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)

The University of British Columbia
(Vancouver)

December 2017

Abstract

Germline alterations can have clinical implications for both cancer patients and their families. Because the tumour genome may contain both germline and somatic variants, the increasingly common practice of clinical tumour sequencing presents an opportunity to also pre-screen for germline variants. This framework is time- and cost-effective because only patients with potential germline variants are referred to downstream confirmatory testing. However, a key challenge is that tumour specimens are commonly formalin-fixed and paraffin-embedded (FFPE), which induces DNA damage that may interfere with molecular testing. Another challenge is distinguishing between germline and somatic variants in the tumour in order to accurately select candidates for follow-up screening. In order to leverage tumour sequencing for identifying germline variants, these challenges must be addressed.

To this end, we retrospectively analyzed clinical amplicon sequencing data from 213 patients with a range of tumours, for whom matched-normal samples were available. We assessed formalin-induced DNA damage by comparing amplicon enrichment and sequencing results of FFPE DNA to the matched-normal DNA isolated from peripheral blood mononuclear cells, a gold standard for germline testing. Although formalin-induced DNA fragmentation and cytosine deamination were detectable, we determined that the discrepancies were minor and could be mitigated by using shorter amplicons and enriching for longer DNA templates. We also found that 98.0% of germline alterations identified in the blood were retained in the tumours, suggesting that FFPE tumour DNA can be a reliable source for germline variant calling. Finally, we applied variant allele frequency (VAF) thresholds to delineate germline and somatic variants in tumour-only analyses. We reported that a VAF cut-off of 15% would correctly identify 99% of germline alterations in FFPE tumours, but erroneously submit 14% of somatic mutations (false positives) for follow-up germline testing. This underscores the high sensitivity and positive predictive value of using VAF to discriminate between germline and somatic variants. Collectively, our results demonstrate that clinical tumour amplicon sequencing could also be used to provide cost-efficient first-line germline testing.

Lay Summary

Hereditary genetic changes have clinical impacts on cancer patients and their families. Tumours contain tumour-specific and inherited genetic variations. Using tumour DNA for pre-screening of hereditary variants is time- and cost-saving because only patients with potential hereditary variants require follow-up. Follow-up testing involves analyzing blood or saliva to confirm the presence of the potential hereditary variants before making clinical decisions. A key challenge in implementing this approach is differentiating between tumour-specific and inherited variants in the tumour DNA. Furthermore, the commonly-used tumour fixative, formalin, induces DNA damage, which interferes with using tumour DNA for genetic testing. We showed that the effects of formalin on tumour DNA were minor, and we established a highly sensitive and precise method for separating hereditary variants from tumour-specific mutations. Our findings imply that extracting hereditary information from tumour DNA analysis could serve as a practical, cost-effective approach to providing hereditary genetic testing in the clinic.

Preface

This dissertation is based on next-generation sequencing data from The OncoPanel Pilot (TOP) study. The TOP study was designed to optimize and validate a clinical targeted NGS panel for its utility as standard of care. Approval of this pilot study was covered by the Human Research Ethics Protocol H14-01212. Sample preparation and sequencing, as well as processing of raw data, read alignment, and variant calling were collaboratively performed by members of Canada's Michael Smith Genome Sciences Centre and the Centre for Clinical Genomics. Data analyses in Chapter 3 and 4 are my original work.

Analysis of pharmacogenomic genes in Chapter 3 and 4 was presented as a poster titled "Comparative Analysis of Pharmacogenomic Variants in Amplicon-sequenced DNA from Peripheral Blood and Formalin-fixed Paraffin-embedded Tumours" at the 2016 Intelligent Systems for Molecular Biology conference and the Translational Medicine (TransMed) Special Interest Group meeting.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACCE	Analytical validity, Clinical validity, Clinical utility, and Ethical, legal and social implications
ACMG	American College of Medical Genetics and Genomics
ALK	Anaplastic lymphoma kinase gene
APC	Adenomatous polyposis coli gene
ASCII	American Standard Code for Information Interchange
BAM	Binary Alignment/Map
BAQ	Base quality score
BCR-ABL1	Breakpoint cluster region and Abelson murine leukemia viral oncogene homolog 1 fusion gene
BRAF	B-Raf proto-oncogene
BRCA1	Breast cancer type 1 susceptibility gene
BRCA2	Breast cancer type 2 susceptibility gene
BWA	Burrows-Wheeler aligner
BWT	Burrows-Wheeler transform
CDC	Centers for Disease Control and Prevention
CI	Confidence interval
CNV	Copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer database
CPG	Cancer predisposing gene

CRC	Colorectal cancer
dbSNP	Single Nucleotide Polymorphism Database
dNTP	Deoxyribonucleotides
dTMP	Deoxythymidine monophosphate
DNA	Deoxyribonucleic acid
DPD	Dihydropyrimidine dehydrogenase enzyme
DPYD	Dihydropyrimidine dehydrogenase gene
EGFR	Epidermal growth factor receptor gene
ExAC	Exome Aggregation Consortium
FAP	Familial adenomatous polyposis
FET	Fisher's exact test
FFPE	Formalin-fixed Paraffin-embedded
GIST	Gastrointestinal stromal tumour
GSTP	Glutathione S-transferases
GSTP1	Glutathione S-transferases gene
HER2	Human epidermal growth factor receptor 2 gene
HGP	Human Genome Project
ICGC	International Cancer Genome Consortium
IGV	Intergrative Genomics Viewer
LOH	Loss of heterozygosity
MAPQ	Mapping quality score
MIT	Massachusetts Institute of Technology
MEN2	Multiple endocrine neoplasia type 2
MR	Methylenetetrahydrofolate reductase enzyme
MTHFR	Methylenetetrahydrofolate reductase gene

NHGRI	National Human Genome Research Institute
NGS	Next-generation sequencing
PGx	Pharmacogenomics
PCR	Polymerase chain reaction
PPV	Positive predictive value
RB1	Retinoblastoma 1 gene
RET	Ret proto-oncogene
SAM	Sequence Alignment/Map
SNV	Single nucleotide variant
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TOP	The OncoPanel Pilot
TP53	Tumour protein p53 gene
TP	Thymidine phosphorylase enzyme
TS	Thymidylate synthase enzyme
TYMP	Thymidine phosphorylase gene
TYMS	Thymidylate synthase gene
UDG	Rracil-DNA glycosylase
UGT1A	Uridine diphosphate glycosyltransferase 1 family enzymes
UGT1A1	Uridine diphosphate glycosyltransferase 1A1 gene
VAF	Variant allele frequency
VCF	Variant Call Format
WES	Whole exome sequencing
WGS	Whole genome sequencing

Acknowledgments

I would like to thank my supervisor, Dr. Aly Karsan for giving me the opportunity to undertake a bioinformatics project for my Master's thesis and funding me throughout the course of my research. My gratitude is extended to my committee members, Dr. Martin Hirst and Dr. Ryan Morin for providing guidance and ideas in data analyses for my project. I am also grateful for the productive discussions I had with my colleagues from the Centre for Clinical Genomics and Karsan lab.

There were many individuals who endured my frustration yet provided me with endless support and encouragement. For that, I would like to express my heartfelt gratitude to my family, my academic mentor, Dr. David Walker, and my friends. Thank you James Topham, Jenny Zhao, Jessica Pilsworth, Joey Ip, Kate Slowski, Ka Ming Nip, Laura Graziano, Patrick Coulombe, Samiah Alam, Samantha Jones, Santina Lin, Tehmina Masud, and Yuko Goto for keeping me motivated and sane throughout this journey. Last but not least, I would like to thank my partner, Florian Krauthan for comforting me in my moments of failures, celebrating my successes, and understanding why dirty dishes were left in the sink.

Chapter 1

Introduction

1.1 The emergence of precision oncology

Cancers are fundamentally a group of genetic disorders. The role of genetic alterations in driving malignant transformation has been implicated in studies dating back to the late nineteenth and early twentieth centuries by David von Hansemann and Theodor Boveri. Von Hansemann suggested that aberrant cell division accounted for unequal chromosome distributions in tumour cells [1, 213]. Motivated by von Hansemann's findings, Boveri explored the outcomes of sea urchin embryos that were induced to divide abnormally. An intriguing observation that drew Boveri's attention was not all chromosomal imbalanced cells proliferated uncontrollably and formed tumours, there were some that resulted in cell death, thus indicating that genetic materials were functionally distinct. This led to Boveri's hypothesis that tumour development is promoted by retention of chromatin parts that stimulate growth or elimination of those that inhibit growth, concepts that manifested in the present-day knowledge of oncogenes and tumour-suppressor genes, respectively [1, 32].

Major strides have been made in understanding the molecular basis of cancer, including the discovery of recurrent gene mutations and elucidation of oncogenic pathways. Some of these findings have been successfully translated into clinical applications wherein patients who harbour actionable somatic mutations benefitted from treatment with targeted anti-cancer drugs. Notable examples include treatment of *HER2*-overexpressed breast cancer with trastuzumab [10, 68, 186, 191, 193, 212], *BCR-ABL1*-translocated chronic myeloid leukemia with imatinib [69, 175], and *BRAF*-mutated melanoma with vemurafenib [42, 168]. The ability to improve clinical outcome by exploiting tumour genetic vulnerabilities has contributed to the advent of precision oncology, a framework that tailors patient care based on tumour genetic makeup.

As more actionable somatic mutations are revealed, precision oncology regimens begin to face limitations caused by single-gene assays, which pose challenges in scaling to meet diagnostic needs. Fortunately, these barriers have been surmounted by advances in next-generation sequencing

(NGS) technologies. By harnessing the high-throughput nature of NGS, traditional gene-by-gene approaches are being rapidly supplanted by targeted gene panels and genome-scale profiling, which surveys the whole exome or genome. The dramatic decline in sequencing cost [217] and low requirement for DNA input [48, 179, 194] have also accelerated the adoption of NGS-based genomic testing in clinical practice (Figure 1.1). Furthermore, concurrent progress in algorithmic development has enabled efficient storage, processing, and interpretation of massive genomic data sets produced by NGS platforms [156, 206]. Automated variant analysis pipelines can be established by integrating these bioinformatics tools to allow accurate reporting of clinically significant genomic alterations [30, 98, 113, 187]. Hence, while the precision oncology framework has been instigated by the discovery of actionable somatic mutations, its translation into clinical use is largely catalyzed by advancements in DNA sequencing technologies and analysis algorithms. Although research efforts are still underway in refining these technological components, it is undeniable that the precision oncology paradigm holds great potential in enhancing disease management and therapeutic intervention for cancer patients.

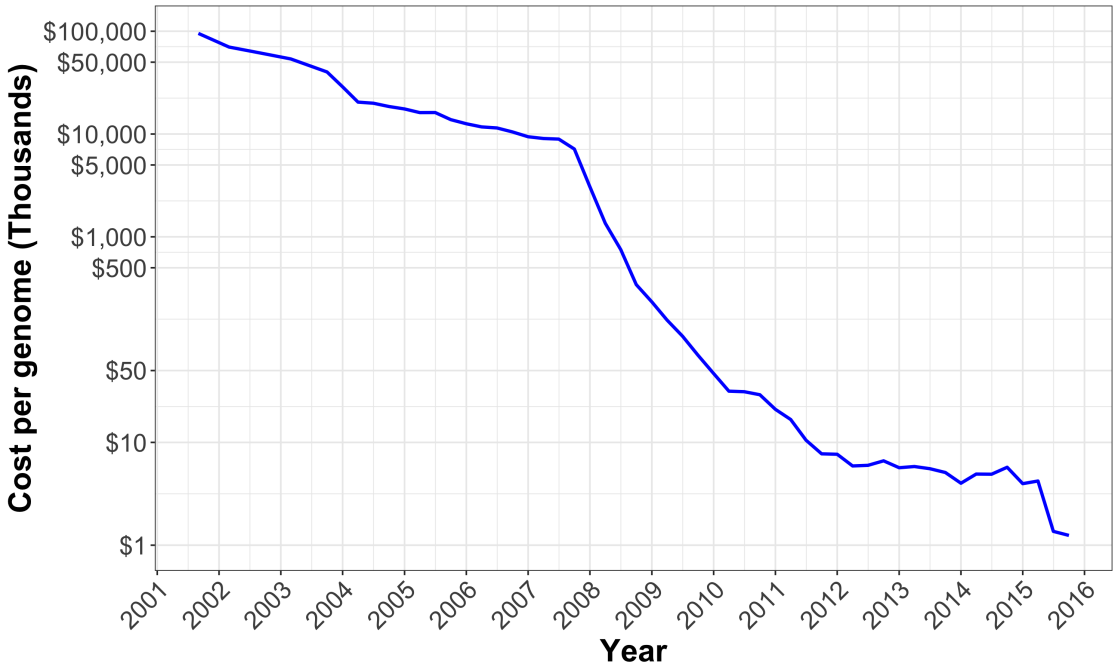


Figure 1.1: Sequencing cost per human-sized genome between 2001 and 2015. Data by courtesy of the National Human Genome Research Institute (NHGRI) [217].

1.2 Overview of next-generation sequencing technologies

The Human Genome Project (HGP) was completed in 2003, approximately 13 years after its launch date, producing the first human reference genome at an estimated expense of US$2.7 billion [2]. While the HGP provided a wealth of information, which led to major breakthroughs in the field of genomics, the completion time and cost of the project were apparent rate limiting steps. The need for more time- and cost-efficient DNA sequencing methods stimulated the development of NGS technologies. NGS is a general term that describes various high-throughout DNA sequencing technologies that can vary based on read length, chemistry, and detection methods (Table 1.1). These differences give rise to the strengths and weaknesses of each NGS platform. Recognition of these system specifications are essential for users to capitalize on the strengths and compensate for the limitations of the different NGS technologies.

In general, the sequencing process of most NGS platforms can be summarized into three steps. The first step involves nucleotide addition, which can be accomplished by DNA polymerase reaction or ligation (sequencing by synthesis *vs.* sequencing by ligation). This is followed by a detection step to identify the nucleotide species that was incorporated on single molecule or clonally amplified DNA templates. Nucleotide detection can be performed using optical or non-optical sensing. Illumina and Pacific Biosciences (PacBio) platforms use optical sensing to detect fluorescence for base calling [18, 36, 70, 72, 89], whereas the Ion Torrent platform uses non-optical sensing to detect change in pH to determine nucleotide identity [173]. Lastly, a wash step re-initiates the cycle for the next base on the DNA templates by removing anchor-probe complexes or fluorophores and blocking groups. A key feature of NGS is its ability to simultaneously carry out this stepwise process for many millions of DNA templates; hence, NGS is also known as massively parallel sequencing. There are also NGS technologies that deviate from this sequencing cycle, such as the Oxford Nanopore Technologies (ONT) platform, which directly measures DNA sequence using current shifts produced as DNA translocates through nanopore sensors [216].

Table 1.1: Different NGS platforms by read length, chemistry, and detection method. Adapted from a table created by [118] under the Creative Commons Attribution 4.0 International License.

Platform	Read length[†]	Amplification	Chemistry	Detection	Website
Complete Genomics	Short	Clonal	Sequencing by ligation	Optical	http://www.completegenomics.com/
Illumina	Short	Clonal	Sequencing by synthesis	Optical	http://www.illumina.com
Ion Torrent	Short	Clonal	Sequencing by synthesis	Solid state	http://www.thermofisher.com/ca/en/home/brands/ion-torrent.html
Oxford Nanopore	Long	Single molecule	Nanopore	Nanopore	https://nanoporetech.com/
Pacific Biosciences	Long	Single molecule	Sequencing by synthesis	Optical	http://www.pacb.com/
Roche 454	Short	Clonal	Sequencing by synthesis	Optical	http://www.454.com
SoLiD, ThermoFisher Applied Biosystems	Short	Clonal	Sequencing by ligation	Optical	http://www.thermofisher.com/ca/en/home/brands/applied-biosystems.html

[†]Short-read platforms range from 35 bp to 700 bp, whereas long-read platforms range from 8 Kbp to 200 Kbp [84].

1.2.1 Illumina sequencing

At present, the most widely used NGS technology is the Illumina short-read platform as evident by its prevalence in the literature and the Sequence Read Archive (SRA). In 2011, 84% of sequence reads in the SRA were generated by Illumina sequencing [109]. The Illumina platform uses a sequencing-by-synthesis approach with reversible dye terminators, which enable base calling through detection of fluorescent signals while blocking the ribose 3'-OH group to prevent addition of the next nucleotide by DNA polymerase [18, 89]. Briefly, an Illumina NGS workflow begins by ligating adapters to the ends of fragmented DNA, followed by hybridizing these templates to complementary adapter sequences on flow cell surfaces. Bridge amplification is then performed to generate clusters of clonally amplified DNA templates. DNA sequencing starts by annealing primers complementary to adapter sequences, which enable DNA polymerase to carry out the elongation process. All four reversible dye terminator-bound deoxyribonucleotides (dNTPs) are simultaneously added during each cycle and are distinguishable by unique fluorophore-labelling. The dNTPs are also terminally blocked, allowing the incorporation of only one dNTP molecule per cycle. Subsequent to dNTPs addition, unbound dNTPs are washed away. The flow cells are then imaged using laser channels and fluorescence corresponding to the incorporated dNTP is emitted at each cluster. Finally, a new cycle is initiated by cleaving the fluorophores and unblocking the 3'-OH groups (Figure 1.2) [84, 118, 135, 136].

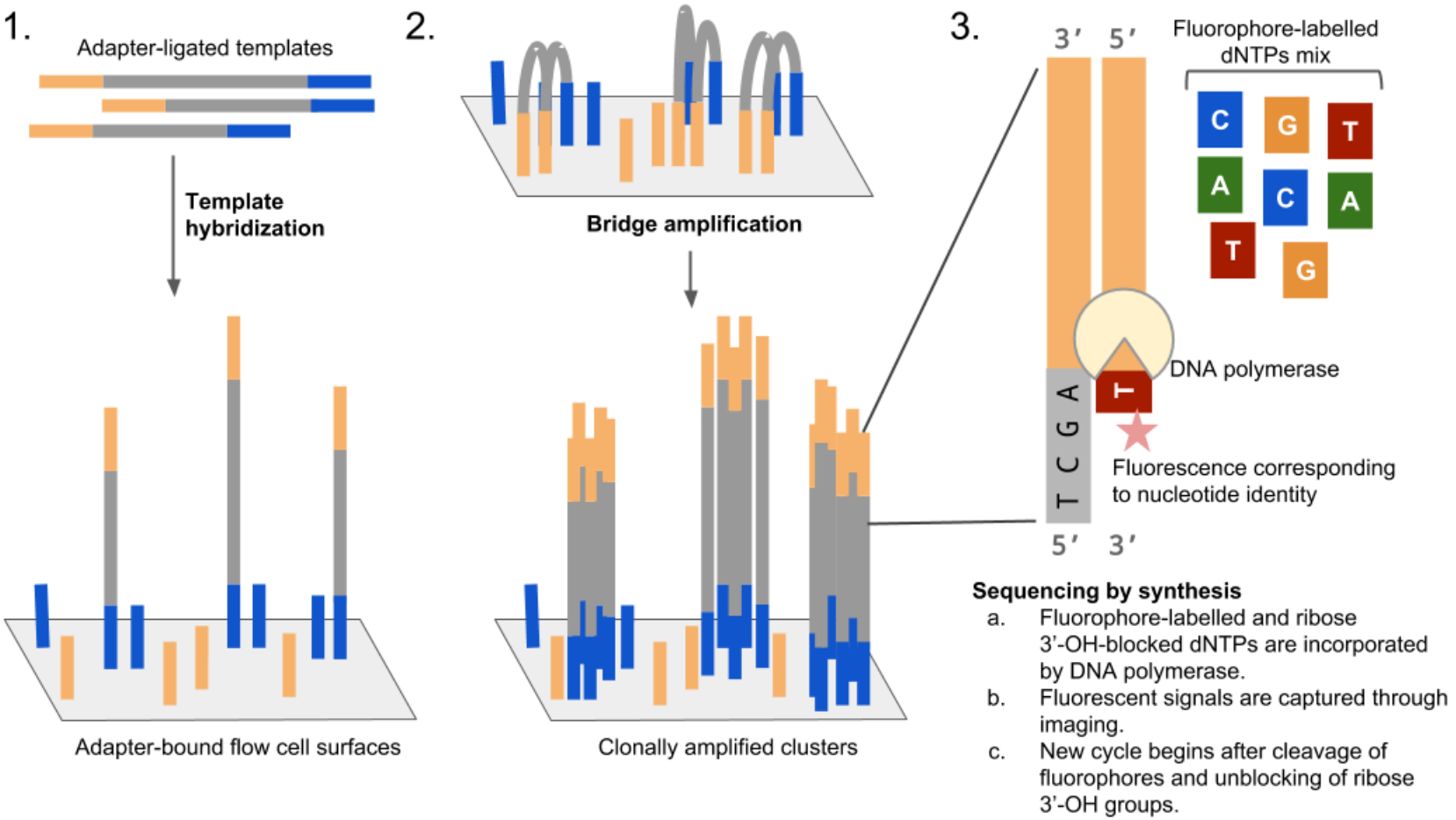


Figure 1.2: Workflow for Illumina sequencing. (1) Hybridization of adapter-ligated templates to flow cell surfaces. (2) Bridge amplification to generate clonally amplified clusters. (3) Sequencing by synthesis.

1.2.2 Clinical applications of NGS

The emergence of NGS has revolutionized biological inquiry, particularly in cancer genomics research. Collaborative efforts such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have leveraged NGS technologies to characterize genomic landscapes of different subtypes of tumours [6, 150, 166, 172, 203]. This has resulted in the identification of novel driver mutations, thereby enhancing knowledge of tumour biology and treatment strategies. The ability to sequence multiple genes and samples in parallel with less DNA and in a cost- and time-effective manner, also makes NGS an attractive clinical tool to complement precision medicine initiatives. These advantages demonstrate that the viability and efficiency of NGS are superior to traditional Sanger sequencing, which is typically limited to sequencing a specific gene region of a given sample per run. Currently, tumour sequencing assays ranging from targeted gene panels up to genome-scale profiling have been employed in clinical oncology to guide diagnosis, prognosis, and therapeutic decision-making (Figure 1.3).

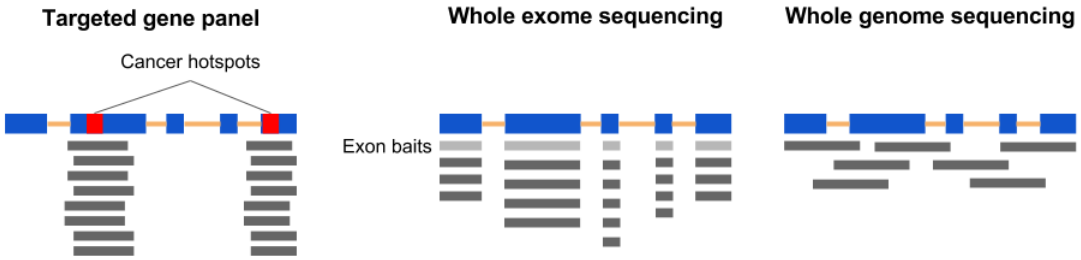


Figure 1.3: Comparison of testing content across targeted gene panels, whole exome sequencing, and whole genome sequencing.

Targeted gene panels

Several considerations must be made when developing clinical-grade tumour genomic tests, including turnaround time, testing cost per patient, and depth of sequencing coverage. For these reasons, many clinical laboratories have resorted to targeted gene panels, which focus on mutational hotspots, actionable genes, or genomic regions of known clinical relevance. Furthermore, despite the growing catalog of tumour genetic variants contributed by large consortium projects, the impact of the majority of variants on cancer development remains unknown [171, 197, 199]. Genomic information of unknown significance not only pose limitations in clinical translation, but also challenges in communicating such results to patients [53, 171, 183, 197]. Hence, targeted gene panels are more practical for prospective clinical use at the present time than whole exome and genome approaches.

Strategies to interrogate genomic regions of interest in targeted NGS assays include amplicon-based and capture-based methods (Figure 1.4). Amplicon-based method enriches target regions

using PCR amplification prior to NGS. PCR can be performed in uniplex, in which a single primer pair is used to generate amplicons within one reaction, or in multiplex, in which multiple primer pairs are used to generate amplicons in a single reaction. Conventional multiplex PCR faces challenges such as interactions between primers and competition for reagents, which could lead to amplification failures. These limitations of conventional multiplex PCR can be circumvented by the microdroplet-based PCR platform developed by RainDance Technologies. In microdroplet-based PCR enrichment, droplets containing a single primer pair are merged with droplets containing fragmented genomic DNA and PCR reagent mix, which includes dNTPs and DNA polymerase. PCR enrichment is then performed for a library of droplets, which simultaneously amplifies several target regions in a single reaction. Within each PCR droplet, amplicon generation is confined to a single primer pair and other reagents. This eliminates interaction between primer pairs and competition for reagents, thereby enhancing amplification uniformity [130].

The amplicon-based approach typically requires lower amount of DNA input and offers quick turnaround relative to hybridization capture methods [41, 79]. However, PCR amplification can distort detection of copy number variation (CNV), although bioinformatics tools, such as ONCOCNV [23], have been developed to perform copy number analysis using amplicon sequencing data. Moreover, amplicon sequencing is prone to enrichment bias, especially in samples with low amounts of DNA templates. For example, Wong et al. [221] reported higher prevalence of formalin-induced sequence artifacts in samples with lower amounts of amplifiable templates. Clinical specimens with low template copies tend to have reduced amplicon enrichment and higher probability of amplifying DNA templates with sequence artifacts [221]. Another limitation of amplicon sequencing is its inability to detect novel gene fusions because PCR primer pairs would fail to amplify the translocated DNA [79, 190]. PCR primers can also mask genetic variants that are located within the primer region, but this disadvantage could be mitigated by designing overlapping amplicons.

Hybridization capture methods involve the use of complementary oligonucleotide probes to bind targeted regions. There are two methods for target capture, namely array-based and in-solution capture (Figure 1.4B). In array-based capture, complementary oligonucleotide probes of targeted regions are fixed on microarrays. Fragmented genomic DNA library is hybridized to the probes on the microarray and subjected to NGS after unbound library DNA is washed away. In the in-solution capture approach, the target-specific probes are biotinylated. The pool of probes is mixed with fragmented genomic DNA library and hybridization occurs "in solution." Hybridized DNA is pulled down using streptavidin-labelled magnetic beads and then subjected to NGS [79, 190]. The in-solution capture method is typically preferred over array-based capture because it omits the necessity for expensive instrument to process microarrays, and it requires lower quantity of DNA as starting material [22, 52]. In contrast to the amplicon-based method, hybridization capture approaches can detect gene fusions, as well as yield more reliable inference of CNVs [79, 190].

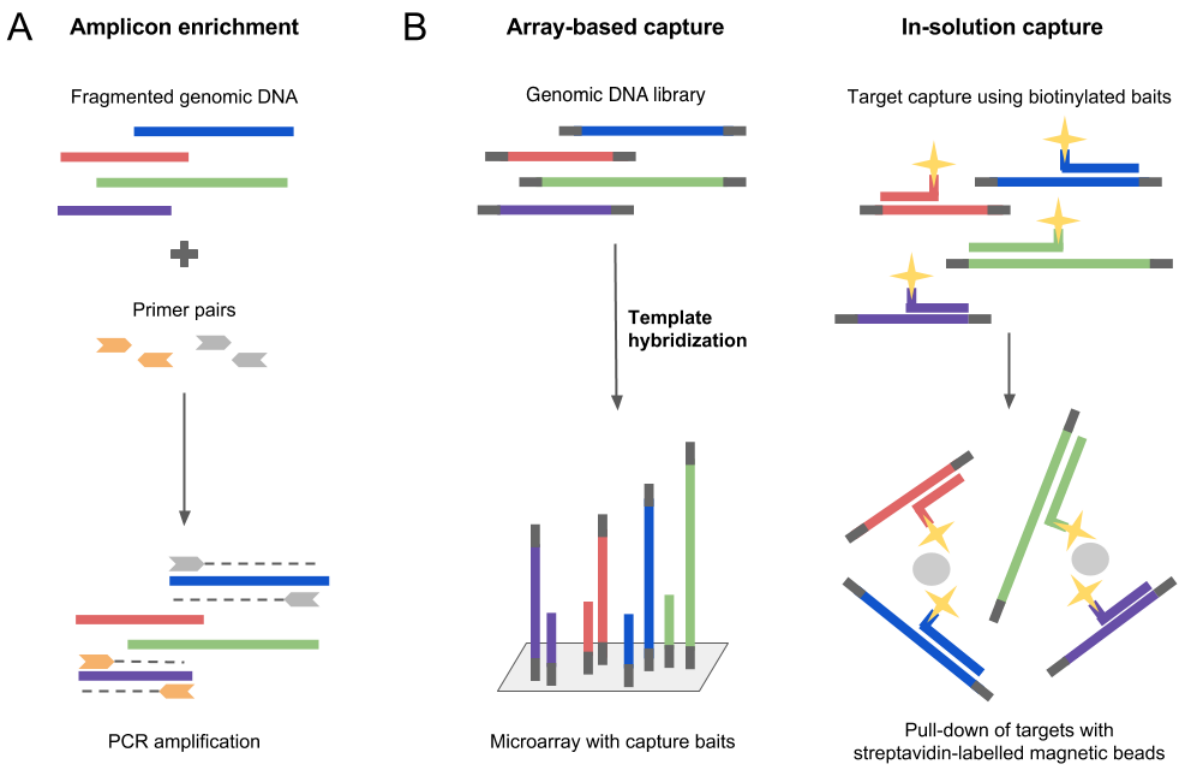


Figure 1.4: Different approaches in target enrichment. (A) Amplicon-based enrichment. (B) Capture-based enrichment, which can be categorized as array-based and in-solution capturing.

Whole exome sequencing

Whole exome sequencing (WES) interrogates all protein-coding regions, which constitute approximately 1% of the genome [164, 190]. Target capture methods are commonly used to enrich coding sequences before massively parallel sequencing. To date, it is estimated that 85% of pathogenic variants are present within exons [164]. Therefore, WES assays have the potential to facilitate prospective medical decision-making, as well as contribute to retrospective studies to uncover the functional and clinical impacts of newly discovered genetic alterations in tumours.

There are several disadvantages associated with WES assays. This includes the inability to detect mutations in non-coding regions and structural variants, which can promote cancer formation. WES assays also tend to achieve lower depth of coverage compared to targeted gene panels, increasing the rates of false positives. Because of the increased testing content, analytical validation of WES assays is more challenging and time-consuming. Nevertheless, whole exome testing has already been offered by a few academic centres such as Broad Institute of MIT and Harvard, Baylor College of Medicine, and Washington University in St. Louis, to assist in precision cancer medicine [164, 190].

Whole genome sequencing

Whole genome sequencing (WGS) scans the entire genome, including coding and non-coding genomic regions. Similar to WES, WGS faces drawbacks in terms of depth of coverage and difficulty in ensuring analytic validity [190]. Although the Illumina HiSeq X Ten System has made it possible to sequence an entire genome at 30x coverage under US$1000 [66], application of WGS in routine clinical testing is still challenging due to limitations in variant interpretation. As a result of limited clinically annotated genetic variants, WGS is expected to yield a high burden of variants of unknown significance, which are problematic in clinical practice [170, 190]. Despite these constraints, Laskin et al. [113] reported that the Personalized OncoGenomics (POG) study, which integrates whole genome analysis in making therapeutic decision, can benefit patients with advanced cancers by matching them with targeted agents that are approved or currently in clinical trials. Thus, while there are challenges that need to be overcome to implement WGS as standard of care, WGS has proven its utility in conducting additional search for actionable mutations that are undetected by less comprehensive sequencing strategies. In particular, follow-up testing with WGS can broaden the treatment options for patients with incurable cancers.

1.3 Variant analysis pipeline

The increase in affordability of NGS technologies has led to a marked surge in data production. For instance, the Illumina HiSeq 2500 platform is capable of sequencing 150–180 whole exomes from human samples at 50x coverage, generating approximately 1TB of raw data in a single run [66]. Processing these enormous data sets and extracting useful results for research and clinical purposes

rely heavily on bioinformatics algorithms and tools. A general workflow for variant analysis in medical genomics consists of four stages: (1) quality control and pre-processing of raw sequencing reads, (2) read alignment to the reference genome and post-alignment processing, (3) variant calling, and (4) variant annotation and interpretation.

1.3.1 Quality control and pre-processing of raw sequencing reads

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

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

Hence, $1/1000$ probability of base error would correspond to a Phred-scaled BAQ score of 30 [114, 149].

The raw output of NGS instruments, which contains information from base calling, are stored in FASTQ and FASTA text-based formats. FASTQ files contain DNA sequences with sequence names and Phred-scaled BAQ scores, whereas FASTA files only contain DNA sequences with sequence names (Figure 1.5) [15, 143]. Quality of raw NGS data can be evaluated using bioinformatics tools such as FastQC, which generates a diagnostic report consisting of various quality control parameters. These include sequence length distribution, GC content distribution, degree of sequence duplication, presence of overrepresented and adapter sequences, and average Phred-scaled BAQ scores at each base across reads [9].

Quality control results are used to assist in pre-processing of raw NGS data before further analysis takes place. For example, NGS libraries with poor quality bases near the 3' ends of reads may require read trimming before alignment. Removal of adapters is also typically performed in the pre-processing step [15]. Computational tools that are commonly used to accomplish these tasks include Trimmomatic [26] and Cutadapt [132]. Moreover, assessment of quality control metrics also enables the recognition of poor quality NGS libraries and those that are potentially contaminated. Flagging of these libraries would allow downstream analyses and result interpretation to be performed with caution.

Figure 1.5: File formats of raw output from NGS instruments. (A) FASTQ format. Sequence name starting with @ is presented in line 1, followed by read sequence in line 2. Line 3 begins with + and may be followed by the sequence name again, and line 4 consists of quality scores corresponding to bases in line 2. Quality scores are represented as ASCII characters. (B) FASTA format. Line 1 contains sequence name, which starts with >, whereas line 2 contains the read sequence.

1.3.2 Read alignment and post-alignment processing

Next, pre-processed reads are aligned to the reference genome. Alignment algorithms essentially match read sequences to sequences in the reference genome, while accounting for sequencing errors and true genomic alterations [15, 123, 143, 149]. Because of the large data output produced by NGS, alignment algorithms must also be time- and memory-efficient.

Widely-used alignment softwares implement either the Burrows-Wheeler transform (BWT) compression algorithm or hash table indexing [15, 123, 149, 156]. Popular BWT-based aligners include BWA [120–122] and Bowtie2 [112], which are well-known for reduced runtimes and memory requirements. Conversely, algorithms based on hash tables such as Novoalign (http://novocraft.com), SHRiMP2 [55], and Stampy [129] have longer computational times but tend to yield more accurate alignments [15, 123, 149]. The standard formats for storing aligned read data are the Sequence Alignment/Map (SAM) text-based format and its compressed binary version, the Binary Alignment/Map (BAM) format [124].

Several post-alignment steps are performed such as removal of read duplicates, which could be introduced by PCR bias. Local realignment is also often performed at genomic regions surrounding insertions-deletions (indels) to reduce errors caused by inaccurate alignments. As raw BAQ scores might be erroneously assigned, recalibration of BAQ scores is sometimes recommended prior to variant calling. Quality scores are adjusted by accounting for differences in quality between machine cycles and neighbouring dinucleotides [15, 143, 149, 155].

1.3.3 Variant calling

Variant calling identifies genomic differences between aligned read sequences and the reference genome. When matched normal samples are sequenced, some variant callers are able to detect

germline, somatic, and loss of heterozygosity (LOH) events through paired analysis of tumour and matched normal samples [108]. Variant callers can be categorized as heuristic or probabilistic [15, 143, 156]. The standard output of variant callers for storing sequence variant data is the text-based Variant Call Format (VCF) (Figure 1.6) [54].

VarScan2 is a variant caller that uses heuristic factors such as cut-offs for coverage depth, BAQ score, and variant allele frequency (VAF), as well as the one-sided Fisher's exact test (FET) on mapped read counts to identify variants [15, 107, 108, 143, 156]. In single sample analysis (sample *vs.* reference genome), a variant is detected through applying the one-sided FET to compare reference- and variant-supporting read counts to an expected read distribution that reflects sequencing error rate at a reference site [107, 108]. For example, the expected read distribution for a reference site with 2000x coverage depth and a 0.1% sequencing error rate would comprise 2 variant-supporting reads, which are reads with errors, and 1998 reference-supporting reads.

Variant callers that are based on probabilistic methods employ Bayes' theorem to measure posterior probabilities of all possible genotypes at a loci. These posterior probabilities are then used to infer genotype. Probabilistic methods also take into account prior information such as population allele frequencies, which can be obtained from the Single Nucleotide Polymorphism Database (dbSNP) or through multi-sample variant calling, and patterns of linkage disequilibrium, as well as genotype likelihood, which can be computed using BAQ scores [143, 149]. Examples of variant callers that implement probabilistic approaches include GATK [138], MuTect [49], and Strelka [182].

```
##fileformat=VCFv4.1
##source=VarScan2
##INFO=<ID=ADP,Number=1,Type=Integer,Description="Average per-sample depth of bases with Phred score >= 20">
##INFO=<ID=WT,Number=1,Type=Integer,Description="Number of samples called reference (wild-type)">
##INFO=<ID=HET,Number=1,Type=Integer,Description="Number of samples called heterozygous-variant">
##INFO=<ID=HOM,Number=1,Type=Integer,Description="Number of samples called homozygous-variant">
##INFO=<ID=NC,Number=1,Type=Integer,Description="Number of samples not called">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=SDP,Number=1,Type=Integer,Description="Raw Read Depth as reported by SAMtools">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Quality Read Depth of bases with Phred score >= 20">
##FORMAT=<ID=RD,Number=1,Type=Integer,Description="Depth of reference-supporting bases (reads1)">
##FORMAT=<ID=AD,Number=1,Type=Integer,Description="Depth of variant-supporting bases (reads2)">
##FORMAT=<ID=FREQ,Number=1,Type=String,Description="Variant allele frequency">
##FORMAT=<ID=PVAL,Number=1,Type=String,Description="P-value from Fisher's Exact Test">
##FORMAT=<ID=RBQ,Number=1,Type=Integer,Description="Average quality of reference-supporting bases (qual1)">
##FORMAT=<ID=ABQ,Number=1,Type=Integer,Description="Average quality of variant-supporting bases (qual2)">
##FORMAT=<ID=RDF,Number=1,Type=Integer,Description="Depth of reference-supporting bases on forward strand (reads1plus)
">
##FORMAT=<ID=RDR,Number=1,Type=Integer,Description="Depth of reference-supporting bases on reverse strand (reads1minus
)">
##FORMAT=<ID=ADF,Number=1,Type=Integer,Description="Depth of variant-supporting bases on forward strand (reads2plus)">
##FORMAT=<ID=ADR,Number=1,Type=Integer,Description="Depth of variant-supporting bases on reverse strand (reads2minus)"
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Sample1
chr1    11205058        .   C       T       .       PASS    ADP=8188;WT=0;HET=0;HOM=1;NC=0      GT:GQ:SDP:DP:RD:AD:FRE
Q:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR    1/1:99:8345:8188:67:8121:99.18%:0E0:38:38:33:34:4082:4039
```

Figure 1.6: VCF format for storing sequence variation data. The VCF header, which begins with ##, contains information describing the data in the file. The line starting with # displays the column names and indicates the beginning of the body of the VCF file. Data columns must include eight mandatory fields: chromosome (CHROM), 1-based genomic position of the variant (POS), variant identifier (ID), reference base (REF), alternate base (ALT), quality score (QUAL), indicator whether the variant passed the specified filtering criteria (FILTER), variant annotations separated by semicolon (INFO). The FORMAT field contains colon-separated descriptors of the values in the genotype column, which is named after the sample(s) reported in the VCF file [54].

1.3.4 Variant annotation and interpretation

Subsequent to variant calling, variant annotation is performed by adding structural and functional information. Structural annotation provides information on the genomic location of the variant (e.g. intronic, intergenic, 5'UTR, 3'UTR, *etc.*), the impact of the variant on the gene transcript (e.g. missense, non-synonymous, synonymous, frameshift, *etc.*), and changes in codon and amino acid [15, 145, 155]. These information are typically reported in the Human Genome Variation Society (HGVS) nomenclature to ensure a consistent format in describing sequence variants [59].

Functional annotation, on the other hand, provides information on the effect of genomic variants on protein function. Pathogenicity of a variant cannot be exclusively presumed based on the predicted variant effect. For example, not all truncating, missense, and frameshift variants lead to deleterious effects. Conversely, not all synonymous variants are benign. Thus, functional prediction algorithms compute functional inference by incorporating additional data such as sequence homology, 3D protein structure, genomic context, protein interaction network, and evolutionary conservation [155, 163]. Examples of functional prediction tools include PolyPhen-2 [3], SIFT

[147], MutPred [119], Condel [83], and PhD-SNP [35].

Population allele frequencies are also taken into account when interpreting variants [15, 145, 155]. For instance, a common variant (minor allele frequency $>1\%$) is unlikely to be involved in cancer predisposition [14, 16]. Population allele frequencies can be obtained by annotating variants with the dbSNP and Exome Aggregation Consortium (ExAC) database. Furthermore, variant databases such as ClinVar and Catalogue of Somatic Mutations in Cancer (COSMIC) are also useful resources that provide information on clinical significance of variants and presence of variants in cancers [145, 155].

The original output of variants is then filtered and prioritized based on the different types of annotations, as well as clinical and biological relevance, generating a list of candidate variants [145, 155]. These variants are typically visualized using a genomic browser such as the Intergrative Genomics Viewer (IGV) for quality control purposes [155, 156]. Finally, a report is produced and reviewed by a board certified medical pathologist or clinical molecular geneticist for approval [145, 197].

1.4 ACCE model process for evaluating genetic tests

Development of a genetic test, including NGS-based genomic testing, for clinical use must be accompanied by an evaluation process to establish robustness and clinical benefits of the test. One approach to assess genetic tests is the ACCE model process, which consists of four criteria that make up its acronym: **A**nalytical validity, **C**linical validity, **C**linical utility, and **E**thical, legal and social implications [181, 232].

Analytic validation ensures that a clinical assay detects the genetic changes it is designed to identify with sufficient sensitivity and specificity [181, 232]. For example, analytic validity of a targeted NGS panel can be determined by running the panel on samples with known mutations that were previously identified using Sanger sequencing. Sensitivity and specificity of the targeted NGS panel can be measured using the results from Sanger sequencing as reference standards. Analytic validity also refers to quality assurance of a clinical assay. For instance, a targeted NGS panel must be capable of producing similar sequencing metrics (e.g. read depth and coverage uniformity) for samples with comparable DNA quantity and integrity.

Clinical validation determines whether results of the genetic test correspond to the clinical condition it is meant to detect [181, 232]. Example of a genetic test with high clinical validity is *RET* mutational testing, which can identify individuals with multiple endocrine neoplasia type 2 (MEN2) at a sensitivity of 95–98%. MEN2 is a heritable disorder transmitted in an autosomal dominant pattern, resulting in increased susceptibility to tumours in endocrine tissues, especially medullary thyroid carcinoma [33]. On the other hand, clinical utility of a genetic test is defined by its ability to enhance clinical outcome, such as survival and progression-free survival, after weighing in the risks, benefits, and economic impact of the test [181, 232]. The clinical utility of *RET* mutational

testing is demonstrated by its efficacy in identifying MEN2 susceptible children who would benefit from prophylactic surgery to remove all or parts of the thyroid gland. This results in reduced risk of developing thyroid cancers, thereby improving survival of these individuals [33].

Lastly, the ACCE model includes evaluation of ethical, legal and social implications of a genetic test. This component considers how the genetic test can lead to ramifications such as violation of privacy and confidentiality, stigmatization and discrimination based on genetic makeup (e.g. accessibility to insurance), and complications pertaining to consent for disclosure and ownership of the data. As well, this component of the framework ensures that safeguards, such as relevant policies and genetic counseling protocols, are implemented to prevent societal repercussions [181, 232].

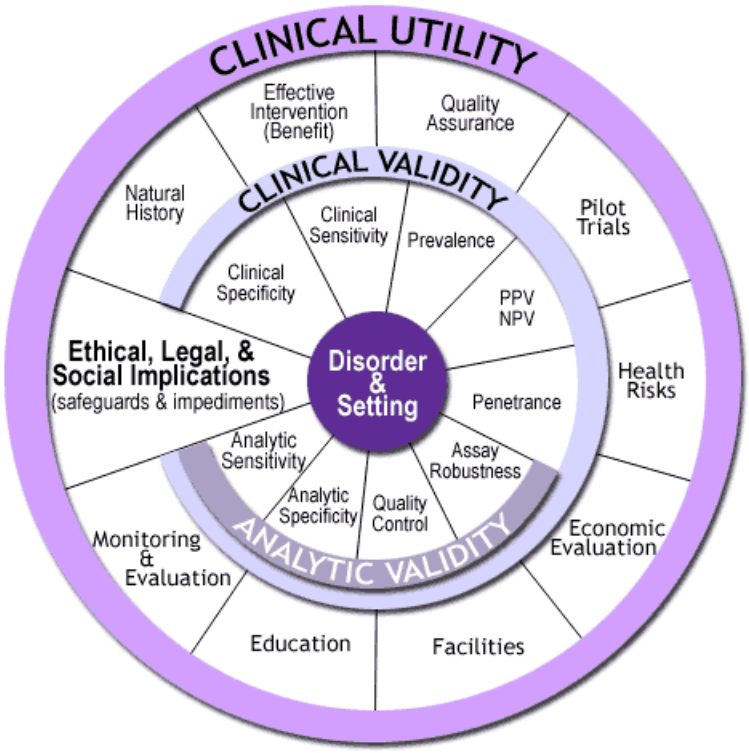


Figure 1.7: Four main criteria of the ACCE model process for evaluating a genetic test: **A**nalytical validity, **C**linical validity, **C**linical utility, and **E**thical, legal and social implications. Image by courtesy of Centers for Disease Control and Prevention (CDC).

1.5 Clinical implications of germline alterations in cancer

Screening for somatic and germline alterations are essential in delivering precision medicine to cancer patients. Somatic mutations can influence disease management and treatment of cancer patients with targeted agents, whereas clinical implications of germline alterations extend beyond the patients, affecting their families as well. Germline variants in cancer predisposing genes (CPGs) can

predict the risk of disease onset, allowing for preventive measures to be administered [165]. Furthermore, germline variants in pharmacogenomic (PGx) genes can predict response to chemotherapeutic drugs, including drug sensitivity and adverse drug reactions [144, 158]. Therefore, germline testing should be offered to ensure more precise cancer care if resources are available to analyze and interpret germline findings, and appropriate protocols are established to communicate results with patients and affected family members.

1.5.1 Cancer predisposition

Germline variants in CPGs can indicate increased cancer risks. Between 1982 and 2014, 114 CPGs have been discovered using approaches such as candidate gene, genome-wide mutation, and linkage analyses. The majority of CPGs act as tumour suppressors; hence, loss-of-function mutations in these genes, which inactivate gene function, predispose carriers to cancer. Genes in this category, including *TP53*, *BRCA1*, *BRCA2*, *APC*, and *RB1*, are usually involved in DNA repair and cell-cycle regulation. Conversely, there are fewer CPGs that promote cancer formation through gain-of-function mutations. These CPGs, typically protein kinases such as *ALK*, *EGFR*, and *RET*, predispose carriers to cancer through activation of gene function [165].

Clinical testing of CPGs can improve various aspects of patient care such as disease management and treatment. For instance, patients with BRCA-deficient breast and ovarian tumours can be treated with PARP inhibitors, which target tumour cells and impair growth through synthetic lethality [57, 91]. Notably, a key benefit of testing for germline variants in CPGs is the window of opportunity to implement cancer preventive measures for patients and affected relatives. Cancer prevention can involve approaches such as early and regular cancer screening, as well as prophylactic surgery and chemotherapy. For example, patients with familial adenomatous polyposis (FAP), which is caused by germline alterations in the *APC* gene and associated with high risk of developing colorectal cancer (CRC), are recommended to begin early colonoscopy-based screening [21].

1.5.2 Pharmacogenomics

Despite the expanding spectrum of targeted anti-cancer drugs, cytotoxic chemotherapy remains the primary treatment for several types of cancers. However, germline variants in PGx genes that affect the function and/or expression of drug targets and drug disposition proteins (proteins involved in drug metabolism and transport) can give rise to chemotherapy-related toxicity [144, 158]. Examples of chemotherapy-related toxicity include hand-foot syndrome, hearing loss, cardiomyopathy, and high-grade neutropenia, diarrhea, nausea and vomiting [58, 95, 110, 115, 178]. Chemotherapy-related toxicity can be debilitating and fatal, as well as culminate significant expenditures towards cancer supportive care [17, 157, 167]. To alleviate the occurrence of chemotherapy-related toxicity, germline PGx testing should be implemented in clinical practice to guide the selection of chemotherapeutic drugs and optimization of drug dosage for cancer patients.

5-fluorouracil

5-fluorouracil (5-FU) is a fluoropyrimidine drug that is commonly administered in chemotherapy regimens for patients with gastrointestinal cancers, including CRC. Inter-patient variability in response to 5-FU treatment can be caused by germline variants in the *TYMS* gene, which encodes for the drug target, the thymidylate synthase enzyme (TS). One of the 5-FU mechanisms of action involves the conversion of 5-FU to fluorodeoxyuridine monophosphate (5-FdUMP). 5-FdUMP then sequesters TS by forming a ternary complex with TS and the 5,10-methylenetetrahydrofolate (CH_2THF) cofactor, thereby impeding DNA synthesis (Figure 1.8). Germline alterations that result in a higher expression of TS such as the triple repeats of a 28 bp sequence upstream of the *TYMS* translational start site (rs45445694) are indicators of reduced likelihood of experiencing 5-FU toxicity. Unfortunately, this also means that treatment with 5-FU might not be effective due to high TS levels in the tumours [144, 158].

Germline variants in *DPYD* and *TYMP* genes, which encode for the 5-FU metabolizing proteins, dihydropyrimidine dehydrogenase (DPD) and thymidine phosphorylase (TP), respectively, can also serve as predictors for 5-FU-induced toxicity. DPD catabolizes 5-FU into dihydrofluorouracil (5-FUH_2), which mainly occurs in the liver, and the inactive products are subsequently excreted in the urine (Figure 1.8). Hence, germline variants resulting in DPD deficiency or total loss contribute to a longer half-life of 5-FU, which can cause severe or fatal toxicity in cancer patients. Several studies implied that TP may play a causative role in tumour growth and metastasis. In fact, higher TP expression was observed in tumours than normal tissues in CRC patients. Consequently, the 5-FU prodrug, capecitabine, is administered to target TP-overexpressed tumours because TP can metabolize capecitabine to the thymidylate synthase inhibitor, 5-FdUMP (Figure 1.8). Hence, this affects tumour growth with minimal toxic effects in normal cells. However, the presence of germline variants that increase expression of TP in normal cells could potentially lead to adverse drug reactions in patients receiving 5-FU-based chemotherapy [144, 158].

Efficacy of 5-FU depends on the intracellular reduced folate, CH_2THF, which together with the 5-FU active metabolite, 5-FdUMP, inhibit TS. This blocks the synthesis of deoxythymidine monophosphate (dTMP), causing imbalanced nucleotide levels in the cell and DNA damage. One of the enzymes that regulates intracellular CH_2THF levels is methylenetetrahydrofolate reductase (MR), which irreversibly converts CH_2THF to CH_3THF (Figure 1.8). Germline variants in the *MTHFR* gene that reduce enzymatic activity such as c.677C>T and c.1298A>C polymorphisms can increase chemosensitivty of tumours to 5-FU through cellular accumulation of CH_2THF. Nevertheless, several studies suggested that the combined presence of *MTHFR* c.1298A>C and *TYMS* 3'UTR indels could serve as predictors for 5-FU toxicity in CRC patients [158].

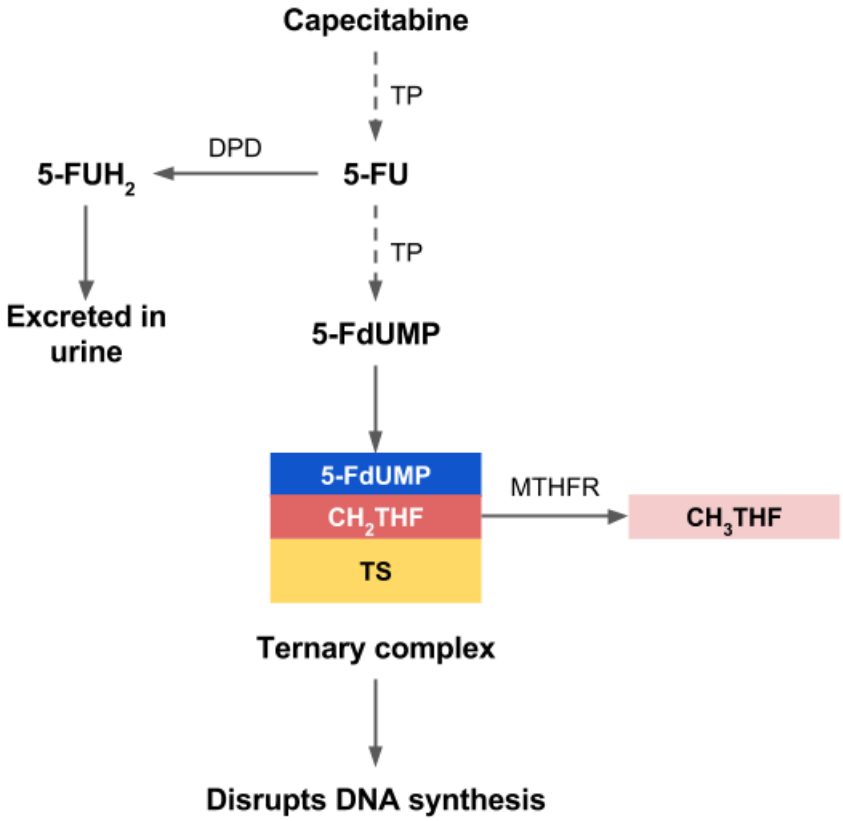


Figure 1.8: Involvement of TS, DPD, TP, and MTHFR in 5-FU mechanism of action. Dashed lines indicate more than one process is involved in producing the output, whereas solid lines indicate direct reactions. 5-FU, fluorouracil; 5-FdUMP, fluorodeoxyuridine monophosphate; 5-FUH_2, dihydrofluorouracil; CH_2THF, 5,10-methylenetetrahydrofolate; CH_3THF, 5-methyltetrahydrofolate.

Oxaliplatin

Oxaliplatin is a platinum derivative commonly used in combination with 5-FU for treating gastric and colorectal cancers. Deactivation of oxaliplatin can be induced by conjugation of the platinum derivative with glutathione (GSH), which is catalyzed by glutathione S-transferases (GSTP). While there are studies suggesting that germline variants in the *GSTP1* gene are associated with increased neurotoxicity in patients treated with oxaliplatin combination therapy, there are also groups reporting conflicting results. Therefore, additional studies are required to confirm the impact of *GSTP1* polymorphisms on oxaliplatin treatment [144, 158].

Irinotecan

Irinotecan is a camptothecin analog widely used in chemotherapy regimens for treating lung cancers and CRC. The active metabolite of irinotecan, SN-38, blocks type I DNA topoisomerase, impairing DNA replication. SN-38 is inactivated in the liver by uridine diphosphate glycosyltransferase 1 family enzymes (UGT1A) through glucuronidation and then excreted. Thus, patients with deficiency in UGT1A, which can be caused by germline variants in the *UGT1A1* gene, are at higher risk of experiencing toxicity due to a longer half-life of SN-38. An example of a germline variant in *UGT1A1* that results in reduced activity is the UGT1A1*28 allele, which corresponds to an extra TA repeat within position -53 and -42 of the translational start codon. Dose reduction is recommended for carriers to prevent toxic effects induced by irinotecan [144, 158].

1.6 Technical challenges in implementing germline testing in clinical oncology

1.6.1 Tumour-only sequencing

One of the challenges in integrating germline testing in clinical oncology is tumours are often sequenced without matched normal samples. Although sequencing of matched normal samples would allow accurate identification of somatic mutations and simultaneous detection of clinically important germline variants, it is common for clinical laboratories to only sequence tumour samples to minimize cost and turnaround time. However, genomic analyses of tumours can reveal clinically relevant germline variants [27, 102, 140, 141, 184]. For examples, Schrader et al. [184] reported that pathogenic germline variants in CPGs were retained in the tumour genomes of 91.9% of patients in the study cohort. Hence, clinical laboratories could leverage tumour genomic testing for identification of germline variants and subsequently refer potential germline variants to downstream confirmatory testing.

A clinical pipeline that leverages tumour genomic testing to perform initial screening for germline alterations could provide germline testing in a cost-effective manner because only selected patients would require follow-up testing. However, as the tumour genome contains both germline and somatic variants, the difficulty remains in devising an approach to accurately separate germline variants from somatic mutations. Jones et al. [102] used public databases such as dbSNP and COSMIC, as well as effect prediction tool to distinguish between germline and somatic variants, but reported high false positive rates. The use of these public databases cannot reliably differentiate between variant statuses because it is possible for a germline variant to occur somatically, and *vice versa*. For instance, an evaluation of 468 genes with known somatic driver mutations recorded in the COSMIC database showed that 49 of these genes were also known to harbour germline alterations that are associated with inherited predisposition to cancer [165].

One possible approach is the use of VAF to discriminate between germline and somatic variants. Because tumour biopsies are typically admixtures of tumour and normal cells, there is a high likelihood that somatic mutations might deviate from diploid zygosity (i.e. heterozygous variants are expected to have VAF close to 50%, whereas homozygous variants are expected to have VAF close to 100%; Figure 1.9A). Moreover, tumour heterogeneity might also give rise to VAF deviations (Figure 1.9B). Therefore, the use of VAF threshold could be a potential solution in distinguishing between germline and somatic alterations in genomic analyses of tumours without matched normal DNA.

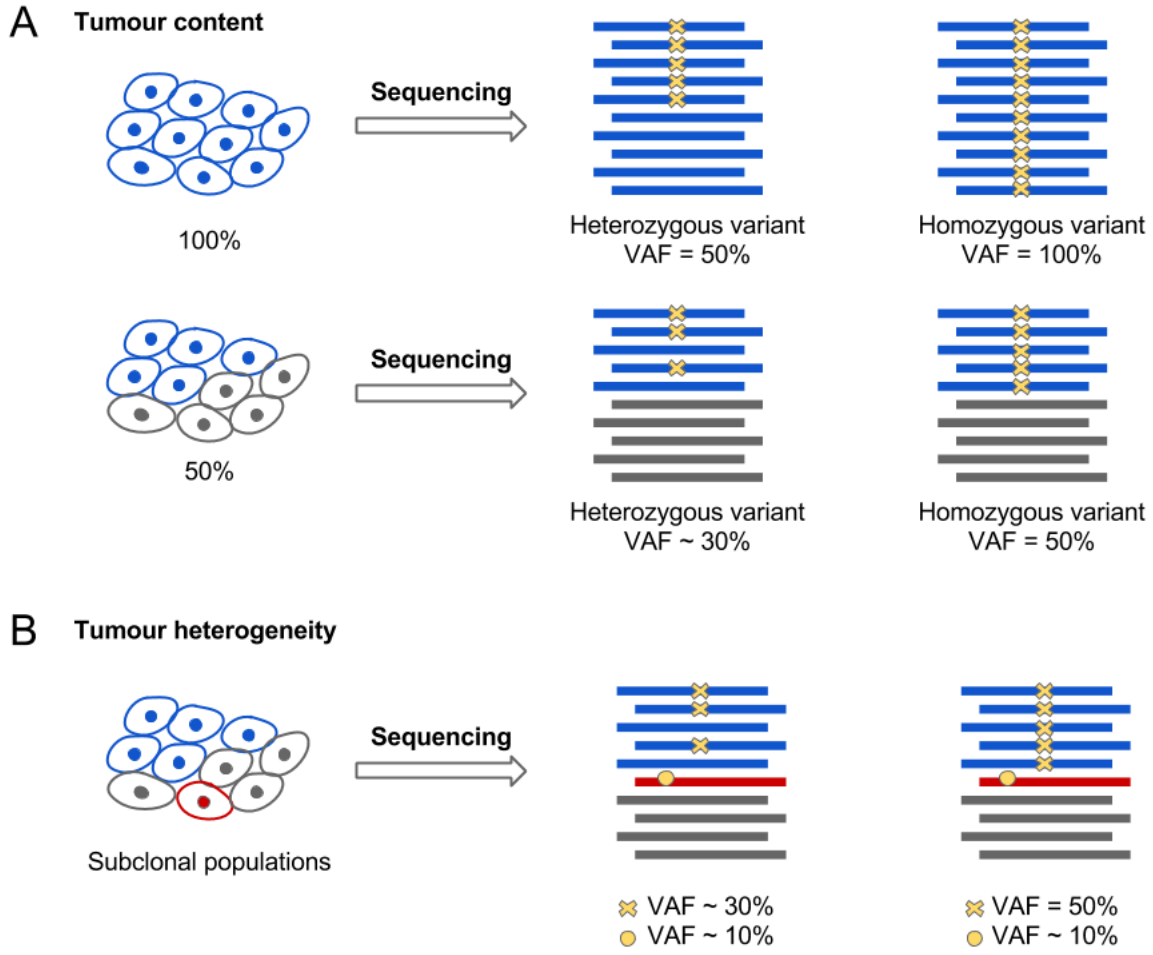


Figure 1.9: Deviations of VAF as a result of tumour content and heterogeneity. Blue and red cells represent tumour cells, whereas grey cells represent normal cells.

1.6.2 Formalin-fixed paraffin-embedded tumours

Another disadvantage of performing germline variant analysis using tumour DNA is tumour samples in the clinic are often formalin-fixed paraffin-embedded (FFPE). Formalin fixation preserves tissue morphology for histological assessment, whereas paraffin embedding enables stable storage of specimens at room temperature, which is cost- and space-saving compared to maintaining fresh frozen specimens in freezers [63, 66]. DNA isolated from FFPE tumours pose technical challenges in molecular testing because formalin fixation induces several types of DNA damage [63]. Therefore, assessment of these different forms of DNA damage is essential to establish quality control for a clinical genomic assay.

The main component of formalin, formaldehyde, can react with DNA bases and proteins, producing DNA-DNA, DNA-protein, and protein-protein crosslinks. Additionally, formaldehyde-DNA adducts can also be generated in formalin-fixed tissues. Crosslinking induced by formaldehyde destabilizes the DNA structure, resulting in degradation and low DNA yields extracted from FFPE tissues [63]. Another predominant form of formalin-induced DNA damage is DNA fragmentation. Hence, FFPE tissues not only produce low quantities of DNA, but also DNA with short fragment sizes. Particularly, this interferes with amplicon-based methods by reducing the amount of amplifiable DNA templates [61, 188, 220]. Severity of DNA fragmentation also increases with age of paraffin blocks and acidity of formalin solution used in tissue fixation [38, 128].

FFPE DNA also constitutes increased frequency of sequence artifacts. This is problematic in clinical practice because there is a high risk of misinterpreting artifactual base changes as true mutations that may influence patient care. Oxidization of formaldehyde, which generates formic acid, creates an acidic environment that catalyzes hydrolytic cleavage of *N*-glycosidic bonds between purines and the sugar backbone [63]. This produces abasic sites at which sequence artifacts can occur as most DNA polymerases tend to selectively incorporate adenines across abasic sites during the extension stage. In fewer cases, guanines and short deletions ranging from 1–3 bases could also be introduced by DNA polymerase when synthesizing through abasic sites [92].

A well-documented source of sequence artifacts in FFPE DNA is cytosine deamination. This generates uracil lesions, which leads to artifactual C>T/G>A transitions because adenines are added opposite of uracils during synthesis of complementary DNA strands [63]. Wong et al. [221] showed increased levels of C>T/G>A artifacts in amplicon sequencing data generated from highly fragmented DNA samples. This observation was attributed to a higher probability of amplifying DNA templates containing sequence artifacts in samples with reduced amount of amplifiable DNA templates as a result of fragmentation damage [221]. While cytosines can be restored by treating FFPE DNA with uracil-DNA glycosylase (UDG) to eliminate uracil lesions, there is currently no method to repair deamination of 5-methylcytosine (5-mC). 5-mC are common at CpG dinucleotides and are more susceptible to deamination in formalin-fixed tissues. Deamination of 5-mC gives rise to thymine instead of uracil, thus cannot be reinstated through treatment with the UDG enzyme [63].

1.7 Objectives

Germline alterations have clinical implications for cancer patients and their family members. Because the tumour genome contains both somatic and germline variants, clinical tumour sequencing presents an opportunity for pre-screening of germline variants. This framework is a time- and cost-effective approach for providing germline testing because only patients with potential germline variants would require downstream confirmatory testing. A primary challenge in implementing this framework for identifying clinically significant germline variants is differentiating between germline and somatic alterations in the tumour genome. Tumour specimens also tend to be FFPE, which causes DNA damage that could affect the use of FFPE DNA for clinical genomic testing.

To date, no study has evaluated the detection of germline alterations in FFPE tumours using NGS-based tests. As formalin fixation is known to cause DNA damage, the usability of FFPE DNA for NGS-based germline testing must be compared to a gold standard such as DNA extracted from the peripheral blood mononuclear cell fraction. To determine reliability of using tumour DNA for germline variant calling, the retention rate of germline variants in the tumour genome must be measured. This is because tumour-specific mutations might result in loss of germline variants. Currently, there is also no standard method in distinguishing between germline and somatic variants in tumour-only analyses. Hence, an approach to distinguish between germline and somatic variants in tumour genomes must be established and assessed for its sensitivity and precision.

In this study, we aimed to determine whether potential germline alterations can be accurately identified through genomic analyses of FFPE tumours. We performed analytic validation of a clinical amplicon-based targeted sequencing panel for FFPE solid tumours by comparison with sequencing of DNA isolated from the peripheral blood mononuclear cell fraction, which is the gold standard for germline testing. We identified three objectives in this study:

1. Assess the degree of formalin-induced DNA damage in FFPE DNA
2. Determine the retention rate of germline alterations in FFPE tumours, and
3. Evaluate the use of VAF thresholds to distinguish germline alterations from somatic mutations in tumour-only analyses.

Through these analyses, we hoped to characterize formalin-induced DNA damage to facilitate quality control and improve robustness of our assay. Finally, we also aimed to measure the sensitivities for identifying potential germline alterations and positive predictive values (PPVs) for referring germline alterations to downstream germline testing at various VAF cut-offs. Establishing these performance parameters would serve as important guidelines in clinical practice.

Chapter 2

Materials and Methods

2.1 Overview of study design

This study examined whether potential germline alterations can be accurately identified in FFPE tumours. Targeted sequencing data from 213 cancer patients with FFPE tumour and matched blood samples were retrospectively analyzed (Figure 2.1). DNA was extracted from the peripheral blood mononuclear cell fraction, hereafter referred to as blood, and FFPE specimens. The DNA samples were sheared and enriched for amplicons in the OncoPanel, a clinical targeted NGS panel for solid tumours. Amplicons were barcoded and subjected to NGS. Sequencing data were processed and analyzed with a custom variant analysis 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.

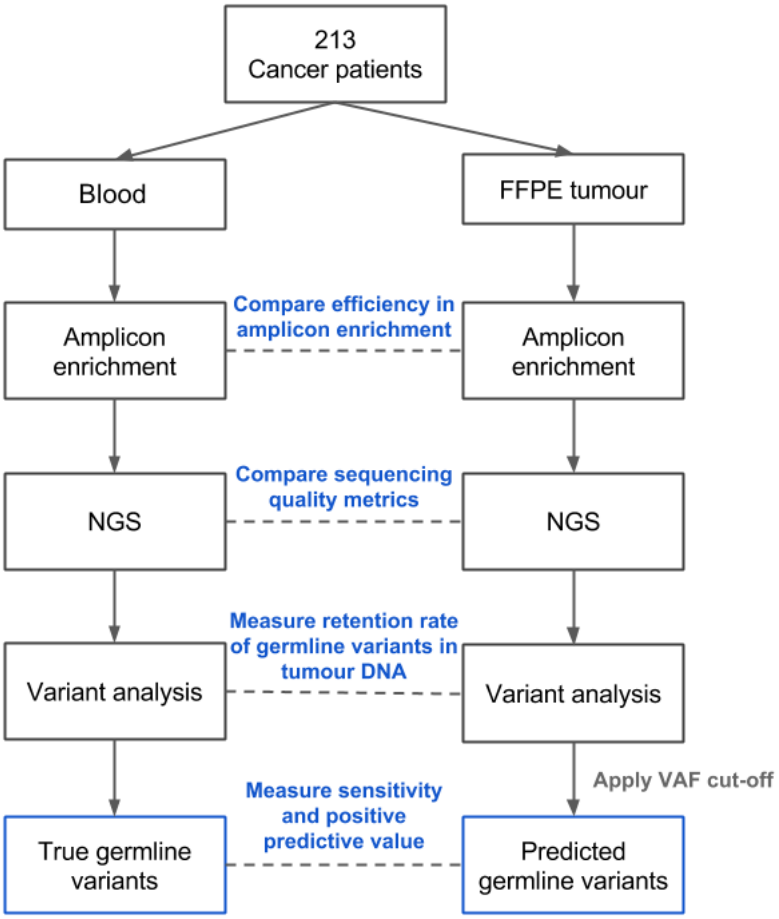


Figure 2.1: Schematic description of study design and data analyses.

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 assessed 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 were those with advanced cancers including CRC, lung cancer, melanoma, gastrointestinal stromal tumour (GIST), and other cancers (Table 2.1). The age of paraffin block for tumour samples ranged 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 the peripheral blood mononuclear cell fraction 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 [30] 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. Germline alterations in the six PGx genes could serve as predictors of susceptibility to 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 $\sim$ 20 Kb of target bases. Complete list of genes and gene reference models for the OncoPanel is presented in Table 2.2, whereas OncoPanel target regions and amplicons are presented in Table A.1.

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

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

2.5 Variant calling pipeline

2.5.1 Read alignment and variant calling

Reads that passed the Illumina Chastity filter were aligned to the hg19 human reference genome using the BWA mem algorithm (version 0.5.9) with default parameters, and the alignments were processed and converted to the BAM format using SAMtools (version 0.1.18). The SAMtools `mpileup` function (`samtools mpileup -BA -d 500000 -L 500000 -q 1`) was used to generate pileup files for all target bases followed by variant calling with the VarScan2 `mpileup2cns` (version 2.3.6) function with parameter thresholds of VAF ≥ 0.1 and Phred-scaled 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 contained 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 was used with parameter thresholds of VAF ≥ 0.25, VAF to call homozygote ≥ 0.9, 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`). 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 reflected percentage of reads with the risk alleles.

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

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

*Description of sequence variants according to the HGVS recommendations.

2.5.2 Variant filtering

Variant calls were filtered using the VarScan2 `fpfilter` function with fraction of variant reads from each strand ≥ 0.1 and default thresholds for other parameters (Table 2.4). The VarScan2 `fpfilter` removed 247 low quality variants. Seventy germline variants in the blood were also excluded from our analysis because these variants in the tumours were filtered by the VarScan2 `fpfilter`. There were also 16 risk allele calls in tumour samples that did not pass the strand filter, causing the removal of 10 risk allele calls in the blood samples from our evaluation. Overall, a total of 343 calls were excluded by the VarScan2 `fpfilter` and strand filter. Eleven low coverage calls ($\leq$ 100x) were also excluded from our analysis. 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 misalignment near indels, sequencing artifacts, primer masking, and primer artifacts (Table 2.5). 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 2.2B).

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

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 2.5: Spurious variants removed by the variant filtering pipeline.

Gene	Chr	Pos	Ref	Alt	Reason
EGFR	chr7	55249112	G	A	Sequencing artifact, alignment of different sized amplicons
EGFR	chr7	55249115	C	T	Sequencing artifact, alignment of different sized amplicons
KIT	chr4	55595593	G	A	Sequencing artifact
KIT	chr4	55599268	C	T	Variant masked by primer in FFPE specimen
MAPK1	chr22	22162126	A	G	Variant masked by primer in FFPE specimen
MTHFR	chr1	11856378	G	A	Sequence artifact
MTOR	chr1	11186783	G	A	Sequencing artifact within primer region
MTOR	chr1	11190646	G	A	Variant masked by primer in FFPE specimen
MTOR	chr1	11199428	G	T	Sequencing artifact
MTOR	chr1	11269434	C	T	Sequencing artifact
MTOR	chr1	11298014	C	T	Sequencing artifact
STAT3	chr17	40476769	C	T	Sequencing artifact, alignment of different sized amplicons
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	Alignment error near the indel, chr18:673443 c.*447_*452delTTAAAG
UGT1A1	chr2	234668879	CAT	C	Sequencing artifact at poly-AT sequence in promoter
UGT1A1	chr2	234668881	T	TAC	Alignment error/sequencing artifact at poly-AT sequence in promoter

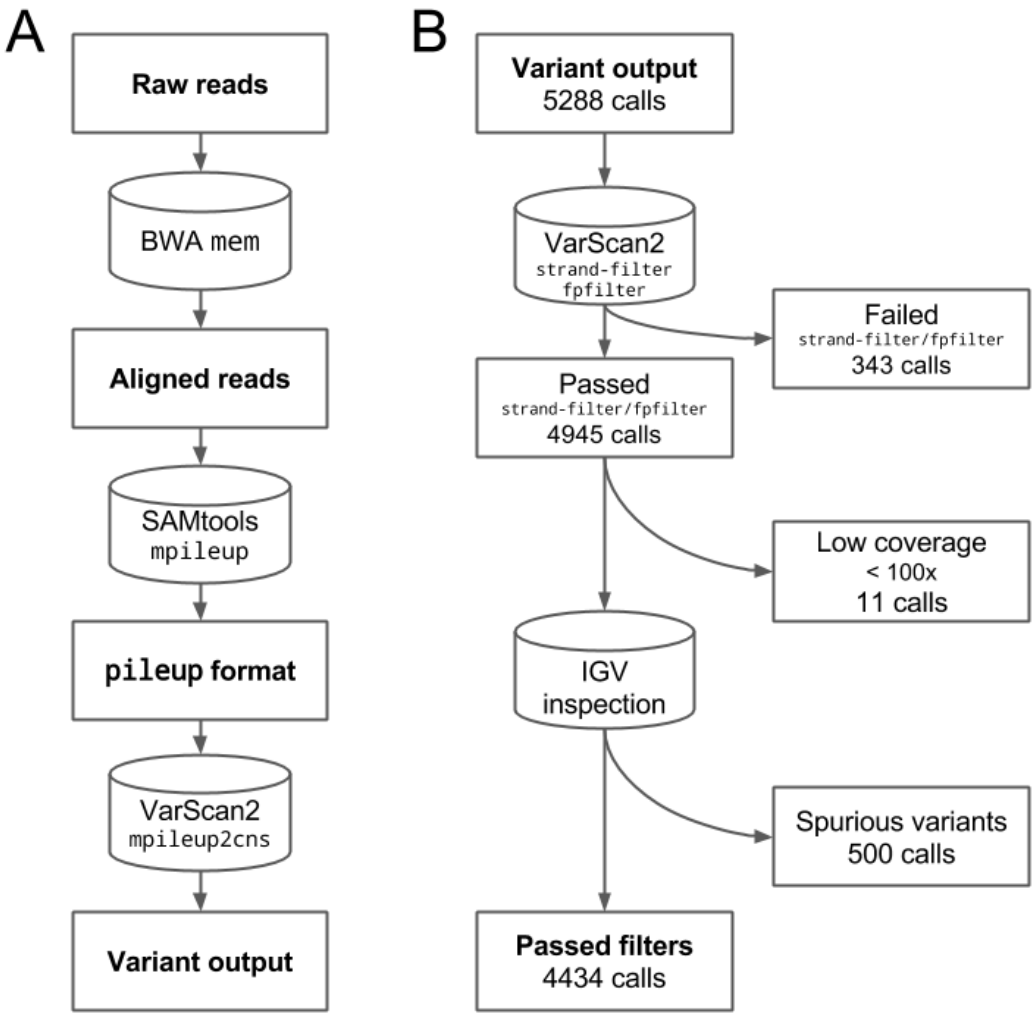


Figure 2.2: Pipelines for (A) variant calling and (B) filtering.

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 did not contain significant amount of microbial contaminants. Coverage depth for target bases with MAPQ ≥ 1 and BAQ ≥ 20 were 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 0.85 -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 excluded true variants identified by our VarScan2 variant calling pipeline and base changes with VAF $< 1\%$, which were 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 2.3). 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 would be referred to follow-up germline testing, PPVs were calculated at each VAF cut-off to evaluate precision of our approach. False positives were identified as predicted germline variants that were absent in the blood, and PPV was calculated by dividing the number of true positives with the sum of true positives and false positives.

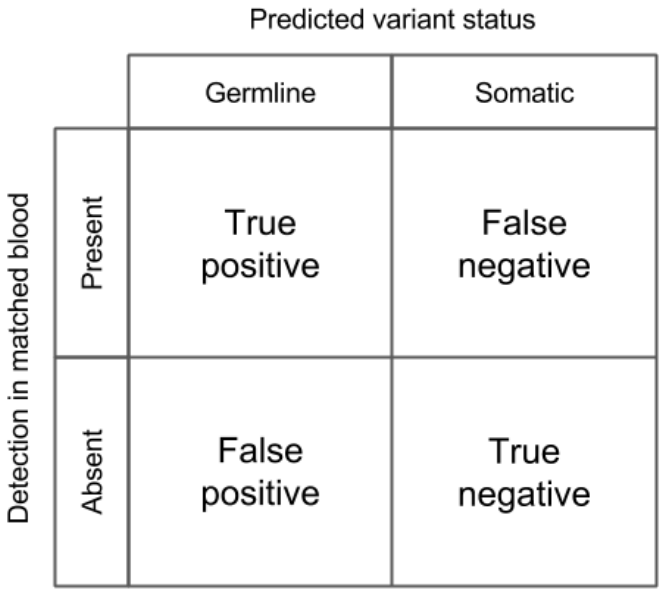


Figure 2.3: 2x2 contingency table for determination of true positive, false positive, true negative, and false negative variant calls in tumour-only analyses.

Chapter 3

Results: Assessment of Formalin-Induced DNA Damage in FFPE Specimens

Tumour biopsies and resections are often FFPE to preserve cellular morphology for pathological review, which is a requirement for standard of care. The FFPE method also enables storage of tissues at room temperature, minimizing cost and mitigating logistical difficulties in storing large archives of clinical specimens [127]. 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 [63, 106, 153, 154, 189, 220, 221]. We characterized formalin-induced DNA damage in our data to assess its impact on the utility of FFPE DNA for germline variant calling. As DNA derived from blood is one of the gold standards for germline testing, 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 [61, 63, 220, 221]. 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 varied across specimens in our study design, and we demonstrated that amplicon yield was weakly correlated with DNA

input for both blood and FFPE specimens (Spearman's rank correlation: blood, r_s = 0.29, 95% CI = 0.16–0.41, $p = 2.1 \times 10^{-5}$; FFPE, r_s = 0.25, 95% CI = 0.12–0.37, $p = 2.5 \times 10^{-4}$; Figure 3.1B). To account for the difference in DNA input across specimens, we derived the $\log_2$ fold change between DNA input and amplicon yield ($\log_2$ (Amplicon Yield/DNA Input)) to measure the efficiency in amplicon enrichment. We compared the $\log_2$ fold change in FFPE specimens to blood, and we found a significant decrease in enrichment efficiency in FFPE specimens compared to blood (Wilcoxon signed-rank test, W = 24754, Z = 12.7, $p = 4.6 \times 10^{-57}$, r = 0.61; Figure 3.1C). This result implies that production of amplicons was 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 = 10 178.5, 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 = 11 494.5, 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 = 19 069, 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 = 14 877, 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 did 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, $W = 20\,864$, $Z = 9.76$, $p = 2.5 \times 10^{-26}$, $r = 0.66$), 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 0x 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 0x and 100x coverage depth cut-off (Wilcoxon signed-rank test, $p < 0.0001$; Figure 3.3), we considered these discrepancies to be minor 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 was 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.

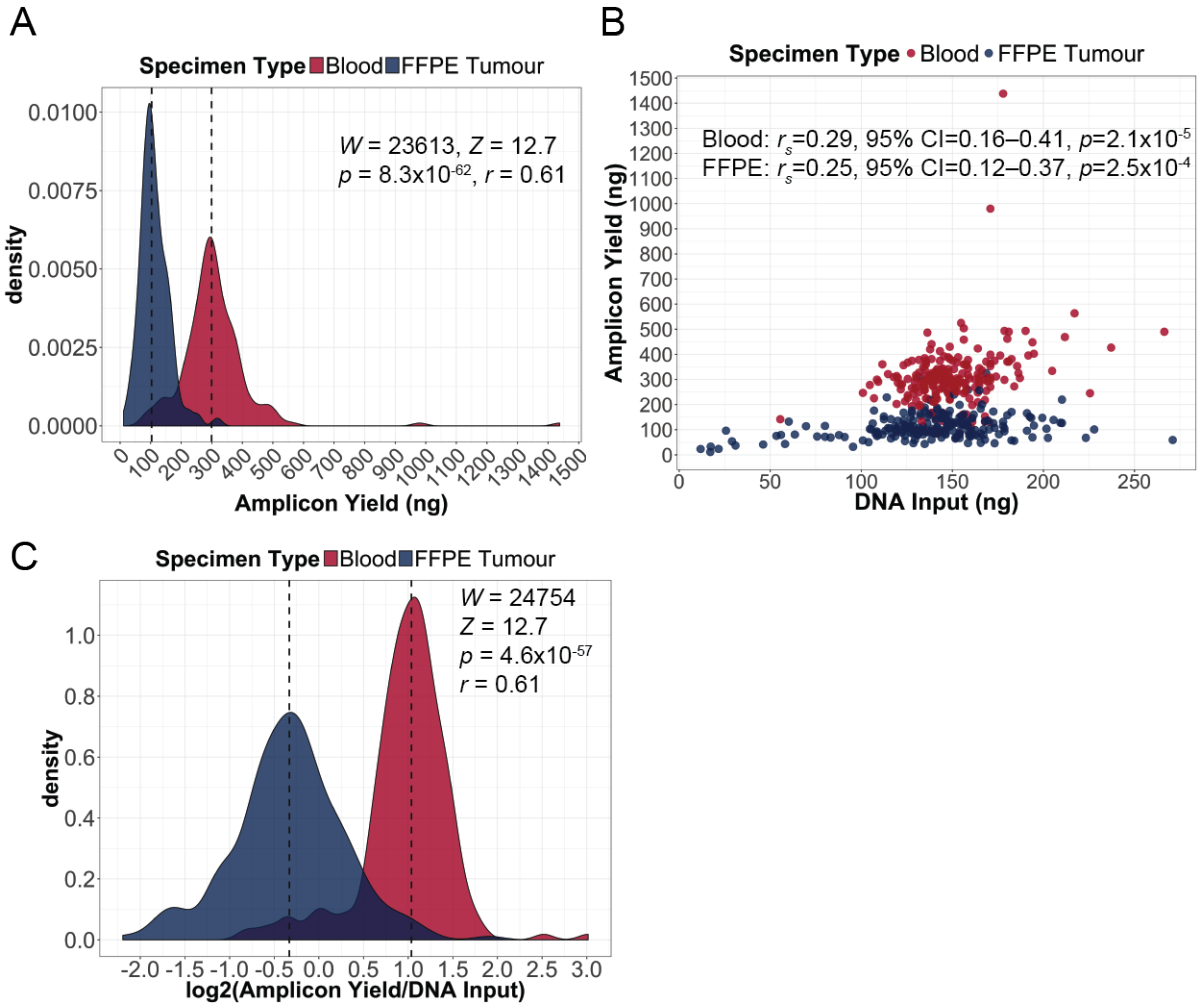


Figure 3.1: Comparison of efficiency in amplicon enrichment between blood and FFPE specimens. (A) Distributions of amplicon yield in blood and FFPE specimens (Wilcoxon signed-rank test). Dashed lines indicate median amplicon yield in blood and FFPE specimens, which are 299.3 ng and 103.6 ng, respectively.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

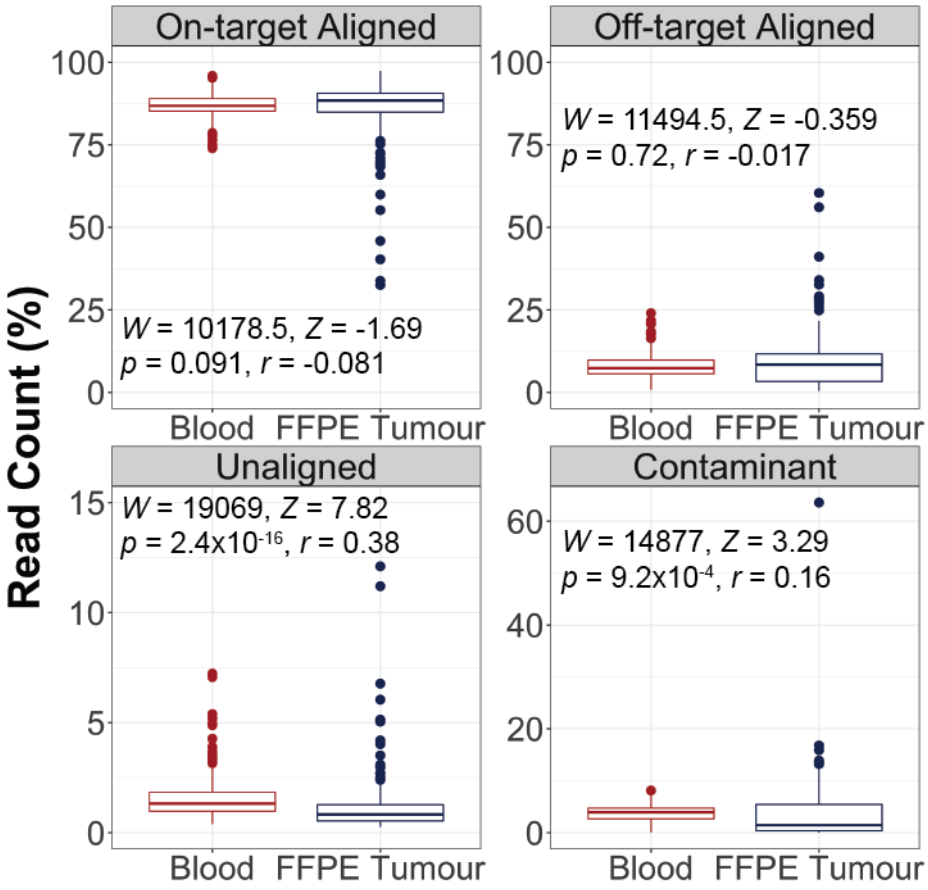


Figure 3.2: Assessment of read alignments between blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and interquartile range (IQR) of percentage of reads, with whiskers representing the range of data $\leq$ 1.5x the IQR and circles indicating outliers.

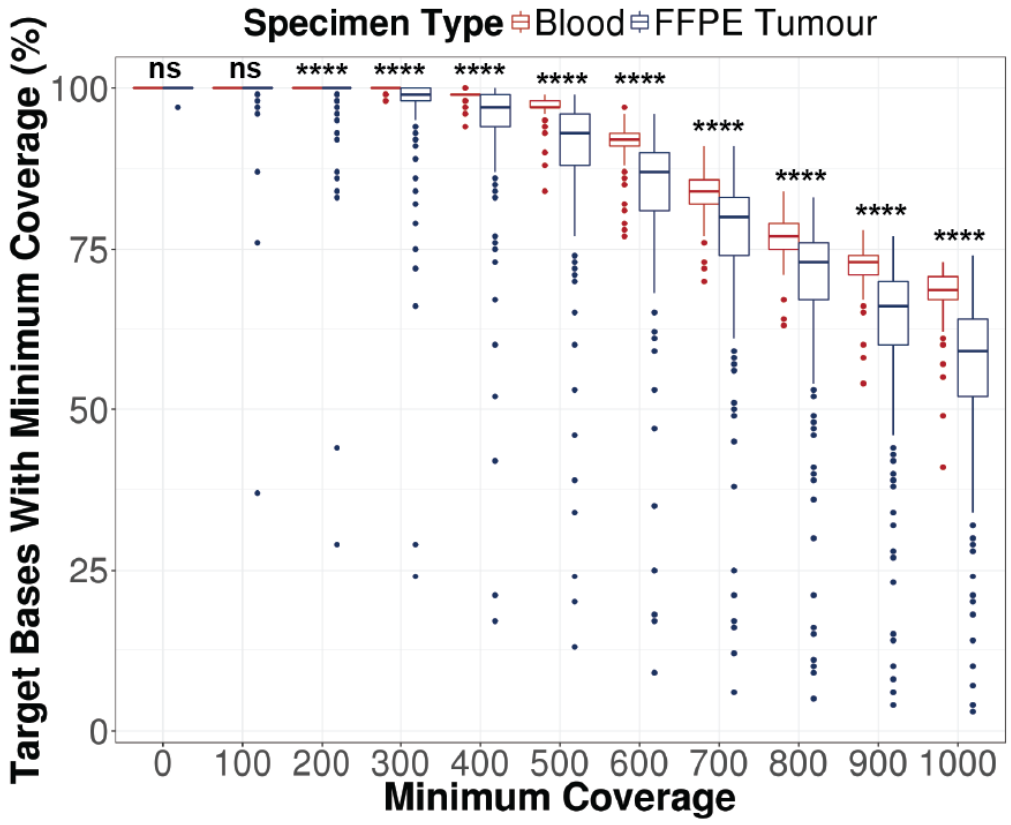


Figure 3.3: Evaluation of coverage uniformity in blood and FFPE specimens (Wilcoxon signed-rank test, ****$p < 0.0001$, ns = not significant). Per base coverage was normalized to account for difference in library size. Percentage of target bases that met various coverage thresholds was calculated. Box 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 $\leq$ 1.5x the IQR and circles indicating outliers.

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

	Blood		FFPE Tumour			
Threshold	Median (%)	Range (%)	Median (%)	Range (%)	$D^{\dagger}$ (%)	p ($< 0.0001^{*}$)
$\geq$ 0x	100	100–100	100	97.0–100	0.0	1.0
$\geq$ 100x	100	100–100	100	37.0–100	0.0	2.4×10^{-4}
$\geq$ 200x	100	100–100	100	29.0–100	0.0	$2.9 \times 10^{-11*}$
$\geq$ 300x	100	98.0–100	99.0	24.0–100	1.0	$4.1 \times 10^{-18*}$
$\geq$ 400x	99.0	94.0–100	97.0	17.0–100	2.0	$5.0 \times 10^{-28*}$
$\geq$ 500x	97.0	84.0–99.0	89.5	13.0–99.0	7.5	$2.1 \times 10^{-38*}$
$\geq$ 600x	92.0	77.0–97.0	87.0	9.0–96.0	5.0	$1.5 \times 10^{-32*}$
$\geq$ 700x	84.0	70.0–91.0	80.0	6.0–91.0	4.0	$5.7 \times 10^{-25*}$
$\geq$ 800x	77.0	63.0-84.0	73.0	5.0–83.0	4.0	$4.7 \times 10^{-27*}$
$\geq$ 900x	73.0	54.0–78.0	66.0	4.0–77.0	7.0	$4.6 \times 10^{-40*}$
$\geq$ 1000x	68.5	41.0–73.0	59.0	3.0-74.0	9.5	$3.6 \times 10^{-42*}$

†Absolute difference between median of blood and FFPE specimens.

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

The OncoPanel consists of 416 amplicons that interrogate coding exons and mutational hotpots of 21 genes, and these amplicons vary in length and GC content. Since we observed discrepancy in sequencing coverage between blood and FFPE specimens, we sought to determine whether this discrepancy was 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 336 out of 416 amplicons (Wilcoxon signed-rank test with Benjamini-Hochberg correction, adjusted $p < 0.0001$; Figure 3.4). To quantify the amplicon-specific differences in coverage depth, we derived the $\log_2$ fold change in the median coverage depth between blood and FFPE specimens ($\log_2$ (Median Coverage$_{\text{FFPE}}$/Median Coverage$_{\text{Blood}}$)) for each amplicon. Hence, a negative fold change indicates lower coverage depth of the amplicon in FFPE specimens relative to blood specimens, whereas a positive fold change indicates higher coverage depth of the amplicon in FFPE specimens relative to blood specimens. The volcano plot shows that 223 out of the 336 amplicons had negative $\log_2$ fold changes, whereas 113 out of the 336 amplicons had positive $\log_2$ fold changes (Figure 3.4). These results indicate that there were differences in coverage depth between FFPE and blood specimens for a large proportion 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 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 existed 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 = -0.77, 95% CI = -0.81– -0.73, $p = 1.4 \times 10^{-82}$; Figure 3.6A). This result indicates that coverage depth in FFPE specimens tended to be lower relative to blood specimens as amplicon length increased. On the other hand, coverage depth tended 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 = -0.32, 95% CI = -0.41– -0.23, $p = 1.8 \times 10^{-11}$; Figure 3.6B). Although the correlation is weak, this finding still implies that coverage depth in FFPE specimens tended to be lower relative to blood specimens as amplicon GC content increased, 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 had 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 ($F(2, 413) = 427.6$, $p = 2.41 \times 10^{-101}$), with an adjusted R^2 of 0.673. Predicted $\log_2$ fold change in amplicon coverage depth between blood and FFPE specimens is equal to

$$1.63 - 6.97 \times 10^{-3}(\textit{Length}) - 1.03 \times 10^{-2}(\textit{GC Content}),$$

in which amplicon length is expressed in base pairs (bp) and GC content is expressed as percentage (%). Both amplicon length and GC content were significant predictors of $\log_2$ fold change in amplicon coverage depth. Based on the standardized coefficients, we compared the strength of predictors within the model to identify the predictor with a greater effect on the response variable. Our assessment showed that one standard deviation increase in amplicon length would lead to a 0.756 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 0.288 standard deviation decrease in $\log_2$ fold change in amplicon coverage depth. This result indicates that amplicon length had 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 resulted in shorter template DNA that would not be amenable to PCR amplification of longer amplicons.

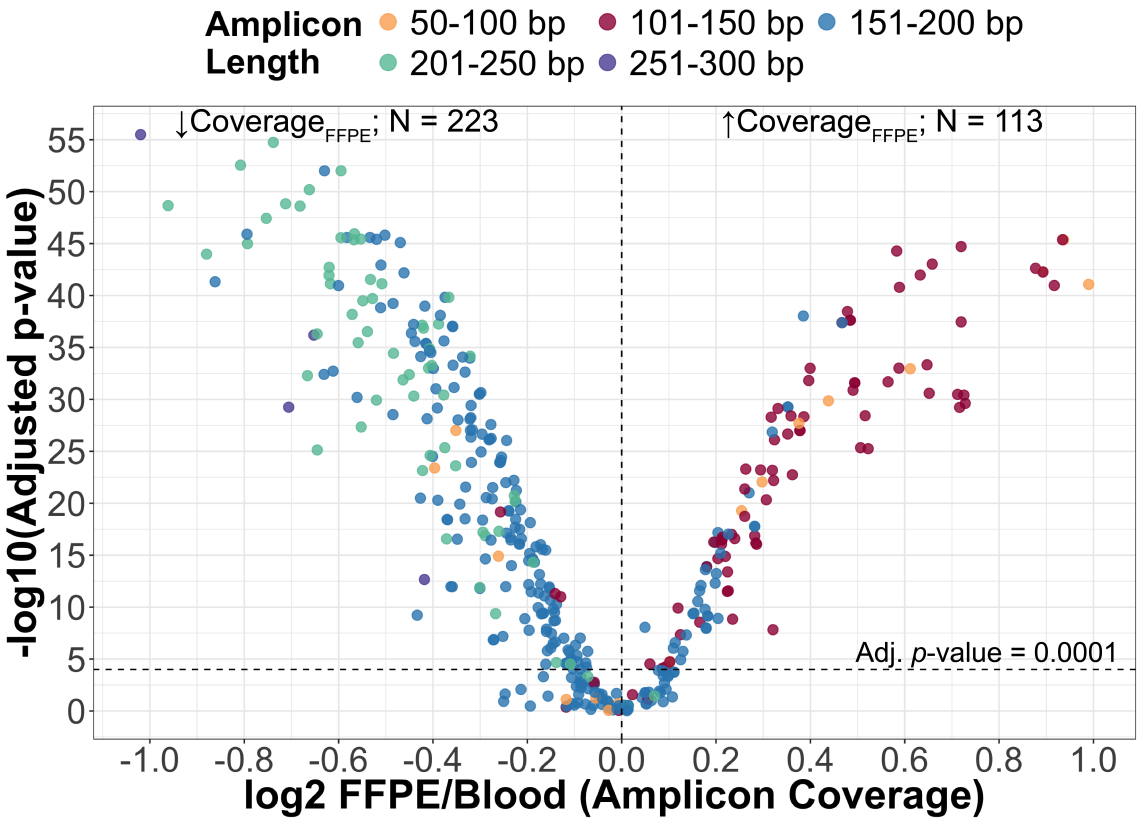


Figure 3.4: Amplicon-specific differences in coverage depth between blood and FFPE specimens. Difference in amplicon coverage depth between specimen types was determined using the Wilcoxon signed-rank test with Benjamini-Hochberg correction (adjusted $p < 0.0001$). Volcano plot illustrates the -$\log_{10}$ adjusted p-value in relation to $\log_2$ fold change between median coverage depth in blood and FFPE specimens ($\log_2$ (Median Coverage$_{FFPE}$/Median Coverage$_{Blood}$)) for amplicons in the panel. Negative $\log_2$ fold change indicates lower coverage depth of the amplicon in FFPE specimens relative to blood (↓ Coverage$_{FFPE}$), whereas positive $\log_2$ fold change indicates higher coverage depth of the amplicon in FFPE specimens relative to blood (↑ Coverage$_{FFPE}$). Color of points represents length of amplicons in base pairs (bp). N = number of amplicons

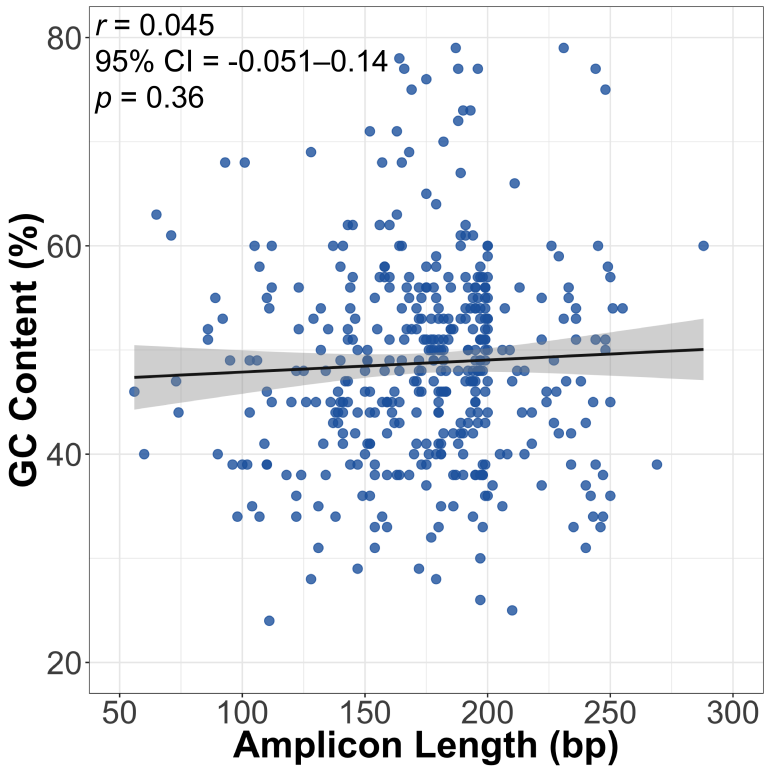


Figure 3.5: The relationship between amplicon GC content and amplicon length (Pearson's correlation). Solid line represents the fitted linear relationship between the two variables, and the shaded band indicates pointwise 95% confidence interval of the fitted linear regression line.

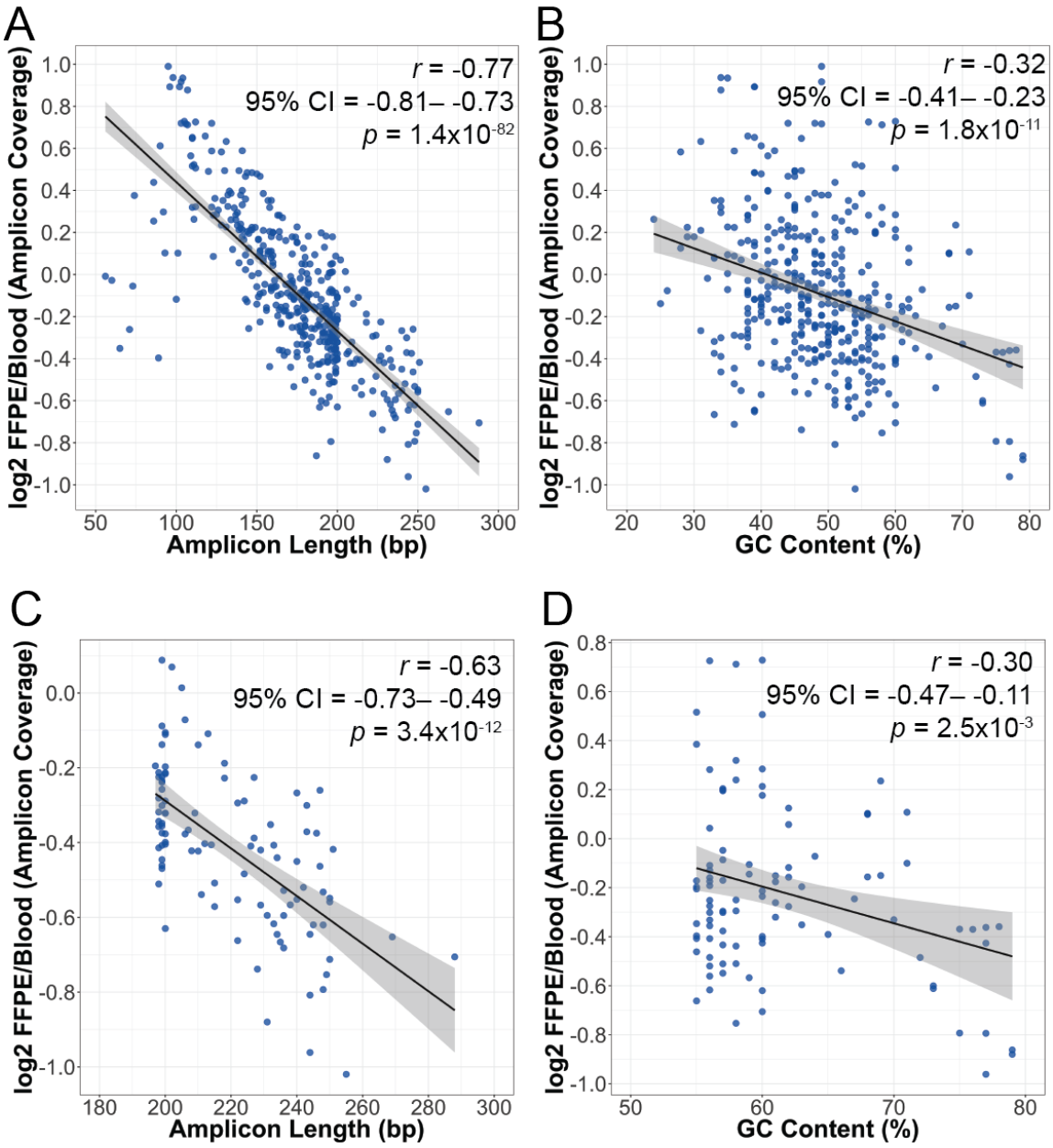


Figure 3.6: Scatter plots showing $\log_2$ fold change between amplicon coverage depth in blood and FFPE specimens ($\log_2$ (Median $\text{Coverage}_{\text{FFPE}}$/Median $\text{Coverage}_{\text{Blood}}$)) in relation to (A) amplicon length, (B) GC content, C) top 100 longest amplicons, and (D) top 100 amplicons with the highest GC content (Pearson's correlation). Solid line represents the fitted linear relationship between the two variables, and the shaded band indicates pointwise 95% confidence interval of the fitted linear regression line.

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

Variable	Unstandardized Coefficient	Standard Error	Standardized Coefficient	p-value
Length (bp)	-6.97×10^{-3}	2.59×10^{-4}	-7.56×10^{-1}	7.45×10^{-93}
GC Content (%)	-1.03×10^{-2}	1.01×10^{-3}	-2.88×10^{-1}	4.71×10^{-22}
				Intercept = 1.63, Adjusted R^2 = 0.673
				$F(2, 413) = 427.6$, p-value = 2.41×10^{-101}

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 [62–64, 94, 106, 153, 154, 221]. A prominent type of formalin-induced sequence artifact is C>T/G>A transitions as a result of deamination of cytosine bases [63, 106, 125, 154, 221]. To measure the level of formalin-induced artifacts in FFPE specimens, we quantified the fraction of base changes that were not identified as true single nucleotide variants (SNVs) by our variant calling pipeline. We only considered high quality bases (BAQ $\geq$ 20) and base changes that were $\geq$ 1% allele frequency to exclude sequencing errors from our analysis. Base changes were categorized into C>T/G>A and A>G/T>C, which are nucleotide transitions, as well as C>A/G>T, A>C/T>G, C>G/G>C, and A>T/T>A, which are nucleotide transversions. We compared the fraction of base changes between specimen types and found significant differences in fraction of C>T/G>A and A>G/T>C between blood and FFPE specimens (Wilcoxon signed rank test, $p < 0.0001$; Figure 3.7A). As blood DNA is not affected by formalin fixation, we evaluated the prevalence of artifactual base changes in FFPE specimens with respect to blood by calculating the fold change between the median fraction of base changes in blood and FFPE specimens (Table 3.3). We noted a substantially larger 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. Increased C>T/G>A artifacts was consistent with cytosine deamination effects that are reportedly predominant in FFPE DNA. On the other hand, A>G/T>C artifacts could be caused by deamination of adenine to generate hypoxanthine, which forms base pairs with cytosine instead of thymine, changing A-T base pairs to G-C base pairs. Deamination of adenine to hypoxanthine can be catalyzed by an acidic environment [214], which can arise in FFPE specimens because formaldehyde can be oxidized to generate formic acid [63].

To assess the relative difference in fraction of base changes in FFPE specimens compared to blood specimens, we calculated the $\log_2$ fold change in fraction of base changes between paired blood and FFPE specimens ($\log_2$ (Fraction of Base Changes$_{\text{FFPE}}$/Fraction of Base Changes$_{\text{Blood}}$)). We compared the relative difference in fraction of base changes across different types of base changes, and a Kruskal-Wallis test indicated that type of base changes had 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 confirmed that deamination of cytosine bases is the most frequent form of sequence artifact in FFPE DNA.

Formalin-induced sequence artifacts often occur at low allele frequency; hence, we examined the prevalence of sequence artifacts at different ranges of allele frequency, including 1–10%, 10-20%, and 20-30%. Because variants were not called within the 1–10% allele frequency range, we did not remove true SNVs detected by our variant calling pipeline to ensure consistency when comparing fraction of base changes across different ranges of allele frequency. Nevertheless, we adhered to the previous criterion of only including base changes with BAQ $\geq$ 20 in this analysis. For all types of base changes, we noted that the range of allele frequency had 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 (Figure 3.7; Table 3.3), 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 transition was present and detectable in FFPE specimens, which indicated that deamination of cytosine is the primary form of formalin-induced sequence artifact, and these artifactual transitions were more prevalent at low, but clinically relevant allele frequency.

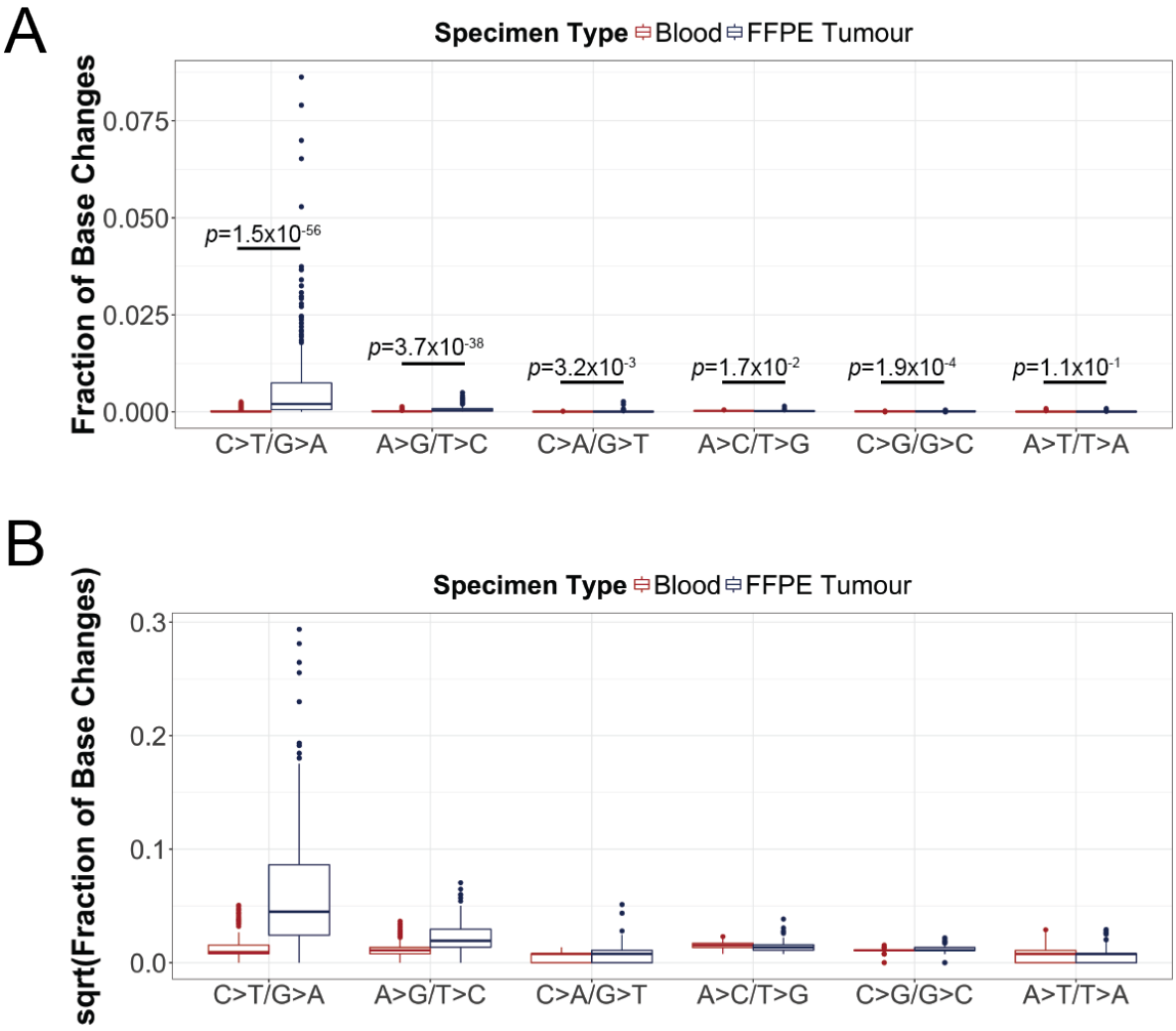


Figure 3.7: Assessment of formalin-induced sequence artifacts in FFPE specimens. (A) Comparison of fraction of base changes in blood and FFPE specimens (Wilcoxon signed-rank test). Box plots show the median (horizontal bar within) and IQR of fraction of base changes for different types of base changes, with whiskers representing the range of data $\leq$ 1.5x the IQR and circles indicating outliers. (B) Box plots showing square root-transformed fraction of base changes on the Y-axis.

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

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

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

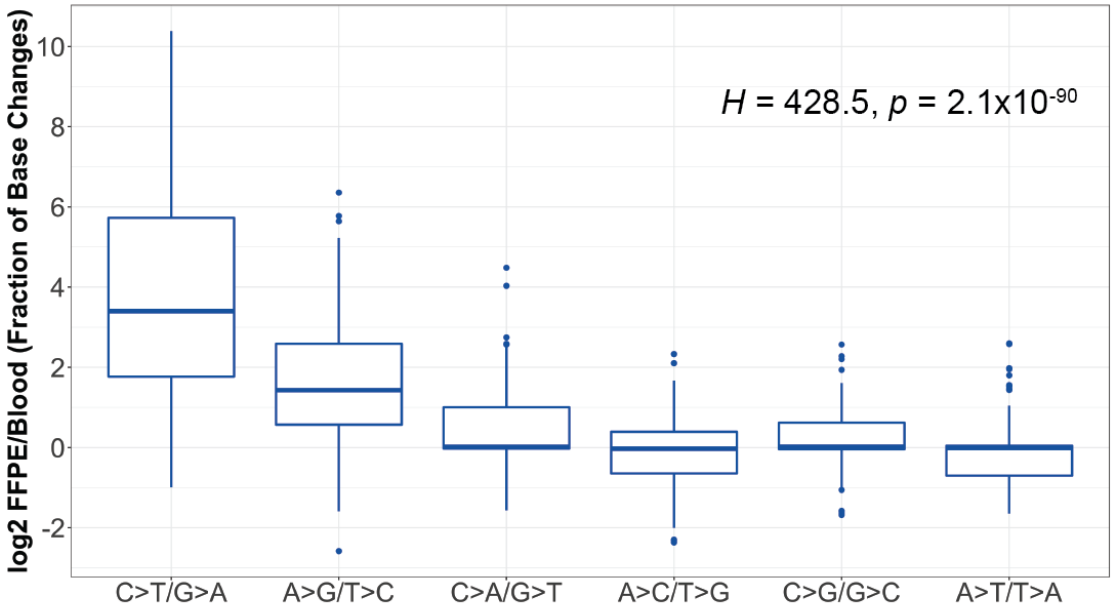
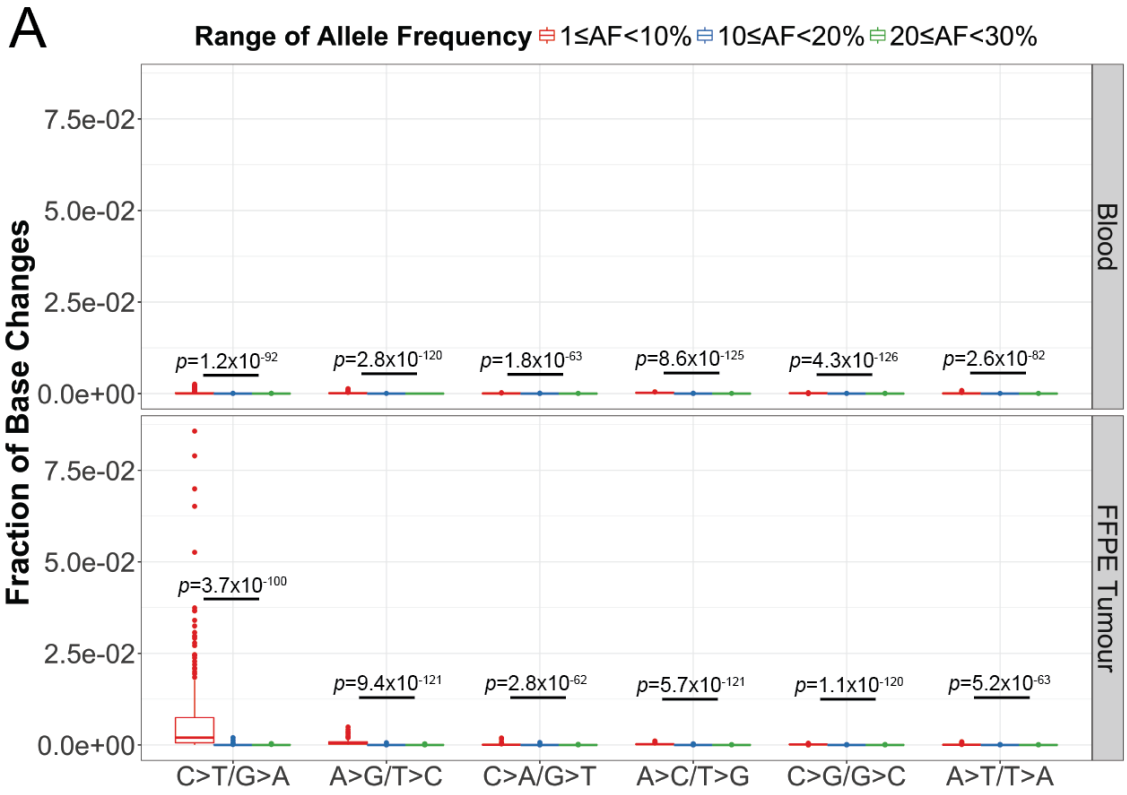


Figure 3.8: Comparison of relative difference in fraction of base changes in FFPE specimens compared to blood (Kruskal-Wallis test). Relative difference was measured as $\log_2$ fold change between fraction of base changes in blood and FFPE specimens ($\log_2$ (Fraction of Base Changes$_{\text{FFPE}}$/Fraction of Base Changes$_{\text{Blood}}$)). Box plots show the median (horizontal bar within) and IQR of $\log_2$ fold change for different types of base changes, with whiskers representing the range of data $\leq$ 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.

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


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

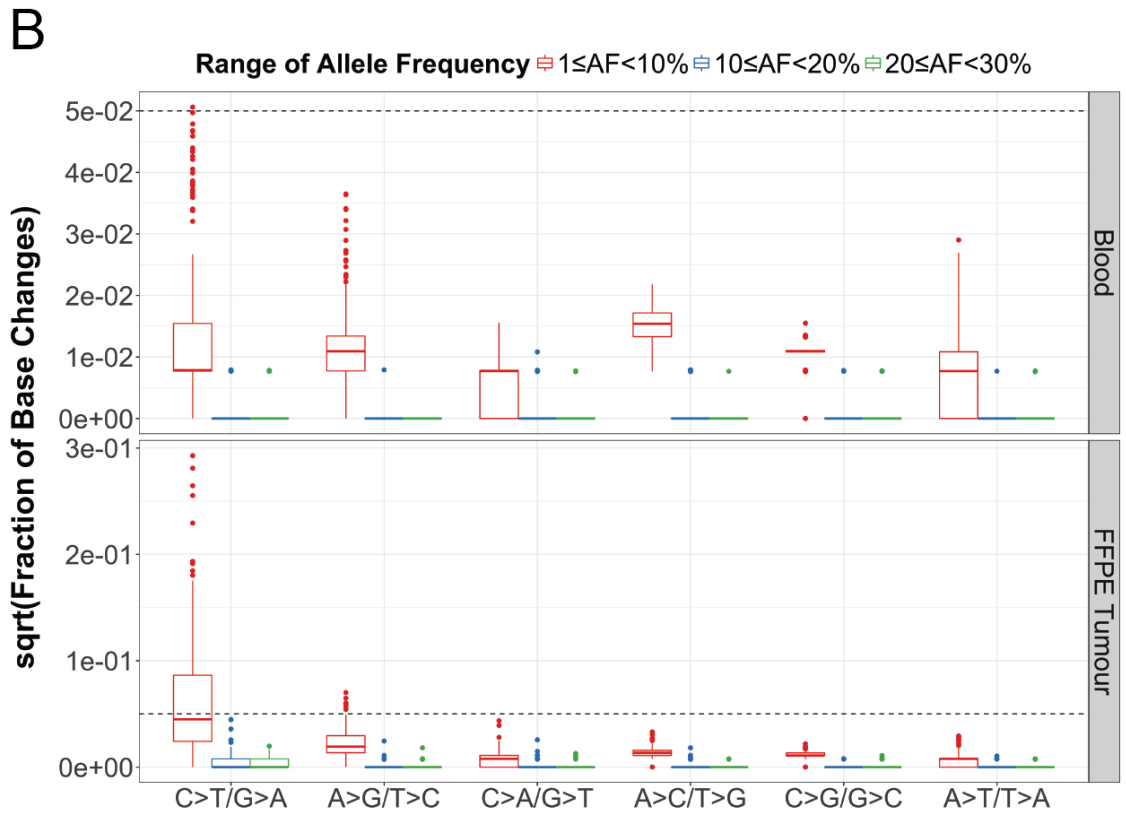

B
Range of Allele Frequency 1≤AF<10% 10≤AF<20% 20≤AF<30%
sqrt(Fraction of Base Changes)
5e-02
4e-02
3e-02
2e-02
1e-02
0e+00
3e-01
2e-01
1e-01
0e+00
Blood
FFPE Tumour
C>T/G>A
A>G/T>C
C>A/G>T
A>C/T>G
C>G/G>C
A>T/T>A

Figure 3.9: Assessment of formalin-induced sequence artifacts in FFPE specimens at different ranges of allele frequency. (A) Comparison of fraction of base changes across different ranges of allele frequency (Kruskal-Wallis test). Box plots show the median (horizontal bar within) and IQR of fraction of base changes for different types of base changes, with whiskers representing the range of data $\leq$ 1.5x the IQR and circles indicating outliers. (B) Box plots demonstrating square root-transformed fraction of base changes across different ranges of allele frequency. Dashed lines equal to 0.05 to indicate that the Y-axis scales are different for blood and FFPE tumour plots.

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

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

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

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

The amount of amplifiable DNA derived from FFPE specimens is dependent on the extent of fragmentation damages. Given two FFPE DNA samples of similar quantity, the sample with more extensive DNA fragmentation would yield reduced amount of PCR amplicons compared to the less fragmented sample [61, 221]. As the age of paraffin blocks in our study ranged 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 tended to yield lower amount of amplicons. Because the amount of DNA input for production of amplicons varied 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. 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 was less efficient in FFPE DNA extracted from older paraffin blocks, which was likely caused by more substantial DNA fragmentation.

As DNA fragmentation results in reduced template DNA for PCR amplification, this leads to a higher probability for enrichment of sequence artifacts [38, 221]. Our previous evaluation indicated that older paraffin blocks were associated with lower efficiency in amplicon enrichment, which was possibly due to increased fragmentation damages in the extracted DNA (Figure 3.10). This led 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, was 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 like 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 were 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, $p = 2.0 \times 10^{-13}$), 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, $p = 1.4 \times 10^{-7}$). 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, $p = 3.8 \times 10^{-17}$), and a weak, positive correlation between percentage of on-target aligned reads and efficiency in amplicon enrichment (Spearman's rank correlation, r_s = 0.34, 95% CI = 0.22–0.45, $p = 1.2 \times 10^{-7}$). Since efficiency in amplicon enrichment was 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, $p = 1.4 \times 10^{-18}$). As reduced efficiency in amplicon enrichment is an indicator for low amount of template DNA, the consequent increase in C>T/G>A changes could be 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 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.

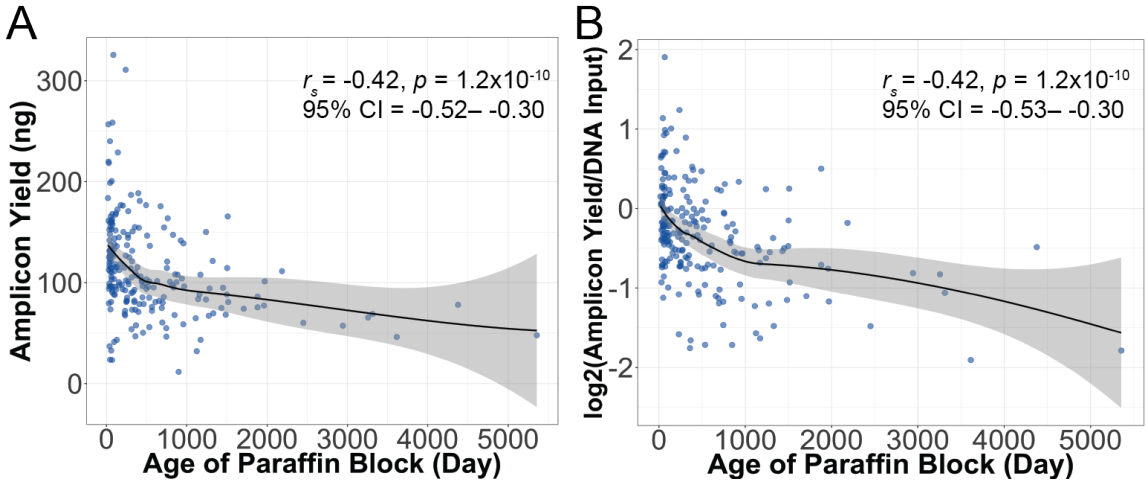


Figure 3.10: Scatter plots showing (A) amplicon yield and (B) efficiency in amplicon enrichment, which is represented by the $\log_2$ fold change between the amount of DNA input for producing amplicons and amplicon yield, in relation to age of paraffin blocks (Spearman's rank correlation). Solid lines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

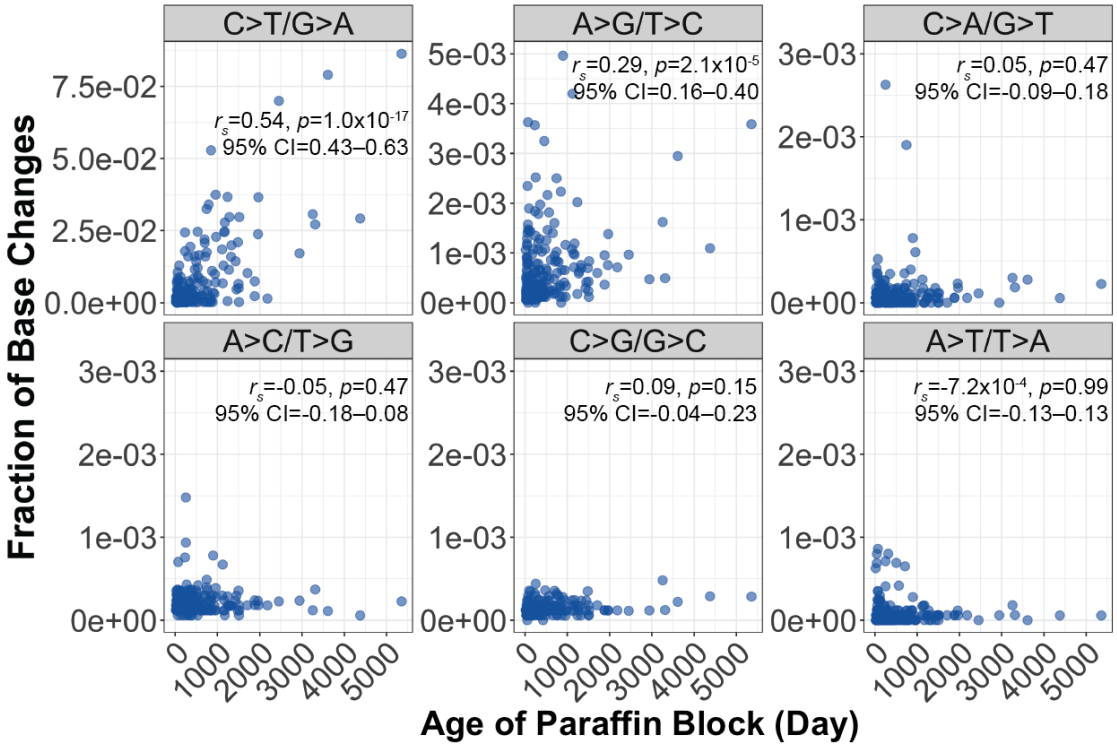


Figure 3.11: The relationship between fraction of base changes and age of paraffin block for different types of base changes (Spearman's rank correlation).

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) 1.2×10^{-10}*			
Fraction of C>T/G>A	-0.55 (-0.64– -0.45) 1.4×10^{-18}*	0.54 (0.43–0.63) 1.0×10^{-17}*		
Average Per Base Normalized Coverage	0.52 (0.42–0.61) 3.8×10^{-17}*	-0.47 (-0.57– -0.36) 2.0×10^{-13}*	-0.80 (-0.84– -0.75) 1.3×10^{-48}*	
On-target Aligned Reads (%)	0.34 (0.22–0.45) 1.2×10^{-7}*	-0.35 (-0.46– -0.23) 1.4×10^{-7}*	-0.57 (-0.65– -0.47) 2.2×10^{-20}*	0.73 (0.66–0.79) 0*

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

Chapter 4

Results: Identification of Germline Alterations in FFPE Tumours

Clinical tumour sequencing can guide disease management and therapeutic intervention for cancer patients. Although simultaneous identification of clinically important germline variants could be performed if tumour-normal pairs are sequenced, matched normal samples are not routinely processed in the clinical setting due to logistical constraints and limited funding and time [77, 78, 125, 219]. Genomic analyses of tumours can reveal both germline and somatic alterations [27, 102, 140, 141, 184]. However, this requires a method to separate germline variants from somatic mutations. The TOP study is comprised of 213 patients with tumour and matched blood specimens. This enabled us to measure the retention rate of germline alterations in tumour DNA to confirm that tumour DNA is a reliable source for identification of germline alterations. We also evaluated the use of VAF thresholds to distinguish between germline and somatic statuses of variants identified in the tumour DNA. Our goal was to establish the sensitivities for identifying potential germline alterations and PPVs for referring germline alterations to downstream germline testing at various VAF cut-offs. These parameters could serve as guidelines in the clinical practice.

4.1 Frequency and interpretation of germline alterations in patients from the TOP cohort

We examined 15 cancer-related genes and six PGx genes in DNA isolated from blood samples from the 213 cancer patients in the TOP cohort. We identified a total of 1990 germline alterations that passed our filtering criteria (Figure 2.2B). 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 $39/50$ of these variants were synonymous, $8/50$ were missense variants, $2/50$ occurred within splice regions, and $1/50$ were frameshift variants. Based on the ExAC database, $28/50$ germline variants had population allele frequency $< 1\%$, whereas $18/50$ germline variants had population allele frequency $\geq 1\%$. Four out of 50 germline variants were not reported 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 $8/50$ benign variants, $8/50$ likely benign variants, $6/50$ annotated as benign/likely benign, $2/50$ with conflicting interpretations of pathogenicity, and $1/50$ with uncertain significance. We were unable to determine the clinical significance of $24/50$ 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 and 5-fluorouracil-based adjuvant chemotherapy [25, 28, 31, 46, 97, 103, 105, 224, 226, 227, 230, 231]. 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 clinical management 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 $13/23$ germline variants were missense variants, $4/23$ were synonymous, $2/23$ occurred within splice regions, $2/23$ occurred upstream of a gene, $1/23$ were located at splice donor sites, and $1/23$ were present at the 3' untranslated region. Based on the ExAC database, $8/23$ germline variants had population allele frequency $< 1\%$, whereas $10/23$ germline variants had population allele frequency $\geq 1\%$. Five out of 23 germline variants were not reported 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 $5/23$ variants were categorized as either benign or likely benign, $4/23$ with conflicting interpretations of pathogenicity, $2/23$ submitted without assessment of clinical significance, and $1/23$ with uncertain significance. There was also $4/23$ 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 $7/23$ germline alterations that were associated with drug response. These alterations were *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 [144, 158].

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 the 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) was present in 97 out of 213 patients (46%), and 208 out of 213 (98%) TOP patients had at least one germline PGx variant that was associated with drug response (Figure 4.1; Figure 4.2).

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

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

Gene	Chr	Pos	ID*	HGVS*	Zygosity wt-var†, var-var††	Total	Pct‡ (%)
	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

Gene	Chr	Pos	ID*	HGVS*	Zygosity wt-var†, var-var††	Total	Pct‡ (%)
	2	191859906	rs61756197	p.Gln275Gln c.825G>A	1, 0	1	0.9
	2	191859935	rs41473544	p.Val266Ile c.796G>A	2, 0	2	0.9
	2	191872307	rs45463799	p.Asn118Asn c.354C>T	3, 0	3	1
	2	191874667	rs386556119; rs2066802	p.Leu21Leu c.63T>C	42, 3	45	21
STAT3	17	40469241	COSM979464	c.2100C>T	1, 0	1	0.5
	17	40475056	rs117691970	p.Gly618Gly c.1854C>T	4, 0	4	2
	17	40486040	rs200098006	p.Leu275Leu c.825T>G	2, 0	2	0.9
	17	40486043	NA	p.Gln274Gln c.822A>G	1, 0	1	0.5
	17	40498635	rs146184566; COSM979479	p.Ser75Ser c.225G>A	1, 0	1	0.5
	17	40498713	NA	p.Lys49Lys c.147A>G	1, 0	1	0.5
	17	40498722	NA	p.Ala46Ala c.138G>T	1, 0	1	0.5
TP53	17	7577069	rs55819519; COSM44017	p.Arg290His c.869G>A	1, 0	1	0.5
	17	7577553	COSM44368	p.Met243fs c.727delA	1, 0	1	0.5
	17	7578210	rs1800372; COSM249885	p.Arg213Arg c.639A>G	1, 0	1	0.5
	17	7578420	COSM1386804	p.Thr170Thr c.510G>A	1, 0	1	0.5
	17	7579472	rs1042522; COSM250061	p.Arg72Pro c.215G>C	73,24	97	46
	17	7579579	rs1800370	p.Pro36Pro c.108G>A	5, 0	5	2

Gene	Chr	Pos	ID*	HGVS†	Zygosity wt-var‡, var-var‡‡	Total	Pct§ (%)
				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: Interpretation 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.	[117]
	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.	[29, 116, 215, 228]

Gene	Chr:Pos	ID*	HGVS[*]	AF[**]	Variant Effect[†]	Clinical Significance[††]	Functional/Clinical Impacts	Ref.
	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.	[215]
	7:55259450	rs2229066; COSM85893; rs17290559	p.Arg836Arg c.2508C>T	1.7	Syn.	Benign/Likely benign	NA	NA
KIT	4:55592059	rs151016327; COSM3760661	p.Thr461Thr c.1383A>G	0.28	Syn.	Benign	NA	NA
	4:55599268	rs55789615; COSM1307	p.Ile798Ile c.2394C>T	2.1	Syn.	Benign/Likely benign	NA	NA
	4:55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	12	Syn.	Benign/Likely benign	NA	NA
MAPK1	22:22162126	rs386488966; rs3729910	p.Tyr43Tyr c.129T>C	4.5	Syn.	NA	NA	NA
	22:22221623	rs201495639	p.Tyr36Tyr c.108C>T	0.052	Syn.	NA	NA	NA

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
MTOR	1:11169420	rs41274506	p.Asp2485Asp c.7455C>T	0.33	Syn.	NA	NA	NA
	1:11172909	NA	p.Glu2456Lys c.7366G>A	0.00082	Missense	NA	NA	NA
	1:11174452	NA	p.Arg2408Gln c.7223G>A	NA	Missense	NA	NA	NA
	1:11181327	rs11121691	p.Leu2303Leu c.6909G>A	22	Syn.	NA	Likely has an effect on exonic splicing enhancer or exonic splicing silencer binding site activity.	[233]
	1:11184593	rs56051835	p.Leu2208Leu c.6624T>C	0.49	Syn.	Benign	NA	NA
	1:11188172	rs370318222	p.Tyr1974Tyr c.5922C>T	0.00082	Syn.	NA	NA	NA
	1:11190646	rs2275527	p.Ser1851Ser c.5553C>T	22	Syn.	Benign	NA	NA
	1:11190730	rs17848553	p.Ala1823Ala c.5469C>T	2.4	Syn.	Benign	NA	NA

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	1:11194521	COSM180791	c.5133C>T	0.029	Splice region	NA	NA	NA
	1:11205058	rs386514433; rs1057079‡	p.Ala1577Ala c.4731A>G	32	Syn.	NA	One study reported improved clinical response and progression-free survival in advanced esophageal squamous cell carcinoma patients with the AG genotype compared to the AA genotype who were treated with paclitaxel plus cisplatin chemotherapy.	[126]
	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

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	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.	[233]
	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

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
PDGFRA	4:55141055	rs1873778; COSM1430082‡	p.Pro567Pro c.1701A>G	99	Syn.	Benign	No association with PDGFRα expression in colorectal cancer.	[73]
	4:55152040	rs2228230; COSM22413	p.Val824Val c.2472C>T	18	Syn.	Benign	NA	NA
STAT1	2:191851646	rs41270237	p.Thr385Thr c.1155G>A	0.42	Syn.	Likely benign	NA	NA
	2:191856001	rs41509946	p.Gln330Gln c.990G>A	0.36	Syn.	Likely benign	NA	NA
	2:191859906	rs61756197	p.Gln275Gln c.825G>A	0.025	Syn.	NA	NA	NA
	2:191859935	rs41473544	p.Val266Ile c.796G>A	0.20	Missense	Likely benign	Functional testing indicated that the variant was not a gain-of-function mutation in STAT1	[60]
	2:191872307	rs45463799	p.Asn118Asn c.354C>T	0.32	Syn.	Likely benign	NA	NA

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	2:191874667	rs386556119; rs2066802	p.Leu21Leu c.63T>C	8.5	Syn.	Benign	High frequency among patients with multiple sclerosis and chronic hepatitis C.	[76]
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

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
TP53	17:7577069	rs55819519; COSM44017	p.Arg290His c.869G>A	0.016	Missense	Conflicting interpretations of pathogenicity	A conservative amino acid substitution that was predicted to be possibly damaging by *in silico* analysis. Reported in patients with Li-Fraumeni syndrome and cancer patients without family histories of Li-Fraumeni syndrome or Li-Fraumeni-like syndrome.	[11, 12, 47, 160, 162, 210]
	17:7577553	COSM44368	p.Met243fs c.727delA	NA	Frameshift	NA	Reported in esophageal squamous cell carcinoma of patients from northern Iran.	[20]
	17:7578210	rs1800372; COSM249885	p.Arg213Arg c.639A>G	1.2	Syn.	Benign/Likely benign	One study demonstrated that this variant was not a predictive biomarker for initiation and progression of gastroesophageal reflux disease, Barrett's Esophagus, and esophageal cancer in the Brazilian population.	[161]

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	17:7578420	COSM1386804	p.Thr170Thr c.510G>A	0.012	Syn.	NA	One study reported that TP53 mutations in exon 5, which include this variant, were associated with the worst prognosis for patients with non-small-cell lung cancer.	[209]
	17:7579472	rs1042522; COSM250061‡	p.Arg72Pro c.215G>C	34	Missense	Drug response	p53 protein with Arg72 was associated with increased apoptosis, while p53 protein with Pro72 demonstrated increased G_1 cell-cycle arrest and activation of p53-dependent DNA repair. Pro/Pro genotype resulted in severe neutropenia in ovarian cancer patients receiving cisplatin-based chemotherapy, and poor survival and treatment response in gastric cancer patients receiving paclitaxel and capecitabine combination chemotherapy, as well as 5-fluorouracil-based adjuvant chemotherapy. Conflicting evidence on risk of predispositon to various cancer types.	[25, 28, 31, 46, 97, 103, 105, 224, 226, 227, 230, 231]

Gene	Chr:Pos	ID*	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	17:7579579	rs1800370	p.Pro36Pro c.108G>A	1.3	Syn.	Benign/Likely benign	NA	NA

*dbSNP and/or COSMIC IDs.

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

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

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

††Clinical significance on ClinVar database.

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

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

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

Gene	Chr	Pos	dbSNP ID	HGVS*	Zygosity wt-var†, var-var††	Total	Pct‡ (%)
	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: Interpretation 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.	[7, 24, 40, 58, 65, 115, 137, 142, 146, 152, 185, 200, 205, 207, 208]
	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.	[24, 58, 81, 185, 207, 208]

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	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.	[7, 40, 58, 81, 115, 142, 146, 185, 200, 205, 207, 208]
	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.	[24, 152]
	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.	[142, 152, 185, 205, 208]

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	1:98039419	rs56038477	p.Glu412Glu c.1236G>A	1.5	Syn.	Benign	Synonymous variant in high linkage disequilibrium with c.1129-5923C>G (rs75017182) in haplotype B3 (HapB3). rs75017182 causes nonsense mutation in exon 11, resulting in reduced DPD activity. Associated with fluoropyrimidine-related toxicity.	[7, 58, 142, 148]
	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.	[58, 81, 152, 200, 207, 208]
	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.	[40, 81, 146, 201, 208]

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
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.	[5, 45, 95, 139, 176, 196]
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.	[74, 75, 100, 134, 176]
	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.	[50, 74, 75, 90, 100, 134, 176, 185, 198]

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
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.	[37, 96, 101]
	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.	[88, 101, 192]
	22:50964493	NA	p.Glu413Lys c.1237G>A	NA	Missense	NA	NA	NA

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
	22:50964907	rs201685922	c.929_932delCCGC	0.49	Splice region	Conflicting interpretations of pathogenicity	Observed in a German American patient with mitochondrial neurogastrointestinal encephalomyopathy (MNGIE), but relation with TP enzymatic defect was not established.	[151]
	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

Gene	Chr:Pos	dbSNP ID	HGVS*	AF**	Variant Effect†	Clinical Significance††	Functional/Clinical Impacts	Ref.
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.	[4, 67, 85, 90, 131, 196]
UGT1A1	2:234668870	rs873478	c.-64G>C	1.1‡‡	Upstream gene	NA	Unknown	[43, 225, 229]
	2:234668879	rs34983651	c.-55_-54insAT	33‡‡	Upstream gene	Conflicting interpretations of pathogenicity, affects, association	Lower UGT1A1 expression and associated with irinotecan-related toxicity.	[8, 56, 82, 99, 111, 133, 139, 174, 177, 204]

[*]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.

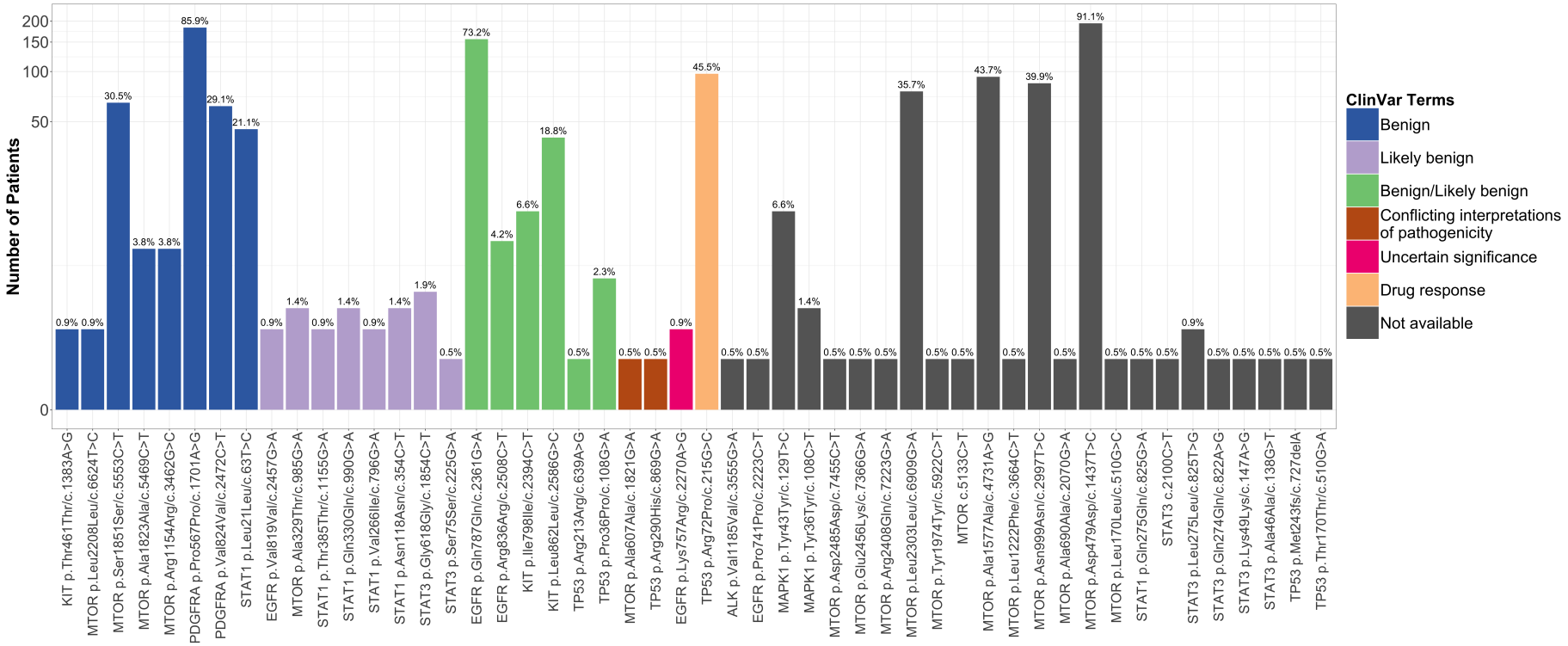


Figure 4.1: Distribution of germline alterations in cancer-related genes in patients from the TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. The TP53 variant, p.Arg72Pro/c.215G>C,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

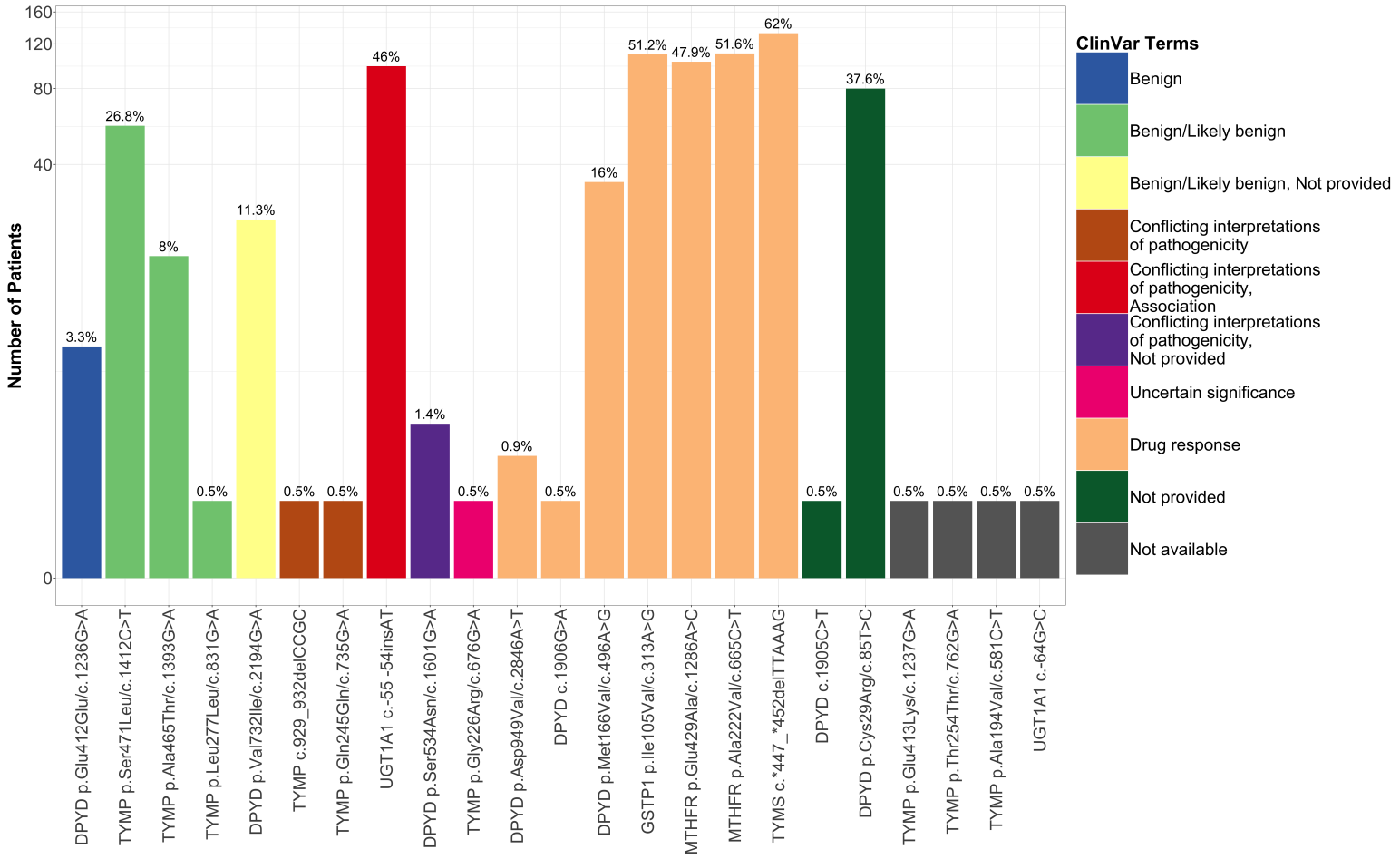


Figure 4.2: Distribution of germline alterations in PGx genes in patients from the TOP study. Percentage of patients is calculated for each variant and annotated above individual bars. Color of bars represent options for clinical significance in the ClinVar database. 208 out of 213 patients in the TOP 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 [102, 141, 169, 184]. 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 the TOP study, we were still able to identify 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 the 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 the 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, 2022 germline variants were detected in the blood. We found that $1981/2022$ germline variants in the blood were retained in the tumours, giving rise to a concordance rate of 98.0% (Figure 4.3). Eighty five out of 1981 germline variants did not retain the same allele status between blood and tumour: $83/85$ were heterozygous in the blood specimens but homozygous in the tumours, whereas $2/85$ were homozygous in the blood specimens but heterozygous in the tumours. We also identified 41 germline variants in the blood that were discordant with the tumours. These germline variants were present in the blood but not detected in the tumours. Thirty four out of 41 discordant variants were heterozygous in the blood specimens but wild type in the tumours. The sum of these 34 germline variants with the 83 germline variants that were heterozygous in the blood specimens but homozygous in the tumours gave rise to a total of 117 LOH events out of 2022 germline variants in the blood, which resulted in a LOH rate of 5.8%. The remaining $7/41$ discordant germline variants were caused by low sequencing depth ($<$ 100x) in the tumours. Table 4.5 shows the 85 germline variants that were discordant between blood and tumour due to allelic status and the 41 discordant germline variants that were present in the blood but not detected in the tumours.

Multiple factors could contribute to the discordant calls aside from LOH events, 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 [13, 87]. 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.

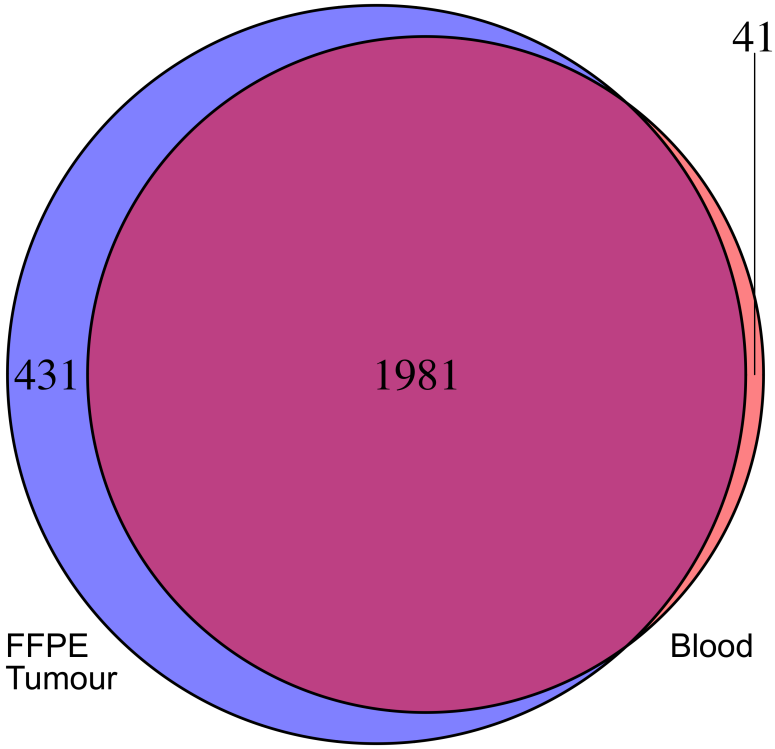


Figure 4.3: Venn diagram demonstrating concordance of variants identified in 217 tumour-blood paired samples.

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

Gene	Chr:Pos	ID*	HGVS[*]	Clinical Significance[†]	Reason for discordance (Blood/Tumour)	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	1
	1:98165091	rs2297595	p.Met166Val c.496A>G	Drug response	Het/Hom	1
	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	3
EGFR	7:55249063	rs1050171; COSM1451600	p.Gln787Gln c.2361G>A	Benign/Likely benign	Het/Hom	1
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	7
KIT	4:55602765	rs3733542; COSM1325	p.Leu862Leu c.2586G>C	Benign/Likely benign	Het/Hom	4
MTHFR	1:11854476	rs1801131	p.Glu429Ala c.1286A>C	Drug response	Het/Hom	6

Gene	Chr:Pos	ID*	HGVS*	Clinical Significance†	Reason for discordance (Blood/Tumour)	Count
	1:11856378	rs1801133	p.Ala222Val c.665C>T	Drug response	Het/Hom	6
	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	1
	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	2
	1:11205058	rs1057079; rs386514433	p.Ala1577Ala c.4731A>G	NA	Het/Hom	4
	1:11205058	rs1057079; rs386514433	p.Ala1577Ala c.4731A>G	NA	Het/WT	4
	1:1272468	rs17036536	p.Arg1154Arg c.3462G>C	Benign	Het/Hom	2
	1:11288758	rs1064261	p.Asn999Asn c.2997T>C	NA	Het/Hom	2

Gene	Chr:Pos	ID*	HGVS*	Clinical Significance†	Reason for discordance (Blood/Tumour)	Count
	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	4
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	2
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	13
	17:7579472	COSM250061; rs1042522	p.Arg72Pro c.215G>C	Drug response	Het/WT	4

Gene	Chr:Pos	ID*	HGVS[*]	Clinical Significance[†]	Reason for discordance (Blood/Tumour)	Count
	17:7579579	rs1800370	p.Pro36Pro c.108G>A	Benign/Likely benign	Het/Hom	1
TYMP	22:50964236	rs11479	p.Ser471Leu c.1412C>T	Benign/Likely benign	Het/Hom	7
TYMS	18:673443	rs151264360	c.*447_*452delTTAAAG	Drug response	Het/Hom	16
	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	2
	2:234668879	rs34983651	c.-55_-54insAT	Conflicting interpretations of pathogenicity, Association	Hom/WT	2
					Total discordant variants = 126	

*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 VAF thresholds to separate germline alterations from somatic mutations in tumour-only analyses

Through variant analysis of DNA from blood specimens, we identified germline alterations that were 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 98.0%. Together, these analyses confirmed that clinically relevant germline alterations were present in our data set and a large proportion of germline alterations could be identified in the tumour DNA. Next, we sought to evaluate the use of VAF thresholds to separate germline alterations from somatic mutations in tumour-only analyses. This assessment would determine whether application of VAF thresholds is an accurate method to identify potential germline alterations in clinical tumour sequencing for follow-up germline testing. While our data set did not contain pathogenic germline variants, we anticipated that this approach could be used to detect germline genetic events associated with cancer predisposition and drug response 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 tended to have VAFs close to 50%, whereas homozygous alterations in the blood tended to have VAFs close to 100%. However, the VAF distribution of germline variants in the tumours tended 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, LOH, or DNA damage as a result of formalin fixation, indicated 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. At each VAF cut-off, we determined the number of true positives by identifying variants in the tumours that overlapped with germline variants in matched blood samples. True positive rate (sensitivity) was then calculated as the fraction of variants that were correctly identified as germline using the VAF threshold over the total number of germline variants in the tumours. We noted that the sensitivity of germline variant detection in FFPE tumours had lowered with the increase of VAF thresholds. At VAF thresholds of 15%, 20%, 25%, and 30%, 0.99 (95% CI = 0.99–1.0), 0.98 (95% CI = 0.97–0.98), 0.96 (95% CI = 0.95–0.97), and 0.94 (95% CI = 0.93–0.95) sensitivities of detecting germline variants in the tumours were achieved, respectively (Table 4.6).

Because clinical genomics requires accurate identification of genetic alterations that are clinically important, potential germline alterations identified through tumour-only analyses must be referred to follow-up testing [27, 86, 169]. 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 caused 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 tended to be concentrated at lower percentages compared to VAFs of germline variants. We measured PPVs at various VAF thresholds to examine the precision of referring germline variants to follow-up testing. 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. PPV was then calculated as the fraction of true positives over total number of variants identified in the tumours, including somatic mutations (false positives). We noted that as VAF thresholds increased, PPV of referring germline variants to follow-up testing increased. At VAF thresholds of 15%, 20%, 25%, and 30%, 0.86 (95% CI = 0.85–0.87), 0.88 (95% CI = 0.86–0.89), 0.89 (95% CI = 0.87–0.90), and 0.90 (95% CI = 0.89–0.91) PPVs of correctly referring predicted germline variants to downstream germline testing were achieved, respectively (Table 4.7).

Despite the difference in VAF distributions between germline alterations in blood and tumour samples, VAF thresholds at 15%, 20%, 25%, and 30% managed to achieve high sensitivity ($\geq$ 0.94) for detecting germline variants in the tumours. 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. Evidently, VAF cut-offs of 15%, 20%, 25%, and 30% could enable precise referral of potential germline variants to follow-up testing (PPV $\geq$ 0.86). Missed cases in detecting clinically important germline variants could lead to severe clinical ramifications. For example, failure to follow-up on a patient with a potential germline variant that is associated with cancer predisposition would mean losing out on an opportunity to inform his or her family members. Because confirmatory testing is mandatory for potential germline variants detected in the tumour, we suggest selecting a VAF cut-off with a high sensitivity to minimize the number of missed cases. Based on our evaluation, 15% VAF threshold would maximize sensitivity by identifying 99% of germline variants in the tumours, although 14% of the predicted germline variants would be somatic mutations (false positives). Overall, we demonstrated that the use of VAF thresholds is a promising approach to accurately identify potential germline alterations in clinical tumour sequencing.

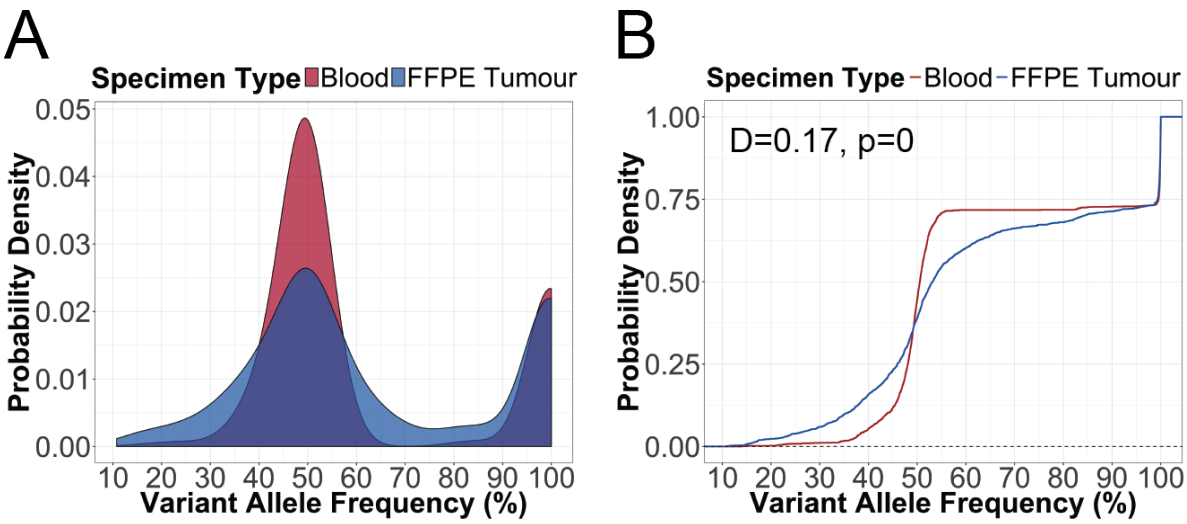


Figure 4.4: Assessment of using a VAF cut-off approach to identify germline alterations in tumour-only analyses. (A) Comparison of VAF distributions of germline alterations between blood and tumour. (B) Empirical cumulative distribution of VAFs of germline alterations in blood and tumour samples (Kolmogorov-Smirnov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

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	Miss rate	95% CI
10	0	1981	1.0	1.0–1.0	0	0–0.0019
15	13	1968	0.99	0.99-1.0	0.0066	0.0035–0.011
20	46	1935	0.98	0.97–0.98	0.023	0.017–0.031
25	77	1904	0.96	0.95–0.97	0.039	0.031–0.048
30	117	1864	0.94	0.93–0.95	0.059	0.049–0.070
35	192	1789	0.90	0.89–0.92	0.097	0.084–0.11
40	313	1668	0.84	0.83–0.86	0.16	0.14–0.17
45	458	1523	0.77	0.75–0.79	0.23	0.21–0.25

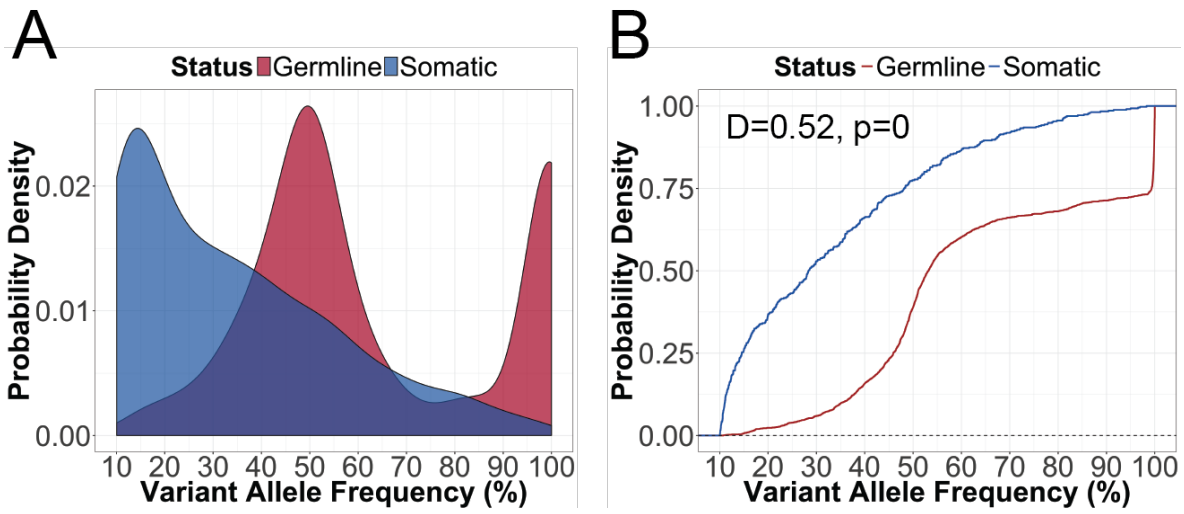


Figure 4.5: Assessment of using a VAF cut-off approach to refer potential germline alterations in tumour-only analyses to follow-up testing. (A) Comparison of VAF distributions between germline and somatic alterations in tumour specimens. (B) Empirical cumulative distribution of VAFs of germline and somatic alterations in tumour samples (Kolmogorov-Smirnov test).

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

Clinical tumour sequencing can inform medical decision-making for cancer patients. While sequencing of tumour-normal pairs would enable accurate identification of somatic mutations and simultaneous detection of clinically important germline variants, matched normal samples are often not obtained in clinical practice. Genomic analyses of tumours can reveal actionable somatic mutations and clinically relevant germline alterations [102, 141, 184]. However, tumour tissues are commonly FFPE, which represents a challenge in clinical genomics. Formalin fixation damages nucleic acid through fragmentation and cytosine deamination, affecting molecular testing with FFPE DNA [63, 106, 153, 154, 189, 220, 221]. 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 were consistent with DNA fragmentation and cytosine deamination being 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 minor and could be minimized using methods such as using shorter amplicons and enriching for longer DNA templates. We also found that 98.0% of germline alterations identified in the blood using our panel test were present in the FFPE tumours. This implies that a high proportion of germline genetic changes were retained in the tumour genome, demonstrating the reliability of using tumour DNA for germline variant calling. Finally, we assessed the application of VAF thresholds to delineate germline and somatic variants in tumour-only analyses. We reported that a VAF cut-off of 15

5.1 Formalin-induced DNA damage has minor effects on sequencing metrics

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. [188], Didelot et al. [61], and Wong et al. [220] demonstrated that shorter amplicons gave rise to better PCR amplification success in FFPE DNA, indicating the presence of fragmentation damage, which yielded 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 [19, 195, 220]. Additionally, Wong et al. [221] and Didelot et al [61] 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 either minor or statistically insignificant. As for the discrepancy in amplicon enrichment, shorter amplicons can be designed to circumvent the drawback of fragmentation damage in FFPE samples.

5.2 Sequence artifacts induced by cytosine deamination tend to occur at low allelic frequency

Cytosine deamination is a major cause of sequence artifacts in formalin-fixed specimens [44, 62, 64, 106, 154, 195, 221]. Herein, we observed increased C>T/G>A artifacts in FFPE DNA compared to blood. Artifactual C>T/G>A changes are formed by incorporation of adenines in the complementary DNA strand at uracil lesions generated by deamination of cytosines [63]. When measuring frequency of sequence artifacts at different allele frequency ranges, Wong et al. [221] reported higher C>T/G>A transitions at a lower allele frequency range (1–10% *vs.* 10–25%). This finding led us to compare the fraction of base changes at different allele frequency ranges, including 1–10%, 10–20%, and 20–30%. Indeed, we observed a substantial increase in C>T/G>A within the 1–10% allele frequency range. However, we were unable to separate FFPE artifacts from low-allelic-fraction somatic mutations within the different allele frequency ranges due to the lack of matched fresh frozen or unfixed tumour tissues. Somatic mutations can occur at VAFs that deviate significantly from a diploid zygosity (i.e. heterozygous variant should have VAF close to 50%, whereas homozygous variant should have VAF close to 100%) because of low tumour content or

tumour heterogeneity [34, 39, 104, 202, 222]. Therefore, further workflow optimization should be performed for the purpose of identifying clinically relevant somatic mutations in the tumour genome. A method to reduce sequence artifacts caused by cytosine deamination is treatment with uracil-DNA glycosylase (UDG) before sequencing. UDG is an enzyme capable of depleting uracil lesions in DNA, giving rise to abasic sites. During PCR amplification, cytosine bases are restored at abasic sites by using the complementary DNA strand as template, which consists of guanine bases opposite of the uracil lesions [63]. Several studies showed that pre-treatment of FFPE DNA with UDG can markedly reduce C>T/G>A sequence artifacts [62, 64, 106]. However, this approach cannot correct sequence artifacts at CpG dinucleotides because these cytosines are typically methylated, and deamination of 5-methyl cytosines generates thymines instead of uracil bases, which are resistant to UDG repair [64].

5.3 Sequence artifacts other than those caused by cytosine deamination are detected

We also observed elevated levels of A>G/T>C artifacts in FFPE DNA, albeit to a lesser extent compared to C>T/G>A artifacts. Likewise, Wong et al. [218] reported that 35% of sequence artifacts in Sanger sequencing of the *BRCA1* gene were A>G/T>C nucleotide changes. We speculate that increase in A>G/T>C artifacts is caused by deamination of adenine to generate hypoxanthine, which forms base pairs with cytosine instead of thymine. This results in transformation of A-T base pairs to G-C base pairs. Deamination of adenine to hypoxanthine can be catalyzed by an acidic environment [214], which can arise in FFPE specimens because formaldehyde can be oxidized to generate formic acid [63].

Acidic conditions also promotes depurination, creating abasic sites. Many DNA polymerases selectively incorporate adenines across abasic sites, while guanines and small deletions are integrated in fewer cases [92]. Despite being statistically insignificant, we observed a subset of FFPE specimens with higher fractions of C>A/G>T artifacts. These artifactual changes could have resulted from depurination of guanines, followed by incorporation of adenines by DNA polymerase in the complementary strand, which alters G-C base pairs to A-T base pairs. Heyn et al. [92] reported that DNA polymerases demonstrated varying bypass rates at abasic sites. For instance, AmpliTaq Gold, *Pfu*, and Platinum Taq HiFi extended across lower frequency of abasic sites compared to Platinum Taq, *Bst* and *Sso*-Dpo4 (<34% *vs.* >77%) [92]. Thus, selection of a high fidelity DNA polymerase could lessen these forms of sequence artifacts.

Costello et al. [51] discovered that C>A/G>T artifacts can also occur due to oxidation of DNA during the shearing process, converting guanines to 8-oxoguanine lesions. This conversion is highly dependent on the surrounding 5' and 3' bases of the guanine, in which guanines within GGC are the most susceptible to oxidation. 8-oxoguanine can form base pairs with cytosine and adenine, and mispairing with adenine would give rise to artifactual C>A/G>T transversions. However, this was

not the cause of C>A/G>T artifacts in our data because both blood and FFPE DNA were sheared, and we did not observe simultaneous C>A/G>T increments in both specimen types compared to other types of base changes.

5.4 Storage time of FFPE blocks correlates with the extent of formalin-induced DNA damage

Ludyga et al. [128] demonstrated that long-term storage of FFPE blocks led to increased DNA fragmentation, producing shorter template DNA for PCR amplification. Furthermore, Carrick et al. [38] showed that increased storage time of FFPE blocks affects sequencing coverage and depth in NGS data. These findings are in agreement with our results, in which we found negative associations between age of paraffin blocks and efficiency in amplicon enrichment, coverage depth of target bases, and percentage of on-target aligned reads. As well, we observed a positive correlation between age of paraffin blocks and fraction of C>T/G>A artifacts, an outcome of stochastic enrichment. Due to exposure to environmental conditions, older FFPE blocks tend to produce increasingly fragmented DNA, which results in lower amounts of amplifiable DNA. Consequently, there is a higher chance of amplifying template DNA with sequence artifacts caused by formalin, yielding increased frequency of artifactual nucleotide changes in older FFPE specimens [221]. These results demonstrating the correlations between storage time of paraffin blocks and sequencing variables suggest that if multiple FFPE blocks are available, the specimen with the shorter storage time should be selected for molecular testing. However, clinical specimens are often limited, making sample selection a rare option in the diagnostic setting. As such, other approaches to eliminate sequence artifacts should be considered such as application of molecular barcodes and hybridization-capture enrichment, which allow tracking of DNA templates [71, 159, 180, 220]. This would enable detection of variants that are only supported by the same template DNA, indicating a higher chance that these variants are sequence artifacts and should be interpreted with caution.

5.5 Germline variants are highly retained in the tumour genome

Various groups have identified clinically significant germline alterations through analyzing tumour genomes [102, 140, 141, 184]. Schrader et al. [184] reported that potential pathogenic germline variants in cancer-predisposing genes were conserved in the tumours of 91.9% of patients in their study cohort (182 of 198 patients), whereas 21.4% of these patients (39 of 182 patients) demonstrated LOH or other forms of mutations in the remaining wild type allele. We found that 98.0% (1981/2022) of germline alterations identified in the blood were retained in the tumour, a finding that is in line with previous work. This suggests that tumour DNA could be a reliable source for detecting germline alterations, implying that a tumour-only sequencing protocol could be leveraged for pre-screening of germline variants before submission to downstream confirmatory testing. A

framework as such could provide germline testing in a time- and cost-effective manner because only selected patients (i.e. those with potential germline alterations that are clinically important) would require follow-up.

Out of the 98.0% of germline alterations that were retained in the tumours, 4.3% (85/1981) of variants demonstrated discordant allelic status between blood and tumour. We also identified discordant germline variants, which were present in the blood but not detected in the tumour, that were caused by LOH (heterozygous variant in the blood but wild type in the tumour; 34/41 variants) and low sequencing coverage ($<$ 100x; 7/41 variants). All tumour specimens in our study were formalin-fixed, therefore it is possible that DNA damage induced by formaldehyde exposure played a role in creating discordant germline variants. Variant discordance can also be caused by mutagenesis in the tumour, such as somatic CNVs in the region of the germline variant. For instance, Gross et al. [87] showed a high prevalence of *DPYD* CNVs in high-grade triple negative breast cancer, particularly in cases with copy number loss of the *BRCA1* DNA-repair gene. The common fragile site FRA1E is located within the *DPYD* gene and its stability is highly dependent on intact *BRCA1* [13]. Hence, deficiency in *BRCA1* protein would result in increased fragility of FRA1E, leading to genomic rearrangements in *DPYD*. As germline variants in the *DPYD* gene can predict susceptibility to 5-FU-related toxicity, somatic CNVs in *DPYD* could affect the detection of these germline variants in tumour genomic sequencing.

5.6 The use of VAF thresholds is feasible for distinguishing between germline and somatic alterations in tumour-only analyses

Although sequencing of tumour-normal pairs would enable accurate identification of germline and somatic variants, this approach is not routinely practiced in clinical genomics due to inadequate funding and facilities to store additional specimens. Methods to distinguish between germline and somatic alterations in tumour-only analyses have been described by different groups [80, 93, 102]. Hiltemann et al. [93] used a virtual normal that was assembled by aggregating whole-genome-sequenced normal samples from 931 healthy and unrelated individuals, whereas Jones et al. [102] resorted to using an unmatched normal sample and public databases such as dbSNP, 1000 Genomes Project, and COSMIC, as well as effect prediction tools. We leveraged the fact that the VAFs of somatic mutations typically deviate from 50% and 100% for heterozygous and homozygous variants, respectively, and employed VAF thresholds to differentiate between variant statuses. Our approach managed to achieve high sensitivity and precision, therefore verifying the feasibility of using VAF threshold to differentiate between germline and somatic alterations in the absence of matched normal samples.

The VAF threshold method takes advantage of genetic impurity and heterogeneity of tumours, which render the deviation of somatic VAFs from diploid zygosity. Jones et al. [102] discovered that performance of the VAF threshold approach was highly dependent on tumour purity. While

the use of VAF can correctly identify all germline and somatic alterations in tumours with $< 50\%$ purity, this accuracy was not observed for specimens with higher tumour content. In fact, only 12.5% of cancer-predisposing germline variants and an average of 48% of somatic mutations were accurately predicted using the VAF method in tumours with more than 50% tumour content [102]. Unfortunately, pathologic estimation of tumour content was not available for our analyses. However, we speculate that the tumour specimens in our data set are highly impure or heterogeneous, thereby contributing to the high sensitivity and precision attained by the VAF threshold approach. While there are bioinformatic algorithms available to infer clonality and impurity estimates of tumours, many of these methods require matched normal controls or are not compatible with targeted sequencing data [223]. Nevertheless, this information should be integrated into clinical pipelines to enhance the performance of using a VAF threshold approach to distinguish between germline and somatic alterations in tumour-only genomic analyses.

5.7 Limitations and future directions

There are several limitations in our study. First, we did not manually review every single variant called by our pipeline. Only variants located within primer regions were manually inspected, while our variant filter also included common artifacts that were curated during clinical assessment. Hence, it is highly possible that sequence artifacts are present in our data set, particularly low-allelic-fraction variants (i.e. $< 10\%$) detected in the blood. Variant inspection using a genome browser is routinely conducted by genomic analysts in clinical practice to decrease the risk of reporting false positive results [80, 197]. However, manual review of variants was not implemented in our study because our analyses were focused on evaluating analytical validity instead of producing genomic information for clinical decision-making. Moreover, the large number of variants in our study would be time-consuming and unfeasible for manual inspection. Our evaluation of the VAF threshold approach in differentiating between germline and somatic variants is favourable of the framework to implement initial screening for germline variants in clinical tumour sequencing before follow-up germline testing.

The small gene panel and patient cohort size are caveats in our study. Although we were able to identify germline variants that can influence drug response, we did not report any pathogenic germline variants that were associated with cancer predisposition in our data set. Hence, we can only speculate that our approach could be applicable to variants in cancer-predisposing genes. Studies that were able to identify pathogenic germline variants were performed with cohort sizes and gene panels that were substantially larger than this study. For instance, the study by Schrader et al. [184], which revealed pathogenic germline variants in 16% of patients, was performed in a cohort of 1566 patients and 341 genes were screened. To determine whether the VAF threshold method can be applied to detect genetic alterations linked to cancer susceptibility, further assessment involving a larger patient cohort and surveying known cancer-predisposing genes should be carried out.

The present study addresses two problems faced by using tumour genomic sequencing to identify germline alterations: the widespread use of FFPE tumours and the challenge in differentiating between germline and somatic variants in tumour genomes. Archival FFPE tissues remain a sizable resource for cancer genomic studies and clinical genomic sequencing. Thus, there is a need to understand the extent of the different forms of DNA damage induced by formalin. Our analyses not only provided insights on the impact of formalin-induced DNA damage on amplicon-based NGS data, but also helped us devised guidelines to minimize these effects. Formalin fixation followed by paraffin embedding is an attractive method to preserve tissue morphology for histologic assessment because it allows storage at ambient temperature, which reduces cost incurred by maintaining freezers required for fresh-frozen samples. Yet, many studies, including ours, have indicated the side effects of formaldehyde exposure on nucleic acid [63, 106, 153, 154, 189, 220, 221]. Instead of investing efforts into mitigating these side effects, a potential solution is to transition from the use of formalin to the UMFIX (Sakura Finetek USA, Inc.) fixative, which is capable of preserving both cellular morphology for pathologic review and macromolecules, including DNA [211].

Most clinical laboratories conduct tumour-only sequencing and apply approaches to distinguish between germline and somatic alterations. Without matched normal samples, interpretation of variants becomes complicated. Jones et al. [102] and Garofalo et al. [80] concluded that sequencing of tumour-normal pairs is the best practice to accurately identify variant statuses. For a center to provide tumour-normal paired sequencing, it must be equipped to collect, analyze, and report germline findings. This includes establishing appropriate pre-test and post-test counseling, protocols to secure patient consent and manage variant of uncertain significance, and frameworks to communicate results that may implicate the patients' relatives. While various groups recommend the sequencing of tumour-normal pairs, some centers simply do not have the funding or infrastructure to implement this as a standard practice. Furthermore, the American College of Medical Genetics and Genomics (ACMG) recommended that clinical laboratories report incidental variants in 56 genes that are associated with disease risk in DNA derived from germline samples, including matched normal samples that only serve the purpose of subtracting germline variants to identify somatic mutations in tumours [86]. Interrogation of these genes suggested by the ACMG guidelines could result in detection of more variants with uncertain significance, which might pose more harm than good to patients. Additionally, cases in which only FFPE tumour blocks exist for a deceased patient would greatly benefit from approaches in differentiating between germline and somatic variants. For example, if the deceased individual is suspected to be a carrier of an inheritable disease, the ability to accurately identify the germline risk allele could prompt germline testing for the individual's relatives and facilitate preventive care. Thus, establishing approaches to tell apart germline and somatic variants in tumour genomic analyses still has its advantages from a clinical and financial perspective.

To summarize, we showed that the common forms of formalin-induced DNA damage in tumour samples were DNA fragmentation and cytosine deamination. Because these effects were

either minor or statistically insignificant compared to DNA extracted from blood, this justified the feasibility of using FFPE DNA for germline testing. Characterization of formalin-induced DNA damage could also assist in establishing recommendations to enhance amplicon enrichment and sequencing results. We also reported a high retention rate of germline alterations in the tumour genomes, suggesting the reliability of using tumour DNA for germline variant calling. Finally, we showed that application of VAF thresholds can achieve high sensitivity and precision in distinguishing germline alterations from somatic mutations in tumour-only analyses. This supports the framework of leveraging clinical tumour sequencing for identification of germline alterations. Subsequently, only patients with potential germline variants are referred to follow-up testing. A framework in which selected patients are referred to downstream confirmatory testing represents a time- and cost-effective approach to deliver germline testing.

Bibliography

[1] Nature Milestones in Cancer. *Nature Reviews Cancer*, 6:S8, 2006. → pages 1

[2] The Human Genome Project Completion: Frequently Asked Questions, 2010. URL https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/. Accessed September 12, 2017. → pages 3

[3] I. A. Adzhubei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010. → pages 14

[4] S. Afzal, M. Gusella, B. Vainer, et al. 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. → pages 85

[5] F. Ali-osman, O. Akande, G. Antoun, et al. Molecular Cloning , Characterization , and Expression in Escherichia coli of Full-length cDNAs of Three Human Glutathione S -Transferase Pi Gene Variants. *The Journal of Biological Chemistry*, 272(15):10004–10012, 1997. → pages 82

[6] A. Ally, M. Balasundaram, R. Carlsen, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, 169(7):1327–1341.e23, 2017. → pages 7

[7] U. Amstutz, S. Farese, S. Aebi, et al. Dihydropyrimidine dehydrogenase gene variation and severe 5-fluorouracil toxicity: a haplotype assessment. *Pharmacogenomics*, 10(6):931–944, 2009. → pages 79, 80, 81

[8] Y. Ando, H. Saka, M. Ando, et al. Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: A pharmacogenetic analysis. *Cancer Research*, 60(24):6921–6926, 2000. → pages 85

[9] S. Andrews. FastQC. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed October 1, 2017. → pages 11

[10] I. L. Andrulis, S. B. Bull, M. E. Blackstein, et al. Neu/erbB-2 amplification identifies a poor-prognosis group of women with node-negative breast cancer. Toronto Breast Cancer Study Group. *Journal of Clinical Oncology*, 16(4):1340–9, 1998. → pages 1

[11] N. Ånensen, J. Skavland, C. Stapnes, et al. 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. → pages 74

[12] S. L. Arcand, C. M. Maugard, P. Ghadirian, et al. Germline TP53 mutations in BRCA1 and BRCA2 mutation-negative French Canadian breast cancer families. *Breast Cancer Research and Treatment*, 108(3):399–408, 2008. → pages 74

[13] M. F. Arlt, B. Xu, S. G. Durkin, et al. BRCA1 Is Required for Common-Fragile-Site Stability via Its G 2 / M Checkpoint Function BRCA1 Is Required for Common-Fragile-Site Stability via Its G 2 / M Checkpoint Function. *Molecular and Cellular Biology*, 24(15): 6701–6709, 2004. → pages 89, 103

[14] S. G. Baker and J. Kaprio. Common susceptibility genes for cancer: search for the end of the rainbow. *BMJ (Clinical research ed.)*, 332(7550):1150–2, 2006. → pages 15

[15] R. Bao, L. Huang, J. Andrade, et al. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Libertas Academica*, 13:67–82, 2014. → pages 11, 12, 13, 14, 15

[16] C. Baynes, C. S. Healey, K. a. Pooley, et al. Common variants in the ATM, BRCA1, BRCA2, CHEK2 and TP53 cancer susceptibility genes are unlikely to increase breast cancer risk. *Breast cancer research: BCR*, 9(2):R27, 2007. → pages 15

[17] C. L. Bennett and E. A. Calhoun. Evaluating the total costs of chemotherapy-induced febrile neutropenia: results from a pilot study with community oncology cancer patients. *The Oncologist*, 12(4):478–83, 2007. → pages 17

[18] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. → pages 3, 5

[19] J. Betge, G. Kerr, T. Miersch, et al. Amplicon Sequencing of Colorectal Cancer: Variant Calling in Frozen and Formalin-Fixed Samples. *Plos One*, 10(5):e0127146, 2015. → pages 100

[20] F. Biramijamal, A. Allameh, P. Mirbod, et al. Unusual profile and high prevalence of p53 mutations in esophageal squamous cell carcinomas from northern Iran. *Cancer Research*, 61(7):3119–23, 2001. → pages 74

[21] G. D. V. Blanco, O. A. Paoluzi, P. Sileri, et al. Familial colorectal cancer screening: When and what to do? *World Journal of Gastroenterology*, 21(26):7944–7953, 2015. → pages 17

[22] K. Bodi, A. G. Perera, P. S. Adams, et al. Comparison of commercially available target enrichment methods for next-generation sequencing. *Journal of Biomolecular Techniques*, 24(2):73–86, 2013. → pages 8

[23] V. Boeva, T. Popova, M. Lienard, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*, 30(24): 3443–3450, 2014. → pages 8

[24] V. Boige, M. Vincent, P. Alexandre, et al. DPYD Genotyping to PredictAdverse Events Following Treatment With Fluorouracil-BasedAdjuvant Chemotherapy in Patients With Stage III Colon Cancer. *JAMA Oncology*, 2(5):655–662, 2016. → pages 79, 80

[25] 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. → pages 61, 75

[26] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. → pages 11

[27] Y. Bombard, M. Robson, and K. Offit. Revealing the Incidentalome When Targeting the Tumor Genome. *The Journal of the American Medical Association*, 310(8):795–6, 2014. → pages 20, 60, 95

[28] M. Bonafé, S. Salvioli, C. Barbi, et al. 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. → pages 61, 75

[29] S. Bonin, M. Donada, G. Bussolati, et al. A synonymous EGFR polymorphism predicting responsiveness to anti-EGFR therapy in metastatic colorectal cancer patients. *Tumor Biology*, 37(6):7295–7303, 2016. → pages 67

[30] I. E. Bosdet, T. R. Docking, Y. S. Butterfield, et al. 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. → pages 2, 26

[31] G. Bougeard, S. Baert-Desurmont, I. Tournier, et al. 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. → pages 61, 75

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

[33] W. Burke. Genetic tests: clinical validity and clinical utility. *Current Protocols in Human Genetics*, 81(9):1–8, 2014. → pages 15, 16

[34] L. Cai, W. Yuan, Z. Zhang, et al. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, 6 (November):36540, 2016. → pages 101

[35] E. Capriotti, R. Calabrese, and R. Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–2734, 2006. → pages 15

[36] M. O. Carneiro, C. Russ, M. G. Ross, et al. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13(1):375, 2012. → pages 3

[37] D. Caronia, M. Martin, J. Sastre, et al. A polymorphism in the cytidine deaminase promoter predicts severe capecitabine-induced hand-foot syndrome. *Clinical Cancer Research*, 17(7): 2006–2013, 2011. → pages 83

[38] D. M. Carrick, M. G. Mehaffey, M. C. Sachs, et al. Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue. *PLoS ONE*, 10(7):3–10, 2015. → pages 22, 56, 102

[39] J. Carrot-Zhang and J. Majewski. LoLoPicker: Detecting Low Allelic-Fraction Variants in Low-Quality Cancer Samples from Whole-exome Sequencing Data. *bioRxiv*, 8(23):043612, 2016. → pages 101

[40] K. E. Caudle, C. F. Thorn, T. E. Klein, et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing. *Clinical Pharmacology & Therapeutics*, 94(6):640–645, 2013. → pages 79, 80, 81

[41] F. Chang and M. M. Li. Clinical application of amplicon-based next-generation sequencing in cancer. *Cancer Genetics*, 206(12):413–419, 2013. → pages 8

[42] P. B. Chapman, A. Hauschild, C. Robert, et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine*, 364(26): 2507–2516, 2011. → pages 1

[43] S. Cheli, F. Pietrantonio, E. Clementi, et al. LightSNiP assay is a good strategy for pharmacogenetics test. *Frontiers in Pharmacology*, 6(JUN):1–5, 2015. → pages 85

[44] G. Chen, S. Mosier, C. D. Gocke, et al. Cytosine Deamination is a Major Cause of Baseline Noise in Next Generation Sequencing. *Molecular Diagnosis & Therapy*, 18(5):587–593, 2014. → pages 100

[45] Y. C. Chen, C. H. Tzeng, P. M. Chen, et al. 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. → pages 82

[46] H. Cheng, B. Ma, R. Jiang, et al. 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. → pages 61, 75

[47] 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. → pages 74

[48] J. Chung, D.-S. Son, H.-J. Jeon, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Scientific Reports*, 6(1):26732, 2016. → pages 2

[49] K. Cibulskis, M. S. Lawrence, S. L. Carter, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3): 213–219, 2013. → pages 13

[50] V. Cohen, V. Panet-raymond, N. Sabbaghian, et al. Methylenetetrahydrofolate Reductase Polymorphism in Advanced Colorectal Cancer : A Novel Genomic Predictor of Clinical

Response to Fluoropyrimidine-based Chemotherapy Advances in Brief Methylenetetrahydrofolate Reductase Polymorphism in Advanced Colorecta. *Clinical Cancer Research*, 9(May):1611–1615, 2003. → pages 82

[51] M. Costello, T. J. Pugh, T. J. Fennell, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):1–12, 2013. → pages 101

[52] H. Cui. Methods of Gene Enrichment and Massively Parallel Sequencing Technologies. In L.-J. C. Wong, editor, *Next Generation Sequencing: Translation to Clinical Diagnostics*, chapter 3, pages 39–58. Springer Science+Business Media, New York, 2013. ISBN 9781461470014. → pages 8

[53] J. Culver, C. Brinkerhoff, J. Clague, et al. Variants of uncertain significance in BRCA testing: Evaluation of surgical decisions, risk perception, and cancer distress. *Clinical Genetics*, 84(5):464–472, 2013. → pages 7

[54] P. Danecek, A. Auton, G. Abecasis, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011. → pages 13, 14

[55] M. David, M. Dzamba, D. Lister, et al. SHRiMP2: Sensitive yet practical short read mapping. *Bioinformatics*, 27(7):1011–1012, 2011. → pages 12

[56] 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. → pages 85

[57] K. J. Dedes, P. M. Wilkerson, D. Wetterskog, et al. Synthetic lethality of PARP inhibition in cancers lacking BRCA1 and BRCA2 mutations. *Cell Cycle*, 10(8):1192–1199, 2011. → pages 17

[58] M. J. Deenen, J. Tol, A. M. Burylo, et al. 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. → pages 17, 79, 80, 81

[59] J. T. den Dunnen, R. Dalgleish, D. R. Maglott, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, 37(6):564–569, 2016. → pages 14

[60] M. Depner, S. Fuchs, J. Raabe, et al. 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. → pages 72

[61] A. Didelot, S. K. Kotsopoulos, A. Lupo, et al. Multiplex picoliter-droplet digital PCR for quantitative assessment of DNA integrity in clinical samples. *Clinical Chemistry*, 59(5): 815–823, 2013. → pages 22, 35, 56, 100

[62] 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. → pages 48, 100, 101

[63] 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. → pages 22, 35, 48, 99, 100, 101, 105

[64] H. Do, S. Q. Wong, J. Li, et al. 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. → pages 48, 100, 101

[65] D. Dobritzsch, G. Schneider, K. D. Schnackerz, et al. 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. → pages 79

[66] L. Dong, W. Wang, A. Li, et al. Clinical Next Generation Sequencing for Precision Medicine in Cancer. *Current Genomics*, 16(4):253–63, 2015. → pages 10, 22

[67] E. Dotor, M. Cuatrecases, M. Martínez-Iniesta, et al. 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. → pages 85

[68] J. A. Drebin, V. C. Link, D. F. Stern, et al. Down-modulation of an oncogene protein product and reversion of the transformed phenotype by monoclonal antibodies. *Cell*, 41(3): 695–706, 1985. → pages 1

[69] B. J. Druker, F. Guilhot, S. G. O'Brien, et al. Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *New England Journal of Medicine*, 355: 2408–2417, 2006. → pages 1

[70] J. Eid, A. Fehr, J. Gray, et al. Real-Time DNA Sequencing from. *Science, New Series*, 323 (5910):133–138, 2009. → pages 3

[71] A. Eijkelenboom, E. J. Kamping, A. W. Kastner-van Raaij, et al. Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *The Journal of Molecular Diagnostics*, 18(6):851–863, 2016. → pages 102

[72] A. C. English, S. Richards, Y. Han, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE*, 7(11):1–12, 2012. → pages 3

[73] P. Estevez-Garcia, A. Castaño, A. C. Martin, et al. PDGFRα/β and VEGFR2 polymorphisms in colorectal cancer: incidence and implications in clinical outcome. *BMC Cancer*, 12:514, 2012. → pages 72

[74] M. C. Etienne, J. L. Formento, M. Chazal, et al. Methylenetetrahydrofolate reductase gene polymorphisms and response to fluorouracil-based treatment in advanced colorectal cancer patients. *Pharmacogenetics*, 14(12):785–792, 2004. → pages 82

[75] M. C. Etienne-Grimaldi, G. Milano, F. Maindrault-Gœbel, et al. Methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms and FOLFOX response in colorectal cancer patients. *British Journal of Clinical Pharmacology*, 69(1):58–66, 2010. → pages 82

[76] G. Fortunato, G. Calcagno, V. Bresciamorra, et al. Multiple_sclerosis_and_hepatit.PDF. *Journal of Interferon & Cytokine Research*, 28:141–152, 2008. → pages 73

[77] G. M. G. Frampton, A. Fichtenholtz, G. a. G. Otto, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature*, 31(October):1–11, 2013. → pages 60

[78] D. Fumagalli, P. G. Gavin, Y. Taniyama, et al. A rapid, sensitive, reproducible and cost-effective method for mutation profiling of colon cancer and metastatic lymph nodes. *BMC Cancer*, 10:101, 2010. → pages 60

[79] J. Gagan and E. M. Van Allen. Next-generation sequencing to guide cancer therapy. *Genome Medicine*, 7(1):80, 2015. → pages 8

[80] A. Garofalo, L. Sholl, B. Reardon, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Medicine*, 8(1):79, 2016. → pages 103, 104, 105

[81] G. Gentile, A. Botticelli, L. Lionetto, et al. Genotypephenotype correlations in 5-fluorouracil metabolism: a candidate DPYD haplotype to improve toxicity prediction. *The Pharmacogenomics Journal*, 16(4):320–325, 2016. → pages 79, 80, 81

[82] B. Glimelius, H. Garmo, A. Berglund, et al. Prediction of irinotecan and 5-fluorouracil toxicity and response in patients with advanced colorectal cancer. *The Pharmacogenomics Journal*, 11(1):61–71, 2011. → pages 85

[83] A. González-Pérez and N. López-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*, 88(4):440–449, 2011. → pages 15

[84] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. → pages 4, 5

[85] F. Graziano, A. Ruzzo, F. Loupakis, et al. 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. → pages 85

[86] R. C. Green, J. S. Berg, W. W. Grody, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, 15(7): 565–74, 2013. → pages 95, 105

[87] E. Gross, C. Meul, S. Raab, et al. Somatic copy number changes in DPYD are associated with lower risk of recurrence in triple-negative breast cancers. *British Journal of Cancer*, 109(9):2347–2355, 2013. → pages 89, 103

[88] V. Guillem, J. C. Hernandez-Boluda, D. Gallardo, et al. 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. → pages 83

[89] J. Guo, N. Xu, Z. Li, et al. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences*, 105(27):9145–9150, 2008. → pages 3, 5

[90] M. Gusella, G. Crepaldi, C. Barile, et al. 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. → pages 82, 85

[91] T. Helleday. The underlying mechanism for the PARP and BRCA synthetic lethality: Clearing up the misunderstandings. *Molecular Oncology*, 5(4):387–393, 2011. → pages 17

[92] P. Heyn, U. Stenzel, A. W. Briggs, et al. Road blocks on paleogenomes–polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Research*, 38(16):e161, 2010. → pages 22, 101

[93] S. Hiltemann, G. Jenster, J. Trapman, et al. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Research*, 25(9):1382–1390, 2015. → pages 103

[94] M. Hofreiter, V. Jaenicke, D. Serre, et al. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, 29(23):4793–4799, 2001. → pages 48

[95] J. Hong, S. W. Han, H. S. Ham, et al. 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. → pages 17, 82

[96] L. Huang, F. Chen, Y. Chen, et al. Thymidine phosphorylase gene variant, platelet counts and survival in gastrointestinal cancer patients treated by fluoropyrimidines. *Scientific Reports*, 4(1):5697, 2014. → pages 83

[97] Z.-H. Huang, D. Hua, L.-H. Li, et al. 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. → pages 61, 75

[98] D. M. Hyman, D. B. Solit, M. E. Arcila, et al. Precision medicine at Memorial Sloan Kettering Cancer Center: Clinical next-generation sequencing enabling next-generation targeted therapy trials. *Drug Discovery Today*, 20(12):1422–1428, 2015. → pages 2

[99] F. Innocenti, S. D. Undevia, L. Iyer, et al. 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. → pages 85

[100] A. Jakobsen, J. N. Nielsen, N. Gyldenkerne, et al. Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphism in normal tissue as predictors of fluorouracil sensitivity. *Journal of Clinical Oncology*, 23(7):1365–1369, 2005. → pages 82

[101] B. A. Jennings, Y. K. Loke, J. Skinner, et al. Evaluating Predictive Pharmacogenetic Signatures of Adverse Events in Colorectal Cancer Patients Treated with Fluoropyrimidines. *PLoS ONE*, 8(10):1–9, 2013. → pages 83

[102] S. Jones, V. Anagnostou, K. Lytle, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Science Translational Medicine*, 7(283ra53), 2015. → pages 20, 60, 89, 99, 102, 103, 104, 105

[103] A. V. Khrunin, A. Moisseev, V. Gorbunova, et al. Genetic polymorphisms and the efficacy and toxicity of cisplatin-based chemotherapy in ovarian cancer patients. *The Pharmacogenomics Journal*, 10(1):54–61, 2010. → pages 61, 75

[104] J. Kim, D. Kim, J. S. Lim, et al. Accurate detection of low-level somatic mutations with technical replication for next-generation sequencing. *bioRxiv*, 2017. → pages 101

[105] J. G. Kim, S. K. Sohn, Y. S. Chae, et al. 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. → pages 61, 75

[106] S. Kim, C. Park, Y. Ji, et al. Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine. *Journal of Molecular Diagnostics*, 19(1): 137–146, 2017. → pages 35, 48, 99, 100, 101, 105

[107] D. C. Koboldt, Q. Zhang, D. E. Larson, et al. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012. → pages 13

[108] D. C. Koboldt, D. E. Larson, and R. K. Wilson. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Current Protocols in Bioinformatics*, 44: 15.4.1–15.4.17, 2013. → pages 13

[109] Y. Kodama, M. Shumway, and R. Leinonen. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):2011–2013, 2012. → pages 5

[110] S. Kumar, R. Marfatia, S. Tannenbaum, et al. Doxorubicin-induced cardiomyopathy 17 years after chemotherapy. *Texas Heart Institute Journal*, 39(3):424–427, 2012. → pages 17

[111] D. M. Kweekel, H. Gelderblom, T. Van der Straaten, et al. 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. → pages 85

[112] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2013. → pages 12

[113] J. Laskin, S. Jones, S. Aparicio, et al. Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Molecular Case Studies*, 1(1):a000570, 2015. → pages 2, 10

[114] C. Ledergerber and C. Dessimoz. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12(5):489–497, 2011. → pages 11

[115] A. M. Lee, Q. Shi, E. Pavey, et al. 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. → pages 17, 79, 80

[116] J. Leichsenring, A.-L. Volckmar, N. Magios, et al. Synonymous EGFR Variant p.Q787Q is Neither Prognostic Nor Predictive in Patients with Lung Adenocarcinoma Jonas. *Genes, Chromosomes & Cancer*, 56(3):214–220, 2017. → pages 67

[117] F. Leone, G. Cavalloni, Y. Pignochino, et al. Somatic mutations of epidermal growth factor receptor in bile duct and gallbladder carcinoma. *Clinical Cancer Research*, 12(6): 1680–1685, 2006. → pages 67

[118] S. E. Levy and R. M. Myers. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 17(1):95–115, 2016. → pages 4, 5

[119] B. Li, V. G. Krishnan, M. E. Mort, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–2750, 2009. → pages 15

[120] H. Li. Exploring single-sample snp and indel calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012. → pages 12

[121] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. → pages

[122] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010. → pages 12

[123] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010. → pages 12

[124] H. Li, B. Handsaker, A. Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. → pages 12

[125] M.-T. Lin, S. L. Mosier, M. Thiess, et al. Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. *American Journal of Clinical Pathology*, 141(6):856–66, 2014. → pages 48, 60

[126] Y. Liu, S. N. Xu, Y. S. Chen, et al. 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. → pages 70

[127] J. J. Lou, L. Mirsadraei, D. E. Sanchez, et al. A review of room temperature storage of biospecimen tissue and nucleic acids for anatomic pathology laboratories and biorepositories. *Clinical Biochemistry*, 47:267–273, 2014. → pages 35

[128] N. Ludyga, B. Grünwald, O. Azimzadeh, et al. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. *Virchows Archiv*, 460(2):131–140, 2012. → pages 22, 102

[129] G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011. → pages 12

[130] L. Mamanova, A. J. Coffey, C. E. Scott, et al. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2):111–118, 2010. → pages 8

[131] M. V. Mandola, J. Stoehlmacher, W. Zhang, et al. 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. → pages 85

[132] Marcel and Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011. → pages 11

[133] E. Marcuello, A. Altés, A. Menoyo, et al. UGT1A1 gene variations and irinotecan treatment in patients with metastatic colorectal cancer. *British Journal of Cancer*, 91(4):678–682, 2004. → pages 85

[134] E. Marcuello, A. Altés, A. Menoyo, et al. Methylenetetrahydrofolate reductase gene polymorphisms: Genomic predictors of clinical response to fluoropyrimidine-based chemotherapy? *Cancer Chemotherapy and Pharmacology*, 57(6):835–840, 2006. → pages 82

[135] E. R. Mardis. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303, 2013. → pages 5

[136] E. R. Mardis. DNA sequencing technologies: 20062016. *Nature Protocols*, 12(2):213–218, 2017. → pages 5

[137] 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. → pages 79

[138] A. McKenna, M. Hanna, E. Banks, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20: 1297–1303, 2010. → pages 13

[139] H. L. McLeod, D. J. Sargent, S. Marsh, et al. 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. → pages 82, 85

[140] S. R. McWhinney and H. L. McLeod. Using germline genotype in cancer pharmacogenetic studies. *Pharmacogenomics*, 10(3):489–493, 2009. → pages 20, 60, 102

[141] F. Meric-Bernstam, L. Brusco, M. Daniels, et al. Incidental germline variants in 1000 advanced cancers on a prospective somatic genomic profiling protocol. *Annals of Oncology*, 27(5):795–800, 2016. → pages 20, 60, 89, 99, 102

[142] D. Meulendijks, L. M. Henricks, G. S. Sonke, et al. 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. → pages 79, 80, 81

[143] M. Mielczarek and J. Szyda. Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics*, 57(1):71–79, 2016. → pages 11, 12, 13

[144] B. Mohelnikova-Duchonova, B. Melichar, and P. Soucek. FOLFOX/FOLFIRI pharmacogenetics: The call for a personalized approach in colorectal cancer therapy. *World Journal of Gastroenterology*, 20(30):10316–10330, 2014. → pages 17, 18, 19, 20, 62

[145] S. Moorthie, A. Hall, and C. F. Wright. Informatics and clinical genome sequencing: opening the black box. *Genetics in Medicine*, 15(3):165–171, 2013. → pages 14, 15

[146] A. Morel, M. Boisdron-Celle, L. Fey, et al. Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide polymorphisms on 5-fluorouracil tolerance. *Molecular Cancer Therapeutics*, 5(11):2895–2904, 2006. → pages 79, 80, 81

[147] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. → pages 15

[148] Q. Nie, S. Shrestha, E. E. Tapper, et al. Quantitative contribution of rs75017182 to dihydropyrimidine dehydrogenase mRNA splicing and enzyme activity. *Clinical Pharmacology & Therapeutics*, 00(00):1–9, 2017. → pages 81

[149] R. Nielsen, J. S. Paul, A. Albrechtsen, et al. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–51, jun 2011. → pages 11, 12, 13

[150] S. Nik-Zainal, H. Davies, J. Staaf, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016. → pages 7

[151] I. Nishino, A. Spinazzola, A. Papadimitriou, et al. Mitochondrial neurogastrointestinal encephalomyopathy: An autosomal recessive disorder due to thymidine phosphorylase mutations. *Annals of Neurology*, 47(6):792–800, 2000. → pages 84

[152] S. M. Offer, C. C. Fossum, N. J. Wegner, et al. Comparative functional analysis of dpyd variants of potential clinical relevance to dihydropyrimidine dehydrogenase activity. *Cancer Research*, 74(9):2545–2554, 2014. → pages 79, 80, 81

[153] R. Ofner, C. Ritter, S. Ugurel, et al. Non-reproducible sequence artifacts in FFPE tissue : an experience report. *Journal of Cancer Research and Clinical Oncology*, 143(7):1199–1207, 2017. → pages 35, 48, 99, 105

[154] E. Oh, Y.-L. Choi, M. J. Kwon, et al. Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PloS One*, 10(12):e0144162, 2015. → pages 35, 48, 99, 100, 105

[155] G. R. Oliver, S. N. Hart, and E. W. Klee. Bioinformatics for clinical next generation sequencing. *Clinical Chemistry*, 61(1):124–135, 2015. → pages 12, 14, 15

[156] S. Pabinger, A. Dander, M. Fischer, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278, 2014. → pages 2, 12, 13, 15

[157] B. J. Paessens, C. von Schilling, K. Berger, et al. Health resource consumption and costs attributable to chemotherapy-induced toxicity in german routine hospital care in lymphoproliferative disorder and NSCLC patients. *Annals of Oncology*, 22(10):2310–2319, 2011. → pages 17

[158] M. Panczyk. Pharmacogenetics research on chemotherapy resistance in colorectal cancer over the last 20 years. *World Journal of Gastroenterology*, 20(29):9775–9827, 2014. → pages 17, 18, 19, 20, 62

[159] Q. Peng, R. Vijaya Satya, M. Lewis, et al. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics*, 16(1):589, 2015. → pages 102

[160] K. P. Pennington, T. Walsh, M. Lee, et al. BRCA1, TP53, and CHEK2 germline mutations in uterine serous carcinoma. *Cancer*, 119(2):332–338, 2013. → pages 74

[161] D. A. Pilger, P. L. Da Costa Lopez, F. Segal, et al. 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. → pages 74

[162] S. Quesnel, S. Verselis, C. Portwine, et al. p53 compound heterozygosity in a severely affected child with Li-Fraumeni syndrome. *Oncogene*, 18(27):3970–3978, 1999. → pages 74

[163] B. Quintáns, A. Ordóñez-Ugalde, P. Cacheiro, et al. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Applied and Translational Genomics*, 3(3):60–67, 2014. → pages 14

[164] B. Rabbani, M. Tekin, and N. Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, 59(1):5–15, 2014. → pages 10

[165] N. Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483): 302–308, 2014. → pages 17, 20

[166] B. J. Raphael, R. H. Hruban, A. J. Aguirre, et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13, 2017. → pages 7

[167] N. Rashid, H. A. Koh, H. C. Baca, et al. Economic burden related to chemotherapy-related adverse events in patients with metastatic breast cancer in an integrated health care system. *Breast Cancer: Targets and Therapy*, 8:173–181, 2016. → pages 17

[168] M. C. Ravnan and M. S. Matalka. Vemurafenib in Patients With BRAF V600E Mutation-Positive Advanced Melanoma. *Clinical Therapeutics*, 34(7):1474–1486, 2012. → pages 1

[169] V. M. Raymond, S. W. Gray, S. Roychowdhury, et al. Germline findings in tumor-only sequencing: Points to consider for clinicians and laboratories. *Journal of the National Cancer Institute*, 108(4):1–5, 2016. → pages 89, 95

[170] H. L. Rehm. Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 14(4):295–300, 2013. → pages 10

[171] S. Richards, N. Aziz, S. Bale, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423, 2015. → pages 7

[172] A. G. Robertson, J. Shih, C. Yau, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell*, 32(2):204–220.e15, 2017. → pages 7

[173] J. M. Rothberg, W. Hinz, T. M. Rearick, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011. → pages 3

[174] E. Rouits, V. Charasson, A. Pétain, et al. Pharmacokinetic and pharmacogenetic determinants of the activity and toxicity of irinotecan in metastatic colorectal cancer patients. *British Journal of Cancer*, 99:1239–45, 2008. → pages 85

[175] J. D. Rowley. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, 243(5405): 290–293, 1973. → pages 1

[176] A. Ruzzo, F. Graziano, F. Loupakis, et al. Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFOX-4 chemotherapy. *Journal of Clinical Oncology*, 25(10):1247–1254, 2007. → pages 82

[177] A. Ruzzo, F. Graziano, F. Loupakis, et al. Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFIRI chemotherapy. *The Pharmacogenomics Journal*, 8(4):278–288, 2008. → pages 85

[178] L. P. Rybak, D. Mukherjea, S. Jajoo, et al. Cisplatin ototoxicity and protection: clinical and experimental studies. *The Tohoku Journal of Experimental Medicine*, 219(3):177–86, 2009. → pages 17

[179] V. N. Rykalina, A. A. Shadrin, V. S. Amstislavskiy, et al. Exome sequencing from nanogram amounts of starting DNA: Comparing three approaches. *PLoS ONE*, 9(7), 2014. → pages 2

[180] E. Samorodnitsky, B. M. Jewell, R. Hagopian, et al. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human Mutation*, 36(9): 903–914, 2015. → pages 102

[181] S. Sanderson, R. Zimmern, M. Kroese, et al. How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. *Genetics in Medicine*, 7(7):495–500, 2005. → pages 15, 16

[182] C. T. Saunders, W. S. W. Wong, S. Swamy, et al. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012. → pages 13

[183] C. L. Scherr, N. M. Lindor, T. L. Malo, et al. A preliminary investigation of genetic counselors' information needs when receiving a variant of uncertain significance result: a mixed methods study. *Genetics in Medicine*, 17(9):739–746, 2015. → pages 7

[184] K. A. Schrader, D. T. Cheng, V. Joseph, et al. Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. *JAMA Oncology*, 2(1):1–8, 2015. → pages 20, 60, 89, 99, 102, 104

[185] M. Schwab, U. M. Zanger, C. Marx, et al. 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. → pages 79, 80, 82

[186] A. D. Seidman, D. Berry, C. Cirrincione, et al. Randomized phase III trial of weekly compared with every-3-weeks paclitaxel for metastatic breast cancer, with trastuzumab for all HER-2 overexpressors and random assignment to trastuzumab or not in HER-2 nonoverexpressors: final results of Cancer and Leu. *Journal of Clinical Oncology*, 26(10): 1642–9, 2008. → pages 1

[187] B. S. Sheffield, B. Tessier-Cloutier, H. Li-Chang, et al. Personalized oncogenomics in the management of gastrointestinal carcinomas-early experiences from a pilot study. *Current Oncology*, 23(6):e571–e575, 2016. → pages 2

[188] S.-R. Shi, R. J. Cote, L. Wu, et al. DNA extraction from archival formalin-fixed, paraffin-embedded tissue sections based on the antigen retrieval principle: heating under the influence of pH. *Journal of Histochemistry and Cytochemistry*, 50(8):1005–1011, 2002. → pages 22, 100

[189] J. A. Sikorsky, D. A. Primerano, T. W. Fenger, et al. DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency. *Biochemical and Biophysical Research Communications*, 355(2):431–437, 2007. → pages 35, 99, 105

[190] R. Simon and S. Roychowdhury. Implementing personalized cancer genomics in clinical trials. *Nature Reviews Drug Discovery*, 12(5):358–369, 2013. → pages 8, 10

[191] S. Sjogren, M. Inganas, A. Lindgren, et al. Prognostic and predictive value of c-erbB-2 overexpression in primary breast cancer, alone and in combination with other prognostic markers. *Journal of Clinical Oncology*, 16:462–469, 1998. → pages 1

[192] E. H. Slager, M. W. Honders, E. D. V. D. Meijden, et al. 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. *Blood*, 107(12):4954–4960, 2006. → pages 83

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

[194] A. So, A. Vilborg, Y. Bouhlal, et al. A Robust Targeted Sequencing Approach for Low Input and Variable Quality DNA from Clinical Samples. *bioRxiv*, 2017. → pages 2

[195] D. H. Spencer, J. K. Sehn, H. J. Abel, et al. 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. → pages 100

[196] J. Stoehlmacher, D. J. Park, W. Zhang, et al. 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. → pages 82, 85

[197] S. P. Strom. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biology & Medicine*, 13(1):3–11, 2016. → pages 7, 15, 104

[198] K. W. Suh, J. H. Kim, D. Y. Kim, et al. 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. → pages 82

[199] M. A. Sukhai, K. J. Craddock, M. Thomas, et al. A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genetics in Medicine*, 18(2):128–136, 2016. → pages 7

[200] J. J. Swen, M. Nijenhuis, A. de Boer, et al. Pharmacogenetics: From Bench to Byte An Update of Guidelines. *Clinical Pharmacology & Therapeutics*, 89(5):662–673, 2011. → pages 79, 80, 81

[201] D. Tanaka, A. Hishida, K. Matsuo, et al. Polymorphism of dihydropyrimidine dehydrogenase (DPYD) Cys29Arg and risk of six malignancies in Japanese. *Nagoya Journal of Medical Science*, 67(Fig 1):117–124, 2005. → pages 81

[202] R. Tian, M. K. Basu, and E. Capriotti. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC genomics*, 16 Suppl 8(Suppl 8):S7, 2015. → pages 101

[203] F. Tirode, D. Surdez, X. Ma, et al. Genomic landscape of ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. *Cancer Discovery*, 4 (11):1342–1353, 2014. → pages 7

[204] G. Toffoli, E. Cecchin, G. Corona, et al. 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. → pages 85

[205] G. Toffoli, L. Giodini, A. Buonadonna, et al. Clinical validity of a DPYD-based pharmacogenetic test to predict severe toxicity to fluoropyrimidines. *International Journal of Cancer*, 137(12):2971–2980, 2015. → pages 79, 80

[206] F. Torri, I. D. Dinov, A. Zamanyan, et al. Next generation sequence analysis and computational genomics using graphical pipeline workflows. *Genes*, 3(3):545–575, 2012. → pages 2

[207] A. B. P. van Kuilenburg, J. Haasjes, D. J. Richel, et al. 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. → pages 79, 80, 81

[208] A. B. P. van Kuilenburg, J. Meijer, M. W. T. Tanck, et al. Phenotypic and clinical implications of variants in the dihydropyrimidine dehydrogenase gene. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1862(4):754–762, 2016. → pages 79, 80, 81

[209] F. J. Vega, P. Iniesta, T. Caldes, et al. 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. → pages 75

[210] A. Villani, U. Tabori, J. Schiffman, et al. 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. → pages 74

[211] V. Vincek, M. Nassiri, M. Nadji, et al. A Tissue Fixative that Protects Macromolecules (DNA, RNA, and Protein) and Histomorphology in Clinical Samples. *Laboratory Investigation*, 83(10):1427–1435, 2003. → pages 105

[212] C. L. Vogel, M. A. Cobleigh, D. Tripathy, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *Journal of Clinical Oncology*, 20(3):719–26, 2002. → pages 1

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

[214] H. Wang and F. Meng. Theoretical study of proton-catalyzed hydrolytic deamination mechanism of adenine. *Theoretical Chemistry Accounts*, 127(5):561–571, 2010. → pages 48, 101

[215] Y. Wang, W. Bao, H. Shi, et al. 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. → pages 67, 68

[216] Y. Wang, Q. Yang, and Z. Wang. The evolution of nanopore sequencing. *Frontiers in Genetics*, 5(DEC):1–20, 2014. → pages 3

[217] K. Wetterstrand. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP), 2016. URL https://www.genome.gov/27541954/dna-sequencing-costs-data/. Accessed September 12, 2017. → pages 2

[218] C. Wong, R. A. DiCioccio, H. J. Allen, et al. Mutations in BRCA1 from fixed, paraffin-embedded tissue can be artifacts of preservation. *Cancer Genetics and Cytogenetics*, 107(1):21–27, 1998. → pages 101

[219] N. A. Wong, D. Gonzalez, M. Salto-Tellez, et al. 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. → pages 60

[220] S. Q. Wong, J. Li, R. Salemi, et al. Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Scientific Reports*, 3(3):3494, 2013. → pages 22, 35, 99, 100, 102, 105

[221] S. Q. Wong, J. Li, A. Y-C Tan, et al. 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. → pages 8, 22, 35, 48, 56, 99, 100, 102, 105

[222] C. Xu, M. Nezami Ranjbar, Z. Wu, et al. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. *BMC Genomics*, 18(1):1–11, 2017. → pages 101

[223] V. K. Yadav and S. De. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in Bioinformatics*, 16(2):232–241, 2015. → pages 104

[224] M. Yang, Y. Guo, X. Zhang, et al. Interaction of P53 Arg72Pro and MDM2 T309G polymorphisms and their associations with risk of gastric cardia cancer. *Carcinogenesis*, 28 (9):1996–2001, 2007. → pages 61, 75

[225] S. S. Yea, S. S. Lee, W.-Y. Kim, et al. Genetic variations and haplotypes of UDP-glucuronosyltransferase 1A locus in a Korean population. *Therapeutic Drug Monitoring*, 30(1):23–34, 2008. → pages 85

[226] T. Yoneda, A. Kuboyama, K. Kato, et al. Association of MDM2 SNP309 and TP53 Arg72Pro polymorphisms with risk of endometrial cancer. *Oncology Reports*, 30(1):25–34, 2013. → pages 61, 75

[227] Y. Zha, P. Gan, Q. Liu, et al. 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. → pages 61, 75

[228] W. Zhang, L. P. Stabile, P. Keohavong, et al. Mutation and polymorphism in the EGFR-TK domain associated with lung cancer. *Journal of Thoracic Oncology*, 1(7):635–47, 2006. → pages 67

[229] X. Zhang, G. Ao, Y. Wang, et al. Genetic variants and haplotypes of the UGT1A9, 1A7 and 1A1 genes in Chinese Han. *Genetics and Molecular Biology*, 35(2):428–434, 2012. → pages 85

[230] Y. Zhang, L. Liu, Y. Tang, et al. 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. → pages 61, 75

[231] Z. Z. Zhu, A. Z. Wang, H. R. Jia, et al. Association of the TP53 codon 72 polymorphism with colorectal cancer in a Chinese population. *Japanese Journal of Clinical Oncology*, 37 (5):385–390, 2007. → pages 61, 75

[232] R. L. Zimmern and M. Kroese. The evaluation of genetic tests. *Journal of Public Health*, 29 (3):246–250, 2007. → pages 15, 16

[233] J. Zining, X. Lu, H. Caiyun, et al. Genetic polymorphisms of mTOR and cancer risk: a systematic review and updated meta-analysis. *Oncotarget*, 7(35), 2016. → pages 69, 71

Appendix A

Supporting Materials

Table A.1: Target regions and amplicons of the OncoPanel. Genomic regions are presented in the 0-based coordinate system.

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
AKT1	AKT1_a	chr14:105246545-105246554	chr14:105246426-105246549	123	56
			chr14:105246508-105246651	143	62
AKT1	AKT1_b	chr14:105246449-105246458	chr14:105246426-105246549	123	56
ALK	ALK_a	chr2:29445212-29445274	chr2:29445125-29445298	173	55
ALK	ALK_b	chr2:29443629-29443703	chr2:29443598-29443783	185	52
BRAF	BRAF_a	chr7:140481395-140481418	chr7:140481369-140481472	103	44
BRAF	BRAF_b	chr7:140453129-140453152	chr7:140453054-140453256	202	37
			chr7:140453084-140453218	134	38
DPYD	DPYD_a	chr1:97915607-97915621	chr1:97915561-97915715	154	39
			chr1:97915565-97915675	110	39
DPYD	DPYD_b	chr1:98348878-98348892	chr1:98348806-98349041	235	33
			chr1:98348854-98349048	194	34
DPYD	DPYD_c	chr1:97981336-97981350	chr1:97981294-97981367	73	47
			chr1:97981302-97981501	199	46
			chr1:97981339-97981478	139	44
DPYD	DPYD_d	chr1:97547940-97547954	chr1:97547867-97547976	109	41
			chr1:97547890-97547980	90	40
DPYD	DPYD_e	chr1:97981417-97981425	chr1:97981302-97981501	199	46
			chr1:97981339-97981478	139	44
DPYD	DPYD_f	chr1:98039415-98039423	chr1:98039299-98039458	159	45
			chr1:98039303-98039444	141	44
DPYD	DPYD_g	chr1:97770916-97770924	chr1:97770882-97771019	137	43

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
DPYD	DPYD_h	chr1:98165087-98165095	chr1:98164981-98165128	147	44
			chr1:98164995-98165134	139	43
EGFR	EGFR_18	chr7:55241611-55241738	chr7:55241576-55241789	213	56
			chr7:55241580-55241751	171	52
			chr7:55241626-55241822	196	55
EGFR	EGFR_19	chr7:55242412-55242515	chr7:55242355-55242548	193	47
			chr7:55242357-55242539	182	46
EGFR	EGFR_20	chr7:55248983-55249173	chr7:55248931-55249219	288	60
			chr7:55248958-55249121	163	63
			chr7:55249036-55249220	184	57
EGFR	EGFR_21	chr7:55259409-55259569	chr7:55259367-55259564	197	54
			chr7:55259367-55259618	251	54
			chr7:55259486-55259683	197	51
ERBB2	ERBB2_20	chr17:37880976-37881166	chr17:37880956-37881156	200	60
			chr17:37881009-37881184	175	58
GSTP1	GSTP1_a	chr11:67352682-67352696	chr11:67352653-67352763	110	55
HRAS	HRAS_a	chr11:534281-534293	chr11:534207-534319	112	56
			chr11:534221-534328	107	58
HRAS	HRAS_b	chr11:533870-533884	chr11:533768-533926	158	57
			chr11:533839-533944	105	60
HRAS	HRAS_c	chr11:533549-533557	chr11:533469-533606	137	60
			chr11:533492-533604	112	60
HRAS	HRAS_d	chr11:533462-533470	chr11:533362-533490	128	69
			chr11:533375-533532	157	68

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr11:533469-533606	137	60
IDH1	IDH1_a	chr2:209113110-209113113	chr2:209113081-209113224	143	45
			chr2:209113081-209113233	152	44
IDH2	IDH2_a	chr15:90631932-90631935	chr15:90631870-90631993	123	52
			chr15:90631870-90632013	143	51
IDH2	IDH2_b	chr15:90631836-90631839	chr15:90631716-90631861	145	62
			chr15:90631730-90631890	160	62
KIT	KIT_11	chr4:55593579-55593710	chr4:55593484-55593753	269	39
			chr4:55593515-55593749	234	39
KIT	KIT_13	chr4:55594174-55594289	chr4:55594134-55594327	193	44
KIT	KIT_14	chr4:55595498-55595653	chr4:55595415-55595657	242	36
			chr4:55595461-55595701	240	37
			chr4:55595591-55595691	100	39
KIT	KIT_17	chr4:55599233-55599360	chr4:55599192-55599390	198	38
			chr4:55599249-55599471	222	37
KIT	KIT_18	chr4:55602661-55602777	chr4:55602617-55602825	208	40
KIT	KIT_9	chr4:55592020-55592218	chr4:55591945-55592179	234	42
			chr4:55592075-55592262	187	38
KRAS	KRAS_a	chr12:25398274-25398291	chr12:25398185-25398329	144	39
			chr12:25398185-25398332	147	39
KRAS	KRAS_b	chr12:25380269-25380283	chr12:25380213-25380316	103	49
KRAS	KRAS_c	chr12:25378554-25378568	chr12:25378521-25378619	98	34
			chr12:25378521-25378625	104	35
KRAS	KRAS_d	chr12:25378644-25378653	chr12:25378617-25378774	157	34

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr12:25378618-25378777	159	33
MAP2K1	MAP2K1_a	chr15:66727449-66727486	chr15:66727398-66727542	144	54
			chr15:66727420-66727512	92	53
MAPK1	MAPK1_1	chr22:22221609-22221732	chr22:22221487-22221662	175	76
			chr22:22221490-22221659	169	75
			chr22:22221595-22221688	93	68
			chr22:22221601-22221702	101	68
			chr22:22221634-22221798	164	78
			chr22:22221637-22221803	166	77
MAPK1	MAPK1_2	chr22:22161950-22162137	chr22:22161907-22162096	189	43
			chr22:22161907-22162136	229	42
			chr22:22161958-22162176	218	41
			chr22:22162016-22162178	162	43
MAPK1	MAPK1_3	chr22:22160136-22160330	chr22:22160021-22160175	154	33
			chr22:22160122-22160286	164	45
			chr22:22160220-22160371	151	41
MAPK1	MAPK1_4	chr22:22153298-22153419	chr22:22153248-22153419	171	44
			chr22:22153264-22153479	215	40
			chr22:22153343-22153540	197	38
MAPK1	MAPK1_5	chr22:22142980-22143099	chr22:22142925-22143098	173	46
			chr22:22143051-22143177	126	45
MAPK1	MAPK1_6	chr22:22142543-22142679	chr22:22142465-22142665	200	36
			chr22:22142499-22142749	250	36
			chr22:22142546-22142745	199	36

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
MAPK1	MAPK1_7	chr22:22127159-22127273	chr22:22127112-22127303	191	47
			chr22:22127112-22127307	195	47
MAPK1	MAPK1_8	chr22:22123490-22123611	chr22:22123448-22123624	176	40
			chr22:22123459-22123654	195	38
			chr22:22123499-22123697	198	38
MTHFR	MTHFR_a	chr1:11856371-11856385	chr1:11856347-11856476	129	53
			chr1:11856349-11856435	86	51
MTHFR	MTHFR_b	chr1:11854469-11854483	chr1:11854433-11854519	86	52
			chr1:11854433-11854568	135	52
MTOR	MTOR_0	chr1:11319302-11319468	chr1:11319264-11319432	168	57
			chr1:11319360-11319552	192	48
			chr1:11319403-11319553	150	48
MTOR	MTOR_1	chr1:11318539-11318652	chr1:11318494-11318653	159	45
			chr1:11318549-11318730	181	40
MTOR	MTOR_10	chr1:11298456-11298676	chr1:11298401-11298570	169	52
			chr1:11298473-11298672	199	56
			chr1:11298608-11298733	125	48
MTOR	MTOR_11	chr1:11297897-11298107	chr1:11297829-11298062	233	56
			chr1:11297830-11298025	195	56
			chr1:11297932-11298128	196	57
			chr1:11297938-11298137	199	56
			chr1:11298015-11298209	194	55
MTOR	MTOR_12	chr1:11294197-11294324	chr1:11294140-11294320	180	54
			chr1:11294179-11294376	197	51

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
MTOR	MTOR_13	chr1:11293452-11293546	chr1:11293416-11293575	159	42
			chr1:11293416-11293587	171	41
			chr1:11293502-11293640	138	34
MTOR	MTOR_14	chr1:11292490-11292587	chr1:11292453-11292614	161	44
			chr1:11292478-11292624	146	42
MTOR	MTOR_15	chr1:11291354-11291493	chr1:11291283-11291482	199	51
			chr1:11291301-11291523	222	51
			chr1:11291395-11291540	145	51
MTOR	MTOR_16	chr1:11290979-11291113	chr1:11290930-11291107	177	50
			chr1:11290995-11291191	196	48
MTOR	MTOR_17	chr1:11288722-11288977	chr1:11288694-11288837	143	52
			chr1:11288783-11288938	155	52
			chr1:11288866-11289048	182	49
MTOR	MTOR_18	chr1:11276202-11276293	chr1:11276136-11276316	180	44
			chr1:11276254-11276394	140	45
MTOR	MTOR_19	chr1:11273453-11273625	chr1:11273386-11273568	182	50
			chr1:11273419-11273657	238	47
			chr1:11273523-11273673	150	46
MTOR	MTOR_2	chr1:11316987-11317224	chr1:11316917-11317167	250	57
			chr1:11316937-11317132	195	56
			chr1:11317039-11317219	180	53
			chr1:11317059-11317290	231	53
			chr1:11317122-11317289	167	52
MTOR	MTOR_20	chr1:11272850-11272967	chr1:11272815-11272975	160	49

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr1:11272871-11273029	158	48
MTOR	MTOR_21	chr1:11272366-11272533	chr1:11272332-11272525	193	54
			chr1:11272412-11272590	178	51
MTOR	MTOR_22	chr1:11270868-11270965	chr1:11270785-11270983	198	39
			chr1:11270830-11271009	179	40
			chr1:11270901-11271087	186	38
MTOR	MTOR_23	chr1:11269366-11269517	chr1:11269329-11269510	181	50
			chr1:11269329-11269544	215	48
			chr1:11269415-11269611	196	43
MTOR	MTOR_24	chr1:11264615-11264762	chr1:11264525-11264725	200	53
			chr1:11264583-11264812	229	59
			chr1:11264668-11264812	144	56
MTOR	MTOR_25	chr1:11259595-11259762	chr1:11259547-11259742	195	49
			chr1:11259621-11259794	173	48
MTOR	MTOR_26	chr1:11259312-11259462	chr1:11259269-11259496	227	49
			chr1:11259283-11259481	198	51
			chr1:11259299-11259490	191	53
MTOR	MTOR_27	chr1:11227496-11227576	chr1:11227454-11227608	154	44
MTOR	MTOR_28	chr1:11217206-11217350	chr1:11217177-11217376	199	54
			chr1:11217177-11217384	207	54
			chr1:11217262-11217457	195	46
MTOR	MTOR_29	chr1:11210180-11210285	chr1:11210155-11210336	181	48
			chr1:11210155-11210351	196	48
MTOR	MTOR_3	chr1:11316046-11316251	chr1:11315980-11316216	236	54

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr1:11315982-11316181	199	55
			chr1:11316099-11316281	182	53
			chr1:11316161-11316373	212	48
MTOR	MTOR_30	chr1:11206730-11206850	chr1:11206694-11206859	165	49
			chr1:11206742-11206922	180	50
MTOR	MTOR_31	chr1:11205022-11205104	chr1:11204986-11205137	151	49
MTOR	MTOR_32	chr1:11204702-11204814	chr1:11204662-11204861	199	52
			chr1:11204675-11204861	186	52
MTOR	MTOR_33	chr1:11199587-11199717	chr1:11199411-11199595	184	51
			chr1:11199557-11199732	175	56
			chr1:11199615-11199777	162	46
MTOR	MTOR_34	chr1:11199358-11199494	chr1:11199316-11199494	178	49
			chr1:11199321-11199531	210	47
			chr1:11199411-11199595	184	51
MTOR	MTOR_35	chr1:11194405-11194525	chr1:11194361-11194533	172	56
			chr1:11194435-11194632	197	53
MTOR	MTOR_36	chr1:11193134-11193256	chr1:11193062-11193253	191	57
			chr1:11193124-11193296	172	53
MTOR	MTOR_37	chr1:11190583-11190836	chr1:11190480-11190665	185	52
			chr1:11190523-11190768	245	60
			chr1:11190618-11190812	194	61
			chr1:11190643-11190869	226	60
			chr1:11190664-11190864	200	60
MTOR	MTOR_38	chr1:11189792-11189897	chr1:11189760-11189906	146	53

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr1:11189828-11189939	111	54
MTOR	MTOR_39	chr1:11188909-11189010	chr1:11188860-11189060	200	44
			chr1:11188881-11189061	180	44
MTOR	MTOR_4	chr1:11313893-11314032	chr1:11313839-11314007	168	55
			chr1:11313920-11314062	142	47
MTOR	MTOR_40	chr1:11188508-11188611	chr1:11188475-11188658	183	46
			chr1:11188475-11188689	214	44
MTOR	MTOR_41	chr1:11188058-11188185	chr1:11188017-11188196	179	55
			chr1:11188090-11188285	195	45
MTOR	MTOR_42	chr1:11187678-11187865	chr1:11187630-11187824	194	48
			chr1:11187644-11187880	236	51
			chr1:11187745-11187945	200	50
			chr1:11187755-11187964	209	50
MTOR	MTOR_43	chr1:11187064-11187203	chr1:11187021-11187193	172	49
			chr1:11187083-11187254	171	47
MTOR	MTOR_44	chr1:11186676-11186855	chr1:11186642-11186789	147	50
			chr1:11186733-11186894	161	45
MTOR	MTOR_45	chr1:11184552-11184692	chr1:11184522-11184710	188	48
			chr1:11184527-11184726	199	49
MTOR	MTOR_46	chr1:11182033-11182185	chr1:11181999-11182166	167	56
			chr1:11182026-11182217	191	54
MTOR	MTOR_47	chr1:11181300-11181427	chr1:11181248-11181448	200	57
			chr1:11181341-11181492	151	49
MTOR	MTOR_48	chr1:11177058-11177145	chr1:11177000-11177154	154	38

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr1:11177010-11177180	170	40
			chr1:11177111-11177167	56	46
			chr1:11177123-11177247	124	38
MTOR	MTOR_49	chr1:11175450-11175527	chr1:11175417-11175568	151	49
MTOR	MTOR_5	chr1:11307873-11308153	chr1:11307713-11307911	198	48
			chr1:11307824-11307989	165	54
			chr1:11307947-11308136	189	53
			chr1:11307996-11308224	228	46
			chr1:11308035-11308224	189	42
MTOR	MTOR_50	chr1:11174867-11174946	chr1:11174829-11174981	152	41
MTOR	MTOR_51	chr1:11174372-11174512	chr1:11174347-11174531	184	53
			chr1:11174366-11174566	200	52
MTOR	MTOR_52	chr1:11172906-11172976	chr1:11172880-11173039	159	45
			chr1:11172882-11173039	157	45
MTOR	MTOR_53	chr1:11169703-11169788	chr1:11169648-11169834	186	42
			chr1:11169670-11169834	164	43
MTOR	MTOR_54	chr1:11169344-11169429	chr1:11169300-11169473	173	39
			chr1:11169300-11169490	190	40
MTOR	MTOR_55	chr1:11168235-11168345	chr1:11168184-11168362	178	47
			chr1:11168230-11168424	194	47
MTOR	MTOR_56	chr1:11167539-11167559	chr1:11167453-11167651	198	38
			chr1:11167457-11167616	159	38
MTOR	MTOR_6	chr1:11307679-11307792	chr1:11307627-11307819	192	50
			chr1:11307713-11307911	198	48

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
MTOR	MTOR_7	chr1:11303168-11303359	chr1:11303118-11303318	200	53
			chr1:11303125-11303369	244	51
			chr1:11303216-11303408	192	50
MTOR	MTOR_8	chr1:11301607-11301740	chr1:11301563-11301748	185	56
			chr1:11301635-11301811	176	50
MTOR	MTOR_9	chr1:11300357-11300606	chr1:11300300-11300533	233	55
			chr1:11300308-11300507	199	56
			chr1:11300388-11300637	249	58
			chr1:11300393-11300590	197	58
			chr1:11300494-11300669	175	51
NRAS	NRAS_a	chr1:115258737-115258754	chr1:115258618-115258778	160	51
			chr1:115258707-115258813	106	49
NRAS	NRAS_b	chr1:115256522-115256536	chr1:115256435-115256573	138	44
			chr1:115256499-115256573	74	44
NRAS	NRAS_c	chr1:115252286-115252294	chr1:115252219-115252339	120	45
NRAS	NRAS_d	chr1:115252199-115252207	chr1:115252152-115252262	110	46
PDGFRA	PDGFRA_12	chr4:55141005-55141142	chr4:55140958-55141182	224	45
PDGFRA	PDGFRA_14	chr4:55144060-55144175	chr4:55144023-55144217	194	47
PDGFRA	PDGFRA_18	chr4:55152005-55152132	chr4:55151966-55152172	206	50
PIK3CA	PIK3CA_a	chr3:178936076-178936102	chr3:178936001-178936244	243	34
			chr3:178936046-178936198	152	36
PIK3CA	PIK3CA_b	chr3:178952078-178952092	chr3:178952032-178952128	96	39
			chr3:178952039-178952141	102	39
PTEN	PTEN_a	chr10:89692985-89692999	chr10:89692941-89693051	110	39

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
PTEN	PTEN_b	chr10:89717666-89717680	chr10:89717587-89717717	130	45
			chr10:89717589-89717728	139	46
PTEN	PTEN_c	chr10:89717765-89717785	chr10:89717709-89717831	122	36
			chr10:89717744-89717891	147	29
STAT1	STAT1_0	chr2:191874599-191874731	chr2:191874528-191874718	190	42
			chr2:191874613-191874793	180	41
STAT1	STAT1_1	chr2:191873686-191873835	chr2:191873624-191873871	247	38
			chr2:191873635-191873803	168	38
			chr2:191873697-191873892	195	38
STAT1	STAT1_10	chr2:191851762-191851796	chr2:191851613-191851799	186	35
			chr2:191851649-191851803	154	31
			chr2:191851690-191851887	197	26
			chr2:191851777-191851888	111	24
STAT1	STAT1_11	chr2:191851577-191851675	chr2:191851498-191851620	122	48
			chr2:191851576-191851671	95	49
			chr2:191851613-191851799	186	35
			chr2:191851649-191851803	154	31
STAT1	STAT1_12	chr2:191850342-191850388	chr2:191850312-191850430	118	38
STAT1	STAT1_13	chr2:191849033-191849121	chr2:191848981-191849156	175	37
			chr2:191849006-191849181	175	39
			chr2:191849069-191849244	175	41
STAT1	STAT1_14	chr2:191848365-191848468	chr2:191848325-191848479	154	55
			chr2:191848356-191848530	174	51
STAT1	STAT1_15	chr2:191847106-191847246	chr2:191847028-191847218	190	38

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr2:191847030-191847274	244	39
			chr2:191847178-191847370	192	43
STAT1	STAT1_16	chr2:191845343-191845397	chr2:191845312-191845453	141	42
STAT1	STAT1_17	chr2:191844495-191844594	chr2:191844415-191844614	199	39
			chr2:191844445-191844651	206	35
			chr2:191844553-191844675	122	34
STAT1	STAT1_18	chr2:191843579-191843729	chr2:191843525-191843725	200	59
			chr2:191843601-191843759	158	58
STAT1	STAT1_19	chr2:191841563-191841753	chr2:191841511-191841754	243	45
			chr2:191841512-191841692	180	46
			chr2:191841629-191841810	181	41
			chr2:191841677-191841810	133	41
STAT1	STAT1_2	chr2:191872286-191872389	chr2:191872197-191872377	180	33
			chr2:191872207-191872447	240	31
			chr2:191872329-191872389	60	40
			chr2:191872342-191872470	128	28
STAT1	STAT1_20	chr2:191840535-191840615	chr2:191840499-191840649	150	40
STAT1	STAT1_21	chr2:191839553-191839660	chr2:191839503-191839699	196	48
			chr2:191839516-191839699	183	48
STAT1	STAT1_22	chr2:191835426-191835445	chr2:191835395-191835502	107	34
STAT1	STAT1_3	chr2:191865797-191865891	chr2:191865727-191865932	205	40
			chr2:191865750-191865938	188	41
STAT1	STAT1_4	chr2:191864349-191864432	chr2:191864320-191864483	163	38
			chr2:191864320-191864484	164	38

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
STAT1	STAT1_5	chr2:191862940-191863036	chr2:191862850-191863096	246	33
			chr2:191862910-191863091	181	35
STAT1	STAT1_6	chr2:191862579-191862735	chr2:191862544-191862710	166	51
			chr2:191862648-191862845	197	30
			chr2:191862668-191862840	172	29
STAT1	STAT1_7	chr2:191859784-191859947	chr2:191859744-191859926	182	42
			chr2:191859744-191859962	218	44
			chr2:191859799-191859994	195	45
			chr2:191859800-191859996	196	44
STAT1	STAT1_8	chr2:191855951-191856048	chr2:191855914-191856111	197	48
			chr2:191855915-191856093	178	50
STAT1	STAT1_9	chr2:191854338-191854402	chr2:191854239-191854418	179	28
			chr2:191854270-191854480	210	25
			chr2:191854374-191854505	131	31
STAT3	STAT3_0	chr17:40500404-40500536	chr17:40500316-40500516	200	47
			chr17:40500411-40500610	199	51
STAT3	STAT3_1	chr17:40498584-40498733	chr17:40498525-40498708	183	46
			chr17:40498525-40498749	224	46
			chr17:40498650-40498844	194	42
			chr17:40498692-40498844	152	41
STAT3	STAT3_10	chr17:40481762-40481796	chr17:40481632-40481773	141	45
			chr17:40481704-40481863	159	45
			chr17:40481731-40481843	112	45
STAT3	STAT3_11	chr17:40481569-40481667	chr17:40481494-40481693	199	50

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr17:40481533-40481703	170	47
			chr17:40481632-40481773	141	45
STAT3	STAT3_12	chr17:40481425-40481477	chr17:40481366-40481506	140	49
			chr17:40481373-40481505	132	50
STAT3	STAT3_13	chr17:40478131-40478219	chr17:40478074-40478256	182	51
STAT3	STAT3_14	chr17:40476978-40477081	chr17:40476949-40477109	160	56
			chr17:40476950-40477142	192	56
STAT3	STAT3_15	chr17:40476726-40476866	chr17:40476640-40476780	140	58
			chr17:40476674-40476910	236	53
			chr17:40476716-40476913	197	49
STAT3	STAT3_16	chr17:40475588-40475645	chr17:40475546-40475687	141	41
STAT3	STAT3_17	chr17:40475275-40475374	chr17:40475235-40475407	172	48
STAT3	STAT3_18	chr17:40475019-40475163	chr17:40474995-40475190	195	54
			chr17:40475030-40475210	180	51
STAT3	STAT3_19	chr17:40474297-40474514	chr17:40474229-40474410	181	46
			chr17:40474229-40474479	250	45
			chr17:40474329-40474561	232	47
			chr17:40474343-40474515	172	46
			chr17:40474468-40474602	134	48
STAT3	STAT3_2	chr17:40497574-40497677	chr17:40497541-40497718	177	51
			chr17:40497542-40497739	197	51
STAT3	STAT3_20	chr17:40469197-40469244	chr17:40469143-40469279	136	45
			chr17:40469162-40469305	143	47
STAT3	STAT3_21	chr17:40468804-40468921	chr17:40468739-40468917	178	49

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr17:40468862-40469026	164	48
STAT3	STAT3_22	chr17:40467760-40467820	chr17:40467728-40467860	132	54
STAT3	STAT3_3	chr17:40491329-40491429	chr17:40491290-40491450	160	57
			chr17:40491313-40491458	145	57
STAT3	STAT3_4	chr17:40490746-40490832	chr17:40490645-40490892	247	34
			chr17:40490692-40490890	198	33
STAT3	STAT3_5	chr17:40489778-40489877	chr17:40489711-40489910	199	53
			chr17:40489732-40489909	177	53
STAT3	STAT3_6	chr17:40489450-40489606	chr17:40489368-40489564	196	48
			chr17:40489411-40489659	248	50
			chr17:40489491-40489686	195	50
STAT3	STAT3_7	chr17:40485906-40486069	chr17:40485844-40486043	199	43
			chr17:40485844-40486071	227	43
			chr17:40485925-40486165	240	43
			chr17:40485962-40486143	181	40
STAT3	STAT3_8	chr17:40485688-40485785	chr17:40485635-40485806	171	54
			chr17:40485696-40485860	164	60
STAT3	STAT3_9	chr17:40483487-40483551	chr17:40483423-40483600	177	32
TP53	TP53_0	chr17:7579836-7579914	chr17:7579781-7579946	165	60
			chr17:7579810-7579966	156	57
TP53	TP53_1	chr17:7579697-7579723	chr17:7579553-7579751	198	57
			chr17:7579599-7579757	158	58
TP53	TP53_2	chr17:7579309-7579592	chr17:7579268-7579457	189	60
			chr17:7579378-7579534	156	62

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr17:7579442-7579617	175	56
			chr17:7579553-7579751	198	57
TP53	TP53_3	chr17:7578368-7578556	chr17:7578326-7578501	175	65
			chr17:7578418-7578597	179	58
TP53	TP53_4	chr17:7578174-7578291	chr17:7578076-7578331	255	54
			chr17:7578084-7578279	195	54
			chr17:7578238-7578327	89	55
TP53	TP53_5	chr17:7577496-7577610	chr17:7577463-7577641	178	56
TP53	TP53_6	chr17:7577016-7577157	chr17:7576986-7577057	71	61
			chr17:7576988-7577210	222	55
			chr17:7576995-7577060	65	63
			chr17:7577041-7577214	173	53
TP53	TP53_7	chr17:7576850-7576928	chr17:7576793-7576973	180	45
TP53	TP53_8	chr17:7573924-7574035	chr17:7573874-7574063	189	61
			chr17:7573894-7574073	179	59
TP53	TP53_9	chr17:7572924-7573010	chr17:7572851-7573045	194	52
			chr17:7572895-7573046	151	50
TYMP	TYMP_0	chr22:50967922-50968140	chr22:50967858-50968047	189	67
			chr22:50967971-50968139	168	69
			chr22:50968080-50968243	163	71
TYMP	TYMP_1	chr22:50967562-50967769	chr22:50967507-50967698	191	62
			chr22:50967604-50967783	179	64
			chr22:50967718-50967883	165	68
TYMP	TYMP_2	chr22:50966938-50967041	chr22:50966860-50967060	200	55

Gene	Target	Target region	Amplicon region	Amplicon length (bp)	Amplicon GC content (%)
			chr22:50966976-50967117	141	60
TYMP	TYMP_3	chr22:50966014-50966148	chr22:50965980-50966177	197	57
TYMP	TYMP_4	chr22:50965591-50965714	chr22:50965553-50965744	191	61
TYMP	TYMP_5	chr22:50965002-50965169	chr22:50964791-50965035	244	77
			chr22:50964973-50965125	152	71
			chr22:50965012-50965194	182	70
			chr22:50965108-50965319	211	66
TYMP	TYMP_6	chr22:50964672-50964907	chr22:50964619-50964809	190	73
			chr22:50964644-50964892	248	75
			chr22:50964739-50964935	196	77
			chr22:50964791-50965035	244	77
TYMP	TYMP_7	chr22:50964427-50964587	chr22:50964354-50964541	187	79
			chr22:50964354-50964585	231	79
			chr22:50964448-50964636	188	77
TYMP	TYMP_8	chr22:50964196-50964349	chr22:50964160-50964348	188	72
			chr22:50964184-50964377	193	73
TYMS	TYMS_a	chr18:673437-673451	chr18:673367-673498	131	35
			chr18:673367-673516	149	36
UGT1A1	UGT1A1_a	chr2:234668849-234668909	chr2:234668782-234668979	197	49
			chr2:234668782-234669030	248	51
			chr2:234668806-234668958	152	46