

**GERMLINE PHARMACOGENOMIC TESTING IN FORMALIN-FIXED
PARAFFIN-EMBEDDED TUMOURS**

by

Shyong Quin Yap

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

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

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Experimental Medicine Program)

The University of British Columbia

(Vancouver)

June 2017

© Shyong Quin Yap, 2017

Abstract

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

Note: Abstracts should generally try to avoid using acronyms.

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

Preface

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

Table of Contents

| | |
|--|-------------|
| Abstract | ii |
| Preface | iii |
| Table of Contents | iv |
| List of Tables | vi |
| List of Figures | vii |
| List of Abbreviations | viii |
| Acknowledgments | ix |
| 1 Introduction | 1 |
| 1.1 The Evolution of Molecular Diagnostics in Cancer | 1 |
| 1.2 Next-generation Sequencing Technologies | 1 |
| 1.3 Applications of Next-generation Sequencing | 1 |
| 1.3.1 Targeted Sequencing | 1 |
| 1.3.2 Whole Exome Sequencing | 1 |
| 1.3.3 Whole Genome Sequencing | 1 |
| 1.4 Bioinformatics Tools for Variant Calling | 1 |
| 1.4.1 Types of Genomic Alterations | 1 |
| 1.4.2 Variant Calling Pipeline | 2 |
| 1.4.3 Variant Calling Algorithms | 2 |
| 1.4.4 Variant Curation and Interpretation | 2 |
| 1.5 Germline Variant Calling in The Tumour Genome | 2 |
| 1.5.1 Incidental Findings | 2 |
| 1.5.2 Pharmacogenomic Variants | 2 |
| 1.5.3 Challenges | 3 |

| | | |
|----------|--|-----------|
| 1.6 | Objectives | 3 |
| 2 | Characterization of Formalin-Induced DNA Damages | 4 |
| 2.1 | Materials and Methods | 4 |
| 2.1.1 | Patient Samples | 4 |
| 2.1.2 | OncoPanel (Solid Tumour NGS Panel) | 5 |
| 2.1.3 | Sample Preparation, Library Construction, and Illumina Sequencing | 5 |
| 2.1.4 | Variant Calling Pipeline | 6 |
| 2.1.5 | Data Analysis | 6 |
| 2.2 | Results | 7 |
| 2.2.1 | Formalin fixation affects amplicon yield and sequencing results | 7 |
| 2.3 | Discussion | 7 |
| 3 | Concordance of Germline PGx Variants Between Blood and Tumours | 8 |
| 3.1 | Materials and Methods | 8 |
| 3.2 | Results | 8 |
| 3.2.1 | Frequency of germline and somatic variants | 8 |
| 3.2.2 | Variant Allele Frequency Thresholds Can Maximize Sensitivity and Positive Predictive Value of Germline Variant Calling in Tumours | 8 |
| 3.3 | Discussion | 8 |
| 4 | Conclusion | 10 |
| | Bibliography | 11 |
| A | Supporting Materials | 12 |

List of Tables

| | | |
|-----------|---|----|
| Table 2.1 | Distribution of cancer types in the TOP cohort. | 5 |
| Table 2.2 | Assessment of Amplicon Generation, Sequencing Results, and Coverage Statistics. | 7 |
| Table 3.1 | Frequency of germline and somatic variants detected in the tumours of 213 patients in the TOP cohort. | 9 |
| Table A.1 | Gene Reference Models for HGVS Nomenclature. | 13 |

List of Figures

List of Abbreviations

| | |
|------------|---|
| GPS | Graduate and Postdoctoral Studies |
| PDF | Portable Document Format |
| URL | Unique Resource Locator, used to describe a means for obtaining some resource on the world wide web |

Acknowledgments

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

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

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

Chapter 1

Introduction

1.1 The Evolution of Molecular Diagnostics in Cancer

Cancer is a

1.2 Next-generation Sequencing Technologies

1.3 Applications of Next-generation Sequencing

1.3.1 Targeted Sequencing

Capture-based, amplicon-based etc.

1.3.2 Whole Exome Sequencing

1.3.3 Whole Genome Sequencing

1.4 Bioinformatics Tools for Variant Calling

1.4.1 Types of Genomic Alterations

There are different types of genomic alterations.

1.4.2 Variant Calling Pipeline

1.4.3 Variant Calling Algorithms

1.4.4 Variant Curation and Interpretation

1.5 Germline Variant Calling in The Tumour Genome

1.5.1 Incidental Findings

The application of next-generation sequencing (NGS) technologies for tumour profiling has been increasingly integrated into oncologic care to detect targetable somatic mutations and personalize treatments for cancer patients. Although analysis of tumour-normal paired samples is required to accurately discriminate between somatic and germline variants, most clinical laboratories only sequence tumour samples to minimize cost and turnaround time [?]. However, genomic analyses of tumours can also reveal secondary genomic findings, which are germline information that may have clinical implications for patients and their family members [?]. In fact, several studies demonstrated that a germline cancer-predisposing variant is present in 3-10% of patients undergoing tumour-normal sequencing [? ? ? ?]. Therefore, clinical laboratories providing tumour genomic testing must be equipped to perform germline confirmatory testing on potential germline variants or be prepared to refer such cases to external services.

1.5.2 Pharmacogenomic Variants

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

1.5.3 Challenges

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

1.6 Objectives

Chapter 2

Characterization of Formalin-Induced DNA Damages

In this chapter, we will address the

2.1 Materials and Methods

2.1.1 Patient Samples

Blood and FFPE tumour specimens were acquired from 213 patients who provided informed consent for The OncoPanel Pilot (TOP) study, a pilot study to optimize the OncoPanel, which is an amplicon-based targeted NGS panel for solid tumours, and assess its application for guiding disease management and therapeutic intervention. Patients in the TOP study are those with advanced cancers including colorectal cancer, lung cancer, melanoma, gastrointestinal stromal tumour (GIST), and other cancers (Table 2.1). The age of paraffin block for tumour specimens ranges from 18 to 5356 days with a median of 274 days.

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

| Cancer Type | Number of Cases | Percentage (%) |
|----------------|-----------------|----------------|
| Colorectal | 97 | 46 |
| Lung | 59 | 28 |
| Melanoma | 18 | 8 |
| Other* | 17 | 8 |
| GIST | 7 | 3 |
| Sarcoma | 4 | 2 |
| Neuroendocrine | 4 | 2 |
| Cervical | 2 | 0.9 |
| Ovarian | 2 | 0.9 |
| Breast | 2 | 0.9 |
| Unknown | 1 | 0.5 |

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

2.1.2 OncoPanel (Solid Tumour NGS Panel)

The OncoPanel is offered by the British Columbia Cancer Agency (BCCA) for clinical genomic testing of coding exons and clinically relevant hotspots of 20 cancer-related genes. The panel also tested for variants in six PGx genes that can predict chemotherapeutic response: *DPYD*, *GSTP1*, *MTHFR*, *TYMP*, *TYMS*, and *UGT1A1*. Full list of genes and gene reference models for the OncoPanel is presented in ???. Primers were designed by RainDance Technologies (Lexington, MA) using the GRCh37/hg19 reference sequence to generate 414 amplicons between 100 bp and 250 bp in size, which interrogate ~ 20 kb of target bases. Complete lists of primers and amplicons are provided in the Supplemental Materials.

2.1.3 Sample Preparation, Library Construction, and Illumina Sequencing

Genomic DNA was extracted from blood and FFPE tumour specimens using the Gentra Autopure LS DNA preparation platform and QIAamp DNA FFPE tissue kit (Qiagen, Hilden, Germany) respectively. The extracted DNA was sheared according to a previously described protocol [?] to attain approximate sizes of 3 kb followed by PCR primer merging, amplification of target regions, and adapter ligation using the Thunderstorm NGS Targeted Enrichment System (RainDance Technologies, Lexington, MA) as per manufacturer's protocol. Barcoded amplicons were sequenced

with the Illumina MiSeq system for paired end sequencing with a v2 250-bp kit (Illumina, San Diego, CA).

2.1.4 Variant Calling Pipeline

Reads that passed the Illumina Chastity filter were aligned to the GRCh37/hg19 human reference genome using the BWA mem algorithm (version 0.5.9) with default parameters, and the alignments were processed and converted to the BAM format using SAMtools (version 0.1.18). Variant calling was performed with the SAMtools mpileup function (`samtools mpileup -BA -d 500000 -L 500000 -q 1`) to generate pileup files for all target bases followed by the VarScan2 mpileup2cns (version 2.3.6) function with parameter thresholds of variant allele frequency $\geq 10\%$ and Phred-scaled base quality score ≥ 20 (`--min-var-freq 0.1 --p-value 0.01 --strand-filter 0 --output-vcf --variants --min-avg-qual 20`). Variant calls were filtered using the VarScan2 fpcfilter function with fraction of variant reads from each strand ≥ 0.1 and default thresholds for other parameters. SnpEff (version 4.2) was used for variant annotation and effect prediction whereas the SnpSift package in SnpEff was used to annotate variants with databases such as dbSNP (b138), COSMIC (version 70), 1000 Genomes Project, ClinVar, and ExAC (release 0.3) for interpretation.

2.1.5 Data Analysis

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

2.2 Results

2.2.1 Formalin fixation affects amplicon yield and sequencing results

Table 2.2: Assessment of Amplicon Generation, Sequencing Results, and Coverage Statistics.

| Parameter | Blood | | FFPE Tumour | | <i>P</i> value |
|------------------------------------|--------|--------------|-------------|--------------|-----------------------|
| | Median | Range | Median | Range | |
| DNA Input (ng) | 147.8 | 55.5–266.4 | 140.9 | 11.8–271.0 | 3.5×10^{-4} |
| Amplicon Yield (ng) | 299.2 | 84.0–1438.0 | 103.6 | 11.6–325.5 | 2.2×10^{-16} |
| On-target Aligned Reads (%) | 86.8 | 74.0–95.9 | 88.4 | 32.5–97.4 | 9.4×10^{-2} |
| Off-target Aligned Reads (%) | 7.3 | 0.8–24.0 | 8.4 | 0.4–60.4 | 0.7 |
| Unaligned Reads (%) | 1.3 | 0.4–7.2 | 0.8 | 0.3–12.1 | 5.2×10^{-15} |
| Contaminant (%) | 3.9 | 0.1–8.1 | 1.4 | 0.03–63.6 | 9.9×10^{-4} |
| Zero Coverage Target Bases (%) | 0.00 | 0.00–0.01 | 0.00 | 0.00–0.36 | – |
| $\geq 100\times$ Target Bases (%) | 100.00 | 99.97–100.00 | 100.00 | 54.95–100.00 | 0.58 |
| $\geq 1000\times$ Target Bases (%) | 100.00 | 99.00–100.00 | 99.98 | 20.61–100.00 | 1.92×10^{-5} |

2.3 Discussion

Chapter 3

Concordance of Germline PGx Variants Between Blood and Tumours

3.1 Materials and Methods

3.2 Results

3.2.1 Frequency of germline and somatic variants

3.2.2 Variant Allele Frequency Thresholds Can Maximize Sensitivity and Positive Predictive Value of Germline Variant Calling in Tumours

3.3 Discussion

Table 3.1: Frequency of germline and somatic variants detected in the tumours of 213 patients in the TOP cohort.

| Gene | Germline (N Patients) | Pathogenic Germline (N Patients) | Somatic (N Patients) |
|----------------------------|--------------------------|-------------------------------------|-------------------------|
| <i>Cancer predisposing</i> | | | |
| AKT1 | 0 | 0 | 2 (2) |
| ALK | 1 (1) | 1 (1) | 2 (1) |
| BRAF | 0 | 0 | 18 (17) |
| EGFR | 170 (164) | 5 (5) | 31 (24) |
| HRAS | 0 | 0 | 1 (1) |
| MAP2K1 | 0 | 0 | 2 (2) |
| MAPK1 | 17 (17) | 3 (3) | 3 (2) |
| MTOR | 763 (213) | 6 (6) | 71 (30) |
| NRAS | 0 | 0 | 8 (8) |
| PDGFRA | 242 (185) | 0 | 8 (4) |
| PIK3CA | 0 | 0 | 15 (4) |
| PTEN | 0 | 0 | 1 (1) |
| STAT1 | 54 (51) | 1 (1) | 7 (6) |
| STAT3 | 10 (10) | 4 (4) | 16 (11) |
| TP53 | 189 (184) | 2 (2) | 131 (109) |
| <i>Pharmacogenomics</i> | | | |
| DPYD | 271 (212) | 1 (1) | 1 (1) |
| GSTP1 | 106 (106) | 0 | 0 |
| MTHFR | 209 (177) | 0 | 0 |
| TYMP | 81 (76) | 2 (2) | 18 (13) |
| TYMS | 131 (131) | 0 | 0 |
| UGT1A1 | 96 (96) | 0 | 1 (1) |
| Total | 2396 (213*) | 25 (23*) | 431 (180*) |

Chapter 4

Conclusion

What are my conclusions?

Bibliography

Appendix A

Supporting Materials

Table A.1: Gene Reference Models for HGVS Nomenclature.

| Gene | Protein | Reference Model |
|--------|--|-----------------|
| AKT1 | Protein kinase B | NM_001014431.1 |
| ALK | Anaplastic lymphoma receptor tyrosine kinase | NM_004304.3 |
| BRAF | Serine/threonine-protein kinase B-Raf | NM_004333.4 |
| DPYD | Dihydropyrimidine dehydrogenase | NM_000110.3 |
| EGFR | Epidermal growth factor receptor | NM_005228.3 |
| ERBB2 | Receptor tyrosine-protein kinase erbB-2 | NM_001005862.1 |
| GSTP1 | Glutathione S-transferase pi 1 | NM_000852.3 |
| HRAS | GTPase HRas | NM_005343.2 |
| IDH1 | Isocitrate dehydrogenase 1 | NM_005896.2 |
| IDH2 | Isocitrate dehydrogenase 2 | NM_002168.2 |
| KIT | Tyrosine-protein kinase Kit | NM_000222.2 |
| KRAS | KRas proto-oncogene GTPase | NM_033360.2 |
| MAPK1 | Mitogen-activated protein kinase 1 | NM_002745.4 |
| MAP2K1 | Mitogen-activated protein kinase kinase 1 | NM_002755.3 |
| MTHFR | Methylenetetrahydrofolate reductase | NM_005957.4 |
| MTOR | Serine/threonine-protein kinase mTOR | NM_004958.3 |
| NRAS | Neuroblastoma RAS viral oncogene homolog | NM_002524.3 |
| PDGFRA | Platelet-derived growth factor receptor alpha | NM_006206.4 |
| PIK3CA | Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha | NM_006218.2 |
| PTEN | Phosphatase and tensin homolog | NM_000314.4 |
| STAT1 | Signal transducer and activator of transcription 1 | NM_007315.3 |
| STAT3 | Signal transducer and activator of transcription 3 | NM_139276.2 |
| TP53 | Tumor protein P53 | NM_000546.5 |
| TYMP | Thymidine phosphorylase | NM_001113755.2 |
| TYMS | Thymidylate synthetase | NM_001071.2 |
| UGT1A1 | Uridine diphosphate (UDP)-glucuronosyl transferase 1A1 | NM_000463.2 |