GERMLINE VARIANT CALLING IN FORMALIN-FIXED PARAFFIN-EMBEDDED TUMOURS

by

Shyong Quin Yap

B.Sc. (Hons), Trent University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Experimental Medicine Program)

The University of British Columbia
(Vancouver)

November 2017

Abstract

This document provides brief instructions for using the `ubcdiss` class to write a **UBC!**-conformant dissertation in LaTeX. This document is itself written using the `ubcdiss` class and is intended to serve as an example of writing a dissertation in LaTeX. This document has embedded Unique Resource Locators (URLs) and is intended to be viewed using a computer-based Portable Document Format (PDF) reader.

Note: Abstracts should generally try to avoid using acronyms.

Note: at **UBC!** (**UBC!**), both the Graduate and Postdoctoral Studies (GPS) Ph.D. defence programme and the Library's online submission system restricts abstracts to 350 words.

Preface

At **UBC!**, a preface may be required. Be sure to check the GPS guidelines as they may have specific content to be included.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

HER2 Human epidermal growth factor receptor 2

GPS Graduate and Postdoctoral Studies

PDF Portable Document Format

URL Unique Resource Locator, used to describe a means for obtaining some resource on the world wide web

Acknowledgments

Chapter 1

Introduction

1.1 The emergence of precision oncology

Cancers are fundamentally a group of genetic disorders. The role of genetic alterations in driving malignant transformation has been implicated in studies dating back to the late nineteenth and early twentieth centuries by David von Hansemann and Theodor Boveri. Both von Hansemann and Boveri observed unequal chromosomal segregations in tumour cells, which prompted their speculations that tumour development is induced by anomalies in hereditary material [22, 138]. Major strides have been made in understanding the molecular basis of cancer, including the discovery of recurrent gene mutations and elucidation of oncogenic pathways. Some of these findings were successfully translated into clinical applications wherein patients who harbour actionable somatic mutations benefitted from treatment with targeted anti-cancer drugs. Notable examples include treatment of *HER2*-overexpressed breast cancer with trastuzumab [6, 46, 115, 119, 121, 137] and treatment of *BCR-ABL1*-translocated chronic myeloid leukemia and *KIT*-activated gastrointestinal stromal tumour (GIST) with imatinib [47, 108]. The ability to improve clinical outcome by exploiting tumour genetic vulnerabilities contributed to the advent of precision oncology, a framework that tailors patient care based on tumour genetic makeup.

As more actionable somatic mutations are revealed, precision oncology regimens begin to face limitations caused by single-gene assays, which pose logistical and technical challenges in scaling to meet diagnostic needs. Fortunately, these barriers were surmounted by advances in next-generation sequencing (NGS) technologies and the development of bioinformatics softwares. By harnessing the high-throughput nature of NGS, single-gene assays are rapidly supplanted by targeted gene panels and genome-scale profiling, which surveys the whole exome or genome. The dramatic decline in sequencing cost [141] and low DNA input requirements [33, 111, 122] also accelerated the adoption of NGS-based genomic testing in clinical practice. Furthermore, the simultaneous progress in algorithmic development enabled efficient storage, processing, and interpretation of massive genomic

data sets produced by NGS platforms []. Automated variant analysis pipelines were established by integrating these bioinformatics tools and subsequent optimization resulted in accurate reporting of clinically significant genomic alterations []. Hence, while the precision oncology framework was instigated by the discovery of actionable somatic mutations, its translation into clinical use was catalyzed by advancements in DNA sequencing technologies and analysis algorithms. Although research efforts are still underway in refining these technological components, it is undeniable that the precision oncology paradigm holds great potential in enhancing disease management and therapeutic intervention for cancer patients.

1.2 Overview of next-generation sequencing technologies

The Human Genome Project was completed in 2003, approximately 13 years after its launch date, producing the first human reference genome at an estimated expense of US$2.7 billion [1]. While the HGP provided a wealth of information, creating a major breakthrough in the field of genomics, the completion time and cost of the project revealed the drawbacks of DNA sequencing methods at that time, thereby stimulating the development of NGS technologies. NGS is a general term that describes various high-throughout DNA sequencing technologies that can vary based on read length, chemistry, and detection method **??**. These differences give rise to the strengths and weaknesses of each NGS platform. Recognition of these system specifications would allow users to capitalize the strengths and compensate for the limitations of the different NGS technologies.

In general, the sequencing process of most NGS platforms can be summarized into three steps. The first step involves nucleotide addition, which can be accomplished by DNA polymerase reaction or ligation (sequencing by synthesis *vs.* sequencing by ligation). This is followed by a detection step to identify the nucleotide species that was incorporated on single molecule or clonally amplified DNA templates. Nucleotide detection can be performed using optical or non-optical sensing. Illumina and Pacific Biosciences (PacBio) platforms use optical sensing to detect fluorescence for base calling [], whereas the Ion Torrent platform uses non-optical sensing to detect change in pH to determine nucleotide identity []. Lastly, a wash step re-initiates the cycle for the next base on the DNA templates by removing anchor-probe complexes, fluorophores, or blocking groups. A key feature of NGS is its ability to simultaneously carry out this process for many millions of DNA templates; hence, NGS is also known as massively parallel sequencing. There are also NGS technologies that deviate from this stepwise sequencing cycle, such as the Oxford Nanopore Technologies (ONT) platform, which directly measures DNA sequence using current shifts produced as DNA translocates through nanopore sensors [].

1.2.1 Illumina sequencing

At present, the most widely used NGS technology is the Illumina short-read platform as evident by its prevalence in the literature and the Sequence Read Archive (SRA). In 2011, 84% of sequence reads in the SRA were generated by Illumina sequencing [77]. The Illumina platform uses a sequencing-by-synthesis approach with reversible dye terminators, which enable base calling through detection of fluorescent signals while blocking the ribose 3'-OH group to prevent addition of the next nucleotide by DNA polymerase. Briefly, an Illumina NGS workflow begins by ligating adapters to the ends of fragmented DNA, followed by hybridizing these templates to complementary adapter sequences on flow cell surfaces. Bridge amplification is then performed to generate clusters of clonally amplified DNA templates. DNA sequencing starts by annealing primers complementary to adapter sequences, which enable DNA polymerase to carry out the elongation process. All four reversible dye terminator-bound deoxyribonucleotides (dNTPs) are simultaneously added during each cycle and are distinguishable by unique fluorophore-labelling. The dNTPs are also terminally blocked, allowing the incorporation of only one dNTP molecule per cycle. Subsequent to dNTPs addition, unbound dNTPs are washed away. The flow cells are then imaged using laser channels and fluorescence corresponding to the incorporated dNTP is emitted at each cluster. Finally, a new cycle is initiated by cleaving the fluorophores and unblocking the 3'-OH groups.

1.2.2 Clinical applications of NGS

The emergence of NGS has revolutionized biological inquiry, particularly in cancer genomics research. Collaborative efforts such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have leveraged NGS technologies to characterize genomic landscapes of different subtypes of tumours. This resulted in the identification of novel driver mutations, thereby enhancing knowledge of tumour biology and treatment strategies. The ability to sequence multiple genes and samples in parallel with less DNA and in a cost- and time-effective manner, also makes NGS an attractive clinical tool to complement precision medicine initiatives. These advantages demonstrate that the viability and efficiency of NGS are superior to traditional Sanger sequencing, which is typically limited to sequencing a specific gene region of a given sample per run. Currently, tumour sequencing assays ranging from targeted gene panels up to genome-scale profiling have been employed in clinical oncology to guide diagnosis, prognosis, and therapeutic decision-making.

Targeted gene panels

Several considerations must be made when developing clinical-grade tumour genomic tests, including turnaround time, testing cost per patient, and depth of sequencing coverage. For these reasons, many clinical laboratories resorted to targeted gene panels, which focus on mutational hotspots, actionable genes, or genomic regions of known clinical relevance. Furthermore, despite

the growing catalog of tumour genetic variants contributed by large consortium projects, the impact of the majority of variants on cancer development remains unknown. Genomic information of unknown significance not only pose limitations in clinical translation, but also challenges in communicating such results to patients. Hence, targeted gene panels are more practical for prospective clinical use at the present time than whole exome and genome approaches.

Strategies to interrogate genomic regions of interest in targeted NGS assays include amplicon-based and capture-based methods. Amplicon-based method enriches targeted regions using PCR amplification prior to NGS. This approach requires lower amount of DNA input and offers quick turnaround relative to hybridization capture methods. However, PCR amplification can distort detection of copy number variation (CNV), although bioinformatics tools, such as ONCOCNV [14], have been developed to perform copy number analysis using amplicon sequencing data. Moreover, amplicon sequencing is prone to enrichment bias, especially in samples with low amounts of DNA templates. For example, Wong et al. [145] reported higher prevalence of formalin-induced sequence artifacts in samples with lower amounts of amplifiable templates. Clinical samples with low template copies tend to have reduced amplicon enrichment and higher probability of amplifying DNA templates with sequence artifacts. Another limitation of amplicon sequencing is its inability to detect novel gene fusions because PCR primer pairs would fail to amplify the translocated DNA.

Hybridization capture methods involve the use of complementary oligonucleotide probes to bind targeted regions. There are two methods for target capture, namely array-based and in-solution capture. In array-based capture, complementary oligonucleotide probes of targeted regions are fixed on microarrays. Fragmented genomic DNA is hybridized to the probes on the microarray and subjected to NGS after unbound DNA is washed away. In the in-solution capture approach, the target-specific probes are biotinylated. The pool of probes is mixed with fragmented genomic DNA and hybridization occurs "in solution." Hybridized DNA is pulled down using streptavidin-labelled magnetic beads and then subjected to NGS. The in-solution capture method is typically preferred over array-based capture because it omits the necessity for expensive instrument to process microarrays, and it requires lower quantity of DNA as starting material. In contrast to the amplicon-based method, hybridization capture approaches can detect gene fusions, as well as yield more reliable inference of CNVs.

Whole exome sequencing

Whole exome sequencing (WES) interrogates all protein-coding regions, which constitute approximately 1% of the genome. Target capture methods are used to enrich coding sequences before massively parallel sequencing. To date, it is estimated that 85% of pathogenic variants are present within exons. Therefore, WES assays have the potential to facilitate prospective medical decision-making, as well as contribute to retrospective studies to uncover the functional and clinical impacts of newly discovered genetic alterations in tumours.

There are several disadvantages associated with WES assays. This includes the inability to detect mutations in non-coding regions and structural variants, which can promote cancer formation. WES assays also tend to achieve lower depth of coverage compared to targeted gene panels, increasing the rates of false positives. Because of the increased testing content, analytical validation of WES assays is more challenging and time-consuming. Nevertheless, whole exome testing has already been offered by a few academic centres such as Broad Institute of MIT and Harvard, Baylor College of Medicine, and Washington University in St. Louis, to assist in precision cancer medicine.

Whole genome sequencing

Whole genome sequencing (WGS) scans the entire genome, including coding and non-coding genomic regions. Similar to WES, WGS faces drawbacks in terms of depth of coverage and difficulty in ensuring analytic validity. Although the Illumina HiSeq X Ten System has made it possible to sequence an entire genome at 30x coverage under US$1000, application of WGS in routine clinical testing is still challenging due to limitations in variant interpretation. As a result of limited clinically annotated genetic variants, WGS is expected to yield a high burden of variants of unknown significance, which are problematic in clinical practice. Despite these constraints, Laskin et al. [79] reported that the Personalized OncoGenomics (POG) study, which integrates whole genome analysis in making therapeutic decision, can benefit patients with advanced cancers by matching them with targeted agents that are approved or currently in clinical trials. Thus, while there are challenges that need to be overcome to implement WGS as standard of care, WGS has proven its utility in conducting additional search for druggable mutations that were undetected by less comprehensive sequencing strategies. In particular, follow-up testing with WGS can broaden the treatment options for patients with incurable cancers.

1.3 Variant analysis pipeline

Advances in NGS technologies and the concomitant decline in sequencing cost have led to a marked surge in data production. For instance, the Illumina HiSeq 2500 platform is capable of sequencing 150–180 whole exomes from human samples at 50x coverage, generating approximately 1TB of raw data in a single run. Processing these enormous data sets and extracting useful results for research and clinical purposes rely heavily on the development of bioinformatics algorithms and tools. A general workflow for variant analysis in medical genomics consists of five stages: (1) quality control and pre-processing of raw data, (2) read alignment to the reference genome and post-alignment processing, (3) variant calling, (4) variant annotation, and (5) variant interpretation and reporting.

1.3.1 Quality control and pre-processing of raw data

Base-calling algorithms convert signals, such as fluorescence, light intensity, or electrical current, captured by NGS instruments to DNA sequences. For each base identified, a measure of uncertainty, known as base quality (BAQ) score, is derived by taking into account background noise. BAQ scores are commonly reported in Phred scale, given by the equation below.

$$BAQ_{\text{Phred}} = -10log_{10}P(error)$$

Hence, $1/100$ probability of base error would correspond to a Phred-scaled BAQ score of 20.

The raw output of NGS instruments, which contains information from base calling, are stored in FASTQ and FASTA text-based formats. FASTQ files contain DNA sequences with sequence names and Phred-scaled BAQ scores, whereas FASTA files only contain DNA sequences with sequence names. Quality of raw NGS data can be evaluated using bioinformatics tools such as FastQC, which generates a diagnostic report consisting of various quality control parameters. These include sequence length distribution, GC content distribution, degree of sequence duplication, presence of overrepresented and adapter sequences, and average Phred-scaled BAQ scores at each base across reads.

Quality control results are used to assist in preprocessing of raw NGS data before further analysis takes place. For example, NGS libraries with poor quality bases near the 3' ends of reads may require read trimming before alignment. Removal of adapters is also typically performed in the preprocessing step. Computational tools that are commonly used to accomplish these tasks include Trimmomatic and Cutadapt. Moreover, assessment of quality control metrics also enables the recognition of poor quality NGS libraries and those that are potentially contaminated. Flagging of these libraries would allow downstream analyses and result interpretation to be performed with caution.

1.3.2 Read alignment and post-alignment processing

Next, pre-processed reads are aligned to the reference genome. Alignment algorithms can be categorized into

After raw data QC and preprocessing, the next step is to map the reads to the reference genome and with high efficiency and accuracy. Alignment mapping is a classical string match task in computer science. For example, most web browsers and text editors provide a Find function to search for the perfect matching string with a given query. However, finding the optimal alignment for a sequence read requires an align- ment algorithm that is tolerant to imperfect matches, where genomic variations may occur. Moreover, the algorithm needs to be able to align millions of reads at a reasonable speed. As a first step to address this challenge, the reference genome is usually indexed in a hash table for efficient querying. Many different tools have been developed for short reads map-

ping. They use BurrowsWheeler Transformation (BWT) compression techniques, SmithWaterman (SW) Dynamic programing algorithm or the combination of both in order to find find the optimal alignment match within an acceptable com- putational time. Bowtie225 and BWA26 are two well-known short reads alignment tools that implement BWT algorithm. SW is a score-based dynamic programming algorithm that provides at least one optimal local alignment even though the solution might not be unique. This algorithm is tolerant to mismatches and gaps at the expense of increased computational time. MOSAIK,27 SHRiMP2,28 and Novoalign (http://www. novocraft.com) are implementations of SW algorithms with increased alignment accuracy. Multithreading and/or MPI implementations are employed in those mapping tools allow- ing significant reduction in the run-time.

Included in VarSeq is functionality similar to SnpEff or Variant Effect Predictor. Each variant is mapped to all overlapping transcripts and information about the region where it is located (exon, intron, intergenic, etc.), sequence ontology (frame shift, synonymous, etc.), and HGVS notation (g dot, c dot, and p dot) is provided. You can chose to filter against the highest-impact annotation for each variant or the entire set of variant-transcript interactions.

briefly mention other methods like assembly

1.3.3 Variant calling

1.3.4 Variant annotation

1.3.5 Variant interpretation and reporting

1.4 ACCE model process for evaluating genetic tests

Development of a genetic test, including NGS-based genomic testing, for clinical use must be accompanied by an evaluation process to establish robustness and clinical benefits of the test. One approach to assess genetic tests is the ACCE model process, which consists of four criteria that make up its acronym: **A**nalytical validity, **C**linical validity, **C**linical utility, and **E**thical, legal and social implications.

Analytic validation ensures that a clinical assay detects the genetic changes it was designed to identify with sufficient sensitivity and specificity. For example, analytic validity of a targeted NGS panel can be determined by running the panel on samples with known mutations that were previously identified using Sanger sequencing. Sensitivity and specificity of the targeted NGS panel can be measured using the results from Sanger sequencing as reference standards. Analytic validity also refers to quality assurance of a clinical assay. For instance, a targeted NGS panel must be capable of producing similar sequencing metrics (e.g. read depth and coverage uniformity) for samples with comparable DNA quantity and integrity.

Clinical validation determines whether results of the genetic test correspond to the clinical condition it was meant to detect. Example of a genetic test with high clinical validity is *RET* mutational testing, which can identify individuals with multiple endocrine neoplasia type 2 (MEN2) at a sensitivity of 95–98%. MEN2 is a heritable disorder transmitted in an autosomal dominant pattern, resulting in increased susceptibility to tumours in endocrine tissues, especially medullary thyroid carcinoma. On the other hand, clinical utility of a genetic test is defined by its ability to enhance clinical outcome, such as survival and progression-free survival, after weighing in the risks, benefits, and economic impact of the test. The clinical utility of *RET* mutational testing is demonstrated by its efficacy in identifying MEN2 susceptible children who would benefit from prophylactic surgery to remove all or parts of the thyroid gland. This results in reduced risk of developing thyroid cancers, thereby improving survival of these individuals.

Lastly, the ACCE model includes evaluation of ethical, legal and social implications of a genetic test. This component considers how the genetic test can lead to ramifications such as violation of privacy and confidentiality, stigmatization and discrimination based on genetic makeup (e.g. accessibility to insurance), and complications pertaining to consent for disclosure and ownership of the data. As well, this component of the framework ensures that safeguards, such as relevant policies and genetic counseling protocols, are implemented to prevent societal repercussions.

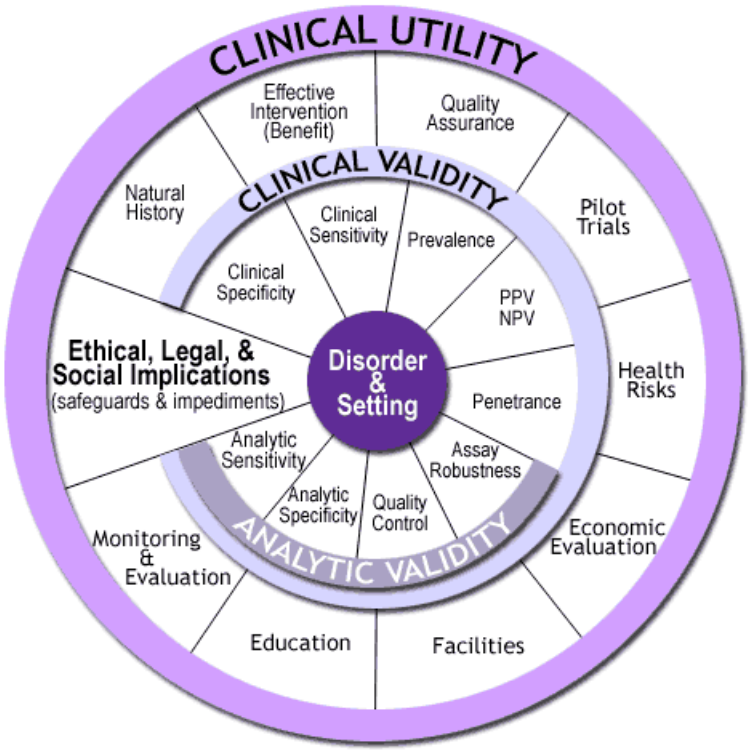


Figure 1.1: Four main criteria of the ACCE model process for evaluating a genetic test: **A**nalytical validity, **C**linical validity, **C**linical utility, and **E**thical, legal and social implications

1.5 Clinical implications of germline alterations in cancer

Screening for somatic and germline alterations are essential in delivering precision medicine to cancer patients. Somatic mutations can influence disease management and treatment of cancer patients, whereas clinical implications of germline alterations extend beyond the patient, affecting their families as well. Germline variants in cancer predisposing genes (CPGs) can predict the risk of disease onset, allowing for preventive measures to be administered. Furthermore, germline variants in pharmacogenomic (PGx) genes can predict response to chemotherapeutic drugs. Therefore, germline testing should be offered to ensure more precise cancer care if resources are available to analyze and interpret germline findings, and appropriate protocols are established to communicate results with patients and affected family members.

1.5.1 Cancer predisposition

Germline variants in CPGs can indicate increased cancer risks. Between 1982 and 2014, 114 CPGs were discovered using approaches such as candidate gene, genome-wide mutation, and linkage analyses. The majority of CPGs act as tumour suppressors; hence, loss-of-function mutations in

these genes, which inactivate gene function, predispose carriers to cancer. Genes in this category, including *TP53*, *BRCA1*, *BRCA2*, *APC*, and *RB1*, are usually involved in DNA repair and cell-cycle regulation. Conversely, there are fewer CPGs that promote cancer formation through gain-of-function mutations. These CPGs, typically protein kinases like *ALK*, *EGFR*, and *RET*, predispose carriers to cancer through activation of gene function.

Clinical testing of CPGs can improve various aspects of patient care such as disease management and treatment. For instance, patients with BRCA-deficient breast and ovarian tumours can be treated with PARP inhibitors, which target tumour cells and impair growth through synthetic lethality. Notably, a key benefit of testing for germline variants in CPGs is the window of opportunity to implement cancer preventive measures for patients and affected relatives. Cancer prevention can involve approaches like early and regular cancer screening, as well as prophylactic surgery and chemotherapy. For example, patients with familial adenomatous polyposis (FAP), which is caused by germline alterations in the *APC* gene and associated with high risk of developing colorectal cancer (CRC), are recommended to begin early colonoscopy-based screening. In particular, patients who have first-degree relatives with CRC should start screening as early as 40 years of age or 10 years earlier than the youngest age of onset of an affected family member [13].

1.5.2 Pharmacogenomics

Despite the expanding spectrum of targeted anti-cancer drugs, cytotoxic chemotherapy remains the primary treatment for several types of cancers. However, germline variants in PGx genes that affect the function and/or expression of drug targets and drug disposition proteins (proteins involved in drug metabolism and transport) can give rise to chemotherapy-related toxicity. Examples of chemotherapy-related toxicity include hand-foot syndrome, hearing loss, cardiomyopathy, and high-grade neutropenia, diarrhea, nausea and vomiting. Chemotherapy-related toxicity can be debilitating and fatal, as well as culminate significant health care expenditures in "treating the treatment." To alleviate the occurrence of chemotherapy-related toxicity, germline PGx testing should be implemented in clinical practice to guide the selection of chemotherapeutic drugs and optimization of drug dosage for cancer patients.

5-fluorouracil

5-fluorouracil (5-FU) is a fluoropyrimidine drug that is commonly administered in chemotherapy regimens for patients with gastrointestinal cancers, including CRC. Inter-patient variability in response to 5-FU treatment can be caused by germline variants in the *TYMS* gene, which encodes for the drug target, the thymidylate synthase enzyme (TS). One of the 5-FU mechanisms of action involves the conversion of 5-FU to fluorodeoxyuridine monophosphate (5-FdUMP). 5-FdUMP then sequesters TS by forming a ternary complex with TS and the 5,10-methylenetetrahydrofolate (CH_2THF) cofactor, thereby impeding DNA synthesis. Germline alterations that result in a higher

expression of TS such as the triple repeats of a 28 bp sequence upstream of the *TYMS* translational start site (rs45445694) are indicators of reduced likelihood of experiencing 5-FU toxicity. Unfortunately, this also means that treatment with 5-FU might not be effective due to high TS levels in the tumours.

Germline variants in (*DPYD*) and (*TYMP*) genes, which encode for the 5-FU metabolizing proteins, dihydropyrimidine dehydrogenase (DPD) and thymidine phosphorylase (TP), respectively, can also serve as predictors for 5-FU-induced toxicity. DPD catabolizes 5-FU into dihydrofluorouracil, which mainly occurs in the liver, and the inactive products are subsequently excreted in the urine. Hence, germline variants resulting in DPD deficiency or total loss contribute to a longer half-life of 5-FU, which can cause severe or fatal toxicity in cancer patients. Several studies implied that TP may play a causative role in tumour growth and metastasis. In fact, higher TP expression was observed in tumours than normal tissues in CRC patients. Consequently, the 5-FU prodrug, capecitabine, is administered to target TP-overexpressed tumours because TP can metabolize capecitabine to the thymidylate synthase inhibitor, 5-FdUMP. Hence, this affects tumour growth with minimal toxic effects in normal cells. However, the presence of germline variants that increase expression of TP in normal cells could potentially lead to adverse drug reactions in patients receiving 5-FU-based chemotherapy.

Efficacy of 5-FU depends on the intracellular reduced folate, CH_2THF, which together with the 5-FU active metabolite, 5-FdUMP, inhibit TS. This blocks the synthesis of deoxythymidine monophosphate (dTMP), causing imbalanced nucleotide levels in the cell and DNA damage. One of the enzymes that regulates intracellular CH_2THF levels is methylenetetrahydrofolate reductase (MTHFR), which irreversibly converts CH_2THF to CH_3THF. Germline variants in the *MTHFR* gene that reduce enzymatic activity such as c.677C>T and c.1298A>C polymorphisms can increase chemosensitivty of tumours to 5-FU through cellular accumulation of CH_2THF. Nevertheless, several studies suggested that the combined presence of *MTHFR* c.1298A>C and *TYMS* 3UTR indels could serve as predictors for 5-FU toxicity in CRC patients.

Oxaliplatin

Oxaliplatin is a platinum derivative commonly used in combination with 5-FU for treating gastric and colorectal cancers. Deactivation of oxaliplatin can be induced by conjugation of the platinum derivative with glutathione (GSH), which is catalyzed by glutathione S-transferases (GSTP). While there are studies suggesting that germline variants in *GSTP1* genes are associated with increased neurotoxicity in patients treated with oxaliplatin combination therapy, there are also groups reporting conflicting results. Therefore, additional studies are required to confirm the impact of *GSTP1* polymorphisms on oxaliplatin treatment.

Irinotecan

Irinotecan is a camptothecin analog widely used in chemotherapy regimens for treating lung cancers and CRC. The active metabolite of irinotecan, SN-38, blocks type I DNA topoisomerase, impairing DNA replication. SN-38 is inactivated in the liver by uridine diphosphate glycosyltransferase 1A1 (UGT1A1) through glucuronidation and then excreted. Thus, patients with deficiency in UGT1A1, which can be caused by germline variants in the *UGT1A1* gene, are at higher risk of experiencing toxicity due to a longer half-life of SN-38. An example of a germline variant in *UGT1A1* that results in reduced activity is the UGT1A1*28 allele, which corresponds to an extra TA repeat within position -53 and -42 of the translational start codon. Dose reduction is recommended for carriers to prevent toxic effects induced by irinotecan.

1.6 Technical challenges in implementing germline testing in clinical oncology

1.6.1 Tumour-only sequencing

One of the challenges in integrating germline testing in clinical oncology is tumours are often sequenced without matched normal samples. Although sequencing of matched normal samples would allow accurate identification of somatic mutations and simultaneous detection of clinically important germline variants, it is common for clinical laboratories to only sequence tumour samples to minimize cost and turnaround time. However, genomic analyses of tumours can reveal clinically relevant germline variants. For examples, Schrader et al. [113] reported that 91.9% of pathogenic germline variants in CPGs were retained in the tumour genome. Hence, clinical laboratories could leverage tumour genomic testing for identification of germline variants and subsequently, refer potential germline variants to downstream confirmatory testing.

A clinical pipeline that leverages tumour genomic testing to perform initial screening for germline alterations could provide germline testing in a cost-effective manner because only selective patients would require follow-up testing. However, as the tumour genome contains both germline and somatic variants, the difficulty remains in devising an approach to accurately separate germline variants from somatic mutations. Jones et al. [72] use public databases like the Single Nucleotide Polymorphism Database (dbSNP) and Catalogue of Somatic Mutations in Cancer (COSMIC) database, as well as effect prediction tool to distinguish between germline and somatic variants, but reported high false positive rates. The use of these public databases cannot reliably differentiate between variant statuses because it is probable for a germline variant to occur somatically, and *vice versa*. For instance, an evaluation of 468 genes with known somatic driver mutations recorded in the COSMIC database showed that 49 of these genes were also known to harbour germline alterations that are associated with inherited predisposition to cancer.

One possible approach is the use of variant allele frequency (VAF) to discriminate between

germline and somatic variants. Because tumour biopsies are typically admixtures of tumour and normal cells, there is a high likelihood that somatic mutations might deviate from diploid zygosity (i.e. heterozygous variants are expected to have VAF close to 50%, whereas homozygous variants are expected to have VAF close to 100%). Moreover, tumour heterogeneity might also give rise to VAF deviations. Therefore, the use of VAF threshold could be a potential solution in distinguishing between germline and somatic alterations in genomic analyses of tumours without matched normal DNA.

1.6.2 Formalin-fixed paraffin-embedded tumours

Another disadvantage of performing germline variant analysis using tumour DNA is tumour samples in the clinic are often formalin-fixed paraffin-embedded (FFPE). Formalin fixation preserves tissue morphology for histological assessment, whereas paraffin embedding enables stable storage of specimens at room temperature, which is cost- and space-saving compared to maintaining fresh frozen specimens in freezers. DNA isolated from FFPE tumours pose technical challenges in molecular testing because formalin fixation induces several types of DNA damage. Therefore, assessment of these different forms of DNA damage is essential to establish quality control for a clinical genomic assay.

The main component of formalin, formaldehyde, can react with DNA bases and proteins, producing DNA-DNA, DNA-protein, and protein-protein crosslinks. Additionally, formaldehyde-DNA adducts can also be generated in formalin-fixed tissues. Crosslinking induced by formaldehyde destabilizes the DNA structure, resulting in degradation and low DNA yields extracted from FFPE tissues. Another predominant form of formalin-induced DNA damage is DNA fragmentation. Hence, FFPE tissues not only produce low quantities of DNA, but also DNA with short fragment sizes. Particularly, this interferes with amplicon-based methods by reducing the amount of amplifiable DNA templates. Severity of DNA fragmentation also increases with age of paraffin blocks and acidity of formalin solution used in tissue fixation.

FFPE DNA also constitutes increased frequency of sequence artifacts. This is problematic in clinical practice because there is a high risk of misinterpreting artifactual base changes as true mutations that may influence patient care. Oxidization of formaldehyde, which generates formic acid, creates an acidic environment that catalyzes hydrolytic cleavage of *N*-glycosidic bonds between purine bases and the sugar backbone. This produces abasic sites at which sequence artifacts can occur as most DNA polymerases tend to selectively incorporate adenines across abasic sites during the extending stage. In fewer cases, guanines and short deletions ranging from 1–3 bases could also be introduced by DNA polymerase when synthesizing through abasic sites.

A well-documented source of sequence artifacts in FFPE DNA is cytosine deamination. This generates uracil lesions, which leads to artifactual C>T/G>A transitions because thymines are added opposite of uracils during synthesis of complementary DNA strands. Wong et al. [145]

showed increased levels of C>T/G>A artifacts in amplicon sequencing data generated from highly fragmented DNA samples. This observation was attributed to a higher probability of amplifying DNA templates containing sequence artifacts in samples with reduced amount of amplifiable DNA templates as a result of fragmentation damage. While cytosines can be restored by treating FFPE DNA with uracil-DNA glycosylase (UDG) to eliminate uracil lesions, there is currently no method to repair deamination of 5-methylcytosine (5-mC). 5-mC are common at CpG dinucleotides and are more susceptible to deamination in formalin-fixed tissues. Deamination of 5-mC gives rise to thymine instead of uracil, thus cannot be reinstated through treatment with the UDG enzyme.

1.7 Objectives

This thesis aims to determine whether potential germline alterations can be accurately identified in FFPE tumours without the use of matched normal samples for follow-up testing. We performed analytic validation of a clinical amplicon-based targeted sequencing panel for FFPE solid tumours by comparison with sequencing of blood DNA, which is the gold standard for germline testing. Our objectives include (1) assessing the degree of formalin-induced DNA damage in FFPE DNA, (2) determining retention rate of germline alterations in FFPE tumours, and (3) evaluating the use of VAF thresholds to distinguish germline alterations from somatic mutations in tumour-only analyses, as well as establishing a VAF cut-off that would maximize true positive rate of identifying germline alterations in FFPE tumours and minimize referral of somatic mutations (false positives) to downstream germline testing.

Chapter 2

Materials and Methods

2.1 Overview of study design

This study examines whether potential germline alterations can be accurately identified in FFPE tumours without the use of matched normal samples for follow-up testing. Targeted sequencing data from 213 cancer patients with FFPE tumour and matched blood samples were retrospectively analyzed. Extracted DNA from samples were sheared, enriched for amplicons in the OncoPanel, barcoded, and subjected to next-generation sequencing. Sequencing data were processed and analyzed with a custom variant calling pipeline. To assess the degree of formalin-induced DNA damage, the efficiency in amplicon enrichment and sequencing results of FFPE samples were compared to blood. Furthermore, variant concordance between blood and FFPE tumours was measured to determine whether tumour DNA is a reliable resource for detecting germline alterations. Lastly, the use of VAF thresholds in distinguishing between germline and somatic alterations in tumour-only analyses was evaluated.

2.2 Patient samples

Blood and FFPE tumour samples were acquired from 213 patients who provided informed consent for The OncoPanel Pilot (TOP) study (Human Research Ethics Protocol H14-01212), a pilot study to optimize the OncoPanel, which is an amplicon-based targeted NGS panel for solid tumours. The TOP study also aims to assess the OncoPanel's application for guiding disease management and therapeutic intervention. One blood sample and four FFPE tumours were sequenced in duplicates, which resulted in 217 tumour-normal paired samples (434 sequencing libraries were included in our analyses). Patients in the TOP study are those with advanced cancers including colorectal cancer, lung cancer, melanoma, gastrointestinal stromal tumour (GIST), and other cancers (Table 2.1). The age of paraffin block for tumour samples ranges from 18 to 5356 days with a median of 274 days.

Table 2.1: Distribution of cancer types in the TOP cohort.

Cancer Type	Number of Cases	Percentage (%)
Colorectal	97	46
Lung	60	28
Melanoma	18	8
Other[†]	16	8
GIST	7	3
Sarcoma	4	2
Neuroendocrine	4	2
Cervical	2	0.9
Ovarian	2	0.9
Breast	2	0.9
Unknown	1	0.5

[†]This category includes thyroid, peritoneum, Fallopian tube, gastric, endometrial, squamous cell carcinoma, anal, salivary gland, peritoneal epithelial mesothelioma, adenoid cystic carcinoma, pancreas, breast, gall bladder, parotid epithelial myoepithelial carcinoma, carcinoid, and small bowel cancers.

2.3 Sample preparation, library construction, and Illumina sequencing

Genomic DNA was extracted from blood and FFPE tumour samples using the Gentra Autopure LS DNA preparation platform and QIAamp DNA FFPE tissue kit (Qiagen, Hilden, Germany), respectively. The extracted DNA was sheared according to a previously described protocol [20] to obtain approximate sizes of 3 kb followed by PCR primer merging, amplification of target regions, and adapter ligation using the Thunderstorm NGS Targeted Enrichment System (RainDance Technologies, Lexington, MA) as per manufacturer's protocol. Barcoded amplicons were sequenced with the Illumina MiSeq system for paired end sequencing with a v2 250-bp kit (Illumina, San Diego, CA).

2.4 OncoPanel (Amplicon-based targeted sequencing panel for solid tumours)

The OncoPanel assesses coding exons and clinically relevant hotpots of 15 cancer-related genes and six PGx genes that can predict risk of developing chemotherapy-induced toxicity. Primers were designed by RainDance Technologies (Lexington, MA) using the GRCh37/hg19 human reference genome to generate 416 amplicons between 56 bp and 288 bp in size, which interrogate ∼20 kb of target bases. Complete list of genes and gene reference models for the OncoPanel is presented in Table 2.2, whereas OncoPanel target regions and amplicons are presented in Table A.1.

Table 2.2: Gene reference models for HGVS nomenclature of OncoPanel genes.

Gene	Protein	Reference Model
Cancer-related		
AKT1	Protein kinase B	NM_001014431.1
ALK	Anaplastic lymphoma receptor tyrosine kinase	NM_004304.3
BRAF	Serine/threonine-protein kinase B-Raf	NM_004333.4
EGFR	Epidermal growth factor receptor	NM_005228.3
HRAS	GTPase HRas	NM_005343.2
MAPK1	Mitogen-activated protein kinase 1	NM_002745.4
MAP2K1	Mitogen-activated protein kinase kinase 1	NM_002755.3
MTOR	Serine/threonine-protein kinase mTOR	NM_004958.3
NRAS	Neuroblastoma RAS viral oncogene homolog	NM_002524.3
PDGFRA	Platelet-derived growth factor receptor alpha	NM_006206.4
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	NM_006218.2
PTEN	Phosphatase and tensin homolog	NM_000314.4
STAT1	Signal transducer and activator of transcription 1	NM_007315.3
STAT3	Signal transducer and activator of transcription 3	NM_139276.2
TP53	Tumor protein P53	NM_000546.5
Pharmacogenomics		
DPYD	Dihydropyrimidine dehydrogenase	NM_000110.3
GSTP1	Glutathione S-rransferase pi 1	NM_000852.3
MTHFR	Methylenetetrahydrofolate reductase	NM_005957.4
TYMP	Thymidine phosphorylase	NM_001113755.2
TYMS	Thymidylate synthetase	NM_001071.2
UGT1A1	Uridine diphosphate (UDP)-glucuronosyl transferase 1A1	NM_000463.2

2.5 Variant calling pipeline

2.5.1 Read alignment and variant calling

Reads that passed the Illumina Chastity filter were aligned to the hg19 human reference genome using the BWA mem algorithm (version 0.5.9) with default parameters, and the alignments were processed and converted to the BAM format using SAMtools (version 0.1.18). The SAMtools `mpileup` function (`samtools mpileup -BA -d 500000 -L 500000 -q 1`) was used to generate pileup files for all target bases followed by variant calling with the VarScan2 `mpileup2cns` (version 2.3.6) function with parameter thresholds of VAF $\geq$ 10% and Phred-scaled base quality (BAQ) score $\geq$ 20 (`--min-var-freq 0.1 --min-avg-qual 20 --strand-filter 0 --p-value 0.01 --output-vcf --variants`).

Four genomic positions at which the hg19 human reference genome contains potential risk alleles were identified (Table 2.3). Hence, patients homozygous for these four risk alleles would not be identified by our standard variant calling procedure. For these four genomic sites, our method for variant calling was modified to provide calls for every patient in the cohort. The VarScan2 `mpileup2cns` function with parameter thresholds of VAF $\geq$ 25%, VAF to call homozygote $\geq$ 90%, BAQ score $\geq$ 20, and fraction of variant reads from each strand $\geq$ 0.1 (`--min-var-freq 0.25 --min-freq-for-hom 0.9 --min-avg-qual 20 --strand-filter 1 --p-value 0.01 --output-vcf`) was used. Next, allelic statuses were re-assigned, in which wild type calls were re-assigned as homozygous variants, while homozygous variants were re-assigned as wild type calls. Corrections to the VAFs of these four genomic sites were also made to ensure that the VAFs reflect percentage of reads with the risk alleles.

Table 2.3: Potential risk alleles in the hg19 human reference genome within the target regions of the OncoPanel.

Gene	Chr	Pos	Risk Allele	dbSNP ID	HGVS*
DPYD	chr1	98348885	C	rs1801265	p.Cys29Arg
					c.85T>C
MTOR	chr1	11205058	G	rs386514433;	p.Ala1577Ala
				rs1057079	c.4731A>G
	chr1	11288758	C	rs1064261	p.Asn999Asn
					c.2997T>C
TP53	chr17	7579472	C	rs1042522	p.Arg72Pro
					c.215G>C

*Description of sequence variants according to the HGVS recommendations.

2.5.2 Variant filtering

Variant calls were filtered using the VarScan2 `fpfilter` function with fraction of variant reads from each strand $\geq$ 0.1 and default thresholds for other parameters (Table 2.4). The VarScan2 `fpfilter` removed 247 low quality variants. Seventy germline variants in the blood were also excluded from our analysis because these variants in the tumours were filtered by the VarScan2 `fpfilter`. There were also 16 risk allele calls in tumour samples that did not pass the strand filter, causing the removal of 10 risk allele calls in the blood samples from our evaluation. Overall, a total of 343 calls were excluded by the VarScan2 `fpfilter` and strand filter. Manual inspection was performed for a subset of variants, including variants detected within primer regions and in PGx genes, using the Intergrative Genomics Viewer (IGV, version 2.3). This resulted in the removal of 500 spurious calls, which stemmed from software bugs, sequencing artifacts, primer masking, and primer artifacts (Table 2.5). Eleven low coverage calls ($\leq$ 100x) were also excluded from our analysis. Implementation of this filtering pipeline reduced the raw variant output of 5288 calls from 217 paired tumour-blood samples (434 sequencing libraries) to 4434 calls (Figure 2.1B).

2.5.3 Variant annotation and interpretation

SnpEff (version 4.2) was used for effect prediction, and the SnpSift package in SnpEff was used to annotate variants with databases such as dbSNP (b138), COSMIC (version 70), 1000 Genomes Project, and ExAC (release 0.3) for interpretation. Clinical significance reported by the ClinVar database and literature review were also used for variant interpretation.

Table 2.4: Thresholds for parameters of VarScan2 `fpfilter` used for filtering raw variant output.

Parameter	Description	Threshold
`--min-var-count`	Min number of var-supporting reads	4
`--min-var-count-lc`	Min number of var-supporting reads when depth below somaticPdepth	2
`--min-var-freq`	Min variant allele frequency	0.1
`--max-somatic-p`	Max somatic p-value	0.05
`--max-somatic-p-depth`	Depth required to test max somatic p-value	10
`--min-ref-readpos`	Min average read position of ref-supporting reads	0.1
`--min-var-readpos`	Min average read position of var-supporting reads	0.1
`--min-ref-dist3`	Min average distance to effective 3' end of ref reads	0.1
`--min-var-dist3`	Min average distance to effective 3' end of variant reads	0.1
`--min-strandedness`	Min fraction of variant reads from each strand	0.1
`--min-strand-reads`	Min allele depth required to perform the strand tests	5
`--min-ref-basequal`	Min average base quality for ref allele	15
`--min-var-basequal`	Min average base quality for var allele	15
`--min-ref-avgrl`	Min average trimmed read length for ref allele	90
`--min-var-avgrl`	Min average trimmed read length for var allele	90
`--max-rl-diff`	Max average relative read length difference (ref - var)	0.25
`--max-ref-mmqs`	Max mismatch quality sum of ref-supporting reads	100
`--max-var-mmqs`	Max mismatch quality sum of var-supporting reads	100
`--max-mmqs-diff`	Max average mismatch quality sum (var - ref)	50
`--min-ref-mapqual`	Min average mapping quality for ref allele	15
`--min-var-mapqual`	Min average mapping quality for var allele	15
`--max-mapqual-diff`	Max average mapping quality (ref - var)	50

Table 2.5: Spurious variants removed by the variant filtering pipeline.

Gene	Chr	Pos	Ref	Alt	Reason
KIT	chr4	55599268	C	T	Variant masked by primer in FFPE specimen
MAPK1	chr22	22162126	A	G	Variant masked by primer in FFPE specimen
MTOR	chr1	11186783	G	A	Sequencing artifact within primer region
MTOR	chr1	11190646	G	A	Variant masked by primer in FFPE specimen
TYMP	chr22	50964446	A	T	Poor target region, alignment of different sized amplicons
TYMP	chr22	50964862	A	T	Poor target region, alignment of different sized amplicons
TYMS	chr18	673449	G	C	VarScan2 bug after chr18:673443 c.*447_*452delTTAAAG
UGT1A1	chr2	234668879	CAT	C	Sequencing artifact at AT repeats in promoter
UGT1A1	chr2	234668881	T	TAC	VarScan2 bug after AT insertion in promoter

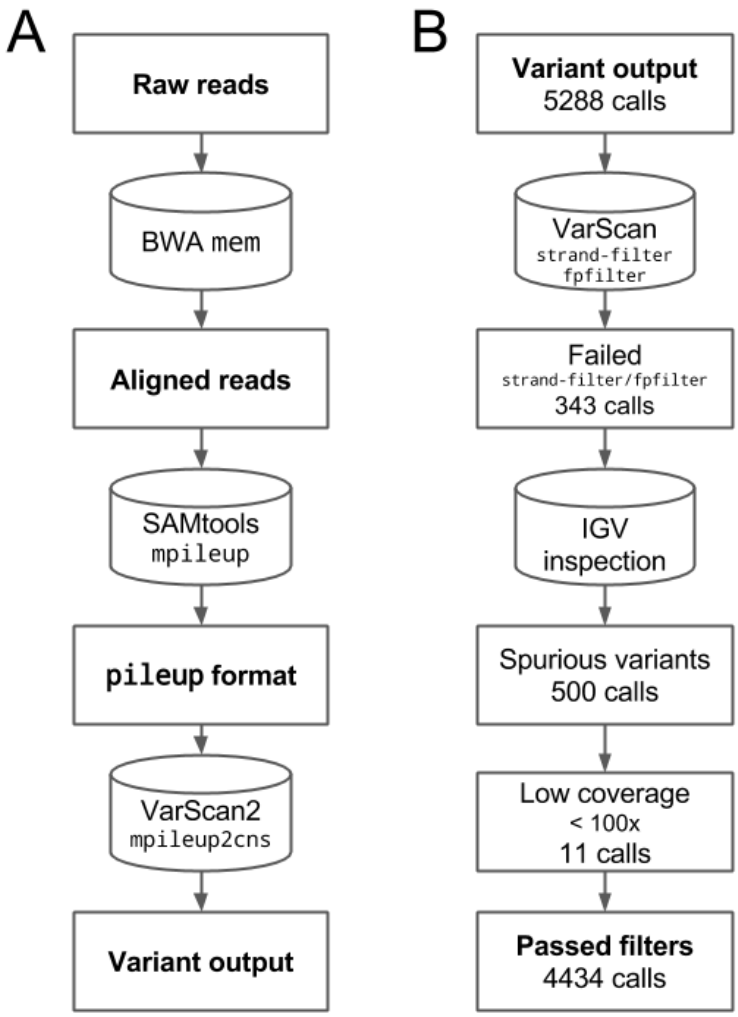


Figure 2.1: Pipelines for (A) variant calling and (B) filtering.

2.6 Sequence analysis

A custom Python script was used to process BAM files to quantify the number of on-target aligned (reads that map to target regions), off-target aligned (reads that map to hg19 but not target regions), and unaligned reads with a Phred-scaled mapping quality (MAPQ) score ≥ 10. Unaligned reads were also screened against microbial sequences, including viruses, archaea, bacteria, and fungi, to ensure that samples do not contain significant amount of microbial contaminants. Coverage depth for target bases with MAPQ ≥ 1 and BAQ ≥ 20 was obtained using bam-readcount (https://github.com/genome/bam-readcount). To measure coverage depth of amplicons, the SAMtools `view` function was used to filter for reads with MAPQ ≥ 1 `(samtools view -b -q 1)` followed by the bedtools `intersect` function (version 2.25.0) to quantify the number of reads that overlap with amplicon positions `(intersect -a $AMPLICON_POSITIONS -b $BAM_FILE -f 0.95 -r -c)`.

Per-base metrics generated using bam-readcount were also used for assessment of sequence artifacts. A custom R script was used to count and categorize the different groups of base changes (i.e. C>T/G>A, A>G/T>C, C>A/G>T, A>C/T>G, C>G/G>C, and A>T/T>A). Unless stated otherwise, analysis of sequence artifacts excludes true variants identified by our VarScan2 variant calling pipeline and base changes with VAF $< 1\%$, which are considered sequencing errors. All statistical analyses and data visualization were performed using the R statistical software package (version 3.3.2) and associated open-source packages.

2.7 Application of VAF thresholds to separate germline alterations from somatic mutations

Variants in the tumours that passed our filtering criteria were subjected to VAF thresholds between 10–45%. At each VAF cut-off, variants that were not filtered out were considered predicted germline variants. Given that all tumour samples have matched blood samples, true positives were identified as predicted germline variants that overlap with variants in the blood (Figure 2.2). Conversely, false negatives were identified as variants that were filtered out by the VAF cut-off (predicted as somatic), but were present in the blood samples. Sensitivity at each VAF threshold was calculated by dividing the number of true positives with the sum of true positives and false negatives. Because predicted germline variants will be referred to follow-up germline testing, positive predictive values (PPVs) were calculated at each VAF cut-off to evaluate precision of our approach. False positives were identified as predicted germline variants that were absent in the blood, and PPV was calculated by dividing the number of true positives with the sum of true positives and false positives.

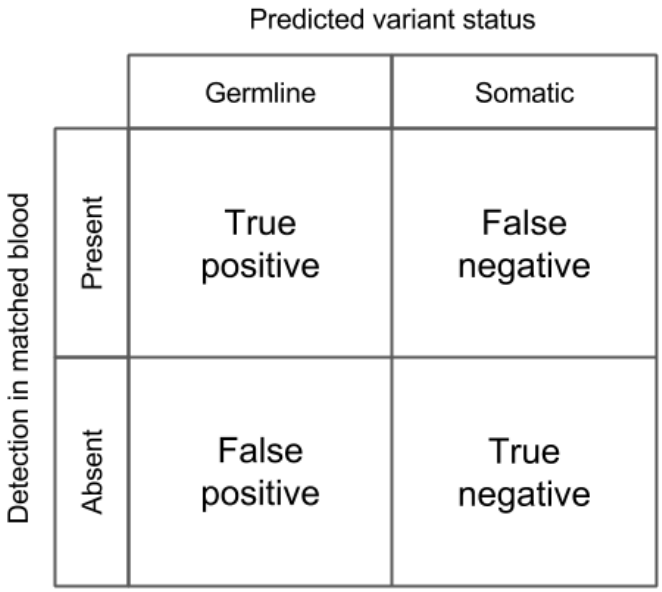


Figure 2.2: 2x2 contingency table for determination of true positive, false positive, true negative, and false negative variant calls in tumour-only analyses.

Chapter 3

Assessment of Formalin-Induced DNA Damage in FFPE Specimens

Tumour biopsies and resections are often formalin-fixed and paraffin-embedded to preserve cellular morphology for pathological review. The FFPE method also enables storage of tissues at room temperature, minimizing cost and mitigating logistical difficulties in procurement of large archives of clinical specimens [85]. However, formaldehyde, the main component of formalin, is known to induce DNA damage such as fragmentation and cytosine deamination, which could affect the use of FFPE DNA in clinical genomic testing [42, 76, 100, 101, 118, 144, 145]. As DNA derived from blood is one of the gold standards for germline testing, we characterized formalin-induced DNA damage in our data to assess its impact on identification of germline alterations in FFPE DNA. With blood specimens serving as non-formalin-fixed controls, we compared efficiency in amplicon enrichment and sequencing results of FFPE specimens to blood.

3.1 Comparison of efficiency in amplicon enrichment and sequencing results between blood and FFPE specimens

Formalin fixation causes DNA fragmentation that would reduce template DNA for PCR amplification, leading to decreased efficiency in amplicon enrichment methods for FFPE DNA [40, 42, 144, 145]. To investigate this effect, we first compared the amplicon yield between blood and FFPE specimens, and a Wilcoxon signed-rank test indicated that amplicon yield in FFPE specimens was significantly lower than blood specimens (W = 23613, Z = 12.7, $p = 8.3 \times 10^{-62}$, r = 0.61; Figure 3.1A). However, the amount of DNA input for amplicon enrichment varies across specimens in our study design, and we demonstrated that amplicon yield was weakly correlated with DNA input for both blood and FFPE specimens (Spearman's rank correlation: blood, r_s = 0.29, 95% CI = 0.16–0.41, $p = 2.1 \times 10^{-5}$; FFPE, r_s = 0.25, 95% CI = 0.12–0.37, $p = 2.5 \times 10^{-4}$; Figure 3.1B). To account for the difference in DNA input across specimens, we derived the $\log_2$ fold change between

DNA input and amplicon yield ($\log_2$ (Amplicon Yield/DNA Input)) to measure the efficiency in amplicon enrichment. We compared the $\log_2$ fold change in FFPE specimens to blood, and we found a significant decrease in enrichment efficiency in FFPE specimens compared to blood (Wilcoxon signed-rank test, W = 24754, Z = 12.7, $p = 4.6 \times 10^{-57}$, r = 0.61; Figure 3.1C). This result implies that production of amplicons is less efficient in FFPE specimens compared to blood, demonstrating the drawback of using FFPE DNA in amplicon-based NGS.

To examine whether blood and FFPE specimens produce comparable sequencing results, we compared read alignments between blood and FFPE specimens. Inspection of on-target aligned reads, which are reads that align to target regions used for variant calling, revealed no significant difference in the percentage of on-target aligned reads between blood and FFPE specimens (Wilcoxon signed-rank test, W = 10178.5, Z = -1.69, p = 0.091, r = -0.081; Figure 3.2). However, there were more outliers with slightly lower percentage of on-target aligned reads ($< 75\%$) in FFPE specimens compared to blood, and the distribution of percentage of on-target aligned reads was also wider in FFPE specimens (range: FFPE = 32.5–97.4%, blood = 74.0–95.9%), suggesting more variability in the rate of on-target alignment in FFPE specimens than blood. Similarly, no significant difference in the percentage of off-target aligned reads, which are reads that map to the human reference genome but not to target regions, was observed between specimen types (Wilcoxon signed-rank test, W = 11494.5, Z = -0.359, p = 0.72, r = -0.017; Figure 3.2). Although a Wilcoxon signed-rank test indicated that the percentage of unaligned reads was significantly different between blood and FFPE specimens (W = 19069, Z = 7.82, $p = 2.4 \times 10^{-16}$, r = 0.38; Figure 3.2), there was only a small decrease in the median percentage of unaligned reads in FFPE specimens compared to blood (median: FFPE = 0.8%, blood = 1.3%). Moreover, our data showed no significant difference in percentage of contaminant reads between specimen types (W = 14877, Z = 3.29, $p = 9.2 \times 10^{-4}$, r = 0.16; Figure 3.2), although there was one extreme outlier in FFPE specimens (range: FFPE = 0.028–64%, blood = 0.082–8.1%). While there were minor differences in percentage of unaligned reads between sequencing libraries generated from blood and FFPE DNA, blood and FFPE libraries resulted in comparable percentage of on-target aligned reads, thereby providing equivalent amount of aligned reads for variant calling.

Although blood and FFPE specimens demonstrated no significant difference in the percentage of on-target aligned reads, this result does not reflect the coverage depth of target regions in blood and FFPE specimens. To examine whether discrepancy in coverage depth exists between specimen types, we obtained coverage depth of target bases for all sequencing libraries and normalized per base coverage depth to account for difference in library size. We derived the average per base coverage depth for each library and compared this sequencing metric between blood and FFPE specimens. The average per base coverage depth was significantly different between FFPE and blood specimens (Wilcoxon signed-rank test, W = 20864, Z = 9.76, $p = 2.5 \times 10^{-26}$, r = 0.47), but there was only a slight decrease in the average per base coverage depth in FFPE specimens compared to

blood (median: FFPE = 1194, blood = 1271). We also calculated the percentages of target bases that met coverage thresholds ranging from zero to 1000x to evaluate coverage uniformity of target bases between blood and FFPE specimens. While coverage uniformity was significantly different between blood and FFPE specimens at coverage levels except at the zero and 100x coverage depth cut-off (Wilcoxon signed-rank test, $p < 0.0001$; Figure 3.3), we considered these discrepancies to be technically insignificant because the absolute difference in median percentage of target bases only exceeded 5% at 500x, 900x, and 1000x coverage thresholds (Table 3.1). Nevertheless, there were more outliers with lower percentage of target bases than median values in FFPE specimens at coverage thresholds between 100x to 1000x, implying that poor coverage uniformity is more profound for a subset of FFPE specimens. Together, our findings reveal that FFPE specimens demonstrated lower efficiency in amplicon enrichment and minor discrepancies in coverage depth and uniformity compared to blood specimens, whereas comparable proportion of on-target read alignments could be attained between specimen types.

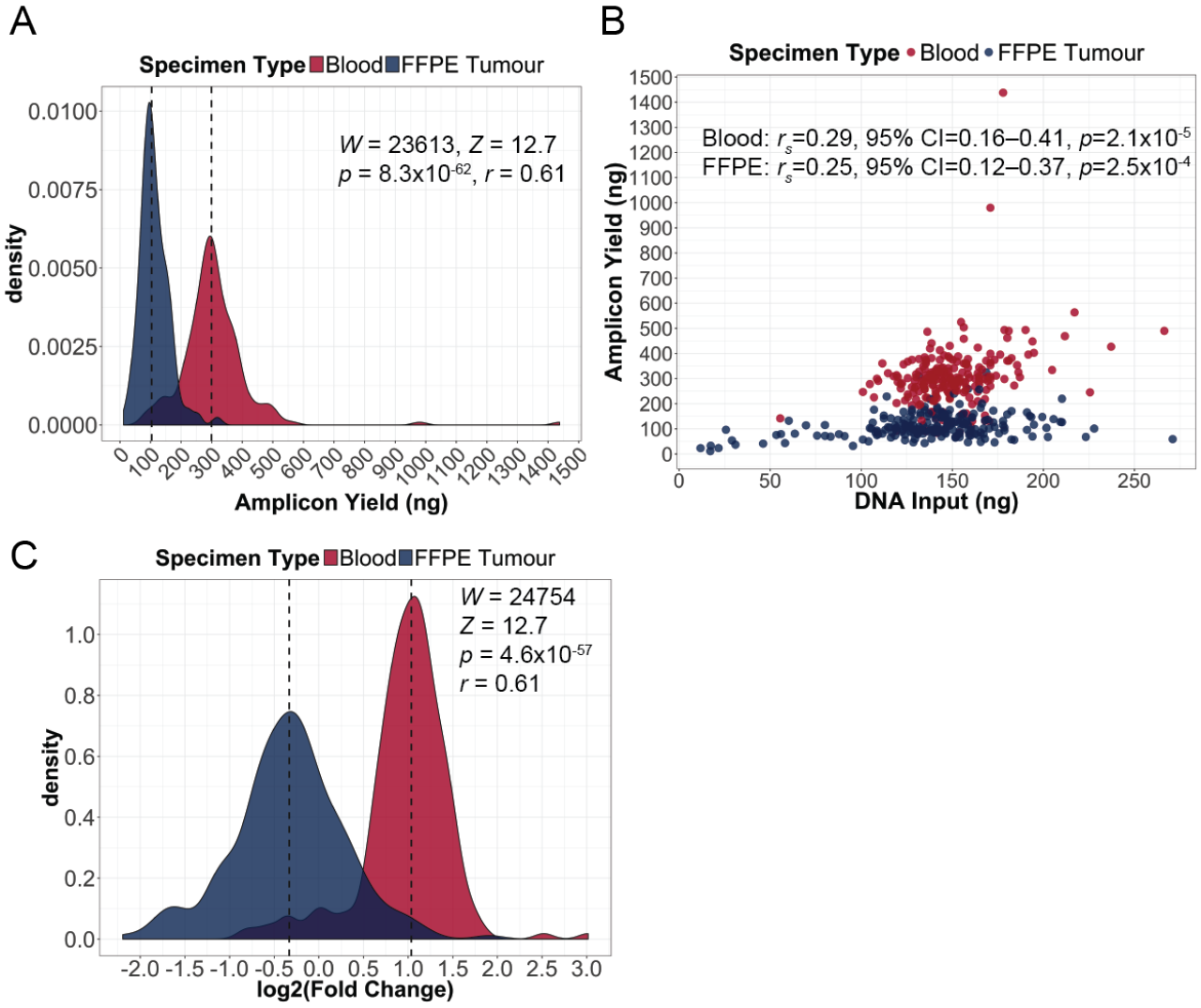


Figure 3.1: Comparison of efficiency in amplicon enrichment between blood and FFPE specimens. (A) Distributions of amplicon yield in blood and FFPE specimens (Wilcoxon signed-rank test). Dashed lines indicate median amplicon yield in blood and FFPE specimens, which are 299.3 ng and 103.6 ng, respectively. (B) Correlations between amplicon yield and the amount of DNA input for amplicon enrichment in blood and FFPE specimens (Spearman's rank correlation). (C) Distributions of fold change between DNA input and amplicon yield ($\log_2$), which is used to measure efficiency in amplicon enrichment in blood and FFPE specimens (Wilcoxon signed-rank test). Dashed lines indicate median $\log_2$ fold change in blood and FFPE specimens, which are 1.04 and -0.332, respectively.

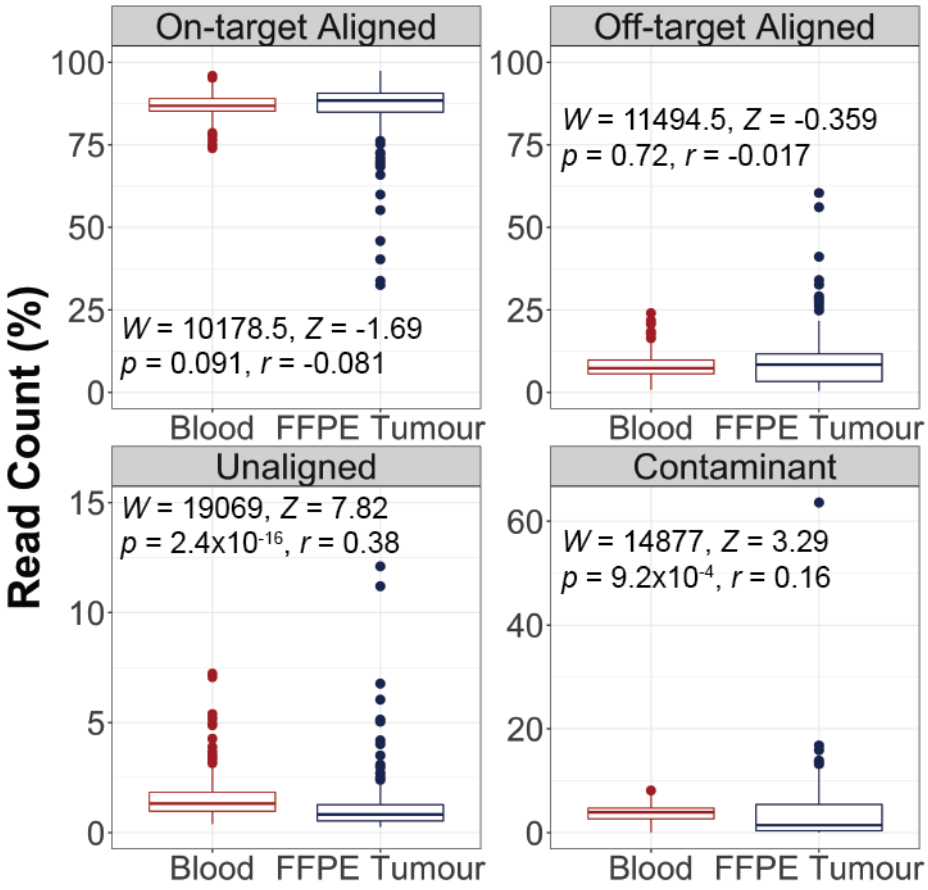


Figure 3.2: Assessment of read alignments between blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and interquartile range (IQR) of percentage of reads, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers.

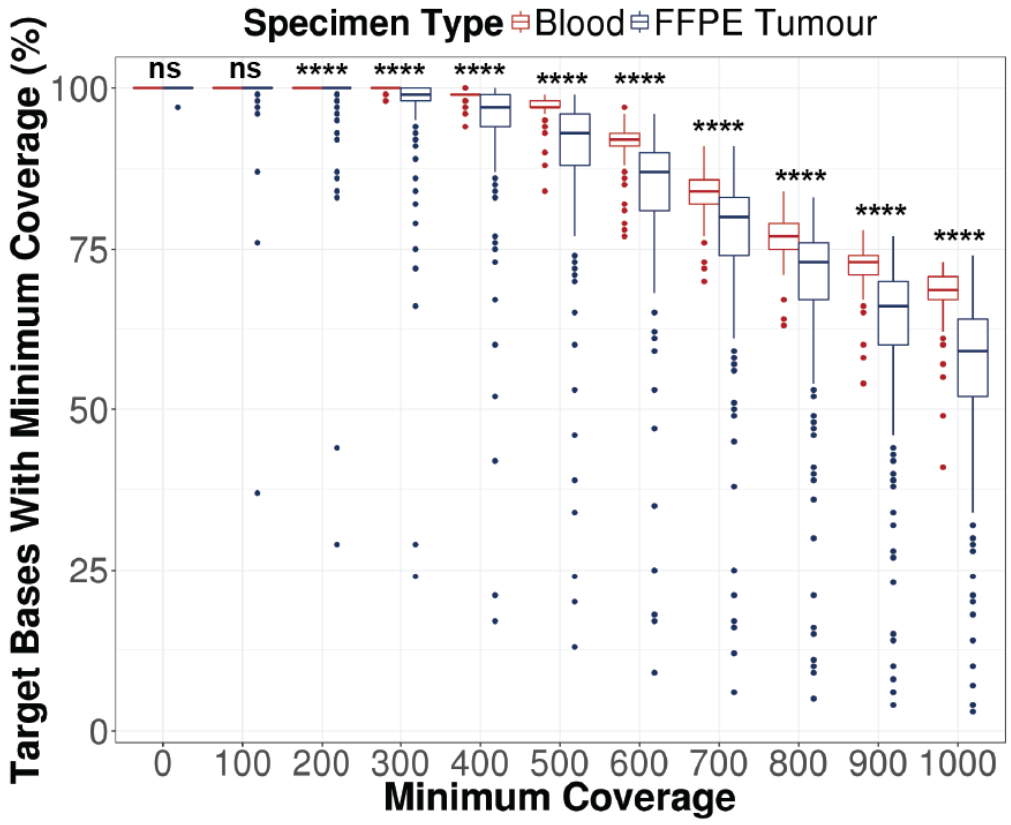


Figure 3.3: Evaluation of coverage uniformity in blood and FFPE specimens (Wilcoxon signed-rank test, ****$p < 0.0001$, ns = not significant). Per base coverage was normalized to account for difference in library size. Percentage of target bases that met various coverage thresholds was calculated. Box plots show the median (horizontal bar within) and IQR of percentage of target bases that met the respective coverage thresholds, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers.

Table 3.1: Comparison of coverage uniformity between blood and FFPE specimens using the Wilcoxon signed-rank test.

	Blood		FFPE Tumour			
Threshold	Median (%)	Range (%)	Median (%)	Range (%)	$D^{\dagger}$ (%)	p ($< 0.0001^{*}$)
$\geq$ 0x	100	100–100	100	97.0–100	0.0	1.0
$\geq$ 100x	100	100–100	100	37.0–100	0.0	2.3×10^{-4}
$\geq$ 200x	100	100–100	100	29.0–100	0.0	$2.9 \times 10^{-11*}$
$\geq$ 300x	100	98.0–100	99.0	24.0–100	1.0	$4.1 \times 10^{-18*}$
$\geq$ 400x	99.0	94.0–100	97.0	17.0–100	2.0	$5.0 \times 10^{-28*}$
$\geq$ 500x	97.0	84.0–99.0	89.5	13.0–99.0	7.5	$2.1 \times 10^{-38*}$
$\geq$ 600x	92.0	77.0–97.0	87.0	9.0–96.0	5.0	$1.5 \times 10^{-32*}$
$\geq$ 700x	84.0	70.0–91.0	80.0	6.0–91.0	4.0	$5.7 \times 10^{-25*}$
$\geq$ 800x	77.0	63.0-84.0	73.0	5.0–83.0	4.0	$4.7 \times 10^{-27*}$
$\geq$ 900x	73.0	54.0–78.0	66.0	4.0–77.0	7.0	$4.6 \times 10^{-40*}$
$\geq$ 1000x	68.5	41.0–73.0	59.0	3.0-74.0	9.5	$3.6 \times 10^{-42*}$

†Absolute difference between median of blood and FFPE specimens.

3.2 Reduced coverage depth in FFPE specimens is more pronounced for longer amplicons

The OncoPanel consists of 416 amplicons that interrogate coding exons and mutational hotpots of 21 genes, and these amplicons vary in length and GC content. Since we observed discrepancy in sequencing coverage between blood and FFPE specimens, we sought to determine whether this discrepancy is influenced by amplicon length and GC content. We obtained the coverage depth for each amplicon and normalized the coverage depth to account for difference in library size. We found significant differences in coverage depth between blood and FFPE specimens for 331 out of 416 amplicons (Wilcoxon signed-rank test with Benjamini-Hochberg correction, adjusted $p < 0.0001$; Figure 3.4). To quantify the amplicon-specific differences in coverage depth, we derived the $\log_2$ fold change in the median coverage depth between blood and FFPE specimens ($\log_2$ (Median Coverage$_{\text{FFPE}}$/Median Coverage$_{\text{Blood}}$)) for each amplicon. Hence, a negative fold change indicates lower coverage depth of the amplicon in FFPE specimens relative to blood specimens, whereas a positive fold change indicates higher coverage depth of the amplicon in FFPE specimens relative to blood specimens. The volcano plot showed that 217 out of the 331 amplicons have negative $\log_2$ fold changes, whereas 114 out of the 331 amplicons have positive $\log_2$ fold changes (Figure 3.4). These results indicate that there are differences in coverage depth between FFPE and blood specimens for a large proportion of amplicons in the panel, with substantially more amplicons exhibiting lower coverage depth in FFPE specimens than blood specimens.

We subsequently examined the impact of amplicon length and GC content on the amplicon-specific differences in coverage depth between specimen types, which we measured as the $\log_2$ fold change in median coverage depth between blood and FFPE specimens. We first confirmed that no significant correlation exists between amplicon GC content and length (Pearson's correlation, r = 0.045, 95% CI = -0.051–0.14, p = 0.36; Figure 3.5). We then evaluated the correlation between $\log_2$ fold change in amplicon coverage depth and amplicon length, and Pearson's correlation demonstrated a strong, negative correlation between the two variables (r = -0.79, 95% CI = -0.82– -0.75, $p = 1.4 \times 10^{-88}$; Figure 3.6A). This result indicates that coverage depth in FFPE specimens tend to be lower relative to blood specimens as amplicon length increases. On the other hand, coverage depth tend to be enriched in FFPE specimens relative to blood for shorter amplicons. We also assessed the correlation between $\log_2$ fold change in amplicon coverage depth and amplicon GC content, and Pearson's correlation demonstrated a weak, negative correlation between the two variables (r = -0.31, 95% CI = -0.40– -0.22, $p = 1.1 \times 10^{-10}$; Figure 3.6B). Although the correlation is weak, this finding still implies that coverage depth in FFPE specimens tend to be lower relative to blood specimens as amplicon GC content increases, whereas enriched coverage depth in FFPE specimens with respect to blood was observed for amplicons with lower GC content.

Because amplicon length and GC content demonstrated significant correlations with amplicon-specific differences in coverage depth, we determined which contributing factor has a greater effect.

We used a multiple linear regression to predict $\log_2$ fold change in amplicon coverage depth based on amplicon length and GC content (Table 3.2). A significant equation was found ($F(2, 411) = 471$, $p = 4.65 \times 10^{-107}$), with an adjusted R^2 of 0.695. Predicted $\log_2$ fold change in amplicon coverage depth between blood and FFPE specimens is equal to

$$1.66 - 7.24 \times 10^{-3}(\textit{Length}) - 9.92 \times 10^{-3}(\textit{GC Content}),$$

in which amplicon length is expressed in base pairs (bp) and GC content is expressed as percentage (%). Both amplicon length and GC content were significant predictors of $\log_2$ fold change in amplicon coverage depth. Based on the standardized coefficients, we compared the strength of predictors within the model to identify the predictor with a greater effect on the response variable. Our assessment showed that one standard deviation increase in amplicon length would lead to a 0.775 standard deviation decrease in $\log_2$ fold change in amplicon coverage depth, whereas one standard deviation increase in amplicon GC content would lead to a 0.277 standard deviation decrease in $\log_2$ fold change in amplicon coverage depth. This result indicates that amplicon length has a stronger association with amplicon-specific differences in coverage depth between specimen types, which we measured as the $\log_2$ fold change in amplicon coverage depth between blood and FFPE specimens, than GC content. Collectively, these findings reveal the challenge imposed by fragmentation damage in FFPE DNA, which results in shorter template DNA that would not be amenable to PCR amplification of longer amplicons.

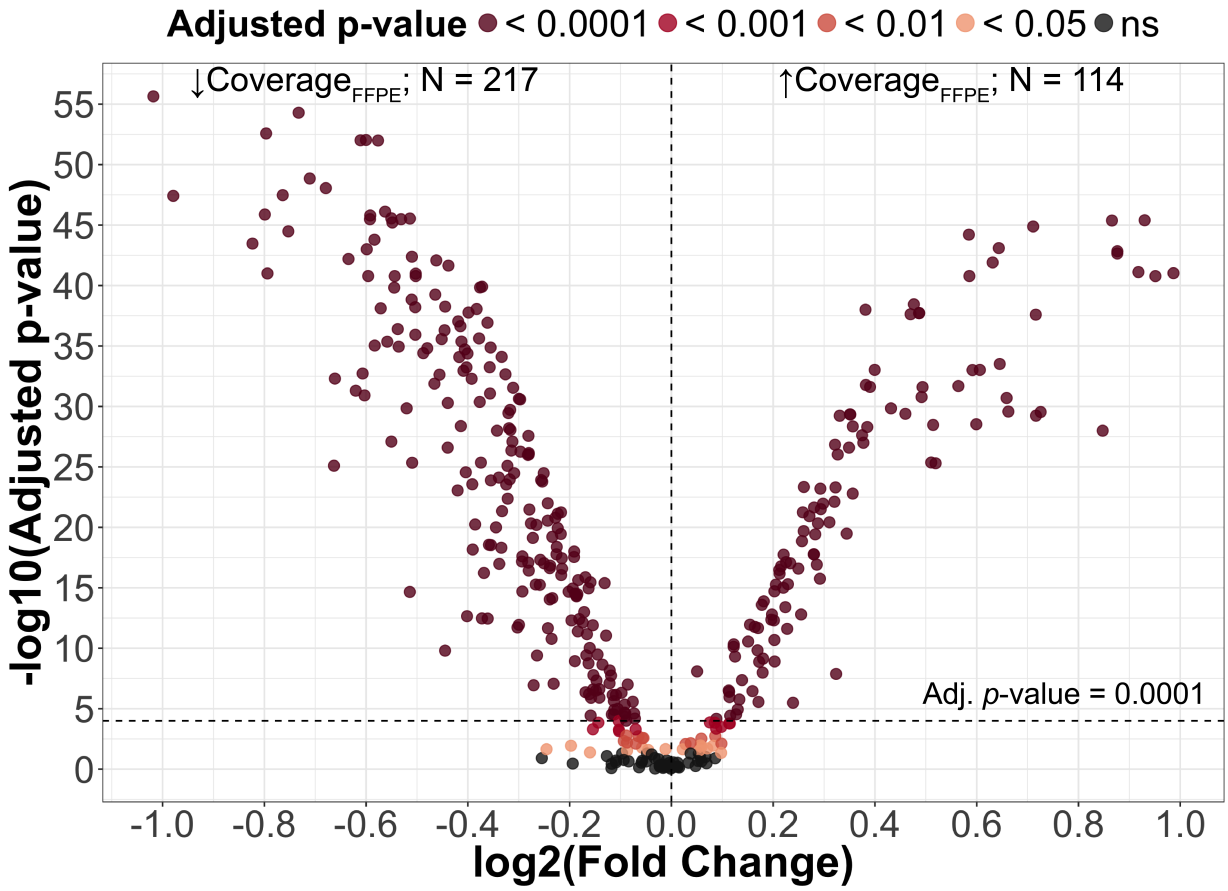


Figure 3.4: Amplicon-specific differences in coverage depth between blood and FFPE specimens. Difference in amplicon coverage depth between specimen types was determined using the Wilcoxon signed-rank test with Benjamini-Hochberg correction (adjusted $p < 0.0001$). Volcano plot illustrates the -$\log_{10}$ adjusted p-value in relation to $\log_2$ fold change between median coverage depth in blood and FFPE specimens ($\log_2$ (Median Coverage$_{FFPE}$/Median Coverage$_{Blood}$)) for amplicons in the panel. Negative $\log_2$ fold change indicates lower coverage depth of the amplicon in FFPE specimens relative to blood (↓ Coverage$_{FFPE}$), whereas positive $\log_2$ fold change indicates higher coverage depth of the amplicon in FFPE specimens relative to blood (↑ Coverage$_{FFPE}$). N = number of amplicons; ns = not significant

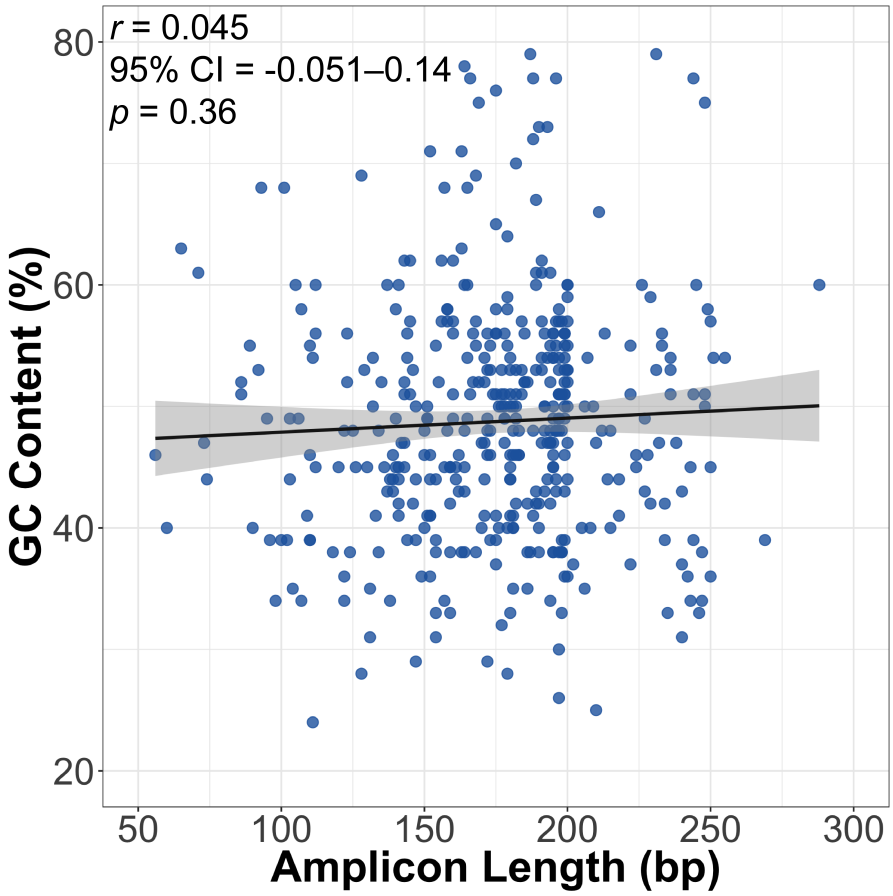


Figure 3.5: The relationship between amplicon GC content and amplicon length (Pearson's correlation). Solid line represents the fitted linear relationship between the two variables, and the shaded band indicates pointwise 95% confidence interval of the fitted linear regression line.

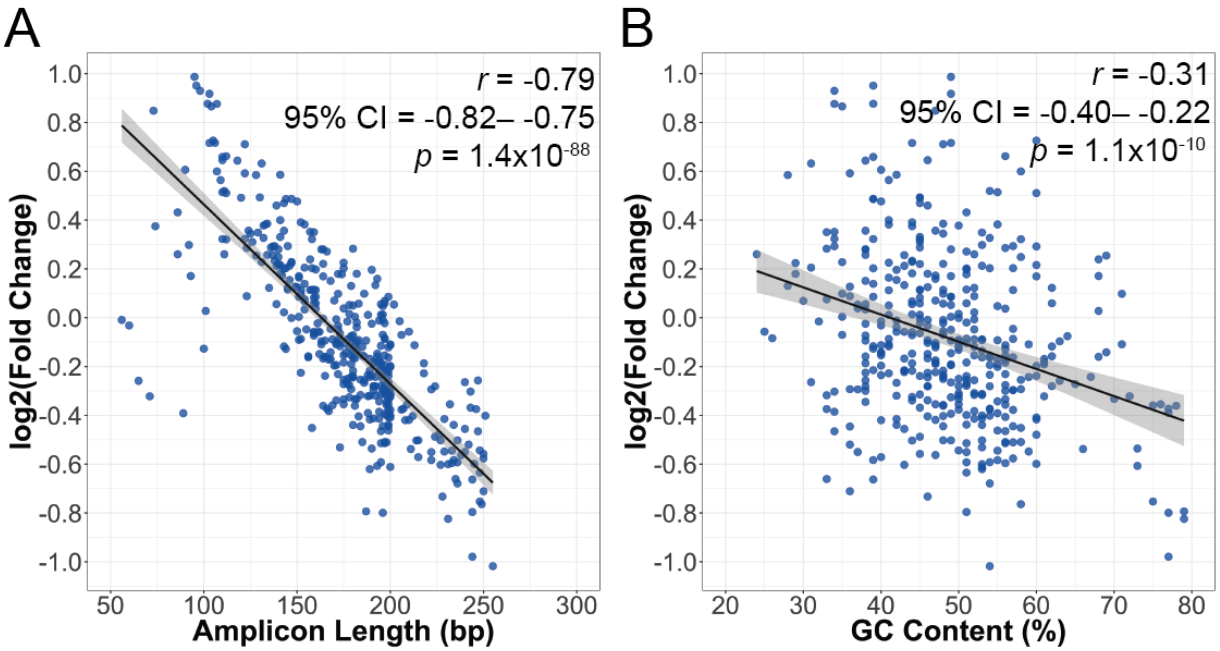


Figure 3.6: Scatter plots showing $\log_2$ fold change between amplicon coverage depth in blood and FFPE specimens ($\log_2$ (Median Coverage$_{\text{FFPE}}$/Median Coverage$_{\text{Blood}}$)) in relation to (A) amplicon length and (B) GC content (Pearson's correlation). Solid line represents the fitted linear relationship between the two variables, and the shaded band indicates pointwise 95% confidence interval of the fitted linear regression line.

Table 3.2: Multiple linear regression to predict $\log_2$ fold change between amplicon coverage depth in blood and FFPE specimens ($\log_2$ (Median Coverage$_{\text{FFPE}}$/Median Coverage$_{\text{Blood}}$)) based on amplicon length and GC content.

Variable	Unstandardized Coefficient	Standard Error	Standardized Coefficient	p-value
Length (bp)	-7.24×10^{-3}	2.54×10^{-4}	-7.75×10^{-1}	2.47×10^{-99}
GC Content (%)	-9.92×10^{-3}	9.77×10^{-4}	-2.77×10^{-1}	8.70×10^{-22}
				Intercept = 1.66, Adjusted R^2 = 0.695
				$F(2, 411) = 471$, p-value = 4.65×10^{-107}

3.3 Deamination effects lead to increased C>T/G>A transitions in FFPE specimens

Formalin fixation not only induces DNA fragmentation, but also base modifications that give rise to sequence artifacts [41–43, 65, 76, 100, 101, 145]. A prominent type of formalin-induced sequence artifact is C>T/G>A transitions as a result of deamination of cytosine bases [42, 76, 83, 101, 145]. To measure the level of formalin-induced artifacts in FFPE specimens, we quantified the fraction of base changes that were not identified as true SNVs by our variant calling pipeline. We only considered high quality bases (BAQ $\geq$ 20) and base changes that were $\geq$ 1% allele frequency to exclude sequencing errors from our analysis. Base changes were categorized into C>T/G>A and A>G/T>C, which are nucleotide transitions, as well as C>A/G>T, A>C/T>G, C>G/G>C, and A>T/T>A, which are nucleotide transversions. We compared the fraction of base changes between specimen types and found significant differences in fraction of C>T/G>A and A>G/T>C between blood and FFPE specimens (Wilcoxon signed rank test, $p < 0.0001$; Figure 3.7A). As blood DNA is not affected by formalin fixation, we evaluated the prevalence of artifactual base changes in FFPE specimens with respect to blood by calculating the fold change between the median fraction of base changes in blood and FFPE specimens (Table 3.3). We noted a substantially higher fold change for C>T/G>A compared to A>G/T>C: fraction of C>T/G>A was 23 times higher in FFPE specimens relative to blood, whereas fraction of A>G/T>C was 3.1 times higher in FFPE specimens relative to blood. This result is consistent with cytosine deamination effects that are reportedly predominant in FFPE DNA. As well, increased A>G/T>C base changes could be caused by incorporation of guanines at abasic sites [63]. In the presence of atmospheric oxygen, formaldehyde can be oxidized into formic acid, causing depurination, which gives rise to abasic sites [42].

To assess the relative difference in fraction of base changes in FFPE specimens compared to blood specimens, we calculated the $\log_2$ fold change in fraction of base changes between paired blood and FFPE specimens ($\log_2$ (Fraction of Base Changes$_{\text{FFPE}}$/Fraction of Base Changes$_{\text{Blood}}$)). We compared the relative difference in fraction of base changes across different types of base changes, and a Kruskal-Wallis test indicated that type of base changes has a significant effect on the relative difference in fraction of base changes ($H = 428.5$, $p = 2.1 \times 10^{-90}$; Figure 3.8). Multiple pairwise comparison of the relative difference in fraction of base changes was performed using a post-hoc Dunn's test with Benjamini-Hochberg correction. Relative difference in fraction of C>T/G>A was significantly different compared to the five other types of base changes, and this was similar for A>G/T>C (adjusted $p < 0.0001$; Table 3.4). Although both C>T/G>A and A>G/T>C were elevated in FFPE specimens compared to the other base transversions, the magnitude of difference was larger for C>T/G>A than A>G/T>C (median $\log_2$ fold change: C>T/G>A = 4.2, A>G/T>C = 1.6), which further confirms that deamination of cytosine bases is the most frequent form of sequence artifact in FFPE DNA.

Formalin-induced sequence artifacts often occur at low allele frequency; hence, we examined

the prevalence of sequence artifacts at different ranges of allele frequency, including 1–10%, 10-20%, and 20-30%. Because variants were not called within the 1–10% allele frequency range, we did not remove true SNVs detected by our variant calling pipeline to ensure consistency when comparing fraction of base changes across different ranges of allele frequency. Nevertheless, we adhered to the previous criterion of only including base changes with BAQ $\geq$ 20 in this analysis. For all types of base changes, we noted that the range of allele frequency has a significant effect on fraction of base changes in blood and FFPE specimens (Kruskal-Wallis test, $p < 0.0001$; Figure 3.9), with increased levels of base changes at the 1-10% allele frequency range compared to 10-20% and 20-30%. Because blood DNA represents good quality DNA that is unaffected by formalin fixation, we also compared the fraction of base changes at the 1-10% allele frequency range in FFPE specimens to blood. Similar to previous analyses, there was a marked increase in C>T/G>A and a modest increase in A>G/T>C in FFPE specimens relative to blood within the 1-10% allele frequency (fold change: C>T/G>A = 33, A>G/T>C = 3.1; Table 3.5). Collectively, our assessment demonstrates that high frequency of C>T/G>A transitions is present and detectable in FFPE specimens, which indicates that deamination of cytosine is the primary form of formalin-induced sequence artifact, and these artifactual transitions are more prevalent at low, but clinically relevant allele frequency.

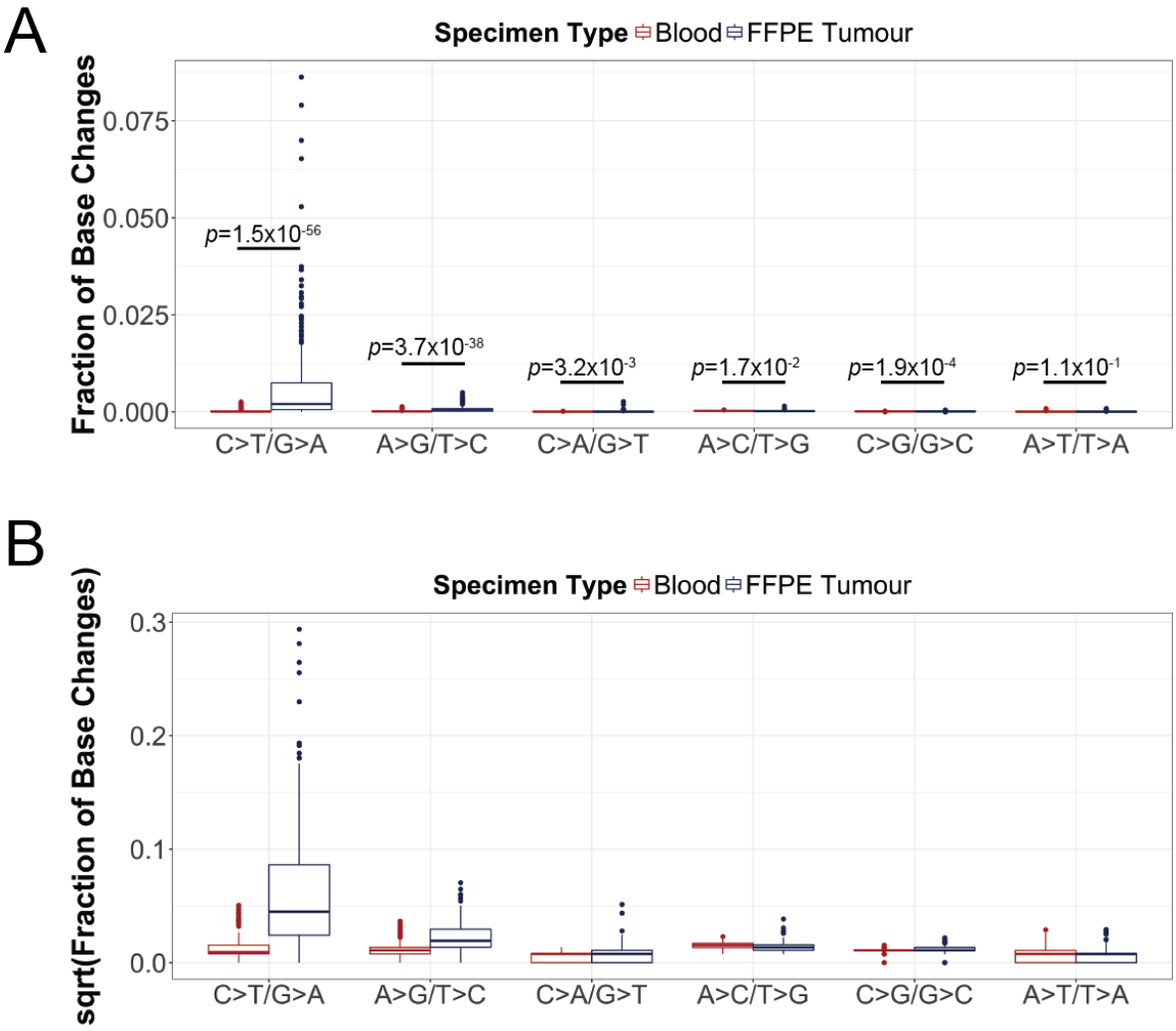


Figure 3.7: Assessment of formalin-induced sequence artifacts in FFPE specimens. (A) Comparison of fraction of base changes in blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and IQR of fraction of base changes for different types of base changes, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers. (B) Box plots showing square root-transformed fraction of base changes on the Y-axis.

Table 3.3: Summary statistics of fraction of base changes in blood and FFPE specimens.

	Blood		FFPE Tumour		
Type of Base Changes	Median	Range	Median	Range	FC[†]
C>T/G>A	8.9×10^{-5}	0–2.6×10^{-3}	2.0×10^{-3}	0–8.6×10^{-2}	23
A>G/T>C	1.2×10^{-4}	0–1.3×10^{-3}	3.7×10^{-4}	0–5.0×10^{-3}	3.1
C>A/G>T	6.0×10^{-5}	0–1.8×10^{-4}	6.0×10^{-5}	0–2.6×10^{-3}	1.0
A>C/T>G	2.4×10^{-4}	5.9×10^{-5}–5.3×10^{-4}	1.8×10^{-4}	5.8×10^{-5}–1.4×10^{-3}	0.77
C>G/G>C	1.2×10^{-4}	0–2.4×10^{-4}	1.2×10^{-4}	0–4.8×10^{-4}	1.0
A>T/T>A	6.0×10^{-5}	0–8.4×10^{-4}	5.9×10^{-5}	0–8.6×10^{-4}	0.99

[†]Fold change (FC) between the median of blood and FFPE specimens.

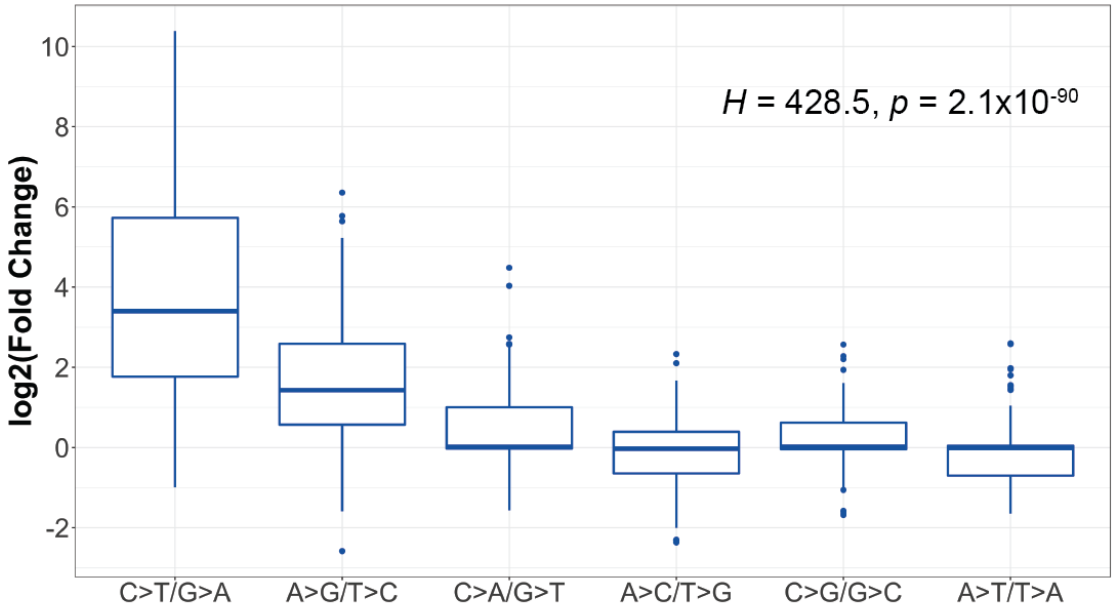
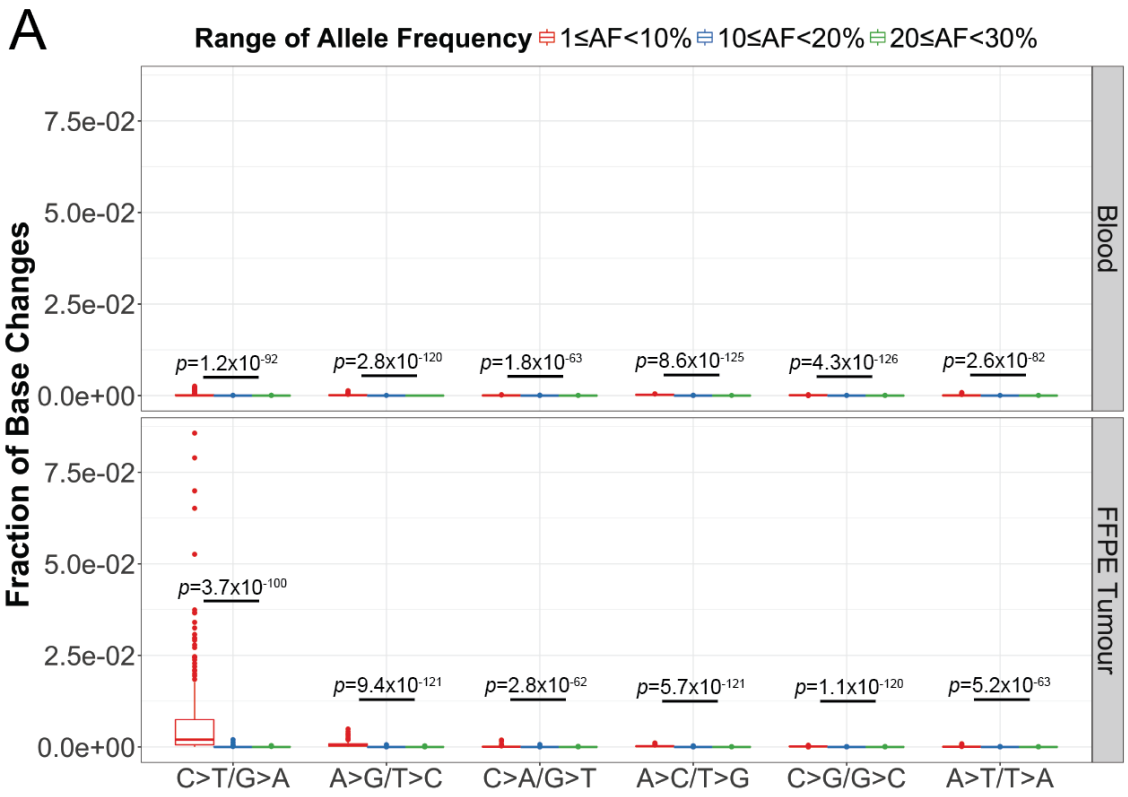


Figure 3.8: Comparison of relative difference in fraction of base changes in FFPE specimens compared to blood (Kruskal-Wallis test). Relative difference was measured as $\log_2$ fold change between fraction of base changes in blood and FFPE specimens ($\log_2$ (Fraction of Base Changes$_{\text{FFPE}}$/Fraction of Base Changes$_{\text{Blood}}$)). Box plots show the median (horizontal bar within) and IQR of $\log_2$ fold change for different types of base changes, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers.

Table 3.4: Multiple pairwise comparison of $\log_2$ fold change in fraction of base changes between blood and FFPE specimens using Dunn's test with Benjamini-Hochberg multiple hypothesis testing correction. Top values represent Dunn's pairwise z statistics, whereas bottom values represent adjusted p-value. Asterisk(*) indicates significance level of adjusted p-value < 0.0001.

Type of Base Changes	A>C/T>G	A>G/T>C	A>T/T>A	C>A/G>T	C>G/G>C
A>G/T>C	-11.7 $4.15 \times 10^{-31}*$				
A>T/T>A	-0.399 3.45×10^{-1}	9.57 $1.31 \times 10^{-21}*$			
C>A/G>T	-3.46 4.00×10^{-4}	6.39 $1.52 \times 10^{-10}*$	-2.73 3.99×10^{-3}		
C>G/G>C	-3.02 1.73×10^{-3}	8.63 $6.76 \times 10^{-18}*$	-2.17 1.71×10^{-2}	0.918 1.92×10^{-1}	
C>T/G>A	-17.1 $7.78 \times 10^{-65}*$	-5.60 $1.76 \times 10^{-8}*$	-14.3 $5.10 \times 10^{-46}*$	-11.1 $1.32 \times 10^{-28}*$	-14.1 $6.46 \times 10^{-45}*$


A
Range of Allele Frequency 1≤AF<10% 10≤AF<20% 20≤AF<30%
Fraction of Base Changes
7.5e-02
5.0e-02
2.5e-02
0.0e+00
Blood
FFPE Tumour
p=1.2x10-92
p=2.8x10-120
p=1.8x10-63
p=8.6x10-125
p=4.3x10-126
p=2.6x10-82
p=3.7x10-100
p=9.4x10-121
p=2.8x10-62
p=5.7x10-121
p=1.1x10-120
p=5.2x10-63
C>T/G>A
A>G/T>C
C>A/G>T
A>C/T>G
C>G/G>C
A>T/T>A

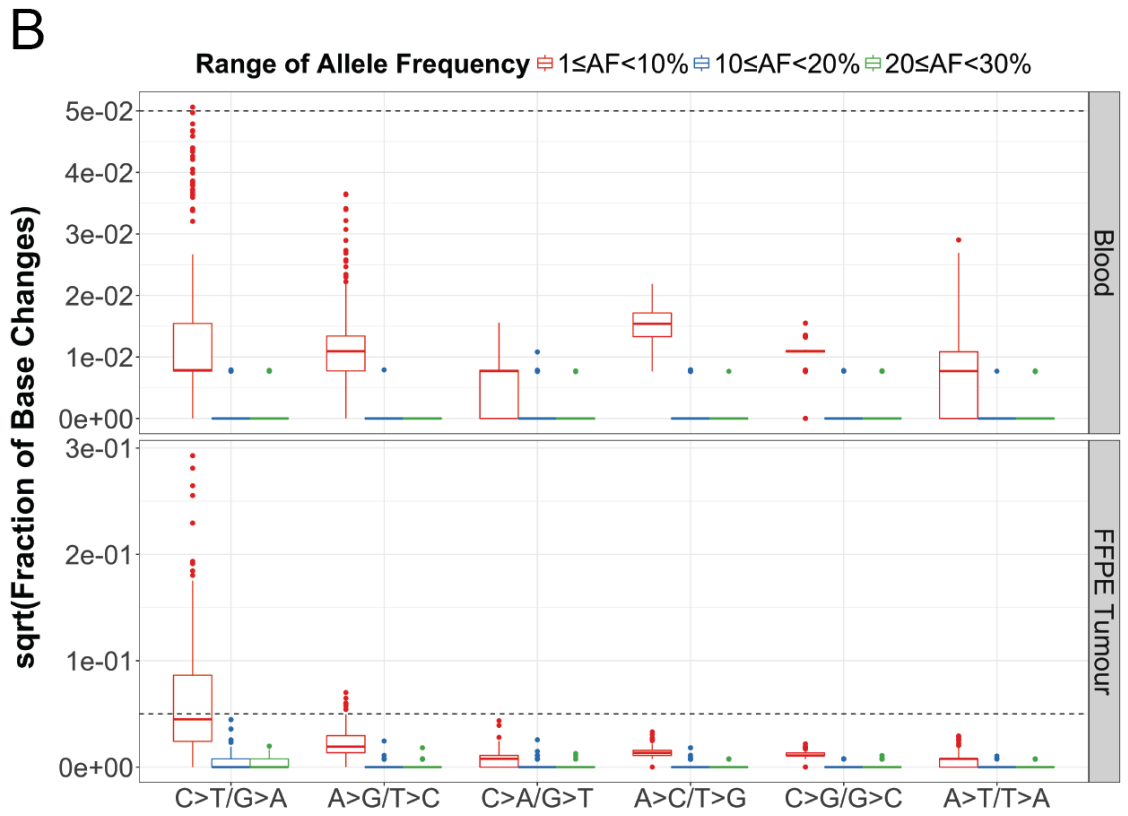

B
Range of Allele Frequency 1≤AF<10% 10≤AF<20% 20≤AF<30%
sqrt(Fraction of Base Changes)
5e-02
4e-02
3e-02
2e-02
1e-02
0e+00
3e-01
2e-01
1e-01
0e+00
Blood
FFPE Tumour
C>T/G>A
A>G/T>C
C>A/G>T
A>C/T>G
C>G/G>C
A>T/T>A

Figure 3.9: Assessment of formalin-induced sequence artifacts in FFPE specimens at different ranges of allele frequency. (A) Comparison of fraction of base changes across different ranges of allele frequency (Kruskal-Wallis test). Box plots show the median (horizontal bar within) and IQR of fraction of base changes for different types of base changes, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers. (B) Box plots demonstrating square root-transformed fraction of base changes across different ranges of allele frequency. Dashed lines equal to 0.05 to indicate that the Y-axis scales are different for blood and FFPE tumour plots.

Table 3.5: Summary statistics of fraction of base changes in blood and FFPE specimens within 1-10% allele frequency.

	Blood		FFPE Tumour		
Type of Base Changes	Median	Range	Median	Range	†FC
C>T/G>A	6.2×10^{-5}	0–2.6×10^{-3}	2.0×10^{-3}	0–8.6×10^{-2}	33
A>G/T>C	1.2×10^{-4}	0–1.3×10^{-3}	3.7×10^{-4}	0–4.9×10^{-3}	3.1
C>A/G>T	6.0×10^{-5}	0–2.4×10^{-4}	6.0×10^{-5}	0–1.9×10^{-3}	1.0
A>C/T>G	2.4×10^{-4}	5.9×10^{-5}–4.8×10^{-4}	1.8×10^{-4}	0–1.1×10^{-3}	0.77
C>G/G>C	1.2×10^{-4}	0–2.4×10^{-4}	1.2×10^{-4}	0–4.8×10^{-4}	1.0
A>T/T>A	6.0×10^{-5}	0–8.4×10^{-4}	5.9×10^{-5}	0–8.6×10^{-4}	0.99

†Fold change (FC) between the median of blood and FFPE specimens.

3.4 Increased age of paraffin block results in reduced amplicon yield and elevated level of C>T/G>A sequence artifacts

The amount of amplifiable DNA derived from FFPE specimens is dependent on the extent of fragmentation damages. Given two FFPE DNA samples of similar quantity, the sample with more extensive DNA fragmentation would yield reduced amount of PCR amplicons compared to the less fragmented sample [40, 145]. Several studies reported increased fragmentation damages in DNA isolated from older paraffin blocks due to longer exposure to environmental conditions [10, 25, 86, 116]. As the age of paraffin blocks in our study ranges from 18 to 5356 days, we hypothesized that older paraffin blocks would result in more extensively fragmented DNA, leading to reduced efficiency in amplicon enrichment. Inspection of the relationship between amplicon yield and age of paraffin block demonstrated a moderate, negative correlation (Spearman's rank correlation, r_s = -0.42, 95% CI = -0.52– -0.30, $p = 1.2 \times 10^{-10}$; Figure 3.10A), suggesting that DNA extraction from older paraffin blocks tend to yield lower amount of amplicons. Because the amount of DNA input for production of amplicons varies across specimens in our study design, a representation of efficiency in amplicon enrichment would be the $\log_2$ fold change between DNA input and amplicon yield. Thus, we assessed the correlation between $\log_2$ fold change and the storage time of FFPE blocks. Similarly, there was a moderate, negative correlation between $\log_2$ fold change and age of paraffin block (Spearman's rank correlation, r_s = -0.42, 95% CI = -0.53– -0.30, $p = 1.2 \times 10^{-10}$; Figure 3.10B), implying that production of amplicons is less efficient in FFPE DNA extracted from older paraffin blocks, which is likely caused by more substantial DNA fragmentation.

There are also studies that revealed increased frequency of sequence artifacts in FFPE DNA that are exceedingly fragmented [25, 145]. As DNA fragmentation results in reduced template DNA for PCR amplification, this leads to a higher probability for enrichment of sequence artifacts. Our previous evaluation indicated that older paraffin blocks were associated with lower efficiency in amplicon enrichment, which is possibly due to increased fragmentation damages in the extracted DNA. This leads to our hypothesis that older paraffin blocks would yield elevated levels of sequence artifacts, particularly C>T/G>A transitions, which are the most prominent type of formalin-induced base modifications. To address our hypothesis, we assessed the relationship between fraction of base changes and age of paraffin blocks for different types of base changes (Figure 3.11). There was a moderate, positive correlation between fraction of C>T/G>A transitions and age of paraffin block (Spearman's rank correlation, r_s = 0.54, 95% CI = 0.43–0.63, $p = 1.0 \times 10^{-17}$). We also noted a positive correlation between fraction of A>G/T>C and age of paraffin block (Spearman's rank correlation, r_s = 0.29, 95% CI = 0.16–0.40, $p = 2.1 \times 10^{-5}$), albeit a weak one. As for transversion base changes (i.e. C>A/G>T, A>C/T>G, C>G/G>C, and A>T/T>A), no significant correlations with age of paraffin block were observed (Spearman's rank correlation, $p < 0.05$). These findings reveal that increased detection of sequence artifacts, especially the common C>T/G>A changes in FFPE specimens, is associated with long term storage of FFPE blocks.

We subsequently examined how pre-sequencing variables such as age of paraffin block and efficiency in amplicon enrichment correlate with sequencing metrics, which include average per base coverage (normalized to account for library size), percentage of on-target alignments, and fraction of C>T/G>A changes (Table 3.6). This assessment would provide insight on how pre-sequencing variables can affect sequencing results, thereby facilitating sample selection if multiple specimens are available before sequencing. We noted a moderate, negative correlation between average per base coverage and age of paraffin block (Spearman's rank correlation, r_s = -0.47, 95% CI = -0.57– -0.36, $p = 4.7 \times 10^{-7}$), and a weak, negative correlation between percentage of on-target aligned reads and age of paraffin block (Spearman's rank correlation, r_s = -0.35, 95% CI = -0.46– -0.23, $p = 8.2 \times 10^{-3}$). Conversely, we observed a moderate, positive correlation between average per base coverage and efficiency in amplicon enrichment (Spearman's rank correlation, r_s = 0.52, 95% CI = 0.42–0.61, $p = 2.3 \times 10^{-11}$), and a weak, positive correlation between percentage of on-target aligned reads and efficiency in amplicon enrichment (Spearman's rank correlation, $r_s = 0.35$, 95% CI = 0.22–0.45, $p = 2.9 \times 10^{-5}$). Since efficiency in amplicon enrichment is inversely correlated with storage time of FFPE blocks, opposing correlations with sequencing metrics were expected for both pre-sequencing variables. Furthermore, there was also a moderate, negative correlation between fraction of C>T/G>A and efficiency in amplicon enrichment (Spearman's rank correlation, r_s = -0.55, 95% CI = -0.64– -0.45, $p = 2.0 \times 10^{-20}$). As reduced efficiency in amplicon enrichment is an indicator for low amount of template DNA, the consequent increase in C>T/G>A changes is the outcome of stochastic enrichment of sequence artifacts. Together, these results reveal that pre-sequencing variables such as age of paraffin block and efficiency in amplicon enrichment could be predictors of sequencing metrics, in which older FFPE blocks are more likely to yield lower efficiency in amplicon enrichment, leading to poorer sequencing results and increased prevalence of artifactual C>T/G>A transitions.

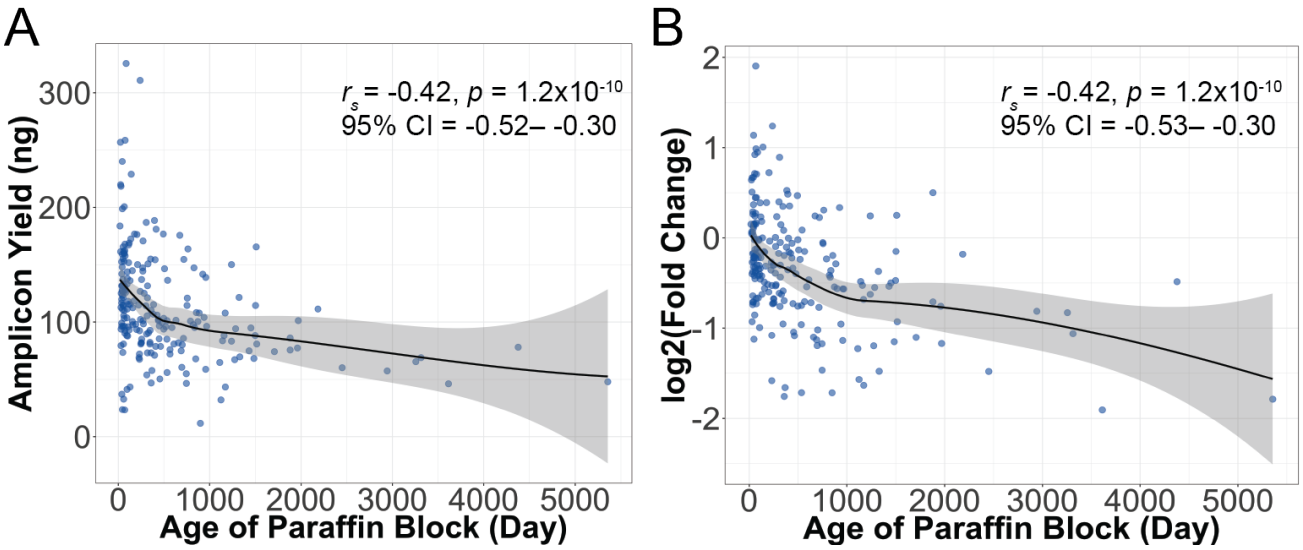


Figure 3.10: Scatter plots showing (A) amplicon yield and (B) efficiency in amplicon enrichment, which is represented by the $\log_2$ fold change between the amount of DNA input for producing amplicons and amplicon yield, in relation to age of paraffin blocks (Spearman's rank correlation). Solid lines represent locally weighted smoothing (LOESS) curves, with shaded bands indicating 95% confidence interval of the LOESS curves.

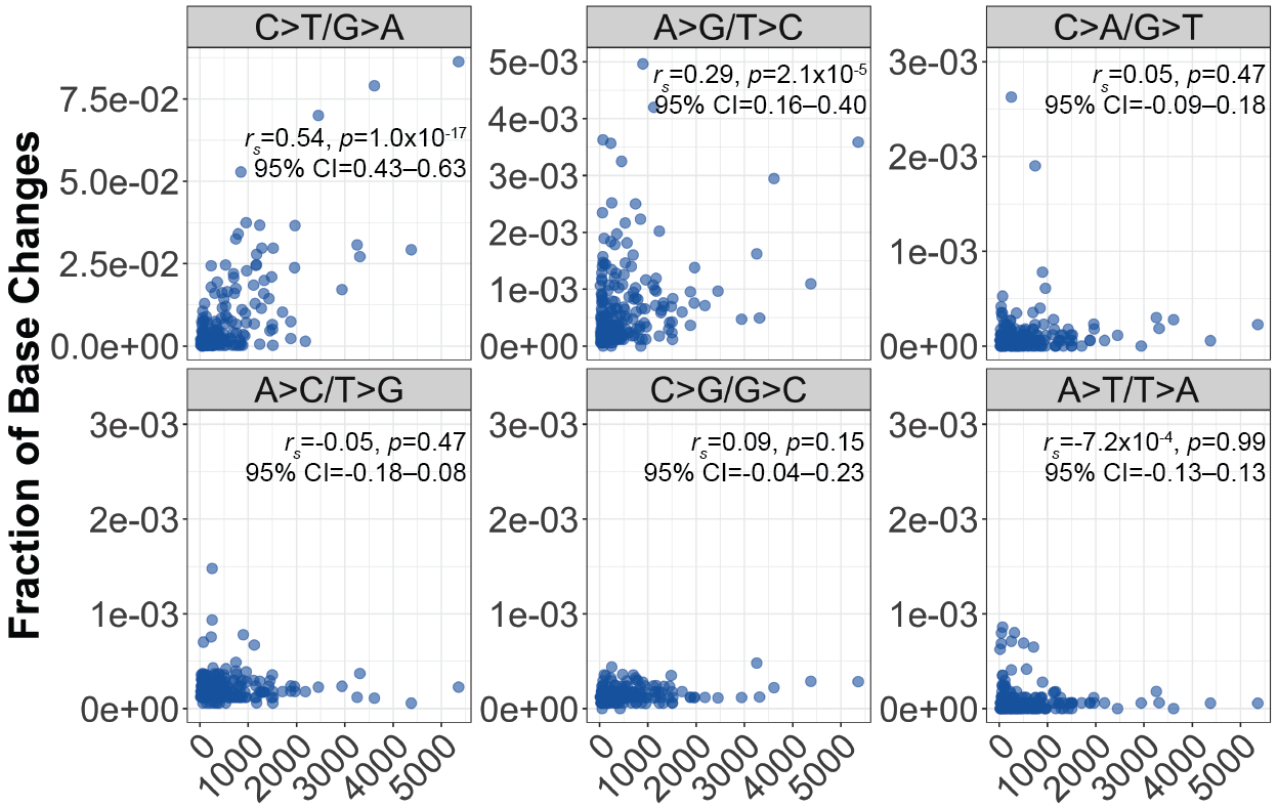


Figure 3.11: The relationship between fraction of base changes and age of paraffin block for different types of base changes (Spearman's rank correlation).

Table 3.6: Spearman's rank correlation between pre-sequencing variables (e.g. enrichment efficiency and age of paraffin block) and sequencing metrics (e.g. fraction of C>T/G>A, average per base normalized coverage, and on-target aligned reads). Top values represent Spearman's *rho* and 95% confidence interval in brackets, whereas bottom values represent *p*-value. Asterisk(*) indicates significance level of p-value < 0.05.

Variable	Enrichment Efficiency[†]	Age of Paraffin Block (Day)	Fraction of C>T/G>A	Average Per Base Normalized Coverage
Age of Paraffin Block (Day)	-0.42 (-0.53– -0.30) $9.3 \times 10^{-10}*$			
Fraction of C>T/G>A	-0.55 (-0.64– -0.45) $2.0 \times 10^{-20}*$	0.54 (0.43–0.63) $1.0 \times 10^{-17}*$		
Average Per Base Normalized Coverage	0.52 (0.42–0.61) $2.3 \times 10^{-11}*$	-0.47 (-0.57– -0.36) $4.7 \times 10^{-7}*$	-0.80 (-0.84– -0.75) $7.5 \times 10^{-17}*$	
On-target Aligned Reads (%)	0.34 (0.22–0.45) $2.9 \times 10^{-5}*$	-0.35 (-0.46– -0.23) $8.2 \times 10^{-3}*$	-0.57 (-0.65– -0.47) $4.2 \times 10^{-8}*$	0.73 (0.66–0.79) $3.1 \times 10^{-58}*$

[†]$\log_2$ fold change between DNA input for amplicon enrichment and amplicon yield.

Chapter 4

Identification of Germline Alterations in FFPE Tumours

Tumour-only sequencing is commonly performed by clinical laboratories to detect targetable somatic mutations, which can inform clinical decision making. Unlike the research setting, matched normal samples such as blood, saliva, or adjacent normal tissues are not routinely processed in the clinical setting due to limited sample availability, funding, and time [53, 54, 83, 143]. The tumour genome also contains germline information that may have clinical implications for patients and their families. For instance, germline alterations in cancer-predisposing genes could facilitate implementation of cancer preventative measures such as early screening and sibling testing [93, 113]. Moreover, germline PGx variants could predict response to drugs like chemotherapeutic agents, thereby preventing adverse drug reactions [36, 51, 71, 80, 91, 95, 96, 133].

Because the tumour genome consists of both germline and somatic alterations, it is important to establish approaches to distinguish between germline and somatic alterations in cancer diagnostic assays that only sequence tumour DNA. In the absence of matched normal samples, approaches such as constructing a virtual normal by combining variants identified in multiple normal samples from healthy individuals and filtering variants using public databases such as such as dbSNP, 1000 Genomes Project, and COSMIC could enable the differentiation of germline variations from somatic mutations [64, 72]. Subsequently, potential germline alterations can be referred to follow-up testing, which involves genetic counseling and collection of germline samples for further sequencing and analysis [17, 59, 106].

The TOP study is comprised of 213 patients with tumour and matched blood specimens. We interpreted the germline variants identified in blood specimens from TOP patients using the effect prediction software, SnpEff (version 4.2), and ExAC and 1000 Genomes databases, which provide information on population frequency. We also annotated the variant calls with the ClinVar database, which enable assessment of clinical significance. Furthermore, we performed manual literature

review to determine the functional and clinical impacts of all germline alterations detected in the blood samples. Because several studies demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [72, 93, 106, 113], we sought to confirm the presence of germline alterations in the tumour genome by measuring variant concordance between blood and tumour DNA. This enables us to determine whether tumour DNA is a reliable substrate for identification of germline alterations.

Lastly, we differentiated between germline and somatic statuses of variants identified in tumour DNA through applying VAF thresholds. While heterozygous germline variants are expected to have VAF of close to 50%, homozygous germline variants are expected to have VAF of close to 100%. In contrast, the VAF of somatic mutations relies on tumour purity. Due to contamination of normal tissues in tumour specimens, it is highly likely that the VAFs of somatic mutations are substantially lower than the expected VAFs for germline alterations []. Furthermore, other factors such as tumour heterogeneity and formlin-induced DNA damage could cause deviation of somatic VAFs from the expected 50% and 100% for heterozygous and homozygous variants, respectively []. As we have matched blood samples for all tumour samples, we were able to evaluate the sensitivity of using VAF thresholds to discriminate between germline and somatic alterations in tumour DNA. Furthermore, we also assessed the positive predictive value of referring potential germline alterations for follow-up testing. Through these analyses, we hope to establish a VAF cut-off that could maximize true positive rate for identification of potential germline alterations, as well as minimize false positive rate to reduce unnecessary follow-up testing, which could cause patients preventable psychological distress and hassles.

Together, our analyses would provide insights on whether application of VAF thresholds is a practical approach to distinguish between germline and somatic alterations in tumour-only sequencing assays. Hence, this will determine whether tumour-only sequencing assays can be leveraged by clinical laboratories for initial screening of germline alterations that are clinically relevant.

4.1 Frequency and variant assessment of germline alterations in patients from TOP cohort

We examined 15 cancer-related genes and six PGx genes in DNA isolated from blood samples from the 213 cancer patients in TOP cohort. We identified a total of 1990 germline alterations that passed our filtering criteria (Figure 2.1B). In 212 out of 213 patients, we detected a total of 1205 variants in the 15 cancer-related genes screened by the OncoPanel, with an average of 5.7 variants per patient (standard error = 0.15, range = 1–11 variants; Table 4.1). These germline alterations were found at 50 genomic positions and interpreted using various bioinformatics approaches and literature review (Table 4.2). Through effect prediction using the SnpEff software, we demonstrated that 78% of these variants were synonymous, 16% were missense variants, 4% occurred within splice regions, and 2% were frameshift variants. Eighteen out of the 50 germline variants were classified as common variants by the 1000 Genomes Project with population frequencies of $\geq 1\%$ in the ExAC database, whereas eight out of the 50 variants were classified as rare variants with population frequencies of $< 1\%$ in the ExAC database.

To assess clinical significance of the 50 germline alterations in cancer-related genes, we used information in the ClinVar database. Our assessment revealed 16% benign variants, 16% likely benign variants, 12% annotated as benign/likely benign, 4% with conflicting interpretations of pathogenicity, and 2% with uncertain significance. We were unable to determine the clinical significance of 48% of the 50 germline variants because these variants were not reported in the ClinVar database. While we found no variants that were pathogenic or likely pathogenic, we identified one TP53 variant, p.Arg72Pro/c.215G>C (rs1042522), that is associated with drug response. Based on literature review, clinical studies revealed that the Pro/Pro genotype results in severe neutropenia in ovarian cancer patients receiving cisplatin-based chemotherapy, and poor survival and treatment response in gastric cancer patients receiving paclitaxel and capecitabine combination chemotherapy, as well as 5-fluorouracil-based adjuvant chemotherapy []. The combination of evidence from our literature review and the ClinVar database suggests that the TP53 p.Arg72Pro/c.215G>C (rs1042522) could be potentially useful in guiding therapeutic intervention for cancer patients.

Furthermore, we identified a total of 785 variants in the six PGx genes screened by the OncoPanel in 212 out of 213 patients, with an average of 3.7 germline alterations per patient (standard error = 0.10, range = 1–8 variants; Table 4.3). These PGx variants occurred at 23 genomic positions and were interpreted using similar methods to the germline alterations identified in cancer-related genes (Table 4.4). Effect prediction using the SnpEff software demonstrated that 57% of these 23 germline variants were missense variants, 17% were synonymous, 9% occurred within splice regions, 9% occurred upstream of a gene, 4% were located at splice donor sites, and 4% were present at the 3' untranslated region. Ten out of the 23 germline variants were classified as common variants by the 1000 Genomes Project with population frequencies of $\geq 1\%$ in the ExAC database, whereas one out of the 23 variants was classified as a rare variant with population frequency of $< 1\%$ in the

ExAC database.

We also assessed clinical significance of the germline alterations in the PGx genes using the ClinVar database. This assessment demonstrated that 21% of the 23 variants were categorized as either benign or likely benign, 17% with conflicting interpretations of pathogenicity, 9% submitted without assessment of clinical significance, and 4% with uncertain significance. There was also 17% of variants that were not reported in the ClinVar database. Although our analysis showed no variants that were pathogenic or likely pathogenic in the PGx genes, we identified seven out of the 23 germline alterations that were associated with drug response. These alterations are DPYD p.Asp949Val/c.2846A>T (rs67376798), c.1906G>A (rs3918290), p.Met166Val/c.496A>G (rs2297595), GSTP1 p.Ile105Val/c.313A>G (rs1695), MTHFR p.Glu429Ala/c.1286A>C (rs1801131), p.Ala222Val/c.665C>T (rs1801133), and TYMS c.*447_*452delTTAAAG (rs151264360), which could serve as predictors for response to chemotherapy. While the germline variants in DPYD, MTHFR, and TYMS are associated with fluoropyrimidine-related toxicities, the germline variant in GSTP1 is associated with adverse drug reactions in response to oxaliplatin treatment [].

Overall, we found an average of 5.7 variants per patient in cancer-related genes and an average of 3.7 variants per patient in PGx genes in TOP cohort. Our assessment also revealed germline alterations at 50 and 23 genomic positions in cancer-related and PGx genes, respectively. While annotation with the ClinVar database did not identified any pathogenic or likely pathogenic germline alterations, this analysis revealed a total of eight variants (one in a cancer-related gene and seven in PGx genes) that could serve as predictors for drug response. We showed that the TP53 p.Arg72Pro/c.215G>C (rs1042522) is present in 97 out of 213 patients (46%), and 208 out of 213 (98%) TOP patients have at least one germline PGx variant that is associated with drug response (Figure 4.1; Figure 4.2).

Table 4.1: Frequency of germline variants in cancer-related genes in blood specimens from TOP patients.

Gene	Chr	Pos	ID*	HGVS[*]	Zygosity wt-var[†], var-var[††]	Total	Pct[‡] (%)
ALK	2	29443662	NA	p.Val1185Val c.3555G>A	1, 0	1	0.5
EGFR	7	55242453	NA	p.Pro741Pro c.2223C>T	1, 0	1	0.5
	7	55242500	COSM133588	p.Lys757Arg c.2270A>G	2, 0	2	0.9
	7	55249063	rs1050171; COSM1451600	p.Gln787Gln c.2361G>A	96, 60	156	73
	7	5524915	rs56183713; COSM13400	p.Val819Val c.2457G>A	2, 0	2	0.9
	7	55259450	rs2229066; COSM85893; rs17290559	p.Arg836Arg c.2508C>T	9, 0	9	4
KIT	4	55592059	rs151016327; COSM3760661	p.Thr461Thr c.1383A>G	2, 0	2	0.9
	4	55599268	rs55789615; COSM1307	p.Ile798Ile c.2394C>T	14, 0	14	7
	4	55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	37, 3	40	18
MAPK1	22	22162126	rs386488966; rs3729910	p.Tyr43Tyr c.129T>C	13, 1	14	7
	22	22221623	rs201495639	p.Tyr36Tyr c.108C>T	3, 0	3	1
MTOR	1	11169420	rs41274506	p.Asp2485Asp c.7455C>T	1, 0	1	0.5
	1	11172909	NA	p.Glu2456Lys c.7366G>A	1, 0	1	0.5
	1	11174452	NA	p.Arg2408Gln c.7223G>A	1, 0	1	0.5
	1	11181327	rs11121691	p.Leu2303Leu c.6909G>A	70, 6	76	36
	1	11184593	rs56051835	p.Leu2208Leu c.6624T>C	2, 0	2	0.9

	1	11188172	rs370318222	p.Tyr1974Tyr c.5922C>T	1, 0	1	0.5
	1	11190646	rs2275527	p.Ser1851Ser c.5553C>T	65, 0	65	31
	1	11190730	rs17848553	p.Ala1823Ala c.5469C>T	8, 0	8	0.5
	1	11194521	COSM180791	c.5133C>T	1, 0	1	0.5
	1	11205058	rs386514433; rs1057079	p.Ala1577Ala c.4731A>G	81, 12	93	44
	1	11269506	NA	p.Leu1222Phe c.3664C>T	1, 0	1	0.5
	1	11272468	rs17036536	p.Arg1154Arg c.3462G>C	8, 0	8	4
	1	11288758	rs1064261	p.Asn999Asn c.2997T>C	85, 0	85	40
	1	11298038	rs55752564	p.Ala690Ala c.2070G>A	1, 0	1	0.5
	1	11298640	rs55881943	p.Ala607Ala c.1821G>A	1, 0	1	0.5
	1	11301714	rs1135172	p.Asp479Asp c.1437T>C	80, 114	194	92
	1	11308007	rs35903812	p.Ala329Thr c.985G>A	3, 0	3	1
	1	11316244	rs12120294	p.Leu170Leu c.510G>C	1, 0	1	0.5
PDGRRA	4	55141055	rs1873778; COSM1430082	p.Pro567Pro c.1701A>G	0, 183	183	86
	4	55152040	rs2228230; COSM22413	p.Val824Val c.2472C>T	57, 5	62	29
STAT1	2	191851646	rs41270237	p.Thr385Thr c.1155G>A	2, 0	2	0.9
	2	191856001	rs41509946	p.Gln330Gln c.990G>A	3, 0	3	1
	2	191859906	rs61756197	p.Gln275Gln c.825G>A	1, 0	1	0.9

	2	191859935	rs41473544	p.Val266Ile c.796G>A	2, 0	2	0.9
	2	191872307	rs45463799	p.Asn118Asn c.354C>T	3, 0	3	1
	2	191874667	rs386556119; rs2066802	p.Leu21Leu c.63T>C	42, 3	45	21
STAT3	17	40469241	COSM979464	c.2100C>T	1, 0	1	0.5
	17	40475056	rs117691970	p.Gly618Gly c.1854C>T	4, 0	4	2
	17	40486040	rs200098006	p.Leu275Leu c.825T>G	2, 0	2	0.9
	17	40486043	NA	p.Gln274Gln c.822A>G	1, 0	1	0.5
	17	40498635	rs146184566; COSM979479	p.Ser75Ser c.225G>A	1, 0	1	0.5
	17	40498713	NA	p.Lys49Lys c.147A>G	1, 0	1	0.5
	17	40498722	NA	p.Ala46Ala c.138G>T	1, 0	1	0.5
TP53	17	7577069	rs55819519; COSM44017	p.Arg290His c.869G>A	1, 0	1	0.5
	17	7577553	COSM44368	p.Met243fs c.727delA	1, 0	1	0.5
	17	7578210	rs1800372; COSM249885	p.Arg213Arg c.639A>G	1, 0	1	0.5
	17	7578420	COSM1386804	p.Thr170Thr c.510G>A	1, 0	1	0.5
	17	7579472	rs1042522; COSM250061	p.Arg72Pro c.215G>C	73,24	97	46
	17	7579579	rs1800370	p.Pro36Pro c.108G>A	5, 0	5	2
					Total variants in cancer-related genes = 1205 Average number of variants per patient = 5.7 Standard error = 0.15		

⋆dbSNP and/or COSMIC IDs.

*Description of sequence variants according to the HGVS recommendations.

†wt-var represents heterozygous variant.

††var-var represents homozygous variant.

‡Percentage of patients with the variant.

Table 4.2: Variant assessment of germline alterations in cancer-related genes detected in blood specimens of TOP patients.

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
ALK	2:29443662	NA	p.Val1185Val c.3555G>A	0.00082	Syn.	NA	NA	NA
EGFR	7:55242453	NA	p.Pro741Pro c.2223C>T	0.0074	Syn.	NA	NA	NA
	7:55242500	COSM133588	p.Lys757Arg c.2270A>G	0.00082	Missense	Uncertain significance	Homozygous mutation was identified in a patient with intrahepatic cholangiocarcinoma, leading to activation of downstream EGFR pathways as demonstrated by MAPK and Akt phosphorylations.	[82]
	7:55249063	rs1050171; COSM1451600‡	p.Gln787Gln c.2361G>A	52	Syn.	Benign/Likely benign	Conflicting evidence on predictive and prognostic values in lung cancer patients. Poorer response to anti-EGFR therapy in colorectal cancer patients compared to patients with the GG genotype.	[19, 81, 140, 152]

	7:5524915	rs56183713; COSM13400	p.Val819Val c.2457G>A	0.035	Syn.	Likely benign	One study reported that this variant in combination with rs1050171 was correlated with TNM stage of squamous cell lung carcinoma.	[140]
	7:55259450	rs2229066; COSM85893; rs17290559	p.Arg836Arg c.2508C>T	1.7	Syn.	Benign/Likely benign	NA	NA
KIT	4:55592059	rs151016327; COSM3760661	p.Thr461Thr c.1383A>G	0.28	Syn.	Benign	NA	NA
	4:55599268	rs55789615; COSM1307	p.Ile798Ile c.2394C>T	2.1	Syn.	Benign/Likely benign	NA	NA
	4:55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	12	Syn.	Benign/Likely benign	NA	NA
MAPK1	22:22162126	rs386488966; rs3729910	p.Tyr43Tyr c.129T>C	4.5	Syn.	NA	NA	NA
	22:22221623	rs201495639	p.Tyr36Tyr c.108C>T	0.052	Syn.	NA	NA	NA
MTOR	1:11169420	rs41274506	p.Asp2485Asp c.7455C>T	0.33	Syn.	NA	NA	NA

1:11172909	NA	p.Glu2456Lys c.7366G>A	0.00082	Missense	NA	NA	NA
1:11174452	NA	p.Arg2408Gln c.7223G>A	NA	Missense	NA	NA	NA
1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	22	Syn.	NA	Likely has an effect on exonic splicing enhancer or exonic splicing silencer binding site activity.	[156]
1:11184593	rs56051835	p.Leu2208Leu c.6624T>C	0.49	Syn.	Benign	NA	NA
1:11188172	rs370318222	p.Tyr1974Tyr c.5922C>T	0.00082	Syn.	NA	NA	NA
1:11190646	rs2275527	p.Ser1851Ser c.5553C>T	22	Syn.	Benign	NA	NA
1:11190730	rs17848553	p.Ala1823Ala c.5469C>T	2.4	Syn.	Benign	NA	NA
1:11194521	COSM180791	c.5133C>T	0.029	Splice region	NA	NA	NA

1:11205058	rs386514433; rs1057079‡	p.Ala1577Ala c.4731A>G	32	Syn.	NA	One study reported improved clinical response and progression-free survival in advanced esophageal squamous cell carcinoma patients with the AG genotype compared to the AA genotype who were treated with paclitaxel plus cisplatin chemotherapy.	[84]
1:11269506	NA	p.Leu1222Phe c.3664C>T	0.00082	Missense	NA	NA	NA
1:11272468	rs17036536	p.Arg1154Arg c.3462G>C	1.8	Syn.	Benign	NA	NA
1:11288758	rs1064261‡	p.Asn999Asn c.2997T>C	26	Syn.	NA	C allele likely influences exonic splicing enhancer or exonic splicing silencer binding site activity or disrupts a protein domain. Meta-analysis found no association with cancer risk.	[156]
1:11298038	rs55752564	p.Ala690Ala c.2070G>A	0.077	Syn.	NA	NA	NA

	1:11298640	rs55881943	p.Ala607Ala c.1821G>A	0.017	Syn.	Conflicting interpretations of pathogenicity	NA	NA
	1:11301714	rs1135172‡	p.Asp479Asp c.1437T>C	72	Syn.	NA	NA	NA
	1:11308007	rs35903812	p.Ala329Thr c.985G>A	0.27	Missense	Likely benign	NA	NA
	1:11316244	rs12120294	p.Leu170Leu c.510G>C	0.36	Syn.	NA	NA	NA
PDGFRA	4:55141055	rs1873778; COSM1430082‡	p.Pro567Pro c.1701A>G	99	Syn.	Benign	No association with PDGFRα expression in colorectal cancer.	[49]
	4:55152040	rs2228230; COSM22413	p.Val824Val c.2472C>T	18	Syn.	Benign	NA	NA
STAT1	2:191851646	rs41270237	p.Thr385Thr c.1155G>A	0.42	Syn.	Likely benign	NA	NA
	2:191856001	rs41509946	p.Gln330Gln c.990G>A	0.36	Syn.	Likely benign	NA	NA

	2:191859906	rs61756197	p.Gln275Gln c.825G>A	0.025	Syn.	NA	NA	NA
	2:191859935	rs41473544	p.Val266Ile c.796G>A	0.20	Missense	Likely benign	Functional testing indicated that the variant was not a gain-of-function mutation in STAT1	[39]
	2:191872307	rs45463799	p.Asn118Asn c.354C>T	0.32	Syn.	Likely benign	NA	NA
	2:191874667	rs386556119; rs2066802	p.Leu21Leu c.63T>C	8.5	Syn.	Benign	High frequency among patients with multiple sclerosis and chronic hepatitis C.	[52]
STAT3	17:40469241	COSM979464	c.2100C>T	NA	Splice region	NA	NA	NA
	17:40475056	rs117691970	p.Gly618Gly c.1854C>T	0.37	Syn.	Likely benign	NA	NA
	17:40486040	rs200098006	p.Leu275Leu c.825T>G	0.066	Syn.	NA	NA	NA
	17:40486043	NA	p.Gln274Gln c.822A>G	0.00082	Syn.	NA	NA	NA

	17:40498635	rs146184566; COSM979479	p.Ser75Ser c.225G>A	0.029	Syn.	Likely benign	NA	NA
	17:40498713	NA	p.Lys49Lys c.147A>G	0.012	Syn.	NA	NA	NA
	17:40498722	NA	p.Ala46Ala c.138G>T	NA	Syn.	NA	NA	NA
TP53	17:7577069	rs55819519; COSM44017	p.Arg290His c.869G>A	0.016	Missense	Conflicting interpretations of pathogenicity	A conservative amino acid substitution that was predicted to be possibly damaging by *in silico* analysis. Reported in patients with Li-Fraumeni syndrome and cancer patients without family histories of Li-Fraumeni syndrome or Li-Fraumeni-like syndrome.	[7, 8, 32, 103, 105, 135]
	17:7577553	COSM44368	p.Met243fs c.727delA	NA	Frameshift	NA	Reported in esophageal squamous cell carcinoma of patients from northern Iran.	[12]

17:7578210	rs1800372; COSM249885	p.Arg213Arg c.639A>G	1.2	Syn.	Benign/Likely benign	One study demonstrated that this variant was not a predictive biomarker for initiation and progression of gastroesophageal reflux disease, Barrett's Esophagus, and esophageal cancer in the Brazilian population.	[104]
17:7578420	COSM1386804	p.Thr170Thr c.510G>A	0.012	Syn.	NA	One study reported that TP53 mutations in exon 5, which include this variant, were associated with the worst prognosis for patients with non-small-cell lung cancer.	[134]

17:7579472	rs1042522; COSM250061‡	p.Arg72Pro c.215G>C	34	Missense	Drug response	p53 protein with Arg72 was associated with increased apoptosis, while p53 protein with Pro72 demonstrated increased G_1 cell-cycle arrest and activation of p53-dependent DNA repair. Pro/Pro genotype resulted in severe neutropenia in ovarian cancer patients receiving cisplatin-based chemotherapy, and poor survival and treatment response in gastric cancer patients receiving paclitaxel and capecitabine combination chemotherapy, as well as 5-fluorouracil-based adjuvant chemotherapy. Conflicting evidence on risk of predispositon to various cancer types.	[16, 18, 21, 31, 68, 73, 75, 148, 150, 151, 154, 155]
17:7579579	rs1800370	p.Pro36Pro c.108G>A	1.3	Syn.	Benign/Likely benign	NA	NA

⋆dbSNP and/or COSMIC IDs.

*Description of sequence variants according to the Human Genome Variation Society (HGVS) recommendations.

**AF = Allele frequency reported by the Exome Aggregation Consortium (ExAC) and presented in percentage.

†Effect of genetic variants as predicted by the SnpEff software.

††Clinical significance on ClinVar database.

‡Human reference genome hg19 contains the minor allele. If the minor allele is associated with functional and/or clinical impacts reported in the literature, this will be indicated in the functional/clinical impacts column.

Table 4.3: Frequency of germline variants in pharmacogenomic genes detected in blood specimens of TOP patients.

Gene	Chr	Pos	dbSNP ID	HGVS*	Zygosity wt-var†, var-var††	Total	Pct‡ (%)
DPYD	1	97547947	rs67376798	p.Asp949Val c.2846A>T	2, 0	2	0.9
	1	97770920	rs1801160	p.Val732Ile c.2194G>A	24, 0	24	11
	1	97915614	rs3918290	c.1906G>A	1, 0	1	0.5
	1	97915615	rs3918289	c.1905C>T	1, 0	1	0.5
	1	97981421	rs1801158	p.Ser534Asn c.1601G>A	3, 0	3	2
	1	98039419	rs56038477	p.Glu412Glu c.1236G>A	7, 0	7	3
	1	98165091	rs2297595	p.Met166Val c.496A>G	34, 0	34	16
	1	98348885	rs1801265	p.Cys29Arg c.85T>C	69, 11	80	37
GSTP1	11	67352689	rs1695	p.Ile105Val c.313A>G	89, 20	109	51
MTHFR	1	11854476	rs1801131	p.Glu429Ala c.1286A>C	86, 16	102	47
	1	11856378	rs1801133	p.Ala222Val c.665C>T	90, 20	110	51
TYMP	22	50964236	rs11479	p.Ser471Leu c.1412C>T	51, 6	57	27
	22	50964255	rs112723255	p.Ala465Thr c.1393G>A	16, 1	17	8
	22	50964493	NA	p.Glu413Lys c.1237G>A	1, 0	1	0.5
	22	50964907	rs201685922	c.929_932delCCGC	1, 0	1	0.5
	22	50965102	rs8141558	p.Leu277Leu c.831G>A	1, 0	1	0.5
	22	50965597	rs373478014	p.Thr254Thr c.762G>A	1, 0	1	0.5
	22	50965624	rs139223629	p.Gln245Gln c.735G>A	1, 0	1	0.5

	22	50965683	rs200497106	p.Gly226Arg c.676G>A	1, 0	1	0.5
	22	50966082	NA	p.Ala194Val c.581C>T	1, 0	1	0.5
TYMS	22	673443	rs151264360	c.*447_*452delTTAAAG	89, 43	132	62
UGT1A1	2	234668870	rs873478	c.-64G>C	1, 0	1	0.5
	2	234668879	rs34983651	c.-55_-54insAT	81, 17	98	46
					Total variants in PGx genes = 785 Average number of variants per patient = 3.7 Standard error = 0.10		

*Description of sequence variants according to the HGVS recommendations.

†wt-var represents heterozygous variant.

††var-var represents homozygous variant.

‡Percentage of patients with the variant.

Table 4.4: Variant assessment of germline alterations in pharmacogenomic genes detected in blood specimens of TOP patients.

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
DPYD	1:97547947	rs67376798	p.Asp949Val c.2846A>T	0.26	Missense	Drug response	Close to iron sulfur motif, which could interfere with electron transport or cofactor binding. Reduced DPD activity with strong clinical evidence indicating association with severe fluoropyrimidine-related toxicity.	[4, 15, 27, 38, 44, 80, 90, 94, 96, 99, 114, 127, 131–133]
	1:97770920	rs1801160	p.Val732Ile c.2194G>A	4.6	Missense	Benign/Likely benign, not provided	Reduced DPD activity and associated with severe fluoropyrimidine-related toxicity.	[15, 38, 56, 114, 132, 133]

1:97915614	rs3918290	c.1906G>A	0.52	Splice donor	Drug response	Exon 14 is skipped, producing an inactive enzyme with no uracil-binding site. Reduced DPD activity with strong clinical evidence indicating association with severe fluoropyrimidine-related toxicity.	[4, 27, 38, 56, 80, 94, 96, 114, 127, 131–133]
1:97915615	rs3918289	c.1905C>T	0.030	Splice region	Not provided	Benign variant as predicted by PolyPhen-2, a functional prediction software. No association with fluoropyrimidine-related toxicity.	[15, 99]
1:97981421	rs1801158	p.Ser534Asn c.1601G>A	1.4	Missense	Conflicting interpretations of pathogenicity, not provided	Conflicting evidence on changes to DPD activity. Conflicting clinical evidence on association with fluoropyrimidine-related toxicity.	[94, 99, 114, 131, 133]

1:98039419	rs56038477	p.Glu412Glu c.1236G>A	1.5	Syn.	Benign	Synonymous variant in high linkage disequilibrium with c.1129-5923C>G (rs75017182) in haplotype B3 (HapB3). rs75017182 causes nonsense mutation in exon 11, resulting in reduced DPD activity. Associated with fluoropyrimidine-related toxicity.	[4, 38, 94, 97]
1:98165091	rs2297595	p.Met166Val c.496A>G	8.6	Missense	Drug response	Conflicting evidence on changes to DPD activity. Associated with fluoropyrimidine-related toxicity.	[38, 56, 99, 127, 132, 133]
1:98348885	rs1801265[‡]	p.Cys29Arg c.85T>C	23	Missense	Not provided	C allele causes reduced DPD activity. Conflicting clinical evidence on association with fluoropyrimidine-related toxicity.	[27, 56, 96, 128, 133]

GSTP1	11:67352689	rs1695	p.Ile105Val c.313A>G	33	Missense	Drug response	Disrupts the enzyme's electrophile-binding active site, thereby lowering catalytic efficiency. Increased risk of oxaliplatin-related toxicity and efficacy of oxaliplatin treatment.	[3, 30, 66, 92, 109, 124]
MTHFR	1:11854476	rs1801131	p.Glu429Ala c.1286A>C	30	Missense	Drug response	Reduced MTHFR activity with conflicting evidence on efficacy of treatment with fluoropyrimidines.	[50, 51, 70, 89, 109]
	1:11856378	rs1801133	p.Ala222Val c.665C>T	30	Missense	Drug response	Reduced MTHFR activity, resulting in stronger inhibition of DNA synthesis. Increased effectiveness of fluoropyrimidine treatment, although conflicting clinical evidence exists. Conflicting evidence on fluoropyrimidine-related toxicity.	[34, 50, 51, 62, 70, 89, 109, 114, 126]

TYMP	22:50964236	rs11479	p.Ser471Leu c.1412C>T	12	Missense	Benign/Likely benign	High expression in tumour cells, correlated with poor overall survival in the presence of high platelet counts. Limited clinical evidence suggesting association with adverse reactions from fluoropyrimidine treatment.	[24, 67, 71]
	22:50964255	rs112723255	p.Ala465Thr c.1393G>A	4.4	Missense	Benign/Likely benign	No association with fluoropyrimidine-related toxicity. Increased risk of transplant-related toxicity from HLA-matched sibling allogeneic stem cell transplantation. Increased risk of chronic graft-versus-host disease when donor is a carrier of the minor allele and recipient is homozygous for the major allele.	[61, 71, 120]
	22:50964493	NA	p.Glu413Lys c.1237G>A	NA	Missense	NA	NA	NA

22:50964907	rs201685922	c.929_932delCCGC	0.49	Splice region	Conflicting interpretations of pathogenicity	Observed in a German American patient with mitochondrial neurogastrointestinal encephalomyopathy (MNGIE), but relation with TP enzymatic defect was not established.	[98]
22:50965102	rs8141558	p.Leu277Leu c.831G>A	0.58	Syn.	Benign/Likely benign	NA	NA
22:50965597	rs373478014	p.Thr254Thr c.762G>A	0.0016	Syn.	NA	NA	NA
22:50965624	rs139223629	p.Gln245Gln c.735G>A	0.26	Syn.	Conflicting interpretations of pathogenicity	NA	NA
22:50965683	rs200497106	p.Gly226Arg c.676G>A	0.0091	Missense	Uncertain significance	NA	NA
22:50966082	NA	p.Ala194Val c.581C>T	NA	Missense	NA	NA	NA

TYMS	22:673443	rs151264360	c.*447_*452delTTAAAG	48‡‡	3' UTR	Drug response	Decreased stability of secondary mRNA structure and lower TS expression. Conflicting evidence on survival, response to fluoropyrimidine treatment, and risk of fluoropyrimidine-related toxicity.	[2, 45, 58, 62, 87, 124]
UGT1A1	2:234668870	rs873478	c.-64G>C	1.1‡‡	Upstream gene	NA	Unknown	[28, 149, 153]
	2:234668879	rs34983651	c.-55_-54insAT	33‡‡	Upstream gene	Conflicting interpretations of pathogenicity, affects, association	Lower UGT1A1 expression and associated with irinotecan-related toxicity.	[5, 37, 57, 69, 78, 88, 92, 107, 110, 130]

*Description of sequence variants according to the Human Genome Variation Society (HGVS) recommendations.

**AF = Allele frequency reported by the Exome Aggregation Consortium (ExAC) and presented in percentage.

†Effect of genetic variants as predicted by the SnpEff software.

††Clinical significance on ClinVar database.

‡Human reference genome hg19 contains the minor allele. If the minor allele is associated with functional and/or clinical impacts reported in the literature, this will be indicated in the functional/clinical impacts column.

‡‡Allele frequency from the 1000 Genomes Project is reported when the allele frequency is unavailable in the ExAC database.

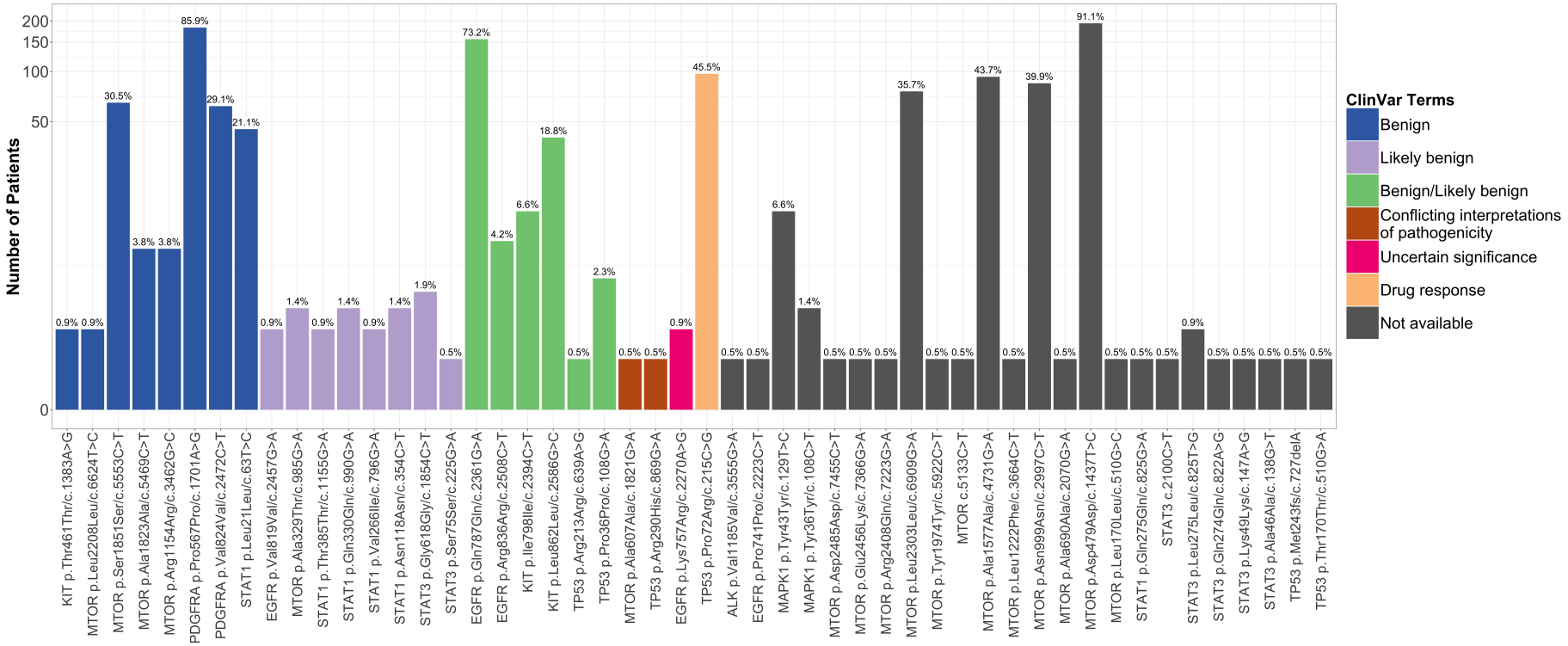


Figure 4.1: Distribution of germline alterations in cancer-related genes in patients from TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. The TP53 variant, p.Arg72Pro/c.215G>C, that is associated with drug response is present in 97 out of 213 (45.5 %) patients in TOP cohort. log(1 + x) transformation is applied to change the scale of set values on the Y-axis.

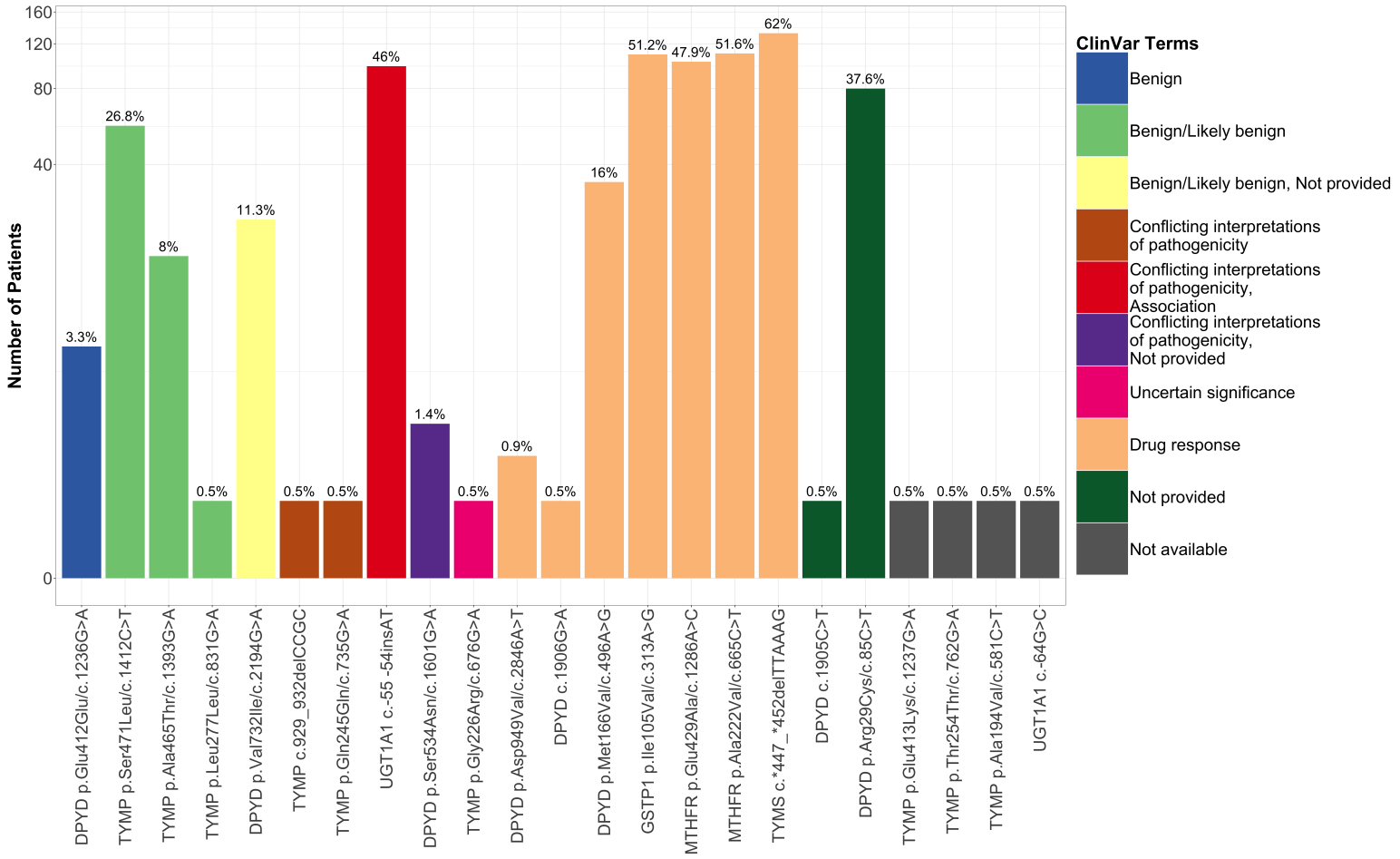


Figure 4.2: Distribution of germline alterations in PGx genes in patients from TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. 208 out of 213 patients in TOP cohort have at least one germline PGx variant that is associated with drug response. log(1 + x) transformation is applied to change the scale of set values on the Y-axis.

4.2 Germline alterations are highly concordant between blood and FFPE specimens

The tumour genome consists of germline and somatic alterations. In fact, several studies demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [72, 93, 106, 113]. While we were unable to detect any pathogenic or likely pathogenic germline variants due to the rarity of these variants and the small cohort size of TOP study, we were still able to identified eight germline alterations that could serve as predictors for drug response, in addition to other germline alterations. Because paired tumour-blood samples were collected for patients in TOP cohort, we sought to determine variant concordance of germline alterations between tumour and blood specimens. This analysis would reveal the extent to which germline alterations can be detected in DNA isolated from tumours.

Because there are four tumour specimens in TOP cohort with duplicates, we examined a total of 217 tumour-normal paired samples. A total of 4434 variants were identified, in which 4003 variants were germline and 431 variants were somatic. Out of the 4003 germline variants, 3792 variants were concordant between tumour and blood specimens, whereas 211 variants were discordant between specimen types (Figure 4.3). Thus, the concordance rate for the 217 tumour-normal paired samples was 93.8%. Out of the 211 discordant germline alterations, 166 (3.7%) demonstrated loss of heterozygosity in the tumours, 34 (0.77%) were heterozygous in the blood specimens but wild type in the tumours, 7 (0.16%) have low sequencing depth ($<$ 100x) in the tumours, and 4 (0.090%) were called as homozygous in the blood specimens but heterozygous in the tumours (Table 4.5).

Multiple factors could contribute to the discordant calls including position of the variant within regions of somatic copy number mutations, genomic rearrangements due to the presence of intragenic fragile sites, and DNA damage caused by formalin fixation []. Nevertheless, despite the presence of discordant germline alterations, our analysis revealed that the majority of germline alterations identified in the blood could be detected in tumour specimens with correct designation of zygosity.

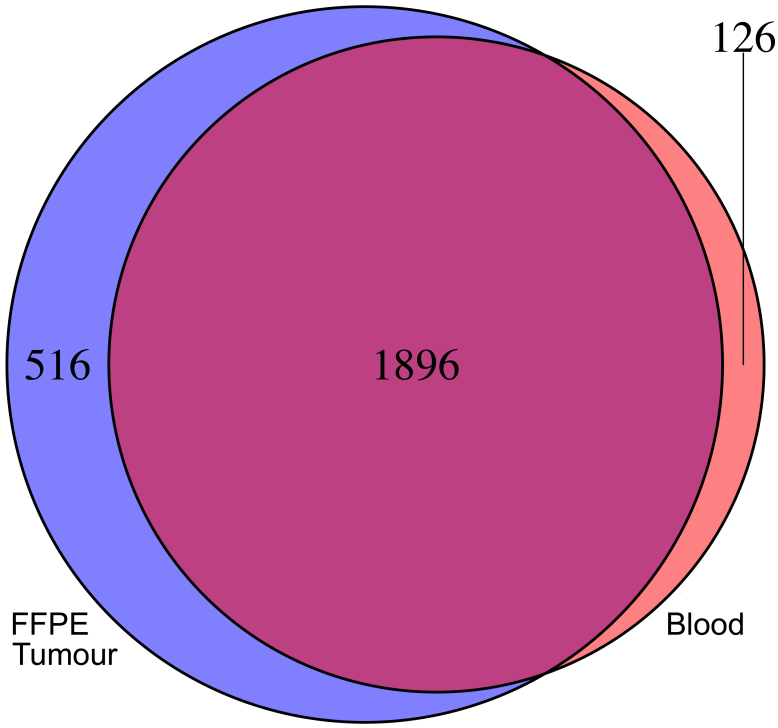


Figure 4.3: Venn diagram demonstrating concordance of variants identified in 217 tumour-blood paired samples.

Table 4.5: Distribution of discordant germline alterations identified in patients from TOP cohort.

Gene	Chr:Pos	ID*	HGVS*	Clinical Significance†	Reason for discordance	Count
DPYD	1:97547947	rs67376798	p.Asp949Val c.2846A>T	Drug response	Het/WT	1
	1:97770920	rs1801160	p.Val732Ile c.2194G>A	Benign/Likely benign, Not provided	Het/Hom	2
	1:98165091	rs2297595	p.Met166Val c.496A>G	Drug response	Het/Hom	2
	1:98348885	rs1801265	p.Cys29Arg c.85T>C	Not provided	Low coverage in tumour	2
	1:98348885	rs1801265	p.Cys29Arg c.85T>C	Not provided	Het/WT	2
	1:98348885	rs1801265	p.Cys29Arg c.85T>C	Not provided	Het/Hom	6
EGFR	7:55249063	rs1050171; COSM1451600	p.Gln787Gln c.2361G>A	Benign/Likely benign	Het/Hom	2
GSTP1	11:67352689	rs1695	p.Ile105Val c.313A>G	Drug response	Het/WT	3
	11:67352689	rs1695	p.Ile105Val c.313A>G	Drug response	Het/Hom	14
KIT	4:55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	Benign/Likely benign	Het/Hom	8
MTHFR	1:11854476	rs1801131	p.Glu429Ala c.1286A>C	Drug response	Het/Hom	12

	1:11856378	rs1801133	p.Ala222Val c.665C>T	Drug response	Het/Hom	12
	1:11856378	rs1801133	p.Ala222Val c.665C>T	Drug response	Het/WT	3
MTOR	1:11169420	rs41274506	p.Asp2485Asp c.7455C>T	NA	Het/WT	1
	1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	NA	Het/Hom	2
	1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	NA	Low coverage in tumour	1
	1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	NA	Het/WT	2
	1:11190646	rs2275527	p.Ser1851Ser c.5553C>T	Benign	Het/WT	1
	1:11190730	rs17848553	p.Ala1823Ala c.5469C>T	Benign	Het/Hom	4
	1:11205058	rs1057079; rs386514433	p.Ala1577Ala c.4731A>G	NA	Het/Hom	8
	1:11205058	rs1057079; rs386514433	p.Ala1577Ala c.4731A>G	NA	Het/WT	4
	1:1272468	rs17036536	p.Arg1154Arg c.3462G>C	Benign	Het/Hom	4
	1:11288758	rs1064261	p.Asn999Asn c.2997T>C	NA	Het/Hom	4
	1:11288758	rs1064261	p.Asn999Asn c.2997T>C	NA	Het/WT	3

	1:11301714	rs1135172	p.Asp479Asp c.1437T>C	NA	Low coverage in tumour	1
	1:11301714	rs1135172	p.Asp479Asp c.1437T>C	NA	Het/Hom	8
PDGFRA	4:55141055	rs1873778; COSM1430082	p.Pro567Pro c.1701A>G	Benign	Low coverage in tumour	3
	4:55152040	rs2228230; COSM22413	p.Val824Val c.2472C>T	Benign	Het/WT	2
	4:55152040	rs2228230; COSM22413	p.Val824Val c.2472C>T	Benign	Het/Hom	4
STAT1	2:191872307	rs45463799	p.Asn118Asn c.354C>T	Likely benign	Het/WT	1
	2:191874667	rs386556119; rs2066802	p.Leu21Leu c.63T>C	Benign	Het/WT	1
STAT3	17:40498713	NA	p.Lys49Lys c.147A>G	NA	Het/WT	1
TP53	17:7577553	COSM44368	p.Met243fs c.727delA	NA	Het/WT	1
	17:7579472	COSM250061; rs1042522	p.Arg72Pro c.215G>C	Drug response	Het/Hom	26
	17:7579472	COSM250061; rs1042522	p.Arg72Pro c.215G>C	Drug response	Het/WT	4
	17:7579579	rs1800370	p.Pro36Pro c.108G>A	Benign/Likely benign	Het/Hom	2
TYMP	22:50964236	rs11479	p.Ser471Leu c.1412C>T	Benign/Likely benign	Het/Hom	14

TYMS	18:673443	rs151264360	c.*447_*452delTTAAAG	Drug response	Het/Hom	32
	18:673443	rs151264360	c.*447_*452delTTAAAG	Drug response	Het/WT	1
UGT1A1	2:234668870	rs873478	c.-64G>C	NA	Het/WT	1
	2:234668879	rs34983651	c.-55_-54insAT	Conflicting interpretations of pathogenicity, Association	Hom/Het	4
	2:234668879	rs34983651	c.-55_-54insAT	Conflicting interpretations of pathogenicity, Association	Hom/WT	2
					Total discordant variants = 211	

⋆dbSNP and/or COSMIC IDs.

*Description of sequence variants according to the HGVS recommendations.

†Clinical significance on ClinVar database.

Het/Hom = Loss of heterozygosity in the tumour

Het/WT = Heterozygous in the blood, but wild type in the tumour

Hom/Het = Homozygous in the blood, but heterozygous in the tumour

4.3 Application of tumour content to separate germline alterations from somatic mutations in tumour-only analyses

Through variant analysis of DNA from blood specimens, we identified germline alterations that are associated with drug response, which could predict risk of developing chemotherapy-induced toxicity. Furthermore, we assessed the concordance of germline variants between blood and tumour samples, which demonstrated a high concordance rate of 93.8%. Together, these analyses confirmed that germline alterations that are clinically relevant are present in our dataset and a large proportion of germline alterations can be identified in tumour DNA with the correct designation of allelic statuses. Next, we sought to evaluate the use of VAF thresholds to separate germline alterations from somatic mutations in tumour-only analyses. Because of the lack of availability of matched normal samples in clinical genomic sequencing, this assessment would determine whether application of VAF thresholds is an accurate method to identify potential germline alterations in clinical tumour sequencing for referral to follow-up testing. While our dataset does not contain pathogenic germline variants, we anticipate that this approach can be used to detect genetic events associated with cancer predisposition for future patients.

We compared the VAF distributions of germline variants detected in blood and tumour specimens, and we found a significant difference (Kolmogorov-Smirnov test, D = 0.17, $p = 0$; Figure 4.4). As expected, we showed that heterozygous alterations in blood tend to have VAFs close to 50%, whereas homozygous alterations in the blood tend to have VAFs close to 100%. However, the VAF distribution of germline variants in the tumours tend to deviate from 50% and 100% for heterozygous and homozygous statuses, respectively. This variation in VAF distributions between blood and tumour samples, which could be caused by tumour content, tumour heterogeneity, or DNA damage as a result of formalin fixation, indicates that the sensitivity of using a VAF cut-off to distinguish between germline and somatic alterations in tumour-only analyses could be compromised. Thus, we explored the sensitivity of identifying germline alterations at various VAF thresholds to select a VAF cut-off that maximizes true positive rate. At each VAF cut-off, we determined the number of true positives by identifying variants in the tumours that overlap with germline variants in matched blood samples. True positive rate (sensitivity) is then calculated as the fraction of variants that are correctly identified as germline using the VAF threshold over the total number of germline variants in the tumours. At a VAF cut-off of 30%, we achieved a sensitivity of 0.94 (95% CI = 0.93–0.95; Figure 4.4; Table 4.6), resulting in 1864 true positives and 117 false negatives out of a total of 1981 calls.

Because clinical genomics require accurate identification of genetic alterations that are clinically important, potential germline alterations identified through tumour-only analyses must be referred to follow-up testing [17, 59, 106]. Hence, not only must our approach for discriminating between germline and somatic alterations be highly sensitive, but also highly precise to minimize submission of somatic mutations (false positives) for downstream germline testing, which could in-

cur additional cost and time. For similar reasons that cause VAFs of germline alterations in tumour samples to differ from germline alterations in the blood, we presumed VAFs of somatic mutations to be lower. We assessed this variation in VAF distributions between germline and somatic alterations in the tumours and found a significant difference (Kolmogorov-Smirnov test, D = 0.52, p = 0; Figure 4.5). Indeed, VAFs of somatic mutations tend to be concentrated at lower percentages compared to VAFs of germline variants. To select a VAF cut-off that would achieve high precision, we measured positive predictive values at various VAF thresholds. At each VAF cut-off, we identified true germline alterations by overlapping the variants in the tumours with germline variants called in matched blood samples. Positive predictive value is then calculated as the fraction of true positives over total number of variants identified in the tumours, including somatic mutations (false positives). At a VAF cut-off of 30%, we achieved a positive predictive value of 0.90 (95% CT = 0.89–0.91; Figure 4.5; Table 4.7), resulting in 1864 true positives and 203 false positives out of a total of 2067 calls.

Despite the difference in VAF distributions between germline alterations in blood and tumour samples, we managed to apply a VAF cut-off of 30% to obtain a sensitivity of 0.94. This also means that this cut-off would result in a miss rate of 0.059 (95% CI = 0.049–0.07), in which approximately 6% of true germline variants will be missed. Moreover, we were also able to leverage the difference in VAFs of germline and somatic variants to distinguish germline variants from somatic mutations in tumour-only analyses. At the 30% VAF cut-off, we were not only able to achieve sensitivity of 0.94, but also a positive predictive value of 0.90, meaning that close to 10% of calls identified using this approach are somatic mutations. Overall, we demonstrated that the use of VAF thresholds to identify potential germline alterations in clinical tumour sequencing is a promising approach towards mitigating challenges caused by the lack of matched normal samples and funding in the clinical setting.

need to address PGx variants

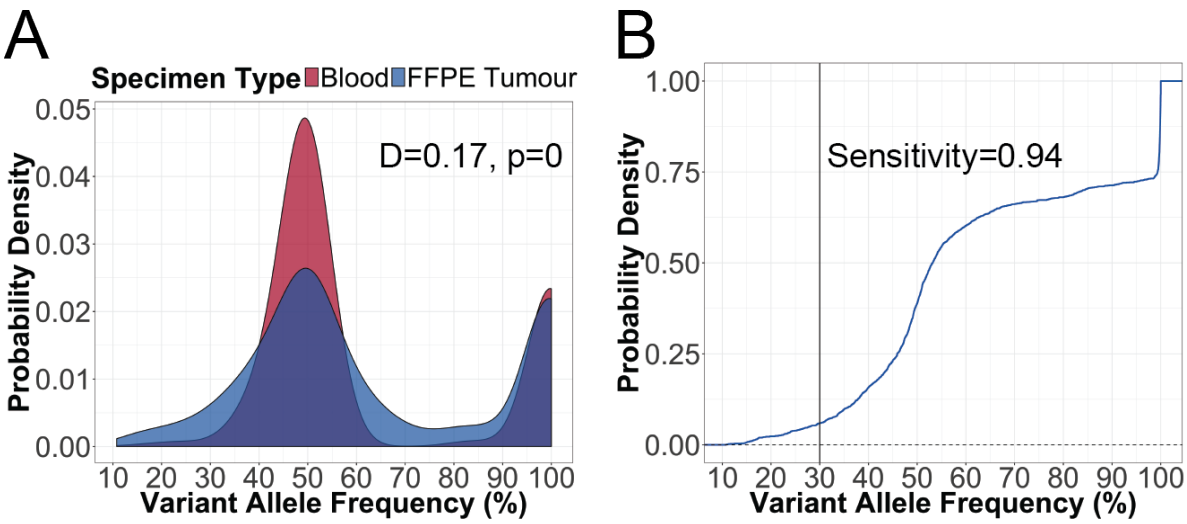


Figure 4.4: Assessment of using a VAF cut-off approach to identify germline alterations in tumour-only analyses. (A) Comparison of VAF distributions of germline alterations between blood and tumour (Kolmogorov-Smirnov test). (B) Empirical cumulative distribution of VAFs of germline alterations in tumour samples. Black line indicates VAF cut-off at 30%, in which sensitivity of identifying germline variants is 0.94.

Table 4.6: Sensitivity of identifying germline variants in tumour-only analyses at various variant allele frequency thresholds. 95% confidence interval is the binomial confidence interval calculated using the Clopper-Pearson method.

VAF (%)	False Negative	True Positive	Sensitivity	95% CI
10	0	1981	1.0	1.0–1.0
15	13	1968	0.99	0.99-1.0
20	46	1935	0.98	0.97–0.98
25	77	1904	0.96	0.95–0.97
30	117	1864	0.94	0.93–0.95
35	192	1789	0.90	0.89–0.92
40	313	1668	0.84	0.83–0.86
45	458	1523	0.77	0.75–0.79

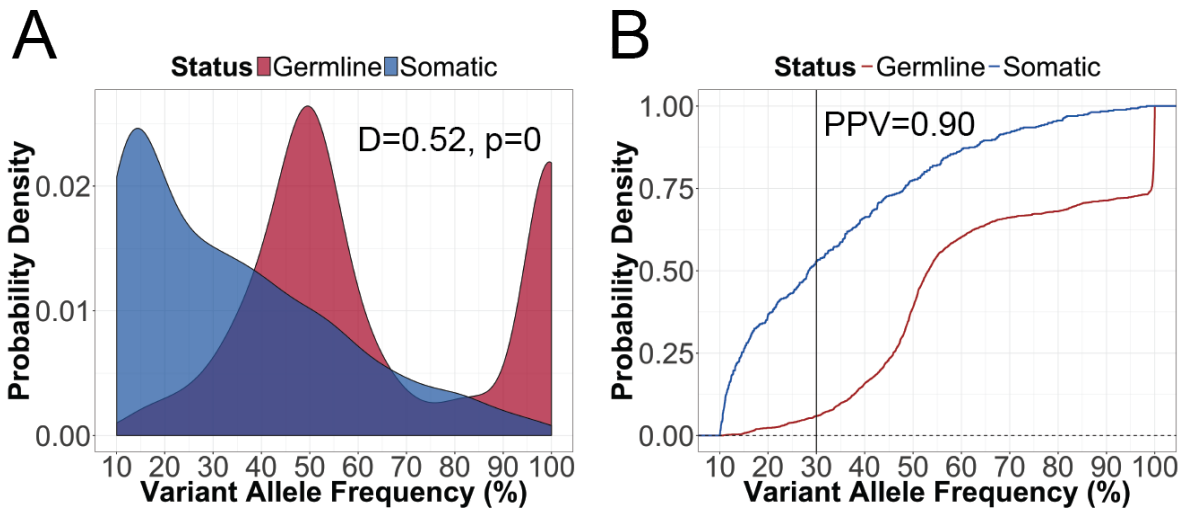


Figure 4.5: Assessment of using a VAF cut-off approach to refer potential germline alterations in tumour-only analyses to follow-up testing. (A) Comparison of VAF distributions between germline and somatic alterations in tumour specimens (Kolmogorov-Smirnov test). (B) Empirical cumulative distribution of VAFs of germline and somatic alterations in tumour samples. Black line indicates VAF cut-off at 30%, in which positive predictive value of referring potential germline variants to follow-up testing is 0.90.

Table 4.7: Positive predictive values for referral of potential germline variants to downstream confirmatory testing at various variant allele frequency thresholds. 95% confidence interval is the binomial confidence interval calculated using the Clopper-Pearson method.

VAF (%)	False Positive	True Positive	Total Calls	Positive Predictive Value	95% CI
10	431	1981	2412	0.82	0.81–0.84
15	319	1968	2287	0.86	0.85–0.87
20	273	1935	2208	0.88	0.86–0.89
25	245	1904	2149	0.89	0.87–0.90
30	203	1864	2067	0.90	0.89–0.91
35	178	1789	1967	0.91	0.90–0.92
40	146	1668	1814	0.92	0.91–0.93
45	118	1523	1641	0.93	0.91–0.94

Chapter 5

Discussion

Genomic analyses of tumours can reveal druggable somatic mutations, as well as clinically relevant germline alterations that are beneficial to patients and their family members [72, 93, 113]. While sequencing of tumour-normal pairs can enable differentiation between germline and somatic variants, matched normal samples are often not obtained in clinical practice. Moreover, FFPE tumour tissues represent another challenge in clinical genomics. Formalin fixation damages nucleic acid through fragmentation and cytosine deamination, which affect molecular testing with FFPE DNA [42, 76, 100, 101, 118, 144, 145]. Hence, usability of FFPE DNA for germline testing and approaches to discriminate between germline and somatic variants in tumour-only analyses must be evaluated. These assessments would facilitate optimization of workflows to identify potential germline alterations using clinical tumour sequencing.

In this study, we retrospectively analyzed targeted sequencing data from tumour and matched blood specimens of 213 cancer patients. Our findings demonstrated that DNA fragmentation and cytosine deamination were common forms of DNA damage in FFPE specimens. While the impact of formalin fixation on amplicon enrichment and sequencing results was detectable, we determined that these discrepancies were either technically negligible or could be minimized using appropriate methods. We also found that the majority of germline alterations identified in blood using our panel test were present with the same allelic statuses in FFPE tumours. This implies that a high proportion of germline genetic changes is retained in the tumour genome, demonstrating the reliability of using tumour DNA for germline variant calling. Finally, we assessed the application of VAF threshold to delineate germline and somatic variants in tumour-only analyses. We reported that a VAF cut-off of 30% would correctly identify 94% of germline alterations, while erroneously submit 10% of false positives, which are somatic mutations, for follow-up germline testing. Because our gene panel and patient cohort are relatively small, we were only able to identify germline variants that are predictive of drug response. However, we surmised that application of this VAF cut-off could be expanded to predict the statuses of pathogenic germline variants such as alterations in *BRCA* genes.

Several studies have reported findings that are consistent with our assessment of formalin-induced DNA damage in FFPE specimens. To assess the usability of FFPE DNA for germline testing, we compared efficiency in amplicon enrichment and sequencing results of FFPE DNA to blood, which is a gold standard for germline testing. We noted lower efficiency in amplicon enrichment in FFPE DNA, with a more pronounced decrease in coverage depth for longer amplicons in the panel. Similarly, Shi et al. [117], Didelot et al. [40], and Wong et al. [144] demonstrated that shorter amplicons gave rise to better PCR amplification success in FFPE DNA, indicating the presence of fragmentation damage, which yields template DNA of shorter fragment lengths. While we observed comparable proportion of on-target aligned reads between FFPE and blood DNA, there were minor discrepancies in coverage depth and uniformity of target bases in FFPE DNA. Various groups have also reported disparities in coverage depth and uniformity in FFPE DNA when compared to DNA extracted from either fresh frozen or unfixed specimens [11, 123, 144]. Additionally, Wong et al. [145] and Didelot et al [40] showed inverse correlations between coverage depth and the degree of DNA fragmentation in FFPE DNA, suggesting that formalin-induced fragmentation damage could be accountable for such discrepancies in sequencing results. Although we detected differences in sequencing results between FFPE and blood DNA, we concluded that these effects were minor and technically insignificant. As for the discrepancy in amplicon enrichment, shorter amplicons can be designed to circumvent the drawback of fragmentation damage in FFPE samples.

Cytosine deamination is a major cause of sequence artifacts in formalin-fixed specimens [29, 41, 43, 76, 101, 123, 145]. Herein, we observed increased C>T/G>A artifacts in FFPE DNA compared to blood. Artifactual C>T/G>A changes are formed by incorporation of adenines in the complementary DNA strand at uracil lesions generated by deamination of cytosines [42]. When measuring frequency of sequence artifacts at different allele frequency ranges, Wong et al. [145] reported higher C>T/G>A transitions at a lower allele frequency range (1–10% *vs.* 10–25%). This finding led us to compare the fraction of base changes at different allele frequency ranges, including 1–10%, 10–20%, and 20–30%. Indeed, we observed a substantial increase in C>T/G>A within the 1–10% allele frequency range. Considering that our goal is to predict germline status, disproportionate base changes between FFPE and blood DNA within these allele frequency ranges suggest that germline calls should be made at $> 30\%$ VAF to avoid false positives that could either arise from true somatic mutations or FFPE artifacts. We were unable to separate FFPE artifacts from low-allelic-fraction somatic mutations within these allele frequency ranges due to the lack of matched fresh frozen or unfixed tumour tissues. Nevertheless, somatic mutations can occur at VAFs that deviate significantly from a diploid zygosity (i.e. heterozygous variant should have VAF close to 50%, whereas homozygous variant should have VAF close to 100%) because of low tumour content or tumour heterogeneity [23, 26, 74, 129, 146]. Therefore, further workflow optimization should be performed for the purpose of identifying clinically relevant somatic mutations in the tumour genome. A method to reduce sequence artifacts caused by cytosine deamination is through treatment with uracil-DNA

glycosylase (UDG) before sequencing. UDG is an enzyme capable of depleting uracil lesions in DNA, giving rise to abasic sites. During PCR amplification, cytosine bases are restored at abasic sites by using the complementary DNA strand as template, which consists of guanine bases opposite of the uracil lesions [42]. Several studies showed that pre-treatment of FFPE DNA with UDG can markedly reduce C>T/G>A sequence artifacts [41, 43, 76]. However, this approach cannot correct sequence artifacts at CpG dinucleotides because these cytosines are typically methylated, and deamination of 5-methyl cytosines generates thymines instead uracil bases, which are resistant to UDG repair [43].

We also observed elevated levels of A>G/T>C artifacts in FFPE DNA, albeit to a lesser extent compared to C>T/G>A artifacts. Likewise, Wong et al. [142] reported that 35% of sequence artifacts in Sanger sequencing of the *BRCA1* gene were A>G/T>C nucleotide changes. We speculate that increase in A>G/T>C artifacts is caused by deamination of adenine to generate hypoxanthine, which forms base pairs with cytosine instead of thymine. This results in transformation of A-T base pairs to G-C base pairs. Deamination of adenine to hypoxanthine can be catalyzed by an acidic environment [139], which can arise in FFPE specimens because formaldehyde can be oxidized to generate formic acid [42]. Acidic conditions also promotes depurination, creating abasic sites. Many DNA polymerases selectively incorporate adenines across abasic sites, while guanines and small deletions are integrated in fewer cases [63]. Despite statistically insignificant, we observed a subset of FFPE specimens with higher fractions of C>A/G>T artifacts. These artifactual changes could be resulted from depurination of guanines, followed by incorporation of adenines by DNA polymerase in the complementary strand, which alters G-C base pairs to A-T base pairs. Heyn et al. [63] reported that DNA polymerases demonstrated varying bypass rates at abasic sites. For instance, AmpliTaq Gold, *Pfu*, and Platinum Taq HiFi extended across lower frequency of abasic sites compared to Platinum Taq, *Bst* and *Sso*-Dpo4 (<34% *vs.* >77%) [63]. Thus, selection of a high fidelity DNA polymerase could lessen these forms of sequence artifacts. Costello et al. [35] discovered that C>A/G>T artifacts can also occur due to oxidation of DNA during the shearing process, converting guanines to 8-oxoguanine lesions. This conversion is highly dependent on the surrounding 5' and 3' bases of the guanine, in which guanines within GGC are the most susceptible to oxidation. 8-oxoguanine can form base pairs with cytosine and adenine, and mispairing with adenine would give rise to artifactual C>A/G>T transversions. However, this was not the cause of C>A/G>T artifacts in our data because both blood and FFPE DNA were sheared, and we did not observe simultaneous C>A/G>T increments in both specimen types compared to other types of base changes.

Ludyga et al. [86] demonstrated that long-term storage of FFPE blocks led to increased DNA fragmentation, producing shorter template DNA for PCR amplification. Furthermore, Carrick et al. [25] showed that increased storage time of FFPE blocks affects sequencing coverage and depth in NGS data. These findings are in agreement with our results, in which we found negative asso-

ciations between age of paraffin blocks and efficiency in amplicon enrichment, coverage depth of target bases, and percentage of on-target aligned reads. As well, we observed a positive correlation between age of paraffin blocks and fraction of C>T/G>A artifacts, an outcome of stochastic enrichment. Due to exposure to environmental conditions, older FFPE blocks tend to produce increasingly fragmented DNA, which results in lower amounts of amplifiable DNA. Consequently, there is a higher chance of amplifying template DNA with sequence artifacts caused by formalin, yielding increased frequency of artifactual nucleotide changes in older FFPE specimens [145]. These results demonstrating the correlations between storage time of paraffin blocks and sequencing variables suggest that if multiple FFPE blocks are available, the specimen with the shorter storage time should be selected for molecular testing. However, clinical specimens are often limited, making sample selection a rare option in the diagnostic setting. As such, other approaches to eliminate sequence artifacts should be considered such as application of molecular barcodes and hybridization-capture enrichment, which allow tracking of DNA templates [48, 102, 112, 144]. This would enable detection of variants that are only supported by the same template DNA, indicating a higher chance that these variants are sequence artifacts and should be interpreted with caution.

Various groups have identified clinically significant germline alterations through analyzing tumour genomes [72, 93, 113]. Schrader et al. [113] reported that potential pathogenic germline variants in cancer-predisposing genes were conserved in the tumours of 91.9% of patients in their study cohort (182 of 198 patients), whereas 21.4% of these patients (39 of 182 patients) demonstrated LOH or other forms of mutations in the remaining wild type allele. We found that 93.8% of germline alterations identified in the blood were retained in the tumour with the same allelic statuses, a finding that is in line with previous work. This suggests that tumour DNA could be a reliable substrate for detecting germline alterations, implying that a tumour-only sequencing protocol could be leveraged for pre-screening of germline variants before submission to downstream confirmatory testing. A framework as such could provide germline testing in a cost-effective manner because only selective patients (i.e. those with potential germline alterations that are clinically important) would require follow-up. We also identified discordant germline variants between blood and tumour DNA, which were caused by various reasons like LOH, low sequencing coverage ($<$ 100x), and loss of variant allele in the tumours. All tumour specimens in our study were formalin-fixed, therefore it is possible that DNA damage induced by formaldehyde exposure played a role in creating discordant germline variants. Variant discordance can also be caused by mutagenesis in the tumour, such as somatic CNVs in the region of the germline variant. For instance, Gross et al. [60] showed a high prevalence of *DPYD* CNVs in high-grade triple negative breast cancer, particularly in cases with copy number loss of the *BRCA1* DNA-repair gene. The common fragile site FRA1E is located within the *DPYD* gene and its stability is highly dependent on intact *BRCA1* [9]. Hence, deficiency in *BRCA1* protein would result in increased fragility of FRA1E, leading to genomic rearrangements in *DPYD*. As germline variants in the *DPYD* gene can predict susceptibility to 5-FU-related toxicity,

somatic CNVs in *DPYD* could affect the detection of these germline variants in tumour genomic sequencing.

Although sequencing of tumour-normal pairs would enable accurate identification of germline and somatic variants, this approach is not routinely practice in clinical genomics due to inadequate funding and facilities to store additional specimens. Methods to distinguish between germline and somatic alterations in tumour-only analyses have been described by different groups [55, 64, 72]. Hiltemann et al. [64] used a virtual normal that was assembled by aggregating whole-genome-sequenced normal samples from 931 healthy and unrelated individuals, whereas Jones et al. [72] resorted to using an unmatched normal sample and public databases such as dbSNP, 1000 Genomes Project, and COSMIC, as well as effect prediction tools. We leveraged the fact that the VAFs of somatic mutations typically deviate from 50% and 100% for heterozygous and homozygous variants, respectively, and employed VAF threshold to differentiate between variant statuses. Our approach managed to achieve high sensitivity and precision, therefore verifying the feasibility of using VAF threshold to differentiate between germline and somatic alterations in the absence of matched normal samples. The VAF threshold method takes advantage of genetic impurity and heterogeneity of tumours, which render the deviation of somatic VAFs from diploid zygosity. Jones et al. [72] discovered that performance of the VAF threshold approach was highly dependent on tumour purity. While the use of VAF threshold can correctly identify germline and somatic alterations in tumours with $< 50\%$ purity, this accuracy was not observed for specimens with higher tumour content. In fact, only 12.5% of cancer-predisposing germline variants and an average of 48% of somatic mutations were accurately predicted [72]. Unfortunately, pathologic estimation of tumour content was not available for our analyses. However, we speculate that the tumour specimens in our dataset are highly impure or heterogeneous, thereby contributing to the high sensitivity and precision attained by the VAF threshold approach. While there are bioinformatic algorithms available to infer clonality and impurity estimates of tumours, many of these methods require matched normal controls or are not compatible with targeted sequencing data [147]. Nevertheless, these information should be integrated into clinical pipelines to enhance the performance of using a VAF threshold approach to distinguish between germline and somatic alterations in the course of analyzing tumour genomes without matched normal samples.

There are several limitations in our study. First, we did not manually review every single variant called by our pipeline. Only variants located within primer regions were manually inspected, while our variant filter also included common artifacts that were curated during clinical assessment. Hence, it is highly possible that sequence artifacts are present in our dataset, particularly low-allelic-fraction variants (i.e. $< 30\%$) detected in the blood. These potentially artifactual variants account for 6% of all germline variants identified in blood DNA, thereby compromising sensitivity of the VAF threshold method. Variant inspection using a genome browser is routinely conducted by genomic analysts in clinical practice to decrease the risk of reporting false positive results [55, 125].

However, manual review of variants was not implemented in our study because our analyses were focused on evaluating analytical validity instead of inferring clinical implications of the variants called. Moreover, the large number of variants in our study would be time-consuming and unfeasible for manual inspection. Our evaluation of the VAF threshold approach in differentiating between germline and somatic variants is favourable of the framework to implement initial screening for germline variants in clinical tumour sequencing before follow-up germline testing. The relatively small gene panel and cohort size of our study are caveats in drawing this conclusion. Although we were able to identify germline variants that can influence drug response, we did not report any pathogenic germline variants that are associated with cancer predisposition in our dataset. Hence, we can only speculate that our approach is scalable to variants in cancer-predisposing genes. Studies that were able to identify pathogenic germline variants were performed with cohort sizes and gene panels that are substantially larger than ours. For instance, the study by Schrader et al. [113], which revealed pathogenic germline variants in 16% of patients, was performed in a cohort of 1566 patients and screened for 341 genes. To determine whether the VAF threshold method can be applied to detect genetic alterations linked to cancer susceptibility, further assessment which involves a larger patient cohort and surveying known cancer-predisposing genes must be carried out.

The present study addresses two problems faced by using tumour genomic sequencing to identify germline alterations: the widespread use of FFPE tumours and the lack of matched normal samples. Archival FFPE tissues remain a sizable resource for cancer genomic studies and clinical genomic sequencing. Thus, there is a need to understand the extent of the different forms of DNA damage induced by formalin. Our analyses not only provide insights on the impact of formalin-induced DNA damage on amplicon-based NGS data, but also help us devise guidelines to minimize these effects. Formalin fixation followed by paraffin embedding is an attractive method to preserve tissues for histologic assessment because it allows storage at ambient temperature, which reduces cost that could be incurred by maintaining freezers required for fresh-frozen samples. Yet, many studies, including ours, have indicated the side effects of the formaldehyde exposure on nucleic acid [42, 76, 100, 101, 118, 144, 145]. Instead of investing efforts into mitigating these side effects, a potential solution is to transition from the use of formalin to the UMFIX (Sakura Finetek USA, Inc.) fixative, which is capable of preserving both cellular morphology for pathologic review and macromolecules, including DNA [136]. Most clinical laboratories conduct tumour-only sequencing and apply approaches to distinguish between germline and somatic alterations. Without matched normal samples, interpretation of variants becomes complicated. Jones et al. [72] and Garofalo et al. [55] concluded that sequencing of tumour-normal pairs is the best practice to accurately identify variant statuses. For a center to provide this service, it must be equipped to collect, analyze, and report germline findings. This includes establishing appropriate pre-test and post-test counseling, protocols to secure patient consent and manage variant of uncertain significance, and frameworks to communicate results that may implicate the patients' relatives. While various groups recommend

the sequencing of tumour-normal pairs, some centers simply do not have the funding or infrastructure to implement this as a standard practice. Furthermore, the ACMG recommended that clinical laboratories report incidental variants in 56 genes that are associated with disease risk in DNA derived from germline samples, including matched normal samples that only serve the purpose of subtracting germline variants to identify somatic mutations in tumours [59]. Interrogation of these genes suggested by the ACMG guidelines could result in detection of more variants with uncertain significance, which might pose more harm than good to patients. Additionally, cases in which only FFPE tumour blocks exist for a deceased patient would greatly benefit from approaches in differentiating between germline and somatic variants. For example, if the deceased individual is suspected to be a carrier of an inheritable disease, the ability to accurately identify the germline risk allele could prompt germline testing for the individual's relatives and facilitate preventive care. Thus, establishing approaches to tell apart germline and somatic variants in tumour genomic analyses still has its advantages from clinical and financial perspectives.

To summarize, we confirmed that the common forms of formalin-induced DNA damage in our data were DNA fragmentation and cytosine deamination. Because these effects were either minor or technically insignificant, this justifies the use of FFPE DNA for germline testing. Characterization of formalin-induced DNA damage also assist in devising recommendations to enhance amplicon enrichment and sequencing results. We also reported a high retention rate of germline alterations in the tumour genome, suggesting the reliability of using tumour DNA for germline variant calling. Finally, we showed that application of VAF threshold can achieve high sensitivity and precision in distinguishing germline alterations from somatic mutations in tumour-only analyses. This supports the framework of leveraging clinical tumour sequencing for initial germline testing. Subsequently, only patients with potential germline variants will be referred to follow-up testing. A framework as such represents a cost-effective way to deliver germline testing because only selective patients will require downstream testing. Nevertheless, scalability of this approach for discriminating between germline and somatic variants in cancer-predisposing genes needs further evaluation.

Chapter 6

Conclusion

What are my main findings?

Bibliography

[1] The Human Genome Project Completion: Frequently Asked Questions, 2010. URL https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/. → pages 2

[2] S. Afzal, M. Gusella, B. Vainer, U. B. Vogel, J. T. Andersen, K. Broedbaek, M. Petersen, E. Jimenez-Solem, L. Bertolaso, C. Barile, R. Padrini, F. Pasini, S. A. Jensen, and H. E. Poulsen. Combinations of polymorphisms in genes involved in the 5-fluorouracil metabolism pathway are associated with gastrointestinal toxicity in chemotherapy-treated colorectal cancer patients. *Clinical Cancer Research*, 17(11):3822–3829, 2011. ISSN 10780432. doi:10.1158/1078-0432.CCR-11-0304. → pages 76

[3] F. Ali-osman, O. Akande, G. Antoun, J.-x. Mao, and J. Buolamwini. Molecular Cloning , Characterization , and Expression in Escherichia coli of Full-length cDNAs of Three Human Glutathione S -Transferase Pi Gene Variants. 272(15):10004–10012, 1997. doi:10.1074/jbc.272.15.10004. → pages 73

[4] U. Amstutz, S. Farese, S. Aebi, and C. R. Largiadèr. Dihydropyrimidine dehydrogenase gene variation and severe 5-fluorouracil toxicity: a haplotype assessment. *Pharmacogenomics*, 10(6):931–944, 2009. ISSN 1462-2416. doi:10.2217/pgs.09.28. URL http://www.futuremedicine.com/doi/10.2217/pgs.09.28. → pages 70, 71, 72

[5] Y. Ando, H. Saka, M. Ando, T. Sawa, K. Muro, H. Ueoka, A. Yokoyama, S. Saitoh, K. Shimokata, and Y. Hasegawa. Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: A pharmacogenetic analysis. *Cancer Research*, 60(24):6921–6926, 2000. ISSN 00085472. → pages 76

[6] I. L. Andrulis, S. B. Bull, M. E. Blackstein, D. Sutherland, C. Mak, S. Sidlofsky, K. P. Pritzker, R. W. Hartwick, W. Hanna, L. Lickley, R. Wilkinson, A. Qizilbash, U. Ambus, M. Lipa, H. Weizel, A. Katz, M. Baida, S. Mariz, G. Stoik, P. Dacamara, D. Strongitharm, W. Geddie, and D. McCready. Neu/erbB-2 amplification identifies a poor-prognosis group of women with node-negative breast cancer. Toronto Breast Cancer Study Group. *J Clin Oncol.*, 16(4):1340–9, 1998. → pages 1

[7] N. Ånensen, J. Skavland, C. Stapnes, A. Ryningen, A.-L. Børresen-Dale, B. T. Gjertsen, and Ø. Bruserud. Acute myelogenous leukemia in a patient with LiFraumeni syndrome treated with valproic acid, theophyllamine and all-trans retinoic acid: a case report. *Leukemia*, 20 (4):734–736, 2006. ISSN 0887-6924. doi:10.1038/sj.leu.2404117. URL http://www.nature.com/doifinder/10.1038/sj.leu.2404117. → pages 64

[8] S. L. Arcand, C. M. Maugard, P. Ghadirian, A. Robidoux, C. Perret, P. Zhang, E. Fafard, A. M. Mes-Masson, W. D. Foulkes, D. Provencher, S. A. Narod, and P. N. Tonin. Germline TP53 mutations in BRCA1 and BRCA2 mutation-negative French Canadian breast cancer families. *Breast Cancer Research and Treatment*, 108(3):399–408, 2008. ISSN 01676806. doi:10.1007/s10549-007-9608-6. → pages 64

[9] M. F. Arlt, B. Xu, S. G. Durkin, A. M. Casper, M. B. Kastan, and T. W. Glover. BRCA1 Is Required for Common-Fragile-Site Stability via Its G 2 / M Checkpoint Function BRCA1 Is Required for Common-Fragile-Site Stability via Its G 2 / M Checkpoint Function. 24(15): 6701–6709, 2004. doi:10.1128/MCB.24.15.6701. → pages 93

[10] B. P. Bass, K. B. Engel, S. R. Greytak, and H. M. Moore. A review of preanalytical factors affecting molecular, protein, and morphological analysis of Formalin-Fixed, Paraffin-Embedded (FFPE) tissue: How well do you know your FFPE specimen? *Archives of Pathology and Laboratory Medicine*, 138(11):1520–1530, 2014. ISSN 15432165. doi:10.5858/arpa.2013-0691-RA. → pages 45

[11] J. Betge, G. Kerr, T. Miersch, S. Leible, G. Erdmann, C. L. Galata, T. Zhan, T. Gaiser, S. Post, M. P. Ebert, K. Horisberger, and M. Boutros. Amplicon Sequencing of Colorectal Cancer: Variant Calling in Frozen and Formalin-Fixed Samples. *Plos One*, 10(5):e0127146, 2015. ISSN 1932-6203. doi:10.1371/journal.pone.0127146. URL http://dx.plos.org/10.1371/journal.pone.0127146. → pages 91

[12] F. Biramijamal, A. Allameh, P. Mirbod, H. J. Groene, R. Koomagi, and M. Hollstein. Unusual profile and high prevalence of p53 mutations in esophageal squamous cell carcinomas from northern Iran. *Cancer research*, 61(7):3119–23, 2001. ISSN 0008-5472. URL http://www.ncbi.nlm.nih.gov/pubmed/11306496. → pages 64

[13] G. D. V. Blanco, O. A. Paoluzi, P. Sileri, P. Rossi, G. Sica, and F. Pallone. Familial colorectal cancer screening: When and what to do? *World Journal of Gastroenterology*, 21 (26):7944–7953, 2015. ISSN 22192840. doi:10.3748/wjg.v21.i26.7944. → pages 10

[14] V. Boeva, T. Popova, M. Lienard, S. Toffoli, M. Kamal, C. Le Tourneau, D. Gentien, N. Servant, P. Gestraud, T. R. Frio, P. Hupé, E. Barillot, and J. F. Laes. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*, 30(24):3443–3450, 2014. ISSN 14602059. doi:10.1093/bioinformatics/btu436. → pages 4

[15] V. Boige, M. Vincent, P. Alexandre, S. Tejpar, S. Landolfi, K. L. Malicot, R. Greil, P. J. Cuyle, M. Yilmaz, R. Faroux, A. Matzdorff, R. Salazar, C. Lepage, J. Taieb, and P. Laurent-puig. DPYD Genotyping to PredictAdverse Events Following Treatment With Fluorouracil-BasedAdjuvant Chemotherapy in Patients With Stage III Colon Cancer. *JAMA oncology*, 2(5):655–662, 2016. doi:10.1001/jamaoncol.2015.5392. → pages 70, 71

[16] S. E. Bojesen and B. G. Nordestgaard. The common germline Arg72Pro polymorphism of p53 and increased longevity in humans. *Cell Cycle*, 7(2):158–163, 2008. ISSN 15514005. doi:10.4161/cc.7.2.5249. → pages 66

[17] Y. Bombard, S.-k. Cancer, S.-k. Cancer, N. Haven, M. Robson, S.-k. Cancer, C. Medical, K. Offit, S.-k. Cancer, C. Biology, G. Program, S.-k. Cancer, and W. Cornell. Revealing the Incidentalome When Targeting the Tumor Genome. pages 7–8, 2014. doi:10.1038/gim.2013.37.Barriers. → pages 50, 86

[18] M. Bonafé, S. Salvioli, C. Barbi, C. Trapassi, F. Tocco, G. Storci, L. Invidia, I. Vannini, M. Rossi, E. Marzi, M. Mishto, M. Capri, F. Olivieri, R. Antonicelli, M. Memo, D. Uberti, B. Nacmias, S. Sorbi, D. Monti, and C. Franceschi. The different apoptotic potential of the p53 codon 72 alleles increases with age and modulates in vivo ischaemia-induced cell death. *Cell Death and Differentiation*, 11(9):962–973, 2004. ISSN 1350-9047. doi:10.1038/sj.cdd.4401415. URL http://www.nature.com/doifinder/10.1038/sj.cdd.4401415. → pages 66

[19] S. Bonin, M. Donada, G. Bussolati, E. Nardon, L. Annaratone, M. Pichler, A. M. Chiaravalli, C. Capella, G. Hoefler, and G. Stanta. A synonymous EGFR polymorphism predicting responsiveness to anti-EGFR therapy in metastatic colorectal cancer patients. *Tumor Biology*, 37(6):7295–7303, 2016. ISSN 14230380. doi:10.1007/s13277-015-4543-3. URL http://dx.doi.org/10.1007/s13277-015-4543-3. → pages 58

[20] I. E. Bosdet, T. R. Docking, Y. S. Butterfield, A. J. Mungall, T. Zeng, R. J. Coope, E. Yorida, K. Chow, M. Bala, S. S. Young, M. Hirst, I. Birol, R. A. Moore, S. J. Jones, M. A. Marra, R. Holt, and A. Karsan. A clinically validated diagnostic second-generation sequencing assay for detection of hereditary BRCA1 and BRCA2 mutations. *Journal of Molecular Diagnostics*, 15(6):796–809, 2013. ISSN 15251578. doi:10.1016/j.jmoldx.2013.07.004. → pages 16

[21] G. Bougeard, S. Baert-Desurmont, I. Tournier, S. Vasseur, C. Martin, L. Brugieres, A. Chompret, B. Bressac-de Paillerets, D. Stoppa-Lyonnet, C. Bonaiti-Pellie, and T. Frebourg. Impact of the MDM2 SNP309 and p53 Arg72Pro polymorphism on age of tumour onset in Li-Fraumeni syndrome. *Journal of medical genetics*, 43(6):531–3, 2006. ISSN 1468-6244. doi:10.1136/jmg.2005.037952. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1904480{&}tool=pmcentrez{&}rendertype=abstract. → pages 66

[22] T. Boveri. *Zur Frage der Entstehung Maligner Tumoren*. Gustav Fischer, 1914. → pages 1

[23] L. Cai, W. Yuan, Z. Zhang, L. He, and K.-C. Chou. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific reports*, 6(November):36540, 2016. ISSN 2045-2322. doi:10.1038/srep36540. URL http://www.ncbi.nlm.nih.gov/pubmed/27874022{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5118795. → pages 91

[24] D. Caronia, M. Martin, J. Sastre, J. De La Torre, J. A. García-Sáenz, M. R. Alonso, L. T. Moreno, G. Pita, E. Díaz-Rubio, J. Benítez, and A. González-Neira. A polymorphism in the cytidine deaminase promoter predicts severe capecitabine-induced hand-foot syndrome. *Clinical Cancer Research*, 17(7):2006–2013, 2011. ISSN 10780432. doi:10.1158/1078-0432.CCR-10-1741. → pages 74

[25] D. M. Carrick, M. G. Mehaffey, M. C. Sachs, S. Altekruse, C. Camalier, R. Chuaqui, W. Cozen, B. Das, B. Y. Hernandez, C. J. Lih, C. F. Lynch, H. Makhlouf, P. McGregor, L. M. McShane, J. P. Rohan, W. D. Walsh, P. M. Williams, E. M. Gillanders, L. E. Mechanic, and S. D. Schully. Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. *PLoS ONE*, 10(7):3–10, 2015. ISSN 19326203. doi:10.1371/journal.pone.0127353. → pages 45, 92

[26] J. Carrot-Zhang and J. Majewski. LoLoPicker: Detecting Low Allelic-Fraction Variants in Low-Quality Cancer Samples from Whole-exome Sequencing Data. *bioRxiv*, 8(23):043612, 2016. ISSN 1949-2553. doi:10.1101/043612. URL http://biorxiv.org/content/early/2016/04/24/043612.abstract. → pages 91

[27] K. E. Caudle, C. F. Thorn, T. E. Klein, J. J. Swen, H. L. McLeod, R. B. Diasio, and M. Schwab. Clinical Pharmacogenetics Implementation Consortium Guidelines for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing. *Clinical Pharmacology & Therapeutics*, 94(6):640–645, 2013. ISSN 0009-9236. doi:10.1038/clpt.2013.172. URL http://doi.wiley.com/10.1038/clpt.2013.172. → pages 70, 71, 72

[28] S. Cheli, F. Pietrantonio, E. Clementi, and F. S. Falvella. LightSNiP assay is a good strategy for pharmacogenetics test. *Frontiers in Pharmacology*, 6(JUN):1–5, 2015. ISSN 16639812. doi:10.3389/fphar.2015.00114. → pages 76

[29] G. Chen, S. Mosier, C. D. Gocke, M.-T. Lin, and J. R. Eshleman. Cytosine Deamination is a Major Cause of Baseline Noise in Next Generation Sequencing. *Mol Diagn Ther.*, 18(5): 587–593, 2014. doi:10.1007/s40291-014-0115-2.Cytosine. → pages 91

[30] Y. C. Chen, C. H. Tzeng, P. M. Chen, J. K. Lin, T. C. Lin, W. S. Chen, J. K. Jiang, H. S. Wang, and W. S. Wang. Influence of GSTP1 I105V polymorphism on cumulative neuropathy and outcome of FOLFOX-4 treatment in Asian patients with colorectal carcinoma. *Cancer Science*, 101(2):530–535, 2010. ISSN 13479032. doi:10.1111/j.1349-7006.2009.01418.x. → pages 73

[31] H. Cheng, B. Ma, R. Jiang, W. Wang, H. Guo, N. Shen, D. Li, Q. Zhao, R. Wang, P. Yi, Y. Zhao, Z. Liu, and T. Huang. Individual and combined effects of MDM2 SNP309 and TP53 Arg72Pro on breast cancer risk: An updated meta-analysis. *Molecular Biology Reports*, 39(9):9265–9274, 2012. ISSN 03014851. doi:10.1007/s11033-012-1800-z. → pages 66

[32] K. N. Chitrala and S. Yeguvapalli. Computational screening and molecular dynamic simulation of breast cancer associated deleterious non-synonymous single nucleotide polymorphisms in TP53 gene. *PLoS ONE*, 9(8), 2014. ISSN 19326203. doi:10.1371/journal.pone.0104242. → pages 64

[33] J. Chung, D.-S. Son, H.-J. Jeon, K.-M. Kim, G. Park, G. H. Ryu, W.-Y. Park, and D. Park. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Scientific Reports*, 6(1):26732, 2016. ISSN 2045-2322. doi:10.1038/srep26732. URL http://www.nature.com/articles/srep26732. → pages 1

[34] V. Cohen, V. Panet-raymond, N. Sabbaghian, I. Morin, and G. Batist. Methylenetetrahydrofolate Reductase Polymorphism in Advanced Colorectal Cancer : A Novel Genomic Predictor of Clinical Response to Fluoropyrimidine-based Chemotherapy Advances in Brief Methylenetetrahydrofolate Reductase Polymorphism in Advanced Colorecta. *Clinical Cancer Research*, 9(May):1611–1615, 2003. → pages 73

[35] M. Costello, T. J. Pugh, T. J. Fennell, C. Stewart, L. Lichtenstein, J. C. Meldrim, J. L. Fostel, D. C. Friedrich, D. Perrin, D. Dionne, S. Kim, S. B. Gabriel, E. S. Lander, S. Fisher, and G. Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):1–12, 2013. ISSN 03051048. doi:10.1093/nar/gks1443. → pages 92

[36] Z. Dai, A. C. Papp, D. Wang, H. Hampel, and W. Sadee. Genotyping panel for assessing response to cancer chemotherapy. *BMC Medical Genomics*, 1(1):24, 2008. ISSN 1755-8794. doi:10.1186/1755-8794-1-24. URL http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-1-24. → pages 50

[37] F. A. de Jong. Prophylaxis of Irinotecan-Induced Diarrhea with Neomycin and Potential Role for UGT1A1*28 Genotype Screening: A Double-Blind, Randomized, Placebo-Controlled Study. *The Oncologist*, 11(8):944–954, 2006. ISSN 1083-7159. doi:10.1634/theoncologist.11-8-944. URL http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.11-8-944. → pages 76

[38] M. J. Deenen, J. Tol, A. M. Burylo, V. D. Doodeman, A. De Boer, A. Vincent, H. J. Guchelaar, P. H. M. Smits, J. H. Beijnen, C. J. A. Punt, J. H. M. Schellens, and A. Cats. Relationship between single nucleotide polymorphisms and haplotypes in DPYD and toxicity and efficacy of capecitabine in advanced colorectal cancer. *Clinical Cancer Research*, 17(10):3455–3468, 2011. ISSN 10780432. doi:10.1158/1078-0432.CCR-10-2209. → pages 70, 71, 72

[39] M. Depner, S. Fuchs, J. Raabe, N. Frede, C. Glocker, R. Doffinger, E. Gkrania-Klotsas, D. Kumararatne, T. P. Atkinson, H. W. Schroeder, T. Niehues, G. D??ckers, A. Stray-Pedersen, U. Baumann, R. Schmidt, J. L. Franco, J. Orrego, M. Ben-Shoshan, C. McCusker, C. M. A. Jacob, M. Carneiro-Sampaio, L. A. Devlin, J. D. M. Edgar, P. Henderson, R. K. Russell, A. B. Skytte, S. L. Seneviratne, J. Wanders, H. Stauss, I. Meyts, L. Moens, M. Jesenak, R. Kobbe, S. Borte, M. Borte, D. A. Wright, D. Hagin, T. R. Torgerson, and B. Grimbacher. The Extended Clinical Phenotype of 26 Patients with Chronic Mucocutaneous Candidiasis due to Gain-of-Function Mutations in STAT1. *Journal of Clinical Immunology*, 36(1):73–84, 2016. ISSN 15732592. doi:10.1007/s10875-015-0214-9. → pages 63

[40] A. Didelot, S. K. Kotsopoulos, A. Lupo, D. Pekin, X. Li, I. Atochin, P. Srinivasan, Q. Zhong, J. Olson, D. R. Link, P. Laurent-Puig, H. Blons, J. B. Hutchison, and V. Taly. Multiplex picoliter-droplet digital PCR for quantitative assessment of DNA integrity in clinical samples. *Clinical Chemistry*, 59(5):815–823, 2013. ISSN 00099147. doi:10.1373/clinchem.2012.193409. → pages 25, 45, 91

[41] H. Do and A. Dobrovic. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. *Oncotarget*, 3 (5):546–58, 2012. ISSN 1949-2553. doi:10.18632/oncotarget.503. URL http://www.ncbi.nlm.nih.gov/pubmed/22643842{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3388184. → pages 37, 91, 92

[42] H. Do and A. Dobrovic. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clinical Chemistry*, 61(1):64–71, 2015. ISSN 15308561. doi:10.1373/clinchem.2014.223040. → pages 25, 37, 90, 91, 92, 95

[43] H. Do, S. Q. Wong, J. Li, and A. Dobrovic. Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clinical Chemistry*, 59(9):1376–1383, 2013. ISSN 00099147. doi:10.1373/clinchem.2012.202390. → pages 37, 91, 92

[44] D. Dobritzsch, G. Schneider, K. D. Schnackerz, and Y. Lindqvist. Crystal structure of dihydropyrimidine dehydrogenase, a major determinant of the pharmacokinetics of the anti-cancer drug 5-fluorouracil. *EMBO Journal*, 20(4):650–660, 2001. ISSN 02614189. doi:10.1093/emboj/20.4.650. → pages 70

[45] E. Dotor, M. Cuatrecases, M. Martínez-Iniesta, M. Navarro, F. Vilardell, E. Guinó, L. Pareja, A. Figueras, D. G. Molleví, T. Serrano, J. De Oca, M. A. Peinado, V. Moreno, J. R. Germà, G. Capellá, and A. Villanueva. Tumor thymidylate synthase 1494del6 genotype as a prognostic factor in colorectal cancer patients receiving fluorouracil-based adjuvant treatment. *Journal of Clinical Oncology*, 24(10):1603–1611, 2006. ISSN 0732183X. doi:10.1200/JCO.2005.03.5253. → pages 76

[46] J. A. Drebin, V. C. Link, D. F. Stern, R. A. Weinberg, and M. I. Greene. Down-modulation of an oncogene protein product and reversion of the transformed phenotype by monoclonal antibodies. *Cell*, 41(3):695–706, 1985. ISSN 00928674. doi:10.1016/S0092-8674(85)80050-7. → pages 1

[47] B. J. Druker, F. Guilhot, S. G. O'Brien, I. Gathmann, H. Kantargian, N. Gattermann, M. W. Deininger, R. T. Silver, J. M. Goldman, R. M. Stone, F. Cervantes, A. Hochhaus, B. L. Powell, J. L. Gabrilove, P. Rousselot, J. Reiffers, J. J. Cornelissen, T. Hughes, H. Agis, T. Fischer, G. Verhoef, J. Shepherd, G. Saglio, A. Gratwohl, J. L. Nielsen, J. P. Radich, B. Simonsson, K. Taylor, M. Baccarani, C. So, L. Letvak, and R. A. Larson. Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *New England Journal of Medicine*, 355:2408–2417, 2006. → pages 1

[48] A. Eijkelenboom, E. J. Kamping, A. W. Kastner-van Raaij, S. J. Hendriks-Cornelissen, K. Neveling, R. P. Kuiper, A. Hoischen, M. R. Nelen, M. J. Ligtenberg, and B. B. Tops. Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *The Journal of Molecular Diagnostics*, 18(6):851–863, 2016. ISSN 15251578. doi:10.1016/j.jmoldx.2016.06.010. URL http://dx.doi.org/10.1016/j.jmoldx.2016.06.010http://linkinghub.elsevier.com/retrieve/pii/S1525157816301416. → pages 93

[49] P. Estevez-Garcia, A. Castaño, A. C. Martin, F. Lopez-Rios, J. Iglesias, S. Muñoz-Galván, I. Lopez-Calderero, S. Molina-Pinelo, M. D. Pastor, A. Carnero, L. Paz-Ares, and R. Garcia-Carbonero. PDGFRα/β and VEGFR2 polymorphisms in colorectal cancer: incidence and implications in clinical outcome. *BMC cancer*, 12:514, 2012. ISSN 1471-2407. doi:10.1186/1471-2407-12-514. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531259{&}tool=pmcentrez{&}rendertype=abstract. → pages 62

[50] M. C. Etienne, J. L. Formento, M. Chazal, M. Francoual, N. Magne, P. Formento, A. Bourgeon, J. F. Seitz, J. R. Delpero, C. Letoublon, D. Pezet, and G. Milano. Methylenetetrahydrofolate reductase gene polymorphisms and response to fluorouracil-based treatment in advanced colorectal cancer patients. *Pharmacogenetics*, 14 (12):785–792, 2004. ISSN 0960-314X. → pages 73

[51] M. C. Etienne-Grimaldi, G. Milano, F. Maindrault-Gœbel, B. Chibaudel, J. L. Formento, M. Francoual, G. Lledo, T. André, M. Mabro, L. Mineur, M. Flesch, E. Carola, and A. De Gramont. Methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms and FOLFOX response in colorectal cancer patients. *British Journal of Clinical Pharmacology*, 69(1):58–66, 2010. ISSN 03065251. doi:10.1111/j.1365-2125.2009.03556.x. → pages 50, 73

[52] G. Fortunato, G. Calcagno, V. Bresciamorra, E. Salvatore, A. Filla, S. Capone, R. Liguori, S. Borelli, I. Gentile, F. Borrelli, G. Borgia, and L. Sacchetti. Multiple_sclerosis_and_hepatit.PDF. *Journal of Interferon & Cytokine Research*, 28: 141–152, 2008. doi:doi:10.1089/jir.2007.0049. → pages 63

[53] G. M. G. Frampton, A. Fichtenholtz, G. a. G. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. a. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. a. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, and R. Yelensky. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature*, 31(October): 1–11, 2013. ISSN 1087-0156. doi:10.1038/nbt.2696. URL http://dx.doi.org/10.1038/nbt.2696{%}5Cnhttp://www.nature.com/nbt/journal/v31/n11/abs/nbt.2696.html{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24142049. → pages 50

[54] D. Fumagalli, P. G. Gavin, Y. Taniyama, S.-I. Kim, H.-J. Choi, S. Paik, and K. L. Pogue-Geile. A rapid, sensitive, reproducible and cost-effective method for mutation profiling of colon cancer and metastatic lymph nodes. *BMC cancer*, 10:101, 2010. ISSN 1471-2407. doi:10.1186/1471-2407-10-101. → pages 50

[55] A. Garofalo, L. Sholl, B. Reardon, A. Taylor-Weiner, A. Amin-Mansour, D. Miao, D. Liu, N. Oliver, L. MacConaill, M. Ducar, V. Rojas-Rudilla, M. Giannakis, A. Ghazani, S. Gray, P. Janne, J. Garber, S. Joffe, N. Lindeman, N. Wagle, L. A. Garraway, and E. M. Van Allen. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Medicine*, 8(1):79, 2016. ISSN 1756-994X.

doi:10.1186/s13073-016-0333-9. URL http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0333-9. → pages 94, 95

[56] G. Gentile, A. Botticelli, L. Lionetto, F. Mazzuca, M. Simmaco, P. Marchetti, and M. Borro. Genotypephenotype correlations in 5-fluorouracil metabolism: a candidate DPYD haplotype to improve toxicity prediction. *The Pharmacogenomics Journal*, 16(4):320–325, 2016. ISSN 1470-269X. doi:10.1038/tpj.2015.56. URL http://www.nature.com/doifinder/10.1038/tpj.2015.56. → pages 70, 71, 72

[57] B. Glimelius, H. Garmo, A. Berglund, L. a. Fredriksson, M. Berglund, H. Kohnke, P. Byström, H. Sørbye, and M. Wadelius. Prediction of irinotecan and 5-fluorouracil toxicity and response in patients with advanced colorectal cancer. *The pharmacogenomics journal*, 11(1):61–71, 2011. ISSN 1470-269X. doi:10.1038/tpj.2010.10. → pages 76

[58] F. Graziano, A. Ruzzo, F. Loupakis, D. Santini, V. Catalano, E. Canestrari, P. Maltese, R. Bisonni, L. Fornaro, G. Baldi, G. Masi, A. Falcone, G. Tonini, P. Giordani, P. Alessandroni, L. Giustini, B. Vincenzi, and M. Magnani. Liver-only metastatic colorectal cancer patients and thymidylate synthase polymorphisms for predicting response to 5-fluorouracil-based chemotherapy. *British journal of cancer*, 99(5):716–21, 2008. ISSN 1532-1827. doi:10.1038/sj.bjc.6604555. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-50249189069{&}partnerID=tZOtx3y1. → pages 76

[59] R. C. Green, J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, R. L. Nussbaum, J. M. O'Daniel, K. E. Ormond, H. L. Rehm, M. S. Watson, M. S. Williams, L. G. Biesecker, and American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(7):565–74, 2013. ISSN 1530-0366. doi:10.1038/gim.2013.73. URL http://www.ncbi.nlm.nih.gov/pubmed/23788249{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3727274. → pages 50, 86, 96

[60] E. Gross, C. Meul, S. Raab, C. Propping, S. Avril, M. Aubele, A. Gkazepis, T. Schuster, N. Grebenchtchikov, M. Schmitt, M. Kiechle, J. Meijer, R. Vijzelaar, A. Meindl, and A. B. P. van Kuilenburg. Somatic copy number changes in DPYD are associated with lower risk of recurrence in triple-negative breast cancers. *British Journal of Cancer*, 109(9): 2347–2355, 2013. ISSN 0007-0920. doi:10.1038/bjc.2013.621. URL http://www.nature.com/doifinder/10.1038/bjc.2013.621. → pages 93

[61] V. Guillem, J. C. Hernandez-Boluda, D. Gallardo, I. Buno, A. Bosch, C. Martinez-Laperche, R. de la Camara, S. Brunet, C. Martin, J. B. Nieto, C. Martinez, A. Perez, J. Montoro, A. Garcia-Noblejas, and C. Solano. A polymorphism in the TYMP gene is associated with the outcome of HLA-identical sibling allogeneic stem cell transplantation. *American Journal of Hematology*, 88(10):883–889, 2013. ISSN 03618609. doi:10.1002/ajh.23523. → pages 74

[62] M. Gusella, G. Crepaldi, C. Barile, A. Bononi, D. Menon, S. Toso, D. Scapoli, L. Stievano, E. Ferrazzi, F. Grigoletto, M. Ferrari, and R. Padrini. Pharmacokinetic and demographic markers of 5-fluorouracil toxicity in 181 patients on adjuvant therapy for colorectal cancer. *Annals of Oncology*, 17(11):1656–1660, 2006. ISSN 09237534. doi:10.1093/annonc/mdl284. URL http://dx.doi.org/10.1038/sj.bjc.6605052. → pages 73, 76

[63] P. Heyn, U. Stenzel, A. W. Briggs, M. Kircher, M. Hofreiter, and M. Meyer. Road blocks on paleogenomes–polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic acids research*, 38(16):e161, 2010. ISSN 13624962. doi:10.1093/nar/gkq572. → pages 37, 92

[64] S. Hiltemann, G. Jenster, J. Trapman, P. Van Der Spek, and A. Stubbs. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Research*, 25(9):1382–1390, 2015. ISSN 15495469. doi:10.1101/gr.183053.114. → pages 50, 94

[65] M. Hofreiter, V. Jaenicke, D. Serre, A. von Haeseler, and S. Pääbo. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, 29(23):4793–4799, 2001. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/29.23.4793. → pages 37

[66] J. Hong, S. W. Han, H. S. Ham, T. Y. Kim, I. S. Choi, B. S. Kim, D. Y. Oh, S. A. Im, G. H. Kang, Y. J. Bang, and T. Y. Kim. Phase II study of biweekly S-1 and oxaliplatin combination chemotherapy in metastatic colorectal cancer and pharmacogenetic analysis. *Cancer Chemotherapy and Pharmacology*, 67(6):1323–1331, 2011. ISSN 03445704. doi:10.1007/s00280-010-1425-7. → pages 73

[67] L. Huang, F. Chen, Y. Chen, X. Yang, S. Xu, S. Ge, S. Fu, T. Chao, Q. Yu, X. Liao, G. Hu, P. Zhang, and X. Yuan. Thymidine phosphorylase gene variant, platelet counts and survival in gastrointestinal cancer patients treated by fluoropyrimidines. *Scientific reports*, 4(1):5697, 2014. ISSN 2045-2322. doi:10.1038/srep05697. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4100023{&}tool=pmcentrez{&}rendertype=abstract. → pages 74

[68] Z.-H. Huang, D. Hua, L.-H. Li, and J.-D. Zhu. Prognostic role of p53 codon 72 polymorphism in gastric cancer patients treated with fluorouracil-based adjuvant chemotherapy. *Journal of cancer research and clinical oncology*, 134(10):1129–1134, 2008. ISSN 0171-5216. doi:10.1007/s00432-008-0380-8. → pages 66

[69] F. Innocenti, S. D. Undevia, L. Iyer, P. X. Chen, S. Das, M. Kocherginsky, T. Karrison, L. Janisch, J. Ramírez, C. M. Rudin, E. E. Vokes, and M. J. Ratain. Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of irinotecan. *Journal of Clinical Oncology*, 22(8):1382–1388, 2004. ISSN 0732183X. doi:10.1200/JCO.2004.07.173. → pages 76

[70] A. Jakobsen, J. N. Nielsen, N. Gyldenkerne, and J. Lindeberg. Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphism in normal tissue as predictors of fluorouracil sensitivity. *Journal of Clinical Oncology*, 23(7):1365–1369, 2005. ISSN 0732183X. doi:10.1200/JCO.2005.06.219. → pages 73

[71] B. A. Jennings, Y. K. Loke, J. Skinner, M. Keane, G. S. Chu, R. Turner, D. Epurescu, A. Barrett, and G. Willis. Evaluating Predictive Pharmacogenetic Signatures of Adverse Events in Colorectal Cancer Patients Treated with Fluoropyrimidines. *PLoS ONE*, 8(10): 1–9, 2013. ISSN 19326203. doi:10.1371/journal.pone.0078053. → pages 50, 74

[72] S. Jones, V. Anagnostou, K. Lytle, S. Parpart-li, M. Nesselbush, D. R. Riley, M. Shukla, B. Chesnick, M. Kadan, E. Papp, K. G. Galens, D. Murphy, T. Zhang, L. Kann, M. Sausen, S. V. Angiuoli, L. A. D. Jr, and V. E. Velculescu. Personalized genomic analyses for cancer mutation discovery and interpretation. *Science Translational Medicine*, 7(283):283ra53, 2015. ISSN 1946-6234. doi:10.1126/scitranslmed.aaa7161. → pages 12, 50, 51, 80, 90, 93, 94, 95

[73] A. V. Khrunin, A. Moisseev, V. Gorbunova, and S. Limborska. Genetic polymorphisms and the efficacy and toxicity of cisplatin-based chemotherapy in ovarian cancer patients. *The Pharmacogenomics Journal*, 10(1):54–61, 2010. ISSN 1470-269X. doi:10.1038/tpj.2009.45. URL http://www.nature.com/doifinder/10.1038/tpj.2009.45. → pages 66

[74] J. Kim, D. Kim, J. S. Lim, J. H. Maeng, H. Son, H.-C. Kang, H. Nam, J. H. Lee, and S. Kim. Accurate detection of low-level somatic mutations with technical replication for next-generation sequencing. 2017. doi:10.1101/179713. URL http://dx.doi.org/10.1101/179713. → pages 91

[75] J. G. Kim, S. K. Sohn, Y. S. Chae, H. S. Song, K. Y. Kwon, Y. R. Do, M. K. Kim, K. H. Lee, M. S. Hyun, W. S. Lee, C. H. Sohn, J. S. Jung, G. C. Kim, H. Y. Chung, and W. Yu. TP53 codon 72 polymorphism associated with prognosis in patients with advanced gastric cancer treated with paclitaxel and cisplatin. *Cancer Chemotherapy and Pharmacology*, 64(2): 355–360, 2009. ISSN 03445704. doi:10.1007/s00280-008-0879-3. → pages 66

[76] S. Kim, C. Park, Y. Ji, D. G. Kim, H. Bae, M. van Vrancken, D. H. Kim, and K. M. Kim. Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine. *Journal of Molecular Diagnostics*, 19(1):137–146, 2017. ISSN 19437811. doi:10.1016/j.jmoldx.2016.09.006. URL http://dx.doi.org/10.1016/j.jmoldx.2016.09.006. → pages 25, 37, 90, 91, 92, 95

[77] Y. Kodama, M. Shumway, and R. Leinonen. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):2011–2013, 2012. ISSN 03051048. doi:10.1093/nar/gkr854. → pages 3

[78] D. M. Kweekel, H. Gelderblom, T. Van der Straaten, N. F. Antonini, C. J. A. Punt, and H.-J. Guchelaar. UGT1A1*28 genotype and irinotecan dosage in patients with metastatic colorectal cancer: a Dutch Colorectal Cancer Group study. *British Journal of Cancer*, 99 (2):275–282, 2008. ISSN 0007-0920. doi:10.1038/sj.bjc.6604461. URL http://www.nature.com/doifinder/10.1038/sj.bjc.6604461. → pages 76

[79] J. Laskin, S. Jones, S. Aparicio, S. Chia, C. Ch'ng, R. Deyell, P. Eirew, A. Fok, K. Gelmon, C. Ho, D. Huntsman, M. Jones, K. Kasaian, A. Karsan, S. Leelakumari, Y. Li, H. Lim, Y. Ma, C. Mar, M. Martin, R. Moore, A. Mungall, K. Mungall, E. Pleasance, S. R. Rassekh,

D. Renouf, Y. Shen, J. Schein, K. Schrader, S. Sun, A. Tinker, E. Zhao, S. Yip, and M. A. Marra. Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Molecular Case Studies*, 1(1):a000570, 2015. ISSN 2373-2865. doi:10.1101/mcs.a000570. URL http://molecularcasestudies.cshlp.org/lookup/doi/10.1101/mcs.a000570. → pages 5

[80] A. M. Lee, Q. Shi, E. Pavey, S. R. Alberts, D. J. Sargent, F. A. Sinicrope, J. L. Berenberg, R. M. Goldberg, and R. B. Diasio. DPYD variants as predictors of 5-fluorouracil toxicity in adjuvant colon cancer treatment (NCCTG N0147). *Journal of the National Cancer Institute*, 106(12):1–12, 2014. ISSN 14602105. doi:10.1093/jnci/dju298. → pages 50, 70, 71

[81] J. Leichsenring, A.-L. Volckmar, N. Magios, C. M. M. de Oliveira, R. Penzel, R. Brandt, M. Kirchner, F. Bozorgmehr, M. Thomas, P. Schirmacher, A. Warth, V. Endris, and A. Stenzinger. Synonymous EGFR Variant p.Q787Q is Neither Prognostic Nor Predictive in Patients with Lung Adenocarcinoma Jonas. *Genes, chromosomes & cancer*, 56(3):214–220, 2017. doi:doi:10.1002/gcc.22427. → pages 58

[82] F. Leone, G. Cavalloni, Y. Pignochino, I. Sarotto, R. Ferraris, W. Piacibello, T. Venesio, L. Capussotti, M. Risio, and M. Aglietta. Somatic mutations of epidermal growth factor receptor in bile duct and gallbladder carcinoma. *Clinical Cancer Research*, 12(6): 1680–1685, 2006. ISSN 10780432. doi:10.1158/1078-0432.CCR-05-1692. → pages 58

[83] M.-T. Lin, S. L. Mosier, M. Thiess, K. F. Beierl, M. Debeljak, L.-H. Tseng, G. Chen, S. Yegnasubramanian, H. Ho, L. Cope, S. J. Wheelan, C. D. Gocke, and J. R. Eshleman. Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. *American journal of clinical pathology*, 141(6):856–66, 2014. ISSN 1943-7722. doi:10.1309/AJCPMWGWGO34EGOD. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4332779{&}tool=pmcentrez{&}rendertype=abstract. → pages 37, 50

[84] Y. Liu, S. N. Xu, Y. S. Chen, X. Y. Wu, L. Qiao, K. Li, and L. Yuan. Study of single nucleotide polymorphisms of FBW7 and its substrate genes revealed a predictive factor for paclitaxel plus cisplatin chemotherapy in Chinese patients with advanced esophageal squamous cell carcinoma. *Oncotarget*, 7(28):44330–44339, 2016. ISSN 1949-2553. doi:10.18632/oncotarget.9736. URL http://www.ncbi.nlm.nih.gov/pubmed/27259248{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5190100. → pages 61

[85] J. J. Lou, L. Mirsadraei, D. E. Sanchez, R. W. Wilson, M. Shabihkhani, G. M. Lucey, B. Wei, E. J. Singer, S. Mareninov, and W. H. Yong. A review of room temperature storage of biospecimen tissue and nucleic acids for anatomic pathology laboratories and biorepositories. *Clinical Biochemistry*, 47(4-5):267–273, 2014. ISSN 18732933. doi:10.1016/j.clinbiochem.2013.12.011. URL http://dx.doi.org/10.1016/j.clinbiochem.2013.12.011. → pages 25

[86] N. Ludyga, B. Gr??nwald, O. Azimzadeh, S. Englert, H. H??fler, S. Tapio, and M. Aubele. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses.

Virchows Archiv, 460(2):131–140, 2012. ISSN 09456317. doi:10.1007/s00428-011-1184-9. → pages 45, 92

[87] M. V. Mandola, J. Stoehlmacher, W. Zhang, S. Groshen, M. C. Yu, S. Iqbal, H.-j. Lenz, and R. D. Ladner. A 6 bp polymorphism in the thymidylate synthase gene causes message instability and is associated with decreased intratumoral TS mRNA levels. *Pharmacogenetics*, 14(5):319–327, 2004. ISSN 0960-314X. doi:10.1097/01.fpc.0000114730.08559.df. → pages 76

[88] E. Marcuello, a. Altés, a. Menoyo, E. Del Rio, M. Gómez-Pardo, and M. Baiget. UGT1A1 gene variations and irinotecan treatment in patients with metastatic colorectal cancer. *British journal of cancer*, 91(4):678–682, 2004. ISSN 0007-0920. doi:10.1038/sj.bjc.6602042. → pages 76

[89] E. Marcuello, A. Altés, A. Menoyo, E. Del Rio, and M. Baiget. Methylenetetrahydrofolate reductase gene polymorphisms: Genomic predictors of clinical response to fluoropyrimidine-based chemotherapy? *Cancer Chemotherapy and Pharmacology*, 57(6): 835–840, 2006. ISSN 03445704. doi:10.1007/s00280-005-0089-1. → pages 73

[90] L. K. Mattison, M. R. Johnson, and R. B. Diasio. A comparative analysis of translated dihydropyrimidine dehydrogenase cDNA; conservation of functional domains and relevance to genetic polymorphisms. *Pharmacogenetics*, 12(2):133–44, 2002. ISSN 0960-314X. doi:10.1097/00008571-200203000-00007. URL http://www.ncbi.nlm.nih.gov/pubmed/11875367. → pages 70

[91] H. L. McLeod. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science*, 339(March):1563–1566, 2013. doi:10.1126/science.1234139. → pages 50

[92] H. L. McLeod, D. J. Sargent, S. Marsh, E. M. Green, C. R. King, C. S. Fuchs, R. K. Ramanathan, S. K. Williamson, B. P. Findlay, S. N. Thibodeau, A. Grothey, R. F. Morton, and R. M. Goldberg. Pharmacogenetic predictors of adverse events and response to chemotherapy in metastatic colorectal cancer: Results from North American Gastrointestinal Intergroup Trial N9741. *Journal of Clinical Oncology*, 28(20):3227–3233, 2010. ISSN 0732183X. doi:10.1200/JCO.2009.21.7943. → pages 73, 76

[93] F. Meric-Bernstam, L. Brusco, M. Daniels, C. Wathoo, A. M. Bailey, L. Strong, K. Shaw, K. Lu, Y. Qi, H. Zhao, H. Lara-Guerra, J. Litton, B. Arun, A. K. Eterovic, U. Aytac, M. Routbort, V. Subbiah, F. Janku, M. A. Davies, S. Kopetz, J. Mendelsohn, G. B. Mills, and K. Chen. Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol. *Annals of Oncology*, 27(5):795–800, 2016. ISSN 15698041. doi:10.1093/annonc/mdw018. → pages 50, 51, 80, 90, 93

[94] D. Meulendijks, L. M. Henricks, G. S. Sonke, M. J. Deenen, T. K. Froehlich, U. Amstutz, C. R. Largiadèr, B. A. Jennings, A. M. Marinaki, J. D. Sanderson, Z. Kleibl, P. Kleiblova, M. Schwab, U. M. Zanger, C. Palles, I. Tomlinson, E. Gross, A. B. P. van Kuilenburg, C. J. A. Punt, M. Koopman, J. H. Beijnen, A. Cats, and J. H. M. Schellens. Clinical relevance of DPYD variants c.1679T>G, c.1236G>A/HapB3, and c.1601G>A as predictors of severe fluoropyrimidine-associated toxicity: A systematic review and

meta-analysis of individual patient data. *The Lancet Oncology*, 16(16):1639–1650, 2015. ISSN 14745488. doi:10.1016/S1470-2045(15)00286-7. → pages 70, 71, 72

[95] B. Mohelnikova-Duchonova, B. Melichar, and P. Soucek. FOLFOX/FOLFIRI pharmacogenetics: The call for a personalized approach in colorectal cancer therapy. *World Journal of Gastroenterology*, 20(30):10316–10330, 2014. ISSN 22192840. doi:10.3748/wjg.v20.i30.10316. → pages 50

[96] A. Morel, M. Boisdron-Celle, L. Fey, P. Soulie, M. C. Craipeau, S. Traore, and E. Gamelin. Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance. *Molecular Cancer Therapeutics*, 5(11): 2895–2904, 2006. ISSN 1535-7163. doi:10.1158/1535-7163.MCT-06-0327. URL http://mct.aacrjournals.org/cgi/doi/10.1158/1535-7163.MCT-06-0327. → pages 50, 70, 71, 72

[97] Q. Nie, S. Shrestha, E. E. Tapper, C. S. Trogstad-Isaacson, K. J. Bouchonville, A. M. Lee, R. Wu, C. R. Jerde, Z. Wang, P. A. Kubica, S. M. Offer, and R. B. Diasio. Quantitative contribution of rs75017182 to dihydropyrimidine dehydrogenase mRNA splicing and enzyme activity. *Clinical Pharmacology & Therapeutics*, 00(00):1–9, 2017. ISSN 00099236. doi:10.1002/cpt.685. URL http://doi.wiley.com/10.1002/cpt.685. → pages 72

[98] I. Nishino, A. Spinazzola, A. Papadimitriou, S. Hammans, I. Steiner, C. D. Hahn, A. M. Connolly, A. Verloes, J. Guimarães, I. Maillard, H. Hamano, M. A. Donati, C. E. Semrad, J. A. Russell, A. L. Andreu, G. M. Hadjigeorgiou, T. H. Vu, S. Tadesse, T. G. Nygaard, I. Nonaka, I. Hirano, E. Bonilla, L. P. Rowland, S. Dimauro, and M. Hirano. Mitochondrial neurogastrointestinal encephalomyopathy: An autosomal recessive disorder due to thymidine phosphorylase mutations. *Annals of Neurology*, 47(6):792–800, 2000. ISSN 03645134. doi:10.1002/1531-8249(200006)47:6⟨792::AID-ANA12⟩3.0.CO;2-Y. → pages 75

[99] S. M. Offer, C. C. Fossum, N. J. Wegner, A. J. Stuflesser, G. L. Butterfield, and R. B. Diasio. Comparative functional analysis of dpyd variants of potential clinical relevance to dihydropyrimidine dehydrogenase activity. *Cancer Research*, 74(9):2545–2554, 2014. ISSN 15387445. doi:10.1158/0008-5472.CAN-13-2482. → pages 70, 71, 72

[100] R. Ofner, C. Ritter, S. Ugurel, L. Cerroni, M. Stiller, T. Bogenrieder, F. Solca, and D. Schrama. Non-reproducible sequence artifacts in FFPE tissue : an experience report. *Journal of Cancer Research and Clinical Oncology*, 143(7):1199–1207, 2017. ISSN 1432-1335. doi:10.1007/s00432-017-2399-1. → pages 25, 37, 90, 95

[101] E. Oh, Y.-L. Choi, M. J. Kwon, R. N. Kim, Y. J. Kim, J.-Y. Song, K. S. Jung, and Y. K. Shin. Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PloS one*, 10(12):e0144162, 2015. ISSN 1932-6203. doi:10.1371/journal.pone.0144162. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4671711{&}tool=pmcentrez{&}rendertype=abstract. → pages 25, 37, 90, 91, 95

[102] Q. Peng, R. Vijaya Satya, M. Lewis, P. Randad, and Y. Wang. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics*, 16(1):589, 2015. ISSN 1471-2164. doi:10.1186/s12864-015-1806-8. URL http://www.biomedcentral.com/1471-2164/16/589. → pages 93

[103] K. P. Pennington, T. Walsh, M. Lee, C. Pennil, A. P. Novetsky, K. J. Agnew, A. Thornton, R. Garcia, D. Mutch, M. C. King, P. Goodfellow, and E. M. Swisher. BRCA1, TP53, and CHEK2 germline mutations in uterine serous carcinoma. *Cancer*, 119(2):332–338, 2013. ISSN 0008543X. doi:10.1002/cncr.27720. → pages 64

[104] D. A. Pilger, P. L. Da Costa Lopez, F. Segal, and S. Leistner-Segal. Analysis of R213R and 13494 g???a polymorphisms of the p53 gene in individuals with esophagitis, intestinal metaplasia of the cardia and Barrett's Esophagus compared with a control group. *Genomic Medicine*, 1(1-2):57–63, 2007. ISSN 18717934. doi:10.1007/s11568-007-9007-4. → pages 65

[105] S. Quesnel, S. Verselis, C. Portwine, J. Garber, M. White, J. Feunteun, D. Malkin, and F. P. Li. p53 compound heterozygosity in a severely affected child with Li-Fraumeni syndrome. *Oncogene*, 18(27):3970–3978, 1999. ISSN 0950-9232. doi:10.1038/sj.onc.1202783. → pages 64

[106] V. M. Raymond, S. W. Gray, S. Roychowdhury, S. Joffe, A. M. Chinnaiyan, D. W. Parsons, and S. E. Plon. Germline findings in tumor-only sequencing: Points to consider for clinicians and laboratories. *Journal of the National Cancer Institute*, 108(4):1–5, 2016. ISSN 14602105. doi:10.1093/jnci/djv351. → pages 50, 51, 80, 86

[107] E. Rouits, V. Charasson, a. Pétain, M. Boisdron-Celle, J.-P. Delord, M. Fonck, a. Laurand, a L Poirier, a. Morel, E. Chatelut, J. Robert, and E. Gamelin. Pharmacokinetic and pharmacogenetic determinants of the activity and toxicity of irinotecan in metastatic colorectal cancer patients. *British journal of cancer*, 99:1239–45, 2008. ISSN 1532-1827. doi:10.1038/sj.bjc.6604673. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570505{&}tool=pmcentrez{&}rendertype=abstract. → pages 76

[108] J. D. Rowley. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, 243(5405):290–293, 1973. ISSN 0028-0836. doi:10.1038/243290a0. URL http://www.nature.com/doifinder/10.1038/243290a0. → pages 1

[109] A. Ruzzo, F. Graziano, F. Loupakis, E. Rulli, E. Canestrari, D. Santini, V. Catalano, R. Ficarelli, P. Maltese, R. Bisonni, G. Masi, G. Schiavon, P. Giordani, L. Giustini, A. Falcone, G. Tonini, R. Silva, R. Mattioli, I. Floriani, and M. Magnani. Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFOX-4 chemotherapy. *Journal of Clinical Oncology*, 25(10):1247–1254, 2007. ISSN 0732183X. doi:10.1200/JCO.2006.08.1844. → pages 73

[110] A. Ruzzo, F. Graziano, F. Loupakis, D. Santini, V. Catalano, R. Bisonni, R. Ficarelli, A. Fontana, F. Andreoni, A. Falcone, E. Canestrari, G. Tonini, D. Mari, P. Lippe, F. Pizzagalli, G. Schiavon, P. Alessandroni, L. Giustini, P. Maltese, E. Testa, E. T.

Menichetti, and M. Magnani. Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFIRI chemotherapy. *The Pharmacogenomics Journal*, 8(4):278–288, 2008. ISSN 1470-269X. doi:10.1038/sj.tpj.6500463. URL http://www.nature.com/doifinder/10.1038/sj.tpj.6500463. → pages 76

[111] V. N. Rykalina, A. A. Shadrin, V. S. Amstislavskiy, E. I. Rogaev, H. Lehrach, and T. A. Borodina. Exome sequencing from nanogram amounts of starting DNA: Comparing three approaches. *PLoS ONE*, 9(7), 2014. ISSN 19326203. doi:10.1371/journal.pone.0101154. → pages 1

[112] E. Samorodnitsky, B. M. Jewell, R. Hagopian, J. Miya, M. R. Wing, E. Lyon, S. Damodaran, D. Bhatt, J. W. Reeser, J. Datta, and S. Roychowdhury. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human Mutation*, 36(9): 903–914, 2015. ISSN 10981004. doi:10.1002/humu.22825. → pages 93

[113] K. A. Schrader, D. T. Cheng, V. Joseph, M. Prasad, M. Walsh, A. Zehir, A. Ni, T. Thomas, R. Benayed, A. Ashraf, A. Lincoln, M. Arcila, Z. Stadler, D. Solit, D. Hyman, L. Zhang, D. Klimstra, M. Ladanyi, K. Offit, M. Berger, and M. Robson. Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. *JAMA oncology*, 2(1):1–8, 2015. ISSN 2374-2445. doi:10.1001/jamaoncol.2015.5208. URL http://oncology.jamanetwork.com/article.aspx?articleid=2469517. → pages 12, 50, 51, 80, 90, 93, 95

[114] M. Schwab, U. M. Zanger, C. Marx, E. Schaeffeler, K. Klein, J. Dippon, R. Kerb, J. Blievernicht, J. Fischer, U. Hofmann, C. Bokemeyer, and M. Eichelbaum. Role of genetic and nongenetic factors for fluorouracil treatment-related severe toxicity: A prospective clinical trial by the German 5-FU toxicity study group. *Journal of Clinical Oncology*, 26 (13):2131–2138, 2008. ISSN 0732183X. doi:10.1200/JCO.2006.10.4182. → pages 70, 71, 73

[115] A. D. Seidman, D. Berry, C. Cirrincione, L. Harris, H. Muss, P. K. Marcom, G. Gipson, H. Burstein, D. Lake, C. L. Shapiro, P. Ungaro, L. Norton, E. Winer, and C. Hudis. Randomized phase III trial of weekly compared with every-3-weeks paclitaxel for metastatic breast cancer, with trastuzumab for all HER-2 overexpressors and random assignment to trastuzumab or not in HER-2 nonoverexpressors: final results of Cancer and Leu. *J Clin Oncol.*, 26(10):1642–9, 2008. → pages 1

[116] C. Seiler, A. Sharpe, J. C. Barrett, E. A. Harrington, E. V. Jones, and G. B. Marshall. Nucleic acid extraction from formalin-fixed paraffin-embedded cancer cell line samples: a trade off between quantity and quality? *BMC Clinical Pathology*, 16(1):17, 2016. ISSN 1472-6890. doi:10.1186/s12907-016-0039-3. URL http://dx.doi.org/10.1186/s12907-016-0039-3. → pages 45

[117] S.-R. Shi, R. J. Cote, L. Wu, C. Liu, R. Datar, Y. Shi, D. Liu, H. Lim, and C. R. Taylor. DNA extraction from archival formalin-fixed, paraffin-embedded tissue sections based on the antigen retrieval principle: heating under the influence of pH. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 50(8): 1005–1011, 2002. ISSN 0022-1554. doi:10.1177/002215540205000802. → pages 91

[118] J. A. Sikorsky, D. A. Primerano, T. W. Fenger, and J. Denvir. DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency. *Biochemical and Biophysical Research Communications*, 355(2):431–437, 2007. ISSN 0006291X. doi:10.1016/j.bbrc.2007.01.169. → pages 25, 90, 95

[119] S. Sjogren, M. Inganas, A. Lindgren, L. Holmberg, and J. Bergh. Prognostic and predictive value of c-erbB-2 overexpression in primary breast cancer, alone and in combination with other prognostic markers. *J Clin Oncol.*, 16:462–469, 1998. → pages 1

[120] E. H. Slager, M. W. Honders, E. D. V. D. Meijden, S. A. P. V. Luxemburg-heijs, F. M. Kloosterboer, M. G. D. Kester, I. Jedema, W. A. E. Marijt, M. R. Schaafsma, J. H. F. Falkenburg, E. H. Slager, M. W. Honders, E. D. V. D. Meijden, S. A. P. V. Luxemburg-heijs, and F. M. Kloosterboer. Identification of the angiogenic endothelial-cell growth factor-1 / thymidine phosphorylase as a potential target for immunotherapy of cancer Identification of the angiogenic endothelial-cell growth factor-1 / thymidine phosphorylase as a potential target. 107(12):4954–4960, 2013. doi:10.1182/blood-2005-09-3883. → pages 74

[121] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER- 2/neu Oncogene. *Science*, 235(4785):177–182, 1987. → pages 1

[122] A. So, A. Vilborg, Y. Bouhlal, R. T. Koheler, S. M. Grimes, D. Mendoza, F. Goodsaid, M. Lucero, F. M. De La Vega, and H. P. Ji. A Robust Targeted Sequencing Approach for Low Input and Variable Quality DNA from Clinical Samples. *bioRxiv*, 2017. doi:https://doi.org/10.1101/123117. → pages 1

[123] D. H. Spencer, J. K. Sehn, H. J. Abel, M. A. Watson, J. D. Pfeifer, and E. J. Duncavage. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *Journal of Molecular Diagnostics*, 15(5):623–633, 2013. ISSN 15251578. doi:10.1016/j.jmoldx.2013.05.004. URL http://dx.doi.org/10.1016/j.jmoldx.2013.05.004. → pages 91

[124] J. Stoehlmacher, D. J. Park, W. Zhang, D. Yang, S. Groshen, S. Zahedy, and H.-J. Lenz. A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. *British Journal of Cancer*, (May):344–354, 2004. ISSN 0007-0920. doi:10.1038/sj.bjc.6601975. URL http://www.nature.com/doifinder/10.1038/sj.bjc.6601975. → pages 73, 76

[125] S. P. Strom. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer biology & medicine*, 13(1):3–11, 2016. ISSN 2095-3941. doi:10.28092/j.issn.2095-3941.2016.0004. URL http://www.ncbi.nlm.nih.gov/pubmed/27144058{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4850126. → pages 94

[126] K. W. Suh, J. H. Kim, D. Y. Kim, Y. B. Kim, C. Lee, and S. Choi. Which Gene is a Dominant Predictor of Response During FOLFOX Chemotherapy for the Treatment of Metastatic Colorectal Cancer, the MTHFR or XRCC1 Gene? *Annals of Surgical Oncology*, 13(11):1379–1385, 2006. ISSN 1068-9265. doi:10.1245/s10434-006-9112-y. URL http://www.springerlink.com/index/10.1245/s10434-006-9112-y. → pages 73

[127] J. J. Swen, M. Nijenhuis, A. de Boer, L. Grandia, A. H. Maitland-van der Zee, H. Mulder, G. A. P. J. M. Rongen, R. H. N. van Schaik, T. Schalekamp, D. J. Touw, J. van der Weide, B. Wilffert, V. H. M. Deneer, and H.-J. Guchelaar. Pharmacogenetics: From Bench to Byte An Update of Guidelines. *Clinical Pharmacology & Therapeutics*, 89(5):662–673, 2011. ISSN 0009-9236. doi:10.1038/clpt.2011.34. URL http://doi.wiley.com/10.1038/clpt.2011.34. → pages 70, 71, 72

[128] D. Tanaka, A. Hishida, K. Matsuo, H. Iwata, M. Shinoda, Y. Yamamura, T. Kato, S. Hatooka, T. Mitsudomi, Y. Kagami, M. Ogura, K. Tajima, M. Suyama, M. Naito, K. Yamamoto, A. Tamakoshi, and N. Hamajima. Polymorphism of dihydropyrimidine dehydrogenase (DPYD) Cys29Arg and risk of six malignancies in Japanese. *Nagoya journal of medical science*, 67(Fig 1):117–124, 2005. ISSN 0027-7622. → pages 72

[129] R. Tian, M. K. Basu, and E. Capriotti. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC genomics*, 16 Suppl 8(Suppl 8):S7, 2015. ISSN 1471-2164. doi:10.1186/1471-2164-16-S8-S7. URL http://www.biomedcentral.com/1471-2164/16/S8/S7. → pages 91

[130] G. Toffoli, E. Cecchin, G. Corona, A. Russo, A. Buonadonna, M. D'Andrea, L. M. Pasetto, S. Pessa, D. Errante, V. De Pangher, M. Giusto, M. Medici, F. Gaion, P. Sandri, E. Galligioni, S. Bonura, M. Boccalon, P. Biason, and S. Frustaci. The role of UGT1A1*28 polymorphism in the pharmacodynamics and pharmacokinetics of irinotecan in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*, 24(19):3061–3068, 2006. ISSN 0732183X. doi:10.1200/JCO.2005.05.5400. → pages 76

[131] G. Toffoli, L. Giodini, A. Buonadonna, M. Berretta, A. De Paoli, S. Scalone, G. Miolo, E. Mini, S. Nobili, S. Lonardi, N. Pella, G. Lo Re, M. Montico, R. Roncato, E. Dreussi, S. Gagno, and E. Cecchin. Clinical validity of a DPYD-based pharmacogenetic test to predict severe toxicity to fluoropyrimidines. *International Journal of Cancer*, 137(12): 2971–2980, 2015. ISSN 10970215. doi:10.1002/ijc.29654. → pages 70, 71

[132] A. B. P. van Kuilenburg, J. Haasjes, D. J. Richel, L. Zoetekouw, H. V. Lenthe, R. A. De Abreu, J. G. Maring, P. Vreken, and A. H. Van Gennip. Clinical implications of dihydropyrimidine dehydrogenase (DPD) deficiency in patients with severe 5-fluorouracil-associated toxicity: Identification of new mutations in the DPD gene. *Clinical Cancer Research*, 6(12):4705–4712, 2000. ISSN 10780432. → pages 70, 72

[133] A. B. P. van Kuilenburg, J. Meijer, M. W. T. Tanck, D. Dobritzsch, L. Zoetekouw, L. L. Dekkers, J. Roelofsen, R. Meinsma, M. Wymenga, W. Kulik, B. Büchel, R. C. M. Hennekam, and C. R. Largiadèr. Phenotypic and clinical implications of variants in the dihydropyrimidine dehydrogenase gene. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1862(4):754–762, 2016. ISSN 1879260X. doi:10.1016/j.bbadis.2016.01.009. URL http://dx.doi.org/10.1016/j.bbadis.2016.01.009. → pages 50, 70, 71, 72

[134] F. J. Vega, P. Iniesta, T. Caldes, A. Sanchez, J. A. Lopez, C. de Juan, E. Diaz-Rubio, A. Torres, J. L. Balibrea, and M. Benito. P53 Exon 5 Mutations as a Prognostic Indicator of Shortened Survival in Non-Small-Cell Lung Cancer. *British journal of cancer*, 76(1):44–51, 1997. ISSN 0007-0920; 0007-0920. doi:Doi10.1038/Bjc.1997.334. → pages 65

[135] A. Villani, U. Tabori, J. Schiffman, A. Shlien, J. Beyene, H. Druker, A. Novokmet, J. Finlay, and D. Malkin. Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: A prospective observational study. *The Lancet Oncology*, 12 (6):559–567, 2011. ISSN 14702045. doi:10.1016/S1470-2045(11)70119-X. → pages 64

[136] V. Vincek, M. Nassiri, M. Nadji, and A. R. Morales. A Tissue Fixative that Protects Macromolecules (DNA, RNA, and Protein) and Histomorphology in Clinical Samples. *Laboratory Investigation*, 83(10):1427–1435, 2003. ISSN 0023-6837. doi:10.1097/01.LAB.0000090154.55436.D1. URL http://www.nature.com/doifinder/10.1097/01.LAB.0000090154.55436.D1. → pages 95

[137] C. L. Vogel, M. A. Cobleigh, D. Tripathy, J. C. Gutheil, L. N. Harris, L. Fehrenbacher, D. J. Slamon, M. Murphy, W. F. Novotny, M. Burchmore, S. Shak, S. J. Stewart, and M. Press. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol*, 20(3):719–26, 2002. → pages 1

[138] D. von Hansemann. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchows Arch. Path. Anat.*, 119(2):299–326, 1890. → pages 1

[139] H. Wang and F. Meng. Theoretical study of proton-catalyzed hydrolytic deamination mechanism of adenine. *Theoretical Chemistry Accounts*, 127(5):561–571, 2010. ISSN 1432881X. doi:10.1007/s00214-010-0747-1. → pages 92

[140] Y. Wang, W. Bao, H. Shi, C. Jiang, and Y. Zhang. Epidermal growth factor receptor exon 20 mutation increased in post-chemotherapy patients with non-small cell lung cancer detected with patients' blood samples. *Translational oncology*, 6(4):504–10, 2013. ISSN 1936-5233. doi:10.1593/tlo.13391. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3730025{&}tool=pmcentrez{&}rendertype=abstract. → pages 58, 59

[141] K. Wetterstrand. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP), 2016. URL https://www.genome.gov/27541954/dna-sequencing-costs-data/. → pages 1

[142] C. Wong, R. A. DiCioccio, H. J. Allen, B. A. Werness, and M. S. Piver. Mutations in BRCA1 from fixed, paraffin-embedded tissue can be artifacts of preservation. *Cancer Genetics and Cytogenetics*, 107(1):21–27, 1998. ISSN 01654608. doi:10.1016/S0165-4608(98)00079-X. → pages 92

[143] N. A. Wong, D. Gonzalez, M. Salto-Tellez, R. Butler, S. J. Diaz-Cano, M. Ilyas, W. Newman, E. Shaw, P. Taniere, and S. V. Walsh. RAS testing of colorectal carcinoma-a guidance document from the Association of Clinical Pathologists Molecular Pathology and Diagnostics Group. *Journal of clinical pathology*, 67(9):751–757, 2014. ISSN 1472-4146. doi:10.1136/jclinpath-2014-202467. URL http://www.ncbi.nlm.nih.gov/pubmed/24996433. → pages 50

[144] S. Q. Wong, J. Li, R. Salemi, K. E. Sheppard, H. Do, R. W. Tothill, G. a. McArthur, and A. Dobrovic. Targeted-capture massively-parallel sequencing enables robust detection of

clinically informative mutations from formalin-fixed tumours. *Scientific reports*, 3(3):3494, 2013. ISSN 2045-2322. doi:10.1038/srep03494. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3861801{&}tool=pmcentrez{&}rendertype=abstract. → pages 25, 90, 91, 93, 95

[145] S. Q. Wong, J. Li, A. Y-C Tan, R. Vedururu, J.-M. B. Pang, H. Do, J. Ellul, K. Doig, A. Bell, G. A. MacArthur, S. B. Fox, D. M. Thomas, A. Fellowes, J. P. Parisot, and A. Dobrovic. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Medical Genomics*, 7(1):1–10, 2014. ISSN 1755-8794. doi:10.1186/1755-8794-7-23. URL http://www.biomedcentral.com/1755-8794/7/23. → pages 4, 13, 25, 37, 45, 90, 91, 93, 95

[146] C. Xu, M. Nezami Ranjbar, Z. Wu, J. DiCarlo, and Y. Wang. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. *BMC Genomics*, 18(1):1–11, 2017. ISSN 1471-2164. doi:10.1186/s12864-016-3425-4. URL http://dx.doi.org/10.1186/s12864-016-3425-4. → pages 91

[147] V. K. Yadav and S. De. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in Bioinformatics*, 16(2):232–241, 2015. ISSN 14774054. doi:10.1093/bib/bbu002. → pages 94

[148] M. Yang, Y. Guo, X. Zhang, X. Miao, W. Tan, T. Sun, D. Zhao, D. Yu, J. Liu, and D. Lin. Interaction of P53 Arg72Pro and MDM2 T309G polymorphisms and their associations with risk of gastric cardia cancer. *Carcinogenesis*, 28(9):1996–2001, 2007. ISSN 01433334. doi:10.1093/carcin/bgm168. → pages 66

[149] S. S. Yea, S. S. Lee, W.-Y. Kim, K.-H. Liu, H. Kim, J.-H. Shon, I.-J. Cha, and J.-G. Shin. Genetic variations and haplotypes of UDP-glucuronosyltransferase 1A locus in a Korean population. *Therapeutic drug monitoring*, 30(1):23–34, 2008. ISSN 0163-4356. doi:10.1097/FTD.0b013e3181633824. URL http://www.ncbi.nlm.nih.gov/pubmed/18223459. → pages 76

[150] T. Yoneda, A. Kuboyama, K. Kato, T. Ohgami, K. Okamoto, T. Saito, and N. Wake. Association of MDM2 SNP309 and TP53 Arg72Pro polymorphisms with risk of endometrial cancer. *Oncology Reports*, 30(1):25–34, 2013. ISSN 1021335X. doi:10.3892/or.2013.2433. → pages 66

[151] Y. Zha, P. Gan, Q. Liu, and Q. Yao. TP53 Codon 72 Polymorphism Predicts Efficacy of Paclitaxel Plus Capecitabine Chemotherapy in Advanced Gastric Cancer Patients. *Archives of Medical Research*, 47(1):13–18, 2016. ISSN 01884409. doi:10.1016/j.arcmed.2015.12.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0188440915002854. → pages 66

[152] W. Zhang, L. P. Stabile, P. Keohavong, M. Romkes, J. R. Grandis, A. M. Traynor, and J. M. Siegfried. Mutation and polymorphism in the EGFR-TK domain associated with lung cancer. *Journal of thoracic oncology : official publication of the International Association*

for the Study of Lung Cancer, 1(7):635–47, 2006. ISSN 1556-1380. doi:01243894-200609000-00007[pii]. URL http://www.ncbi.nlm.nih.gov/pubmed/17409930. → pages 58

[153] X. Zhang, G. Ao, Y. Wang, W. Yan, M. Wang, E. Chen, F. Yang, and J. Yang. Genetic variants and haplotypes of the UGT1A9, 1A7 and 1A1 genes in Chinese Han. *Genetics and Molecular Biology*, 35(2):428–434, 2012. ISSN 14154757. doi:10.1590/S1415-47572012005000036. → pages 76

[154] Y. Zhang, L. Liu, Y. Tang, C. Chen, Q. Wang, J. Xu, C. Yang, X. Miao, S. Wei, J. Chen, and S. Nie. Polymorphisms in TP53 and MDM2 contribute to higher risk of colorectal cancer in Chinese population: A hospital-based, case-control study. *Molecular Biology Reports*, 39 (10):9661–9668, 2012. ISSN 03014851. doi:10.1007/s11033-012-1831-5. → pages 66

[155] Z. Z. Zhu, A. Z. Wang, H. R. Jia, X. X. Jin, X. L. He, L. F. Hou, and G. Zhu. Association of the TP53 codon 72 polymorphism with colorectal cancer in a Chinese population. *Japanese Journal of Clinical Oncology*, 37(5):385–390, 2007. ISSN 03682811. doi:10.1093/jjco/hym034. → pages 66

[156] J. Zining, X. Lu, H. Caiyun, and Y. Yuan. Genetic polymorphisms of mTOR and cancer risk: a systematic review and updated meta-analysis. *Oncotarget*, 7(35), 2016. ISSN 1949-2553 (Electronic). doi:10.18632/oncotarget.10805. → pages 60, 61

Appendix A

Supporting Materials

Table A.1: Target regions and amplicons of the OncoPanel.

Gene	Target	Target Region	Amplicon	Length (bp)	GC Content (%)
ALK					
DPYD					
EGFR					
GSTP1					
KIT					
MAPK1					
MTHFR					
MTOR					
PDGFRA					
STAT1					
STAT3					
TP53					
TYMP					
TYMS					
UGT1A1					