Ollscoil na Gaillimhe
University of Galway

Computational approaches to identify and explain sources of error in cancer somatic mutation data

Brian O'Sullivan B.Sc., M.Sc.

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisor: Professor Cathal Seoighe

School of Mathematical and Statistical Sciences
University of Galway
Galway City, Ireland

Table of Contents

Declaration of Authorship

I, Brian O'Sullivan, declare that this thesis titled, 'Computational approaches to identify and explain sources of error in cancer somatic mutation data' submitted to the Discipline of Bioinformatics, School of Mathematical & Statistical Science, University of Galway in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD.) in Bioinformatics is entirely my own work and I have acknowledged any assistance or contributions and cited the published work of others where applicable. The research contained within this thesis has been conducted with the financial support of Science Foundation Ireland under Grant Number 16/IA/4612. This work has not been submitted, in whole or in part, by me or another person, for the purpose of obtaining any other degree.

I agree freely that the library may lend or copy this thesis upon request.

Brian O'Sullivan
1st January, 2024

Acknowledgements

I would like to convey my sincere gratitude to Prof. Cathal Seoighe, my supervisor. Thank you for providing me with the opportunity to embark on this PhD program; it has been an incredible journey. Over the past five years, including one year of my MSc. and the subsequent four years of my PhD., your support, guidance, and insight while steering me along the road has helped me over the line. I recognise and appreciate the time and effort you put in. Thank you.

I would like to express my gratitude to my GRC members, Dr. Haixuan Yang, Dr. Andrew Flaus, and Dr. Derek Morris. Your advice has consistently been constructive and helpful and I was very glad of it. Thank you.

To Prof. Aaron Golden, I express my gratitude for your wit, perspective, insight, wisdom, reassurance, humour and big heart. Thank you.

To my fellow Bioinformatic PhD students and colleagues, particularly the crew of ADB1018, Declan Bennet PhD., Siobhán Cleary PhD., Noor Kherreh PhD., and Sumaira Malik. It was a privilege to share the office with and to get to know you all. Thanks for your friendship, company, help, chats, encouragement, humour, advice and coffee! Thanks to Declan and Siobhán for putting up with my numerous and tedious questions during my time in the MSc. (and also for quite a while after that too). Declan, thank you for your constructive comments following your careful review of this manuscript, and over the last number of years also. I really appreciate your patience and your help. Siobhán thank you for your help also, and for your sound advice in helping me navigate the stresses of life as a PhD student. Noor, thank you for broadening my perspective of the world beyond these shores, and in particular, for my new found interest in cricket. Thank you Sumaira for your help too.

Thanks to my family for their patience and encouragement over the last 5 years. Without their support, my return to college would not have been possible. A special thanks to Caoimhe, our daughter, who will undoubtedly be relieved to know her father is no longer a fellow student. I am very grateful for her skill in transforming my clumsy sketches into the beautiful graphics in figures 1.1 and 1.2 in this thesis. Thanks to Rory, our son, for his friendship and determination that was an example for me in my return to education. You both make Martina and I very proud. Thank you Martina, my wife, for picking up the slack over the last number of years. Thanks for your encouragement, your help, understanding and your advice. Our retirement is surely sorted now that I am nearly a Doctor of philosophy.

To Mum and Dad, thank you for getting me here, and for all our memories together.

To the brave survivors,
and the good friends we've had and lost along the way.

Abstract

Errors in the identification of somatic mutations in cancer samples can have critical implications in both research and clinical applications. Failure to detect potential variants of interest can lead to missed opportunities in patient treatment or scientific research. Incorrectly identifying a somatic variant may result in inaccurate prognosis, unsuitable treatment selection, or misleading research. By understanding the sources of error in somatic mutation calling, we are better placed to mitigate these risks. The reevaluation of variants that have been excluded from analysis by mutation calling methodologies can provide valuable insights in this regard. By considering the allele frequency, nucleotide context, and potential impact on protein of a mutation that has been discarded from analysis, we can incorporate the overall biological context into our assessment of the variant call. This approach enables us to identify putative somatic variants that were overlooked by the caller and, importantly, investigate the reason for their omission.

In Chapter 2, we outline vcfView, an interactive R Shiny tool designed to support the evaluation and exploratory analysis of somatic mutation records from cancer sequencing data. We use vcfView to reevaluate the TCGA acute myeloid leukaemia data and identify clinically actionable mutation records in patients that were incorrectly excluded from analysis due to the presence of tumour sample DNA in the matched normal sample.

The validation of somatic mutation calling pipelines is a critical step in ensuring the accuracy and reliability of the results obtained from the analysis of cancer genomic sequencing data. However, the trustworthiness of the validation results is directly linked to the quality of the truth set used for validation. In Chapter 3, we introduce a simulation framework designed to generate comprehensive and realistic tumour genomic sequencing data. This framework takes into account the inherent randomness of genomic sequencing, providing an accurate representation of the frequency profile as it is observed in real sequencing data. It generates a corresponding truth set alongside the simulated sequencing data, documenting the true source of each non-reference base in the data. Unlike existing validation methods, this truth set not only identifies variant caller errors but, crucially, enables us to understand the reasons behind the erroneous calls. Using the GATK Mutect2 variant calling pipeline, we apply this framework to highlight and explain sources of error in somatic mutation data and biases in the estimation of somatic allele frequency.

Finally in Chapter 4, we analyse tumour-only sequencing and somatic variant data from an unpublished dataset comprising 60 individuals diagnosed with early-onset and aggressive pancreatic ductal adenocarcinoma. We apply the tools and methods we have developed previously to recover somatic variant information from sequence data obtained from heavily damaged FFPE samples. We provide an improved estimate of the true incidence of pathogenic KRAS variants within the cohort that accounts for the sequencing strategy and sample preparation methods used. We also highlight recurrent mutations in several other cancer associated genes that may have played a role in disease progression in these patients.

List of Figures

List of Tables

1 Introduction

1.1 Cancer evolution and development

1.1.1 Hallmarks of cancer

Cancer is a disease that can be traced back to the early stages of recorded history. One reference dating back to around 3000 BCE in Egypt records treatments for breast cancer[1], while Hippocrates, around 400 BCE, described the invasive nature of the disease and distinguished between benign and malignant tumours[2]. In the fossil record, the true extent of the ancient origins of this disease becomes evident. In South Africa, scans of a 1.7 million year old fossilised toe bone uncovered an osteosarcoma in an ancient species of hominid[3], while metastatic bone cancer has been identified in the polished cross-section of a 150 million-year-old Jurassic dinosaur fossil[4]. Hallmarks of this disease transcend species and have roots that extend far back into our evolutionary past[3,4,5,6].

In recent decades, progress in the field of personalised medicine has given rise to the development of novel therapies and treatment modalities targeting the unique set of factors driving an individual tumour. Alongside these advancements, an awareness has emerged that, while no two cancers are identical, they all share a core set of characteristics or hallmarks that define the disease. This list of hallmarks continues to evolve as our understanding of the disease progresses. A formalised proposal of the hallmarks of cancer was first published at the turn of the 21st century[7], listing six core characteristics of the disease: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Cell growth and division are strictly regulated in normal cells. Tissue homeostasis is maintained by a complex network of intracellular signals that tell the cell when to enter the replicative cell cycle. In the absence of external signals, the cell remains in G0, the default phase of a normal cell. Signalling autonomy, the ability for cells to grow and proliferate independently, without relying on external signalling, is a distinctive hallmark of cancer cells. Signalling autonomy may be acquired in various ways, through the autocrine production of growth factors by cancer cells, the overexpression of cell surface signalling proteins, or abnormal signal transduction resulting from mutations in cell surface proteins or further downstream in the signalling pathway[8,9].

Anti-growth signalling pathways are crucial in counterbalancing proliferative signalling. They regulate cellular processes such as growth and cell division, often interworking with tumour suppressor proteins such as p53, phosphatase and tensin homolog (PTEN), and retinoblastoma protein (pRb) that detect abnormalities and trigger growth inhibition or apoptosis when necessary. Signals that trigger a cellular response may originate from outside the cell via ligands or molecules that bind to specific receptors. Signals may originate from within the cytoplasm, for example, signalling pathways involved in regulating cell cycle progression, DNA damage, and responses to cellular stress.[10,9]. Anti-growth signal dysregulation is common in cancer. For example, the growth-inhibiting cytokine TGF-β is produced by white blood cells to instruct cells to halt the cell cycle (via Rb). Cancer cells block this signal through mutations that disrupt TGF-β receptors or other downstream components

in the signalling pathway. TGF-β also causes immunosuppression and angiogenesis within the tumour microenvironment. In the absence of anti-growth signals within the cancer cells, this enables the tumour to grow and invade[11].

Apoptosis is the process of orderly self-destruction of the cell. Morphologically, this is characterised by cell shrinkage, nuclear condensation, and fragmentation of the cell into smaller membrane-bound apoptotic bodies, which are ultimately disposed of by phagocytosis. Apoptosis is an essential characteristic of multicellular organisms that allows for the removal of damaged cells that have the potential to become malignant. Apoptosis is triggered either extrinsically via the death receptor cell surface protein (tumour necrosis factor receptor) or intrinsically, primarily mediated by the tumour suppressor p53[12]. P53 responds to signals of cell stress or DNA damage, initiating repair if possible, or apoptosis if repair is not achievable. Due to their critical role in preventing malignancy, apoptotic pathways are frequently targeted in cancer. The tumour suppressor gene TP53 has been identified as the most frequently mutated gene in human cancers[13]. Viral proteins that bind to and inactivate p53, along with other tumour suppressors, are also exploited by cancer to suppress apoptosis. Mutations in the death receptor pathway, receptor downregulation, or the expression of decoy receptors are other mechanisms by which tumour cells may evade extrinsic apoptotic signals[14].

Most normal cells have a finite lifespan. With each round of mitosis, the telomeres get progressively shorter. Once the telomeres reach a critical length the cell enters senescence, a state of irreversible growth arrest and will no longer divide[15]. Stem and germ cells are the exception however. These cells express telomerase reverse transcriptase (TERT) which repairs the ends of the telomeres allowing the cell to continue to replicate. The progressive shortening of the telomeres with each round of cell division presents a significant barrier to cancer growth. Telomerase, typically not expressed in somatic cells, has been detected in approximately 85% of malignant tumours[16,17], demonstrating that, for the majority of malignancies, TERT activation is hijacked by cancer to continue proliferating. In a minority of cancers the homologous recombination pathway, alternative lengthening of telomeres (ALT) is dysregulated to maintain telomere length[18]. Both mechanisms result in the cancer gaining the capability for unrestricted proliferation and progression of the malignancy.

Angiogenesis, the formation of new blood vessels, is a crucial milestone in the development of malignancy. Without a blood supply, a tumour cannot attain significant growth, invade surrounding tissues, or metastasize to other parts of the body. Angiogenesis is a complex and tightly regulated process involving degradation of the basement membrane and activation, proliferation, and migration of the endothelial cells. As the tumour evolves it begins to manipulate the surrounding microenvironment releasing growth factors and cytokines to secure a blood supply. Pro-angiogenic factor VEGF-A, expressed by nearly all malignant tumours binds to and activates both VEGFR-1 and VEGFR-2 on vascular endothelial cells, promoting permeability, proliferation, migration, survival, and angiogenesis[19]. Hypoxia in the tumour microenvironment typically plays an important role in this regard, up-regulating the transcription factor hypoxia-inducible factor-1 (HIF-1) which in turn increases the expression of many angiogenesis inducers while suppressing angiogenesis inhibitors[20].

Arguably, the most destructive characteristic of cancer is its capacity to invade

and metastasize, with the latter being the primary cause of cancer-related deaths[21]. The ability to infiltrate surrounding tissues distinguishes a benign, typically manageable growth from a malignant and potentially fatal tumour. An essential aspect contributing to a tumour's invasive potential is the physiological process known as epithelial-mesenchymal transition (EMT), where epithelial cells undergo transformation, acquiring motile and invasive characteristics of mesenchymal cells. The triggers for EMT activation in cancer are complex and not fully understood, with various factors implicated, including epigenetic modifications[22], the tumour microenvironment[23], microRNAs[24], and mutations in signalling pathways[25]. Once acquired, EMT allows cancer to breach the basement membrane and dissolve the extracellular matrix below. Migration of cancer cells continues until intravasation occurs, with cancer cells entering a blood or lymphatic vessel, resulting in the transportation of the cancer from the primary site. Extravasation follows, leading to the establishment of secondary tumours in other locations in the body.

Eleven years after their initial review of six biological capabilities acquired during tumour progression[7], the authors released a second publication summarising four additional emerging hallmarks of the disease[26]: tumour immune evasion, inflammation, deregulation of cellular energetics, and mutation and genomic instability. Within the tumour microenvironment, inflammation and immune activity are often considered a double-edged sword in combating cancer[27,28,29]. On the one hand, they play a crucial role in restraining cancer, limiting its growth upon detection, and surveilling and destroying malignant cells throughout the body to prevent the establishment of a cancer niche. However cancer may evolve capabilities to manipulate immune and other cells within this environment leading to immune cell anergy, tumour tolerance and continued angiogenesis and tumour growth. A critical factor in the tumour acquiring these capabilities is instability and somatic mutation within the cancer genome, ultimately producing a clone with the necessary capabilities to persistently evade and grow. Tumour hypoxia and the deregulation of energy releasing metabolic pathways in cancer (the warburg effect), another important and long standing hallmark of the disease[30] is also considered to play a pivotal role in progression. Our understanding of these characteristics has advanced significantly in recent decades, and ongoing research is expected to further clarify the mechanisms underlying the hallmarks of cancer and their therapeutic implications.

1.1.2 Genomic damage and instability

Genome instability has been proposed as the force that generates genetic diversity, expediting the acquisition of the hallmarks of cancer[26]. Cancer is commonly described as a disease of the genome. Tumours typically contain a range of genetic and epigenetic lesions, as well as structural alterations that dysregulate cellular metabolism, mutate proteins, and transform healthy, normal cells into malignant tissue. Determining the cause of this damage can reveal the origin of the cancer itself. Damage to DNA can occur as a result of normal cellular metabolism, a process typically referred to as intrinsic DNA damage. For example, oxidative phosphorylation within the mitochondria, which produces cellular energy in the form of ATP, also generates reactive oxygen species (ROS) as byproducts that have the potential to damage DNA. Furthermore, during DNA replication, the genome encounters numerous challenges, collectively known as replication stress, where it is exposed to several potentially DNA-damaging processes that cause replication fork progression

to slow or stall[31]. The correct partitioning of replicated sister chromatids during mitosis is also critical. Errors in this process lead to aneuploidy and potentially malignant transformation. Transcription-associated mutagenesis is also considered to play a role in intrinsic DNA damage[32]. Transcription usually copies only one DNA strand, leaving the other non-transcribed strand unpaired and susceptible to damage until the transcription process completes. This may result in localised DNA lesions within the transcribed genomic region[32].

Although intrinsic DNA damage is unavoidable, only a very small proportion of the lesions caused by this damage result in somatic mutations. The cellular DNA repair machinery plays a critical role in this regard. ROS damage, along with other common lesions caused by deamination and alkylation are typically repaired through base excision repair (BER). In this process, enzymes first remove the damaged bases and create a nick in the phosphodiester backbone, allowing DNA polymerase I and DNA ligase to repair the missing bases and reseal the nick[33]. Throughout various stages of the cell cycle, additional DNA damage responses (DDR) and checkpoints play crucial roles. For example, during replication, in the S and G2 phases, the Homologous Recombination Repair (HRR) pathway comes into play for the restoration of double-strand breaks. This pathway capitalises on the inherent homology within the sister chromatid, using it as a template to replace damaged or missing DNA segments flanking the break in the affected chromosome[34]. DNA damage and replication stress during the cell cycle trigger a checkpoint response that prevents progression until DNA synthesis has completed and DNA damage repaired. The mitotic checkpoint complex also plays a crucial role in maintaining genome integrity by preventing the separation of the duplicated chromosomes until each chromosome is properly attached to the spindle[35]. Transcription-coupled repair (TCR) plays a vital role in intrinsic DNA damage response by recruiting nucleotide excision repair (NER) to excise and repair the DNA containing the lesion on the transcribed strand[36].

Genomic damage and instability can result not only from DNA-damaging agents produced during normal cellular metabolism but also from extrinsic sources of damage originating from external processes, posing a potential threat to DNA integrity and contributing to the development of cancer. Numerous epidemiological studies have established significant correlations between cancer and exposure to carcinogens[37]. Such exposure can arise from occupational or other environmental factors, including, but not limited to, diesel exhaust, asbestos, radon gas, ionising radiation (i.e., X-rays), as well as lifestyle choices such as cigarette smoking or alcohol consumption. The extent of extrinsic DNA damage resulting in somatic mutation or chromosomal alteration, and the corresponding risk of malignant progression, has been linked to the nature and duration of exposure to the carcinogen[38].

One of the most common examples is extrinsic DNA damage caused by exposure to ultraviolet (UV) light. UV radiation can induce pyrimidine dimers in DNA. High-energy radiation, such as X-rays, can also result in significant damage and instability in the genome. As the radiation traverses the cell, it liberates electrons that damage and sever DNA molecules[39]. If left unrepaired, these types of lesions can impede transcription and replication, leading to cell death or malignancy. Chemicals such as benzene or aflatoxins that form adducts with DNA bases also pose a significant threat to genome integrity. Similarly, conditions linked to chronic inflammation expose cells to DNA-damaging agents[40].

DNA damage repair processes within the cell are essential for preserving genomic integrity in the face of extrinsic damage. NER typically addresses the restoration of DNA containing crosslinks or adducts. Double-strand breaks, when homologous recombination repair (HRR) is not an option, are repaired through non-homologous end joining (NHEJ)[41]. NHEJ involves ligating the broken ends of the chromosome together without reference to homology, presenting a notable risk. Executing this repair without attempting to restore the original sequence can lead to gene inactivation or the potential formation of oncoproteins.

The vital significance of DNA damage response pathways in preserving genomic integrity becomes particularly apparent when these pathways experience inactivating somatic or germline mutations. Genetic disorders, such as xeroderma pigmentosum, an autosomal recessive condition impacting pathways crucial for UV damage repair, underscore the magnitude of genomic damage that will result if UV-induced lesions are not effectively repaired. Individuals afflicted with the disease experience severe sunburn even with minimal sun exposure, and many succumb to early-onset skin cancers[42]. Lynch syndrome, an autosomal dominant cancer predisposition syndrome caused by defects in the DNA mismatch repair (MMR) pathway, is typically marked by the early onset of colorectal cancers and various other cancer types[43]. The defective MMR pathway usually results in elevated microsatellite instability, leading to changes in the length of short repeated DNA sequences within the genome. Genomic damage and instability have also been linked to the theory of ageing, which posits that the accumulation of somatic alterations with increasing age contributes to increasing cellular dysfunction and the effects of ageing[44].

Extrinsic biological agents, such as retroviruses, may also pose a risk to genome integrity. A retrovirus disrupts the cell's genome by inserting a DNA copy of its RNA genome into the DNA of a host cell, potentially damaging cellular proto-oncogenes, leading to further genomic instability. Additionally, some oncogenic retroviruses, known as acute transforming viruses, carry oncogenes within their genome that can transform the cell following insertion into the host's DNA[45]. Several intrinsic cellular defences, collectively termed restriction factors, attempt to prevent or restrict genomic damage resulting from retroviral infection[46]. Examples of these factors include the APOBEC3[47] protein family, a group of cytidine deaminases that induce hypermutation in the viral genome, inhibiting its replication, and TRIM5α[48], which targets the incoming viral capsid, preventing reverse transcription.

1.1.3 Clonal evolution in cancer

The role of genomic damage and instability in cancer evolution began with zoologist Theodor Boveri's research[49] in the early 1900's linking abnormal chromosomal alterations during mitosis with carcinogenesis. Boveri's work was validated by Hungerford and Nowell with the discovery of the Philadelphia chromosome in 1959[50]. This theory was further developed over the course of the 20th century culminating in Nowell research into the clonal evolution of tumour cell populations[51]. Nowell proposed cancer as an evolutionary process involving the sequential selection of mutant subpopulations derived from a common progenitor. Within this process, genetic instability facilitates the acquisition of biological traits associated with tumour progression. In recent years, the emergence of high-throughput sequencing (HTS) datasets has facilitated the validation of numerous aspects of Nowell's model. Truncal mutations, signalling a common ancestral clone, have been detected in most

cancer types[52], while other research has affirmed cancer as an adaptive evolutionary process[53,54]. This analysis together with improved methods for the identification of somatic mutations has prompted further refinements to the clonal model of cancer evolution.

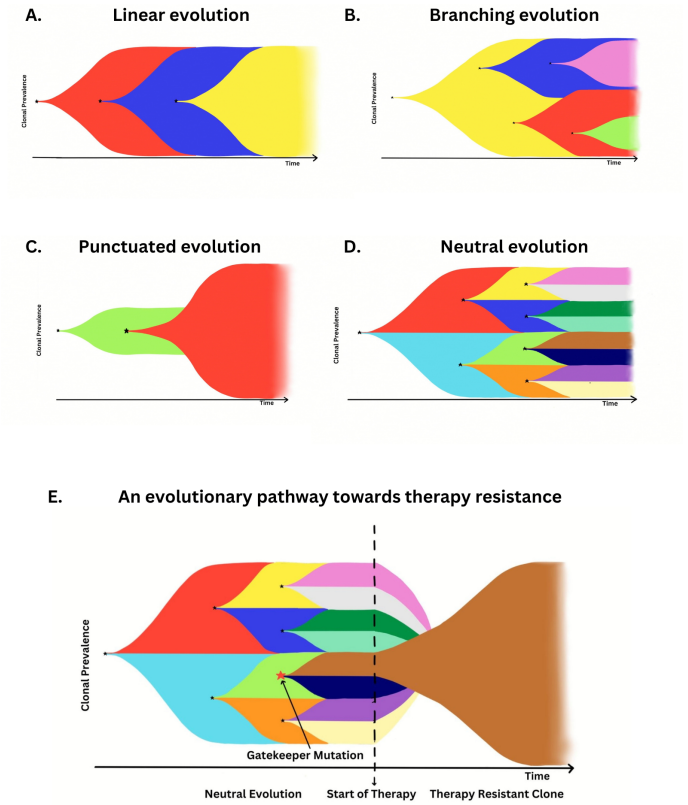

Figure 1.1: Various modes of cancer evolution. **A:** Linear evolution, as characterised by a sequential, 'selective sweep' evolutionary model[51]. **B:** Branching evolution. This follows the classical Darwinian model of evolution, of 'descent with modification' and 'natural selection'[55]. **C:** Punctuated evolution, where a period of stasis is interrupted by a single catastrophic cellular event and a significant increase in the malignant potential of the resulting clone. Examples of this mode of evolution have been identified in numerous cancer types[56,57,58]. **D:** Neutral evolution. All mutations acquired during this mode of evolution have no impact on the clone's ability to survive and reproduce.**E:** One possible evolutionary pathway towards therapy resistance. A gatekeeper mutation acquired prior to the application of therapy confers a strong selective advantage to the clone after therapy commences[59].

Earlier models of clonal evolution, like Nowell's, suggest a predominantly linear evolutionary trajectory (Figure 1.1A) characterised by sequentially dominant clones that drive disease progression. In this linear model of evolution, a clone acquires a new mutation that enables a selective sweep, where the new clone outcompetes all others that do not have the mutation, replacing them in the neoplasm. This

model has been used to explain the linear progression of colon cancer through a series of stepwise mutations leading to sequentially more malignant stages of tumour growth[60]. However, the analysis of numerous large-scale cancer genomic sequencing datasets reveals that many tumours do not exhibit the predominantly homogeneous tumour cell population predicted by this model[61]. The branching model of tumour evolution provides an alternative perspective that accommodates intratumour heterogeneity (Figure 1.1B). In this model, clones diverge from a common ancestor and evolve in parallel within the tumour, giving rise to multiple competing lineages that mirror a Darwinian evolutionary tree[55]. Clones that are more adapted to their microenvironment within the tumour increase their number while others remain stable or die out. The heterogeneity that this model implies may potentially be exploited as a strategy for delaying or preventing the development of treatment resistance. Aggressive treatments that maximise tumour cell kill may inadvertently favour the selection of treatment-resistant clones within the tumour population. In contrast, adaptive therapies exploit the competition between treatment sensitive and resistant populations by adjusting the treatment dose or duration to manage the resistant population and extend the duration of the therapeutic response[62].

In recent years, the application of neutral evolutionary models in cancer genomics has been employed to explain the extent of intratumour heterogeneity observed in many tumours (Figure 1.1D). This hypothesis, borrowed from the field of population genetics, posits that at the molecular level, the majority of polymorphism and substitutions arise from selectively neutral mutations and subsequent genetic drift[63,64,65](Figure 1.1C). The model's application in the context of cancer was formalised by *Williams et al.*[66] who determined that, in a neutrally evolving tumour cell population under exponential growth, the number of mutations at a frequency greater than f should be proportional to $1/f$. This model was employed in analysing a large pan-cancer cohort from the TCGA Consortium, revealing that over 30% of all tumours exhibited evidence of neutral evolution. Neutral evolution is compatible with natural selection as postulated by Darwin. Darwin's theory[55] is concerned with mutations resulting in phenotypic change that confer selective advantage. Neutral evolution applies only at the molecular level to alterations that are selectively neutral. Despite the absence of any selective advantage, these mutations are important in cancer research and clinical application as they increase intratumour heterogeneity[67]. As the tumour microenvironment changes, mutations that previously accumulated neutrally may become targets of strong selection. In effect, these mutations provide a source of standing genetic variation upon which future selective pressures may act. For instance, they could lead to the development of therapy resistance, where a once neutral somatic mutation confers a level of intrinsic resistance, gradually resulting in therapy failure over time[59] (Figure 1.1E). Neutral somatic mutations also contribute to the Tumour Mutation Burden (TMB), a clinically actionable biomarker, due to its relevance to the immune response against cancer[68].

Classic Darwinian models of tumour evolution typically assume that mutations are acquired sequentially in a steady, stepwise fashion, sometimes requiring years to accumulate the necessary drivers to initiate cancer progression. However, with the comparatively recent availability of high-depth tumour sequencing, researchers have uncovered evidence suggesting that, in some cases, a single catastrophic event can result in a high number of genomic aberrations giving rise to malignancy[69]. Examples

of such events include chromothripsis[70], a phenomenon characterised by chromosomal shattering, causing hundreds of genomic breakpoints and aberrant reassembly. Chromoplexy, a related phenomenon, occurs when multiple chromosomes fragment and rejoin, causing extensive and complex restructuring across the genome[57]. Cells that do not undergo apoptosis following such events are at risk of acquiring multiple driver alterations, potentially leading to malignancy. In the case of cells that are already cancerous, this may result in aggressive transformation. From a cancer evolutionary perspective, this process is sometimes modelled as a punctuated evolution (Figure 1.1C). In line with this model, tumour cells undergo extended phases of mutational equilibrium interspersed with brief intervals of intense evolution. During these periods, tumour cells have the potential to accumulate multiple impactful driver events and acquire a more aggressive phenotype.

1.1.4 Germline-somatic variant interactions

The genetic context in which a somatic variant associated with cancer occurs can have a significant impact on tumourigenesis. For the last 50 years, this potential interplay between germline (inherited) and somatic (acquired) variants in cancer progression continues to be the subject of ongoing scientific research. In 1971, Knudson observed that individuals inheriting a mutant *RB1* allele were more frequently diagnosed with early-onset, bilateral retinoblastoma than wild-type patients. In what later became known as the 'two-hit' hypothesis, Knudson proposed that individuals with the mutant germline allele needed only one cell to undergo an inactivating somatic mutation in the remaining wild-type allele for malignant transformation to occur[71]. The Rb gene was considered to be haplosufficient, implying that the quantity of protein generated by a single functional Rb allele was adequate to inhibit tumourigenesis. In effect, the inactivating somatic event in the wild type Rb allele acts as the trigger for an underlying germline vulnerability in the mutant allele that can lead to cancer.

Twenty years after Knudson's research on the Rb gene, the newly named *BRCA1* gene was mapped to chromosome 17 by Dr. Mary King and colleagues[72]. The research marked the culmination of a substantial international effort that began with the genetic epidemiological analysis in 1988 of early-onset familial breast cancer data, proposing the existence of a dominant gene linked to susceptibility to breast cancer[73]. This achievement was all the more remarkable considering that, at that time, in the absence of a human reference genome and modern sequencing techniques, their research relied on painstaking linkage analysis and overlapping DNA clones to map the locus[74]. In 1994, Michael Stratton and Richard Wooster mapped *BRCA2* on chromosome 13[75]. The *BRCA* genes code for tumour suppressor proteins and, like many such genes, are generally believed to follow the Knudson 'two-hit' model, where cancers that develop in carriers of the mutant allele almost always exhibit loss of the wild-type[76,77]. Both *BRCA1* and *BRCA2* were cloned and patented by Myriad Genetics, who enforced the patent on commercial testing for hereditary breast and ovarian cancer susceptibility until it was invalidated by the US Supreme Court in 2013[78].

The discovery of the *BRCA* genes played a pivotal role in advancing cancer research and enhanced our understanding of cancer predisposition. Since then, subsequent advancements in genotyping technology for identifying both copy number and single nucleotide polymorphism (SNP) variation, the compilation of large-scale

cancer genomics datasets, and the application of genome-wide association studies (GWAS) have identified additional germline variants associated with cancer predisposition[79]. These include tumour suppressor genes associated with familial cancer syndromes, such as Li Fraumeni Syndrome[80] (p53), Cowden Syndrome[81] (PTEN), and Peutz-Jeghers Syndrome[82] (STK11). Highly penetrant germline variants in DNA repair pathways, in particular, may pose a significant risk of DNA damage and cancer progression. For example, individuals with Lynch syndrome, an autosomal dominant condition that affects DNA mismatch repair, commonly harbor germline defects in one or more of four repair genes (MLH1, PMS2, MSH2, MSH6), leading to an increased risk of colorectal and other cancers[83]. Genes of interest are typically grouped together for testing on a gene panel, enabling simultaneous sequencing of genes relevant to cancer predisposition. Identifying an individual with a cancer predisposition syndrome enables them to access frequent cancer screening for early detection, significantly enhancing the chances of successful treatment if cancer develops. In cases of high cancer predisposition risk, radical surgical options like the preemptive removal of breasts and/or ovaries before malignancy onset may be considered to minimise the risk[84].

Genomic research continues to advance our understanding of the genetic mechanisms influencing cancer predisposition. Notably, recent studies propose that, in certain instances, tumourigenesis can persist despite the presence of a single functional tumour suppressor allele. Haploinsufficiency in tumour suppressor genes signifies that inactivating mutations in one allele may result in the reduced expression level of the corresponding protein, thereby contributing to tumour development. Tumour suppressor haploinsufficiency has been observed in both cell line and mouse models, including Rb[85], and more recently in BRCA1/2[86]. However, the precise mechanisms underlying this phenomenon remain the subject of ongoing research. Studies have also identified germline variants that result in a decrease in cancer risk. For example, the germline regulatory variant rs3903072 is linked to an increased expression of the tumour suppressor gene CTSW, which is associated with a reduced risk of breast cancer and prolonged survival among patients diagnosed with the condition[87]. Other variants rs10497520-T and rs2242442-G have also been associated with decreased risk in patients with a family history of the disease[88]. Finally, a recent study into the effects of *HLA* polymorphisms in UK Biobank data observed the protective effect of HLA diversity in lung, head and neck, and B cell carcinomas[89]. A better understanding of the role played by variants associated with a reduction in cancer risk could contribute to tailoring personalised approaches for cancer screening and therapy planning.

1.1.5 The tumour microenvironment

For a tumour to grow and invade, it needs to navigate a complex series of interactions with normal cells in surrounding tissues and the supporting structures. It must acquire a blood supply, evade attacks from the immune system, and break through surrounding membranes and fibrous proteins to metastasize to other parts of the body. The tumour microenvironment (TME) encompasses a heterogeneous collection of hematopoietic, mesenchymal, and tumour cells, along with the surrounding tissues and connective structures collectively known as the extracellular matrix (ECM) that supports them. In the context of a tumour, the stroma, a subset of the TME, typically refers to the surrounding ECM and the mesenchymal cells,

such as endothelial cells and fibroblasts, within it[90]. The TME can play a key role in promoting or preventing tumour growth[91,92]. For a tumour to survive, it must manipulate this environment to its own advantage.

A common factor in tumour manipulation of the TME is hypoxia. Hypoxia is the condition in which cells or tissues experience insufficient oxygen levels, impairing their function. As the malignancy progresses, local hypoxia within the TME occurs as a result of oxygen demand in the growing tumour cell mass outstriping supply, and from abnormal or obstructed blood vessel anatomy that disrupts microcirculation[93]. Cancer cells which survive this increasingly harsh environment typically respond by upregulating hypoxia-inducible factor HIF-1 expression and activating alternative metabolic pathways such as glycolysis to continue to grow[94]. HIF-1 has a profound effect on the surrounding TME, upregulating a number of pro-angiogenic factors such as the vascular endothelial growth factor (VEGF). In the absence of a blood supply, the size of a tumour is typically restricted to 2 to 3mm[95]. However, the presence of the VEGF signalling protein in the TME stimulates angiogenesis by enhancing vascular permeability and recruiting neighbouring vascular endothelial cells. This promotes the development of new blood vessels to supply oxygen and nutrients to the tumour, thereby providing it with nearly unlimited growth potential.

Another crucial element influencing the manipulation of the TME during tumour progression is the infiltration of cancer associated fibroblasts. In normal tissue, fibroblast cells primarily produce connective tissue in the extracellular matrix (ECM) and play essential roles in communication with various other cell types during normal tissue homeostasis and wound healing[96]. However, within the TME, fibroblasts undertake distinct functions, prompting the characterization of these cells as cancer associated fibroblasts (CAFs). Various factors like TGF-β1, osteopontin (OPN), and interleukin-1β (IL-1β), released from cancer or immune cells, induce stromal fibroblast transition to cancer associated fibroblasts (CAFs) by modulating the TGF-β and NF-κB signalling pathways[97]. Cancer-associated fibroblasts (CAFs), frequently comprising a substantial part of the tumour stroma[93], play a crucial role in remodelling the extracellular matrix (ECM) to facilitate tumour growth and angiogenesis[90]. CAFs secrete matrix metalloproteinases, enabling ECM degradation, followed by its resynthesis to facilitate the invasion of neighbouring tissues and the establishment of a blood supply to the tumour[98]. Additionally, they can promote angiogenesis directly by secreting growth factors such as VEGF[99,100].

CAFs further interact with tumour infiltrating lymphocytes and other immune cells within the TME. Through the secretion of various cytokines, growth factors, chemokines, and exosomes, CAFs actively suppress the immune response within the TME, allowing the tumour to evade destruction by the immune system. In prostate cancer, CAFs recruit monocytes to the tumour through stromal-derived growth factor-1, promoting their differentiation into the protumourigenic M2 macrophage phenotype. This interaction between CAFs and M2 macrophages may enhance tumour cell motility and potentially lead to metastatic spread[101]. Mouse models have demonstrated that CAFs can suppress the recruitment of cytotoxic T cells into the tumour, conferring resistance to Immune-Checkpoint Blockade[102]. Additionally, CAFs have been observed to stimulate the differentiation and migration of Treg cells, leading to their recruitment and activation within the TME and the induction of tumour immune tolerance[103]. The complex crosstalk among cancer, stromal, and immune cells within the TME is pivotal in cancer progression and remains an ongoing

focus of research.

1.1.6 The immune system and cancer

The immune system has long been suspected of playing a significant role in defending the body against cancer. In the late 19th century, bacterial pathogens were used to trigger an anticancer immune response and achieve disease remission in some patients[104]. However, during that period, the limited understanding of the immune system's role in cancer prevention constrained further progress. Throughout the first half of the 20th century, an enhanced understanding of the immune response and experimental evidence demonstrating immune suppression of transplanted tumour models culminated in the formalisation of the theory of immune surveillance by Burnet and Thomas in 1957[105]. In relation to cancer, immune surveillance refers to ongoing monitoring by the immune system to detect and eliminate premalignant or malignant cells in the body. Enhanced understanding of mechanisms of adaptive immunity prompted the validation of immune surveillance in knockout mice models[106] and a revised concept of tumour immunoediting[107]. The immunoediting theory posits that a tumour's progression, ultimately leading to its escape from suppression by the immune system, can be categorised into three distinct phases: elimination, equilibrium, and escape[108].

The elimination phase of immunoediting is initiated when the body's immune defences detect the presence of malignant cells. Cytotoxic T cells are activated by detecting oncogenic peptides attached to MHC class II receptors on the cell surface of antigen-presenting cells, such as macrophages or B cells, or by their direct presentation on MHC class I receptors of cancer cells themselves. Activation of natural killer cells may also occur if they detect the presence of oncogenic proteins or protein dysregulation from a wide variety of antigen targets present on a cancer cell's surface[109]. Alongside the direct destruction of cancer cells by the secretion of perforin and granzymes to initiate apoptosis, these immune cells also release cytokines that help modulate the immune response, such as IFN-γ, to promote macrophage activation, enhance antigen presentation, regulate other T cells within the tumour and effect the clearing of the tumour cell population[110].

Although, in the majority of cases, the immune system will succeed in clearing the cancer cells, in some instances, a tumour population will persist and progress to the next phase of immunoediting. The second phase, known as equilibrium, results in tumour stasis. During this phase, the rate at which cancer cells are replicating approximately equals the rate at which they are cleared by the immune system. Despite the malignancy being kept in check by the immune system, it is not fully cleared. Further changes within the tumour microenvironment and the cancer cell population may also occur, leading to the survival and growth of cancer cell variants that can evade immune recognition and destruction. Although there is no appreciable change in tumour size during this period, there is an important and ongoing interplay between the immune system, tumour microenvironment, and tumour cells that will define the final outcome of the immune response. The balance of effector, helper, regulatory T, and other immune cells that is set during equilibrium will dictate how the disease progresses. A robust immune response will ensure cancer regression, while the induction of immune tolerance will ensure continued tumour growth[108].

During tumour escape, the final phase of immunoediting, the balance in the

immune response to cancer shifts to one of immune tolerance, resulting in tumour progression. Increased heterogeneity within the tumour population that accumulates during the equilibrium phase results in a subpopulation that is less immunogenic and escapes immune surveillance. This may be achieved by cancer cell secretion of cytokines and chemokines that recruit immunosuppressive cells into the tumour microenvironment[111], or IDO to suppress macrophages and effector T cell responses[112]. An established tumour may also express cytokines such as TGF-β and IL-10 to modulate immune and stromal cell functions, creating tumour immune tolerance[113]. PD-L1 overexpression by cancer cells may also be a factor, promoting T cell anergy or apoptosis and tumour tolerance[114]. Mutation or dysregulation of antigen-presenting pathways, leading to the loss of MHC elements[115], is another mechanism by which tumour cells attempt to avoid detection by the immune system. An additional evasion strategy may include the shedding of MICA/B surface proteins by cancer cells. MICA/B is a cell surface protein indicative of cell stress that engages with NKG2D, activating natural killer cells or providing costimulatory signals to T cells. Shedding of MICA/B by cancer cells induces endocytosis and degradation of NKG2D, desensitising natural killer cells, and also impairing effector T cell responses[116,117].

In recent years, research in this field has explored longitudinal whole-exome sequencing data of cell-free DNA from a single patient. This data tracks the evolutionary dynamics of the tumour and immune evasion across a 12 year trajectory, identifying a significant number of somatic mutations believed to play a role in immune evasion and disease progression[118]. Ongoing research in this area may in future assist clinicians in selecting appropriate therapies to react to the evolving cancer in the patient.

1.1.7 The role of epigenetics in cancer

The term 'epigenetics' was coined by the developmental biologist Conrad Waddington in 1942 to describe an unknown mechanism that would explain how the same genome can give rise to different cell types in a multicellular organism. The word 'epigenetics' is derived from Greek, literally meaning 'over genetics'. In its most general sense, the term now relates to modifications of gene expression that are independent of DNA sequence[119]. It is now known that, in somatic cells, epigenetic modifications are dynamic and reversible, may be heritable and play a crucial role in cell differentiation, determining cell fate[120]. The first epigenetic mechanism involving DNA methylation was proposed in the mid-1970s[121] and its role in gene regulation and cell differentiation validated in 1981[122]. DNA methylation regulates gene expression by either recruiting proteins associated with gene silencing or by inhibiting the binding of transcription factors. This enables differentiated cells to develop a stable DNA methylation pattern that regulates tissue-specific gene transcription[120]. Methylation of histones may also occur and can activate or repress gene expression depending on which residue is methylated[123].

In the mid-1990's, David Allis and his team outlined an additional epigenetic mechanism involving the acetylation of histone residues and its corresponding impact on chromatin structure[124]. Histone proteins act as a scaffold around which DNA is packaged into what is called a nucleosome. A section of DNA packaged into a series of nucleosomes is referred to as euchromatin, a structure that resembles beads on a string. Multiple nucleosomes may in turn be packed tightly into arrays in a

compact structure referred to as heterochromatin. The condensed DNA structure in heterochromatin helps to protect the DNA from damage and, importantly, prevents transcription of any genes within genomic regions packaged in this way. In contrast the decondensed DNA structure in euchromatin does not act as a barrier to transcription. The research[124] concluded that the epigenetic enzyme GCN5, a histone acetyltransferases adds acetyl groups onto lysine residues of cellular proteins such as histones, to remodel chromatin and regulate transcription. Acetylation reduces the positive charge on histones, relaxing the chromatin structure and is associated with increased gene transcription.[125].

In recent decades, our understanding of the mechanisms governing epigenetic regulation, particularly concerning the dysregulation of gene expression in cancer, has advanced significantly. Of particular concern is the dysregulation in methylation patterns, which may provide cancer with a mechanism to repress tumour suppressor genes or express genes linked to immune escape[126]. First described in the 1980s[127], DNA hypomethylation, characterised by a decrease in the epigenetic methylation of cytosine and adenosine residues in DNA, has been observed across a number of cancer types and is associated with poor prognosis and the risk of relapse[128,129]. Removal of methylation patterns can have a significant impact, resulting in gene activation and dysregulation of gene expression[130,131]. Similarly, hypermethylation, an increase in the epigenetic methylation of cytosine and adenosine residues in DNA, is frequently observed in cancer[132]. Hypermethylation of DNA in cancer often manifests in regulatory regions such as promoters and enhancers of tumour suppressor genes, suggesting that cancer-related epigenetic modifications may function as a driver of the disease[133].

Issues related to histones or chromatin remodelling may also have serious and widespread implications for the cell. Somatic mutations in chromatin remodelling genes, particularly within the *SWI/SNF* complex, have been identified as affecting more than 20% of human cancers across various tumour types[134]. Additionally, mutations in genes encoding histone acetyltransferases have been recognized as playing a significant role in several cancer types. These mutations may function as tumour suppressors or oncogenes in cancer transformation and serve as biomarkers in predicting patient survival[135]. Finally, mutations within genes that code for histone modifier enzymes related to chromatin methylation have also been linked with tumour development and metastasis[136].

1.2 Somatic mutations and their clinical relevance

1.2.1 Types of somatic mutations

It has been estimated that the genetic makeup of any two individuals on earth is on average 99.9% identical[137]. However, this minor variation and the environment within which it occurs is often a key aspect of a person's health and susceptibility to disease. At the molecular level, differences in the genome of two cells taken from the same individual are even more infrequent and estimated to be orders of magnitude less than typical inter-individual differences[138]. Although subtle, these differences also have the potential to impact a person's health and an in-depth understanding of their consequences is a key objective in many health-care settings. In genetics, a variant is defined as any alteration in DNA sequence relative to a reference genome. Variants are typically classified as either germline or somatic.

Each of us was born with our own unique set of germline variants that were present at zygote formation and are now contained within the DNA of every cell in our body. In contrast somatic mutations are alterations to a cell's genome that occur after the formation of the zygote, in somatic (non-germ) cells. As a result somatic mutations are unique to a cell or group of cells within the body. Somatic mutations usually arise through misrepaired DNA damage and once acquired may be passed on to all lineage descendants of the cell. They occur in both healthy and diseased cells and vary significantly in abundance across all tissue types[139]. Somatic mutations have been linked with disease[140] and the ageing process[44] however most are assumed to be silent, having minimal phenotypic consequences for both the cell and organism[141].

The types of genetic alterations that arise from somatic mutations and the consequences for both the cell and organism also vary significantly. The simplest type of modification, a single-based substitution (or 'point' mutation), is also one of the most frequently occurring somatic variants. Genomic location is crucial when assessing clinical impact and somatic variants are further classified as occurring in a coding (i.e., within exons) or non-coding region of the genome. Coding regions only account for approximately 1.5% of the human genome[142] however the potential functional impact on protein of somatic variants within coding regions is significant. The degeneracy of the genetic code implies that nearly a quarter of all possible single base substitutions (SBSs) listed in the codon table are essentially interchangeable, and as such do not result in amino acid changes in translated protein[143]. These 'synonymous' SBSs are generally considered to have no biological impact, although there is a growing awareness that in a minority of cases, effects on protein due to other mechanisms such as splice site modifications may occur[143]. Nonsynonymous mutations are of particular interests although assessing their functional impact is not always straightforward. A missense mutation is a type of nonsynonymous mutation that results in an amino acid change in the translated protein. The biological impacts of missense mutations also vary significantly depending on the chemical properties of the amino acid change and the location within the peptide where to alteration occurred. A missense mutation is said to be conservative if the amino acid change has no functional effect on protein. Non-conservative mutations, in particular those involving hydrophobicity changes within core peptide regions may deform the molecule disrupting binding affinities that result in loss of normal function or gain of toxic function. The biological implications of missense variants are not always obvious and *in-silico* algorithms such as SIFT[144] and PolyPhen-2[145] which compare sequence homology and the physical properties of amino acids are often used to predict functional impact. SBSs may also result in the insertion of a stop codon into a coding sequence. This creates a 'nonsense' mutation (or protein truncated variant PTV) that usually implies incomplete protein product and loss of function. Indels or short insertions and deletions within the genome are another type of mutation commonly identified in variant caller output. These are further classified as 'in-frame' if the amount of DNA gained or lost is divisible by 3 (the number of bases in a single codon) or alternately they are referred to as a 'frameshift'. Frameshifts usually have a significant impact, typically resulting in a new amino acid sequence downstream of the mutation including premature stop codons that truncate the protein. Similarly to indels, microsatellite instability; the extension or contraction of short repeated DNA sequences (usually less than 50bp) may also lead to a frame shift within some exons.

Although less likely to cause disruption, SBSs and indels that occur outside of exonic regions may also have functional implications for both cell and organism. Somatic mutations that modify core splice sites can cause introns to be retained during splicing or modification or skipping of exons, leading to significantly altered protein isoforms. Mutations at promoter sites that change the binding affinity of transcription factors and dysregulate gene expression are also associated with several diseases[146]. Larger and more complex somatic alterations, referred to as structural variation may also occur. Structural variants (SVs) are genomic rearrangements generally defined as encompassing at least 50 bp and may traverse both coding and non-coding regions of the genome. The affected region is typically <1 kb and may take many different forms such as copy number alterations (amplification or deletion of copies of a DNA segment) , inversions (a broken segment reattaches with reverse orientation) and translocations (a broken segment reattaches within a different chromosome). Somatic structural variation may result in gene fusions, the amplification of oncogenes or the deletion of tumour suppressors and is associated with a number of malignancies[50] as well as developmental, neurological or neuromuscular disorders involving the somatic extension of a critical germline triplet repeat variant that causes complications as the patient ages[147,148]. Targeted assays for SV detection are used in clinical diagnostics, however detecting structural variation from Next Generation Sequencing (NGS) data poses a number of challenges, which in turn complicates research although a number of dedicated software tools for detecting copy number alterations (CNAs) and SVs are available[149,150].

1.2.2 Mutational signatures

The biological DNA damage processes that give rise to somatic mutations are of particular interest in cancer research and clinical practice. Establishing the cause of somatic mutations that give rise to cancer ultimately allows us to explain the origin of the cancer itself. As we age DNA damage accumulates in our cells as a result of exposure to numerous mutagenic processes at various stages throughout life. Exposure may be transient (ie., UV damage) or it may take the form of a constant threat from reactive oxygen species (ROS) produced as a by-product of normal cellular metabolism. Cellular DNA repair machinery ensures the integrity of the genome in the face of DNA damage, but it is not 100% effective. Misrepaired lesions can become integrated into the cell's genome, establishing a record of mutational events traced right back through the cell's lineage. The types of genomic changes brought about by these somatic mutations depend on the mutational process that gave rise to them with each process leaving its own signature on the type of DNA damage it creates. Researchers have exploited these differences to break down the overall somatic burden found in various cancer types and identify the mutagenic sources involved and their proportional contributions to the overall tumour somatic burden[151].

The most common approach used to identify signatures of mutational processes is to focus on single base substitutions. Each process will typically be associated with a set of base substitution types (from original to mutated base) and sequence contexts (the 3' and 5' bases adjacent to the damage) within which the damage occurs. These are usually categorised according to the base change, referred to by the pyrimidine of the mutated base pair (i.e. C>A, C>G, C>T, T>A, T>C and T>G), and adjacent 5′ and 3′ bases into 96 types[152]. The relative contributions

of mutations generated within each substitution class and sequence context for a given genome and target region are often represented graphically in a mutational profile plot. Individual signatures of mutational processes that have left their mark on mutational profiles across a cohort of individuals with a particular type of cancer are usually extracted from large scale cancer genome datasets using non-negative matrix factorisation (NMF). The analysis considers the mutations in a sample are the result of the activity of a number of distinct processes, each with a characteristic mutational signature (described by the relative contributions of each of the 96 mutation types) and that these processes make additive contributions to the overall mutational load in the sample. To resolve these signatures and their relative contributions in each sample the somatic mutation data for a specific cancer type is compiled into a mutation count matrix that records the number of each of the 96 mutation types for every cancer sample in the cohort. This is then factored into the product of a mutational signatures matrix and exposure matrix (representing the relative contribution of each mutational signature in each sample). The process iterates until it converges on the set of mutational signatures and corresponding exposures that best describe the data[152].

Mutational signatures have provided unique insights into disease aetiology in a number of cancer types[152]. COSMIC Signature 7, primarily composed of C>T mutations, is most commonly found in malignant melanomas and retains the hallmarks of pyrimidine dimers and other lesions induced by exposure to ultraviolet light[153]. Signature 4 is found mainly in cancers associated with smoking and indicative of failed nucleotide excision repair of DNA adducts likely formed as a result of exposure to tobacco carcinogens[154], while signature 24 found exclusively in liver cancer has been associated with exposure to the carcinogen aflatoxin B1[155]. More recently research has focused on translational applications of mutation signatures in oncology to aid in cancer prognosis and identify therapy sensitivity. A significant contribution from sig 4 has been associated with shorter survival and higher tumour mutational burden[156] in a large cohort of non-small cell lung cancer patients (500 individuals). Similarly the APOBEC signature (COSMIC signature 13) has also been associated with poor prognosis and high mutation burden in multiple myeloma[157] while signature 17b has been associated with a decline in progression-free survival in EGFRi treated colorectal cancer[158]. Mutational signature analysis has also been used to inform therapy selection in cancer. Biomarkers of homologous recombination DNA-repair deficiency (HRD) are an important consideration in clinical decision making and may indicate tumour sensitivity to a number of HR-targeted therapies such as poly(ADP-ribose) polymerase (PARP) inhibitors[159]. Using machine learning techniques (supervised lasso logistic regression) researchers have identified a set of mutational signatures predictive of HR deficiency (BRCA1/BRCA2) that has been validated on independent cohorts of breast, ovarian and pancreatic cancers[160]. A separate study in pancreatic cancer used signature analysis to identify MMR or HR deficient tumours and increased CD8+ T cell infiltration within the cohort. Traditional methods of identifying DNA repair deficiency have relied on the detection of loss of function mutations in key driver genes. Mutational signature analysis provides an important orthogonal approach in this instance as issues such as epigenetic silencing of DNA repair[161], variant caller false negatives or mutations in parallel pathways of unknown significance may confound clinical decision making. Indeed both studies noted a significant number of HR deficient patent tumours without

the detection of a corresponding loss of function mutation in *BRCA1*, *BRCA2*, or *PALB2*.

There are however a number of issues preventing the widespread use of mutation signatures in clinical oncology. Mutation signature analysis requires whole genome or at a minimum whole exome sequencing data. Targeted gene panel sequencing is currently the most commonly used cancer NGS assay in clinical practice and usually interrogates only a small fraction of the exome and is therefore unsuitable for signature analysis. Formalin fixation, paraffin embedding (FFPE) of patient samples, a standard practice in clinical pathology also presents challenges. FFPE sample treatment may lead to a significant number of low frequency false positive calls in somatic variant data[162,163]. Similar to true somatic mutation calls, FFPE artefacts leave a characteristic mutational footprint across the genome that may obscure or confound the detection of true mutational signatures[164]. Similar issues may be caused by other artefacts associated with sample preparation (for example oxidative damage). In addition, reliable *de novo* extraction of mutational signatures in small cohorts or in cancers with a low mutational burden is challenging[165]. Issues such as these have resulted in half of the 79 SBS signatures detailed in COSMIC being listed under unknown aetiology or possible sequencing artefact, while in fourteen of the remaining signatures it is unclear if the evidence for the signature supports the aetiology proposed[165]. The increased use of WGS (whole-genome sequencing) and WEX (whole-exome sequencing) in clinical practice, large-cohort based validation of mutational signature aetiologies and continuing advances in bioinformatic methods of artefact removal are likely to give rise to new clinical applications of mutational signatures in the future.

1.2.3 Cancer, molecular diagnosis and prognosis

The clinical use of somatic mutation detection began in the 1970s with the development of Southern blot to detect gene duplication and rearrangements in DNA isolated from cancer cells. By the early 1990s a similar technique known as Fluorescence *in situ* hybridization (FISH) based on earlier work by Pardue and Gall[166,167] began to emerge that allowed for fluorescent staining of a specific DNA sequence on a human chromosome in metaphase or interphase cells leading to improved diagnosis in malignancies such as leukaemia[168]. Both southern hybridization and FISH techniques continue to be routinely used in cancer diagnostics. Since then, translational advances in cancer genomics have led to the emergence of molecular pathology as an integral component of cancer screening, diagnosis and management. Molecular pathology has been defined as the testing of nucleic acids within a clinical context[169]. In the broadest sense, the types of molecular testing involved need not necessarily be genomic and may encompass techniques such as immunohistochemistry (IHC). Molecular pathology augments traditional histopathological and immunohistochemistry techniques to provide diagnostic information to oncologists and other clinical professionals to aid prognostication and treatment stratification. The field continues to evolve as new biomarkers for targeted therapies are discovered and our understanding of the pathogenicity of cancer-related germline and somatic variants improves.

The primary clinical modalities for somatic variant identification in DNA or RNA are PCR-based detection, targeted NGS (gene panel) and ctDNA (circulating tumour DNA) assays. In addition, a limited number of clinics offer whole exome or

rarely whole cancer genome sequencing tests. PCR-based detection and quantification methods (such as real-time / quantitative PCR, or digital PCR) are generally a first choice strategy that provide simple low-cost detection suitable for a diverse range of DNA/cDNA targets. They are used to detect genetic alterations, somatic biomarkers and viral transformations related to cancer. The first clinical application of the technology, for the detection of leukaemia, dates back to 1988[170,171]. PCR-based detection methods are often single target assays and suited to screening for disease[172] and risk stratification[173]. The assay is highly sensitive and can potentially be turned around in a short time frame[174]. There are drawbacks however. Single target molecular assays like PCR need to be run sequentially when required to interrogate multiple loci within the tumour. In most instances however there will simply not be enough tumour DNA to support a piecemeal approach. Follow-up biopsy, typically performed with fine-needle aspiration or core needle yields a limited amount of tissue that may need to be further divided between histopathology, immunohistochemistry and molecular diagnostics. Furthermore, high tumour cellularity in the biopsy is not guaranteed. These issues are reflected in the main reasons cited for molecular testing failure in lung cancer; insufficient amount of tumour cells (83%), inadequate tissue quality (55%)[175], confirming the long standing mantra among pathologists and oncologists that 'tissue is the issue'. Multiplex PCR (mPCR), which can simultaneously detect a number of sequences using multiple primers in a single reaction may be an option to conserve lab resources, and most importantly biopsy tissue. The assay is significantly more complex than single target (singleplex) PCR, requiring sophisticated instrumentation and comes with an additional set of multiplex-specific challenges (for example primer incompatibility, particularly as the number of targets increases)[176,177]. It is used successfully in some clinical settings. However, a bespoke mPCR assay for each set of targets relating to a specific cancer type is not a viable option in many laboratories.

Gene panel or targeted NGS assays provide an alternative method of interrogating multiple genomic targets from a single sample by performing hybridization-based enrichment of a specific subset of clinically relevant gene and regulatory regions followed by sequencing and variant analysis. In effect, it provides a cost effective approach that focuses limited sequencing resources directly on disease relevant genomic regions enhancing both the sensitivity and resolution of clinically actionable variant detection[178,179]. Gene panels can also detect fusion events and copy number variation[180], although these may also be obtained with IHC or FISH depending on the gene panel available to the clinician. Unlike PCR-based tests which confirm the presence or absence of a particular target variant, gene panels capture the entire content of target regions to significantly enhance the discovery power of the biopsy. This is often a key factor when, for example, deciding on an assay to identify the mechanism of drug resistance in an EGFR+ patient who is no longer responding to treatment. Third generation of osimertinib, a potential treatment option, has 14 associated point mutations each conferring drug resistance, and is just one of a number of drugs in its class. Resistance may also arise in parallel or downstream of the inhibited EGFR, requiring comprehensive and accurate molecular diagnostics to inform the next choice of treatment. A number of off-the-shelf cancer gene panel diagnostics (or oncopanels) are available, including FDA approved solutions such as F1CDx and MSK-IMPACT[181] and other targeted NGS methods that have been analytically validated in many clinical laboratories[182,183] or received approval

by other regulatory authorities[184,185]. However, targeted NGS approaches have their drawbacks. Low DNA content may be an issue. Although high sensitivity assays are available[186] a typical oncogene panel requires at least 200 nanograms of DNA and a minimum tumour purity of 20%[187] while other targeted PCR methods require only a fraction of that amount[186]. Turnaround time may be an issue in some cases with typical estimates at 14 days[187] .

Recent years have seen an exciting development in translational oncology with the emergence of cell free DNA (cfDNA) as a biomarker in a number of clinical applications. The diagnostic assessment of circulating nucleic acids found in blood, urine, and other body fluids is commonly referred to as liquid biopsy. The first identification of cfDNA in human blood was made by Mandel and Metais in 1948[188]. Circulating tumour DNA (ctDNA) is now being used to diagnose drug resistance or disease recurrence following therapy[189,190]. Liquid biopsy is cheaper and less invasive than traditional biopsy methods. Disease surveillance and minimal residual disease (MRD) detection are an important application of liquid biopsy in cancer care and a number of important trials are currently in progress to evaluate its efficacy in guiding cancer treatment post resection[191]. The use of liquid biopsy in cancer is still an evolving technology and a number of challenges remain. The half-life of ctDNA in the bloodstream is brief[192] and many factors may affect the quantity of ctDNA released[193,194,195] causing low sensitivity in liquid biopsy assays and generally requiring initial diagnosis to be obtained from a tissue biopsy. This issue is somewhat offset by the fact that, given its low invasiveness, the test can be repeated. Liquid biopsies are generally classified as tumour-informed or uninformed. With a tumour-informed assay, a molecular profile of clonal variants from the primary tumour is first recovered by NGS from a tissue biopsy and used to inform the presence of ctDNA in later plasma-based tests when testing for other biomarkers[196]. Tumour-uninformed assays on the other hand do not sequence the primary tumour[197]. Subsequent somatic variant identification from ctDNA may be performed by a quantitative PCR or NGS assay as required. Research is also ongoing into potential future applications of liquid biopsy including the use of fragment characteristics such as DNA molecule length to identify the presence of ctDNA within a cfDNA sample as an early detection mechanism for cancer[198].

1.2.4 Targeted therapies for cancer

Targeted therapy as defined by the US National Cancer Institute, is a treatment that targets specific proteins that cancer cells use to grow, divide and spread[199]. Signalling pathways in cancer cells involved in growth, the cell cycle and damage repair pathways often contain vulnerabilities that can be exploited by drugs to stop the malignancy from progressing. Research into targeted therapy dates back to the 1970s with the drug tamoxifen, a selective oestrogen receptor modulator (SERM) that was originally intended for use as a contraceptive[200] (Figure 1.2). A long established link between oestrogen and breast cancer[201] ultimately led to it being repurposed as the first successful endocrine therapy for cancer.

Oestrogen stimulates the proliferation of breast epithelial cells by binding to intracellular (ERα and ERβ) or cell surface (GPER1) receptors, which in turn activate transcriptional processes and/or signalling cascades controlling gene expression[202]. Oestrogen signalling pathways are commonly dysregulated by receptor overexpression (ER+) in breast cancer[203]. Tamoxifen, a selective oestrogen receptor modula-

tor has tissue selective actions that is used in the treatment of breast cancer[204]. As with many targeted therapies, accurate detection of somatic variants from patient biopsy can significantly inform the clinical use of tamoxifen. Somatic mutations that cause amino acid changes in the oestrogen receptor alpha ligand binding domain (Y537S and E380Q) give rise to a constitutively active and antagonist resistant receptor[205,206] while genetic polymorphisms impacting tamoxifen metabolic activity have also been shown to play a role in resistance[207].

Growth and angiogenic signalling pathways, in particular those mediated by receptor tyrosine kinase, are commonly dysregulated in cancer and frequently the subject of targeted therapy research and clinical application[208]. Receptor tyrosine kinases (RTKs) are a group of single-pass membrane-bound signalling receptors that regulate many normal cellular processes. They contain an extracellular ligand-binding domain and an intracellular kinase domain (Figure 1.2). Targeted drugs are typically classified into two categories, small molecule inhibitors and monoclonal antibodies. Small molecule inhibitors typically weigh less than 900 daltons[209]). This allows the molecule to diffuse across cell membranes and reach intracellular targets whereas monoclonal antibodies can only target the extracellular domain.

The epidermal growth factor receptor family (EGFR, HER2 etc.) is frequently dysregulated across a large number of cancer types and a common target for tyrosine kinase inhibition therapy (TKI). Cetuximab (typically used in the treatment of bowel and colorectal cancer, non-small cell lung cancer and unresectable squamous cell skin cancer) and panitumumab (a single-agent treatment of metastatic colorectal carcinoma) are examples of monoclonal antibody therapies that target EGFR receptor tyrosine kinases by out competing epidermal growth factor ligands for the extracellular binding site and preventing EGFR receptor dimerisation that triggers the signalling cascade[210,211] (Figure 1.2). Both antibodies bind to different sites on the EGFR and somatic mutations associated with these regions (leading to amino acid changes S492R or S468R that confer cetuximab resistance) can play an important role in selecting between both therapy options[212,213].

Overexpression of the EGFR ligand or somatic mutation giving rise to amino acid change G465R in EGFR can result in a loss of sensitivity and treatment resistance with both cetuximab and panitumumab while methylation changes and mutations in the *EGFR* kinase domain are correlated with disease progression in the presence of cetuximab[214,213]. Similarly HER2 monoclonal antibody therapies such as pertuzumab and trastuzumab are also vulnerable to ligand overexpression[215] and somatic mutations in HER2 that confer resistance to trastuzumab have been observed and may guide the selection of other more appropriate treatments[216]. All epidermal growth factor receptor family targeted therapies are vulnerable to activating mutations in *RAS* or inactivating mutations in *PTEN* which stimulate signalling downstream of the EGFR and acquired kinase inhibitor therapy resistance[217,218].

Additional monoclonal antibody targets include the VEGF pathway to inhibit tumour angiogenesis and growth. Bevacizumab acts by selectively binding circulating VEGF, inhibiting binding to VEGF receptors[219] (Figure 1.2). This inhibition leads to a reduction in tumour neovascularization and growth. Mechanisms of resistance to VEGF monoclonal antibody inhibition are complex[220] although recent research has highlighted the possible role of somatic mutations in *KRAS* in conferring therapy resistance[221].

Small molecule inhibitors that can traverse the cell membrane have opened up

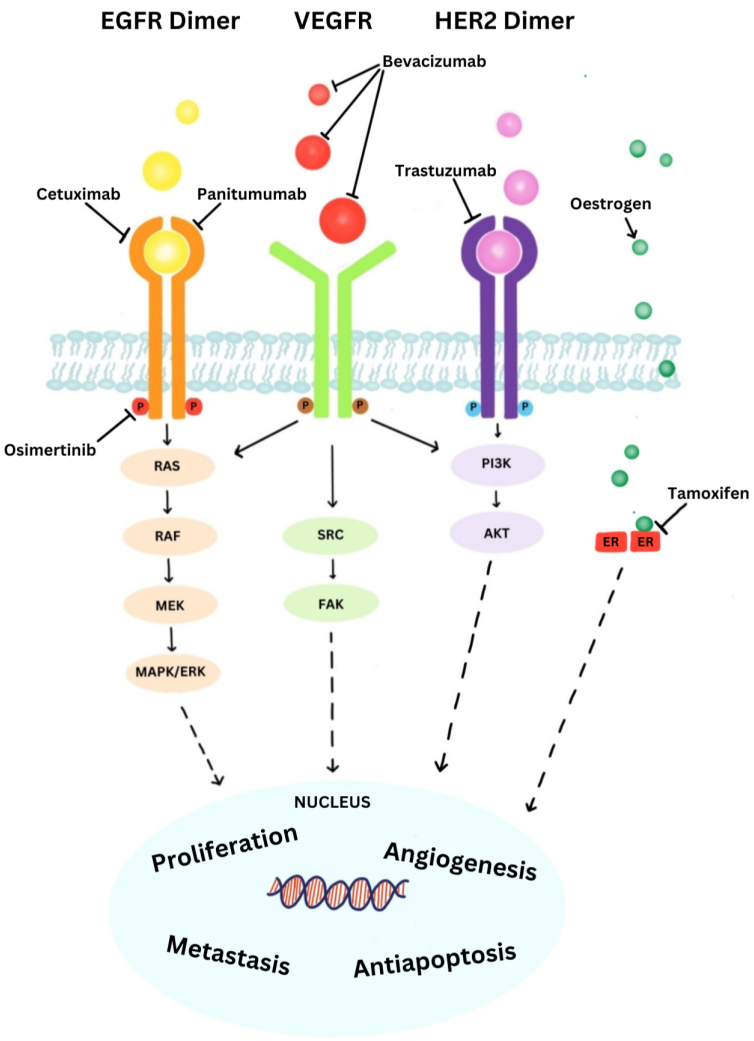

Figure 1.2: An overview of the signalling pathways targeted with monoclonal antibodies therapies cetuximab, panitumumab[210,211,212,213], bevacizumab[219], trastuzumab[215] and small molecule inhibitor tamoxifen[204].

new targets for kinase inhibition. In patients harbouring mutations in the epidermal growth factor receptor family or demonstrating receptor overexpression these molecules effectively target the intracellular ATP binding site of the tyrosine kinase domain, out competing ATP and preventing cascade activation[222] (Figure 1.2). Significant improvements in the potency and specificity of these therapies plus their ability to target truncated forms of the EGFR and HER2 receptors have made them a valuable option for kinase inhibitor therapy[223]. Three generations of small molecule tyrosine kinase inhibitors have been developed, each targeting one or more members of the epidermal growth factor receptor family with different mechanisms aimed at specific activating mutations and mechanisms of resistance[224]. Patients generally undergo screening for somatic mutations in *EGFR*[225] and other somatic variants known to confer therapy resistance before being assigned therapy. An important biomarker in small molecule EGFR inhibitor treatment is a somatic mutation in the *EGFR* gene, causing the amino acid change T790M[226], within the

tyrosine kinase domain of the EGFR that confers resistance to first and second generation therapies. Third generation therapy osimertinib overcomes this resistance by selectively binding to and inhibiting the mutant form of *EGFR* to prevent EGFR-mediated signalling, although further resistance is often acquired[227]. The ATP-binding pocket in the VEGF receptor is also a target of small molecule mediated inhibition in cancer therapy. As with monoclonal antibody VEGF inhibition, resistance to VEGF receptor targeted therapy is complex, while some studies have indicated that downstream mutations in the AKT pathway may also contribute to resistance[228].

In recent years, new therapies have emerged that target other pathways commonly exploited by cancer cells to grow and evade therapy. Cyclin-dependent kinases (CDKs) are enzymes that regulate critical checkpoints during cell cycle progression. CDK4 and CDK6 (CDK4/6) inhibitor therapy has demonstrated significant efficacy in combination with endocrine therapy, in many cases overcoming endocrine therapy resistance[229]. The mechanism of action of CDK inhibitors is similar to that of other small molecule inhibitor therapies. CDK inhibitors target the APT binding site or adjacent hydrophobic pocket of CDK, preventing phosphorylation, inactivating the kinase and preventing cell cycle progression. Somatic mutation screening for loss of function of the *RB1* gene that codes for the retinoblastoma protein is an important prerequisite in therapy selection to identify treatment resistant patients[230].

Deficiencies in DNA repair pathways in cancer cells are also exploited as therapeutic targets. Tumour suppressor genes in the homologous recombination repair pathway that are involved in the repair of double strand breaks (for example BRCA1 and BRCA2) are commonly mutated in breast and ovarian cancers rendering them sensitive to cytotoxic treatments such as radiotherapy[231]. PARP inhibitors use a synthetic lethality approach to exploit deficiencies in homologous repair by targeting the alternative repair pathways used by cancer cells. The small molecule inhibitor binds to the PARP protein following a single strand break, preventing the recruitment of base excision repair machinery[232] and resulting in the conversion of DNA single-strand breaks (SSBs) to DSBs. This results in an increase in the number of DSBs resulting from radiotherapy and a significant improvement in the therapeutic response. Patients are pre-screened for loss of function somatic mutations in the homologous repair pathway to ensure they are suitable for PARP combination therapy.

1.2.5 Immunotherapy and tumour mutation burden

The mobilisation of the body's own immune system to attack malignant cells is perhaps the first nonsurgical intervention to demonstrate efficacy in cancer treatment. At the end of the nineteenth century, building on previous research by German physicians, Fehleisen and Busch[233], William Coley successfully treated inoperable cancer patients by injecting tumours with immunoadjuvants derived from heat-inactivated bacterial pathogens to elicit a potent antitumour immune response[234]. As a result of his work, Coley is now recognized as the 'Father of Cancer Immunotherapy'[104]. In recent decades a more comprehensive understanding of the immune system has provided a mechanistic explanation of Coley work and its relevance to modern day cancer treatment. The discovery of T cells in particular and their critical role in the adaptive immune response[235] helped to pave the way for modern immunotherapy regimes in use today. The concept of adoptive cell transfer therapy (ACT), where

tumour infiltrating lymphocytes (TIL), usually isolated from a tumour biopsy are cultured in large numbers *ex-vivo* and reintroduced into the body to combat the malignancy (a process known as adoptive transfer) began to emerge in the late 1980s for the treatment of metastatic melanoma[236]. Significant improvements were achieved in 2002 with the addition of neoadjuvant chemotherapy to deplete the patient's immune system prior to adoptive transfer that achieved persistent clonal repopulation with antitumour T cells[237] and an objective durable response in over 50% of patients. In recent years additional adoptive cell strategies have emerged using other immune cells, such as natural killer cells or genetically re-engineered T lymphocytes to overcome previous limitations. One approach (TCR-T) modifies T-cell receptors (TCRs) to recognise tumour specific antigens prior to ex-vivo clonal expansion and adoptive transfer thereby enhancing T cell specificity. T cell cytotoxicity is normally mediated by MHC antigen presentation on the cell surface, and this pathway is frequently downregulated in cancer[238] preventing treatment response. To overcome this restriction chimeric antigen receptors or CAR T cells have been engineered that link the antigen-binding domain and intracellular T cell signalling domain. This enables CAR T cells to directly identify oncoproteins on the malignant cell surface and achieve T-cell activation in the absence of MHC signalling restrictions[239].

Despite impressive clinical results in particular with CD19+ B-cell malignancies and in the treatment of virus-associated solid tumours, a number of challenges remain. ACT has been associated with severe side effects, limited tumour infiltration and malignant cytotoxicity. Newly acquired somatic mutations may cause modification or loss of target antigen expression (antigen escape) and therapy resistance[240]. Somatic variant identification from transcriptome and whole-exome sequencing continues to inform research in this area and improve outcomes by identifying expressed mutations that may be targeted by ACT therapy[241,242,243] and in guiding clinical decision making regarding therapy selection and prognosis[244,130]. Tumour mutation burden (TMB) is a numeric estimate of the overall mutational burden in the tumour. It is considered to be a proxy for neoantigen burden and tumour immunogenicity and is the subject of ongoing research as a potential biomarker of ACT response. A recent study of ACT in anti-PD-1 naïve and experienced patients with metastatic melanoma found that higher TMB was associated with TIL ACT response[245] while a separate study in CD19+ CAR T cell treatment of chronic lymphocytic leukaemia found TMB was not predictive of clinical outcomes[246]. Further studies will be important to help clarify the potential role of somatic mutations and mutational burden in ACT.

Another novel class of immunotherapy drugs that have had significant impact on survival rates across a broad spectrum of cancers are immune checkpoint inhibitors (ICIs, or checkpoint blockade). The immune response against invading pathogens is regulated by a series of checkpoints that maintain immune homeostasis and prevent autoimmunity. As the response progresses, cytokines expression triggers the upregulation of surface protein CTLA4 in activated T cells which gradually rolls back the immune response by binding to CD80 and CD86 on antigen presenting cells, outcompeting CD28 and limiting T cell effector response[247]. Another checkpoint mediated by T cell PD1/PD2 surface proteins helps to maintain peripheral tolerance by triggering anergy or deletion of self-reactive T cells. The activating ligand, PD-L1/2 is usually expressed by tolerogenic antigen presenting cells and is found in a variety of normal tissues (including the heart, lung, placenta and thymus)[248].

However both of these pathways are subject to dysregulation and manipulation in cancer. The upregulation of CTLA4 has been observed in tumours[249] while the PD-L1 pathway may be hijacked by the malignancy resulting in ligand overexpression and suppression of the immune response.

Pivotal research in the 1990s by James P. Allison and his team[250,251] that uncovered the mechanism of action of CTLA4 and its potential as a candidate for therapeutic blockade would eventually result in FDA approval of the groundbreaking checkpoint inhibitor ipilimumab, a monoclonal antibody CTLA4 antagonist for the treatment of stage IV melanoma[252]. In parallel, research by Tasuku Honjo and his team identified PD-1[253], paving the way for antibody-mediated PD-1 and PD-L1 inhibitor therapies that have emerged in recent years[254]. Both of these therapies have resulted in a significant improvement in outcomes for many cancer patients and have helped elevate immunotherapy to the forefront of cancer research. Despite impressive response rates for some cancers (estimated at 50% for melanoma, including 20% complete responses[255]), a significant number of patients will regress or show no response and some will experience significant side effects. In this regard biomarkers play an important role in ICI therapy screening, including the identification of somatic mutations in key driver genes that confer treatment resistance (for example, loss-of-function mutation of *PTEN*[256] or activation of *PI3K-AKT*[257]). Perhaps the most widely recognized biomarker for ICI efficacy is TMB, as evidenced by the FDA-approved ICI companion diagnostic FoundationOne CDx[258]. TMB is an imperfect predictor of tumour immunogenicity and ICI efficacy. Current bioinformatic methods do not account for a mutation's neoantigen potential when estimating TMB. There is also a lack of consensus on requirements regarding sequencing target size, depth of coverage, bioinformatic pipelines and tumour purity of the test sample[259]. In this regard, TMB and other ICI biomarkers are likely to continue to be the focus of significant research effort given their importance to ICI patient outcomes.

A number of other promising immunotherapeutic strategies are emerging for the treatment of malignant disease. The Bacillus Calmette-Guérin (tuberculosis) vaccine has been standard of care for patients with bladder cancer since the 1970s[260,261], however subsequent progress in the field has been limited by the absence of suitable vaccine targets and delivery mechanisms. In recent years, the routine and widespread application of genomic sequencing technology enabling the identification of personalised tumour neoantigens, and significant advances in adoptive cell transfer technologies and nucleic acid-based vaccine platforms[262] have opened up new possibilities for the development of personalised and general therapeutic cancer vaccines. A number of different approaches are currently undergoing clinical trials including the adoptive cell transfer of dendritic cells primed with target tumour associated antigens that function as a vaccine carrier and a number of mRNA based cancer vaccines across a diverse range of cancer types[263]. The field continues to evolve and improved bioinformatic methods of identifying personalised tumour associated antigens capable of eliciting a potent anticancer immune response are the subject of ongoing research[264]. As in other areas of immunotherapy, accurate and reliable identification of somatic mutations continues to play a key role in this regard.

1.2.6 Epigenetic inhibitor therapy

The growing use of therapeutic agents such as DNA methylation or histone deacetylation inhibitors, often in conjunction with immunotherapy are showing promising results in cancer treatment. Epigenetic mechanisms such as methylation and acetylation play an important role in regulating transcription by controlling the structural properties of chromatin. Methylation involves the addition of one or more methyl groups to a histone protein or directly onto the DNA molecule to package (or condense) the surrounding DNA into nucleosomes preventing transcription. Conversely histone acetylation (the addition of an acetyl group to a histone lysine residue) leads to relaxation of chromatin, 'unwinding' the DNA and enabling transcription[265]. DNA methylation generally occurs throughout CpG genomic regions, punctuated by unmethylated sections known as CPG Island that usually extend for 300–3000 base pairs and often contain gene promoters[266]. Aberrant or 'hyper' methylation within CPG Islands, can lead to the silencing of tumour suppressor genes while hypomethylation of oncogenes or aberrant acetylation[267] is also an important factor in many malignancies.

Clinical studies involving drugs that target aberrant DNA methylation in cancer began in the late 1960s and early 1970s[268], however the role played by DNA methylation inhibition in these treatments did not become clear until 1980[269]. Azacytidine, an analog of cytidine that prevents methylation when incorporated into DNA was the first DNA methyltransferase inhibitor (DNMTi) to be used in cancer for the treatment of myelodysplastic syndromes and acute myeloid leukaemia. Additional DNMTi's including second generation drugs with improved metabolic and chemical stability have also been developed in recent years[270]. In addition to the reactivation of tumour suppressor genes azacytidine has also demonstrated the ability to sensitise tumour cells to T cell-mediated cytotoxicity possibly by the upregulation of neo-antigens that trigger an immune response[271,272]. Histone acetyltransferases (HATs) and deacetylases (HDACs) are other promising epigenetic targets for cancer therapy. HATs and HDACs regulate gene expression by adding or removing an acetyl group on a histone lysine residue and dysregulation of this process can lead to aberrant gene expression in certain malignancies. Research into HDAC inhibitors (HDACi) in cancer therapy dates back to the 1970s when n-butyrate, a potent inhibitor of mammalian histone deacetylase was used to reversibly transform HeLa and Friend erythroleukaemia cancer cell lines into morphologically normal cells[273]. A number of HDACi cancer treatments have been approved in the last decade including combinatorial approaches with radio, chemo and immunotherapy[274].

There are however a number of issues that limit the clinical application of non-selective DNMT and HDAC inhibitors. These medications affect all regions of the genome in both tumour and normal cells and are associated with toxicity and, with HDAC inhibitors in particular, serious side effects that require treatment termination[275]. In addition there are associated pharmacokinetic challenges. Additional research is required to expand the limited number of biomarkers available to predict a clinical response to these treatments as it may be difficult to weigh expected clinical benefit against potential toxicity. Mutations in *DNMT3A* which codes for the DNA methyltransferase 3 alpha, an enzyme responsible for *de novo* methylation patterns in embryogenesis and germ cell development and TET2, a critical regulator of DNA methylation, have been associated with improved response to DNMT inhibitors and progression free survival in myelodysplasia and related neoplasms[276]. Similarly mu-

tations in IDH1 and IDH2 (isocitrate dehydrogenase isozymes) resulting in the production of mutant protein 2-hydroxyglutarate (2HG) that inhibits histone and DNA demethylases leading to hypemethylation and tumourigenesis[277] are linked with improved DNMTi response in patients with acute myeloid leukaemia[278]. In relation to HDAC inhibition, a truncating mutation of *HDAC2* has been identified that confers HDACi treatment resistance in colorectal cancer cell lines[279] while HDACi have also been used to suppress tumour growth and promote apoptosis in *ARID1A* mutated ovarian and urothelial cancers[280,281]. Further research is required to identify other predictive biomarkers of epigenetic inhibitor therapies. In addition, a selective, small molecule inhibitor treatment that directly targets oncoproteins causing hypermethylation is now available. AG-221 or enasidenib selectively inhibits the mutated IDH2 enzyme, preventing accumulation of the oncometabolite 2HG and restoring demethylases activity[282]. In parallel, research into selective, isoform specific HDAC2 inhibitors is also yielding promising results[283] It is hoped that advances such as these will produce more effective inhibition of epigenetic dysregulation in cancer treatments with substantially reduced side effects. Accurate identification of treatment relevant somatic mutations will likely play an increasingly important role in stratifying patient care in this area.

1.2.7 Cytotoxic therapies

1.2.7.1 Chemotherapy

Cytotoxic therapy is a treatment that kills cancer cells directly (as opposed to targeted therapies that inhibit oncogenic proteins required for cancer cell replication or elicit an anti-cancer immune response). In cytotoxic chemotherapy (usually abbreviated to just 'chemotherapy') this is achieved by a chemical agent that damages DNA causing cell death or apoptosis. The cytotoxic properties of some chemicals have long been recognized after the exposure of soldiers to mustard gas and alkylating agents during both world wars. Subsequent research into the effects of these agents on high turnover cell types noted their therapeutic potential to suppress the division of cancer cells[284]. This work culminated with the groundbreaking NCI MOPP (nitrogen mustard, oncovin, procarbazine, prednisone) program in the late 1960s for patients with previously untreatable Hodgkin's disease[285]. The program, which completed its 50th year follow-up in 2013, resulted in an 80% complete remission rate of patients with advanced stage disease[286].

The primary mechanism of DNA damage used by MOPP is provided by the two alkylating agents, nitrogen mustard and procarbazine. An alkylating (or crosslinking) agent covalently bonds adjacent bases (usually guanine) on opposite DNA strands tightly together through a linker molecule, preventing DNA replication and transcription[287]. The repair process process is complicated by the fact that the lesion affects both strands and may include nucleotide / basic decision repair, mismatch repair and homologous recombination repair pathways. Due to specific vulnerabilities within the cell cycle and DNA repair deficiencies often present in cancer cells, the greatest effect of chemotherapy is seen in frequently dividing cancer cell populations. Although still used in chemotherapy regimes, alkylating agents such as nitrogen mustard and procarbazine have been augmented by second and third generation treatments that can be used to target different cancer types and reduce toxicity[288]. New therapies with novel mechanisms of action such as gemcitabine

(anticancer analog of deoxycytidine) have also been added. The incorporation of gemcitabine triphosphate on the end of the elongating DNA strand halts DNA polymerases inhibiting synthesis.

The effects of chemotherapy are systemic. As with radiotherapy the concept of therapeutic ratio applies where a chemotherapy regimen is selected to ensure toxicity is outweighed by the benefits of tumour control. A patient's response to chemotherapy and the toxicity they experience varies between individuals and across tumour types and stages of progression. The decision to administer chemotherapy, and the selection of a specific treatment protocol are not always clearcut. In this context, research has looked to molecular profiling to better inform clinical decision making regarding the use of chemotherapy. Somatic variants associated with chemotherapy sensitivity have been identified within DNA damage repair pathways, in particular in homologous recombination repair (HRR) and NER. HRR provides high-fidelity, template-dependent repair of double strand breaks during the S phase or G2 phase of cell cycle and is an important repair mechanism in frequently dividing cell populations, such as cancer cells. Tumour defects in HRR, such as those caused by somatic mutations leading to loss of function of the BRCA2 BRC[289] domain make the cancer cells particularly sensitive to damage caused by chemotherapy agents and may predict improved overall survival associated with this treatment[290]. Somatic mutations in *ERCC2*, a gene associated with NER have also been linked to improved response[291]. Perhaps the most significant development in this area to date is the ongoing TRACC (Tracking mutations in cell free DNA to predict Relapse in eArly Colorectal Cancer), a large scale, randomised control trial (scheduled for completion in 2026)[292]. Post-operative colorectal cancer patients are routinely offered adjuvant chemotherapy as it is unknown if the malignancy has been cured by surgery. Advances in the detection of somatic variants from circulating tumour DNA (ctDNA) allow screening for biomarkers of minimal residual disease to determine if chemotherapy is required. Many patients experience significant adverse effects (or toxicity) from chemotherapy[293]. There is also a risk of secondary cancers related to the treatment[294]. The success of the study would positively impact the quality of life of many patients and provide significant cost savings to the public health service.

1.2.7.2 Radioherapy

Radiation plays a key role in cancer management. It is estimated that half of all cancers including 40% of those cured of the disease receive radiotherapy during the course of their treatment[295]. Modern clinical radiation treatment began to emerge in the 1970s when Douglas and Fowler proposed a mechanistic approach to model the cytotoxic effects of radiation on tumour and normal tissue[296]. Radiotherapy is the use of ionising radiation to cause DNA damage and cell death in cancer cells. It usually takes the form of an X-rays beam from a linear accelerator directed into the body at the tumour. As the radiation traverses both tumour and healthy tissue it liberates electrons that damage and sever DNA molecules, halting cell division and causing cell death[39]. Similar to chemotherapy, the concept of a therapeutic ratio applies where the treatment is viable only if risk due to radiotoxicity (damage to normal tissue) is outweighed by the benefits of tumour control. In order to maximise the therapeutic ratio the total dose of radiation is broken up into a number of smaller doses or fractions with the angle of delivery rotated after each fraction, sparing normal tissue while still intersecting the tumour to inflict maximum damage.

Modern 3D computer imaging and conformal radiation therapy shape the beams to match the exact contour of the tumour from a given angle of delivery. The time between fractions is calculated to allow normal tissue to recover while maximising the number of tumour cells that are in the M phase of the cell cycle (the most radiosensitive phase) when the next fraction is delivered, thereby ensuring the greatest possible tumour cell kill[297].

Despite significant advances over the last 50 years, predicting patient specific radiobiological response in both tumour and normal tissue remains a significant challenge. Clinical best practice is based primarily on tumour site and histology[298] however the oncologist and radiotherapist will also exercise clinical judgement when planning each specific treatment schedule. Molecular determinants of tumour radiosensitivity may also indicate potential radiotoxicity if the variant is also present in the patient's gemline. Somatic mutations impacting DSB repair in *ATM* and members of the *MRN* complex have been identified as predictive of excellent radiotherapeutic response[299,300]. These variants, when present in the germline, are also associated with DNA repair disorders such as Ataxia-telangiectasia and Nijmegen Breakage Syndrome and indicate an increased chance of toxicity in response to radiotherapy[301]. On the other hand, loss of function somatic mutations in *NRF2*, *KEAP1*, *KRAS* and *P53* have been reported to confer resistance to radiotherapy[302,303]. Gene expression data has also been used in the prediction of tumour radiosensitivity. *Scott et al.* demonstrated a microarray based, gene expression classifier (the radiation-sensitivity index, RSI) to derive the personalised genomic-adjusted radiation dose for an individual patient tumour[304,305]. RSI has also been applied to RNA-seq data, including adjacent normal tissue to assess if the revised biological effective dose would increase normal tissue toxicity and adverse events[306]. Recent advancements in the detection of ctDNA from plasma samples have opened new opportunities for personalised radiotherapy. Liquid biopsy can potentially monitor patient response to radiotherapy in real time allowing the radiation dose to be adapted according to prognosis. The occurrence of treatment resistant biomarkers may be monitored and acted on without delay. Indicators of tumour radiosensitivity may be accounted for during treatment planning or by the selection of other complementary therapies (for example PARP or ICI) to exploit specific tumour radiological vulnerabilities. However more research, including large-scale randomised trials, are necessary before these biomarkers are integrated into routine clinical practice.

1.2.8 Therapy resistance

Resistance to therapy is by far the most common cause of death among patients receiving cancer treatment[307]. Some patients experience primary (or intrinsic) resistance, and do not respond to the initial therapy. Among patients who improve, acquired resistance, causing disease progression following initial positive response to therapy, is a significant issue. Cancer treatments place strong evolutionary selection pressures on surviving cells and often lead to the development of a subpopulation displaying a therapy resistant phenotype. As therapy progresses, the resistant phenotype ultimately dominates the tumour population and therapy fails. Advances in tumour sequencing have increased our understanding of the diverse molecular mechanisms governing resistance to targeted therapies and in some cases helped to develop strategies to overcome them. Mechanisms of cancer resistance to targeted therapy are further classified as on-target or off-target resistance. On-target resis-

tance occurs when the primary molecular drug target (ie, the targeted oncoprotein) sustains a somatic mutation that prevents therapeutic response, while off-target resistance develops by the activation of parallel signalling pathways or malignant pathway reactivation due to a somatic mutation downstream of the drug target. A detailed understanding of these factors is critical in restoring therapeutic response.

Research over the last two decades into acquired resistance to targeted therapies has observed that mechanisms of resistance often converge to reactivate the original pathway targeted by the drug, usually by a secondary mutation(s) at the drug target or downstream within the same pathway[308,309,310,311]. On-target resistance is typically caused by an acquired secondary mutation(s) in the target oncoprotein that triggers a loss of therapeutic response, for example, by reducing the relative binding affinity of a drug to its 'gatekeeper' residue and restoring mutant kinase activity[312,313,314]. Systematic molecular profiling of patient tumours is therefore crucial to detect the onset of therapy resistance and to adapt treatment strategy accordingly. A 'gatekeeper' mutation giving rise to amino acid change T790M is common in mutant *EGFR* non-small cell lung cancer and confers resistance to first and second generation TKI therapy[312]. However if T790M is detected, progression can potentially be arrested by changing therapy to third generation osimertinib which specifically binds to mutated forms of EGFR proteins, including T790M[315]. In prostate cancer therapy involving the anti-androgen medication bicalutamide, non detection of certain 'gatekeeper' somatic mutations may have serious implications. The somatic mutation W741C in the ligand-binding domain of the androgen receptor turns the nonsteroidal androgen antagonist drug bicalutamide, into an androgen receptor agonist which drives the cancer and accelerates disease progression[316]. Other mechanisms of on-target resistance identified in both anti-androgen[317] and protein kinase inhibitor therapy[318] include increased expression of the targeted oncoprotein.

As the disease progresses, the causes of therapy resistance become more complex[319] and may involve a range of mechanisms beyond individual mutations within the region targeted by a specific drug. Off-target resistance may be acquired through dysregulation of an alternative (or 'bypass') signalling pathway that is not affected by the original targeted therapy, allowing tumour growth to continue. Bypass mechanisms of resistance typically involve off-target somatic mutations, gene fusions or amplification in other pathways associated with cancer. In non small cell lung cancer therapy resistance may arise in response to amplification and overexpression of the MET growth receptor[320] leading to ligand independent dimerization, activation of the PI3K, STAT, mTOR pathway and disease progression[321]. Acquired mutations or amplification of other genes such as *HER2*, *KRAS*, *BRAF*, *PIK3CA* and cyclin-dependent kinases have also been implicated in TKI resistance[59]. Re-initiation of oncogenic signalling in the drug targeted pathway (EGFR) due to downstream reactivating somatic mutations in *BRAF*[322] or *PIK3CA*[323] has also been observed. Histologic transformation to the more aggressive form of small cell lung cancer, possibly mediated by epigenetic gene suppression[324] also leads to loss of therapeutic response. The efflux (ABC) transporter may also play an important role in multi drug resistance to some cancer therapies. The ABC transporter is responsible for pumping a broad range of compounds including anticancer drugs out of cells and is overexpressed particularly in cytotoxic chemotherapy resistance tumours[325,326]. Research into use of ABC transporter inhibition to sensitise tumours during chemotherapy is ongoing[327]. Other factors associated with cytotoxic treatment failure include

tumour induced hypoxia which plays a significant role in resistance to radiotherapy. Radiation induced DNA damage may be rendered permanent by reaction with molecular oxygen[328]. Hypoxic cells within oxygen deprived regions of the tumour mass are significantly more resistant to radiation damage and quickly repopulate the tumour after therapy[329].

Understanding the role of intratumour heterogeneity (ITH) in acquired resistance and its clinical implications remain a significant challenge in cancer research. ITH, the tumour microenvironment and cancer therapy provide the diversity and selection pressures necessary for a drug resistance population to evolve [REF]. Tumour heterogeneity has been associated with poor prognosis and outcome in cancer treatment[330,331,332], however a lack of consensus on quantifying ITH has prevented the translation of these methods into clinical practice[333,334,330]. Genomic intratumour heterogeneity may be characterised from somatic variant caller output. The cancer cell fraction associated with each variant is calculated from VAF, copy number and tumour purity estimates and cluster analysis is used to determine the number of tumour cell populations (or subclones) contained within the biopsy. In addition, morphological[335] and epigenetic features[336,337], and their interplay with the tumour microenvironment have also been used to characterise ITH. Different methods of sample collection (for example surgical resection or needle biopsy) may also influence results and it is uncertain how accurately one needle biopsy of a single lesion would represent overall tumour heterogeneity[338]. Quantification techniques involving liquid biopsy or imaging may record a more representative estimate of ITH[339,340]. Perhaps a consensus-based approach to quantification that captures the diverse feature classes involved will enable the future use of ITH as a prognostic indicator in clinical practice.

1.2.9 Implications beyond cancer

The primary focus of scientific research into somatic mutations has been in relation to human carcinogenesis. This is perhaps unsurprising given its significant and widespread impact on public health. The increasing application of genomic sequencing and somatic variant detection outside of cancer research however is raising awareness that cancer is not the only illness caused by somatic mutations. Somatic mutations that occur during cell proliferation and embryogenesis may give rise to two or more cell lineages with different genotypes; a process known as somatic mosaicism[341]. Occasionally these mutations may also cause phenotypic differences. In most cases there is little or no impact for the individual concerned. However, in rare instances somatic mosaicism may cause a range of physical, intellectual and neurological disorders. Somatic mosaicism can prove difficult to confirm as the mutation concerned only occurs in the affected tissue and may not be identifiable from a blood sample. This is particularly true of mosaicism within the brain, where obtaining a tissue sample for analysis may be rare or impossible. Despite these challenges however, a significant amount of research sequencing data has been collected over the last number of decades. A recent study of brain tissue from 105 neurosurgically treated patients with drug resistant temporal lobe epilepsy collected over a 30 year period showed gain of function somatic mutations in *RAS*, *RAF* and *MAPK* signalling pathways in tissue from the affected region[342] indicating a possible causal role in the disease also identifying potential therapeutic targets. Promising research into the detection of mosaic somatic mutations from cerebrospinal fluid derived

cell-free DNA and its application to research and diagnosis of non-malignant brain diseases[343] may provide a much sought after alternative to somatic variant detection from brain autopsy or biopsy.

The developmental timing of mosaic somatic mutations is of critical importance. Generally speaking the earlier in embryogenesis the pathogenic variant arises, the greater the extent of mosaicism and the more significant the impact to the individuals health. Alzheimer's disease shows increased prevalence in the elderly and the accumulation of somatic mutational burden in the ageing brain (a concept known as genosenium) is suspected to play a role[344]. However sporadic, early-onset disease has also been attributed to a brain related somatic mosaic in the PSEN1 gene[345]. The range of potentially harmful mosaic somatic mutations is more diverse than pathogenic variants found in the germline as mutations which would otherwise be lethal in utero may persist as mosaic. This gives rise to a number of disorders that only occur in mosaic form[346,347,348] (proteus, Sturge–Weber, or McCune–Albright]. In recent years the application of genomic sequencing technology to affected tissue has uncovered the mutations causing these diseases and continues to inform new treatment strategies. Overgrowth disorders such as Proteus syndrome and PIK3CA-Related Overgrowth Spectrum (PROS) have now been linked with somatic mosaicism relating to genes involved in the PI3K/AKT/mTOR pathway[349] leading to the repurposing of miransertib, a small molecule protein kinase inhibitor initially developed for cancer therapy to suppress the AKT pathway. A number of trials are underway and initial results are encouraging[350,351,352]. Similarly the immunosuppressant sirolimus which also targets the mTOR pathway is now being evaluated in the treatment of Sturge-Weber Syndrome[353] (a sporadic vascular malformation syndrome caused by a somatic mutation in the *GNAQ* gene[354]).

The role of somatic mutations in non-malignant disease is not only confined to mosaicism. Outside of cancer perhaps the most well-known illness driven by somatic mutation is Huntington's disease[355], an incurable, heritable neurodegenerative disorder. Although generally described as an autosomal dominant disease, onset of symptoms is caused by somatic mutations that accumulate during DNA replication. The Huntington gene usually contains a section of 26 or fewer cytosine-adenine-guanine (CAG) trinucleotide repeats. However, individuals who inherit one copy of a pathogenic allele (with approximately >40 trinucleotide repeats) will develop symptoms of Huntington's over the course of a natural lifespan. In individuals which exceed this critical repeat number, during each round of replication, the section of CAG repeats starts to form a stable hairpin structure during replication, resulting in polymerase slippage and progressive increase in the total number of the triplets due to somatic mutation, ultimately leading to mutant protein formation and disease onset. A recent GWAS Study of 4,082 affected individuals has identified six genes involved in DNA maintenance and other genetic modifier loci, apart from the uninterrupted CAG repeat length of the affected individual, that influence age of onset[356]. Further research in this field is ongoing with the ultimate aim of creating a therapy to delay or prevent onset.

Somatic mutations in frequently dividing cell populations are often of particular research interest. Mutations acquired by individual progenitors may give rise to non-malignant clonal expansion within a number of human tissues such as skin[357], oesophagus[358] liver[359] and colorectal epithelial tissue[360] often with strong positive selection of clones carrying mutations in genes normally associated with cancer. In

particular, clonal hematopoiesis, where the progeny of a small number of hematopoietic stem and progenitor cells are significantly overrepresented in the circulatory system is the subject of ongoing research. These hematopoietic clones may also contain somatic mutations in a number of recognized cancer driver genes[361] and the condition is frequently presented as a precursor state for haematological neoplasms[362]. However, recent evidence confirms that, although it does represent a significant increase in risk, the vast majority of individuals with clonal hematopoiesis will not progress to develop hematopoietic malignancies[363,364]. As such, the condition is often referred to as Clonal Hematopoiesis of Indeterminate Potential (CHIP). New research which has re-examined drivers associated with CHIP and their prevalence has given us a broader understanding of the overall clinical implications concerned. Cardiovascular diseases (CVDs) are the leading cause of death globally and place a significant economic burden on health care[365]. Clonal hematopoiesis has recently emerged as a major independent risk factor in CVD with a number of studies demonstrating a link between clonal hematopoiesis, in particular, involving somatic mutations in *PPM1D* and *TET2* (which may indirectly mediate a number of inflammatory responses) and increased risk of stroke and other cardiovascular events[366,367,368]. The research also identifies potential preventative measures involving cholesterol-lowering medications or targeting of specific inflammatory pathways and underlines the broader importance of accurate and reliable identification of somatic mutations in healthcare and scientific research.

1.3 Detecting somatic mutations

1.3.1 A brief history of DNA and sequencing

Although first isolated by Friedrich Miescher in 1869 little was known about the function of DNA until 1944 when Avery, MacLeod and McCarthy's research[369] into bacterial transformation determined that DNA encoded the genetic information governing 'the biochemical activities and specific characteristics of cells'. Less than a decade later (1953), building on the X-ray crystallography research of Franklin and Gosling[370], Watson and Crick, proposed their double-helix model of DNA[371]. This work formed the basis of subsequent research into mechanisms of DNA replication[372] that over the course of the following decade culminated in the identification of mRNA[373,374], its role in protein synthesis and regulation in bacteria[375] and the interpretation of genetic information stored in DNA as a degenerate, non overlapping triplet code involved in the production of protein[376]. By 1966 Nirenberg and his team[377] had deciphered the remaining RNA codons encoding all 20 amino acids. In spite of these advances however, the absence of a practical method for determining the order of nucleotides in DNA was still proving a significant barrier to progress.

In 1977 however, Frederick Sanger achieved a significant breakthrough that would help shape the potential of DNA research for decades to come. Using a low concentration of radiolabelled chain terminating (dideoxy) nucleotides in separate *in-vitro* DNA replication assays with each of the four bases, Sanger stopped synthesis in a small percentage of strands at each base. The truncated strands were then fractionated according to size via gel electrophoresis revealing order of the bases in the DNA molecule. The sequence was read directly off an X-ray film containing a mark left by the radiolabeled chain terminator at the end of each molecule. Sanger used this method to confirm the DNA sequence for the genome of bacteriophage ΦX174

and published his results in 1977[378]. Together with some refinements Sanger's chain termination method of DNA sequencing became the mainstay of genomic research culminating with the initial sequence of the human genome in 2001[379]. In 1983 Mullins and Smith developed the thermal cycling DNA amplification method Polymerase Chain Reaction (PCR)[380]. By 1986 further improvements to this method (in particular the use of temperature-resistant taq polymerase[381]) began to open up new possibilities in DNA sequencing and analysis. The technique was expanded by Higuchi 1992 who introduced fluorescent dyes in the reaction that allowed them to estimate the number of DNA molecules of the amplified sequence that were initially present in the sample[382]. Their work continues to play a significant role in molecular diagnostics and quantification.

Advances in technology such as the development of PCR helped lay the foundations for the NGS techniques that began to emerge over the course of the next decade[383,384,385]. NGS methods typically involve a library preparation stage, where the DNA is fragmented into a large number of smaller sections (inserts) each labelled with a short synthetic DNA adaptor (oligo) annealed to the 5' and 3' ends and denatured to form a template. The adaptor contains an identifiable sequence (or barcode) indicating the source and orientation of the original DNA strand. This technique enables large quantities of DNA templates from different sources to be amplified (via PCR), multiplexed together and sequenced simultaneously to reduce cost. Although the exact method of sequencing the order of bases varies, most NGS technologies employ a sequencing by synthesis approach where each base in the DNA fragment is read as the complementary strand is synthesised. The base type is deduced from light emitted by a fluorescently labelled nucleotide analogue (illumina), or a secondary reaction between an additional ATP sulfurylase (ATPS) enzyme and the pyrophosphate molecule released as a byproduct of the reaction that adds the base to the new strand (454). Cluster amplification of the template (using either bridge PCR or emulsion PCR) is required prior to sequencing to ensure that the intensity of the light generated can be detected by the sequencer's optical systems. Each template cluster is located at unique spots (illumina) or wells (454) on an NGS sequencer flow cell. As each new base is incorporated, DNA synthesis is paused using reversible dye-terminators (illumina) or by the stepwise addition of each nucleotide type (454) to allow the sequencer optics to record the base order of each cluster.

In the past decade third and fourth generation sequencing Technologies have begun to emerge. Single molecule sequencing systems such as PacBio SMRT and Oxford Nanopore are sensitive enough to detect individual nucleotides in a DNA strand without the requirement of cluster amplification, providing continuous read lengths an order of magnitude greater than previously possible. This feature has been crucial to the *de novo* assembly of difficult to map regions of the genome. Together with NGS these technologies are often collectively referred to as high throughput sequencing (HTS) systems. HTS has had a profound impact on the study of genomics and its clinical application. The dramatic reduction in sequencing costs coupled with greater sensitivity in somatic variant detection has brought about the routine use of HTS in clinical oncology[386]. Large scale cancer genomic sequencing programs like the Cancer genome Atlas continue to contribute to the development of targeted therapies and diagnostics across a broad range of cancer types.

1.3.2 Somatic variant calling from high-throughput sequencing data

The detection of somatic mutations by high-throughput sequencing has become a critical tool in scientific research and its use in clinical molecular pathology continues to increase. Unlike targeted variant detection methods that confirm the presence or absence of a specific pathogenic variant, somatic variant calling can simultaneously interrogate thousands of potentially harmful variants across all genomic regions of interest and continues to play a key role in scientific research, the evolution of new treatment strategies, drug development and increasingly, by informing clinical decision making in daily practice. In the past decade new techniques for isolating circulating normal and tumour DNA from blood plasma have been developed[189] using a procedure known as liquid biopsy and analysis methods that apply WEX somatic variant calling to the DNA recovered are starting to emerge[387]. However these assays require specific methods of DNA extraction[388], may be complicated by low levels of tumour DNA and may require specialised bioinformatics analysis[389]. The majority of approaches to somatic variant calling however consists of five basic steps: sample collection and storage, DNA extraction, library preparation, sequencing and somatic variant calling.

The first stage in any variant calling assay involves the removal of a small amount of tumour tissue usually by means of a needle or surgical biopsy. A blood sample is used in the case of myeloproliferative neoplasm. Typically a control sample of normal tissue (skin) or more commonly, a blood sample from the same patient is also required, although this will depend on the choice of downstream bioinformatic processing. After the sample has been collected it is generally stored as formalin-fixed paraffin embedded (FFPE) or fresh frozen (FF) prior to sequencing. FFPE is the most cost effective and commonly available sample preparation method for clinical testing[390] that preserves morphology and enables samples to be stored at room temperature almost indefinitely. Samples are first fixed with a formaldehyde solution to stop cell metabolism, and then paraffin is used to seal the tissue and reduce the rate of oxidation[391]. Unfortunately FFPE may also create significant DNA damage that usually manifests as a high burden of low frequency artefacts in sequencer output[162]. If the resolution of low frequency somatic variants is a priority then fresh frozen (FF) sample storage is generally the preferred option. FF storage requires that the sample is frozen in liquid nitrogen 30-60 minutes after surgery[392] and kept frozen thereafter as once it starts to thaw the DNA or RNA starts to degrade[393]. Although FF does not preserve morphology, DNA/RNA is preserved better than FFPE and is less susceptible to artefacts introduced during pre-analytical processing (the sample storage and DNA library preparation stages). However the overheads involved in storing a frozen sample are obviously much more significant in comparison.

A biological sample removed from storage for sequencing is first treated with reagents to dissolve the cell membrane and release nucleic acids. Organic solvents degrade other cellular proteins and debris which are then removed by centrifugation. RNase (ribonuclease) treatment is performed as required for the removal of unwanted RNA after which the DNA is precipitated and purified[394]. A similar process is followed if extracting RNA analyte after which it is converted to cDNA using a reverse transcriptase enzyme for downstream analysis. Short read sequencing technologies (typically Illumina) can not effectively form clusters with long DNA molecules (typically >1 kb)[395]. This necessitates the creation of a DNA library prior to sequencing

by breaking the DNA up into a series of smaller fragments typically between 200 and 500 bp in length. Fragmentation is performed via acoustic shearing or enzymatic reaction and is followed by DNA end repair and the ligation of adaptor sequences to fragment ends for identification during DNA multiplexing and flow cell binding etc. If required, target enrichment is performed by in-solution hybridization and removal of the desired genomic DNA using magnetic probes containing sequences that hybridise within target regions. Alternatively amplicon based techniques that use primers to amplify specific regions of interest prior to fragmentation and adaptor ligation may also be used. In preparation for downstream sequencing the library contents may be PCR amplified if required, size select is performed (where fragments outside the required length range are removed) and the final percentage DNA content validated[396].

For short read technology (such as illumina), sequencing is performed by denaturing the DNA library and loading it onto a flow cell where adapter sequences on either end of the templates hybridise to oligonucleotides on the cell surface forming a bridge structure. Next, a polymerase synthesises the reverse strand after which both strands release from one end and straighten resulting in the creation of a forward and reverse clone. The process (termed 'bridge amplification' by illumina) is repeated resulting in a cluster of forward and reverse strand clones of the original templates grouped beside each other on the flow cell surface. Once bridge amplification is complete and the clusters are formed, all reverse strands are washed off the flow cell, leaving clusters of forward strands only. The reverse strands are then re-synthesised (up to at least one read length which is typically about 50 to 150bp) by stepwise elongation of the template primer using fluorescently labelled reversible terminators[397]. The fluorescence signal from each cluster is recorded by high resolution sequencer optics and converted to DNA base calls by onboard analysis. If paired-end sequencing is required, after reading the forward DNA template strands, the reads are washed away, and the process repeats for the reverse strand, generating a second 'paired' read from the opposite end of the original template.

Sequence data recorded during HTS serves as input to bioinformatics analysis pipelines that attempt to recover somatic variants of interest. DNA read data stored as unaligned nucleotide sequences, usually in binary base call (BCL) format by the sequencer is converted to FASTQ for data pre-processing. After passing an initial quality control check (FASTQC[398]) the raw FASTQ files are used as input to the alignment stage. During alignment the short read sequences in the FASTQ files are aligned against a reference genome using BWA-MEM[399] or any of a number of sequence alignment tools to create a Sequence Alignment Map[400] (SAM), usually stored in compressed binary (BAM) format. SAM records containing read duplicates, as a result of PCR amplification events or optical duplications (where a single cluster has falsely been called as two separate clusters by the base caller software) that occurred during the sequencing process are marked (using GATK[401] or SAMTOOLS[400]) to exclude them from downstream variant calling analysis. Depending on the type of somatic variant caller used, an additional alignment step (for example using GATK[401]) may be employed at this stage which locally realigns reads detected near indels to reduce alignment artefacts (this is not usually required for haplotype based caller algorithms). In the final preprocessing step base quality score recalibration[402] (BQSR) may be applied to empirically adjust quality scores in the sequence alignment file. This removes the effect of sequence dependent base call and

systematic technical errors (associated with lane, tile and machine cycle etc) from base quality estimates in the alignment. The data is now suitable for use as input to somatic variant calling.

A somatic variant caller is a bioinformatics software application that examines the aligned contents of covering reads (ie., the pile-up) at each locus in a sequence alignment file(s) to check for evidence of putative somatic mutations. Somatic variant callers are generally classified as position-based or haplotype-based. Position-based callers (or 'pile-up' callers) only check for evidence of a somatic variant at the target locus under consideration, while haplotype-based callers use adjacent variation to phase the sequencing data (split reads into groups supporting different haplotypes) thereby improving the accuracy of the call. For example Mutect2[403] the GATK haplotype-based caller applies a two pass method to variant calling. The first pass applies a position-based approach to identify regions of interest while the second proceeds with local reassembly and realignment of reads traversing those regions before calling variants. Somatic variant calling methods are generally classified as tumour only or matched tumour normal depending on whether or not they require a control sample of normal DNA from the patient (sequenced on the same run as the tumour) to exclude germline and other alignment or sequencing artefacts. Most variant callers designed to work with WEX or WGS data usually require a matched normal while tumour only methods are typically associated with clinical gene panel analysis pipelines where only the tumour sample is sequenced to reduce cost and turn around time[386]. A range of open source somatic variant callers[404,405,406,407] are used in various clinical or research applications and many publications have assessed the advantages and disadvantages associated with various approaches[408,409,410,411].

1.3.3 Artefacts in somatic variant calling

Variants identified in somatic variant caller output are not always true somatic mutations. Along with somatic mutations from the tumour sample there is also technical variation, introduced by non-biological events during sequencing and data processing. A diverse range of computational approaches that target specific aspects of technical variation have been devised to remove (referred to as variant filtering) artefacts in somatic variant calling. Collectively these algorithms form part of an essential step in somatic variant calling known as variant filtering. Perhaps the most common artefacts (and best example of filtering) in somatic variant calling are germline variants. In somatic variant calling germline variants are considered to be artefacts of the sequencing process and are typically excluded from analysis means of a control sample of normal DNA from the patient. Putative somatic variants that are common to both the tumour and matched normal samples are filtered as potential germline or other artefact of the sequencing process (for example due to alignment issues). In general there are two approaches to removing putative variants from analysis. Soft filtering annotates the filter field of the Variant Call Format (VCF) record (field 7) with the reason why the variant was filtered while hard filtered variants are excluded from the file. Soft filtering affords the researcher the opportunity to reevaluate putative variants that have been identified as artefacts. However it may also increase the size of the VCF and caller runtime. The Mutect2[403] somatic variant caller for example, which uses a normal control, ensures putative variants that are clearly present in the matched normal are removed at an early stage and not recorded in the VCF to avoid spending computational resources on germline arte-

facts. In borderline cases Mutect2 will record and filter the variant in the VCF in case further curation may be required. Tumour-only calling pipelines (which do not use a normal control) generally employ published databases of known germline polymorphisms, for example Single Nucleotide Polymorphism database (dbSNP)[412], the 1000 Genomes Project[413] and Exome Aggregation Consortium (ExAC)[414], sometimes in conjunction with computational modelling to predict germline status[415]. These approaches however may lack the ability to exclude other artefacts of the sequencing process that are highlighted by the normal control (for example alignment artefacts).

Preanalytical processing and library preparation can have a critical impact on the quality and integrity of DNA and somatic mutations recovered from it. FFPE treatment in particular may have significant implications for downstream somatic mutation analysis. Formaldehyde, (an active component of FFPE) generates crosslinks between nucleic acids that block PCR[416] and causes the removal by hydrolysis, of purine bases[417] from DNA, leaving an abasic site that weakens the strand and reduces insert size during sequencing library preparation[162]. This in turn causes a reduction in coverage or uniformity of coverage as less DNA fragments make it through size selection during library preparation. Attempts to compensate for this by selecting for a shorter fragment length may cause alignment issues in downstream bioinformatic analysis. FFPE can also cause direct alterations in DNA sequence likely due to the deamination of cytosine to uracil which in turn pairs with adenine during replication leading to C>T transitions. A number of chemical approaches are available that attempt to limit damage or restore DNA in FFPE. Prolonged formalin fixation during FFPE sample preparation increases DNA damage[418,419] and should be limited. Some methods have reported success in reversing crosslinks in FFPE samples[420] however they still remain a common issue. Pretreatment with uracil DNA glycosylase has proven successful in degrading DNA molecules containing uracil prior to sequencing library preparation and reducing FFPE artefacts[421,422].

Bioinformatic methods such as GATK[403] or illumina[423] orientation bias filters are also effective at removing FFPE artefacts from variant caller output. In paired end sequencing, read 1 is taken from the 5' end of the original fragment and read 2 the 3' end (i.e, the 5' of the reverse strand) . The order in which both reads in the pair align to the reference genome enables us to infer the orientation of the original DNA fragment (i.e, whether it originated from the forward or reverse genomic strand). DNA alterations induced by FFPE are likely to occur on one genomic strand only. For example the deamination of cytosine to uracil will, in general, not change the guanine it was paired with on the opposite strand. A number of bioinformatic filtering algorithms exploit this asymmetry to check if the evidence of a putative mutation is biassed to a particular orientation, indicating that it is likely to be an artefact. This method is also effective at removing oxidative damage (another form of orientation bias artefact) that may occur during library preparation, in particular from shearing of DNA, causing G>T transversions[424]. Although these filtering algorithms provide additional means of guarding against orientation bias, no method, chemical or computational, is 100% successful at preventing false positive mutation calls as a result of FFPE or oxidative damaged DNA, particularly at low allelic frequencies. Depending on assay requirements and the amount present, such damage may still prove a significant confounding factor in somatic mutation analysis.

A wide range of other bioinformatic filters are used to prevent false positive

mutation calls arising from sequencing artefacts. During the sequencing process, in some instances, reversible terminators may not be correctly removed from DNA fragments before the start of the next step of the PCR cycle, leaving some out of sync with the rest of the cluster, creating a problem known as phasing[425]. This dilutes the optical signal making it harder for the base caller to discern which base that was added. The base caller uses the quality of the signal detected to calculate the probability that a given base is called incorrectly and this information, known as a Phred quality score or Q score, is included in the sequencer output. Somatic variant callers use this information to filter artefacts caused by poor quality base calls. Another common metric used to filter artefacts is mapping quality. Mapping quality is an attribute associated with a read (field 5 in the SAM record). It is a measure of confidence that the primary alignment listed for the read is correct (i.e. specifies its true location in the genome). It is calculated from secondary alignment scores associated with the read and indicates if the read mapping is ambiguous[399]. Typically, the median mapping quality of reads that provide evidence of a putative somatic variant is filtered to exclude possible alignment artefacts. A further example of an implementation of artefact filtering is a panel of normals. A panel of normals is a list of variation (usually in VCF format) derived from a cohort of healthy individuals using the same library preparation and sequencing workflow that was used for the case samples[403,423]. It is considered to represent common sequencing, germline and alignment artefacts unrelated to the condition (for example cancer) that should be excluded from analysis. A number of other filters are employed in somatic variant analysis that are beyond the scope of this manuscript. Most play an important role in reducing false positives in variant caller output. Despite this however, artefacts may occasionally be misidentified as somatic variants in caller output and continue to pose significant issues in somatic mutation detection[426,163]. Artefact filters may also contribute to type II errors and impact sensitivity[163]. Quantifying the extent of these issues is not always straightforward.

1.3.4 Somatic variant calling accuracy and limit-of-detection

Reliable and accurate detection of somatic mutations is a key requirement in research and clinical application. The choice of sequencing strategy and variant calling pipeline can have a significant effect on detection sensitivity that may lead to discordance in results across studies[427,428] and failure to detect clinically actionable variants in oncology[429,430]. There are also cost and patient implications, if for example a biopsy needs to be resequenced as the original depth of coverage was inadequate. There are a number of considerations when assessing a proposed mutation analysis pipeline in line with analytical requirements and an estimation of the assay's Limit-of-Detection (LoD) plays an important role in this regard. Limit of detection may be defined as the lowest allele frequency which would result in the variant being detected in 95% or more of the samples where it is present[431]. Sensitivity and limit of detection are generally modelled using the binomial distribution[431] with size (number of trials) taken as the average depth of coverage, and probability, the alternate allele frequency. Given the minimum number of reads containing the alternative allele that the calling pipeline requires to detect (annotate as PASS in the VCF) a variant it is possible to express sensitivity (ie., the probability a variant with a true allele frequency f, gets detected) as,

```
sensitivity = 1-pbinom(q    = Alt.allele_depth_calling_threshold,
                       size = depth_of_coverage,
                       prob = f)
```

(where the pbinom term lists the probability that a true somatic variant will not be covered by the required number of reads to be detected by the caller). The limit of detection is therefore the value of allele frequency that yields a sensitivity of 0.95 (Figure 1.3). The minimum number of reads required to detect a variant may be found from variant caller documentation or estimated using a value for caller specificity and the average base error rate of the sequencing run.

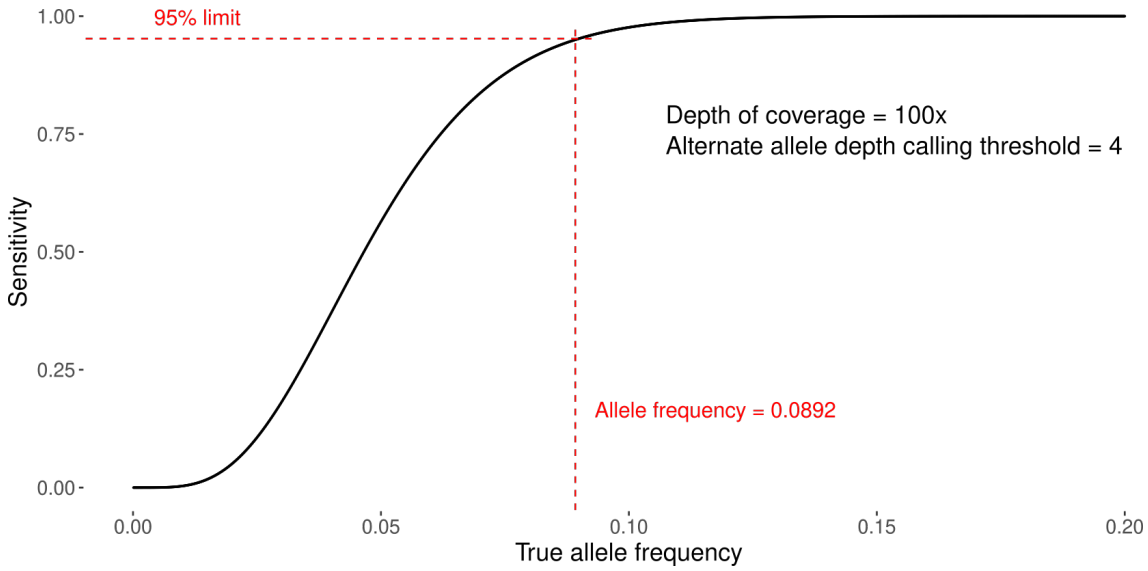

Figure 1.3: Theoretical somatic variant detection sensitivity plot for a variant calling pipeline implementing a depth calling threshold of four reads containing the alternative allele with 100x sequencing data. The theoretical Limit-of-Detection (LoD) of this system is indicated as 0.0892 allele frequency.

This binomial LoD model contains a number of simplifications that limit its application however. The model does not account for base quality in assessing sensitivity. Many variant callers scale evidence of the mutant allele at the pile-up with the quality scores of bases that support it[403]. This implies that the assumption there is a minimum number of reads containing the alternative allele required to call a variant is incorrect. Variant filtering can also impact caller sensitivity. Putative variants which demonstrate evidence of an alternative allele at the pile-up are routinely removed from analysis if they fail to pass a number of filters, for example median mapping quality and panel of normals among others. This is also not accounted for by the binomial model used to predict sensitivity. Tumour purity and high target GC content may also negatively impact somatic variant caller sensitivity. Low tumour purity decreases the allele frequency at which important clonal mutations will appear in the spectrum, in some instances causing them to fall below the minimum detection limit of the system. PCR amplification of sequencing libraries may also lead to coverage bias in regions with extreme GC and AT content resulting in fragments from these regions being underrepresented in sequencing results[432]. The resulting decrease in coverage can confound efforts to predict the

sensitivity of somatic variant detection within these regions that use overall sample estimates for depth of coverage. PCR-free library preparation protocols may be used to address issues associated with GC and AT rich library content[433,434] however in many instances the initial DNA content from the sample will be insufficient for this approach. Coverage biases associated with target capture may also lead to sensitivity issues[427].

The diverse range of computational and statistical approaches employed by somatic variant calling tools can be challenging to assess. The absence of an incontrovertible mutation set from real tumour data means that some differences in output between callers are never fully resolved and although a number of callers have proven effective in a diverse set of clinical and research applications no single caller has emerged as the industry standard. In this environment a consensus approach to somatic variant calling forms the basis for a number of studies aimed at improving accuracy[435,436]. Essentially this approach accepts that while each individual variant caller has its own strengths and weaknesses, there is no single best algorithm to identify somatic variants, and a mutation set based on the consensus of several proven and trusted somatic variant callers provides the most accurate representation of tumour somatic variation. Consensus methods typically employ an intersection threshold on the minimum number of callers that must detect a putative somatic variant before it is passed by the ensemble. Perhaps the most comprehensive implementation of consensus-based variant calling has been the Multi-Center Mutation Calling in Multiple Cancers (MC3)[437] working group which comprises a consortium of researchers across multiple institutions, to enable pancancer analysis of data from more than 10K patients across multiple different sequencing centres as part of the TCGA project. MC3 employs an ensemble of 7 variant calling methods[404,438,439,406,440,441,442] and includes a robust set of additional filters to implement a meta caller approach where consensus is required among variant callers and also between sequencing centres to remove batch effects and reduce false positives.

1.3.5 Validating somatic variant caller output

Several aspects of the mutation calling process significantly affect variant caller output. Sequencing metrics such as depth[443] and targeted region[444] lead to assay discordance, as do aspects of sample preparation such as tumour purity and the option of a matched normal. The selection of bioinformatics pipelines and associated somatic variant caller software also has significant impact. The objective of validation is to demonstrate that a given somatic variant calling assay, in research or clinical application, is suitable for its intended purpose. Quantifying the effect technical aspects of the calling process have on the somatic landscape recovered from variant calling requires sequencing reference data containing a 'ground truth' set of somatic mutations; data in which the location and details of all non-reference sites (a genomic location containing a base or bases that do not match the reference genome) is know. Ground truth data (or truth set) is also a key component in benchmarking somatic caller pipelines and its integrity underlies all conclusions drawn from this research. In general there are two types of ground truth data that may be used for validation. These are actual, well characterised sequencing data from a real patient tumour or cancer cell line, or synthetic data that has been created by mixing samples or sequencing data from different sources at various concentrations to create virtual somatic variants. A synthetic truth set may also be created by modifying

the contents of a BAM file to create 'somatic' variants. This is achieved using a software application to edit a subset of reads from each target pileup to include the required alternate allele (a process known as spiking-in variants). Each method has its own advantages and limitations and both are used extensively in mutation calling pipeline validation.

A common approach to validating somatic mutation detection pipelines with real sequencing data is to use consensus-based calling to derive the somatic variant truth set. Similar to an ensemble variant calling approach, a somatic variant is added to the truth set if it is detected by at least a minimum number of callers in the ensemble. When completed this consensus truth set is usually confirmed by orthogonal methods such as high depth or PCR free libraries[445], qPCR or Sanger sequencing[446,447]. Truth sets derived from well characterised, matched, tumour normal cell lines that yield high DNA content for sequencing are often preferred. Sequencing data derived from these methods are accompanied by high-confidence somatic variant calls that can easily be used to validate other mutation calling pipelines. However there are limitations. It may not be possible to unambiguously resolve all variant calls in the data, despite extensive orthogonal validation. The truth set is also biassed towards the consensus of the variant callers selected for the ensemble. For example, it is possible that a more recent somatic variant caller release will score worse on a consensus truth set that was compiled using an older software version of the same caller. Despite this, consensus truth sets play an important role in validation and in benchmarking variant callers against each other. A variant call that differs from the consensus should be investigated and used to guide further developments in the field.

Creating a truth set by combining DNA containing germline variants from different cell line sources is another frequently employed method of validating mutation calling pipelines. Sequencing data from different sources is mixed either *in-silico* (by computationally combining records from BAM files sequenced from different cell lines) or *in-vitro* by creating a titration dataset (generated using a mixture of DNA from different cell lines which is then sequenced)[448]. In this process, a selection of loci containing homozygous germline alleles in one cell line that are not present in the other are typically labelled as somatic. The allele frequency of these pseudo-somatic variants in the mixture is regulated by downsampling one of the BAM files before combining the data into a 'tumour' BAM, or by adjusting the levels of titration between the two DNA sources prior to sequencing when using an *in-vitro* approach. It is recommended to employ well characterised cell lines for this purpose[431]. The Genome in a Bottle Consortium (GIAB)[449] line NA12878, originally generated for the CEPH/HapMap project, is frequently selected (often in conjunction with NA12877) due to the availability of a comprehensive set of phased, high-confidence variant calls. Prior to creating the validation truth set, the presence in the cell line of germline alleles, designated for use as pseudo-somatic variants, should be confirmed by the application of a germline variant caller. Cell line add-mixture methods offer a practical means of validating somatic variant caller pipelines without the need for dedicated simulation software. The generation of a titration truth set also enables the comprehensive validation of the entire pipeline, from flow cell through to the list of mutations in the VCF output. However these methods also have limitations. The use of germline alleles to mimic somatic variants is a particular concern, especially considering that somatic variant callers often employ filtering algorithms to remove

sites that may be germline in origin. Disabling these filters (for example, removing the panel-of-normals (PON) prior to validation[450]) may also have other unintended impacts on caller specificity. In addition, the occurrence of real somatic variants are not restricted solely to well defined loci containing germline alleles. Cell line genotypes diverge[451,452]. They also contain a distribution of somatic variants, sequence and alignment errors at unknown locations. This ambiguity may lead to conflicting validation results. Haplotype-based somatic variant callers locally realign reads around a putative somatic mutation into phased groups to increase the accuracy of the call. Consequently, it is important to account for haplotype structure (ie., to phase sequencing data prior to combining the truth set) when creating simulated tumour sequencing data. This may prove difficult or impossible with data from cell line admixture.

An alternative method for generating a synthetic truth set involves the use of specialised software to manipulate the content of sequencing reads within a BAM file in a process known as spiking-in variants. This approach, most recently pioneered by the ICGC-TCGA DREAM Somatic Mutation Calling Challenge[408] addresses many of the shortcomings of admixture simulation methods. BAMSurgeon, ICGA-TCGA's tool used to spike-in somatic variants, provides greater control over genomic location, allele type and depth than can be achieved with *in-silico* mixing of reads from different sources. The researcher populates a config file, containing the locus, alternative allele and allele frequency and BAMSurgeon ensures the pileup at each target locus is modified accordingly, creating a virtual tumour BAM. Loci that do not have coverage to support the required allele depth are recorded and skipped. BAMSurgeon is typically used to insert variants in a BAM file created from a normal human sample. By default, BAMSurgeon also ensures that no variants are spiked-in to regions adjacent to pre-existing SNPs to avoid interfering with haplotype structure. The software tool can also be employed with entirely synthetic BAM files generated by a read simulator like ART[453], although, in this scenario, reads simulated from a haploid reference will usually not represent the diversity of variation present within an actual tumour sample. Similar to cell line admixtures however, a truth set created by spiking variants into real sequencing data will inherently contain a distribution of somatic variants, sequence and alignment errors, the precise locations of which are unknown.

1.3.6 Relative and absolute, PCR-based quantification

Starting with developments in the mid-1980s[380] to enable researchers to amplify a specific DNA template (a target genomic region of interest), the application of Polymerase Chain Reaction (PCR) has progressively evolved to become an indispensable tool in the field of molecular diagnostics. In its basic application PCR involves the combination of a sample of purified, denatured DNA with primers containing complementary sequences that anneal to the 5' and 3' boundaries of the target template and provide a starting point for DNA synthesis. The addition of a polymerase and DNA nucleotides results in synthesis of the template's complementary strand and thermocycling (the repeated heating and cooling of the reaction mixture) repeats the process producing billions of copies of the original DNA segment for further analysis. Refinements to the process made in the early 1990s resulted in the development of the first qualitative PCR method[382]. This research demonstrated the potential of employing PCR with fluorescent primers to successfully detect specific

alleles of human beta globin and Y chromosome-specific sequences and also described its potential application in quantitative analysis. The availability of protein engineered chimeric polymerases in 2003[454] led to significant advancements in PCR technologies. The DNA-binding domain of Sso7 which naturally functions in chromatin remodelling was fused to the low-processive, proofreading polymerase domain of *Pyrococcus* resulting in a high-fidelity, high processivity enzyme ideally suited to quantitative analysis methods (Pfu polymerase). Quantitative real-time PCR or qPCR involves the addition of special marker molecules to primers that increase in fluorescence with each positive PCR template amplification[455,456,457]. The intensity of the fluorescent signal is measured in real-time with each round of PCR amplification until it ultimately reaches a plateau phase at the end of the reaction. The quantification cycle (Cq, also referred to as the cycle threshold Ct) is defined as the number of cycles needed for the fluorescence intensity to exceed a detectable threshold. This value is directly proportional to the initial quantity of template DNA molecules present in the sample. An estimation of this quantity can be derived by referencing a calibration curve constructed from a series of standard dilutions, each with known concentrations or copy numbers. A combination of reverse transcription, which generates complementary DNA (cDNA), and qPCR may also be employed for the detection of RNA.

Since its inception towards the end of the last century real-time PCR/qPCR has become established as the preferred method for rapid and sensitive quantitation of nucleic acid in numerous clinical and research applications. However qPCR has its limitations. The accuracy of relative quantification (qPCR) relies on the quality of the standard curve from which the results are extrapolated. If the amplification efficiency within the sample undergoing analysis significantly deviates from that of the reference samples, the reliability of the results will be impacted[458]. Some applications also require a higher level of sensitivity than can be consistently achieved through qPCR. This shift in analytical demands has prompted a resurgence in an alternative quantification methodology, also conceived in the early 1990s[459]. Digital PCR (dPCR) is a method for the absolute quantification of nucleic acid concentrations based on end-point PCR, where the reaction continues until it reaches a plateau characterised by a robust fluorescent signal (indicating the presence of the target molecule), or absence of signal (indicating the target was not detected). Quantification is achieved by first isolating individual wild type and target DNA molecules in the sample into individual compartments prior to PCR (using cylindrical microchambers, or emulsion-assisted microdroplets) in a procedure known as partitioning. The relative concentration of wild type and mutant sequences is estimated from the Poisson distribution by counting the number of positive and negative fluorescence signals across all partitions upon PCR completion[460].

Unlike qPCR, dPCR does not rely on standard curves for quantification and provides precise and highly sensitive quantification of nucleic acids. The exceptional level of accuracy exhibited by dPCR makes it particularly suitable for use in the analysis of liquid biopsy to track minute variations in the levels of target DNA against a complex background of other circulating nucleic acids, particularly over the course of a treatment regimen[461]. In gene expression analysis dPCR has been shown to outperform qPCR in terms of precision and reproducibility, particularly when analysing low abundant targets[462] while somatic Copy Number Alterations[463,464] and CNV[465] have also been quantified using this method. Although quantitative PCR

methods do not have the capacity to simultaneously interrogate somatic munitions across a wide target region, qPCR and dPCR nevertheless play an important role in this regard by providing highly sensitive methods for orthogonal validation of NGS based somatic variant analysis[466]. The field continues to evolve with emerging technologies in multiplex real-time PCR such as microfluidics-based methods for detection and variant discrimination of SARS-CoV-2[467] and new applications of dPCR in Single-Cell Analysis[468] and environmental microbiology[469]. It is likely PCR techniques will continue to play a significant role in targeted quantification for many years to come.

1.3.7 Ethical and data protection obligations linked to genomic data

The volume of research data generated by clinical trials and population-based observational studies has experienced exponential growth in recent years. It has emerged as an invaluable research asset that may be subject to reanalysis by multiple research teams worldwide over periods often spanning decades. Given its transformative impact on research, its role in innovating treatment development, and its potential to generate significant revenue, the question frequently arises regarding the ownership of a patient-derived dataset. Perhaps more importantly from a regulatory perspective, there is also the question of responsibility for it. Research involving human subjects is generally reviewed by an ethics committee associated with the relevant university or institution to ensure that the appropriate ethical standards and legal requirements are being upheld. The regulations governing the data protection rights of participants in clinical research vary significantly across different regions of the world. In the U.S., the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[470] stands as the primary federal law safeguarding protected health (including genetic) information. It establishes limits and conditions for the use and disclosure of such information without an individual's consent. In the EU, the protection of health data is governed by the General Data Protection Regulation (GDPR)[471], while the UK operates a specific data protection framework (UK GDPR) tailored for the UK context[472]. While there is substantial overlap between these standards, they are not entirely equivalent. GDPR extends its jurisdiction to all international organisations processing personal data of individuals in the EU. This presents an important consideration, as compliance with GDPR may be necessary, even if the organisation concerned is not located within the EU.

The initial step in evaluating GDPR implications associated with a research dataset is to determine whether the data allows for the identification of individuals participating in the study (commonly referred to as 'data subjects'). GDPR explicitly states that the principles of data protection do not apply to information that has been fully anonymized in a manner that ensures the data subject is no longer identifiable[473]. However, achieving complete anonymization of genetic data is not always straightforward. Studies have shown that using fewer than 100 single nucleotide polymorphisms (SNPs) can be sufficient to distinguish an individual's genetic record[474]. Furthermore, GDPR expands its definition of genetic data to include somatic variation, categorising genetic information as encompassing both inherited and acquired genetic characteristics that provide information about an individual's physiology or health[475]. Somatic mutation data may also include variant records filtered as germline, potentially aiding in the re-identification of the individual concerned. For these reasons, somatic mutation data generally falls within the

scope of GDPR.

A key requirement of most regulatory frameworks for human research involves the process of obtaining informed, explicit consent from each individual participating in the study. Participants should receive comprehensive information regarding any personal risks and benefits associated with their participation. They should be informed about who is collecting the data, the purpose behind data collection, and how it will be used. When applicable, agreement on the management of incidental or pedigree-sensitive findings should be documented, and access to genetic counselling provided. GDPR specifies that this consent be explicit (i.e., in writing or by equivalent electronic means) and that the data subject has the right to withdraw their consent at any time and have their data deleted[476]. After data collection, the institution conducting the study typically takes on the role of the data controller. The data controller is responsible for providing all or a subset of the data to third parties (such as institutions, research groups, etc.) referred to as data processors, who analyse the data and conduct research. The data is shared under the terms of a data processing agreement (or data use agreement), which serves as a contract between the data controller and processor. This agreement typically outlines conditions such as the permissible types of research for which the data may be used and provides detailed requirements regarding the secure storage and handling of the data. Data protection legislation primarily places responsibilities on data controllers, holding them accountable for compliance and exposing them to substantial fines in case of non-compliance. Consequently, data processing agreements are comprehensive and binding contracts that delineate operational procedures and responsibilities for data processors. These agreements also outline the specific actions that data processors are expected to undertake in the event of suspected breaches of protocol. Importantly, these agreements also impose significant penalties on data processors in case of failure to adhere to the stipulated terms.

Ethical and data protection considerations associated with somatic mutation data typically involve a balance between the imperative to inform treatment or scientific research and the obligation to safeguard sensitive, patient-specific clinical data. Research ethics and data protection processes continue to evolve as regulatory bodies strive to streamline the process while simultaneously safeguarding the rights of individuals. However, ensuring compliance with GDPR results in an increased workload[477]. Significant effort may be expended to ensure that the network hosting the research data is secure, access to the data is controlled, and data sharing is fully compliant. Careful consideration of data processing agreements is essential, as misinterpretations could render detailed scientific research ineligible for publication if the stipulated conditions on data usage have been breached. Indeed, an inherent conflict between critical care research and the obligations of data protection has been observed[478]. In this context it is important to highlight the potential for severe patient consequences if data protection and ethical practices are not adhered to.

In 1996 building on publicly available research[73,72,479] and genetic data from patients under their care Myriad Genetics patented the sequence and detection of *BRCA1* and *BRCA2* genes[78] for use in the diagnosis of predisposition to breast and ovarian cancer. The granting of this patient raised significant concerns. Genetic research of this nature usually depends on open, international collaboration involving numerous research teams and patients who willingly share their genetic data and

family history to contribute to research. Myriad Genetics made significant use of the results of publically available and publically funded research in developing their patent and prevented patients under their care from accessing their genetic data. Myriad Genetics enforced this patent over a period of 17 years and claimed exclusive rights to BRCA testing and to patient genetic data. During this period clinical genetic testing for cancer was often conducted using a multigene panel for numerous other cancer variants of interest along with the more expensive Myriad test for BRCA. Many patients were unable to afford to get tested for BRCA predisposition. In 2013, the Supreme Court overturned Myriad's patent in a case pursued by the American Civil Liberties Union (ACLU), determining that genes, being products of nature, could not be patented. In 2016, the ACLU lodged a complaint with HIPAA, to require Myriad to release the complete genotype data generated from its patients[480]. As of 2023, it has been reported that Myriad Genetics has committed to submitting hereditary cancer risk variants to ClinVar[481]. Recent advancements in data protection legislation and ethical standards for dealing with genetic data aim to prevent a situation similar to this from recurring in the future.

1.4 Thesis overview and research question

The identification of somatic mutations plays a crucial role in cancer care. The routine application of molecular diagnostics to align therapies with target mutations has significantly enhanced patient outcomes and quality of life[482]. This improvement in patient outcomes continues to grow as more therapies become accessible. In research, somatic variant detection has uncovered key drivers of the disease and identified other vulnerabilities that can be clinically targeted to impede the progression of malignancy. The emphasis on somatic variant detection has spurred extensive research, yielding a diverse array of tools and methods designed to identify mutations. However, methods designed to validate the output of these tools have remained remarkably limited in comparison. Ambiguities in truth sets used to validate these methods lead to inconsistent results. Resolving these inconsistencies is challenging. Limited tools are available to explore somatic mutation data. It is often difficult to ascertain whether the error lies with the variant caller or in the ground truth data by which it is assessed. Ideally, a comprehensive truth set should not only provide a definitive identification of incorrect calls but also, critically, enable us to explain why the variant caller made an error. Unfortunately, current validation methods lack such insight.

The task of identifying sources of error in mutation calling pipelines becomes more challenging due to the complex artefact filtering protocols employed by these pipelines. It is challenging to quantify the impact of such filters in terms of the trade-off between sensitivity and specificity or to assess if the thresholds on which they are based are suitable for the particular research or clinical requirements at hand. The mutant allele frequency spectrum is also of particular interest both in research and clinical practice. It shapes our understanding of tumour evolutionary dynamics and, in precision oncology, helps identify the clonal variants driving the malignancy. Furthermore, the allele frequency spectrum, along with other details such as mutation context and the biological impacts on proteins, offers valuable insights into the sources of errors in somatic mutation data. However, research in this area has been hindered by a lack of tools to effectively explore the impacts of analytical choices

and thresholds involved in variant filtering. In Chapter 2, we introduce vcfView, a graphical tool that enables researchers to investigate the impact of different filtering approaches on the mutant allele frequency spectrum, mutational signatures, and functional effects on proteins inferred from a somatic VCF file. We demonstrate the utility of this tool in the TCGA AML cohort by uncovering evidence of tumour DNA in the control sample and re-evaluating candidate somatic variants that were previously excluded from the analysis to recover clinically actionable information.

In Chapter 3 we outline methods to create a comprehensive and realistic truth set from tumour genomic sequencing data. To accomplish this, we devised a novel computational framework for simulating tumour sequencing data to match the phased, personalised genome of a donor from the 1000 Genomes project and the chosen sequencing strategy. We use this framework to explore the impact of somatic variant caller filters on sensitivity and provide a definitive assessment of false positive and false negative mutation calls. We predict caller sensitivity as a function of allele frequency for an individual sequencing configuration and determine the empirical mutant frequency spectrum corresponding to the neutral model of tumour evolution. Our simulations highlight biases in caller-estimated allele frequencies as a function of sequencing depth and explain sources of false positive mutation calls in FFPE and oxidative-damaged tumour sequencing data. Finally, in Chapter 4, we apply these methods to analyse FFPE damaged, tumour-only, low depth sequencing and somatic variant data from a cohort comprising 60 individuals diagnosed with early-onset and aggressive pancreatic ductal adenocarcinoma, recovering additional clinically actionable and research-relevant somatic mutation information.

Our research questions can be summarised as follows:

1. What impacts does variant filtering have on type II errors in somatic mutation data? Can we gain new insights into cancer treatment and research by re-evaluating variant records that have been removed from analysis by filtering? (Chapter 2)

2. Can we create a comprehensive and realistic truth set from tumour genomic sequencing data that not only definitively identifies errors in somatic mutation data but also, crucially, allows us to explain why the variant caller made the incorrect call? (Chapter 3)

3. Can we apply the research tools and methods that we have developed to overcome the challenges presented by a low-depth, tumour-only, heavily DNA-damaged dataset and recover clinically and research-relevant information? (Chapter 2, Chapter 4)

2 vcfView

The contents of this chapter has been published[483] *as:*
O'Sullivan, B.; Seoighe, C. vcfView: An Extensible Data Visualization and Quality Assurance Platform for Integrated Somatic Variant Analysis. *Cancer Inform* **2020**, *19*, 1176935120972377

Brian O'Sullivan developed the concept and wrote the code. Cathal Seoighe supervised the research and suggested additional features and analyses. The authors co-wrote the paper.

2.1 Abstract

Motivation: Somatic mutations can have critical prognostic and therapeutic implications for cancer patients. Although targeted methods are often used to assay specific cancer driver mutations, high throughput sequencing is frequently applied to discover novel driver mutations and to determine the status of less-frequent driver mutations. The task of recovering somatic mutations from these data is nontrivial as somatic mutations must be distinguished from germline variants, sequencing errors, and other artefacts. Consequently, bioinformatics pipelines for recovery of somatic mutations from high throughput sequencing typically involve a large number of analytical choices in the form of quality filters.

Results: We present vcfView, an interactive tool designed to support the evaluation of somatic mutation calls from cancer sequencing data. The tool takes as input a single variant call format (VCF) file and enables researchers to explore the impacts of analytical choices on the mutant allele frequency spectrum, on mutational signatures and on annotated somatic variants in genes of interest. It allows variants that have failed variant caller filters to be re-examined to improve sensitivity or guide the design of future experiments. It is extensible, allowing other algorithms to be incorporated easily.

Availability: The shiny application can be downloaded from GitHub (https://github.com/BrianOSullivanGit/vcfView). All data processing is performed within R to ensure platform independence. The app has been tested on RStudio, version 1.1.456, with base R 3.6.2 and Shiny 1.4.0. A vignette based on a publicly available data set is also available on GitHub.
Keywords: R, cancer, shiny, visualization, VCF, variant filtering, SNV, indel, SBS, COSMIC v3

2.2 Introduction

A broad array of variant callers and computational pipelines serving a diverse range of research requirements has been developed to identify somatic mutations from DNA sequencing data. Each method comes with its own set of performance characteristics[484]. Despite differences in calling algorithms and applications, most use tumour and normal next-generation sequencing (NGS) data, aligned to a reference as input and output detailed tumour-specific single nucleotide variants and indel

records in variant call format (VCF)[485]. One of the most popular callers in clinical oncology[486], Mutect2[404], has been shown to perform well in terms of overall balanced accuracy[484]. It employs a series of filters to flag likely false-positive variants, resulting from biases, artefacts, or failure to meet confidence thresholds.

The frequency spectrum of mutations is often of particular interest. Somatic mutations with relatively high frequency (>0.25, accounting for ploidy and sample purity)[66] are often clonal (i.e. they occur in every cancer cell) and some of these may be cancer driver mutations and therefore of particular interest for precision oncology. The somatic mutation spectrum can also provide information about the evolutionary dynamics of the tumour[66,487]. In the case of blood cancers, variant frequencies are used for risk stratification and prognosis for a number of myeloid malignancies[488,489]. High throughput sequencing can be used as an alternative to polymerase chain reaction (PCR)-based clinical analysis of mutant allele burden[490], with the advantage that it has the potential to provide an accurate estimate of the mutant frequency and can detect clinically relevant mutations that are not targeted a priori. Moreover, mutational signatures that can be recovered from high throughput sequencing data have been associated with distinct clinical outcomes and are emerging as potential biomarkers for novel targeted therapies[491].

There are significant technical challenges that inhibit the application of next-generation sequencing in cancer treatment, including the lack of user-friendly tools and data analysis pipelines[492]. The data derived from cancer sequencing is complex, making it difficult to extract information on relevant variants. Candidate variants flagged by the caller as having failed quality filters are routinely removed from analysis; however, they have the potential to highlight sources of technical artefacts. These variants may also contain false negatives that are of clinical or research interest. Command-line based, hard-filtering approaches such as VCFtools[485], GATK VariantFiltration[401], and SnpSift[493] use variant attribute values combined with logical operators to further subset a VCF file. Such tools are complex to configure and lack a means to review the impact of analytical choices involved when subsetting the data. This significantly limits the scope and pace of exploratory analysis of VCF data.

We developed vcfView as an interactive graphical tool to support filtering of putative somatic variants. vcfView displays the allele frequency spectrum as well as mutation patterns and signatures inferred from all putative mutations so that the user can assess the impact of different filtering choices on the variants discovered. All displays are updated dynamically as the user adjusts the filters that are applied to the data. Users can also display somatic mutations on a gene of interest, gaining insights about which mutations are lost from the gene as different filters are applied. To demonstrate its utility, we use vcfView to isolate putative tumour-in-normal (TiN) variants in acute myeloid leukaemia (AML) samples from The Cancer Genome Atlas (TCGA). These variants, which were removed from analysis in TCGA, were significant enrichment for known AML driver mutations.

2.3 Features

vcfView enables users to analyse mutations that fall within a region of the mutation frequency spectrum and that pass or fail user-specified quality filters. Analyses available by default include mutation patterns, mutational signatures and the func-

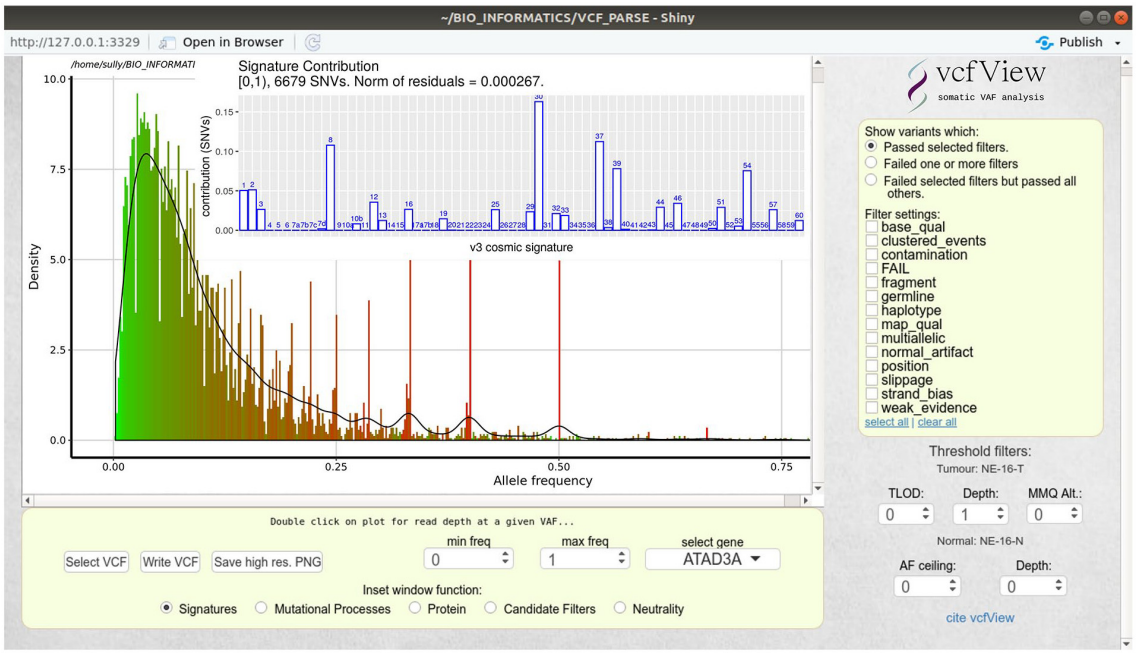

Figure 2.1: vcfView user interface. Display shows the VAF Density plot with signatures inset plot active, filter panel to the right and inset function selection below.

tional effects of the mutations on proteins of interest. The user can explore how these change in different regions of the frequency spectrum or when different filters or other user configurable thresholds are applied. The interface features a variant allele frequency (VAF) density plot within which a region of interest may be selected for further filtering or analysis. The results are displayed inset within the top corner of the density plot. An additional feature allows for a summary file created from a number of individual VCF files to be analysed to identify patterns that may exist across a collection of samples. High-resolution publication standard plots and a filtered copy of the original VCF file may be saved at any stage.

2.3.1 Density plot, thresholds, and filters

The central window of the vcfView user interface is the VAF Density plot (Figure 2.1) that displays the frequency spectrum of somatic mutations. To the right, a series of checkboxes present the researcher with the available quality filters, which are parsed directly from the input file, as well as user configurable thresholds for calling somatic mutations. These provide a means to threshold on the evidence for a somatic mutation at a site. Variants that pass all these filters are displayed in the VAF density plot. The colour in the plot indicates the median sequencing depth of variants in the VAF bin shown on the x-axis. All plot data are updated when the user modifies any filter or threshold settings. The density plot also includes interactive click and drag functionality, enabling the researcher to extract a region of interest from the allelic spectrum for further analysis.

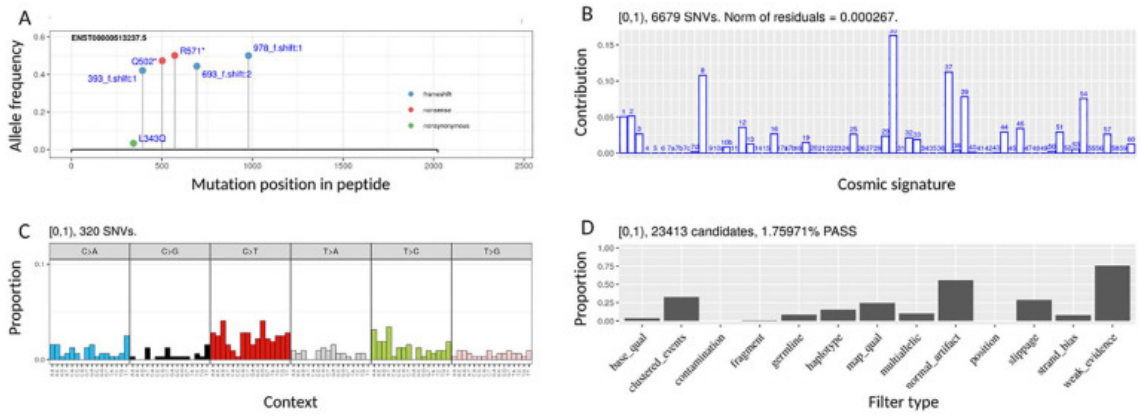

Figure 2.2: vcfView inset analysis function plots showing **(A)** protein analysis plot, **(B)** mutational signatures, **(C)** trinucleotide contexts, and **(D)** candidate filters plot.

2.3.2 Inset plots

Variants that are included within the selected frequency range can be visualized in several alternative inset plots on the main VAF display (Figure 2.2). The required inset function is specified using the radio button list below the density plot. The inset plot is triggered when one of these buttons is activated and redrawn when a region of the allelic spectrum is selected (either with the mouse or by updating the relevant numeric inputs on the user interface) or when filter/threshold settings are updated. vcfView is configured with four inset plots by default, displaying mutation signatures, mutation processes, functional impacts of variants on a selected protein, and candidate filters. The architecture is extensible allowing integration of third-party or custom algorithms to produce alternative inset plots. The signatures inset plot displays the estimated contribution of each of the Catalogue of Somatic Mutations in Cancer (COSMIC)[152] reference signatures to the selected mutations. The mutation processes plot shows a bar plot of the fraction of the mutations in each of the 96 trinucleotide contexts[152]. The protein function inset plot is a lollipop diagram indicating the functional impacts of mutations in the selected frequency range on the specified gene. The dropdown list from which this gene is chosen is updated dynamically so that only genes with a protein impacting variant in the selected frequency range are displayed. Finally, the candidate filters function shows the number of variants that have failed each variant caller filter within the selected range.

2.3.3 Package vignette

The package vignette demonstrates the functionality of vcfView using publicly available data from The Texas Cancer Research Biobank (TCRB). It is composed of three main sections. It first demonstrates exploratory analysis of a single VCF file from the data set. Mutational processes and signatures within a somatic allele frequency window are identified. Candidate variants within that region are subsetted according to various thresholds and filter settings, and the impact of selected variants on proteins of interest is visualized. This enables the evidence for the presence of

potential driver variants to be assessed.

The vignette also describes how to generate a summary of all VCF files within a data set to identify patterns that may exist across that cohort. A summary file is loaded into vcfView and analysed in the same way as an individual VCF file. It may be used, for example, to identify mutational signatures or processes across a data set from patients with the same cancer type or who have undergone the same therapy. It can help to determine if variants are impacting a putative driver gene across multiple patients. A subset of candidate variants that have failed specific filters may be re-examined in an attempt to recover information previously excluded from analysis. In the vignette, a summary file is created and used to check for the existence of putative tumour-in-normal variants within the TCRB data set that may have resulted in failure to call some cancer somatic variants.

Finally, the vignette shows how to integrate other packages into vcfView to produce additional inset plots. It provides a simple wrapper example that integrates a third-party algorithm with vcfView's extensible function set. This enables the researcher to use that library within the inset window of vcfView and take advantage of its preprocessing capabilities.

2.4 vcfView Architecture

Interactive visualization is implemented in R[494] with Shiny[495]. All data processing is performed within the R environment to facilitate platform independence. It leverages several Bioconductor packages[496,497,498,499,500] to functionally annotate VCF-formatted data and drive exploratory visualizations within a user-selectable allele frequency window.

VCF-based record details together with amino acid and trinucleotide context annotation is maintained in two internal data structures for ease of manipulation and subsetting. An index into the original VCF object is used to save a subsetted version of the VCF as required. Functional annotation of protein impact (enumerated as frameshift, nonsense, nonsynonymous, or synonymous) provided by the UCSC transcript annotation library[498] is displayed by the protein inset function using ggplot[501] and labelled using ggrepel[502].

Single base substitution (SBS) trinucleotide sequence context annotation for inset function signatures and mutational processes is provided by the Bsgenome annotation library[497]. In the signatures inset function, a summary of this annotation is used in conjunction with the lsqnonneg method (from pracma library)[503] to create the optimal combination of COSMIC v3 SBS signatures required to reconstruct the variant subset. This returns a non-negative linear least-squares fit of the 65 COSMIC mutational signatures (version 3). Processing is similar to that used in the MutationalPatterns[504] library but streamlined to reduce demands on system memory. All plots are rendered using the ggplot2[501] library.

2.5 Visualization of Putative Tumour in Normal Variants in Leukaemia Samples With vcfView

To demonstrate the utility of vcfView, we used it to re-examine TCGA AML data to determine whether potentially relevant prognostic or diagnostic information could be recovered from candidate variants that have previously been excluded from analysis.

Mutect2 filters potential false positives resulting from germline variants by scoring the confidence that a mutation is present in the tumour sample and absent from matched normal, typically a skin sample in the case of blood cancers. This could result in the failure to detect true somatic variants in the cancer if the same mutation is found in a subset of the cells in the skin sample or if the skin sample includes a proportion of cancer cells or cell-free DNA. Many cancer driver mutations have been found to be relatively common in normal skin cells[357]. Moreover, 3% to 5% of all nucleated cells in the epidermis are myeloid derived[505]. Therefore, it is possible that some of the driver mutations that are critical for gaining an insight into the cancer may also occur, albeit potentially at a lower level, in the skin sample. The use of skin as a normal sample for the identification of somatic mutations in leukaemia cells risks removing these variants from analysis. Mutect2 filters variants that are identified in a panel of normal samples to reduce the effects of recurrent sequencing artefacts and common genetic variation. This variant blacklist is usually derived from blood samples; however, a substantial proportion of blood samples may be affected by clonal haematopoiesis and contain somatic mutations that are relevant for blood cancers[506]. Here, we used vcfView to re-examine data from TCGA LAML samples for evidence of somatic variant exclusion due to the presence in the matched normal sample or in the panel of normals. We refer to these variants as putative tumour-in-normal (pTiN).

Exploratory analysis with vcfView highlighted a significant number of candidate variants that had been removed solely because they failed the allele in normal filter despite being present in extremely low amounts in the normal sample relative to the tumour. Further inspection with the protein inset function revealed a large number of these were in known AML drivers. We tested for enrichment of pTiN variants among known AML driver genes and isolated a significant number of pTiN AML driver variants previously excluded from analysis in TCGA LAML.

We downloaded the protected mutation annotation format (MAF) file containing variants previously identified in the TCGA LAML data set of 149 AML patient samples from the NCI's Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/files/66124158-7feb-4b8e-8fc4-393a5e641fea). We retrieved all protein-truncating variants that had failed the allele in normal and/or panel of normal Mutect2 filters for which the VAF was greater than 0.1 and at least 10 times the frequency of the variant in the normal samples. We further restricted to variants at loci with a read depth of at least 20 in both the tumour and normal samples and that were not found in dbSNP.

We next obtained a list of canonical transcript lengths for all protein coding genes in the exome from Ensembl. Further annotation was added to each gene in this list identifying it as AML driver or non-AML driver (as indicated by IntOGen[507]) and recording the number of pTiN variants contained per base of coding sequence. We performed Fisher's exact test for enrichment of known AML drivers among genes containing pTiN variants. To account for coding sequence length, we also compared the number of pTiN variants per base between AML driver genes and non-AML drivers using the Wilcoxon rank sum test.

A total of 3549 protein-truncating variants were flagged as occurring in the normal sample or panel of normals and were, therefore, not called as cancer somatic variants. Of these, 129 met all of the criteria we used for identifying pTiN variants (Table 2.1: see Methods for details). Among these pTiN variants, 16 occurred in

pTiN Gene	Patient	Reason for variant removal	Clinical relevance
ASXL1	TCGA-AB-2917	Allele in normal	Prats-Martin et al[509]
EZH2	TCGA-AB-2817	Allele in normal	Mechaal et al[510]
FAT4	TCGA-AB-2863	Panel of normal	Garg et al[511]
KMT2C	TCGA-AB-2940	Allele in normal	Rampias et al[512]
NPM1	TCGA-AB-2900, TCGA-AB-2924	Allele in normal	Lachowiez et al[513]
PHF6	TCGA-AB-2912	Allele in normal	Przychodzen et al[514]
RUNX1	TCGA-AB-2805, TCGA-AB-2890, TCGA-AB-2927	Allele in normal & PON	Jalili et al[515]
TET2	TCGA-AB-2876, TCGA-AB-2882	Allele in normal	Wang et al[516]
TP53	TCGA-AB-2820, TCGA-AB-2860, TCGA-AB-2878	Allele in normal	Barbosa et al[517]

Table 2.1: Genes and patients affected by pTiN variants relevant to AML in TCGA AML.

AML driver genes obtained from IntOGen29 (Figure 2.3A). One additional variant, RUNX11, not listed as a driver in IntOGen, was reported to be relevant to AML[508] and is included in (Figure 2.3A).

AML driver genes were significantly enriched among the set of genes affected by these 129 pTiN variants ($P = 3 \times 10^{-11}$, Fisher's Exact Test). This difference was also highly statistically significant when we compared the number of pTiN variants per base pair between AML driver genes and other genes to account for differences in length between genes in the two groups ($P = 1 \times 10^{-53}$, Wilcoxon rank-sum test). In all, variants in 9 genes of prognostic or diagnostic value were identified in 15 patients (Table 2.1). Thus, 10% of TCGA LAML patients had potential clinically actionable variants overlooked due to pTiN. For example, two patients had pTiN variants in NPM1 (Figure 2.3B). This would be of potential prognostic value for these patients, as mutations in NPM1, which occur in approximately 30% of patients with AML, are associated with favourable response to standard intensive chemotherapy[513].

This tool is intended for use with matched cancer/normal VCF files only (not for germline VCFs). Although it has been used with VCFs generated by other callers, it is recommended for use with Mutect2 VCFs. It uses the Mutect2-specific 'TLOD' subfield to select within records that contain multiple possible alternative alleles. As other callers do not provide this value in their VCFs non-Mutect2 records containing multiple possible alternative alleles are currently discarded by vcfView. We intend to add an option in a future release allowing the user to specify the field and condition used to calculate the index required to select among multiple possible alternative alleles when using vcfView with VCFs generated by callers other than Mutect2. vcfView has been tested with tumour exome data from MuTect2 GATK3 and Mutect2 GATK4 VCF files[404]. We cannot guarantee it will work with all future versions of GATK without requiring modification. Mitochondrial DNA variants are currently excluded. We intend to add an option for their inclusion in a future release.

2.6 Acknowledgments

The author would like to acknowledge the TCGA data as it relates to data analyzed in the study.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This publication has emanated from research conducted with the financial support of Science Foundation Ireland (grant no. 16/IA/4612).

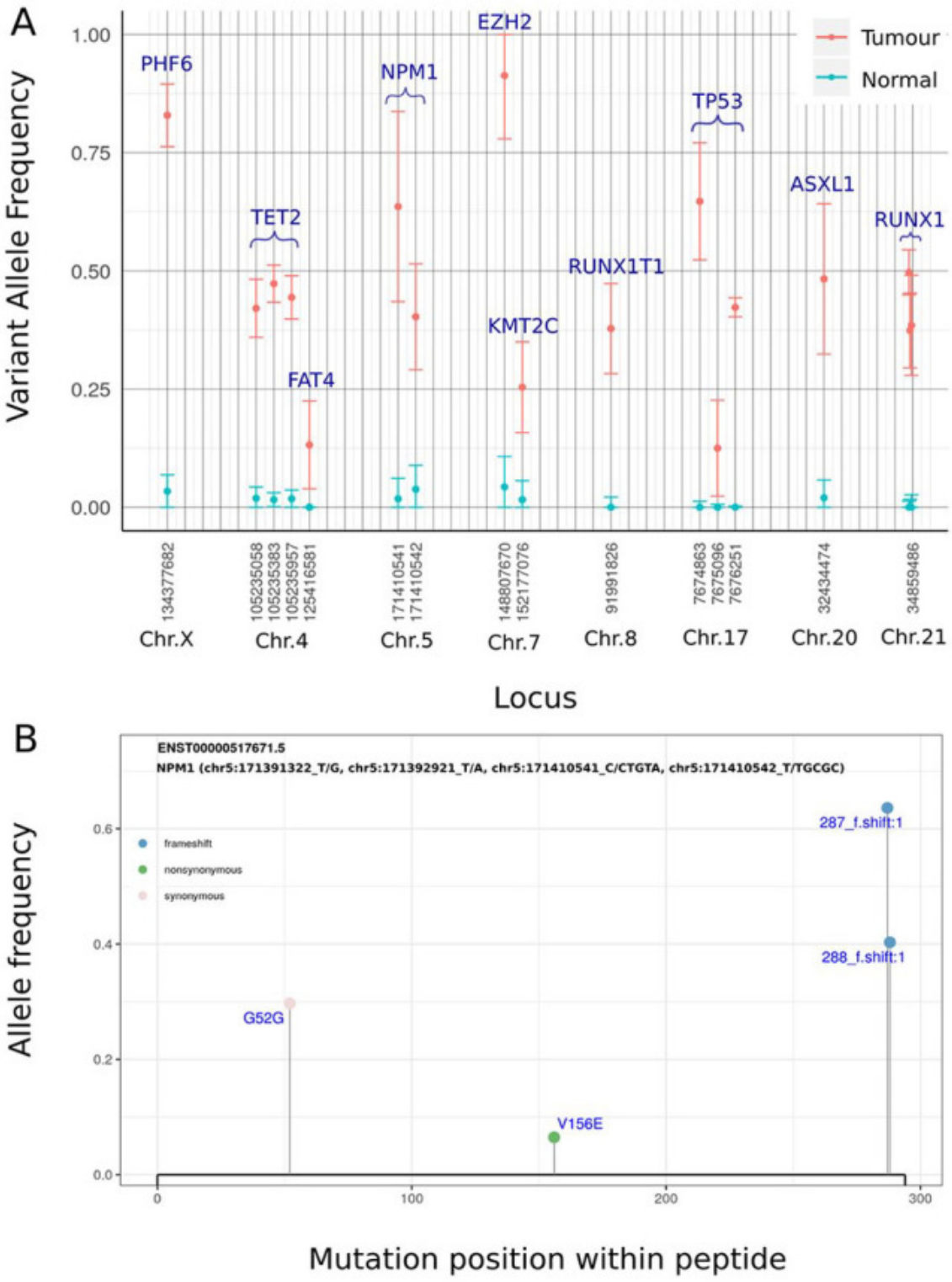

Figure 2.3: **(A)** Alternative allele frequency plot of tumour (blue) and matched normal (red) TCGA LAML pTiN variants affecting AML-relevant genes. Error bars show 95% confidence intervals for the alternative allele frequency (estimated from reference and alternative read counts). **(B)** vcfView protein plot for NPM1 derived from a VCF summary of the TCGA LAML data set.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Contributed by
Author Contributions: BOS developed the concept and wrote the code. CS supervised the research and suggested additional features and analyses. The authors co-wrote the paper.

3 Comprehensive and realistic simulation of tumour genomic sequencing data

The contents of this chapter has been published[163] *as:*
O'Sullivan, B.; Seoighe, C. Comprehensive and realistic simulation of tumour genomic sequencing data. *NAR Cancer* **2023**, *5*, zcad051

Brian O'Sullivan developed the concept and wrote the code. Cathal Seoighe supervised the research and suggested additional features and analyses. The authors co-wrote the paper.

3.1 Abstract

Accurate identification of somatic mutations and allele frequencies in cancer has critical research and clinical applications. Several computational tools have been developed for this purpose but, in the absence of comprehensive 'ground truth' data, assessing the accuracy of these methods is challenging. We created a computational framework to simulate tumour and matched normal sequencing data for which the source of all loci that contain non-reference bases is known, based on a phased, personalised genome. Unlike existing methods, we account for sampling errors inherent in the sequencing process. Using this framework we assess accuracy and biases in inferred mutations and their frequencies in an established somatic mutation calling pipeline. We demonstrate bias in existing methods of mutant allele frequency estimation and show, for the first time, the observed mutation frequency spectrum corresponding to a theoretical model of tumour evolution. We highlight the impact of quality filters on detection sensitivity of clinically actionable variants and provide definitive assessment of false positive and false negative mutation calls. Our simulation framework provides an improved means to assess the accuracy of somatic mutation calling pipelines and a detailed picture of the effects of technical parameters and experimental factors on somatic mutation calling in cancer samples.

3.2 Introduction

The identification of somatic mutations from high-throughput sequencing (HTS) data plays a critical role in scientific research and clinical oncology. Cancer driver mutations continue to inform prognosis, guide therapy and shape our understanding of how cancer develops and evolves. Experimental design and analytical decisions, such as sequencing depth[443], target[444], and the choice of bioinformatics pipelines[404,403,405,407,438,406], all influence the power and accuracy of somatic mutation detection. Assessing their effects on the recovered somatic mutation landscape requires HTS reference data containing a 'ground truth' set of somatic mutations for which the location and source of all loci that contain non-reference bases (a base call that does not match the reference genome at that locus) are known. Such data is a key component in the validation and benchmarking of mutation calling pipelines. The accuracy of the results returned by somatic mutation calling pipelines is critical for many research and clinical applications; however, studies benchmarking these pipelines often publish inconsistent results[484,447,518,446], indicative of the many challenges faced in this area. The variant allele frequency (VAF) spectrum across all

somatic mutations is also relevant for understanding cancer origin and evolution[66] and the development of treatment resistance[519], as well as for inferring clinically important metrics such as tumour mutation burden (TMB), purity and ploidy[520]. Significantly, despite the clinical and research relevance of individual mutation frequencies and the VAF spectrum as a whole, no studies to date have attempted to assess the accuracy with which the frequencies of somatic variants are inferred by mutation callers and this represents a significant knowledge gap.

A number of computational tools have been developed to provide ground truth data to assess the accuracy of somatic mutation callers. Sequencing read simulators, such as ART[453], can be used to generate reads from a reference genome. These are then modified to introduce 'somatic' variants (a process known as spiking-in variants) using software such as BAMSurgeon[408]. Although convenient, a reference based approach does not reflect the diverse sources of variation within real sequencing data. Increasingly, this is being addressed through a 'hybrid' solution employing both real and synthetic sequences[408]. Real sequencing data is subsampled into two sets, corresponding to a virtual matched tumour and normal pair, and somatic variants are then spiked into the tumour reads. However, in addition to the variants that are purposely spiked in, this data also contains low-frequency somatic variants present in the source sample from which they were derived, as well as sequencing errors and other artefacts. The precise origin of this additional variation is likely to be unknown, creating difficulties for the evaluation of mutation callers.

The computational tools that are currently used to spike in somatic mutations also have significant limitations. The number of reads that are edited to introduce the mutant allele at the required locus is typically determined by the product of specified mutation frequency and the sequencing depth at the site. This fails to take account of stochastic aspects of the sequencing process. In reality, sequence reads are a random sample of the DNA at a locus and the observed number of reads containing the alternate allele is, therefore, a random variable. Failure to take this into account can result in spike-in bias, where a variant allele is always spiked-in if the product of the VAF and the sequencing depth is greater than one and never otherwise. For example, a site designated to contain a somatic mutation with a frequency of 10%, which is sequenced to a depth of thirty reads will always contain exactly three reads with the alternate allele if the mutation is spiked in with the widely used tool, BAMSurgeon[408]. However, the number of sequence reads containing the alternate allele for a somatic mutation of this frequency may be greater or less than three in real data.

Here we describe a comprehensive and stochastic tool for simulating tumour sequencing data which, unlike existing methods, enables precise determination of the accuracy and power to detect a somatic mutation as a function of its actual frequency within the cancer sample. Analysis of simulated data generated using this framework provides novel insights into the relationship between the true somatic mutation frequency spectrum and the empirical frequency spectrum obtained following application of a mutation caller. Using our simulations we assess mutation caller bias in variant allele frequency estimates and demonstrate the empirical somatic mutation frequency distribution corresponding to somatic mutations derived from a neutral model of tumour evolution. We also perform a comprehensive assessment of false positive and false negative somatic mutation calls, made possible by the fact that our simulation tool provides the source of all non-reference alleles

in the dataset.

3.3 Materials and methods

A personalised reference genome containing all germline single nucleotide variants (SNVs) and indels annotated for 1000 genomes donor HG00110 (female of English and Scottish ancestry) was created. This was used as a base to simulate normal and pre-tumour (i.e., before the somatic variants have been spiked in) genomic sequencing data using the ART read simulator configured with default empirical error profile and corresponding to depths of coverage, 100x, 200x, 350x and 600x. All reads in the SAM output generated by ART are aligned to their true location within the personalised phased genome from which they were simulated. The target simulated was an exome capture consisting of all hg38 exons plus an additional 100 base pairs at the 3' and 5' ends of each capture range. This data was then used as a base to spike in the required somatic distribution of SNVs. Once the spike-in process was completed the phased data (a BAM pair corresponding to the maternal/paternal haplotype set) was merged and realigned against a standard reference (Figure 3.1). Realignment was performed using BWA (v0.7.17) [400] with the hg38 reference genome. Somatic variant calling was performed using Mutect2 according to GATK (4.2.2.0) Best Practices for somatic short variant discovery [521]. Cross-sample contamination and Base Quality Score Recalibration stages were not run as the data was simulated without contamination or systematic biases. Mutect2 was called with arguments set to ensure all filtered variants were recorded in the Variant Call Format (VCF) file (`tumor-lod-to-emit=0`). The stochastic simulation framework is written in C, on the HTSlib 1.13 API [522].

3.3.1 Simulation of mutation allele frequency spectrum

We used our simulation framework to simulate somatic mutations with defined frequencies. These simulations included the full mutant allele frequency spectrum of a diploid tumour expected under a neutral evolutionary model [66]. We also simulated a low frequency, high burden point-mass at a fixed frequency and finally a uniform distribution of somatic mutation frequencies to investigate caller detection rate and inferred allele frequency as a function of true allele frequency. The first two simulations were repeated across different depths of coverage (100x, 200x, 350x and 600x) to explore the effect of depth on somatic mutation detection and inferred allele frequency. The uniform distribution was simulated at a fixed depth of 100x. Finally, to illustrate the distinction between somatic variant simulation using BAMSurgeon and our simulation framework the point-mass simulations were repeated using BAMSurgeon to spike-in the required burden.

3.3.2 Complete allele spectrum simulation of a diploid tumour derived from a neutral evolutionary model

A total burden of 2,681 somatic variants within a true frequency range 0.01 to 0.25 was simulated, with a variant allele frequency spectrum for subclonal mutations corresponding to the neutral model of evolution [66], Figure 3.2**A**, (i.e. cumulative distribution function of the mutation frequency, f, proportional to $\frac{1}{f}$). The clonal mutations were simulated with a fixed frequency of 0.5 (the simulation had 100%

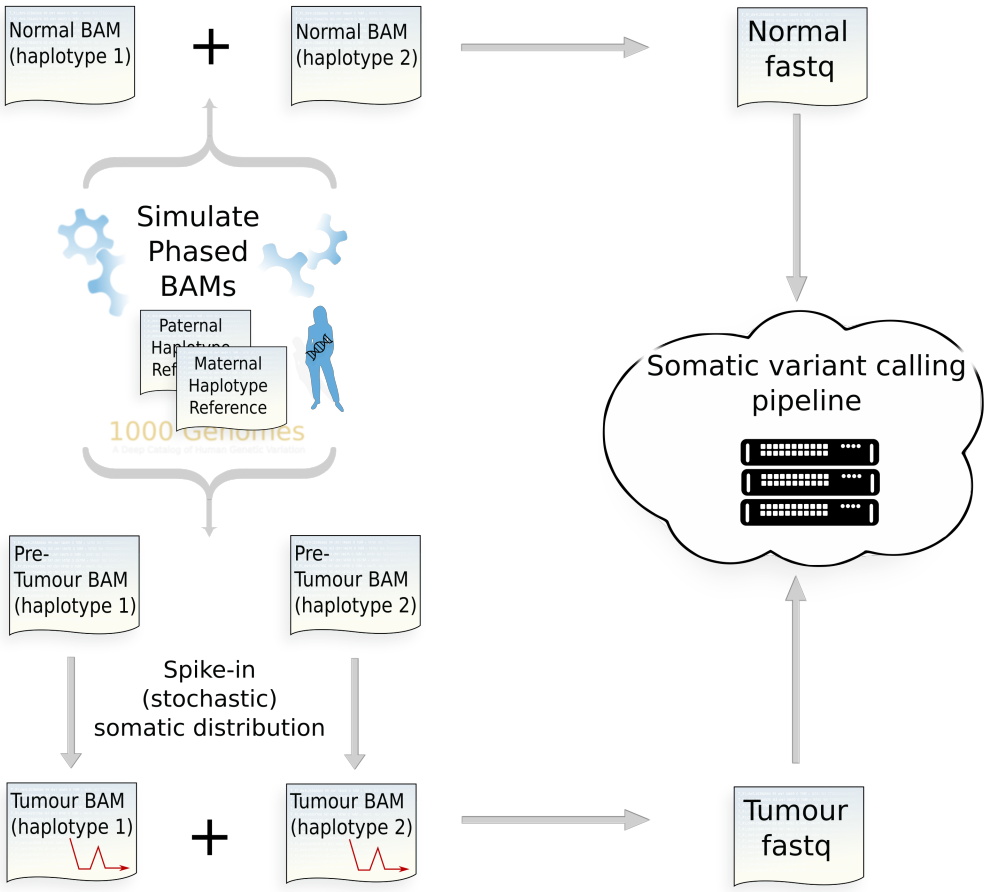

Figure 3.1: Tumour stochastic HTS simulation framework. Personalised phased donor genome incorporates all SNPs and indels recorded from any 1000 genomes donor.

tumour purity). Mutations were spiked into each of the four pre-tumour phased BAM pairs (with depths of 100x, 200x, 350x and 600x). Mutect2 was then run on the resulting matched tumour-normal pairs. The resulting variant output (VCF) was compared to the ground truth values.

3.3.3 Low frequency, high burden point-mass somatic distribution

A low frequency burden of 10,000 somatic variants, all with a true allele frequency of 0.035, was spiked into each of the four pre-tumour phased BAM pairs (100x, 200x, 350x and 600x) using the stochastic simulation framework. Mutect2 was then run on the matched tumour-normal pairs. The resulting variant output (VCF) was correlated against its ground truth values (again using the simulation framework). The observed VAF spectrum was plotted showing the dispersion of allele frequencies about their ground truth at each of the four sequencing depths.

3.3.4 Uniform somatic distribution

We divided the first half of the allele frequency range (i.e. frequencies from 0 to 0.5) into 100 semi-centile bins. Into each bin we spiked in a uniform distribution

of 10,000 somatic variants at loci randomly distributed across the target region in 100x pre-tumour HTS data. For each semi-centile we recorded the percentage of the ground truth burden that was passed (considered true somatic) or filtered (considered artifactual) and its associated allele frequencies, as inferred by Mutect2. From this we created a matrix detailing the detection rate for each semi-centile and the regions of the spectrum in which that burden was detected, as annotated by Mutect2. This enables us to predict how the caller performs in detecting a true burden and identifying its associated allele frequencies. The reasons candidate somatic variants were filtered by the caller in each semi-centile were also recorded. The simulations were carried out in groups of 4 semi-centiles per simulation, with each simulation containing a total burden of 40,000 somatic variants, yielding 25, 100x, tumour/normal pairs which were subsequently analysed with Mutect2.

3.3.5 Simulations of FFPE and 8-oxoG artefacts

We performed additional simulations that included artefacts that are typical of formalin-fixed, paraffin-embedded (FFPE) samples and oxidative DNA damage. To simulate artefacts associated with FFPE samples, we downloaded high coverage (minimum depth $\geq$ 500) colorectal variant call data[523] from three patients, each containing two samples, one fresh frozen (FF), the other FFPE (48 hour fixation time), both of which had been resected from the same tumour. Using variants identified only in the FFPE sample, we estimated the FFPE SBS signature (based on the conventional 96 triplet mutation types) associated with this data and created a context-specific FFPE distribution of simulated artefacts. We then created a second distribution based on COSMIC signature SBS45[524] to simulate oxidative damage during sample preparation. Both distributions were combined with an empirical set of DNA damage artefact allele frequencies[523]. To preserve the required orientation the 100x pre-tumour BAMs were each split into two separate files, one with reads from inserts that aligned to the forward genomic strand and the other containing the remaining reads. The target burden, totaling 7332528 DNA damage artefacts, was then divided evenly between forward and reverse strand alignment BAM files and spiked-in using the stochastic simulation framework. All files were merged back into the final tumour BAM on completion and realigned against the hg38 reference before being subjected to variant calling, using Mutect2 in tumour-normal mode. These simulations did not include any somatic mutations and, consequently, any variants identified by the caller were false positives.

3.4 Results

The simulation framework developed here, which we refer to as stochasticSim, has significantly enhanced functionality compared to existing methods (Table 3.1). A key feature is the fully comprehensive truth set. Truth sets based on data derived from controlled mixtures of distinct samples (usually cell lines) or by spiking in mutations into a single sample contain germline variants, sequencing errors and alignment errors, among other artefacts. They also contain true somatic mutations present, usually at low frequencies, in the original samples. As a consequence, these simulation methods do not provide an accurate and complete set of the true somatic mutations in the sample[452,451]. This is required, for example, for an accurate assessment of the performance of methods to detect somatic mutations. In contrast,

Method	Features												
	Germline simulation			Somatic simulation			Alignment		Haplotype aware	Context aware	Orientation aware	Stochastic	Comprehensive truth set
	SV /CNV	SNP	INDEL	SV /CNV	SBS	INDEL	pre-spikein	post-spikein					
Stochastic Simulation	NO	YES	YES	NO	YES	NO	True	hg38	YES	YES	YES	YES	YES
Bamsurgeon 2018[408]	NO	NO	NO	YES	YES	YES	Estimate	hg19	YES	NO	NO	NO	NO
Cell line admixture[450]	NO	NO	NO	NO	YES	NO	Estimate	hg19	NO	NO	NO	YES	NO

Table 3.1: Comparison of the functionality of tumour simulation methods. Somatic indel simulation is not yet supported by stochasticSim. All germline indels, which account for the vast majority of indels in tumour samples, are simulated however.

a comprehensive truth set not only allows us to identify true and false positives definitively, but it also enables us to identify the cause of all false positive calls.

Our simulation framework provides a complete record of the source of all non-reference bases in the data. This allows us to assess the sources of all false positive and false negative mutation calls. In the first simulation a total of 2,681 somatic mutations were simulated across a range of frequencies corresponding to a neutral model of tumour evolution[66] (Figure 3.2**A**), with sequencing depths ranging from 100x to 600x. The false positive rate was extremely low, with just one false positive (due to sequencing error) being detected over four simulations at different sequencing depths. Overall detection rates at each of the four sequencing depths (100x, 200x, 350x and 600x) were 22%, 28%, 32% and 36%, respectively. The relatively low proportion of mutations detected in this simulation reflects the large proportion of low-frequency variants implied by the neutral model of tumour evolution (about 75% of the mutations had a true allele frequency less than 0.05). As the number of reads carrying the alternative allele is a random variable a number of mutant alleles were not found at all in the simulated data and therefore would not be detected by any mutation calling pipeline. At each of the four sequencing depths (100x, 200x, 350x and 600x), 24%, 13%, 8% and 4% respectively of the true somatic burden received no coverage of the alternative allele at the variant locus. The majority of the remaining true somatic variants were present at too low frequencies to be considered by Mutect2 for filtering and were dropped without leaving any record in the VCF file (Figures 3.2).

A small proportion of all somatic variants (approximately 2% at 600x), most of which were also of low allelic frequency were missed as the reads supporting the alternative allele were incorrectly aligned against the reference genome. A substantial number of the true somatic mutations were removed by mutation caller filters which incorrectly identified them as artefacts (Figure 3.2**B**). The contribution of different filters varied across sequencing depths. Interestingly, the number of true somatic mutations that failed these filters increased with increasing sequencing depth. This was mainly due to the alternative allele being incorrectly flagged as a germline or other artefact common to both the tumour and normal sample (`normal_artifact`). As sequencing depth increases so too does the probability of a read error in the normal sample occurring at the same locus and with the same base as a true somatic variant in the tumour thereby increasing the number of variants filtered in this way.

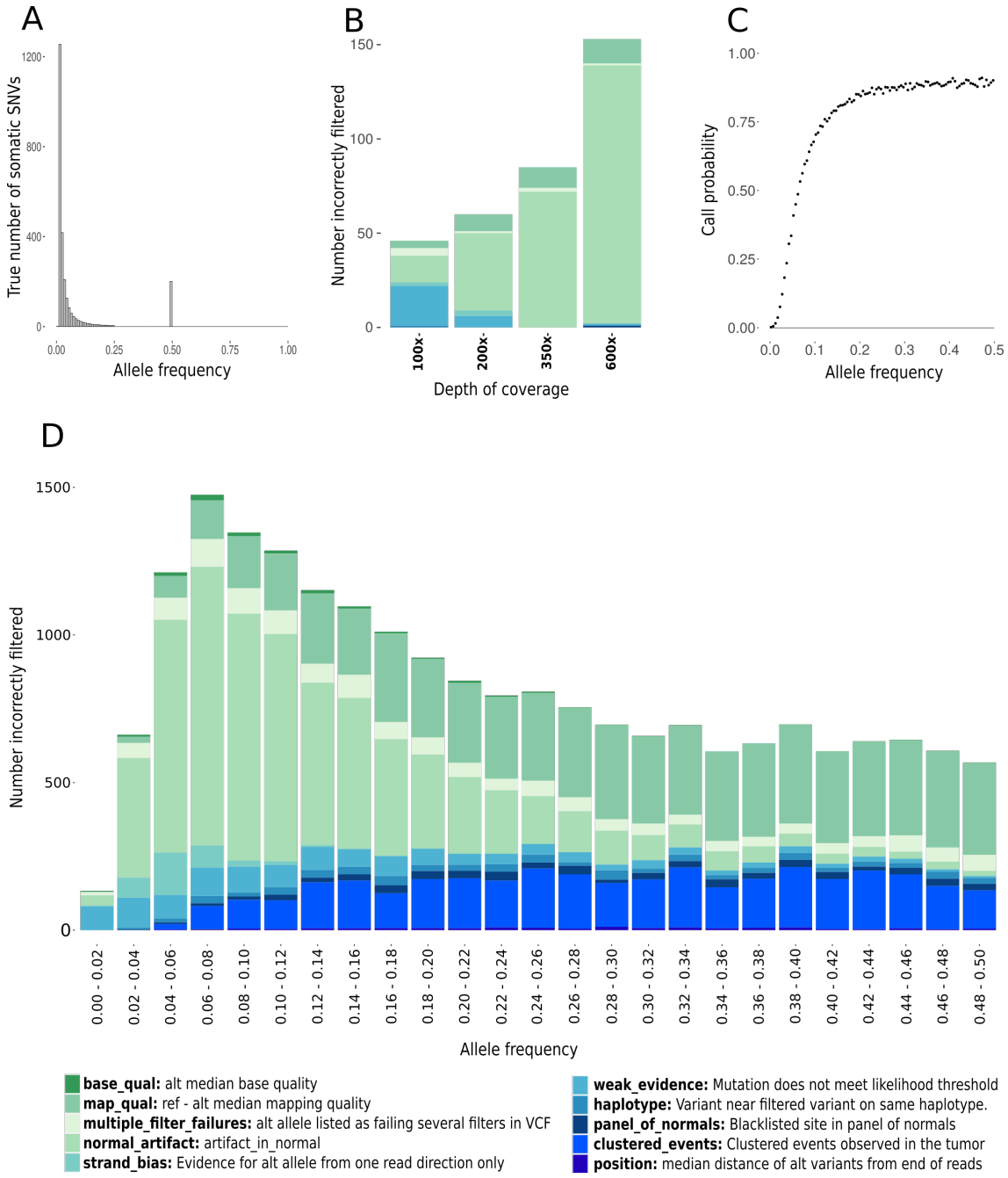

Figure 3.2: **A.** The 'ground truth' or true frequency spectrum of somatic mutations in our simulation of a diploid tumour derived from a neutral evolutionary model. The total true burden was 2,681 somatic variants. **B.** Number of true somatic variants incorrectly filtered by Mutect2 as artefacts, stacked by filter type, from neutral model simulations at each of the four sequencing depths, 100x, 200x, 350x and 600x. **C.** Probability that a true somatic mutation is passed by Mutect2 as a function of its true allele frequency for 100x sequencing data with 100% tumour purity. This simulation was also repeated over a range of depths on a reduced target size (Supplementary Figure 1). **D.** Number of true somatic variants incorrectly filtered as artefacts in the uniform frequency simulations, stacked by filter type. Each bar represents the fractions of false negatives incorrectly excluded by the caller from a total somatic burden of 40,000 variants.

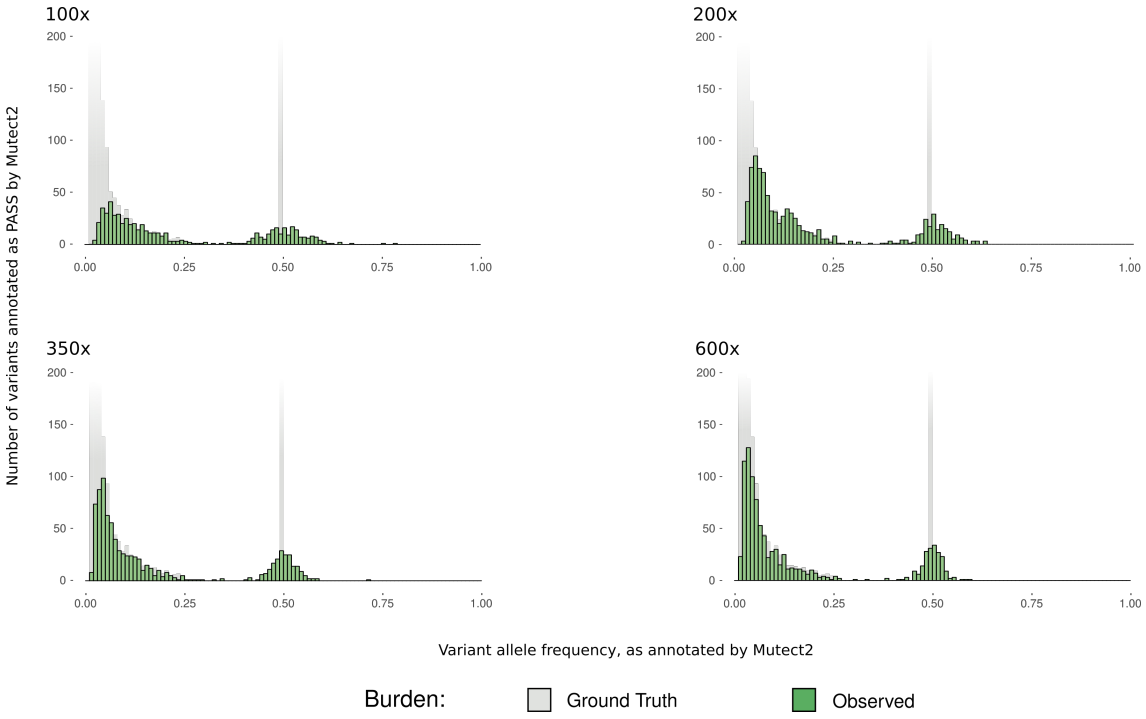

Figure 3.3: VAF plots from Mutect2 output of a diploid tumour derived from a neutral evolutionary model, overlayed on the ground truth. Ground truth burden is faded where it starts to extend beyond the y-axes. Depth of coverage is as indicated on each plot.

3.4.1 Probability of somatic mutation detection as a function of mutation frequency

To investigate the relationship between the probability of detecting a somatic mutation and its frequency in the tumour sample, we simulated somatic mutations distributed uniformly over a range of frequencies from 0 to 0.5. The true number of somatic variants together with the total number detected by the caller in each semi centile were recorded, allowing us to assess the sensitivity to detect somatic variants over a range of allele frequencies. As expected, the detection rate (defined as the probability of a true somatic mutation being passed by the caller) was a strong function of the simulated variant frequency (Figure 3.2**C**). The `normal_artifact` filter accounted for approximately 34% of true somatic variants incorrectly filtered by Mutect2 with a significant contribution at lower frequencies (< 0.2) (Figure 3.2). The number of true somatic mutations removed by the `clustered_events` filter increased with increasing frequency (Figure 3.2**D**). This number was relatively high in these simulations due to the high burden of mutations simulated in each frequency band (40,000 somatic variants, randomly distributed across a 76 megabase target region). This high burden increased the probability of multiple somatic mutations being spiked-in in close proximity, resulting in them being flagged by this filter. A small number of true somatic mutations were filtered due to strand bias but this was not noticeable beyond an allele frequency of 0.1 and it should be noted that no strand specific artefacts were simulated.

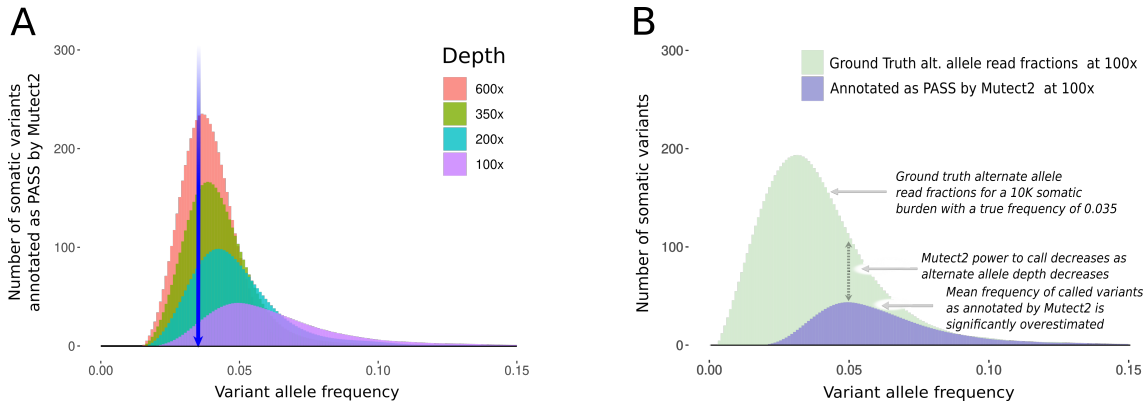

Figure 3.4: **A.** The VAF distribution as inferred by Mutect2 from simulated data consisting of 10,000 somatic mutations each with a true allele frequency of 0.035. The blue arrow indicates the true allele frequency at which the somatic burden is located (the ground truth). Each of the overlay plots indicates what is inferred by Mutect2 at the sequencing depths indicated. The data has been processed using the smooth.spline function from the base R stats package[525]. The same data are also available without smoothing (Supplementary Figure 4). **B.** Explanation of the somatic variant low-frequency caller bias, as annotated by Mutect2 for the 100x data from the previous panel.

3.4.2 Empirical mutation frequency spectrum corresponding to a neutral model of tumour evolution

Information on how tumours evolve is relevant for gaining a better understanding of cancer origin, the development of immune evasion and resistance to treatment. Models of tumour evolution have implications for the frequency spectrum of somatic mutations observed in a cancer sample; however, the relationship between a theoretical frequency spectrum and the empirical spectrum that is observed when mutations are called using existing computational pipelines is unclear, particularly in the case of moderate sequencing depth. We assessed the mutation frequency spectrum recovered by the caller for the simulations corresponding to the neutral model of evolution[66] over a range of sequencing depths (100x, 200x, 350x and 600x). These empirical distributions differ qualitatively for different sequencing depths, with lower depth simulations having a much higher proportion of mutations at intermediate frequencies than predicted by the neutral model of tumour evolution (Figure 3.3). As expected, the observed frequency spectrum resembled more closely the expected form (with a cumulative distribution function proportional to the reciprocal of the frequency) at the higher sequencing depths. The inferred frequencies of somatic mutations are also relevant for the calculation of TMB, which is typically defined as the number of somatic single nucleotide variants per megabase (mut/MB) with an inferred frequency ≥ 0.05. The ground truth TMB for the diploid tumour in this simulation was 8.5 mut/MB (equivalent to 645 variants with a true frequency ≥ 0.05, at 100% tumour purity for a 76MB target). TMB was estimated at each of the four sequencing depths (100x, 200x, 350x and 600x) as 6.99, 7.70, 7.84 and 7.84 mut/MB respectively, with a 12% increase in estimated TMB between 100x and 600x.

3.4.3 Misestimation of mutation frequencies

To illustrate the impact of stochastic effects on the estimation of somatic mutation frequencies we simulated 10,000 somatic mutations at a fixed frequency of 0.035. The detection rate (i.e. percentage of the somatic mutations annotated as PASS by Mutect2) at sequencing depths of 100x, 200x, 350x and 600x was 20%, 33%, 45% and 54%, respectively, with very small numbers of false positive somatic mutations at each depth (6, 2, 1 and 3, respectively with one of the false positives resulting from incorrect read alignment and the remainder from sequencing error). The mean inferred frequencies returned by the caller were 0.065, 0.050, 0.044 and 0.040, illustrating an upward bias (relative to the true frequency of 0.035; Figure 3.4**A**), which decreases with increasing sequencing depth. The bias results from the relationship between the probability of detecting a somatic mutation and the number of reads containing the mutation. As we move to the right of the spectrum (Figure 3.4**B**), the fraction of the variants recovered (ratio of the height of the Pass to Ground Truth histograms) increases. This results in an allele frequency distribution for Pass variants with a mode that is shifted to the right, relative to the Ground Truth distribution. We have identified similar biases in variant allele frequencies previously, in the context of the use of a mutation frequency threshold in the calculation of TMB[526]. However, as seen in these simulations, a threshold is not required to observe a bias in the inferred variant frequencies (which are estimated from the read fractions). We also used these simulations to compare our stochastic simulation toolkit with read fraction methods of simulating HTS data by repeating the simulation using BAMSurgeon[408] to spike-in the required distribution. A similar total burden was detected by Mutect2 from both simulations (BAMSurgeon totals: 2130, 3016, 4376, 5303, stochastic simulation totals: 2008, 3349, 4473, 5398); however, there were substantial differences in the VAF spectrum associated between the two cases (Supplementary Figure 2).

3.4.4 Simulations of FFPE and 8-oxoG artefacts

FFPE is a method of tissue preservation enabling samples to be stored at room temperature almost indefinitely[390]. The procedure also creates asymmetric DNA damage such as deamination of cytosine to uracil (resulting in the detection of C>T transitions)[523]. Similarly, oxidative DNA damage introduced during sample preparation, for example as a by-product of acoustic shearing[527], generates 8-oxoguanine (8-oxoG) leading to G>T transversions. Both types of DNA damage usually manifest as a high burden of low frequency artefacts in sequencer output. Only a small proportion of the simulated FFPE and 8-oxoG burdens (11283 DNA damage artefacts from a true total of 7332528, or .0015) met the required (GATK) significance threshold to be considered for filtering. The remainder was either ignored by Mutect2 or no reads carrying the artefacts were recovered from the simulation output. Of these, 7072 were correctly removed by standard Mutect2 filters (such as `weak_evidence` or `base_qual`). An additional optional filter in Mutect2, which checks for evidence of orientation bias in the variant call, removed most (99%) of the artefacts that made it through the standard set of filters. However, even after this optional filter, 42 variants corresponding to simulated FFPE and 8-oxoG artefacts were incorrectly annotated as PASS. Although, the 42 artefacts that made it through all filters represented only a very small proportion of the original number

of sites at which artefacts were simulated, this number could have a substantial impact on results in many studies. A large-scale empirical analysis of TMB in over 100 tumour types indicated a median value of 2.7 mut/MB[528]. This would translate as 205 somatic mutations on the simulation target used in this study. Together with 42 artefacts incorrectly annotated as PASS this would imply a false positive rate of 20%.

As expected, the mutational profile detected by Mutect2 resembles an FFPE and 8-oxoG DNA damage signature (Supplementary Figure 3). Interestingly, the total median depth at true negative loci (where the DNA damage artefact was correctly filtered by Mutect2) was 95, as opposed to 50 at false positive loci (where the damage was incorrectly passed). The median number of reads supporting the alternative allele recorded by the caller was 5 at true negative loci and 3 at false positive loci, with median alternative allele frequencies of 0.054 and 0.084, respectively. In the case of true negatives, 4043 out of a total of 4169 true negatives contained evidence supporting the candidate somatic allele on one genomic strand only. The remaining records contained evidence from both genomic strands with the evidence from one strand caused by sequence error. In the case of false positives however only 30 out of 42 recorded evidence of the alternate allele from one genomic strand only, suggesting interaction between false positives arising from DNA damage and the occurrence of the same substitution due to sequencing error on the opposite strand (P = 5e-09, from Fisher's exact test). In effect, in the case of 12 false positives, a sequence error enabled the DNA damage artefact to escape the orientation bias filter as evidence of the alternative allele was present on both genomic strands. To explore the scenario in which histologically normal tissue adjacent to the FFPE tumour sample is used as a control we spiked-in the same level of FFPE and 8-oxoG burden to the normal sample and re-ran the somatic variant calling pipeline. This simulation yielded similar results (34 false positives with 12 showing evidence of the alternate allele on both strands).

3.5 Discussion

We have developed a computational framework for simulating personalised, phased, cancer genome sequencing data that creates a somatic SBS cancer distribution in a base BAM file containing all germline indel and SNV variation from a 1000 Genomes donor. Cancer indel and structural variants are not yet simulated by this framework. Our framework provides a comprehensive report on the sources of all non-reference sites in the simulated data and accounts for the randomness in the number of reads that contain the non-reference allele at somatic mutation sites. We have applied this framework to assess the performance of a widely used pipeline to call somatic mutations. In agreement with previous reports[484,404,529] our initial analysis indicated that the GATK4 Mutect2 pipeline had a very low rate of false positive mutation calls. However, preanalytical factors, particularly those associated with sample storage and preparation can significantly impact downstream somatic variant analysis. Artefacts introduced in these stages are overlooked by current bioinformatic simulation methods (Table 3.1) and not accounted for in their assessment of caller specificity. To illustrate this we tested the GATK orientation bias filter against simulated FFPE and 8-oxoG damaged sequencing data and demonstrated a mechanism by which orientation bias artefacts escape GATK filtering leading to

additional caller false positives. We also quantified the number of false negatives, corresponding to true somatic variants that were incorrectly filtered by Mutect2 and investigated biases in allele frequency estimation.

Misestimation of allele frequency may be of particular scientific and clinical relevance. We have previously reported a bias in the inferred mutation frequency when only mutations with an observed frequency greater than a threshold are considered[526]. Despite the absence of an explicit threshold, simulations in this paper reveal a similar allele frequency bias resulting from the dependence of mutation detection probability on the number of reads that support the mutant allele (Figure 3.4). The mutation frequencies at which this bias is observed decrease with increasing depth of coverage (Figure 3.4). The novel simulation framework enabled us to investigate such biases in realistic simulated data using a commonly applied somatic mutation calling pipeline. It also allowed us, for the first time, to determine the observed frequency spectrum that results when mutations from a theoretical spectrum corresponding to a model of tumour evolution are called from cancer sequencing data.

We found that a substantial number of true somatic variants were excluded by the caller as a consequence of being incorrectly identified as an artefact common to both tumour and normal samples. Mutect2 filters candidate variants based on minimal evidence of their presence in the normal sample (`normal_artifact`), even when the variant is present at much greater frequency in the cancer sample. This means, along with filtering a number of (primarily germline) true negatives, true somatic variants can also be removed due, for example to sequencing errors in the normal sample. Our simulations, which used the same depth of coverage in both tumour and normal, demonstrated this can be a significant issue, particularly at high depths. For example, one variant was incorrectly flagged as `normal_artifact` based on its detection by Mutect2 at an allele frequency of 0.00054 in the 600x normal sample. The allele in the normal was, in fact, a sequencing error. The median allele frequencies in the normal for which `normal_artifact` false negatives were excluded at depths 100x, 200x, 350x and 600x were 0.0064, 0.0039, 0.0022 and 0.0016 respectively. In practice, the normal sample is often sequenced to a lower depth than the cancer sample. This would reduce the number of mutations that are lost in this way. However, some assays require the same depth of coverage in the normal (for example, copy number analysis) while another publication recommends as high a depth of coverage as possible in both tumour and normal samples[62]. We recommend manual curation of variants filtered solely as `normal_artifact`, particularly where they may be of clinical relevance.

3.6 Conclusion

High-confidence identification of somatic mutations in tumour samples and accurate inference of their frequencies is important for clinical decision making and in cancer research. Realistic simulations continue to play a key role in this regard, improving our understanding of the performance of computational pipelines that have been designed to identify somatic mutations. The extremely low false positive rates achieved by somatic variant callers such as Mutect2[484,404,529] are in part enabled by an extensive set of filtering steps designed to remove artefacts. However, we have demonstrated that strict thresholds enforced by some of these filters come

at a price, in terms of power, with some true mutations being flagged by the filters. We have highlighted limitations in existing methods of assessing the false positive rates of mutation callers and demonstrated a mechanism through which DNA damage introduced during the preanalytical phase of the sequencing process can lead to false somatic mutation calls. We have also quantified the extent of the bias in the estimated frequencies of the somatic mutations that are identified, as a function of sequencing depth and determined the empirical mutant frequency spectrum corresponding to the neutral model of tumour evolution. Our simulations also allow us to predict caller detection rate as a function of allele frequency. This novel simulation tool can be applied to evaluate the accuracy with which individual mutations or mutation burdens are calculated and to compare the observed frequencies of somatic mutations to their expected distribution under competing models of tumour evolution.

3.7 Data Availability

The software and results relating to this publication are available at Zenodo with DOI 10.5281/zenodo.8155004. The stochastic simulation framework is also available from https://github.com/BrianOSullivanGit/stochasticSim.

3.8 Supplementary data

Supplementary Data are available at NAR Cancer Online at .

3.9 Acknowledgements

The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga and the 1000 Genomes Project, made available through The International Genome Sample Resource[530]. The authors would like to thank the participants of both studies. The authors would also like to acknowledge Harrison Anthony for his help with the installation and testing of the simulation framework GitHub. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (grant no. 16/IA/4612).

3.10 Funding

Science Foundation Ireland [16/IA/4612].
Conflict of interest statement. None declared.

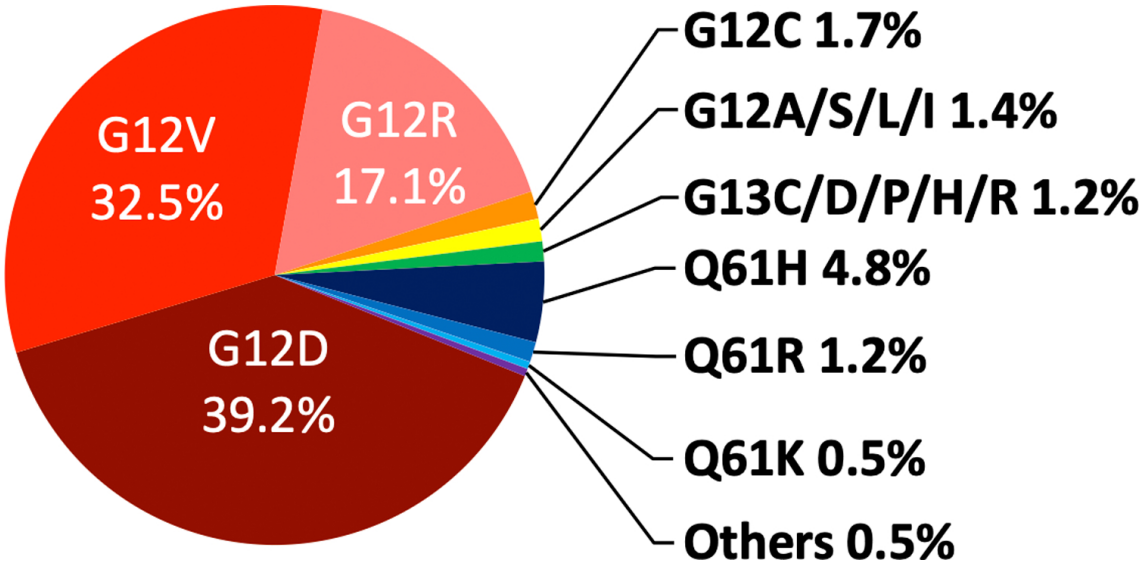

Figure 4.1: Distribution of *KRAS* mutations in pancreatic cancer[536]. Figure courtesy of KRAS mutation in Pancreatic Cancer[536].

4 Pancreatic ductal adenocarcinoma, St James's Hospital cohort study

4.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC), is a severe and highly prevalent form of pancreatic cancer. It is the seventh leading cause of cancer death worldwide[531]. Despite significant advancements in cancer treatment, the 5-year survival rate for PDAC has remained remarkably consistent over the past six decades, at just 12%[531]. The human RAS oncogene family, identified in the early 1980s, is the most frequently mutated oncogene in human cancers and *KRAS*[532], in particular, plays a significant role in PDAC with the predominant driver KRAS-G12, along with less common Q61, G13, and G15 mutations, found in approximately 85% of all PDAC cases (Figure 4.1). Despite considerable research spanning several decades, attempts to develop a KRAS inhibitor proved unsuccessful, leading to its characterization as 'undruggable'. However, a significant breakthrough in 2013 identified a novel binding site in oncogenic mutant KRAS G12C, presenting a promising drug target for inhibition[533]. The discovery culminated in the development of sotorasib[534], the first FDA-approved KRAS inhibitor, indicated for the treatment of G12C mutated non-small cell lung cancer (NSCLC), which effectively stabilises the G12C oncoprotein in an inactive GDP-bound state. This breakthrough has encouraged PDAC research efforts in other targeted therapies for KRAS G12 oncoproteins. Recent findings in murine models involving inhibitor therapy targeting KRAS G12D[535], the most prevalent driver variant in PDAC, have exhibited substantial and enduring tumour regression, instilling fresh hope for a potential breakthrough therapy for the disease.

As part of a collaborative research effort on pancreatic cancer in association with Trinity College Dublin and St James's Hospital, we obtained tumour-only sequencing and somatic variant data from a cohort comprising 60 individuals diagnosed

with early onset and aggressive pancreatic ductal adenocarcinoma, which was suspected to have a hereditary basis. Matched normal control samples were unavailable. The data was described as 30x WGS of FFPE tissue samples that had been subset to a 196MB target region of interest, which included the exome and other regulatory regions of the genome. The mean tumour purity for the cohort was stated as 45%. Somatic variant calling had been performed using the GATK 4.2.6.1 Mutect2 tumour only variant calling pipeline including GATK orientation bias filtering for FFPE treated samples. Initial analysis of variant calling output performed by the bioinformatics team at Trinity College Dublin and St James's Hospital highlighted a higher than expected burden of somatic variants annotated as PASS in the filtered Mutect2 VCF output. The burden was suspected to be primarily germline in origin and concerns of possible DNA damage to the samples were noted. In an attempt to exclude a substantial number of presumed artefactual records from the data, additional filtering strategies were implemented. Supplementary population databases were consulted to further reduce germline artefacts. To mitigate the impact of sequencing artefacts records with fewer than 2 reads supporting the alternate allele and those with an allele frequency below 0.05 were also excluded. However these methods did not achieve the expected reduction in burden. The low incidence of KRAS G12D/V (compared to the prevalence noted in the literature[536]) recovered from somatic variant calling was noted. The team expressed concerns about the suitability of the data for somatic variant analysis in PDAC and the ability to infer the true incidence of *KRAS* within the cohort from the data available. Based on the data, it remained unclear what proportion of the variants detected by Mutect2 represented genuine tumour mutations.

As part of the initial analysis conducted at the University of Galway, somatic variant data from the PDAC cohort was aggregated, and the allele frequency spectrum was examined. Tumour-only somatic variant calling pipelines typically retain a significant number of germline artefacts that persist despite the application of filtering algorithms dependent on germline databases or panels of normals to exclude germline variation from variant caller output. Given that these artefacts exist in the DNA of both tumour and normal cells, we anticipated their distribution in the allele frequency spectrum to centre around 0.5 for heterozygous alleles and one for homozygous alleles. We also expected a smaller proportion of putative tumour mutations centred around an allele frequency of 0.225, considering the reported average tumour purity of 45%. However, the observed burden differed significantly from our expectations.

The burden detected by Mutect2 was notably higher than anticipated (over 50 times the level expected for PDAC[528]), predominantly characterised by a substantial number of variants with frequencies around 0.18. Considering that the cohort samples underwent FFPE treatment and were subject to low-depth sequencing, we opted to explore supplementary variant filtering strategies. Our objective was to determine whether the predominant burden identified by Mutect2 resulted from orientation bias artefacts introduced during sample preparation.

The standard GATK orientation filter applied to this cohort systematically examines the F1R2 and F2R1 attributes located within the format field of each VCF record, in conjunction with a set of context-specific priors, to filter the record if evidence for the allele comes primarily from inserts that align to one genomic strand (forward or reverse) only. However, with this method, it is not always possible to

detect a statistical signal of orientation bias in the presence of FFPE or 8-oxoG damage, particularly at the low alternative allele depths found in heavily damaged, low-coverage, FFPE samples. This may result in a large number of artefacts making it through the GATK orientation bias filtering stage[163].

In this chapter we highlight the presence of somatic and germline variants in the sequencing data of each PDAC cohort member. We implement supplementary filtering strategies, specifically targeting variant records that were deemed to be the result of FFPE or oxidative (8-oxoG) damage. Subsequently, we investigate whether the incidence of *KRAS* mutations among this PDAC cohort was significantly less than would be expected, considering the prevalence of *KRAS* in the broader PDAC patient population. Using methods described in Chapter 3, we estimate the average sensitivity of detecting somatic variants, and the sensitivity to detect KRAS G12D in particular for the specific method of sample preparation, target region, sequencing strategy and variant calling pipeline employed in creating the PDAC dataset. We use this information to reassess the incidence of *KRAS* mutations identified by Mutect2 in the PDAC dataset. In addition we reevaluate evidence for the presence of mutations at pileup loci associated with known *KRAS* driver mutations in cohort members where *KRAS* had not been identified by Mutect2 or where the relevant *KRAS* variant records had been removed from the analysis by variant filtering.

4.2 Methods

4.2.1 Additional orientation bias filtering strategies to remove FFPE/8-oxoG damage variants

To confirm the presence of a somatic burden in the variant caller output the filtered VCFs from each member of the cohort were subjected to an additional level of filtering to remove variant records that may have been as a result of FFPE or 8-oxoG damage. This was achieved by implementing an additional filtering method (Appendix A 5.2) to remove records from analysis that did not demonstrate evidence of absence of FFPE and 8-oxoG damage. Evidence of absence was defined as the presence of two or more reads that contained the alternate allele in aligned inserts from to both the forward and reverse genomic strand. The VAF spectrum post additional FFPE filtering was then reevaluated.

4.2.2 Estimation of average sensitivity of somatic variant detection

Several independent methods were employed to evaluate the average sensitivity of somatic variant detection across the PDAC target region and the significance of the detected incidence of pathogenic *KRAS* (24 out of 60 patients or 40%), in particular. The observed incidence of *KRAS* is not only influenced by its prevalence in the broader PDAC patient population (assumed to be 85%) but also by the sensitivity to detect it, which in turn depends on the sequencing and variant calling strategies employed when creating the dataset. A typical approach to estimating the sensitivity of somatic variant detection used in molecular diagnostics is based on the binomial distribution[431,537,443]. In this case, the sensitivity (s) of detecting a mutant allele, which has a true variant frequency (f) at a mean depth of coverage (d), where the minimum alternative allele depth required to call a somatic variant

is (m), is estimated as $s = 1 - pbinom(m, d, f)$, with $pbinom()$ representing the cumulative binomial probability function.

Employing empirical values derived from cohort sequencing and somatic variant data for the minimum alternative allele calling threshold and average depth of coverage, we applied a binomial model to estimate sensitivity as a function of allele frequency for this dataset. Subsequently, we used this model to estimate the sensitivity of somatic variant detection at 45% tumour purity, using it as a basis to assess the significance of the 40% incidence of pathogenic *KRAS* within this PDAC cohort. The binomial approach however presents a number of challenges, particularly due to the absence of consensus on establishing the minimum alternative allele depth required for calling a somatic variant, an essential parameter within this model. In contrast, instead of implementing a minimum alternative allele depth threshold, somatic variant callers typically consider base and mapping qualities at the pileup when assessing evidence for an alternative allele, and putative variants also undergo stringent filtering. These factors impact somatic variant detection sensitivity, yet none are represented in the binomial model used in its estimation. Mindful of these limitations, we compared the binomial estimate of *KRAS* detection sensitivity with other methods for estimating sensitivity.

As an alternative strategy to employing the binomial model for assessing the average sensitivity to detect somatic variants across the target region as a function of allele frequency within this cohort, we conducted a series of simulations at 26x, the mean depth of coverage in the St. James's PDAC cohort. These simulations incorporated mean fragment length (122 bp), read length (100 bp) and PDAC target region as well as a sequence error profile generated using 'art_profiler_illumina' based on sequencing data from the PDAC cohort. The allele frequency spectrum was divided into 200 semi-centile bins, and a uniform distribution of 50,000 somatic variants was created at loci randomly distributed across the target region in each bin. A phased, pre-tumour pair of BAM files with a combined total of 26x depth of coverage was created using the stochastic simulation framework, ART read simulator, and the 1000 Genomes germline VCF of donor HG00110 (a female of English and Scottish ancestry). These files were then used as a base for simulations to estimate the average variant detection sensitivity.

The simulations were conducted in groups of four semi-centiles of the VAF spectrum per simulation, with each simulation containing a uniformly distributed burden of 200,000 somatic variants. This process resulted in 50 virtual tumour BAM files, which were subsequently analysed with Mutect2 using default parameters. For each semi-centile, we recorded the percentage of the ground truth burden that passed (considered true somatic) or was filtered (deemed artefactual), along with its associated allele frequencies from the filtered VCF output as recorded by Mutect2. This data was used to represent the average detection sensitivity for a member of this cohort as a function of allele frequency (Figure 4.4). This plot was compared with a theoretical estimate of sensitivity of somatic variant detection based on the binomial distribution[431] and was also used as a basis to assess the significance of the 40% incidence of pathogenic *KRAS* within the PDAC cohort. The number of true germline and alignment artefacts, false positives misidentified as somatic during variant filtering, was also recorded.

4.2.3 Estimation of *KRAS* sensitivity of detection using cohort simulation

The first step in establishing a comprehensive simulation of the PDAC dataset involves generating a set of BAM files that represent each patient in the cohort. Simulations based on a randomly distributed burden within the target genomic region are essential for attaining a balanced assessment of the sensitivity in detecting somatic variants from a particular sequencing assay. However, in this instance, the sensitivity of detection at specific pathogenic *KRAS* loci in PDAC is of particular interest. The sensitivity of somatic variant detection varies based on the genomic location of a given variant. Additionally, the average depth of coverage, which varies significantly across this PDAC cohort, and DNA damage artefacts may also influence sensitivity. To compare the specific incidence of *KRAS* identified by Mutect2 in this PDAC cohort with the prevalence of *KRAS* in the broader PDAC patient population, with greater precision than would be obtained from previous detection sensitivity estimates, we generated 60 BAM files with varying depths of coverage 20x (1 BAM), 22x (8 BAMs), 24x (20 BAMs), 26x (16 BAMs), 28x (11 BAMs), 32x (3 BAMs), 36x (1 BAMs) using the stochastic simulation framework described in Chapter 3, ART read simulator and HG00110. Each BAM file represents an individual in the PDAC cohort. The BAM files serve as a base for spiking-in both somatic and artefactual distributions used in simulating this cohort. The selected depths of coverage were deliberately chosen to closely align (within a 2x margin) with those observed in the actual PDAC cohort. Furthermore, a read length of 100bp and a mean fragment length of 122bp were selected to mirror the sequencing strategy used in the St. James's PDAC cohort. A custom ART read quality profile was also derived from the PDAC dataset's BAM files and employed in the generation of BAM files for this cohort simulation.

Following the generation of the set of base BAM files, an artefactual distribution mirroring that observed within the PDAC cohort was simulated. Patient VCF data was subsetted to extract SBS records where evidence for the alternate allele was supported exclusively by five or more reads originating from inserts aligned to the same genomic strand. These records were assumed to be due to orientation bias, possibly induced by factors such as FFPE or oxidative (8-oxoG) damage. Subsequently, the data was further stratified based on the strand on which the damage was observed (forward or reverse), and annotated with the corresponding tri-allelic context for each record (using bedtools getfasta). The proportion of the total burden, specific to each tri-nucleotide sequence context, for the target region under investigation, was recorded. This empirical DNA damage profile was then used to simulate a distinct set of orientation bias artefacts within the target region of interest for each cohort member[163]. The true artefactual burden that represents the cohort average was estimated through simulation to be approximately $5.5\text{x}10^6$ artefacts, distributed within a region of the allele frequency spectrum ranging between 0.005 and 0.01. This artefactual burden was spiked-into each of the 60 base BAM files of the PDAC simulation group. To preserve the required orientation, the simulated pre-tumour BAMs were each split into two separate files, one with reads from inserts that aligned to the forward genomic strand and the other containing the remaining reads. The target artefactual burden, totalling $5.5\text{x}10^6$ DNA damage artefacts, was then spiked-in as required to the forward and reverse alignments using the stochastic simulation framework[163] after which the alignments were remerged.

With the artefactual burden in place, the somatic variant KRAS G12D was then introduced into all pre-tumour BAMs. The spike-in process was performed within alternate haplotypes in each BAM. Following this, the pairs of pre-tumour haplotype BAMs were merged and realigned against the hg38 reference. Somatic variant calling was performed by applying Mutect2 in tumour-only mode to the simulated tumour BAM files, and the incidence of detected KRAS G12D was evaluated. Subsequently, the DNA damage trinucleotide profile and reference and alternative allele depths were compared between the simulation and PDAC patient cohort.

4.2.4 PDAC patient cohort, *KRAS* incidence and further analysis

The incidence of known *KRAS* driver mutations in the St. James's PDAC cohort was determined from Mutect2 VCF output. Following this, variant records that had been removed by GATK filtering were re-evaluated. The reference and alternative allele depths at recognized *KRAS* hotspot loci were assessed to identify potential driver mutations overlooked by Mutect2. Other recurrently observed mutations among patients in this cohort, in genes associated with cancer, were also noted and examined.

4.3 Results

4.3.1 The somatic allele frequency spectrum and potential driver variants in the PDAC patient cohort

The incidence of known *KRAS* driver mutations in the St. James's PDAC cohort was determined from variant records annotated as 'PASS' in Mutect2 VCF output (Figure 4.2). In total, pathogenic *KRAS* variants were identified in 24 out of the 60 patients in the cohort, corresponding to an incidence of 40%. A number of other *KRAS* VCF records were filtered due to insufficient evidence of the alternate allele or as a result of detecting orientation bias at the variant locus (Table 4.1, 4.2 and Figure 4.3). The implementation of an additional orientation bias filtering strategy, specifically designed to exclude variant records that did not exhibit a clear absence of FFPE or oxidative (8-oxoG) damage, led to a substantial reduction in the somatic burden identified. The total count of somatic mutations within the cohort decreased by 98%, from 1,462,363 to 25,113. The additional variant filtering also had a significant influence on the allele frequency spectrum recovered from the data, revealing distinct patterns of both germline heterozygous and homozygous distributions, while also highlighting a discernible low frequency somatic burden in the majority of samples (Appendix B 5.2). Unfortunately, this enhancement in specificity was accompanied by a notable decrease in variant detection sensitivity, especially in the region of the frequency spectrum expected to contain somatic variants (estimated to be centred at 0.225 in this cohort). With additional variant filtering the number of individuals in whom pathogenic *KRAS* mutations were detected fell from 24 (40%) to just 5 (13%).

In addition to *KRAS*, there were recurrent mutations at several other cancer associated loci recorded in patient VCF files across the cohort. Recurrent mutations in mucin genes *MUC3A* and *MUC5AC* are of particular interest. Mucins, a family of multifunctional glycoproteins, are believed to fulfill various roles in pancreatic biological processes, primarily contributing to the protection, hydration, and lubrication of

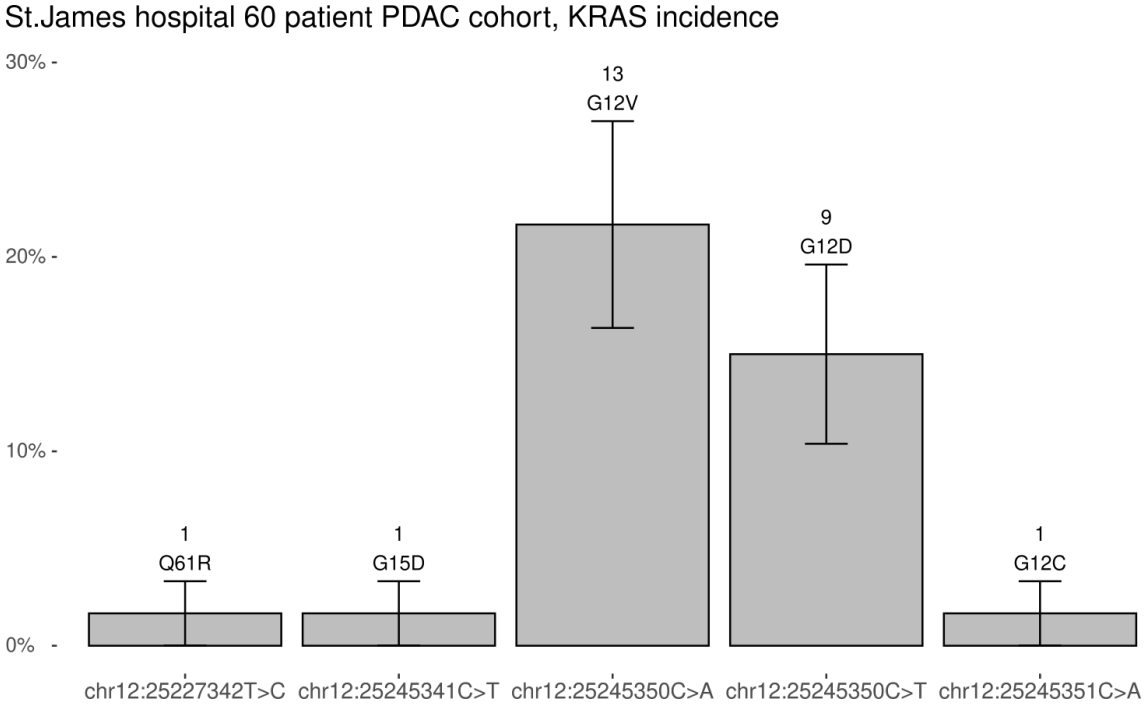

Figure 4.2: Occurrence of five pathogenic *KRAS* variants annotated as 'PASS' by Mutect2 in St. James's Hospital PDAC cohort. The error bars indicate the standard error of the proportions.

epithelial tissues. Mucins have a long association with various cancer types, particularly pancreatic cancer, and have been used as both diagnostic biomarkers and therapeutic targets[538]. It is common, in tumour only mode, for Mutect2 to incorrectly classify a substantial number of less common germline variants that are not present in the germline resource file as somatic. In an attempt to minimise this issue, recurring variants in mucin genes were cross-checked against an additional germline database (1000 genomes), leading to the removal of a further 11 suspected germline variants from the analysis. *MUC5AC* has been associated with carcinogenesis and an aggressive, chemoresistant pancreatic cancer phenotype.[539,540,541,542]. In total, recurrent mutations in *MUC5AC* were found in 15 patients. The co-occurrence of *MUC5AC* P1569L (chr11:1182851C>T) and T1570T (chr11:1182855G>T) was identified in four patients (P24A, P31A, P59A, P5A). *MUC5AC* p.P1919L (chr11:1183901C>T) was present in six patients (P16A, P29A, P3A, P44A, P50A, P9A). Additionally, *MUC5AC* p.P1919L (chr11:1188915C>T) was found in seven patients (P10A, P21A, P32A, P50A, P59A, P5A, P7A), and *MUC5AC* p.A4396A (chr11:1191333C>A) in five patients (P21A, P23A, P32A, P3A, P7A). Recurring mutations in *MUC3A*, a gene also associated with pancreatic cancer[543], were identified in 35 members of the PDAC cohort (Appendix C 5.2).

4.3.2 Sensitivity of somatic variant detection in PDAC

The median alternative allele depth of variants annotated as 'PASS' by Mutect2 in the PDAC cohort was derived from patient VCF files. This value (2) served as a working estimate for the alternate allele depth calling threshold when applying the

Patient	Variant	Protein change	Ref. allele depth	Alt. allele depth
P2A	chr12:25227342T>C	Q61R	11	2
P3A	chr12:25245350C>A	G12V	26	7
P6A	chr12:25245350C>T	G12D	19	4
P9A	chr12:25245350C>T	G12D	18	2
P15A	chr12:25245350C>T	G12D	19	8
P17A	chr12:25245350C>T	G12D	22	3
P18A	chr12:25245350C>A	G12V	14	4
P20A	chr12:25245350C>T	G12D	19	5
P22A	chr12:25245350C>A	G12V	23	7
P23A	chr12:25245350C>A	G12V	23	6
P27A	chr12:25245350C>A	G12V	19	6
P30A	chr12:25245341C>T	G15D	9	1
P30A	chr12:25245350C>A	G12V	11	1
P31A	chr12:25245350C>T	G12D	20	5
P32A	chr12:25245350C>A	G12V	37	7
P34A	chr12:25245350C>A	G12V	12	3
P36A	chr12:25245350C>A	G12V	10	3
P38A	chr12:25245350C>A	G12V	1	1
P42A	chr12:25245350C>T	G12D	10	2
P43A	chr12:25245350C>A	G12V	19	16
P48A	chr12:25245350C>A	G12V	26	10
P49A	chr12:25245350C>A	G12V	27	4
P51A	chr12:25245350C>T	G12D	20	5
P54A	chr12:25245351C>A	G12C	18	4
P60A	chr12:25245350C>T	G12D	9	2

Table 4.1: Pathogenic variants identified in *KRAS* using Mutect2. In total, pathogenic *KRAS* variants were identified in 24 members of the PDAC cohort, all of which were confirmed to be missense mutations. Of note, two pathogenic *KRAS* mutations, G12V and G15D, were detected in patient P30A, while patient P54A, with KRAS G12C, would currently be considered a candidate for the recently developed KRAS G12C covalent inhibitor therapy. The depth of coverage at the variant locus and depth of coverage of the alternative allele were evaluated using SAMtools with overlap handling enabled (by default) and minimum thresholds for base and mapping quality set at 20. The mean depth of coverage at the variant locus was 17.68 (the Mutect2-adjusted depth values, compiled from informative reads after local reassembly, are accessible from the corresponding VCF files).

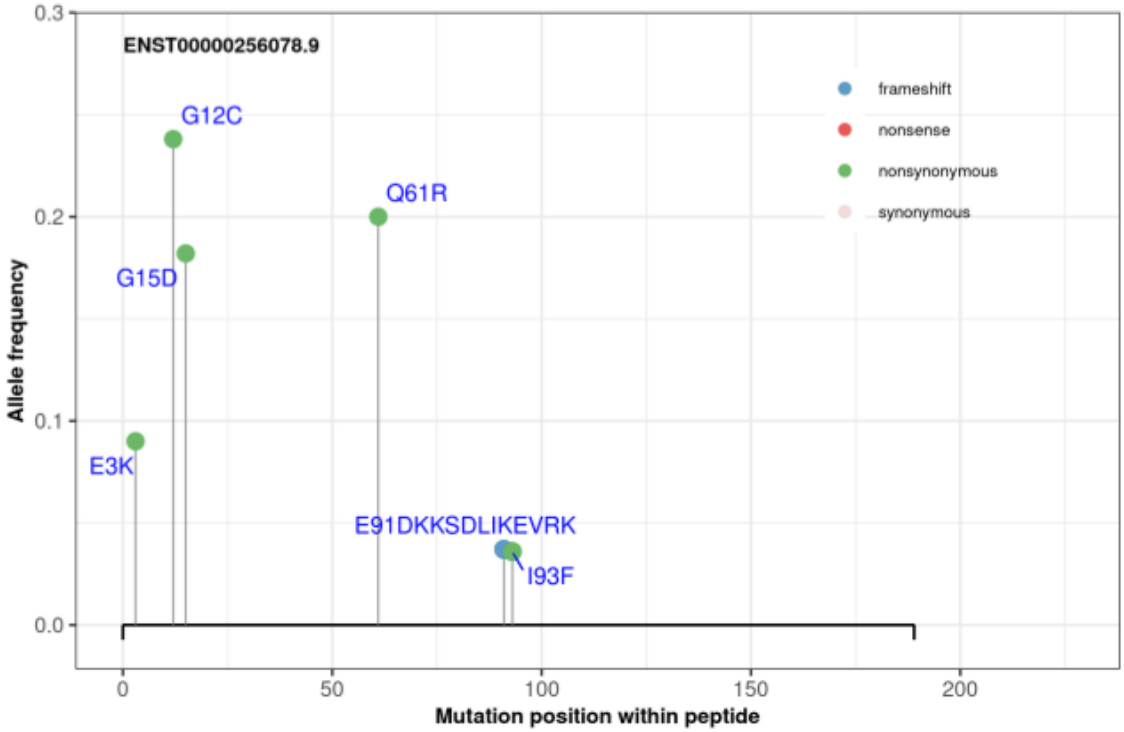

Figure 4.3: Selection of protein plots from vcfView for *KRAS* variant records annotated as 'PASS' in patients P54A (G12C), P30A (G15D), P2A (Q61R), and filtered variants in P61A (E3K, filtered as orientation bias) and P12A (I93F and E91 insertion, both filtered as weak evidence).

binomial model to assess the sensitivity of somatic variant detection in the PDAC dataset. In this simplified model, it is assumed that all variants with an alternative allele depth greater than two will be passed by the caller. The median insert length of all properly paired alignments in the PDAC cohort was determined to be 122 bp, with a read length of 100 bp, using SAMtools. FFPE induces DNA damage that can necessitate the selection of shorter fragment lengths during library preparation. In this instance, the combination of a short insert length and a relatively long read length resulted in a very high proportion of overlapping read pairs in the sequencing data. This required the adjustment of the effective depth of coverage estimation to 17x when calculating the sensitivity of detection using the binomial model. Consequently, the projected detection sensitivity using this model was 0.77, implying a significantly lower incidence of pathogenic *KRAS* (40%, p=5.0x10^{-5}, from the binomial test) in the PDAC cohort than anticipated, given its prevalence in the general PDAC population.

The sensitivity of somatic variant detection and the significance of the detected incidence of pathogenic *KRAS*, both determined using the binomial approach, were reassessed by comparing them against an empirically derived sensitivity profile. This profile was generated through simulations using an average depth of coverage, average fragment and read length, and sequence error profile that matched that observed in the PDAC cohort (Figure 4.4). The empirically derived detection sensitivity exhibited a notable decline compared to the value anticipated by the binomial model at allele frequencies greater than 0.2. Notably, the 95% Limit of Detection (LoD)

Patient	Filter reason	Variant type	Variant	Protein change	Ref. allele depth	Alt. allele depth
P4A	orientation	MISSENSE	chr12:25245350C>T	G12D	14	2
P5A	weak_evidence	MISSENSE	chr12:25245350C>T	G12D	13	2
P12A	weak_evidence	IN_FRAME_INS	chr12:25227251_25227252ins	E91DKKSDLIKEVRK	45	0
P40A	weak_evidence	MISSENSE	chr12:25245350C>A	G12V	15	2
P50A	weak_evidence	MISSENSE	chr12:25245351C>G	G12R	22	2
P61A	orientation	MISSENSE	chr12:25245378C>T	E3K	19	1

Table 4.2: Members of the PDAC cohort for whom all pathogenic *KRAS* records identified by Mutect2 failed variant filtration. The depth of coverage at the variant locus and depth of coverage of the alternative allele were evaluated using SAMtools with overlap handling enabled (by default) and minimum thresholds for base and mapping quality set at 20.

projected by the binomial model at an allele frequency of 0.33 was not attained at any allele frequency in the empirical simulations. The predicted sensitivity of detection at the expected somatic clonal centre (0.225 at a tumour purity of 45%) was 0.72, a small decrease on the previous estimate. The incidence of pathogenic mutations in *KRAS* observed in the cohort (40%) was significantly lower than expected, given this detection sensitivity ($p = 0.0007$, from the binomial test). In addition, a number of false positives, including approximately 2000 germline artefacts and 60 alignment artefacts, were incorrectly identified as somatic during variant filtering.

Simulations based on a randomly distributed burden within the target genomic region are key to obtaining a balanced assessment of the sensitivity to detect somatic variants from a particular sequencing assay. However, in this PDAC dataset, assessing the sensitivity to detect a set of key pathogenic *KRAS* variants is of particular interest. Out of the 60 individual simulations representing each cohort member, matching the depth of coverage, error and DNA damage profile (Figure 4.5), and average tumour purity in the PDAC cohort, Mutect2 successfully detected (annotated as 'PASS') the KRAS G12V variant in 53 simulations. In the seven simulations where *KRAS* variants were not detected, four were attributed to low coverage at the variant locus, resulting in no reads containing the alternate allele.

In two of the remaining simulations, reads containing the alternative allele were of low base or mapping quality and were discarded by Mutect2 or lost due to incorrect alignment. In the final simulation, Mutect2 generated a VCF record for the *KRAS* allele; however, this was subsequently filtered as weak evidence. This result indicates a sensitivity of detection at the KRAS G12V locus of 88% at 45% tumour purity, also implying a significantly lower incidence of pathogenic *KRAS* in the PDAC cohort than expected (40%, p=1.85e-06, from the binomial test), given the 85% *KRAS* prevalence in the general PDAC population.

The mean total SBS burden across the simulated cohort was 18,990 variants annotated as PASS by Mutect2, compared to 19,423 per patient in the PDAC cohort. Approximately 90% of the simulated burden consisted of false positives due to DNA damage artefacts, with the remainder caused by germline artefacts and a small number resulting from alignment and sequence errors.

We then compared the reference and alternative allele depths at the KRAS G12 locus between the simulated data and the actual sequencing data from the PDAC cohort. Both the simulated and real datasets exhibited the same average total depth of coverage of 25x across the mutant *KRAS* locus. The median base quality across all reads at the pileup loci of KRAS G12 was slightly higher in the simulated data

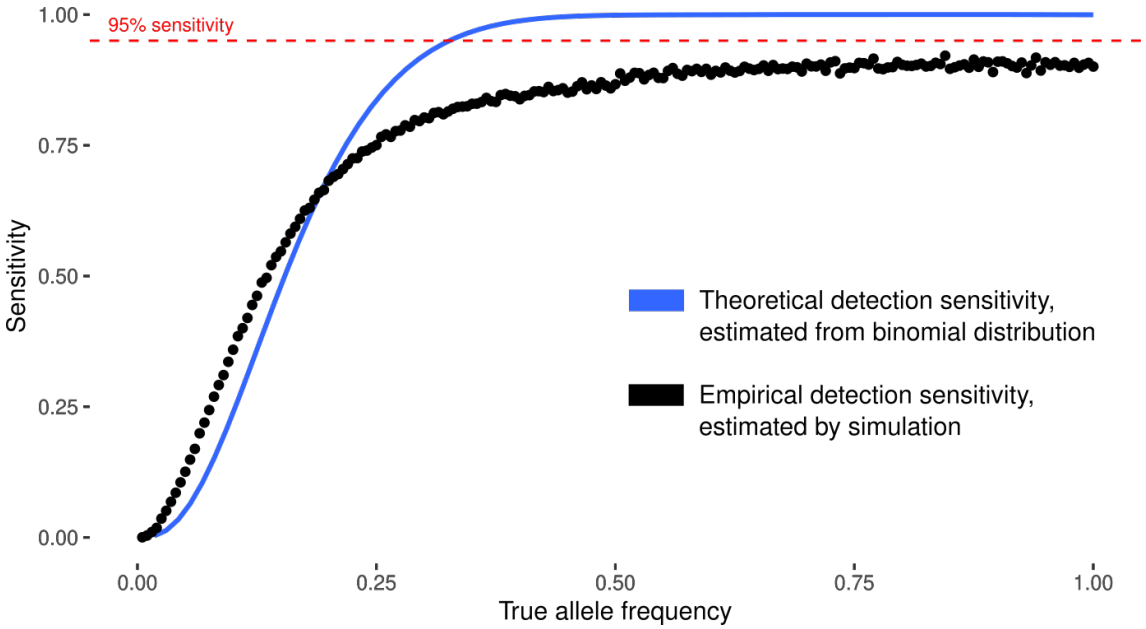

Figure 4.4: Sensitivity to detect somatic variants as a function of allele frequency. The theoretical sensitivity plot is estimated using the binomial distribution at a 17x depth of coverage, accounting for read pair overlap, and with a minimum alternate allele depth required to call a somatic variant set at 3 or more reads. The empirical plot is derived from simulations at 26x, representing the average depth of coverage in St. James's PDAC cohort.

(37 vs 32). Interestingly, evidence of pathogenic *KRAS* alleles in the PDAC cohort was detected at low alternate allele depth in reads at the variant pile-up in 15 out of 30 patients, where Mutect2 output did not identify any variant record (Table 4.3). To assess whether the evidence of pathogenic *KRAS* alleles in these patients might be attributed to sequencing or other artefacts, we compared the total number of reads containing non-reference bases at pathogenic *KRAS* loci chr12:25245350 and chr12:25245351 against randomly selected loci close to the pathogenic *KRAS* variant, specifically chr12:25251748 and chr12:25251749, across all 30 PDAC patients. These adjacent loci share the same sequence context and approximate coverage as the pathogenic *KRAS* loci. In total, only one read containing a non-reference base was detected at these adjacent loci, indicating a significant association between reads containing evidence of an alternate allele and pathogenic KRAS G12 related loci (p=1.455e-08, from Fisher's exact test). Mutant *KRAS* records in the PDAC cohort that had failed variant filtration were also re-evaluated (Table 4.2). In total, six members of the PDAC cohort had *KRAS* records identified by Mutect2 that failed variant filtration. Among these, four were filtered due to weak evidence. The log_{10} likelihood of the existence of pathogenic *KRAS* in the tumour, as noted in the Mutect2 VCF output for each of the four patients, P5A, P12A, P40A, and P50A, was 3.43, 3.41, 3.05, and 3.78, respectively, implying a probability of pathogenic *KRAS* greater than 0.999 in each patient. Two additional *KRAS* records, G12V and E3K in P4A and P61A, were filtered due to orientation bias.

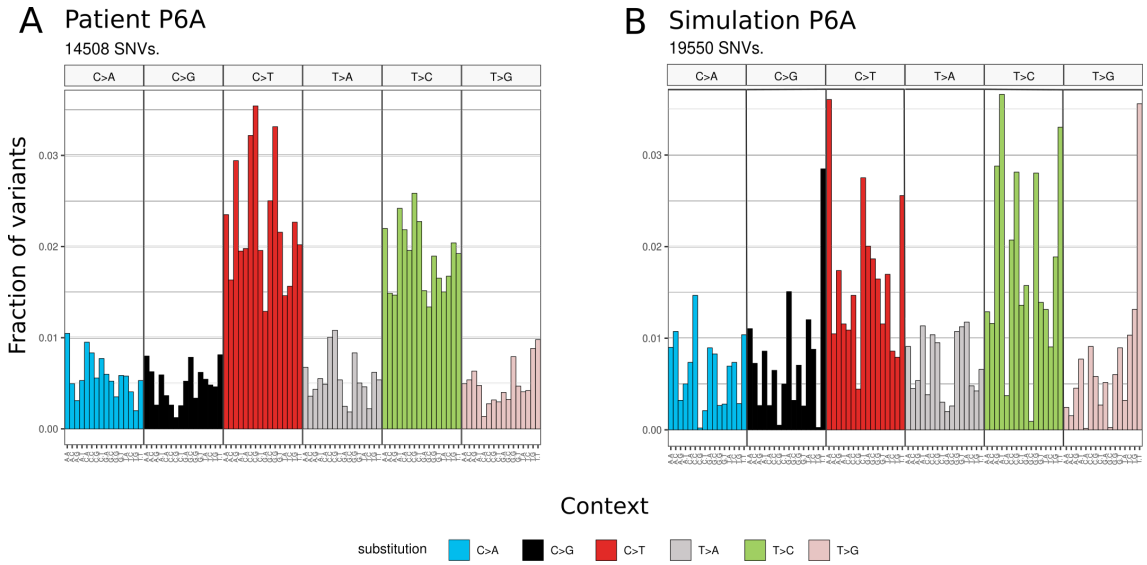

Figure 4.5: Mutational profile of simulated and real data representing one of the patients in the PDAC cohort. **A:** Actual mutational profile observed from variant records annotated as 'PASS' by Mutect2 in patient P6A with a depth of coverage of 26x. **B:** The mutational profile observed in variant records annotated as 'PASS' by Mutect2 in simulated data, with a depth of coverage of 26x. The mutational profile chosen for the simulated data was derived from all SBS records in the PDAC dataset where evidence for the alternate allele was supported exclusively by five or more reads originating from inserts aligned to the same genomic strand. The simulated burden represents an estimate of the average artefactual burden of the cohort.

4.4 Discussion

Highly degraded DNA from FFPE samples poses significant challenges for somatic variant recovery. In this dataset, severely fragmented FFPE-DNA has led to reduced library insert sizes and a lower effective depth of coverage. Despite applying specific GATK filtering intended for orientation bias artefact removal, a substantial artefactual burden persisted in the somatic variant caller output, complicating the identification of biological variants of interest. The sequencing strategy, involving low-depth tumour-only sequencing data, further complicates efforts to distinguish true somatic variants from artefacts. However, while the sequencing data in this dataset is less than ideal, it remains usable. By analysing the allele frequency spectrum and implementing additional orientation bias filtering, we were able to validate the existence of both germline and somatic mutations within the data, some of which may be potentially relevant to cancer.

The advent of high-throughput sequencing has enabled researchers to establish the prevalence of cancer drivers in many cancer types[544]. Mutations detected (ie., annotated as 'PASS') by somatic variant callers such as Mutect2 are routinely employed to validate expected instances of known drivers within a cohort and to deduce additional variants contributing to the cancer. However, considering the high prevalence and penetrance of pathogenic *KRAS* in the general PDAC population, and the challenges posed by the low depth of coverage and a substantial proportion

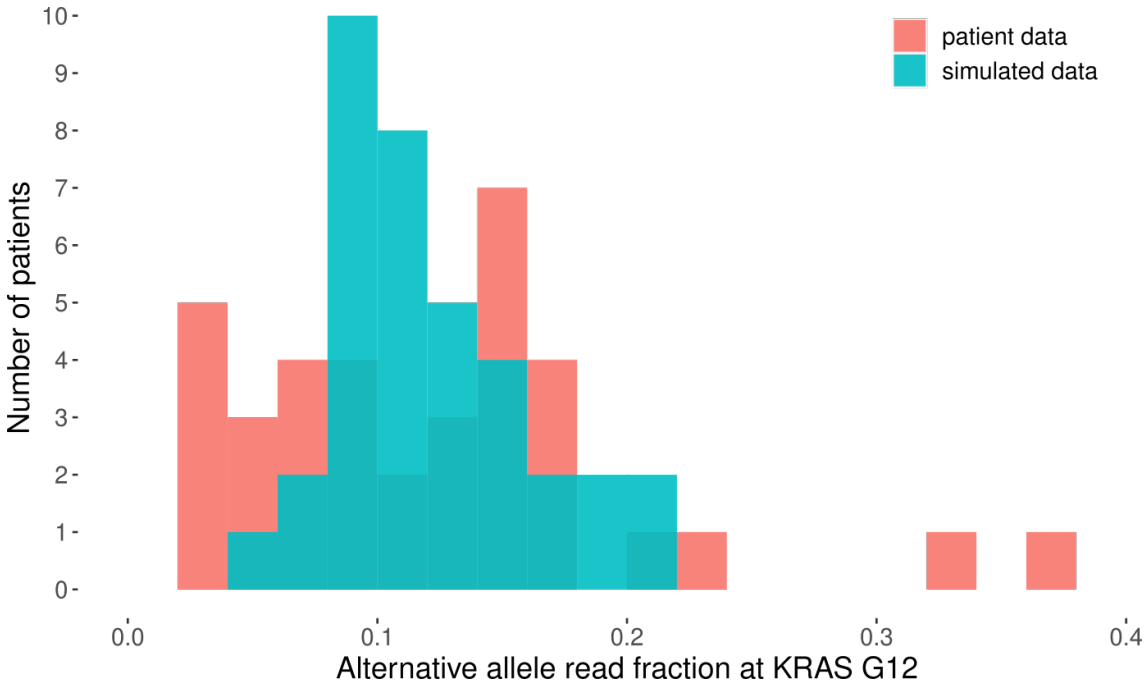

Figure 4.6: Comparison of alternate allele read fractions at KRAS G12 related loci between patient sequencing data, where reads containing alternative alleles were identified, and the corresponding data from the cohort simulation. While the mean read fraction remains consistent between both simulated and real datasets at 0.12, the increased variability in read fractions within the PDAC patient data suggests broader than anticipated variation from the stated 45% mean tumour purity. Significant variation in tumour purity across the cohort could result in a reduction of detection sensitivity, as some somatic variants may fall below the minimum detectable frequency.

of overlapping read pairs observed in sequencing data from this cohort, we advise against determining the incidence of *KRAS* solely based on the number of patients where Mutect2 detected a pathogenic *KRAS* variant. To address this question using the available data, a better approach would be to assess how many members of this cohort demonstrate the absence of pathogenic *KRAS*. If we accept, for example, the presence of any read containing a pathogenic *KRAS* allele as evidence for the presence of a *KRAS* variant, the resulting incidence would be 45 of 60 patients (75%) which is no longer significantly below the expected prevalence (85%) of pathogenic mutations in *KRAS* in PDAC ($p = 0.32$) . Uncertainty regarding the extent of variance in tumour purity across the PDAC cohort is also a matter of concern. Cohort simulations confirm substantial variation in alternative allele read fractions at the KRAS G12 related loci in the PDAC patient data compared to what would be expected from a fixed 45% purity (Figure 4.6). This discrepancy may be a result of variability in tumour purity across samples. Significant variation in tumour purity across the cohort would result in reduced detection sensitivity in several patient samples, as some somatic variants may fall below the minimum allele frequency detectable by Mutect2. This variation may also explain the lower than anticipated incidence of *KRAS* observed in the PDAC dataset.

Analysis of Mutect2 output in St. James's PDAC cohort emphasises the im-

Patient	Variant	Protein change	Ref. allele depth	Alt. allele depth
P1A	chr12:25245350C>A	G12V	14	1
P8A	chr12:25245350C>T	G12D	19	2
P13A	chr12:25245350C>A	G12V	12	1
P14A	chr12:25245350C>T	G12D	27	1
P26A	chr12:25245350C>T	G12D	15	1
P28A	chr12:25245351C>G	G12R	15	1
P29A	chr12:25245350C>T	G12D	17	2
P33A	chr12:25245351C>G	G12R	28	1
P37A	chr12:25245350C>T	G12D	13	1
P39A	chr12:25245350C>T	G12D	19	1
P41A	chr12:25245350C>T	G12D	26	1
P47A	chr12:25245350C>A	G12V	34	1
P52A	chr12:25245351C>G	G12R	14	1
P56A	chr12:25245350C>T	G12D	11	1
P57A	chr12:25245350C>A	G12V	5	1

Table 4.3: Evidence of pathogenic *KRAS* variants in reads at the variant locus pile up was detected in 15 of the 30 patients where Mutect2 output did not identify any *KRAS* variant record. The mean depth of coverage at the variant locus was 17.93. The depth of coverage at the variant locus and depth of coverage of the alternative allele were evaluated using SAMtools with overlap handling enabled (by default) and minimum thresholds for base and mapping quality set at 20.

portance of reassessing filtered VCF records at key loci crucial for therapeutic decision making. In particular, records for four patients exhibited strong evidence of pathogenic *KRAS* variants (p>0.999), despite being filtered by GATK as weak evidence. Considering the overall prevalence of *KRAS* mutations in the general PDAC patient population, we may assume a high prior probability of *KRAS* association with the patient's cancer. This assumption should prompt a reevaluation of these records from the perspective of their potential to inform clinical decision making, particularly in light of recent advances in KRAS inhibitor therapies. Despite evidence of *KRAS* mutations, Mutect2 can not accept these variant records as somatic. This is because accepting them based on that level of probability across a typical Whole Exome Sequencing (WEX) or Whole Genome Sequencing (WGS) target would result in a large increase in the number of false positives in its output. Mutect2 is unable to incorporate an understanding of biological context into its calling algorithm while recovering somatic variants. However, this does not prevent clinicians armed with this knowledge from doing so.

The simulations conducted to evaluate the sensitivity of detection across the target area revealed a large number (approximately 40) of false positives that originated from alignment issues attributed to germline SNPs and indels. While these variants are unlikely to be found in common germline databases, they would typically be removed from analysis in a tumour-normal calling pipeline during the variant filtering stage, as the allele in question would also be detected in the matched normal sample. However, in the absence of a matched normal in the St. James's PDAC dataset, these artefacts are incorrectly annotated as 'PASS'. Though the number of false

positives generated in this manner is relatively small compared to the substantial burden of FFPE/8-oxoG artefacts, these alignment artefacts are noteworthy. They hold the potential for co-occurrence within a cohort of individuals from the same population, possibly leading to their incorrect identification as contributors to cancer.The absence of a comprehensive database representing germline variation within the Irish population hinders the identification of these artefacts. Further analysis, particularly examining secondary alignments of the variant pileup at these loci, is necessary to clarify the source of variants in mucin genes and other mutations that recur at the same genomic site across a significant proportion of tumours within this cohort.

4.5 Conclusion

For over a century, formalin fixation and paraffin embedding (FFPE) has remained the preferred method for tissue storage and biobanking, preserving patients' tissues worldwide. With an estimated repository of over 50 million clinically annotated FFPE specimens available globally for somatic variant analysis, FFPE represents a potentially invaluable research asset[545,162]. The retrieval of sequence data from FFPE samples dating back almost a century has played a key role in public health research[546], demonstrating its potential in other areas of research. However, along with this potential, sequencing FFPE-treated samples presents significant challenges. One of these challenges involves the need for innovative bioinformatics approaches to address the DNA damage artefacts inherent in FFPE-sequenced data. The severity of this damage is often such that the resulting sequencing data is considered unsuitable for variant analysis.[546]. We have demonstrated methods for recovering somatic variant information from sequence data taken from heavily damaged FFPE samples. These methods have enabled us to confirm the existence of somatic and germline distributions in the sequence data and detect the occurrence of crucial pathogenic driver alleles, even at average depths of coverage well below the typically accepted minimum for FFPE samples[537,547]. To achieve this, we employed the stochastic simulation framework[163] as described in Chapter 3, allowing us to computationally reconstruct the dataset with detailed simulations for each member of the PDAC cohort. These simulations considered sampling variations in alternative allele depth and the individual sequencing strategy employed in creating each patient's sequence data. They allowed us to validate important assumptions about LoD, average tumour purity and the incidence of pathogenic *KRAS* within the PDAC dataset. The analysis of variant calling data from patients across the PDAC cohort also revealed mutational hotspots in several other cancer-associated genes. However, a more comprehensive analysis is necessary to confirm these variants. This is due to the significant number of germline, and to a lesser extent, alignment false positives that the simulations have predicted to be among the somatic variants identified by the PDAC tumour-only variant calling pipeline. Conducting additional simulations that account for germline variation within the Irish population would be beneficial in addressing this concern. The methods outlined here, developed using the stochastic simulation framework, can be applied to other heavily damaged FFPE datasets to recover useful information from them.

5 Conclusions

5.1 Overview

Accuracy in cancer molecular profiling is paramount in treatment and research. The failure to detect clinically actionable somatic variants can profoundly impact patient care, leading to missed opportunities for targeted therapy, inaccurate prognosis, and suboptimal treatment selection and planning. Additionally, it has the potential to misguide scientific research, resulting in erroneous assumptions about the significance of disease drivers within a cohort[548]. Incorrectly identified somatic variants may introduce noise into research studies, obscuring the aetiology of the cancer. In oncology, incorrectly identifying a clinically actionable variant may expose a patient to unnecessary risks and side effects of treatments they do not need. The aim of this thesis was to develop a set of computational methods to identify and explain the sources of error in cancer somatic mutation data and to apply them in the validation and analysis of real patient data. The solution was developed in two stages. Firstly, software was created to facilitate the re-analysis of somatic variant records previously excluded by variant filtering. Secondly, a computational framework was established to enable comprehensive and realistic simulation of tumour genomic sequencing data. This framework facilitated the validation of hypotheses regarding specific sources of error associated with the individual sequencing strategy used in patient somatic mutation analysis.

Chapter 2 of this thesis focused on the development and application of vcfView, an interactive Rshiny tool designed to support the evaluation of somatic mutation calls from cancer sequencing data. In somatic variant calling, the majority of variant records are commonly excluded from analysis through variant filtering. For instance, in the TCGA lung cancer dataset, only 12% of variant records identified by Mutect2 were annotated as 'PASS' and thus included in the analysis. Despite the substantial influence that variant filtering exerts on the somatic mutations retrieved from clinical studies, there has been limited research on its impact on false negatives. Initially, the absence of suitable analysis software substantially constrained the scope and pace of our research aimed at quantifying this impact. Both sequencing artefacts and genuine somatic variants leave distinct patterns on the allele frequency spectrum, offering insights into their true origin. Considerations of mutational signatures and effects on proteins are also of significant interest when re-evaluating variant records. However, tools for assessing such impacts are cumbersome to configure and lack the capability to easily review the effects of different analytical choices when subsetting somatic mutation data.

We addressed these challenges by developing vcfView, an interactive data visualisation application designed for the exploratory analysis of somatic mutation data. This application integrates comprehensive variant annotation, mutational signature, variant filtering, and allele frequency spectrum analysis. Through its graphical user interface, vcfView enables users to subset VCF mutational data and observe, in real-time, the impact of these selections on the analysis output. This significantly streamlines the investigation into how different filtering options and configurations impact the mutations extracted from VCF data. For instance, users can easily select a region of interest in the allele frequency spectrum from the display, enabling real-time examination of variant filters, mutational signatures, and protein impacts

within that specific region. Furthermore, this analysis dynamically updates while adjusting variant filtering thresholds, parameters, and selections.

Using vcfView, we re-examined the somatic variant caller output from the TCGA Acute Myeloid Leukaemia (AML) project. While exploring various subsetting options across filtered TCGA-LAML VCF records, we observed the substantial removal of records due to the presence of the alternative allele in the normal sample, despite its extremely low alternate allele frequency in that sample. A large number of these filtered VCF records relate to high-impact variants associated with cancer. Comparing TCGA-AML against other TCGA datasets, we noticed this occurrence was far more pronounced in TCGA-AML. In somatic variant calling a matched normal blood sample is typically used as a control to rule out germline and other artefacts of the sequencing process. However in blood cancers like AML this is generally replaced with a matched normal a skin sample. Clonal populations in normal skin samples, which frequently harbour well-known cancer driver mutations, have been identified in various sources in the literature. Moreover, 3% to 5% of all nucleated cells in the epidermis are myeloid derived. However, the potential of such issues to elevate false negatives in clinical diagnosis has not been addressed. We confirmed that AML drivers were significantly enriched among the set of putative Tumour in Normal (TiN) records filtered by Mutect2. We therefore conclude that the biological contexts of variants filtered solely based on their presence at very low allele frequencies in the normal sample should be carefully considered before excluding them from clinical decision making. The exploratory analysis conducted using vcfView played a pivotal role in this research.

Although filtered records provide insight into errors in mutation calls, they offer an incomplete picture. They do not identify sequencing artefacts passed by the caller and provide limited insight into false negatives, only considering filtered variants in the caller VCF output. Many loci containing putative somatic variants are routinely excluded during variant caller preprocessing to save computational resources, potentially leading to decreased sensitivity in detection. Other potential errors in somatic mutation data, such as inaccuracies in the estimation of alternative allele frequency, cannot be addressed solely by analysing filtered variants. To thoroughly investigate these issues, it is essential to have comprehensive reference sequencing data that includes a 'ground truth' set of somatic mutations, indicating the locations and sources of all loci containing non-reference bases within the dataset. However, initial efforts to overcome these challenges using existing methods for ground truth sequencing data creation proved unsuccessful. The true sets for data created using these methods are incomplete as they typically contain other somatic variants as well as sequence and alignment errors at unknown locations. In many instances, we could not determine whether an issue stemmed from an error on the part of the somatic variant caller or, in fact, originated from inaccuracies in the truth set used for validation. These methods also fail to take into account stochastic aspects of the sequencing process, leading to unrealistic simulation of the somatic variant alternate allele depth.

Chapter 3 of this thesis detailed how challenges associated with somatic mutation truth sets were addressed by creating a simulation framework to generate comprehensive and realistic tumour sequencing data. In contrast, truth sets created using this framework not only provide definitive identification of variant caller errors but also, critically, enable us to explain why the caller mistakenly made the incorrect

call, thereby eliminating ambiguity associated with existing simulation methods. This information has the potential to predict and avoid scenarios where false positives arise, improve caller detection algorithms, and predict sensitivity across a range of sequencing strategies, including coverage and target selection. This simulation framework also accounts for the randomness in the number of reads that contain the non-reference allele at somatic mutation sites, thereby providing an effective means of assessing the impact of sequencing strategy on the variant caller estimation of somatic variant allele frequency, a capability not available with other simulation methods.

In Chapter 3, we applied this novel simulation framework to evaluate the performance of the GATK4 Mutect2 mutation caller across a range of sequencing depths, somatic mutational frequencies, and a diverse set of sequencing artefacts. We confirmed the GATK4 pipeline's minimal type I error rate when applied to high quality sequencing data. Furthermore, we quantified the impact of various GATK variant filters on type II errors, specifically highlighting the potential influence of the 'allele in normal' quality filter on the sensitivity of detecting clinically actionable variants. We identified bias in Mutect2's estimation of the mutant allele frequency which decreases with increasing sequencing depth. In addition, we used the simulation framework to determine the observed frequency spectrum that results when mutations from a theoretical spectrum corresponding to a model of tumour evolution are called from cancer sequencing data. The results of applying Mutect2 to heavily FFPE and 8-oxoG damaged simulated sequencing data are also of particular interest. Despite the GATK4 orientation bias filter successfully removing the vast majority of this artificial burden, a very small percentage (1%) was incorrectly passed by the variant caller. Pre-analytical factors, such as FFPE treatment, may cause significant DNA damage, resulting in a very large burden of orientation bias artefacts in the sequencing data output. These simulations suggest that this damage could lead to a substantial number of type I errors.

In Chapter 4, we applied the simulation tools and analytical techniques we developed previously to examine an unpublished pancreatic dataset comprising 60 patients. The cohort included individuals diagnosed with early-onset and aggressive pancreatic ductal adenocarcinoma. Researchers at St. James's Hospital and Trinity College Dublin commissioned the study to validate the role of KRAS and explore other potential drivers of PDAC within this cohort. However, the somatic mutation data were obtained from low-depth sequencing of heavily DNA-damaged FFPE samples without the availability of normal control samples, posing significant challenges in the analysis. Despite the application of GATK4 orientation bias filtering, the initial analysis of all patients was complicated by an extremely high burden detected by Mutect2, suspected to be of artefactual origin. The application of bespoke orientation bias filtering and examination of the frequency spectrum enabled us to conclude that this burden was composed primarily of orientation bias artefacts that had made it past GATK4 filtering, obscuring the germline and somatic burdens. To avoid a decrease in detection sensitivity, the additional filtering was removed, and two orthogonal methods, using the binomial model and the simulation framework, were applied to estimate the sensitivity of somatic variant detection in the dataset. Notably, the binomial approach predicted a sensitivity that was significantly higher than what was empirically observed in the simulation, suggesting it would be inappropriate to use the binomial model in this case.

The incidence of pathogenic *KRAS*, detected by Mutect2 within the PDAC dataset, was also reassessed using the simulation framework in Chapter 4. This framework enabled a detailed replication of the individual sequencing strategy used in creating each patient's sequence data. It also facilitated the validation of key assumptions regarding Limit of Detection (LoD) and average tumour purity in the dataset, leading to an improved estimation of the incidence of pathogenic *KRAS* across the cohort. The simulations also drew attention to an increase in the number of type I errors associated with alignment issues in tumour-only mutation calling, which is not observed in tumour-normal pipelines. Alignment artefacts during somatic variant calling, arising from disparities between the patient and reference genome, are typically eliminated because they are also present in the matched normal sample. However, in tumour-only pipelines, these artefacts are mistakenly identified as somatic variants. Finally, further analysis of somatic mutation data also revealed recurrent mutations in several other cancer-associated genes that may have played a role in disease progression in these patients. Our analysis also underscored the crucial importance of considering biological context when analysing filtered variants at clinically relevant loci. It revealed evidence of pathogenic *KRAS* variants in several patients that had been previously excluded from the analysis by variant filtering.

5.2 Future perspectives

While both vcfView and the stochastic simulation framework have proven highly useful in recovering clinically and research-relevant information from somatic mutation data, there are several areas in which these methods could be further enhanced. As outlined in Chapter 2, vcfView relies on multiple Bioconductor packages to functionally annotate VCF data each time a new VCF file is loaded. Depending on the system running the application and the target size, this process can lead to prolonged load times and increased demands on system resources. The introduction of an option to use pre-existing annotations from a pre-annotated VCF file would expedite analyses using this tool. Potential improvements to the simulation framework outlined in Chapter 3 include improving the replication of the coverage profile, reflecting how the depth of coverage varies across the target, when simulating individual BAM files. Additionally, adding functionality to spike-in somatic indels would further contribute to improving the accuracy of the simulation.

There are several other potential applications of the simulation framework that warrant investigation. Tumour-only methods of mutational profiling have seen increased use in clinical oncology in recent years. This simplified assay does not require a matched normal sample, reducing costs associated with sample collection and sequencing. Tumour-only calling pipelines generally employ databases of known germline polymorphisms and computational modelling using high depth of coverage to predict germline status. These approaches however lack the ability to exclude other artefacts of the sequencing process, such as alignment artefacts, that are typically highlighted by the normal control and this may result in an increase in type I errors. Similarly, disparities between the patient genome and the standard reference genome used in variant analysis may result in the misalignment of reads containing evidence of the alternate allele, leading to the failure to detect clinically relevant somatic variants. The full extent of these issues has not been thoroughly investi-

gated in the literature. The simulation framework facilitates a comprehensive and realistic simulation of tumour genomic sequencing data based on the phased, personalised genome of any 1000 Genomes donor. This makes it the ideal computational approach to explore the potential scope of these issue.

Finally, the simulations involving a uniform somatic distribution outlined in Chapters 3 and 4 have the potential to further inform our understanding of the allele frequency spectrum inferred by the variant caller from a specific sequencing strategy. In these simulations, we divided the frequency spectrum into a specific number of equal intervals and introduced a uniform, constant somatic burden in each interval. Observing the actual burden in each interval, we recorded the fraction of that burden detected across all intervals as estimated by Mutect2. For example, with 200 intervals, this resulted in a 200x200 detection matrix, enabling us to predict where in the frequency spectrum the true burden in any semi-centile will be observed by the caller. This matrix allows us to accurately simulate any observed allele frequency spectrum. This can be achieved using multiple least squares regression to decompose the spectrum into a series of semi-centiles, each containing an individual uniform burden, which may then serve as a basis for further simulations. As these semi-centile burdens represent an estimate of where in the frequency spectrum the true somatic burden is located, this method may also be applicable as a means of estimating TMB, the total burden present in the sample above a 5% VAF threshold.

Bibliography

[1] Breasted, J. H. The Edwin Smith Surgical Papyrus: published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes. **1930**,

[2] Hajdu, S. I. Greco-Roman thought about cancer. *Cancer* **2004**, *100*, 2048–2051.

[3] Odes, E. J.; Randolph-Quinney, P. S.; Steyn, M.; Throckmorton, Z.; Smilg, J. S.; Zipfel, B.; Augustine, T. N.; De Beer, F.; Hoffman, J. W.; Franklin, R. D.; Berger, L. R. Earliest hominin cancer: 1.7-million-year-old osteosarcoma from Swartkrans Cave, South Africa. *South African Journal of Science* **2016**, *112*, 5.

[4] Rothschild, B. M.; Witzke, B. J.; Hershkovitz, I. Metastatic cancer in the Jurassic. *Lancet* **1999**, *354*, 398.

[5] Rothschild, B. M.; Tanke, D. H.; Helbling, M.; Martin, L. D. Epidemiologic study of tumors in dinosaurs. *Naturwissenschaften* **2003**, *90*, 495–500.

[6] Aktipis, C. A.; Boddy, A. M.; Jansen, G.; Hibner, U.; Hochberg, M. E.; Maley, C. C.; Wilkinson, G. S. Cancer across the tree of life: cooperation and cheating in multicellularity. *Philos Trans R Soc Lond B Biol Sci* **2015**, *370*.

[7] Hanahan, D.; Weinberg, R. A. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70.

[8] Chigira, M.; Noda, K.; Watanabe, H. Autonomy in tumor cell proliferation. *Med Hypotheses* **1990**, *32*, 249–254.

[9] Sever, R.; Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb Perspect Med* **2015**, *5*.

[10] Amin, A. R. M. R. *et al.* Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. *Semin Cancer Biol* **2015**, *35 Suppl*, S55–S77.

[11] Derynck, R.; Turley, S. J.; Akhurst, R. J. TGF-β biology in cancer progression and immunotherapy. *Nat Rev Clin Oncol* **2021**, *18*, 9–34.

[12] Kari, S.; Subramanian, K.; Altomonte, I. A.; Murugesan, A.; Yli-Harja, O.; Kandhavelu, M. Programmed cell death detection methods: a systematic review and a categorical comparison. *Apoptosis* **2022**, *27*, 482–508.

[13] Chen, X.; Zhang, T.; Su, W.; Dou, Z.; Zhao, D.; Jin, X.; Lei, H.; Wang, J.; Xie, X.; Cheng, B.; Li, Q.; Zhang, H.; Di, C. Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis* **2022**, *13*, 974.

[14] Leonard, B. C.; Johnson, D. E. Signaling by cell surface death receptors: Alterations in head and neck cancer. *Adv Biol Regul* **2018**, *67*, 170–178.

[15] Shay, J. W.; Wright, W. E.; Hayflick, L. Hayflick, his limit, and cellular ageing. *Nat Rev Mol Cell Biol* **2000**, *1*, 72–76.

[16] Kim, N. W.; Piatyszek, M. A.; Prowse, K. R.; Harley, C. B.; West, M. D.; Ho, P. L.; Coviello, G. M.; Wright, W. E.; Weinrich, S. L.; Shay, J. W. Specific association of human telomerase activity with immortal cells and cancer. *Science* **1994**, *266*, 2011–2015.

[17] Shay, J. W.; Bacchetti, S. A survey of telomerase activity in human cancer. *Eur J Cancer* **1997**, *33*, 787–791.

[18] Dilley, R. L.; Greenberg, R. A. ALTernative Telomere Maintenance and Cancer. *Trends Cancer* **2015**, *1*, 145–156.

[19] Sitohy, B.; Nagy, J. A.; Dvorak, H. F. Anti-VEGF/VEGFR therapy for cancer: reassessing the target. *Cancer Res* **2012**, *72*, 1909–1914.

[20] Al-Ostoot, F. H.; Sherapura, A.; V, V.; Basappa, G.; H K, V.; B T, P.; Khanum, S. A. by newly synthesized Indolephenoxyacetamide (IPA) analogs to induce anti-angiogenesis-mediated solid tumor suppression. *Pharmacol Rep* **2021**, *73*, 1328–1343.

[21] s, H.; Rogers, M. S.; Straume, O. Are 90metastases? *Cancer Med* **2019**, *8*, 5574–5576.

[22] Chung, V. Y.; Tan, T. Z.; Ye, J.; Huang, R. L.; Lai, H. C.; Kappei, D.; Wollmann, H.; Guccione, E.; Huang, R. Y. The role of GRHL2 and epigenetic remodeling in epithelial-mesenchymal plasticity in ovarian cancer cells. *Commun Biol* **2019**, *2*, 272.

[23] Gao, D.; Vahdat, L. T.; Wong, S.; Chang, J. C.; Mittal, V. Microenvironmental regulation of epithelial-mesenchymal transitions in cancer. *Cancer Res* **2012**, *72*, 4883–4889.

[24] Somarelli, J. A.; Shetler, S.; Jolly, M. K.; Wang, X.; Bartholf Dewitt, S.; Hish, A. J.; Gilja, S.; Eward, W. C.; Ware, K. E.; Levine, H.; Armstrong, A. J.; Garcia-Blanco, M. A. Mesenchymal-Epithelial Transition in Sarcomas Is Controlled by the Combinatorial Expression of MicroRNA 200s and GRHL2. *Mol Cell Biol* **2016**, *36*, 2503–2513.

[25] Lin, K.; Baritaki, S.; Militello, L.; Malaponte, G.; Bevelacqua, Y.; Bonavida, B. The Role of B-RAF Mutations in Melanoma and the Induction of EMT via Dysregulation of the NF- B/Snail/RKIP/PTEN Circuit. *Genes & Cancer* **2010**, *1*, 409–420.

[26] Hanahan, D.; Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646–674.

[27] Hagemann, T.; Balkwill, F.; Lawrence, T. Inflammation and cancer: a double-edged sword. *Cancer Cell* **2007**, *12*, 300–301.

[28] Musa, M. Immune mechanism: a 'double-edged sword'. *Malays J Med Sci* **2013**, *20*, 61–67.

[29] Lakshmi Narendra, B.; Eshvendar Reddy, K.; Shantikumar, S.; Ramakrishna, S. Immune system: a double-edged sword in cancer. *Inflamm Res* **2013**, *62*, 823–834.

[30] WARBURG, O. On the origin of cancer cells. *Science* **1956**, *123*, 309–314.

[31] Mazouzi, A.; Velimezi, G.; Loizou, J. I. DNA replication stress: causes, resolution and disease. *Exp Cell Res* **2014**, *329*, 85–93.

[32] Jinks-Robertson, S.; Bhagwat, A. S. Transcription-associated mutagenesis. *Annu Rev Genet* **2014**, *48*, 341–359.

[33] Liu, Y.; Prasad, R.; Beard, W. A.; Kedar, P. S.; Hou, E. W.; Shock, D. D.; Wilson, S. H. Coordination of steps in single-nucleotide base excision repair mediated by apurinic/apyrimidinic endonuclease 1 and DNA polymerase beta. *J Biol Chem* **2007**, *282*, 13532–13541.

[34] Wright, W. D.; Shah, S. S.; Heyer, W. D. Homologous recombination and the repair of DNA double-strand breaks. *J Biol Chem* **2018**, *293*, 10524–10535.

[35] Lara-Gonzalez, P.; Pines, J.; Desai, A. Spindle assembly checkpoint activation and silencing at kinetochores. *Semin Cell Dev Biol* **2021**, *117*, 86–98.

[36] Spivak, G. Transcription-coupled repair: an update. *Arch Toxicol* **2016**, *90*, 2583–2594.

[37] Wogan, G. N.; Hecht, S. S.; Felton, J. S.; Conney, A. H.; Loeb, L. A. Environmental and chemical carcinogenesis. *Semin Cancer Biol* **2004**, *14*, 473–486.

[38] Shala, N. K. *et al.* Exposure to benzene and other hydrocarbons and risk of bladder cancer among male offshore petroleum workers. *Br J Cancer* **2023**, *129*, 838–851.

[39] Borrego-Soto, G.; pez, R.; nez, A. Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer. *Genet Mol Biol* **2015**, *38*, 420–432.

[40] Kay, J.; Thadhani, E.; Samson, L.; Engelward, B. Inflammation-induced DNA damage, mutations and cancer. *DNA Repair (Amst)* **2019**, *83*, 102673.

[41] Chang, H. H. Y.; Pannunzio, N. R.; Adachi, N.; Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **2017**, *18*, 495–506.

[42] Brambullo, T.; Colonna, M. R.; Vindigni, V.; Piaserico, S.; Masciopinto, G.; Galeano, M.; Costa, A. L.; Bassetto, F. Xeroderma Pigmentosum: A Genetic Condition Skin Cancer Correlated-A Systematic Review. *Biomed Res Int* **2022**, *2022*, 8549532.

[43] ki, P.; m, M.; Mecklin, J. P.; ä, T. T. Lynch Syndrome Genetics and Clinical Implications. *Gastroenterology* **2023**, *164*, 783–799.

[44] Kriebs, A. Somatic mutation rate tracks with lifespan. *Nature Aging* **2022**, *2*, 273–273.

[45] Maeda, N.; Fan, H.; Yoshikai, Y. Oncogenesis by retroviruses: old and new paradigms. *Rev Med Virol* **2008**, *18*, 387–405.

[46] Blanco-Melo, D.; Venkatesh, S.; Bieniasz, P. D. Intrinsic cellular defenses against human immunodeficiency viruses. *Immunity* **2012**, *37*, 399–411.

[47] Sheehy, A. M.; Gaddis, N. C.; Choi, J. D.; Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **2002**, *418*, 646–650.

[48] Stremlau, M.; Owens, C. M.; Perron, M. J.; Kiessling, M.; Autissier, P.; Sodroski, J. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* **2004**, *427*, 848–853.

[49] Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci* **2008**, *121 Suppl 1*, 1–84.

[50] Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J Clin Invest* **2007**, *117*, 2033–2035.

[51] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **1976**, *194*, 23–28.

[52] Davis, A.; Gao, R.; Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta Rev Cancer* **2017**, *1867*, 151–161.

[53] Merlo, L. M.; Pepper, J. W.; Reid, B. J.; Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **2006**, *6*, 924–935.

[54] Pepper, J. W.; Scott Findlay, C.; Kassen, R.; Spencer, S. L.; Maley, C. C. Cancer research meets evolutionary biology. *Evol Appl* **2009**, *2*, 62–70.

[55] Darwin, C. *On the origin of species by means of natural selection; or, the preservation of favored races in the struggle for life*; John murray, 1876.

[56] Field, M. G.; Durante, M. A.; Anbunathan, H.; Cai, L. Z.; Decatur, C. L.; Bowcock, A. M.; Kurtenbach, S.; Harbour, J. W. Punctuated evolution of canonical genomic aberrations in uveal melanoma. *Nat Commun* **2018**, *9*, 116.

[57] Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **2013**, *153*, 666–677.

[58] Cross, W. C. h.; Graham, T. A.; Wright, N. A. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J Pathol* **2016**, *240*, 126–136.

[59] Passaro, A.; nne, P. A.; Mok, T.; Peters, S. Overcoming therapy resistance in EGFR-mutant lung cancer. *Nat Cancer* **2021**, *2*, 377–391.

[60] Fearon, E. R.; Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **1990**, *61*, 759–767.

[61] Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **2021**, *184*, 2239–2254.

[62] Hansen, E.; Read, A. F. Modifying Adaptive Therapy to Enhance Competitive Suppression. *Cancers (Basel)* **2020**, *12*.

[63] Kimura, M. *The neutral theory of molecular evolution*; Cambridge University Press, 1983.

[64] Ohta, T. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics* **1992**, *23*, 263–286.

[65] Hurst, L. D. Genetics and the understanding of selection. *Nature Reviews Genetics* **2009**, *10*, 83–93.

[66] Williams, M. J.; Werner, B.; Barnes, C. P.; Graham, T. A.; Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat Genet* **2016**, *48*, 238–244.

[67] Marusyk, A.; Janiszewska, M.; Polyak, K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell* **2020**, *37*, 471–484.

[68] Klempner, S. J.; Fabrizio, D.; Bane, S.; Reinhart, M.; Peoples, T.; Ali, S. M.; Sokol, E. S.; Frampton, G.; Schrock, A. B.; Anhorn, R.; Reddy, P. Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence. *Oncologist* **2020**, *25*, e147–e159.

[69] Wang, Y.; Zhang, M.; Shi, J.; Zhu, Y.; Wang, X.; Zhang, S.; Wang, F. Cracking the pattern of tumor evolution based on single-cell copy number alterations. *Brief Bioinform* **2023**, *24*.

[70] Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **2011**, *144*, 27–40.

[71] Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **1971**, *68*, 820–823.

[72] Hall, J. M.; Lee, M. K.; Newman, B.; Morrow, J. E.; Anderson, L. A.; Huey, B.; King, M. C. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **1990**, *250*, 1684–1689.

[73] Newman, B.; Austin, M. A.; Lee, M.; King, M. C. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc Natl Acad Sci U S A* **1988**, *85*, 3044–3048.

[74] Hurst, J. H.; King, M. Pioneering geneticist Mary-Claire King receives the 2014 Lasker Koshland Special Achievement Award in Medical Science. *J Clin Invest* **2014**, *124*, 4148–4151.

[75] Wooster, R.; Neuhausen, S. L.; Mangion, J.; Quirk, Y.; Ford, D.; Collins, N.; Nguyen, K.; Seal, S.; Tran, T.; Averill, D. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **1994**, *265*, 2088–2090.

[76] Cornelis, R. S.; Neuhausen, S. L.; Johansson, O.; Arason, A.; Kelsell, D.; Ponder, B. A.; Tonin, P.; Hamann, U.; Lindblom, A.; Lalle, P. High allele loss rates at 17q12-q21 in breast and ovarian tumors from BRCAl-linked families. The Breast Cancer Linkage Consortium. *Genes Chromosomes Cancer* **1995**, *13*, 203–210.

[77] Collins, N.; McManus, R.; Wooster, R.; Mangion, J.; Seal, S.; Lakhani, S. R.; Ormiston, W.; Daly, P. A.; Ford, D.; Easton, D. F. Consistent loss of the wild type allele in breast cancers from a family linked to the BRCA2 gene on chromosome 13q12-13. *Oncogene* **1995**, *10*, 1673–1675.

[78] Inc., M. G. 17Q-linked breast and ovarian cancer susceptibility gene. U.S. Patent US5747282A, May. 1998; `https://patents.google.com/patent/US5747282A/en`, Accessed on November 28, 2023.

[79] Price, A. L.; Spencer, C. C.; Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* **2015**, *282*, 20151684.

[80] Varley, J. M. Germline TP53 mutations and Li-Fraumeni syndrome. *Hum Mutat* **2003**, *21*, 313–320.

[81] Mester, J.; Eng, C. Cowden syndrome: recognizing and managing a not-so-rare hereditary cancer syndrome. *J Surg Oncol* **2015**, *111*, 125–130.

[82] Vogel, T. *Encyclopedia of Cancer*; Springer Berlin Heidelberg, 2008; p 2296–2298.

[83] Li, X.; Liu, G.; Wu, W. Recent advances in Lynch syndrome. *Exp Hematol Oncol* **2021**, *10*, 37.

[84] Berger, E. R.; Golshan, M. Surgical Management of Hereditary Breast Cancer. *Genes (Basel)* **2021**, *12*.

[85] Ishak, C. A.; Dick, F. A. Conditional haploinsufficiency of the retinoblastoma tumor suppressor gene. *Mol Cell Oncol* **2015**, *2*, e968069.

[86] Minello, A.; Carreira, A. BRCA1/2 Haploinsufficiency: Exploring the Impact of Losing one Allele. *J Mol Biol* **2024**, *436*, 168277.

[87] Zhang, Y.; Manjunath, M.; Yan, J.; Baur, B. A.; Zhang, S.; Roy, S.; Song, J. S. The Cancer-Associated Genetic Variant Rs3903072 Modulates Immune Cells in the Tumor Microenvironment. *Front Genet* **2019**, *10*, 754.

[88] Fernandez-Moya, A.; Morales, S.; Arancibia, T.; Gonzalez-Hormazabal, P.; Tapia, J. C.; Godoy-Herrera, R.; Reyes, J. M.; Gomez, F.; Waugh, E.; Jara, L. Germline Variants in Driver Genes of Breast Cancer and Their Association with Familial and Early-Onset Breast Cancer Risk in a Chilean Population. *Cancers (Basel)* **2020**, *12*.

[89] Wang, Q. L.; Wang, T. M.; Deng, C. M.; Zhang, W. L.; He, Y. Q.; Xue, W. Q.; Liao, Y.; Yang, D. W.; Zheng, M. Q.; Jia, W. H. Association of HLA diversity with the risk of 25 cancers in the UK Biobank. *EBioMedicine* **2023**, *92*, 104588.

[90] Hughes, C. C. Endothelial-stromal interactions in angiogenesis. *Curr Opin Hematol* **2008**, *15*, 204–209.

[91] Salmon, H.; Remark, R.; Gnjatic, S.; Merad, M. Host tissue determinants of tumour immunity. *Nat Rev Cancer* **2019**, *19*, 215–227.

[92] Quail, D. F.; Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat Med* **2013**, *19*, 1423–1437.

[93] Petrova, V.; Annicchiarico-Petruzzelli, M.; Melino, G.; Amelio, I. The hypoxic tumour microenvironment. *Oncogenesis* **2018**, *7*, 10.

[94] Masson, N.; Ratcliffe, P. J. Hypoxia signaling pathways in cancer metabolism: the importance of co-selecting interconnected physiological pathways. *Cancer Metab* **2014**, *2*, 3.

[95] Folkman, J. Tumor angiogenesis: therapeutic implications. *N Engl J Med* **1971**, *285*, 1182–1186.

[96] Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer* **2020**, *20*, 174–186.

[97] Wu, F.; Yang, J.; Liu, J.; Wang, Y.; Mu, J.; Zeng, Q.; Deng, S.; Zhou, H. Signaling pathways in cancer-associated fibroblasts and targeted therapy for cancer. *Signal Transduct Target Ther* **2021**, *6*, 218.

[98] Glentis, A.; Oertle, P.; Mariani, P.; Chikina, A.; El Marjou, F.; Attieh, Y.; Zaccarini, F.; Lae, M.; Loew, D.; Dingli, F.; Sirven, P.; Schoumacher, M.; Gurchenkov, B. G.; Plodinec, M.; Vignjevic, D. M. Cancer-associated fibroblasts induce metalloprotease-independent cancer cell invasion of the basement membrane. *Nat Commun* **2017**, *8*, 924.

[99] Fukumura, D.; Xavier, R.; Sugiura, T.; Chen, Y.; Park, E. C.; Lu, N.; Selig, M.; Nielsen, G.; Taksir, T.; Jain, R. K.; Seed, B. Tumor induction of VEGF promoter activity in stromal cells. *Cell* **1998**, *94*, 715–725.

[100] Weber, C. E.; Kuo, P. C. The tumor microenvironment. *Surg Oncol* **2012**, *21*, 172–177.

[101] Comito, G.; Giannoni, E.; Segura, C. P.; Barcellos-de Souza, P.; Raspollini, M. R.; Baroni, G.; Lanciotti, M.; Serni, S.; Chiarugi, P. Cancer-associated fibroblasts and M2-polarized macrophages synergize during prostate carcinoma progression. *Oncogene* **2014**, *33*, 2423–2431.

[102] Jenkins, L.; Jungwirth, U.; Avgustinova, A.; Iravani, M.; Mills, A.; Haider, S.; Harper, J.; Isacke, C. M. Cancer-Associated Fibroblasts Suppress CD8+ T-cell Infiltration and Confer Resistance to Immune-Checkpoint Blockade. *Cancer Res* **2022**, *82*, 2904–2917.

[103] Mao, X.; Xu, J.; Wang, W.; Liang, C.; Hua, J.; Liu, J.; Zhang, B.; Meng, Q.; Yu, X.; Shi, S. Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives. *Mol Cancer* **2021**, *20*, 131.

[104] Decker, W. K.; Safdar, A.; Coley, W. B. Bioimmunoadjuvants for the treatment of neoplastic and infectious disease: Coley's legacy revisited. *Cytokine Growth Factor Rev* **2009**, *20*, 271–281.

[105] BURNET, M. Cancer: a biological approach. III. Viruses associated with neoplastic conditions. IV. Practical applications. *Br Med J* **1957**, *1*, 841–847.

[106] Shankaran, V.; Ikeda, H.; Bruce, A. T.; White, J. M.; Swanson, P. E.; Old, L. J.; Schreiber, R. D. IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **2001**, *410*, 1107–1111.

[107] Dunn, G. P.; Bruce, A. T.; Ikeda, H.; Old, L. J.; Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol* **2002**, *3*, 991–998.

[108] Dunn, G. P.; Old, L. J.; Schreiber, R. D. The three Es of cancer immunoediting. *Annu Rev Immunol* **2004**, *22*, 329–360.

[109] Nutt, S. L.; Huntington, N. D. In *Clinical Immunology (Fifth Edition)*, fifth edition ed.; Rich, R. R., Fleisher, T. A., Shearer, W. T., Schroeder, H. W., Frew, A. J., Weyand, C. M., Eds.; Elsevier: London, 2019; pp 247–259.e1.

[110] Jorgovanovic, D.; Song, M.; Wang, L.; Zhang, Y. Roles of IFN-γ in tumor progression and regression: a review. *Biomark Res* **2020**, *8*, 49.

[111] Pernot, S.; Evrard, S.; Khatib, A. M. The Give-and-Take Interaction Between the Tumor Microenvironment and Immune Cells Regulating Tumor Progression and Repression. *Front Immunol* **2022**, *13*, 850856.

[112] Zhai, L. *et al.* Immunosuppressive IDO in Cancer: Mechanisms of Action, Animal Models, and Targeting Strategies. *Front Immunol* **2020**, *11*, 1185.

[113] Mirlekar, B. Tumor promoting roles of IL-10, TGF-β, IL-4, and IL-35: Its implications in cancer immunotherapy. *SAGE Open Med* **2022**, *10*, 20503121211069012.

[114] Yi, M.; Niu, M.; Xu, L.; Luo, S.; Wu, K. Regulation of PD-L1 expression in the tumor microenvironment. *J Hematol Oncol* **2021**, *14*, 10.

[115] Dhatchinamoorthy, K.; Colbert, J. D.; Rock, K. L. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Front Immunol* **2021**, *12*, 636568.

[116] Groh, V.; Wu, J.; Yee, C.; Spies, T. Tumour-derived soluble MIC ligands impair expression of NKG2D and T-cell activation. *Nature* **2002**, *419*, 734–738.

[117] Xing, S.; Ferrari de Andrade, L. NKG2D and MICA/B shedding: a 'tag game' between NK cells and malignant cells. *Clin Transl Immunology* **2020**, *9*, e1230.

[118] Hastings, R. K. *et al.* Longitudinal whole-exome sequencing of cell-free DNA for tracking the co-evolutionary tumor and immune evasion dynamics: longitudinal data from a single patient. *Ann Oncol* **2021**, *32*, 681–684.

[119] Tirado, C. In *Pathobiology of Human Disease*; McManus, L. M., Mitchell, R. N., Eds.; Academic Press: San Diego, 2014; pp 3399–3407.

[120] Moore, L. D.; Le, T.; Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **2013**, *38*, 23–38.

[121] Holliday, R.; Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **1975**, *187*, 226–232.

[122] Compere, S. J.; Palmiter, R. D. DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. *Cell* **1981**, *25*, 233–240.

[123] Miller, J. L.; Grant, P. A. The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell Biochem* **2013**, *61*, 289–317.

[124] Brownell, J. E.; Zhou, J.; Ranalli, T.; Kobayashi, R.; Edmondson, D. G.; Roth, S. Y.; Allis, C. D. Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* **1996**, *84*, 843–851.

[125] Gujral, P.; Mahajan, V.; Lissaman, A. C.; Ponnampalam, A. P. Histone acetylation and the role of histone deacetylases in normal cyclic endometrium. *Reprod Biol Endocrinol* **2020**, *18*, 84.

[126] Zhang, C.; Sheng, Q.; Zhao, N.; Huang, S.; Zhao, Y. DNA hypomethylation mediates immune response in pan-cancer. *Epigenetics* **2023**, *18*, 2192894.

[127] Feinberg, A. P.; Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **1983**, *301*, 89–92.

[128] Irizarry, R. A.; Ladd-Acosta, C.; Wen, B.; Wu, Z.; Montano, C.; Onyango, P.; Cui, H.; Gabo, K.; Rongione, M.; Webster, M.; Ji, H.; Potash, J.; Sabunciyan, S.; Feinberg, A. P. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **2009**, *41*, 178–186.

[129] Widschwendter, M.; Jiang, G.; Woods, C.; ller, H. M.; Fiegl, H.; Goebel, G.; Marth, C.; ller Holzner, E.; Zeimet, A. G.; Laird, P. W.; Ehrlich, M. DNA hypomethylation and ovarian cancer biology. *Cancer Res* **2004**, *64*, 4472–4480.

[130] Zhang, Q.; Zhu, X.; Liu, B.; Zhang, Y.; Xiao, Y. Case report: Sandwich therapy of CAR-T combined with ASCT: Sequential CAR-T cell therapy with ASCT after remission with CAR-T therapy caused long-term survival in a patient with relapsed/refractory Burkitt's lymphoma with TP53 mutations. *Front Immunol* **2023**, *14*, 1127868.

[131] Cao, W. *et al.* Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun* **2020**, *11*, 3675.

[132] Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **2012**, *13*, 484–492.

[133] Nishiyama, A.; Nakanishi, M. Navigating the DNA methylation landscape of cancer. *Trends Genet* **2021**, *37*, 1012–1027.

[134] Ribeiro-Silva, C.; Vermeulen, W.; Lans, H. SWI/SNF: Complex complexes in genome stability and cancer. *DNA Repair (Amst)* **2019**, *77*, 87–95.

[135] Di Cerbo, V.; Schneider, R. Cancers with wrong HATs: the impact of acetylation. *Brief Funct Genomics* **2013**, *12*, 231–243.

[136] Yang, Y.; Zhang, M.; Wang, Y. The roles of histone modifications in tumorigenesis and associated inhibitors in cancer therapy. *Journal of the National Cancer Center* **2022**, *2*, 277–290.

[137] Ahmed, Z.; Zeeshan, S.; Mendhe, D.; Dong, X. Human gene and disease associations for clinical-genomics and precision medicine research. *Clin Transl Med* **2020**, *10*, 297–318.

[138] Ren, P.; Dong, X.; Vijg, J. Age-related somatic mutation burden in human tissues. *Front Aging* **2022**, *3*, 1018119.

[139] Fowler, J. C.; Jones, P. H. Somatic Mutation: What Shapes the Mutational Landscape of Normal Epithelia? *Cancer Discov* **2022**, *12*, 1642–1655.

[140] Li, C.; Williams, S. M. Human Somatic Variation: It's Not Just for Cancer Anymore. *Current Genetic Medicine Reports* **2013**, *1*, 212–218.

[141] Olafsson, S.; Anderson, C. A. Somatic mutations provide important and unique insights into the biology of complex diseases. *Trends Genet* **2021**, *37*, 872–881.

[142] Nurk, S. *et al.* The complete sequence of a human genome. *Science* **2022**, *376*, 44–53.

[143] Vihinen, M. When a Synonymous Variant Is Nonsynonymous. *Genes (Basel)* **2022**, *13*.

[144] Sim, N. L.; Kumar, P.; Hu, J.; Henikoff, S.; Schneider, G.; Ng, P. C. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **2012**, *40*, W452–457.

[145] Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat Methods* **2010**, *7*, 248–249.

[146] Lee, T. I.; Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **2013**, *152*, 1237–1251.

[147] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **2015**, *526*, 75–81.

[148] McColgan, P.; Tabrizi, S. J. Huntington's disease: a clinical review. *Eur J Neurol* **2018**, *25*, 24–34.

[149] Sohn, J. I. *et al.* Ultrafast prediction of somatic structural variations by filtering out reads matched to pan-genome k-mer sets. *Nat Biomed Eng* **2023**, *7*, 853–866.

[150] Rausch, T.; Zichner, T.; Schlattl, A.; tz, A. M.; Benes, V.; Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, i333–i339.

[151] Koh, G.; Degasperi, A.; Zou, X.; Momen, S.; Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* **2021**, *21*, 619–637.

[152] Alexandrov, L.; Nik-Zainal, S.; Wedge, D. C.; et al., Signatures of mutational processes in human cancer. *Nature* **2013**, *500*, 415–421.

[153] Pfeifer, G. P.; You, Y. H.; Besaratinia, A. Mutations induced by ultraviolet light. *Mutat Res* **2005**, *571*, 19–31.

[154] South, A. P.; den Breems, N. Y.; Richa, T.; Nwagu, U.; Zhan, T.; Poojan, S.; Martinez-Outschoorn, U.; Johnson, J. M.; Luginbuhl, A. J.; Curry, J. M. Mutation signature analysis identifies increased mutation caused by tobacco smoke associated DNA adducts in larynx squamous cell carcinoma compared with oral cavity and oropharynx. *Sci Rep* **2019**, *9*, 19256.

[155] Huang, M. N. *et al.* Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res* **2017**, *27*, 1475–1486.

[156] Cai, X.; Chen, Z.; Deng, M.; Li, Z.; Wu, Q.; Wei, J.; Dai, C.; Wang, G.; Luo, C. Unique genomic features and prognostic value of COSMIC mutational signature 4 in lung adenocarcinoma and lung squamous cell carcinoma. *Ann Transl Med* **2020**, *8*, 1176.

[157] Walker, B. A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat Commun* **2015**, *6*, 6997.

[158] Woolston, A.; Barber, L. J.; Griffiths, B.; Pich, O.; Lopez-Bigas, N.; Matthews, N.; Rao, S.; Watkins, D.; Chau, I.; Starling, N.; Cunningham, D.; Gerlinger, M. Mutational signatures impact the evolution of anti-EGFR antibody resistance in colorectal cancer. *Nat Ecol Evol* **2021**, *5*, 1024–1032.

[159] Audeh, M. W.; Carmichael, J.; Penson, R. T.; Friedlander, M.; Powell, B.; Bell-McGuinn, K. M.; Scott, C.; Weitzel, J. N.; Oaknin, A.; Loman, N.; Lu, K.; Schmutzler, R. K.; Matulonis, U.; Wickens, M.; Tutt, A. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* **2010**, *376*, 245–251.

[160] Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **2017**, *23*, 517–525.

[161] Sun, T.; Ruscito, I.; Dimitrova, D.; Chekerov, R.; Kulbe, H.; Baron, U.; Blanchard, V.; Panici, P. B.; Darb-Esfahani, S.; Sehouli, J.; Olek, S.; Braicu, E. I. Genetic Versus Epigenetic BRCA1 Silencing Pathways: Clinical Effects in Primary Ovarian Cancer Patients: A Study of the Tumor Bank Ovarian Cancer Consortium. *Int J Gynecol Cancer* **2017**, *27*, 1658–1665.

[162] Steiert, T. A. *et al.* A critical spotlight on the paradigms of FFPE-DNA sequencing. *Nucleic Acids Res* **2023**, *51*, 7143–7162.

[163] O'Sullivan, B.; Seoighe, C. Comprehensive and realistic simulation of tumour genomic sequencing data. *NAR Cancer* **2023**, *5*, zcad051.

[164] Guo, Q.; Lakatos, E.; Bakir, I. A.; Curtius, K.; Graham, T. A.; Mustonen, V. The mutational signatures of formalin fixation on the human genome. *Nat Commun* **2022**, *13*, 4487.

[165] Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **2017**, *45*, D777–D783.

[166] Pardue, M. L.; Gall, J. G. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A* **1969**, *64*, 600–604.

[167] Huber, D.; von Voithenberg, L. V.; Kaigala, G. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering* **2018**, *1*, 15–24.

[168] Nolte, M.; Werner, M.; Ewig, M.; von Wasielewski, R.; Wilkens, L.; Link, H.; Ganser, A.; Georgii, A. Fluorescence in situ hybridization (FISH) is a reliable diagnostic tool for detection of the 9;22 translocation. *Leuk Lymphoma* **1996**, *22*, 287–294.

[169] Warshawsky, I. In *Cell and Tissue Based Molecular Pathology*; Tubbs, R. R., Stoler, M. H., Eds.; Elsevier, 2009; pp 3–9.

[170] Kawasaki, E. S.; Clark, S. S.; Coyne, M. Y.; Smith, S. D.; Champlin, R.; Witte, O. N.; McCormick, F. P. Diagnosis of chronic myeloid and acute lymphocytic leukemias by detection of leukemia-specific mRNA sequences amplified in vitro. *Proc Natl Acad Sci U S A* **1988**, *85*, 5698–5702.

[171] Roth, M. S.; Terry, V. H. Application of the polymerase chain reaction for detection of minimal residual disease of hematologic malignancies. *Henry Ford Hosp Med J* **1991**, *39*, 112–116.

[172] nik, H.; nik, U. Blood-Based mRNA Tests as Emerging Diagnostic Tools for Personalised Medicine in Breast Cancer. *Cancers (Basel)* **2023**, *15*.

[173] Surriabre, P.; Torrico, A.; Vargas, T.; Ugarte, F.; Rodriguez, P.; Fontaine, V. Assessment of a new low-cost, PCR-based strategy for high-risk human papillomavirus DNA detection for cervical cancer prevention. *BMC Infect Dis* **2019**, *19*, 842.

[174] Bustin, S. A.; Nolan, T. RT-qPCR Testing of SARS-CoV-2: A Primer. *Int J Mol Sci* **2020**, *21*.

[175] Smeltzer, M. P. *et al.* The International Association for the Study of Lung Cancer Global Survey on Molecular Testing in Lung Cancer. *J Thorac Oncol* **2020**, *15*, 1434–1448.

[176] of Medicine, I.; Council, N. R. *BioWatch PCR Assays: Building Confidence, Ensuring Reliability: Abbreviated Version*; The National Academies Press: Washington, DC, 2015.

[177] Commission, E. *et al. Guidance document on multiplex real-time PCR methods*; Publications Office, 2021.

[178] Klee, E. W.; Hoppman-Chaney, N. L.; Ferber, M. J. Expanding DNA diagnostic panel testing: is more better? *Expert Rev Mol Diagn* **2011**, *11*, 703–709.

[179] Myllykangas, S.; Ji, H. P. Targeted deep resequencing of the human cancer genome using next-generation technologies. *Biotechnol Genet Eng Rev* **2010**, *27*, 135–158.

[180] Heydt, C. *et al.* Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation. *BMC Med Genomics* **2021**, *14*, 62.

[181] NCI, Genomic Profiling Tests Cleared by FDA Can Help Guide Cancer Treatment, Clinical Trial Enrollment. `https://www.cancer.gov/news-events/cancer-currents-blog/2017/genomic-profiling-tests-cancer`, Accessed: 2023-09-02.

[182] FDA, Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing (NGS) – Based In Vitro Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases. `https://www.fda.gov/media/99208/download`, Accessed: 2023-09-02.

[183] of Health, N. Y. D. Next Generation Sequencing (NGS) guidelines for somatic genetic variant detection. `https://www.wadsworth.org/sites/default/files/WebDoc/NextGenSeqONCOGuidelines%20_April_2021.pdf`, Accessed: 2023-09-02.

[184] Gong, B. *et al.* Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. *Genome Biol* **2021**, *22*, 109.

[185] Sunami, K. *et al.* Feasibility and utility of a panel testing for 114 cancer-associated genes in a clinical setting: A hospital-based study. *Cancer Sci* **2019**, *110*, 1480–1490.

[186] Openshaw, M. R.; Mohamed, A. A.; Ottolini, B.; Fernandez-Garcia, D.; Richards, C. J.; Page, K.; Guttery, D. S.; Thomas, A. L.; Shaw, J. A. Longitudinal monitoring of circulating tumour DNA improves prognostication and relapse detection in gastroesophageal adenocarcinoma. *British Journal of Cancer* **2020**, *123*, 1271–1279.

[187] Banyi, N.; Alex, D.; Hughesman, C.; McNeil, K.; N Ionescu, D.; Ma, C.; Yip, S.; Melosky, B. EGFR Testing Platform. *Curr Oncol* **2022**, *29*, 7900–7911.

[188] MANDEL, P.; METAIS, P. [Nuclear Acids In Human Blood Plasma]. *C R Seances Soc Biol Fil* **1948**, *142*, 241–243.

[189] Diagnostics, R. cobas® EGFR Mutation Test v2. *Originally published by Roche on their Roche Diagnostics product information website*

[190] Natera, Signatera™, molecular residual disease assay. *Originally published by Natera on their oncology product information website*

[191] Hasenleithner, S. O.; Speicher, M. R. A clinician's handbook for using ctDNA throughout the patient journey. *Mol Cancer* **2022**, *21*, 81.

[192] Diehl, F.; Schmidt, K.; Choti, M. A.; Romans, K.; Goodman, S.; Li, M.; Thornton, K.; Agrawal, N.; Sokoll, L.; Szabo, S. A.; Kinzler, K. W.; Vogelstein, B.; Diaz, L. A. Circulating mutant DNA to assess tumor dynamics. *Nat Med* **2008**, *14*, 985–990.

[193] Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **2014**, *6*, 224ra24.

[194] van der Pol, Y.; Mouliere, F. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell* **2019**, *36*, 350–368.

[195] Chen, X.; Dong, Z.; Hubbell, E.; Kurtzman, K. N.; Oxnard, G. R.; Venn, O.; Melton, C.; Clarke, C. A.; Shaknovich, R.; Ma, T.; Meixiong, G.; Seiden, M. V.; Klein, E. A.; Fung, E. T.; Liu, M. C. Prognostic Significance of Blood-Based Multi-cancer Detection in Plasma Cell-Free DNA. *Clin Cancer Res* **2021**, *27*, 4221–4229.

[196] Coombes, R. C. *et al.* Personalized Detection of Circulating Tumor DNA Antedates Breast Cancer Metastatic Recurrence. *Clinical Cancer Research* **2019**, *25*, 4255–4263.

[197] Sorensen, B. S.; Wu, L.; Wei, W.; Tsai, J.; Weber, B.; Nexo, E.; Meldgaard, P. Monitoring of epidermal growth factor receptor tyrosine kinase inhibitor-sensitizing and resistance mutations in the plasma DNA of patients with advanced non-small cell lung cancer during treatment with erlotinib. *Cancer* **2014**, *120*, 3896–3901.

[198] Qi, T.; Pan, M.; Shi, H.; Wang, L.; Bai, Y.; Ge, Q. Cell-Free DNA Fragmentomics: The Novel Promising Biomarker. *Int J Mol Sci* **2023**, *24*.

[199] NCIinfo@nih.gov, Targeted Therapy to Treat Cancer. *Targeted Therapy to Treat Cancer was originally published by the National Cancer Institute on their website, https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies*

[200] Quirke, V. M. Tamoxifen from Failed Contraceptive Pill to Best-Selling Breast Cancer Medicine: A Case-Study in Pharmaceutical Innovation. *Front Pharmacol* **2017**, *8*, 620.

[201] Beatson, G. T. On the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases. *Trans Med Chir Soc Edinb* **1896**, *15*, 153–179.

[202] Fuentes, N.; Silveyra, P. Estrogen receptor signaling mechanisms. *Adv Protein Chem Struct Biol* **2019**, *116*, 135–170.

[203] Harrell, J. C.; Dye, W. W.; Allred, D. C.; Jedlicka, P.; Spoelstra, N. S.; Sartorius, C. A.; Horwitz, K. B. Estrogen receptor positive breast cancer metastasis: altered hormonal sensitivity and tumor aggressiveness in lymphatic vessels and lymph nodes. *Cancer Res* **2006**, *66*, 9308–9315.

[204] Gallo, M. A.; Kaufman, D. Antagonistic and agonistic effects of tamoxifen: significance in human cancer. *Semin Oncol* **1997**, *24*, 1–71.

[205] Fanning, S. W. *et al.* Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *Elife* **2016**, *5*.

[206] Brett, J. O.; Spring, L. M.; Bardia, A.; Wander, S. A. ESR1 mutation as an emerging clinical biomarker in metastatic hormone receptor-positive breast cancer. *Breast Cancer Res* **2021**, *23*, 85.

[207] de Souza, J. A.; Olopade, O. I. CYP2D6 genotyping and tamoxifen: an unfinished story in the quest for personalized medicine. *Semin Oncol* **2011**, *38*, 263–273.

[208] Paul, M. K.; Mukhopadhyay, A. K. Tyrosine kinase - Role and significance in Cancer. *Int J Med Sci* **2004**, *1*, 101–115.

[209] Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002**, *45*, 2615–2623.

[210] Bou-Assaly, W.; Mukherji, S. Cetuximab (erbitux). *AJNR Am J Neuroradiol* **2010**, *31*, 626–627.

[211] Gemmete, J. J.; Mukherji, S. K. Panitumumab (vectibix). *AJNR Am J Neuroradiol* **2011**, *32*, 1002–1003.

[212] Sickmier, E. A.; Kurzeja, R. J.; Michelsen, K.; Vazir, M.; Yang, E.; Tasker, A. S. The Panitumumab EGFR Complex Reveals a Binding Mechanism That Overcomes Cetuximab Induced Resistance. *PLoS One* **2016**, *11*, e0163366.

[213] a Foncillas, J.; Sunakawa, Y.; Aderka, D.; Wainberg, Z.; Ronga, P.; Witzler, P.; Stintzing, S. Distinguishing Features of Cetuximab and Panitumumab in Colorectal Cancer and Other Solid Tumors. *Front Oncol* **2019**, *9*, 849.

[214] Zhou, J.; Ji, Q.; Li, Q. Resistance to anti-EGFR therapies in metastatic colorectal cancer: underlying mechanisms and reversal strategies. *J Exp Clin Cancer Res* **2021**, *40*, 328.

[215] Motoyama, A. B.; Hynes, N. E.; Lane, H. A. The efficacy of ErbB receptor-targeted anticancer therapeutics is influenced by the availability of epidermal growth factor-related peptides. *Cancer Res* **2002**, *62*, 3151–3158.

[216] Xu, X. *et al.* Breast Cancer. *Clin Cancer Res* **2017**, *23*, 5123–5134.

[217] Siravegna, G. *et al.* Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat Med* **2015**, *21*, 795–801.

[218] Van Emburgh, B. O. *et al.* Acquired RAS or EGFR mutations and duration of response to EGFR blockade in colorectal cancer. *Nat Commun* **2016**, *7*, 13665.

[219] Kazazi-Hyseni, F.; Beijnen, J. H.; Schellens, J. H. Bevacizumab. *Oncologist* **2010**, *15*, 819–825.

[220] Itatani, Y.; Kawada, K.; Yamamoto, T.; Sakai, Y. Resistance to Anti-Angiogenic Therapy in Cancer-Alterations to Anti-VEGF Pathway. *Int J Mol Sci* **2018**, *19*.

[221] Hosaka, K.; Andersson, P.; Wu, J.; He, X.; Du, Q.; Jing, X.; Seki, T.; Gao, J.; Zhang, Y.; Sun, X.; Huang, P.; Yang, Y.; Ge, M.; Cao, Y. KRAS mutation-driven angiopoietin 2 bestows anti-VEGF resistance in epithelial carcinomas. *Proc Natl Acad Sci U S A* **2023**, *120*, e2303740120.

[222] Yamaoka, T.; Kusumoto, S.; Ando, K.; Ohba, M.; Ohmori, T. Receptor Tyrosine Kinase-Targeted Cancer Therapy. *Int J Mol Sci* **2018**, *19*.

[223] Cohen, P.; Cross, D.; nne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat Rev Drug Discov* **2021**, *20*, 551–569.

[224] Lin, Y.; Wang, X.; Jin, H. EGFR-TKI resistance in NSCLC patients: mechanisms and strategies. *Am J Cancer Res* **2014**, *4*, 411–435.

[225] Wu, J. Y.; Shih, J. Y.; Chen, K. Y.; Yang, C. H.; Yu, C. J.; Yang, P. C. Gefitinib therapy in patients with advanced non-small cell lung cancer with or without testing for epidermal growth factor receptor (EGFR) mutations. *Medicine (Baltimore)* **2011**, *90*, 159–167.

[226] Takahama, T. *et al.* Plasma screening for the T790M mutation of EGFR and phase 2 study of osimertinib efficacy in plasma T790M-positive non-small cell lung cancer: West Japan Oncology Group 8815L/LPS study. *Cancer* **2020**, *126*, 1940–1948.

[227] Gomatou, G.; Syrigos, N.; Kotteas, E. Osimertinib Resistance: Molecular Mechanisms and Emerging Treatment Options. *Cancers (Basel)* **2023**, *15*.

[228] Liu, Y.; Li, Y.; Wang, Y.; Lin, C.; Zhang, D.; Chen, J.; Ouyang, L.; Wu, F.; Zhang, J.; Chen, L. Recent progress on vascular endothelial growth factor receptor inhibitors with dual targeting capabilities for tumor therapy. *J Hematol Oncol* **2022**, *15*, 89.

[229] Watt, A. C.; Goel, S. Cellular mechanisms underlying response and resistance to CDK4/6 inhibitors in the treatment of hormone receptor-positive breast cancer. *Breast Cancer Res* **2022**, *24*, 17.

[230] Condorelli, R.; Spring, L.; O'Shaughnessy, J.; Lacroix, L.; Bailleux, C.; Scott, V.; Dubois, J.; Nagy, R. J.; Lanman, R. B.; Iafrate, A. J.; Andre, F.; Bardia, A. Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. *Ann Oncol* **2018**, *29*, 640–645.

[231] Barker, C. A.; Powell, S. N. Enhancing radiotherapy through a greater understanding of homologous recombination. *Semin Radiat Oncol* **2010**, *20*, 267–273.

[232] Min, A.; Im, S. A. PARP Inhibitors as Therapeutics: Beyond Modulation of PARylation. *Cancers (Basel)* **2020**, *12*.

[233] Dobosz, P.; tkowski, T. The Intriguing History of Cancer Immunotherapy. *Front Immunol* **2019**, *10*, 2965.

[234] Decker, W. K.; da Silva, R. F.; Sanabria, M. H.; Angelo, L. S.; es, F.; Burt, B. M.; Kheradmand, F.; Paust, S. Cancer Immunotherapy: Historical Perspective of a Clinical Revolution and Emerging Preclinical Animal Models. *Front Immunol* **2017**, *8*, 829.

[235] Miller, J. F.; Mitchell, G. F. The thymus and the precursors of antigen reactive cells. *Nature* **1967**, *216*, 659–663.

[236] Rosenberg, S. A.; Packard, B. S.; Aebersold, P. M.; Solomon, D.; Topalian, S. L.; Toy, S. T.; Simon, P.; Lotze, M. T.; Yang, J. C.; Seipp, C. A. Use of tumor-infiltrating lymphocytes and interleukin-2 in the immunotherapy of patients with metastatic melanoma. A preliminary report. *N Engl J Med* **1988**, *319*, 1676–1680.

[237] Dudley, M. E. *et al.* Cancer Regression and Autoimmunity in Patients after Clonal Repopulation with Antitumor Lymphocytes. *Science* **2002**, *298*, 850–854.

[238] Angell, T. E.; Lechner, M. G.; Jang, J. K.; LoPresti, J. S.; Epstein, A. L. MHC class I loss is a frequent mechanism of immune escape in papillary thyroid cancer that is reversed by interferon and selumetinib treatment in vitro. *Clin Cancer Res* **2014**, *20*, 6034–6044.

[239] Keane, J. T.; Posey, A. D. Chimeric Antigen Receptors Expand the Repertoire of Antigenic Macromolecules for Cellular Immunity. *Cells* **2021**, *10*.

[240] Sterner, R. C.; Sterner, R. M. CAR-T cell therapy: current limitations and potential strategies. *Blood Cancer J* **2021**, *11*, 69.

[241] Kim, S. P. *et al.* Adoptive Cellular Therapy with Autologous Tumor-Infiltrating Lymphocytes and T-cell Receptor-Engineered T Cells Targeting Common p53 Neoantigens in Human Solid Tumors. *Cancer Immunol Res* **2022**, *10*, 932–946.

[242] Tran, E.; Robbins, P. F.; Lu, Y. C.; Prickett, T. D.; Gartner, J. J.; Jia, L.; Pasetto, A.; Zheng, Z.; Ray, S.; Groh, E. M.; Kriley, I. R.; Rosenberg, S. A. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N Engl J Med* **2016**, *375*, 2255–2262.

[243] Tran, E.; Turcotte, S.; Gros, A.; Robbins, P. F.; Lu, Y. C.; Dudley, M. E.; Wunderlich, J. R.; Somerville, R. P.; Hogan, K.; Hinrichs, C. S.; Parkhurst, M. R.; Yang, J. C.; Rosenberg, S. A. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **2014**, *344*, 641–645.

[244] Zhang, Z.; Chen, X.; Tian, Y.; Li, F.; Zhao, X.; Liu, J.; Yao, C.; Zhang, Y. Point mutation in CD19 facilitates immune escape of B cell lymphoma from CAR-T cell therapy. *Journal for ImmunoTherapy of Cancer* **2020**, *8*.

[245] Levi, S. T. *et al.* Neoantigen Identification and Response to Adoptive Cell Transfer in Anti–PD-1 Naïve and Experienced Patients with Metastatic Melanoma. *Clinical Cancer Research* **2022**, *28*, 3042–3052.

[246] Fraietta, J. A. *et al.* Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia. *Nat Med* **2018**, *24*, 563–571.

[247] Jago, C. B.; Yates, J.; mara, N. O.; Lechler, R. I.; Lombardi, G. Differential expression of CTLA-4 among T cell subsets. *Clin Exp Immunol* **2004**, *136*, 463–471.

[248] Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J Exp Med* **2000**, *192*, 1027–1034.

[249] Yu, G. T.; Bu, L. L.; Zhao, Y. Y.; Mao, L.; Deng, W. W.; Wu, T. F.; Zhang, W. F.; Sun, Z. J. CTLA4 blockade reduces immature myeloid cells in head and neck squamous cell carcinoma. *Oncoimmunology* **2016**, *5*, e1151594.

[250] Krummel, M. F.; Allison, J. P. CD28 and CTLA-4 have opposing effects on the response of T cells to stimulation. *J Exp Med* **1995**, *182*, 459–465.

[251] Leach, D. R.; Krummel, M. F.; Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **1996**, *271*, 1734–1736.

[252] Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* **2010**, *363*, 711–723.

[253] Ishida, Y.; Agata, Y.; Shibahara, K.; Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J* **1992**, *11*, 3887–3895.

[254] Alsaab, H. O.; Sau, S.; Alzhrani, R.; Tatiparti, K.; Bhise, K.; Kashaw, S. K.; Iyer, A. K. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Front Pharmacol* **2017**, *8*, 561.

[255] Rosenberg, S. A.; Yang, J. C.; Sherry, R. M.; Kammula, U. S.; Hughes, M. S.; Phan, G. Q.; Citrin, D. E.; Restifo, N. P.; Robbins, P. F.; Wunderlich, J. R.; Morton, K. E.; Laurencot, C. M.; Steinberg, S. M.; White, D. E.; Dudley, M. E. Durable complete responses in heavily pretreated patients with metastatic melanoma using T-cell transfer immunotherapy. *Clin Cancer Res* **2011**, *17*, 4550–4557.

[256] Peng, W. *et al.* Loss of PTEN Promotes Resistance to T Cell-Mediated Immunotherapy. *Cancer Discov* **2016**, *6*, 202–216.

[257] Gao, Y.; Feng, Y.; Liu, S.; Zhang, Y.; Wang, J.; Qin, T.; Chen, P.; Li, K. Immune-independent acquired resistance to PD-L1 antibody initiated by PD-L1 upregulation via PI3K/AKT signaling can be reversed by anlotinib. *Cancer Med* **2023**, *12*, 15337–15349.

[258] Marcus, L.; Fashoyin-Aje, L. A.; Donoghue, M.; Yuan, M.; Rodriguez, L.; Gallagher, P. S.; Philip, R.; Ghosh, S.; Theoret, M. R.; Beaver, J. A.; Pazdur, R.; Lemery, S. J. FDA Approval Summary: Pembrolizumab for the Treatment of Tumor Mutational Burden-High Solid Tumors. *Clin Cancer Res* **2021**, *27*, 4685–4689.

[259] Merino, D. M. *et al.* Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project. *J Immunother Cancer* **2020**, *8*.

[260] Rentsch, C. A.; user, F. D.; Biot, C.; Gsponer, J. R.; Bisiaux, A.; Wetterauer, C.; Lagranderie, M.; Marchal, G.; Orgeur, M.; Bouchier, C.; Bachmann, A.; Ingersoll, M. A.; Brosch, R.; Albert, M. L.; Thalmann, G. N. rin strain differences have an impact on clinical outcome in bladder cancer immunotherapy. *Eur Urol* **2014**, *66*, 677–688.

[261] Brandau, S.; Suttmann, H. Thirty years of BCG immunotherapy for non-muscle invasive bladder cancer: a success story with room for improvement. *Biomed Pharmacother* **2007**, *61*, 299–305.

[262] Baghban, R.; Ghasemian, A.; Mahmoodi, S. Nucleic acid-based vaccine platforms against the coronavirus disease 19 (COVID-19). *Arch Microbiol* **2023**, *205*, 150.

[263] Liu, J.; Fu, M.; Wang, M.; Wan, D.; Wei, Y.; Wei, X. Cancer vaccines as promising immuno-therapeutics: platforms and current progress. *J Hematol Oncol* **2022**, *15*, 28.

[264] Charoenkwan, P.; Schaduangrat, N.; Shoombuatong, W. StackTTCA: a stacking ensemble learning-based framework for accurate and high-throughput identification of tumor T cell antigens. *BMC Bioinformatics* **2023**, *24*, 301.

[265] An, W. In *Chromatin and Disease*; Kundu, T. K., Bittman, R., Dasgupta, D., Engelhardt, H., Flohe, L., Herrmann, H., Holzenburg, A., Nasheuer, H.-P., Rottem, S., Wyss, M., Zwickl, P., Eds.; Springer Netherlands: Dordrecht, 2007; pp 355–374.

[266] Huang, W.; Yan, J. Profiling selective binding to promoter CpG islands by a single-DNA mechanical footprinting assay. *Biophys J* **2021**, *120*, 3235–3236.

[267] Hu, M.; He, F.; Thompson, E. W.; Ostrikov, K. K.; Dai, X. Lysine Acetylation, Cancer Hallmarks and Emerging Onco-Therapeutic Opportunities. *Cancers (Basel)* **2022**, *14*.

[268] Von Hoff, D. D.; Slavik, M.; Muggia, F. M. 5-Azacytidine. A new anticancer drug with effectiveness in acute myelogenous leukemia. *Ann Intern Med* **1976**, *85*, 237–245.

[269] Jones, P. A.; Taylor, S. M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **1980**, *20*, 85–93.

[270] Zhang, Z.; Wang, G.; Li, Y.; Lei, D.; Xiang, J.; Ouyang, L.; Wang, Y.; Yang, J. Recent progress in DNA methyltransferase inhibitors as anticancer agents. *Front Pharmacol* **2022**, *13*, 1072651.

[271] Gang, A. O.; sig, T. M.; Brimnes, M. K.; Lyngaa, R.; Treppendahl, M. B.; k, K.; Dufva, I. H.; Straten, P. T.; Hadrup, S. R. 5-Azacytidine treatment sensitizes tumor cells to T-cell mediated cytotoxicity and modulates NK cells in patients with myeloid malignancies. *Blood Cancer J* **2014**, *4*, e197.

[272] Becker, J. P.; Helm, D.; Rettel, M.; Stein, F.; Hernandez-Sanchez, A.; Urban, K.; Gebert, J.; Kloor, M.; Neu-Yilik, G.; von Knebel Doeberitz, M.; Hentze, M. W.; Kulozik, A. E. NMD inhibition by 5-azacytidine augments presentation of immunogenic frameshift-derived neoepitopes. *iScience* **2021**, *24*, 102389.

[273] Riggs, M. G.; Whittaker, R. G.; Neumann, J. R.; Ingram, V. M. n-Butyrate causes histone modification in HeLa and Friend erythroleukaemia cells. *Nature* **1977**, *268*, 462–464.

[274] Suraweera, A.; O'Byrne, K. J.; Richard, D. J. Combination Therapy With Histone Deacetylase Inhibitors (HDACi) for the Treatment of Cancer: Achieving the Full Therapeutic Potential of HDACi. *Front Oncol* **2018**, *8*, 92.

[275] Ribrag, V.; Kim, W. S.; Bouabdallah, R.; Lim, S. T.; Coiffier, B.; Illes, A.; Lemieux, B.; Dyer, M. J. S.; Offner, F.; Felloussi, Z.; Kloos, I.; Luan, Y.; Vezan, R.; Graef, T.; Morschhauser, F. Safety and efficacy of abexinostat, a pan-histone deacetylase inhibitor, in non-Hodgkin lymphoma and chronic lymphocytic leukemia: results of a phase II study. *Haematologica* **2017**, *102*, 903–909.

[276] Traina, F. *et al.* Impact of molecular mutations on treatment response to DNMT inhibitors in myelodysplasia and related neoplasms. *Leukemia* **2014**, *28*, 78–87.

[277] Liu, X.; Ling, Z. Q. Role of isocitrate dehydrogenase 1/2 (IDH 1/2) gene mutations in human tumors. *Histol Histopathol* **2015**, *30*, 1155–1160.

[278] Emadi, A.; Faramand, R.; Carter-Cooper, B.; Tolu, S.; Ford, L. A.; Lapidus, R. G.; Wetzler, M.; Wang, E. S.; Etemadi, A.; Griffiths, E. A. Presence of isocitrate dehydrogenase mutations may predict clinical response to hypomethylating agents in patients with acute myeloid leukemia. *Am J Hematol* **2015**, *90*, E77–79.

[279] Ropero, S. *et al.* A truncating mutation of HDAC2 in human cancers confers resistance to histone deacetylase inhibition. *Nat Genet* **2006**, *38*, 566–569.

[280] Fukumoto, T.; Park, P. H.; Wu, S.; Fatkhutdinov, N.; Karakashev, S.; Nacarelli, T.; Kossenkov, A. V.; Speicher, D. W.; Jean, S.; Zhang, L.; Wang, T. L.; Shih, I. M.; Conejo-Garcia, J. R.; Bitler, B. G.; Zhang, R. Repurposing Pan-HDAC Inhibitors for ARID1A-Mutated Ovarian Cancer. *Cell Rep* **2018**, *22*, 3393–3400.

[281] Gupta, S.; Albertson, D. J.; Parnell, T. J.; Butterfield, A.; Weston, A.; Pappas, L. M.; Dalley, B.; O'Shea, J. M.; Lowrance, W. T.; Cairns, B. R.; Schiffman, J. D.; Sharma, S. -Mutated Advanced Urothelial Carcinoma. *Mol Cancer Ther* **2019**, *18*, 185–195.

[282] Yen, K. *et al.* Mutations. *Cancer Discov* **2017**, *7*, 478–493.

[283] Shetty, M. G.; Pai, P.; Deaver, R. E.; Satyamoorthy, K.; Babitha, K. S. Histone deacetylase 2 selective inhibitors: A versatile therapeutic strategy as next generation drug target in cancer therapy. *Pharmacol Res* **2021**, *170*, 105695.

[284] GOODMAN, L. S.; WINTROBE, M. M. Nitrogen mustard therapy; use of methyl-bis (beta-chloroethyl) amine hydrochloride and tris (beta-chloroethyl) amine hydrochloride for Hodgkin's disease, lymphosarcoma, leukemia and certain allied and miscellaneous disorders. *J Am Med Assoc* **1946**, *132*, 126–132.

[285] Devita, V. T.; Serpick, A. A.; Carbone, P. P. Combination chemotherapy in the treatment of advanced Hodgkin's disease. *Ann Intern Med* **1970**, *73*, 881–895.

[286] DeVita, V. T. A selective history of the therapy of Hodgkin's disease. *Br J Haematol* **2003**, *122*, 718–727.

[287] Ralhan, R.; Kaur, J. Alkylating agents and cancer therapy. *Expert Opinion on Therapeutic Patents* **2007**, *17*, 1061–1075.

[288] Noonan, K. L.; Ho, C.; Laskin, J.; Murray, N. The Influence of the Evolution of First-Line Chemotherapy on Steadily Improving Survival in Advanced Non-Small-Cell Lung Cancer Clinical Trials. *J Thorac Oncol* **2015**, *10*, 1523–1531.

[289] Zhang, G.; Zhang, J.; Zhu, Y.; Liu, H.; Shi, Y.; Mi, K.; Li, M.; Zhao, Q.; Huang, Z.; Huang, J. Association of somatic mutations in BRCA2 BRC domain with chemotherapy sensitivity and survival in high grade serous ovarian cancer. *Exp Cell Res* **2021**, *406*, 112742.

[290] Yang, D.; Khan, S.; Sun, Y.; Hess, K.; Shmulevich, I.; Sood, A. K.; Zhang, W. Association of BRCA1 and BRCA2 mutations with survival, chemotherapy

sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* **2011**, *306*, 1557–1565.

[291] Van Allen, E. M. *et al.* Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. *Cancer Discov* **2014**, *4*, 1140–1153.

[292] Slater, S. *et al.* ctDNA guided adjuvant chemotherapy versus standard of care adjuvant chemotherapy after curative surgery in patients with high risk stage II or stage III colorectal cancer: a multi-centre, prospective, randomised control trial (TRACC Part C). *BMC Cancer* **2023**, *23*, 257.

[293] Ingrand, I.; Defossez, G.; Lafay-Chebassier, C.; Chavant, F.; Ferru, A.; Ingrand, P.; rault Pochat, M. C. Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey. *Br J Clin Pharmacol* **2020**, *86*, 711–722.

[294] Aidan, J. C.; Priddee, N. R.; McAleer, J. J. Chemotherapy causes cancer! A case report of therapy related acute myeloid leukaemia in early stage breast cancer. *Ulster Med J* **2013**, *82*, 97–99.

[295] of Health Cancer Policy Team, D. Radiotherapy in England 2012. *Published to Delphi* **2012**,

[296] Douglas, B. G.; Fowler, J. F. The Effect of Multiple Small Doses of X Rays on Skin Reactions in the Mouse and a Basic Interpretation. *Radiation Research* **1976**, *66*, 401–426.

[297] Orth, M.; Lauber, K.; Niyazi, M.; Friedl, A. A.; Li, M.; fer, C.; ttrumpf, L.; Ernst, A.; ller, O. M.; Belka, C. Current concepts in clinical radiation oncology. *Radiat Environ Biophys* **2014**, *53*, 1–29.

[298] van Leeuwen, C. M.; Oei, A. L.; Crezee, J.; Bel, A.; Franken, N. A. P.; Stalpers, L. J. A.; Kok, H. P. The alfa and beta of tumours: a review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies. *Radiat Oncol* **2018**, *13*, 96.

[299] Ma, J.; Setton, J.; Morris, L.; Albornoz, P. B.; Barker, C.; Lok, B. H.; Sherman, E.; Katabi, N.; Beal, K.; Ganly, I.; Powell, S. N.; Lee, N.; Chan, T. A.; Riaz, N. Genomic analysis of exceptional responders to radiotherapy reveals somatic mutations in ATM. *Oncotarget* **2017**, *8*, 10312–10323.

[300] Bian, L.; Meng, Y.; Zhang, M.; Li, D. MRE11-RAD50-NBS1 complex alterations and DNA damage response: implications for cancer treatment. *Mol Cancer* **2019**, *18*, 169.

[301] Pollard, J. M.; Gatti, R. A. Clinical radiation sensitivity with DNA repair disorders: an overview. *Int J Radiat Oncol Biol Phys* **2009**, *74*, 1323–1331.

[302] Duldulao, M. P.; Lee, W.; Nelson, R. A.; Li, W.; Chen, Z.; Kim, J.; Garcia-Aguilar, J. Mutations in specific codons of the KRAS oncogene are associated with variable resistance to neoadjuvant chemoradiation therapy in patients with rectal adenocarcinoma. *Ann Surg Oncol* **2013**, *20*, 2166–2171.

[303] Jeong, Y. *et al.* Role of KEAP1/NRF2 and TP53 Mutations in Lung Squamous Cell Carcinoma Development and Radiation Resistance. *Cancer Discov* **2017**, *7*, 86–101.

[304] Eschrich, S.; Zhang, H.; Zhao, H.; Boulware, D.; Lee, J. H.; Bloom, G.; Torres-Roca, J. F. Systems biology modeling of the radiation sensitivity network: a biomarker discovery platform. *Int J Radiat Oncol Biol Phys* **2009**, *75*, 497–505.

[305] Scott, J. G. *et al.* A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. *Lancet Oncol* **2017**, *18*, 202–211.

[306] Nolan, B.; O'Sullivan, B.; Golden, A. Exploring breast and prostate cancer RNA-seq derived radiosensitivity with the Genomic Adjusted Radiation Dose (GARD) model. *Clin Transl Radiat Oncol* **2022**, *36*, 127–131.

[307] Bukowski, K.; Kciuk, M.; Kontek, R. Mechanisms of Multidrug Resistance in Cancer Chemotherapy. *Int J Mol Sci* **2020**, *21*.

[308] Pagliarini, R.; Shao, W.; Sellers, W. R. Oncogene addiction: pathways of therapeutic response, resistance, and road maps toward a cure. *EMBO Rep* **2015**, *16*, 280–296.

[309] Yu, H. A.; Arcila, M. E.; Rekhtman, N.; Sima, C. S.; Zakowski, M. F.; Pao, W.; Kris, M. G.; Miller, V. A.; Ladanyi, M.; Riely, G. J. Analysis of tumor specimens at the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers. *Clin Cancer Res* **2013**, *19*, 2240–2247.

[310] Choi, P. S.; Li, Y.; Felsher, D. W. Addiction to multiple oncogenes can be exploited to prevent the emergence of therapeutic resistance. *Proc Natl Acad Sci U S A* **2014**, *111*, E3316–3324.

[311] Amin, A. D.; Rajan, S. S.; Groysman, M. J.; Pongtornpipat, P.; Schatz, J. H. Oncogene Overdose: Too Much of a Bad Thing for Oncogene-Addicted Cancer Cells. *Biomark Cancer* **2015**, *7*, 25–32.

[312] Kobayashi, S.; Boggon, T. J.; Dayaram, T.; nne, P. A.; Kocher, O.; Meyerson, M.; Johnson, B. E.; Eck, M. J.; Tenen, D. G.; Halmos, B. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med* **2005**, *352*, 786–792.

[313] Gibbons, D. L.; Pricl, S.; Kantarjian, H.; Cortes, J.; s Cardama, A. The rise and fall of gatekeeper mutations? The BCR-ABL1 T315I paradigm. *Cancer* **2012**, *118*, 293–299.

[314] Whittaker, S.; Kirk, R.; Hayward, R.; Zambon, A.; Viros, A.; Cantarino, N.; Affolter, A.; Nourry, A.; Niculescu-Duvaz, D.; Springer, C.; Marais, R. Gatekeeper mutations mediate resistance to BRAF-targeted therapies. *Sci Transl Med* **2010**, *2*, 35ra41.

[315] Yver, A. Osimertinib (AZD9291)-a science-driven, collaborative approach to rapid drug design and development. *Ann Oncol* **2016**, *27*, 1165–1170.

[316] Chen, S.; Gulla, S.; Cai, C.; Balk, S. P. Androgen receptor serine 81 phosphorylation mediates chromatin binding and transcriptional activation. *J Biol Chem* **2012**, *287*, 8571–8583.

[317] Chen, C. D.; Welsbie, D. S.; Tran, C.; Baek, S. H.; Chen, R.; Vessella, R.; Rosenfeld, M. G.; Sawyers, C. L. Molecular determinants of resistance to antiandrogen therapy. *Nat Med* **2004**, *10*, 33–39.

[318] Corcoran, R. B.; Dias-Santagata, D.; Bergethon, K.; Iafrate, A. J.; Settleman, J.; Engelman, J. A. BRAF gene amplification can promote acquired resistance to MEK inhibitors in cancer cells harboring the BRAF V600E mutation. *Sci Signal* **2010**, *3*, ra84.

[319] Shi, K.; Wang, G.; Pei, J.; Zhang, J.; Wang, J.; Ouyang, L.; Wang, Y.; Li, W. Emerging strategies to overcome resistance to third-generation EGFR inhibitors. *J Hematol Oncol* **2022**, *15*, 94.

[320] Qin, K.; Hong, L.; Zhang, J.; Le, X. MET Amplification as a Resistance Driver to TKI Therapies in Lung Cancer: Clinical Challenges and Opportunities. *Cancers (Basel)* **2023**, *15*.

[321] Cheng, H.; Shcherba, M.; Pendurti, G.; Liang, Y.; Piperdi, B.; Perez-Soler, R. Targeting the PI3K/AKT/mTOR pathway: potential for lung cancer treatment. *Lung Cancer Manag* **2014**, *3*, 67–75.

[322] Aboubakar Nana, F.; Ocak, S. -Mutant Non-Small-Cell Lung Cancer. *Pharmaceutics* **2021**, *13*.

[323] Engelman, J. A. *et al.* Allelic dilution obscures detection of a biologically significant resistance mutation in EGFR-amplified lung cancer. *J Clin Invest* **2006**, *116*, 2695–2706.

[324] Lin, S. H. *et al.* Genes suppressed by DNA methylation in non-small cell lung cancer reveal the epigenetics of epithelial-mesenchymal transition. *BMC Genomics* **2014**, *15*, 1079.

[325] Sun, Y. L.; Patel, A.; Kumar, P.; Chen, Z. S. Role of ABC transporters in cancer chemotherapy. *Chin J Cancer* **2012**, *31*, 51–57.

[326] Huang, Q.; Cai, T.; Bai, L.; Huang, Y.; Li, Q.; Wang, Q.; Chiba, P.; Cai, Y. State of the art of overcoming efflux transporter mediated multidrug resistance of breast cancer. *Transl Cancer Res* **2019**, *8*, 319–329.

[327] Xiao, H.; Zheng, Y.; Ma, L.; Tian, L.; Sun, Q. Clinically-Relevant ABC Transporter for Anti-Cancer Drug Resistance. *Front Pharmacol* **2021**, *12*, 648407.

[328] GRAY, L. H.; CONGER, A. D.; EBERT, M.; HORNSEY, S.; SCOTT, O. C. The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy. *Br J Radiol* **1953**, *26*, 638–648.

[329] Gray, M.; Turnbull, A. K.; Ward, C.; Meehan, J.; rez, C.; Bonello, M.; Pang, L. Y.; Langdon, S. P.; Kunkler, I. H.; Murray, A.; Argyle, D. Development and characterisation of acquired radioresistant breast cancer cell lines. *Radiat Oncol* **2019**, *14*, 64.

[330] Mroz, E. A.; Rocco, J. W. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol* **2013**, *49*, 211–215.

[331] Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **2013**, *152*, 714–726.

[332] Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **2014**, *346*, 256–259.

[333] Smits, H. J. G.; Ruiter, L. N.; Breimer, G. E.; Willems, S. M.; Philippens, M. E. P. Using Intratumor Heterogeneity of Immunohistochemistry Biomarkers to Classify Laryngeal and Hypopharyngeal Tumors Based on Histologic Features. *Mod Pathol* **2023**, *36*, 100199.

[334] Li, M.; Zhang, Z.; Li, L.; Wang, X. An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. *Commun Biol* **2020**, *3*, 505.

[335] Ieni, A.; Vita, R.; Pizzimenti, C.; Benvenga, S.; Tuccari, G. Intratumoral Heterogeneity in Differentiated Thyroid Tumors: An Intriguing Reappraisal in the Era of Personalized Medicine. *J Pers Med* **2021**, *11*.

[336] Beyes, S.; Bediaga, N. G.; Zippo, A. An Epigenetic Perspective on Intra-Tumour Heterogeneity: Novel Insights and New Challenges from Multiple Fields. *Cancers (Basel)* **2021**, *13*.

[337] Guo, M.; Peng, Y.; Gao, A.; Du, C.; Herman, J. G. Epigenetic heterogeneity in cancer. *Biomark Res* **2019**, *7*, 23.

[338] n Y Cajal, S.; é, M.; Capdevila, C.; Aasen, T.; De Mattos-Arruda, L.; Diaz-Cano, S. J.; ndez Losa, J.; í, J. Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med (Berl)* **2020**, *98*, 161–177.

[339] Alic, L.; Niessen, W. J.; Veenland, J. F. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PLoS One* **2014**, *9*, e110300.

[340] Hemenway, G.; Tierno, M. B.; Nejati, R.; Sosa, R.; Zibelman, M. Clinical Utility of Liquid Biopsy to Identify Genomic Heterogeneity and Secondary Cancer Diagnoses: A Case Report. *Case Rep Oncol* **2022**, *15*, 78–85.

[341] Freed, D.; Stevens, E. L.; Pevsner, J. Somatic mosaicism in the human genome. *Genes (Basel)* **2014**, *5*, 1064–1094.

[342] Khoshkhoo, S. *et al.* Contribution of Somatic Ras/Raf/Mitogen-Activated Protein Kinase Variants in the Hippocampus in Drug-Resistant Mesial Temporal Lobe Epilepsy. *JAMA Neurol* **2023**, *80*, 578–587.

[343] Ye, Z. *et al.* Cerebrospinal fluid liquid biopsy for detecting somatic mosaicism in brain. *Brain Commun* **2021**, *3*, fcaa235.

[344] Li, Z.; Min, S.; Alliey-Rodriguez, N.; Giase, G.; Cheng, L.; Craig, D. W.; Faulkner, G. J.; Asif, H.; Liu, C.; Gershon, E. S. Single-neuron whole genome sequencing identifies increased somatic mutation burden in Alzheimer's disease related genes. *Neurobiol Aging* **2023**, *123*, 222–232.

[345] Beck, J. A.; Poulter, M.; Campbell, T. A.; Uphill, J. B.; Adamson, G.; Geddes, J. F.; Revesz, T.; Davis, M. B.; Wood, N. W.; Collinge, J.; Tabrizi, S. J. Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. *Hum Mol Genet* **2004**, *13*, 1219–1224.

[346] Biesecker, L. The challenges of Proteus syndrome: diagnosis and management. *Eur J Hum Genet* **2006**, *14*, 1151–1157.

[347] Comi, A. M. Sturge-Weber syndrome. *Handb Clin Neurol* **2015**, *132*, 157–168.

[348] Boyce, A. M.; Collins, M. T. s Activation. *Endocr Rev* **2020**, *41*, 345–370.

[349] Keppler-Noreuil, K. M.; Parker, V. E.; Darling, T. N.; Martinez-Agosto, J. A. Somatic overgrowth disorders of the PI3K/AKT/mTOR pathway and therapeutic strategies. *Am J Med Genet C Semin Med Genet* **2016**, *172*, 402–421.

[350] Ours, C. A.; Sapp, J. C.; Hodges, M. B.; de Moya, A. J.; Biesecker, L. G. Case report: five-year experience of AKT inhibition with miransertib (MK-7075) in an individual with Proteus syndrome. *Cold Spring Harb Mol Case Stud* **2021**, *7*.

[351] Biesecker, L. G.; Edwards, M.; O'Donnell, S.; Doherty, P.; MacDougall, T.; Tith, K.; Kazakin, J.; Schwartz, B. Clinical report: one year of treatment of Proteus syndrome with miransertib (ARQ 092). *Cold Spring Harb Mol Case Stud* **2020**, *6*.

[352] Forde, K.; Resta, N.; Ranieri, C.; Rea, D.; Kubassova, O.; Hinton, M.; Andrews, K. A.; Semple, R.; Irvine, A. D.; Dvorakova, V. Clinical experience with the AKT1 inhibitor miransertib in two children with PIK3CA-related overgrowth syndrome. *Orphanet J Rare Dis* **2021**, *16*, 109.

[353] Sebold, A. J. *et al.* Sirolimus Treatment in Sturge-Weber Syndrome. *Pediatr Neurol* **2021**, *115*, 29–40.

[354] Van Trigt, W. K.; Kelly, K. M.; Hughes, C. C. W. GNAQ mutations drive port wine birthmark-associated Sturge-Weber syndrome: A review of pathobiology, therapies, and current models. *Front Hum Neurosci* **2022**, *16*, 1006027.

[355] et al, M. E. M. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **1993**, *72*, 971–983.

[356] Lee, J. M. *et al.* CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell* **2019**, *178*, 887–900.

[357] Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **2015**, *348*, 880–886.

[358] Martincorena, I.; Fowler, J. C.; Wabik, A.; Lawson, A. R. J.; Abascal, F.; Hall, M. W. J.; Cagan, A.; Murai, K.; Mahbubani, K.; Stratton, M. R.; Fitzgerald, R. C.; Handford, P. A.; Campbell, P. J.; Saeb-Parsy, K.; Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *Science* **2018**, *362*, 911–917.

[359] Brunner, S. F.; Roberts, N. D.; Wylie, L. A.; Moore, L.; Aitken, S. J.; Davies, S. E.; Sanders, M. A.; Ellis, P.; Alder, C.; Hooks, Y.; Abascal, F.; Stratton, M. R.; Martincorena, I.; Hoare, M.; Campbell, P. J. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **2019**, *574*, 538–542.

[360] Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **2019**, *574*, 532–537.

[361] Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **2017**, *130*, 742–752.

[362] Sacks, D. *et al.* Multisociety Consensus Quality Improvement Revised Consensus Statement for Endovascular Therapy of Acute Ischemic Stroke. *Int J Stroke* **2018**, *13*, 612–632.

[363] Weeks, L. D. *et al.* Prediction of risk for myeloid malignancy in clonal hematopoiesis. *NEJM Evid* **2023**, *2*.

[364] Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet* **2012**, *44*, 1179–1181.

[365] Mensah, G. A.; Roth, G. A.; Fuster, V. Factors: 2020 and Beyond. *J Am Coll Cardiol* **2019**, *74*, 2529–2532.

[366] Arends, C. M. *et al.* Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke. *Blood* **2023**, *141*, 787–799.

[367] Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med* **2017**, *377*, 111–121.

[368] Tall, A. R.; Fuster, J. J. Clonal hematopoiesis in cardiovascular disease and therapeutic implications. *Nat Cardiovasc Res* **2022**, *1*, 116–124.

[369] Avery, O. T.; Macleod, C. M.; McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med* **1944**, *79*, 137–158.

[370] FRANKLIN, R. E.; GOSLING, R. G. Molecular configuration in sodium thymonucleate. *Nature* **1953**, *171*, 740–741.

[371] WATSON, J. D.; CRICK, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **1953**, *171*, 964–967.

[372] MESELSON, M.; STAHL, F. W. The replication of DNA. *Cold Spring Harb Symp Quant Biol* **1958**, *23*, 9–12.

[373] BRENNER, S.; JACOB, F.; MESELSON, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **1961**, *190*, 576–581.

[374] GROS, F.; HIATT, H.; GILBERT, W.; KURLAND, C. G.; RISEBROUGH, R. W.; WATSON, J. D. Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. *Nature* **1961**, *190*, 581–585.

[375] Jacob, F.; Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **1961**, *3*, 318–356.

[376] CRICK, F. H.; BARNETT, L.; BRENNER, S.; WATTS-TOBIN, R. J. General nature of the genetic code for proteins. *Nature* **1961**, *192*, 1227–1232.

[377] Nirenberg, M.; Caskey, T.; Marshall, R.; Brimacombe, R.; Kellogg, D.; Doctor, B.; Hatfield, D.; Levin, J.; Rottman, F.; Pestka, S.; Wilcox, M.; Anderson, F. The RNA code and protein synthesis. *Cold Spring Harb Symp Quant Biol* **1966**, *31*, 11–24.

[378] Sanger, F.; Nicklen, S.; Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **1977**, *74*, 5463–5467.

[379] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.

[380] Mullis, K. B.; Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **1987**, *155*, 335–350.

[381] Saiki, R. K.; Gelfand, D. H.; Stoffel, S.; Scharf, S. J.; Higuchi, R.; Horn, G. T.; Mullis, K. B.; Erlich, H. A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **1988**, *239*, 487–491.

[382] Higuchi, R.; Dollinger, G.; Walsh, P. S.; Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)* **1992**, *10*, 413–417.

[383] Chamberlain, J. S.; Gibbs, R. A.; Ranier, J. E.; Nguyen, P. N.; Caskey, C. T. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res* **1988**, *16*, 11141–11156.

[384] Tsien, Y., Roger; Ross, P.; Fahnestock, M.; Johnston, J., Allan DNA SEQUENCING (Sequencing with removable 3' blockers). WO9106678A1·1991-05-16.

[385] Mayer, P.; Farinelli, L.; Kawashima, E. METHOD OF NUCLEIC ACID AMPLIFICATION (DNA colony sequencing). WO9844151A1·1998-10-08.

[386] Foundation Medicine, I. FoundationOne®CDx (F1CDx). 2020-6-16, P170019/S016.

[387] Lin, L. H.; Chang, K. W.; Cheng, H. W.; Liu, C. J. Identification of Somatic Mutations in Plasma Cell-Free DNA from Patients with Metastatic Oral Squamous Cell Carcinoma. *Int J Mol Sci* **2023**, *24*.

[388] Diefenbach, R. J.; Lee, J. H.; Kefford, R. F.; Rizos, H. Evaluation of commercial kits for purification of circulating free DNA. *Cancer Genet* **2018**, *228-229*, 21–27.

[389] Huang, C. C.; Du, M.; Wang, L. Bioinformatics Analysis for Circulating Cell-Free DNA in Cancer. *Cancers (Basel)* **2019**, *11*.

[390] Gaffney, E. F.; Riegman, P. H.; Grizzle, W. E.; Watson, P. H. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotech Histochem* **2018**, *93*, 373–386.

[391] Xie, R.; Chung, J. Y.; Ylaya, K.; Williams, R. L.; Guerrero, N.; Nakatsuka, N.; Badie, C.; Hewitt, S. M. Factors influencing the degradation of archival formalin-fixed paraffin-embedded tissue sections. *J Histochem Cytochem* **2011**, *59*, 356–365.

[392] Mager, S. R.; Oomen, M. H.; Morente, M. M.; Ratcliffe, C.; Knox, K.; Kerr, D. J.; Pezzella, F.; Riegman, P. H. Standard operating procedure for the collection of fresh frozen tissue samples. *Eur J Cancer* **2007**, *43*, 828–834.

[393] Auer, H. *et al.* The effects of frozen tissue storage conditions on the integrity of RNA and protein. *Biotech Histochem* **2014**, *89*, 518–528.

[394] Gupta, N. DNA Extraction and Polymerase Chain Reaction. *J Cytol* **2019**, *36*, 116–117.

[395] Quail, M. A.; Swerdlow, H.; Turner, D. J. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **2009**, *Chapter 18*, Unit 18.2.

[396] Singh, R. R. Target Enrichment Approaches for Next-Generation Sequencing Applications in Oncology. *Diagnostics (Basel)* **2022**, *12*.

[397] Chen, F.; Dong, M.; Ge, M.; Zhu, L.; Ren, L.; Liu, G.; Mu, R. The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics Proteomics Bioinformatics* **2013**, *11*, 34–40.

[398] Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2015; `https://qubeshub.org/resources/fastqc`.

[399] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

[400] Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.

[401] McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **2010**, *20*, 1297–1303.

[402] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **2011**, *43*, 491–498.

[403] Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* **2019**,

[404] Cibulskis, K.; Lawrence, M. S.; Carter, S. L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E. S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **2013**, *31*, 213–219.

[405] Kim, S.; Scheffler, K.; Halpern, A. L.; Bekritsky, M. A.; Noh, E.; llberg, M.; Chen, X.; Kim, Y.; Beyter, D.; Krusche, P.; Saunders, C. T. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **2018**, *15*, 591–594.

[406] Larson, D. E.; Harris, C. C.; Chen, K.; Koboldt, D. C.; Abbott, T. E.; Dooling, D. J.; Ley, T. J.; Mardis, E. R.; Wilson, R. K.; Ding, L. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **2012**, *28*, 311–317.

[407] Koboldt, D. C.; Zhang, Q.; Larson, D. E.; Shen, D.; McLellan, M. D.; Lin, L.; Miller, C. A.; Mardis, E. R.; Ding, L.; Wilson, R. K. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **2012**, *22*, 568–576.

[408] Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **2015**, *12*, 623–630.

[409] Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* **2015**, *6*, 10001.

[410] rd, A. B.; Thomassen, M.; nkholm, A. V.; Kruse, T. A.; Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS One* **2016**, *11*, e0151664.

[411] de Schaetzen van Brienen, L.; Larmuseau, M.; Van der Eecken, K.; De Ryck, F.; Robbe, P.; Schuh, A.; Fostier, J.; Ost, P.; Marchal, K. Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples. *BMC Med Genomics* **2020**, *13*, 94.

[412] Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **2001**, *29*, 308–311.

[413] Auton, A. *et al.* A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.

[414] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291.

[415] Sun, J. X.; He, Y.; Sanford, E.; Montesion, M.; Frampton, G. M.; Vignot, S.; Soria, J. C.; Ross, J. S.; Miller, V. A.; Stephens, P. J.; Lipson, D.; Yelensky, R. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput Biol* **2018**, *14*, e1005965.

[416] Yudkina, A. V.; Dvornikova, A. P.; Zharkov, D. O. Variable termination sites of DNA polymerases encountering a DNA-protein cross-link. *PLoS One* **2018**, *13*, e0198480.

[417] Rait, V. K.; Zhang, Q.; Fabris, D.; Mason, J. T.; O'Leary, T. J. Conversions of formaldehyde-modified 2'-deoxyadenosine 5'-monophosphate in conditions modeling formalin-fixed tissue dehydration. *J Histochem Cytochem* **2006**, *54*, 301–310.

[418] Maraschin, B. J.; Silva, V. P.; Rock, L.; Sun, H.; Visioli, F.; Rados, P. V.; Rosin, M. P. Optimizing Fixation Protocols to Improve Molecular Analysis from FFPE Tissues. *Braz Dent J* **2017**, *28*, 82–84.

[419] Matsuo, Y.; Yoshida, T.; Yamashita, K.; Satoh, Y. Reducing DNA damage by formaldehyde in liquid-based cytology preservation solutions to enable the molecular testing of lung cancer specimens. *Cancer Cytopathol* **2018**, *126*, 1011–1021.

[420] Gracia Villacampa, E. *et al.* Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genom* **2021**, *1*, 100065.

[421] Serizawa, M.; Yokota, T.; Hosokawa, A.; Kusafuka, K.; Sugiyama, T.; Tsubosa, Y.; Yasui, H.; Nakajima, T.; Koh, Y. The efficacy of uracil DNA glycosylase pretreatment in amplicon-based massively parallel sequencing with DNA extracted from archived formalin-fixed paraffin-embedded esophageal cancer tissues. *Cancer Genet* **2015**, *208*, 415–427.

[422] Kim, S.; Park, C.; Ji, Y.; Kim, D. G.; Bae, H.; van Vrancken, M.; Kim, D. H.; Kim, K. M. Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine. *J Mol Diagn* **2017**, *19*, 137–146.

[423] Scheffler, K.; Catreux, S.; O'Connell, T.; Jo, H.; Jain, V.; Heyns, T.; Yuan, J.; Murray, L.; Han, J.; Mehio, R. Somatic small-variant calling methods in Illumina DRAGEN™ Secondary Analysis. **2023**,

[424] Park, G.; Park, J. K.; Shin, S. H.; Jeon, H. J.; Kim, N. K. D.; Kim, Y. J.; Shin, H. T.; Lee, E.; Lee, K. H.; Son, D. S.; Park, W. Y.; Park, D. Characterization of background noise in capture-based targeted sequencing data. *Genome Biol* **2017**, *18*, 136.

[425] Ledergerber, C.; Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief Bioinform* **2011**, *12*, 489–497.

[426] Chen, L.; Liu, P.; Evans, T. C.; Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **2017**, *355*, 752–756.

[427] Wang, V. G.; Kim, H.; Chuang, J. H. Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. *PLoS One* **2018**, *13*, e0204912.

[428] Wang, Q.; Kotoula, V.; Hsu, P. C.; Papadopoulou, K.; Ho, J. W. K.; Fountzilas, G.; Giannoulatou, E. Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. *BMC Med Genomics* **2019**, *12*, 181.

[429] Garcia-Prieto, C. A.; nez, F.; Valencia, A.; Porta-Pardo, E. Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. *Bioinformatics* **2022**, *38*, 3181–3191.

[430] Tack, V.; Spans, L.; Schuuring, E.; Keppens, C.; Zwaenepoel, K.; Pauwels, P.; Van Houdt, J.; Dequeker, E. M. C. Describing the Reportable Range Is Important for Reliable Treatment Decisions: A Multiple Laboratory Study for Molecular Tumor Profiling Using Next-Generation Sequencing. *J Mol Diagn* **2018**, *20*, 743–753.

[431] Jennings, L. J.; Arcila, M. E.; Corless, C.; Kamel-Reid, S.; Lubin, I. M.; Pfeifer, J.; Temple-Smolkin, R. L.; Voelkerding, K. V.; Nikiforova, M. N. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* **2017**, *19*, 341–365.

[432] Benjamini, Y.; Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **2012**, *40*, e72.

[433] Aird, D.; Ross, M. G.; Chen, W. S.; Danielsson, M.; Fennell, T.; Russ, C.; Jaffe, D. B.; Nusbaum, C.; Gnirke, A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **2011**, *12*, R18.

[434] Ross, M. G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N. J.; Hegarty, R.; Nusbaum, C.; Jaffe, D. B. Characterizing and measuring bias in sequence data. *Genome Biol* **2013**, *14*, R51.

[435] Chiara, M.; Gioiosa, S.; Chillemi, G.; D'Antonio, M.; Flati, T.; Picardi, E.; Zambelli, F.; Horner, D. S.; Pesole, G.; ò, T. CoVaCS: a consensus variant calling system. *BMC Genomics* **2018**, *19*, 120.

[436] Trevarton, A. J.; Chang, J. T.; Symmans, W. F. Simple combination of multiple somatic variant callers to increase accuracy. *Sci Rep* **2023**, *13*, 8463.

[437] Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems* **2018**, *6*, 271–281.e7.

[438] Fan, Y.; Xi, L.; Hughes, D. S.; Zhang, J.; Zhang, J.; Futreal, P. A.; Wheeler, D. A.; Wang, W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **2016**, *17*, 178.

[439] Radenbaugh, A. J.; Ma, S.; Ewing, A.; Stuart, J. M.; Collisson, E. A.; Zhu, J.; Haussler, D. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One* **2014**, *9*, e111516.

[440] 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* **2009**, *25*, 2865–2871.

[441] Institute, T. B. Indelocator. `https://software.broadinstitute.org/cancer/cga/indelocator`, Accessed: 2019-11-28.

[442] Koboldt, D. C.; Larson, D. E.; Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinformatics* **2013**, *44*, 1–17.

[443] Starks, E. R.; Swanson, L.; Docking, T. R.; Bosdet, I.; Munro, S.; Moore, R. A.; Karsan, A. Assessing Limit of Detection in Clinical Sequencing. *J Mol Diagn* **2021**, *23*, 455–466.

[444] Barbitoff, Y. A.; Abasov, R.; Tvorogova, V. E.; Glotov, A. S.; Predeus, A. V. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics* **2022**, *23*, 155.

[445] Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci Rep* **2016**, *6*, 24607.

[446] Wang, Q.; Jia, P.; Li, F.; Chen, H.; Ji, H.; Hucks, D.; Dahlman, K. B.; Pao, W.; Zhao, Z. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* **2013**, *5*, 91.

[447] Cai, L.; Yuan, W.; Zhang, Z.; He, L.; Chou, K. C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep* **2016**, *6*, 36540.

[448] Denroche, R. E.; Mullen, L.; Timms, L.; Beck, T.; Yung, C. K.; Stein, L.; McPherson, J. D.; Brown, A. M. A cancer cell-line titration series for evaluating somatic classification. *BMC Res Notes* **2015**, *8*, 823.

[449] Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **2016**, *3*, 160025.

[450] Chen, Z.; Yuan, Y.; Chen, X.; Chen, J.; Lin, S.; Li, X.; Du, H. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep* **2020**, *10*, 3501.

[451] Kleensang, A.; Vantangoli, M. M.; Odwin-DaCosta, S.; Andersen, M. E.; Boekelheide, K.; Bouhifd, M.; Fornace, A. J.; Li, H. H.; Livi, C. B.; Madnick, S.; Maertens, A.; Rosenberg, M.; Yager, J. D.; Zhao, L.; Hartung, T. Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Sci Rep* **2016**, *6*, 28994.

[452] Noronha, N.; Ehx, G.; Meunier, M. C.; Laverdure, J. P.; riault, C.; Perreault, C. Major multilevel molecular divergence between THP-1 cells from different biorepositories. *Int J Cancer* **2020**, *147*, 2000–2006.

[453] Huang, W.; Li, L.; Myers, J. R.; Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **2012**, *28*, 593–594.

[454] Wang, Y.; Prosen, D. E.; Mei, L.; Sullivan, J. C.; Finney, M.; Vander Horn, P. B. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Res* **2004**, *32*, 1197–1207.

[455] Kubista, M.; Andrade, J. M.; Bengtsson, M.; Forootan, A.; k, J.; Lind, K.; Sindelka, R.; back, R.; green, B.; mbom, L.; hlberg, A.; Zoric, N. The real-time polymerase chain reaction. *Mol Aspects Med* **2006**, *27*, 95–125.

[456] Bustin, S. A.; Benes, V.; Garson, J. A.; Hellemans, J.; Huggett, J.; Kubista, M.; Mueller, R.; Nolan, T.; Pfaffl, M. W.; Shipley, G. L.; Vandesompele, J.; Wittwer, C. T. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **2009**, *55*, 611–622.

[457] Yang, S.; Rothman, R. E. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect Dis* **2004**, *4*, 337–348.

[458] Bivins, A.; Kaya, D.; Bibby, K.; Simpson, S. L.; Bustin, S. A.; Shanks, O. C.; Ahmed, W. Variability in RT-qPCR assay parameters indicates unreliable SARS-CoV-2 RNA quantification for wastewater surveillance. *Water Res* **2021**, *203*, 117516.

[459] Morley, A. A. Digital PCR: A brief history. *Biomol Detect Quantif* **2014**, *1*, 1–2.

[460] Commission, E. *et al. Overview and recommendations for the application of digital PCR*; Publications Office, 2019.

[461] n Aliana, I.; a Romero, N.; Asensi-Puig, A.; n Navarro, J.; lez Rumayor, V.; Ayuso Sacido, A. Clinical Utility of Liquid Biopsy-Based Actionable Mutations Detected via ddPCR. *Biomedicines* **2021**, *9*.

[462] Taylor, S. C.; Laperriere, G.; Germain, H. Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Sci Rep* **2017**, *7*, 2409.

[463] Hughesman, C. B.; Lu, X. J.; Liu, K. Y.; Zhu, Y.; Poh, C. F.; Haynes, C. A Robust Protocol for Using Multiplexed Droplet Digital PCR to Quantify Somatic Copy Number Alterations in Clinical Tissue Specimens. *PLoS One* **2016**, *11*, e0161274.

[464] Oscorbin, I. P.; Smertina, M. A.; Pronyaeva, K. A.; Voskoboev, M. E.; Boyarskikh, U. A.; Kechin, A. A.; Demidova, I. A.; Filipenko, M. L. Genes Amplification in Non-Small Cell Lung Cancer. *Cancers (Basel)* **2022**, *14*.

[465] Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* **2011**, *83*, 8604–8610.

[466] Dong, L.; Wang, S.; Fu, B.; Wang, J. Evaluation of droplet digital PCR and next generation sequencing for characterizing DNA reference material for KRAS mutation detection. *Sci Rep* **2018**, *8*, 9650.

[467] Ho, K. L.; Ding, J.; Fan, J. S.; Tsui, W. N. T.; Bai, J.; Fan, S. K. Digital Microfluidic Multiplex RT-qPCR for SARS-CoV-2 Detection and Variants Discrimination. *Micromachines (Basel)* **2023**, *14*.

[468] Korotkaja, K.; Zajakina, A. Recombinant Virus Quantification Using Single-Cell Droplet Digital PCR: A Method for Infectious Titer Quantification. *Viruses* **2023**, *15*.

[469] Borchardt, M. A.; Boehm, A. B.; Salit, M.; Spencer, S. K.; Wigginton, K. R.; Noble, R. T. The Environmental Microbiology Minimum Information (EMMI) Guidelines: qPCR and dPCR Quality and Reporting for Environmental Microbiology. *Environ Sci Technol* **2021**, *55*, 10210–10223.

[470] 2009; http://dx.doi.org/10.4135/9781412971942.n180.

[471] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). 2016; http://data.europa.eu/eli/reg/2016/679/oj, Accessed on November 28, 2023.

[472] Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019. 2019; https://www.legislation.gov.uk/uksi/2019/419, Accessed on November 28, 2023.

[473] Regulation (EU) 2016/679, GDPR, Recital 26. 2016; http://data.europa.eu/eli/reg/2016/679/oj, Accessed on November 28, 2023.

[474] Shabani, M.; Marelli, L. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep* **2019**, *20*.

[475] Regulation (EU) 2016/679, GDPR, Article 4(13). 2016; http://data.europa.eu/eli/reg/2016/679/oj, Accessed on November 28, 2023.

[476] Regulation (EU) 2016/679, GDPR, Article 7. 2016; `http://data.europa.eu/eli/reg/2016/679/oj`, Accessed on November 28, 2023.

[477] Vukovic, J.; Ivankovic, D.; Habl, C.; Dimnjakovic, J. Enablers and barriers to the secondary use of health data in Europe: general data protection regulation perspective. *Arch Public Health* **2022**, *80*, 115.

[478] Timmers, M.; Van Veen, E. B.; Maas, A. I. R.; Kompanje, E. J. O. Will the Eu Data Protection Regulation 2016/679 Inhibit Critical Care Research? *Med Law Rev* **2019**, *27*, 59–78.

[479] King, M. C. "The race" to clone BRCA1. *Science* **2014**, *343*, 1462–1465.

[480] Zeughauser, B.; Deutsch, K.; Limary, R.; Ciccarella, A. Health Information Privacy Complaint against Myriad Genetics Laboratories, Inc. 2016; `https://www.aclu.org/sites/default/files/field_document/2016.5.19_hipaa_complaint.pdf`, Accessed on November 28, 2023.

[481] Ray, T. Myriad Genetics to Submit Hereditary Cancer Risk Variants to ClinVar in 2023. 2023; `https://www.precisiononcologynews.com/molecular-diagnostics/myriad-genetics-submit-hereditary-cancer-risk-variants-clinvar-2023`, Accessed on November 28, 2023.

[482] Bustin, S. A.; Jellinger, K. A. Advances in Molecular Medicine: Unravelling Disease Complexity and Pioneering Precision Healthcare. *Int J Mol Sci* **2023**, *24*.

[483] O'Sullivan, B.; Seoighe, C. vcfView: An Extensible Data Visualization and Quality Assurance Platform for Integrated Somatic Variant Analysis. *Cancer Inform* **2020**, *19*, 1176935120972377.

[484] Bian, X.; Zhu, B.; Wang, M.; et al., Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **2018**, *19*, 429.

[485] Danecek, P.; Auton, A.; Abecasis, G.; et al., The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **2011**, *27*, 2156–2158.

[486] Bartha, A.; Gyorffy, B. Comprehensive outline of whole exome sequencing data analysis tools available in clinical oncology. *Cancers* **2019**, *11*, 1725.

[487] Williams, M.; Werner, B.; Heide, T.; et al., Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet* **2018**, *50*, 895–903.

[488] Bagheropur, S.; Ehsanpour, A.; Birgani, M. T.; Saki, N. JAK2V617F allele burden: innovative concept in monitoring of myeloproliferative neoplasms. *Memo* **2018**, *11*, 152–157.

[489] Pratcorona, M.; Brunet, S.; Nomdedéu, J.; et al., Favorable outcome of patients with acute myeloid leukemia harboring a low-allelic burden FLT3-ITD mutation and concomitant NPM1 mutation: relevance to post-remission therapy. *Blood* **2013**, *121*, 2734–2738.

[490] Sallman, D. A.; Padron, E. Integrating mutation variant allele frequency into clinical practice in myeloid malignancies. *Hematol Oncol Stem Cell Ther* **2016**, *9*, 89–95.

[491] Maura, F.; Degasperi, A.; Nadeu, F.; et al., A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun* **2019**, *10*, 2969.

[492] Messner, D. A.; Al Naber, J.; Koay, P.; et al., Barriers to clinical adoption of next generation sequencing: perspectives of a policy Delphi panel. *Appl Transl Genom* **2016**, *10*, 19–24.

[493] Cingolani, P.; Patel, V. M.; Coon, M.; et al., Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* **2012**, *3*, 35.

[494] R Core Team, R: A Language and Environment for Statistical Computing. R Core Team: Vienna, 2019.

[495] Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. RStudio, Inc.: Vienna, 2019.

[496] Obenchain, V.; Lawrence, M.; Carey, V.; Gogarten, S.; Shannon, P.; Morgan, M. VariantAnnotation: a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **2014**, *30*, 2076–2078.

[497] Bsgenome.hsapiens.ucsc.hg38: Full Genome Sequences for Homo Sapiens (UCSC Version Hg38). Bioconductor, 2015; https://bioconductor.riken.jp/packages/3.8/data/annotation/html/BSgenome.Hsapiens.UCS

[498] Txdb.hsapiens.ucsc.hg38.knowngene: Annotation Package for Txdb Object(s). Bioconductor, 2019; http://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg3

[499] Org.hs.eg.db: Genome Wide Annotation for Human. Bioconductor, 2019; https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

[500] Gentleman, R. Annotate: Annotation for Microarrays. Springer-Verlag: New York, NY, 2019.

[501] Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; 2016.

[502] Slowikowski, K. Ggrepel: Automatically Position Non-overlapping Text Labels with 'Ggplot2'. 2019; https://rdrr.io/cran/ggrepel/.

[503] Borchers, H. Pracma: Practical Numerical Math Functions. 2019; https://rdrr.io/cran/pracma/.

[504] Blokzijl, F.; Janssen, R.; van Boxtel, R.; Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine* **2018**, *10*, 33.

[505] Deckers, J.; Hammad, H.; Hoste, E. Langerhans cells: sensing the environment in health and disease. *Front Immunol* **2018**, *9*, 93–93.

[506] Shlush, L. Age-related clonal hematopoiesis. *Blood* **2018**, *131*, 496–504.

[507] Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, e. a., J IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **2013**, *10*, 1081–1082.

[508] Haferlach, T.; Meggendorfer, M. More than a fusion gene: the RUNX1-RUNX1T1 AML. *Blood* **2019**, *133*, 1006–1007.

[509] Prats-Martín, C.; Burillo-Sanz, S.; Morales-Camacho, e. a., RM ASXL1 mutation as a surrogate marker in acute myeloid leukemia with myelodysplasia-related changes and normal karyotype. *Cancer Med* **2020**, *9*, 3637–3646.

[510] Mechaal, A.; Menif, S.; Abbes, S.; Safra, I. EZH2, new diagnosis and prognosis marker in acute myeloid leukemia patients. *Adv Med Sci* **2019**, *64*, 395–401.

[511] Garg, M.; Nagata, Y.; Kanojia, e. a., D Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse. *Blood* **2015**, *126*, 2491–2501.

[512] Rampias, T.; Karagiannis, D.; Avgeris, e. a., M The lysine-specific methyltransferase KMT2C/MLL3 regulates DNA repair components in cancer. *EMBO Rep* **2019**, *20*, e46821.

[513] Lachowiez, C.; Loghavi, S.; Kadia, e. a., TM Outcomes of older patients with NPM1-mutated AML: current treatments and the promise of venetoclax-based regimens. *Blood Advances* **2020**, *4*, 1311–1320.

[514] Przychodzen, B.; Gu, X.; You, e. a., D PHF6 somatic mutations and their functional role in the pathophysiology of myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML). *Blood* **2016**, *128*, 2736–2736.

[515] Jalili, M.; Yaghmaie, M.; Ahmadvand, e. a., M Prognostic value of RUNX1 mutations in AML: a meta-analysis. *APJCP* **2018**, *19*, 325–329.

[516] Wang, R.; Gao, X.; Yu, L. The prognostic impact of tet oncogene family member 2 mutations in patients with acute myeloid leukemia: a systematic-review and meta-analysis. *BMC Cancer* **2019**, *19*, 389.

[517] Barbosa, K.; Li, S.; Adams, P.; Deshpande, A. The role of TP53 in acute myeloid leukemia: challenges and opportunities. *Genes Chromosomes Cancer* **2019**, *58*, 875–888.

[518] Xu, H.; DiCarlo, J.; Satya, R. V.; Peng, Q.; Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **2014**, *15*, 244.

[519] Fittall, M. W.; Van Loo, P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Med* **2019**, *11*, 20.

[520] Frampton, G. M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* **2013**, *31*, 1023–1031.

[521] Caetano-Anolles, D. GATK Best Practices Workflows. Somatic short variant discovery (SNVs + Indels). 2022; `https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-`, Accessed on March 3, 2023.

[522] Bonfield, J. K.; Marshall, J.; Danecek, P.; Li, H.; Ohan, V.; Whitwham, A.; Keane, T.; Davies, R. M. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **2021**, *10*.

[523] Prentice, L. M.; Miller, R. R.; Knaggs, J.; Mazloomian, A.; Aguirre Hernandez, R.; Franchini, P.; Parsa, K.; Tessier-Cloutier, B.; Lapuk, A.; Huntsman, D.; Schaeffer, D. F.; Sheffield, B. S. Formalin fixation increases deamination mutation signature but should not lead to false positive mutations in clinical practice. *PLoS One* **2018**, *13*, e0196434.

[524] Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101.

[525] 2020, R. C. T. R: A Language and Environment for Statistical Computing. 2020; `https://www.R-project.org`, Accessed on March 3, 2023.

[526] Makrooni, M. A.; O'Sullivan, B.; Seoighe, C. Bias and inconsistency in the estimation of tumour mutation burden. *BMC Cancer* **2022**, *22*, 840.

[527] Costello, M.; Pugh, T. J.; Fennell, T. J.; Stewart, C.; Lichtenstein, L.; Meldrim, J. C.; Fostel, J. L.; Friedrich, D. C.; Perrin, D.; Dionne, D.; Kim, S.; Gabriel, S. B.; Lander, E. S.; Fisher, S.; Getz, G. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **2013**, *41*, e67.

[528] Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* **2017**, *9*, 34.

[529] Park, J. Y.; Kricka, L. J.; Fortina, P. Next-generation sequencing in the clinic. *Nat Biotechnol* **2013**, *31*, 990–992.

[530] Clarke, L.; Fairley, S.; Zheng-Bradley, X.; Streeter, I.; Perry, E.; Lowy, E.; é, A. M.; Flicek, P. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res* **2017**, *45*, D854–D859.

[531] Siegel, R. L.; Miller, K. D.; Wagle, N. S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J Clin* **2023**, *73*, 17–48.

[532] Der, C. J.; Krontiris, T. G.; Cooper, G. M. Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc Natl Acad Sci U S A* **1982**, *79*, 3637–3640.

[533] Ostrem, J. M.; Peters, U.; Sos, M. L.; Wells, J. A.; Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **2013**, *503*, 548–551.

[534] Skoulidis, F. *et al.* p.G12C Mutation. *N Engl J Med* **2021**, *384*, 2371–2381.

[535] Kemp, S. B. *et al.* Efficacy of a Small-Molecule Inhibitor of KrasG12D in Immunocompetent Models of Pancreatic Cancer. *Cancer Discov* **2023**, *13*, 298–311.

[536] Luo, J. KRAS mutation in pancreatic cancer. *Semin Oncol* **2021**, *48*, 10–18.

[537] Petrackova, A.; Vasinek, M.; Sedlarikova, L.; Dyskova, T.; Schneiderova, P.; Novosad, T.; Papajik, T.; Kriegova, E. Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics. *Front Oncol* **2019**, *9*, 851.

[538] Wang, S.; You, L.; Dai, M.; Zhao, Y. Mucins in pancreatic cancer: A well-established but promising family for diagnosis, prognosis and therapy. *J Cell Mol Med* **2020**, *24*, 10279–10289.

[539] Yamazoe, S.; Tanaka, H.; Sawada, T.; Amano, R.; Yamada, N.; Ohira, M.; Hirakawa, K. RNA interference suppression of mucin 5AC (MUC5AC) reduces the adhesive and invasive capacity of human pancreatic cancer cells. *J Exp Clin Cancer Res* **2010**, *29*, 53.

[540] Hoshi, H.; Sawada, T.; Uchida, M.; Iijima, H.; Kimura, K.; Hirakawa, K.; Wanibuchi, H. MUC5AC protects pancreatic cancer cells from TRAIL-induced death pathways. *Int J Oncol* **2013**, *42*, 887–893.

[541] Krishn, S. R.; Ganguly, K.; Kaur, S.; Batra, S. K. Ramifications of secreted mucin MUC5AC in malignant journey: a holistic view. *Carcinogenesis* **2018**, *39*, 633–651.

[542] Ganguly, K.; Krishn, S. R.; Jahan, R.; Atri, P.; Rachagani, S.; Rauth, S.; Xi, H.; Lu, Y.; Batra, S.; Kaur, S. Abstract 65: Gel-forming mucin MUC5AC as the nexus for cell-adhesion molecules governing pancreatic cancer aggressiveness and chemoresistance. *Cancer Research* **2019**, *79*, 65–65.

[543] Jonckheere, N.; Skrypek, N.; Van Seuningen, I. Mucins and pancreatic cancer. *Cancers (Basel)* **2010**, *2*, 1794–1812.

[544] Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **2019**, *47*, D941–D947.

[545] Sah, S.; Chen, L.; Houghton, J.; Kemppainen, J.; Marko, A. C.; Zeigler, R.; Latham, G. J. Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies. *Genome Med* **2013**, *5*, 77.

[546] Xiao, Y. L.; Kash, J. C.; Beres, S. B.; Sheng, Z. M.; Musser, J. M.; Taubenberger, J. K. High-throughput RNA sequencing of a formalin-fixed, paraffin-embedded autopsy lung tissue sample from the 1918 influenza pandemic. *J Pathol* **2013**, *229*, 535–545.

[547] Kerick, M.; Isau, M.; Timmermann, B.; ltmann, H.; Herwig, R.; Krobitsch, S.; Schaefer, G.; Verdorfer, I.; Bartsch, G.; Klocker, H.; Lehrach, H.; Schweiger, M. R. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med Genomics* **2011**, *4*, 68.

[548] Rheinbay, E.; Nielsen, M. M.; Abascal, F.; Wala, J. A.; Shapira, O. e. a. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **2020**, *578*, 102–111.

Appendices

Appendix A

```
#!/bin/bash
# Filter script that allow through only PASS variants that demonstrate evidence of absence of FFPE
# Run it as follows..
# ./hFilterFfpe.bash P15A/results/VCF/P15A_MUTECT2.filtered.annotated.vcf.gz
#
# Output is in P15A_MUTECT2.filtered.annotated.extraFiltered_ffpe_OxoG.vcf
outfile=‘echo ${1}| sed -e ’s/.*\///1’ -e ’s/vcf.gz$/extraFiltered_ffpe_OxoG.vcf/1’‘
echo "Hard filtering "${1}
zcat ${1} | egrep -v ’^#’ | \
awk ’{
/* Pass and SNVs only */
if($7=="PASS" && $4 ~ /^[GCAT]$/ && $5 ~ /^[GCAT]$/ )
{
  split($9,format,":")
  split($10,formatContents,":")
  /* Pull out format attributes from VCF sample field */
  for(i in format)
    {
      formatAttribute[format[i]]=formatContents[i]
    }
  /* Pair Orientation */
  /* Pull out num of reads in forward and reverse directions */
  /* F1R2 means first read in the pair aligns to the forward strand, second reverse, etc. */
  /* F1R2 means the first read is a forward one and comes from the reads 1.fq file and */
  /* the second read is a reverse one and comes from the reads 2.fq */
  /* ref=1, alt=2 */
  split(formatAttribute["F1R2"],fwdReads,",")
  split(formatAttribute["F2R1"],revReads,",")
  /* If there is solid evidence for absence of FFPE, */
  /* ie., if there is more than one supporting read from both strands */
  /* btw., supporting read => read that provides evidence of the variant at this locus */
  if(fwdReads[2]>2 && revReads[2]>2)
     print $0
}
}’ > ${outfile}
```

Appendix B

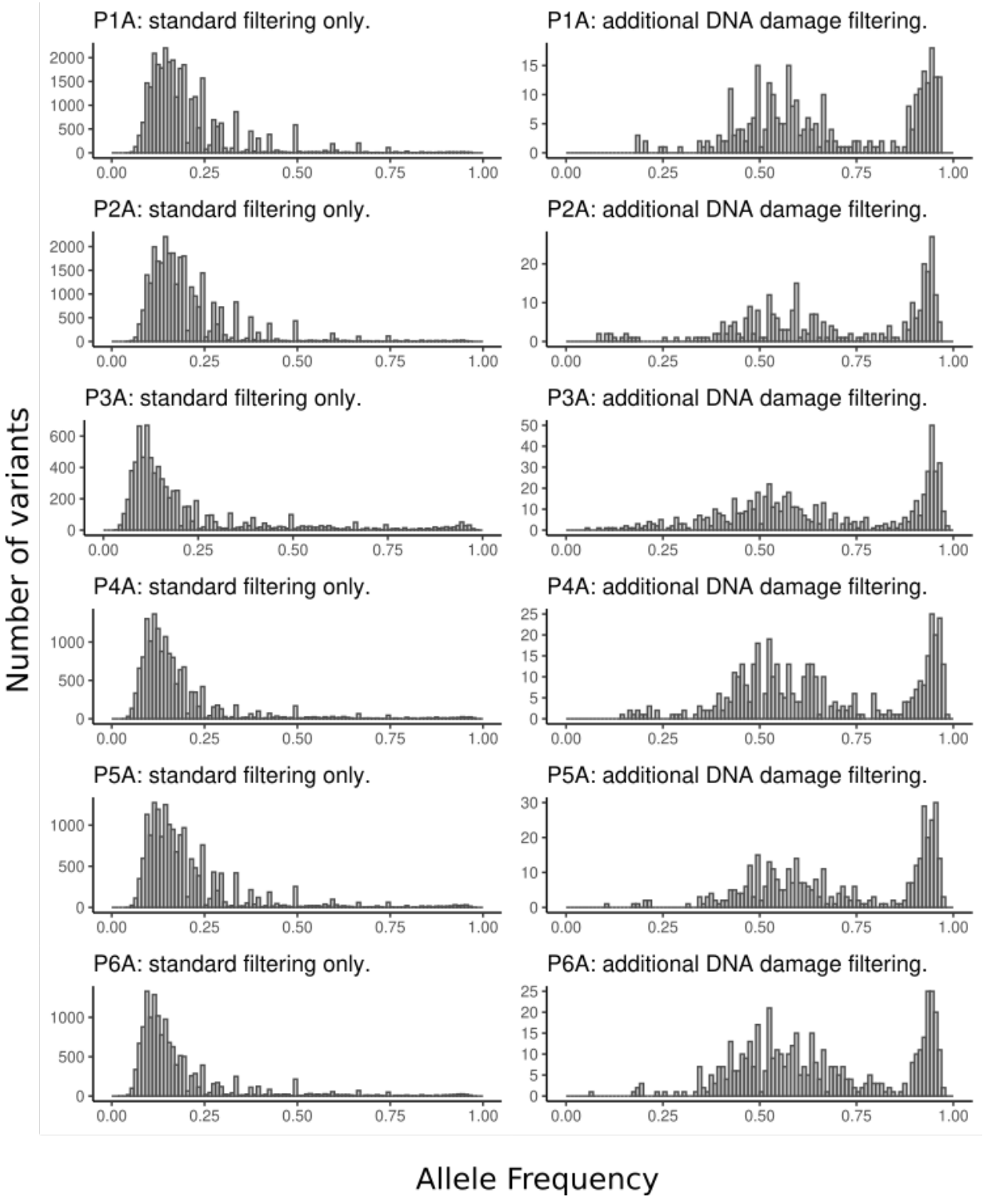

Figure 5.1: Illustration of allele frequency spectra for patients P1A to P6A before and after applying additional DNA damage filtering.

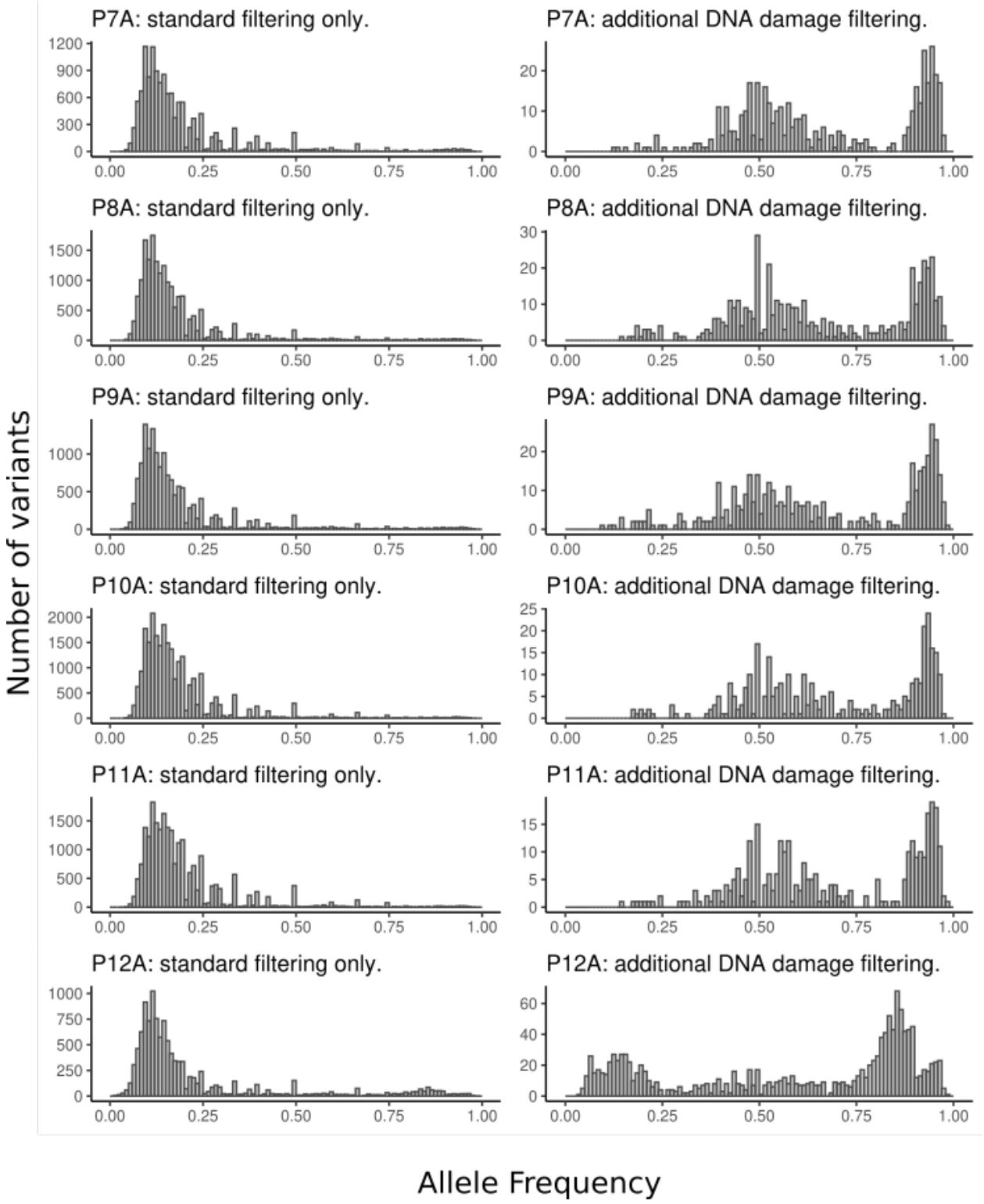

Figure 5.2: Illustration of allele frequency spectra for patients P7A to P12A before and after applying additional DNA damage filtering.

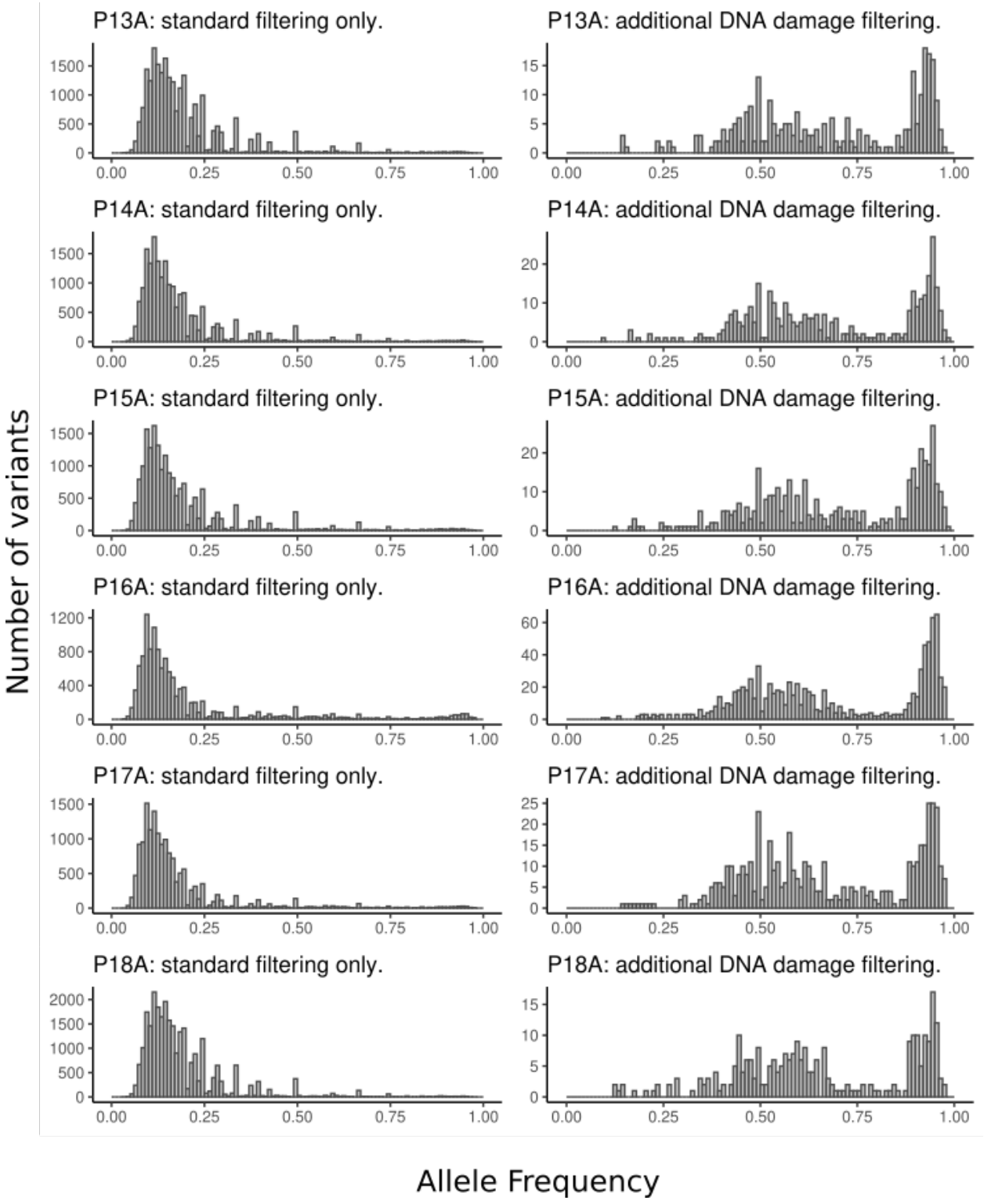

Figure 5.3: Illustration of allele frequency spectra for patients P13A to P18A before and after applying additional DNA damage filtering.

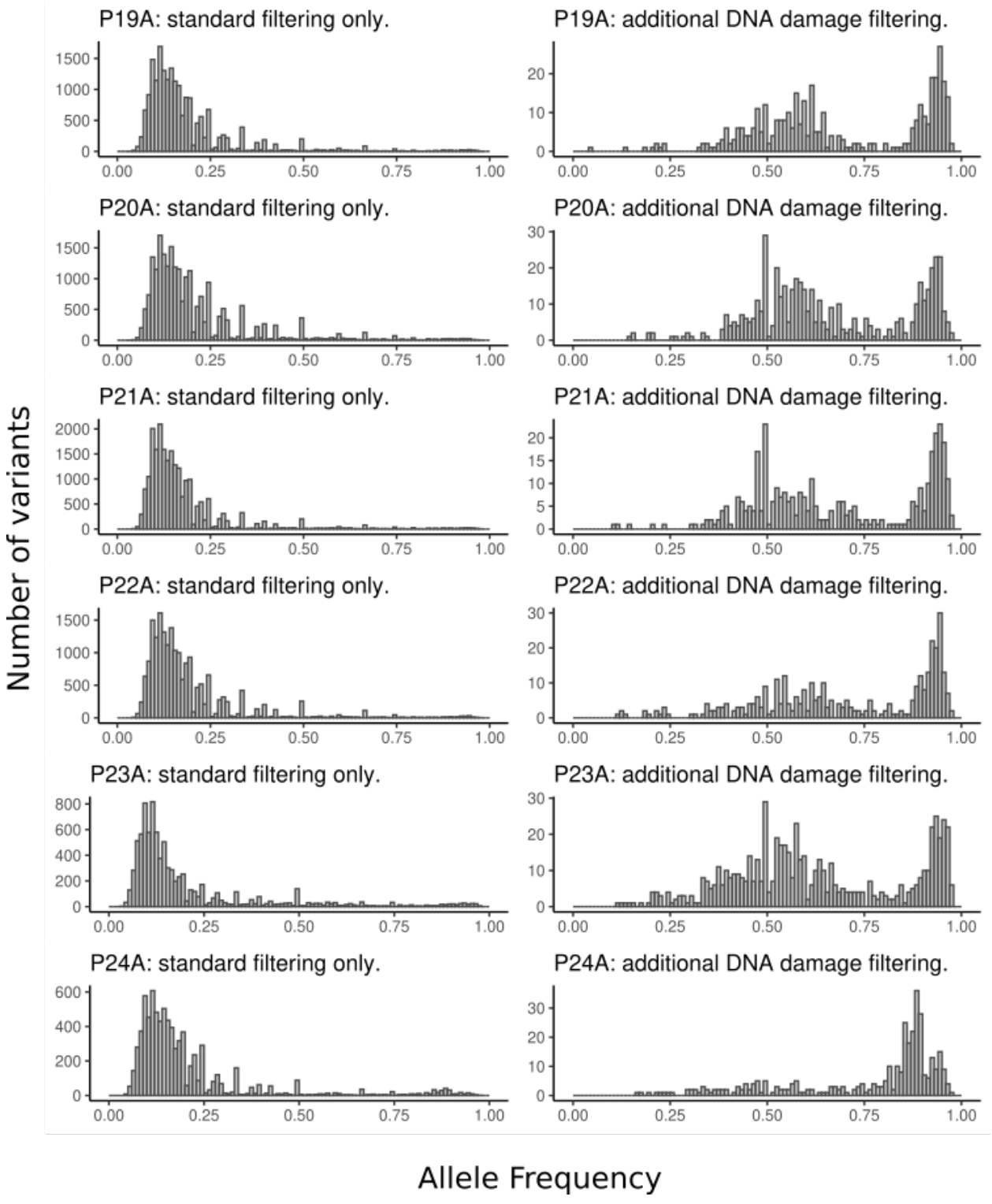

Figure 5.4: Illustration of allele frequency spectra for patients P19A to P24A before and after applying additional DNA damage filtering.

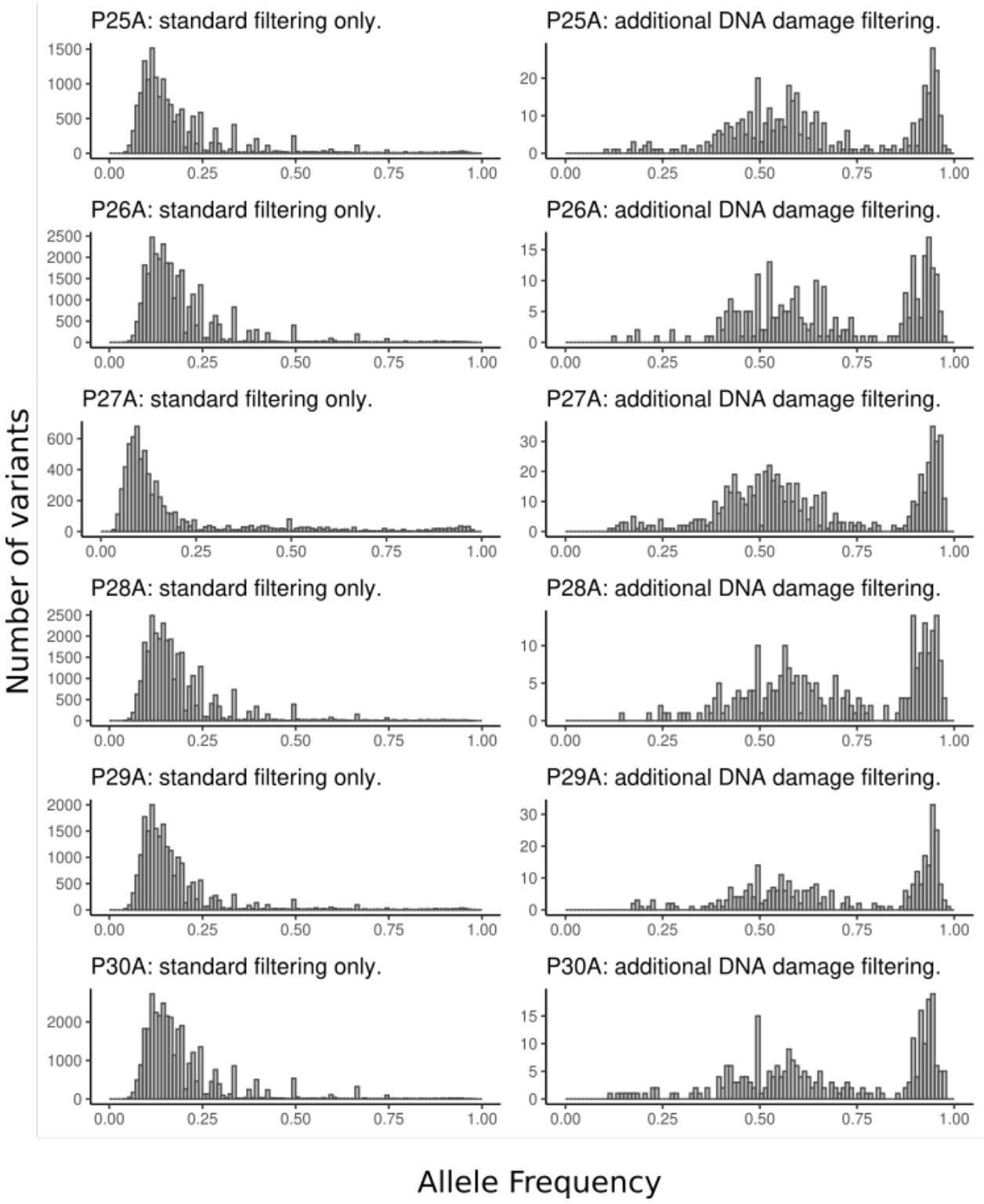

Figure 5.5: Illustration of allele frequency spectra for patients P25A to P30A before and after applying additional DNA damage filtering.

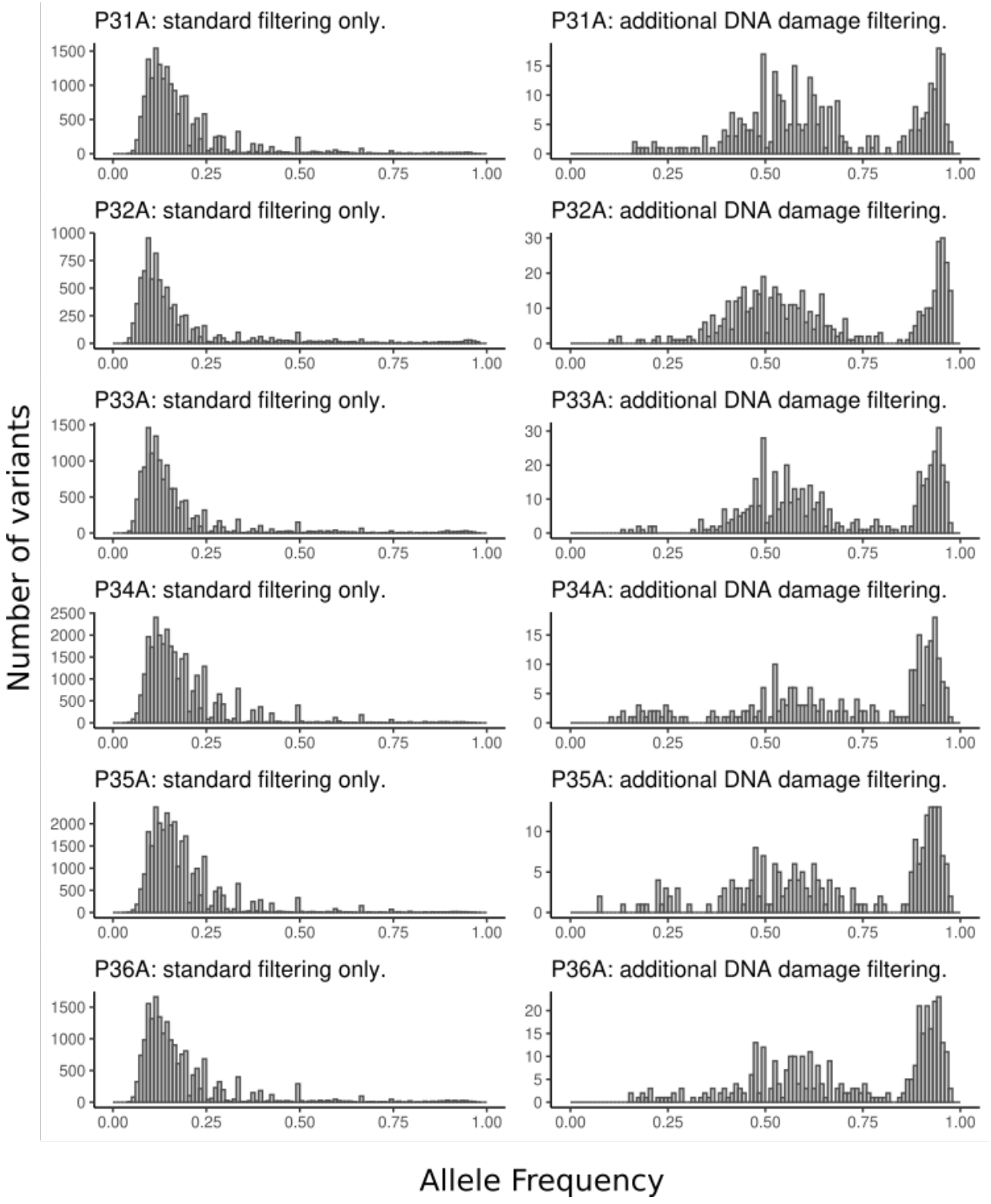

Figure 5.6: Illustration of allele frequency spectra for patients P31A to P36A before and after applying additional DNA damage filtering.

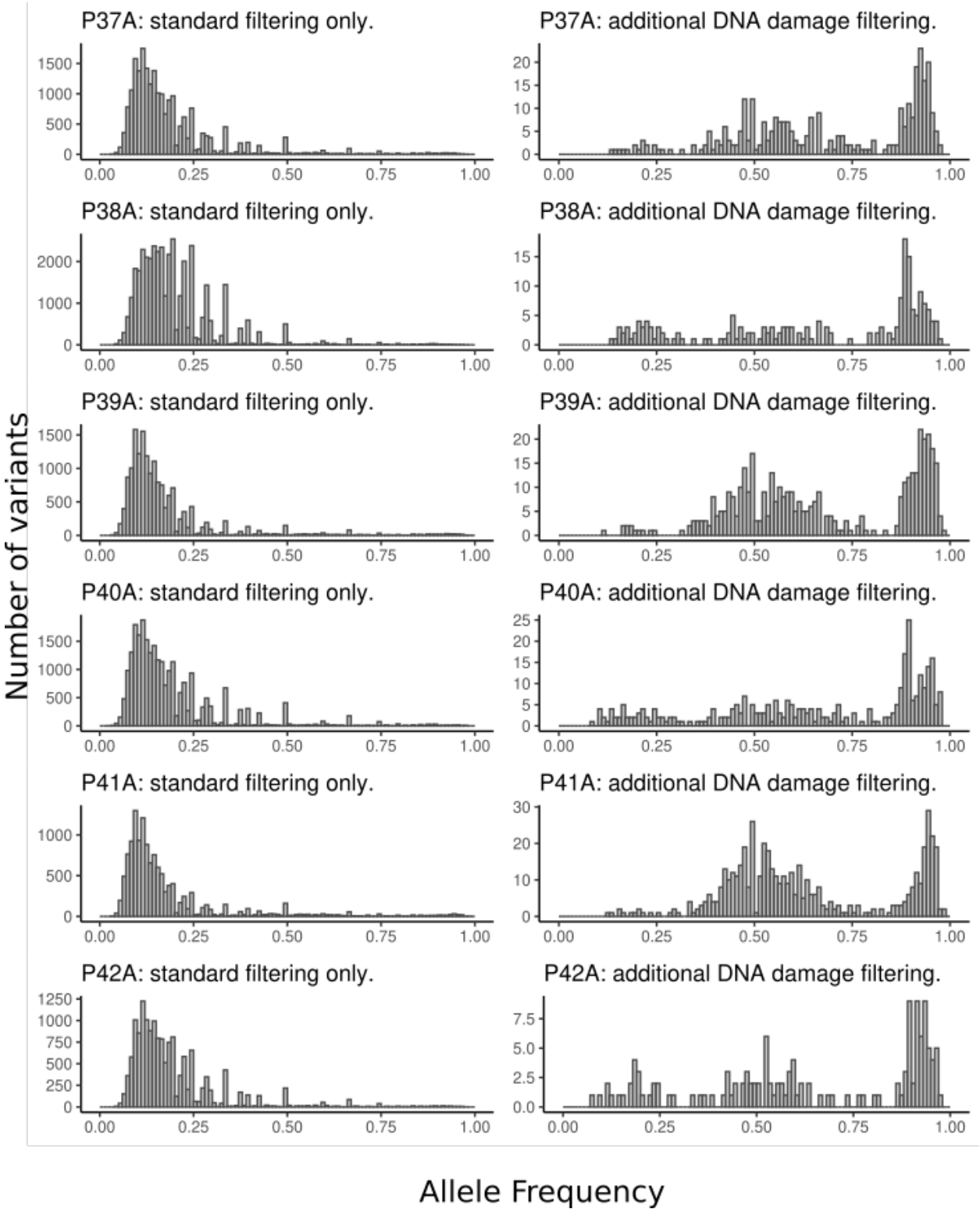

Figure 5.7: Illustration of allele frequency spectra for patients P37A to P42A before and after applying additional DNA damage filtering.

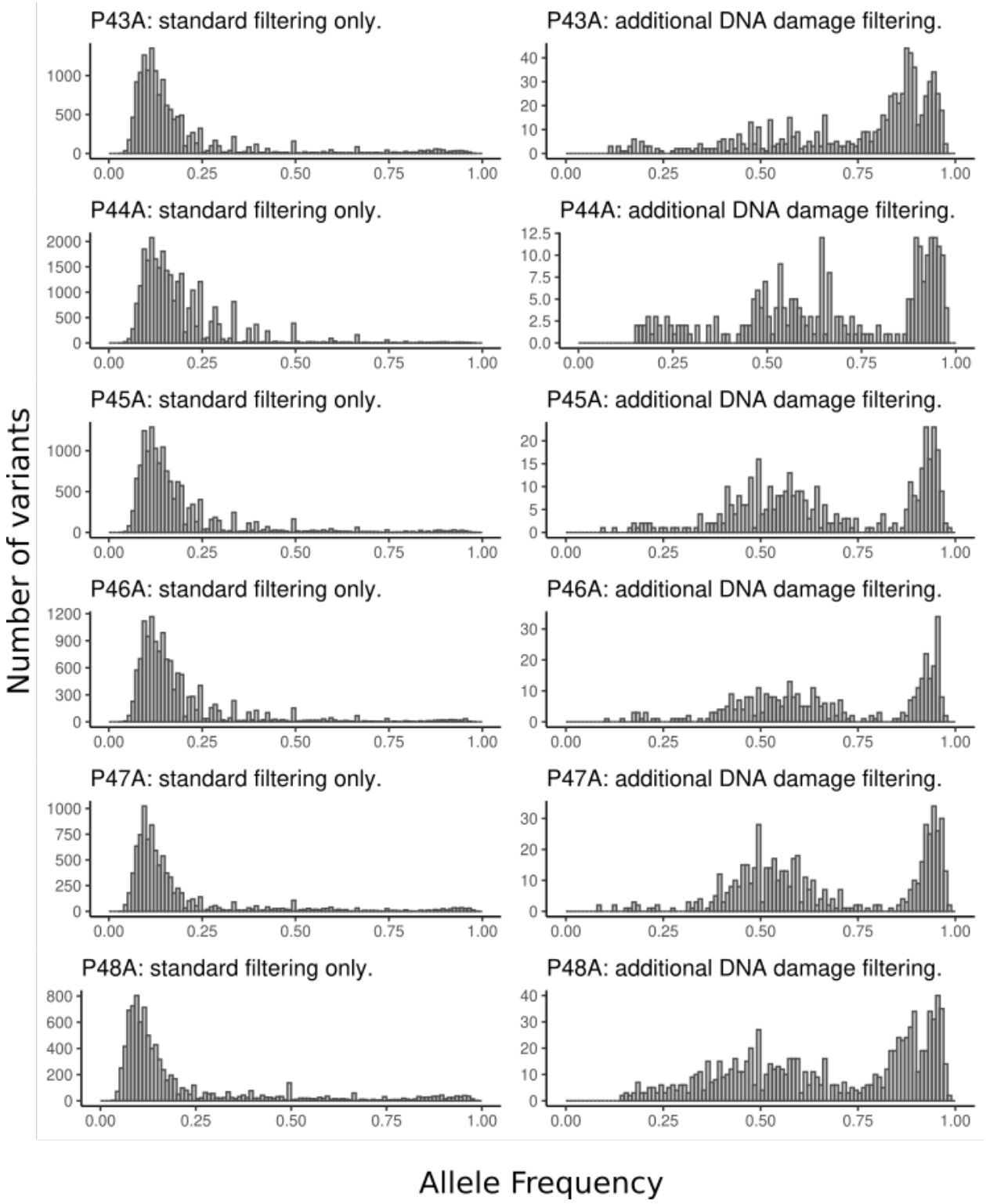

Figure 5.8: Illustration of allele frequency spectra for patients P43A to P48A before and after applying additional DNA damage filtering.

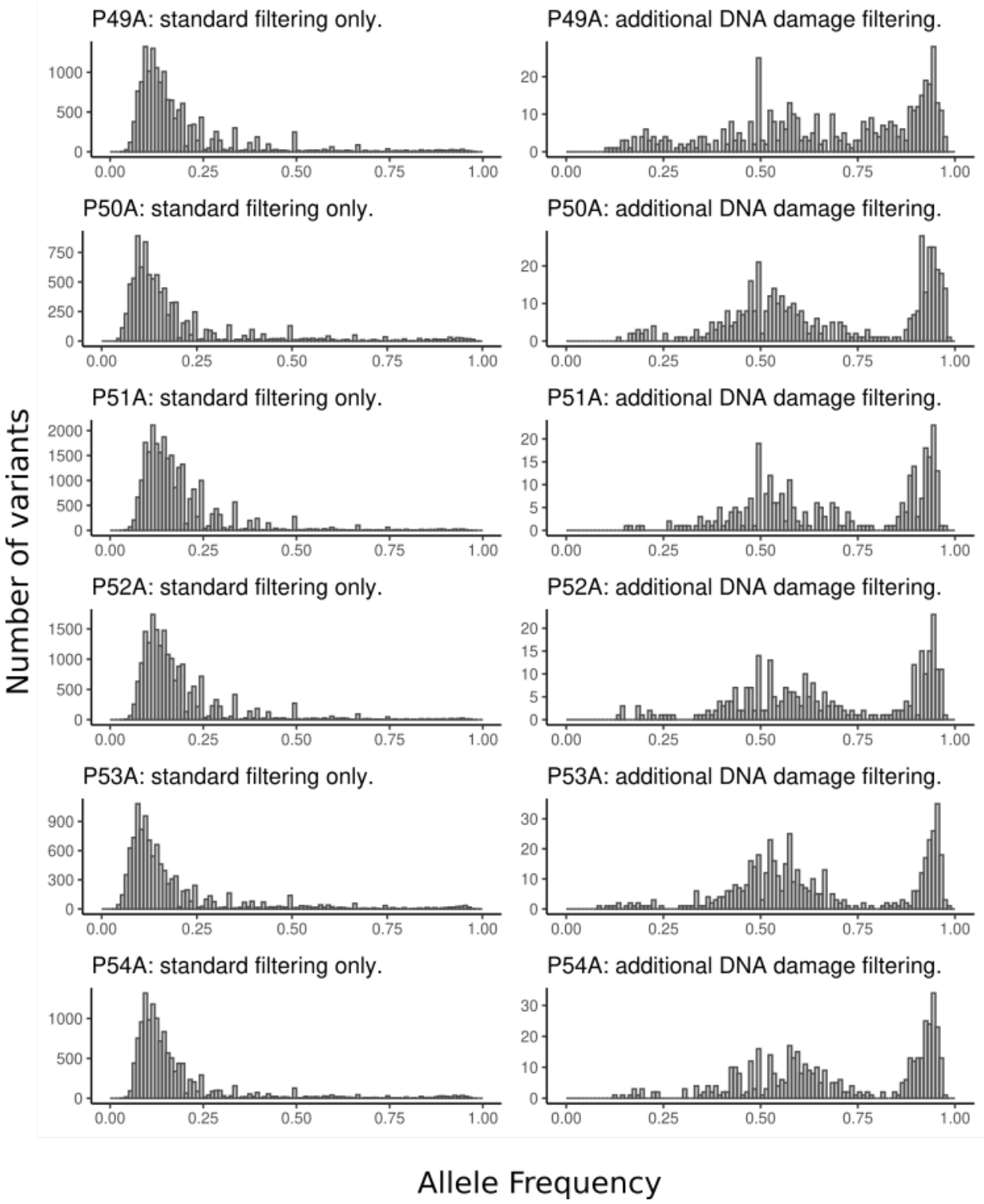

Figure 5.9: Illustration of allele frequency spectra for patients P49A to P54A before and after applying additional DNA damage filtering.

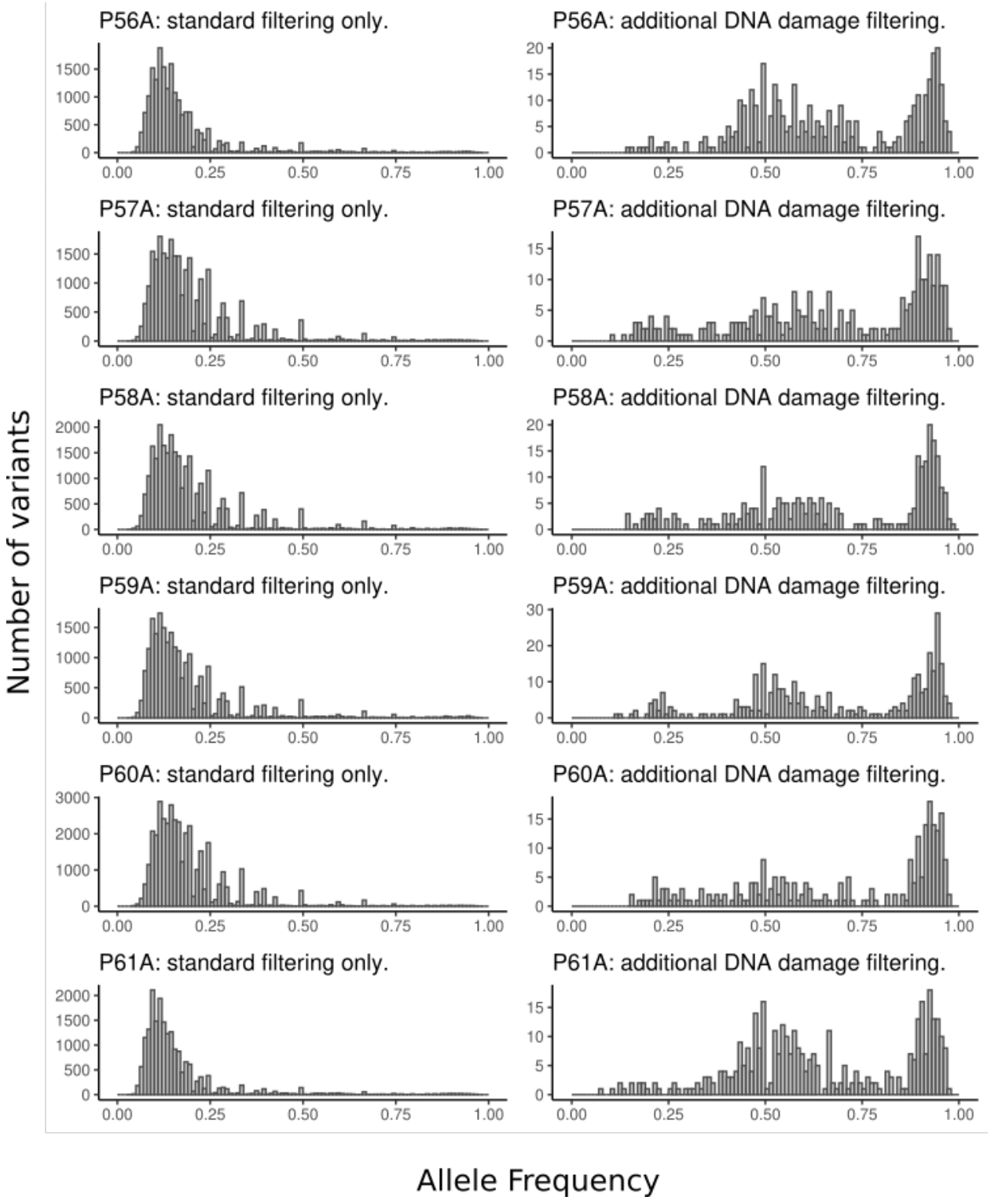

Figure 5.10: Illustration of allele frequency spectra for patients P56A to P61A before and after applying additional DNA damage filtering.

Appendix C

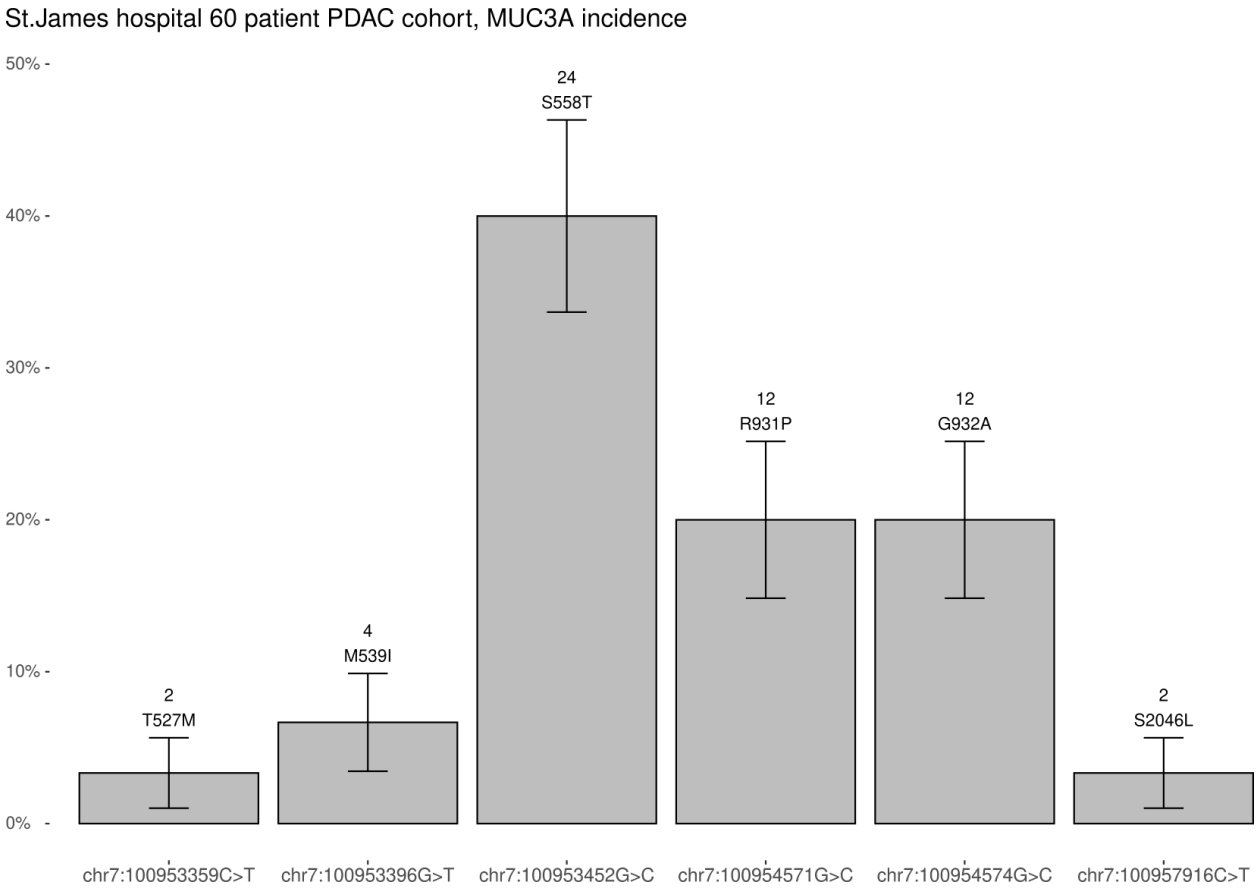

Figure 5.11: Incidence of recurring MUC3A mutations in PDAC cohort.