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

Abstract

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

Note: Abstracts should generally try to avoid using acronyms.

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

Preface

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

Table of Contents

List of Tables

List of Figures

List of Abbreviations

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

Although this thesis only bears one name, its completion would be impossible without the contribution of many individuals. First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Aly Karsan, for the opportunity to work with his team of diverse talents as well as his patience, guidance, and extensive knowledge in clinical informatics.

I would also like to thank my labmates from the bioinformatics team, Kieran and Rod, and members from the Centre of Clinical Genomics, Liz and Jill, for their insightful comments and help throughout my time in the lab. I would like to extend my gratitude to my supervisory committee members, Dr. Ryan Morin and Dr. Martin Hirst, for their knowledgeable feedback and continuous effort in asking me difficult questions which motivated me to widen my research perspective.

My sincere thanks also goes to my friends, who supported me and lifted my spirits through the tough times. Last but not least, I would like to thank my family: mom and dad, for always encouraging my interest in science and listening to my endless science talks and my sisters, for believing in my ability even when I doubt myself. This thesis is yours as much as it is mine.

Chapter 1

Introduction

1.1 The Evolution of Molecular Diagnostics in Cancer

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 received precision medicine that is the right drug at the right dose.

1.2 Next-generation Sequencing Technologies

1.3 Applications of Next-generation Sequencing

1.3.1 Targeted Sequencing

Capture-based, amplicon-based etc.

1.3.2 Whole Exome Sequencing

1.3.3 Whole Genome Sequencing

1.4 Bioinformatics Tools for Variant Calling

1.4.1 Types of Genomic Alterations

There are different types of genomic alterations.

1.4.2 Variant Calling Pipeline

1.4.3 Variant Calling Algorithms

1.4.4 Variant Curation and Interpretation

1.5 Germline Variant Calling in The Tumour Genome

1.5.1 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 [**?**]. 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 [**?**]. In fact, several studies demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [**? ? ? ?**]. 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.

1.5.2 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 [**?**]. 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 [**?**].

1.5.3 Challenges

Detection of genomic alterations in tumour DNA is also faced with technical challenges conferred by formalin-fixed paraffin-embedded (FFPE) tumour specimens [**? ?**]. 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 [**? ?**]. Fragmentation damage caused by formalin fixation leads to reduced template DNA for PCR amplification, thereby affecting the efficiency of amplicon-based NGS testing [**? ?**]. 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.

1.6 Objectives

Chapter 2

Materials and Methods

2.1 Patient Samples

Blood and FFPE tumour specimens were acquired from 213 patients who provided informed consent for The OncoPanel Pilot (TOP) study, 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. 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 specimens 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	59	28
Melanoma	18	8
Other*	17	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, lung sarcomatoid carcinoma, Fallopian tube, gastric, endometrial, squamous cell carcinoma, anal, salivary gland, peritoneal epithelial mesothelioma, adenoid cystic carcinoma, pancreas, breast, gall bladder, parotid epithelial myoepithelial carcinoma, and small bowel cancers.

2.2 Sample Preparation, Library Construction, and Illumina Sequencing

Genomic DNA was extracted from blood and FFPE tumour specimens 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 [**?**] to attain 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).

2.3 OncoPanel (Targeted NGS Panel for Solid Tumours)

The OncoPanel assesses coding exons and clinically relevant hotpots of 15 cancer predisposing genes and six PGx genes that can predict chemotherapeutic response. Primers were designed by RainDance Technologies (Lexington, MA) using the GRCh37/hg19 reference sequence to generate 416 amplicons between 56 bp and 288 bp in size, which interrogate $\sim$ 20 kb of target bases. Complete list of genes and gene reference models for the OncoPanel is presented in Table A.1, whereas

OncoPanel target regions and amplicons are presented in Table A.2.

Table 2.2: Gene Reference Models for Genes in the OncoPanel.

Gene	Protein	Reference Model
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
DPYD	Dihydropyrimidine dehydrogenase	NM_000110.3
EGFR	Epidermal growth factor receptor	NM_005228.3
ERBB2	Receptor tyrosine-protein kinase erbB-2	NM_001005862.1
GSTP1	Glutathione S-rransferase pi 1	NM_000852.3
HRAS	GTPase HRas	NM_005343.2
IDH1	Isocitrate dehydrogenase 1	NM_005896.2
IDH2	Isocitrate dehydrogenase 2	NM_002168.2
KIT	Tyrosine-protein kinase Kit	NM_000222.2
KRAS	KRas proto-oncogene GTPase	NM_033360.2
MAPK1	Mitogen-activated protein kinase 1	NM_002745.4
MAP2K1	Mitogen-activated protein kinase kinase 1	NM_002755.3
MTHFR	Methylenetetrahydrofolate reductase	NM_005957.4
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
TYMP	Thymidine phosphorylase	NM_001113755.2
TYMS	Thymidylate synthetase	NM_001071.2
UGT1A1	Uridine diphosphate (UDP)-glucuronosyl transferase 1A1	NM_000463.2

2.4 Variant Calling Pipeline

Reads that passed the Illumina Chastity filter were aligned to the GRCh37/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). To assess the true positive rate of germline variant calling in FFPE tumours, variant calling was per-

formed separately for blood and tumour sequencing libraries. 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 the VarScan2 `mpileup2cns` (version 2.3.6) function with parameter thresholds of variant allele frequency $\geq$ 10% and Phred-scaled base quality score $\geq$ 20 `(--min-var-freq 0.1 --p-value 0.01 --strand-filter 0 --output-vcf --variants --min-avg-qual 20)`. As for comparison of variant calls between FFPE tumour replicates, variant calling was performed on tumour-normal pairs using similar parameters for the SAMtools `mpileup` function, but followed by the VarScan2 `somatic` (version 2.3.6) function. Variant calls were filtered using the VarScan2 `fpfilter` function with fraction of variant reads from each strand $\geq$ 0.1 and default thresholds for other parameters. SnpEff (version 4.2) was used for variant annotation and effect prediction whereas the SnpSift package in SnpEff was used to annotate variants with databases such as dbSNP (b138), COSMIC (version 70), 1000 Genomes Project, ClinVar, and ExAC (release 0.3) for interpretation.

2.5 Data Analysis

Coverage depth was measured using bedtools (version 2.25.0) and per-base metrics were obtained using bam-readcount (https://github.com/genome/). Statistical analyses and data visualization were performed using R (version 3.3.2) and associated open-source packages. Manual review of PGx variants were carried out using the Intergrative Genomics Viewer (IGV, version 2.3). *Note: be more specific on how the data is generated*

Chapter 3

Assessment of Formalin-Induced DNA Damages in FFPE Specimens

The main component of formalin is formaldehyde, which is known to induce DNA damages such as fragmentation and sequence artifacts. Our study design is comprised of 213 patients with FFPE tumour and matched blood specimens that were subjected to evaluation with the OncoPanel. Sequencing data of tumour-normal paired samples were processed and analyzed with a custom variant calling pipeline. With blood specimens serving as non-formalin-fixed controls, we assessed formalin-induced DNA damages by comparing efficiency in amplicon enrichment, sequencing results, and prevalence of sequencing artifacts between FFPE and blood specimens. As DNA derived from blood is the gold standard for germline testing, our assessment would determine whether FFPE DNA is a reliable resource for detection of germline variants.

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. We investigated this effect by comparing amplicon yield between blood and FFPE specimens, and a Wilcoxon signed-rank test indicated that the median amplicon yield (median = 299.3 ng) in blood specimens was significantly higher than the median amplicon yield (median = 103.6 ng) in FFPE specimens ($W = 23613$, $Z = 12.7$, $p = 8.3 \times 10^{-62}$, $r = 0.86$; Figure 3.1A, Table 3.1). However, the DNA input for amplicon enrichment varies across specimens, and a Wilcoxon signed-rank test indicated that the median DNA input (median = 147.8 ng) in blood specimens was significantly higher than the median DNA input (median = 140.9 ng) in FFPE specimens ($W = 15004$, $Z = 3.57$, $p = 3.2 \times 10^{-4}$, $r = 0.24$; Table 3.1). We also determined the correlation between the amount of DNA input for amplicon enrichment and amplicon yield using Spearman's rank correlation, which demonstrated

weak correlations between the two variables for both blood and FFPE specimens (blood, r_s = 0.29, 95% CI = 0.16–0.41, $p = 2.1 \times 10^{-5}$; FFPE tumour, 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 ($\log2\frac{\text{Amplicon Yield}}{\text{DNA Input}}$), which represents the efficiency in amplicon enrichment. We compared the log2 fold change in FFPE specimens to blood specimens, and a Wilcoxon signed-rank test indicated that the median log2 fold change (median = 1.04) in blood specimens was significantly higher than the median log2 fold change (median = -0.332) in FFPE specimens (W = 24754, Z = 12.7, $p = 4.6 \times 10^{-57}$, r = 0.86; Figure 3.1C, Table 3.1). This result implies that amplicon generation is less efficient in FFPE specimens in comparison to blood specimens, demonstrating the impact of formalin fixation on an amplicon enrichment approach.

To examine whether blood and FFPE specimens produce comparable sequencing results, we compared read alignments between blood and FFPE specimens. For the percentage of on-target aligned reads, which are reads that align to target regions used for variant calling, we found no significant difference between blood and FFPE specimens (Wilcoxon signed-rank test, W = 10 178.5, Z = -1.69, p = 0.091, r = -0.11; Figure 3.2, Table 3.2). 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 demonstrated between specimen types (Wilcoxon signed-rank test, W = 11 494.5, Z = -0.359, p = 0.72, r = -0.024; Figure 3.2, Table 3.2). However, a Wilcoxon signed-rank test indicated that the distribution of percentage of unaligned reads was significantly different between blood and FFPE specimens (W = 19 069, Z = 7.82, $p = 2.4 \times 10^{-16}$, r = 0.53; Figure 3.2, Table 3.2). We also found a significant difference in the percentage of contaminant reads between specimen types (W = 14 877, Z = 3.29, $p = 9.2 \times 10^{-4}$, r = 0.22; Figure 3.2, Table 3.2). Although sequencing libraries generated from blood and FFPE DNA differed in percentages of unaligned and contaminant reads, blood and FFPE libraries resulted in comparable percentage of on-target aligned reads, which provides equivalent amount of aligned reads for variant calling.

While 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 we 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. A Wilcoxon signed-rank test indicated that the median of average per base coverage (median = 1271) in blood specimens was significantly higher than the median of average per base coverage (median = 1194) in FFPE specimens (W = 20 864, Z = 9.76, $p = 2.5 \times 10^{-26}$, r = 0.66). 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. We found significant differences in coverage uniformity between blood and FFPE specimens at all cov-

erage levels except at the zero coverage cut-off (Wilcoxon signed-rank test, $p < 0.05$; Figure 3.3, Table 3.3). These findings reveal that the coverage depth of target bases in FFPE specimens is lower and less uniform compared to blood specimens, indicating that formalin fixation could interfere with the ability to attain comparable sequencing coverage between FFPE and blood DNA.

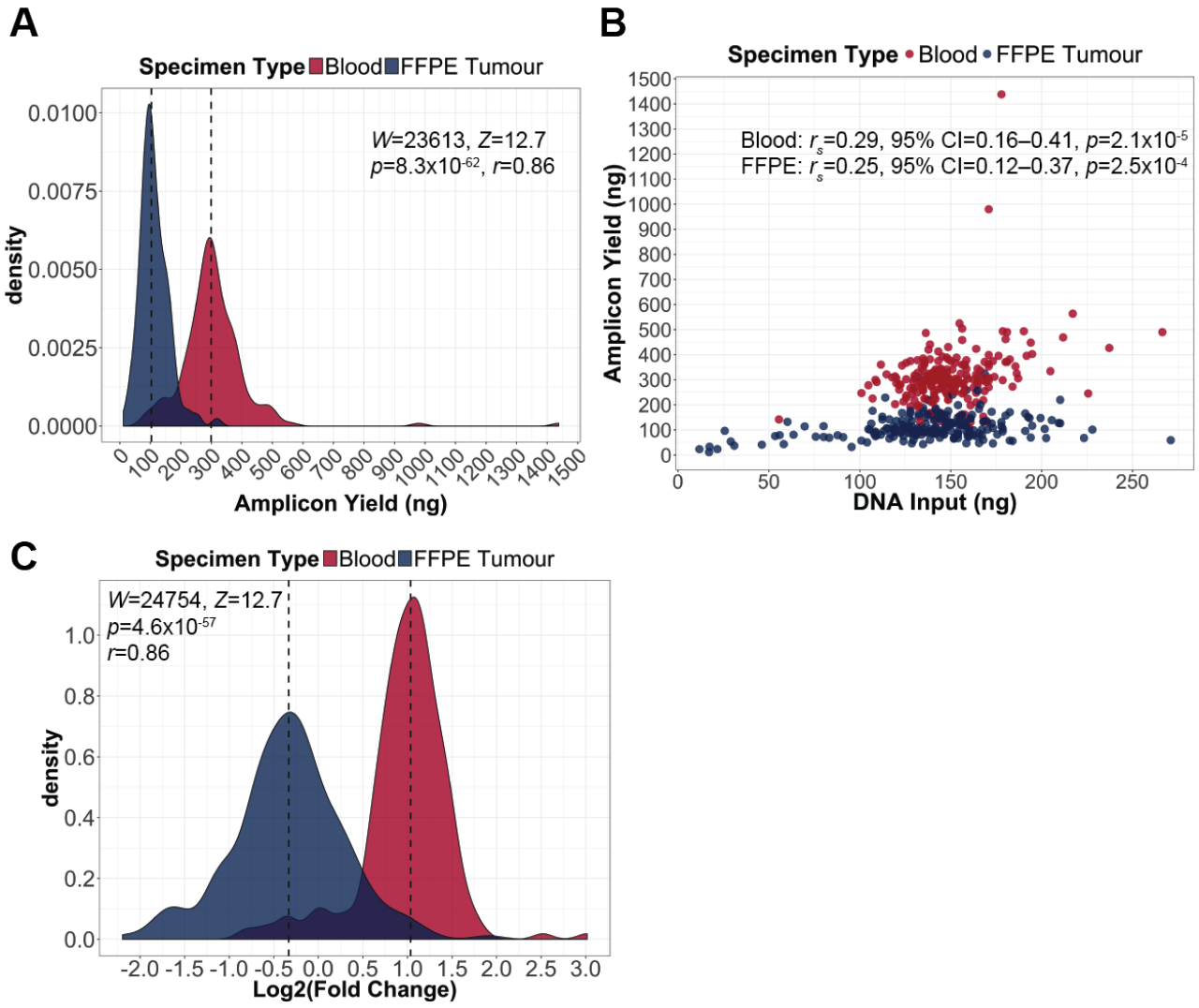

Figure 3.1: Comparison of efficiency in amplicon enrichment between blood and FFPE specimens. (A) Median amplicon yield is significantly lower in FFPE specimens than blood based on the Wilcoxon signed-rank test ($p < 0.05$). Dashed lines indicate median amplicon yield for blood and FFPE specimens. (B) Amplicon yield and DNA input for amplicon generation are weakly correlated in blood and FFPE specimens as demonstrated by Spearman's rank correlation ($p < 0.05$). (C) Efficiency of amplicon generation is represented by the log2 fold change between DNA input and amplicon yield (ratio of amplicon yield to DNA input). Median log2 fold change is significantly lower in FFPE specimens compared to blood as shown by the Wilcoxon signed-rank test ($p < 0.05$). Dashed lines indicate median log2 fold change for blood and FFPE specimens.

Table 3.1: Assessment of amplicon enrichment results in blood and FFPE specimens. p value indicates significance level for Wilcoxon signed-rank test.

	Blood		FFPE Tumour		
Parameter	Median	Range	Median	Range	p ($< 0.05^*$)
Amplicon Yield (ng)	299.3	84.0–1438.0	103.6	11.6–325.5	$8.3 \times 10^{-62*}$
DNA Input (ng)	147.8	55.5–266.4	140.9	11.8–271.0	$3.2 \times 10^{-4*}$
$\log2(\frac{\text{Amplicon Yield}}{\text{DNA Input}})$	1.04	-0.845–3.01	-0.332	-2.20–1.90	$4.6 \times 10^{-57*}$

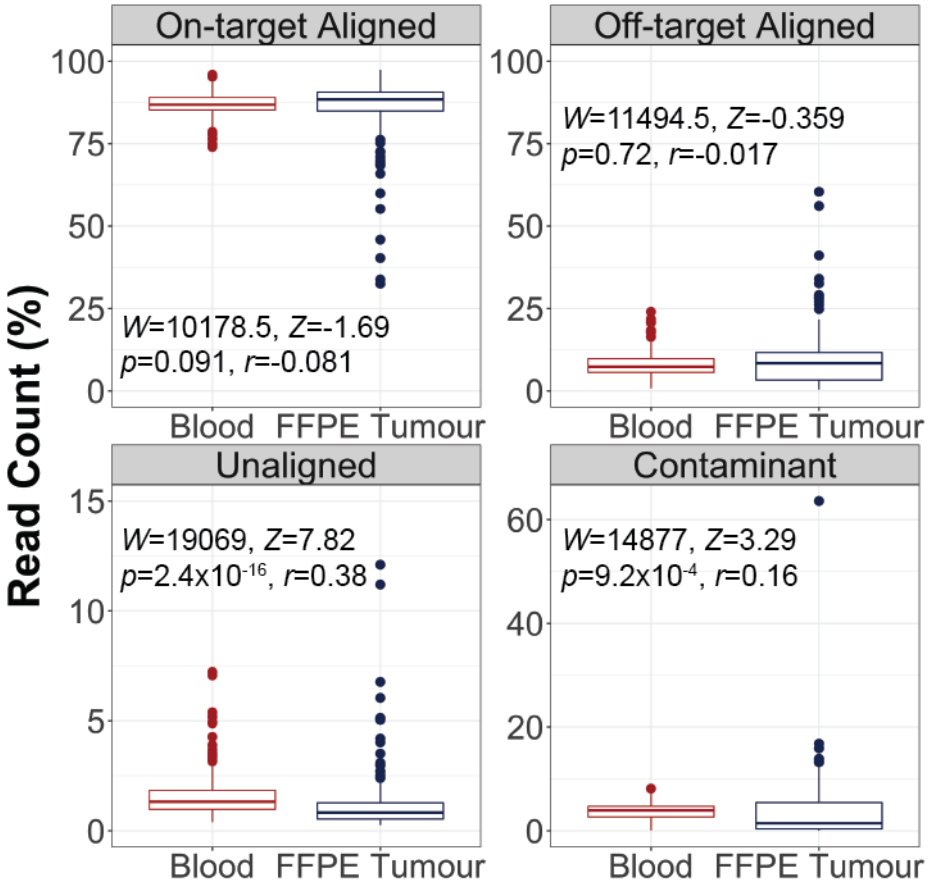

Figure 3.2: Assessment of read alignments between blood and FFPE specimens. Percentages of on-target and off-target aligned reads are not significantly different between specimen types as shown by the Wilcoxon signed-rank test, whereas percentages of unaligned and contaminant reads are significantly different ($p < 0.05$). 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.

Table 3.2: Comparison of read alignments between blood and FFPE specimens. p value indicates significance level for Wilcoxon signed-rank test.

	Blood		FFPE Tumour		
Parameter	Median	Range	Median	Range	p ($< 0.05^*$)
On-target Aligned Reads (%)	86.8	74.0–95.9	88.4	32.5–97.4	0.091
Off-target Aligned Reads (%)	7.3	0.8–24.0	8.4	0.4–60.4	0.72
Unaligned Reads (%)	1.3	0.4–7.2	0.8	0.3–12.1	$2.4 \times 10^{-16*}$
Contaminant (%)	3.9	0.1–8.1	1.4	0.03–63.6	$9.2 \times 10^{-4*}$

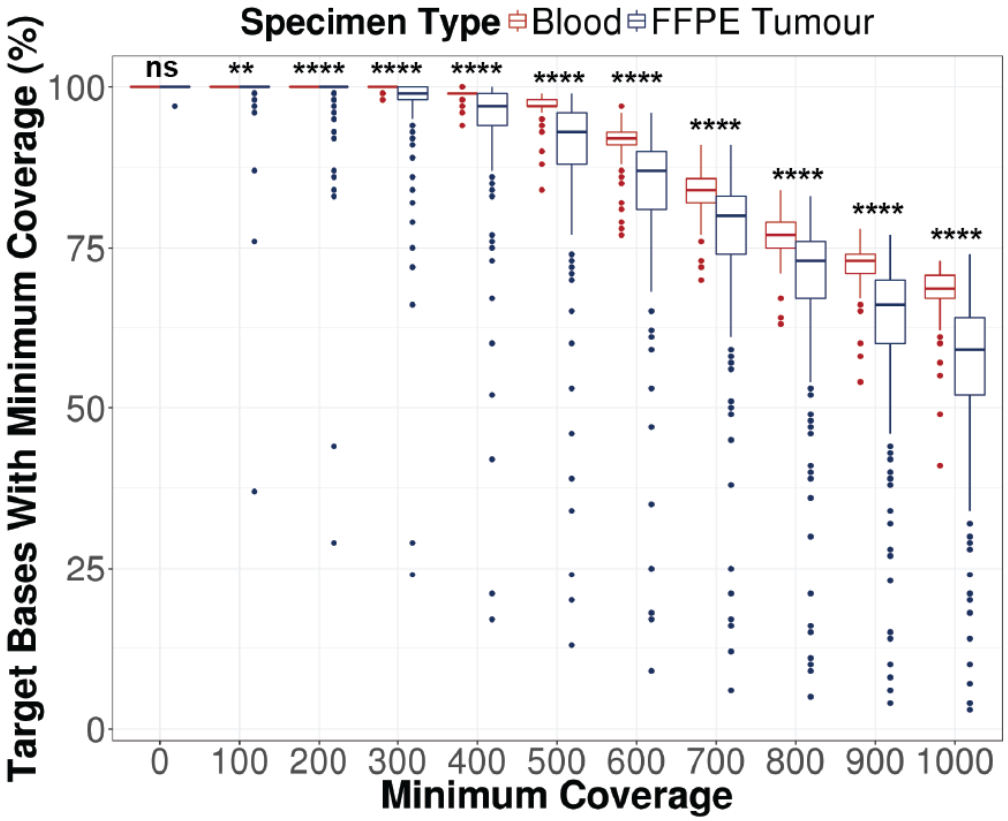

Figure 3.3: Evaluation of coverage uniformity in blood and FFPE specimens. Per base coverage was normalized to account for difference in library size. Percentage of target bases that met various coverage thresholds was calculated. Wilcoxon signed-rank test showed significant differences in percentage of target bases at all coverage thresholds except at the zero coverage cut-off (****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns = not significant). Box plots show the median (horizontal bar within) and interquartile range (IQR) of percentage of target bases that meet the respective coverage thresholds, with whiskers representing the range of data not exceeding 1.5x the IQR and circles indicating outliers.

Table 3.3: Assessment of coverage uniformity between blood and FFPE specimens using the Wilcoxon signed-rank test.

	Blood		FFPE Tumour					
Threshold	Median (%)	Range (%)	Median (%)	Range (%)	W	Z	p ($< 0.05^*$)	r
$\geq$ 0x	100.0	100.0–100.0	100.0	97.0–100.0	1	1.00	1.0	0.068
$\geq$ 100x	100.0	100.0–100.0	100.0	37.0–100.0	91	3.61	$2.3 \times 10^{-4*}$	0.25
$\geq$ 200x	100.0	100.0–100.0	100.0	29.0–100.0	666	5.99	$2.9 \times 10^{-11*}$	0.41
$\geq$ 300x	100.0	98.0–100.0	99.0	24.0–100.0	7696	8.17	$4.1 \times 10^{-18*}$	0.55
$\geq$ 400x	99.0	94.0–100.0	97.0	17.0–100.0	13934	10.0	$5.0 \times 10^{-28*}$	0.68
$\geq$ 500x	97.0	84.0–99.0	89.5	13.0–99.0	19880.5	11.3	$2.1 \times 10^{-38*}$	0.77
$\geq$ 600x	92.0	77.0–97.0	87.0	9.0–96.0	20762	10.6	$1.5 \times 10^{-32*}$	0.72
$\geq$ 700x	84.0	70.0–91.0	80.0	6.0–91.0	18458.5	9.54	$5.7 \times 10^{-25*}$	0.65
$\geq$ 800x	77.0	63.0-84.0	73.0	5.0–83.0	18127	9.87	$4.7 \times 10^{-27*}$	0.67
$\geq$ 900x	73.0	54.0–78.0	66.0	4.0–77.0	20706	11.5	$4.6 \times 10^{-40*}$	0.78
$\geq$ 1000x	68.5	41.0–73.0	59.0	3.0-74.0	21054.5	11.7	$3.6 \times 10^{-42*}$	0.79

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 clinically relevant hotpots of 21 genes, and these amplicons vary in length and GC content. Since we observed discrepancy in sequencing coverage between blood and FFPE specimens, we sought to determine whether this discrepancy is influenced by amplicon length and GC content. We attained the coverage depth for each amplicon and normalized the coverage depth to account for difference in library size. We found significant difference 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 discrepancy in amplicon coverage depth between specimen types, we derived the log2 fold change from the median coverage depth in blood specimens to the median coverage depth in FFPE specimens ($\log2\frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) 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. Our assessment showed that 217 out of the 331 amplicons have negative log2 fold changes, whereas 114 out of the 331 amplicons have positive log2 fold changes (Figure 3.4). These results indicate that there are differences in coverage depth between FFPE and blood specimens for the majority of amplicons in the panel, with 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 discrepancy

in amplicon coverage depth between FFPE and blood 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). We then explored the correlation between the median coverage depth of amplicons and amplicon length in both blood and FFPE specimens (Figure 3.6A). While we found no significant correlation between median coverage depth and amplicon length in blood specimens (Pearson's correlation, r = 0.042, 95% CI = -0.054–0.14, p = 0.39), we observed a negative correlation in FFPE specimens (Pearson's correlation, r = -0.28, 95% CI = -0.37– -0.19, $p = 7.4 \times 10^{-9}$). We evaluated the relationship between the discrepancy in amplicon coverage depth between FFPE and blood specimens, which was calculated as the log2 fold change from the median coverage depth in blood specimens to the median coverage depth in FFPE specimens ($\log2\frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$), and amplicon length (Figure 3.6B). Pearson's correlation demonstrated a strong, negative correlation between the log2 fold change in amplicon coverage depth and amplicon length (r = -0.79, 95% CI = -0.82– -0.75, $p = 1.4 \times 10^{-88}$), indicating that the decline in amplicon coverage depth in FFPE specimens relative to blood specimens becomes larger as amplicon length increases.

To assess the effect of amplicon GC content on coverage depth of amplicons, we explored the correlation between the median coverage depth of amplicons and amplicon GC content in blood and FFPE specimens (Figure 3.7A). Using Pearson's correlation, we identified weak, negative correlations between median coverage depth and amplicon GC content in both blood and FFPE specimens (blood: r = -0.16, 95% CI = -0.25– -0.067, $p = 8.7 \times 10^{-4}$; FFPE: r = -0.29, 95% CI = -0.38– -0.20 $p = 1.0 \times 10^{-9}$). We next assessed the relationship between the discrepancy in amplicon coverage depth between FFPE and blood specimens ($\log2\frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) and amplicon GC content (Figure 3.7B). In contrast to the strong, negative correlation observed for the log2 fold change in amplicon coverage depth in relation to amplicon length, Pearson's correlation demonstrated a weak, negative correlation between the log2 fold change in amplicon coverage depth and amplicon GC content (r = -0.31, 95% CI = -0.40– -0.22, $p = 1.1 \times 10^{-10}$). While the correlation is weak, this finding still implies that increased amplicon GC content has a significant impact on the decrease in amplicon coverage depth in FFPE specimens relative to blood specimens.

Because amplicon length and GC content demonstrated significant effects on the discrepancy in amplicon coverage depth between FFPE and blood specimens, we sought to determine which contributing factor has a greater effect. We used a multiple linear regression to predict log2 fold change in amplicon coverage depth between blood and FFPE specimens ($\log2\frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) based on amplicon length and GC content (Table 3.4). A significant equation was found ($F(2, 411) = 471$, $p = 4.65 \times 10^{-107}$), with an adjusted R^2 of 0.695. Predicted fold change in amplicon coverage depth between blood and FFPE specimens (log2) is equal to $1.66\text{-}7.24 \times 10^{-3}$(*Amplicon Length*)-$9.92 \times 10^{-3}$(*Amplicon 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 log2 fold change in amplicon coverage depth between blood and FFPE specimens. 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 log2 fold change in amplicon coverage depth between blood and FFPE specimens, whereas one standard deviation increase in amplicon GC content would lead to a -0.277 standard deviation decrease in the response variable. This result indicates that amplicon length has a stronger association with the fold change in amplicon coverage depth between blood and FFPE specimens (log2) than amplicon GC content. Together, these findings reflect the challenges imposed by fragmentation damages in FFPE DNA: formalin fixation induces DNA fragmentation, resulting in shorter template DNA that would be more difficult for PCR amplification to generate longer amplicons.

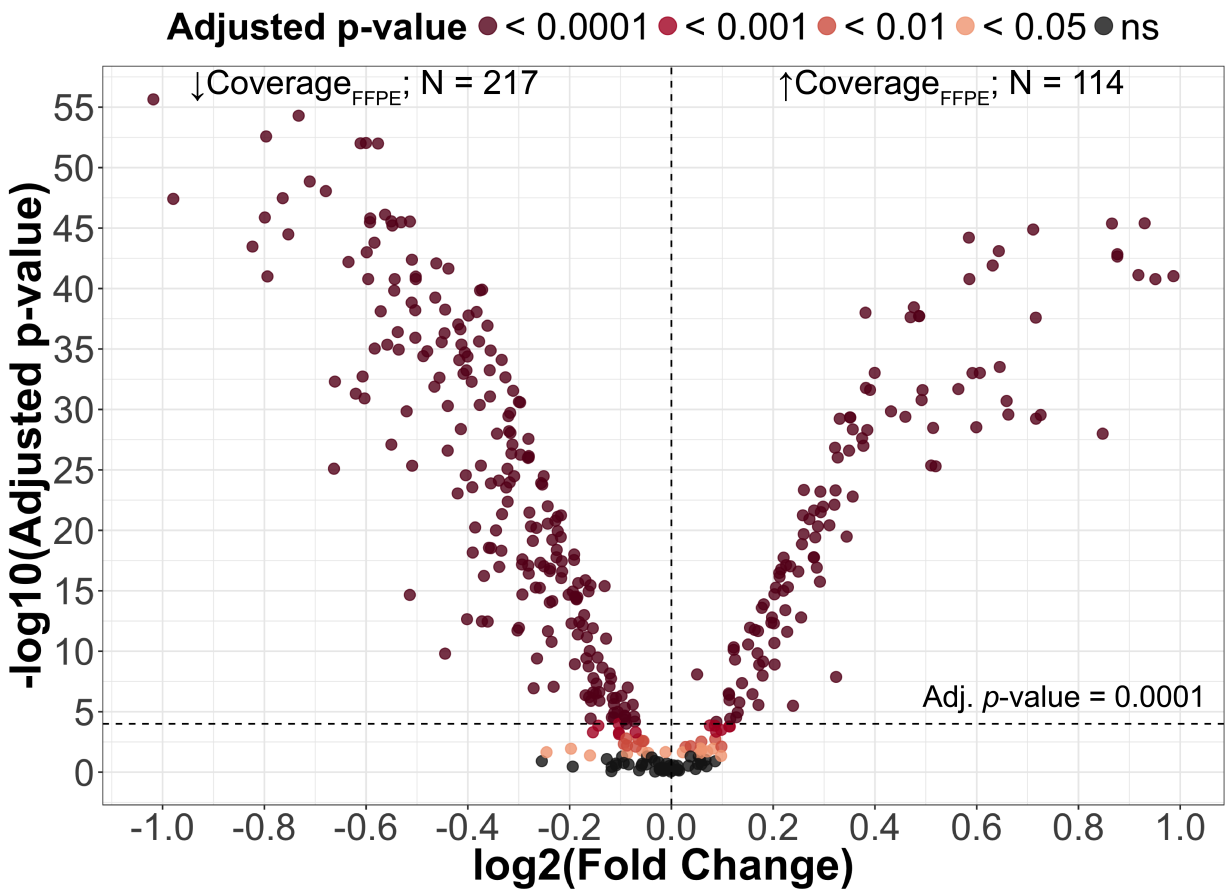

Figure 3.4: Coverage depth is significantly different between FFPE and blood specimens for the majority of amplicons in the OncoPanel (Wilcoxon signed-rank test with Benjamini-Hochberg correction, adjusted $p < 0.0001$). Amplicon coverage depth was normalized to account for difference in library size and log2 fold change between the median coverage depth in blood and FFPE specimens ($\log2\frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) was calculated for each amplicon. Volcano plot illustrates -log10 adjusted p-value in relation to log2 fold change. Negative log2 fold change indicates lower amplicon coverage depth in FFPE specimens than blood specimens (↓ Coverage$_{FFPE}$), whereas positive log2 fold change indicates higher amplicon coverage depth in FFPE specimens than blood specimens (↑ Coverage$_{FFPE}$). N = number of amplicons.

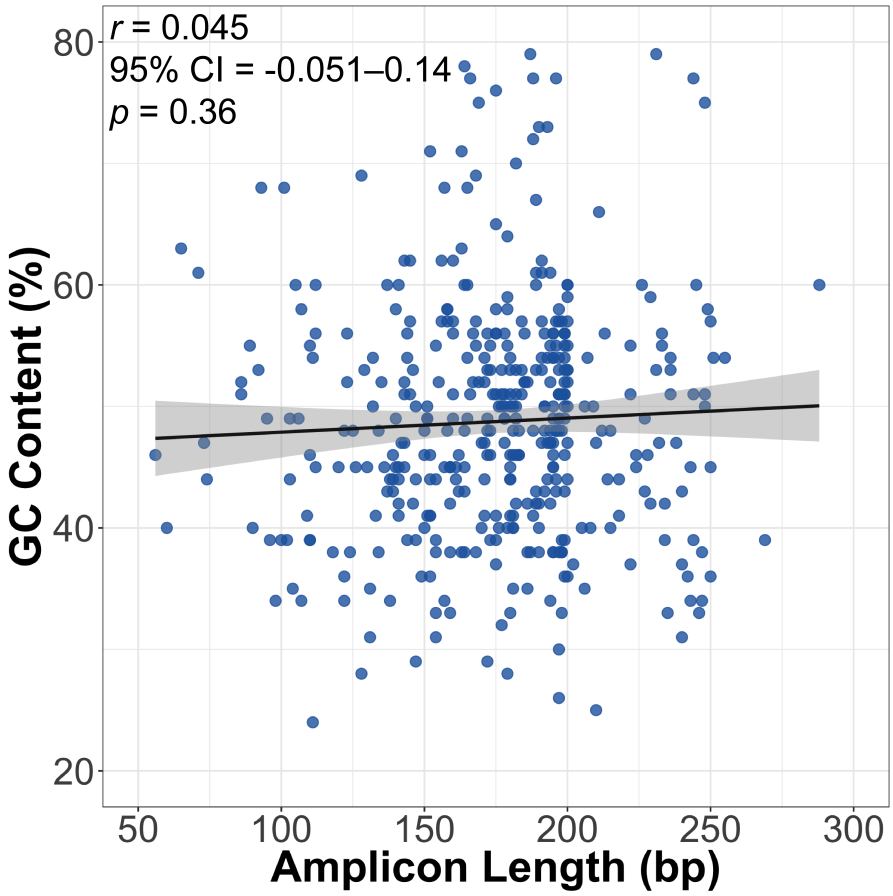

Figure 3.5: The relationship between amplicon GC content (%) and amplicon length (bp). Pearson's correlation showed that amplicon GC content is not significantly correlated with amplicon length ($p < 0.05$).

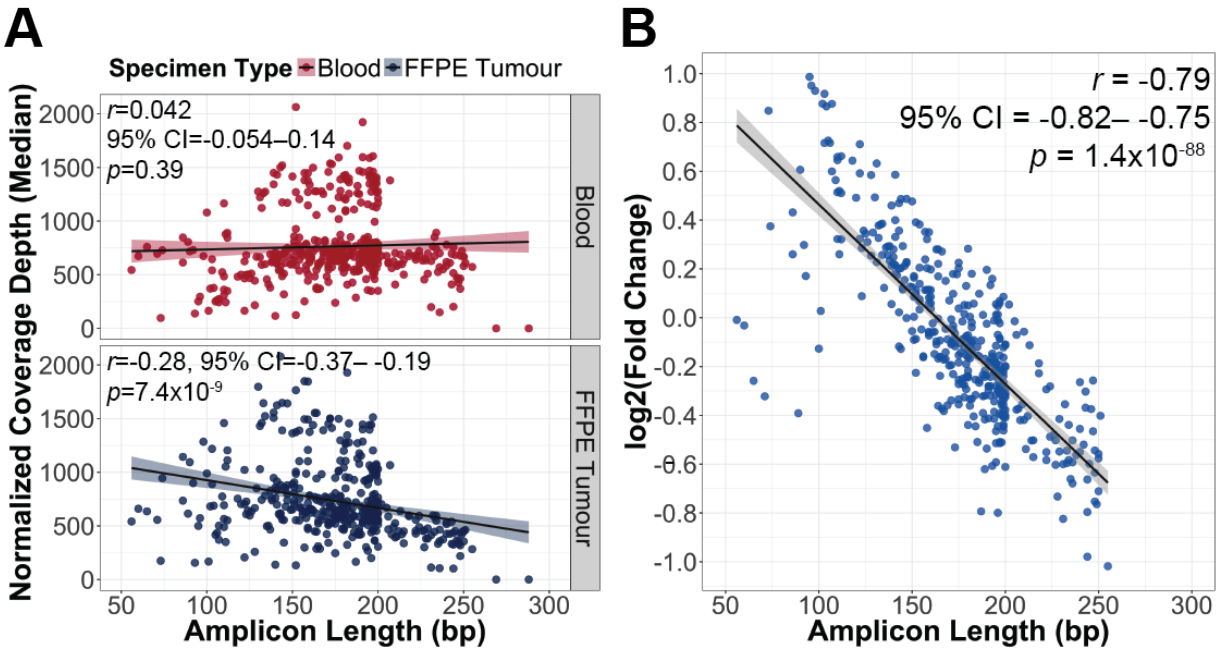

Figure 3.6: The effect of amplicon length on coverage depth of amplicons. Coverage depth of amplicons was normalized to account for difference in library size and log2 fold change between the median coverage depth in blood and FFPE specimens ($\log 2 \frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) was calculated for each amplicon. (A) No significant correlation between coverage depth of amplicons and amplicon length was demonstrated in blood specimens, whereas coverage depth of amplicons is negatively correlated with amplicon length in FFPE specimens (Pearson's correlation, $p < 0.05$). (B) Increased in amplicon length leads to lower log2 fold change in amplicon coverage depth between blood and FFPE specimens (Pearson's correlation, $p < 0.05$).

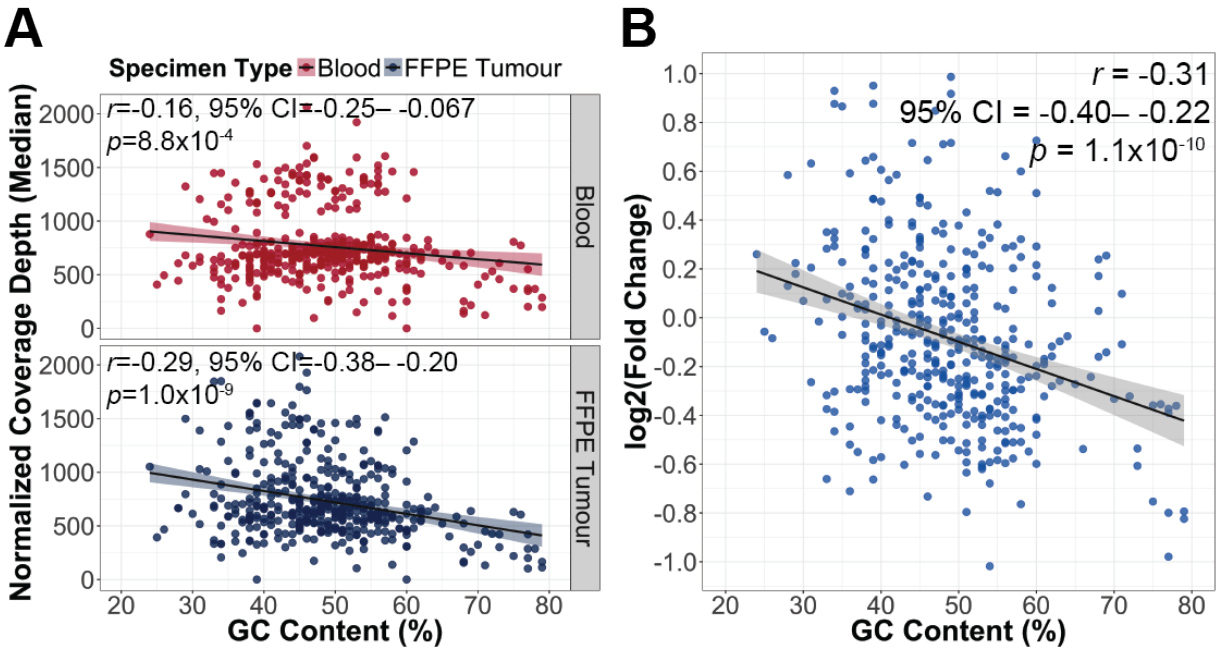

Figure 3.7: The effect of amplicon GC content on coverage depth of amplicons. Coverage depth of amplicons was normalized to account for difference in library size and log2 fold change between the median coverage depth in blood and FFPE specimens ($\log 2 \frac{\text{Median Coverage Depth in FFPE}}{\text{Median Coverage Depth in Blood}}$) was calculated for each amplicon. (A) Coverage depth of amplicons is negatively correlated with amplicon GC content in both blood and FFPE specimens (Pearson's correlation, $p < 0.05$). (B) Increased in amplicon GC content leads to lower log2 fold change in amplicon coverage depth between blood and FFPE specimens (Pearson's correlation, $p < 0.05$).

Table 3.4: Multiple linear regression to predict log2 fold change in amplicon coverage depth between blood and FFPE specimens 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 causes deamination of cytosine bases, leading to increase in C>T/G>A transitions. 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 (Phred-scaled base quality score $\geq$ 20) and base changes that were $\geq$ 1% allele frequency to avoid inclusion of sequencing errors in our analysis. Base changes were categorized into 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, and we compared the fraction of base changes between blood and FFPE specimens.

Formalin fixation not only induces DNA fragmentation, but also causes deamination of cytosine bases, leading to increase in C>T/G>A transitions. 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 (Phred-scaled base quality score $\geq$ 20) and base changes that were $\geq$ 1% allele frequency to avoid inclusion of sequencing errors in our analysis. We found significant difference in fraction of base changes

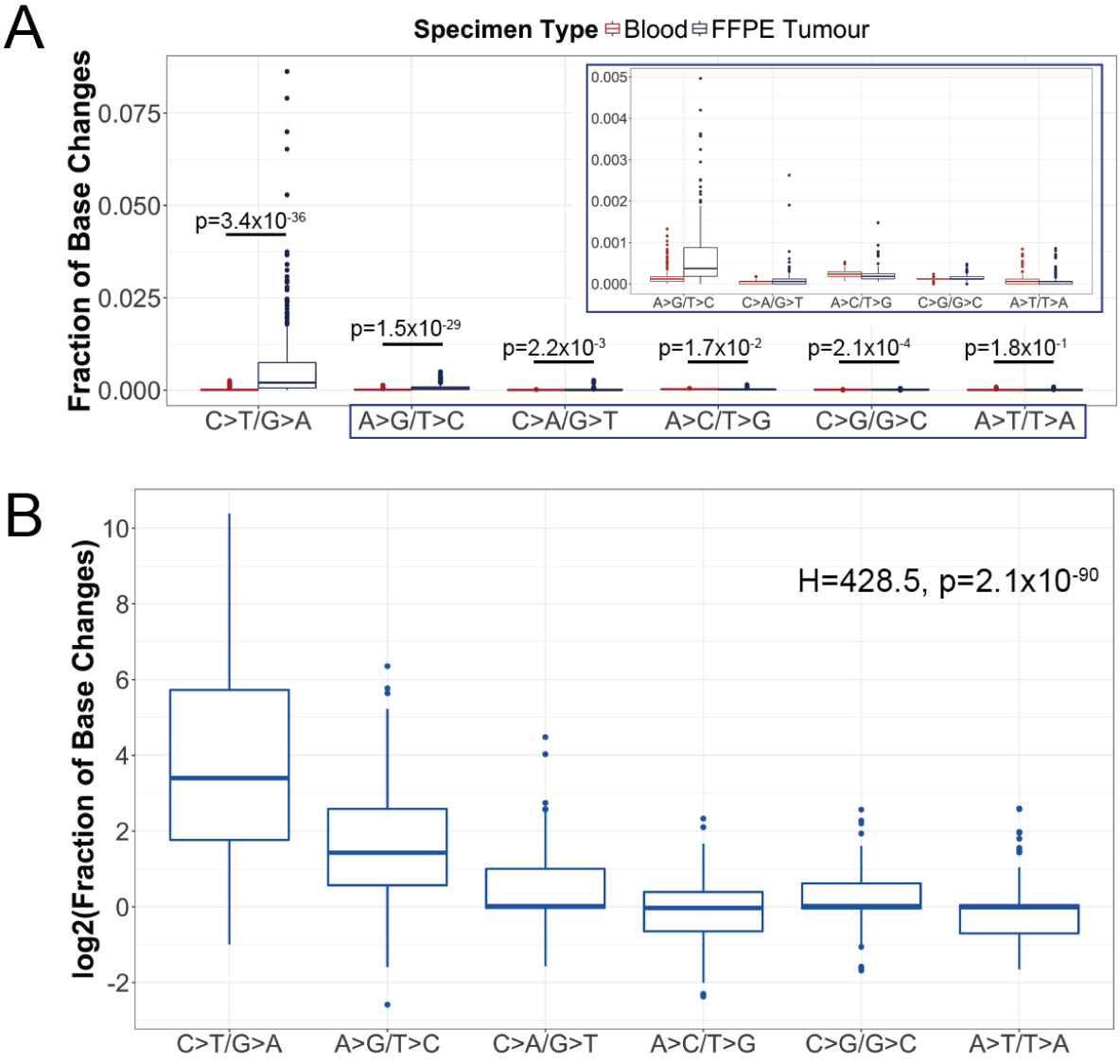

Figure 3.8: Detection of formalin-induced sequence artifacts in FFPE specimens. (A)

Table 3.5: Multiple pairwise comparison of log2 fold change in fraction of base changes between FFPE and blood 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 < 0.05.

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 $4.00 \times 10^{-4}*$	6.39 $1.52 \times 10^{-10}*$	-2.73 $3.99 \times 10^{-3}*$		
C>G/G>C	-3.02 $1.73 \times 10^{-3}*$	8.63 $6.76 \times 10^{-18}*$	-2.17 $1.71 \times 10^{-2}*$	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 $6.46 \times 10^{-45}*$

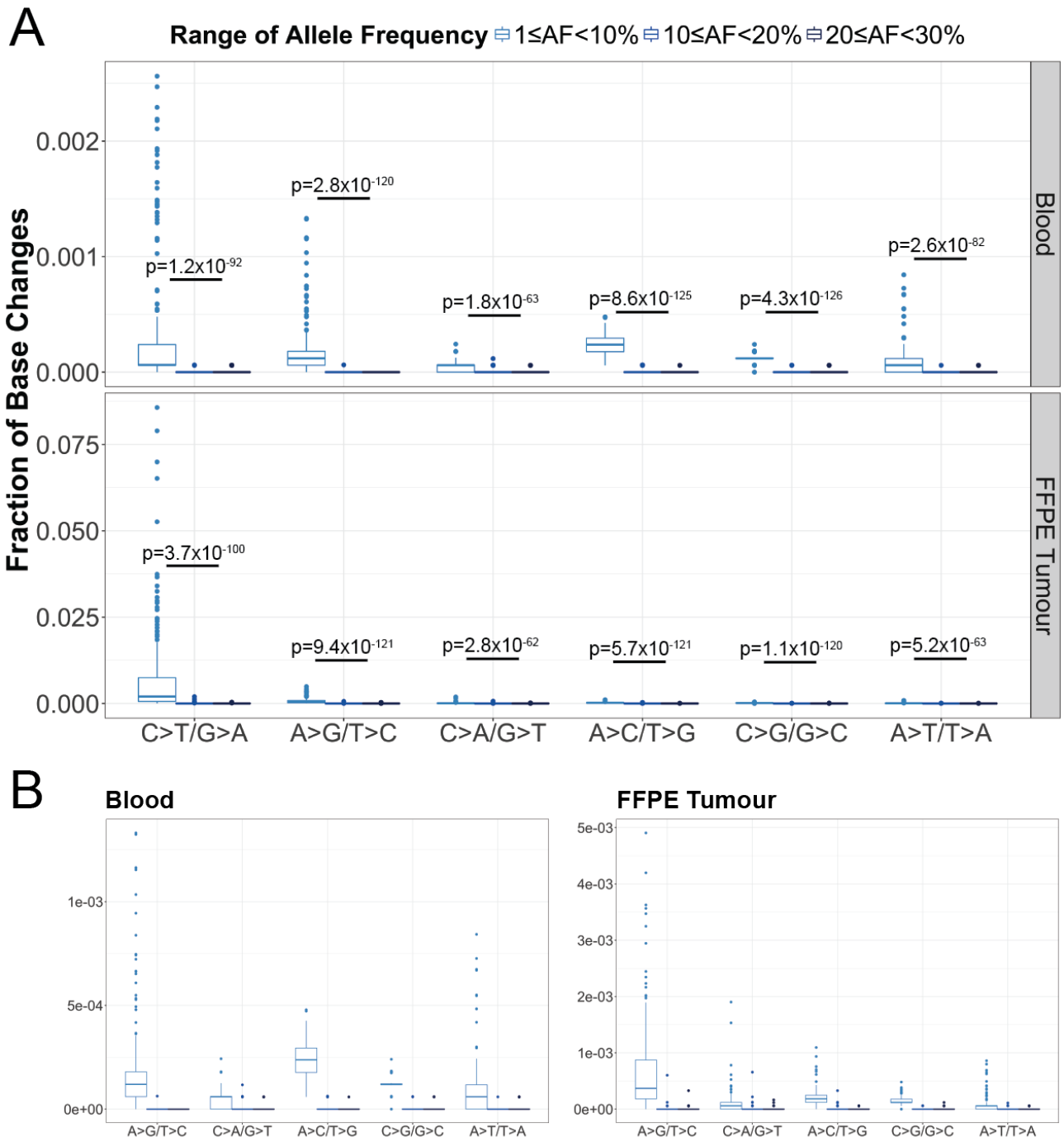

Figure 3.9: Add caption.

3.4 Increased age of paraffin block results in reduced amplicon yield and elevated events 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 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.

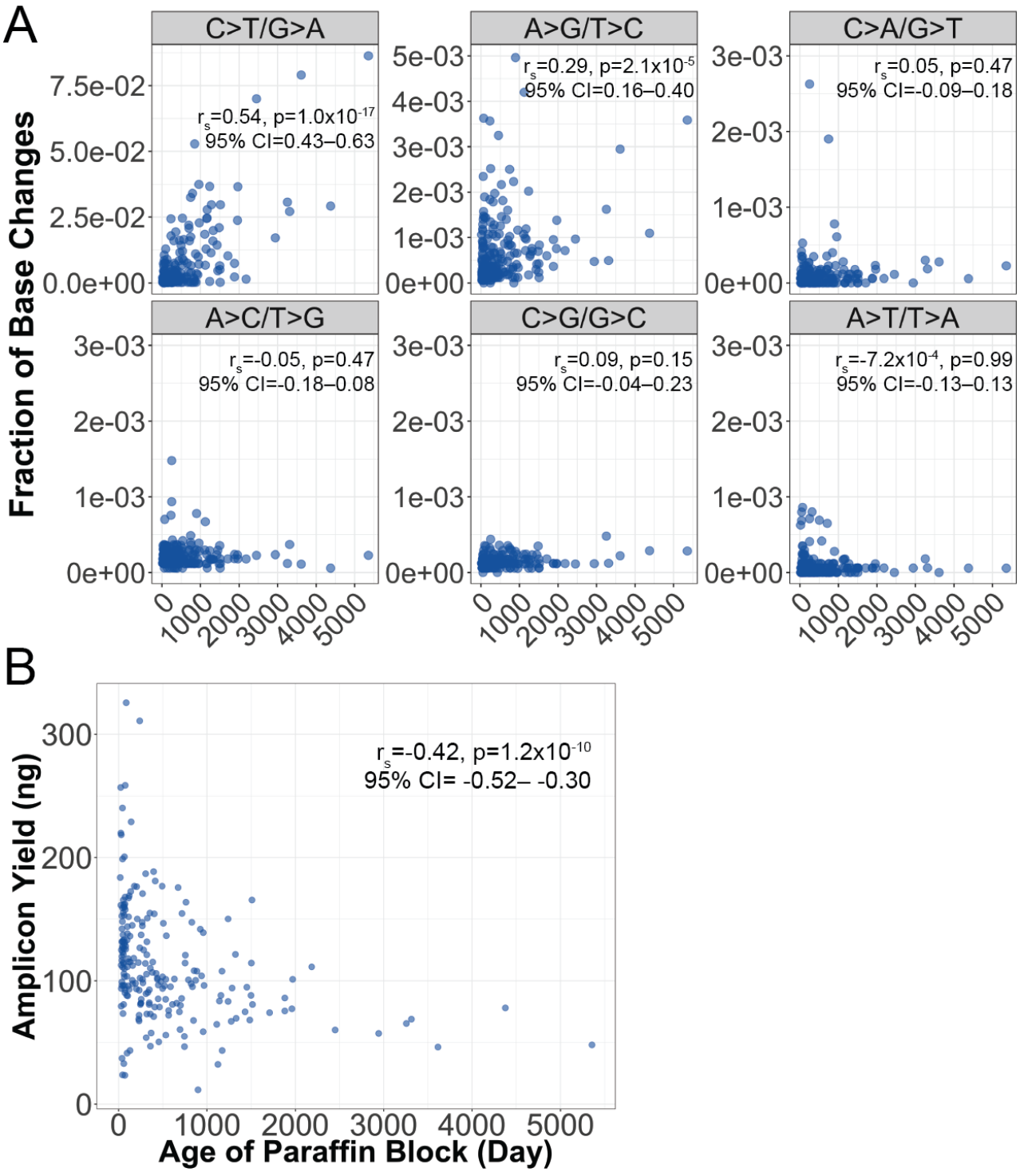

Figure 3.10: Add caption.

Table 3.6: Determination of correlation between pre-sequencing variables and sequencing results using Spearman's correlation. 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	Amplicon Yield (ng)	Age of Paraffin Block (Day)	Fraction of C>T/G>A	Average Per Base Normalized Coverage
Age of Paraffin Block (Day)	-0.42 (-0.52– -0.30) $5.2 \times 10^{-7}*$			
Fraction of C>T/G>A	-0.72 (-0.77– -0.65) $1.9 \times 10^{-11}*$	0.54 (0.61–0.75) $6.3 \times 10^{-35}*$		
Average Per Base Normalized Coverage	0.69 (0.61–0.75) $8.5 \times 10^{-20}*$	-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 (%)	0.58 (0.48–0.66) $2.1 \times 10^{-13}*$	-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) $3.1 \times 10^{-58}*$

3.5 Non-reproducible variant calls are detected in sequencing replicates of FFPE specimens

3.6 Discussion

Chapter 4

True Positive Rates of Germline Variant Calling in FFPE Tumours

Short recap of objectives ...

4.1 Frequency of germline and somatic variants

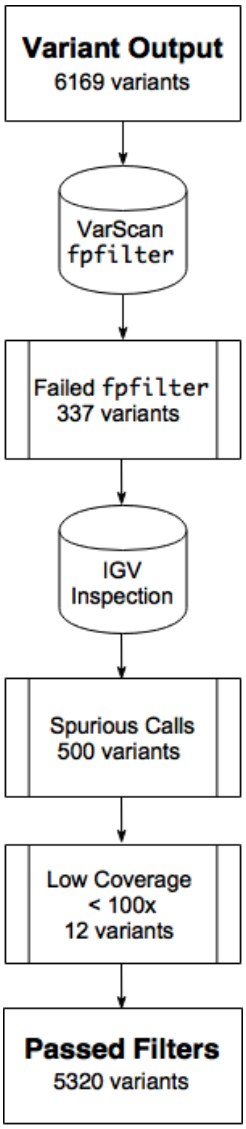

Figure 4.1: Add caption.

4.2 Germline variants are highly concordant between blood and FFPE specimens

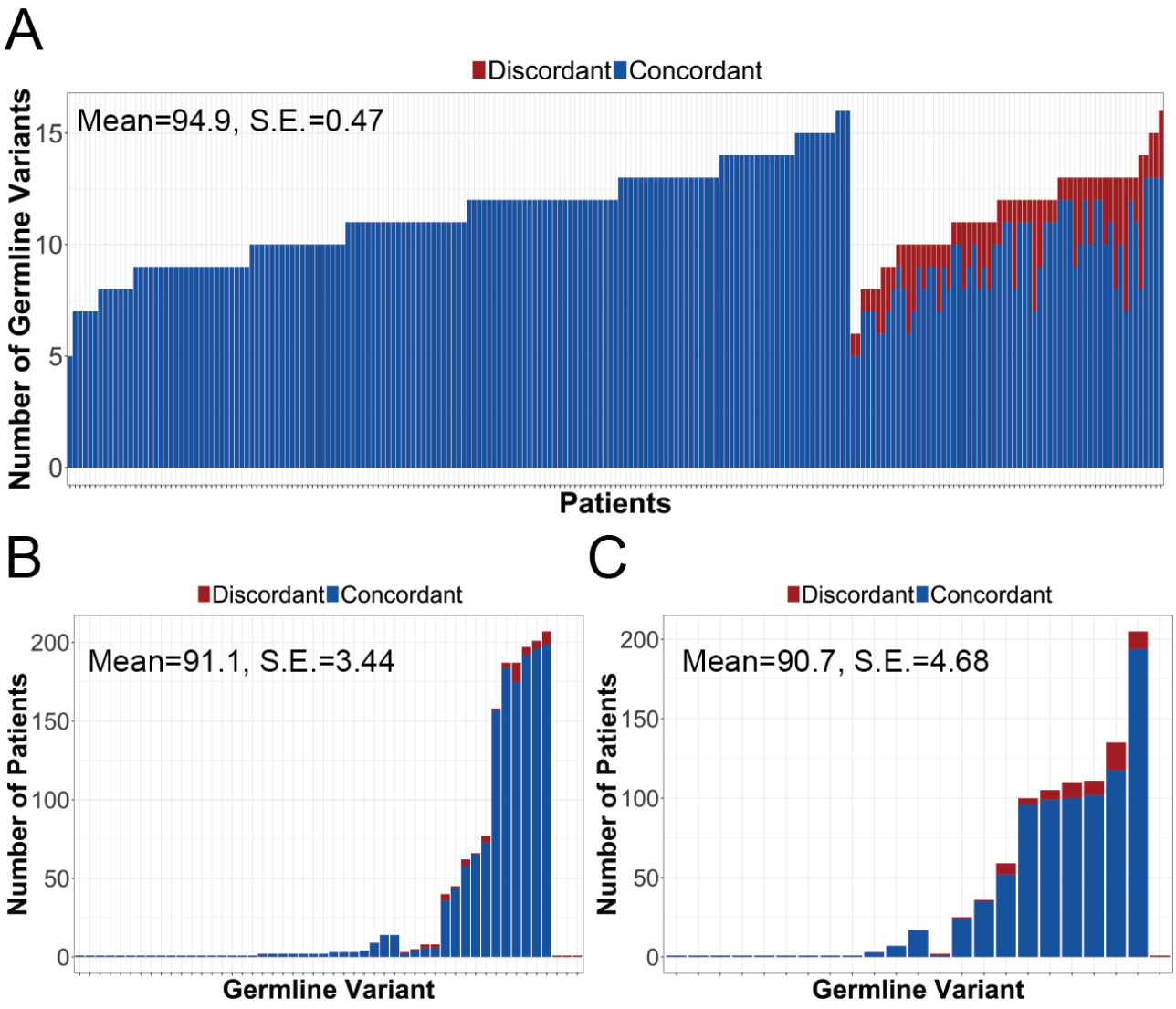

Figure 4.2: Add caption.

4.3 Reduced sensitivity is observed for detection of germline variants in FFPE specimens compared to blood

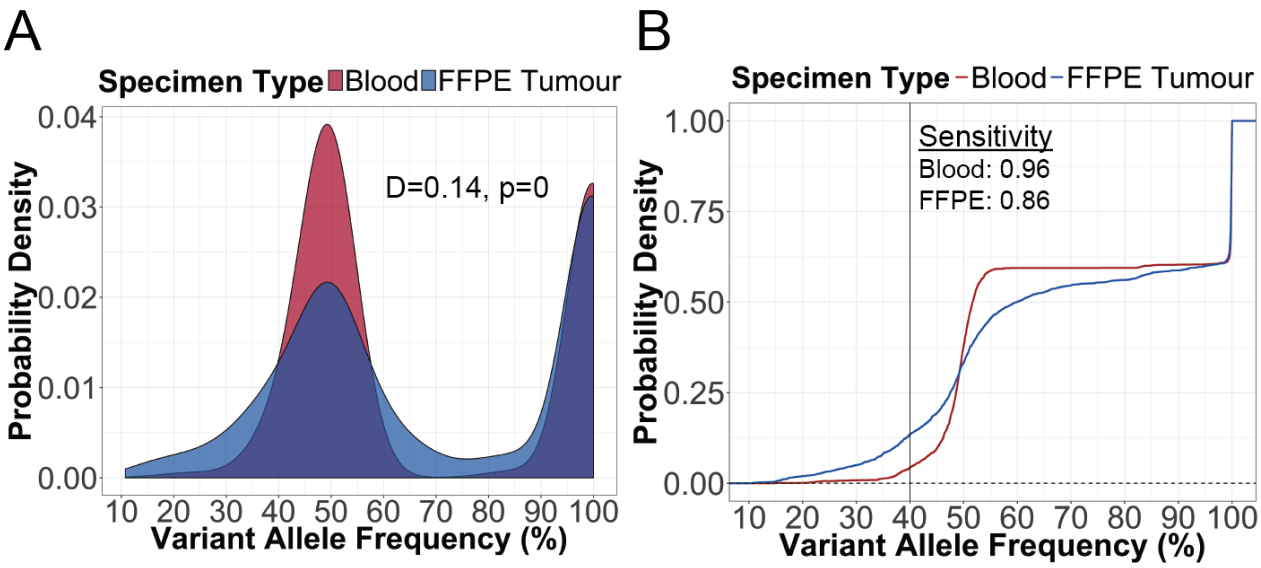

Figure 4.3: Add caption.

Table 4.1: Sensitivity of detecting germline variants in blood and FFPE specimens at various variant allele frequency thresholds.

	Blood				FFPE Tumour			
VAF (%)	FN*	TP**	Sensitivity	95% CI	FN*	TP**	Sensitivity	95% CI
10	0	2461	1.0	1.0–1.0	0	2428	1.0	1.0–1.0
15	2	2459	1.0	1.0-1.0	12	2416	1.0	0.99–1.0
20	3	2458	1.0	1.0–1.0	48	2380	0.98	0.97–0.99
25	15	2446	0.99	0.99–1.00	79	2349	0.97	0.96–0.97
30	20	2441	0.99	0.99–1.00	121	2307	0.95	0.94–0.96
35	33	2428	0.99	0.98–0.99	197	2231	0.92	0.91–0.93
40	107	2354	0.96	0.95–0.96	328	2100	0.86	0.85–0.88
45	234	2227	0.90	0.89–0.92	470	1958	0.81	0.79–0.82

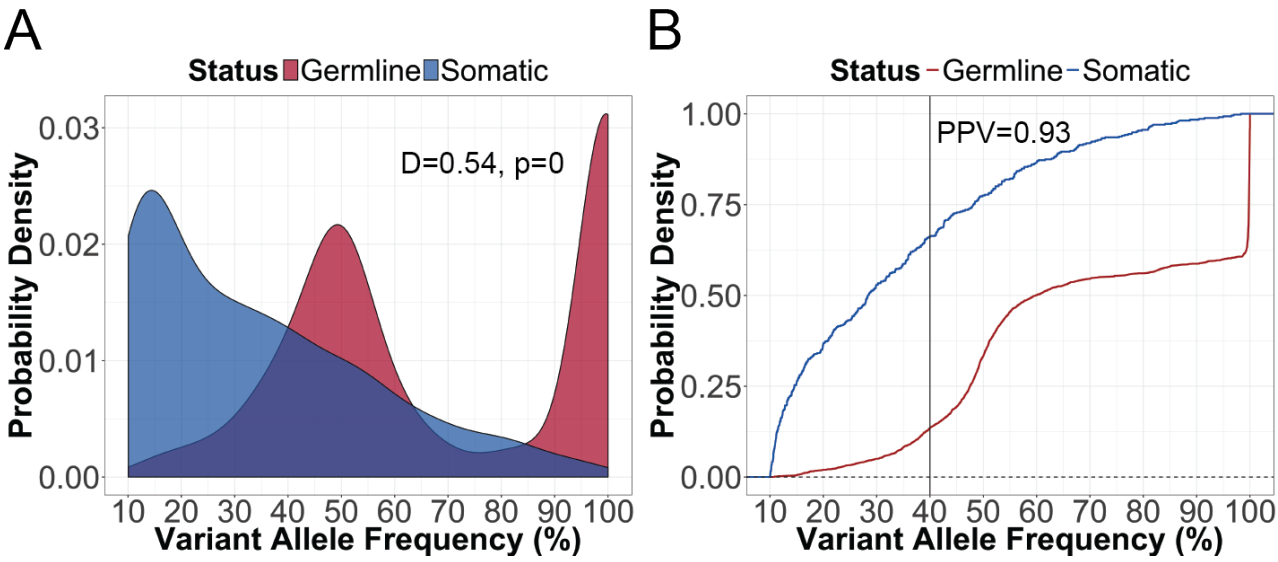

Figure 4.4: Add caption.

Table 4.2: Positive predictive value for referral of potential germline variants for downstream confirmatory testing.

VAF (%)	False Positive	True Positive	Total Calls	Positive Predictive Value	95% CI
10	431	2428	2859	0.85	0.84–0.86
15	319	2416	2735	0.88	0.87–0.90
20	273	2380	2653	0.90	0.88–0.91
25	245	2349	2594	0.91	0.89–0.92
30	203	2307	2510	0.92	0.91–0.93
35	178	2231	2409	0.93	0.91–0.94
40	146	2100	2246	0.93	0.92–0.94
45	118	1958	2076	0.94	0.93–0.95

4.4 Factors underlying reduced sensitivity of germline variant calling in FFPE specimens

4.5 Discussion

Chapter 5

Conclusion

What are my conclusions?

Bibliography

Appendix A

Supporting Materials

Table A.1: Gene reference models for HGVS nomenclature.

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

Table A.2: Target regions and amplicons of the OncoPanel.