	20 % in N2
	53	0.994	0.606	0.002	0.001	0	49	143	49	1.1	100% in T1, 61% in T2
	55	0	0.019	0.013	0.272	0.338	41	4	0	1	27% in N2, 34% in N3

Table B.11: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants for case 0008 (prostate cancer patient with multiple normal and a tumour samples). Two subclones in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs, and indels with predicted functional significance are annotated for each cluster.

	cluster	N	BPH	T	SNVs	Indels	Svs	Total SNVs (%)	CCF values
0065	1	0.00	0.00	0.53	301	208	16	6.66	52% in Tb
	2	0.00	0.00	1.02	2858	209	16	60.34	100% in Ta
	3	0.00	0.29	0.00	571	658	0	6.68	29 % in BPH
	6	0.42	0.00	0.00	262	447	0	11.76	41% in N
	7	0.47	0.48	0.00	176	150	0	1.5	46% in N, 48% n BPH
	10	0.76	0.74	0.00	61	150	0	1.84	73% in N, 73% in BPH

	cluster	N	BPH	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0073	1	0.00	0.00	0.48	77	104	32	2.71	48% in Tb
	2	0.00	0.00	0.98	1951	103	32	68.75	100% in Ta
	4	0.00	0.29	0.00	619	181	0	21.81	29 % in BPH
	6	0.00	0.62	0.00	148	181	0	5.21	61 % in BPH
	8	0.46	0.00	0.00	460	213	0	16.2	46% in N
	11	0.48	0.44	0.00	123	203	0	4.33	48% in N, 45% in BPH

	cluster	N	BPH	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0077	1	0.002	0.001	0.376	129	274	93	3.6	37% in Tb
	2	0.002	0.001	1.000	2938	274	94	8.2	100% in Ta
	3	0.088	0.338	0.012	152	0	0	4.24	8% in N, 34% in BPH, 1% in T
	4	0.346	0.317	0.013	232	0	0	6.48	34 % in N, 31 % in BPH, 1 % in T
	5	0.563	0.571	0.012	112	4	0	3.12	56% IN n, 57% in BPH

Table B.12: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants from prostate cancer patients with a normal, a BPH and a tumour sample. Two subclones in the normal sample were not considered due to suspected evidence of neutral evolution (green). Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs and indels with predicted functional significance are annotated for each cluster.

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0066	1	0.001	0.516	337	148	41	9.81	51% in Tb
	2	0.002	1.078	2038	148	42	59.36	100% in Ta
	3	0.529	0	1058	469	0	30.81	53% in N
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0072	1	0.003	0.500	173	171	122	6.19	50% in Tb
	2	0.004	1.056	2412	171	122	86.35	100% in Ta
	3	0.425	0.036	208	7	0	7.44	42% in N,3% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0074	1	0.002	0.502	1062	98	12	32.32	50% in Tb
	2	0.002	0.948	1091	99	13	33.21	95% in Ta
	3	0.456	0.007	1109	680	0	33.75	46% in N
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0076	1	0.002	0.901	2453	302	88	89.95	90% in T
	2	0.390	0.019	258	6	0	9.46	39% in N, 2% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0115	1	0.004	0.397	84	148	20	3.55	39% in Tb
	2	0.000	1.045	1901	148	21	80.38	100% in Ta
	3	0.353	0.011	341	569	0	14.41	35% in N
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0120	1	0.003	0.361	694	191	18	21.11	36% in Tb
	2	0.002	1.038	2376	191	19	72.28	100% in Ta
	3	0.391	0.024	217	2	0	6.6	39% in N, 2% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0146	1	0.000	0.423	172	94	6	7.49	42% in Tb
	2	0.003	0.997	1780	93	7	77.52	100% in Ta
	3	0.308	0.013	218	446	0	9.49	31% in N
	5	0.617	0.069	99	5		4.31	61 % in N, 7 % in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0149	1	0.002	1.608	2408	179	2	92.15	100% in T
	2	0.348	0.052	177	4	0	6.77	34% in N and 5% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0156	1	0.000	1.208	2724	263	53	87.92	100% in T
	2	0.270	0.076	176	2	0	5.68	27% in N, 7% in T
	4	0.595	0.151	162	2	0	5.22	60% in N, 15% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0159	1	0	0.511	101	147	16	4.1	51% in Tb
	2	0.003	0.982	2115	146	16	85.87	100% in Ta
	3	0.432	0.034	236	5	0	9.58	43% in N, 3% in T
	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0162	1	0	0.615	141	129	239	3.19	61% in Tb
	2	0.003	0.996	3025	129	240	68.6	100% in Ta
	3	0.387	0.014	293	5	0	6.64	38% in N, 1.5% in T

Table B.13: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants from prostate cancer patients with a normal and a tumour sample. Subclones in the normal sample marked in green were not considered due to suspected evidence of neutral evolution. Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs and indels with predicted functional significance are annotated for each cluster.

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0063	1	0.00	0.48	605	256	20	10.74	47% in Tb
	2	0.00	1.01	4106	257	21	72.9	100 % in Ta
	3	0.25	0.00	921	635	1	16.35	25 % in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0069	1	0.00	0.61	226	85	12	11.33	60% in Tb
	2	0.00	1.00	1478	85	12	74.12	100 % in Ta
	3	0.41	0.32	290	4	0	14.54	41 % in N, 3% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0116	1	0.00	0.40	2325	214	97	70	39% in T
	2	0.33	0.00	957	416	0	28.79	32% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0122	1	0.02	0.62	1890	157	19	54	2% in N, 62% in Tb
	2	0.06	1.02	1240	158	20	35.42	5% in N, 100% in Ta
	3	0.34	0.03	339	8	0	9.68	32% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0124	1	0.00	0.42	305	89	10	18.61	42% in Tc
	2	0.00	0.75	618	89	10	69.13	74% in Tb
	3	0.00	1.06	1644	90	11	55.57	100% in Ta
	4	0.32	0.00	328	423	0	11.08	32% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0127	1	0.00	0.46	533	189	16	18.61	46% in Tb
	2	0.00	1.03	1980	188	16	69.13	100% in Ta
	3	0.35	0.03	289	5	16	10.09	35% in N,, 3% in T
	3	0.44	0.54	62	5	0	2.16	44% in N, 54% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0140	1	0.02	0.94	2461	284	126	87.67	95% in T
	2	0.31	0.00	290	346	0	10.33	31% in N

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0144	1	0.00	0.37	162	128	37	7.27	38% in Tb
	2	0.00	1.04	1658	129	38	74.48	100% in Ta
	3	0.32	0.02	382	4	0	17.16	32% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0145	1	0.00	0.43	847	103	27	28.23	43% in Tc
	2	0.00	0.72	326	102	28	10.86	72% in Tb
	3	0.00	1.06	1372	102	28	45.73	100% in Ta
	4	0.30	0.03	439	9	0	14.63	30% in N, 2% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0152	1	0.00	0.57	718	97	29	31.1	57% in Tb
	2	0.00	0.99	1248	97	29	54	100% in Ta
	3	0.34	0.03	286	6	0	12.39	33% in N, 4% in T

	cluster	N	T	SNVs	Indels	SVs	Total SNVs (%)	CCF values
0160	1	0.000	0.990	3227	165	12	92.27	100 % in T
	2	0.330	0.120	253	9	0	7.23	343% in N, 12 % in T

Table B.14: Subclonal hierarchies identified by the Bayesian Dirichlet process including SNP variants from prostate cancer patients with a normal and a tumour sample. Subclones in the normal sample were not considered due to suspected evidence of neutral evolution. Because there was only one normal subclonal cluster for these patients the whole phylogeny was not constructed in Figures 4.6 and 4.7. Each row represents a cluster identified by the Bayesian Dirichlet process. SNVs and indels with predicted functional significance are annotated for each cluster.

APPENDIX C

SUPPLEMENTARY TABLES ANF FIGURES FOR CHAPTER 5

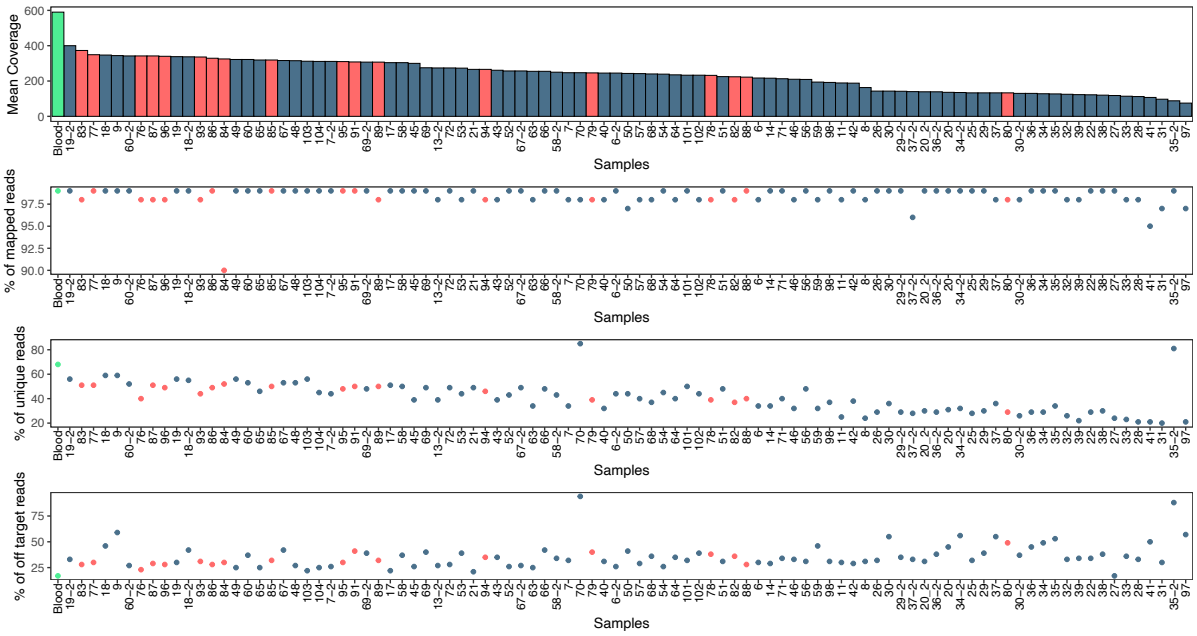
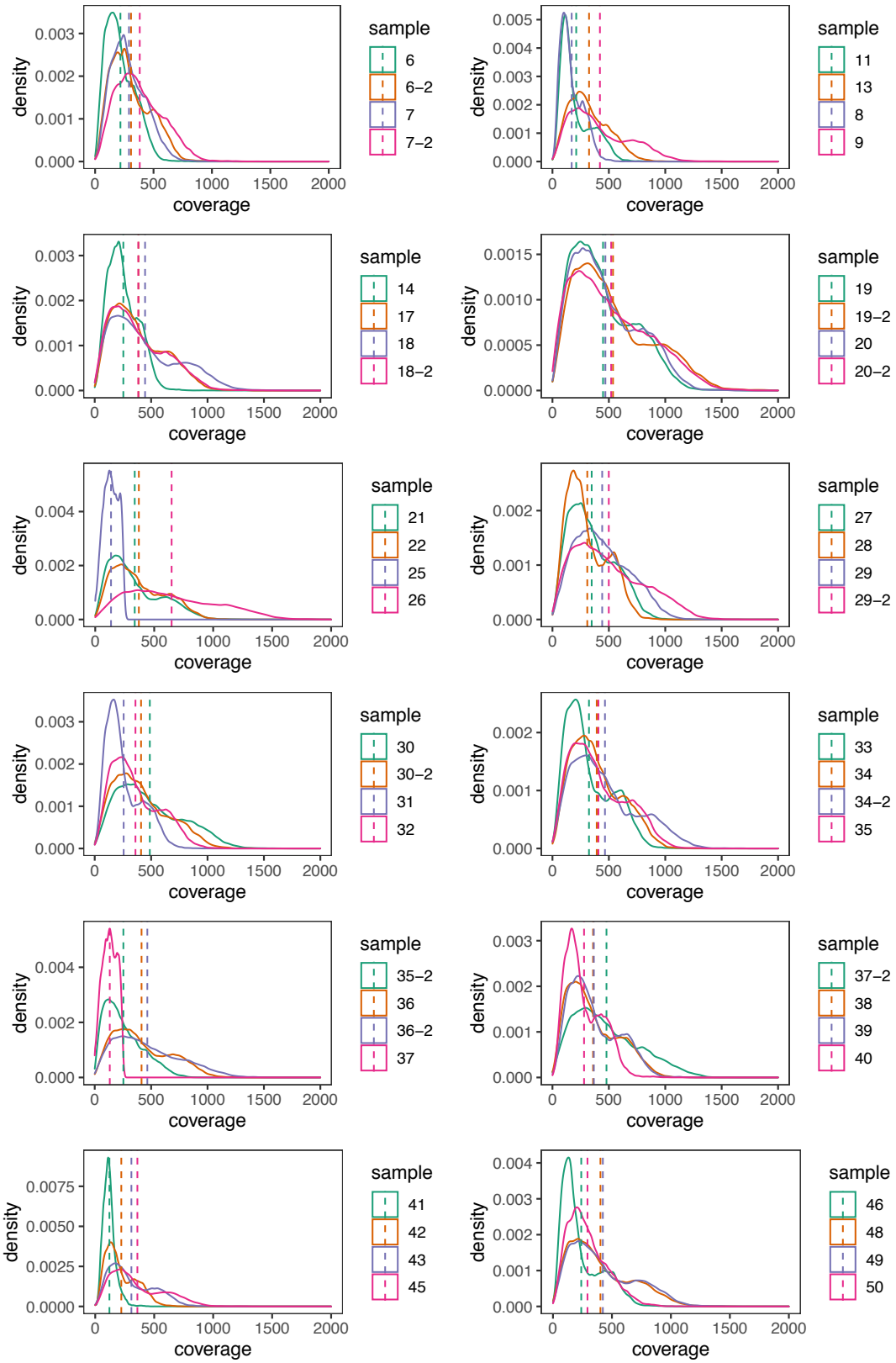


Figure C.1: Coverage and alignment metrics for the first targeted-sequencing run only: The % of mapped reads represents the reads that aligned successfully to the reference genome. The % of unique reads show moderate to high levels of duplication.


density
coverage
sample
6
6–2
7
7–2
11
13
8
9
14
17
18
18–2
19
19–2
20
20–2
21
22
25
26
27
28
29
29–2
30
30–2
31
32
33
34
34–2
35
35–2
36
36–2
37
37–2
38
39
40
41
42
43
45
46
48
49
50

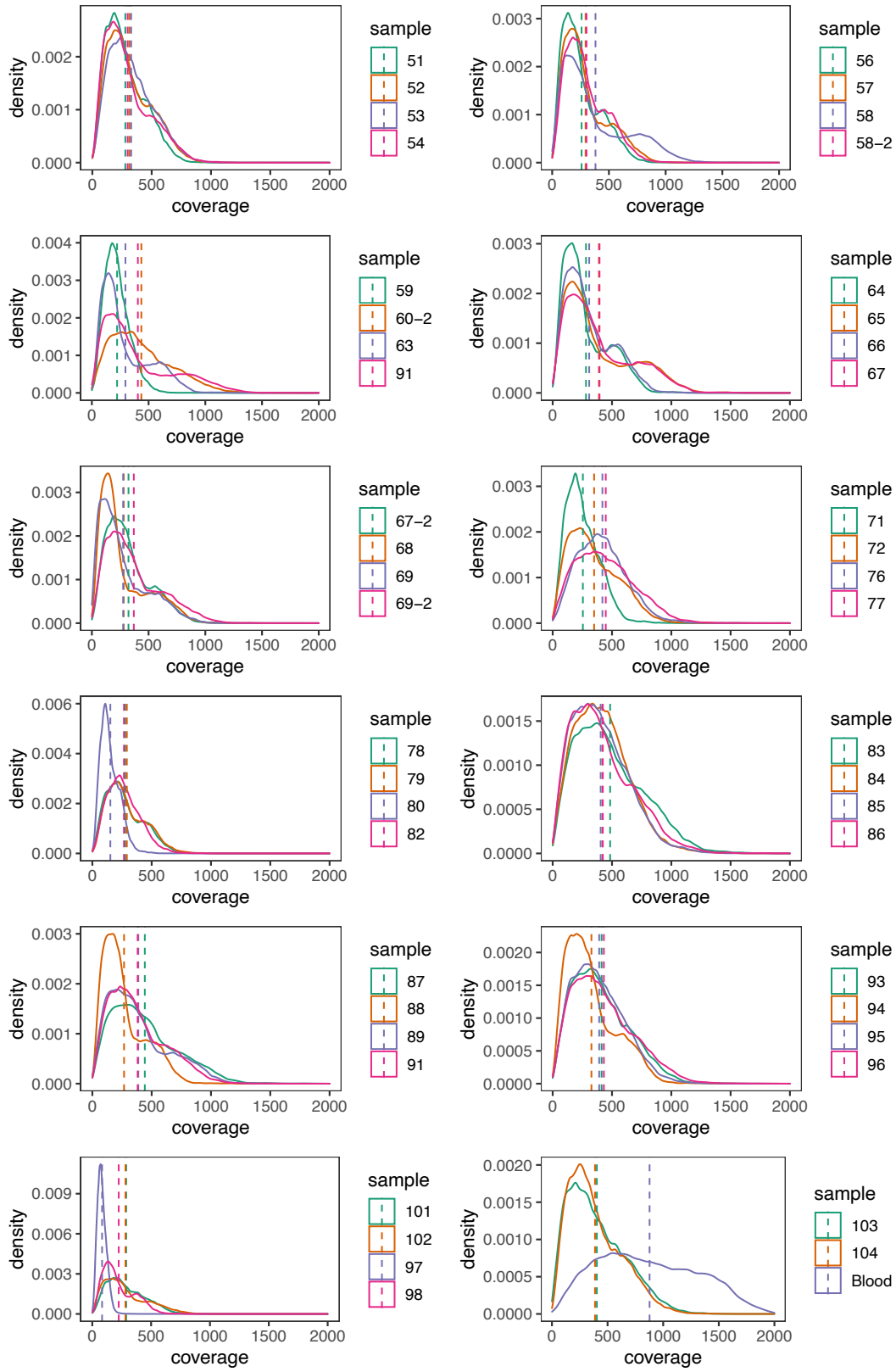


Figure C.2: Per base coverage across the total target region (98 genes.) Each line represents a sample and the matched dashed lines indicate the average coverage for that sample.

Sample	Chr	Postion	Ref	Alt	Vaf	Gene	Type	Mutation Type
36	11	533874	T	C	0.02425579	HRAS	Normal	Missense
87	11	1247919	G	A	0.01321138	MUC5B	Tumour	Missense
87	11	1248959	C	T	0.03291714	MUC5B	Tumour	Missense
88	11	1248959	C	T	0.30458716	MUC5B	Tumour	Missense
89	11	1248959	C	T	0.22972973	MUC5B	Tumour	Missense
94	11	1248959	C	T	0.14574315	MUC5B	Tumour	Missense
7	11	1253905	G	A	0.01129944	MUC5B	Normal	Missense
82	11	1261570	C	T	0.05149052	MUC5B	Tumour	Missense
78	11	1262534	G	A	0.01486989	MUC5B	Tumour	Missense
59	11	1262935	A	G	0.03196347	MUC5B	Normal	Missense
80	11	1264150	C	T	0.02412869	MUC5B	Tumour	Missense
59	11	1264453	G	A	0.0178282	MUC5B	Normal	Missense
28	11	1265018	C	T	0.02862254	MUC5B	Normal	Missense
80	11	1265741	C	T	0.01481482	MUC5B	Tumour	Missense
93	11	1266188	C	T	0.01353503	MUC5B	Tumour	Missense
84	11	1266787	G	A	0.01002735	MUC5B	Tumour	Missense
14	11	1266856	G	A	0.0130039	MUC5B	Normal	Missense
59	11	1266856	G	A	0.01949318	MUC5B	Normal	Missense
87	11	1266950	G	A	0.00873138	MUC5B	Tumour	Missense
11	11	1267210	G	A	0.02330509	MUC5B	Normal	Missense
57	11	1267352	C	T	0.01955307	MUC5B	Normal	Missense
94	11	1267649	C	T	0.01877934	MUC5B	Tumour	Missense
57	11	1267691	A	C	0.02219321	MUC5B	Normal	Missense
57	11	1267694	T	A	0.02113606	MUC5B	Normal	Missense
94	11	1267874	C	T	0.01547988	MUC5B	Tumour	Missense
80	11	1267901	C	T	0.01949861	MUC5B	Tumour	Missense
59	11	1268224	G	A	0.02199662	MUC5B	Normal	Missense
78	11	1268356	C	G	0.02071563	MUC5B	Tumour	Missense
78	11	1268359	C	A	0.01950355	MUC5B	Tumour	Missense
40	11	1268857	T	C	0.01468625	MUC5B	Normal	Missense
84	11	1268887	G	A	0.01456311	MUC5B	Tumour	Missense
80	11	1269835	G	A	0.02681992	MUC5B	Tumour	Missense
82	11	1269938	C	T	0.02569593	MUC5B	Tumour	Missense
93	11	1270253	C	T	0.01198872	MUC5B	Tumour	Missense
80	11	1270361	C	T	0.02250804	MUC5B	Tumour	Missense
78	11	1270543	C	T	0.01218274	MUC5B	Tumour	Missense
31	11	1270603	G	A	0.01582279	MUC5B	Normal	Missense
07-Feb	11	1270841	C	T	0.02097902	MUC5B	Normal	Missense
93	11	1270964	C	T	0.01824818	MUC5B	Tumour	Missense
87	11	1272768	G	A	0.02264151	MUC5B	Tumour	Missense
88	5	1278788	G	A	0.0256776	TERT	Tumour	Missense
40	5	1294888	G	C	0.07746479	TERT	Normal	Missense
69-2	19	10602443	C	T	0.01579467	KEAP1	Normal	Missense
77	1	10699508	G	T	0.05762712	CASZ1	Tumour	Missense
86	1	10699741	G	A	0.01310401	CASZ1	Tumour	Missense
68	1	10725274	A	C	0.0206044	CASZ1	Normal	Missense
79	8	13356839	T	A	0.03167421	DLC1	Tumour	Missense
80	17	16001803	G	A	0.10280374	NCOR1	Tumour	Missense
31	17	16004833	C	A	0.03636364	NCOR1	Normal	Missense
59	1	16260749	G	A	0.02493075	SPEN	Normal	Missense
87	1	16261434	C	T	0.01805054	SPEN	Tumour	Missense
68	1	16261655	C	T	0.01807229	SPEN	Normal	Nonsense
11	1	16263971	T	A	0.02507837	SPEN	Normal	Missense
59	1	16264087	T	G	0.05588235	SPEN	Normal	Missense
46	8	17611830	T	C	0.06818182	MTUS1	Normal	Missense
83	8	17612484	C	T	0.021611	MTUS1	Tumour	Missense
41	18	19154252	C	T	0.05487805	ESCO1	Normal	Missense
56	21	19685332	A	T	0.10869565	TMPRSS15	Normal	Missense
60	21	19685332	A	T	0.11188811	TMPRSS15	Normal	Missense
60-2	21	19685332	A	T	0.08753316	TMPRSS15	Normal	Missense
59	16	23646819	G	T	0.03703704	PALB2	Normal	Missense
82	4	23803389	G	A	0.07531381	PPARGC1A	Tumour	Missense
41	4	23814699	T	A	0.11981567	PPARGC1A	Normal	Missense
80	4	23815403	C	T	0.06214689	PPARGC1A	Tumour	Missense
88	1	24019162	A	G	0.04573171	RPL11	Tumour	Missense
60	8	24192995	G	A	0.0487106	ADAM28	Normal	Missense
58-2	8	24192995	G	A	0.28615385	ADAM28	Normal	Missense
60-2	8	24192995	G	A	0.31717172	ADAM28	Normal	Missense
18	12	24989522	G	T	0.12749004	BCAT1	Normal	Missense

22	12	24989522	G	T	0.07983193	BCAT1	Normal	Missense
31	12	24989522	G	T	0.12781955	BCAT1	Normal	Missense
34	12	24989522	G	T	0.22123894	BCAT1	Normal	Missense
56	12	24989522	G	T	0.0982659	BCAT1	Normal	Missense
60	12	24989522	G	T	0.17894737	BCAT1	Normal	Missense
65	12	24989522	G	T	0.17777778	BCAT1	Normal	Missense
68	12	24989522	G	T	0.10958904	BCAT1	Normal	Missense
69	12	24989522	G	T	0.23870968	BCAT1	Normal	Missense
19-2	12	24989522	G	T	0.10354223	BCAT1	Normal	Missense
34-2	12	24989522	G	T	0.0899654	BCAT1	Normal	Missense
58-2	12	24989522	G	T	0.26436782	BCAT1	Normal	Missense
60-2	12	24989522	G	T	0.3003413	BCAT1	Normal	Missense
69-2	12	24989522	G	T	0.09401709	BCAT1	Normal	Missense
41	2	25965700	G	A	0.08994709	ASXL2	Normal	Missense
79	2	25966190	C	T	0.02419355	ASXL2	Tumour	Missense
80	2	25973185	C	T	0.02973978	ASXL2	Tumour	Missense
78	1	27057818	C	T	0.01692708	ARID1A	Tumour	Missense
80	1	27057871	C	T	0.02914798	ARID1A	Tumour	Missense
7	1	27087497	G	A	0.02030457	ARID1A	Normal	Missense
8	1	27101367	C	T	0.02517483	ARID1A	Normal	Missense
53	2	27324413	G	A	0.03640257	CGREF1	Normal	Missense
82	2	27324413	G	A	0.03378378	CGREF1	Tumour	Missense
88	2	27324491	T	G	0.02277433	CGREF1	Tumour	Missense
87	6	32163797	C	T	0.0100365	NOTCH4	Tumour	Missense
78	6	32166698	C	T	0.02835052	NOTCH4	Tumour	Essential_Splice
11	6	32190386	A	T	0.02722772	NOTCH4	Normal	Missense
31	17	37627772	C	A	0.0302267	CDK12	Normal	Missense
68	17	37646947	C	T	0.04484305	CDK12	Normal	Missense
78	17	37687312	G	A	0.01677149	CDK12	Tumour	Missense
18	14	38061207	C	A	0.03333333	FOXA1	Normal	Missense
36-2	14	38061334	G	A	0.02766477	FOXA1	Normal	Missense
59	14	38061705	C	G	0.04849885	FOXA1	Normal	Missense
13	14	38061715	C	A	0.02334152	FOXA1	Normal	Missense
80	3	41278198	G	A	0.05147059	CTNNB1	Tumour	Missense
80	7	45122562	T	G	0.04926108	NACAD	Tumour	Missense
41	7	45122943	G	A	0.16666667	NACAD	Normal	Nonsense
41	7	45123366	G	A	0.01190476	NACAD	Normal	Missense
41	7	45123606	G	A	0.0083682	NACAD	Normal	Missense
80	7	45123618	C	T	0.00746826	NACAD	Tumour	Missense
79	7	45123668	G	A	0.00641368	NACAD	Tumour	Missense
58-2	7	45124337	G	A	0.04040404	NACAD	Normal	Missense
78	7	45124745	C	T	0.02003643	NACAD	Tumour	Missense
84	7	45125114	G	A	0.02362205	NACAD	Tumour	Missense
80	18	45422958	C	T	0.03112841	SMAD2	Tumour	Missense
94	12	46244829	T	G	0.02466793	ARID2	Tumour	Missense
80	12	46244872	C	T	0.02762431	ARID2	Tumour	Missense
14	3	47098411	G	A	0.03058104	SETD2	Normal	Missense
86	3	47125634	C	T	0.02123894	SETD2	Tumour	Missense
98	3	47163854	C	T	0.04017857	SETD2	Normal	Missense
41	3	47165452	G	A	0.09178744	SETD2	Normal	Missense
59	13	48955485	G	A	0.10752688	RB1	Normal	Missense
89	12	49424741	G	A	0.01950719	KMT2D	Tumour	Nonsense
78	12	49424788	G	A	0.02195609	KMT2D	Tumour	Missense
94	12	49431879	C	T	0.01697531	KMT2D	Tumour	Missense
58-2	12	49436359	G	A	0.03828829	KMT2D	Normal	Missense
6	12	49443578	G	A	0.02772277	KMT2D	Normal	Missense
67-2	12	49443578	G	A	0.02013423	KMT2D	Normal	Missense
94	12	49444220	C	T	0.02842809	KMT2D	Tumour	Missense
78	12	49444771	C	G	0.0212766	KMT2D	Tumour	Missense
46	12	49444804	C	T	0.02071823	KMT2D	Normal	Missense
86	12	49444932	C	T	0.01902174	KMT2D	Tumour	Missense
77	12	49445146	G	T	0.01845019	KMT2D	Tumour	Missense
82	12	49448810	C	T	0.02424242	KMT2D	Tumour	Essential_Splice
83	19	54649392	G	A	0.01575533	CNOT3	Tumour	Missense
71	18	56001095	G	A	0.04132231	NEDD4L	Normal	Missense
59	18	56035013	G	A	0.03181818	NEDD4L	Normal	Missense
14	17	56435261	G	A	0.0256917	RNF43	Normal	Missense
58-2	5	56526784	C	T	0.2137931	GPBP1	Normal	Missense
60-2	5	56526784	C	T	0.22878229	GPBP1	Normal	Missense
60	2	61575498	C	T	0.02439024	USP34	Normal	Missense

36-2	X	70463806	T	C	0.05367232	ZMYM3	Normal	Missense
80	16	72821167	G	A	0.02662722	ZFHX3	Tumour	Nonsense
82	16	72821931	G	A	0.01634684	ZFHX3	Tumour	Missense
78	16	72821934	G	A	0.01401542	ZFHX3	Tumour	Missense
80	16	72829379	G	A	0.11721612	ZFHX3	Tumour	Missense
82	16	72830121	C	T	0.02355073	ZFHX3	Tumour	Missense
79	16	72831230	G	A	0.02443281	ZFHX3	Tumour	Missense
80	16	72831317	T	G	0.03189066	ZFHX3	Tumour	Missense
80	16	72831366	G	T	0.02267003	ZFHX3	Tumour	Missense
11	16	72831371	T	A	0.03494624	ZFHX3	Normal	Missense
80	16	72831844	T	A	0.02580645	ZFHX3	Tumour	Missense
82	16	72832556	C	T	0.02292264	ZFHX3	Tumour	Missense
80	16	72863687	C	T	0.03414634	ZFHX3	Tumour	Missense
78	16	72991494	C	T	0.01373626	ZFHX3	Tumour	Missense
57	16	72991509	G	A	0.0195599	ZFHX3	Normal	Missense
80	16	72992349	T	A	0.05069124	ZFHX3	Tumour	Missense
18-2	6	87865458	A	G	0.09516838	ZNF292	Normal	Missense
28	4	88766952	A	G	0.04100946	MEPE	Normal	Missense
59	9	98209631	G	A	0.02504174	PTCH1	Normal	Missense
97	9	98231349	T	C	0.07881773	PTCH1	Normal	Missense
59	7	1.01E+08	C	T	0.01689189	MUC3A	Normal	Missense
71	7	1.01E+08	C	T	0.01554404	MUC3A	Normal	Missense
31	7	1.01E+08	A	T	0.02047244	MUC3A	Normal	Missense
79	7	1.01E+08	C	T	0.01546392	MUC3A	Tumour	Missense
58-2	7	1.01E+08	C	T	0.015625	MUC3A	Normal	Missense
41	7	1.01E+08	T	C	0.02512563	MUC3A	Normal	Missense
79	7	1.01E+08	C	T	0.01600985	MUC3A	Tumour	Missense
71	7	1.01E+08	C	T	0.01532567	MUC3A	Normal	Missense
89	7	1.01E+08	C	T	0.01496479	MUC3A	Tumour	Missense
67-2	7	1.01E+08	C	T	0.01884701	MUC3A	Normal	Missense
46	7	1.01E+08	C	T	0.01265823	MUC3A	Normal	Missense
50	7	1.01E+08	C	T	0.01368301	MUC3A	Normal	Missense
80	7	1.01E+08	C	T	0.01530612	MUC3A	Tumour	Missense
64	7	1.01E+08	C	T	0.01908397	MUC3A	Normal	Missense
68	7	1.01E+08	C	T	0.01217533	MUC3A	Normal	Missense
88	5	1.12E+08	C	T	0.2625	APC	Tumour	Missense
82	11	1.14E+08	A	G	0.04219409	USP28	Tumour	Missense
64	12	1.15E+08	T	A	0.03030303	TBX3	Normal	Missense
79	1	1.2E+08	C	T	0.02014652	NOTCH2	Tumour	Missense
85	1	1.2E+08	C	T	0.01751825	NOTCH2	Tumour	Missense
64	1	1.2E+08	G	A	0.02430134	NOTCH2	Normal	Missense
97	1	1.21E+08	A	G	0.13953488	NOTCH2	Normal	Missense
97	12	1.25E+08	A	G	0.06321839	NCOR2	Normal	Missense
68	12	1.25E+08	G	A	0.01492537	NCOR2	Normal	Missense
40	4	1.26E+08	C	T	0.02247191	FAT4	Normal	Missense
59	4	1.26E+08	G	A	0.04597701	FAT4	Normal	Missense
41	4	1.26E+08	T	-	0.03088803	FAT4	Normal	no-SNV
85	4	1.26E+08	G	A	0.02201835	FAT4	Tumour	Missense
85	4	1.26E+08	T	C	0.02301255	FAT4	Tumour	Missense
88	4	1.26E+08	C	T	0.03238866	FAT4	Tumour	Missense
84	X	1.29E+08	A	C	0.2716763	SMARCA1	Tumour	Nonsense
82	9	1.39E+08	C	T	0.02259887	NOTCH1	Tumour	Missense
89	9	1.39E+08	G	A	0.01114206	NOTCH1	Tumour	Missense
87	9	1.39E+08	C	T	0.01395349	NOTCH1	Tumour	Missense
83	9	1.39E+08	C	T	0.01211454	NOTCH1	Tumour	Missense
79	9	1.39E+08	C	G	0.01923077	NOTCH1	Tumour	Missense
79	9	1.39E+08	G	C	0.01953125	NOTCH1	Tumour	Missense
78	9	1.39E+08	G	A	0.0177305	NOTCH1	Tumour	Missense
89	9	1.39E+08	A	T	0.05317919	NOTCH1	Tumour	Missense
80	7	1.4E+08	C	T	0.0372093	BRAF	Tumour	Missense
80	2	1.41E+08	C	T	0.14	LRP1B	Tumour	Missense
54	2	1.42E+08	C	T	0.05027933	LRP1B	Normal	Missense
79	2	1.42E+08	C	T	0.03529412	LRP1B	Tumour	Missense
80	2	1.42E+08	C	T	0.05421687	LRP1B	Tumour	Missense
76	4	1.45E+08	C	T	0.18518519	GYPA	Tumour	Missense
77	4	1.45E+08	C	T	0.08805031	GYPA	Tumour	Missense
60	5	1.51E+08	A	T	0.03740158	FAT2	Normal	Missense
58-2	5	1.51E+08	A	T	0.28878648	FAT2	Normal	Missense
60-2	5	1.51E+08	A	T	0.28060046	FAT2	Normal	Missense
68	5	1.51E+08	G	A	0.03125	FAT2	Normal	Missense

41	5	1.51E+08	G	A	0.05594406	FAT2	Normal	Missense
91	5	1.51E+08	G	A	0.03731343	FAT2	Tumour	Nonsense
46	5	1.51E+08	C	T	0.02803738	FAT2	Normal	Missense
80	7	1.52E+08	C	T	0.03550296	KMT2C	Tumour	Missense
41	7	1.52E+08	T	C	0.05194805	KMT2C	Normal	Missense
78	7	1.52E+08	C	T	0.01986755	KMT2C	Tumour	Missense
31	7	1.52E+08	G	A	0.03436426	KMT2C	Normal	Missense
78	7	1.52E+08	C	G	0.01778656	KMT2C	Tumour	Missense
78	7	1.52E+08	G	A	0.02016129	KMT2C	Tumour	Missense
41	7	1.52E+08	G	A	0.04411765	KMT2C	Normal	Missense
78	7	1.52E+08	C	T	0.02835052	KMT2C	Tumour	Missense
80	7	1.52E+08	G	A	0.03501946	KMT2C	Tumour	Missense
39	7	1.52E+08	C	T	0.06349206	KMT2C	Normal	Missense
41	7	1.52E+08	C	A	0.04316547	KMT2C	Normal	Missense
41	7	1.52E+08	A	T	0.021611	KMT2C	Normal	Missense
64	1	1.53E+08	A	T	0.02507837	LCE2B	Normal	Missense
64	1	1.53E+08	T	C	0.02551834	LCE2B	Normal	Missense
39	1	1.55E+08	G	A	0.03236797	ASH1L	Normal	Nonsense
71	1	1.55E+08	G	A	0.022	ASH1L	Normal	Nonsense
39	1	1.55E+08	G	A	0.03005008	ASH1L	Normal	Missense
68	1	1.55E+08	G	A	0.025	ASH1L	Normal	Nonsense
83	1	1.55E+08	C	T	0.02290076	ASH1L	Tumour	Missense
79	1	1.55E+08	C	T	0.03503185	ASH1L	Tumour	Missense
42	1	1.55E+08	C	T	0.05208333	ASH1L	Normal	Missense
68	1	1.55E+08	C	T	0.03829787	ASH1L	Normal	Missense
46	1	1.55E+08	G	T	0.04402516	ASH1L	Normal	Missense
85	1	1.55E+08	C	T	0.02028081	ASH1L	Tumour	Missense
41	1	1.55E+08	C	T	0.07222222	ASH1L	Normal	Missense
66	1	1.55E+08	C	T	0.03184713	ASH1L	Normal	Missense
54	1	1.57E+08	C	T	0.02511416	ETV3	Normal	Missense
33	1	1.57E+08	A	T	0.0282187	ETV3	Normal	Missense
60	4	1.88E+08	C	T	0.04545455	FAT1	Normal	Missense
53	4	1.88E+08	C	T	0.02826087	FAT1	Normal	Missense
78	4	1.88E+08	G	A	0.02422908	FAT1	Tumour	Nonsense
79	4	1.88E+08	G	A	0.0259366	FAT1	Tumour	Missense
82	4	1.88E+08	C	T	0.02232143	FAT1	Tumour	Missense
85	2	1.98E+08	T	C	0.38848921	SF3B1	Tumour	Missense
86	2	1.98E+08	T	C	0.41578947	SF3B1	Tumour	Missense
87	2	1.98E+08	T	C	0.32209738	SF3B1	Tumour	Missense
88	2	1.98E+08	T	C	0.30711611	SF3B1	Tumour	Missense
89	2	1.98E+08	T	C	0.33488372	SF3B1	Tumour	Missense
91	2	1.98E+08	T	C	0.35261708	SF3B1	Tumour	Missense
93	2	1.98E+08	T	C	0.33874239	SF3B1	Tumour	Missense
94	2	1.98E+08	T	C	0.22251309	SF3B1	Tumour	Missense
95	2	1.98E+08	T	C	0.41153846	SF3B1	Tumour	Missense
96	2	1.98E+08	T	C	0.30382294	SF3B1	Tumour	Missense
96	2	2.26E+08	C	T	0.09774436	DOCK10	Tumour	Missense
78	1	2.27E+08	C	T	0.05976096	CDC42BPA	Tumour	Missense
41	1	2.27E+08	A	T	0.05217391	CDC42BPA	Normal	Missense

Table C.1: Total number of SNVs detected using “deepSNV”.

Genes	Targeted sequencing (TS)	WGS	Tissue Type
ADAM28	Mutated	Mutated	Normal (WGS)/Normal (TS)
ADAM29	-	-	-
AKT1	-	-	-
ANTXR2	-	-	-
APC	Mutated	-	Tumour
AR	-	-	-
ARID1A	-	-	-
ARID2	Mutated	-	Tumour
ARID4A	-	-	-
ARID4B	-	-	-
ASH1L	Mutated	-	Normal/Tumour
ASXL2	Mutated	-	Normal/Tumour
ATAD3B	-	-	-
ATM	-	-	-
BCAT1	Mutated	Mutated	Normal (WGS)/Normal (TS)
BRAF	Mutated	-	Tumour
BRCA2	-	-	-
CASZ1	Mutated	-	Normal/Tumour
CDC42BPA	Mutated	-	Normal/Tumour
CDK12	Mutated	-	Normal/Tumour
CDKN1B	-	-	-
CGREF1	Mutated	-	Normal/Tumour
CHD1	Mutated	-	Normal
CNOT3	Mutated	-	Tumour
CTNNB1	Mutated	-	Tumour
CUL3	-	-	-
DLC1	Mutated	-	Tumour
DOCK10	Mutated	-	Tumour
ESCO1	Mutated	-	Normal
ETV3	Mutated	-	Normal
FAM149A	-	-	-
FAT1	Mutated	-	Normal/Tumour
FAT2	Mutated	Mutated	Normal (WGS)/Normal (TS)
FAT4	Mutated	-	Normal/Tumour
FOXA1	Mutated	-	Normal
GATA1	-	-	-
GPBP1	Mutated	Mutated	Normal (WGS)/Normal (TS)
GYPA	Mutated	-	Tumour
HRAS	Mutated	-	Normal
IDH1	-	-	-
KDM6A	-	-	-
KEAP1	Mutated	-	Normal
KMT2C	Mutated	-	Normal/Tumour
KMT2D	Mutated	-	Normal/Tumour
KRAS	-	-	-
LCE2B	Mutated	-	Normal
LRP1B	Mutated	-	Normal/Tumour
MEPE	Mutated	-	Normal
MRE11A	-	-	-
MTUS1	Mutated	-	Normal/Tumour
MUC3A	Mutated	-	Normal/Tumour
MUC5B	Mutated	-	Normal/Tumour
NACAD	Mutated	-	Normal/Tumour

NCOR1	Mutated	-	Normal/Tumour
NCOR2	Mutated	-	Normal/Tumour
NDST4	-	-	-
NEDD4L	Mutated	-	Normal
NOTCH1	Mutated	-	Normal/Tumour
NOTCH2	Mutated	-	Normal/Tumour
NOTCH3	Mutated	-	Normal/Tumour
NOTCH4	Mutated	-	Normal/Tumour
NRAS	-	-	-
PALB2	Mutated	-	Normal
PCSK2	Mutated	-	Normal
PIK3CA	-	-	-
PIK3CB	-	-	-
PIK3R1	-	-	-
PIK3R2	-	-	-
PPARG	-	-	-
PPARGC1A	Mutated	-	Normal/Tumour
PTCH1	Mutated	-	Normal
PTEN	-	-	-
RB1	Mutated	-	Normal
RBM10	-	-	-
RNF43	Mutated	-	Normal
RPL11	Mutated	Mutated	Tumour (WGS)/Tumour (TS)
SETD2	Mutated	-	Normal/Tumour
SF3B1	Mutated	Mutated	Tumour (WGS)/Normal (TS)/Tumour (TS)
SMAD2	Mutated	-	Tumour
SMAD4	-	-	-
SMARCA1	Mutated	-	Tumour
SOX2	-	-	-
SPANXC	-	-	-
SPEN	Mutated	-	Normal/Tumour
SPOP	-	-	-
TBL1XR1	Mutated	-	Normal
TBX3	Mutated	-	Normal
TERT	Mutated	-	Normal/Tumour
TMPRSS15	Mutated	Mutated	Normal (WGS)/Normal (TS)
TP53	-	-	-
TRAT1	-	-	-
U2AF1	-	-	-
USP28	Mutated	-	Tumour
USP34	Mutated	-	Normal
USP7	-	-	-
ZFHX3	Mutated	-	Normal/Tumour
ZMYM3	Mutated	-	Normal
ZNF292	Mutated	-	Normal/Tumour

Table C.2: List of genes that were targeted sequenced for the Patchwork experiment. Mutated/Non-mutated (-) genes are indicated (including synonymous and non-synonymous mutations) for both the patchwork experiment and the whole genome sequenced slice for the same patient (0007).