The qRT-PCR Assay and the Genomic/Epigenomic Properties of the 10-Gene Yin Yang Expression Ratio Signature in Non-Small Cell Lung Cancer

by

Shavira Narrandes

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics

University of Manitoba

Winnipeg

Abstract

Lung cancer is the leading cause of cancer-related death in North America with a 5-year survival rate less than 20%. The ability to understand which lung cancer patients will progress and predict treatment responses will improve its management. Previously, our lab has shown a 10-gene YMR signature (*GRM1, RECQL4, NRAS,* and *IGFBP5* are over-expressed and *HOXA5, TNNC1, SOSTDC1, CRIP2, CD83,* and *GATA2* are under-expressed in tumor cells) correlates with prognosis and treatment prediction for non-small cell lung cancer (NSCLC). To further develop our signature for clinical use, other factors that regulate gene expression, such as changes in genomic sequences (mutations or copy number) or epigenomic factors (methylation of DNA), need to be investigated. Multiple linear regression models demonstrated that the combination of DNA methylation and copy number variation (CNV) correlate with gene expression for *RECQL4, NRAS, IGFBP5, HOXA5, TNNC1, SOSTDC1*, and *CRIP2*. Next, we evaluated six gene expression assay systems (qRT-PCR, DNA microarray, NanoString nCounter, RNA-seq, FISH, and tissue microarray) in a literature review to obtain our signature; qRT-PCR was determined to be the most feasible in a clinical setting. To validate our signature using qRT-PCR, we used an A549 cell line and lung tumor FFPE test samples obtained from the Manitoba Tumor Bank. *IGFBP5* had the lowest mRNA expression level compared to *TNNC1, CRIP2,* and *GATA2* and *CRIP2* had the highest mRNA expression level in the A549 cell line, contrary to our expected signature. However, the expression levels of these genes correlated with the signature in lung tumor FFPE samples. To further confirm these results, I assessed 37 and 29 NSCLC cell lines from the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) databases, respectively, for the YMR signature. Only four cell lines (NCIH2066, NCIH810, NCIH292, and IA-LM) were similar to the expected signature expression pattern. These gene expression results indicate that cell lines may not be ideal for assessing the YMR

signature. In the future, we will need to determine the 10-gene signature in NSCLC tumor samples in a prospective study of lung cancer patients using qRT-PCR to predict prognosis and treatment response.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Spencer Gibson, for welcoming me into his lab as a graduate student to continue and finish my Master of Science degree when my past supervisor was no longer able to. I greatly appreciate the advice and guidance given to me throughout my training and work, as well as the continued support of my career goals. I would also like to thank my past supervisor, Dr. Wayne Xu, for training me in Bioinformatics since 2015, thus beginning my career in research. I am grateful for being pushed out of my comfort zone at an early stage, allowing me opportunities to do presentations and publish manuscripts.

I am very thankful for the time and constructive criticism, as well as support of my future career goals, given to me by my committee members and mentors, Dr. Shantanu Banerji and Dr. Leigh Murphy. Their guidance helped me understand the diverse areas of my research and allowed me to learn from my mistakes and improve on my work.

A special thank you to my lab manager, Elizabeth Hensen, for encouraging me throughout my time as a graduate student and teaching me all the lab techniques I used, as well as helping me troubleshoot the many errors I came across. Thank you to all the current and past members of the Gibson lab for creating such a friendly and fun learning environment, teaching me lab techniques and tricks, and patiently listening to my frustrations.

I am appreciative of my past lab mate, Shujun Huang, for helping me in the Bioinformatics area over the years, teaching me the many tricks, scripts, and codes used in my computational biology work, and being a kind and supportive friend. I would also like to give a special thanks to Dr. Sanzida Jahan for helping me with my initial experiments and Andrea Fristensky for promptly preparing the many FFPE samples I used.

Lastly, I would like to thank my family and friends for their continued support, encouragement, and love, making these last couple of years easier as a graduate student.

I acknowledge the financial support received from the CancerCare Manitoba Foundation and Research Manitoba/CancerCare Manitoba Masters Studentship.

Dedication

This thesis is dedicated to my late father, Dr. Rohitsingh Narrandes, a lover of science and medical innovations whose footsteps I hope to follow in.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5-Aza-2’-deoxycytidine (5-Aza-CdR)

5-methylcytosine (m^5C)

Acute myeloid leukemia (AML)

Adenocarcinoma (ADC)

Adjuvant chemotherapy (ACT)

American Joint Committee on Cancer (AJCC)

Anaplastic lymphoma kinase (ALK)

Basal cell carcinoma (BCC)

Bone morphogenic protein (BMP)

Cancer Cell Line Encyclopedia (CCLE)

Chronic lymphocytic leukemia (CLL)

Chronic obstructive pulmonary disease (COPD)

Colorectal cancer (CRC)

Complimentary DNA (cDNA)

Computed tomography (CT)

Copy number variation (CNV)

Cycle threshold (Ct)

Cysteine rich protein 2 (CRIP2)

Cytosine-guanine (CpG)

Database of Genomic Variants (DGV)

Death-associated protein kinase (DAPK)

Diffuse large B-cell lymphoma (DLBCL)

Double-stranded DNA (dsDNA)

DNA methyltransferase (DNMT)

Epidermal growth factor receptor (EGFR)

Epithelial-to-mesenchymal transition (EMT)

Estrogen receptor (ER)

Fluorescent *in situ* hybridization (FISH)

Formalin-fixed, paraffin-embedded (FFPE)

Fresh frozen (FF)

GATA binding protein 2 (GATA2)

Genomics of Drug Sensitivity in Cancer (GDSC)

Glutamate metabotropic receptor 1 (GRM1)

Hematoxylin and eosin (H&E)

Hepatocyte growth factor (HGF)

Histone acetyltransferase (HAT)

Histone deacetylase (HDAC)

Histone methyltransferase (HMT)

Homeobox A5 (HOXA5)

In situ hybridization (ISH)

Insulin-like growth factor (IGF)

Insulin-like growth factor binding protein 5 (IGFBP5)

Kilobase (kb)

Kirsten rat sarcoma virus (KRAS)

Long non-coding RNA (LncRNA)

Magnetic resonance imaging (MRI)

Merkel cell carcinoma (MCC)

MicroRNA (miRNA)

Multiple linear regression (MLR)

Multiple permutation process (MPP)

Non-small cell lung cancer (NSCLC)

Optimal cutting temperature (OCT)

Phospholipase C (PLC)

Positron emission tomography (PET)

Protein Kinase C (PKC)

Quantitative reverse transcription polymerase chain reaction (qRT-PCR)

Risk of relapse (ROR)

Rolling circle amplification (RCA)

RNA sequencing (RNA-Seq)

Sclerostin domain containing 1 (SOSTDC1)

Sequencing by synthesis (SBS)

Severe acute respiratory syndrome (SARS)

Simple linear regression (SLR)

Single nucleotide polymorphism (SNV)

Small cell lung cancer (SCLC)

Squamous cell carcinoma (SQC)

Tissue microarray (TMA)

Troponin (Tn)

Troponin C (Tn-C)

Troponin C1 (TNNC1)

Tumor-educated blood platelet (TEP)

Vascular endothelial growth factor (VEGF)

Vascular endothelial growth factor receptor (VEGFR)

World Health Organization (WHO)

Yin Yang Mean Ratio (YMR)

Chapter 1: Introduction

1.1 Lung Cancer Background

Lung cancer is the most common cancer-related cause of death worldwide, having a 5-year survival rate of 15% and being the leading cause for men and second highest for women.[1–3] These rates vary depending on patient sex, age, race or ethnicity, socioeconomic status, and geography. Smoking is attributable to 80-90% of all lung cancers and has been shown to increase the risk of development by five- to ten-fold, while non-smokers have a 20% increased risk when exposed to environmental tobacco smoke. Therefore, there is a variation in global lung cancer rates according to the history of smoking and tobacco uptake and reduction patterns. The highest rates of lung cancer occur where smoking uptake began the earliest, such as North America and Europe, and are increasing in countries where smoking uptake occurred at a later period. Other factors that may cause lung cancer include genetic predispositions (e.g. family history of lung cancer), infections, history of chronic obstructive pulmonary disease (COPD), second-hand smoke and exposure to asbestos, radon, ionizing radiation, diesel, and environmental toxins (e.g. pesticides, exhaust).[4–6]

1.1.1 Lung Cancer Stratification

The World Health Organization (WHO) has classified 50 different heterogeneous lung cancer variants based on certain genetic and biologic characteristics to ensure patients receive the proper treatment for the best disease outcome.[7] By recognizing specific tumor morphology properties under a light microscope, four main histological lung cancer subtypes emerged. Squamous cell carcinomas (SQCs) and small cell tumors arise from epithelial cells lining larger airways and account for 30% and 18%, respectively, of lung cancers. Adenocarcinomas (ADCs) arise from

epithelial cells lining small peripheral airways and account for 30% of lung cancers.[8] They are the most common lung cancer subtype found in non-smokers.[9] Large cell carcinomas contribute to about 10% of lung cancers, are poorly differentiated, and generally located in the periphery of the lung. Adenocarcinomas, squamous cell carcinomas, and large cell carcinomas together comprise the non-small cell lung cancer (NCSLC) subtype that encompasses about 85% of all lung cancers. NSCLCs and small cell lung cancers (SCLCs) present with unique morphology, treatment, and patient clinical course and outcome from each other. SCLCs demonstrate a predictable clinical course of an initial favourable chemotherapy response associated with several months of remission, while its short-term regression is followed by cancer recurrence, development of resistance to chemotherapy, and metastasis. NSCLCs are more complex in their diagnosis and prediction of patient outcome. Although their initial tumor diagnosis is based on small bronchoscopic biopsy specimens, the diagnosis may change following their removal and most patients are diagnosed after the disease has metastasized from the originating site.[8, 10] NSCLC patients have a relapse rate of 40% within five years of treatment and 50% die from the metastatic disease after surgical removal of the tumor.[11, 12] Surgical resection proves to be the most effective treatment for patients in their early stages of the disease, while adjuvant chemotherapy (ACT) increases the survival rate by 4-15%.[13] However, 30-60% of patients presenting with stages IB to IIIA may relapse and die within 5 years after diagnosis.[10, 14]

It is generally accepted that driver gene mutations initialize cancer development. Proto-oncogenes, such as *Ras*, are transcribed into products for cellular proliferation such as receptors, growth factors, transcription factors, and signaling enzymes.[15] Gain-of-function mutations in

proto-oncogenes result in dominant oncogenes that differ from their proto-oncogenes or are over-expressed and occur due to point mutations, localized reduplication, or chromosomal translocation. An oncogene therefore disrupts a cell's normal activity and can lead to uncontrolled cell division, and ultimately cancerous cells. Genes for the inhibition of cell growth and survival are termed tumor-suppressor genes and include *APC* and *TP53*; the loss or under-expression of these genes also results in an uncontrolled cell division and survival.[15–17]

Therefore, by analyzing gene expression levels and the biological pathways associated with the genes involved in a cancer, one can study the difference between normal cell and cancerous cell pathways to determine the genetic origin of the faulty pathway, thereby identifying potential targets for treating cancer. The potential subtypes of that cancer can be identified through class discovery, the identification of novel cancer subtypes, and class prediction, the assignment of tumor samples to pre-defined classes in order to aid in predicting the outcome.[18] Gene expression analysis also allows for biomarker and gene signature discovery. The use of gene expression profiling and development of gene biomarkers/signatures for cancer allows for the diagnosis, progression and aggressiveness analyses, prognosis, prediction of therapeutic treatment, and/or identification of patients who would benefit from therapeutic treatment and to better understand the disease and its biology.[17, 19]

1.1.2 Lung Cancer Staging

The American Joint Committee on Cancer (AJCC) developed the commonly used TNM staging system for the most effective treatment selection and prognosis prediction of cancers which includes the size of the tumor, tumor location, lymph node involvement, and whether and where

to the cancer has metastasized. T describes the size of the primary tumor and is given a ranking of 0 to 4 (T0-T4), with T0 meaning no indication of primary tumor and a higher number indicating the cancer has grown deeper into an organ or spread to nearby tissues. Other stages include TX (primary tumor cannot be assessed) and Tis (early in situ carcinoma that has not spread to nearby tissue). N represents whether the cancer has spread to the regional lymph nodes of the organ and is given a ranking of 0 to 3 (N0-N3). N0 indicates there is no involvement of lymph nodes, a value of 1-3 represents the number and/or extent of the lymph node involvement, and NX means the lymph node involvement cannot be assessed. Lastly, M specifies the degree of cancer metastasis via blood or the lymphatic system, with M0 being no distant metastasis and M1 spreading to other areas of the body. After the determination of the T, N, and M rankings, the values are combined and an overall cancer staging of 0, I, II, III, and IV is output.

For NSCLC, Stage 0 signifies a localized, in situ carcinoma; Stage I a tumor that is between 3 and 4 cm or less in size; Stage II a tumor that is no more than 5 cm in size and may have spread to regional lymph nodes or bronchi; Stage III a tumor that may be larger than 5 cm in size, spread to some degree, and/or more than 1 tumor present; and lastly Stage IV a cancer that has metastasized and possibly grown into two or more tumors outside the chest. In some cases, the stages are subdivided using letters to specify the tumor size or degree of metastasis. For example, Stage IIIA NSCLC represents a cancer that has spread to lymph nodes on the same side of the chest as the primary tumors and Stage IIIB NSCLC a cancer that has metastasized to lymph nodes on the opposite side or above the clavicle. Stage I cancers are simpler in terms of minimal size, lymph node involvement, and metastasis, allowing patients to have better prognoses. Cancers of higher stages are more complex but still have the ability to be treated successfully.[20–23]

1.1.3 Lung Cancer Diagnosis and Treatment

The diagnosis and staging of NSCLCs are based on a variety of tests that include physical examinations, imaging (X-rays, computed tomography (CT), positron emission tomography (PET), and/or magnetic resonance imaging (MRI) scans), laboratory tests (blood, urine, and/or tissue analyses), pathology reports, and/or surgical reports or removed samples. CT scans of the thorax and upper abdomen are commonly performed for clinical staging; however, they are limited in their detection of microscopic metastasis. PET scans using fluorine 18-labelled fluorodeoxyglucose are very sensitive in detecting metabolically active and malignant cancers and their results are usually confirmed by mediastinoscopy.[24–26]

The main forms of treatment for NSCLC patients include surgery, chemotherapy, radiation therapy, and targeted therapy. NSCLC patients of Stages I and II tend to benefit from surgical resection, while patients presenting with a more advanced form of the disease are favourable candidates for non-surgical treatments.[13, 27] Surgery includes pneumonectomy, the removal of an entire lung, lobectomy, a section or lobe of the lung is removed, segmentectomy, part of a lung lobe is removed, and laser surgery, the use of a high-energy beam to destroy cancer cells. Chemotherapy is beneficial in advanced and metastatic disease states and may be used before or after surgical resection or to prevent relapse.[28–30] NSCLCs are commonly treated using a combination of two drugs, which can be platinum-based or non-platinum-based. Platinum-based chemotherapies are considered the standard of care and include cisplatin and carboplatin, which can crosslink with DNA purine bases, obstructing DNA repair mechanisms and causing DNA damage, ultimately inducing apoptosis. Non-platinum-based therapies are used to avoid the undesirable toxicities obtained from platinum-based chemotherapies.[29, 31–34] Radiation

therapy encompasses external beam radiation therapy, which uses high energy beams to kill cancer cells, and brachytherapy, the placement of a radioactive material in or near a tumor that gradually kills the cancer cells. It is commonly combined with chemotherapy in chemoradiation. Studies have shown that cisplatin-based chemotherapy combined with radiation therapy increases patient survival rates compared to those who solely receive radiation therapy.[29, 35] Lastly, targeted therapies use drugs specific to the genetic mutations found in the cancerous cells or tissues. For example, an epidermal growth factor receptor (EGFR) over-expression confers of poor prognosis to 40-80% of NSCLC patients. This mutation can be targeted by EGFR inhibitors such as Erlotinib.[29, 36, 37] The activation of vascular endothelial growth factor receptor 2 (VEGFR2) by vascular endothelial growth factor (VEGF) affects mitogenesis, angiogenesis, and vascular permeability. Therefore, the exploitation of this interaction may have cancerous effects. The combination of chemotherapy and bevacizumab targeting VEGF has demonstrated in increased survival time in patients with advanced NSCLCs.[29, 38]

1.2 Biomarkers and Biology in Lung Cancer

The heterogeneous nature of NSCLC makes it difficult to classify patients. The evaluation of prognostic biomarkers, mutations statuses of genes, and identification of gene signatures are key to overcome the heterogeneity in lung cancer. To further develop the field and provide the best outcomes for patients, lung cancer biomarkers are needed to enable early diagnosis of potentially curable tumors, for the selection of early and late stage patients for effective therapies, and for the stratification of patients with an unfavourable prognostic outcome to identify additional or more efficient therapies.[18, 19] A number of studies have assessed various genetic biomarkers in lung cancers. *EGFR*, a tyrosine kinase, plays a role in regulating cell proliferation, apoptosis,

and motility. Upon the binding of one of six ligands to EGFR, its C-terminal tail becomes phosphorylated, resulting in interactions between the receptor and its downstream effectors related to the PI3K and MAPK signalling pathways.[39] Over-expression and/or improper activation of *EGFR* has the potential to prompt mechanisms related to carcinogenesis, such as increased cell proliferation, survival, and metastasis. Adenocarcinomas harbour 95% of *EGFR* mutations, with it being associated with overall survival in these patients.[7] FDA-approved *EGFR* inhibitors for NSCLC patients include Erlotinib, Afatinib, and Gefitinib.[36, 37, 40, 41] Therapies are also available or in development for patients presenting with abnormal expression of Anaplastic Lymphoma Kinase (*ALK*) and Kirsten Rat Sarcoma Virus proto-oncogene (*KRAS*), which can predict shorter prognoses in patients. Unfortunately, these targeted therapies are only partially effective or patients often relapse.[42, 43]

1.3 Biology of Biomarker Expression

1.3.1 Cancer Cell Traits

A number of cancerous traits have been described that allow the cancer cells to avoid standard processes involving proliferation, differentiation, and apoptosis, among others. Normal cells bind mitogenic growth signals via their cell transmembrane receptors, allowing them to proliferate. Antiproliferative signals, such as immobile and soluble growth factors, also bind to normal cells to inhibit cellular proliferation. Cells are either forced from the mitotic cycle into the G_0, or quiescent, state with the ability to re-enter the mitotic cycle in the future or induced into post-mitotic states. These normal cells can develop into cancerous cells that continually divide, proliferate, and may invade other areas of the body through a variety of mechanisms.[44, 45] Gene mutations may produce in abnormal cells; mutations that result in cancer development are

called driver mutations while passenger mutations are those that do not result in the cancer phenotype.[46, 47] Genome instability and high rates of gene mutations and chromosome alterations also alter the gene products that regulate the diversity of cellular functions. For example, tumor cells are able to evade apoptosis, or programmed cell death, have been shown to be self-sufficient in the generation of their own growth signals, reducing their dependence on growth factors from the environment, and have the ability to avoid antiproliferative signals, making them capable of continued mitotic divisions. Cancer cells also display a sustained angiogenesis, or development of vasculature, to provide nutrients to a growing cancer. In more advanced stages of cancers, the cells have the potential to invade other tissues and metastasize to other areas of the body. Several key proteins are altered to provide cancer cells with this ability. For example, cell-to-cell adhesion molecules (CAMs) and integrins regulate cell-to-cell interactions and bind cells to the extracellular matrix, respectively.[44, 45, 47]

1.3.2 Gene Expression Modulators

1.3.2.1

Various DNA mutations may occur that can change the presence or dosage of a gene, ultimately affecting the expression of a gene with the potential to lead to disease. Silent mutations are those that result in the coding of the same amino acid, making the change redundant. Nonsense mutations result in a stop codon, likely leading to a non-functional protein. Missense mutations change the nucleotide sequence and its corresponding codon and amino acid, while frameshift mutations change the reading frame of a nucleotide sequence, also resulting in unintended amino acids. They may be the result of point mutations which encompass base pair substitution (missense mutation that replaces a nucleotide, altering the resulting code for an amino acid),

insertion (the insertion of a nucleotide that usually results in a frameshift and misreading of a codon), and deletion (the removal of a nucleotide leading to a frameshift). Chromosomal mutations affect a portion of a chromosome and include inversions (a region of a chromosome is flipped), deletions (a region of a chromosome is lost, resulting in an absence of genes), duplication (a region of a chromosome is multiplied, resulting in a gene dose increase), and translocation (a region of a chromosome is moved to another chromosome). Lastly, copy number variations result in an increase or decrease in the dosage of a gene.[48, 49]

1.3.2.2 Copy Number Variation

Single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), inversions, deletions, insertions, duplications, and translocations are alterations in a human genome that modify gene expression levels. This, in turn, can affect cellular proliferation, differentiation, fitness, and clonal selection to potentially result in diseased phenotypes such as cancer, diabetes, HIV-1, and heart disease.[50, 51] CNVs are DNA segments equal to or greater than 1 kilobase (≥1 kb) in size that vary between individuals in comparison to the human reference genome. Insertions, deletions, and amplifications can be referred to as CNVs and whole genes may be affected. Gene expression can increase or decrease by deletions and amplifications, changing the dosage of a gene, and a CNV overlapping or disrupting a gene.[52–54] Several studies have assessed the affects of CNVs of certain genes in different cancers. Shlien et al. mapped CNVs with loci that are related to those of cancer-related genes, calling them cancer CNVs, and demonstrated that 49 cancer genes were affected by CNVs.[55] Forbes et al. assessed the Database of Genomic Variants (DVG) and found that 40% of cancer-related genes are affected by CNVs, many being oncogenes and tumor suppressor genes with functions in apoptosis, cell cycle regulation, and

DNA repair.[56] Studies have also shown that lung adenocarcinoma and squamous cell carcinoma subtypes have different patterns of CNVs, indicating that different expression levels of some genes are implicated between the lung cancer subtypes.[57]

1.3.2.3 DNA Methylation

Epigenetics refers to heritable, yet reversible, stable alterations in gene expression through DNA methylation and histone modification.[58, 59] DNA methylation involves a covalent interaction between a methyl group and carbon 5 of the cytosine ring in a cytosine-guanine (CpG) dinucleotide pair via DNA methyltransferase (DNMT) enzymes; the resultant structure is called 5-methylcytosine (m^5C). The methylation is not uniform throughout the human genome but occurs in 0.5-5 kb regions called CpG islands.[58, 60] The presence of m^5C affects gene expression in several ways. It may inhibit proteins from binding to DNA; for example, transcription factors that bind to CpG pairs will be prevented from binding to their designated sites by the presence of a methyl moiety. It also results in transcriptionally active euchromatin changing into inactive heterochromatin as the chromatin condenses, making genes inaccessible for transcription and expression. Other consequences include genomic imprinting and tissue-specific silencing of gene expression.[58–62] Histone deacetylases (HDACs), histone acetyltransferases (HATs), and histone methyltransferases (HMTs) also affect the epigenetic regulation of genes. For example, the methylation of the K4 amino acid residue of the H3 histone via an HMT results in euchromatin formation, or the de-condensation of chromatin to allow for gene transcription.[58, 63] Discrepancies have been found between the DNA methylation patterns of normal and malignant cells.[60] Through altering the accessibility of genes from transcription factors, encoded proteins that function in genome stability, cell metastasis, and

healthy cell functioning may be affected, resulting in cancer disease states. Hypermethylation of the promoter regions of tumor suppressor genes results in gene silencing and the promotion of oncogenesis. Alternatively, hypomethylation, generally of repeated DNA sequences, allows for cancer-related genes to be expressed, sometimes at greater than normal levels.[58, 60] *RARß, RASSF1A, CDNK2A, CHD13, APC* are some commonly methylated genes implicated in lung cancer.[64] Similarly, *ER, PR,* and *BRCA1*, are well-known genes commonly presenting in abnormal levels in breast cancers that have shown hypermethylation.[65] In relation to a larger number of reports of hypermethylation than hypomethylation in cancers, it has been shown that DNMTs are greatly increased in breast, colon, and prostate cancers, to name a few. DNMT1 and DNMT3B have elevated levels in malignant cells of varying cancer types. Therefore, through the reversal of hypermethylation, therapeutic treatments can be produced targeting the implicated regions. For example, 5-Aza-2'-deoxycytidine (5-Aza-CdR) and antisense oligonucleotides are inhibitors for DNMT1 and DNMT3B. 5-Aza-CdR incorporates into DNA to inactivate DNMTs while antisense oligonucleotides induce degradation of DNMT1 mRNA, reducing its active levels in cells.[58, 66, 67]

1.4 Yin Yang Mean Ratio (YMR) Model

1.4.1 Gene Signatures in Lung Cancer

Many studies have previously established prognostic gene signatures based on gene expression levels in NSCLC patients. Chen et al. assessed *DUSP6, MMD, STAT1, ERBB3,* and *LCK* genes in surgically resected frozen samples from 125 adenocarcinoma and squamous cell carcinoma patients, which were randomly assigned into the training or testing datasets, using microarray analysis and/or real-time RT-PCR. By means of risk scores and decision-tree analyses, they were

able to demonstrate a correlation between the expression levels of their signature genes and relapse-free and overall patient survival. Although validation with an independent cohort of 60 patients proved significant, further analyses are needed to assess the benefit of Cisplatin-based adjuvant chemotherapy in those patients stratified with this 5-gene signature.[12] Shahid et al. developed an 8-gene signature through the use of a Cox proportional hazard regression model to determine its prognostic significance. The *STAT1, CLU, GTSE1, NUSAP1, ABCA8, TNNT1, ENTPD3,* and *CPA3* genes were found in both the training and testing datasets. Patients designated into low- and high-risk scores demonstrated differing heatmaps expression patterns and overall survival based on the 8-gene signature.[10] Similarly, Yu et al. separated 112 patients into training and testing datasets then used a Cox regression to identify expressed microRNAs (miRNAs) imperative in determining the prognosis of NSCLC patients. Patient tumor samples presenting with a high-risk score of the 5-miRNA signature (hsa-let-7a, hsa-miR-221, hsa-miR-137, hsa-miR-372, and hsa-miR-182) demonstrated increased cancer relapse and decreased survival.[68] The common approach is to determine the correlation coefficients between gene expression and patient survival time using training datasets and then using testing datasets to validate or normalize to the trained data. This method, however, tends to result in problems of low reproducibility, disallowing the signature(s) to be used in a clinical setting. Our previously established Yin Yang Mean Ratio (YMR) gene signature avoids using data training and instead hypothesizes that the opposing effects of two groups of genes, the Yin genes (over-expressed in lung tumor cells: *GRM1, RECQL4, NRAS, IGFBP5*) and Yang genes (over-expressed in lung normal cells: *HOXA5, TNNC1, SOSTDC1, CRIP2, CD83, GATA2*), determine a patient's prognosis and can guide treatment selection.[69, 70]

1.4.2 Construction and Validation

The Yin and Yang genes were first selected using unsupervised clustering and pathway analyses to compare gene expression data from normal and tumor lung tissue samples; 31 Yin and 32 Yang genes resulted. The YMR was calculated as patient risk scores, with normal tissues demonstrating values less than 1.0 and lung cancer tissue values greater than 1.0. Patients were significantly stratified into high- and low-risk groups when four independent datasets were used to assess the YMR's applicability in predicting lung cancer prognosis. The YMR also predicted chemotherapy outcomes for cancer stages II and III and was a better predictor of clinical outcomes than the commonly used clinical factors, excluding tumor stage.[69] The multiple permutation process (MPP) was used to reduce the initial 63 Yin and Yang genes to an optimal number that would generate the most beneficial results. The MPP first produced 10 000 combinations of Yin and Yang genes, with gene lists ranging from 2 to 32 genes having at most a difference of two genes between the Yin and Yang lists. All Yin and Yang gene list combinations were tested using 1000 randomly permutated gene expression datasets with a sampling of 200 patient cases; combinations that gave a Cox regression p-value of less than 0.05 were kept. Permutations of 1 million Yin and Yang lists of the fixed gene size (that which could produce the highest number of p-values less than 0.05) were then tested against 1 million random samplings of gene expression datasets. Again, combinations that gave a Cox regression p-value of less than 0.05 were kept. The genes were ranked based on how often they appeared in the lists and those genes with the best rankings were used to test 1 million randomly permutated gene expression datasets. The chosen YMR had the lowest p-value and hazard ratio greater than 1.0. Since the gene expression levels can be measured in individual patients, the YMR signature demonstrates extensive application in clinical settings.[70]

1.4.3 YMR Genes

Glutamate metabotropic receptor 1 (GRM1) activates phospholipase C (PLC) through its binding to L-glutamate, an excitatory neurotransmitter in the central nervous system.[71, 72] PLC, as well as its phospholipase family members (PLA and PLD), are key mediators of intracellular and intercellular signalling. Activated PLC hydrolyzes phosphatidyl inositol 1,4-bisphosphate into inositol triphosphate and diacylglycerol, which in turn regulate calcium concentration in the cytosol and protein kinase C (PKC) activity, respectively. It may ultimately affect cellular proliferation and differentiation.[73, 74] Irregularities in the expression of *GRM1* have been implicated in various diseases, such as schizophrenia, bipolar disorder, and breast cancer.[72] RECQL4 is a member of the family of RECQ helicases, which aid in maintaining the structure and integrity of DNA and may modulate chromosome segregation.[75] The helicases bind double-stranded DNA (dsDNA) then unwind the strands during DNA replication in cell division and repair of DNA damage.[76, 77] RECQL4 specifically functions in the initiation of DNA replication, is mainly expressed in the enterocytes lining the intestines, thymus, and testis, and may play a role in cell development in bones and skin.[75, 78] The Ras family contains the NRAS, HRAS, and KRAS proteins that function in cell division, differentiation, and apoptosis. NRAS is a GTPase that is activated by bound GTP, converts the GTP to GDP, then is turned off by the bound GDP. It transmits signals via signal transduction from outside the cell to the cell's nucleus to ultimately affect cell proliferation and differentiation.[79–83] The family of insulin-like growth factor binding proteins (IGFBPs) bind to and regulate the stability of insulin-like growth factors I and II (IGF-I, IGF-II). Once bound to cell surface receptors, such as the IGF-1 receptor, IGF-II receptor, and insulin receptor, IGFs regulate cellular growth, differentiation, and apoptosis and general development, and metabolism. IGFBP5 has been shown to function in cell

growth and adhesion, determination of cell fate, apoptosis, and metastasis during cancer development.[84–86]

The homeobox genes are found in four clusters (A, B, C, D) on different chromosomes and their products comprise a family of DNA-binding transcription factors. The homeobox A5 (HOXA5) transcription factor from cluster A on chromosome 7 regulates gene expression and cell morphogenesis and differentiation. It has also been shown to upregulate the tumor suppressor *p53*, implying that a downregulation of *HOXA5* may cause an increase in cancer formation and tumorigenesis.[87, 88] Troponin (Tn) regulates striated muscle contraction and is found on the actin filament with tropomyosin. It is made of three subunits: troponin I (Tn-I) which inhibits actomyosin ATPase, troponin T (Tn-T) which holds the tropomyosin binding site, and troponin C (Tn-C). Tn-C binds to calcium to stop the inhibitory action of Tn so that myosin can interact with the actin filaments and ATP can be hydrolyzed, resulting in tension generation of the muscle. In non-muscle cells, troponin C type 1 (TNNC1) functions in cell motility, cytoplasmic streaming, and cytokinesis. [89, 90] Sclerostin domain containing 1 (*SOSTDC1*), a member of the sclerostin family, is translated into a N-glycosylated protein. It then acts as a bone morphogenic protein (BMP) antagonist by binding to and inhibiting BMP from interacting with its receptors, affecting its signalling in cell proliferation, differentiation, and apoptosis.[91, 92] Cysteine rich protein 2 (*CRIP2*) is a transcription factor containing two LIM zinc-binding domains that has the potential to regulate the differentiation of smooth muscle tissues. It also regulates actin-rich structures, with a possible role in actin dynamics and/or cell migration.[93, 94] CD83 is a part of the immunoglobulin family of receptors, found on the cell membrane. It has the potential to solubilize and bind to dendritic cells to inhibit their maturation and plays a

role in regulating immune system development via antigen presentation.[95–97] The GATA family of zinc-finger transcription factors contains GATA binding protein 2 (GATA2) which functions in the development and proliferation of hematopoietic and endocrine cells. Abnormal functioning of GATA2 has been found in myelodysplastic syndrome, acute myeloid leukemia (AML), and KRAS-driven NSCLC.[98, 99]

1.5 Biomarker Expression Techniques for Clinical Use

Through the analysis of gene expression levels and the biological pathways associated with the genes involved in a cancer, one can study the difference between normal cell and cancerous cell pathways to determine the genetic origin of the faulty pathway, thereby identifying potential targets for treating cancer. The possible subtypes of that cancer can be identified through class discovery, the identification of novel cancer subtypes, and class prediction, the assignment of tumor samples to pre-defined classes in order to aid in predicting the outcome. Gene expression analysis also allows for biomarker and gene signature discovery. The use of gene expression profiling and development of gene biomarkers/signatures for cancer allows for the diagnosis, progression and aggressiveness analyses, prognosis, prediction of therapeutic treatment, and/or identification of patients who would benefit from therapeutic treatment to better understand the disease and its biology.[17]

Various assay technologies have been developed for gene expression analysis. For example, qRT-PCR amplifies the RNA expression of a gene of interest and uses fluorescent probes or dyes to depict the gene expression; NanoString nCounter uses probe pairs to anneal to a region of RNA and detect gene expression; DNA microarrays allow cDNA targets to hybridize to probes

on a solid slide and can use fluorescence to detect genes; Illumina MiSeq RNA-Seq bridge amplifies nucleic acid samples to create clusters, which are then interpreted by the MiSeq system; and lastly, tissue microarrays (TMAs) study circular punches from tissue sample blocks with labelled probes or antibodies to determine the gene expression.[100–109] Each of these assays has been used in diagnosis, prognosis, and treatment prediction for a variety of cancers. Some assays have been commercialized for cancer clinical use. For example, Afirma® is a microarray test for thyroid cancer diagnosis and the Oncotype DX qRT-PCR test is for guiding breast cancer treatment. Many assays are under investigation in clinical trials or studies.[17, 110, 111]

However, numerous challenges involved with tumor sample collection, experimental design and determining the proper assay to use, analytical and diagnostic factors (such as the interpretation of samples and biomarker performance analysis), assay detection limits and specificity, drug development for rare cancers, and clinical distribution of a significant biomarker must be overcome to develop a novel clinical assay for cancer patients. Biomarker development is also naturally affected by a patient's natural biology and history, intratumor heterogeneity, cancer progression, and germ-line mutations. This results in a prolonged period from the biomarker discovery to patenting and clinical translation stages.[17, 19, 112–114]

1.6 Thesis Rationale, Hypothesis, and Aims

1.6.1 Rationale

Since the 10-gene signature showed clinical relevance in prognosis and prediction, discovering other genomic and epigenomic factors, such as DNA mutation, CNV, and DNA methylation, that

can regulate these 10 genes may provide targets or modulators for this YMR signature. To further develop our 10-gene signature, we first aimed to choose a gene expression detection assay most feasible for clinical oncology use. A comprehensive assessment on the qRT-PCR, DNA microarray, nCounter, RNA-Seq, FISH, and tissue microarray (TMA) assays will aid in the selection of the most suitable method for each clinical application. The qRT-PCR assay involves parameters including primer design, RNA quality, and PCR conditions. Assay optimization will ensure the assay will work for FFPE samples, as well as assay reproducibility. To further demonstrate the reproducibility of our YMR model, the assay will be tested on some test FFPE samples from the Manitoba Tumor Bank and to determine if the low- and high-risk early stage NSCLC patients can be significantly stratified. Gene expression data of NSCLC cell lines downloaded from the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) databases will be assessed for the YMR gene signature expression pattern and compared to the qRT-PCR results to further validate the signature.

1.6.2 Hypothesis

We hypothesize the 10-gene signature *in-silico* correlates with genomic and epigenomic factors, such as DNA mutation, copy number variation (CNV), and DNA methylation using computational methods and the 10-gene signature can be reproduced using qRT-PCR in lung cancer cell lines and tumor samples.

1.6.3 Aims

This research aims to:

1) Explore the genomic and epigenomic covariates that are correlated to and modulate the expression levels of the 10 YMR signature genes.

2) Identify a biomarker expression technique that is feasible for clinical use.
3) Validate the YMR gene signature with the chosen biomarker expression technique on the A549 cell line and lung tumor FFPE samples.

Chapter 2: Methods and Materials

2.1 Simple and Multiple Linear Regressions to Determine the Covariables Associated with Gene Expression Levels

2.1.1 Data Collection and Preprocessing

Multi-dimensional data (level 3) including gene expression, mutation, and DNA methylation for the lung adenocarcinoma (LUAD) cohort were downloaded from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) and copy number variation (CNV) from the University of California Santa Cruz (UCSC) Xena browser (https://xenabrowser.net/) for each of the ten YMR genes (*GRM1*, *RECQL4*, *NRAS*, *IGFBP5*, *HOXA5*, *TNNC1*, *SOSTDC1*, *CRIP2*, *CD83*, and *GATA2*) (Table 1). mRNA expression data was downloaded in the form of aligned sequencing reads from the Illumina HiSeq 2000 RNA Sequencing (RNA-Seq) platform, preprocessed by HTSeq for Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values. The gene-level estimates were presented as log2(RSEM+1)-transformed RNA-Seq by Expectation Maximization (RSEM) counts. Gene mutation data was downloaded in the form of nonsense, missense, frame shift deletion, splice site, and silent mutations. Silent mutations were designated as "no" and given a value of 0 since they are neutral mutations that do not result in a change in phenotype; the remaining mutations were designated as "yes" for causing a change in phenotype and given a value of 1. Sequencing data calls were produced on an IlluminaGA system at the Broad Institute Genome Sequencing Center using the MuTect method. Methylation beta values from the Illumina Infinium HumanMethylation27 platform were downloaded for use as DNA methylation data. The beta values were recorded for each array probe per sample using the BeadStudio software. The values range from 0 to 1 and represent the ratio of intensity of the methylated bead to the intensity of the combined locus, with lower levels signifying

hypomethylation and higher values hypermethylation. For CNV data production, the copy number profile (CNP) was first measured at a TCGA genome characterization center using the whole genome microarray (Affymetrix Genome-Wide Human SNP Array 6.0 platform). Segmented CNV data output by the application of the GISTIC2 method to the CNP measurements was then mapped to genes to provide gene-level estimates.

Table 1: Information pertaining to downloaded TCGA LUAC patient data.

Data Type	Platform	Sample Size
mRNA expression	Illumina HiSeq 2000	512
Gene mutation	IlluminaGA	34
DNA methylation	Illumina Infinium HumanMethylation27	455
CNV	Affymetrix Genome-Wide Human SNP Array 6.0	518

2.1.2 Data Cleaning

Data cleaning was done using the UNIX server via PuTTY Secure Shell (SSH).

2.1.2.1 Gene Expression

mRNA expression data was downloaded in a matrix with TCGA patient IDs set as column headings and gene ID numbers and names as row names. A file with gene ID numbers and their corresponding gene names were downloaded. The gene ID numbers were removed from the row names so that only the gene names remained. The *GRM1*, *RECQL4*, *NRAS*, *IGFBP5*, *HOXA5*, *TNNC1*, *SOSTDC1*, *CRIP2*, *CD83*, and *GATA2* genes were selected from the Gene Expression file using the "grep" command. From the resulting file, patients with "01A" in the "sample" portion of their IDs, corresponding to primary solid tumor from sample vial A, were selected

using the “grep” command. After the patient IDs were filtered, only the first three portions of the ID were kept (project, tissue source site (TSS), and participant) to match the patient IDs in the mutation, methylation, and CNV datasets. The final matrix was constructed with the 512 patient IDs as row names and gene symbols as column headings.

2.1.2.2 Mutation

The *GRM1*, *RECQL4*, *NRAS*, *IGFBP5*, *HOXA5*, *TNNC1*, *SOSTDC1*, *CRIP2*, *CD83*, and *GATA2* mutation files were selected from the downloaded TCGA mutation files using the command “grep “[gene symbol]” *.maf.txt”. The separate files were then placed into one large file and any genes not apart of the 10-gene signature were removed using the command “grep -v [gene symbol]”. Using the “sed” command, genes with nonsense mutations, missense mutations, frame shift deletions, and splice sites were set as positive (“Yes”) for mutations and genes with silent mutations were set as negative (“No”) for mutations. The “Yes” and “No” were ultimately set to values of 1 and 0, respectively. The final matrix was constructed with the 34 patient IDs as row names and gene symbols as column headings.

2.1.2.3 Methylation

The *GRM1*, *RECQL4*, *NRAS*, *IGFBP5*, *HOXA5*, *TNNC1*, *SOSTDC1*, *CRIP2*, *CD83*, and *GATA2* genes were selected from the Methylation file using the “grep” command. From the resulting file, patients with “01A” in the “sample” portion of their IDs were selected using the “grep” command. After the patient IDs were filtered, only the first three portions of the ID were kept to match the patient IDs in the gene expression, mutation, and CNV datasets. Methylation Beta values of NA were changed to 0 before the mean Beta value was calculated per gene. The final

matrix was constructed with the 455 patient IDs as row names and gene symbols as column headings.

2.1.2.4 CNV

The CNV data pertaining to the 10 Yin Yang genes were downloaded in the proper matrix format, with the patient IDs as row names and gene symbols in their correct order as column headings. The patient ID column was separated and only the first three portions of the ID were kept to match the patient IDs in the gene expression, mutation, and methylation datasets. The resulting file of the edited 518 patient IDs were then combined with the CNV data file into the final matrix.

2.1.3 Simple Linear Regression Model

Simple linear regression (SLR) models were constructed for each of the 10 YMR signature genes to test the correlation of an independent genomic or epigenomic factor with gene expression. For each gene, three SLR models were built to assess the correlation of the following explanatory factors with gene expression for all patients with available data: (1) gene mutation, (2) DNA methylation, and (3) CNV. The mutation, methylation, and CNV data matrices were filtered to keep the patient IDs common with the gene expression data matrix. The regression model can be formulated as equation (1), where Y_i represents the expression in the tumor *i* for a given gene and X_i the gene's mutation/DNA methylation/CNV value in tumor *i*. For SLRs based on DNA methylation and CNV, a random sampling of 2/3 of the dataset was first used to train the SLR model in R using the *lm* function and the remaining 1/3 of the dataset was used to test the model. SLRs based on mutation were only conducted on the entire dataset due to small sample size.

SLRs based on DNA methylation and CNV were also trained on patients from different disease stages (stage I, stage IA, stage IB, stage II, stage IIA, stage IIB, stage III, stage IIIA, stage IV, stage I to II, and stage I to III).

Equation 1: $Y_i = b_0 + b_1X_i$

Y_i = response variable for the predicted gene expression value

b_0 = regression coefficient; intercept of the regression line on the Y-axis; mean of the probability distribution when X = 0

b_1 = regression coefficient; slope of the regression line; change in the mean of the probability distribution of Y per unit increase in X

X_i = value of the predictor variable

2.1.4 Multiple Linear Regression Model

Using the *lm* R function, multiple linear regression (MLR) models were constructed for each of the 10 YMR signature genes to test the correlation of a combination of predictor variables with the gene expression data. For each given gene, the following predictor variable combinations were used to construct two MLR models with gene expression as the response variable: (1) mutation and methylation and (2) CNV and methylation. This can be formulated as equation (2), where Y_i is the expression in the tumor *I* for the given gene, X_{1i} and X_{2i} are DNA methylation and mutation or CNV for the given gene in tumor *i*. The mutation, methylation, and CNV data matrices were filtered to keep the patient IDs common with the gene expression data matrix. A random sampling of 2/3 of the dataset was first used to train the SLR model. The model was then tested using the remaining 1/3 of the dataset. MLRs of the first model form was only conducted on the entire dataset due to small sample size. Each MLR of the second model form was first conducted on the entire dataset, and then varying combinations of disease stages (stage I, stage

IA, stage IB, stage II, stage IIA, stage IIB, stage III, stage IIIA, stage IV, stage I to II, and stage I to III).

Equation 2: $Y_i = b_0 + b_1X_{1i} + b_2X_{2i}$

Y_i = response variable for the predicted gene expression value

b_0 = regression coefficient; intercept of the regression line on the Y-axis; mean of the probability distribution when X = 0

b_1 = regression coefficient; slope of the regression line; change in the mean of the probability distribution of Y per unit increase in X

X_i = value of the predictor variable

2.1.5 Model Performance

A model's statistical significance was assessed based on its p-value, mean absolute percentage error (MAPE) of the training dataset, and MAPE of the testing dataset. The p-value was first assessed to determine if probability of the results is due to chance is less than 5% ($p < 0.05$). The MAPE assess the difference between the measured and predicted data, with ideal values being minimal and similar between the training and testing datasets. This can be formulated as equation (3). Summaries of the regression models' results were called using the *summary* R function. Included were the distribution of residuals (minimum, first quantile, median, third quantile, maximum), residual standard error of the model, multiple R^2, adjusted R^2, F-statistic and p-value of the model, and estimated values, standard errors, and t-values with their associated probabilities of the regression coefficients.

Equation 3:

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

N = number of fitted points

A_t = actual expression values

F_t = fitted/predicted expression values (Y_i)

2.2 Determination of Gene Expression Detection Assay Feasible for Clinical Use

Six gene expression assay systems (qRT-PCR, DNA microarray, nCounter, RNA-Seq, FISH, and tissue microarray) that are currently being used in clinical cancer studies were investigated in a literature review to provide a guideline for choosing an assay method with respect to its oncological applications in a clinical setting. Some of these methods are also commonly used in a modified way; for example, detection of DNA content or protein expression. Their principles, sample preparation, design, quantification and sensitivity, data analysis, time for sample preparation and processing, and cost are discussed. The methods were also compared according to their sample selection, particularly for the feasibility of using FFPE samples, which are routinely archived for clinical cancer studies.

2.3 qRT-PCR Validation of the YMR 10-Gene Signature

2.3.1 Reagents

All reagents were of analytical or research grade. Reagents were obtained from one of the following sources: Bio-Rad Laboratories Inc., Biotium, Integrated DNA Technologies (IDT)

Inc., Millipore Sigma (Roche), New England BioLabs Ltd., Qiagen, and Thermo Fisher Scientific (Applied Biosystems, Fisher Scientific, and Invitrogen).

2.3.2 RNA Extraction from A549 Cell Line

2.3.2.1 A549 Cell Culture

The A549 cell line was cultured to 80% confluence in Dulbecco's Modified Eagle Medium (DMEM) with high glucose concentration (4500 mg/L) and (400 mML L-glutamine, sodium pyruvate). DMEM was supplemented with 10% Fetal Bovine Serum (FBS) and 5 mL of penicillin/streptomycin from a 100 mL bottle penicillin/streptomycin that contained 10 000 units/mL penicillin and 10 000 µg/mL streptomycin. Cell culture pellets were stored at -80˚C prior to RNA extraction.

2.3.2.2 RNA Extraction Protocol

Reagents, materials, and protocol were retrieved from the Qiagen RNeasy Plus Mini Kit for RNA Extraction. For cell lysis, 600 µL Buffer RLT was added to the A549 cell pellet; the sample was vortexed to dissolve the pellet. The lysate was pipetted into a QIAshredder spin column placed in a 2 mL collection tube, which was then centrifuged for 2 minutes at 10 000 rpm. An aliquot of 600 µL 70% ethanol was added to the lysate and the resulting solution was mixed by pipetting up and down, followed by transferring 700 µL of the sample to a RNeasy spin column in a 2 mL collection tube. The solution was centrifuged for 20 seconds at 8000 x g and the flow through was discarded; this step was repeated with the remainder of the sample. To bind the RNA to the column and wash the sample of non-bound particles, such as proteins and carbohydrates, 700 µL Buffer RW1 was added and the column was centrifuged at 10 000 rpm for 20 seconds; the flow through was discarded. To remove excess salts from the sample and wash

the spin column membrane, 500 µL Buffer RPE was added and the column was centrifuged at 8000 x g for 20 seconds; the flow through was discarded and this step was repeated with another aliquot of Buffer RPE. The spin column was placed in a new 2 mL collection tube and centrifuged for 1 minute at 16 000 x g. The spin column was again placed in a new 2 mL collection tube; 50 µL of RNase-free water was added and the column was centrifuged for 1 minute at 10 000 rpm for RNA elution. For the last RNA extraction from the A549 cell line performed, the lysate from the last step was pipetted back onto the spin column and centrifuged again for 1 minute at 10 000 rpm to provide a more concentrated RNA sample.

2.3.3 RNA Extraction from FFPE Samples

2.3.3.1 Paraffin Removal Protocol

Three lung cancer formalin-fixed, paraffin-embedded (FFPE) test tissue samples were collected from the Manitoba Tumor Bank in 1.5 mL Eppendorf tubes on four different occasions (total of 12 samples) (Table 2). For paraffin removal, 800 µL xylene was added to each tube then vortexed. Aliquots of 400 µL, followed by 1000 µL, of absolute ethanol were added to the tubes and centrifuged at 16 000 x g for 2 minutes to wash the sample. The supernatant was discarded using a pipette following each centrifugation step, leaving a pellet. The tubes were opened, blotted on a paper towel, and left to dry in a heating block set to 55˚C for 10 to 20 minutes.

Table 2: Clinical information of lung cancer test FFPE samples.

Sample	Sex	Patient Age	Surgery/Fixation Date	Cancer Subtype
13137	Female	NA	1995/1996	NA
L0304	Female	67 years	2014	Squamous cell carcinoma
L0379	Male	79 years	2017	Adenocarcinoma

2.3.3.2 RNA Extraction Protocol

Reagents and protocol were retrieved from the Roche High Pure FFPET RNA Isolation Kit. RNA extraction began with adding a mixture of 100 µL RNA tissue lysis buffer to break the cell membranes, 16 µL 10% SDS, and 40 µL 44.44 mg/mL Proteinase K. The tubes were briefly spun in a centrifuge, placed in a shaking heating block set to 85˚C at 600 rpm for 30 minutes, and again briefly spun in a centrifuge. The tubes were cooled to 55˚C before adding an aliquot of 80 µL Proteinase K. The tubes were briefly spun in a centrifuge, placed in the shaking heating block set to 55˚C at 600 rpm for 30 minutes, and again briefly spun in the centrifuge. At this point, the lysate was generally clear. If portions of the pellet remained, as was the case with some extractions, the incubation at 55˚C at 600 rpm was extended for 10 minutes.

To bind RNA to the column and wash the sample, 325 µL RNA binding buffer and 325 µL absolute ethanol were added, respectively, to each tissue lysate. The mixture was vortexed in three second intervals then centrifuged briefly. This lysate was pipetted onto the upper reservoir of a High Pure Filter Tube placed in a High Pure Collection Tube. The tubes were centrifuged for 30 seconds at 6000 x g. The lysate was re-pipetted onto the Filter Tube and the Collection Tube was spun again for 30 seconds at 6000 x g. To effectively dry the Filter Tube fleece, the Filter Tube was placed in a new Collection Tube and centrifuged for 2 minutes at 16 000 x g.

The Filter Tube was placed onto a new Collection Tube. For the degradation of contaminating DNA, 103 µL of DNase I (Invitrogen Amplification Grade) was prepared and pipetted onto the Filter Tube fleece. The sample was left to incubate for 15 minutes at room temperature (15˚C to 25˚C). To wash the sample, 500 µL Wash Buffer I was pipetted onto the Filter Tube and the sample was centrifuged at 6000 x g for 30 seconds. The lysate was re-pipetted onto the Filter

Tube and the centrifugation was repeated, followed by discarding the flow through. An aliquot of 500 µL Wash Buffer II was pipetted onto the Filter Tube and the sample was centrifuged at 6000 x g for 30 seconds. The lysate was re-pipetted onto the Filter Tube and the centrifugation was repeated, followed by discarding the flow through. The sample was then centrifuged for 2 minutes at 16 000 x g to dry the Filter Tube Fleece. The Filter Tube was then placed in a fresh 1.5 mL reaction tube. For the recovery of the RNA sample, 40 µL (out of a designed 25 to 50 µL) RNA Elution Buffer was pipetted onto the Filter Tube; the sample was incubated for 1 minute at room temperature (15˚C to 25˚C) then centrifuged for 1 minute at 6000 x g. The lysate was re-pipetted onto the Filter Tube, left to incubate for 1 minute, then the centrifugation step was repeated.

2.3.4 Measurement of RNA Quality

The NanoDrop ND-1000 Spectrophotometer was used to measure the RNA quality extracted from the A549 cell culture and FFPE tissue samples. The quality was assessed by the output A260/A280 and A260/A230 values from a 2 µL sample placed on the instrument pedestal. The final sample concentration was also noted.

2.3.5 cDNA Synthesis

2.3.5.1 iScript™ Reverse Transcriptase Supermix for RT-PCR Kit

For cDNA synthesis, 150 ng of RNA extracted from the A549 cell line and FFPE lung tumor test samples (L0379B, L0304F, 13137A) was reverse transcribed with 4.0 µL of 5x Reverse Transcriptase (RT) and enough nuclease-free, RNase-free double-distilled water (ddH_2O) to make a 20.0 µL total solution. A no RT control was produced by replacing the 4.0 µL RT with

an equivalent volume of no RT solution. The thermal cycler was set to 25˚C for 5 minutes, 42˚C for 30 minutes, 85˚C for 5 minutes, and then held at 4˚C. Remaining cDNA was stored at -20˚C for future use.

2.3.5.2 Applied Biosystems™ High Capacity cDNA Reverse Transcription Kit

For cDNA synthesis, 150 ng of RNA extracted from the A549 cell line and FFPE lung tumor test samples (L0379B, L0304F, 13137A) was reverse transcribed with a master mix of 2.0 µL 10x reverse transcriptase buffer, 0.8 µL 25x dNTP mix (100 mM), 2.0 µL 10x reverse transcription random primers, 1.0 µL MultiScribe™ Reverse Transcriptase, 1.0 µL RNase inhibitor, and enough nuclease-free, RNase-free ddH_2O to make a 20.0 µL total solution. A no RT control was produced by replacing the 1.0 µL RT with an equivalent volume of nuclease-free, RNase-free ddH_2O. The thermal cycler was set to 25˚C for 10 minutes, 37˚C for 120 minutes, 85˚C for 5 minutes, and then held at 4˚C. Remaining cDNA was stored at -20˚C for future use.

2.3.6 qRT-PCR Protocol

2.3.6.1 Primer Design

Linear mRNA accession numbers associated with the 10 testing and 3 housekeeping genes were acquired from the National Center for Biotechnology Information (NCBI) through an advanced nucleotide search specifying for gene name and homo sapiens species. The accession numbers were then used as input for the NCBI Primer-BLAST primer designing tool. The product (amplicon) size was initially set to 80-100 base pairs (bps), however, when ideal primers were not found, the amplicon size was set to 80-120 bps. The primer melting temperature (Tm) was set to 60˚C-64˚C, with an optimum value of 62˚C and maximum T_m difference of 2˚C. When ideal primers were not output, the T_m was set to 58˚C-64˚C, with an optimum value of 61˚C. An

exon-exon junction span was chosen. The chosen exon at 5' side and exon at 3' side ("number of bases that must anneal to exons at the 5' or 3' side of the junction") values varied according to which output the most favourable primer(s). Primer size was set to 20-25 bps, with an optimum value of 23 bps. GC content was set to 40%-60% and expanded to 35%-65% when no ideal primers were output. The maximum self complementarity and maximum pair complementarity were varied to obtain primers with low or no hairpin, self-dimer, or hetero-dimer formation capabilities.

The chosen primer pairs were checked for primer positioning, amplicon size, and presence of one polymerase chain reaction (PCR) product using the UCSC Genome Browser *In silico* PCR. The human reference genome hg19 was used. The possibility of hairpin, self-dimer, and hetero-dimer formation was checked using the Integrated DNA Technologies (IDT) OligoAnalyzer.

Ideal primers demonstrated a single product on intended target and low self complementarity values (≤2.0). Greater self-complementarity values were accepted when the primer pair showed a low chance of hairpin, self-dimer, and hetero-dimer formation. Final primer designs chosen for qRT-PCR tests are displayed in Tables 3-5.

Table 3: Sequence and parameters of IDT custom-designed qRT-PCR primers for Yin genes.

Gene, primer orientation	Amplicon Size (bp)	T_m Range (˚C)	Primer Size (bp)	GC Content (%)	Self-complementarity	Self-3'-complementarity
GRM1, forward	99	60.69	20	55.00	5.00	3.00
GRM1, reverse		61.96	20	66.00	5.00	2.00
Forward primer sequence: GGGCAGGGAATGCCAATTCT Reverse primer sequence: CAGAGAGGCGGTGCCACATA						
RECQL4, forward	80	60.62	20	60.00	2.00	0.00
RECQL4, reverse		61.28	20	60.00	5.00	2.00
Forward primer sequence: AGGAAGAGGAAGGGCAGGAG Reverse primer sequence: CCAATCCTGGAGTCTGGCCT						
NRAS, forward	84	62.21	22	54.55	2.00	0.00
NRAS, reverse		61.95	21	52.38	4.00	0.00
Forward primer sequence: GCTTGAGGTTCTTGCTGGTGTG Reverse primer sequence: TGTCAGTGCGCTTTTCCCAAC						
IGFBP5, forward	95	63.72	20	60.00	2.00	2.00
IGFBP5, reverse		63.96	23	52.17	5.00	2.00
Forward primer sequence: CTCAAAGCCAGCCCACGCAT Reverse primer sequence: CGGGAAGGTTTGCACTGCTTTCT						

Table 4: Sequence and parameters of IDT custom-designed qRT-PCR primers for Yang genes.

Gene, primer orientation	Amplicon Size (bp)	T_m Range (˚C)	Primer Size (bp)	GC Content (%)	Self-complementarity	Self-3'-complementarity
HOXA5, forward	80	61.12	25	44.00	8.00	3.00
HOXA5, reverse		62.61	20	65.00	4.00	0.00
Forward primer sequence: CAAGCTGCACATAAGTCATGACAAC Reverse primer Sequence: GGGTCTGGTAGCGCGTGTAG						
TNNC1, forward	93	61.81	20	55.00	4.00	0.00
TNNC1, reverse		61.61	25	40.00	4.00	0.00
Forward primer sequence: TCGGTGCATGAAGGACGACA Reverse primer Sequence: TCGATGTAGCCATCAGCATTTTTGT						
SOSTDC1, forward	81	61.38	20	50.00	4.00	3.00
SOSTDC1, reverse		61.19	20	55.00	4.00	1.00
Forward primer sequence: AGCAGCAACAGCACGTTGAA Reverse primer Sequence: AACCCGAGTGTTCCGATCCA						
CRIP2, forward	92	63.72	22	59.09	4.00	2.00
CRIP2, reverse		63.96	22	59.09	2.00	0.00
Forward primer sequence: GACCACCATGAAAGCCAGGAGC Reverse primer Sequence: CTCACCTTCTCGGCTGTTCCCT						
CD83, forward	91	60.35	22	40.91	4.00	0.00
CD83, reverse		60.70	22	50.00	4.00	0.00
Forward primer sequence: TTTTCACTTGTTTTGCACGGCT Reverse primer sequence: TTTGGGGAGGTAACTGGGAGAA						
GATA2, forward	97	63.27	21	57.14	5.00	2.00
GATA2, reverse		61.37	22	54.55	4.00	0.00
Forward primer sequence: GCCACTACCTGTGCAATGCCT Reverse primer sequence: GTGGTGGTTGTCGTCAGTCTTC						

Table 5: Sequence and parameters of IDT custom-designed qRT-PCR primers for Housekeeping genes.

Gene, primer orientation	Amplicon Size (bp)	T_m Range (˚C)	Primer Size (bp)	GC Content (%)	Self-complementarity	Self-3'-complementarity
TBP, forward	86	61.20	22	50.00	5.00	1.00
TBP, reverse		61.69	22	54.55	3.00	1.00
Forward primer sequence: ATCTTTGCAGTGACCCAGCATC Reverse primer sequence: CCAGCACACTCTTCTCAGCAAC						
ACTB, forward	61	62.91	19	63.16	5.00	1.00
ACTB, reverse		62.41	19	63.16	6.00	1.00
Forward primer sequence: ACAGAGCCTCGCCTTTGCC Reverse primer sequence: ATCCATGGTGAGCTGGCGG						
GAPDH, forward	95	61.24	20	50.00	4.00	0.00
GAPDH, reverse		61.02	21	52.38	3.00	1.00
Forward primer sequence: AGCCGCATCTTCTTTTGCGT Reverse primer sequence: GCCCAATACGACCAAATCCGT						

2.3.6.2 Commercial Primers

Commercial primer master mix for qRT-PCR included 7 µL nuclease-free, RNase-free ddH_2O, 0.4 µL 20x PrimePCR assay, and 10 µL 2x SsoAdvanced SYBR green per one reaction. Master mixes were typically made per 10 reactions.

2.3.6.3 Custom Primers

Custom primer master mix for qRT-PCR included 7 µL nuclease-free, RNase-free ddH_2O, 0.2 µL forward primer, 0.2 µL reverse primer, and 10 µL 2x SsoAdvanced SYBR green per one reaction. Master mixes were typically made per 10 reactions. Forward and reverse primers were ordered from IDT and diluted to 100 µM.

18 µL of master mix and 2 µL of sample product (testing sample, no RT, ddH_2O) were pipetted into the 96-well PCR plate. The plate set-up per each sample being tested included 3 wells designated for the sample (testing cDNA), 3 wells for the testing sample lacking RT (no RT control), and 3 wells for the negative control of ddH_2O instead of testing or noRT sample.

2.3.6.4 DNA Gel

The 4% agarose gel was made by combining 100 mL of 0.5X TBE buffer with 4.0 g agarose in an Erlenmeyer flask, and warming the solution in a microwave for 1.5 minutes. An aliquot of 5.0 µL gel red nucleic acid gel stain was pipetted into the flask and then swirled, before pouring the solution into the DNA gel casting tray with comb placed inside (for well formation). The gel was allowed to cool for 20 minutes, followed by placing the tray with gel into the tank, and submerging it in 0.5X TBE buffer. On a piece of paraffin film, 3 µL of gel loading dye purple and 10 µL of qRT-PCR sample product was mixed by pipetting up and down. The dye-sample solution was pipetted into the agarose gel well. After placement of all samples in the gel wells, 5 µL Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was pipetted into the first well. The apparatus was set to run at a constant 200 volts for 15 minutes. The gels were imaged with a UV transilluminator.

2.3.7 Cell Line Database Comparison

Assessment of gene expression levels obtained from cell line databases was done to compare the qRT-CPCR results to. Thirty-seven and twenty-nine NSCLC cell lines from the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) databases, respectively, were downloaded. Using UNIX, the 10 signature genes were ranked from 1-10

according to their expression levels (the gene with the highest expression level was given a rank of 1, the gene with the lowest expression level was given a rank of 10). The rankings of each gene were tallied to gain an understanding of which genes were generally the most and least expressed in NSCLC cell lines. The cell lines were also studied to determine if any mirrored the YMR signature (over-expression of *GRM1, RECQL4, NRAS, IGFBP5*; under-expression of *HOXA5, TNNC1, SOSTDC1, CRIP2, CD83, GATA2*).

Chapter 3: Results

3.1 AIM 1 To explore the genomic and epigenomic properties of signature genes.

3.1.1 Rationale for AIM 1

Since the 10 genes are differentially expressed between normal and tumor lung tissue samples and the YMR signature derived from the 10 genes showed clinical relevance in prognosis and prediction of adjuvant chemotherapy benefit in stage I NSCLC patients, discovering genomic and epigenomic factors that correlate with the properties and expression of the 10 genes will help elucidate their importance in lung cancer. Gene expression, DNA mutation, CNV, and DNA methylation data from the TCGA LUAD dataset were used. The UNIX command prompt and R were used to construct regression models assessing the correlation between YMR genes' expressions with the genomic and epigenomic covariates.

3.1.2 Assessment of TCGA Data

To determine whether any single or combination of NSCLC disease stages had better correlations between the YMR genes expression levels and the genomic and/or epigenomic factors, SLR and MLR tests were performed using the entire dataset and the dataset divided into Stages I, IA, IB, II, IIA, IIB, III, IIIA, IIIB, IV, I and II, and I to III. Stage groupings with samples sizes under 50 patients were considered non-significant (Table 6). Boxplots were also constructed to provide a visual of the gene expression levels of the TCGA data to see whether it mirrors our expected YMR signature (Figure 1).

Table 6: Patient sample sizes of the TCGA LUAD gene expression data separated into their respective disease stages.

Disease Stage	Patient Sample Size
All data	512
Stage I	279
Stage IA	133
Stage IB	139
Stage II	123
Stage IIA	49
Stage IIB	72
Stage III	84
Stage IIIA	73
Stage IIIB	10
Stage IV	25
Stage I to II	403
Stage I to III	488

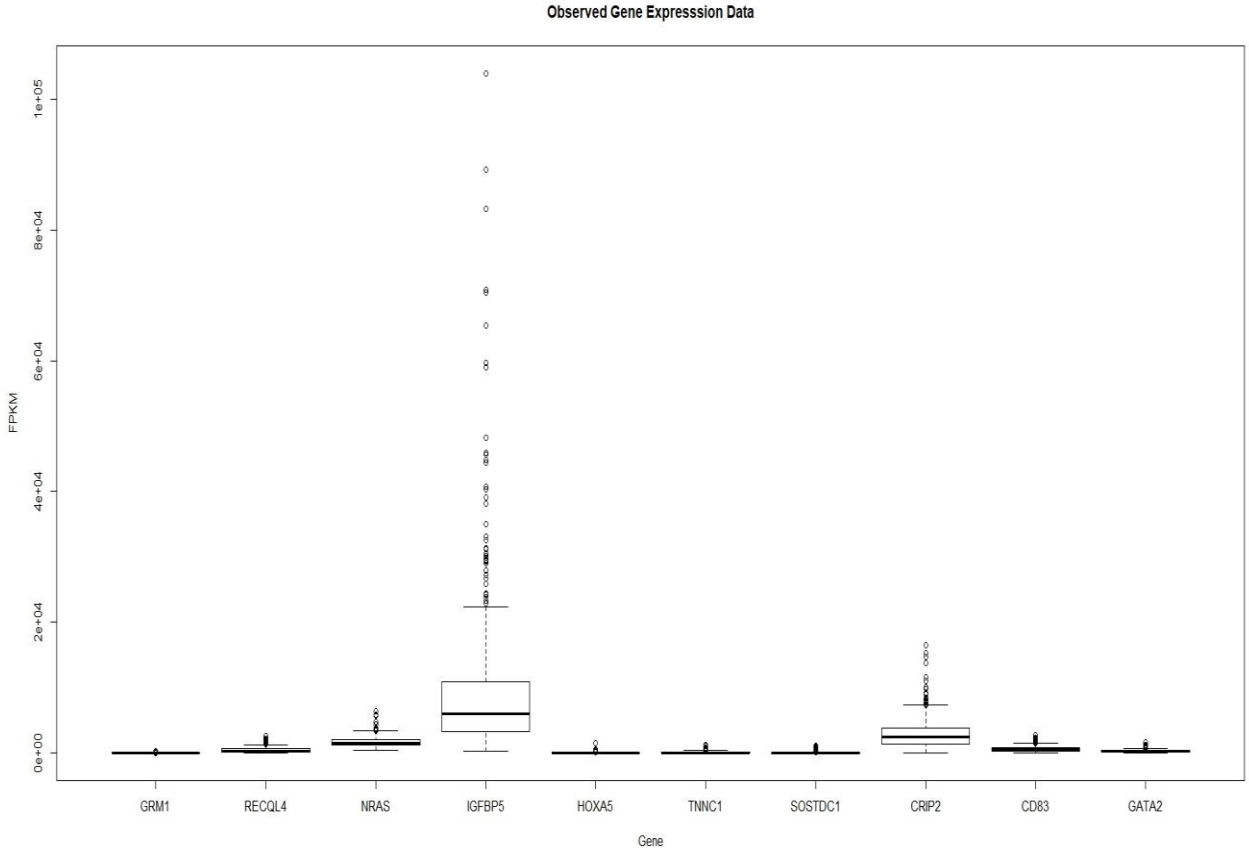

Figure 1: TCGA LUAD YMR gene expression pattern.
Boxplot representation of the 10 YMR gene signature expression data from 512 patients representing lung adenocarcinoma stages I, II, III, and IV. Gene expression data is measured in FPKM and downloaded from TCGA.

Only the boxplot of the entire dataset is shown since the pattern of expression was similar between the entire dataset and the dataset split into stages I, IA, IB, II, IIA, IIB, III, IIIA, IIIB, IV, I to II, and I to III. *IGFBP5* has the highest expression level, followed by *CRIP2* and *NRAS*. The remaining genes had minimal expression levels. As the dataset was further separated into the disease stages (IA, IIA, and IIIB), *RECQL4* and *CD83* demonstrated a more elevated expression level than *GRM1*, *HOXA5*, *TNNC1*, *SOSTDC1*, and *GATA2*. However, the sample sizes for stage IIA and IIIB were small, ranging from 10 to 49 patients. This pattern does not follow the expected gene expression data of *GRM1*, *RECQL4*, *NRAS*, and *IGFBP5* having a higher expression level than *HOXA5*, *TNNC1*, *SOSTDC1*, *CRIP2*, *CD83*, and *GATA2*. Since the YMR signature was constructed using TCGA data in combination with additional datasets and has been shown to be effective in separating stage IA and IB patients into high- and low-risk groups, larger datasets of lung adenocarcinoma patients split into their respective disease stages may be required to observe the expected gene expression data.[69, 70] Filtering the data of clinical factors, such as smoking and treatment statuses, may also provide the expected gene signature expression pattern.

3.1.3 Regression Models

Three simple linear regression (SLR) models were constructed to assess the correlation of the following genomic and/or epigenomic variables with the 10 YMR genes expression levels: (1) DNA mutation, (2) CNV, and (3) DNA methylation. Since there was a sample size of 34 for the gene mutation data, the results obtained for its SLR were non-significant and limited, and therefore is not included (Table 1). Results from the tests using the entire TCGA LUAD dataset are shown since there was no significant difference when the dataset was divided into disease

stages. Significance of linear regression models were first assessed by their p-values, followed by their training and testing dataset mean absolute percentage errors (MAPEs). P-values less than 0.05 are ideal and indicate that the probability the results obtained is due to chance is less than 5% (Tables 7 and 8). The MAPE measures the size of error between the observed and predicted values. It takes the sum of absolute errors (the difference between the observed and predicted gene expression value per patient) to the observed gene expression data and displays the output as a percentage. Therefore, minimal MAPE values are desired.

Table 7: Simple linear regression model and MAPE training and testing dataset values for assessing the correlation between gene CNV and expression of the 10 YMR signature genes.

Gene	p-value	MAPE training	MAPE testing
GRM1	0.0324	Infinite	Infinite
RECQL4	NA	1.445931	1.706748
NRAS	0.06849	0.3892762	0.3765924
IGFBP5	0.127	1.761533	1.626643
HOXA5	0.02804	1.113848	1.170493
TNNC1	0.03253	Infinite	Infinite
SOSTDC1	0.3868	Infinite	Infinite
CRIP2	0.03507	1.009158	1.741447
CD83	0.04498	0.9451412	0.7797078
GATA2	0.6781	1.136364	1.30343

Table 8: Simple linear regression model and MAPE training and testing dataset values for assessing the correlation between DNA methylation and expression of the 10 YMR signature genes.

Gene	p-value	MAPE training	MAPE testing
GRM1	0.02726	Infinite	Infinite
RECQL4	0.007491	1.338019	1.871707
NRAS	0.0299	0.3777232	0.3772146
IGFBP5	0.001299	1.58919	1.907339
HOXA5	0.02213	1.194029	0.9366368
TNNC1	1.35e-06	Infinite	4.942312
SOSTDC1	1.403e-08	Infinite	Infinite
CRIP2	0.0004174	1.320225	1.198741
CD83	0.1007	0.9171166	0.9074103
GATA2	0.03786	1.212465	1.114379

The SLRs for *GRM1* of CNV correlated with gene expression (Figure 2A) and DNA methylation correlated with gene expression (Figure 2B) both output significant p-values of less than 0.05. As expected, an increase in CNV is seen to be correlated with an increase in gene expression while an increase in methylation Beta value correlated with a decrease in gene expression. The MAPEs, however, had infinite values, indicating the presence of many zero values (Tables 7 and 8). The removal of the outlier in the CNV data in attempt to improve the MAPE value increased the p-value of the model.

The TCGA CNV dataset downloaded via UCSC Xena did not have any values for *RECQL4*, resulting in a NA (not available) p-value for the correlation of CNV with gene expression (Figure 3A). The DNA methylation correlated with gene expression (Figure 3B) regression had a significant p-value of 0.007491. As expected, an increase in methylation Beta value was related to a decrease in gene expression. The MAPEs had minimal values below 2.0, indicating there is little difference between the measured and predicted values (Tables 7 and 8).

The SLRs for *NRAS* and *IGFBP5* of CNV correlated with gene expression (Figures 4A and 5A, respectively) output non-significant p-values of 0.06849 and 0.127, respectively. Although the scatterplot for *NRAS* demonstrated the expected trend of an increase in CNV being related to an increase in gene expression, the points were seen to be centered and not spread equally about the regression line. The DNA methylation correlated with gene expression SLRs (Figures 4B and 5B, respectively) had p-values of 0.0299 and 0.001299. As expected, plots depicted an increase in methylation Beta value correlating with a decrease in gene expression. The MAPEs for *NRAS* between the two SLR tests were similar in value and below 0.5, and they were below 2.0 for

IGFBP5 (Tables 7 and 8). Therefore, the SLR models using CNV data is non-significant, as proven by the p-values, while the SLR model using DNA methylation is shown to be significant.

The SLRs for *HOXA5* and *CRIP2* of CNV correlated with gene expression and DNA methylation correlated with gene expression output significant p-values of less than 0.04. As expected, an increase in CNV is seen to be correlated with an increase in gene expression (Figures 6A and 9A, respectively) while an increase in methylation Beta value correlated with a decrease in gene expression (Figures 6B and 9B, respectively). The MAPEs had minimal values less than 2.00 (Tables 7 and 8). These results suggest there is little error in the SLR models, making CNV and DNA methylation a strong factor associated with gene expression.

Although *SOSTDC1* had a non-significant p-value of 0.3868 for its SLR model assessing the relationship between CNV and gene expression, the remainder SLR models for *TNNC1* and *SOSTDC1* had significant p-values below 0.05. The scatterplots of the regression models demonstrated the expected patterns: as CNV increases, gene expression decreases (Figures 7A and 8A), and as methylation Beta value increases, gene expression decreases (Figures 7B and 8B). The MAPEs for the training and testing datasets of both sets of SLRs had infinite or larger values than was seen for the other genes, indicating that the model may need to be modified or a different dataset should be used to gain a better fit of the data to the model (Tables 7 and 8).

While the SLR of CNV and gene expression for *CD83* had a significant p-value of 0.04498, the regression of DNA methylation and gene expression was non-significant with a p-value of 0.1007. The MAPEs were minimal with values below 1.00, indicating little variance between the

data values (Tables 7 and 8). Conversely, the SLR of CNV and gene expression for *GATA2* was non-significant with a p-value of 0.6781 while the SLR of DNA methylation and gene expression had a p-value of 0.03786. The MAPEs were minimal (Tables 7 and 8). The scatterplots of the SLRs assessing CNV demonstrated the expected pattern of increasing CNV being related to an increase in gene expression (Figures 10A and 11A). The SLR assessing DNA methylation scatterplots suggest that an increase in methylation Beta value is correlated to an increase in gene expression (Figures 10B and 11B). This is contrary to the expected pattern and suggests the dataset may need to be further filtered or a different dataset may need to be used.

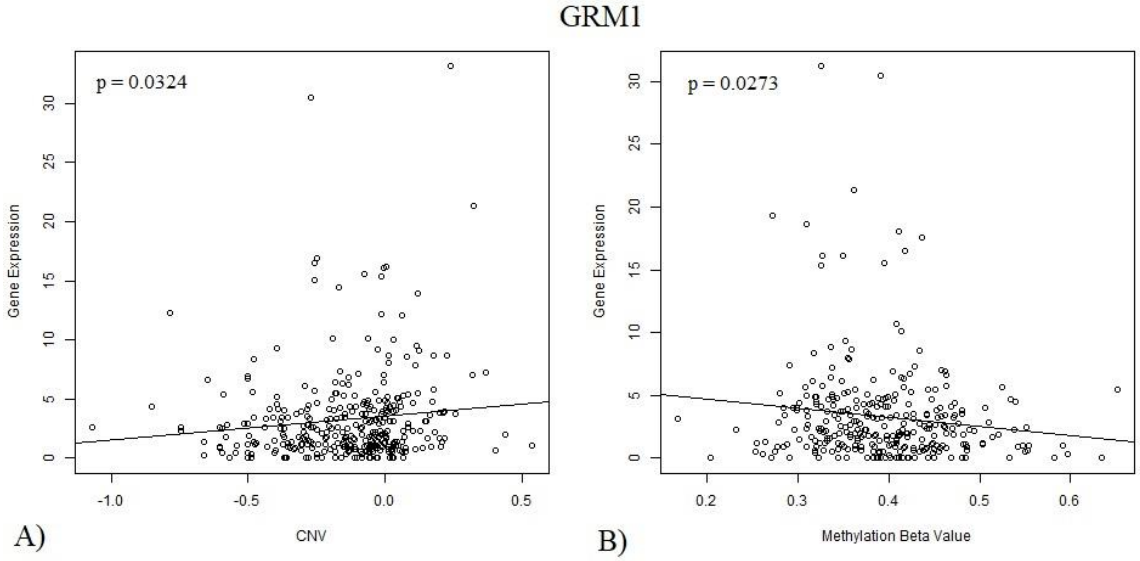

Figure 2: Scatterplots of SLR models based on CNV and DNA methylation for GRM1. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yin *GRM1* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

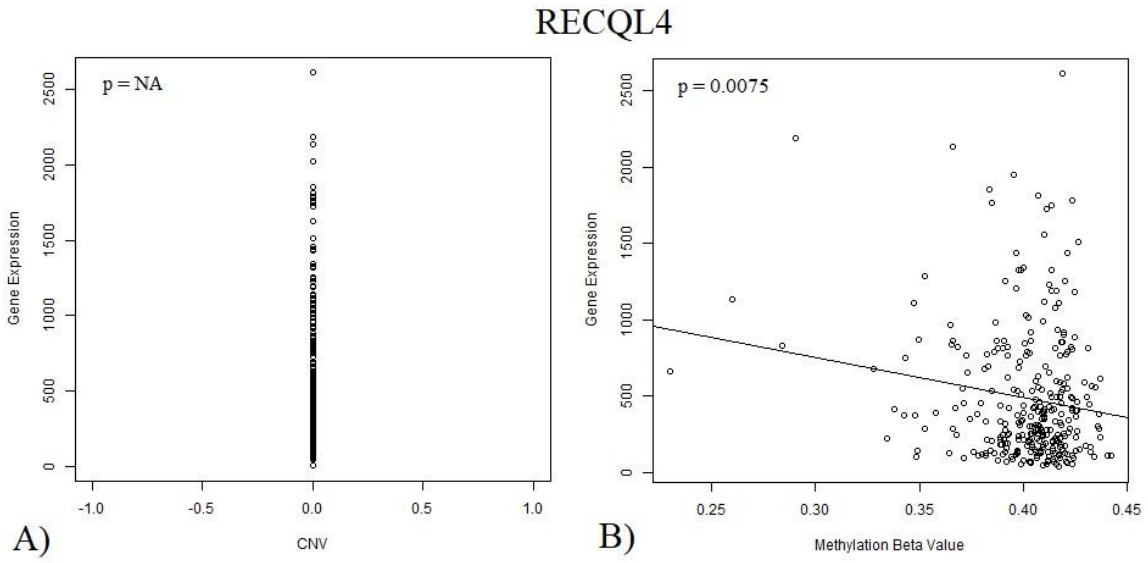

Figure 3: Scatterplots of SLR models based on CNV and DNA methylation for RECQL4. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yin *RECQL4* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

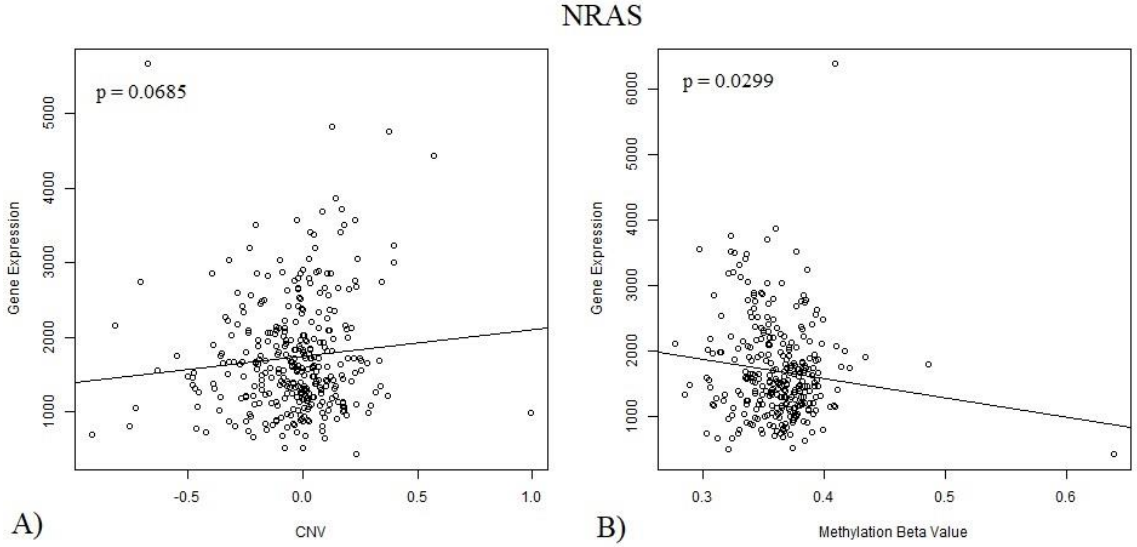

Figure 4: Scatterplots of SLR models based on CNV and DNA methylation for NRAS. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yin *NRAS* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

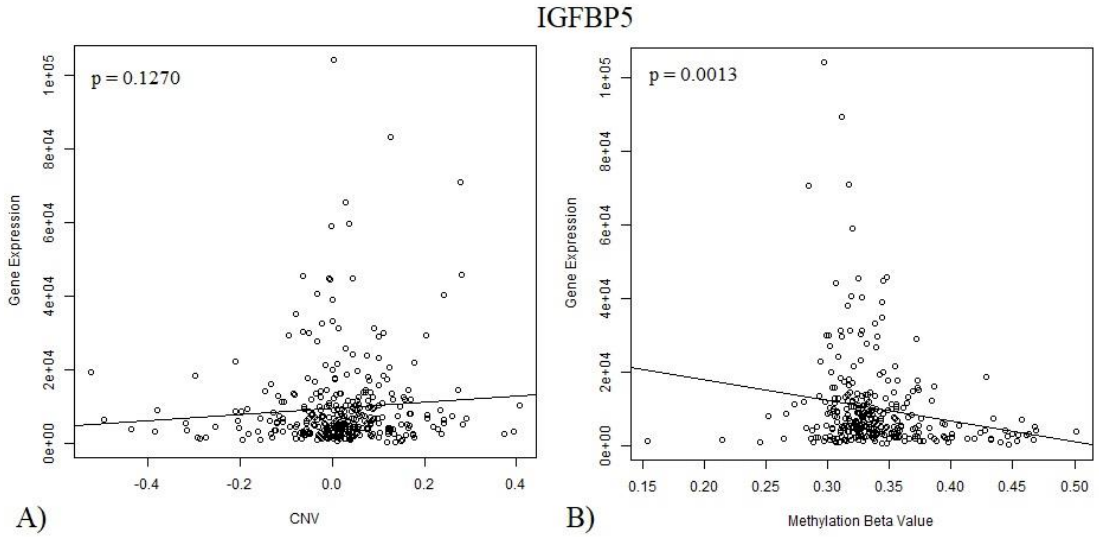

Figure 5: Scatterplots of SLR models based on CNV and DNA methylation for IGFBP5. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yin *IGFBP5* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

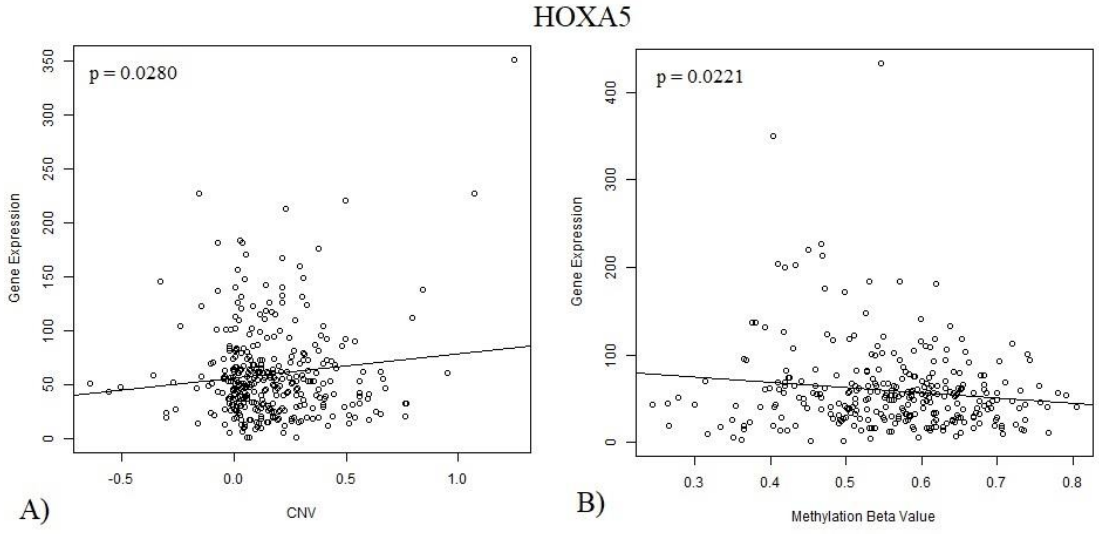

Figure 6: Scatterplots of SLR models based on CNV and DNA methylation for HOXA5. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *HOXA5* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

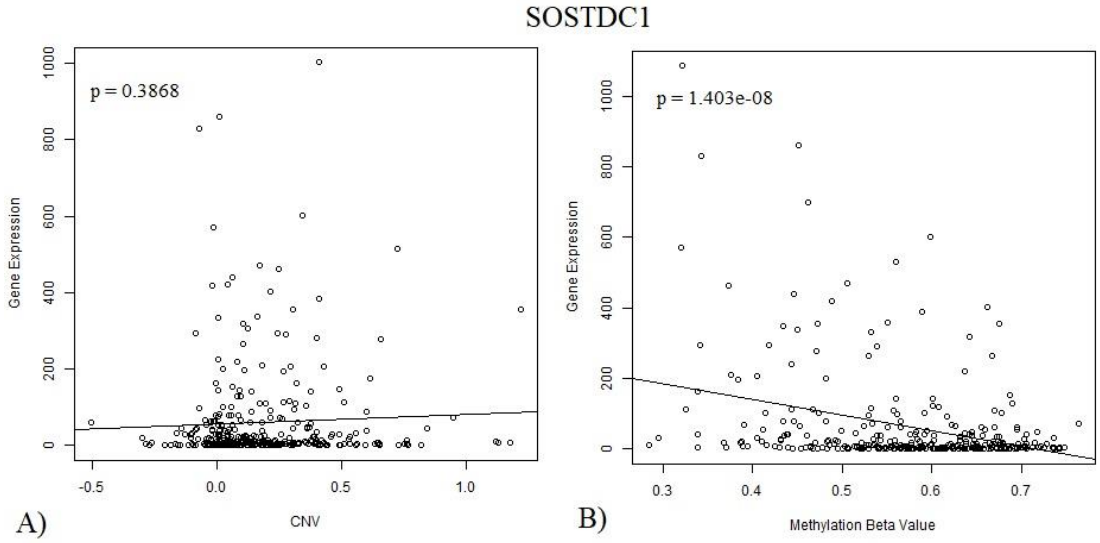

Figure 7: Scatterplots of SLR models based on CNV and DNA methylation for SOSTDC1. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *SOSTDC1* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

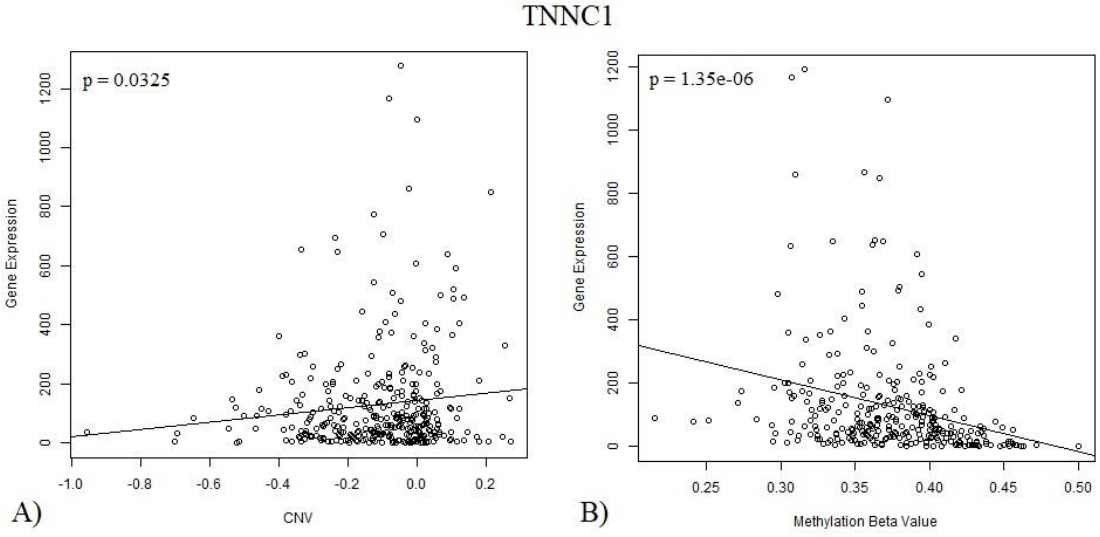

Figure 8: Scatterplots of SLR models based on CNV and DNA methylation for TNNC1. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *TNNC1* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

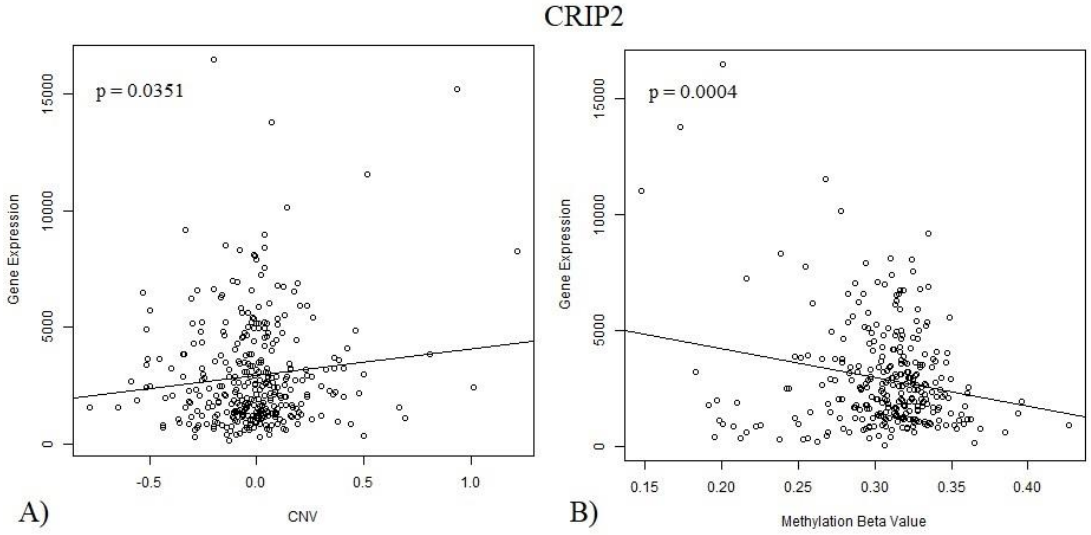

Figure 9: Scatterplots of SLR models based on CNV and DNA methylation for CRIP2. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *CRIP2* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

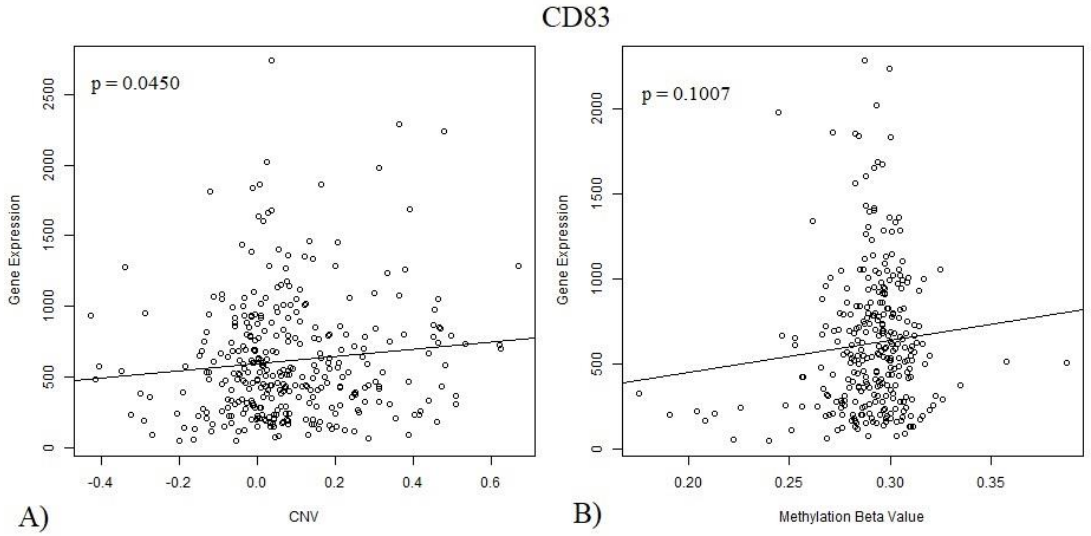

Figure 10: Scatterplots of SLR models based on CNV and DNA methylation for CD83. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *CD83* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

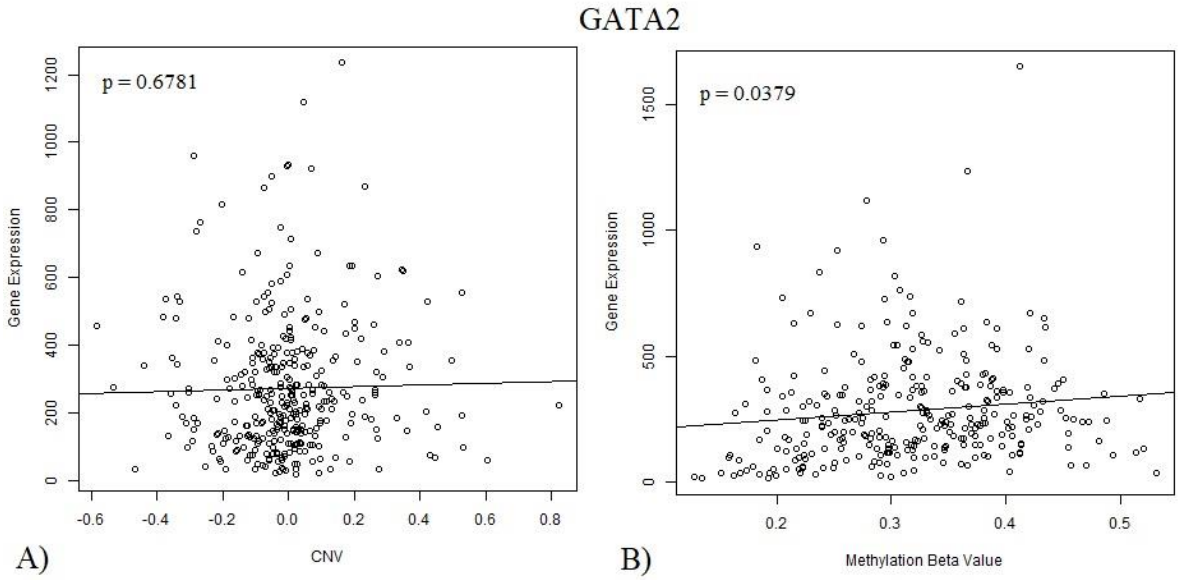

Figure 11: Scatterplots of SLR models based on CNV and DNA methylation for GATA2. Scatterplots with regression line and p-value output for the SLRs of CNV correlated with gene expression (A) and DNA methylation correlated with gene expression (B) of the Yang *GATA2* gene. Gene expression data is measured in FPKM, CNV values are estimated using the GISTIC2 method, and DNA methylation data is presented as Beta values. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

Two multiple linear regression (MLR) models were constructed to assess the correlation of the following genomic and/or epigenomic variables with the 10 YMR genes expression levels: (1) DNA mutation and DNA methylation and (2) gene CNV and DNA methylation. Since there was a sample size of 34 for the gene mutation data, the results obtained for its MLR were non-significant and limited, and therefore is not included (Table 1). Results from the tests using the entire TCGA LUAD dataset are shown since there was no significant difference when the dataset was divided into disease stages. Model significance was assessed by the output p-value and MAPE (as per equation 3) (Table 9).

To visually assess the fit and significance of the regression models, a variety of plots were constructed that displayed the predicted and observed expression values and residual errors in different forms. However, none of the YMR signature genes displayed the ideal output from the Predicted versus Observed Gene Expression and four regression model diagnostic plots. The predicted gene expression value was plotted against the observed gene expression value; since the expected and observed gene expression values are expected to be similar, a plot of the points falling along a 45˚ angle line would be ideal. Multiple linear regression summaries output the residuals vs. fitted, normal Q-Q, scale-location, and residuals vs. leverage plots. The MLR model results were assessed on all plots, however, the residuals vs. fitted plots, which detects a regression's non-linearity, unequal error variances, and outliers, provide the clearest interpretations. Ideal residuals vs. fitted plots demonstrate a random scattering of points around the y = 0 axis, indicating a linear relationship, and a horizontal band of residual points around the y = 0 axis, indicating the error variances are equal and the dataset has a normal distribution. Points that stray from the common grouping of residual points are identified as outliers and may be influential or non-important to the dataset and therefore their removal should be evaluated. The plot for *GRM1* had a band of points around the y = 0 axis, indicating the error variances may be equal. However, the plot indicated that the model needs to be improved since the points were not randomly scattered, suggesting a possible non-linear relationship, and the removal of the outlier increased the p-value of the regression (Figure 12). *RECQL4* and *NRAS* were determined to be non-significant as their plots did not meet the three criteria of significance through assessment of their residuals (Figure 13 and 14). The *HOXA5* data initially had an outlier whose removal increased the p-value of the regression model. The significance of its residuals vs. fitted plot requires slight improvement (Figure 15). *SOSTDC1, TNNC1, CRIP2,* and *CD83* have

unideal plots, indicating their data do not meet the linearity and equal variance criteria. Therefore, these models require improvement (Figures 16, 17, 18, and 19). Similar to *HOXA5*, the *GATA2* plot demonstrates a slight random scattering of points forming the indicative horizontal band around the y = 0 axis, signifying the model requires minor improvement (Figure 20).

Although a number of the regression models had significant p-values, further analyses must be done to determine whether the models and their data are significant. After assessment of the MAPEs and output scatter (for SLR) and residuals vs. fitted (MLR) plots, it is evident that the SLR and MLR models may be modified and improved upon. Corrective measures for improving linear regression models include deleting observations, transforming variables, adding or deleting variables, and using another regression approach or type. Transforming variables have the potential to improve regression models when they do not meet the normality, linearity, homoscedasticity, or heteroscedasticity (non-constant error variance) assumptions. Transformations typically include the replacement of Y with Y^{α}, where Y^{α} may represent $1/Y^2$, $1/Y$, $1/Y^{1/2}$, $\log(Y)$, Y^2, et cetera. Increasing or decreasing the number of variables has the potential to improve the regression model and its fit by including more predictor factors. Therefore, in addition to CNV and DNA methylation, transcription factors and miRNA have been shown to have an impact on gene expression and should be included in the regression models in hopes of improving their significances.[115, 116] In both the simple and multiple linear regression models, the disease staging could have been added to the model as another variable instead of dividing the dataset into the stages since the former uses the entire dataset while the latter decreases the sample size per each stage dataset. Similarly, clinical factors such

as patient sex and treatment and smoking statuses can be included in the regression models as additional variables.

Table 9: Multiple linear regression model MAPE and p-values.
P-value and MAPE training and testing dataset values for assessing how strongly the combination of gene CNV and DNA methylation correlates with expression statistical values for the 10 YMR signature genes.

Gene	p-value	MAPE training	MAPE testing
GRM1	0.1105	Infinite	Infinite
RECQL4	0.001279	1.380921	1.902762
NRAS	0.003221	0.3941123	0.3166736
IGFBP5	0.0004657	1.490869	2.197751
HOXA5	0.02024	1.18767	1.553433
TNNC1	6.875e-06	Infinite	Infinite
SOSTDC1	6.773e-07	Infinite	Infinite
CRIP2	0.01188	1.378425	0.9925978
CD83	0.4864	0.9019874	0.8165663
GATA2	0.5251	1.089769	1.120874

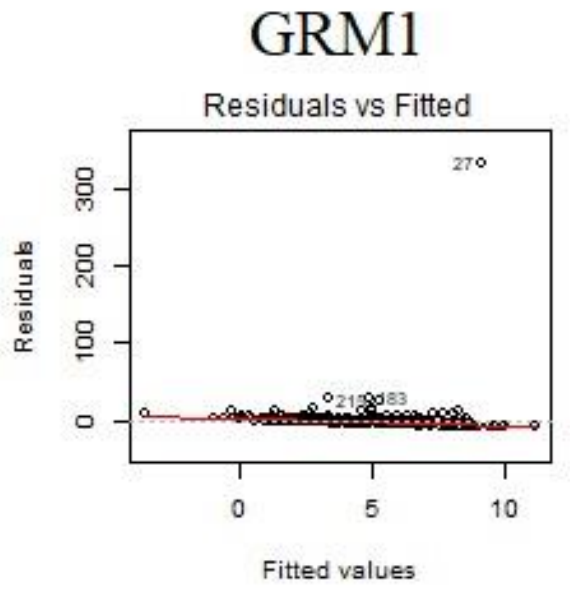

Figure 12: Residual vs. Fitted diagnostic plot for the GRM1 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *GRM1* expression, after the removal of an outlier value. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

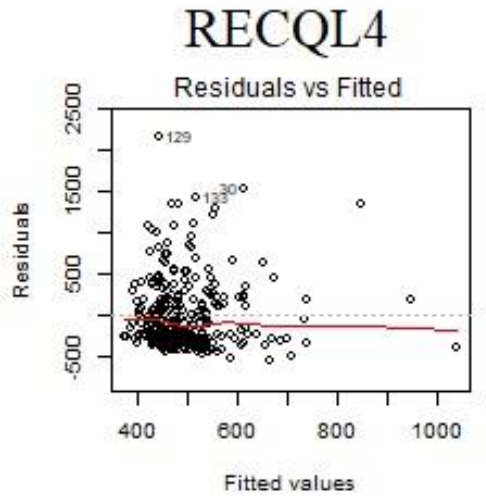

Figure 13: Residual vs. Fitted diagnostic plot for the RECQL4 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *RECQL4* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

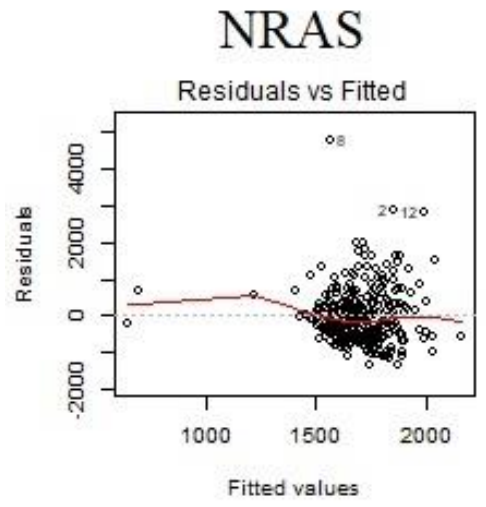

Figure 14: Residual vs. Fitted diagnostic plot for the NRAS multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *NRAS* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

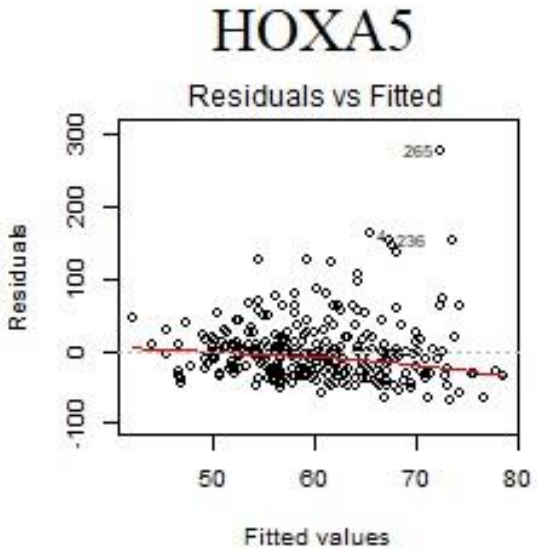

Figure 15: Residual vs. Fitted diagnostic plot for the HOXA5 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *HOXA5* expression, after the removal of an outlier value. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

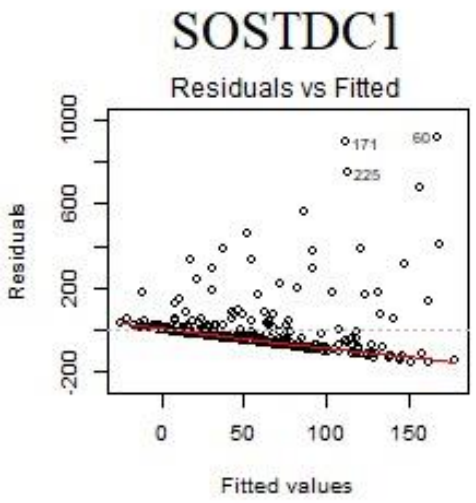

Figure 16: Residual vs. Fitted diagnostic plot for the SOSTDC1 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *SOSTDC1* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

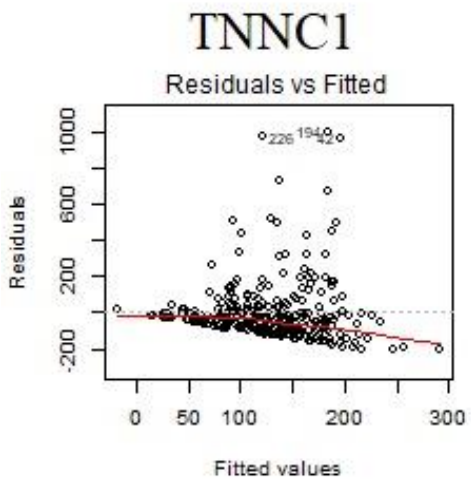

Figure 17: Residual vs. Fitted diagnostic plot for the TNNC1 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *TNNC1* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

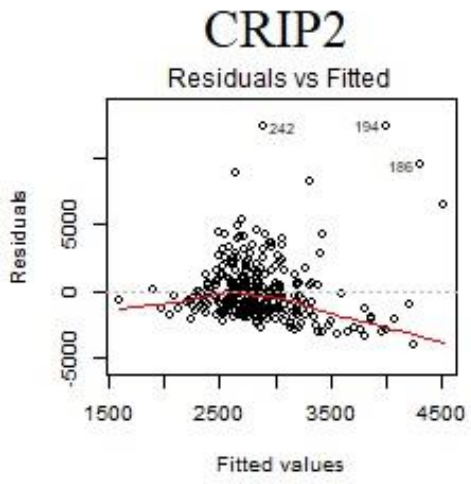

Figure 18: Residual vs. Fitted diagnostic plot for the CRIP2 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *CRIP2* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

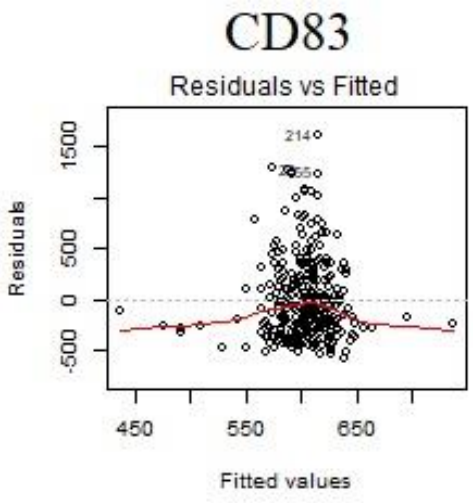

Figure 19: Residual vs. Fitted diagnostic plot for the CD83 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *CD83* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

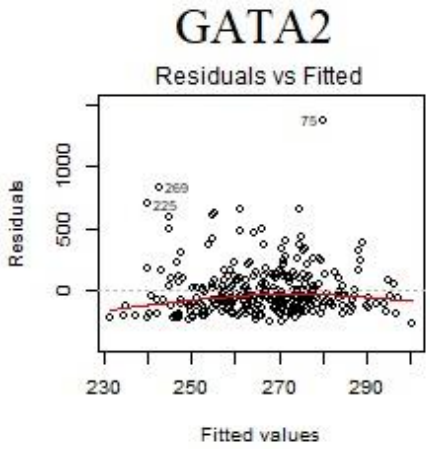

Figure 20: Residual vs. Fitted diagnostic plot for the GATA2 multiple linear regression model.
Residual vs. fitted diagnostic plot assessing how strongly the combination of gene CNV and DNA methylation correlates with *GATA2* expression. The Residuals versus Fitted plot assesses whether there is a linear relationship between the dependent and independent variables. TCGA LUAD data was downloaded and used; the UNIX commend prompt was used to filter the data; R was used to construct and run the linear model.

3.2 AIM 2 Determination of gene expression detection assay feasible for clinical use.

3.2.1 Rationale for AIM 2

To further develop our 10-gene signature, we aimed to choose a gene expression detection assay most feasible for clinical oncology use. A comprehensive literature review assessment on the qRT-PCR, DNA microarray, nCounter, RNA-Seq, FISH, and tissue microarray (TMA) assays would help to choose the most suitable method for each clinical application. The systems are evaluated based on parameters including, but not limited to, sensitivity, specificity, cost of reagents and machine, and time required from sample preparation to data analysis.

3.2.2 Comparison of Gene Detection Assay Systems

Sensitivity refers to the minimum amount of substance, such as RNA, detected per sample in an experimental assay. qRT-PCR and real-time PCR have a reported sensitivity of 10 copies mRNA per cell, with as many as 200 copies per cell having been detected for PCR.[117, 118] As little as 3 copies per cell have been detected in PCR, implying that this number could be similar for qRT-PCR.[119] Similarly, DNA microarrays can detect from 1-10 copies of mRNA per cell.[120] Both nCounter and MiSeq have the ability to sense a minimum of 1 copy of mRNA per cell; nCounter utilizes region-specific colour-coded probes to bind to a single transcript, whereas MiSeq can generate a cluster of 1000 copies from a single transcript.[105, 121] Lastly, sensitivity for TMAs can be analyzed in a number of ways as it may depend on the size of the tissue spot in the array or the sensitivity of the tests (e.g. FISH, ISH) being used. For example, TMAs can array up to 1000 different sample spots on a single array, allowing 0.6 mm as the smallest sample size that can be taken with efficient detection results; *in situ* hybridization is shown to have a sensitivity of 10-20 mRNA copies per cell. (Table 10).[17, 109, 122, 123]

The specificity for each of the qRT-PCR, DNA microarray, NanoString nCounter, MiSeq RNA-Seq, and TMA depends on the primers/probes and/or reagents being used. As previously discussed, a strict design is required to ensure the qRT-PCR primers tightly anneal to the mRNA transcript and amplify only the section of interest. Therefore, the specificity of qRT-PCR depends on the forward and reverse primer design, as well as the oligonucleotide probe used in detection.[124–126] The specificity of DNA microarrays depends on a number of factors, such as the density of probes annealed to the slide, the single probe (hybridization of a probe to a single target), single spot (a spot of multiple probes may contain probes that are perfectly or partially hybridized to a target), and spot-set (multiple spots represent different sections of the same sequence).[127] Similar to PCR, the specificity may also depend on the probe design; for example, cDNA microarrays can employ probes up to a few thousand base pairs in length while oligonucleotide arrays perform best with shorter probes that can be 25-30mer or 60-70mer in length.[120] The specificity of NanoString nCounter is due to its target-specific Capture and Reporter probes, as these are designed to anneal to a nucleic acid region of interest. These combine with the internal controls used to form the CodeSet, which confers the overall specificity.[128] TMA quality is dependent on the type of analysis being done. For example, FISH, RNA FISH, and immunohistochemistry each confer different specificities to the TMA assay.[129, 130] RNA probes about 250-1500 nucleotides in length, with a common recommendation of 800 nucleotides, tend to demonstrate high specificity.[131] Chemically-synthesized oligonucleotides labelled with fluorophores and calibrated to a certain region tend to generate RNA FISH probes with high specificity. (Table 10).[17, 129]

Table 10: Comparison of qRT-PCR, DNA Microarray, NanoString nCounter, Illumina MiSeq RNA-Seq, and Tissue Microarray assay properties.[17]

Assay	**qRT-PCR**	**DNA Microarray**	**NanoString nCounter**	**Illumina MiSeq RNA-Seq**	**Tissue microarray & FISH**
Primer/probe design	Gene-specific primer with attached quencher and reporter fluorophores; SYBR green	DNA probes complementary to cDNA samples	Capture probe with 3' affinity tag and Reporter probe with colour-coded tag	Primers on flow cell and adaptors to ligate to ends of sample	Gene-specific RNA probes; gene-specific fluorochrome-labelled probes; monoclonal antibodies
Sample preparation	RNA extraction; reverse transcribe sample	RNA extraction; reverse transcribe sample	RNA extraction	RNA extraction; reverse transcribe sample; fragmentation	Map donor block; place into recipient block; make TMA
Instrument	Thermal cycler	Microarray scanner	Prep Station and Digital Analyzer	MiSeq benchtop sequencer	Tissue arrayer; microscope or array scanner
Reproducible	Yes	Yes	Yes	Yes	Yes
Specificity	Forward and reverse primer design, oligonucleotide probe	Density of probes annealed to the slide, probe design	Design of Capture and Reporter probes	Rely on data analysis	Rely on probes to be used
Sensitivity	10-200 copies/cell	1-10 copies/cell	<1 copy/cell	<1 copy/cell	1-10 copies/cell
Clinic study	Yes	Yes	Yes	Yes	Yes
Commercialized	Oncotype DX	MammaPrint	Prosigna	No	No
Number of genes or transcripts detected	1-100	50 000	800	700 – several thousand	3
Number of samples	1-96	1-12/array	12	96	1000
Processing steps	Prep reaction mixture, PCR cycles, Result analysis	Label cDNAs, hybridization to array, Data analysis	Label probes, hybridization to array, Data analysis	cDNA lib prep, sequencing, data analysis	Make TMA, Slide TMA Staining, Analysis

Raw Data analysis	by machine in 30 minutes	by machine in 1 hour 40 minutes	by machine in 2.7 hours	by machine in 3 hours	by machine or microscopy in 6 minutes
Normalization	3-5 housekeeping genes	Housekeeping genes; RMA; LOWESS method	Housekeeping genes; positive controls	RPKM	Tissue array co-occurrence matric analysis
Data analysis	Absolute and relative quantification; standard dissociation curve; statistical tests	Visualization; statistical tests	Colour-coded images are taken and output as code counts	Data output as sequenced reads with quality scores or read alignments	PCR; H and E staining; FISH, ISH; fluorescent microscopy

A qRT-PCR reaction per sample can be considered cheap in cost. Various kits can be obtained according to one's experimental design to reverse transcribe the mRNA to cDNA and carry out the remainder qRT-PCR steps for about $300-$550. According to BioSearch Technologies, after all reagent costs have been accounted for, a single qRT-PCR reaction totals under $1.00. When probes are being used as the detection method, the total cost is about $0.82 per reaction; when SYBR is being used, the total cost is about $0.56 per reaction. These increase to about $0.89 and $1.13 if probes and SYBR are being used, respectively, when additional targets are present.[132] A sample processing of $0.50 has been reported. A qRT-PCR reaction per sample has been reported to cost as little as $0.50 per reaction.[133] A whole microarray slide can cost between $150-$400, with a full genome array being $500, and a processing cost of about $325 per array. Some studies have reported a cost of $0.025 per data point (or gene being analyzed).[133, 134] The NanoString nCounter provides a Master Kit and CodeSet to be purchased separately, with all necessary consumables and reagents. The Master Kit costs about $280-350 per 1 cartridge for 12 samples.[105] Custom or pre-designed CodeSets are available for specific gene detection. Pre-designed CodeSets can be purchased for $3600-$4800 per 12 assays for analyzing samples

of human stem cells, human cancer reference, or leukemia, to name a few. A custom CodeSet for 48 assays assessing 25 genes can cost about $6000, resulting in a cost of $125 per assay and $5 per data point. Doubling the target to 50 genes nearly doubles the cost to $10 000 for 48 assays, $208.33 per assay, and $4.17 per data point.[135] Several institutes provide services to carry out an nCounter experiment for up to $530 (University of Kentucky and the Whitehead Institute, respectively) per cartridge of 12. A sample processing for targeted gene sequencing costs $90 on the MiSeq Benchtop sequencer, with a 50-500 base pair sequencing kit valued at $1400-$2000.[136, 137] Some institutes provide data analysis for $160-$175.[137, 138] Prepared tissue microarrays may cost as much as $6000 (stage I breast cancer tissue array).[139] Studies have reported the cost to build and analyze a TMA to be $255 (18 cases) and $12 240 (48 cases), resulting to $12.50 per case.[140] In addition, the instruments are valued at $25 000-$95 000 (thermal cycler for PCR), $50 000-$110 000 (microarray scanner), $149 000-$285 000 (NanoString Prep Station and Digital Analyzer), $128 000 (MiSeq Benchtop sequencer), and $55 000-$98 000 (tissue arrayers), respectively. (Table 11).[17, 141–147]

qRT-PCR, DNA microarrays, NanoString nCounter, and Illumina MiSeq all require the isolation and purification of RNA. Following RNA extraction, qRT-PCR requires the mixing of various reagents before allowing the sample to reverse transcribe. A well-plate must be carefully prepared prior to subjecting the sample for the PCR and dissociation curve analysis steps in the thermal cycler machine and by the computer, respectively; this can take about 5 hours when including an RNA extraction step.[148] Similarly, DNA microarrays require a reverse transcription step after the RNA extraction. Time is then required to allow the sample to hybridize to the array; the arrays can be pre-ordered and do not have to be constructed in the

laboratory. The microarray scanner analyzes and slides and computer software can be purchased for data analysis, resulting in a total time of 20 hours.[149] NanoString nCounter only requires the user to prepare the initial sample and then transfer the sample to and from the Prep Station and Digital Analyzer machines. The machines carry out any processing and data analysis that needs to be done, resulting in a total of about 5.5 hours.[150] Similar to the nCounter, Illumina MiSeq only requires the user to prepare the library and load it onto the reagent cartridge; the rest of the sample processing and data analysis is carried out by the MiSeq benchtop sequencer, with the MiSeq Reporter analysis software launching following completion of the trial. A total of about 40 hours is required.[136] TMA construction is a tedious procedure and therefore requires the most work, demanding an average of up to two days, depending on the exact procedure followed. If donor tissue blocks are not available, one must collect samples prior to preparing the block. Then, the tissue block is sectioned and slides of the tissue sample must be stained so that one can use the tissue arrayer to accurately remove cores from the donor block and place them on the slide. Following sectioning of the TMA, the resulting slide is interrogated using reagents for FISH, ISH, or immunohistochemistry tests must be prepared and applied to the slide. The slides can then be analyzed by eye or with a (digital) slide scanner (Table 11).[17, 151]

Table 11: Cost and time of qRT-PCR, DNA Microarray, NanoString nCounter, Illumina MiSeq RNA-Seq, and Tissue Microarray assays. The prices may vary between facilities.[17]

Assay	qRT-PCR	DNA Microarray	NanoString nCounter	Illumina MiSeq RNA-Seq	Tissue microarray & FISH
Cost on Sample preparation	$0.56 (SYBR) $0.82 (probe)/Single plex reaction	~$50/array	$20	~$200-300	$75-2000/array, depending on cancer type
Cost on Kit	$1416 (SYBR) $1834 (probe)/100 preps	~$350/plate	$280-350/1 cartridge for 12 samples (Master Kit); $3600-4800/12 assays (Custom CodeSet)	$1200-2320/~12 reactions	$515 for FISH
Cost on Processing	$0.50/sample	$0.025/data point; ~$100/array	$4.17/data point; $35-41.67/sample	$90/sample	$12.50/sample
Cost on Data analysis	$55	~$100	$65-250	$160-175	$95/slide using 3 antibodies
Cost on Instrument	$25 000-$95 000 Or ~$25/run	$50 000-$110 000	$149 000-$285 000	$128 000	$55 000-$98 000
Time on Sample preparation	1 hour	20 minutes	5 minutes	8 hours	30 minutes
Time on Sample processing	2-4 hours	60 minutes to 17-18 hours or overnight	5 minutes + 2.5 hours	24 hours	24 hours
Time on Data Analysis	30 minutes	1 hour 40 minutes	5 minutes + 2.7 hours	3 hours	6 minutes

3.3 AIM 3 qRT-PCR assay optimization and YMR signature validation in formalin-fixed, paraffin-embedded (FFPE) samples.

3.3.1 Rationale for AIM 3

qRT-PCR involves parameters including primer design, reagent kit used, RNA extraction, RNA quality, reverse transcription, and PCR conditions. Assay optimization ensures assay consistency

and that the assay will work for FFPE samples. To further demonstrate the reproducibility of our YMR model, we will test the assay in a cohort of FFPE samples from the Manitoba Tumor Bank to assess the expression levels of the YMR signature genes in NSCLC tissue samples. The expression level of *EGFR* will be used as a positive control since it is highly expressed in lung cancers to compare the expression levels of the YMR genes to and *TBP* will be used to normalize the all gene expression levels using the ΔC_t method.

3.3.2 Measurement of RNA Quality

As observed from the A549 cell line and three FFPE lung tissue RNA extractions, the L0304F sample provided RNA with the best quality, having an A260/A280 value of 1.93 and A260/A230 of 2.03. This was comparable to RNA extracted from the A549 cell, which had a larger A260/A280 value of 1.99, yet smaller A260/A230 value of 1.98. The 13137A sample had the next best quality, with A260/A280 and A260/A230 values of 1.84 and 1.56, respectively. The L0379B sample was comparable to the 13137A sample, with an A260/A280 of 1.84 and A260/A230 of 1.44. However, studies have shown that the cell line is expected to provide RNA with the greatest quality, followed by FFPE samples that have been recently prepared (fixed and embedded). Therefore, when assessing the dates of the samples, L0379B should provide the greatest quality RNA, next to the A549 cell line, since it was prepared in 2017. L0304F RNA should have the second-greatest quality, followed by 13137A, since their preparation dates are 2014 and 1996, respectively. A number of factors may have affected the tissue sample and its nucleic acid quality, such as the exact process of fixation and embedding, depth of penetration of formalin fixative into the tissue, volume and concentration of fixative used, amount of tissue used, and temperature of incubation atmosphere and tissue sample during fixation and

embedding. The time between sample removal from patient and fixation is very critical and must be kept to a minimum as tissue autolysis and degradation of nucleic acid by enzymes can occur. The tissue fixation process should be kept to below 24 hours as over-fixation can produce more irreversible cross-links between nucleic acid and proteins and result in a faster fragmentation of nucleic acids.[152–154] Knowledge of specific information regarding these factors may provide insight into why the L0379B sample provided RNA with the poorest quality, instead of the greatest. The L0379B and L0304F samples, however, did provide consistent qRT-PCR results when assessing the expression levels of the YMR signature genes. The 13137A sample provided qRT-PCR results that were inconsistent with the other two samples, likely due to its old age and more degraded nucleic acid (Figure 21).

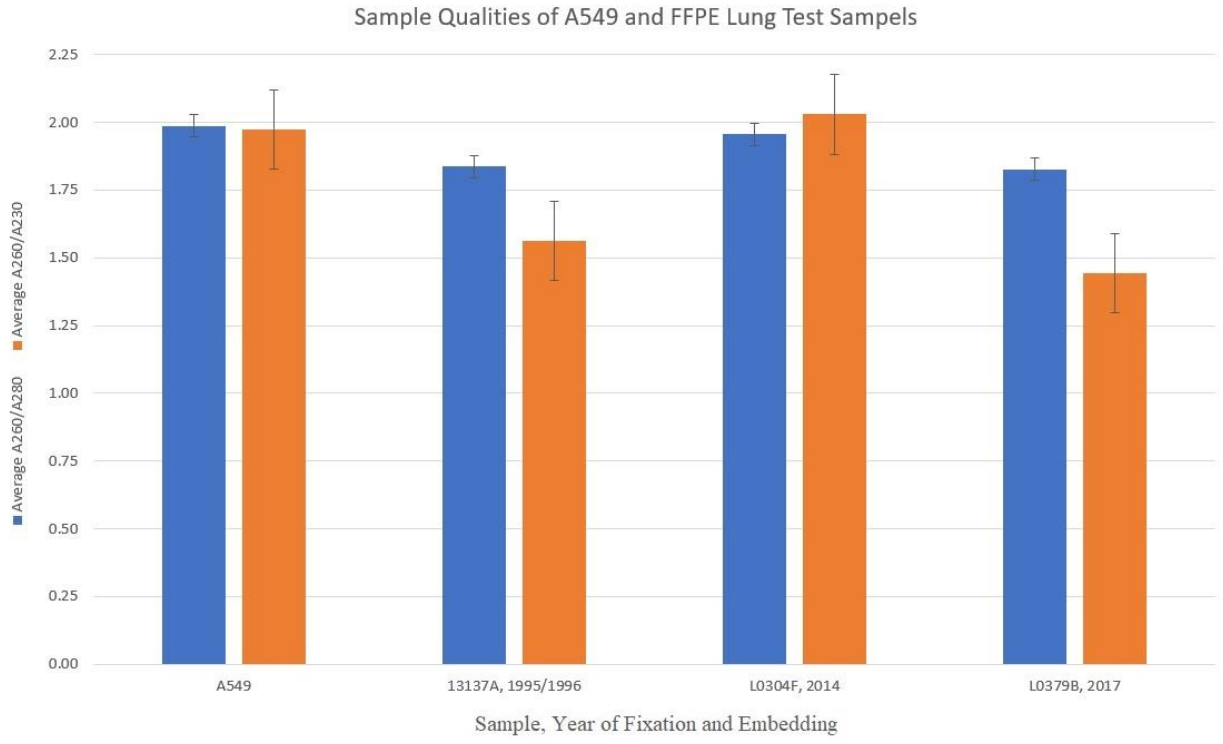

Figure 21: Comparison of RNA quality between the A549 cell line and FFPE samples. RNA quality was measured by the A260/A280 and A260/A230 values output from the NanoDrop BioAnalyzer and compared between the A549 cell line and three lung test FFPE samples. The values from three extractions were averaged. Standard error is depicted.

3.3.3 DNA Gel of Yin and Yang Gene qRT-PCR Products

Although PCR amplicons using material from FFPE samples have been shown to exceed 1000 base pairs (bp), primers designed for amplicons of smaller sizes of about 60-100 bp tend to produce smaller C_q values, the more favourable result.[154–156] Primers for the qRT-PCR tests to measure the expression levels of the 10 YMR signature and three housekeeping genes were first designed for amplicon sizes of 80 – 100 bp. Only *CRIP2* from the first set of designed primers produced a single qRT-PCR product and/or did not produce product(s) in the no Reverse Transcriptase (no RT) and double-distilled water (ddH_2O) negative controls. Product presence in the no RT and ddH_2O imply the presence of contaminating DNA and/or primer dimers. Therefore, the second set of primers were designed with an expanded amplicon range of 80 – 120 bp to allow for more primer pair options. The primer self-complementarity at the 3' and 5' ends were kept below 3.00; however, when primer pairs were not returned from the NCBI Primer-BLAST primer designing tool, the self-complementarity limit was increased to 5.00. IDT OligoAnalyzer Tool confirmed whether the primer pairs would form hairpin structures, self-dimers, and/or heterodimers. From the second design, primers for *IGFBP5*, *TNNC1*, *GATA2*, and *TBP* produced a single qRT-PCR product and/or did not produce product(s) in the no RT and ddH_2O negative controls.

The DNA agarose gels indicated the TNNC1, CRIP2, CD83, GAPDH, ACTB, and TBP primers were of the expected size (measured in base pairs) and worked well to produce a single qRT-PCR product (Figure 22). Further qRT-PCR tests using the CD83, GAPDH, and ACTB primers had products present in the no RT and ddH_2O negative controls, indicating these primers need to be redesigned. Additional DNA agarose gels assessing the qRT-PCR products from the sample

testing, no RT, and ddH_2O trials confirmed the IGFBP5, TNNC1, CRIP2, GATA2, and TBP primers to be feasible for further use in the FFPE samples (Figure 23). The TBP and EGFR commercial primers were also determined to be feasible for further use (Figure 24). To improve the primer designs, other parameters may be adjusted, such as melting size and a more stringent melting temperature. It is also preferred that the primer pairs and test and reference amplicons be of equal lengths.[155] Therefore, to ensure the qRT-PCR results are reproducible, and to improve the qRT-PCR results from the 13137A sample, a more limited amplicon length and equal primer lengths within pairs per gene can be set during primer designing.

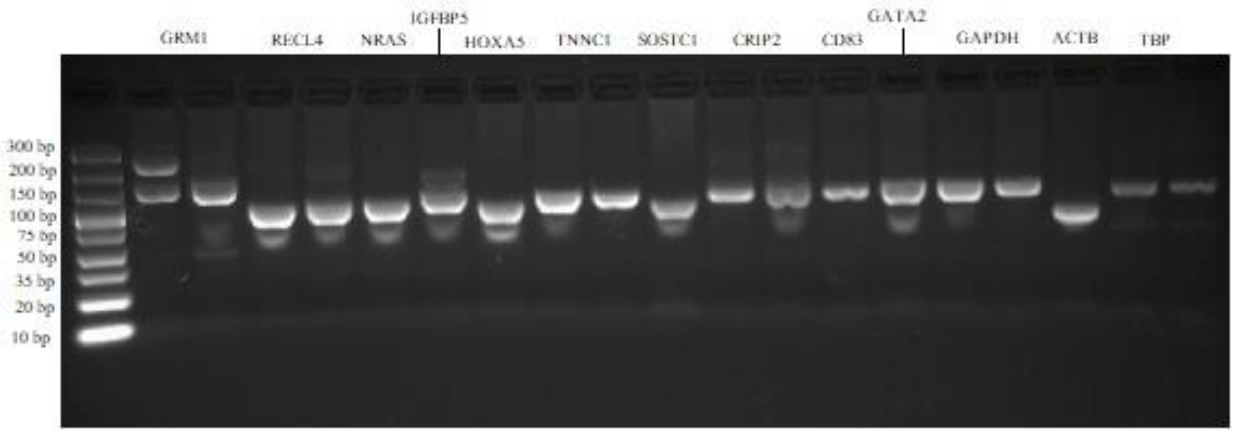

Figure 22: DNA agarose gel of qRT-PCR GRM1, RECQL4, NRAS, IGFBP5, HOXA5, TNNC1, SOSTDC1, CRIP2, CD83, GATA2, GAPDH, ACTB, and TBP products.
The products are the result of the melting temperatures (T_ms) (from left to right) 56.3˚C and 52.0˚C (*GRM1*), 60.0˚C and 56.3˚C (*RECQL4*), 50.7˚C (*NRAS*), 56.3˚C (*IGFBP5*), 52.0˚C (*HOXA5*), 58.3˚C and 56.3˚C (*TNNC1*), 52.0˚C (*SOSTDC1*), 53.9˚C and 52.0˚C (*CRIP2*), 53.9˚C (*CD83*), 52.0˚C (*GATA2*), 58.3˚C and 52.0˚C (*GAPDH*), 52.0˚C (*ACTB*), and 52.0˚C and 50.7˚C (*TBP*) on custom-designed primers. Gradient qRT-PCRs were first performed on each primer to determine their optimum T_m. The products from the T_ms with the lowest qRT-PCR C_q value were run on the DNA gel. qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The primers are designed for amplicon sizes 99 bp (GRM1), 80 bp (RECQL4), 84 bp (NRAS), 95 bp (IGFBP5), 80 bp (HOXA5), 93 bp (TNNC1), 81 bp (SOSTDC1), 92 bp (CRIP2), 91 bp (CD83), 97 bp (GATA2), 95 bp (GAPDH), 61 bp (ACTB), and 86 bp (TBP).

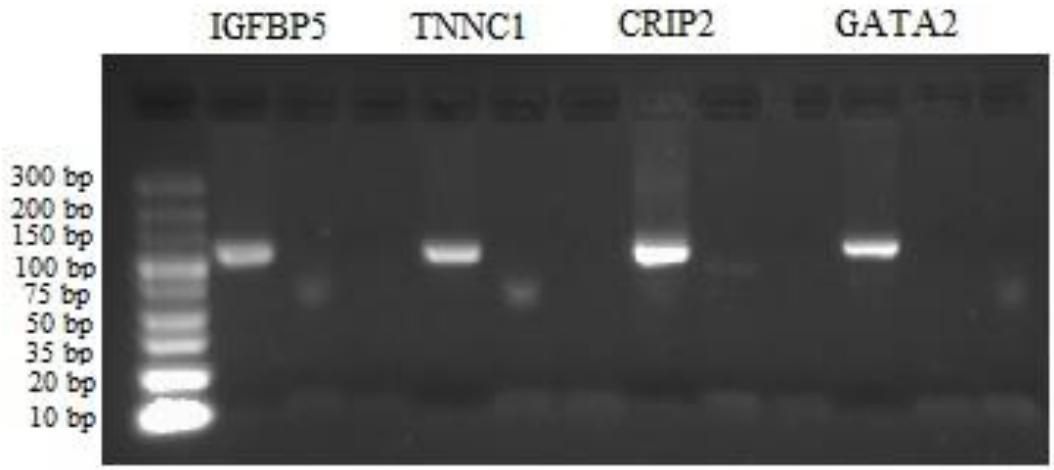

Figure 23: DNA agarose gel of qRT-PCR IGFBP5, TNNC1, CRIP2, and GATA2 products from the A549 cell line.

Gradient qRT-PCRs were first performed on each primer to determine their optimum melting temperatures (T_m), which was taken as 56.3˚C for each custom design primer. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The primers are designed for amplicon sizes 95 bp (IGFBP5), 93 bp (TNNC1), 92 bp (CRIP2), and 97 bp (GATA2).

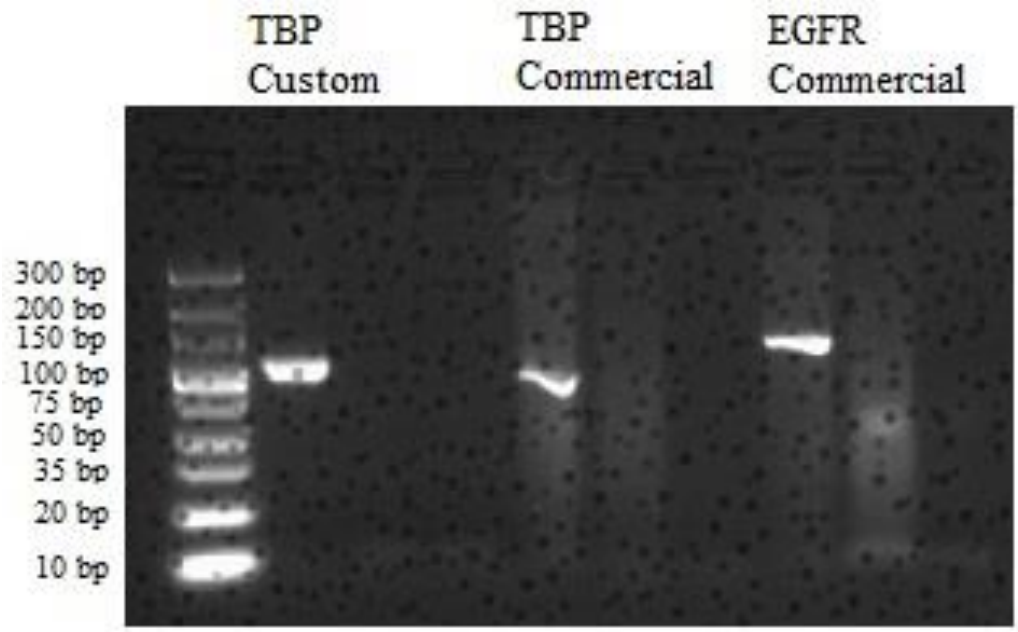

Figure 24: DNA agarose gel of qRT-PCR TBP and EGFR products from the A549 cell line. A gradient qRT-PCR was first performed on the custom primer to determine its optimum melting temperature (T_m). The products are the result of 56.3˚C and 60˚C T_ms on custom design and commercial primers, respectively. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The TBP custom primer is designed for amplicon size 86 bp.

The remainder of the DNA agarose gels confirmed the custom-designed primers were working in the FFPE samples (Figures 25, 26, 27, 28, 29, and 30). The gel for *TNNC1* did not have any product from the 13137A FFPE sample (Figure 26), which correlates to its low-expression level from the qRT-PCR results (Figure 32). The gel for *GATA2* had some lowly-expressed products in the negative controls, which were considered non-significant (Figure 28).

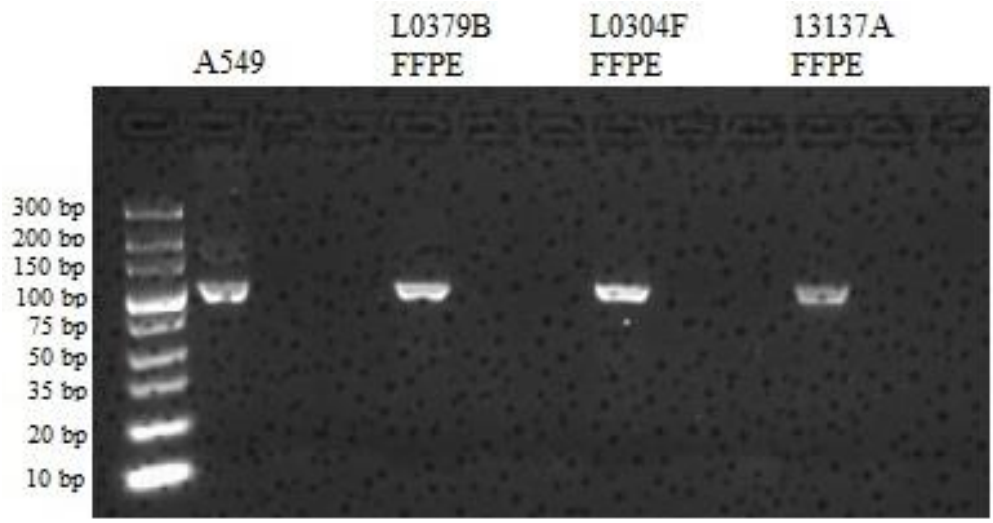

Figure 25: DNA agarose gel of qRT-PCR IGFBP5 products from the A549 cell line and FFPE test samples.

Gradient qRT-PCRs were first performed on the custom primer to determine its optimum melting temperature (T_m), which was taken as 56.3˚C. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The IGFBP5 custom primer is designed for amplicon size 95 bp.

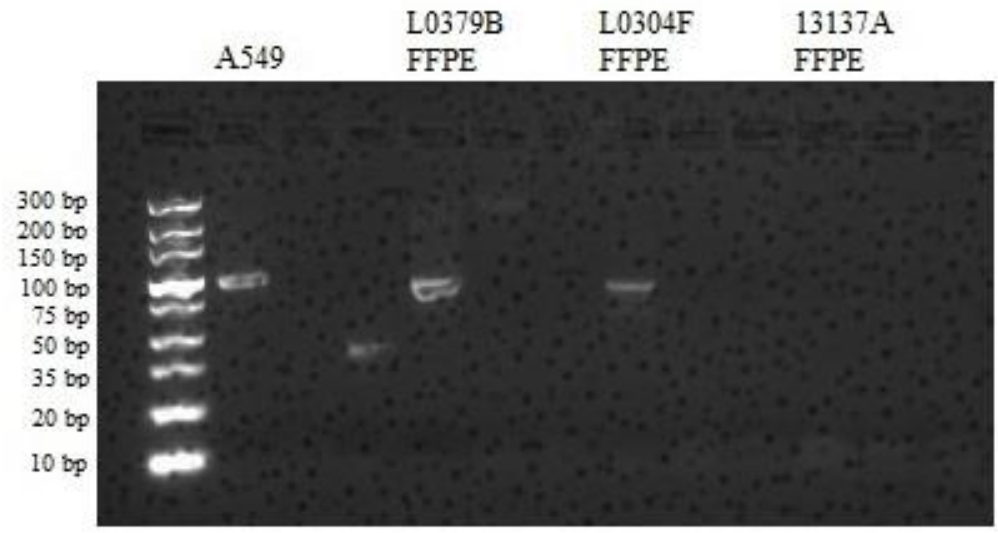

Figure 26: DNA agarose gel of qRT-PCR TNNC1 products from the A549 cell line and FFPE test samples.

Gradient qRT-PCRs were first performed on the custom primer to determine its optimum melting temperature (T_m), which was taken as 56.3˚C. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The TNNC1 custom primer is designed for amplicon size 93 bp.

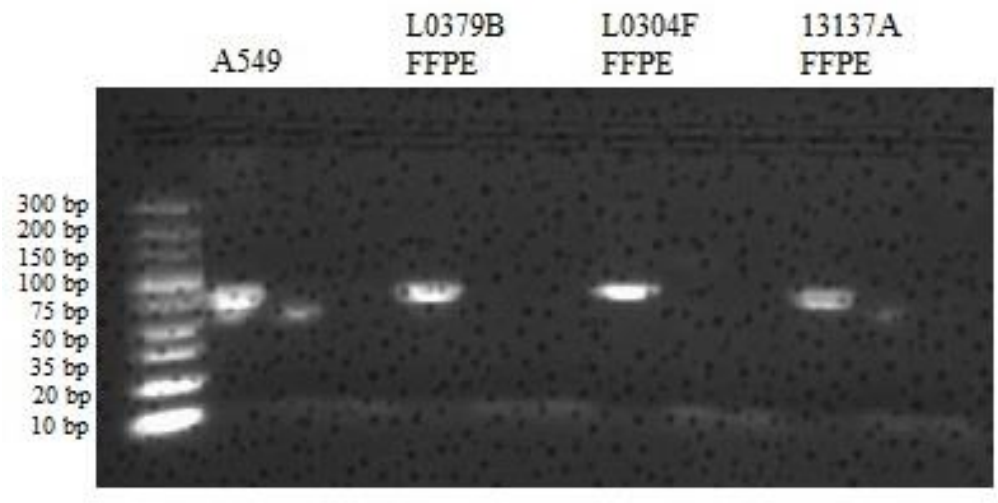

Figure 27: DNA agarose gel of qRT-PCR CRIP2 products from the A549 cell line and FFPE test samples.
Gradient qRT-PCRs were first performed on the custom primer to determine its optimum melting temperature (T_m), which was taken as 56.3˚C. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The CRIP2 custom primer is designed for amplicon size 92 bp.

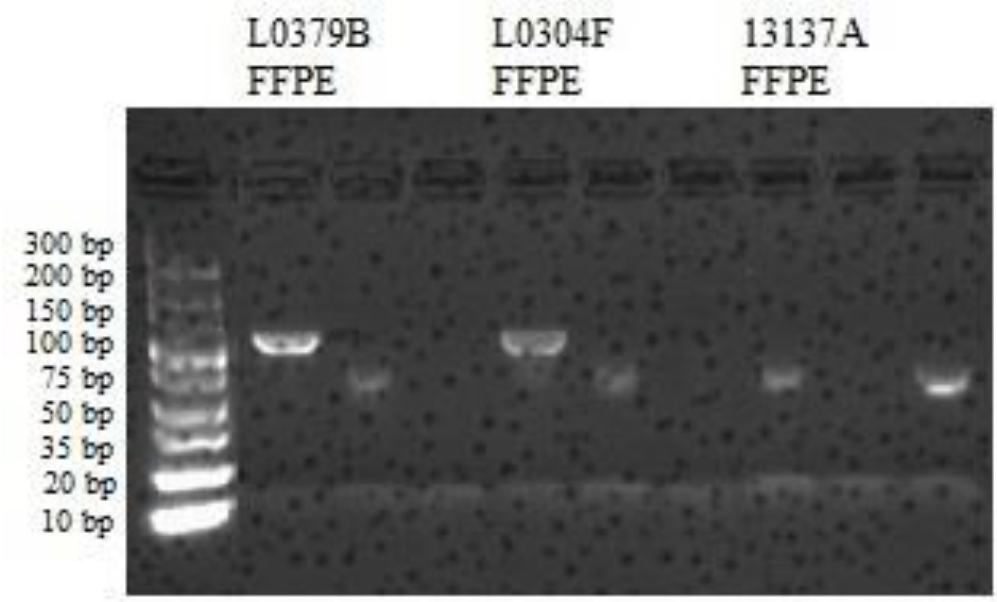

Figure 28: DNA agarose gel of qRT-PCR GATA2 products from the FFPE test samples. Gradient qRT-PCRs were first performed on the custom primer to determine its optimum melting temperature (T_m), which was taken as 56.3˚C. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes. The GATA2 custom primer is designed for amplicon size 97 bp.

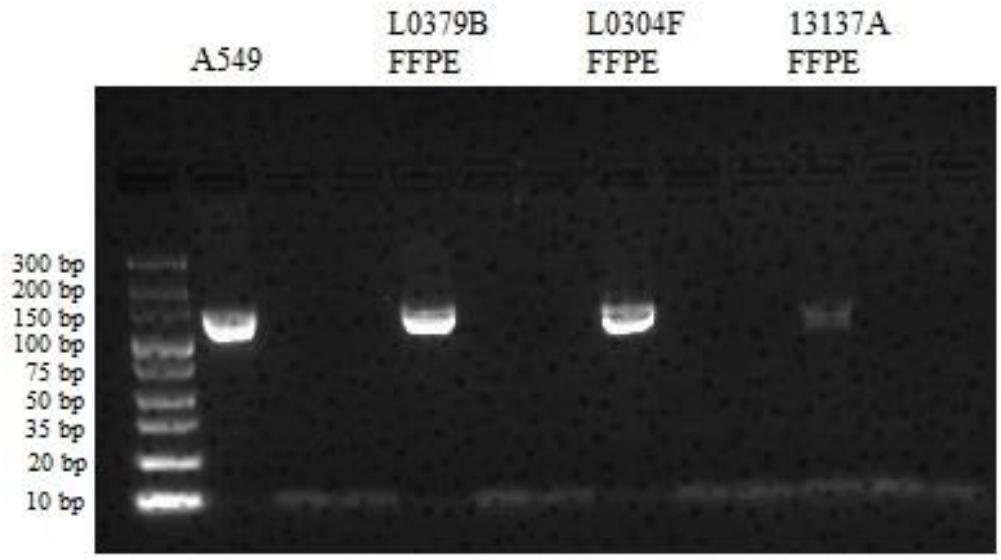

Figure 29: DNA agarose gel of qRT-PCR TBP products from the A549 cell line and FFPE test samples.

An optimum melting temperature (T_m) of 60.0˚C for the commercial primer was used. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes.

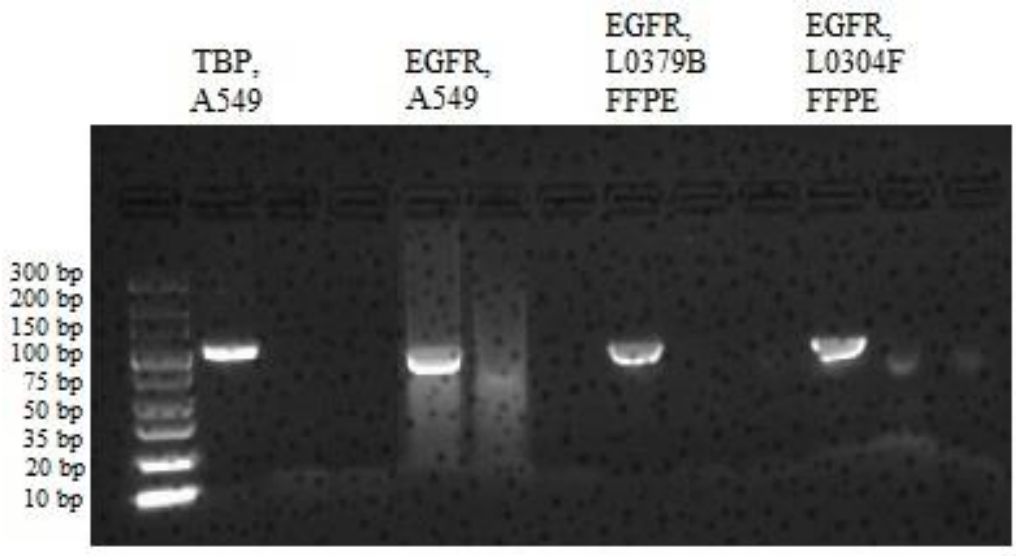

Figure 30: DNA agarose gel of qRT-PCR TBP products from the A549 cell line and FFPE test samples.
An optimum melting temperature (T_m) of 60.0˚C for the commercial primer was used. The products from the T_ms with the lowest qRT-PCR quantitation cycle (C_q) values were run on the DNA gel. The first lane per gene tested contains qRT-PCR product produced from cDNA reverse transcribed with Reverse Transcriptase (RT), the second lane qRT-PCR product produced from RNA that underwent the cDNA synthesis protocol lacking RT (no RT control), and third lane qRT-PCR product of double-distilled H_20 instead of cDNA sample (negative control). qRT-PCR was performed on approximately 15 ng RNA extracted from the A549 cell line and FFPE test samples. Invitrogen™ TrackIT™ Ultra Low Range DNA Ladder was used to measure the product sizes.

3.3.4 qRT-PCR of Yin and Yang Genes

According to the 10-gene YMR signature, *IGFBP5* should be over-expressed in tumor cells while *CRIP2*, *TNNC1*, and *GATA2* should be under-expressed when compared to normal cells. Therefore, *IGFBP5* may have a relatively higher expression level than the other three genes. The gene expression from conducting qRT-PCRs using the A549 cell line, however, resulted in *CRIP2* displaying high expression level while *IGFBP5* had the lowest (Figure 31). However, when the L0379B and L0304F FFPE samples were used, the expression of *IGFBP5* was much greater than *TNNC1*, *CRIP2*, and *GATA2*, as expected (Figure 32).

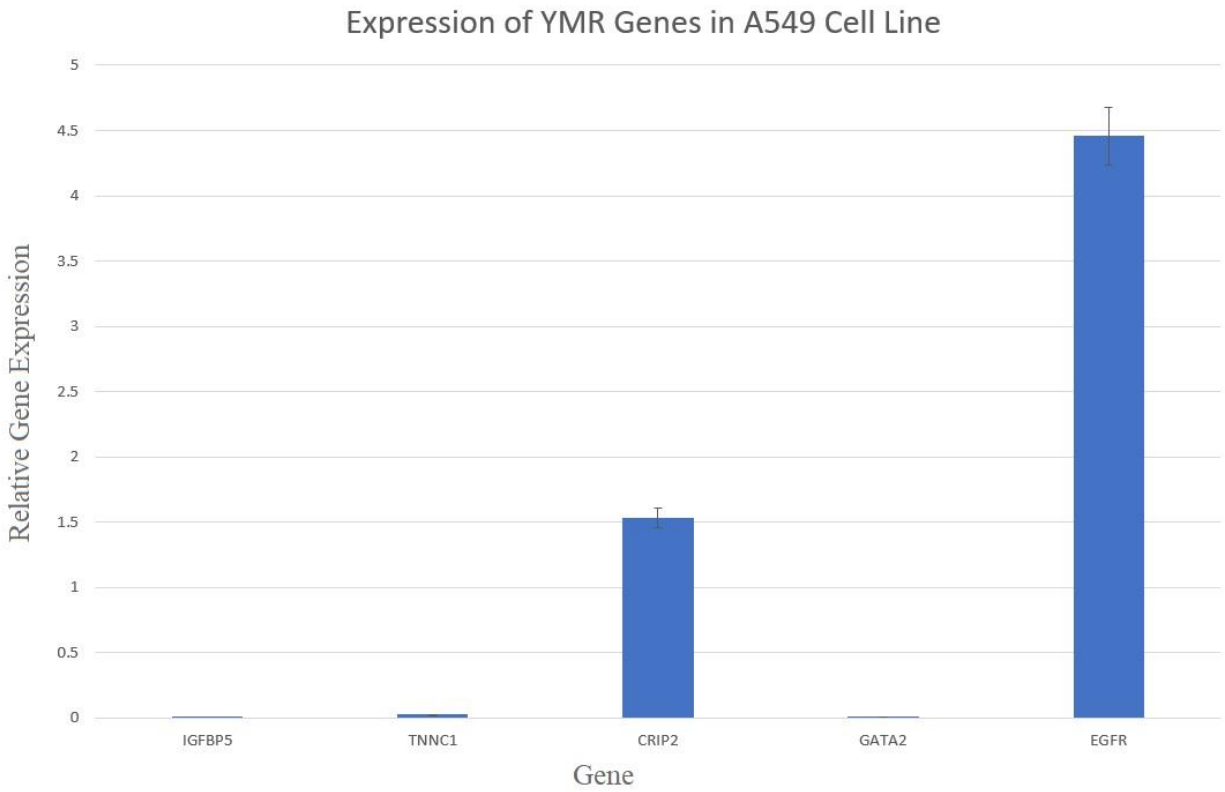

Figure 31: Comparison of IGFBP5, TNNC1, CRIP2, and GATA2 YMR and EGFR positive control gene expression in the A549 cell line.
Custom-designed primers were used for the YMR genes and a commercial primer was used for the positive control gene. *IGFBP5* is expected to be over-expressed in NSCLC, while *TNNC1*, *CRIP2*, and *GATA2* are expected to be under-expressed. An average of three tests are shown; each test well used 15 ng of sample cDNA and each test included three replicates of testing sample, no reverse transcriptase control, and negative double-distilled water control. Standard error is shown. Data was normalized using the $2^{-\Delta Cq}$ normalization method with the *TBP* control gene.

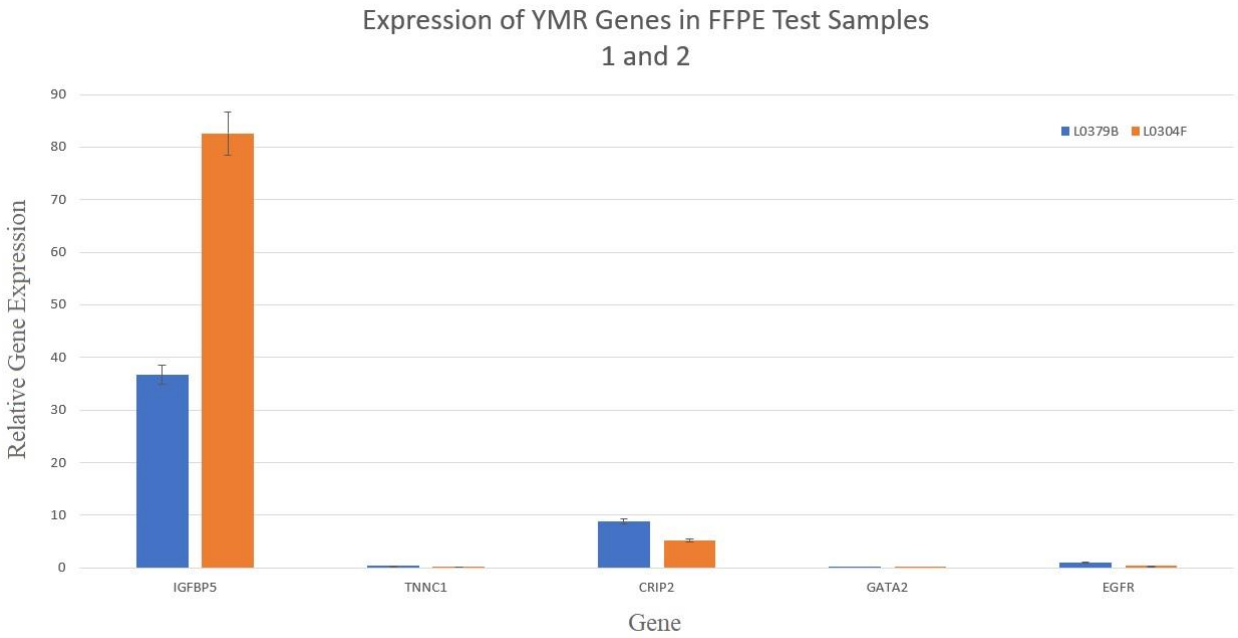

Figure 32: Comparison of IGFBP5, TNNC1, CRIP2, and GATA2 YMR and EGFR positive control gene expression in lung FFPE test samples L0379B and L0304F.
Custom-designed primers were used for the YMR genes and a commercial primer was used for the positive control gene. *IGFBP5* is expected to be over-expressed in NSCLC, while *TNNC1*, *CRIP2*, and *GATA2* are expected to be under-expressed. An average of three tests are shown; each test well used 15 ng of sample cDNA and each test included three replicates of testing sample, no reverse transcriptase control, and negative double-distilled water control. Data was normalized using the $2^{-\Delta Cq}$ normalization method with the *TBP* control gene.

3.3.5 Cell Line Database Comparison

Since the expression levels of the YMR genes from the qRT-PCR results using the A549 did not mirror the expected YMR gene signature, NSCLC gene expression data from cell lines on publicly available databases were downloaded to compare to the qRT-PCR results and assess for the expected YMR signature expression pattern. YMR gene expression data from 37 cell lines were downloaded from the Cancer Cell Line Encyclopedia (CCLE) database. The expression levels were compared between the genes per cell line, followed by ranking of the genes on a scale of 1 to 10, with 1 representing the highest expression level in a cell line and 10 the lowest expression level. Contrary to our expectations of *GRM1* having the highest expression level, it was ranked in the 9th and 10th positions in 36 of the cell lines. *RECQL4* and *NRAS* had rankings

in the 1st and 2nd places for a majority of the cell lines, similar to the YMR signature pattern. *SOSTDC1* had a low expression ranking in 27 cell lines, as expected, whereas *GATA2* was highly expressed in many cell lines, opposing our expected pattern. The remainder of the genes varied in their expression rankings across the cell lines (Table 12). Similar expression patterns were seen in the 28 cell lines downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) database, specifically, *GRM1* had the lowest and second lowest expression levels out of the YMR genes in 28 of the cell lines and *RECQL4* and *NRAS* had high expression levels (Table 13). With the exception of *GRM1*, three CCLE cell lines and one GDSC cell line mirrored the expression pattern of the YMR signature, with *IGFBP5, NRAS,* and *RECQL4* having the highest expression levels out of the 10 genes (Table 14).

Table 12: Yin and Yang gene rankings across the 37 NSCLC cell lines from the Broad Institute Cancer Cell Line Encyclopedia online database.
Each gene was ranked in each cell line, with 1 representing the highest expression level and 10 representing the lowest expression level. The total counts per each gene ranking were summed. RNA-Seq gene expression data was downloaded.

Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
GRM1	0	0	0	0	0	0	1	0	14	22
RECQL4	5	21	5	4	2	0	0	0	0	0
NRAS	18	8	9	2	0	0	0	0	0	0
IGFBP5	4	0	3	2	0	4	8	13	3	0
HOXA5	0	0	2	3	7	9	8	4	2	2
TNNC1	4	1	3	5	3	6	8	5	0	2
SOSTDC1	1	0	0	0	0	2	0	7	17	10
CRIP2	5	5	7	4	2	7	5	2	0	0
CD83	0	0	4	3	16	6	4	4	0	0
GATA2	0	2	6	14	7	3	3	2	0	0

Table 13: Yin and Yang gene rankings across the 29 NSCLC cell lines from the Genomics of Drug Sensitivity in Cancer online database.

Each gene was ranked in each cell line, with 1 representing the highest expression level and 10 representing the lowest expression level. The total counts per each gene ranking were summed. RNA-Seq gene expression data was downloaded. HOXA4 data was not available.

Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
GRM1	0	0	0	0	0	0	1	6	22
RECQL4	2	10	9	7	1	0	0	0	0
NRAS	21	6	2	0	0	0	0	0	0
IGFBP5	4	1	0	0	1	2	4	13	4
TNNC1	0	2	5	1	4	7	9	1	0
SOSTDC1	1	1	0	1	4	8	6	6	2
CRIP2	1	7	7	10	3	1	0	0	0
CD83	0	1	3	9	4	5	4	2	1
GATA2	0	1	3	1	12	6	5	1	0

Table 14: YMR signature gene expression levels in the NCI-H2066, NCI-H810, NCI-H292 (CCLE), and IA-LM (GDSC) NSCLC cell lines.

RNA-Seq gene expression data was downloaded from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) online databases.

Cell Line	**NCI-H2066**	**NCI-H810**	**NCI-H292**	**IA-LM**
GRM1	-1.73398	-6.55673	-8.80399	3.003504
RECQL4	5.384078	4.907503	3.286756	6.55568
NRAS	5.093513	5.195477	4.608559	8.298138
IGFBP5	8.843143	4.671668	1.979979	6.564764
HOXA5	2.744343	2.43317	0.494151	NA
TNNC1	-2.86832	-2.9962	1.400408	3.29924
SOSTDC1	-1.49269	-6.06018	-13	3.259721
CRIP2	1.449656	1.776649	0.11251	4.258763
CD83	2.594576	0.023134	1.38373	3.76698
GATA2	2.537922	3.748687	1.515416	4.304125

Chapter 4: Discussion and Conclusion

4.1 Discussion

Earlier studies aiming to establish gene signatures for predicting the outcome of NSCLC patients have been shown to be unsuccessful in their reproducibility and progression into clinical use. We have previously established a 10-gene signature that assesses the opposing effects of two groups of genes: Yin (over-expressed in tumor cells) and Yang (over-expressed in normal cells). This is different from the common approach of deriving correlation coefficients between gene expression data and patient survival time in training datasets, followed by validation in testing datasets.[69, 70] The Yin and Yang genes have been implicated in a variety of cancers and play roles in DNA replication, cytoskeleton maintenance and regulation, antigen presentation, and cell migration, proliferation, differentiation, and apoptosis. Assessment of CNV as a covariate correlated to the expression of the YMR signature genes revealed it to be significant for *GRM1, HOXA5, TNNC1, CRIP2,* and *CD83,* while DNA methylation was pertinent for all genes except *CD83*. MLR analyses demonstrated the combination of CNV and DNA methylation to be significant in the modulation of *RECQL4, NRAS, IGFBP5, HOXA5, TNNC1, SOSTDC1,* and *CRIP2* expression. Although the expected expression pattern for the *IGFBP5, TNNC1, CRIP2,* and *GATA2* was not seen in the A549 cell line using the chosen biomarker expression technique, qRT-PCR, assessments using FFPE lung tumor samples revealed the expected pattern. The assessment of covariates associated with the YMR signature genes' expression levels and validation of the YMR signature using qRT-PCR formed the basis of my research.

Copy number variations (CNVs) are DNA segments equal to or greater than 1 kb in size that present in varying copy numbers when compared to a reference genome. Therefore, there may be

more or less copy numbers of the DNA segments than the reference genome, resulting in alterations of gene dosage and expression and interrupting coding sequences, and ultimately affecting an individual's phenotype.[54, 157, 158] The increased copy number implies an increased gene expression. For example, studies have shown *PI3KCA* and *EGFR* to have an increased copy number and expression in squamous cell carcinomas, conferring a growth advantage to cancer cells and poor prognosis to patients, respectively.[159–161] All simple linear regression models for the correlation of CNV with gene expression of the YMR signature genes demonstrated the anticipated trend of an increase in CNV being correlated with an increase in gene expression (Figures 2, 4-11). We expect an increased copy number and expression of Yin genes and decreased copy number and expression of Yang genes in the NSCLC patients.

Various studies have demonstrated the anticipated individual YMR gene expression levels in a range of cancer samples. Overexpression and irregular mutations of *GRM1* have been implicated in various types of cancer such as breast, renal cell carcinoma, and melanomas, resulting in tumor progression, development, and growth. Wangari-Talbot et al. further assessed findings that a decrease in GRM1 activity is associated with reduction of *in vitro* cell growth and *in vivo* tumor progression by suppressing *GRM1* expression in a few human melanoma cell lines. As expected, a decrease in viable cell number and tumor progression were observed.[71] RECQL4 is a tumor promoter that is over-expressed in cancers such as hepatocellular carcinoma, gastric, prostate, breast, and colorectal.[162–167] For example, Arora et al. linked increased copy number, mRNA levels, and protein levels of RECQL4 with an aggressive phenotype in breast cancers that includes lymph node metastasis, large tumor size, and poor survival, whereas its depletion resulted in increased sensitivity to chemotherapy in cultured cells. [166] Although the

expression levels of *NRAS* have not been extensively studied in cancers, the RAS family of GTPases have been shown to be commonly mutated in a variety of cancers, such as colorectal and thyroid cancers and advanced melanoma.[168–170] *IGFBP5* is up-regulated in more advanced estrogen receptor negative breast cancers with lymph node metastases, conferring poor prognoses to these patients.[171] Estrogen receptor positive patients presenting with lower levels of *IGFBP5* had better prognoses over patients with higher levels.[172] In a study comparing *IGFBP5* expression levels, samples of intrahepatic cholangiocarcinoma, cancer that arises in the bile ducts, were seen to have the highest expression over samples of normal liver, chronic liver disease, and other cancers (breast, colon, stomach, ovary, and lung).[173] Similarly, *IGFBP5* is has been shown to be over-expressed in pancreatic adenocarcinomas.[174] These studies confirm that the Yin genes are up-regulated in certain cancers and have the potential to be up-regulated in NSCLCs.

As expected, HOXA5 has been shown to be down-regulated and/or act as a tumor suppressor in several cancer types. Ordóñez-Moràn et al. confirmed the decreased expression of *HOXA5* in colon cancer and demonstrated that its re-expression halts tumor progression and metastasis.[175] In lung adenocarcinoma, its *in vitro* ectopic expression in invasive cancer cells reduced cell invasion, migration, filopodia formation, *in vivo* expression repressed metastasis, and its knockdown promoted lung cancer cell invasiveness.[176] For example, NSCLC samples tend to have decreased expression levels of *HOXA5* over their adjacent normal tissues and those with low levels of HOXA5 or lacking expression of both *p53* and *HOXA5* demonstrated the poorest prognosis.[177] In non-muscle cells, TNNC1 is involved in cellular migration and locomotion, cytoplasmic streaming, cytokinesis, and the cell-matrix adherens junction.[178]

Although the role of TNNC1 has not been extensively studied in lung cancers, several studies have assessed its gene expression levels in NSCLC and found it to be one of the top down-regulated genes.[179–181] *SOSTDC1* ultimately affects cell proliferation, differentiation, and apoptosis. Liu et al. demonstrated that the expression levels of *SOSTDC1* in NSCLCs may aid in the prediction of patient outcome since higher levels of the expressed gene confer a better prognosis. In a series of experiments, they were able to show that *SOSTDC1* is downregulated in NSCLCs and its ectopic expression in the A549 and NCO-H520 cell lines halted cell proliferation and resulted in smaller tumors with slow growth rates.[182] Chen et al. also revealed low expression levels of *SOSTDC1* confers a poor prognosis to patients while its over-expression reduces NSCLC cell proliferation, invasion, and migration.[183] *CRIP2* functions in cytoskeletal modulation and localization of actin-rich structures, with a possible role in regulation of actin dynamics and/or cell migration. Lo et al. and Cheung et al. demonstrated that *CRIP2* acts as a tumor suppressor and has reduced expression in esophageal squamous cell carcinoma (ESCC) and nasopharyngeal carcinoma cell line and tumor samples, respectively. Expression of *CRIP2* in ESCC can cause reduction of cell colony formation, growth, and invasion, whereas its over-expression results in apoptosis.[184, 185] Although *CRIP2* has not been extensively studied in lung cancers, it has been shown to be differentially expressed between lung adenocarcinomas and squamous cell carcinomas and have reduced expression in the NCI-H460-LNM35 lung cancer cell line.[186, 187] *CD83* functions in antigen presentation and enhancing T-cell activation and immunity. Several studies have shown that CD83 is implicated in cervical, breast, gastric, and lung cancers.[188–191] Kaplan-Meier analyses in human breast carcinomas associated increased levels CD83+ being expressed in tumor-infiltrating dendritic cells (TIDCs) with longer relapse-free patient survival, whereas it is

negatively correlated to lymph node metastasis. Those patients with lymph node metastasis and higher numbers of CD83+ TIDCs in breast and gastric cancers demonstrated a better outcome than those with fewer numbers.[189, 190] Although *CD83* is expected to be under-expressed in NSCLC tissue samples according to the YMR gene signature, Baleerio et al. assessed *CD83* expression in a number of lung cancer cell lines, showing it to be highly expressed.[191] Abnormal functioning of GATA2 has been found in renal, prostate, breast, and lung cancers.[192–197] As expected from the YMR gene signature, GATA2 is down-regulated in hepatocellular carcinoma cell lines and tissues compared to normal hepatocellular cells, resulting in tumor sizes greater than 5 cm, advanced TNM staging, increased rate of recurrence, and decreased overall survival in patients.[192] Similarly, GATA2 has been shown to be decreased in clear renal cell carcinomas, conferring advanced tumor staging, distant metastasis, and lymph node metastasis to patients.[193] Tessema et al. proved *GATA2* is down-regulated in mutant NSCLC cell lines and primary tumors compared to the respective normal lung tissue.[197] These studies confirm that the Yang genes have repressed expression in certain cancers and have the potential to be down-regulated in NSCLCs.

Methylation of a gene's promoter region results in chromatin condensation and the transcriptional silencing of that gene. Many tumor suppressor genes within cancers have been shown to be hypermethylated, and therefore transcriptionally silenced with little or no gene product. This implies that an increase in DNA methylation results in a decrease in gene expression. Within the YMR gene signature, this expected trend was seen in the SLR models depicting the correlation between DNA methylation and gene expression for all genes, with the exception of *CD83* and *GATA2* (Figures 2-9). DNA methylation has been studied with respect to

a few of the YMR signature genes. For example, Raman et al. have linked the lack of *HOXA5* to cause the loss of *p53* expression in breast cancer, possibly due to *HOXA5* promoter hypermethylation.[88] Methylation impacting *SOSTDC1* expression was observed by Rawat et al., who assessed the epigenetic regulation of *SOSTDC1* in breast cancers and found its down-regulation to be associated with the hypermethylation of its promotor region.[198] Similarly, Gopal et al. found reduced mRNA expression of *SOSTDC1* to be related to its promoter methylation in gastric cancer cell line and tissue samples when compared to normal gastric tissues.[199] Tessema et al. associated the down-regulation of *GATA2* in mutant NSCLC cell lines and primary tumors compared to the respective normal lung tissue with methylation of its promoter region as the normal lung tissue was unmethylated.[197] These studies demonstrate the role of DNA methylation in the transcriptional silencing of the *HOXA5*, *SOSTDC1*, and *GATA2* Yang genes and suggest that the regression model results demonstrating a correlation between increasing DNA methylation and decreasing gene expression observed may be valid.

To further evaluate the YMR signature genes, a feasible gene expression detection assay must be chosen. Assessment of the biomarker expression detection assays for sensitivity, specificity, cost, and time, among other parameters, revealed qRT-PCR to be a feasible assay for the assessment of the YMR gene signature in a clinical setting (Tables 10 and 11). With a sample processing fee of as little as $0.50 for a single gene, an experimental time of about 5.5 hours that includes RNA isolation and purification, as well as PCR and data analysis, qRT-PCR offers the cheapest and quickest assay for processing an individual clinical sample for 1-50 genes.[133] Small laboratories such as a hospital lab or a clinic lab suited to a clinical processing environment can purchase a thermal cycler for as little as $25 000 to process and analyze their sample.[141] Other

techniques employ machines that can cost up to $285 000, which are better suited for larger research institutes and feasible if clinical samples are sent to these institutes or a national lab to be processed.[144, 145] With the cost constraint, the nCounter is suited for detection of 50-500 genes and microarray is good for more than 500 genes or genome-wide gene expression profiling.[150, 200] RNA-Seq can detect a large number of genes with the additional benefit of gene discovery, which is suited for cancer research institutions.[201, 202] TMAs are appropriate for protein level detection in thousands of samples at one time, which is ideal for tumor archive centers at provincial or national facilities.[17, 109, 122]

In addition to the above proposed guidelines, developing an assay into clinical use must pass stringent Clinical Laboratory Improvement Amendments (CLIA) standards and a College of American Pathology certification. Diagnostic tests can result in large costs due to demonstrating improved patient mortality outcomes through replicated and randomized clinical trials.[112] The cost of an assay, then, is an important parameter in experimental design and should be minimized for the desire to translate results into clinical use. Medical costs are increasing in the current healthcare system as cures are being sought for all illnesses and the general population ages and increases in life expectancy, resulting in more money being put towards the elderly population's care and medical costs. Cost-effectiveness studies may be performed on an assay to determine its economic utility for a consumer's and society's well-being.[113, 203] In addition to the necessity of developing an assay with low cost, the FDA guidelines for approval must be met. Once FDA-approved, it is then up to clinicians to accept and employ the assay.[17, 113]

While using qRT-PCR to further assess the YMR gene signature, comparison of the YMR genes' expression levels is inconsistent between the A549 cell lines and FFPE samples. According to the 10-gene YMR signature, *IGFBP5* should be over-expressed in tumor cells while *CRIP2*, *TNNC1*, and *GATA2* should be under-expressed when compared to normal cells. The gene expression from conducting qRT-PCRs using the A549 cell line, however, resulted in *CRIP2* displaying a relatively high expression level while *IGFBP5* had the lowest. However, when the L0379B and L0304F FFPE samples were used, the expression of *IGFBP5* was relatively higher than *TNNC1*, *CRIP2*, and *GATA2*, as expected. This indicates that the A549 cell line may not be a feasible model when assessing the YMR gene expression levels from lung tumor FFPE samples. Several studies have compared the differences between cell lines and their corresponding tumor tissue types and shown that cell lines are not entirely representative of their respective *in vivo* cancers. van Staveren et al. demonstrated that thyroid cell lines present with a p53 inactivation and do not express the functional genes for thyroglobulin, thyroperoxidase, and thyroid stimulating hormone receptor.[204] Studies by Ertel et al. showed that breast, colon, kidney, nervous system, ovary, and prostate cell lines had a larger number of differentially expressed genes in comparison to their respective normal tissues. Breast cancer cell lines have also displayed more activating mutations of kinases than their cancers.[205] Wistuba et al. conducted a comprehensive study assessing the morphological features, aneuploidy, mutation of *TP53*, and expression of estrogen receptor, progesterone receptor, epidermal growth factor receptor, and p53 proteins between breast cancer cell lines cultured for up to 60 months and their corresponding archived tissue samples. Great similarity was seen between the sample types as they demonstrated correlations of 100%, 87%, 75%, 87%, 73%, 93%, and 100%, implying that cell lines are useful in studying breast cancer.[206] In another study, Witsuba et al. analyzed the

differences between lung cancer cell lines and tissues samples on a number of factors. Similarly, great correlations were seen, such as 100% for morphology, p53 expression, and *K-ras* mutations. However, some discrepancies noted include a greater number of aneuploid subpopulations and incidences of *TP53* mutations in the cell lines.[207] In addition to genetic drift being present in cell lines, tumor samples are more representative of the tumor environment and have a heterogeneous representation of other cells, whereas cell lines are cultured in an artificial environment and have a more homogeneous composition of tumor cells.

Although limited in supply, human tissue samples are advantageous in research as they represent the complex state of a tumor *in vivo* at a certain time point or staging of a disease and can be used to study disease pathology, gene expression, and metabolism.[204] Alternatively, immortal cancer cell lines, which have been established from a cancer sample in the past, are cost-effective, more likely to offer reproducible results, and simple to use in terms of providing unlimited material and not requiring the same ethical principles and guidelines for use as animal and human tissues samples. They are expected to mimic and retain the features from primary cells to provide an appropriate model of human pathophysiology.[207, 208] For example, hallmark gene mutations, deletions, insertions, amplification, and silencing have been seen in both cell lines and tumors. As a cell culture passage number increases, the cell line may acquire additional genetic or epigenetic changes, stem-cell like features, and activation of telomerase, which can potentially confer immortality to the cell line. Genetic drift where the frequency of a gene variant is altered over time through gene mutations is also likely to occur, resulting in heterogeneity of the cell culture. There is skepticism around the validity of cell line use in research due to contamination, genomic instability, absence of tumor environment components

(such as stromal and immune cells, vasculature, and inflammatory factors), and the belief that some phenotypic features have been lost and molecular changes occurred through passages of cell lines since establishment from the tumor source.[208, 209] Tumor samples are also advantageous over cell lines in retaining the inter-tumor variability and diverse environment as seen in the body and include both malignant and non-malignant cells. Therefore, cell signalling between these cells cannot be studied using cell lines. Patient clinical information such as ethnicity, geographical locations, and sex can also be assessed more in-depth using tumor samples.[207] Since studies have shown that there may not be a correlation in results observed from cell line and tissue samples, and that these discrepancies may be attributable to genetic drift and differing environments between the samples, the inconsistencies observed in the YMR gene qRT-PCR tests imply that cell lines may not be optimal when assessing the YMR genes' expression. Therefore, the use of tissue samples, such as FFPE samples, are warranted for further assessing the YMR signature genes.

4.2 Conclusion

In conclusion, this study has shown that the gene expression levels of the *GRM1, HOXA5, TNNC1, CRIP2,* and *CD83* YMR genes are correlated to their CNVs, while the expression levels of all the YMR genes, with the exception of *CD83*, are correlated to their DNA methylation. These two covariates in combination were demonstrated to be closely correlated with *RECQL4, NRAS, IGFBP5, HOXA5, TNNC1, SOSTDC1,* and *CRIP2* expression. Furthermore, the chosen biomarker expression technique for clinical use, qRT-PCR, confirmed that *IGFBP5* has a high expression level in NSCLC FFPE samples, whereas *TNNC1, CRIP2*, and *GATA2* had lower expression levels. However, there was minimal expression of *IGFBP5* in the A549 cell line and a

higher expression level of *CRIP2* in the FFPE samples, implying that use of cell lines have significant limitations when assessing the expression levels of the YMR signature genes. Further research is required to validate these claims. Although none of the NSCLC cell line data downloaded from the CCLE and GDSC demonstrated the expected YMR signature gene expression pattern, the YMR genes have independently demonstrated their expected expression levels in other various cancers.

Chapter 5: Future Directions

This research sought to validate the previously established YMR 10-gene signature using FFPE samples and further establish it for clinical use. The results of our current studies have confirmed the expected expression level pattern for *IGFBP5*, *TNNC1*, *CRIP2*, and *GATA2* using FFPE samples and shown that they are correlated to the gene's DNA methylation, and with the exception of *GATA2*, CNV and DNA methylation. Further studies will aim to assess additional covariates that are associated with the YMR gene expression levels, such as miRNA and transcription binding factors. Incorporation of clinical information (patient sex, smoking status, and treatment status, to name a few) will evaluate whether the tailoring of the YRM gene signature to specific traits will be beneficial for certain groups of patients. The addition of more covariates will modify the regression models and may improve them so that they are significant for all genes in assessing which factors correlate with their expression levels. Moreover, completion of the qRT-PCR primer designs for *GRM1*, *RECQL4*, *NRAS*, *HOXA5*, *SOSTD1*, and *CD83* and expression level measurements of all the YMR genes using a larger cohort of FFPE samples will further validate the signature. Since the identified cell lines NCI-H292, NCI-H810, NCI-H2066, and IA-LM demonstrated similarities to our expected YMR signature gene expression levels, qRT-PCR experiments using the custom-designed YMR gene primers with these cell lines will confirm the validity of the primers and provide an affordable medium to further assess the biochemical pathways and implications of these genes in NSCLCs.

Chapter 6: References

1. Jemal A, R S, Ward E, Murray T, Xu J, Smigal C, et al. Cancer Statistics, 2006. CA Cancer J Clin. 2006;56:106–30.

2. Torre L, Bray F, Siegel R, Ferlay J, Lortet-Tieulent J, Jemal A. Global Cancer Statistics, 2012. CA Cancer J Clin. 2015;65:87–108.

3. Yang P. Epidemiology of Lung Cancer Prognosis: Quantity and Quality of Life. Methods Mol Biol. 2009;471:469–86.

4. Torre L, Siegel R, Jemal A. Lung Cancer Statistics. In: Lung Cancer and Personalized Medicine: Current Knowledge and Therapies. 2016. p. 1–19.

5. Schwartz A, Cote M. Epidemiology of Lung Cancer. In: Lung Cancer and Personalized Medicine: Current Knowledge and Therapies. 2016. p. 21–41.

6. Rivera G, Wakelee H. Lung Cancer in Never Smokers. In: Lung Cancer and Personalized Medicine: Current Knowledge and Therapies. 2016. p. 43–57.

7. Cafarotti S, Lococo F, Froesh P, Zappa F, Andre D. Target Therapy in Lung Cancer. In: Lung Cancer and Personalized Medicine: Current Knowledge and Therapies. 2016. p. 127–36.

8. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of Gene Expression in Adenocarcinoma of the Lung. Proc Natl Acad Sci U S A. 2001;98:13784–9. doi:10.1073/pnas.241500798.

9. Li X, Li J, Wu P, Zhou L, Lu B, Ying K, et al. Smoker and Non-Smoker Lung Adenocarcinoma is Characterized by Distinct Tumor Immune Microenvironments. Oncoimmunology. 2018;7:e149677.

10. Shahid M, Choi TG, Nguyen MN, Matondo A, Jo YH, Yoo JY, et al. An 8-gene Signature for Prediction of Prognosis and Chemoresponse in Non-Small Cell Lung Cancer. Oncotarget. 2016;7:86561–72.

11. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. Nat Med. 2002;8:816–24. doi:10.1038/nm.

12. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. N Engl J Med. 2007;356:11–20.

13. Kelsey CR, Marks LB, Hollis D, Hubbs JL, Ready NE, D'Amico TA, et al. Local Recurrence

After Surgery for Early Stage Lung Cancer: An 11-Year Experience with 975 Patients. Cancer. 2009;115:5218–27.

14. Hoffman PC, Mauer AM, Vokes EE. Lung Cancer. Lancet. 2000;355:479–85.

15. Cotter MB, Loda M. Introduction to Pathology. In: Loda M, Mucci LA, Mittelstadt ML, Van Hemelrijk M, Cotter MB, editors. Pathology and Epidemiology of Lung Cancer. Springer International Publishing Switzerland; 2017. p. 27–38.

16. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Proto-Oncogenes and Tumor-Suppressor Genes. In: Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000.

17. Narrandes S, Xu W. Gene Expression Detection Assay for Cancer Clinical Use. J Cancer. 2018;9:2249–65.

18. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (80-). 1999;286:531–7.

19. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer Biomarker Discovery and Validation. Transl Cancer Res. 2015;4:256–69.

20. Canadian Cancer Society. Staging Cancer. Canadian Cancer Society. 2019. http://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/stage-and-grade/staging/?region=ab.

21. American Joint Committee on Cancer. Cancer Staging System. American Joint Committee on Cancer. 2019. https://cancerstaging.org/references-tools/pages/what-is-cancer-staging.aspx.

22. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. Chest. 2017;151:193–203. doi:10.1016/j.chest.2016.10.010.

23. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016;11:39–51.

24. Canadian Cancer Society. Diagnosis of Lung Cancer. Canadian Cancer Society. 2019.

25. Mayo Clinic. Lung Cancer Diagnosis. Mayo Foundation for Medical Education and Research. 2019. https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627.

26. Annema JT, van Meerbeeck JP, Rintoul RC, Dooms C, Deschepper E, Dekkers OM, et al. Mediastinoscopy vs Endosonography for Mediastinal nodal Staging of Lung Cancer: A Randomized Trial. JAMA. 2010;304:2245–52.

27. Uramoto H, Tanaka F. Recurrence After Surgery in Patients with NSCLC. Transl Lung Cancer Res. 2014;3:242–9.

28. Prevention C for DC and. How is Lung Cancer Diagnosed and Treated? U.S. Department of Health & Human Services. 2018. https://www.cdc.gov/cancer/lung/basic_info/diagnosis_treatment.htm.

29. Canadian Cancer Society. Treatments for Non-Small Cell Lung Cancer. Canadian Cancer Society. 2019.

30. Chang A. Chemotherapy, Chemoresistance and the Changing Treatment Landscape for NSCLC. Lung Cancer. 2011;71:3–10.

31. Rossi A, Di Maio M. Platinum-based chemotherapy in advanced non-small-cell lung cancer: optimal number of treatment cycles. Expert Rev Anticancer Ther. 2016;16:653–60.

32. Georgoulias V, Papadakis E, Alexopoulos A, Tsiafaki X, Rapti A, Veslemes M, et al. Platinum-Based and Non-Platinum-Based Chemotherapy in Advanced Non-Small-Cell Lung Cancer: A Randomised Multicentre Trial. Lancet. 2001;357:1478–84.

33. Dasari S, Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. Eur J Pharmacol. 2014;740:364–78.

34. Fidia PM, Dakhil SR, Lyss AP, Loesch DM, Waterhouse DM, Bromund JL, et al. Phase III Study of Immediate Compared With Delayed Docetaxel After Front-Line Therapy With Gemcitabine Plus Carboplatin in Advanced Non–Small-Cell Lung Cancer. J Clin Oncol. 2009;27:591–8.

35. Arrigada R, Bergman B, Dunant A, Le Chevalier T, Pignon JP, Vansteenkiste J. Cisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non–Small-Cell Lung Cancer. N Engl J Med. 2004;350:351–60.

36. Shepherd FA, Pereira JR, Ciuleanu T, Tan EH, Hirsh V, Thongprasert S, et al. Erlotinib in Previously Treated Non-Small-Cell Lung Cancer. N Engl J Med. 2005;353:123–32.

37. Tsao M-S, Sakurada A, Cultz J-C, Zhu C-Q, Kamel-Reid S, Squire J, et al. Erlotinib in Lung Cancer — Molecular and Clinical Predictors of Outcome. N Engl J Med. 2005;353:133–44.

38. Herbst RS, O'Neill VJ, Fehrenbacher L, Belani CP, Bonomi PD, Hart L, et al. Phase II Study

of Efficacy and Safety of Bevacizumab in Combination With Chemotherapy or Erlotinib Compared With Chemotherapy Alone for Treatment of Recurrent or Refractory Non–Small-Cell Lung Cancer. J Clin Oncol. 2007;25:4743–50.

39. Hynes NE, Lane HA. ERBB Receptors and Cancer: The Complexity of Targeted Inhibitors. Nat Rev Cancer. 2005;5:341–54.

40. Sequist LV, Yang JC, Yamamoto N, O'Byrne K, Hirsch V, Mok T, et al. Phase III Study of Afatinib or Cisplatin Plus Pemetrexed in Patients with Metastatic Lung Adenocarcinoma with EGFR Mutations. J Clin Oncol. 2013;31:3327–34.

41. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. Science (80-). 2004;304:1497–500.

42. Rothenstein JM, Chooback N. ALK Inhibitors, Resistance Development, Clinical Trials. Curr Oncol. 2018;25 Supplement 1:S59–67.

43. Roman M, Baraibar I, Lopez I, Nadal E, Rolfo C, Vicent S, et al. KRAS Oncogene in Non-Small Cell Lung Cancer: Clinical Perspectives on the Treatment of an Old Target. Mol Cancer. 2018;17:33.

44. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100:57–70.

45. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646–74. doi:10.1016/j.cell.2011.02.013.

46. Marx V. Cancer Genomes: Discerning Drivers from Passengers. Nat Methods. 2014;11:375–9. doi:10.1038/nmeth.2891.

47. McFarland CD, Yaglom JA, Wojtkowiak JW, Scott JG, Morse DL, Sherman MY, et al. The Damaging Effect of Passenger Mutations on Cancer Progression. Cancer Res. 2017;77:4763–72.

48. Clancy S. Genetic Mutation. Nat Educ. 2008;1:187.

49. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Mutations: Types and Causes. In: Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J, editors. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000.

50. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15 Spec No:57–66.

51. Zhang N, Wang M, Zhang P, Huang T. Classification of Cancers Based on Copy Number

Variation Landscapes. Biochim Biophys Acta. 2016;1860 11 Pt. B:2750–5.

52. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97.

53. Gamazon ER, Stranger BE. The Impact of Human Copy Number Variation on Gene Expression. Brief Funct Genomics. 2015;14:352–7.

54. Freeman J, Perry G, Feuk L, Redon R, McCarroll S, Altshuler D, et al. Copy Number Variation: New Insights in Genome Diversity. Genome Res. 2006;16:949–61. doi:10.1101/gr.3677206.16.

55. Shlien A, Malkin D. Copy Number Variations and Cancer. Genome Med. 2009;1:62.

56. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of Large-Scale Variation in the Human Genome. Nat Genet. 2004;36:949–51.

57. Li BQ, You J, Huang T, Cai YD. Classification of Non-Small Cell Lung Cancer Based on Copy Number Alterations. PLoS One. 2014;9:e88300.

58. Laird PW, Jaenisch R. The Role of DNA Methylation in Cancer Genetics and Epigenetics. Annu Rev Genet. 1996;30:441–64.

59. Sawan C, Vaissiere T, Murr R, Herceg Z. Epigenetic Drivers and Passengers on the Road to Cancer. Mutat Res. 2008;642:1–13.

60. Taberlay PC, Jones PA. DNA Methylation and Cancer. Prog Drug Res. 2011;67:1–23.

61. Reik W, Dean W. DNA Methylation and Mammalian Epigenetics. Electrophoresis. 2001;22:2838–43.

62. Bacolla A, Pradhan S, Roberts R, Wells R. Recombinant Human DNA (Cytosine-5) Methyltransferase II. Steady-State Kinetics Reveal Allosteric Activation by Methylated DNA. J Biol Chem. 1999;274:33011–9.

63. Geiman TM, Sankpal UT, Robertson AK, Zhao Y, Zhao Y, Robertson KD. DNMT3B Interacts with hSNF2H Chromatin Remodeling Enzyme, HDACs 1 and 2, and Components of the Histone Methylation System. Biochem Biophys Res Commun. 2004;318:544–55.

64. Tsou JA, Hagen JA, Carpenter CL, Laird-Offringa A. DNA Methylation Analysis: A Powerful New Tool for Lung Cancer Diagnosis. Oncogene. 2002;21:5450–61.

65. Yang X, Yan L, Davidson NE. DNA Methylation in Breast Cancer. Endocr Relat Cancer. 2001;8:115–27.

66. Fahy J, Jeltsch A, Arimondo PB. DNA Methyltransferase Inhibitors in Cancer: A Chemical and Therapeutic Patent Overview and Selected Clinical Studies. Expert Opin Ther Patients. 2012;22:1427–42.

67. Leu YW, Rahmatpanah F, Shi H, Wei SH, Liu JC, Yan PS, et al. Double RNA Interference of DNMT3b and DNMT1 Enhances DNA Demethylation and Gene Reactivation. Cancer Res. 2003;63:6110–5.

68. Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, et al. MicroRNA Signature Predicts Survival and Relapse in Lung Cancer. Cancer Cell. 2008;13:48–57.

69. Xu W, Banerji S, Davie JR, Kassie F, Yee D, Kratzke R. Yin Yang Gene Expression Ratio Signature for Lung Cancer Prognosis. PLoS One. 2013;8.

70. Xu W, Jia G, Davie JR, Murphy L, Kratzke R, Banerji S. A 10-Gene Yin Yang Expression Ratio Signature for Stage IA and IB Non-Small Cell Lung Cancer. J Thorac Oncol. 2016;11:2150–60.

71. Wangari-Talbot J, Wall BA, Goydos JS, Chen S. Functional Effects of GRM1 Suppression in Human Melanoma Cells. Mol Cancer Res. 2012;10:1440–50.

72. National Center for Biotechnology Information. GRM1 Glutamate Metabotropic Receptor 1. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/2911.

73. Guillon G, Balestre MN, Mouillac B, Berrada R, Kirk CJ. Mechanisms of Phospholipase C Activation: A Comparison with the Adenylate Cyclase System. Biochimie. 1987;69:351–63.

74. Park JB, Lee CS, Jang JH, Ghim J, Kim YJ, You S, et al. Phospholipase Signalling Networks in Cancer. Nat Rev Cancer. 2012;12:782–92. doi:10.1038/nrc3379.

75. National Center for Biotechnology Information. RECQL4 RecQ Like Helicase 4. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/9401.

76. Sharma S, Doherty K, Brosh R. Mechanisms of RecQ Helicases in Pathways of DNA Metabolism and Maintenance of Genomic Stability. Biochem J. 2006;398 Part 3:319–37.

77. Bachrati C, Hickson ID. RecQ Helicases: Guardian Angels of the DNA Replication Fork. Chromosoma. 2008;117:219–33.

78. Sangrithi MN, Bernal JA, Madine M, Philpott A, Lee J, Dunphy WG, et al. Initiation of

DNA Replication Requires the RECQL4 Protein Mutated in Rothmund-Thomson Syndrome. Cell. 2005;121:887–98.

79. Ahmadian MR, Hoffman U, Goody RS, Wittinhofer A. Individual Rate Constants for the Interaction of Ras Proteins with GTPase-Activating Proteins Determined by Fluorescence Spectroscopy. Biochemistry. 1997;36:4535–4531.

80. Vetter IR, Wittinhofer A. The Guanine Nucleotide-Binding Switch in Three Dimensions. Science (80-). 2001;294:1299–304.

81. Karnoub AE, Weinberg RA. Ras oncogenes: split personalities. Nat Rev Mol Cell Biol. 2008;9:517–31.

82. National Center for Biotechnology Information. NRAS NRAS Proto-Oncogene, GTPase. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/4893.

83. Health NI of. NRAS Gene. U.S. National Library of Medicine. 2019. https://ghr.nlm.nih.gov/gene/NRAS.

84. National Center for Biotechnology Information. IGFBP5 Insulin-Like Growth Factor Binding Protein 5. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/3488.

85. Sureshbabu A, Okajima H, Yamanaka D, Tonner E, Shastri S, Maycock J, et al. IGFBP5 Induces Cell Adhesion, Increases Cell Survival, and Inhibits Cell Migration in MCF-7 Human Breast Cancer Cells. J Cell Sci. 2012;125 Part 7:1693–705.

86. Gullu G, Karabulut S, Akkiprik M. Functional Roles and Clinical Values of Insulin-Like Growth Factor-Binding Protein-5 in Different Types of Cancers. Chin J Cancer. 2012;31:266–80.

87. National Center for Biotechnology Information. HOXA5 Homeobox A5. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/3202.

88. Raman V, Martensen SA, Reisman D, Evron E, Odenwald WF, Jaffee E, et al. Compromised HOXA5 Function Can Limit p53 Expression in Human Breast Tumours. Nature. 2000;405:974–8. doi:10.1038/35016125.

89. National Center for Biotechnology Information. TNNC1 Troponin C1, Skeletal and Cardiac Type. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/7134.

90. Pinto JR, Siegfried JD, Parvatiyar MS, Li D, Norton N, Jones MA, et al. Functional Characterization of TNNC1 Rare Variants Identified in Dilated Cardiomyopathy. J Biol Chem.

2011;286:34404–12.

91. National Center for Biotechnology Information. SOSTDC1 Sclerostin Domain Containing 1. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/25928.

92. Lintern KB, Guidato S, Rowe A, Saldanha JW, Itasaki N. Characterization of Wise Protein and its Molecular Mechanism to Interact with both Wnt and BMP Signals. J Biol Chem. 2009;284:23159–68.

93. National Center for Biotechnology Information. CRIP2 Cysteine Rich Protein 2. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/1397.

94. van Ham M, Croes H, Schepens J, Fransen J, Wieringa B, Hendriks W. Cloning and Characterization of mCRIP2, a Mouse LIM-Only Protein that Interacts with PDZ Domain IV of PTP-BL. Genes to Cells. 2003;8:631–44.

95. National Center for Biotechnology Information. CD83 CD83 Molecule. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/9308.

96. Scholler N, Hayden-Ledbetter M, Dahlin A, Hellstrom I, Hellstrom KE, Ledbetter JA. Cutting Edge: CD83 Regulates the Development of Cellular Immunity. J Immunol. 2002;168:2599–602.

97. Prechtel AT, Steinkasserer A. CD83: An update on Functions and Prospects of the Maturation Marker of Dendritic Cells. Arch Dermatol Res. 2007;299:59–69.

98. National Center for Biotechnology Information. GATA2 GATA2 Binding Protein 2. U.S. National Library of Medicine. 2019. https://www.ncbi.nlm.nih.gov/gene/2624.

99. Rodrigues NP, Boyd AS, Fugazza C, May GE, Guo Y, Tipping AJ, et al. GATA-2 Regulates Granulocyte-Macrophage Progenitor Cell Function. Blood. 2008;112:4862–73.

100. Arya M, Shergill IS, Williamson M, Gommersall L, Arya N, Patel HR. Basic Principles of Real-Time Quantitative PCR. Expert Rev Mol Diagn. 2005;5:209–19.

101. ThermoFisher Scientific. Basic Principles of RT-aPCR. ThermoFisher Scientific. https://www.thermofisher.com/ca/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/basic-principles-rt-qpcr.html.

102. Wiltgen M, Tilz GP. DNA Microarray Analysis: Principles and Clinical Impact. Hematology. 2007;12:271–87.

103. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and Its Applications. J Pharm Bioallied Sci. 2012;4 Supplement 2:S310–2.

104. Kulkarni MM. Digital Multiplexed Gene Expression Analysis Using the NanoString nCounter System. Curr Protoc Mol Biol. 2011;25:25B.

105. NanoString Technologies Inc. Direct Digital Detection with nCounter Technology. nanoString. https://www.nanostring.com/scientific-content/technology-overview/ncounter-technology.

106. Kukurba KR, Montgomery ST. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015:951–69.

107. Wang Z, Gerstein M, Snyder M. RNA-Seq: A Revolutionary Tool for Transcriptomics. Nat Rev Genet. 2009;10:57–63.

108. Tzankov A, Went P, Zimpfer A, Dirnhofer S. Tissue Microarray Technology: Principles, Pitfalls, and Perspectives - Lessons Learned from Hematological Malignancies. Exp Gerontol. 2005;40:737–44.

109. Kononen J, Bubendorf L, Kallioniemi A, Bärlund M, Schraml P, Leighton S, et al. Tissue Microarrays for High-Throughput Molecular Profiling of Tumor Specimens. Nat Med. 1998;4:844–7.

110. Wong KS, Angell TE, Strickland KC, Alexander EK, Cibas ES, Krane JF, et al. Noninvasive Follicular Variant of Papillary Thyroid Carcinoma and the Afirma Gene-Expression Classifier. Thyroid. 2016;26:911–5.

111. Cronin M, Sangli C, Liu ML, Pho M, Dutta D, Nguyen A, et al. Analytical Validation of the Oncotype DX Genomic Diagnostic Test for Recurrence Prognosis and Therapeutic Response Prediction in Node-Negative, Estrogen Receptor-Positive Breast Cancer. Clin Chem. 2007;53:1084–91.

112. Damodaran S, Berger MF, Roychowdhury S. Clinical Tumor Sequencing: Opportunities and Challenges for Precision Cancer Medicine. Am Soc Clin Oncol Educ B. 2015;:e175-182.

113. A.K. F, J. L, M.M. C, D.W. C. Translation of Proteomic Biomarkers into FDA Approved Cancer Diagnostics: Issues and Challenges. Clin Proteomics. 2013;10:13.

114. Brooks JD. Translational Genomics: The Challenge of Developing Cancer Biomarkers. Genome Res. 2012;22:183–7.

115. Lambert M, Jambon S, Depauw S, David-Cordonnier M. Targeting Transcription Factors

for Cancer Treatment. Molecules. 2018;23:1479.

116. Cannell IG, Kong YW, Bushell M. How do microRNAs Regulate Gene Expression? Biochem Soc Trans. 2008;36 Part 6:1224–31.

117. Labrenz M, Brettar I, Christen R, Flavier S, Bötel J, Höfle MG. Development and Application of a Real-Time PCR Approach for Quantification of Uncultured Bacteria in the Central Baltic Sea. Appl Environ Biol. 2004;70:4971–9.

118. Wagatsuma A, Sadamoto H, Kitahashi T, Lukowiak K, Urano A, Ito E. Determination of the Exact Copy Numbers of Particular mRNAs in a Single Cell by Quantitative Real-Time RT-PCR. The Journal of Experimental Biology. J Exp Biol. 2005;208 Part 2:2389–98.

119. Bustin S, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clin Chem. 2009;55:611–22.

120. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and Reproducibility Issues in DNA Microarray Measurements. Trends Genet. 2006;22:101–9.

121. Illumina Inc. Illumina Sequencing Technology. Technology Spotlight: Illumina Sequencing. 2010.

122. Dan HL, Zhang YL, Zhang Y, Wang YD, Lai ZS, Yang YJ, et al. A Novel Method for Preparation of Tissue Microarray. World J Gastroenterol. 2004;10:579–82.

123. GeneDetect. In Situ Hybridization. GeneDetect.com. 2018. http://www.genedetect.com/insitu.htm.

124. Vaerman JL, Saussoy P, Ingargioloa I. Evaluation of Real-Time PCR Data. J Biol Regul Homeost Agents. 2004;18:212–4.

125. Wong ML, Medrano JF. Real-Time PCR for mRNA Quantitation. Biotechniques. 2005;39:75–85.

126. Udvardi MK. Eleven Golden Rules of Quantitative RT-PCR. Plant Cell. 2008;20:1736–7.

127. Koltai H, Weingarten-Baror C. Specificity of DNA Microarray Hybridization: Characterization, Effectors, and Approaches for Data Correction. Nucleic Acids Res. 2008;36:2395–405.

128. NanoString Technologies Inc. Strategies for Successful Gene Expression Assays. TECH NOTE nCounter Gene Expression. 2015;:1–6.

129. Arvey A, Hermann A, Hsia CC, Ie E, Freund Y, McGinnis W. Minimizing Off-Target Signals in RNA Fluorescent In situ Hybridization. Nucleic Acids Res. 2010;38:e115.

130. Danati V, Faviana P, Dell'omodarme M, Prati MC, Camacci T, De Ieso K, et al. Applications of Tissue Microarray Technology in Immunohistochemistry: A Study on c-kit Expression in Small Cell Lung Cancer. Hum Pathol. 2004;35:1347–52.

131. ThermoFisher Scientific. Ten Tips for In Situ Hybridization of Tissue Microarrays. ThermoFisher Scientific. https://www.thermofisher.com/ca/en/home/references/ambion-tech-support/microrna-studies/general-articles/ten-tips-for-in-situ-hybridization-.html.

132. BioSearch Technologies. How Much is Your qPCR Assay Really Costing You? LGC BioSearch Technologies. 2015.

133. Croy BA, Yamada AT, DeMayo FJ, Adamson SL. The Guide to Investigation of Mouse Pregnancy. Academic Press; 2013.

134. Whitehead Institute. Pricing. Whitehead Institute Genome Core. 2018. http://genomecore.wi.mit.edu/index.php/Pricing.

135. Centre PMG. NanoString nCounter™ System. University Health Network. 2015. https://www.pmgenomics.ca/pmgenomics/services/nanostring.html.

136. Illumina Inc. A Wide Breadth of Sequencing Applications. Illumina. 2019. https://www.illumina.com/systems/sequencing-platforms/miseq/applications.html.

137. Cornell University Institute of Biotechnology. Price List. Cornell University. http://www.biotech.cornell.edu/brc/genomics/services/price-list.

138. Arizona State University. Illumina Sequencing Pricing. DNASU Next Generation Sequencing at Arizona State University. http://dnasusequencing.org/nextgen/pricingnextgen.html.

139. Cooperative Human Tissue Network. Currently Available Tissue Microarrays. University of Virginia. https://chtn.sites.virginia.edu/tissue-microarrays.

140. Thomson TA, Zhou C, Ceballos K, Knight B. Tissue Microarray for Routine Clinical Breast Biomarker Analysis. The British Columbia Cancer Agency 2008 Experience. Am J Clin Pathol. 2010;133:909–14.

141. Dharmaraj S. The Basics: RT-PCR. ThermoFisher Scientific. https://www.thermofisher.com/ca/en/home/references/ambion-tech-support/rtpcr-analysis/general-articles/rt--pcr-the-basics.html.

142. laboratory-equipment.com. InnoScan 710 Series Microarray Scanners. Terra Universal. 2019. https://www.laboratory-equipment.com/microarray/innoscan-710-series-microarray-scanners-arrayit.php.

143. VWR International. GenePix 4300 and 4400 Microarray Scanner, Molecular Devices. Avantar. 2019. https://us.vwr.com/store/product/14491801/genepix-4300-and-4400-microarray-scanner-molecular-devices.

144. MarketWatch. 10-K: NanoString Technologies Inc. 2014. https://www.marketwatch.com/press-release/10-k-nanostring-technologies-inc-2014-03-27.

145. Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, et al. A Tale of Three Next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers. BMC Genomics. 2012;13:341.

146. Srinath S, Kendole R, Gopinath P, Krishnappa S, Vishwanath S. Economic Methods Used in Fabrication of Tissue Microarray: A Pilot Study. J Oral Maxillofac Pathol. 2016;20:86–90.

147. IHC World LLC. Tissue Microarray Instruments and Kits. IHC World LLC Life Science Products & Services. 2011. http://www.ihcworld.com/products/Tissue-Microarray-Instrument.htm.

148. Fan H, Blumenfeld Y, El-Sayed Y, Chueh J, Quake S. Microfluidic Digital PCR Enables Rapid Prenatal Diagnosis of Fetal Aneuploidy. Am J Obstet Gynecol. 2009;200:543e1-543e7.

149. Raymond F, Carbonneau J, Boucher N, Robitaille L, Boisvert S, Wu WK, et al. Comparison of Automated Microarray Detection with Real-Time PCR Assays for Destection of Respiratory Viruses in Specimens Obtained from Children. J Clin Microbiol. 2009;47:743–50.

150. NanoString Technologies. Product Data Sheet: nCounter Analysis System. NanoString Technologies, Inc. 2015.

151. Kampf C, Olsson I, Ryberg U, Sjostedt E, Ponten F. Production of Tissue Microarrays, Immunohistochemistry Staining and Digitalization Within the Human Protein Atlas. J Vis Exp. 2012;31:3620.

152. Canene-Adams K. Preparation of Formalin-Fixed Paraffin-Embedded Tissue for Immunohistochemistry. Methods Enzymol. 2013;533.

153. Sengüven B, Baris E, Oygur T, Berktas M. Comparison of methods for the extraction of DNA from formalin-fixed, paraffin-embedded archival tissues. Int J Med Sci. 2014;11:494–9.

154. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. Determinants of RNA quality from

FFPE samples. PLoS One. 2007;2:e1261.

155. Cronin M, Pho M, Dutta D, Stephans JC, Shak S, Kiefer MC, et al. Measurement of Gene Expression in Archival Paraffin-Embedded Tissues: Development and Performance of a 92-Gene Reverse Transcriptase-Polymerase Chain Reaction Assay. Am J Pathol. 2004;164:35–42.

156. Kong H, Zhu M, Cui F, Wang S, Gao X, Lu S, et al. Quantitative Assessment of Short Amplicons in FFPE-Derived Long-Chain RNA. Sci Rep. 2014;4.

157. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science (80-). 2007;315:848–53.

158. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18:1–8.

159. Yamamoto H, Shigematsu H, Nomura M, Lockwood WW, Sato M, Okumura N, et al. PIK3CA mutations and copy number gains in human lung cancers. Cancer Res. 2008;68:6913–21.

160. Hirsch FR, Varella-Garcia M, Bunn PA, Di Maria M V., Veve R, Bremnes RM, et al. Epidermal growth factor receptor in non-small-cell lung carcinomas: Correlation between gene copy number and protein expression and impact on prognosis. J Clin Oncol. 2003;21:3798–807.

161. Hirsch FR, Varella-Garcia M, Cappuzzo F, McCoy J, Bemis L, Xavier AC, et al. Combination of EGFR gene copy number and protein expression predicts outcome for advanced non-small-cell lung cancer patients treated with gefitinib. Ann Oncol. 2007;18:752–60.

162. Li J, Jin J, Liao M, Dang W, Chen X, Wu Y, et al. Upregulation of RECQL4 expression predicts poor prognosis in hepatocellular carcinoma. Oncol Lett. 2018;15:4248–54.

163. Chen H, Yuan K, Wang X, Wang H, Wu Q, Wu X, et al. Overexpression of RECQL4 is associated with poor prognosis in patients with gastric cancer. Oncol Lett. 2018;16:5419–25.

164. Su Y, Meador JA, Calaf GM, De-Santis LP, Zhao Y, Bohr VA, et al. Human RecQL4 helicase plays critical roles in prostate carcinogenesis. Cancer Res. 2010;70:9207–17.

165. Fang H, Nie L, Chi Z, Liu J, Guo D, Lu X, et al. RecQL4 Helicase Amplification is Involved in Human Breast Tumorigenesis. PLoS One. 2013;8:e69600.

166. Arora A, Agarwal D, Abdel-Fatah T, Lu H, Croteau DL, Moseley P, et al. RECQL4 helicase has oncogenic potential in sporadic breast cancers. J Pathol. 2016;238:495–501.

167. Lao VV, Welcsh P, Luo Y, Carter KT, Dzieciatkowski S, C D, et al. Altered RECQ Helicase Expression in Sporadic Primary Colorectal Cancers. Transl Oncol. 2013;6:458–69.

168. Stephens R, Yi M, Kessing B, Nissley D, McCormick F. Tumor RAS Gene Expression Levels are Influenced by the Mutational Status of RAS Genes and Both Upstream and Downstream RAS Pathway Genes. Cancer Inform. 2017;16.

169. Johnson DB, Lovly CM, Flavin M, Panageas KS, Ayers GD, Zhao Z, et al. Impact of NRAS Mutations for Patients with Advanced Melanoma Treated with Immune Therapies. Cancer Immunol Res. 2015;3:288–95.

170. Vaughn CP, Zobell SD, Furtado LV, Baker CL, Samowitz WS. Frequency of KRAS, BRAF, and NRAS Mutations in Colorectal Cancer. Genes Chromosomes Cancer. 2011;50:307–12.

171. Li X, Cao X, Li X, Zhang W, Feng Y. Expression level of insulin-like growth factor binding protein 5 mRNA is a prognostic factor for breast cancer. Cancer Sci. 2007;98:1592–6.

172. Mita K, Zhang Z, Ando Y, Toyama T, Hamguchi M, Kobayashi S, et al. Prognostic Significance of Insulin-Like Growth Factor Binding Protein (IGFBP)-4 and IGFBP-5 Expression in Breast Cancer. Jpn J Clin Oncol. 2007;37:575–82.

173. Nishino R, Honda M, Yamashita T, Takatori H, Minato H, Zen Y, et al. Identification of Novel Candidate Tumour Marker Genes for Inhrahepatic Cholangiocarcinoma. J Hepatol. 2008;49:207–16.

174. Johnson SK, Dennis RA, Barone GW, Lamps LW, Haun RS. Differential Expression of Insulin-Like Growth Factor Binding Protein-5 in Pancreatic Adenocarcinomas: Identification Using DNA Microarray. Mol Carcinog. 2006;45:814–27.

175. Ordóñez-Morán P, Dafflon C, Imajo M, Nishida E, Huelsken J. HOXA5 Counteracts Stem Cell Traits by Inhibiting Wnt Signaling in Colorectal Cancer. Cancer Cell. 2015;28:815–29.

176. Wang CC, Su KY, Chen HY, Chang SY, Shen CF, Hsieh CH, et al. HOXA5 Inhibits Metastasis via Regulating Cytoskeletal Remodelling and Associates with Prolonged Survival in Non-Small-Cell Lung Carcinoma. PLoS One. 2015;10:e0124191. doi:10.1371/journal.pone.0124191.

177. Chang C-J, Chen Y-L, Hsieh C-H, Liu Y-J, Yu S-L, Chen JJW, et al. HOXA5 and p53 Cooperate to Suppress Lung Cancer Cell Invasion and Serve as Good Prognostic Factors in Non-Small Cell Lung Cancer. J Cancer. 2017;8:1071–81.

178. Yang X, Wu K, Li S, Hu L, Han J, Zhu D, et al. MFAP5 and TNNC1: Potential Markers for

Predicting Occult Cervical Lymphatic Metastasis and Prognosis in Early Stage Tongue Cancer. Oncotarget. 2017;8:2525–35.

179. Urgard E, Vooder T, Võsa U, Välk K, Liu M, Luo C, et al. Metagenes Associated with Survival in Non-Small Cell Lung Cancer. Cancer Inform. 2011;10:175–83. doi:10.4137/CIN.S7135.

180. Huang C-Y, Huang C-H, Chang PM-H, Wu M-Y, Ng K-L. In silico Identification of Potential Targets and Drugs for Non-Small Cell Lung Cancer. IET Syst Biol. 2014;8:56–66.

181. Han S-S, Kim WJ, Hong Y, Hong S-H, Lee S-J, Ryu DR, et al. RNA Sequencing Identifies Novel Markers of Non-Small Cell Lung Cancer. Lung Cancer. 2014;84:229–35. doi:10.1016/j.lungcan.2014.03.018.

182. Liu L, Wu S, Yang Y, Cai J, Zhu X, Wu J, et al. SOSTDC1 is Down-Regulated in Non-Small Cell Lung Cancer and Contributes to Cancer Cell Proliferation. Cell Biosci. 2016;6:24.

183. Chen G, Gong H, Wang T, Wang J, Han Z, Bai G, et al. SOSTDC1 Inhibits Bone Metastasis in Non-Small Cell Lung Cancer and May Serve as a Clinical Therapeutic Target. Int J Mol Med. 2018;42:3424–36.

184. Lo PHY, Ko JMY, Yu ZY, Law S, Wang LD, Li JL, et al. The LIM Domain Protein, CRIP2, Promotes Apoptosis in Esophageal Squamous Cell Carcinoma. Cancer Lett. 2012;316:39–45. doi:10.1016/j.canlet.2011.10.020.

185. Cheung AK, Ko JMY, Lung HL, Chan KW, Stanbridge EJ, Zabarovsky E, et al. Cysteine-Rich Intestinal Protein 2 (CRIP2) Acts as a Repressor of NF-B-Mediated Proangiogenic Cytokine Transcription to Suppress Tumorigenesis and Angiogenesis. Proc Natl Acad Sci. 2011;108:8390–5.

186. Liang Y. An Expression Meta-Analysis of Predicted microRNA Targets Identifies a Diagnostic Signature for Lung Cancer. BMC Med Genomics. 2008;1:61.

187. Kozaki K, Koshikawa K, Miyaishi O, Saito H, Hida T, Takahashi T, et al. Multi-Faceted Analyses of a Highly Metastatic Human Lung Cancer Cell Line NCI-H460-LNM35 Suggest Mimicry of Inflammatory Cells in Metastasis. Oncogene. 2001;20:4228–34.

188. Zhang Z, Borecki I, Nguyen L, Ma D, Smith K, Huettner PC, et al. CD83 Gene Polymorphisms Increase Susceptibility to Human Invasive Cervical Cancer. Cancer Res. 2007;67:11202–8.

189. Iwamoto M, Shinohara H, Miyamoto A, Okuzawa M, Mabuchi H, Nohara T, et al. Prognostic Value of Tumor-Infiltrating Dendritic Cells Expressing CD83 in Human Breast

Carcinomas. Int J Cancer. 2003;104:92–7.

190. Ananiev J, Gulubova MV, Manolova I. Prognostic Significance of CD83 Positive Tumor-Infiltrating Dendritic Cells and Expression of TGF-Beta 1 in Human Gastric Cancer. Hepatogastroenterology. 2011;58:1834–40.

191. Baleeiro RB, Bergami-Santos PC, Tomiyoshi MY, Gross JL, Haddad F, Pinto CA, et al. Expression of a Dendritic Cell Maturation Marker CD83 on Tumor Cells from Lung Cancer Patients and Several Human Tumor Cell Lines: Is there a biological meaning behind it? Cancer Immunol Immunother. 2008;57:265–70.

192. Li YW, Wang JX, Yin X, Qiu SJ, Wu H, Liao R, et al. Decreased Expression of GATA2 Promoted Proliferation, Migration and Invasion of HepG2 in vitro and Correlated with Poor Prognosis of Hepatocellular Carcinoma. PLoS One. 2014;9:e87505.

193. Peters I, Dubrowinskaja N, Tezval H, Kramer MW, von Klot CA, Hennenlotter J, et al. Decreased mRNA Expression of GATA1 and GATA2 is Associated with Tumor Aggressiveness and Poor Outcome in Clear Cell Renal Cell Carcinoma. Target Oncol. 2015;10:267–75.

194. Chiang YT, Wang K, Fazli L, Qi RZ, Gleave ME, Collins CC, et al. GATA2 as a Potential Metastasis-Driving Gene in Prostate Cancer. Oncotarget. 2014;5:451–61.

195. Vidal SJ, Rodriguez-Bravo V, Quinn SA, Rodriguez-Barrueco R, Lujambio A, Williams E, et al. A Targetable GATA2-IGF2 Axis Confers Aggressiveness in Lethal Prostate Cancer. Cancer Cell. 2015;27:223–39. doi:10.1016/j.ccell.2014.11.013.

196. Wang Y, He X, Ngeow J, Eng C. GATA2 Negatively Regulates PTEN by Preventing Nuclear Translocation of Androgen Receptor and by Androgen-Independent Suppression of PTEN Transcription in Breast Cancer. Hum Mol Genet. 2012;21:569–76.

197. Tessema M, Yingling CM, Snider AM, Do K, Juri DE, Picchi MA, et al. GATA2 is Epigenetically Repressed in Human and Mouse Lung Tumors and is not Requisite for Survival of KRAS Mutant Lung Cancer. J Thorac Oncol. 2014;9:784–93. doi:10.1097/JTO.0000000000000165.

198. Rawat A, Gopisetty G, Thangarajan R. E4BP4 is a Repressor of Epigenetically Regulated SOSTDC1 Expression in Breast Cancer Cells. Cell Oncol (Dordrecht_. 2014;37:409–19.

199. Gopal G, Raja UM, Shirley S, Rajalekshmi KR, Rajkumar T. SOSTDC1 Down-Regulation of Expression Involves CpG Methylation and is a Potential Prognostic Marker in Gastric Cancer. Cancer Genet. 2013;206:174–82. doi:10.1016/j.cancergen.2013.04.005.

200. King HC, Sinha AA. Gene Expression Profile Analysis by DNA Microarrays: Promise and

Pitfalls. JAMA. 2001;286:2280–8.

201. Illumina. MiSeq System Applications. Illumina. 2014.

202. Illumina. Benefits of NGS Targeted Resequencing. Illumina. 2017.

203. Scott MG. When Do New Biomarkers Make Economic Sense? Scand J Clin Lab Invest. 2010;242:90–5.

204. van Staveren WC, Solís DY, Hébrant A, Detours V, Dumont JE, Maenhaut C. Human Cancer Cell Lines: Experimental Models for Cancer Cells in situ? For Cancer Stem Cells? Biochim Biophys Acta. 2009;1795:92–103. doi:10.1016/j.bbcan.2008.12.004.

205. Ertel A, Verghese A, Byers SW, Ochs M, Tozeren A. Pathway-Specific Differences Between Tumor Cell Lines and Normal and Tumor Tissue Cells. Mol Cancer. 2006;5:55.

206. Wistuba II, Behrens C, Milchgrub S, Syed S, Ahmadian M, Virmani AK, et al. Comparison of Features of Human Breast Cancer Cell Lines and their Corresponding Tumors. Clin Cancer Res. 1998;4:2931–8.

207. Wistuba II, Bryant D, Behrens C, Milchgrub S, Virmani AK, Ashfaq R, et al. Comparison of Features of Human Lung Cancer Cell Lines and their Corresponding Tumors. Clin Cancer Res. 1999;5:991–1000.

208. Kaur G, Dufour JM. Cell Lines: Valuable Tools of Useless Artifacts. Spermatogenesis. 2012;2:1–5. http://www.landesbioscience.com/journals/nucleus/article/19513/.

209. Gazdar AF, Gao B, Minna JD. Lung Cancer Cell Lines: Useless Artifacts of Invaluable Tools for Medical Science? Lung Cancer. 2010;68:309–18.