

**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)

June 2017

© Shyong Quin Yap, 2017

Abstract

This document provides brief instructions for using the `ubcdiss` class to write a **UBC!**-conformant dissertation in **L^AT_EX**. This document is itself written using the `ubcdiss` class and is intended to serve as an example of writing a dissertation in **L^AT_EX**. 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

Abstract	ii
Preface	iii
Table of Contents	iv
List of Tables	vi
List of Figures	viii
List of Abbreviations	xii
Acknowledgments	xiii
1 Introduction	1
1.1 Cancer as a Genetic Disease	1
1.2 The Evolution of Molecular Diagnostics in Cancer	2
1.3 The Era of Precision Oncology	4
1.4 Next-generation Sequencing Technologies	4
1.5 Applications of Next-generation Sequencing	4
1.5.1 Targeted Sequencing	4
1.5.2 Whole Exome Sequencing	4
1.5.3 Whole Genome Sequencing	4
1.6 Variant Calling Pipeline	4
1.7 Germline Variants in The Tumour Genome	4
1.7.1 Incidental Findings	4
1.7.2 Pharmacogenomic Variants	5
1.7.3 Challenges	5
1.8 ACCE Model Process for Evaluating Genetic Tests	6
1.9 Objectives	6

2 Materials and Methods	7
2.1 Overview of study design	7
2.2 Patient samples	7
2.3 Sample preparation, library construction, and Illumina sequencing	8
2.4 OncoPanel (Amplicon-based targeted sequencing panel for solid tumours)	9
2.5 Variant calling pipeline	10
2.5.1 Read alignment and variant calling	10
2.5.2 Variant filtering	11
2.5.3 Variant annotation and interpretation	11
2.6 Sequence analysis	15
2.7 Application of VAF thresholds to separate germline alterations from somatic mutations	15
3 Assessment of Formalin-Induced DNA Damage in FFPE Specimens	17
3.1 Comparison of efficiency in amplicon enrichment and sequencing results between blood and FFPE specimens	17
3.2 Reduced coverage depth in FFPE specimens is more pronounced for longer amplicons	24
3.3 Deamination effects lead to increased C>T/G>A transitions in FFPE specimens	29
3.4 Increased age of paraffin block results in reduced amplicon yield and elevated level of C>T/G>A sequence artifacts	37
4 Identification of Germline Alterations in FFPE Tumours	42
4.1 Frequency and variant assessment of germline alterations in patients from TOP cohort	44
4.2 Germline alterations are highly concordant between blood and FFPE specimens	71
4.3 Application of tumour content to separate germline alterations from somatic mutations in tumour-only analyses	77
5 Discussion	81
6 Conclusion	87
Bibliography	88
A Supporting Materials	104

List of Tables

Table 2.1	Distribution of cancer types in the TOP cohort.	8
Table 2.2	Gene reference models for HGVS nomenclature of OncoPanel genes.	9
Table 2.3	Potential risk alleles in the hg19 human reference genome within the target regions of the OncoPanel.	10
Table 2.4	Thresholds for parameters of VarScan2 <code>fpfilter</code> used for filtering raw variant output.	12
Table 2.5	Spurious variants removed by the variant filtering pipeline.	13
Table 3.1	Comparison of coverage uniformity between blood and FFPE specimens using the Wilcoxon signed-rank test.	23
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 _{FFPE} /Median Coverage _{Blood})) based on amplicon length and GC content.	28
Table 3.3	Summary statistics of fraction of base changes in blood and FFPE specimens. .	32
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	33
Table 3.5	Summary statistics of fraction of base changes in blood and FFPE specimens within 1-10% allele frequency.	36
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 ρ and 95% confidence interval in brackets, whereas bottom values represent p -value. Asterisk(*) indicates significance level of p -value < 0.05	41

Table 4.1	Frequency of germline variants in cancer-related genes in blood specimens from TOP patients.	46
Table 4.2	Variant assessment of germline alterations in cancer-related genes detected in blood specimens of TOP patients.	50
Table 4.3	Frequency of germline variants in pharmacogenomic genes detected in blood specimens of TOP patients.	60
Table 4.4	Variant assessment of germline alterations in pharmacogenomic genes detected in blood specimens of TOP patients.	62
Table 4.5	Distribution of discordant germline alterations identified in patients from TOP cohort.	73
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.	79
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.	80
Table A.1	Target regions and amplicons of the OncoPanel.	104

List of Figures

Figure 1.1	6	
Figure 2.1	Pipelines for (A) variant calling and (B) filtering.	14
Figure 2.2	2x2 contingency table for determination of true positive, false positive, true negative, and false negative variant calls in tumour-only analyses.	16
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.	20
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.	21
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.	22

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 ($\log_2 (\text{Median Coverage}_{\text{FFPE}}/\text{Median Coverage}_{\text{Blood}})$) for amplicons in the panel. Negative \log_2 fold change indicates lower coverage depth of the amplicon in FFPE specimens relative to blood ($\downarrow \text{Coverage}_{\text{FFPE}}$), whereas positive \log_2 fold change indicates higher coverage depth of the amplicon in FFPE specimens relative to blood ($\uparrow \text{Coverage}_{\text{FFPE}}$). N = number of amplicons; ns = not significant	26
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.	27
Figure 3.6	Scatter plots showing \log_2 fold change between amplicon coverage depth in blood and FFPE specimens ($\log_2 (\text{Median Coverage}_{\text{FFPE}}/\text{Median Coverage}_{\text{Blood}})$) 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.	28
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.	31
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 ($\log_2 (\text{Fraction of Base Changes}_{\text{FFPE}}/\text{Fraction of Base Changes}_{\text{Blood}})$). 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.	32

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.	35
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.	39
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).	40
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.	69
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.	70
Figure 4.3	Venn diagram demonstrating concordance of variants identified in 217 tumour-blood paired samples.	72

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.	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.	80

List of Abbreviations

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 Cancer as a Genetic Disease

Cancers are diseases defined by unrestrained proliferation of cells that are capable of invading normal tissues and metastasizing to other parts of the body. Early studies between the late nineteenth and early twentieth centuries by David von Hansemann and Theodor Boveri suggested that genetic alterations may contribute to oncogenesis. In von Hansemann's analysis of tumour samples, he observed features of aberrant cell division, which he speculated to be contributing factors of unequal distribution of chromosomes in the tumour cells. Boveri explored the connection between defective cell division and tumour formation by inducing abnormal chromosome segregation in sea urchin eggs and observing the outcome of these cells. While most cases of chromosomal imbalance resulted in cell death, Boveri reported that there were cases in which cell survival was followed by uncontrolled cell growth. These findings led Boveri to surmise that the improper combination of genetic materials could sustain the proliferative ability of tumour cells. In particular, tumour cells were likely to retain chromatin parts with growth stimulatory effects or remove those with growth inhibitory effects. Boveri's speculations pertaining genetic materials that function as stimulators or inhibitors of cell growth were consistent with the modern understanding of oncogenes and tumour-suppressor genes, respectively.

Oncogenes and tumour-suppressor genes are two categories of cancer-causing genes that play a central role in cancer initiation and progression. Prototype oncogenes (proto-oncogenes), the normal counterparts of oncogenes, encode proteins that promote cell growth and survival, as well as inhibit cell differentiation. When proto-oncogenes sustain dominant gain-of-function mutations, they become oncogenes, giving rise to constitutively active or overexpressed protein products that induce malignant transformation of the cells. The first oncogene, *v-src*, was identified by Peyton Rous in a retrovirus that causes sarcoma in chickens, and the virus was later named the Rous sarcoma virus after its discoverer. Subsequently, the proto-oncogene *c-src*, a homologue of *v-src*, was discovered,

but unlike its mutant form, the protein encoded by *c-src* was not constitutively active. The discovery that proto-oncogenes exist in healthy cells led to the recognition that normal cellular genes are capable of gaining oncogenic potential through acquiring gene mutations. This breakthrough in cancer research consequently catalyzed the identification of more proto-oncogenes, providing an enhanced understanding of cellular signalling pathways.

Tumour-suppressor genes, the other important type of cancer-causing genes, can be separated into gatekeepers and caretakers. Gatekeepers are involved in regulating cell-cycle checkpoints and mutations in gatekeepers would directly result in cancer development. On the other hand, caretakers are DNA repair proteins, and mutant caretakers could indirectly cause malignant transformation of cells by inducing accumulation of mutations, thereby increasing the probability that mutations would occur in oncogenes and tumour-suppressor genes. Unlike oncogenes, mutations in tumour-suppressor genes are recessive-acting, meaning that two mutant alleles are required for the tumour-suppressor gene to become oncogenic. Alfred Knudson was the first to propose that mutant tumor-suppressor genes function in a recessive fashion, a notable concept later known as Knudson's two-hit hypothesis. Knudson's statistical model demonstrated that familial retinoblastoma, a pediatric eye cancer, was consistent with a one-hit curve, meaning that a single mutation was sufficient to cause tumour formation. Conversely, non-familial retinoblastoma was consistent with a two-hit curve, meaning that two mutations were involved in the disease. These findings implied that disease carriers, who inherited one mutant allele, only require the loss of the remaining functional allele to drive tumour formation, whereas non-carriers require two mutation events to trigger the development of tumours.

Some bullshit about stepwise progression...

1.2 The Evolution of Molecular Diagnostics in Cancer

Although early studies have implicated the causative role of genetic alterations in cancer development, initial classification of cancers were based on the primary site of the tumour. This long-standing classification was caused by limitations in technologies and tools, as well as by the clinical classification required by surgical management, which was the initial mainstay of clinical oncology. Subsequently, microscopy-based classification of disease further delineated cancer subsets based on histologic differences. For example, aggressiveness or risk of relapse was retrospectively linked to histologic grading as a prognostic biomarker, such as Gleason and Bloom-Richardson for prostate and breast cancer, respectively. The histologic classification was advanced with assessment of prototypic surface markers (immunohistochemistry), gross markers for lymphoid subsets have heavily influenced the diagnostic classification of lymphoma. Characterization of chromosomal abnormalities in leukemia and sarcoma has aided in the diagnosis of disease subsets as well as prognostic assessment.

In conclusion, the diagnosis of cancer has undergone a paradigm shift. No longer is cancer

diagnosed only based on morphological parameters. More and more the diagnostic algorithm is supported by immunohistochemical and molecular alterations at the DNA, mRNAs, miRNAs and proteomic level. Multiple platforms and high throughput technological advances enable faster and cheaper analysis of all these as well as the whole genome. This is having a significant impact on how medicine is now being practiced in a personalized approach leading to the development of precision medicine based on pharmacogenomics. It is being realized that a tumor may not be characterized by a single gene alteration but by a panel of signature genomic alterations leading to targeted therapeutic strategies and surveillance based on the tumor specific alterations. The ultimate goal of cancer diagnosis in personalized medicine would be to identify the correct diagnosis and guide the therapy so that every patient receives precision medicine that is the right drug at the right dose.

Cancer is a group of diseases characterized by uncontrolled proliferation of cells that are capable of normal tissue invasion and metastasis to distant organs.

Oncogenes encode for proteins that stimulate cell proliferation and survival as well as inhibit cell differentiation leading to oncogenesis. The first human oncogene was identified based on homology with

Mutations in oncogenes are typically dominant, which means . The first human oncogene was discovered Conversely, tumour-suppressor genes encode for proteins that inhibit cell proliferation and survival as well as stimulate cell differentiation. Mutations in oncogenes are typically dominant whereas mutations in tumour-suppressor genes are recessive.

In the normal cell, proto-oncogenes stimulate proliferation and inhibit differentiation and apoptosis while the opposite is true for tumour suppressors. Proto-oncogenes are usually dominant, meaning that only one gain-of-function mutation is required to activate the oncogene, thereby causing cancer. Conversely, tumour suppressor genes are usually recessive and two loss-of-function events are required.

Boveri and von Hansemann (1890-1914) - oncogenes and tumour suppressors Philadelphia translocation (1960) Two hit hypothesis, retinoblastoma (1971) - germline and somatic mutations HRAS point mutation (1982) Eric Fearon and Bert Vogelstein find specific sequential mutations in carcinoma (1990) - multi-step process, caretakers and gatekeepers Types of mutations/gene changes - SNVs, indels, SVs Driver vs. passenger mutations - evolutionary process, selective growth advantage, CSCs Frequency and pathway-based: three main pathways

The pathogenesis of cancer is caused by genetic abnormalities Although fundamentally known to arise from genetic mutations, the disease paradigm has expanded to include aberrant epigenetic mechanisms as a contributing factor to oncogenesis. The understanding of cancer pathogenesis has expanded been increasing over the years and a disorder that was fundamentally known to arise from genetic mutations this group of disorders which have been fundamentally known Cancer has been fundamentally known as a genetic disease defined by abnormal proliferation of cells. Our

understanding of cancer pathogenesis has been expanding. Although the understanding of cancer pathogenesis has been expanding, Cancer has been fundamentally known as a genetic disease.

1.3 The Era of Precision Oncology

Cancer diagnosis and treatment have been revolutionized by advances in next-generation sequencing and bioinformatics tools, which contributed to an enhanced understanding of the mutational landscapes of various cancers. In the era of precision oncology, tumours can be sequenced, somatic mutations can be detected, and patients can be treated with targeted therapies or referred to on-going clinical trials. However, the tumour genome also consists of germline information that may have clinical implications for patients and their families.

1.4 Next-generation Sequencing Technologies

Next-generation sequencing

1.5 Applications of Next-generation Sequencing

1.5.1 Targeted Sequencing

Capture-based, amplicon-based etc.

1.5.2 Whole Exome Sequencing

1.5.3 Whole Genome Sequencing

1.6 Variant Calling Pipeline

1.7 Germline Variants in The Tumour Genome

1.7.1 Incidental Findings

The application of next-generation sequencing (NGS) technologies for tumour profiling has been increasingly integrated into oncologic care to detect targetable somatic mutations and personalize treatments for cancer patients. Although analysis of tumour-normal paired samples is required to accurately discriminate between somatic and germline variants, most clinical laboratories only sequence tumour samples to minimize cost and turnaround time [87]. However, genomic analyses of tumours can also reveal secondary genomic findings, which are germline information that may have clinical implications for patients and their family members [87]. In fact, several studies

demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [75, 87, 91?]. Therefore, clinical laboratories providing tumour genomic testing must be equipped to perform germline confirmatory testing on potential germline variants or be prepared to refer such cases to external services.

1.7.2 Pharmacogenomic Variants

MMQS higher means more mismatches in the supporting reads Because the tumour genome contains germline information, clinical laboratories can leverage tumour genomic testing to perform initial screening for clinically relevant germline variants such as variants in pharmacogenomic (PGx) genes. Subsequently, a similar framework for validating secondary germline findings can be applied, in which only patients with potential germline PGx variants are subjected to downstream germline testing. This procedure for germline PGx testing is more cost-effective because it does not require processing, sequencing, and analysis of normal DNA for every patient. The ability to implement germline PGx testing at a reduced cost can significantly benefit patient care because these variants cause functional changes in drug targets and drug disposition proteins (proteins involved in drug metabolism and transport), thereby contributing to inter-patient differences in chemotherapeutic response [73]. Hence, such genomic information can be used to guide the selection of chemotherapeutic drugs and optimization of drug dosage for cancer patients, leading to improved safety and efficacy of treatment and reduced risk of toxicity [73].

1.7.3 Challenges

Detection of genomic alterations in tumour DNA is also faced with technical challenges conferred by formalin-fixed paraffin-embedded (FFPE) tumour specimens [111?]. Tumour biopsies are often formalin-fixed to preserve tissue morphology for histological examination and to enable storage at room temperature; however, formalin fixation causes DNA fragmentation and base modifications, which pose difficulties in using DNA extracted from FFPE tumours for clinical genomic testing [111?]. Fragmentation damage caused by formalin fixation leads to reduced template DNA for PCR amplification, thereby affecting the efficiency of amplicon-based NGS testing [111?]. Furthermore, the degree of DNA fragmentation was shown to be higher in tissues from older FFPE blocks and tissues fixed with formalin of lower pH [?]. Formalin fixation is also problematic because it gives rise to depurination, which generates abasic sites, and cytosine deamination resulting in C>T/G>A transitions [?]. These forms of formalin-induced DNA damage contribute to the presence of sequence artifacts in FFPE specimens, which can be inaccurately identified as real genomic alterations.

1.8 ACCE Model Process for Evaluating Genetic Tests

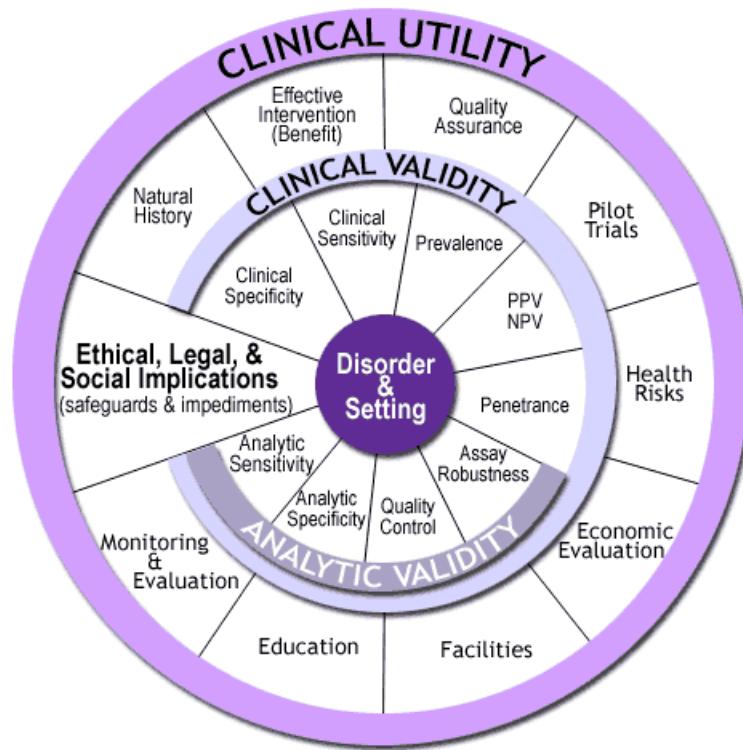


Figure 1.1

1.9 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 the concordance of germline alterations between blood and 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 [15] 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 hotspots 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
<i>Cancer-related</i>		
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
<i>Pharmacogenomics</i>		
DPYD	Dihydropyrimidine dehydrogenase	NM_000110.3
GSTP1	Glutathione S-transferase 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 ≥ 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 ≥ 20 , and fraction of variant reads from each strand ≥ 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
TP53	chr1	11288758	C	rs1064261	p.Asn999Asn c.2997T>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 `fppfilter` function with fraction of variant reads from each strand ≥ 0.1 and default thresholds for other parameters (Table 2.4). The VarScan2 `fppfilter` 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 `fppfilter`. 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 `fppfilter` and strand filter. Manual inspection was performed for a subset of variants, including variants detected within primer regions and in PGx genes, using the Integrative 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

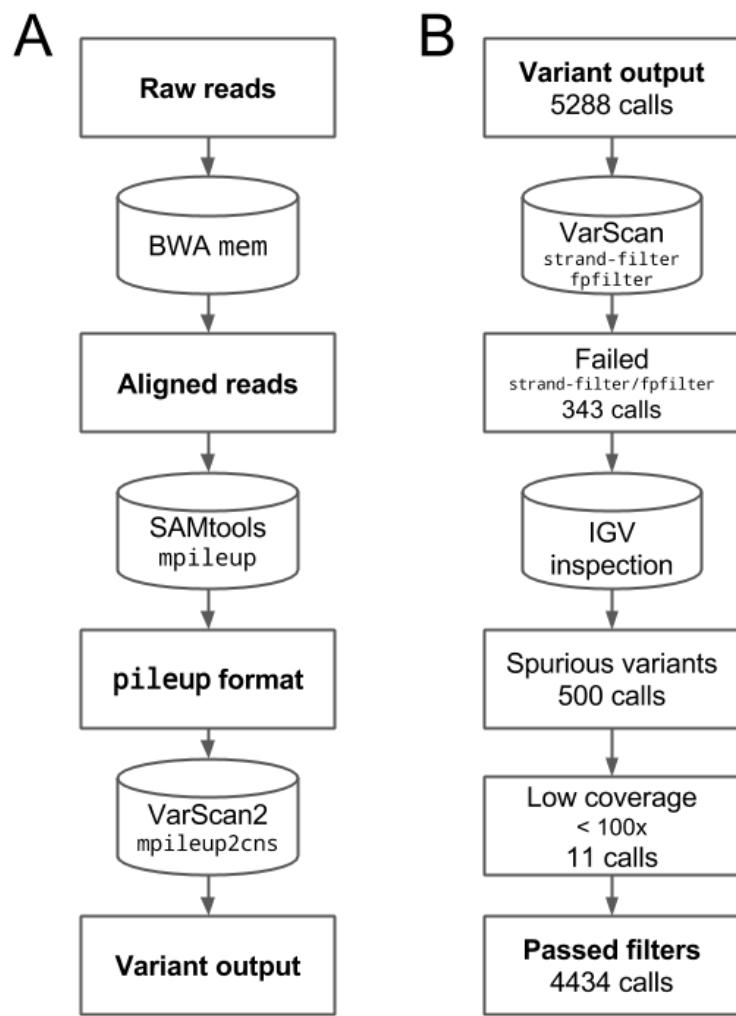


Figure 2.1: Pipelines for (A) variant calling and (B) filtering.

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.

		Predicted variant status	
		Germline	Somatic
Detection in matched blood	Present	True positive	False negative
	Absent	False positive	True negative

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 [67]. 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 [32, 60, 82, 83, 95, 110, 111]. 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 [30, 32, 110, 111]. 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 log₂ 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.

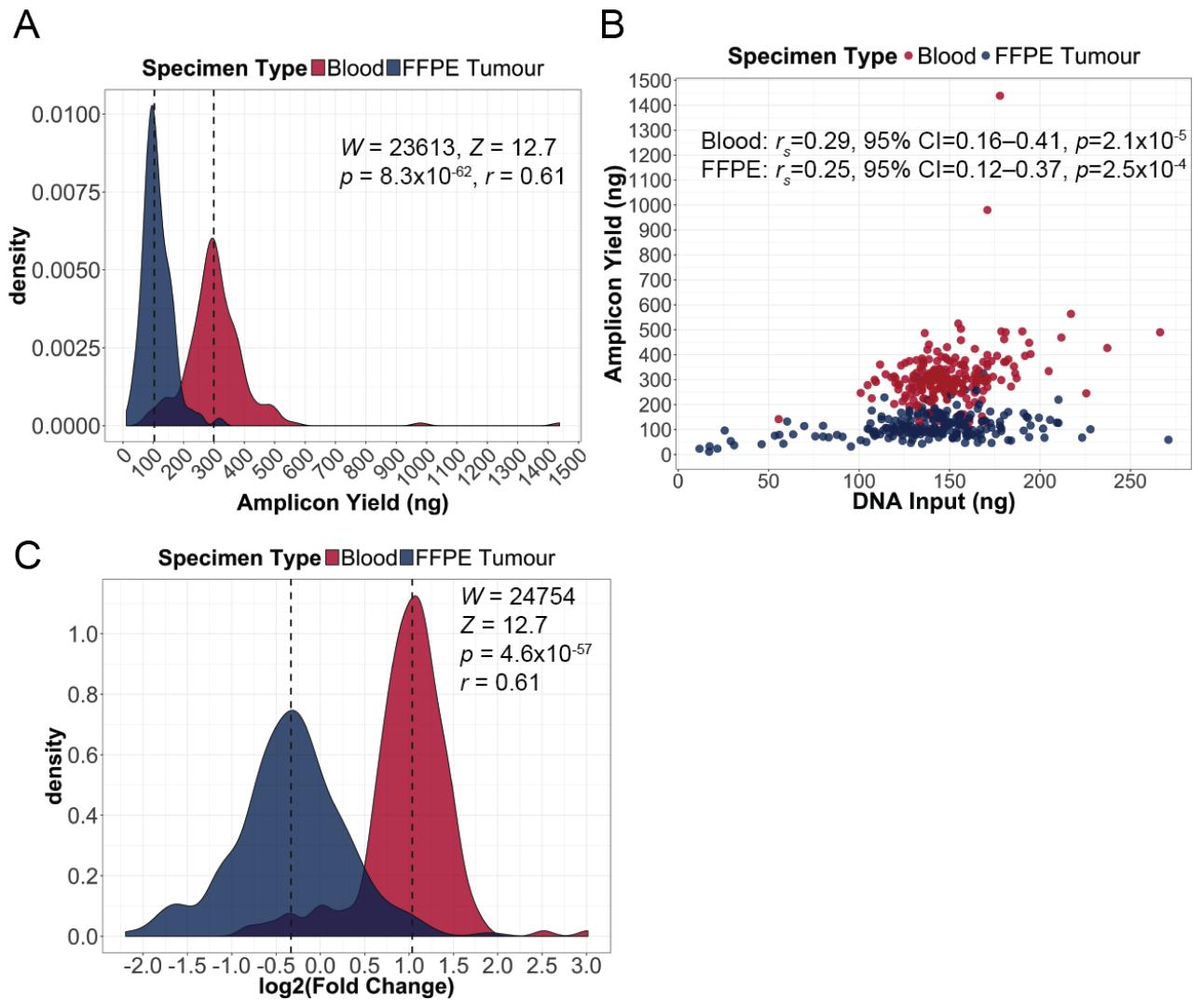


Figure 3.1: Comparison of efficiency in amplicon enrichment between blood and FFPE specimens. (A) Distributions of amplicon yield in blood and FFPE specimens (Wilcoxon signed-rank test). Dashed lines indicate median amplicon yield in blood and FFPE specimens, which are 299.3 ng and 103.6 ng, respectively. (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.

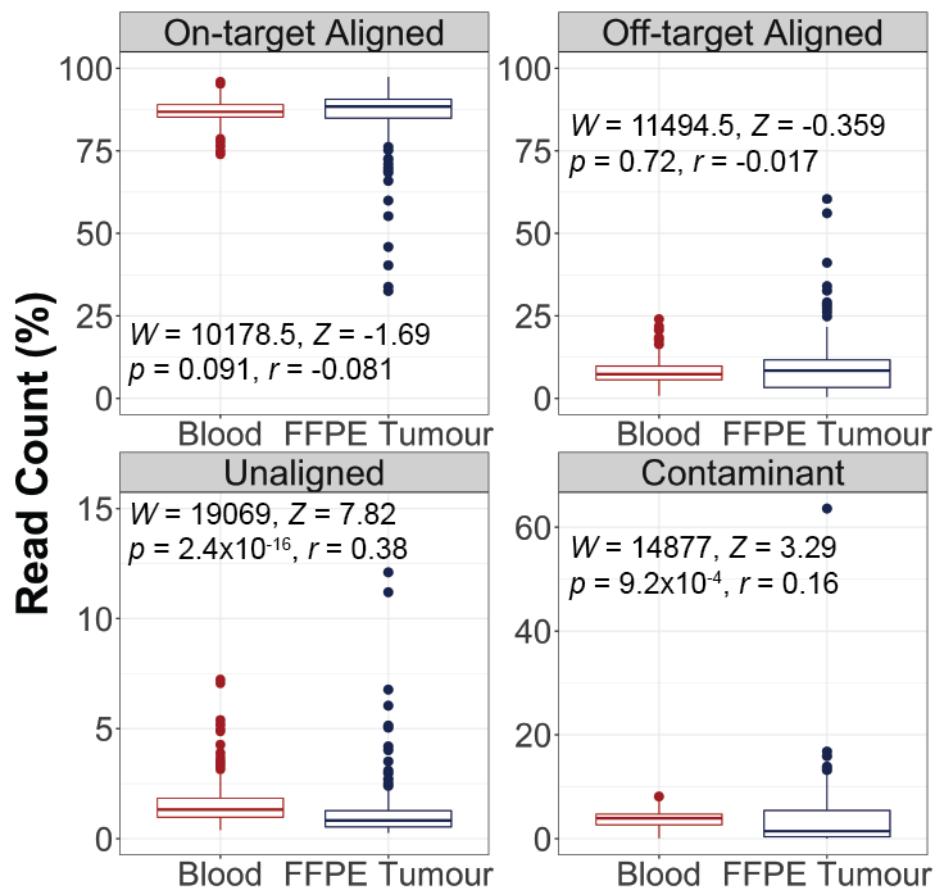


Figure 3.2: Assessment of read alignments between blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and interquartile range (IQR) of percentage of reads, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating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

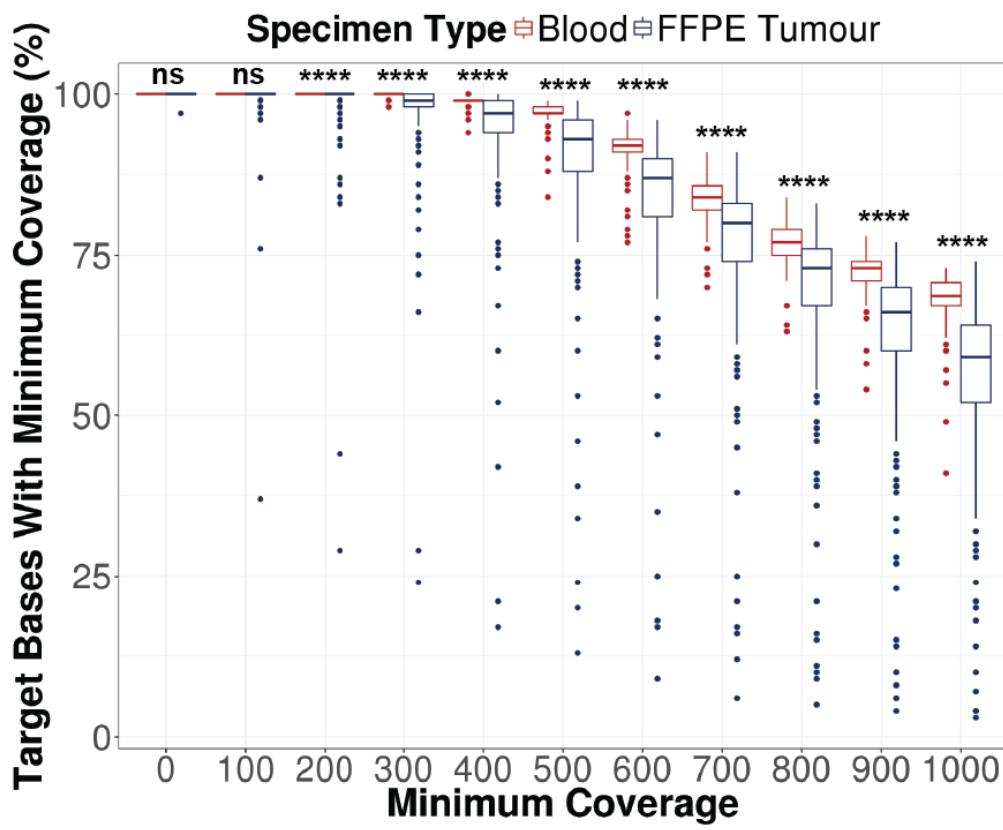


Figure 3.3: Evaluation of coverage uniformity in blood and FFPE specimens (Wilcoxon signed-rank test, *** $p < 0.0001$, ns = not significant). Per base coverage was normalized to account for difference in library size. Percentage of target bases that met various coverage thresholds was calculated. Box 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.

Threshold	Blood		FFPE Tumour		D^{\dagger} (%)	$p (< 0.0001^*)$
	Median (%)	Range (%)	Median (%)	Range (%)		
$\geq 0x$	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 hotspots 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 ($\log_2(\text{Median Coverage}_{\text{FFPE}}/\text{Median Coverage}_{\text{Blood}})$) 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

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

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.

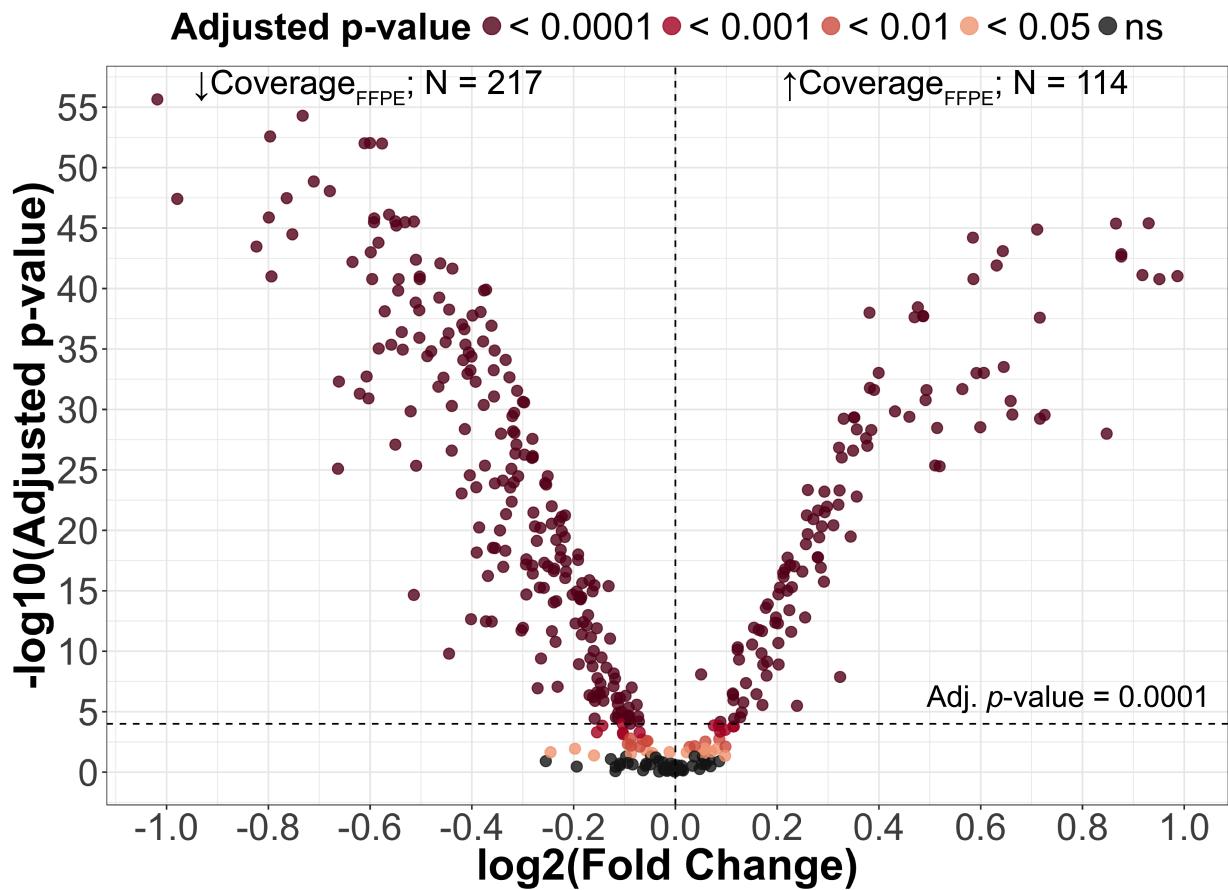


Figure 3.4: Amplicon-specific differences in coverage depth between blood and FFPE specimens. Difference in amplicon coverage depth between specimen types was determined using the Wilcoxon signed-rank test with Benjamini-Hochberg correction (adjusted $p < 0.0001$). Volcano plot illustrates the $-\log_{10}$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

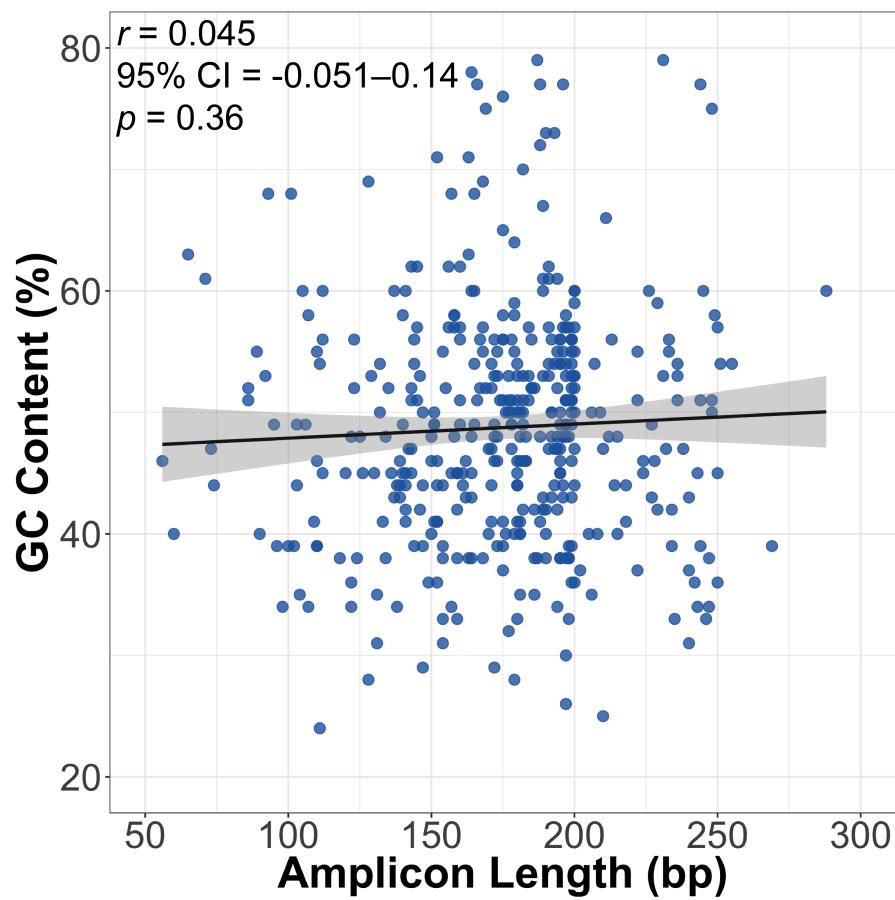


Figure 3.5: The relationship between amplicon GC content and amplicon length (Pearson's correlation). Solid line represents the fitted linear relationship between the two variables, and the shaded band indicates pointwise 95% confidence interval of the fitted linear regression line.

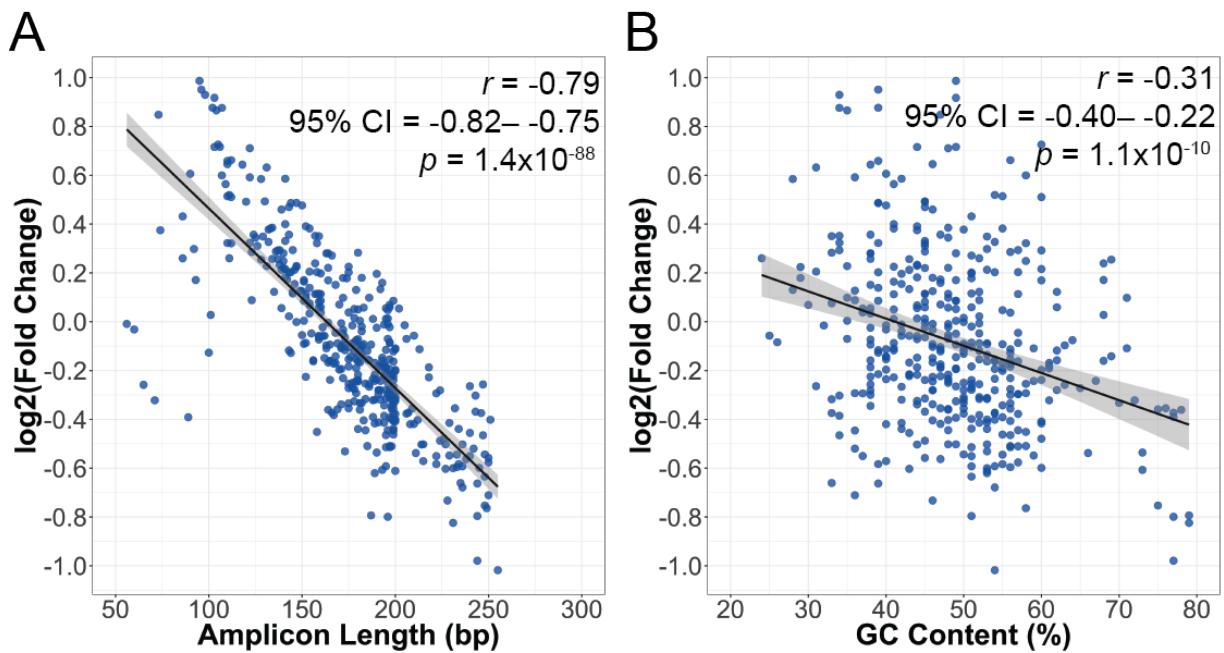


Figure 3.6: Scatter plots showing \log_2 fold change between amplicon coverage depth in blood and FFPE specimens (\log_2 (Median Coverage_{FFPE}/Median Coverage_{Blood})) in relation to (A) amplicon length and (B) GC content (Pearson's correlation). Solid line represents the fitted linear relationship 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_{FFPE}/Median Coverage_{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 [31–33, 50, 60, 82, 83, 111]. A prominent type of formalin-induced sequence artifact is C>T/G>A transitions as a result of deamination of cytosine bases [32, 60, 65, 83, 111]. 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 ($\text{BAQ} \geq 20$) 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 [48]. In the presence of atmospheric oxygen, formaldehyde can be oxidized into formic acid, causing depurination, which gives rise to abasic sites [32].

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 ($\log_2(\text{Fraction of Base Changes}_{\text{FFPE}}/\text{Fraction of Base Changes}_{\text{Blood}})$). 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 $\text{BAQ} \geq 20$ 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.

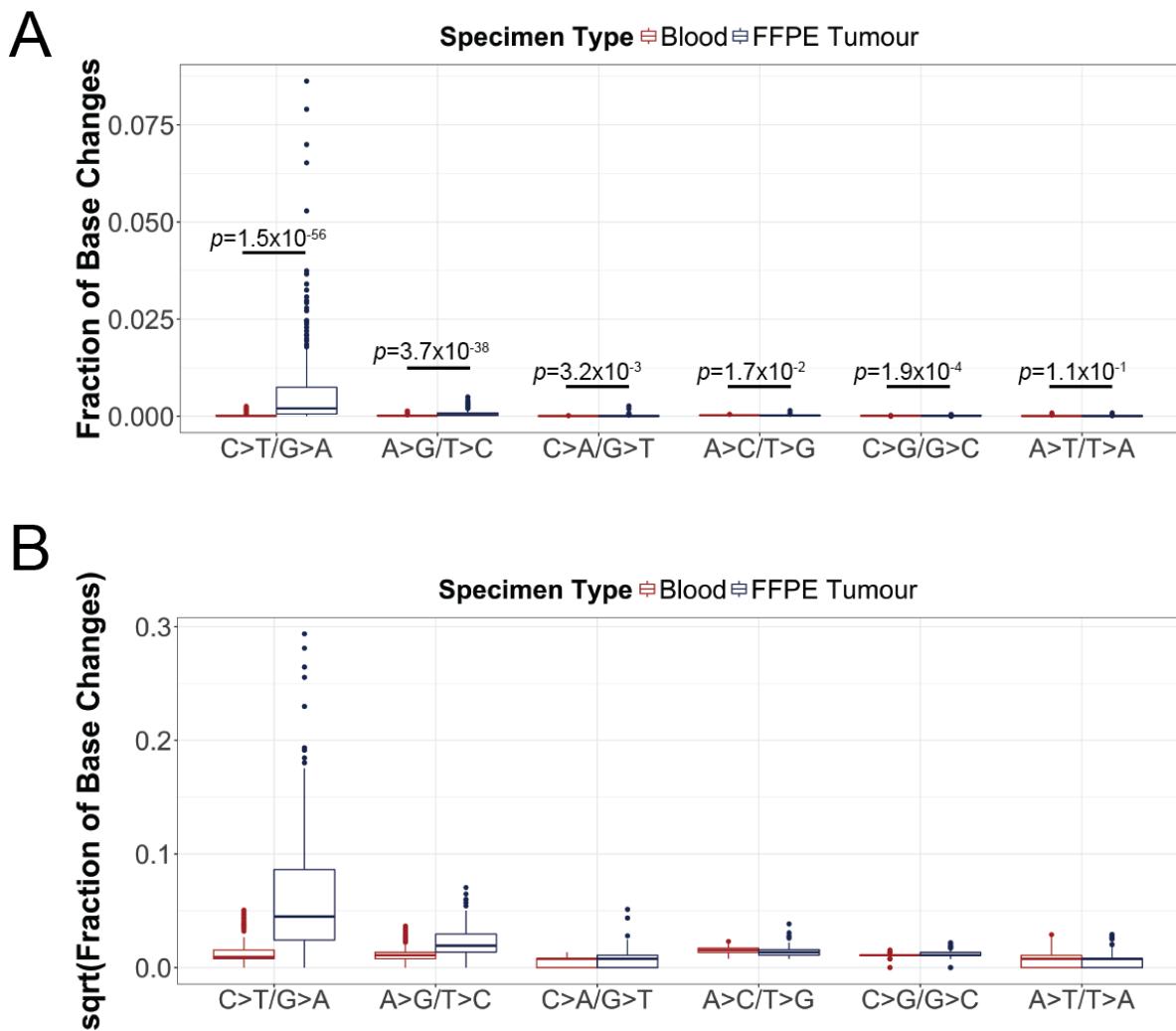


Figure 3.7: Assessment of formalin-induced sequence artifacts in FFPE specimens. (A) Comparison of fraction of base changes in blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and IQR of fraction of base changes for different types of base changes, 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.

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

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

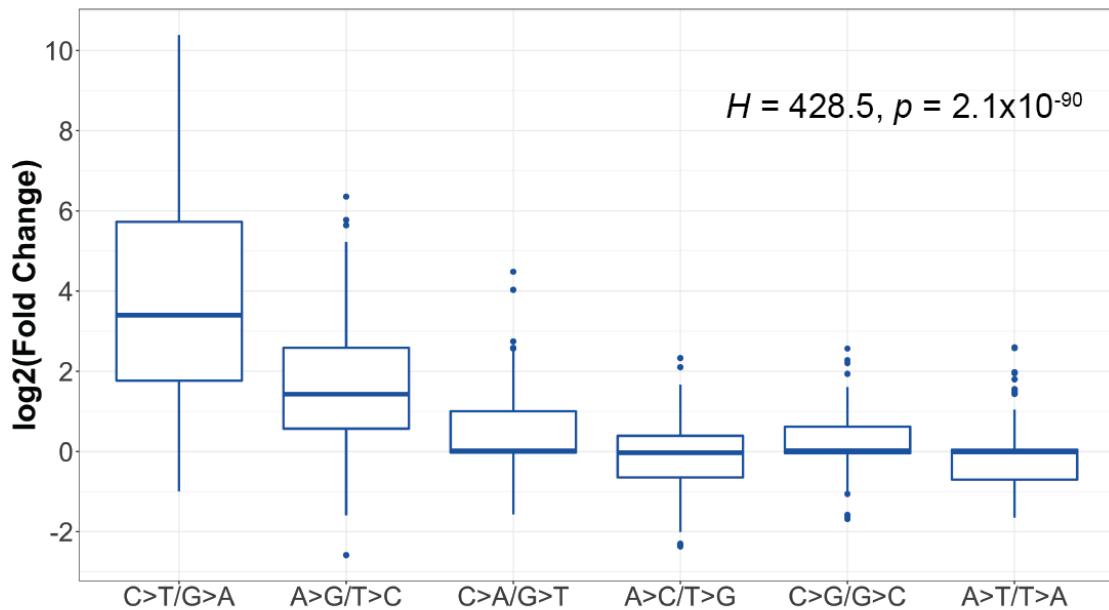


Figure 3.8: Comparison of relative difference in fraction of base changes in FFPE specimens compared to blood (Kruskal-Wallis test). Relative difference was measured as \log_2 fold change between fraction of base changes in blood and FFPE specimens ($\log_2(\text{Fraction of Base Changes}_{\text{FFPE}}/\text{Fraction of Base Changes}_{\text{Blood}})$). 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 $6.46 \times 10^{-45}*$

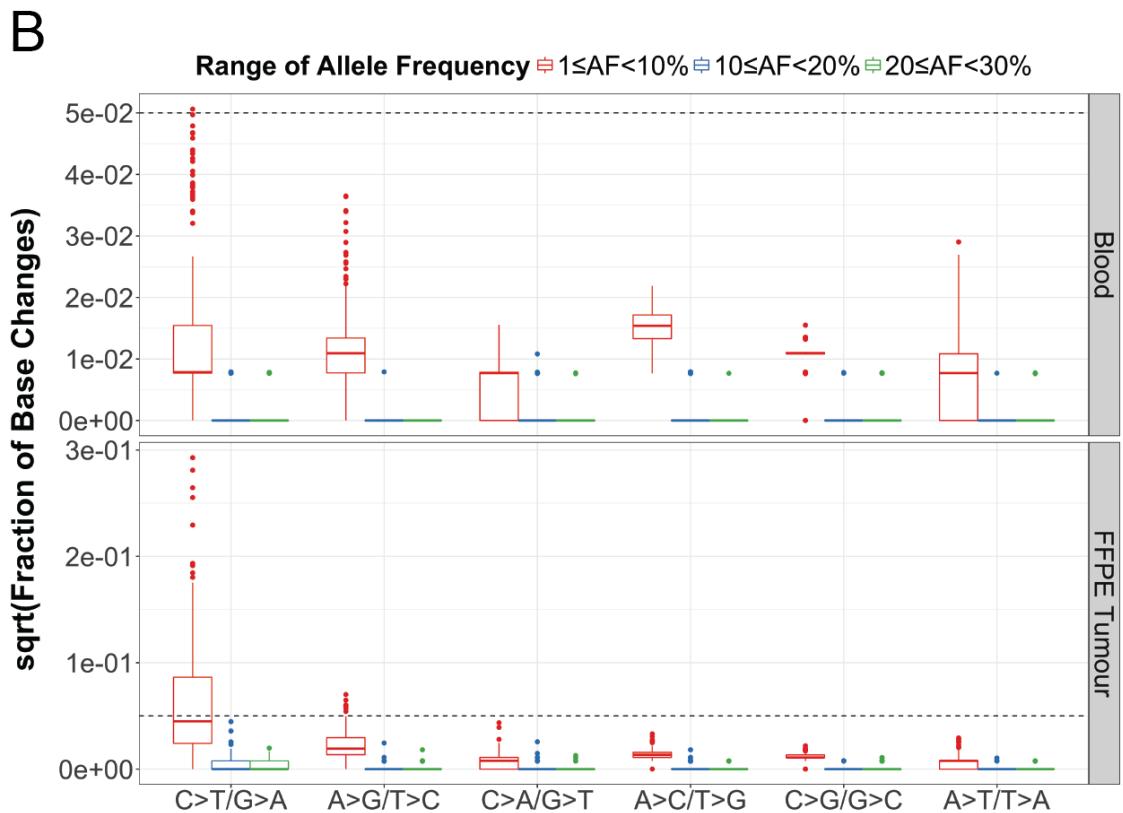
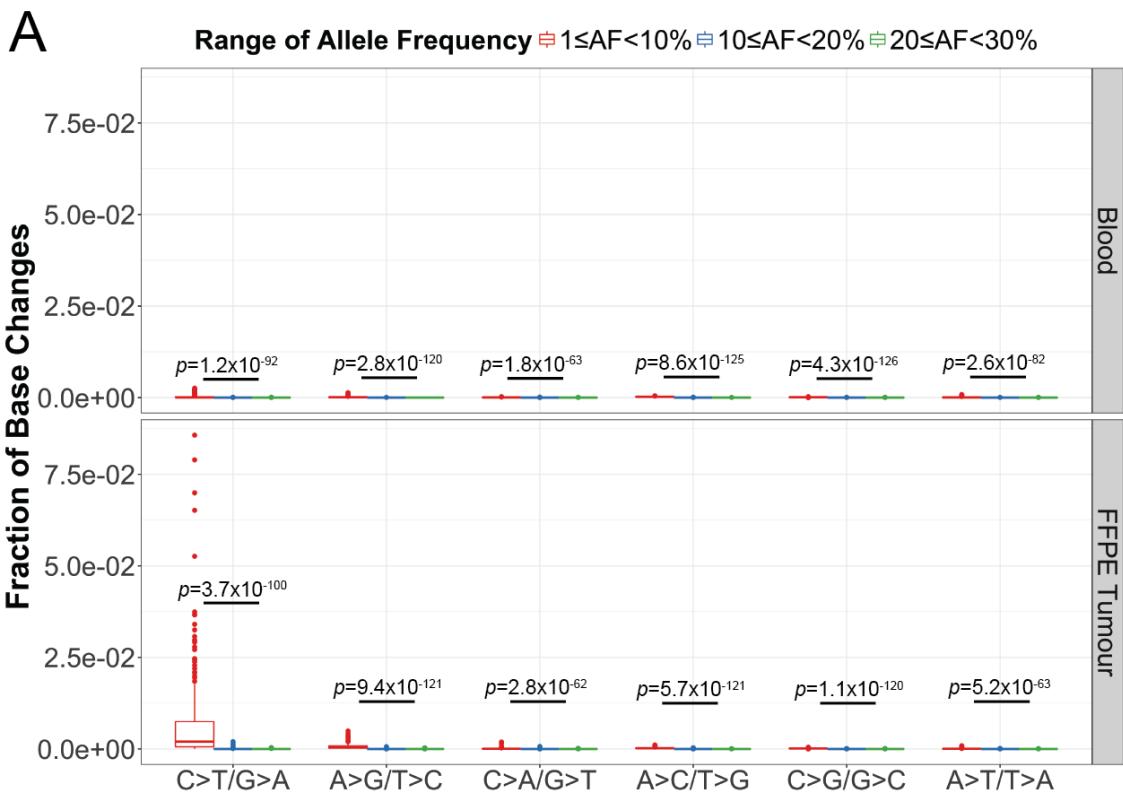


Figure 3.9: Assessment of formalin-induced sequence artifacts in FFPE specimens at different ranges of allele frequency. (A) Comparison of fraction of base changes across different ranges of allele frequency (Kruskal-Wallis test). Box plots show the median (horizontal bar within) and IQR of fraction 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.

Type of Base Changes	Blood		FFPE Tumour		†FC
	Median	Range	Median	Range	
C>T/G>A	6.2×10^{-5}	$0-2.6 \times 10^{-3}$	2.0×10^{-3}	$0-8.6 \times 10^{-2}$	33
A>G/T>C	1.2×10^{-4}	$0-1.3 \times 10^{-3}$	3.7×10^{-4}	$0-4.9 \times 10^{-3}$	3.1
C>A/G>T	6.0×10^{-5}	$0-2.4 \times 10^{-4}$	6.0×10^{-5}	$0-1.9 \times 10^{-3}$	1.0
A>C/T>G	2.4×10^{-4}	$5.9 \times 10^{-5}-4.8 \times 10^{-4}$	1.8×10^{-4}	$0-1.1 \times 10^{-3}$	0.77
C>G/G>C	1.2×10^{-4}	$0-2.4 \times 10^{-4}$	1.2×10^{-4}	$0-4.8 \times 10^{-4}$	1.0
A>T/T>A	6.0×10^{-5}	$0-8.4 \times 10^{-4}$	5.9×10^{-5}	$0-8.6 \times 10^{-4}$	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 [30, 111]. Several studies reported increased fragmentation damages in DNA isolated from older paraffin blocks due to longer exposure to environmental conditions [7, 18, 68, 93]. 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 [18, 111]. 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.

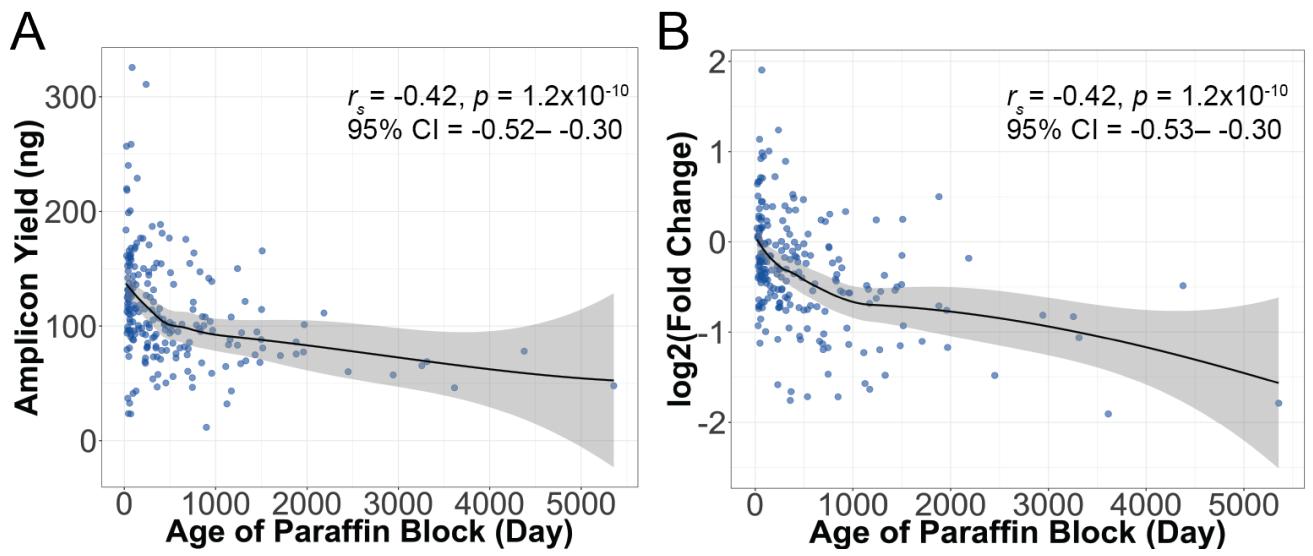


Figure 3.10: Scatter plots showing (A) amplicon yield and (B) efficiency in amplicon enrichment, which is represented by the \log_2 fold change between the amount of DNA input for producing amplicons and amplicon yield, in relation to age of paraffin blocks (Spearman's rank correlation). Solid lines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

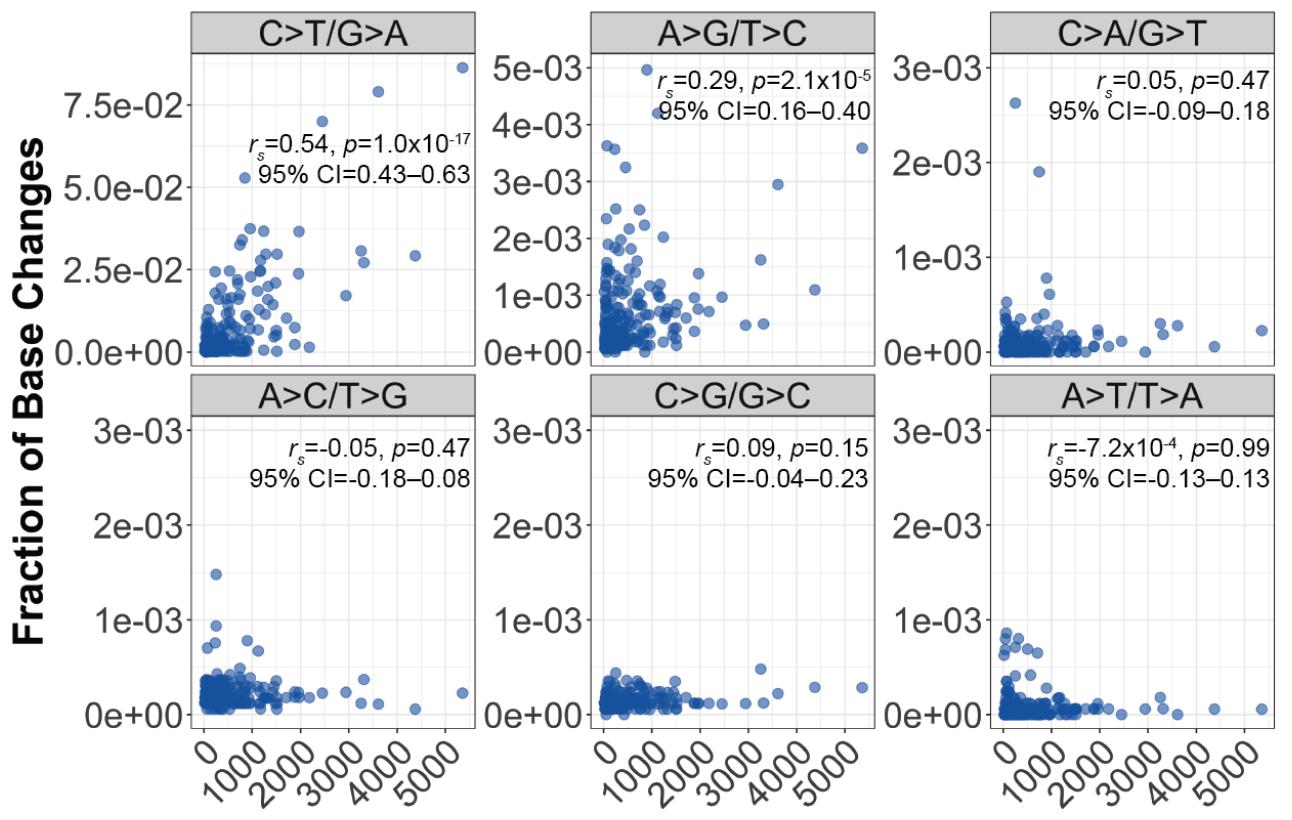


Figure 3.11: The relationship between fraction of base changes and age of paraffin block for different types of base changes (Spearman's rank correlation).

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 [40, 41, 65, 109]. 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 routine screening and sibling testing [75, 91]. Moreover, germline PGx variants could predict response to drugs like chemotherapeutic agents, thereby preventing adverse drug reactions [26, 38, 56, 62, 73, 77, 78, 105].

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 dbSNP, 1000 Genomes Project, and COSMIC could enable the differentiation of germline variations from somatic mutations [49, 57]. 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 [12, 45, 87].

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 [57, 75, 87, 91], 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 formolin-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 identify 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
	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
MAPK1	4	55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	37, 3	40	18
	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
	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.	[64]
	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.	[14, 63, 108, 116]

	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.	[108]
	7:55259450	rs2229066; COSM85893; rs17290559	p.Arg836Arg c.2508C>T	1.7	Syn.	Benign/Likely benign	NA	NA
51	KIT	4:55592059	rs151016327; COSM3760661	p.Thr461Thr c.1383A>G	0.28	Syn.	Benign	NA
		4:55599268	rs55789615; COSM1307	p.Ile798Ile c.2394C>T	2.1	Syn.	Benign/Likely benign	NA
		4:55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	12	Syn.	Benign/Likely benign	NA
	MAPK1	22:22162126	rs386488966; rs3729910	p.Tyr43Tyr c.129T>C	4.5	Syn.	NA	NA
		22:22221623	rs201495639	p.Tyr36Tyr c.108C>T	0.052	Syn.	NA	NA
	MTOR	1:11169420	rs41274506	p.Asp2485Asp c.7455C>T	0.33	Syn.	NA	NA

	1:11172909	NA	p.Glu2456Lys c.7366G>A	0.00082	Missense	NA	NA
	1:11174452	NA	p.Arg2408Gln c.7223G>A	NA	Missense	NA	NA
	1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	22	Syn.	NA	Likely has an effect on exonic [120] splicing enhancer or exonic splicing silencer binding site activity.
	1:11184593	rs56051835	p.Leu2208Leu c.6624T>C	0.49	Syn.	Benign	NA
	1:11188172	rs370318222	p.Tyr1974Tyr c.5922C>T	0.00082	Syn.	NA	NA
	1:11190646	rs2275527	p.Ser1851Ser c.5553C>T	22	Syn.	Benign	NA
	1:11190730	rs17848553	p.Ala1823Ala c.5469C>T	2.4	Syn.	Benign	NA
	1:11194521	COSM180791	c.5133C>T	0.029	Splice region	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.	[66]
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.	[120]
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.	[36]
	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

5

	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	[29]
	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.	[39]
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 <i>in silico</i> analysis. Reported in patients with Li-Fraumeni syndrome and cancer patients without family histories of Li-Fraumeni syndrome or Li-Fraumeni-like syndrome.	[5, 6, 24, 84, 86, 107]
	17:7577553	COSM44368	p.Met243fs c.727delA	NA	Frameshift	NA	Reported in esophageal squamous cell carcinoma of patients from northern Iran.	[9]

17:7578210	rs1800372; COSM249885	p.Arg213Arg c.639A>G	1.2	Syn.	Benign/Likely benign	One study demonstrated that [85] this variant was not a predictive biomarker for initiation and pro- gression of gastroesophageal reflux disease, Barrett's Esophagus, and esophageal cancer in the Brazilian population.
17:7578420	COSM1386804	p.Thr170Thr c.510G>A	0.012	Syn.	NA	One study reported that TP53 mu- [106] tations in exon 5, which include this variant, were associated with the worst prognosis for patients with non-small-cell lung cancer.

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 predisposition to various cancer types.	[11, 13, 16, 23, 53, 58, 59, 112, 114, 115, 118, 119]
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	Total	Pct [‡] (%)
					wt-var [†] , var-var ^{††}		
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.	[3, 10, 19, 28, 34, 62, 72, 76, 78, 81, 92, 100, 103–105]
	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.	[10, 28, 42, 92, 104, 105]
	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.	[3, 19, 28, 42, 62, 76, 78, 92, 100, 103–105]

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.	[10, 81]
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.	[76, 81, 92, 103, 105]
1:98039419	rs56038477	p.Glu412Glu c.1236G>A	1.5	Syn.	Benign	Synonymous variant in 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.	[3, 28, 76, 79]

	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.	[28, 42, 81, 100, 104, 105]
	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.	[19, 42, 78, 101, 105]
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.	[2, 22, 51, 74, 89, 98]
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.	[37, 38, 55, 71, 89]

	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.	[25, 37, 38, 47, 55, 71, 89, 92, 99]
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.	[17, 52, 56]

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.	[46, 56, 96]
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 neuro-gastrointestinal encephalomyopathy (MNGIE), but relation with TP enzymatic defect was not established.	[80]
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.	[1, 35, 44, 47, 69, 98]	[1, 35, 44, 47, 69, 98]
UGT1A1	2:234668870	rs873478	c.-64G>C	1.1 ^{##}	Upstream gene	NA	Unknown	[20, 113, 117]

2:234668879	rs34983651	c.-55_-54insAT	33 ^{‡‡}	Upstream gene	Conflicting interpretations of pathogenicity, affects, association	Lower UGT1A1 expression and associated with irinotecan-related toxicity.	[4, 27, 43, 54, 61, 70, 74, 88, 90, 102]
-------------	------------	----------------	------------------	---------------	--	--	--

*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.

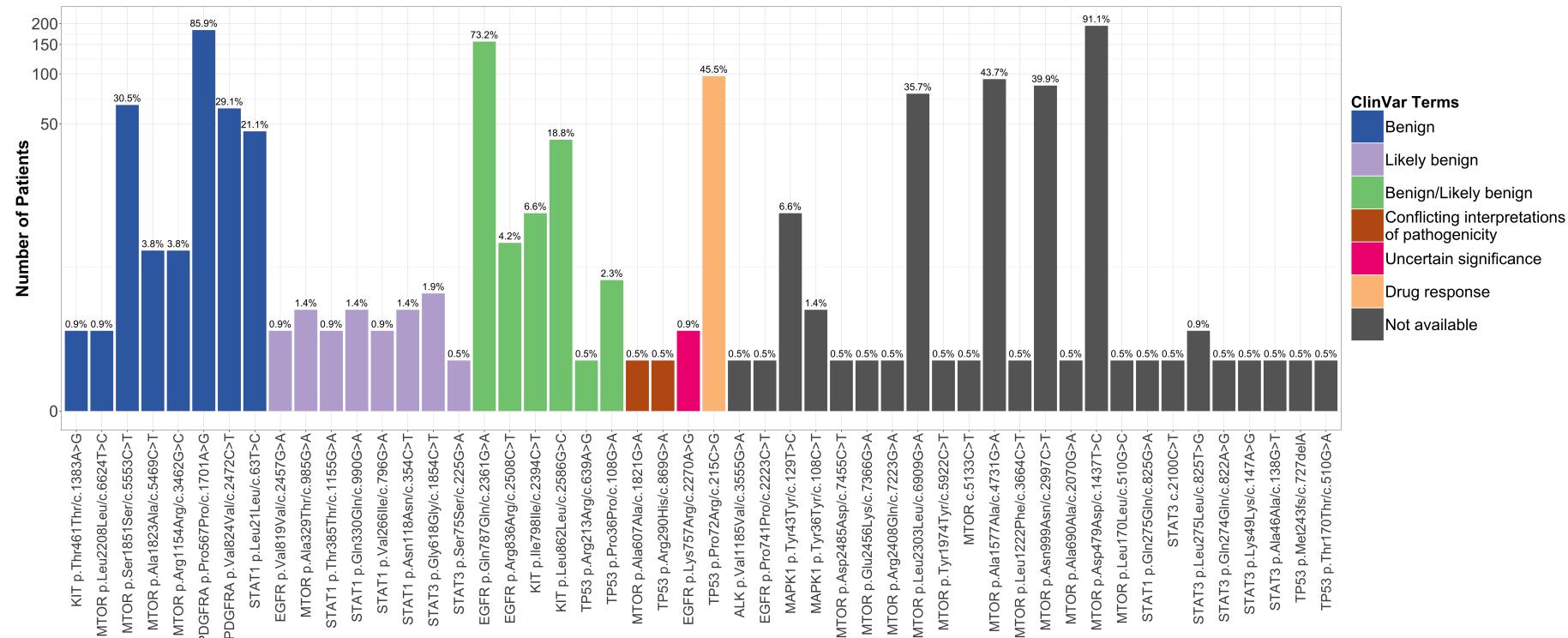


Figure 4.1: Distribution of germline alterations in cancer-related genes in patients from TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. The TP53 variant, p.Arg72Pro/c.215G>C, 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.

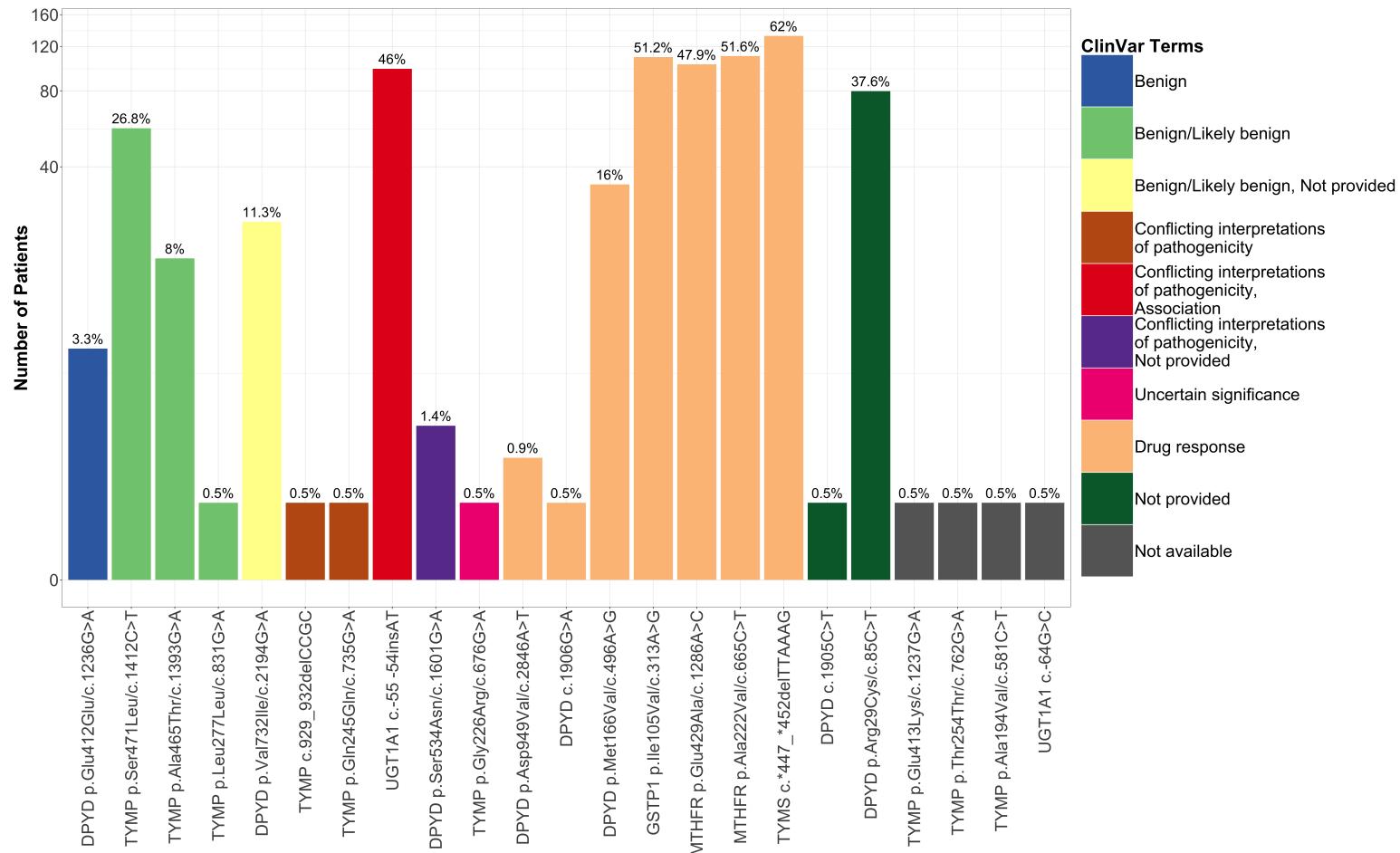


Figure 4.2: Distribution of germline alterations in PGx genes in patients from TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. 208 out of 213 patients in TOP cohort 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 [57, 75, 87, 91]. 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.

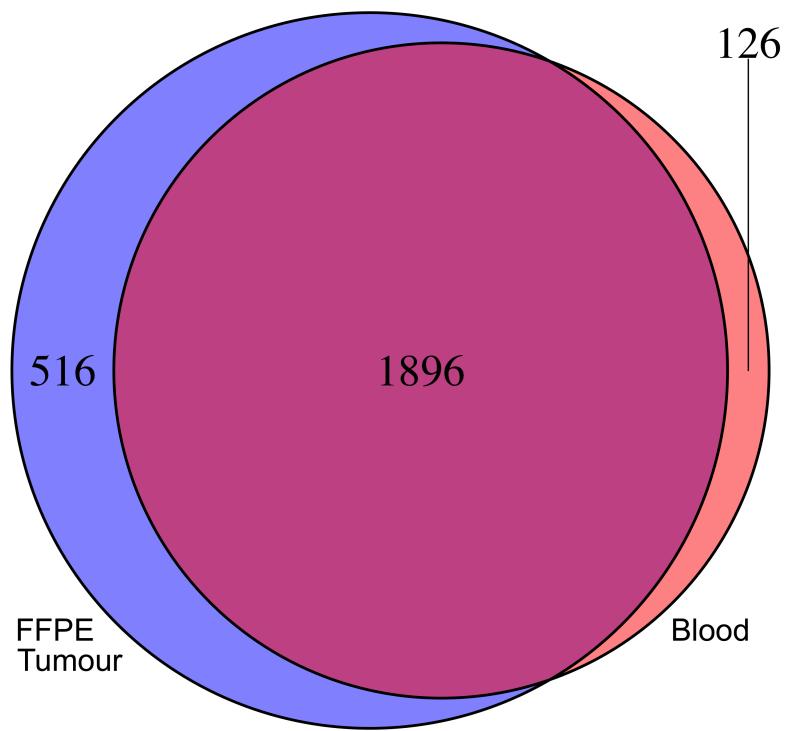


Figure 4.3: Venn diagram demonstrating concordance of variants identified in 217 tumour-blood paired samples.

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	

76

*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 [12, 45, 87]. 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% CI = 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

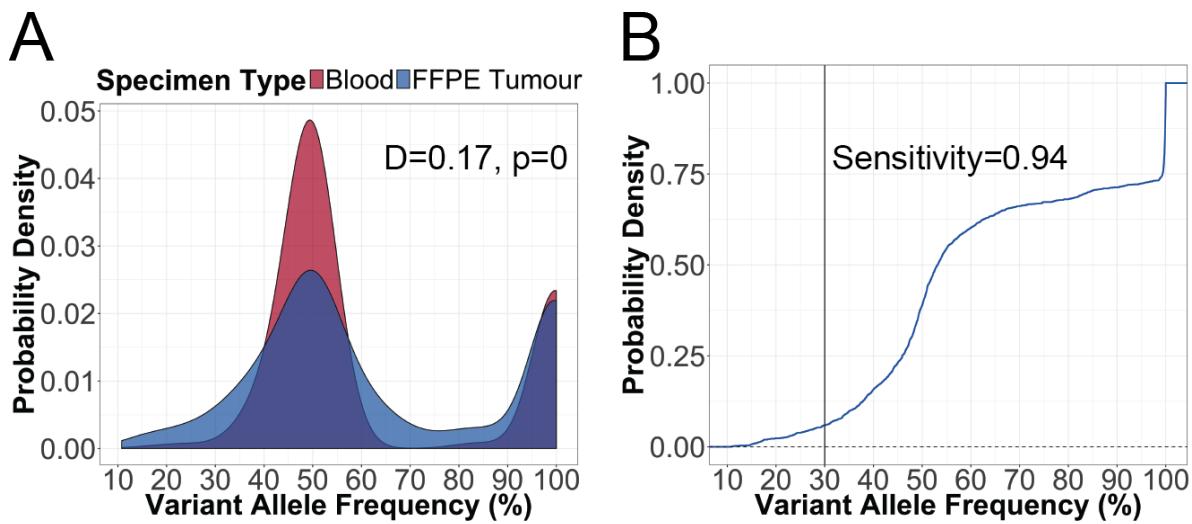


Figure 4.4: Assessment of using a VAF cut-off approach to identify germline alterations in tumour-only analyses. (A) Comparison of VAF distributions of germline alterations between blood and tumour (Kolmogorov-Smirnov test). (B) Empirical cumulative distribution of VAFs of germline alterations in tumour 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

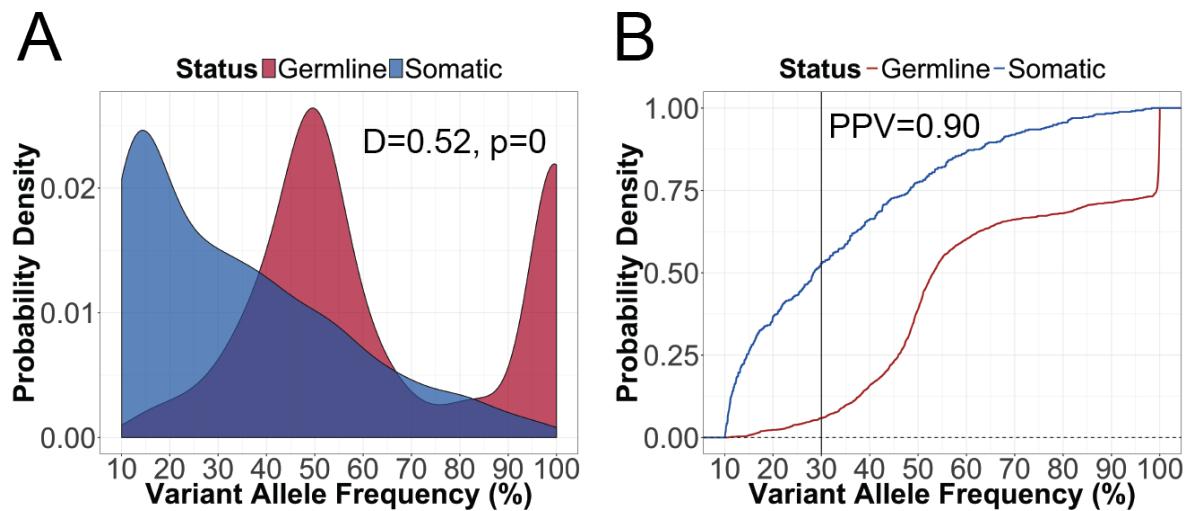


Figure 4.5: Assessment of using a VAF cut-off approach to refer potential germline alterations in tumour-only analyses to follow-up testing. (A) Comparison of VAF distributions between germline and somatic alterations in tumour specimens (Kolmogorov-Smirnov test). (B) Empirical cumulative distribution 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 families [57, 75, 91]. While sequencing of tumour-normal pairs can enable differentiation between germline and somatic variants, matched normal samples are often not obtained in the clinical setting. 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 [32, 60, 82, 83, 95, 110, 111]. 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 feasibility 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[94], Didelot et al[30], and Wong et al[110] 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 [8, 97, 110]. Additionally, Wong et al [111] and Didelot et al [30] 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 should be designed to circumvent the drawback of fragmentation damage in FFPE samples.

Cytosine deamination is the major cause of sequence artifacts in formalin-fixed specimens [21, 31, 33, 60, 83, 97, 111]. Herein, we observed increased C>T/G>A artifacts in FFPE DNA compared to blood. Formation of artifactual C>T/G>A changes is caused by incorporation of adenine in the complementary DNA strand during PCR amplification at uracil lesions that are resulted by deamination of cytosines [32]. When measuring frequency of sequence artifacts at different allele frequency ranges, Wong et al [111] 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 allele frequency ranges, including 1–10%, 10–20%, and 20–30%. Likewise, we observed a substantial increase in C>T/G>A within the 1–10% allele frequency range. We were unable to separate FFPE artifacts from low frequency somatic mutations within these allele frequency ranges because of the lack of matched fresh frozen tumour tissues. As our goal is to predict germline status, disproportionate base changes between FFPE and blood DNA within these allele frequency ranges suggest that germline call should be made at > 30% VAF to avoid false positives that could either arise from true somatic mutations or FFPE artifacts. Nevertheless, somatic mutations can occur at VAFs that deviate significantly from a diploid zygosity due to tumour admixture samples, meaning that. Therefore, genetic alterations at < 30% VAF are clinically significant and further optimization of protocols must be performed to subtract FFPE artifacts from true somatic mutations. One method to reduce sequence artifacts as a result of cytosine deamination is through treatment with UDG

Furthermore, we observed increased C>T/G>A base transitions in FFPE DNA, which are

caused by cytosine deamination, a predominant type of sequence artifact induced by formalin. We also observed differences in frequency of base changes between FFPE and blood DNA within the allele frequency ranges of 1–10%, 10–20%, and 20–30%, implying high prevalence of FFPE artifacts or somatic mutations within these allele frequency ranges. Based on this result, it is recommended that germline calls are made at > 30% VAF to avoid false positives arising from true somatic mutations or FFPE artifacts. Additionally, we showed that increased storage time of FFPE blocks resulted in poorer amplicon enrichment and sequencing results, as well as elevated levels of C>T/G>A artifacts. Therefore, if multiple FFPE specimens are available, the paraffin block with the shorter storage time should be selected for genomic testing. Overall, discrepancies in enrichment efficiency and sequencing results between FFPE and blood DNA were detectable in our dataset. However, we found that these differences were technically negligible or can be minimized using appropriate methods, thereby confirming the feasibility of using FFPE DNA for germline testing.

Furthermore, we showed elevated levels of C>T/G>A base changes in FFPE specimens compared to blood. This finding is consistent with several studies demonstrating cytosine deamination as the main source of sequence artifact in formalin-fixed specimens. In addition to C>T/G>A base changes, we observed increased A>G/T>C in FFPE DNA, although this difference was smaller than C>T/G>A. Our findings also demonstrated that base changes at the 1-10%, 10-20%, and 20-30% allele frequency ranges were more prevalent in FFPE DNA than blood. This result implies that germline calls should be made at >30% VAF to remove the majority of formalin-induced artifactual changes and somatic mutations. Finally, we showed that sequencing metrics are dependent on storage time of paraffin blocks. Older FFPE blocks were more likely to yield lower efficiency in amplicon enrichment, leading to poorer sequencing results and increased prevalence of artifactual C_↓T/G_↓A transitions. Thus, this indicates that if multiple FFPE specimens are available, the specimen with the lower storage time should be selected for molecular analysis.

Oh *et al.*[83] and Spencer *et al.*[97] reported shorter insert sizes in FFPE DNA that were subjected to whole exome and hybridization capture sequencing, respectively. Additionally, Didelot *et al.*[30] and Wong *et al.*[110] examined the integrity of DNA isolated from FFPE samples and demonstrated that FFPE DNA have shorter fragment lengths.

We observed lower efficiency in amplicon enrichment in FFPE DNA, with a more pronounced decrease in coverage depth for longer amplicons in the panel. While we noted 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. These results are consistent with several studies reporting extensive fragmentation of DNA extracted from FFPE samples [8, 30, 83, 97, 110, 111].

Various studies also showed that increased fragmentation damage in FFPE DNA is associated with low coverage depth and uniformity, which is in agreement with our findings [8, 30, 97, 110, 111]. For instance, Betge *et al.*[8] examined the amount of template DNA in FFPE samples using

a qPCR assay and found an inverse correlation between amount of template DNA and coverage depth. Although our findings demonstrated lower efficiency in amplicon enrichment of FFPE DNA, this disparity only led to negligible effects in coverage depth and uniformity. Nevertheless, shorter amplicons could be designed for the panel test to ensure improved enrichment and sequencing outcomes of FFPE DNA.

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. Our findings demonstrated that DNA fragmentation and cytosine deamination were common forms of DNA damage in FFPE specimens. We noted lower efficiency in amplicon enrichment in FFPE DNA, with a more pronounced reduction in coverage depth for longer amplicons in the panel. This suggests that the use of shorter amplicons could achieve improved enrichment and sequencing coverage, mitigating formalin-induced fragmentation effects.

Our findings were consistent with several studies reporting the effects of DNA fragmentation something about yes we noticed lower coverage depth too

This is which is evident of fragmentation damage in FFPE DNA, which leads to shorter fragment sizes that are not amenable to PCR amplification for longer amplicons, suggesting that

A key constraint in clinical genomics is the lack of matched normal samples to enable discrimination between germline and somatic alterations in tumour-only analyses. Analysis of variant concordance between

A key concern of using FFPE DNA for clinical genomic testing is the prevalence of sequence artifacts caused by cytosine deamination.

other forms of sequence artifacts

Our findings confirmed that DNA fragmentation and cytosine deamination were common forms of DNA damage in FFPE specimens.

age of paraffin blocks

can facilitate optimization of diagnostic workflow.

concordance, how reliable is the tumour DNA for germline testing

approach using variant allele frequency

Conclusion: As MPS becomes increasingly incorporated into clinical diagnostic work- flows, it is important to assess DNA damage caused by formalin fixation, as this will greatly optimise diagnostic workflows, increase accuracy of results and lead to better outcomes for patients.

This enabled characterization of formalin-induced DNA damage in our data and its impact on germline variant calling in FFPE DNA.

We observed lower , which can be caused by reduced template DNA as a result of formalin-induced DNA fragmentation. Although FFPE and blood DNA managed to achieve comparable proportions of on-target read alignments, which are aligned reads used for variant calling, our results showed minor discrepancies in coverage depth and uniformity in FFPE DNA compared to blood.

We also evaluated amplicon-specific differences in coverage depth and observed reduced coverage depth for longer amplicons in FFPE DNA. Furthermore, we showed elevated levels of C>T/G>A base changes in FFPE specimens compared to blood. This finding is consistent with several studies demonstrating cytosine deamination as the main source of sequence artifact in formalin-fixed specimens. In addition to C>T/G>A base changes, we observed increased A>G/T>C in FFPE DNA, although this difference was smaller than C>T/G>A. Our findings also demonstrated that base changes at the 1-10%, 10-20%, and 20-30% allele frequency ranges were more prevalent in FFPE DNA than blood. This result implies that germline calls should be made at >30% VAF to remove the majority of formalin-induced artifactual changes and somatic mutations. Finally, we showed that sequencing metrics are dependent on storage time of paraffin blocks. Older FFPE blocks were more likely to yield lower efficiency in amplicon enrichment, leading to poorer sequencing results and increased prevalence of artifactual C_iT/G_iA transitions. Thus, this indicates that if multiple FFPE specimens are available, the specimen with the lower storage time should be selected for molecular analysis.

EGFR-overexpressing cancers are highly aggressive and have a higher tendency to metastasize. Currently, available drugs specifically target the EGFR and elicit high response rates. However, the majority of patients eventually develop progressive disease. The mechanisms through which cancers escape EGFR-targeted therapies remain unclear. Identification of specific molecules that mediate resistance to EGFR-directed treatments will facilitate the development of novel therapies and may improve responses to currently available therapies.

In this study, we measured secreted factors in the media of sensitive and resistant cell lines to identify differentially expressed cytokines that may mediate resistance. Through a combination of ELISAs and mass spectrometry-based assays, we identified cytokine A as being significantly up-regulated in resistant cells. Cytokine A is a major activator of the ABCD signaling cascade (literature citations). The ABCD cascade is a known target of EGFR signaling and is usually blocked in response to EGFR inhibition (literature citations). A previous study demonstrated that exogenous stimulation of ABCD signaling reduces the response to EGFR-targeted drugs (literature citations). This report is consistent with our finding that a major stimulus of ABCD signaling is overexpressed in resistant cells. Based on these data, we propose that hyperactive ABCD signaling is a major mechanism of resistance to EGFR-targeted therapies (Figure XX, schematic of proposed mechanism of resistance). This section will be greatly expanded in a real Discussion section to place your finding in the context of multiple published studies.

In this study, we retrospectively analyzed targeted sequencing data from tumour-blood paired samples of 213 cancer patients to investigate whether potential germline alterations can be accurately identified in FFPE tumours without the use of matched normal samples. Because blood DNA is one of the gold standards for germline testing, we characterized formalin-induced DNA damage in our data to evaluate the quality and usability of FFPE DNA for germline variant calling. Using

blood DNA as non-formalin-fixed controls, we compared efficiency in amplicon enrichment and sequencing results of FFPE DNA to blood. We observed lower efficiency in amplicon enrichment in FFPE DNA, which can be caused by reduced template DNA as a result of formalin-induced DNA fragmentation. Although FFPE and blood DNA managed to achieve comparable proportions of on-target read alignments, which are aligned reads used for variant calling, our results showed minor discrepancies in coverage depth and uniformity in FFPE DNA compared to blood. We also evaluated amplicon-specific differences in coverage depth and observed reduced coverage depth for longer amplicons in FFPE DNA.

This result reveals the impact of DNA fragmentation caused by formalin fixation, which gives rise to shorter template DNA that are less amplifiable for longer amplicons.

In addition to cytosine deamination, formaldehyde can react with atmospheric oxygen to result in formic acid, which causes depurinationTo a lesser extent, which could be caused by incorporation of guanines at abasic sites induced by reaction

Although both 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 log₂ 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.

Caveats: did not review every single variant

Chapter 6

Conclusion

What are my main findings?

Bibliography

- [1] 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 67
- [2] 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 64
- [3] 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 62, 63
- [4] 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 68
- [5] N. Ånensen, J. Skavland, C. Stapnes, A. Rynningen, 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 56
- [6] 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 56
- [7] 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 37

- [8] 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 82, 83
- [9] 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 56
- [10] 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 Predict Adverse Events Following Treatment With Fluorouracil-Based Adjuvant Chemotherapy in Patients With Stage III Colon Cancer. *JAMA oncology*, 2(5):655–662, 2016. doi:10.1001/jamaoncol.2015.5392. → pages 62, 63
- [11] 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 58
- [12] 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 42, 77
- [13] 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. Olivier, R. Antonicelli, M. Memo, D. Uberti, B. Naemias, 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 58
- [14] S. Bonin, M. Donada, G. Bussolati, E. Nardon, L. Annaratone, M. Pichler, A. M. Chiaravalli, C. Capella, G. Hoefer, 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 50
- [15] 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 8

- [16] 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.ncbi.nlm.nih.gov/article/?render.fcgi?artid=1904480&tool=pmcentrez&rendertype=abstract>. → pages 58
- [17] 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. *Cancer Research*, 17(7):2006–2013, 2011. ISSN 10780432. doi:10.1158/1078-0432.CCR-10-1741. → pages 65
- [18] D. M. Carrick, M. G. Mehaffey, M. C. Sachs, S. Altekroose, C. Camalier, R. Chuquui, W. Cozen, B. Das, B. Y. Hernandez, C. J. Lih, C. F. Lynch, H. Makhlof, 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 37
- [19] 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 62, 64
- [20] 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 67
- [21] 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 82
- [22] 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 64
- [23] 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 58

- [24] 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 56
- [25] 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 Colorectal. *Clinical Cancer Research*, 9(May):1611–1615, 2003. → pages 65
- [26] 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 42
- [27] 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 68
- [28] 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 62, 63, 64
- [29] 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 55
- [30] 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 17, 37, 82, 83

- [31] 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 29, 82
- [32] 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 17, 29, 81, 82
- [33] 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 29, 82
- [34] 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 62
- [35] 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 67
- [36] 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. PDGFR α/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.ncbi.nlm.nih.gov/articlerender.fcgi?artid=3531259&tool=pmcentrez&rendertype=abstract>. → pages 54
- [37] 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 64, 65
- [38] M. C. Etienne-Grimaldi, G. Milano, F. Maindrault-Goebel, 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 42, 64, 65

- [39] G. Fortunato, G. Calcagno, V. Bresciamorra, E. Salvatore, A. Fillà, 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 55
- [40] 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>{%}5Cn<http://www.nature.com/nbt/journal/v31/n11/abs/nbt.2696.html>{%}5Cn<http://www.ncbi.nlm.nih.gov/pubmed/24142049>. → pages 42
- [41] 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 42
- [42] 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 62, 64
- [43] 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 68
- [44] 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 67
- [45] 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 42, 77

- [46] 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 66
- [47] 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 65, 67
- [48] 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 29
- [49] 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 42
- [50] 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 29
- [51] 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 64
- [52] 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 65
- [53] 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 58

- [54] F. Innocenti, S. D. Undeva, L. Iyer, P. X. Chen, S. Das, M. Kocherginsky, T. Garrison, 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 68
- [55] 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 64, 65
- [56] 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 42, 65, 66
- [57] 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 42, 43, 71, 81
- [58] 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 58
- [59] 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 58
- [60] 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 17, 29, 81, 82
- [61] 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 68
- [62] A. M. Lee, Q. Shi, E. Pavely, 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 42, 62

- [63] 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 50
- [64] 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 50
- [65] 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/AJCPMWGWO34EGOD. URL <http://www.ncbi.nlm.nih.gov/article/abstract?tool=pmcentrez&rendertype=abstract>. → pages 29, 42
- [66] 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> → pages 53
- [67] J. J. Lou, L. Mirsadraei, D. E. Sanchez, R. W. Wilson, M. Shabikhani, 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 17
- [68] 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 37
- [69] 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 67

- [70] 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 68
- [71] 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 64, 65
- [72] 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 62
- [73] H. L. McLeod. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science*, 339(March):1563–1566, 2013. doi:10.1126/science.1234139. → pages 5, 42
- [74] 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 64, 68
- [75] 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 5, 42, 43, 71, 81
- [76] 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 62, 63
- [77] B. Mohelníková-Duchonová, 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 42

- [78] A. Morel, M. Boisdror-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 42, 62, 64
- [79] 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 63
- [80] I. Nishino, A. Spinazzola, A. Papadimitriou, S. Hammans, I. Steiner, C. D. Hahn, A. M. Connolly, A. Verloes, J. Guimaraes, 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 66
- [81] 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 62, 63, 64
- [82] 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 17, 29, 81
- [83] 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.ncbi.nlm.nih.gov/article/fcgi?artid=4671711&tool=pmcentrez&rendertype=abstract>. → pages 17, 29, 81, 82, 83
- [84] 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 56
- [85] 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 57

- [86] 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 56
- [87] 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 4, 5, 42, 43, 71, 77
- [88] E. Rouits, V. Charasson, a. Pétain, M. Boisdrone-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 68
- [89] 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 64, 65
- [90] 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 68
- [91] 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 5, 42, 43, 71, 81
- [92] 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 62, 63, 65
- [93] 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 37
- [94] 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 82
- [95] 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 17, 81
- [96] 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 66
- [97] 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 82, 83
- [98] 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 64, 67
- [99] 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 65
- [100] 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 62, 64

- [101] 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 64
- [102] 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 68
- [103] 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 62, 63
- [104] 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 62, 64
- [105] 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.bbadi.2016.01.009. URL <http://dx.doi.org/10.1016/j.bbadi.2016.01.009>. → pages 42, 62, 63, 64
- [106] 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 57
- [107] 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 56
- [108] 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.ncbi.nlm.nih.gov/pmc/articles/PMC3730025/>; tool=pmcentrez;rendertype=abstract. → pages 50, 51

- [109] 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 42
- [110] 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.ncbi.nlm.nih.gov/pmc/articles/PMC3718011/>;tool=pmcentrez;rendertype=abstract. → pages 17, 81, 82, 83
- [111] 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 5, 17, 29, 37, 81, 82, 83
- [112] 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 58
- [113] 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 67
- [114] 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 58
- [115] 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 58
- [116] 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 50
- [117] 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 67
- [118] 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 58
- [119] 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 58
- [120] 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 52, 53

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					