

Evaluating Additional Genomic Variants Identified by a 52 Gene Panel Used for Identification of Actionable Mutations in Non-squamous Non-small Cell Lung Cancers

by

Kathleen Mary Varty

Bachelor of Science, Biology-Chemistry, University of New Brunswick, 2022

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the Graduate Academic Unit of Biological Sciences

Supervisor: Tony Reiman, MD, Department of Biological Sciences

Examining Board: Bryan Crawford, PhD, Department of Biology

Petra Kienesberger, PhD, Department of Biochemistry and
Molecular Biology, Dalhousie University

Tobias Karakach, PhD, Department of Pharmacology, Dalhousie
University

This thesis is accepted by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

August 2024

©Kathleen Mary Varty, 2024

ABSTRACT

Lung cancer leads global cancer mortality. Much of North America now employs next-generation sequencing (NGS) for patient cancer mutational profiling. This approach identifies driver mutations, typically non-synonymous somatic alterations altering protein coding. Recent findings suggest clinical relevance of synonymous mutations, traditionally disregarded for not altering protein sequences. Yet, their prevalence and clinical implications remain underexplored. This study analyzes genomic data and clinical outcomes for non-squamous non-small cell lung cancer patients sequenced at the Saint John Regional Hospital in New Brunswick, Canada from January 2019 to January 2023. Nine possibly pathogenic synonymous variants are identified. Additionally, the clinical variant distribution and impact of an expanded gene panel are explored. This research underscores the potential significance of synonymous mutations in cancer and expands on the current knowledge base.

DEDICATION

For my extended family, whose lives have been significantly improved by the benefits of genetic sequencing in cancer, enabling them to make informed decisions that their parents could not. And to my family and friends, whose unwavering support sustains me in all my endeavors—thank you.

ACKNOWLEDGEMENTS

Mum, Dad, and Anna thank you so much for your never-ending support and ability to always help me put things into perspective. This entire experience would have been much more difficult without you. Love you all.

Dr. Reiman, thank you for making my research experience incredibly positive. Your flexibility and support of my endeavors have not gone unnoticed, and I am deeply grateful. Thank you for trusting me and allowing me to take the lead in my research.

Mallory, although we are often apart and on opposing schedules thank you for always being around when I needed you most. I am eternally grateful for our life-long friendship.

Megan, Tori, Will, Aimee, and Maggie thank you for making Saint John feel like home. Your supportive comments, willingness to help, and readiness for a coffee run have been the highlight of my time here. I am beyond lucky to know you and call you friends.

Leah Maclean, thank you for guiding me with bioinformatics analysis and working alongside me. We made a great team, and it was a pleasure to work with you. I hope our paths continue to cross in the future.

Lastly to my lab members, committee members, faculty, and administrative staff who have made my experience at UNB and DMNB so positive. Your guidance was and is so appreciated and valued.

Table of Contents

Abstract.....	ii
Dedication.....	iii
Acknowledgements.....	iv
Table of Contents.....	v
List of Tables.....	viii
List of Figures.....	ix
List of Symbols, Nomenclature, or Abbreviations.....	xi
1.0 Introduction	1
1.1 Prevalence, risk factors, & pathology of lung cancer	1
1.2 The genomic landscape of lung cancer	6
1.2.1 Driver & actionable mutations.....	7
1.2.2 Synonymous mutations	10
1.3 Mutational profiling & Next-Generation Sequencing	12
1.4 Annotation of Variant Call Format files.....	14
1.5 Aim of study.....	15
1.5.1 Objectives	16
2.0 Methods	17
2.1 Study design & population.....	17
2.2 Clinical & demographic variables.....	18
2.3 Genomic Analysis.....	20
2.3.1 Clinical analysis with the IonTorrent platform	20
2.3.2 Pre-processing for analysis	22
2.3.3 Variant Effect Predictor and additional plug-ins	24
2.3.4 Post-annotation processing	26
2.4 Survival analyses	28
3.0 Results	31

3.1 Demographics & clinical characteristics of cohort	31
3.2 Clinical reporting of variants	33
3.3 Comparison of expanded gene panel to predecessor	35
3.4 Novel variants of interest.....	37
3.5 Synonymous variants	43
3.5.1 Synonymous variants with possible splicing impact	43
3.5.2 Synonymous variants with predicted pathogenicity.....	45
3.6 Survival analyses	50
3.6.1 Prevalent variants	54
3.6.2 Synonymous variants	55
4.0 Discussion	59
4.1 Cohort demographics & clinical characteristics	59
4.2 Clinical reporting.....	61
4.3 Benefits of expanded panel	62
4.4 Novel variants of interest.....	63
4.4.1 Prevalent synonymous variants	64
4.4.2 Exploring synonymous variants with predicted splicing impact	65
4.4.3 Predicted pathogenic synonymous variants	66
4.4.4 Variants in MAPK/ERK and PI3K/AKT pathways	67
4.5 Strengths & limitations of the study	71
4.5.1 Strengths.....	71
4.5.1.1 Cohort demographics	71
4.5.1.2 Reflexive sequencing	72
4.5.1.3 Reproducible, inexpensive bioinformatics pipeline	72
4.5.2 Limitations	72
4.5.2.1 Comparison with the predecessor panel	72
4.5.2.2 Cohort size, diversity, and survival analyses	73
4.5.2.3 Splicing predictors.....	74
4.5.2.4 Pathogenic predictors	75

4.6 Future work	75
Bibliography.....	78
Appendix A – Synonymous variants.....	98
Appendix B – Schoenfeld residuals.....	99
Curriculum Vitae	

List of Tables

Table 1.2.1: Examples of clinically actionable mutations.....	9
Table 2.3.1: 52 genes sequenced in the OncoPrint Focus Assay. Genes with asterisks are those that the smaller 12-gene panel could have identified.....	22
Table 3.3.1: Variants captured with the 52-gene panel that would not have been captured with the 12-gene panel.....	36
Table 3.4.2: Ten most prevalent variants identified in cohort.....	39
Table 3.4.4: Variants divided by consequence type.....	42
Table 3.5.1: Synonymous variants in cohort with at least one SpliceAI score.....	44
Table 3.5.2.1: Synonymous variants with possible pathogenicity in cancer as defined by CScape score.....	46
Table 3.5.2.2: Percentile scale for TRaP scores.....	46
Table 3.5.2.3: Synonymous variants with potential pathogenicity as denoted by TRaP score.....	47
Table 3.5.2.4: Synonymous variants with CADD Phred-like score >10.....	48
Table 3.6.2.1: Genes with variants of unknown significance identified, grouped by primary pathway and function.....	56
Table 3.6.2.2: Novel variants in MAPK/ERK and PI3K/AKT pathway with 2 or more predictions of pathogenicity.....	57
Table A1: All synonymous variants in cohort with prediction for splicing impact.....	98

List of Figures

Figure 2.1.1: Schematic indicating workflow of methods and analysis.....	18
Figure 2.3.4: Visual depiction of the bioinformatics methods used.....	28
Figure 3.1.1: Subtypes of NSCLC within cohort.....	31
Figure 3.1.2: TNM staging assigned at diagnosis and treatments received within cohort.....	32
Figure 3.1.3: Survival curves for overall and progression-free survival.....	33
Figure 3.2.1: Clinically reported variants in cohort that were reported using the 52-gene panel.....	34
Figure 3.2.2: Breakdown of clinically reported mutation in cohort found in KRAS and EGFR.....	35
Figure 3.4.1: Number of unique variants identified in each gene.....	38
Figure 3.4.3: OncoPrint displaying the distribution of variants consequences across patients and genes.....	40
Figure 3.5.2.5: Synonymous variants of interest divided into four classes by support for possible pathogenicity.....	49
Figure 3.6.1: Significant covariates in overall survival.....	52
Figure 3.6.2: Significant covariates in progression-free survival.....	53

Figure 3.6.1.1: Kaplan-Meier curves with p-values from log-rank testing comparing patients based on presence of the novel synonymous BRAF variant.....55

Figure 3.6.2.3: Kaplan-Meier curves with p-values from log-rank testing comparing patients based on presence of synonymous variant in gene involved in the MAPK/ERK and PI3K/AKT pathways.....58

Figure 4.4.4.1: Variants of interest and their involvement in the MAPK/ERK and PI3K/AKT pathways.....69

Figure B1: Schoenfeld residuals for significant variables in overall and progression-free survival which do not violate the proportional hazard assumption.....99

List of Symbols, Nomenclature, or Abbreviations

ALK	Anaplastic Lymphoma Kinase
BAM	Binary Alignment Map
CADD	Combined Annotation Dependent Depletion
CNV	Copy Number Variant
COSMIC	Catalogue of Somatic Mutations in Cancer
CScape	Cancer Specific Annotation of Pathogenicity
dbNSFP	Database for Non-synonymous SNPs' Functional Predictions
dbscSNV	Database for Single Nucleotide Variant within Splicing Consensus Regions
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
EMBL-EBI	European Molecular Biology Laboratory-European Bioinformatics Institute
EPV	Events per variable
ExAC	Exome Aggregation Consortium
FASTQ	Fast Quality Score

FuPA	Fragmentase Universal Primer Assay
gnomAD	Genome Aggregation Database
GRCh	Genome Reference Consortium Human Build
HER2	Human Epidermal Growth Factor Receptor 2
IGV	Integrative Genomics Viewer
KM	Kaplan-Meier
KRAS	Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
MANE	Matched Annotation from NCBI and EMBL-EBI
MET	Mesenchymal epithelial transition factor
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NSCLC	Non-small cell lung cancer
OS	Overall survival
PCR	Polymerase chain reaction
PFS	Progression-free survival
PIK3CA	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha

RNA	Ribonucleic acid
SCLC	Small cell lung cancer
SJRH	Saint John Regional Hospital
SNP	Single Nucleotide Polymorphisms
SNV	Single Nucleotide Variant
SpliceAI	Splice Artificial Intelligence
TCGA	The Cancer Genome Atlas Program
TKI	Tyrosine kinase inhibitor
TMB	Tumour mutational burden
TNM	Tumour, Node, Metastasis
TRaP	Transcript-inferred Pathogenicity
UniProt	Universal Protein Resource
VAF	Variant Allele Frequency
VCF	Variant Call Format
VEP	Variant Effect Predictor
YAP1	Yes-associated protein

1.0 Introduction

1.1 Prevalence, risk factors, & pathology of lung cancer

With an estimated 1.76 million deaths per year lung cancer is the leading cause of cancer-related deaths worldwide (Thai et al., 2021). In Canada, lung cancer is the most common cancer, and by far the leading killer of all cancers. Lung cancer kills more than 20,000 Canadians each year, more than breast, prostate, and colon cancer combined. The five-year survival rate for all lung cancers is 22%. For context, prostate, breast, and colorectal cancer – the next most common cancers – all have 5-year survival rate greater than 50%. The Atlantic provinces generally have higher incidence and mortality rates for lung cancer compared to the rest of the Canadian provinces (Canadian Cancer Statistics Advisory Committee et al., 2023).

Lung cancer is traditionally divided into two broad categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). This classification is determined based on the histology of the tumour (Hayashi & Inomata, 2022). Accounting for 85% of all lung cancers, NSCLC is the more prevalent of the two subtypes (Padinharayil et al., 2023). There are three main subtypes within NSCLC: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma is the most prevalent accounting for 40% of all lung cancers (Travis et al., 2015).

Regardless of the cancer subtype, staging – the process of classifying the extent of the cancer within the body – is typically completed. Staging for NSCLC is done with

reference to the tumour, node, and metastasis (TNM) staging system developed by the American Joint Committee on Cancer and approved by the International Association for the Study of Lung Cancer. The size of the primary tumour (T), the spread of the tumour to regional lymph nodes (N), and the presence of distant metastasis (M) all determine the stage of the cancer, which ranges from IA to IVB (Rami-Porta et al., 2017). Outcomes for NSCLC are correlated with the stage of disease at diagnosis with more advanced stages having lower survival rates. Lung cancer 5-year survival rates can range from 62% at stage I to 3% at stage IV (Canadian Cancer Statistics Advisory Committee, 2021; Ellison & Saint-Jacques, 2023).

Most patients will be diagnosed with lung cancer once it is already in the advanced stages when the prognosis is poorer. Cough, hemoptysis, chest pain, and dyspnea are some common symptoms that patients may present with prior to diagnosis at which point the patient may be referred for imaging (Hamilton et al., 2005). Unfortunately, these symptoms are also commonly encountered with other illnesses, making early diagnosis challenging. Currently, low-dose CT is recommended for lung cancer screening in high-risk populations defined as persons who are 55 to 74 years of age with a minimum smoking history of 30 pack-years or more (pack-years = number of cigarette packs smoked per day × the number of years smoked), who currently smoke, or have quit in the past 15 years and are disease-free at the time of screening (Canadian Task Force on Preventive Health Care, 2016). If abnormalities are found in imaging studies a biopsy or cytology sample is

usually obtained for further investigations, classification, and treatment decisions (Planchard et al., 2018)

Length of survival is highly variable and variables such as a patient's tumour genome and available treatments can have a significant impact. For example, the presence of mutation in the Anaplastic Lymphoma Kinase (ALK) or Epidermal Growth Factor Receptor (EGFR) genes have a positive impact on survival time for patients with stage IV disease (Jeon et al., 2023). Immunotherapy, such as immune checkpoint inhibitors, can significantly increase survival depending on the patient's mutational status (De Mello et al., 2021). For example, patients with EGFR mutant NSCLC gain minimal benefit from immunotherapies (Shi et al., 2022). Patients with early-stage disease and who can tolerate surgery typically have the tumour resected. Some of these patients may require no further interventions and surgery is curative. Patients with early-stage NSCLC at high risk of recurrence are typically offered peri-operative systemic therapy if they are fit enough to tolerate it. The aim of this is to control occult metastatic disease and improve survival (Pisters et al., 2022). For patients with early-stage cancer that is non-resectable or where surgery is not curative, radiotherapy, chemotherapy, immunotherapy, and targeted therapy are some of the available treatments (Falkson et al., 2017; Zappa & Mousa, 2016). With approximately 40% of all newly diagnosed patients presenting with stage IV, surgery is often not an option for many patients.

For patients with life-limiting disease, the goal of care is to increase the time of survival and reduce disease-related adverse events while achieving an acceptable quality

of life. For these patients, all previously mentioned treatment modalities may be used however if a patient is eligible for targeted treatment the side effects are typically less, the response rate is higher, and progression-free survival (PFS) is longer (Guo et al., 2022). For a minority of patients, these advances in treatment modalities have led to long-term remissions which may be shifting the paradigm that NSCLC more closely resembles a chronic disease than an incurable one (Gupta et al., 2021; Rigney, 2019).

There are many well established risk factors for lung cancer, the majority of which are environmental. Identifying a genetic component of lung cancer risk has proven difficult over the years. There is an increased risk of developing lung cancer if there is a family history, however, this can partially be accounted for due to shared environment and lifestyle (e.g. smoking, second-hand smoke, asbestos exposure, etc). Germline mutations such as those in the EGFR, Human Epidermal Growth Factor Receptor 2 (HER2), and yes-associated protein 1 (YAP1) genes have begun to appear as inherited mutations that may increase lung cancer risk but screening is not routinely done as no guidelines exist (de Alencar et al., 2020; Kanwal et al., 2017).

Tobacco smoke, both direct and secondary exposure, is one of the greatest risk factors for lung cancer. An estimated 70% of lung cancers in Canada are attributed to the use of tobacco products (Bade & Dela Cruz, 2020; Poirier et al., 2019). Prevention of tobacco use – rather than screening – is the most effective means of reducing lung cancer burden. The effectiveness of this strategy is seen in declining lung cancer rates post-implementation of smoking cessation programs (Ho et al., 2019). Although tobacco is a

serious risk factor, lung cancer in people who have never smoked is now the 7th leading cause of cancer-linked death worldwide (Corrales et al., 2020; “Release Notice - Canadian Cancer Statistics 2019,” 2019; Subramanian & Govindan, 2007). Exposure to arsenic, asbestos, radon, air pollution, and non-tobacco related polycyclic aromatic hydrocarbons have all been linked to the development of lung cancer (Bradley et al., 2019; Planchard et al., 2018).

Atlantic Canada exhibits a unique location of study as there is array of factors that may contribute to the increased lung cancer incidence and mortality rates observed in the region. Compared to the rest of the country, Atlantic Canada displays a slightly higher rate of smoking (Canadian Tobacco and Nicotine Survey (CTNS): Summary of Results, 2019; Pelekanakis et al., 2021). The role of radon in lung cancer development is of particular interest in Canada. Compared to Sweden, a country with similar smoking status and living conditions, Canada reports a higher annual rate of new lung cancer cases, after adjusting for population and age profiles (Khan et al., 2021). Over 40% of Atlantic Canada’s population are considered rural residents, much greater than the national average of 17.8% (Population Growth in Canada’s Rural Areas, 2016 to 2021, 2022). People living in rural communities can experience as much as 31.2% greater average residential radon levels relative to urban equivalents, partially due to wells acting as conduits for radon exposure (Khan et al., 2024). Another environmental carcinogen that is found in hyperabundance in the Atlantic regions is arsenic (Dummer et al., 2015; Grosz et al., 2004).

Exposure to environmental carcinogens such as radon and asbestos is of interest as they are known genotoxic carcinogens. This means they can induce genetic alteration or directly damage DNA which may account for their carcinogenicity (Martinez et al., 2019). The mechanism for these genetic aberrations and the development of lung cancer is distinct compared to those that occur from tobacco smoke (Corrales et al., 2020; Sun et al., 2007).

1.2 The genomic landscape of lung cancer

The development of lung cancer is a stepwise process that involves the acquisition of genetic mutations and epigenetic changes that alter cellular processes (Gomperts et al., 2011). Prior to the 21st century the characteristics and treatment of cancer focused on the macroscopic. Cancers were classified solely on morphology and location with surgery being a primary treatment modality (Faguet, 2015). In 1970 the first oncogene, SRC – a mutated gene with the potential to cause cancer – was discovered in a chicken retrovirus (Duesberg & Vogt, 1970). This discovery provided one explanation for the initiation of carcinogenesis and launched molecular cancer research as we know it today.

With the molecular revolution and the rapid expansion of the knowledge around cancer, it became increasingly important to define universal themes that exist in all cancers. It is easiest to define cancer by a selection of acquired cellular capabilities – or hallmarks – as proposed by Hanahan and Weinberg. The six original hallmarks are as follows: self-sufficiency in growth signals, evading apoptosis, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, and tissue invasion and

metastasis (Hanahan & Weinberg, 2000). Eventually additional hallmarks have been added, such as genome instability and mutation which underlies the acquisition of many of the other hallmarks (Hanahan & Weinberg, 2011).

1.2.1 Driver & actionable mutations

Virtually all cancer genomes are abnormal in some respect. These mutations and variations can initiate cancer development, drive the cancer's behaviour, and will continue to accumulate as the cancer progresses (Hanahan & Weinberg, 2011). Somatic mutations – which are not inherited – accumulate over an individual's lifespan. Most of these mutations are neutral, meaning they are not believed to have any impact on the cancer; these mutations are referred to as passenger mutations (Draetta et al., 2022; Martincorena et al., 2017).

There exist certain genetic variants which can drive cancer initiation and development, appropriately named driver mutations (Luo & Lam, 2013). Driver mutations can be found either in oncogenes, where the mutation is activating, or tumour suppressor genes, where the mutation results in a loss of function (Pon & Marra, 2015). Driver mutations provide an advantage to the cancer and are therefore usually under positive selection. In some cancers driver mutations are subject to historical contingency, meaning that how a tumour progresses can be dependent on all the mutations acquired throughout its history. This means that a mutation can be beneficial in one genetic background and detrimental in another. Additionally, certain genotypes may only be

developed if certain mutations were gained earlier in the cancer's evolution (Porta-Pardo et al., 2020).

A subset of driver mutations have associated targeted therapies and are referred to as actionable mutations. One of the first identified actionable mutations was found in the EGFR gene which encodes a receptor tyrosine kinase. Kinases are involved in many cell regulation pathways and deregulated expression of such kinases can lead to angiogenesis, proliferation, and evasion of apoptosis. This initial mutation was identified when small molecule inhibitors of EGFR were being evaluated in clinical trials. In these trials it was noted that a subset of patients had a lasting progression-free response to EGFR inhibitors, much greater than that of traditional chemotherapy. This led to further investigation, which found that this subset of patients possessed a mutation within EGFR (Garassino et al., 2009). In the case of deregulation of tyrosine kinase pathways, a common treatment is the use of tyrosine kinase inhibitors (TKIs) (Carr et al., 2016; Chevallier et al., 2021). The availability of a treatment is what classifies these mutations as actionable.

The use of mutation detection and associated targeted treatment such as TKIs have significantly improved survival rates (Travis et al., 2015). Since the discovery of EGFR mutants, many more driver mutations and subsequent therapies have been developed, some of which show better outcomes than traditional platinum-based chemotherapy (Lindeman et al., 2013; Mok et al., 2009). Unlike traditional chemotherapy which is

indiscriminate, targeted therapies specifically target molecules or pathways involved in the growth and progression of cancers (Zhou & Li, 2022).

Table 1.2.1: Examples of clinically actionable mutations.

Gene	Description	Clinical significance	Targeted therapies	References
ALK	ALK rearrangements occur in approximately 5% of patients. ALK is a tyrosine kinase.	Fusions with ALK result in dysregulation of ALK expression. Patients with ALK fusions tend to be younger with minimal to no history of smoking.	Crizotinib Ceritinib Alectinib	(Arbour & Riely, 2017; Du et al., 2018; Shreenivas et al., 2023; Vollbrecht et al., 2018)
EGFR	EGFR mutations are the most common mutations in NSCLC. EGFR is a receptor tyrosine kinase.	Deletions and point mutations in EGFR result in constitutive activation. This results in hypersensitivity to TKIs, slowing cancer growth.	Gefitinib Erlotinib	(Fu et al., 2022; C. O'Leary et al., 2020; Rosell et al., 2009)
KRAS	KRAS encodes a small GTPase transductor protein. KRAS mutations occur in approximately 30% of NSCLCs.	A G12C mutation results in the constitutive activation of the KRAS protein leading to hyperactivation of downstream oncogenic pathways.	Sotorasib Adagrasib	(Cascetta et al., 2022; Jančík et al., 2010; Puneekar et al., 2022; Reita et al., 2022)
MET	Exon skipping mutations and amplifications are two types of mutations that can occur in MET. These mutations occur in less than 10% of patients.	Exon skipping mutations result in ligand independent signaling. Amplification results in an increase in the number of copies of the MET gene and is often associated with cancer progression.	Capmatinib Tepotinib	(Blaquier & Recondo, 2022; Peng et al., 2021; Wolf et al., 2020)

Outside of targeted therapies, genetic variants can also guide the selection of other treatments such as immunotherapy. An example of patients who receive minimal benefit from immune checkpoint inhibitors are patients whose cancers harbour EGFR mutations or ALK rearrangements. The mechanism underlying the poor immunotherapy response for patients with EGFR mutations or ALK rearrangements is unclear (Qi et al.,

2022; Sankar et al., 2021; Shi et al., 2022). On the contrary, patients with NSCLC whose tumours harbour mutation in the Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS) gene can benefit from immune checkpoint inhibitors (Watterson & Coelho, 2023). Patients who have KRAS mutant tumours and receive treatment with immune checkpoint inhibitors typically have longer overall survival (OS) than patients with KRAS wild-type who receive the same treatment (Song et al., 2020). Recent evidence suggests that tumour mutational burden (TMB), the number of mutations a tumour harbours, may have potential as a predictive indicator for the use of immunotherapies. A higher TMB should increase the probability of tumor neoantigen production and, therefore, the likelihood of immune recognition and tumor cell killing. Recognition of tumor neoantigens by host T cells is one of the critical factors predicting immunotherapy response (Marcus et al., 2021; Strickler et al., 2021). Many challenges remain, such as identifying a definitive cut-off point, before the use of TMB as a predictive biomarker is implemented in clinical practice (Galvano et al., 2021; Palmeri et al., 2022).

1.2.2 Synonymous mutations

Cancer research has primarily focused on non-synonymous mutations; this work has led to the development of resources such as Catalogue of Somatic Mutations in Cancer (COSMIC). Historically, driver mutations have been non-synonymous in nature, meaning that a mutation results in a change in the encoded protein sequence (Diederichs et al., 2016).

Recent evidence suggests that synonymous mutations – which occur in exons but do not result in a change in the encoded protein sequence (Oelschlaeger, 2024) – may have an oncogenic impact similar to the well-established driver mutation which all alter the protein sequence (Elliott & Larsson, 2021; Supek et al., 2014). These mutations were initially thought to be silent, as they were expected to have no effect on cell activity. Defining mutations as silent varies drastically across databases in part due to interpretation differences (George et al., 2008). Although no change occurs in the protein form, synonymous mutations can influence the function of the gene. For example these mutations can affect the speed and accuracy of mRNA translation, folding, and splicing (Sauna & Kimchi-Sarfaty, 2011). Synonymous variants can lead to changes in epigenetic modifications. When a synonymous variant alters a nucleotide sequence, it can create or remove a site that DNA methyltransferases recognize, resulting in a different methylation pattern (Oelschlaeger, 2024). Roughly 17% of mutational effects reported by different laboratories carry contradictory clinical significance. This is to say that labels of pathogenicity are not fixed, switching from benign to disease based on what condition is being referenced and as evidence accumulates (Zeng & Bromberg, 2019). Furthermore, in the realm of cancer care “silent” mutations hold predictive value over cancer classification and prognosis (Gutman et al., 2021).

The study conducted by Fran Supek and colleagues in 2014 (Supek et al., 2014) marked a significant milestone in cancer research by providing the first comprehensive screening for synonymous mutations across various cancer types. Their findings revealed

a 23-30% enrichment of synonymous mutations in known missense activated oncogenes compared to non-oncogenic matched control genes. These control gene sets shared similarities with oncogenes in terms of genomic characteristics such as higher expression levels, increased heterochromatin content, elevated mutation rates, and later replication times.

Supek et al also estimated that 6-8% of all driver mutations attributed to single nucleotide changes in oncogenes are synonymous. This estimate was determined by dividing the observed number of synonymous mutations by the enrichment ratio, calculated through comparisons with control gene sets. The excess of observed mutations over the expected were considered possible driver mutations.

Synonymous mutations can modify the effects of known clinically significant oncogenes such as KRAS. In KRAS, synonymous mutations may have an impact on factors such as increased protein expression and drug resistance (Kobayashi et al., 2022; Sharma et al., 2019; Waters et al., 2016). In clinical practice, synonymous mutations are excluded from mutational screening analysis as they are less easily interpreted and typically do not yet have applicable targeted treatments (Haimovich, 2011).

1.3 Mutational profiling & Next-Generation Sequencing

Over 50% of patients with adenocarcinoma, the most common form of NSCLC, possess a driver mutation (Kris et al., 2014). Consequently, the current standard of care in Canada has shifted toward using mutational screening to provide personalized therapies

and treatments (Cheema et al., 2020). Clinicians now routinely screen for established core actionable mutations (e.g., EGFR, ALK, KRAS, MET, RET, ROS1, BRAF, NTRK) as part of the standard care for advanced lung cancer. It is important to note that none of these are synonymous mutations. Despite having established which mutations to screen for, the methods of profiling remain the choice of the laboratory (Ionescu et al., 2022; Melosky et al., 2018).

Next-generation sequencing (NGS) has become the standard for identifying genomic alterations such as driver and actionable mutations in NSCLC. There are two main benefits of NGS compared to traditional Sanger sequencing: number of genes sequenced and time. Unlike Sanger sequencing, NGS can be used to sequence multiple genes simultaneously, in a process known as massively parallel sequencing. The time required to perform NGS is also shorter because synthesis and sequencing occur in the same step (Mardis, 2013).

Similar to Sanger sequencing, NGS typically relies on sequencing by synthesis chemistry. For each deoxynucleotide triphosphate (dNTP) a fluorescently labeled reversible terminator is added and subsequently imaged. Each nucleotide (A, T, G, C) is associated with a fluorescent terminator that emits a unique wavelength, allowing the sequencer to identify which dNTP is present. This leads to a human-readable electronic output file known as a Fast Quality Score (FASTQ) file. The FASTQ file lists the sequence along with a quality score, which is the confidence the base pair identified in the read is correct. These files are sometimes referred to as the raw sequence data. In a typical

variant analysis pipeline, the sequence is then aligned to a reference genome such as the Genome Reference Consortium Human Build (GRCh) 38. This alignment process produces a Binary Alignment MAP (BAM) file on which routine quality control is performed by the in-house NGS software.

Once a BAM file is created variant calling is performed. Variant calling can be completed in a variety of ways and varies by program, pipeline, and company (Kamps et al., 2017; Koboldt, 2020; Larson et al., 2023; Zhong et al., 2021). Variant calling produces a Variant Call Format (VCF) file which contains information about variants at different positions in the sequence. A variety of genetic alterations such as single nucleotide variants, insertions, deletions, copy number alterations, structural variants, gene fusions, and chromosomal rearrangements are generated. This data is incredibly information dense and as such, the results of NGS can be complex and time consuming to utilize (Lazzari et al., 2020). Conversely, this provides a wealth of data which could be mined and re-annotated to identify additionally useful variants or variants that may be more prevalent in certain subpopulations.

1.4 Annotation of Variant Call Format files

Variant annotation, the process of assigning functional information to DNA and RNA variants, is dependent on the tools used. The two major producers of commercially available NGS platforms (ThermoFisher Scientific and Illumina) both offer an in-house software to indicate variants of clinical significance (Quail et al., 2012). A variety of

additional variant annotation tools are available, each using different algorithms and pulling from varying collections of databases.

Ensembl's Variant Effect Predictor (VEP) is a tool that has been developed for the analysis, annotation, and prioritization of genomic variants in both coding and non-coding regions. Unlike the in-house software used for clinical analysis, VEP produces and output with all variants annotated. VEP is open source, free to use, regularly updated, offers plugins for additional annotations, and compatible with many other tools used with VCFs. By compiling current knowledge relating to a variant and using predictive algorithms to evaluate the consequence of the variant, VEP is able to predict the effects of variants on genes, transcripts, regulatory regions, and proteins (Hunt et al., 2022; McLaren et al., 2016). With respect to nomenclature standards and clinical integrity, VEP is known to produce some of the most accurate variant annotations compared to other variant annotation tools (Tuteja et al., 2022).

1.5 Aim of study

Mutational profiling in NSCLC is a growing field with clinical, economic and resource implications (Forsythe et al., 2020; Kuang et al., 2022). Previous studies looking at mutational panels in lung cancer have primarily focused on already established driver and actionable mutations and response to a variety of treatments. The prevalence and possible clinical implication of synonymous mutations in lung cancer and cancer more broadly is under-investigated (Sharma et al., 2019).

Here I fill some of these gaps by performing a comprehensive analysis of clinical lung cancer NGS data at an Atlantic Canadian centre, including synonymous mutations, and their relation to clinical outcomes. Recommendations from an expert committee selected by Lung Cancer Canada advocate for the standardization of molecular screening across Canada (Melosky et al., 2018). This research supports these recommendations by adding to the body of research on the clinical trade-offs surrounding expanded mutation panels.

1.5.1 Objectives

- I. Identify and annotate variants that are not reported clinically focusing on predicted synonymous variants. Annotate variants with a variety of predictors to determine possible pathogenic significance and predicted biology.
- II. Explore possible clinical and prognostic impact for identified variants of interest by utilizing clinical data when possible.
- III. Describe the clinical (including reported variants) and demographic characteristics of patients with non-small cell lung cancer who underwent genomic sequencing at the Saint John Regional Hospital. Compare the expanded gene panel currently used to what theoretically would have been reported had the smaller previous panel been used.
- IV. Establish a bioinformatics pipeline at the SJRH that can be used in future analysis of variants that are not clinically reported in NSCLC genomic sequencing.

2.0 Methods

2.1 Study design & population

The study is a retrospective case series using information previously collected for clinical purposes. Participants included in the study had to have received a diagnosis of NSCLC and had a 52-gene next-generation sequencing panel performed at the Saint John Regional Hospital (SJRH) from January 2019 to January 2023. Patient data sources included electronic medical charts, pathology reports, and raw genomic data from NGS. Collection of clinical variables was completed at the end of August 2023. For this study, I obtained Research Ethics Board approval from both Horizon Health Network and the University of New Brunswick. The approval included a waiver of consent according to Articles 5, 3, and 12.3 of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (*Tri-Council Policy Statement, 2022*) and Horizon Health Policies for this to occur. As a retrospective study, patients' confidential information was not re-identified. The data are kept on a secure hospital sever which is password protected. Access to identifying patient information is limited to the research team. Prior to statistical analysis, all data was deidentified. Data will be retained for 7 years after the study. An overview of the methods and analysis is depicted in the schematic (Figure 2.1).

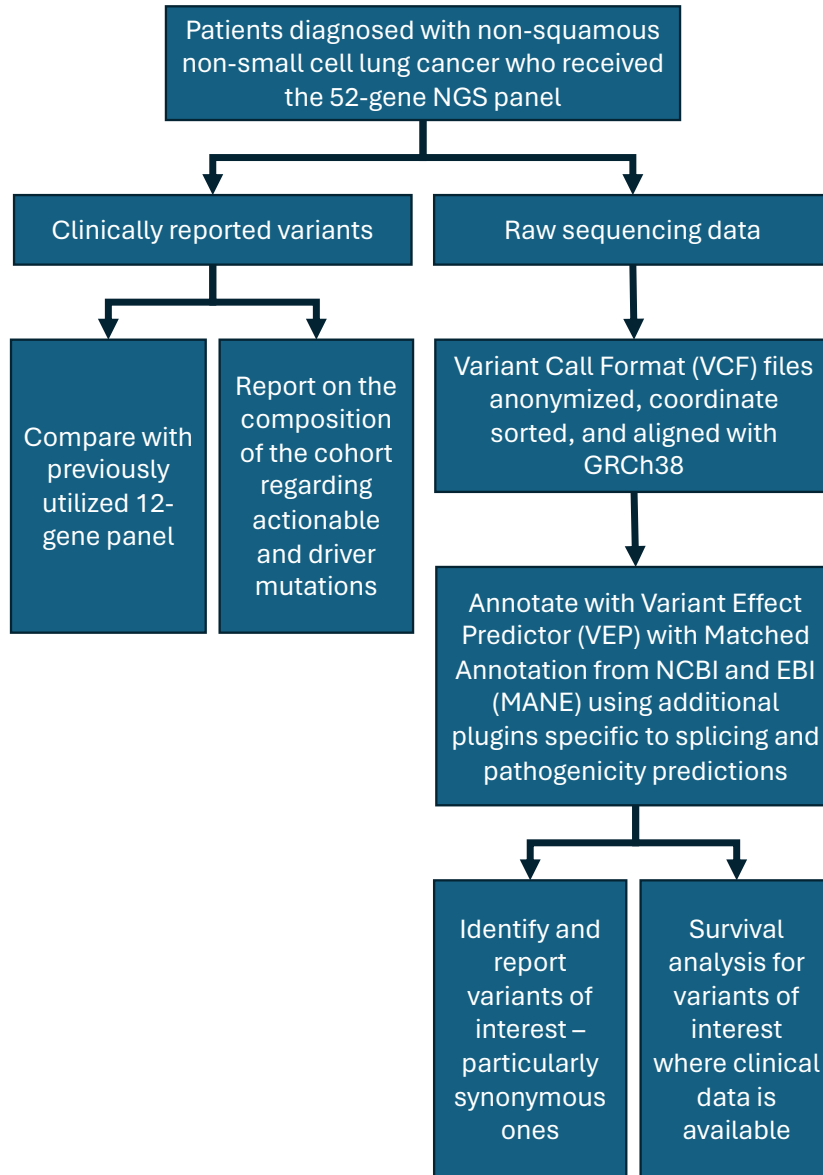


Figure 2.1: Schematic indicating workflow of methods and analysis.

2.2 Clinical & demographic variables

Clinical variables were collected from patient electronic medical charts. Interzone access to electronic medical charts for other Horizon Network zones was obtained and patients who received additional care outside of the Saint John catchment area were also

captured. The clinical and demographic variables obtained from medical records were as follows: sex, location of care, date of birth, cancer diagnosis and TNM staging, variants detected from gene panel, treatment received, last date known alive or date of death, date of diagnosis, date genetic panel received, the date of first cancer progression following diagnosis (if applicable), and smoking status.

Smoking status was divided into three categories: never-smoker, former smoker, and current smoker. Never smoker describes someone who had never smoked or smoked less than 100 cigarettes in their life. Former smokers describe those who smoked 100 or more cigarettes in their lifetime but had stopped smoking at least one month or longer before their diagnosis. Current smokers were smoking upon diagnosis or stopped less than one month before their diagnosis (Ban et al., 2020; Gemine et al., 2019; Lee et al., 2014).

Length of progression-free and overall survival was measured in days. Time zero was defined as the date that tissue was first obtained which led to a diagnosis of NSCLC. The end point for OS was the date of death. The end point for PFS was the date the patient experienced either disease progression or death (whichever was first documented). Health records and publicly available obituaries were reviewed for each patient to ascertain their mortality status and the date of their passing if applicable. The proportion of patients deceased is likely an underestimate as I could not unequivocally determine mortality status for all patients (ex. no obituary is available) (Vena et al., 1987). For those patients who had not yet reached the PFS or OS endpoint, data were censored at the date last known to be alive as determined by the patient's last documented visit to the hospital.

Progression was determined from diagnostic imaging reports, and as documented in the treating physician's notes.

Descriptive statistics for the cohort were determined using Prism (GraphPad Software, 2024). Categorical variables collected were as follows: sex, location of care, cancer diagnosis and staging, variants detected from gene panel, treatment received, and smoking status. Continuous variables included age at diagnosis, length of OS, and length of PFS. Only patients with accessible medical records could be included in the descriptive statistics for the cohort.

2. 3 Genomic Analysis

2.3.1 Clinical analysis with the IonTorrent platform

At the SJRH, non-squamous NSCLC biopsies are reflexively subjected to NGS testing before initiation of first-line therapies, as per the guidelines from the National Comprehensive Cancer Network (Ettinger et al., 2022). All genetic data used were initially sequenced and processed using the ThermoFisher IonTorrent Oncomine Focus Assay, an NGS platform.

Thermofisher's IonTorrent Oncomine Focus Assay covers 52 genes related to cancer (Table 2.3.1). Only 10ng of either DNA or RNA is needed for analysis, lending itself to the analysis of traditionally difficult samples such as fine-needle aspiration biopsies. The genes are categorized by somatic alteration type into four categories: hotspot genes,

focal copy number variant (CNV) gains, full coding sequence for deletion mutations, and fusion drivers.

Genomic sequencing was completed by technicians at the SJRH as part of the reflexive clinical work-up. To prepare samples to be sequenced DNA and RNA must first be extracted and quantified by fluorometer. RNA then must be reverse transcribed to cDNA. Prior to sample analysis libraries must be prepared. Three pools are amplified using polymerase chain reaction (PCR), two for DNA and one for RNA. Both DNA and RNA are then treated with the Fragmentase Universal Primer Assay (FuPA) enzyme to digest the primers, ligated with IonXpress barcode oligonucleotides, purified, and quantified using the Ion Library Quantification Kit. Ligation of IonXpress barcode oligonucleotides allows for multiplexed sequencing analysis by pooling multiple libraries prior to PCR emulsion, the step that amplifies DNA and cDNA.

Table 2.3.1: 52 genes sequenced in the Oncomine Focus Assay. Genes with asterisks are those that the smaller 12-gene panel could have identified.

Hotspot genes				Copy number genes		Gene fusions	
AKT1*	ERBB3	IDH2	MTOR	AKT1	FGFR2	ABL1	FGFR2
ALK	ERBB4	JAK1	NRAS*	ALK	FGFR3	AKT3	FGFR3
AR	ESR1	JAK2	PDGFRA	AR	FGFR4	ALK	MET
BRAF*	FGFR2	JAK3	PIK3CA	BRAF	KIT	AXL	NTRK1
CDK4	FGFR3	KIT	RAF1	CCND1	KRAS	BRAF	NTRK2
CTNNB1	GNA11	KRAS*	RET	CDK4	MET	EGFR	NTRK3
DDR2	GNAQ	MAP2K1	ROS1	CDK6	MYC	ERBB2*	PDGFRA
EGFR*	HRAS	MAP2K2	SMO	EGFR	MYCN	ERG	PPARG
ERBB2	IDH1	MET		ERBB2	PDGFRA	ETV1	RAF1
				FGFR1	PIK3CA	ETV4	RET
						ETV5	ROS1
						FGFR1	

IonTorrent sequencing has two major differences compared to other NGS platforms. IonTorrent uses proton-based sequencing. Instead of measuring fluorescence this type of sequencing measures the direct release of H⁺ from the incorporation of individual bases by DNA polymerase. Each type of nucleotide (A, G, C, T) is added one by one and pH change – if any – is measured to determine how many bases were added. Additionally, unlike NGS assays which sequence only genomic DNA, the Oncomine Focus Assay detects targeted gene fusions by sequencing cDNA converted directly from specifically targeted RNA transcripts (Lih et al., 2017). This sequencing produces a raw sequencing, BAM, and VCF file.

2.3.2 Pre-processing for analysis

I completed none of the sequencing, but all the analysis and pre-processing of the sequencing data was done by me, including all bioinformatics scripts which were written in Python. All scripts are accessible on GitHub (<https://github.com/kathleenvarty/Lung-cancer-mutation-profiling.git>). I obtained the VCFs created by the NGS platform from the clinical server. The NGS data is routinely kept on file in case of patient relapse, at which point additional actionable mutations and targeted treatments may need to be identified to better patient treatment. The NGS data is also kept or archived after a patient has died. The VCFs collected from the clinical server were the starting point for the analysis, acting as the initial input.

To prepare the initial VCF files for analysis, I performed a series of pre-processing steps. I de-identified all genomic data by assigning a randomly generated 6-digit identifier to each patient. For this study, I removed the RNA sequencing data. I then coordinate-sorted the VCFs from 1 through Y as required by the annotation programs. Next, I lifted over the files from reference genome GRCh37, against which the VCF data were originally aligned, with the more recent reference genome GRCh38 using CrossMap (Zhao et al., 2014). I decomposed and normalized the VCFs using bcftools (Danecek et al., 2021). Normalization standardizes and simplifies variant representation for easier analysis and comparison. Decomposition, a component of normalization, involves separating multi-allelic variants into individual records.

The VCFs contained all variants identified, regardless of their quality. Ion Torrent by default added an annotation denoting the quality of the variant. All variants with an

annotation of “PASS” were retained. The “PASS” annotation is indicative of adequate quality of the variant to be subjected for analysis with standard variant finding parameters provided in the Ion Reporter™ Software (Vestergaard et al., 2021). The exact parameters that define a PASS from the system are not publicly available. Finally, variants with a variant allele frequency (VAF) of less than or equal to 2.5% were filtered out to remove suspected artifacts. VAF is the percentage of sequence reads observed matching a specific DNA variant divided by the overall coverage at that locus. VAF is used as a surrogate measure of the proportion of DNA molecules in the sample that carry the variant of interest. This is possible because NGS provides a near random sample (Melosky et al., 2018; Strom, 2016). I chose the threshold of 2.5% because it is the same threshold used for clinical reporting.

2.3.3 Variant Effect Predictor and additional plug-ins

I added additional annotations to each variant transcript using Ensembl’s VEP. The annotations come from numerous sources, and VEP can be modified to pull from additional databases and scoring systems (McLaren et al., 2016). I used the Docker software platform to run VEP to proactively avoid issues such as reproducibility and varying dependencies across tools. I ran VEP using the ‘--everything’ flag and set the output to tab-separated values (tsv) files. To gain additional annotations, I utilized four plugins: dbNSFP, dbSNV, SpliceRegion, and SpliceAI. For pathogenicity predictions, I used the web interfaces CScape and TRaP. All plugins utilized ran the most recent versions.

The Database for Non-synonymous Single Nucleotide Polymorphisms' (SNPs') Functional Predictions (dbNSFP) compiles predictions and annotations from different algorithms providing deleterious prediction and functional annotation for all potential non-synonymous and splice-site SNPs. Version 4.6 was utilized and draws from over 30 algorithms including: AlphaMissense, FATHMM, CADD, EVE, and PrimateAI (Liu et al., 2011, 2020). The Database for Single Nucleotide Variants within Splicing Consensus Regions (dbscSNV) is attached to dbNSFP and provides a whole genome level resource for identifying splice-altering single nucleotide variants (SNVs) discovered from large-scale sequencing studies. The database is modeled off the existing predictive tools Position Weight Matrix and MaxEnt Scan, which were found to have the best predictive performance compared to similar tools (Jian et al., 2014).

Three annotation sources were of particular interest for predicting pathogenicity. Combined Annotation Dependent Depletion (CADD) (v1.7) score – which is included in the dbNSFP – outputs a score incorporating diverse genomic annotations, including conservation scores, functional predictions, and allele frequencies, to evaluate the potential pathogenicity of genetic variants (Kircher et al., 2014; Schubach et al., 2024). Another pathogenicity predictor I utilized was the Transcript-inferred Pathogenicity (TraP) (v3.0) score. The TraP score was developed to predict a SNV's ability to cause disease by impacting the final mRNA transcript (Gelfman et al., 2017). The third pathogenicity predictor I utilized was CScape which predicts the likelihood that somatic point mutations are cancer drivers in both coding and non-coding regions. CScape pulls from genomic

databases, conservation data, protein databases, and predictive tools to provide a singular score (Rogers et al., 2017). This is the only pathogenicity predictor utilized that was specific to cancer.

For splicing predictions, I utilized two annotations primarily. SpliceRegion by Ensembl provided me with more specific predictions of splicing effects. SpliceRegion adds three additional terms: splice donor 5th base variant, splice donor region variant, and splice polypyrimidine tract variant. To predict possible impact on splicing I used SpliceAI. Developed by Illumina in 2019, Splice AI is an open-source deep learning algorithm that assesses how likely a genetic variant is to disrupt the normal RNA splicing process and cause errors in gene expression. The most recent version (v1.3) was utilized (Jaganathan et al., 2019).

2.3.4 Post-annotation processing

To ease filtering, I used only Matched Annotation from NCBI and EMBL-EBI (MANE) to select transcripts for annotation and to predict variant effects and impact. After filtering using MANE, two annotation outputs remained for identical transcripts, one from NCBI and another from EMBL-EBI. I merged the annotation outputs and retained a single transcript. Due to alternative splicing, a single genomic location can produce multiple mRNA transcripts. MANE select transcripts are predicted to be the most representative of the biology at that locus (Morales et al., 2022).

To reduce noise, I filtered out common single nucleotide polymorphisms (SNPs). I did this by removing variants that had an allele frequency greater than 0.5% in the Genome Aggregation Database (gnomAD) which provides a catalogue of genetic variation in humans across different populations (Chen et al., 2024).

I examined the VAF for each variant of interest across all samples that contained the variant. Focusing on somatic mutations, I devoted additional efforts to excluding germline variants. A VAF around 50% or 100% indicates a germline variant. I flagged and removed variants if most samples had a VAF of 50% or 100% \pm 5%. However, I retained variants if they had an annotation in COSMIC. Given the exploratory nature of the study, I conservatively filtered out germline variants to avoid the risk of inadvertently excluding potentially pathogenic somatic variants.

It was not feasible to manually inspect all identified variants to verify that they were not artifacts or germline polymorphisms. Any variant appearing in \geq 20% of the cohort or found to have a significant impact on survival was reviewed using the Broad Institute's Integrative Genomics Viewer (IGV) to identify any signatures suspected to be artifacts or polymorphisms (Robinson et al., 2011). I also reviewed any variant with significant prediction of pathogenicity or splicing impact.

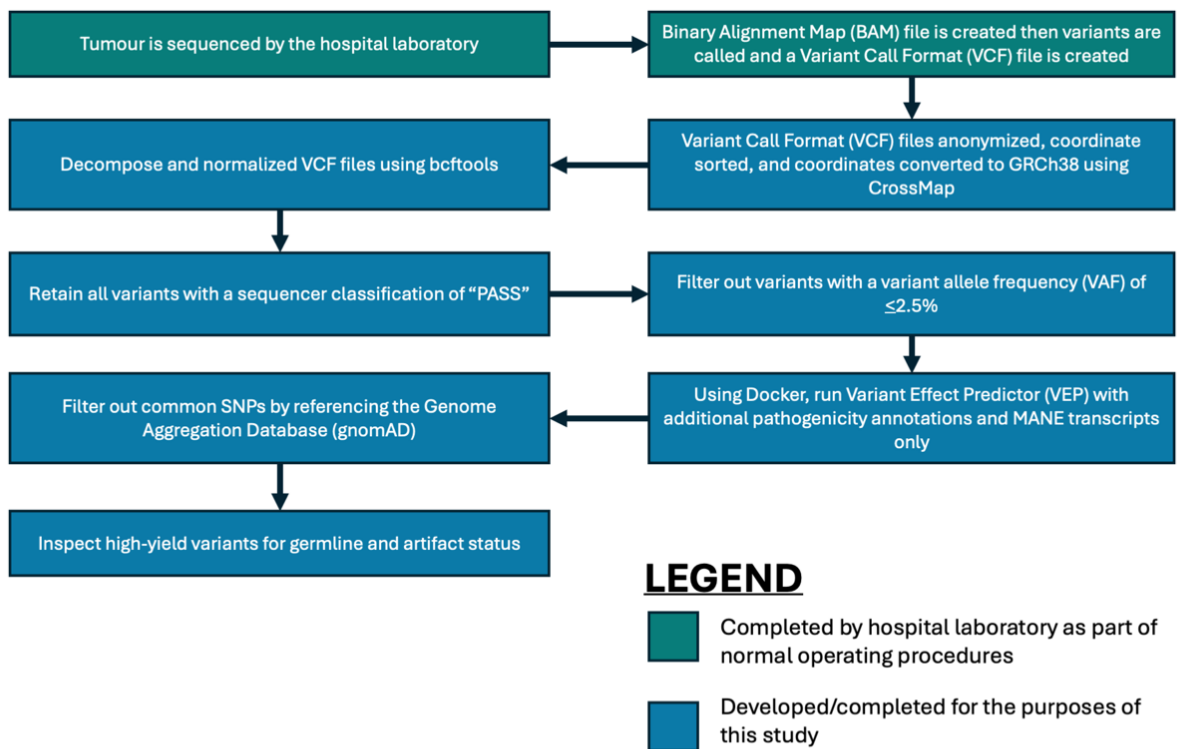


Figure 2.3.4: Visual depiction of the bioinformatics methods used.

2.4 Survival analyses

Cox regression, also known as Cox proportional hazards regression analysis, is a statistical model used to analyze survival data. Unlike a Kaplan-Meier analysis it takes into account multiple variables that may contribute to cancer progression or death. Kaplan-Meier analyses are used to estimate survival probabilities and compare survival curves, but they do not assess the association between multiple variables and survival rate, which is possible with the Cox regression model.

All survival analyses were completed using the statistical software R in conjunction with the RStudio interface, which was used for the Kaplan-Meier analyses (Posit team,

2024; R Core Team, 2023). I assessed the following covariates for their impact on survival: age, sex, TNM stage, presence of EGFR or KRAS clinical mutation, and smoking history. For categorical variables I plotted Kaplan-Meier curves and performed the log-rank test. For categorical variables that were not binary (i.e., TNM stage), I performed the log-rank test using a pairwise comparison with the Bonferroni correction to account for multiple comparisons. An two-sided p-value of 0.05 or less was used as the threshold for statistical significance. For continuous variables such as age, I performed a univariable Cox regression. Variables that were significant were tested to determine if they violated the proportional hazards assumption of the Cox regression model by plotting and visually inspecting the Schoenfeld residuals. If the covariate violated the proportional hazards assumption, I did not include it in the model.

Determination of an adequate sample size for Cox regression suggests there should be a minimum of 10 events per variable (EPV). This rule of thumb was derived from simulation studies. Recent evidence suggests that the EPV should be > 20 , so I opted to adhere to this more stringent standard (Ogundimu et al., 2016). The covariates that I found to be significant and appropriate to fit to the model were smoking history for both overall and PFS in addition to age for OS. To calculate the EPV the number of events (168 for OS 219 for PFS) is divided by the number of predictor variables in the analysis (2); this left us with an EPV of 84 and 109.5 respectively indicating I did have a large enough sample size to perform a Cox regression with the variants I found to be significant.

I analyzed variants of interest by plotting Kaplan-Meier curves and comparing the curves with the log-rank test. The p-value was not adjusted for multiple comparisons due to the exploratory nature of the study. Post-hoc power was also calculated for all log-rank tests using a two-tailed test and a significance level of 0.05. The variants of interest were also fitted to a Cox regression model with appropriate covariates depending on which survival outcome was analyzed.

3.0 Results

3.1 Demographics & clinical characteristics of cohort

A total of 353 patients were included in the study. I analyzed the genomic data for all patients however only 344 were included in the demographic and survival analysis as nine patients either had incomplete or inaccessible chart information. Age at diagnosis ranged from 28 to 89 years, the average age of patients at the time of diagnosis was 68 years. Females were marginally more prevalent in the cohort (52.90%). The three most common subtypes of NSCLC were adenocarcinoma (72.97%), squamous cell carcinoma (17.73%), and poorly differentiated NSCLC (6.69%) (Figure 3.1.1).

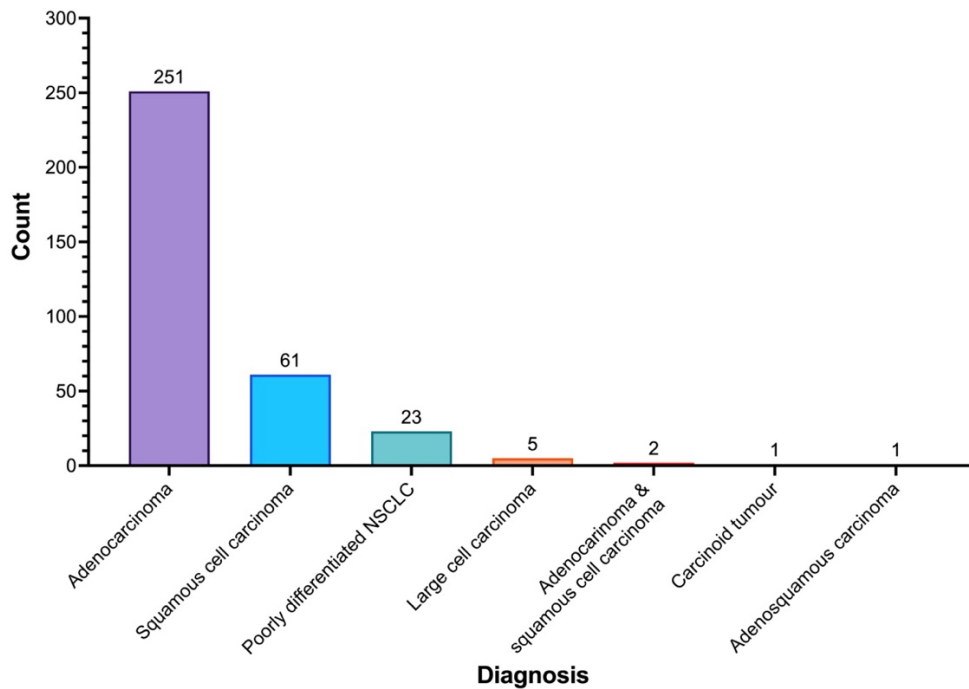


Figure 3.1.1: Subtypes of NSCLC within cohort.

Most patients had a TNM stage at diagnosis of I or IV and four patients not staged. Radiation, surgical resection, and chemotherapy were the three most common treatments received (Figure 3.1.2). Most patients were either former smokers (49.56%) or current smokers (37.90%). One patient was of an unknown smoking status and 12.54% of patients never smoked.

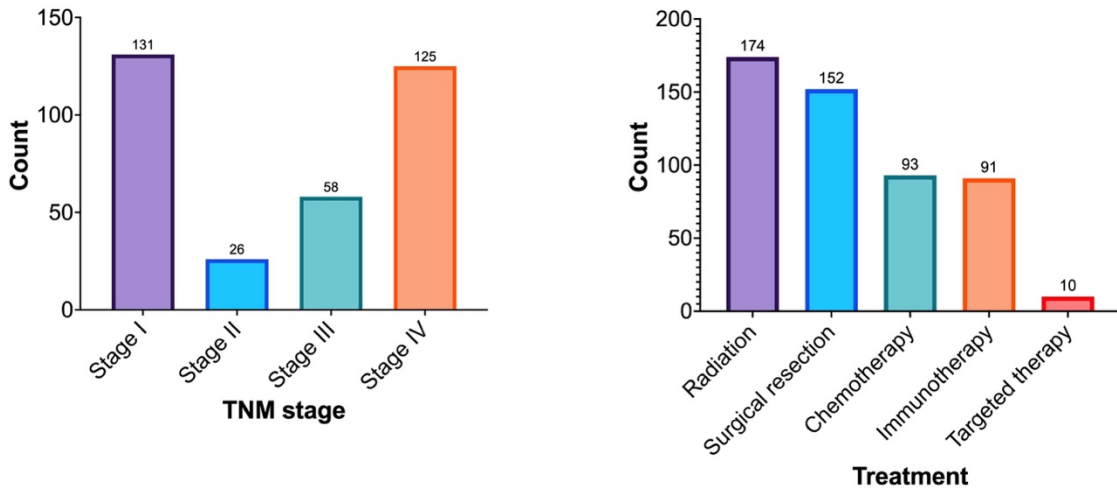


Figure 3.1.2: TNM staging assigned at diagnosis and treatments received within cohort.

At the time of data collection, 63.7% (N=219) of patients had reached the PFS endpoint. Of these 43.9% (N=155) experienced progression of their disease since diagnosis and 48.7% (N=168) of patients were deceased. The median OS and PFS were 906 and 727 days respectively (Figure 3.1.3).

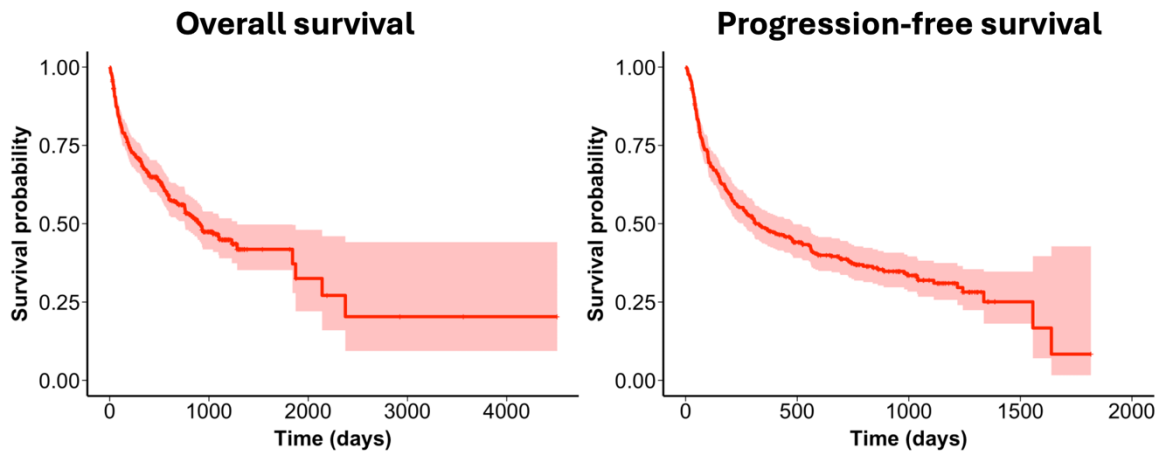


Figure 3.1.3: Survival curves for overall and progression-free survival. Shaded area indicates 96% confidence interval.

3.2 Clinical reporting of variants

Of the 353 patients included in the genomics analysis, more than half had clinically reported variants (62.32%) with an average of one each. The most common clinically reported DNA variants were found in KRAS, PIK3CA, and BRAF. For RNA variants MET exon 14 skipping, AR amplification, and EGFR amplification were most prevalent (Figure 3.2.1).

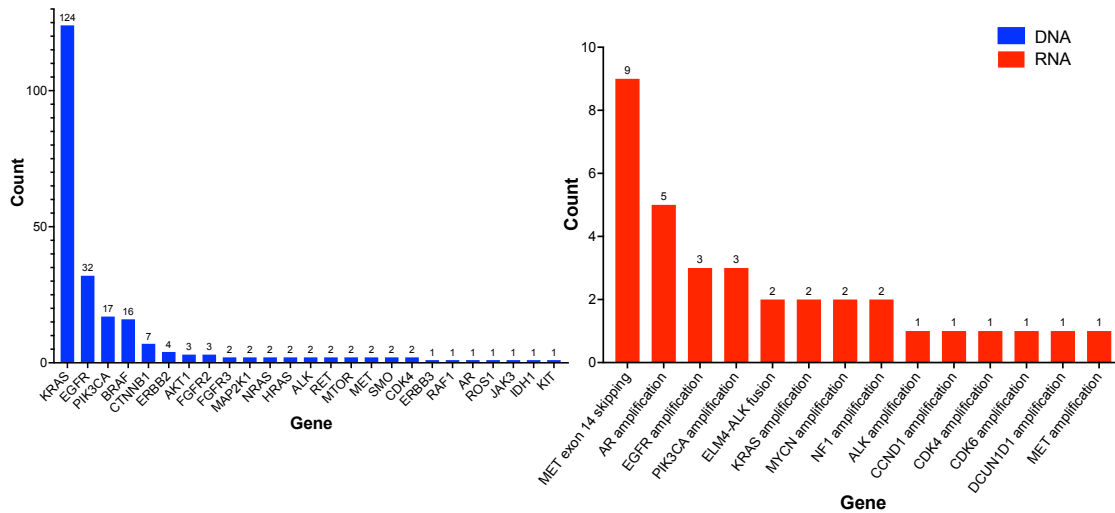


Figure 3.2.1: Clinically reported variants in cohort that were reported using the 52-gene panel.

Interestingly, two patients had co-occurring KRAS and EGFR mutations. Although not unheard of this is a relatively rare phenomenon. One patient had an EGFR 18 variant co-occurring with a KRAS S122F variant and the other patient had an EGFR21 variant co-occurring with a KRAS G12V variant (Figure 3.2.2). The patient with the EGFR 18 variant received immunotherapy. Neither patient received a targeted therapy. One patient had an interesting co-occurrence of mutations presenting with MYCN, ALK, PIK3CA, and DCUN1D1 amplifications.

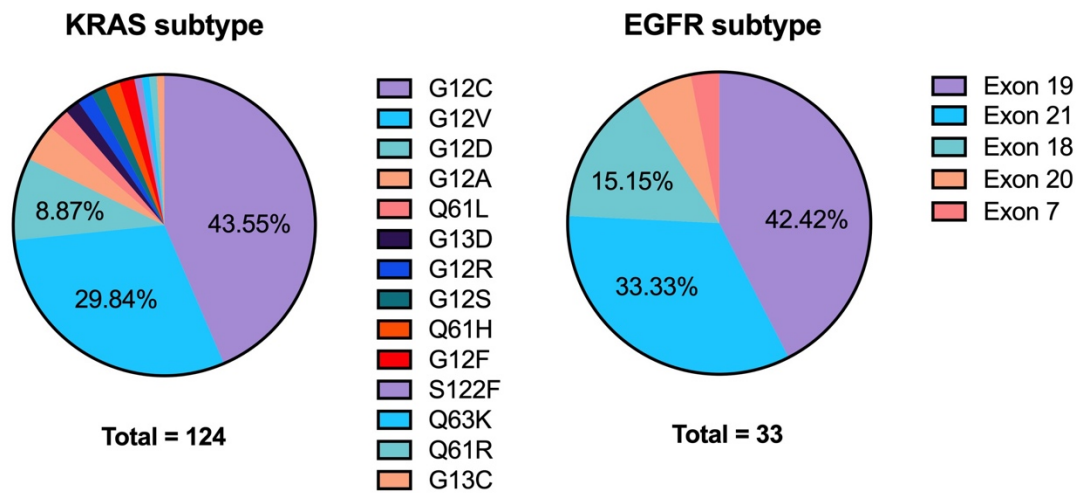


Figure 3.2.2: Breakdown of clinically reported mutation in cohort found in KRAS and EGFR.

Five mutations occurred exclusively within patients who never smoked, these were MAP2K1, ALK, RET, EML4-ALK fusions, and MET amplifications. A small proportion of patients who had never smoked (5.6%) had KRAS mutations. Approximately a third (33.33%) of EGFR positive patients had never smoked. Patients with a smoking history had a lower incidence rate of EGFR mutations.

3.3 Comparison of expanded gene panel to predecessor

A total of 30 variants were detected using the 52-gene panel that would not have been detected using the 12-gene panel. An additional 3 variants were captured by the 52-gene panel that would have been captured by the 12-gene panel but would not have been reported for pathogenicity (Table 3.3.1).

Table 3.3.1: Variants captured with the 52-gene panel that would not have been captured with the 12-gene panel or not assessed for pathogenicity and clinically reported. Those that were captured but not reported are underlined. Italicised variants have known targeted therapies.

Variant	Occurrence in cohort	Variant	Occurrence in cohort
<i>MET exon 14 skipping</i>	9	AR	1
CTNNB1	7	<i>ROS1</i>	1
AR amplification	5	JAK3	1
PIK3CA amplification	3	IDH1	1
RET	2	KIT	1
MTOR	2	FGFR1 amplification	1
HRAS	2	<i>MET amplification</i>	1
SMO	2	CDK6 amplification	1
KRAS amplification	2	CCND1 amplification	1
<i>EML4-ALK fusion</i>	2	CDK4 amplification	1
MYCN amplification	2	ALK amplification	1
NF1 amplification	2	DCUN1D1 amplification	1
CDK4	1	MAP2K1 exon 3	1
EGFR amplification	1	<u>FGFR2 (exon 8)</u>	3
EGFR exon 7	1	<u>FGFR3 (exon 7)</u>	2
ERBB3	1	<u>MAP2K1 (exon 2)</u>	1
RAF1	1		

Two patients received targeted therapies for mutations that were found with the 52-gene panel that either would not have been found with the 12-gene panel or would

not have been reported. One patient received Crizotinib as part of their treatment for a MET amplified mutated NSCLC. The other patient received Alectinib, an ALK specific targeted therapy. The patient with the ALK mutation may have been identified by ALK immunohistochemistry had the patient been treated in the era of the 12-gene panel however this would have been an additional test.

3.4 Novel variants of interest

Utilizing the VCF files and annotating with VEP, a total of 372 variants were identified in the cohort. After removing variants with well-established clinical significance in NSCLC a total of 311 unique variants were identified which occurred 366 times in the cohort. These 311 unique variants compromised the focus of the study. It was not feasible to manually verify all variants therefore I prioritized files of patients with predicted pathogenic variants or prevalent variants. In the end, 246 variants were manually verified and determined not to be sequencing artifacts; I contend this is a reasonable level of scrutiny and all specific variants reported were included in the 246 that were manually verified. The genes with the greatest number of unique variants identified were AR, EGFR, and FGFR2 (Figure 3.4.1).

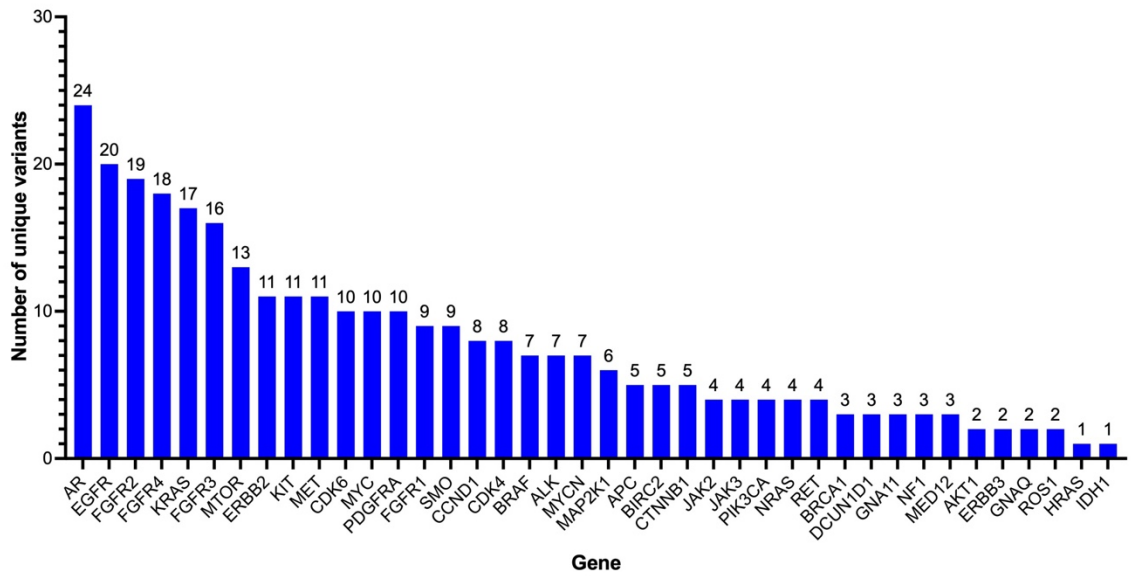


Figure 3.4.1: Number of unique variants identified in each gene.

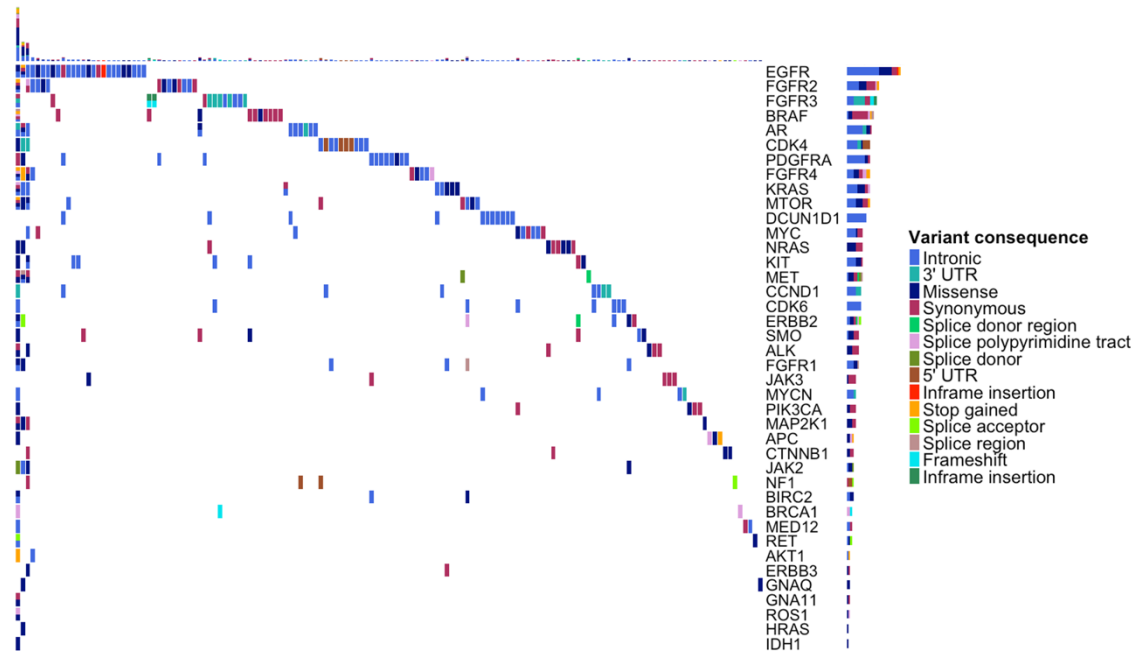
Although the greatest number of variants occurred in the AR gene, the majority of variants were found in two patients – both of whom had an overall high TMB – skewing the results. Across the cohort, the most prevalent individual mutation was a synonymous SNV in BRAF (Table 3.4.2); followed by a deletion and insertion-deletion in EGFR, both of which were intronic.

Table 3.4.2: Ten most prevalent novel variants identified in cohort.

Gene	Location	Variant	Class	Consequence	Occurrences in cohort
BRAF	Chr7:140801453	A>G	SNV	Synonymous	8
EGFR	Chr7:55131288-55131289	TT>_	Deletion	Intronic	8
EGFR	Chr7:55131286-55131290	TCTTT>CTTTC	Insertion-deletion	Intronic	6
DCUN1D1	Chr3:182955043-182955044	_>T	Insertion	Intronic	5
DCUN1D1	Chr3:182973452	G>A	SNV	Intronic	5
NRAS	Chr1:114713886	T>C	SNV	Synonymous	4
PDGFRA	Chr4:54257555	G>T	SNV	Intronic	4
CDK4	Chr12: 57751159	A>G	SNV	Intronic	3
CDK6	Chr7:92644005	C>G	SNV	Intronic	3
FGFR4	Chr10:121546542	T>C	SNV	Intronic	3

Of the 353 patients, 148 (41.9%) had at least one novel mutation. The average number of variants per patient was 2.46 ± 8.60 . The cohort exhibited considerable diversity, with most patients harbouring a single variant. Three patients had high tumour mutational burdens with 99, 31, and 25 novel variants identified. None of the novel variants of interest occurred exclusively in patients with no smoking history.

A



B

Gene	AR	EGFR	MET	FGFR1	FGFR2	FGFR3	FGFR4	MTOR
# of patients with >1 variant	4	4	3	2	2	2	2	2

Gene	ALK	APC	BIRC2	BRAF	CDK6	CTNNB1	ERBB2	GNA11
# of patients with >1 variant	1	1	1	1	1	1	1	1

Gene	KIT	KRAS	MYC	MYCN	RET	ROS2	SMO	
# of patients with >1 variant	1	1	1	1	1	1	1	

Figure 3.4.3: (A) OncoPrint displaying the distribution of variant consequences across patients and genes. Each column is a patient and each row a gene. The type of variant consequence is displayed but not the number of variants with that consequence in a gene per patient. (B) Number of patients with more than one novel variant by gene.

All variants had associated consequence terms annotated by VEP which are defined by sequence ontology – a set of terms and relationships used to describe the features and attributes of biological sequence in a structured and controlled manner (Eilbeck et al., 2005). Consequence terms are predictions of the effects that the variant may have on the genes function or the resulting protein. Intron variants are also labelled using consequence terms. The most prevalent consequences (excluding intronic) were missense, synonymous, and 3' UTR variants (Figure 3.4.3 & Table 3.4.4).

Table 3.4.4: Variants divided by consequence type. The number of unique variants in each gene is denoted in brackets.

Stop gained	Intron		Synonymous		Missense	
MTOR (4)	AR (19)	FGFR3 (3)	FGFR2 (6)	ERBB2	EGFR (6)	ERBB2 (2)
FGFR4 (3)	CDK6 (10)	MED12 (2)	ALK (4)	ERBB3	FGFR2 (6)	GNA11 (2)
AKT1	EGFR (10)	MET (2)	FGFR3 (4)	GNA11	KIT (5)	GNAQ (2)
APC	KRAS (10)	RET (2)	MTOR (4)	KIT	KRAS (5)	JAK2 (2)
BRAF	FGFR4 (8)	AKT1	SMO (4)	KRAS	FGFR1 (4)	PIK3CA (2)
EGFR	PDGFRA (7)	BRAF	CTNNB1 (3)	MED12	MAP2K1 (4)	BRAF
FGFR2	MYC (6)	JAK2	JAK3 (3)	NF1	MET (4)	CDK4
	MYCN (6)	MTOR	MYC (3)	NRAS	MTOR (4)	ERBB3
	CCND1	SMO	BRAF (2)		SMO (4)	HRAS
	ERBB2		EGFR (2)		ALK (3)	IDH1
	FGFR2 (5)		FGFR4 (2)		APC (3)	JAK3
	KIT (5)		PDGFRA (2)		FGFR4 (3)	MYC
	FGFR1 (4)		PIK3CA (2)		NRAS (3)	PDGFRA
	BIRC2 (3)		MAP2K1 (2)		AR (2)	RET
	CDK4 (3)		MET (2)		BIRC2 (2)	ROS1
	DCUN1D1 (3)		AR		CTNNB1 (2)	
Splice polypyrimidine tract		3' UTR	Splice Region	Frameshift	5' UTR	Inframe insertion
BRCA1 (2)	ERBB2	FGFR3 (7)	BRAF	BRCA1	CDK4 (2)	EGFR
FGFR4 (2)	FGFR2	CCND1 (3)	FGFR1	FGFR3	NF1	FGFR
APC	KRAS	AR (2)	MET			
BRAF	ROS1	CDK4 (2)				
		MYCN				
Splice donor region	Splice acceptor	Splice donor				
ERBB2	ERBB2	JAK2				
MET	NF1	MET				
	RET					

3.5 Synonymous variants

Synonymous variants were defined by consequence terms assigned by VEP. These consequence terms are sourced from the sequence ontology. VEP only defines synonymous variants as those that are in the coding region of the gene that result in no change in the amino acid sequence. Two synonymous variants were included in the ten most prevalent variants, occurring a combined twelve times within the cohort. There are no substantial ClinVar or COSMIC submissions specific to lung cancer for either the BRAF or NRAS variants. A total of 53 variants identified in the cohort were synonymous, occurring a combined 67 times (Appendix A - Table A1).

3.5.1 Synonymous variants with possible splicing impact

The only splicing predictor that consistently provided a prediction for all the synonymous variants was SpliceAI. Scores for SpliceAI range from 0-1 with values closer to one indicating a greater likelihood of having an impact on splicing. A threshold of 0.2 is recommended by the developers as a “permissive” threshold to retain variants with high sensitivity. A threshold of 0.5 is recommended for a balanced approach (Jaganathan et al., 2019).

SpliceAI provides four scores for each variant: donor gain indicates the likelihood of creating a new splice donor site, donor loss predicts the disruption of an existing donor site, acceptor gain suggests the creation of a new splice acceptor site, and acceptor loss assesses the disruption of an existing acceptor site.

Table 3.5.1: Synonymous variants in cohort with at least one SpliceAI score.

Gene	Variant	SpliceAI				Occurrence	Allele frequency/range
		Acceptor loss	Donor loss	Acceptor gain	Donor gain		
FGFR3	chr4:1804418 G>A	0.44	0.01	0	0	1	0.37302
MYC	chr8:127738274 C>T	0.12	0.19	0	0	1	0.626
FGFR3	chr4:1801857 C>T	0.14	0.01	0	0	1	0.242424
FGFR3	chr4:1799396 G>A	0.1	0	0	0	1	0.0959128
JAK3	chr19:17835201 C>T	0.08	0	0	0	2	0.237164 - 0.405108
MET	chr7:116763082 C>T	0.06	0.05	0	0	1	0.0544629
PDGFRA	chr4:54274846 G>C	0.06	0.02	0.01	0	1	0.41362
MED12	chrX:71129386 G>A	0	0	0	0.06	1	0.2755
MET	chr7:116763091 C>T	0.06	0	0	0	1	0.0565553
EGFR	chr7:55191841 G>A	0	0	0	0.05	2	0.047 - 0.488924
ALK	chr2:29213992 G>A	0	0	0	0.05	1	0.0440613
JAK3	chr19:17835147 A>G	0.04	0	0	0	1	0.320905
BRAF	chr7:140794401 T>C	0.03	0	0	0	1	0.0554562
NF1	chr17:31358501 G>A	0	0.03	0	0	1	0.0566038
SMO	chr7:129209384 C>A	0	0.01	0	0.02	1	0.0460526
ERBB2	chr17:39725096 C>T	0.02	0	0	0	1	0.356955
FGFR2	chr10:121515279 G>A	0	0.02	0	0	1	0.0861423
MTOR	chr1:11129809 C>A	0.02	0	0	0	1	0.0706215
GNA11	chr19:3115037 C>T	0	0.01	0.01	0.01	1	0.106818
FGFR2	chr10:121515246 G>A	0.01	0	0	0	1	0.0643821
FGFR3	chr4:1799444 C>G	0.01	0	0	0	1	0.0574074
FGFR4	chr5:177091038 G>A	0	0.01	0	0	1	0.049459
JAK3	chr19:17838052 C>T	0	0.01	0	0	1	0.5675
KIT	chr4:54727255 G>A	0.01	0	0	0	1	0.304
MAP2K1	chr15:66435123 G>A	0	0.01	0	0	1	0.0539773

A total of 25 synonymous variants with at least one Splice AI score were identified (Table 3.5.1). The synonymous variant most likely to impact splicing was a G to A substitution in FGFR3 with a SpliceAI acceptor loss score of 0.44. This indicates that the substitution may prevent the normal recognition and use of the acceptor site during splicing. This can lead to exon skipping or the use of alternative splicing sites.

The next most likely synonymous variant to impact splicing was a C to T substitution in MYC with a SpliceAI donor loss score of 0.19, slightly below the recommended threshold. Similarly to the acceptor, the donor loss suggests the variant may prevent the normal functioning of the donor site during splicing. These variations in splicing can lead to abnormal RNA transcripts which may potentially alter the structure and function of their associated proteins.

Neither the FGFR3 mutation nor the MYC mutation co-occurred with established clinical variants in their respective genes. Neither variant in FGFR3 has been reported in ClinVar for clinical significance.

3.5.2 Synonymous variants with predicted pathogenicity

CScape assigns a probability score to each SNV indicating the likelihood that the variant is cancer-associated and pathogenic. The score ranges from 0 to 1 with scores closer to one indicating a high likelihood that the variant is pathogenic. A score greater than 0.5 is labeled oncogenic but cautious classification thresholds are also available which for coding is 0.89 (Rogers et al., 2017). Five variants were identified as possibly oncogenic by CScape, all with low confidence (Table 3.5.2.1). Low confidence means that the score does not pass the cautious classification threshold of 0.89. The NRAS variant is part of the top ten most prevalent variants in the cohort.

Table 3.5.2.1: Synonymous variants with possible pathogenicity in cancer as defined by CScape score. Scores for additional predictive algorithms, occurrence, and allele frequency or range are also included.

Gene	Variant	Cscape			TraP Score	CADD Phred-like score	Occurrence	Allele frequency/range
		Score	Prediction	Confidence				
ALK	chr2:29213992 G>A	0.656541	Oncogenic	Low	0.131	11.45	1	0.0440613
FGFR2	chr10:121515246 G>A	0.641029	Oncogenic	Low	0.104	12.63	1	0.0643821
NF1	chr17:31358501 G>A	0.667882	Oncogenic	Low	0.104	8.931	1	0.0566038
NRAS	chr1:114713886 T>C	0.505196	Oncogenic	Low	0.009	12.59	4	0.0503597 - 0.0715686
PIK3CA	chr3:179204568 A>G	0.520249	Oncogenic	Low	0.027	12.35	1	0.162

The second algorithm utilized was the Transcript-inferred Pathogenicity (TraP) score. Like CScape, the TraP score ranges from 0-1 with higher numbers being more likely to cause disease. TraP uses a percentile scale to define which scores are damaging with separate scales for coding versus non-coding regions. The developers denote a score above the top 90th percentile as possibly damaging and above the 97.5th percentile as probably damaging (Table 3.5.2.2). All variants were coding, therefore the coding percentile scale was used to define a threshold. I chose a score of 0.100, slightly below the 75th percentile, for this project.

Table 3.5.2.2: Percentile scale for TRaP scores (Gelfman et al., 2017).

Percentile	10%	25%	50%	75%	90%	92.5%	95%	97.5%	99%	99.9%
TraPv3 score	0.004	0.015	0.049	0.166	0.221	0.256	0.307	0.416	0.676	0.981

Eleven variants were identified with a TraP score greater than 0.100. Two variants - one in SMO and the other in FGFR3 - had very high scores in the 0.400 range. All other variants had a TraP score in the low 0.100 range (Table 3.5.2.3).

Table 3.5.2.3: Synonymous variants with potential pathogenicity as denoted by TraP score. Scores for additional predictive algorithms, occurrence, and allele frequency or range are also included.

Gene	Variant	Cscape			TRaP Score	CADD Phred-like score	Occurrence	Allele frequency/range
		Score	Prediction	Confidence				
SMO	chr7:129209384 C>A	0.261754	Neutral	Low	0.429	12.79	1	0.0460526
FGFR3	chr4:1804418 G>A	0.148019	Neutral	Low	0.427	18.47	1	0.37302
MYC	chr8:127738274 C>T	0.159967	Neutral	Low	0.131	12.48	1	0.626
ALK	chr2:29213992 G>A	0.656541	Oncogenic	Low	0.131	11.45	1	0.0440613
ERBB3	chr12:56085090 C>G	0.239366	Neutral	Low	0.123	9.763	1	0.332666
EGFR	chr7:55191841 G>A	0.188506	Neutral	Low	0.114	2.405	2	0.047 - 0.488924
SMO	chr7:129210486 C>A	0.272064	Neutral	Low	0.105	5.765	1	0.234281
FGFR2	chr10:121515246 G>A	0.641029	Oncogenic	Low	0.104	12.63	1	0.0643821
NF1	chr17:31358501 G>A	0.667882	Oncogenic	Low	0.104	8.931	1	0.0566038
ALK	chr2:29220757 G>A	0.274993	Neutral	Low	0.103	3.457	1	0.635135
EGFR	chr7:55181331 G>A	0.158837	Neutral	Low	0.101	15.34	1	0.0503751

The Combined Annotation Dependent Depletion (CADD) score was the final prediction score used. Higher scores indicate a greater likelihood that the variant is functionally significant and potentially deleterious. The CADD score, which can be represented on a logarithmic scale, is commonly referred to as a CADD Phred-like score. A Phred score is the most common metric used to assess the accuracy of sequencing, it is a quality measure that estimates the probability that a base was called incorrectly. The CADD score is Phred-like because like a true Phred score it is also given on a negative log scale. The authors recommend a ranking system rather than a hard cut-off but an explanation of score is available. A CADD Phred-like score of 10 or greater indicates that the variant analyzed is predicted to be in the 10% most deleterious substitutions possible in the human genome (Schubach et al., 2024).

Table 3.5.2.4: Synonymous variants with a CADD Phred-like score > 10. Scores for additional predictive algorithms, occurrence, and allele frequency or range are also included.

Gene	Variant	Cscape			TraP Score	CADD Phred-like score	Occurrence	Allele frequency/range
		Score	Prediction	Confidence				
FGFR3	chr4:1804418 G>A	0.148019	Neutral	Low	0.427	18.47	1	0.37302
FGFR3	chr4:1801857 C>T	0.271777	Neutral	Low	0.068	16.03	1	0.242424
BRAF	chr7:140794401 T>C	0.203643	Neutral	Low	0.056	15.81	1	0.0554562
EGFR	chr7:55181331 G>A	0.158837	Neutral	Low	0.101	15.34	1	0.0503751
ERBB2	chr17:39725096 C>T	0.200551	Neutral	Low	0.004	15.33	1	0.356955
GNA11	chr19:3115037 C>T	0.330139	Neutral	Low	0.043	13.43	1	0.106818
SMO	chr7:129209384 C>A	0.261754	Neutral	Low	0.429	12.79	1	0.0460526
FGFR2	chr10:121515246 G>A	0.641029	Oncogenic	Low	0.104	12.63	1	0.0643821
NRAS	chr1:114713886 T>C	0.505196	Oncogenic	Low	0.009	12.59	4	0.0503597 - 0.0715686
MYC	chr8:127738274 C>T	0.159967	Neutral	Low	0.131	12.48	1	0.626
PIK3CA	chr3:179204568 A>G	0.520249	Oncogenic	Low	0.027	12.35	1	0.162
FGFR2	chr10:121488024 G>A	0.453985	Neutral	Low	0.055	12.3	2	0.0671551 - 0.157079
CTNNB1	chr3:41224602 C>T	0.297301	Neutral	Low	0.075	11.56	1	0.0466165
KRAS	chr12:25227398 C>T	0.366425	Neutral	Low	0.069	11.52	1	0.0662139
ALK	chr2:29213992 G>A	0.656541	Oncogenic	Low	0.131	11.45	1	0.0440613
CTNNB1	chr3:41224648 C>T	0.376643	Neutral	Low	0.07	11.42	1	0.0398162
PIK3CA	chr3:179199143 C>A	0.283186	Neutral	Low	0.026	11.38	1	0.0580871
MYC	chr8:127740673 C>T	NA	NA	NA	0.2	11.34	1	0.201601
MED12	chrX:71129386 G>A	NA	NA	NA	0.18	11.13	1	0.2755
MAP2K1	chr15:66481774 C>T	0.291842	Neutral	Low	0.031	10.25	1	0.08742

A total of 20 synonymous variants were identified with a CADD Phred-like score of ten or greater, this included four of the variants identified as potentially oncogenic using CScape (Table 3.5.2.4). The variant with the highest CADD Phred-like score was a G to A substitution in FGFR3.

Variants were organized into four classes based on their predicted potential pathogenicity in lung cancer. I developed the classification strategy as a means of integrating algorithms and databases that identify the variant as being potentially

pathogenic. The classification itself does not have any predictive significance. Variants with the strongest support had indicator scores in all the predictors. Next were variants with only two-supportive scores, one of which was CScape. The third class are variants that have two-supportive scores neither of which are CScape. Class four are variants with only one supportive score.

Class I	Class II	Class III	Class IV
ALK chr2:29213992 G>A FGFR2 chr10:121515246 G>A	NF1 chr17:31358501 G>A NRAS chr1:114713886 T>C PIK3CA chr3:179204568 A>G	SMO chr7:129209384 C>A FGFR3 chr4:1804418 G>A MYC chr8:127738274 C>T EGFR chr7:55181331 G>A	ERBB3 chr12:56085090 C>G EGFR chr7:55191841 G>A SMO chr7:129210486 C>A ALK chr2:29220757 G>A FGFR3 chr4: 1801857 C>T BRAF chr7:140794401 T>C ERBB2 chr17:39725096 C>T GNA11 chr19:3115037 C>T FGFR2 chr10:121488024 G>A

Figure 3.5.2.5: Synonymous variants of interest divided into four classes by support for possible pathogenicity.

Only two variants had support from all the predictors used, they were in ALK and FGFR2 (Figure 3.5.2.5). Interestingly, the most prevalent synonymous variant (BRAF Chr7:140801453) had no indication for pathogenicity. The variants in FGFR3 and MYC with possible splicing impact as denoted by SpliceAI, both fall into Class III. All but one of the variants in classes I, II, and III are either not reported in ClinVar/COSMIC or they are reported but with low support or not specific to lung cancer. The class III EGFR variant is

recorded in the COSMIC database in lung cancer however there is only one sample for this report.

3.6 Survival analyses

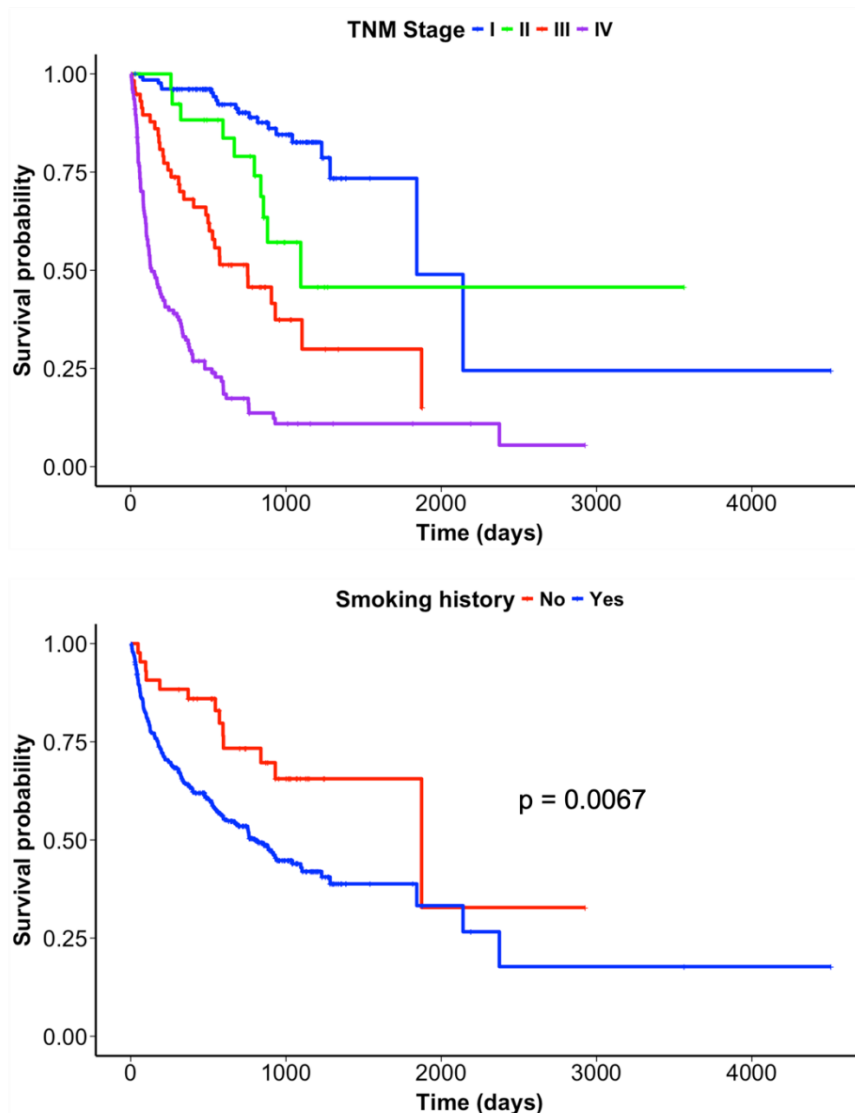
For both overall and progression-free survival the covariates examined were: sex, age, smoking history, TNM stage, and therapies received. For OS, the significant covariates were smoking history, age, and TNM stage (Figure 3.6.1). The pair-wise log-rank analysis revealed significant differences in survival between Stage I and Stage III, Stage I and Stage IV, Stage II and Stage IV, and Stage III and Stage IV. However, no significant differences were observed between Stage I and Stage II or between Stage II and Stage III. This indicates that survival outcomes are notably worse in later stages (III and IV) compared to earlier stage disease (stage I and II).

For PFS, the significant covariates were smoking history and TNM stage (Figure 3.6.2). The pair-wise log-rank analysis revealed significant differences in survival between all stages except stage I and II.

All covariates of interest were analyzed to see if they violate the proportional hazards assumption of the Cox regression. This was done by plotting the Schoenfeld residuals. TNM stage was grouped into early stage (stage I and II) and late stage (stage III and IV) based on the results of the pair-wise log-rank analysis. For both overall and PFS the TNM stage violated the proportional hazards assumption and so I did not perform Cox

regression for stage. For OS and PFS smoking history did not violate the assumption (Appendix B – Figure B1). Age also did not violate the assumption for OS.

Overall survival

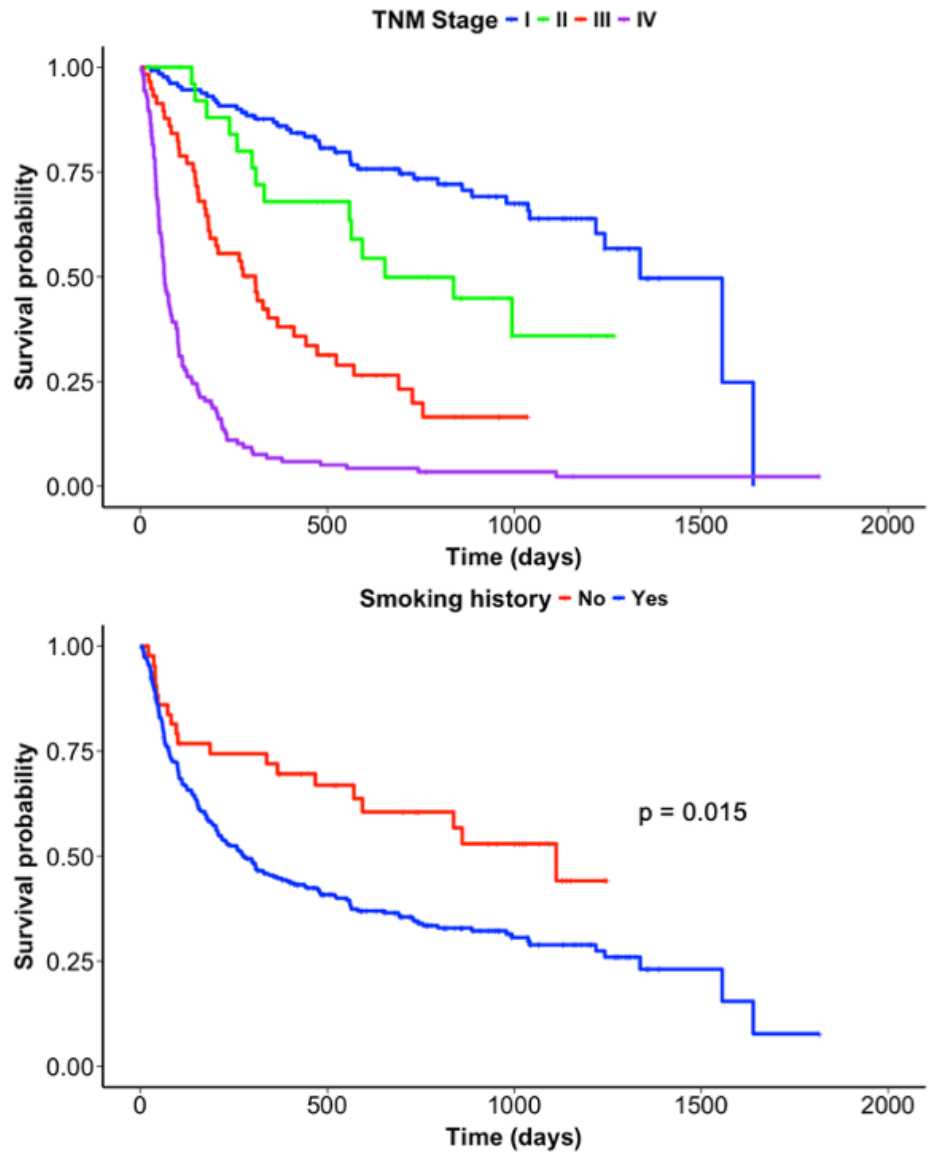


	Stage I	Stage II	Stage III
Stage II	0.085	-	-
Stage III	2.6×10^{-10}	0.193	-
Stage IV	$< 2.6 \times 10^{-10}$	7.4×10^{-7}	2.0×10^{-5}

- Age**
- Hazard ratio: 1.0339
 - Wald: 0.005
 - Log-rank: 0.005
 - Concordance: 0.571

Figure 3.6.1: Significant covariates in overall survival. Smoking history and TNM stage are plotted on Kaplan-Meier curves and compared using the log-rank test. The p-values for the pair-wise comparisons with the Bonferroni correction are displayed in the table. Hazard ratio, Wald test, Log-rank, and concordance values for univariate Cox regression of age are also shown.

Progression-free survival



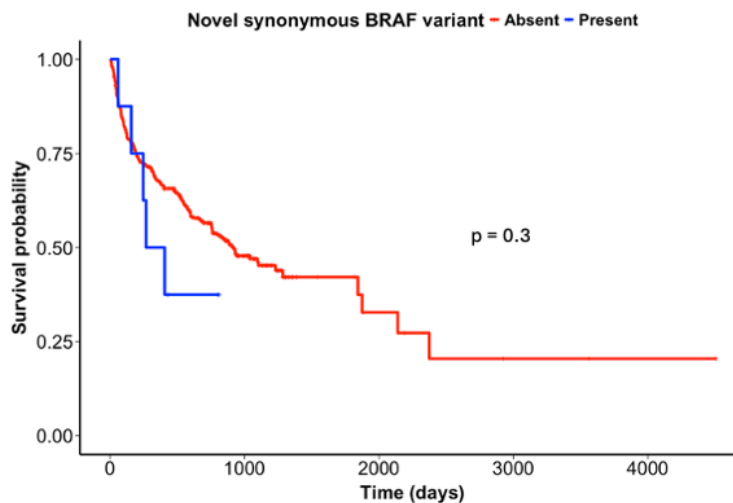
	Stage I	Stage II	Stage III
Stage II	0.093	-	-
Stage III	4.9×10^{-14}	0.028	-
Stage IV	$< 2.6 \times 10^{-16}$	1.5×10^{-10}	6.8×10^{-9}

Figure 3.6.2: Significant covariates in progression-free survival. Smoking history and TNM stage are plotted on Kaplan-Meier curves and compared using the log-rank test. The p-values for the pair-wise comparisons with the Bonferroni correction are displayed in the table.

3.6.1 Prevalent variants

The only synonymous variant prevalent enough to be assessed alone for impact on survival was the A to G substitution in BRAF (Figure 3.6.1.1). This BRAF variant was not mutually exclusive with any known driver mutation, the variant occurred alone and in the presence of varying driver mutations. Presence of the synonymous BRAF variant was associated with decreased rates of both OS and PFS. However, the difference between patients with and without the variant was not significant when compared using the log-rank test.

Overall survival



Progression-free survival

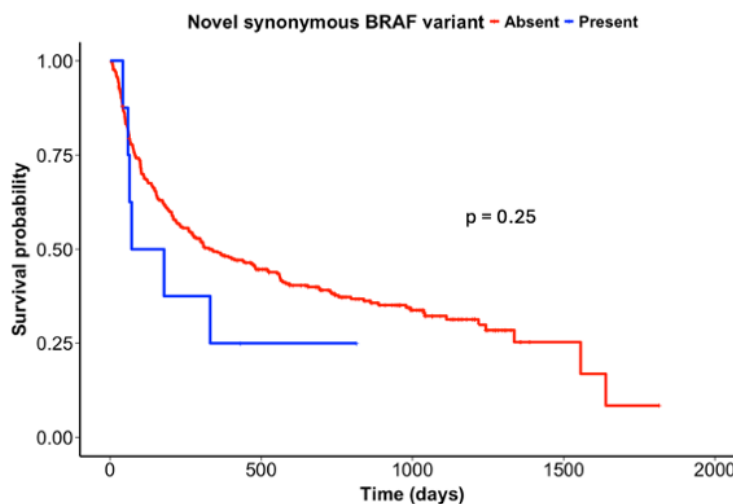


Figure 3.6.1.1: Kaplan-Meier curves with p-values from log-rank testing comparing patients based on presence of the novel synonymous BRAF variant.

3.6.2 Synonymous variants

No other variants had a high enough prevalence within the cohort to be assessed alone. Variants in classes 1-3 were grouped by their associated gene into groups that affect

similar pathways (Table 3.6.2.1). The pathway most enriched with novel variants was the MAPK/ERK or PI3K/AKT with a total of four variants, each occurring once in the cohort (Table 3.6.2.2). These pathways are involved in various cellular processes such as cell proliferation, differentiation, survival, and apoptosis.

Table 3.6.2.1: Genes with variants of unknown significance identified, grouped by primary pathway and function.

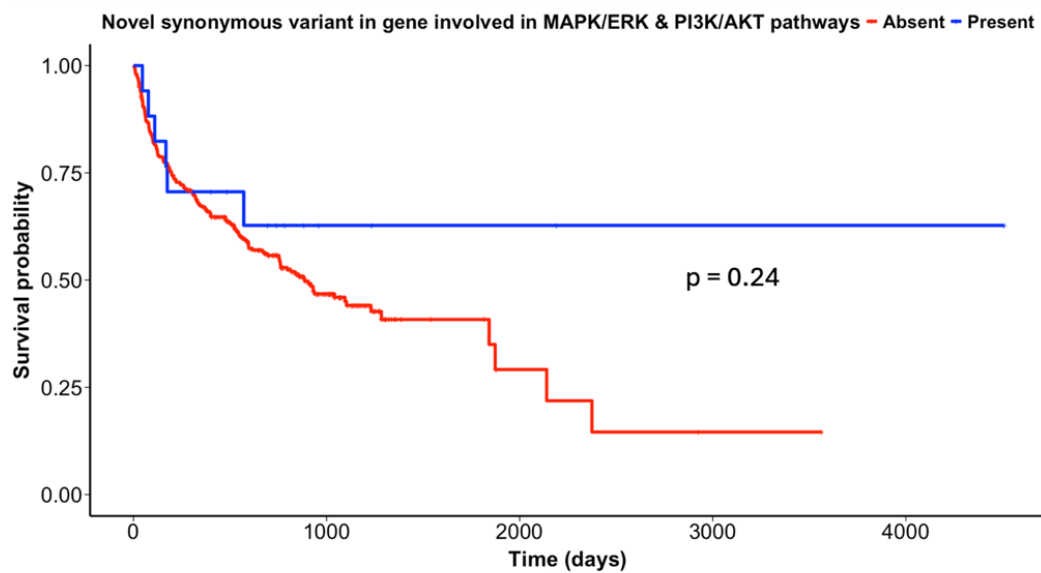
MAPK/ERK	MAPK/ERK and PI3K/AKT		PI3K/AKT	WNT/ β -catenin	JAK/STAT
BRAF	ALK	KIT	AKT1	APC	JAK2
GNA11	EGFR	MET	MTOR	CTNNB1	JAK3
GNAQ	ERBB2	PDGFRA	NF1		
HRAS	ERBB3	ROS1	PIK3CA		
KRAS	FGFR1				
MAP2K1	FGFR2				
NRAS	FGFR3				
RET	FGFR4				
Hedgehog	Transcription regulation	Cell cycle regulation	Protein degradation	Cellular metabolism	Apoptosis regulation
SMO	AR BRCA1 MED12 MYC MYCN	CCND1 CDK4 CDK6	DCUN1D1	IDH1	BIRC2

Table 3.6.2.2: Novel variants in MAPK/ERK and PI3K/AKT pathway with 2 or more predictions of pathogenicity.

Gene	Variant	Class	Cscape			TRaP score	CADD Phred-like score	Patient TNM stage	Clinical variant present
			Score	Prediction	Confidence				
ALK	chr2:29213992 G>A	1	0.6565 41	Oncogenic	Low	0.131	11.45	4	PIK3CA
FGFR2	chr10:121515246 G>A	1	0.6410 29	Oncogenic	Low	0.104	12.63	1	ERBB2
FGFR3	chr4:1804418 G>A	3	0.1480 19	Neutral	Low	0.427	18.47	4	KRAS (G12C)
EGFR	chr7:55181331 G>A	3	0.1588 37	Neutral	Low	0.101	15.34	1	ROS1

The presence of a synonymous variant in a gene associated with the MAPK/ERK and PI3K/AKT pathways resulted in an increased survival probability for both OS and PFS (Figure 3.6.2.3). Again, significance was determined by examining PFS as a greater number of events occurred. The difference between patients with and without one of the four identified variants was not significant when compared using the Log-rank test.

Overall survival



Progression-free survival

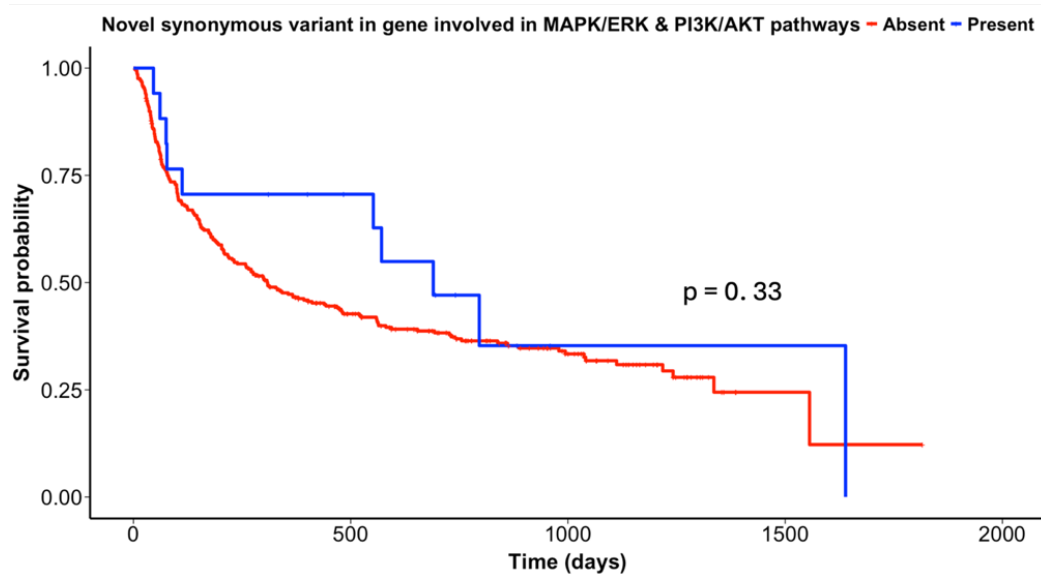


Figure 3.6.2.3: Kaplan-Meier curves with p-values from log-rank testing comparing patients based on presence of synonymous variant in gene involved in the MAPK/ERK and PI3K/AKT pathways.

4.0 Discussion

4.1 Cohort demographics & clinical characteristics

Although small, this cohort is reasonably representative of the general population of non-squamous NSCLC patients in the Western world whose tumours undergo NGS sequencing. As a referral centre for thoracic surgery the SJRH provides an interesting study site with an elevated number of early-stage NSCLCs. The European Society for Medical Oncology indicates the use of NGS for advanced NSCLC (Mosele et al., 2020), in this cohort all non-squamous NSCLC are reflexively sequenced upon diagnosis regardless of stage. Recent molecular testing guidelines advise that although reasonable, reflexive NGS testing of NSCLC is an institutional decision and as such is not routinely utilized (Aggarwal et al., 2021). As a result, the literature has historically focused exclusively on advanced stage disease.

In alignment with worldwide and national statistics, adenocarcinoma was the primary lung cancer subtype in the cohort, followed by squamous cell and large cell carcinoma (Canadian Cancer Statistics Advisory Committee, 2020; Travis et al., 2015). Comparing to national and international cohorts of patients with NSCLC who had NGS sequencing done, adenocarcinoma is still the most prevalent histology but this cohort has a lower percentage of adenocarcinoma (Choudhury et al., 2023; Fan et al., 2022; Tsoulos et al., 2017; Zacharias et al., 2021). Canada-wide demographic statistics for NSCLC are limited, and most reporting combines NSCLC with SCLC. Two studies in Canada provide

demographic data specific to NSCLC patients, one in Ontario and the other in Alberta (Akhtar-Danesh et al., 2019; Brenner et al., 2023). These studies found the mean age of diagnosis to be approximately 70 years, similar to the mean age of diagnosis of 68 years in this cohort. The sex distribution in both Alberta and Ontario leaned slightly male, while this cohort leaned slightly female. This slight male dominance is also seen internationally in both US and Korean cohorts (Chi et al., 2023; Primm et al., 2022).

Only the Ontario study had data available regarding treatment modalities received, matching the trends seen in this cohort with radiation being the most commonly received treatment, followed by surgical resection and chemotherapy (Akhtar-Danesh et al., 2019). Like this cohort, the Alberta study found that most patients present with either stage I or stage IV disease. Alternatively, Ontario found most patients presented with advanced disease (stage III and IV). Worldwide, patients are generally diagnosed at advanced stages when symptoms become evident (Polanco et al., 2021).

The demographics of this cohort align with national and international statistics in terms of subtype, mean age at diagnosis, and treatment modality. This cohort is slightly more female dominant but not to a significant level. Nationally and internationally, most patients present with advanced stage disease upon diagnosis. In this cohort, more patients present with stage I disease than any other stage, likely because the SJRH is a referral center for thoracic surgery for the other parts of the province. Most surgical resections occur in stage I patients, meaning this cohort is likely enriched for patients with early-stage disease. The amount of tissue available for sequencing is also higher in

resected patients, as a result there are patients with advanced stage disease who do not receive NGS because there is insufficient tissue, further enriching this cohort for patients with early-stage disease.

4.2 Clinical reporting

The most common non-synonymous DNA clinical mutations found within the cohort were KRAS (35.12%), EGFR (9.07%), PIK3CA (4.82%), and BRAF (4.53%). The low prevalence of PIK3CA mutations and BRAF mutations in the cohort were consistent with the literature (C. G. O’Leary et al., 2019; Scheffler et al., 2015; Yan et al., 2022).

More common in Caucasians, KRAS is known to be the most frequently mutated oncogene in NSCLC with a frequency around 30%, with KRAS G12C being the most common alteration (Fois et al., 2021; Huang et al., 2021). These trends were represented in the cohort. A small proportion of patients with KRAS mutation had no smoking history, but most patients with these mutations had a smoking history. This aligns with the literature that KRAS mutations are enriched in smokers (Wang et al., 2021).

In this cohort, EGFR was the most prevalent currently actionable mutation, with exon 19 in-frame deletions and exon 21 point mutations being the most common, aligning with the literature (Li et al., 2008). EGFR mutations are detected in more than 50% of patients among Asian populations, but the prevalence is much lower, around 12-15%, in Western populations (Cooper et al., 2013; Melosky et al., 2022). Although data regarding ethnicity was not collected for the cohort, over 80% of New Brunswick’s population is

Caucasian (Statistics Canada, 2023). Therefore, the low number of patients with EGFR mutations in my cohort is expected.

4.3 Benefits of expanded panel

A total of 30 additional patients with clinically reportable variants were detected using the 52-gene panel that would not have been detected had the predecessor 12-gene panel been used. An additional three variants were reported using the 52-gene that would have been captured with the predecessor panel but would not have been reported for pathogenicity.

In the entire cohort ten patients received targeted therapies for their actionable mutations. Two of these patients would not have had their mutations identified had the predecessor panel been utilized and it is probable they would not have received the targeted therapy. One patient received Crizotinib as part of their treatment for a MET amplified NSCLC. The other patient received Alectinib, an ALK specific targeted therapy. Had the patient with the ALK mutation been treated in the era of the 12-gene panel, their mutation may have still been captured using ALK immunohistochemistry; however this additional testing would have required extra time, materials, and effort.

The most prevalent mutation that was additionally identified as a result of the 52-gene panel was MET exon 14 skipping. Exon 14 of MET encodes an element responsible for the negative regulation of the tyrosine kinase met receptor. The loss of this element leads to an overactivation of the MET signalling pathway, promoting tumour growth and

progression. As a result, TKIs such as Capmatinib, Tepotinib, and Savolitinib have been proposed for the treatment of patients harbouring a MET exon 14 skipping mutation. These TKIs have demonstrated durable response in both untreated and pre-treated patients (Blaquier & Recondo, 2022; Drusbosky et al., 2021). It is encouraging that the most additionally captured variant has associated targeted therapies available.

4.4 Novel variants of interest

All the synonymous variants identified occurred in coding – exonic – regions of their respective genes. To discuss the proteins encoded by the genes where novel variants were identified I utilized the Universal Protein Resource (UniProt). A comprehensive resource for protein sequences and functional information, UniProt is an open and accessible database that provides a centralized repository for protein sequences and associated data from a variety of organisms (The UniProt Consortium et al., 2023).

One synonymous variant was identified in KRAS. This variant was not identified in previous literature and occurred at cDNA position 129 (c.129), whereas those existing in the literature occurred at c.30, c.60, and c.180 (Kobayashi et al., 2022; Sharma et al., 2019; Waters et al., 2016). The authors of these studies did not work on primary patient samples, they instead modified cell lines to host these synonymous mutations. These positions were chosen in part because of how frequently they occur in The Cancer Genome Atlas Program (TCGA) pan-cancer cohort. The synonymous mutation in this cohort seen at c.129 is also found in the TCGA but to my knowledge, its functional

significance has not been assessed in the literature (Kobayashi et al., 2022; Waters et al., 2016).

4.4.1 Prevalent synonymous variants

I identified two synonymous variants to be in the top ten most prevalent novel variants within the cohort. One variant was an A to G transition in the BRAF gene. This variant occurred eight times in the cohort and was not identified in the later analyses looking at pathogenicity. I was not able to find this variant in any other databases or publications. The Kaplan-Meier curves I plotted indicated that the presence of the BRAF variant showed a decrease in survival probability, but this was not statistically significant.

The BRAF gene encodes for a protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. The BRAF protein primarily phosphorylates MAP2K1 and as a result activates the MAP kinase signal transduction pathway (Brennan et al., 2011; Lavoie et al., 2018).

This specific variant occurs in the region that encodes for the zinc fingers of the BRAF protein within conserved region 1 (CR1) (UniProt: P15056). In the inactive state, CR1 inhibits the kinase domain of B-Raf. Upon activation, Ras binds to the Ras-binding domain of B-Raf, releasing this inhibition. The zinc fingers then facilitate the docking of B-Raf to the membrane, where they bind to lipids, helping localize B-Raf to specific membrane regions. This localization is crucial for B-Raf's activation and function in the signaling pathway (Cope et al., 2018; Simanshu & Morrison, 2022).

Synonymous mutations may affect mRNA stability or translational efficiency, leading to changes in protein expression levels. In the case of BRAF, alterations in protein expression could dysregulate the MAPK/ERK signaling pathway, contributing to uncontrolled cell proliferation, a hallmark of cancer. Software programs exist that can predict the effects of these mutations in-silico (Jian et al., 2014) but there is also the potential for experimental testing – such as introducing the variant in cell lines and measuring mRNA and/or protein levels.

4.4.2 Exploring synonymous variants with predicted splicing impact

My examination of synonymous variants with predicted splicing impact revealed a singular instance: a G to A transition within the FGFR3 gene. Despite lacking representation in ClinVar, this variant garnered a notable acceptor loss score from SpliceAI. FGFR3, encoding the fibroblast growth factor receptor 3, is a transmembrane receptor tyrosine-protein kinase that orchestrates critical cellular processes including proliferation and apoptosis regulation. The protein FGFR3 that plays an essential role in the regulation of cell proliferation, differentiation, and apoptosis. FGFR3 is involved in both the MAPK/ERK and PI3K/AKT pathway. Its role is to recruit proteins and activate RAS and PI3K respectively (Haugsten et al., 2010; Ornitz & Itoh, 2015).

The variant is located between the exons encoding an immunoglobulin-like domain – which is responsible for binding to fibroblast growth factors – and the exon encoding the protein kinase domain (UniProt: P22607). Loss of the splice acceptor site may lead to aberrant splicing patterns, resulting in the production of alternative mRNA

isoforms. These isoforms may lack the sequences encoding the kinase domain or contain insertions/deletions that disrupt its proper structure or function. Alternatively, loss of the splice acceptor site may result in intron retention, where the intron adjacent to the splice site is retained in the mature mRNA transcript. This could introduce premature termination codons or frameshift mutations, leading to non-functional or truncated proteins. A variant that affects splicing in this region may lead to improper function of FGFR3, theoretically leading to abnormal cell proliferation and apoptosis, which are key characteristics of cancer.

4.4.3 Predicted pathogenic synonymous variants

One of the possible predicted pathogenic variants was also one of the top ten most prevalent variants. This variant was a T to C transition in NRAS that occurred four times in the cohort. Due to such a low prevalence, this variant was not analyzed alone with a survival analysis. The variant in NRAS fell into Class II for variants with a CScape score of 0.50 and a CADD score of 12.59.

The NRAS gene encodes for the GTPase NRAS which functions to propagate signaling cascades, it is a key regulator of the MAPK/ERK and PI3K/AKT signaling pathway. NRAS in a GTP bound state activates RAF kinases such as BRAF (Mendoza et al., 2011). The variant occurs just outside the effector region of the protein in a helix (UniProt: P01111). The location of the variant near the effector region suggests that in theory it might affect mRNA secondary structure or regulatory elements important for translation or protein folding leading to changes in protein structure. Theoretically, structural alterations to the

mRNA in this region can stabilize or destabilize mRNA based on the secondary structure, dependent on the secondary structure of the mRNA. If the mRNA is stabilized this generally increases the cellular concentration of the mRNA as it slows the degradation rate (Gaither et al., 2021; Nachtergaele & He, 2018; Ross, 1995). Increased cellular concentration can result in increased protein expression which may lead to hyperactivation of downstream signaling pathways, such as the MAPK/ERK pathway, promoting uncontrolled cell proliferation, survival, and metastasis. The other variants listed in class I, II, III would also be of interest for further investigation as they have predicted significance and little or nothing in the literature regarding potential prognostic impact in NSCLC. The NF1 variant in class II would be the variant I am next most interested in investigating.

4.4.4 Variants in MAPK/ERK and PI3K/AKT pathways

Several variants were found in genes implicated in the MAPK/ERK and PI3K/AKT pathways, namely ALK, FGFR2, FGFR3, and EGFR. The PI3K/AKT pathway regulates cell survival and inhibits apoptosis (Sarris et al., 2012) while the MAPK/ERK pathway regulates gene expression involved in proliferation, differentiation, and survival. The MAPK/ERK pathway also regulates cellular responses to growth factors and mitogens (Zhang & Liu, 2002). These variants, and the most prevalent variant in BRAF are shown in their respective pathways as bolded and underlined (Figure 4.4.4.1).

The ALK and FGFR2 variants were class I variants with support from three predictors. The variant in FGFR3 is the variant with the predicted splicing impact. All the

variants excluding EGFR are either not present in ClinVar/COSMIC or are not specific to lung cancer. The EGFR variant is reported in COSMIC, but it is reported as part of a larger insertion which was not seen in this case. The EGFR variant also has a submission in ClinVar denoting it as benign in lung cancer, but the classification has low support with only one submission.

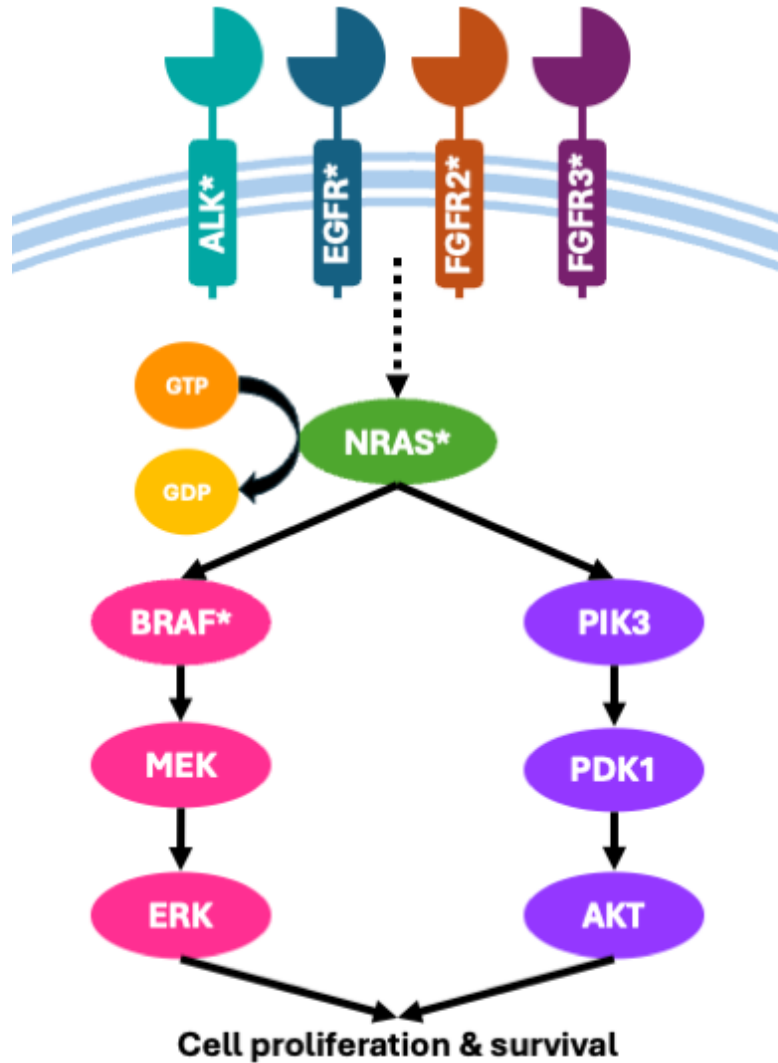


Figure 4.4.4.1: Variants of interest and their involvement in the MAPK/ERK (pink) and PI3K/AKT (purple) pathways. Dotted arrows indicate indirect signaling. Proteins with an asterisk are those with variants of interest detected in their respective genes.

Like FGFR3, the FGFR2 gene encodes a transmembrane receptor tyrosine-protein kinase but is more widely expressed in various tissues. The synonymous FGFR2 variant I have identified is a G to A transition substitution. Like the FGFR3 variant, this variant occurs between the domains that encode for the immunoglobulin-like region that directly interacts with proteins in the extracellular region and the protein kinase domain (UniProt: P21802). This variant is not predicted to alter splicing. One of the ways this variant could

have an impact is through codon usage bias. If this variant changes a codon to one that is less utilized in the cell it could slow down translation or cause premature termination, potentially affecting protein folding and function. Alternatively, it could lead to a codon that is more utilized in the cell. Regardless, downregulation or upregulation of mRNA can both ultimately lead to increased expression either directly or through compensatory upregulation. Like FGFR3, a variant in FGFR2 that may result in improper function of the protein would theoretically lead to abnormal cell proliferation and apoptosis which again, are key markers of cancer.

The other synonymous variant, supported by three pathogenicity prediction scores but no significant scores for splicing impact, is a G to A transition in the ALK gene. The ALK gene encodes a transmembrane receptor tyrosine kinase that like the FGFR family initiates signaling cascades in both the MAPK/ERK and PI3K/AKT pathways (Della Corte et al., 2018). This variant occurs in the region of the gene that encodes for the protein kinase domain (UniProt: Q9UM73). Like the variants in the FGFR variants a variant in this region could alter mRNA secondary structure through codon usage bias – such as leading to protein truncation – potentially affecting the efficiency of translation of the ALK protein kinase domain.

The final of the four variants is in the EGFR gene which encodes for the epidermal growth factor receptor, a transmembrane receptor tyrosine kinase. Like many tyrosine kinases EGFR dimerizes upon ligand binding and autophosphorylates leading to recruitment of adaptor proteins and initiation of signaling cascades in MAPK/ERK and

PI3K/AKT (Hsu et al., 2019). This variant – a G to A transition – occurs in the region that encodes for the protein kinase domain (UniProt: P00533). Like the variant in ALK, which also encodes for the protein kinase domain, this variant has no predicted splicing impact. The possible impacts of this variant are similar to those mentioned for ALK above.

Abnormal function or expression of these proteins can contribute to the dysregulation of these pathways, promoting abnormal cell proliferation and survival, key characteristics of cancer.

4.5 Strengths & limitations of the study

4.5.1 Strengths

4.5.1.1 Cohort demographics

Although the limited size of the cohort does have its disadvantages there also exist significant advantages. As a smaller province with only two referral centres for thoracic surgery and radiation oncology, this study is able to capture a significant portion of New Brunswickers, not only those who live in the city of Saint John. This includes rural patients who may have differing environmental exposures compared to their urban counterparts.

A smaller size also makes it much more feasible to collect and analyze clinical variables through chart reviews. As a referral centre for both thoracic surgery and radiation oncology, I was able to capture both early-stage – which is often lacking in NGS cohort studies – and advanced disease.

4.5.1.2 Reflexive sequencing

As a smaller centre it is much more feasible to reflexively sequence all patients with a diagnosis of non-squamous NSCLC, which is not the case for all institutions. The indication for NGS in NSCLC is for advanced-stage disease therefore much of the literature that exists is primarily representative of patients with this status. In conjunction with the increased number of patients with early-stage disease – due to the presence of the thoracic surgery department – this cohort is representative of cases that have historically been excluded from sequencing.

4.5.1.3 Reproducible, inexpensive bioinformatics pipeline

The bioinformatics pipeline I developed for this study is completely reproducible for any Next-Generation sequencing data. Regardless of the sequencing platform utilized – be it ThermoFisher or Illumina – a bioinformatician should be able to gain the same additional insight into a cohort’s genomics assuming the VCF files are available. All code for this project will be stored on a GitHub repository and freely accessible. GitHub is a platform commonly used by developers to create, share and manage code. I was able to utilize real-world clinical genomic data and the only major expense was one commercially available laptop with no special modifications. All the software I used for the project was publicly available.

4.5.2 Limitations

4.5.2.1 Comparison with the predecessor panel

The retrospective nature of the study limits our understanding of the utility and potential advantages of employing the extended 52-gene panel over the previously employed 12-gene panel. However, the detection of actionable variants like MET offers promise regarding the influence of the expanded gene panel on patient care. There is also the question of access to targeted therapies. If a patient had received care during the era of the 12-gene panel, there is no guarantee that they would have access to the same treatments – particularly targeted therapies – that they were able to access in the period that the study spanned. Although drugs such as Crizotinib have been available as early as 2011, they were initially only approved for ALK rearrangements and not MET variants (Kazandjian et al., 2014).

4.5.2.2 Cohort size, diversity, and survival analyses

The results of this study are limited by the cohort size. Having only 344 patients with available chart information significantly limited the power of the study. With the most prevalent synonymous variant occurring a mere eight times in the cohort it is unsurprising that none of the survival analyses had a high enough power when I performed a post-hoc power analysis to instill confidence in the results.

The cohort exhibits limited diversity. Genetic alterations in NSCLC are known to vary significantly across different regions, influenced by epidemiological factors such as geography, ethnicity, smoking, or gender (Laguna et al., 2024). Consequently, the findings from this cohort cannot be universally applied but rather serve to enrich the existing knowledge base particularly for Western countries.

4.5.2.3 Splicing predictors

The reliability of results in bioinformatics analysis is dependent on type and number of tools used. For this analysis the only splice predictor available was SpliceAI, which is unfortunate as it is advisable to utilize more than one predictor or algorithm where available. One limitation of SpliceAI is that the delta score provided is the difference between the score for the reference allele and the variant. This is a particular challenge when the reference value is in the intermediate range of interpretation (0.2-0.8). Using a threshold of 0.2 as recommended by the authors means that one may filter out variants if the difference between the reference and variant allele is subtle and in the intermediate range. To overcome this, the authors provide a visual interface (De Sainte Agathe et al., 2023); however this is difficult to use when processing large volumes of variants such is the case with this study.

SpliceAI also utilizes a neural network for its predictions, meaning that understanding how it scores a variant can be a black box. The algorithm is trained on data from GENCODE. The benefit of being trained on data from GENCODE is that there is a particular focus on manual validation. This is useful because if one exclusively used data that was generated computationally and may have been subject to coding errors, the generated computational output of SpliceAI would further propagate the errors. In short, the algorithm is only as good as the data it is trained on. Since the inception of SpliceAI the algorithm has been retrained for deficiencies with a particular focus on including more manually validated splice sites and excluding unverified splice sites (Strauch et al., 2022).

SpliceAI was developed with splice site recognition and intron-exon boundaries and does not explicitly account for branch points and therefore may not fully capture variants in this region. Branch points are critical sequences within the intron that interact with the spliceosome. Variation at the branch point could disrupt interaction with the spliceosome, potentially leading to improper spliceosome assembly resulting in aberrant splicing.

These issues with the algorithm are not exclusive to SpliceAI, hence the importance of utilizing multiple predictors. The development of more splicing prediction algorithms that evaluate variants regardless of consequence terms defined would be of benefit.

4.5.2.4 Pathogenic predictors

Fortunately, there are more predictors available when assessing pathogenicity. A total of three tools were utilized. Unfortunately, only one predictor utilized – Cscape – was specific to pathogenicity in cancer. This was a significant limitation of the study. It would be a benefit if more pathogenicity predictors were specific to cancer or possibly provided two scores, one for pathogenicity in cancer and the other for general pathogenicity. This way, variants could still be identified that are pathogenic for certain diseases and impact survival but may not be involved in cancer pathogenesis.

4.6 Future work

It would be of great benefit to validate this bioinformatics approach in an independent cohort, ideally one which uses a Next-Generation Sequencer and reporter

different from ours. This would further confirm that the variants identified are not artifacts of the sequencer. In an ideal world a patient's normal tissue would also be sequenced and analyzed to confirm that the variants identified are specific to the tumour. This technique is known as tumour-normal sequencing and is becoming more significant in cancer care (Mandelker & Ceyhan-Birsoy, 2020).

As I only focused on synonymous mutations, future work could use this data and a similar workflow to look at non-coding regions such as introns and regulatory elements. In future studies I would also recommend replicating this analysis on larger cohorts such as The Cancer Genome Atlas.

I present a list of synonymous variants with their associated predicted impact on splicing and possible pathogenicity. I suggest that further work should focus on the nine variants in classes I, II, and III particularly, those involved in the MAPK/ERK and PI3K/AKT pathways.

With the genetic sequencing of cancers becoming rapidly commonplace, the potential for impactful research is plentiful. Continued collection of genomic sequencing will provide enormous repositories and allow for greater statistical power and more meaningful results. The increasing number of *in silico* methods for elucidating variants makes much of this research possible even in smaller centres such as ours. Data sharing agreements between institutions could also allow for independent validation of variants of interest, adding to their validity. In the future it is possible that all NGS sequencing undergoes as in-depth of an analysis as performed in this study, hopefully allowing patients more insight and resources for their disease.

Bibliography

- Aggarwal, C., Bubendorf, L., Cooper, W. A., Illei, P., Borralho Nunes, P., Ong, B.-H., Tsao, M.-S., Yatabe, Y., & Kerr, K. M. (2021). Molecular testing in stage I–III non-small cell lung cancer: Approaches and challenges. *Lung Cancer*, *162*, 42–53. <https://doi.org/10.1016/j.lungcan.2021.09.003>
- Akhtar-Danesh, N., Akhtar-Danesh, G.-G., Seow, H. Y., Shakeel, S., & Finley, C. (2019). Trends in Survival Based on Treatment Modality in Non-Small Cell Lung Cancer Patients: A Population-Based Study. *Cancer Investigation*, *37*(8), 355–366. <https://doi.org/10.1080/07357907.2019.1653465>
- Arbour, K. C., & Riely, G. J. (2017). Diagnosis and Treatment of ALK Positive NSCLC. *Hematology/Oncology Clinics of North America*, *31*(1), 101–111. <https://doi.org/10.1016/j.hoc.2016.08.012>
- Bade, B. C., & Dela Cruz, C. S. (2020). Lung Cancer 2020: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine*, *41*(1), 1–24. <https://doi.org/10.1016/j.ccm.2019.10.001>
- Ban, W. H., Yeo, C. D., Han, S., Kang, H. S., Park, C. K., Kim, J. S., Kim, J. W., Kim, S. J., Lee, S. H., & Kim, S. K. (2020). Impact of smoking amount on clinicopathological features and survival in non-small cell lung cancer. *BMC Cancer*, *20*(1), 848. <https://doi.org/10.1186/s12885-020-07358-3>
- Blaquier, J. B., & Recondo, G. (2022). Non-small-cell lung cancer: How to manage MET exon 14 skipping mutant disease. *Drugs in Context*, *11*, 2022-2–2. <https://doi.org/10.7573/dic.2022-2-2>
- Bradley, S. H., Kennedy, M. P. T., & Neal, R. D. (2019). Recognising Lung Cancer in Primary Care. *Advances in Therapy*, *36*(1), 19–30. <https://doi.org/10.1007/s12325-018-0843-5>
- Brennan, D. F., Dar, A. C., Hertz, N. T., Chao, W. C. H., Burlingame, A. L., Shokat, K. M., & Barford, D. (2011). A Raf-induced allosteric transition of KSR stimulates phosphorylation of MEK. *Nature*, *472*(7343), 366–369. <https://doi.org/10.1038/nature09860>
- Brenner, D. R., O’Sullivan, D. E., Jarada, T. N., Yusuf, A., Boyne, D. J., Mather, C. A., Box, A., Morris, D. G., Cheung, W. Y., & Mirza, I. (2023). The impact of population-based

- EGFR testing in non-squamous metastatic non-small cell lung cancer in Alberta, Canada. *Lung Cancer*, 175, 60–67. <https://doi.org/10.1016/j.lungcan.2022.11.017>
- Canadian Cancer Statistics Advisory Committee. (2020). *Canadian Cancer Statistics: A 2020 special report on lung cancer*. Canadian Cancer Society.
- Canadian Cancer Statistics Advisory Committee. (2021). *Canadian Cancer Statistics 2021*. Canadian Cancer Society.
- Canadian Cancer Statistics Advisory Committee, Canadian Cancer Society, Statistics Canada, & Public Health Agency of Canada. (2023). *Canadian Cancer Statistics 2023*. Canadian Cancer Society.
- Canadian Task Force on Preventive Health Care. (2016). Recommendations on screening for lung cancer. *CMAJ*, 188(6), 425–432. <https://doi.org/10.1503/cmaj.151421>
- Canadian Tobacco and Nicotine Survey (CTNS): Summary of results*. (2019). Government of Canada. <https://www.canada.ca/en/health-canada/services/canadian-tobacco-nicotine-survey/2019-summary.html>
- Carr, T. H., McEwen, R., Dougherty, B., Johnson, J. H., Dry, J. R., Lai, Z., Ghazoui, Z., Laing, N. M., Hodgson, D. R., Cruzalegui, F., Hollingsworth, S. J., & Barrett, J. C. (2016). Defining actionable mutations for oncology therapeutic development. *Nature Reviews Cancer*, 16(5), 319–329. <https://doi.org/10.1038/nrc.2016.35>
- Cascetta, P., Marinello, A., Lazzari, C., Gregorc, V., Planchard, D., Bianco, R., Normanno, N., & Morabito, A. (2022). KRAS in NSCLC: State of the Art and Future Perspectives. *Cancers*, 14(21), 5430. <https://doi.org/10.3390/cancers14215430>
- Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Grant, R., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., ... Karczewski, K. J. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993), 92–100. <https://doi.org/10.1038/s41586-023-06045-0>
- Chevallier, M., Borgeaud, M., Addeo, A., & Friedlaender, A. (2021). Oncogenic driver mutations in non-small cell lung cancer: Past, present and future. *World Journal of Clinical Oncology*, 12(4), 217–237. <https://doi.org/10.5306/wjco.v12.i4.217>
- Chi, S. A., Yu, H., Choi, Y.-L., Park, S., Sun, J.-M., Lee, S.-H., Ahn, J. S., Ahn, M.-J., Choi, D.-H., Kim, K., Jung, H. A., & Park, K. (2023). Trends in Survival Rates of Non-Small

Cell Lung Cancer With Use of Molecular Testing and Targeted Therapy in Korea, 2010-2020. *JAMA Network Open*, 6(3), e232002.
<https://doi.org/10.1001/jamanetworkopen.2023.2002>

Choudhury, N. J., Lavery, J. A., Brown, S., De Bruijn, I., Jee, J., Tran, T. N., Rizvi, H., Arbour, K. C., Whiting, K., Shen, R., Hellmann, M., Bedard, P. L., Yu, C., Leighl, N., LeNoue-Newton, M., Micheel, C., Warner, J. L., Ginsberg, M. S., Plodkowski, A., ... on behalf of the AACR GENIE BPC Core Team. (2023). The GENIE BPC NSCLC Cohort: A Real-World Repository Integrating Standardized Clinical and Genomic Data for 1,846 Patients with Non-Small Cell Lung Cancer. *Clinical Cancer Research*, 29(17), 3418–3428. <https://doi.org/10.1158/1078-0432.CCR-23-0580>

Cooper, W. A., Lam, D. C. L., O'Toole, S. A., & Minna, J. D. (2013). Molecular biology of lung cancer. *Journal of Thoracic Disease*, 5 Suppl 5(Suppl 5), S479-490.
<https://doi.org/10.3978/j.issn.2072-1439.2013.08.03>

Cope, N., Candelora, C., Wong, K., Kumar, S., Nan, H., Grasso, M., Novak, B., Li, Y., Marmorstein, R., & Wang, Z. (2018). Mechanism of BRAF Activation through Biochemical Characterization of the Recombinant Full-Length Protein. *Chembiochem: A European Journal of Chemical Biology*, 19(18), 1988–1997.
<https://doi.org/10.1002/cbic.201800359>

Corrales, L., Rosell, R., Cardona, A. F., Martín, C., Zatarain-Barrón, Z. L., & Arrieta, O. (2020). Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Critical Reviews in Oncology/Hematology*, 148, 102895.
<https://doi.org/10.1016/j.critrevonc.2020.102895>

D. Martinez, V., P. Sage, A., A. Marshall, E., Suzuki, M., A. Goodarzi, A., Dellaire, G., & L. Lam, W. (2019). Oncogenetics of Lung Cancer Induced by Environmental Carcinogens. In P. Erkekoglu (Ed.), *Oncogenes and Carcinogenesis*. IntechOpen.
<https://doi.org/10.5772/intechopen.81064>

de Alencar, V. T. L., Formiga, M. N., & de Lima, V. C. C. (2020). Inherited lung cancer: A review. *Ecancermedicalscience*, 14, 1008.
<https://doi.org/10.3332/ecancer.2020.1008>

De Mello, R. A. B., Voscaboinik, R., Luciano, J. V. P., Cremonese, R. V., Amaral, G. A., Castelo-Branco, P., & Antoniou, G. (2021). Immunotherapy in Patients with Advanced Non-Small Cell Lung Cancer Lacking Driver Mutations and Future Perspectives. *Cancers*, 14(1), 122. <https://doi.org/10.3390/cancers14010122>

- De Sainte Agathe, J.-M., Filser, M., Isidor, B., Besnard, T., Gueguen, P., Perrin, A., Van Goethem, C., Verebi, C., Masingue, M., Rendu, J., Cossée, M., Bergougnoux, A., Frobert, L., Buratti, J., Lejeune, É., Le Guern, É., Pasquier, F., Clot, F., Kalatzis, V., ... Baux, D. (2023). SpliceAI-visual: A free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*, *17*(1), 7. <https://doi.org/10.1186/s40246-023-00451-1>
- Della Corte, C. M., Viscardi, G., Di Liello, R., Fasano, M., Martinelli, E., Troiani, T., Ciardiello, F., & Morgillo, F. (2018). Role and targeting of anaplastic lymphoma kinase in cancer. *Molecular Cancer*, *17*(1), 30. <https://doi.org/10.1186/s12943-018-0776-2>
- Diederichs, S., Bartsch, L., Berkmann, J. C., Fröse, K., Heitmann, J., Hoppe, C., Iggena, D., Jazmati, D., Karschnia, P., Linsenmeier, M., Maulhardt, T., Möhrmann, L., Morstein, J., Paffenholz, S. V., Röpenack, P., Rückert, T., Sandig, L., Schell, M., Steinmann, A., ... Wullenkord, R. (2016). The dark matter of the cancer genome: Aberrations in regulatory elements, untranslated regions, splice sites, non-coding <sc>RNA</sc> and synonymous mutations. *EMBO Molecular Medicine*, *8*(5), 442–457. <https://doi.org/10.15252/emmm.201506055>
- Draetta, E. L., Lazarević, D., Provero, P., & Cittaro, D. (2022). The frequency of somatic mutations in cancer predicts the phenotypic relevance of germline mutations. *Frontiers in Genetics*, *13*, 1045301. <https://doi.org/10.3389/fgene.2022.1045301>
- Drusbosky, L. M., Rodriguez, E., Dawar, R., & Ikpeazu, C. V. (2021). Therapeutic strategies in RET gene rearranged non-small cell lung cancer. *Journal of Hematology & Oncology*, *14*(1), 50. <https://doi.org/10.1186/s13045-021-01063-9>
- Du, X., Shao, Y., Qin, H., Tai, Y., & Gao, H. (2018). ALK-rearrangement in non-small-cell lung cancer (NSCLC). *Thoracic Cancer*, *9*(4), 423–430. <https://doi.org/10.1111/1759-7714.12613>
- Duesberg, P. H., & Vogt, P. K. (1970). Differences between the Ribonucleic Acids of Transforming and Nontransforming Avian Tumor Viruses*. *Proceedings of the National Academy of Sciences of the United States of America*, *67*(4), 1673–1680.
- Dummer, T. J. B., Yu, Z. M., Nauta, L., Murimboh, J. D., & Parker, L. (2015). Geostatistical modelling of arsenic in drinking water wells and related toenail arsenic concentrations across Nova Scotia, Canada. *Science of The Total Environment*, *505*, 1248–1258. <https://doi.org/10.1016/j.scitotenv.2014.02.055>

- Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology*, 6(5), R44. <https://doi.org/10.1186/gb-2005-6-5-r44>
- Elliott, K., & Larsson, E. (2021). Non-coding driver mutations in human cancer. *Nature Reviews Cancer*, 21(8), 500–509. <https://doi.org/10.1038/s41568-021-00371-z>
- Ellison, L. F., & Saint-Jacques, N. (2023). Five-year cancer survival by stage at diagnosis in Canada. *Health Reports*, 34(82).
- Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., Bruno, D. S., Chang, J. Y., Chirieac, L. R., D'Amico, T. A., DeCamp, M., Dilling, T. J., Dowell, J., Gettinger, S., Grotz, T. E., Gubens, M. A., Hegde, A., Lackner, R. P., Lanuti, M., ... Hughes, M. (2022). Non–Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network*, 20(5), 497–530. <https://doi.org/10.6004/jnccn.2022.0025>
- Faguet, G. B. (2015). A brief history of cancer: Age-old milestones underlying our current knowledge database. *International Journal of Cancer*, 136(9), 2022–2036. <https://doi.org/10.1002/ijc.29134>
- Falkson, C. B., Vella, E. T., Yu, E., El-Mallah, M., Mackenzie, R., Ellis, P. M., & Ung, Y. C. (2017). Guideline for radiotherapy with curative intent in patients with early-stage medically inoperable non-small-cell lung cancer. *Current Oncology (Toronto, Ont.)*, 24(1), e44–e49. <https://doi.org/10.3747/co.24.3358>
- Fan, Z., Tudor, R., Le, L., Law, J., Kuang, S., Meti, N., Fung, A., Perdrizet, K., Chen, K., Li, J., Ghumman, N., Ranich, L., Wei, C., Sabatani, P., Tsao, M.-S., Leighl, N., & Cabanero, M. (2022). 1151P Evolution of biomarker testing among non-squamous/non-small cell lung cancer (NSCLC) patients (Pts) and impact on turnaround times (TAT). *Annals of Oncology*, 33, S1076. <https://doi.org/10.1016/j.annonc.2022.07.1275>
- Fois, S. S., Paliogiannis, P., Zinellu, A., Fois, A. G., Cossu, A., & Palmieri, G. (2021). Molecular Epidemiology of the Main Druggable Genetic Alterations in Non-Small Cell Lung Cancer. *International Journal of Molecular Sciences*, 22(2), Article 2. <https://doi.org/10.3390/ijms22020612>

- Fu, K., Xie, F., Wang, F., & Fu, L. (2022). Therapeutic strategies for EGFR-mutated non-small cell lung cancer patients with osimertinib resistance. *Journal of Hematology & Oncology*, *15*(1), 173. <https://doi.org/10.1186/s13045-022-01391-4>
- Gaither, J. B. S., Lammi, G. E., Li, J. L., Gordon, D. M., Kuck, H. C., Kelly, B. J., Fitch, J. R., & White, P. (2021). Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *GigaScience*, *10*(4), giab023. <https://doi.org/10.1093/gigascience/giab023>
- Galvano, A., Gristina, V., Malapelle, U., Pisapia, P., Pepe, F., Barraco, N., Castiglia, M., Perez, A., Rolfo, C., Troncione, G., Russo, A., & Bazan, V. (2021). The prognostic impact of tumor mutational burden (TMB) in the first-line management of advanced non-oncogene addicted non-small-cell lung cancer (NSCLC): A systematic review and meta-analysis of randomized controlled trials. *ESMO Open*, *6*(3), 100124. <https://doi.org/10.1016/j.esmoop.2021.100124>
- Garassino, M. C., Borgonovo, K., Rossi, A., Mancuso, A., Martelli, O., Tinazzi, A., Di Cosimo, S., La Verde, N., Sburlati, P., Bianchi, C., Farina, G., & Torri, V. (2009). Biological and clinical features in predicting efficacy of epidermal growth factor receptor tyrosine kinase inhibitors: A systematic review and meta-analysis. *Anticancer Research*, *29*(7), 2691–2701.
- Gelfman, S., Wang, Q., McSweeney, K. M., Ren, Z., La Carpia, F., Halvorsen, M., Schoch, K., Ratzon, F., Heinzen, E. L., Boland, M. J., Petrovski, S., & Goldstein, D. B. (2017). Annotating pathogenic non-coding variants in genic regions. *Nature Communications*, *8*(1), 236. <https://doi.org/10.1038/s41467-017-00141-2>
- Gemine, R. E., Ghosal, R., Collier, G., Parry, D., Campbell, I., Davies, G., Davies, K., & Lewis, K. E. (2019). Longitudinal study to assess impact of smoking at diagnosis and quitting on 1-year survival for people with non-small cell lung cancer. *Lung Cancer*, *129*, 1–7. <https://doi.org/10.1016/j.lungcan.2018.12.028>
- George, R. A., Smith, T. D., Callaghan, S., Hardman, L., Pierides, C., Horaitis, O., Wouters, M. A., & Cotton, R. G. H. (2008). General mutation databases: Analysis and review. *Journal of Medical Genetics*, *45*(2), 65–70. <https://doi.org/10.1136/jmg.2007.052639>
- Gomperts, B., Spira, A., Massion, P., Walser, T., Wistuba, I., Minna, J., & Dubinett, S. (2011). Evolving Concepts in Lung Carcinogenesis. *Seminars in Respiratory and Critical Care Medicine*, *32*(01), 032–043. <https://doi.org/10.1055/s-0031-1272867>

- GraphPad Software. (2024). *Prism* (Version 10.2.1) [macOS]. www.graphpad.com
- Grosz, A. E., Grossman, J. N., Garrett, R., Friske, P., Smith, D. B., Darnley, A. G., & Vowinkel, E. (2004). A preliminary geochemical map for arsenic in surficial materials of Canada and the United States. *Applied Geochemistry*, *19*(2), 257–260. <https://doi.org/10.1016/j.apgeochem.2003.09.012>
- Guo, Q., Liu, L., Chen, Z., Fan, Y., Zhou, Y., Yuan, Z., & Zhang, W. (2022). Current treatments for non-small cell lung cancer. *Frontiers in Oncology*, *12*, 945102. <https://doi.org/10.3389/fonc.2022.945102>
- Gupta, R., Smalley, M., Anusim, N., Jindal, V., Singh Rahi, M., Gupta, S., Gupta, S., & Jaiyesimi, I. (2021). Paradigm shift in the management of metastatic nonsmall cell lung cancer. *International Journal of Clinical Practice*, *75*(11), e14533. <https://doi.org/10.1111/ijcp.14533>
- Gutman, T., Goren, G., Efroni, O., & Tuller, T. (2021). Estimating the predictive power of silent mutations on cancer classification and prognosis. *Npj Genomic Medicine*, *6*(1), Article 1. <https://doi.org/10.1038/s41525-021-00229-1>
- Haimovich, A. D. (2011). Methods, challenges, and promise of next-generation sequencing in cancer biology. *The Yale Journal of Biology and Medicine*, *84*(4), 439–446.
- Hamilton, W., Peters, T. J., Round, A., & Sharp, D. (2005). What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax*, *60*(12), 1059–1065. <https://doi.org/10.1136/thx.2005.045880>
- Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. [https://doi.org/10.1016/S0092-8674\(00\)81683-9](https://doi.org/10.1016/S0092-8674(00)81683-9)
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. <https://doi.org/10.1016/j.cell.2011.02.013>
- Haugsten, E. M., Wiedlocha, A., Olsnes, S., & Wesche, J. (2010). Roles of Fibroblast Growth Factor Receptors in Carcinogenesis. *Molecular Cancer Research*, *8*(11), 1439–1452. <https://doi.org/10.1158/1541-7786.MCR-10-0168>
- Hayashi, R., & Inomata, M. (2022). Small cell lung cancer; recent advances of its biology and therapeutic perspective. *Respiratory Investigation*, *60*(2), 197–204. <https://doi.org/10.1016/j.resinv.2021.10.008>

- Ho, C., Lefresne, S., Liberman, M., McGuire, A., Palma, D., Pender, A., Snow, S., Tremblay, A., & Myers, R. (2019). Lung Cancer in Canada. *Journal of Thoracic Oncology*, *14*(7), 1128–1133. <https://doi.org/10.1016/j.jtho.2019.02.012>
- Hsu, P.-C., Jablons, D. M., Yang, C.-T., & You, L. (2019). Epidermal Growth Factor Receptor (EGFR) Pathway, Yes-Associated Protein (YAP) and the Regulation of Programmed Death-Ligand 1 (PD-L1) in Non-Small Cell Lung Cancer (NSCLC). *International Journal of Molecular Sciences*, *20*(15), 3821. <https://doi.org/10.3390/ijms20153821>
- Huang, L., Guo, Z., Wang, F., & Fu, L. (2021). KRAS mutation: From undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy*, *6*(1), 386. <https://doi.org/10.1038/s41392-021-00780-4>
- Hunt, S. E., Moore, B., Amode, R. M., Armean, I. M., Lemos, D., Mushtaq, A., Parton, A., Schuilenburg, H., Szpak, M., Thormann, A., Perry, E., Trevanion, S. J., Flicek, P., Yates, A. D., & Cunningham, F. (2022). Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor-A tutorial. *Human Mutation*, *43*(8), 986–997. <https://doi.org/10.1002/humu.24298>
- Ionescu, D. N., Stockley, T. L., Banerji, S., Couture, C., Mather, C. A., Xu, Z., Blais, N., Cheema, P. K., Chu, Q. S.-C., Melosky, B., & Leighl, N. B. (2022). Consensus Recommendations to Optimize Testing for New Targetable Alterations in Non-Small Cell Lung Cancer. *Current Oncology*, *29*(7), 4981–4997. <https://doi.org/10.3390/curroncol29070396>
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglu, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535–548.e24. <https://doi.org/10.1016/j.cell.2018.12.015>
- Jančík, S., Drábek, J., Radzioch, D., & Hajdúch, M. (2010). Clinical Relevance of KRAS in Human Cancers. *Journal of Biomedicine and Biotechnology*, *2010*, 150960. <https://doi.org/10.1155/2010/150960>
- Jeon, D. S., Kim, H. C., Kim, S. H., Kim, T.-J., Kim, H. K., Moon, M. H., Beck, K. S., Suh, Y.-G., Song, C., Ahn, J. S., Lee, J. E., Lim, J. U., Jeon, J. H., Jung, K.-W., Jung, C. Y., Cho, J. S., Choi, Y.-D., Hwang, S.-S., Choi, C.-M., ... Korea Central Cancer Registry. (2023). Five-Year Overall Survival and Prognostic Factors in Patients with Lung Cancer: Results from the Korean Association of Lung Cancer Registry (KALC-R) 2015.

Cancer Research and Treatment, 55(1), 103–111.
<https://doi.org/10.4143/crt.2022.264>

- Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22), 13534–13544. <https://doi.org/10.1093/nar/gku1206>
- Kamps, R., Brandão, R. D., Bosch, B. J. van den, Paulussen, A. D. C., Xanthoulea, S., Blok, M. J., & Romano, A. (2017). Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *International Journal of Molecular Sciences*, 18(2), 308. <https://doi.org/10.3390/ijms18020308>
- Kanwal, M., Ding, X.-J., & Cao, Y. (2017). Familial risk for lung cancer. *Oncology Letters*, 13(2), 535–542. <https://doi.org/10.3892/ol.2016.5518>
- Kazandjian, D., Blumenthal, G. M., Chen, H.-Y., He, K., Patel, M., Justice, R., Keegan, P., & Pazdur, R. (2014). FDA approval summary: Crizotinib for the treatment of metastatic non-small cell lung cancer with anaplastic lymphoma kinase rearrangements. *The Oncologist*, 19(10), e5-11.
<https://doi.org/10.1634/theoncologist.2014-0241>
- Khan, S. M., Pearson, D. D., Eldridge, E. L., Morais, T. A., Ahanonu, M. I. C., Ryan, M. C., Taron, J. M., & Goodarzi, A. A. (2024). Rural communities experience higher radon exposure versus urban areas, potentially due to drilled groundwater well annuli acting as unintended radon gas migration conduits. *Scientific Reports*, 14(1), 3640. <https://doi.org/10.1038/s41598-024-53458-6>
- Khan, S. M., Pearson, D. D., Rönnqvist, T., Nielsen, M. E., Taron, J. M., & Goodarzi, A. A. (2021). Rising Canadian and falling Swedish radon gas exposure as a consequence of 20th to 21st century residential build practices. *Scientific Reports*, 11(1), 17551. <https://doi.org/10.1038/s41598-021-96928-x>
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315. <https://doi.org/10.1038/ng.2892>
- Kobayashi, Y., Chhoeu, C., Li, J., Price, K. S., Kiedrowski, L. A., Hutchins, J. L., Hardin, A. I., Wei, Z., Hong, F., Bahcall, M., Gokhale, P. C., & Jänne, P. A. (2022). Silent mutations reveal therapeutic vulnerability in RAS Q61 cancers. *Nature*, 603(7900), 335–342. <https://doi.org/10.1038/s41586-022-04451-4>

- Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1), 91. <https://doi.org/10.1186/s13073-020-00791-w>
- Laguna, J. C., García-Pardo, M., Alessi, J., Barrios, C., Singh, N., Al-Shamsi, H. O., Loong, H., Ferriol, M., Recondo, G., & Mezquita, L. (2024). Geographic differences in lung cancer: Focus on carcinogens, genetic predisposition, and molecular epidemiology. *Therapeutic Advances in Medical Oncology*, 16, 17588359241231260. <https://doi.org/10.1177/17588359241231260>
- Larson, N. B., Oberg, A. L., Adjei, A. A., & Wang, L. (2023). A Clinician's Guide to Bioinformatics for Next-Generation Sequencing. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 18(2), 143–157. <https://doi.org/10.1016/j.jtho.2022.11.006>
- Lavoie, H., Sahmi, M., Maisonneuve, P., Marullo, S. A., Thevakumaran, N., Jin, T., Kurinov, I., Sicheri, F., & Therrien, M. (2018). MEK drives BRAF activation through allosteric control of KSR proteins. *Nature*, 554(7693), 549–553. <https://doi.org/10.1038/nature25478>
- Lazzari, C., Bulotta, A., Cangì, M. G., Bucci, G., Pecciarini, L., Bonfiglio, S., Lorusso, V., Ippati, S., Arrigoni, G., Grassini, G., Doglioni, C., & Gregorc, V. (2020). Next Generation Sequencing in Non-Small Cell Lung Cancer: Pitfalls and Opportunities. *Diagnostics*, 10(12), Article 12. <https://doi.org/10.3390/diagnostics10121092>
- Lee, S. J., Lee, J., Park, Y. S., Lee, C.-H., Lee, S.-M., Yim, J.-J., Yoo, C.-G., Han, S. K., & Kim, Y. W. (2014). Impact of smoking on mortality of patients with non-small cell lung cancer. *Thoracic Cancer*, 5(1), 43–49. <https://doi.org/10.1111/1759-7714.12051>
- Li, A. R., Chitale, D., Riely, G. J., Pao, W., Miller, V. A., Zakowski, M. F., Rusch, V., Kris, M. G., & Ladanyi, M. (2008). EGFR mutations in lung adenocarcinomas: Clinical testing experience and relationship to EGFR gene copy number and immunohistochemical expression. *The Journal of Molecular Diagnostics: JMD*, 10(3), 242–248. <https://doi.org/10.2353/jmoldx.2008.070178>
- Lih, C.-J., Harrington, R. D., Sims, D. J., Harper, K. N., Bouk, C. H., Datta, V., Yau, J., Singh, R. R., Routbort, M. J., Luthra, R., Patel, K. P., Mantha, G. S., Krishnamurthy, S., Ronski, K., Walther, Z., Finberg, K. E., Canosa, S., Robinson, H., Raymond, A., ... Williams, P. M. (2017). Analytical Validation of the Next-Generation Sequencing Assay for a Nationwide Signal-Finding Clinical Trial. *The Journal of Molecular Diagnostics : JMD*, 19(2), 313–327. <https://doi.org/10.1016/j.jmoldx.2016.10.007>

- Lindeman, N. I., Cagle, P. T., Beasley, M. B., Chitale, D. A., Dacic, S., Giaccone, G., Jenkins, R. B., Kwiatkowski, D. J., Saldivar, J.-S., Squire, J., Thunnissen, E., & Ladanyi, M. (2013). Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *Journal of Thoracic Oncology*, *8*(7), 823–859. <https://doi.org/10.1097/JTO.0b013e318290868f>
- Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, *32*(8), 894–899. <https://doi.org/10.1002/humu.21517>
- Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020). dbNSFP v4: A comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, *12*(1), 103. <https://doi.org/10.1186/s13073-020-00803-9>
- Luo, S. Y., & Lam, D. C. (2013). Oncogenic driver mutations in lung cancer. *Translational Respiratory Medicine*, *1*, 6. <https://doi.org/10.1186/2213-0802-1-6>
- Mandelker, D., & Ceyhan-Birsoy, O. (2020). Evolving Significance of Tumor-Normal Sequencing in Cancer Care. *Trends in Cancer*, *6*(1), 31–39. <https://doi.org/10.1016/j.trecan.2019.11.006>
- Marcus, L., Fashoyin-Aje, L. A., Donoghue, M., Yuan, M., Rodriguez, L., Gallagher, P. S., Philip, R., Ghosh, S., Theoret, M. R., Beaver, J. A., Pazdur, R., & Lemery, S. J. (2021). FDA Approval Summary: Pembrolizumab for the treatment of tumor mutational burden-high solid tumors. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *27*(17), 4685–4689. <https://doi.org/10.1158/1078-0432.CCR-21-0327>
- Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. <https://doi.org/10.1146/annurev-anchem-062012-092628>
- Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029-1041.e21. <https://doi.org/10.1016/j.cell.2017.09.042>

- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. <https://doi.org/10.1186/s13059-016-0974-4>
- Melosky, B., Blais, N., Cheema, P., Couture, C., Juergens, R., Kamel-Reid, S., Tsao, M.-S., Wheatley-Price, P., Xu, Z., & Ionescu, D. N. (2018). Standardizing Biomarker Testing for Canadian Patients with Advanced Lung Cancer. *Current Oncology*, *25*(1), 73–82. <https://doi.org/10.3747/co.25.3867>
- Melosky, B., Kambartel, K., Häntschel, M., Bennetts, M., Nickens, D. J., Brinkmann, J., Kayser, A., Moran, M., & Cappuzzo, F. (2022). Worldwide Prevalence of Epidermal Growth Factor Receptor Mutations in Non-Small Cell Lung Cancer: A Meta-Analysis. *Molecular Diagnosis & Therapy*, *26*(1), 7–18. <https://doi.org/10.1007/s40291-021-00563-1>
- Mendoza, M. C., Er, E. E., & Blenis, J. (2011). The Ras-ERK and PI3K-mTOR pathways: Cross-talk and compensation. *Trends in Biochemical Sciences*, *36*(6), 320–328. <https://doi.org/10.1016/j.tibs.2011.03.006>
- Mok, T. S., Wu, Y.-L., Thongprasert, S., Yang, C.-H., Chu, D.-T., Saijo, N., Sunpaweravong, P., Han, B., Margono, B., Ichinose, Y., Nishiwaki, Y., Ohe, Y., Yang, J.-J., Chewaskulyong, B., Jiang, H., Duffield, E. L., Watkins, C. L., Armour, A. A., & Fukuoka, M. (2009). Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. *New England Journal of Medicine*, *361*(10), 947–957. <https://doi.org/10.1056/NEJMoa0810699>
- Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., ... Murphy, T. D. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, *604*(7905), 310–315. <https://doi.org/10.1038/s41586-022-04558-8>
- Mosele, F., Remon, J., Mateo, J., Westphalen, C. B., Barlesi, F., Lolkema, M. P., Normanno, N., Scarpa, A., Robson, M., Meric-Bernstam, F., Wagle, N., Stenzinger, A., Bonastre, J., Bayle, A., Michiels, S., Bièche, I., Rouleau, E., Jezdic, S., Douillard, J.-Y., ... André, F. (2020). Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: A report from the ESMO Precision Medicine Working Group. *Annals of Oncology*, *31*(11), 1491–1505. <https://doi.org/10.1016/j.annonc.2020.07.014>

- Nachtergaele, S., & He, C. (2018). Chemical Modifications in the Life of an mRNA Transcript. *Annual Review of Genetics*, 52, 349–372.
<https://doi.org/10.1146/annurev-genet-120417-031522>
- Oelschlaeger, P. (2024). Molecular Mechanisms and the Significance of Synonymous Mutations. *Biomolecules*, 14(1), 132. <https://doi.org/10.3390/biom14010132>
- Ogundimu, E. O., Altman, D. G., & Collins, G. S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, 76, 175–182.
<https://doi.org/10.1016/j.jclinepi.2016.02.031>
- O’Leary, C. G., Anelkovic, V., Ladwa, R., Pavlakis, N., Zhou, C., Hirsch, F., Richard, D., & O’Byrne, K. (2019). Targeting BRAF mutations in non-small cell lung cancer. *Translational Lung Cancer Research*, 8(6).
<https://doi.org/10.21037/tlcr.2019.10.22>
- O’Leary, C., Gasper, H., Sahin, K. B., Tang, M., Kulasinghe, A., Adams, M. N., Richard, D. J., & O’Byrne, K. J. (2020). Epidermal Growth Factor Receptor (EGFR)-Mutated Non-Small-Cell Lung Cancer (NSCLC). *Pharmaceuticals*, 13(10), 273.
<https://doi.org/10.3390/ph13100273>
- Ornitz, D. M., & Itoh, N. (2015). The Fibroblast Growth Factor signaling pathway. *Wiley Interdisciplinary Reviews. Developmental Biology*, 4(3), 215–266.
<https://doi.org/10.1002/wdev.176>
- Padinharayil, H., Varghese, J., John, M. C., Rajanikant, G. K., Wilson, C. M., Al-Yozbaki, M., Renu, K., Dewanjee, S., Sanyal, R., Dey, A., Mukherjee, A. G., Wanjari, U. R., Gopalakrishnan, A. V., & George, A. (2023). Non-small cell lung carcinoma (NSCLC): Implications on molecular pathology and advances in early diagnostics and therapeutics. *Genes & Diseases*, 10(3), 960–989.
<https://doi.org/10.1016/j.gendis.2022.07.023>
- Palmeri, M., Mehnert, J., Silk, A. W., Jabbour, S. K., Ganesan, S., Popli, P., Riedlinger, G., Stephenson, R., de Meritens, A. B., Leiser, A., Mayer, T., Chan, N., Spencer, K., Girda, E., Malhotra, J., Chan, T., Subbiah, V., & Groisberg, R. (2022). Real-world application of tumor mutational burden-high (TMB-high) and microsatellite instability (MSI) confirms their utility as immunotherapy biomarkers. *ESMO Open*, 7(1), 100336. <https://doi.org/10.1016/j.esmoop.2021.100336>

- Pelekanakis, A., O’Loughlin, J. L., Gagné, T., Callard, C., & Frohlich, K. L. (2021). Initiation or cessation: What keeps the prevalence of smoking higher in Quebec than in the rest of Canada? *Health Promotion and Chronic Disease Prevention in Canada*, 41(10), 306–314. <https://doi.org/10.24095/hpcdp.41.10.05>
- Peng, L.-X., Jie, G.-L., Li, A.-N., Liu, S.-Y., Sun, H., Zheng, M.-M., Zhou, J.-Y., Zhang, J.-T., Zhang, X.-C., Zhou, Q., Zhong, W.-Z., Yang, J.-J., Tu, H.-Y., Su, J., Yan, H.-H., & Wu, Y.-L. (2021). MET amplification identified by next-generation sequencing and its clinical relevance for MET inhibitors. *Experimental Hematology & Oncology*, 10(1), 52. <https://doi.org/10.1186/s40164-021-00245-y>
- Pisters, K., Kris, M. G., Gaspar, L. E., Ismaila, N., & Panel, for the A. S. T. and A. R. T. for S. I. to I. N. G. E. (2022). Adjuvant Systemic Therapy and Adjuvant Radiation Therapy for Stage I-III A Completely Resected Non–Small-Cell Lung Cancer: ASCO Guideline Rapid Recommendation Update. *Journal of Clinical Oncology*. <https://doi.org/10.1200/JCO.22.00051>
- Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E. F., Faivre-Finn, C., Mok, T. S., Reck, M., Van Schil, P. E., Hellmann, M. D., & Peters, S. (2018). Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29, iv192–iv237. <https://doi.org/10.1093/annonc/mdy275>
- Poirier, A. E., Ruan, Y., Grevers, X., Walter, S. D., Villeneuve, P. J., Friedenreich, C. M., & Brenner, D. R. (2019). Estimates of the current and future burden of cancer attributable to active and passive tobacco smoking in Canada. *Preventive Medicine*, 122, 9–19. <https://doi.org/10.1016/j.ypmed.2019.03.015>
- Polanco, D., Pinilla, L., Gracia-Lavedan, E., Mas, A., Bertran, S., Fierro, G., Seminario, A., Gómez, S., & Barbé, F. (2021). Prognostic value of symptoms at lung cancer diagnosis: A three-year observational study. *Journal of Thoracic Disease*, 13(3), 1485–1494. <https://doi.org/10.21037/jtd-20-3075>
- Pon, J. R., & Marra, M. A. (2015). Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10(1), 25–50. <https://doi.org/10.1146/annurev-pathol-012414-040312>
- Population growth in Canada’s rural areas, 2016 to 2021: Census of population, 2021.* (2022). Statistics Canada.

- Porta-Pardo, E., Valencia, A., & Godzik, A. (2020). Understanding oncogenicity of cancer driver genes and mutations in the cancer genomics era. *FEBS Letters*, *594*(24), 4233–4246. <https://doi.org/10.1002/1873-3468.13781>
- Posit team. (2024). *RStudio* (Version 2023.12.1.402) [Computer software]. Posit Software, PBC.
- Primm, K. M., Zhao, H., Hernandez, D. C., & Chang, S. (2022). Racial and Ethnic Trends and Disparities in NSCLC. *JTO Clinical and Research Reports*, *3*(8), 100374. <https://doi.org/10.1016/j.jtocrr.2022.100374>
- Punekar, S. R., Velcheti, V., Neel, B. G., & Wong, K.-K. (2022). The current state of the art and future trends in RAS-targeted cancer therapies. *Nature Reviews Clinical Oncology*, *19*(10), Article 10. <https://doi.org/10.1038/s41571-022-00671-9>
- Qi, R., Yu, Y., Shen, M., Lv, D., & He, S. (2022). Current status and challenges of immunotherapy in ALK rearranged NSCLC. *Frontiers in Oncology*, *12*, 1016869. <https://doi.org/10.3389/fonc.2022.1016869>
- Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. <https://doi.org/10.1186/1471-2164-13-341>
- R Core Team. (2023). *R: A Language and Environment for Statistical* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rami-Porta, R., Asamura, H., Travis, W. D., & Rusch, V. W. (2017). Lung cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, *67*(2), 138–155. <https://doi.org/10.3322/caac.21390>
- Reita, D., Pabst, L., Pencreach, E., Guérin, E., Dano, L., Rimelen, V., Voegeli, A.-C., Vallat, L., Mascaux, C., & Beau-Faller, M. (2022). Direct Targeting KRAS Mutation in Non-Small Cell Lung Cancer: Focus on Resistance. *Cancers*, *14*(5), 1321. <https://doi.org/10.3390/cancers14051321>
- Release notice—Canadian Cancer Statistics 2019. (2019). *Health Promotion and Chronic Disease Prevention in Canada*, *39*(8/9), 255–255. <https://doi.org/10.24095/hpcdp.39.8/9.04>

- Rigney, M. (2019). ES05.03 From Living Longer to Also Living Better; Managing Lung Cancer as a Chronic Disease—The Principle of Survivorship. *Journal of Thoracic Oncology*, 14(10), S25. <https://doi.org/10.1016/j.jtho.2019.08.089>
- Rogers, M. F., Shihab, H. A., Gaunt, T. R., & Campbell, C. (2017). CScape: A tool for predicting oncogenic single-point mutations in the cancer genome. *Scientific Reports*, 7(1), 11597. <https://doi.org/10.1038/s41598-017-11746-4>
- Rosell, R., Moran, T., Queralt, C., Porta, R., Cardenal, F., Camps, C., Majem, M., Lopez-Vivanco, G., Isla, D., Provencio, M., Insa, A., Massuti, B., Gonzalez-Larriba, J. L., Paz-Ares, L., Bover, I., Garcia-Campelo, R., Moreno, M. A., Catot, S., Rolfo, C., ... Taron, M. (2009). Screening for Epidermal Growth Factor Receptor Mutations in Lung Cancer. *New England Journal of Medicine*, 361(10), 958–967. <https://doi.org/10.1056/NEJMoa0904554>
- Ross, J. (1995). mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3), 423–450. <https://doi.org/10.1128/mr.59.3.423-450.1995>
- Sankar, K., Nagrath, S., & Ramnath, N. (2021). Immunotherapy for ALK-Rearranged Non-Small Cell Lung Cancer: Challenges Inform Promising Approaches. *Cancers*, 13(6), 1476. <https://doi.org/10.3390/cancers13061476>
- Sarris, E., Saif, M., & Syrigos, K. (2012). The Biological Role of PI3K Pathway in Lung Cancer. *Pharmaceuticals*, 5(11), 1236–1264. <https://doi.org/10.3390/ph5111236>
- Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews. Genetics*, 12(10), 683–691. <https://doi.org/10.1038/nrg3051>
- Scheffler, M., Bos, M., Gardizi, M., König, K., Michels, S., Fassunke, J., Heydt, C., Künstlinger, H., Ihle, M., Ueckerth, F., Albus, K., Serke, M., Gerigk, U., Schulte, W., Töpelt, K., Nogova, L., Zander, T., Engel-Riedel, W., Stoelben, E., ... Wolf, J. (2015). PIK3CA mutations in non-small cell lung cancer (NSCLC): Genetic heterogeneity, prognostic impact and incidence of prior malignancies. *Oncotarget*, 6(2), 1315–1326. <https://doi.org/10.18632/oncotarget.2834>
- Schubach, M., Maass, T., Nazaretyan, L., Röner, S., & Kircher, M. (2024). CADD v1.7: Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, 52(D1), D1143–D1154. <https://doi.org/10.1093/nar/gkad989>

- Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R., & Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. *Nature Communications*, *10*(1), 2569. <https://doi.org/10.1038/s41467-019-10489-2>
- Shi, C., Wang, Y., Xue, J., & Zhou, X. (2022). Immunotherapy for EGFR-mutant advanced non-small-cell lung cancer: Current status, possible mechanisms and application prospects. *Frontiers in Immunology*, *13*. <https://www.frontiersin.org/articles/10.3389/fimmu.2022.940288>
- Shreenivas, A., Janku, F., Gouda, M. A., Chen, H.-Z., George, B., Kato, S., & Kurzrock, R. (2023). ALK fusions in the pan-cancer setting: Another tumor-agnostic target? *Npj Precision Oncology*, *7*(1), Article 1. <https://doi.org/10.1038/s41698-023-00449-x>
- Simanshu, D. K., & Morrison, D. K. (2022). A Structure is Worth a Thousand Words: New Insights for RAS and RAF Regulation. *Cancer Discovery*, *12*(4), 899–912. <https://doi.org/10.1158/2159-8290.CD-21-1494>
- Song, P., Yang, D., Wang, H., Cui, X., Si, X., Zhang, X., & Zhang, L. (2020). Relationship between the efficacy of immunotherapy and characteristics of specific tumor mutation genes in non-small cell lung cancer patients. *Thoracic Cancer*, *11*(6), 1647–1654. <https://doi.org/10.1111/1759-7714.13447>
- Statistics Canada. (2023). *2021 Census of Population* (Nos. 98-316-X2021001) [Table]. <https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E>
- Strauch, Y., Lord, J., Niranjana, M., & Baralle, D. (2022). CI-SpliceAI—Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *PLOS ONE*, *17*(6), e0269159. <https://doi.org/10.1371/journal.pone.0269159>
- Strickler, J. H., Hanks, B. A., & Khasraw, M. (2021). Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better? *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *27*(5), 1236–1241. <https://doi.org/10.1158/1078-0432.CCR-20-3054>
- Strom, S. P. (2016). Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biology & Medicine*, *13*(1), 3–11. <https://doi.org/10.28092/j.issn.2095-3941.2016.0004>

- Subramanian, J., & Govindan, R. (2007). Lung Cancer in Never Smokers: A Review. *Journal of Clinical Oncology*, 25(5), 561–570. <https://doi.org/10.1200/JCO.2006.06.8015>
- Sun, S., Schiller, J. H., & Gazdar, A. F. (2007). Lung cancer in never smokers—A different disease. *Nature Reviews Cancer*, 7(10), 778–790. <https://doi.org/10.1038/nrc2190>
- Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, 156(6), 1324–1335. <https://doi.org/10.1016/j.cell.2014.01.051>
- Thai, A. A., Solomon, B. J., Sequist, L. V., Gainor, J. F., & Heist, R. S. (2021). Lung cancer. *The Lancet*, 398(10299), 535–554. [https://doi.org/10.1016/S0140-6736\(21\)00312-3](https://doi.org/10.1016/S0140-6736(21)00312-3)
- The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., ... Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. <https://doi.org/10.1093/nar/gkac1052>
- Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., Chirieac, Lucian. R., Dacic, S., Duhig, E., Flieder, D. B., Geisinger, K., Hirsch, F. R., Ishikawa, Y., Kerr, K. M., Noguchi, M., Pelosi, G., Powell, C. A., Tsao, M. S., & Wistuba, I. (2015). The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of Thoracic Oncology*, 10(9), 1243–1260. <https://doi.org/10.1097/JTO.0000000000000630>
- Tri-council policy statement: Ethical conduct for research involving humans.* (2022). Secretariat on Responsible Conduct of Research.
- Tsoulos, N., Papadopoulou, E., Metaxa-Mariatou, V., Tsaousis, G., Efstathiadou, C., Tounta, G., Scapeti, A., Bourkoula, E., Zarogoulidis, P., Pentheroudakis, G., Kakolyris, S., Boukovinas, I., Papakotoulas, P., Athanasiadis, E., Floros, T., Koumariou, A., Barbounis, V., Dinischiotu, A., & Nasioulas, G. (2017). Tumor molecular profiling of NSCLC patients using next generation sequencing. *Oncology Reports*, 38(6), 3419–3429. <https://doi.org/10.3892/or.2017.6051>

- Tuteja, S., Kadri, S., & Yap, K. L. (2022). A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. *Journal of Pathology Informatics*, *13*, 100130. <https://doi.org/10.1016/j.jpi.2022.100130>
- Vena, J. E., Sultz, H. A., Carlo, G. L., Fiedler, R. C., & Barnes, R. E. (1987). Sources of Bias in Retrospective Cohort Mortality Studies: A Note on Treatment of Subjects Lost to Follow-up. *Journal of Occupational Medicine*, *29*(3), 256–261.
- Vestergaard, L. K., Oliveira, D. N. P., Poulsen, T. S., Høgdall, C. K., & Høgdall, E. V. (2021). OncoPrint™ Comprehensive Assay v3 vs. OncoPrint™ Comprehensive Assay Plus. *Cancers*, *13*(20), 5230. <https://doi.org/10.3390/cancers13205230>
- Vollbrecht, C., Lenze, D., Hummel, M., Lehmann, A., Moebs, M., Frost, N., Jurmeister, P., Schweizer, L., Kellner, U., Dietel, M., & von Laffert, M. (2018). RNA-based analysis of ALK fusions in non-small cell lung cancer cases showing IHC/FISH discordance. *BMC Cancer*, *18*, 1158. <https://doi.org/10.1186/s12885-018-5070-6>
- Wang, X., Ricciuti, B., Nguyen, T., Li, X., Rabin, M. S., Awad, M. M., Lin, X., Johnson, B. E., & Christiani, D. C. (2021). Association between Smoking History and Tumor Mutation Burden in Advanced Non–Small Cell Lung Cancer. *Cancer Research*, *81*(9), 2566–2573. <https://doi.org/10.1158/0008-5472.CAN-20-3991>
- Waters, A. M., Bagni, R., Portugal, F., & Hartley, J. L. (2016). Single Synonymous Mutations in KRAS Cause Transformed Phenotypes in NIH3T3 Cells. *PLoS ONE*, *11*(9), e0163272. <https://doi.org/10.1371/journal.pone.0163272>
- Watterson, A., & Coelho, M. A. (2023). Cancer immune evasion through KRAS and PD-L1 and potential therapeutic interventions. *Cell Communication and Signaling*, *21*(1), 45. <https://doi.org/10.1186/s12964-023-01063-x>
- Wolf, J., Seto, T., Han, J.-Y., Reguart, N., Garon, E. B., Groen, H. J. M., Tan, D. S. W., Hida, T., de Jonge, M., Orlov, S. V., Smit, E. F., Souquet, P.-J., Vansteenkiste, J., Hochmair, M., Felip, E., Nishio, M., Thomas, M., Ohashi, K., Toyozawa, R., ... Heist, R. S. (2020). Capmatinib in MET Exon 14–Mutated or MET-Amplified Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, *383*(10), 944–957. <https://doi.org/10.1056/NEJMoa2002787>
- Yan, N., Guo, S., Zhang, H., Zhang, Z., Shen, S., & Li, X. (2022). BRAF-Mutated Non-Small Cell Lung Cancer: Current Treatment Status and Future Perspective. *Frontiers in Oncology*, *12*, 863043. <https://doi.org/10.3389/fonc.2022.863043>

- Zacharias, M., Absenger, G., Kashofer, K., Wurm, R., Lindenmann, J., Terbuch, A., Konjic, S., Sauer, S., Gollowitsch, F., Gorkiewicz, G., & Brcic, L. (2021). Reflex testing in non-small cell lung carcinoma using DNA- and RNA-based next-generation sequencing—A single-center experience. *Translational Lung Cancer Research, 10*(11), 4221–4234. <https://doi.org/10.21037/tlcr-21-570>
- Zappa, C., & Mousa, S. A. (2016). Non-small cell lung cancer: Current treatment and future advances. *Translational Lung Cancer Research, 5*(3), 288–300. <https://doi.org/10.21037/tlcr.2016.06.07>
- Zeng, Z., & Bromberg, Y. (2019). Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Frontiers in Genetics, 10*. <https://www.frontiersin.org/articles/10.3389/fgene.2019.00914>
- Zhang, W., & Liu, H. T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research, 12*(1), 9–18. <https://doi.org/10.1038/sj.cr.7290105>
- Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2014). CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics, 30*(7), 1006–1007. <https://doi.org/10.1093/bioinformatics/btt730>
- Zhong, Y., Xu, F., Wu, J., Schubert, J., & Li, M. M. (2021). Application of Next Generation Sequencing in Laboratory Medicine. *Annals of Laboratory Medicine, 41*(1), 25–43. <https://doi.org/10.3343/alm.2021.41.1.25>
- Zhou, Z., & Li, M. (2022). Targeted therapies for cancer. *BMC Medicine, 20*(1), 90, s12916-022-02287–3. <https://doi.org/10.1186/s12916-022-02287-3>

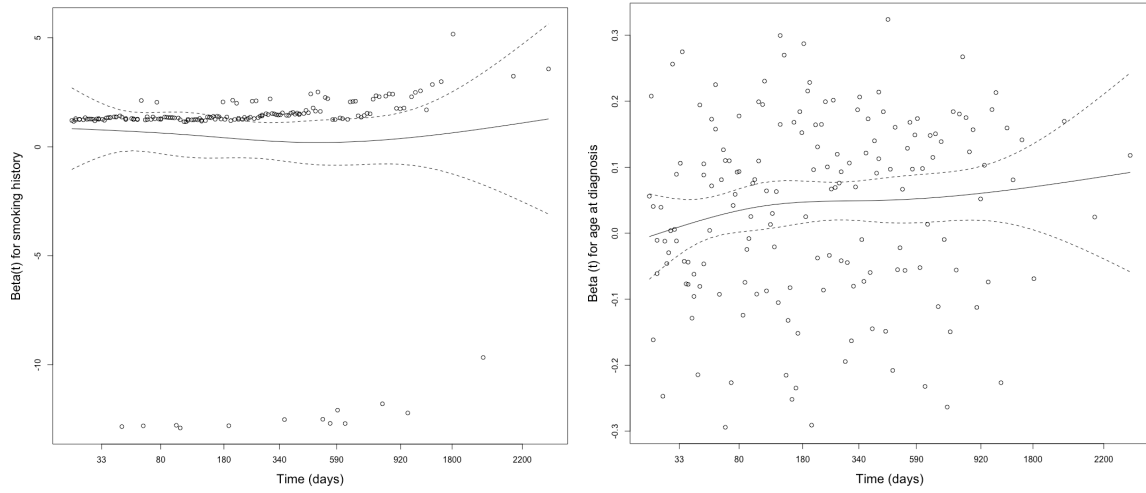
Appendix A – Synonymous variants

Table A1: All synonymous variants in cohort with predictions for splicing impact. Pathogenicity scores are included.

Gene	Variant	SpliceAI		Cscope			CADD Phred score	TRaP Score	Occurrence	Allele frequency/range	
		Donor loss	Acceptor gain	Donor gain	Score	Prediction					Confidence
ALK	chr2:29193302 G>A	0	0	0	0.15779	Neutral	Low	7.804	0.001	2	0.482966 - 0.58041
ALK	chr2:29193230 C>G	0	0	0	0.328795	Neutral	Low	8.922	0.071	1	0.0492754
ALK	chr2:29213992 G>A	0	0	0.05	0.656541	Oncogenic	Low	11.45	0.131	1	0.0440613
ALK	chr2:29220757 G>A	0	0	0	0.274993	Neutral	Low	3.457	0.103	1	0.635135
AR	chrX:67721932 G>A	0	0	0	0NA	NA	NA	6.096	0.052	1	0.0667362
BRAF	chr7:140801453 A>G	0	0	0	0.208679	Neutral	Low	8.451	0.076	8	0.0435 - 0.0520916
BRAF	chr7:140794401 T>C	0.03	0	0	0.203643	Neutral	Low	15.81	0.056	1	0.0554562
CTNNA1	chr3:41224602 C>T	0	0	0	0.297301	Neutral	Low	11.56	0.075	1	0.0466165
CTNNA1	chr3:41224648 C>T	0	0	0	0.376643	Neutral	Low	11.42	0.07	1	0.0398162
CTNNA1	chr3:41224656 T>C	0	0	0	0.299543	Neutral	Low	5.055	0.015	1	0.06893
EGFR	chr7:55191841 G>A	0	0	0.05	0.188506	Neutral	Low	2.405	0.114	2	0.047 - 0.488924
EGFR	chr7:55181331 G>A	0	0	0	0.158837	Neutral	Low	15.34	0.101	1	0.0503751
ERBB2	chr17:39725096 C>T	0.02	0	0	0.200651	Neutral	Low	15.33	0.004	1	0.356955
ERBB3	chr12:56085090 C>G	0	0	0	0.239366	Neutral	Low	9.763	0.123	1	0.332666
FGFR2	chr10:121488024 G>A	0	0	0	0.453985	Neutral	Low	12.3	0.055	2	0.0671551 - 0.157079
FGFR2	chr10:121488060 C>T	0	0	0	0.268496	Neutral	Low	9.221	0.007	1	0.0540541
FGFR2	chr10:121488075 T>C	0	0	0	0.185595	Neutral	Low	1.254	0.036	1	0.298047
FGFR2	chr10:121515246 G>A	0.01	0	0	0.641029	Oncogenic	Low	12.63	0.104	1	0.0643821
FGFR2	chr10:121515279 G>A	0	0.02	0	0.243083	Neutral	Low	7.093	0.155	1	0.0861423
FGFR2	chr10:121520099 G>A	0	0	0	0.227988	Neutral	Low	3.912	0.039	1	0.0789474
FGFR3	chr4:1799396 G>A	0.1	0	0	0.157056	Neutral	Low	2.922	0.04	1	0.0959128
FGFR3	chr4:1799444 C>G	0.01	0	0	0.168004	Neutral	Low	7.438	0.005	1	0.0574074
FGFR3	chr4:1801857 C>T	0.14	0.01	0	0.271777	Neutral	Low	16.03	0.068	1	0.242424
FGFR3	chr4:1804418 G>A	0.44	0.01	0	0.148019	Neutral	Low	18.47	0.427	1	0.37302
FGFR4	chr5:177091038 G>A	0	0.01	0	0.173331	Neutral	Low	1.716	0.019	1	0.049459
FGFR4	chr5:177097325 G>T	0	0.01	0	0.172533	Neutral	Low	0.046	0.019	1	0.233746
GNA11	chr19:3115037 C>T	0	0.01	0.01	0.330139	Neutral	Low	13.43	0.043	1	0.106818
JAK3	chr19:17835201 C>T	0.08	0	0	0.212693	Neutral	Low	5.292	0.01	2	0.237164 - 0.405108
JAK3	chr19:17835147 A>G	0.04	0	0	0.14592	Neutral	Low	0.553	0.075	1	0.320905
JAK3	chr19:17838052 C>T	0	0.01	0	0.157897	Neutral	Low	7.105	0.053	1	0.5675
KIT	chr4:54727255 G>A	0.01	0	0	0.246892	Neutral	Low	8.102	0.006	1	0.304
KRAS	chr12:25227398 C>T	0	0	0	0.366425	Neutral	Low	11.52	0.069	1	0.0662139
MAP2K1	chr15:66435123 G>A	0	0.01	0	0.189629	Neutral	Low	9.135	0.065	1	0.0539773
MAP2K1	chr15:66481774 C>T	0	0	0	0.291842	Neutral	Low	10.25	0.031	1	0.08742
MED12	chrX:71129386 G>A	0	0	0	0NA	NA	NA	11.13	0.18	1	0.2755
MET	chr7:116763082 C>T	0.06	0.05	0	0.339185	Neutral	Low	8.614	0	1	0.0544629
MET	chr7:116763091 C>T	0.06	0	0	0NA	NA	NA	3.817	0	1	0.0565553
MTOR	chr1:11124536 A>G	0	0	0	0.281485	Neutral	Low	2.293	0.014	1	0.561
MTOR	chr1:11129809 C>A	0.02	0	0	0.202674	Neutral	Low	0.097	0.011	1	0.0706215
MTOR	chr1:11130766 C>A	0	0	0	0.285252	Neutral	Low	0.02	0.01	1	0.209605
MYC	chr8:127738274 C>T	0.12	0.19	0	0.159967	Neutral	Low	12.48	0.131	1	0.626
MYC	chr8:127740673 C>T	0	0	0	0NA	NA	NA	11.34	0.2	1	0.201601
MYC	chr8:12774091 G>C	0	0	0	0NA	NA	NA	8.578	0.027	1 (0)	0.0505506
NF1	chr17:31358501 G>A	0	0.03	0	0.667882	Oncogenic	Low	8.931	0.104	1	0.0566038
NRAS	chr1:114713886 T>C	0	0	0	0.505196	Oncogenic	Low	12.59	0.009	4	0.0503597 - 0.0715686
PDGFRA	chr4:54274846 G>C	0.06	0.02	0.01	0.387299	Neutral	Low	1.325	0.038	1	0.41362
PDGFRA	chr4:54274924 G>A	0	0	0	0.289236	Neutral	Low	5.57	0.004	1	0.0461165
PIK3CA	chr3:179199143 C>A	0	0	0	0.283186	Neutral	Low	11.38	0.026	1	0.0580871
PIK3CA	chr3:179204568 A>G	0	0	0	0.520249	Oncogenic	Low	12.35	0.027	1	0.162
SMO	chr7:129206556 G>T	0	0	0	0.20412	Neutral	Low	7.827	0.016	1	0.15421
SMO	chr7:129210460 C>T	0	0	0	0.160916	Neutral	Low	9.94	0.006	1	0.16325
SMO	chr7:129210486 C>A	0	0	0	0.272064	Neutral	Low	5.765	0.105	1	0.234281
SMO	chr7:129209384 C>A	0	0.01	0	0.02	0.261754	Neutral	12.79	0.429	1	0.0460526

Appendix B – Schoenfeld residuals

Overall survival



Progression-free survival

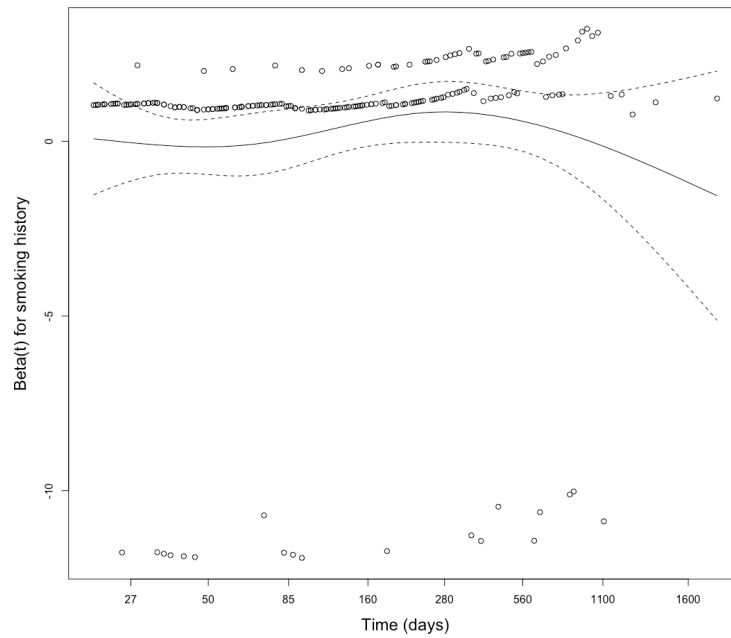


Figure B1: Schoenfeld residuals for significant variables in overall and progression-free survival which do not violate the proportional hazard assumption.

Curriculum Vitae

Candidate's Full Name: Kathleen Mary Varty

Universities Attended (with dates and degrees obtained): University of New Brunswick, Bachelor of Science, Biology – Chemistry, 2022

Publications:

Varty, K., O'Brien, C., & Ignaszak, A. (2021). Breast Cancer Aptamers: Current Sensing Targets, Available Aptamers, and Their Evaluation for Clinical Use in Diagnostics. *Cancers*, 13(16), 3984. <https://doi.org/10.3390/cancers13163984>.

O'Brien, C., Varty, K., & Ignaszak, A. (2021). The electrochemical detection of bioterrorism agents: a review of the detection, diagnostics, and implementation of sensors in biosafety programs for Class A bioweapons. *Microsystems & nanoengineering*, 7, 16. <https://doi.org/10.1038/s41378-021-00242-5>.

Conference Presentations:

MacLean, L., Varty, K., Cutler, S., Xu, Z., Reiman, T., Gaston, D., & Boudreau, J. (2024). Natural Killer cells for precision therapy in non-small cell lung carcinomas with KRAS, TP53, and STK11 co-mutated tumours. Dalhousie University Pathology Research Day, Halifax, Canada.

Varty, K., Itani, D., Hossain, M., Acar, C., Michael, J., Daigle-Maloney, T., Thompson, R., Johnston, B., Russell, C., Gaston, D., & Reiman, T. (2023). Evaluating genomic variants identified by a 50 gene panel used for identification of actionable mutations in non-squamous non-small cell lung cancers (NSCLC). Canadian Cancer Research Conference, Halifax, Canada.