ALTERATION OF NON-CODING RNAS AS A MECHANISM OF LUNG CANCER GENE DEREGULATION

by

GREGORY STEWART

B.Sc, The University of Victoria, 2010

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

DOCTOR OF PHILOSOPHY

In

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Interdisciplinary Oncology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Alteration of non-coding RNAs as a mechanism of lung cancer gene deregulation

submitted by Greg Stewart in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Interdisciplinary Oncology

Examining Committee:

Dr. Wan Lam (Interdisciplinary Oncology)
Supervisor

Dr. Calum MacAulay (Pathology and Laboratory Medicine)
Supervisory Committee Member

Dr. Isabella Tai (Experimental Medicine)
University Examiner

Dr. Decheng Yang (Pathology and Laboratory Medicine)
University Examiner

Additional Supervisory Committee Members:

Dr. Stephen Lam (Medicine)
Supervisory Committee Member

Abstract

Lung cancer remains the deadliest form of cancer, and less than half of lung adenocarcinoma (LUAD) patients harbour clinically actionable driver genes, emphasizing the need to explore alternative mechanisms of cancer gene deregulation. The advent of next generation sequencing has begun to reveal the functional importance of long non-coding RNAs (lncRNAs) in human cell biology, which can be exploited by tumours to drive the hallmarks of cancer. Due to their complex tertiary structure and unknown binding motifs there is a growing disparity between number of lncRNAs identified and those that have been functionally characterized. As such, lncRNAs deregulated in cancer may represent critical members of cancer pathways that could hold therapeutic applicability.

The goal of this thesis is to identify lncRNAs important to LUAD biology, discover shared features and mechanisms used to regulate cancer driving protein coding genes, and evaluate the clinical relevance of these non-coding genes. We discover and investigate three major mechanisms harnesses by lncRNAs in LUAD: (i) cis-acting regulation of neighbouring genes, (ii) trans-acting regulation through sequence homology and (iii) regulation through shared miRNAs.

This work uncovers evidence to suggest that alteration of lncRNAs is a major mechanism of cancer gene regulation in LUAD. Further characterization of these understudied gene regulatory mechanisms could lead to novel therapies that silence oncogenes or reactivate tumour suppressor genes.

Lay Summary

Lung cancer is a deadly disease with a 5-year survival rate of less than 15%. The biological mechanism of approximately 50% of lung cancer remains unknown and there is a need to identify new clinical targets. With recent advances in sequencing technology, we can now investigate previously unexplored areas of the tumour DNA, specifically the many overlooked genes that do not encode proteins called non-coding RNA genes. While several non-coding RNAs (ncRNAs) have been recently implicated in lung cancer development, the function of the vast majority of these ncRNAs remains unexplored. The research described here describes a methodology for identification and characterization of three different classes of ncRNA. Specifically, we find that these three classes of lncRNA are altered in lung cancer, where they regulate known cancer associated genes, and affect patient outcome.

Preface

The UBC Ethics Board approval was obtained for the research presented in this thesis work. The ethics certificate numbers are as follows: EDRN H09-00008, CCSRI H09-00934, and W81XW-10-1-0634.

The work in Chapter 3 has represents a manuscript in preparation, and has been presented at local, national, and international conferences during its preparation. The current author list for this manuscript in preparation is as follows: **Greg L. Stewart**, Adam P. Sage, Katey S.S. Enfield, Erin A. Marshall & Wan L. Lam (2019).

As first author I was responsible for study design, and implementation, including primary data analysis, as well as *in vitro* analysis. Adam P. Sage assisted in data analysis, cell culture, construction of Figure 1.3 and editing. Katey S.S. Enfield and Erin A. Marshall provided invaluable advice for both *in silico* and *in vitro* analysis. This project was overseen and guided by Wan L. Lam.

During its preparation Chapter 4 was presented at local, national, and international conferences and a version of Chapter 4 has been published in *Frontiers in Genetics*: **Stewart GL**, Enfield KSS, Sage AP, Martinez VD, Minatel BC, Pewarchuk ME, Marshall EA, Lam WL (2019). Aberrant expression of pseudogene-derived lncRNAs as an alternative mechanism of cancer gene regulation in lung adenocarcinoma. *Front Genet*. 6;10:138

As first author I was responsible for primary study design and implementation as well as the majority of all data analysis and writing. Katey S.S. Enfield provided vital feedback during design of the study, as well as advice for data analysis, and edited the manuscript. Adam P. Sage assisted in data analysis and figure design of Figure 2, in addition to editing. Victor D. Martinez assisted in processing of RAW RNA seq data files. Brenda C. Minatel assisted in analysis and

generation of Figure 3. Michelle E Pewarchuk assisted with design of Figure 1. Erin A. Marshall provided vital feedback and editing. This project was guided and overseen by Wan L. Lam.

During its preparation Chapter 5 was presented at local, national, and international conferences and a version of Chapter 5 has been published in *PLOS One*:

Erin A. Marshall*****, **Greg L. Stewart***, Adam P. Sage, Wan L. Lam, Carolyn J. Brown (2019). Beyond sequence homology: Cellular biology limits the potential of XIST to act as a miRNA sponge. *PLoS One*. 14(8):e0221371.

***Co-First Authorship**

As co-first author I was responsible for primary study design and implementation as well as the majority of all data analysis and writing. Erin A Marshall and I combined our respective expertise to design and implement this project, and both were heavily involved in data analysis, *in vitro* work, and writing. We were supported by Adam P. Sage who assisted in data analysis, figure design and manuscript editing. This project was overseen and guided by both Wan L Lam and Carolyn J. Brown.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3’ UTR	3 prime untranslated region
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
ATCC	American Type Culture Collection
ATS	American Thoracic Society
B-H	Benjimini-Hochberg
BCCA	British Columbia Cancer Agency (Renamed as BC Cancer in 2019)
BP	Base pair (nucleotides)
BRAF	B-Raf Proto-Oncogene
CN	Copy number
DMX genes	Defended from miRNA by XIST genes
EGFR	Epidermal Growth Factor Receptor
ERS	European Respiratory Society
FC	Fold change
GISTIC	Genomic Identification of Significant Targets in Cancer
HER2	Tyrosine Kinase-Type Cell Surface Receptor HER2
HMGA1	High Mobility Group AT-Hook 1
IASLC	International Association for the Study of Lung Cancer
LNAs	Locked nucleic acids
lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
MAP2K1	Mitogen-Activated Protein Kinase Kinase 1

MET	MET Proto-Oncogene, Receptor Tyrosine Kinase
miRNA	MicroRNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
NRAS	Neuroblastoma RAS Viral Oncogene Homolog
NSCLC	Non small-cell lung cancer
NTRK1	Neurotrophic Receptor Tyrosine Kinase 1
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
piRNA	PIWI-interacting RNA
RET	Ret Proto-Oncogene
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
ROS1	ROS Proto-Oncogene 1, Receptor Tyrosine Kinase
RPKM	Reads per kilobase of exon model per million mapped reads
RPM	Reads per million mapped reads
RT-qPCR	Reverse Transcription Quantitative Polymerase Chain Reaction
TCGA	The Cancer Genome Atlas
UTR	Untranslated region
XIST	X Inactive Specific Transcript
Ψ-lncRNA	Pseudogene overlapping long non-coding RNA

Acknowledgements

The many current and past members of the Lam lab were immensely helpful over the course of my program, providing multi-disciplinary expertise, advice, and friendship. This unique cast of characters were a treat to work with, and I thank them for their support over the years.
Additionally, I am thankful and appreciative of the funding and scholarship support provided to me, and would like to acknowledge this support. The UBC Graduate entrance scholarship was a helpful start to my degree and the Interdisciplinary Oncology Program provided scholarship support multiple times over the course of my degree. I would also like to express my gratitude for The Canadian Institutes of Health Research Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award and the UBC Four Year Fellowship Award.
I would like to thank my supervisory committee, including Dr. Stephen Lam and Dr. Calum MacAulay for their expertise, helpful advice, and insight over the course of my degree.
Additionally, this work would not have been possible if not for the immense support, encouragement, and guidance provided by my supervisor Dr. Wan Lam.

Dedication

I would like to dedicate this thesis to my family, particularly my father Gary Stewart, who passed away from lung cancer before the start of my degree. His encouragement and genuine interest in the sciences instilled a curiosity in me that led me to become the person I am today. Dad, I hope this work would make you proud. Additionally, this wouldn't have been possible without my mother and sister, who are major inspirational figures in my life, and remind me to work hard, eat well, and enjoy life.

…Also my friends. They are pretty great too.

Chapter 1: Introduction

1.1 Biology of the human lung

All cells in the human body require oxygen to function and therefore rely on the lungs for oxygen absorption into the bloodstream, and the removal of waste gasses, such as carbon dioxide. The lungs begin in the upper airways with the trachea (composed of columnar, squamous, goblet, basal and neuroendocrine cells) before dividing into the two main bronchi. Gas exchange takes place in the lower airways and lung periphery, where type I and II pneumocytes and club (Clara) cells make up the alveoli. The lungs are supported by an extensive vascular network to allow for large-scale transportation of oxygen rich cells to the heart, as well as the flow of unoxygenated blood back to the lungs. To accomplish this cellular gas exchange, the lungs have a massive surface area that is organized in a tree-like system of airways.

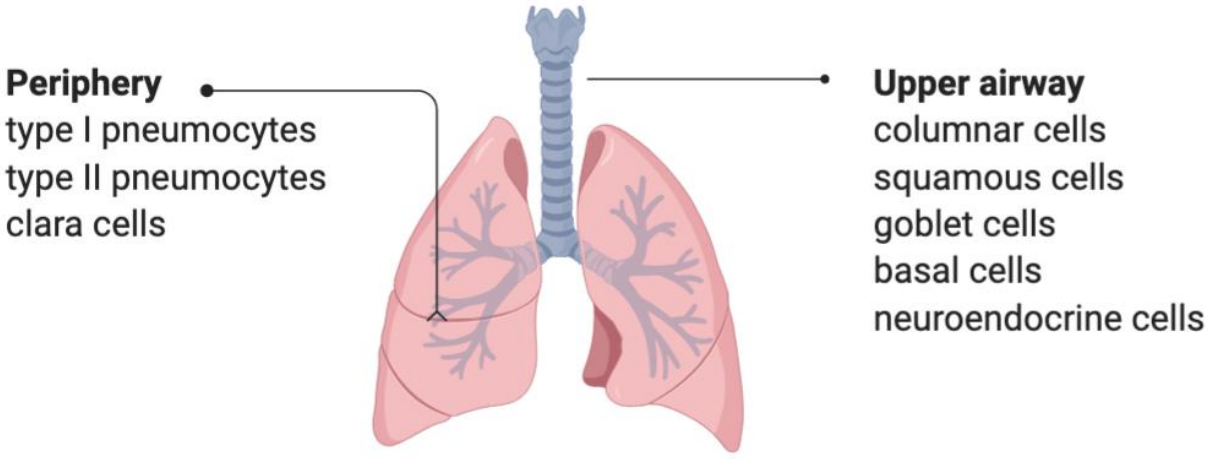


Figure 1.1 Cellular anatomy of the human lungs

The different anatomical regions of the lungs are composed of unique sets of cells.

1.2 Lung cancer

Lung cancer is the most commonly diagnosed cancer for both men and women in Canada, and alarmingly, 78 Canadians are diagnosed with lung cancer every day. In 2017 alone, 28600 Canadians were diagnosed with the disease. In addition to being widespread, lung cancer is one of the deadliest forms of cancer, with 21100 Canadians succumbing to the disease in 2017 alone. Furthermore, lung cancer is also a very aggressive disease, as those diagnosed have an average 5-year survival rate of only 19% (http://www.cancer.ca/en/cancer-information/cancer-type/lung/statistics/?region=pe#ixzz5ijvFJcpl).

Cancer is typically a disease of the elderly and lung cancer is no exception. Lung cancer cases under the age of 45 are rare and the typical age of diagnosis is over 60. Smoking is the major lung cancer risk factor and is preventable, but even if excluded, never smoker lung cancer is the 7th largest cause of cancer death, and therefore remains a huge health concern [1] [2]. Additionally, never smoker lung cancer is on the rise and former smokers, who make up a large portion of patients (35%), remain at increased risk of disease. The lungs lack pain receptors so many patients are asymptomatic until the tumour size begins to impact lung function and may present with shortness of breath or presence of blood when coughing. As such, the majority of lung cancer patients are diagnosed with late stage disease (Figure 1.1), where tumours are invasive or have metastasized to other areas of the body. Lung cancer typically metastasizes to the bones, liver, and brain, which are all challenging regions to perform surgery [3]. Tumours that have metastasized are notoriously challenging to identify, resect, and effectively treat. Additionally, metastatic tumours may have developed new molecular drivers thus not respond to the same therapy as the primary tumour. Metastasis and complications arising from it are the most common cause of cancer associated death.

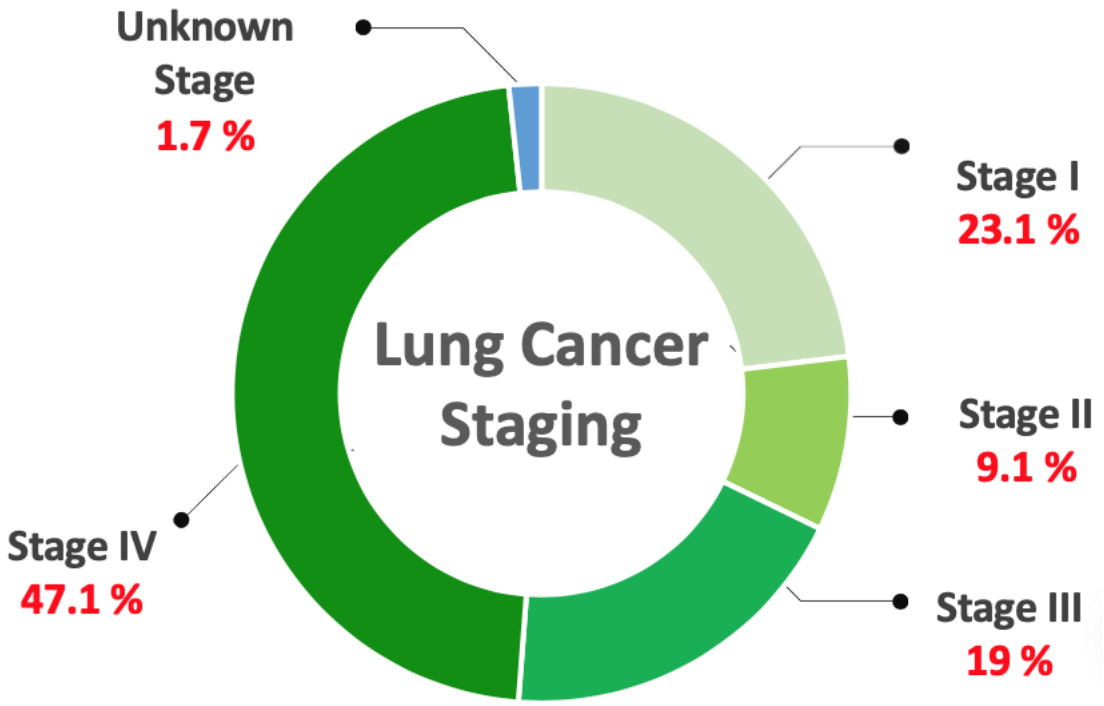


Figure 1.2 Tumour stage in lung cancer patients at diagnosis.

Lung Cancer is typically discovered at more aggressive disease stages at the time of diagnosis. Stage III and IV disease are the most commonly diagnosed, when the disease has become invasive, and the patient presents with symptoms. Adapted from Canadian cancer society (http://www.cancer.ca/en/cancer-information/cancer-type/lung/statistics/?region=pe#ixzz5ijvFJcpl).

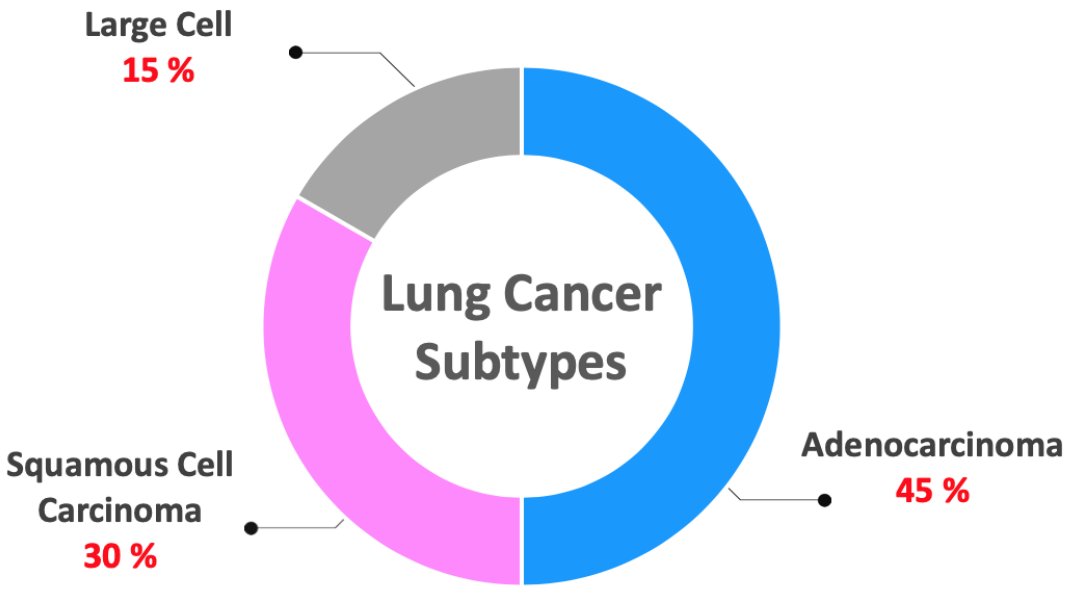


Figure 1.2 Relative proportion of non-small cell lung cancer subtypes at diagnosis

Non-small cell lung cancer is composed of adenocarcinoma (LUAD), squamous cell carcinoma (SqCC), and large cell carcinoma. Small cell lung cancer (not shown) makes up roughly 15% of all diagnosed lung cancers. Adapted from Canadian cancer society (http://www.cancer.ca/en/cancer-information/cancer-type/lung/statistics/?region=pe#ixzz5ijvFJcpl).

There are two main types of lung cancer; small cell (15%), and non-small cell lung cancer (85%). Small cell lung cancer is almost entirely composed of patients with smoking history and is an incredibly aggressive disease, with a 5 year survival of under 10% [4,5]. The most common type of lung cancer, Non-small cell lung cancer (NSCLC) is comprised of 3 subtypes; large cell carcinoma, squamous cell carcinoma (SCC), and lung adenocarcinoma (LUAD). SCC is located in the upper airways and may be composed of neuroendocrine cells [6] (Figure 1.2). LUAD is the most common type of lung cancer and is the sub-type focused on in this thesis work. LUAD is found in the lung periphery and is thought to arise from club cells (a.k.a Clara cells) or type II pneumocytes. In 2014 a new international multidisciplinary classification system (IASLC/ATS/ERS) revised the classifications for LUAD to include 11 sub-subtypes [7] (Table 1.1). The many histological subtypes of LUAD reflect the diverse genetic background of these tumours.

Table 1.1 Histological classification of lung adenocarcinoma patient specimens

Adenocarcinoma *in situ* (≤3 cm formerly BAC)
Nonmucinous
Mucinous
Mixed mucinous/nonmucinous
Minimally invasive adenocarcinoma (≤3 cm lepidic predominant tumor with ≤5 mm invasion)
Nonmucinous
Mucinous
Mixed mucinous/nonmucinous
Invasive adenocarcinoma
Lepidic predominant (formerly nonmucinous BAC pattern, with >5 mm invasion)
Acinar predominant
Papillary predominant
Micropapillary predominant
Solid predominant with mucin production
Variants of invasive adenocarcinoma
Invasive mucinous adenocarcinoma (formerly mucinous BAC)
Colloid
Fetal (low and high grade)

1.3 Genetics of lung cancer

Cancer cell genomes can be drastically altered from those of normal cells, and lung cancer is no exception. Lung cancer tumours have widespread genetic alterations and while late stage diagnosis is one reason for the high mortality rate of lung cancer, another is due to the heterogenous molecular background of the disease, which makes these tumours difficult to treat. Genetic and epigenetic alterations are selected for in order to favour the activation of pro-growth genes, and the de-activation of regulatory genes. LUAD tumours can deploy many mechanisms to achieve this, such as mutations, DNA methylation, changes in DNA copy number, fusion genes, and deregulated gene expression.

Lung cancer has one of the highest mutational burdens of any cancer, which can make it difficult to identify functional driver mutations. Nevertheless, many genetic drivers of LUAD have been identified, including activating mutations to oncogenes, such as EGFR, as well as de-activating mutations to tumour suppressor genes, such as *TP53*. Fusion genes are created from translocations of the genome that cause the merging of previously separated protein domains, and this process can result in new oncogenes. *ALK*, *ROS1*, and *RET* are all examples of oncogenic fusion genes present in NSCLC.

DNA copy number alterations involve amplifications or deletions of regions of DNA. These alterations can range from small focal events to large scale events, effecting entire chromosome arms. Oncogenes frequently have gains (>2 copies) or amplifications (>4 copies) in copy number resulting in increased gene expression. In lung cancer, frequent recurrent gene amplifications are seen for oncogenes such as *EGFR*, *MET*, and *ERBB2*. Growth suppressive genes are also regularly lost (less than 2 copies) or deleted entirely. Many key growth control

genes must be fully deleted for a cell to become cancerous, as loss of one gene copy is not sufficient for tumourigenesis.

In addition to genetic alterations, epigenetic alterations are also common. Epigenetic alterations are heritable features of cancer cells which can deregulate genes without changing the DNA sequence. DNA methylation is a common epigenetic alteration in many tumour types that functions through silencing genomic loci (hypermethylation) and can be reversed to release repression of a region (hypomethylation). In cancer, we typically see widespread hypomethylation of DNA, allowing for expression of many genes and regions that are usually silenced. Additionally, promoter specific hypermethylation of genes that may act as tumour suppressors is a common feature in many cancers. In NSCLC, frequent promoter hypermethylation is observed for the tumour suppressor genes *CDKN2A* (*p16*), *MGMT*, and *RB1* [8] [9].

Deregulated expression of genes can occur through a vast number of mechanisms and is not limited to DNA copy number alterations or changes in the methylome. For example, changes in transcription factor expression can lead to the deregulation of hundreds of genes. In general, tumours favor increased oncogene expression, such as *EGFR*, and *MET* and decreased expression of tumour suppressive genes, such as *LKB1* in NSCLC [10]. Some genes that are affected by multiple different genetic forms of alteration may not appear to be frequently altered if only one type of analysis is performed. By combining different dimensions of genetic data, we are able to identify genes that are recurrently altered by a variety of mechanisms. For example, in previously published work we performed a multiplatform analysis on the gene *FBXW4*. We analyzed *FBXW4* DNA copy number, mutation, and expression in a number of cancers including lung cancer. While in any single dimension of data alteration of *FBXW4* does not appear to be a

frequent event in NSCLC, using a multi-platform approach, we find that *FBXW4* is mutated, lost and under-expressed in a variety of human cancer cell lines and clinical patient samples, and is associated with lung cancer patient survival (Figure 1.3) [11]. This study highlights the utility of combining multidimensional data to identify a novel tumor suppressor gene that would not have met discovery thresholds by using one analysis method alone. Other studies have taken this approach to identify cancer driving genes in a number of other tumour types, and recently, many of these have been the subject of interest for the development of targeted therapies [12] [13].

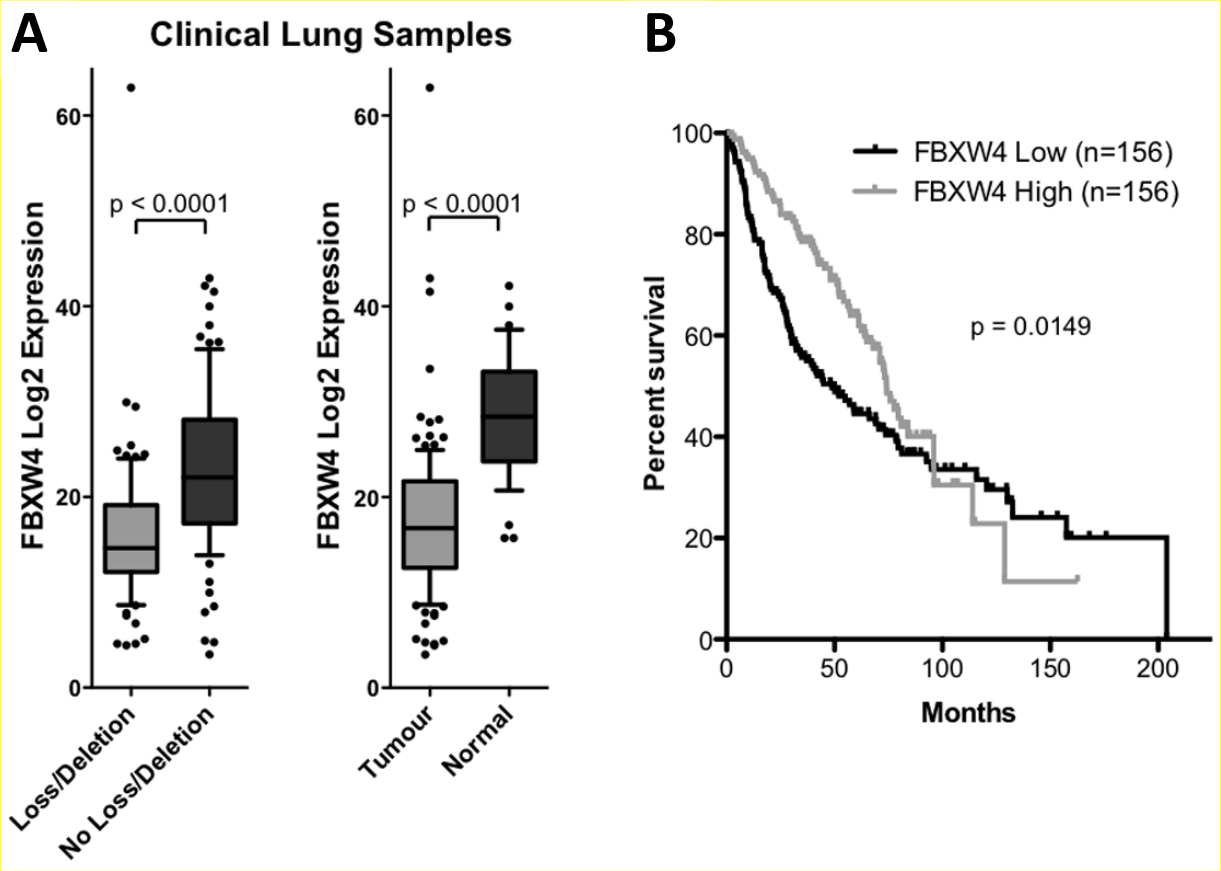


Figure 1.3 FBXW4 is deregulated in multiple genetic levels and is associated with patient survival.

FBXW4 expression is significantly lower in tumours with FBXW4 copy number loss or deletion (A). Loss of FBXW4 expression is associated with poor survival (B). Adapted from our manuscript published in PLoS ONE "The novel ubiquitin ligase complex, SCF(Fbxw4), interacts with the COP9 signalosome in an F-box dependent manner, is mutated, lost and under-expressed in human cancers"[11].

1.4 Molecular background and treatment

After sequencing the lung cancer genome, it was discovered that while many tumours may appear to be of similar histology, they can have vastly different molecular subtypes. While the high burden of genomic alterations in lung cancer complicates the separation of molecular drivers of the disease from passenger alterations, several recurrent cancer driving alterations have been identified. Traditional treatment has relied on non-specific targeting of fast growing cells, such as platinum therapy or demethylating agents. The identification of molecular drivers has lead to the development of targeted therapies that have greatly benefited subsets of patients who harbor these specific alterations. LUAD is defined by alterations that activate oncogenes such as *EGFR*, *KRAS*, *ERBB4*, *KDR*, *FGFR4*, and *NTRK* (Figure 1.4) [14] [15]. Additionally, growth suppressive genes such as *RB1*, *TP53*, *KEAP1*, *NF1*, *STK11*, and *CDKN2A*, are commonly de-activated or abrogated.

Interestingly, many of these molecular drivers are associated with patient characteristics such as race and gender, and can be defined by smoking status. For example, *EGFR* is frequently mutated in Asian females with no smoking history, while patients who smoke are less likely to have tumours harbouring *EGFR* mutations and more likely to harbor mutations in *KRAS* [15]. Additionally, smoking is associated with a significant increase in genome wide mutational burden in patients with LUAD [16] [17].

These molecular driver genes, and by association the up and downstream components of their signaling pathways, have become the focus for targeted drug development. For example, several successful targeted therapies include antibodies and tyrosine kinase inhibitors that were developed against mutant EGFR. In addition, fusion genes are attractive therapeutic targets due

to their cancer-specific expression, such as therapies developed to target the EML4-ALK fusion protein.

While therapies targeting oncogenic protein coding genes have been successful for many patients, this type of approach may not work for all tumour types. Some driver mutations, such as *KRAS*, remain un-druggable despite numerous attempts. Additionally, only a fraction of patients harbor actionable druggable targets; targetable *EGFR* mutations are present in 15-20% of patients and EML4-ALK fusions are only present in 5-7%. Furthermore, there are currently no effective clinical strategies to turn back on silenced tumour suppressor genes. This highlights the need to explore alternative options, both to identify molecular drivers of the large subset of tumours with no known drivers, as well as to identify unknown members of known, un-targeted pathways that may serve as novel clinical targets.

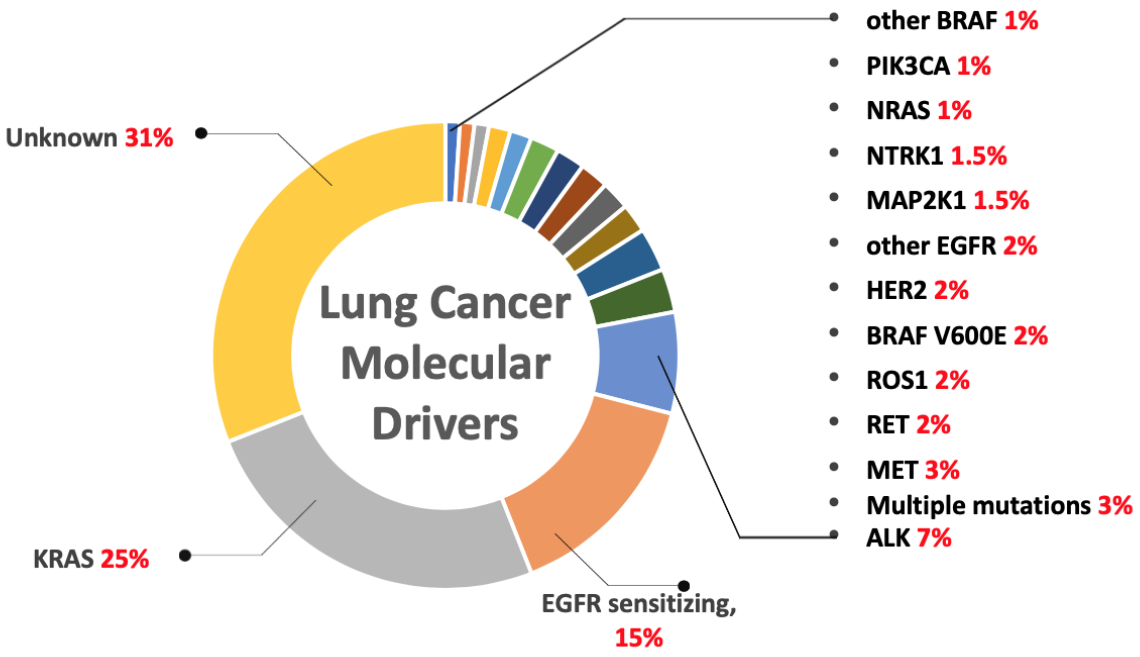


Figure 1.4 Molecular subtypes lung adenocarcinoma defined by mutations and fusions of known driver genes.

While lung adenocarcinoma is made up of many histological subtypes, the molecular alterations driving these tumours are equally as diverse. Several of these molecular driver genes are now targettable and used in the clinic, however the vast majority of tumours still remain un-druggable. Adapted from: 2018 Molecular Profiling of Lung Cancer, My Cancer Genome Website: https://www.mycancergenome.org/content/disease/lung-cancer. [18].

1.5 Non-coding RNAs: a new frontier in gene regulation

The central dogma of genetics defines the flow of genetic information in the cell, beginning with DNA transcribed into messenger RNA, which in turn is translated into a functional protein product. Because of this core tenant of genetics, when RNA transcripts that did not code for protein were observed, they were generally regarded as non-functional "genomic junk." Although recurring alterations were observed in regions that had no protein coding genes, which lead to the discovery of a handful of functional non-coding genes, these were thought to be the exception to this rule [19] [20] [21]. The advent of next generation sequencing has led to the discovery of large-scale non-coding RNA transcription. It has become apparent that while up to 90% of the human genome is transcribed, only about 1.2% of the genome codes for proteins [22]. Since the discovery of their prevalent transcription, these non-coding RNAs (ncRNAs) have been observed to play critical roles in all aspects of cell biology.

Defined simply as transcribed RNA that does not code for protein, non-coding RNAs have revealed themselves to be involved in the regulation of DNA, proteins, and other RNA species. They are often tissue and cell type specific, and have been implicated in numerous cellular processes, including many involved in disease pathology. This includes cancer, where non-coding RNAs have been associated with multiple disease phenotypes, as well as appearing as promising biomarkers and potential therapeutic targets [22]. These ncRNAs are split up into two main classes based entirely on sequence length, long, and short ncRNAs.

1.6 Small non-coding RNAs

1.6.1 What are small non-codings RNAs?

Small non-coding RNAs (sncRNAs) are generally defined as any ncRNA under 200 nucleotides in length. There are many classes of sncRNAs, usually defined based on their length, expression, or folding patterns. These classes include micro RNAs (miRNAs), piwi-interacting RNAs (piRNAs), and small nucleolar and nuclear RNAs (snoRNAs and snRNAs, respectively).

SnoRNAs and snRNAs are the least well studied class of sncRNAs. While small nucleolar have been studied much less than miRNAs, they are lowly expressed and have functions pertaining to RNA regulation and have been shown to play a role in the early stage of pre-mRNA processing. They can be 150 nucleotides long, are transcribed by pol II or III, and are named for their localization to cajal bodies and speckles within the cell nucleus. In contrast, snoRNAs are named after their expression in the nucleolus. Transcribed by RNA pol II, they contain sequences complementary to other RNAs and are thought to act as RNA guide sequences, leading to modification of other RNA species, including ribosome components. snoRNAs have been associated with carcinogenesis, but the mechanisms remain mostly unknown [23] [24].

PiRNAs are shorter than snoRNAs and snRNAs, having lengths 25–30 bp long [25]. PiRNAs were discovered in germ cells, where they were thought to have functioned exclusively, but recently they have been observed in somatic tissues with conserved biological functions [26] [27]. PiRNAs function similarly to several other classes of small RNA, regulating gene expression through a small RNA guided mechanism that targets complementary DNA sequences. PiRNAs bind a class of proteins called PIWI proteins that are part of the RISC complex and guide this complex to complementary transcripts where the complex performs its canonical function to

inhibit target sequences [28]. The function of piRNAs in germ cells was shown to be silencing of transposable elements and maintenance of genomic stability [29]. Recent work has shown additional functions such as epigenetic activation and mRNA transcript silencing, as well as cellular responses to certain environmental agents [30] [31] [32] [33].

MiRNAs are the most well studied and widely expressed class of small ncRNA. First discovered by accident in the early 90's, Scientists studying *C. elegans,* found a vital development gene called *lin-4,* and were surprised when they could not identify a protein product. Initially assumed to be an inhibitory protein coding gene, they discovered that it encoded two small non-coding genes. It wasn't until 7 years later that the first human miRNA was discovered (let-7). Since then over 2500 miRNAs have been discovered in the human genome [34] [35]. Ranging from 19–24 nucleotides in length, they are shorter than piRNAs, snoRNAs, or snRNAs, and have a conserved hairpin structure. miRNAs also have a 7-8 base long "seed" sequence at the 5' end of the miRNA transcript. MiRNAs are evenly spread across the genome with 50% in non-coding transcripts and 50% located in introns of protein coding regions [36]. MiRNAs have become widespread in mammalian cell biology for their ability to regulate protein coding genes, and they will be the sncRNA most focused on in this thesis work [37].

Table 1.2 Small non-coding RNA size and function

TYPE	SIZE (NT)	KNOWN FUNCTION
MIRNA	19-24	mRNA inhibitors
PIRNA	25-30	DNA mediated gene regulation
SNRNA	~150	Pre-mRNA processing
SNORNA	60+	RNA guide sequences

1.6.2 microRNA function

MiRNAs function to inhibit translation of protein coding mRNA targets. They do this by guiding a protein complex called RISC to specific mRNA target sequences. Each miRNA targets mRNAs with complementary sequences in their 3' untranslated regions (UTR). However, full miRNA sequence complementarity is not required, as the miRNAs "seed" sequence sufficient to target mRNAs. As the seed sequence is only 7-8 bases long, the number of potential target genes for a particular miRNA in the genome is immense. In fact, each miRNA may be able to target hundreds of genes, and they have been described as master regulators of the genome because of their ability to target a wide variety of genes simultaneously [38]. MiRNAs with identical sequences are expressed from multiple places in the genome simultaneously. In these cases, they are referred to as the same miRNA, but are given a number to denote their location in the genome. For example, miR-29b-1 is on chromosome 7, and miR-29b-2 is on chromosome 1, and both have the same sequence. MiRNAs can also be referred to as the same "family," which is when they are derived from the same ancestor and have the same seed sequence. As such, families can target overlapping and similar sets of target genes [39].

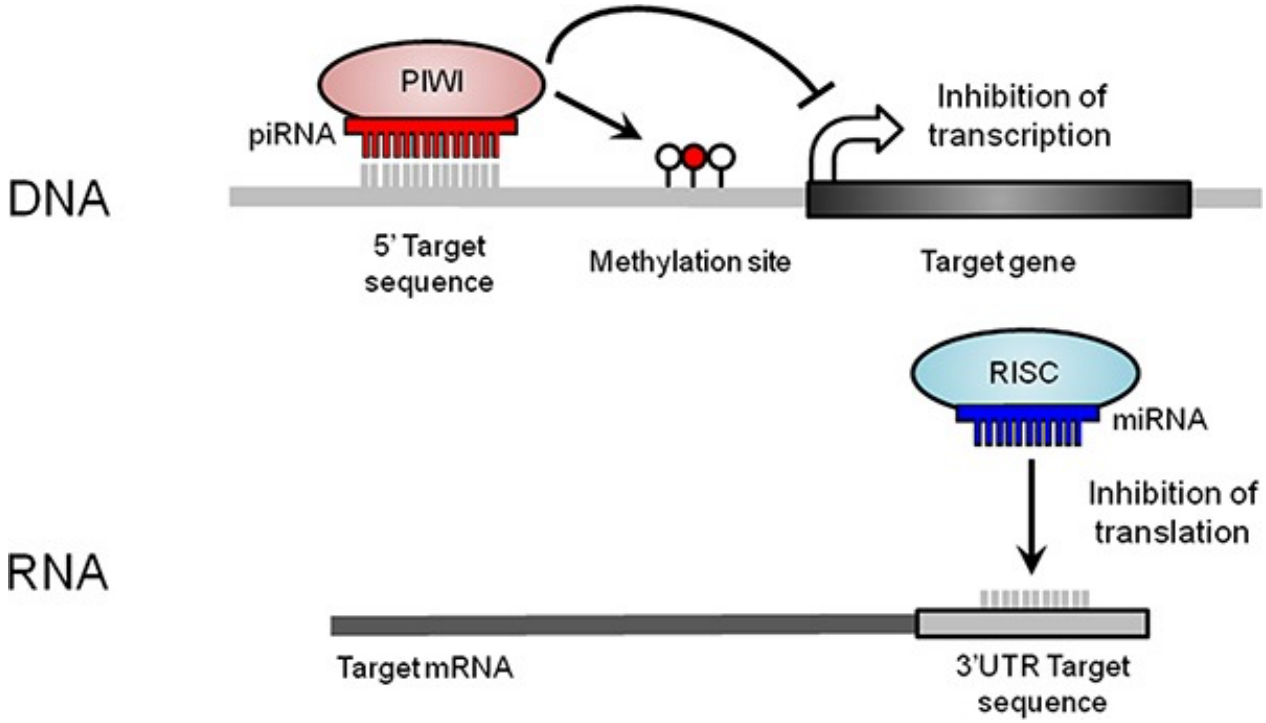


Figure 1.5 Mechanism of inhibition for miRNAs and piRNAs.

miRNAs join with the RISC complex and bind to mRNAs containing complementary sequences in the 3'UTR region. The mRNA is then either degraded or prevented from being translated into protein (below). PiRNAs join with the PIWI complex and bind to complementary DNA sequences to enact their function. This can result in changing the epigenomic state of the DNA, leading to effects like inhibition of translation through DNA methylation (above). Adapted from our manuscript "Deregulation of small non-coding RNAs at the DLK1-DIO3 imprinted locus predicts lung cancer patient outcome", published in Oncotarget [40].

1.6.3 microRNA biogenesis

MiRNA biogenesis has been studied extensively and is described in the following steps:

1. miRNA transcription: miRNAs are transcribed by polymerase 2 into 100+ nucleotide long primary transcripts with poly A tails known as pri-miRNAs.

2. pre-miRNA processing: pri-miRNAs are processed by a large protein RNaseIII complex called DGCR8-DROSHA into 70-120 nucleotide sequences known as pre-miRNAs.

3. Cytoplasmic export: pre-miRNAs are shuttled into the cytoplasm by Exportin 5

4. Cleavage into miRNA duplex: RNase III enzyme DICER cleaves pre-miRs into mature double stranded miRNAs.

5. Incorporation into RISC: 1 of 2 mature strands becomes a guide RNA for RISC complex and guides the complex to specific targets based on the miRNAs sequence [41].

6. Binding of RISC to target mRNA: Once bound, RISC causes the degradation of the mRNA transcript or prevents translation through steric hindrance.

1.6.4 Predictive tools

The short length and conservation of miRNAs seven base target sequences make them ideal for computational prediction of mRNA binding targets, and since their discovery, there have been great advancements in the field of miRNA target prediction. Programs have been designed to search for targets by identifying 6-8mer sequences complementary to miRNAs seeds in the 3'UTRs of mRNAs. These *in silico* analyses consider sequence complementarity and conservation, as well as binding energy of complementary bases to predict which miRNAs will target which mRNAs. These tools have become vital for researchers studying miRNA-mRNA biology. More recent versions include increased stringency and can take experimentally

validated interactions into account [42]. However, miRNAs are promiscuous, therefore it can be difficult to tell which of the many predicted interactions are prioritized in a complex biological system.

1.6.5 Small non-coding RNA in cancer and disease

Because miRNAs are inhibitory molecules that target a vast number of genes in the genome, they can be harnessed by tumours to provide growth advantages. As such, they can be both oncogenic and tumour suppressive. MiRNAs that target tumour suppressor genes are often overexpressed in human cancers. Alternatively, miRNAs that target growth genes are often downregulated in tumours.

MiRNAs can be affected by DNA level alterations similarly to protein coding genes, and tumours can take advantage of this. For example, the first tumour suppressor miRNA in humans was found by researchers searching for a canonical protein coding tumour suppressor. While searching for a tumour suppressor in the frequently deleted 13q14 region of B-cell leukemias, Calin and Croce discovered two tumour suppressive miRNAs, miR-15a and miR-16-1, instead of a protein coding gene [34 43]. Shortly after this the first oncogenic miRNA was discovered in human lymphomas when a frequently overexpressed region was found to contain a cluster of miRNAs. This cluster of miRNAs, known as miR-17-92, is one of the best known examples of sncRNAs driving cancer [44].

Since then miRNAs have been implicated in each of the hallmarks of cancer, and validated examples in a variety of cancer types are detailed below:

Sustaining proliferative signaling: Unchecked cell growth is a key feature of cancer cells, and the oncogenes that drive this growth are often the most sought-after clinical targets. *EGFR* is one

of the most commonly deregulated oncogenes in LUAD and happens to be targeted by a number of miRNAs including miR-7, and miR-34 [45] [46].

Activating invasion and metastasis*:* miRNAs have been shown to play a role in the spread of cancers from local invasion to systemic metastasis. Downregulation of the *E-cadherin* gene has been associated with the loss of cell adhesion leading to increased cell motility, and miR-200 has been shown to target this gene and be upregulated in lung cancer cells [47].

Resisting cell-death: The caspase family of protease enzymes are a vital part of programmed cell death, and miRNAs have been shown to be utilized by tumours to circumvent this process. For example, miR-519a-3p targets Caspase-8 in breast cancer cells to avoid cell death induced by apoptotic stimuli and may also be associated with resistance to treatment [48].

Genome instability: Compared to normal cells, cancer cells are saturated with genomic alterations and mutations, commonly a result of deficiency in the DNA damage response pathway (DDR). miRNAs have been both shown to be involved in the maintenance of genome integrity, as well as harnessed by cancers to downregulate genes involved in this process. In ovarian cancer miRNAs have been shown to both regulate DDR and predict the outcome of patients treated with DNA adduct causing platinum treatment [49].

Evading growth suppressors: Another key tenant of cancer cells is the inactivation of cellular growth checkpoints. The well-known tumour suppressor gene *TP53* is mutated, deleted, or downregulated frequently across many cancer types, including lung cancer. In lung cancer cells the miRNAs miR-641 and miR-660 have been shown to impact TP53 through downregulation of an upstream regulator, MDM2 [50] [51].

Inducing angiogenesis: Increased vascularization is necessary in order to support the increased oxygen and nutrient requirements of rapidly growing tumours. The VEGF family of genes is

most commonly associated with this phenotype and miRNAs have been shown to be involved this process. For example, miR-494 downregulates the well-known tumour suppressor PTEN, which has anti-VEGF functions [52].

Avoiding immune destruction: The complex interplay between tumour cells and the immune system has been recognized as a vital aspect of cancer biology and inspired a new generation of cancer drugs. Recently in lung cancer the PD-L1/PD-1 pathway has become a clinical target for certain immunogenic tumours. Interestingly, miRNAs have been recently shown to be involved in this pathway. For example, miR-34 targets the 3'UTR of PD-L1 and may suppress its translation into protein [53].

Enabling replicative immortality: While normal cells have limited growth potential due to the shortening of telomeres after each replication, cancer cells find ways to skirt this limitation. hTERT, the catalytic subunit of telomerase has been shown to be regulated by miRNAs in several tumour types, such as miR-512-5p in head and neck cancer [54].

Deregulation of cellular energetics: Fast growing and energy demanding cancer cells can reprogram metabolic pathways to bypass oxidative phosphorylation and favor glycolysis. One of the first steps in glycolysis revolves around the entry of glucose into the cells, which is primarily done by glucose transporter proteins known as GLUTs. miRNAs have been known to modulate the expression of these transporters in many cancer types. For example, in oral squamous carcinoma miR-340 targets GLUT1 and is frequently upregulated [55].

1.6.6 small non-coding RNAs and drug resistance

A major issue in the treatment of cancer is a tumour becoming resistant to chemotherapeutic drugs. Small non-coding RNAs, including miRNAs, have been recently shown

to be associated with resistance or sensitivity to many cancers, including lung cancer. For example, miR-92a-2 has been proposed to be predictive of chemotherapy resistance in small cell lung cancer (SCLC) [56]. While measuring pre-treatment expression of miRNAs is commonplace and can be used to predict outcome, drug resistance is hard to quantify within a patient. Post-treatment samples are infrequently taken as they are rarely of clinical benefit to the patient. As such cell models can be a useful tool to directly quantify molecular features that affect a cancer cells response to chemotherapeutics.

In previously published work our laboratory has described the use of cell line data from multiple institutions to identify recurrent alterations to miRNA loci that were associated with resistance to chemotherapy. By integrating DNA CN and miRNA expression profiles we identified recurrent alterations in lung cancer cell lines that were either highly sensitive or highly resistant to 18 different chemotherapeutics. This lead to the discovery of several miRNAs that were frequently gained and overexpressed in chemotherapy resistant cell lines [57]. For example, we identified miR-10b to display frequent DNA copy number gains, be overexpressed, as well as resistant to the chemotherapeutic proteasome inhibitor MG-132. Interestingly a predicted target of miR-10b, RAD1, is a component of the nucleotide excision repair complex, which is known to play an important role in chemotherapy response. We found that RAD1 is significantly downregulated in the same resistant cell lines where miR-10b is upregulated (Figure 1.7). Additionally, predicted targets of other resistance-associated miRNAs included genes in DNA replication and repair pathways, which are pathways known to be involved in chemotherapy resistance. This study emphasizes the potential impact of miRNAs in cellular response to therapies used in the clinic. Chemotherapy resistance-associated miRNAs could potentially be used as predictive biomarkers for selecting patients most likely to respond to specific therapies.

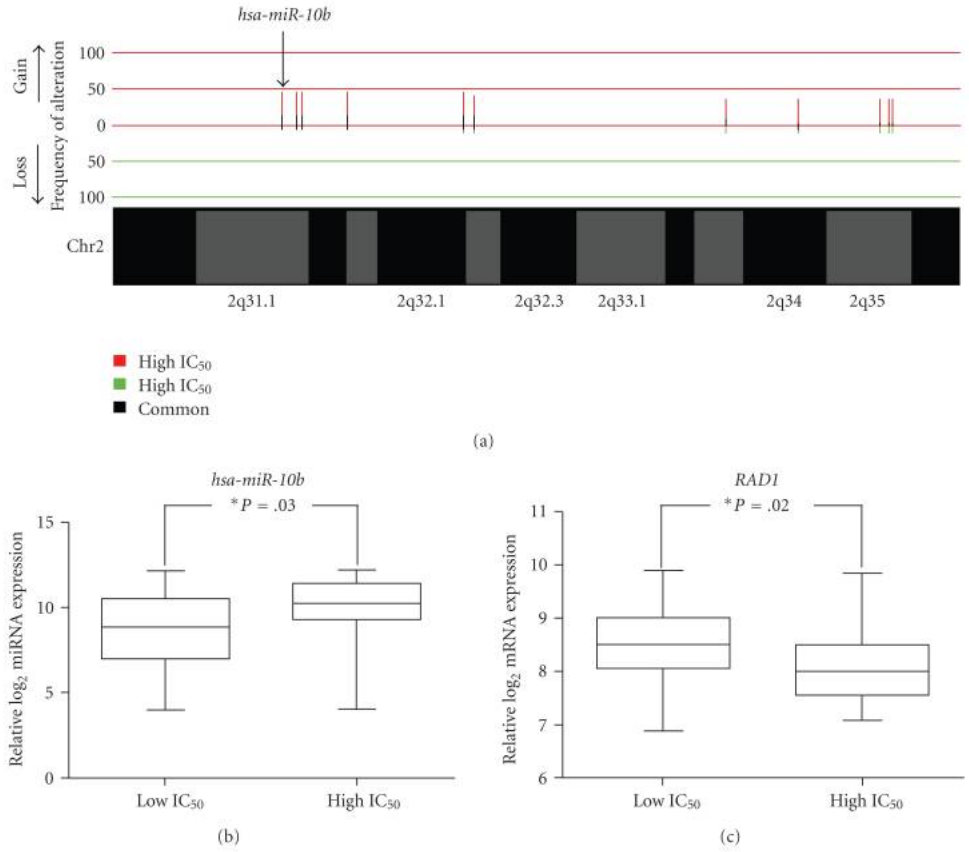


Figure 1.7 Changes in miRNA DNA copy number affect drug resistance in cancer cells.

Copy number alterations in cell lines resistant to MG-132 (above) reveal miR-10b to be frequently gained. Similarly, miR-10b expression is significantly increased in resistant cell lines, while target gene RAD1 expression is significantly decreased, suggesting it may be inhibited by miR-10b (below). This figure is from our manuscript "MicroRNA Gene Dosage Alterations and Drug Response in Lung Cancer" published in the Journal of Biomedicine and Biotechnology [57].

1.6.7 small non-coding RNAs as clinical markers

Biomarkers have become a vital tool in the clinic, as a way to both non-invasively predict disease severity as well as predict response to treatment. In addition to being present in primary tumours, sncRNAs have been found in the circulation, where they can be detected less invasively. They can be released through a variety of mechanisms, including exomes, apoptosis, and necrosis. MiRNAs have also been found to be more stable than protein coding genes and can remain stable in paraffin embedded tissue [58] [59]. As such, there are many examples of miRNAs being used as clinical markers. For example, miR-7 has been proposed as a prognostic biomarker in LUAD, where low levels of the tumour suppressive miRNA let-7 were shown to be associated with significantly shorter survival time [60]. While expression of a single miRNA can be useful, one major limitation is sequence overlap between healthy and diseased tissue. High sensitivity and specificity are the most important criteria for clinical application of diagnostic or prognostic biomarkers, as without this there is an increased chance of false negative or false positive diagnosis, which could negatively affect patient care [61]. Panels of miRNAs have been shown to be more effective biomarkers. For example, individual genes from the *DLK1-DIO3* locus have been associated with LUAD outcome and a combined signature of three miRNAs has been shown to better predict overall survival after tumour resection [62].

In a previous study our laboratory sought to determine whether consideration of other sncRNAs could improve predictive panels of miRNAs. We interrogated two independent datasets of LUAD and LUSC sequencing data for sncRNA expression at the *DLK1-DIO3* locus. We found that seven of the 138 piRNAs encoded at the locus were expressed. When we incorporated four of the seven piRNAs into the existing 3-miRNA survival signature we discovered that it was able to better stratify patients into risk groups (Figure 1.8). While the three

miRNA signature could not stratify patients into risk groups in LUSC, we found the new seven sncRNA signature were able to effectively predict risk. These results highlight the clinical utility of incorporating piRNAs and other understudied sncRNA species into miRNA predictive panels.

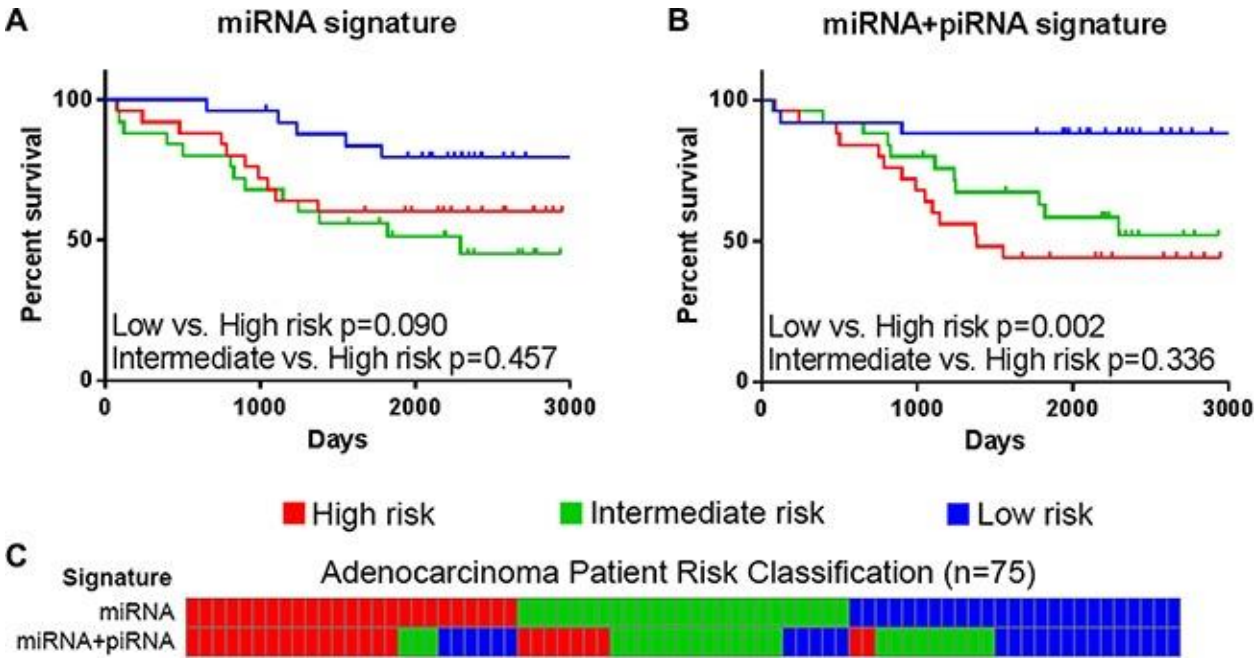


Figure 1.8 Addition of piRNAs improves risk stratification in 75 lung adenocarcinoma patients.

Kaplan-Meier curves of (A) a miRNA miRNA only signature and (B) a signature with the addition of piRNAs. High risk is shown as red, intermediate risk as green, and low risk groups are shown as blue. Log-rank *p*-values of select survival comparisons are shown. (C) Patients re-classified by the addition of the miRNA+piRNA signature. Adapted from our manuscript "Deregulation of small non-coding RNAs at the DLK1-DIO3 imprinted locus predicts lung cancer patient outcome", published in Oncotarget [40].

1.6.8 small non-coding RNAs as clinical targets

In addition to their use as potential biomarkers and indicators of chemotherapy resistance, snRNAs have shown promise as direct clinical targets. As the most well studied class of sncRNAs, miRNAs have been the primary focus for therapeutic targets. Thus far, most miRNA based therapeutics fit into two categories: miRNA inhibitors and miRNA replacement therapy. MiRNAs inhibitors work based on the hypothesis that targeting oncogenic miRNAs for subsequent downregulation will either kill or slow down the growth of tumour cells. Some examples of types of miRNA inhibitors include: miRNA locked nucleic acids (LNAs), miRNA agonists, and miRNA specific antisense oligonucleotides. These therapies target miRNAs for degradation or inhibition of function. A therapeutic antagonist has been developed against miR-10b for use in patients in glioblastoma and is currently in pre-clinical development [63].

MiRNA replacement therapy comprises of delivery of synthetic or exogenous tumour suppressive miRNAs into cancer cells. As these miRNAs have been downregulated in order to prevent inhibition of important oncogenes, these therapies attempt to re-introduce these miRNAs to downregulate oncogenes, and re-activate growth control signaling. Several of these miRNA replacement therapies have entered clinical trials. For example, a miR-34 mimic entered human clinical trials for patients with advanced or metastatic liver cancer, and let-7 mimics have been developed to treat a variety of solid cancers including LUAD [64] [65]. There are still problems to overcome with these novel therapies, such as delivery to cancer cells, accumulation in the liver, low bioavailability, and off-target effects.

1.7 Long non-coding RNAs

1.7.1 An introduction to the wild world of long non-coding RNAs

The longer and more complex cousins of small ncRNAs are known as long non-coding RNAs (lncRNAs). They are defined as transcripts that do not encode for protein that exceed 200 base pairs in length. While a handful of lncRNAs were characterized before the human genome was decoded, such as the female specific X-chromosome silencing *XIST*, these were thought to be rare. Once sequencing studies revealed that only 2% of genes in the human genome were protein coding, the widespread expression of lncRNAs became clear. lncRNAs have many functions and are able to regulate proteins, DNA, and other RNA species. They have been found to be more tissue specific than protein coding genes, localize to different cellular compartments, and can contain multiple functional domains. lncRNAs have been shown to be involved in many vital cellular processes, including chromatin remodeling, splicing, protein complex formation, transport and localization, translation, miRNA sequestration, and regulation of transcription (Figure 1.9) [66] [67].

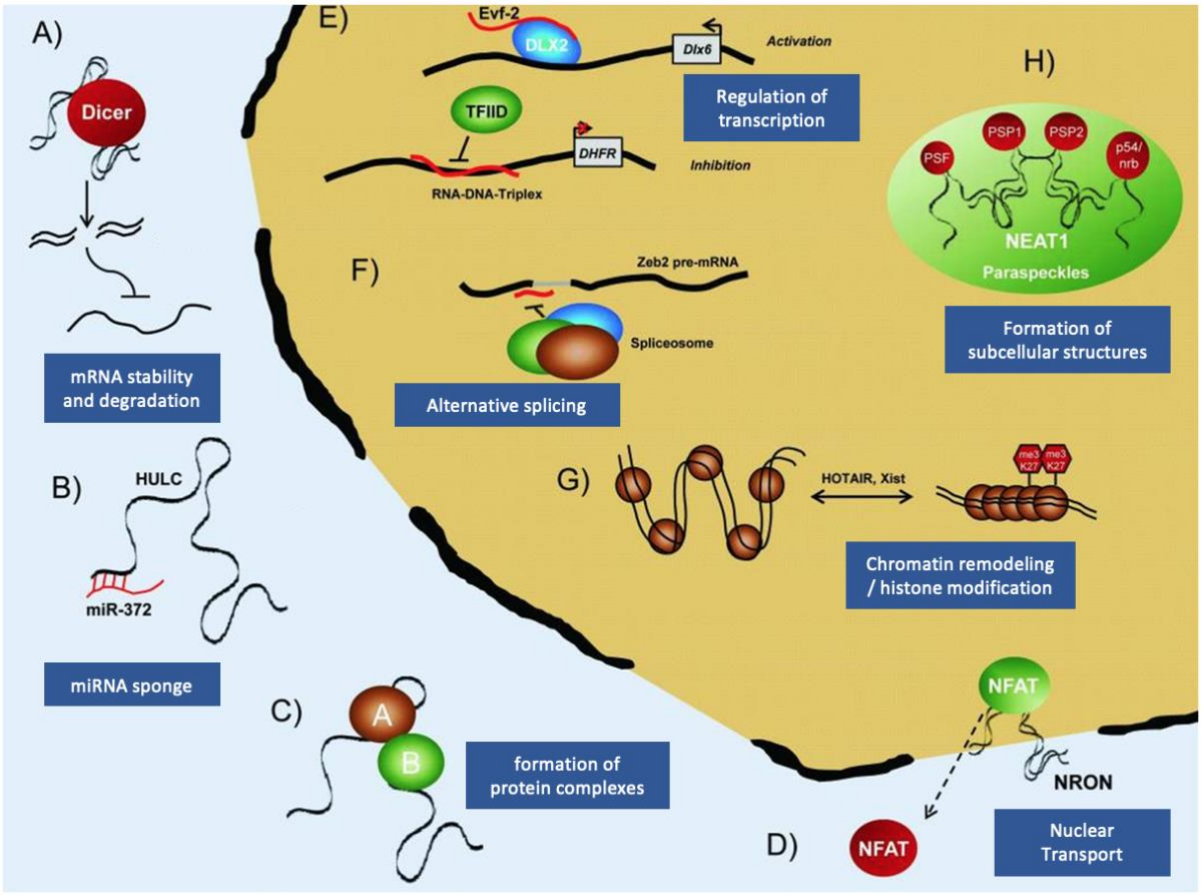


Figure 1.9 The many functions of long non-coding RNAs

Since their discovery lncRNAs have been observed to be involved in regulating DNA, proteins, and other RNAs. Several examples of lncRNA functions are seen above. Adapted from Gutschner *et al*. 2012 [66].

1.7.2 Classes of long non-coding RNAs

Due to their complexity, lncRNA classes are more fluid than small ncRNAs, and new functions are continually being discovered. However, some common functions are shared and can useful to classify lncRNAs. The two primary categories of lncRNAs focused on in this thesis are; *cis*-acting: those that enact function on their transcriptional loci or neighboring genes, and *trans*-acting: those that enact function on genome wide targets.

The first *cis*-acting lncRNAs were a subgroup called natural antisense transcripts (NATs). Their function was thought to be derived from the protein coding gene they overlapped with. An example of this is the lncRNA *Kcnq1ot1* which was discovered to silence overlapping protein coding *Kcnq1* when transcribed [68]. Later it was discovered that lncRNAs could function to not only regulate overlapping genes, but neighbouring genes and entire transcriptional loci, and that these lncRNAs are better defined as *cis*-acting. For example, the lncRNA *XIST*, originally thought of as a NAT, silences the entire duplicated X chromosome in females. To positively or negatively regulate their neighbors, *cis-acting* lncRNAs have been observed to use a handful of different mechanisms, such as head-to-head transcription blockage and recruitment of protein complexes, such as methylation and chromatin remodeling complexes. Head to head transcriptional interference occurs when the transcription of a lncRNA itself causes inhibition of overlapping genes. An example of this is the lncRNA *AIRN*; when *AIRN* recruits RNAPII, it reduces the ability of overlapping *IGF2R* to be transcribed [69]. Interactions with chromatin remodeling complexes can result in selective activation or repression of specific gene loci. For example, the lncRNA *ANRIL* is responsible for recruiting the PRC2 chromatin remodeling complex to silence the neighboring *CDKN2A/B* locus [70]. Another feature of cis-acting lncRNAs

is that they commonly localize to the nucleus and may not require many transcripts to cause large regulatory affects [71].

In contrast, trans-acting lncRNAs regulate genes not associated with their transcriptional loci. This can refer to many possible types of gene regulation including: interacting with other RNAs, miRNA-based regulation, and recruiting protein complexes to distant transcriptional loci. Similarly, to lncRNAs acting in cis, the recruitment of protein complexes to transcriptional loci is a major method of regulation in trans. For example, one of the most widely studied lncRNAs, *HOTAIR*, is transcribed from the *HOXC* locus on chromosome 12 and travels to the *HOXD* locus on chromosome 2, where it recruits PRC2 to repress transcription [72]. lncRNAs can also interact with other RNAs in order to improve or decrease stability of a transcript. For example, the lncRNA *LAST* stabilizes Cyclin D1 by recruiting the CNBP protein to the Cyclin D1 mRNA transcript [73]. Another example is the lncRNA *BACE1-AS*, which binds to the *BACE1* mRNA in the cytoplasm to enhance the stability of the protein coding gene, in part by blocking binding of miR-485-5p [74] [75].

Interactions between lncRNAs and miRNAs are thought to be a common mechanism of gene regulation. Numerous studies have demonstrated that a lncRNA can regulate an mRNA transcript by binding a shared miRNA, thereby releasing the mRNA from miRNA-mediated inhibition. This type of regulation is known as miRNA "sponging". LncRNAs that may be the best candidates for miRNA sponging are lncRNAs expressed from pseudogene loci. Originally formed through duplication events or reverse transcription of an mRNA transcript, pseudogenes are relatives of protein coding genes littered throughout the human genome. Throughout the course of human evolution these redundant genes have lost the ability to code for protein due to mutations resulting in frameshifts and premature stop codons. However, these pseudogenes can

maintain a high degree of similarity to their protein coding relatives when they are expressed, and this degree of similarity means that they may contain several of the same miRNA binding sites as their parent genes. For example, the pseudogene *PTENP1* shares many of the same miRNA binding sites as its parent protein coding gene *PTEN*. When expressed, *PTENP1* positively regulates *PTEN* expression by acting as a decoy for these shared miRNAs. Pseudogene derived lncRNAs have also been observed to recruit protein complexes to the transcriptional loci of their parent gene. For example, the *PTENP1-AS1* is able to recruit PRC2 to the *PTEN* locus transcriptional loci and silence the gene [76].

Sponge-based regulation is an increasingly popular topic. Recently, researchers have speculated that all RNAs and miRNAs may regulate each other; this theory has been referred to as the competing endogenous RNAs (CERNA) hypothesis. According to CERNA, all miRNAs and their target genes have a reciprocal relationship. miRNAs are titrated by target binding sites, and that just as genes containing target sites are regulated by miRNAs, miRNAs are regulated by genes containing their target sites [77]. However, this theory is controversial as the multi gene-miRNA-target site relationships predicted through CERNA largely lack sufficient validation in biological systems.

1.7.3 Long non-coding RNAs in cancer

With such a wide variety of functions, it is perhaps unsurprising that lncRNAs are involved in many key processes in cancer cells. lncRNAs can function both as oncogenes, as well as tumour suppressor genes. For example, the oncogenic lncRNA *ANRIL* is overexpressed in prostate cancer tissue in order to silence the tumour suppressive INK4b/ARF/INK4a locus [70]. In contrast, the lncRNA *TARID* acts as a tumour suppressor. *TARID* recruits GADD45A to

demethylate the tumour suppressive gene *TCF21*, and has been shown to be downregulated in a number of cancers, including lung cancer [78].

In our previous work we performed the first pan-cancer transcriptional analysis of lncRNA expression. We re-mined serial analysis of gene expression (SAGE) data from 272 sequence libraries in order to interrogate the non-coding transcriptome of 19 different cancer types, as well as samples from 26 non-malignant tissues [79]. We noted that the chromosomal expression patterns of lncRNAs across the genome did not match the patterns of expression seen for protein coding genes. Additionally, we found increased tissue-specific expression of lncRNAs compared to protein coding genes, and each cancer type had a significant number of tumour specific lncRNAs. For example, the majority of deregulated lncRNAs in each cancer type were specific, with moderate overlap between two tissues. Additionally, we found only 8 deregulated lncRNAs overlapped between our three largest cancer datasets, which were breast, brain and lung cancers (Figure 1.9). In addition to finding many novel cancer-associated lncRNAs, we confirmed the expression of known oncogenic lncRNAs, such as such as *MALAT1*, *GAS5*, and *NEAT1*. This study generated the first expression atlas of lncRNAs across multiple cancer types. Our results emphasized the widespread tissue-specific deregulation of lncRNAs in human cancers, which were later confirmed in subsequent sequencing studies [80].

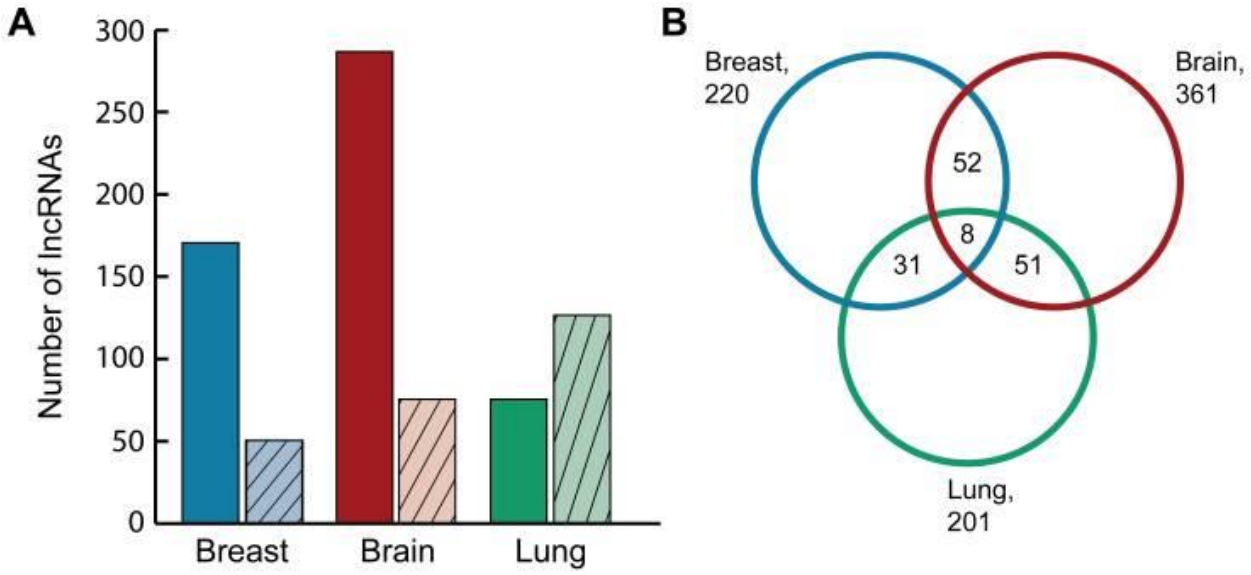


Figure 1.10 Aberrantly expressed lncRNAs across human cancers.

(A) Number of lncRNAs showing significant expression changes (>2 fold upregulated, BH p-value <0.05) across three major cancer types. Upregulated genes are indicated by solid bars while hatched bars indicate downregulated genes. (B) Venn diagram showing the overlap of these differentially expressed lncRNAs in the three major cancer types. Adapted from our manuscript "*Human Cancer Long Non-Coding RNA Transcriptomes*" published in *PLoS One* [79].

Since the discovery of their widespread deregulation in cancer, lnRNAs have been shown to play a role in each of the various hallmarks of cancer:

Activating invasion and metastasis: one of the first lncRNAs found in the context of cancer was *MALAT1,* which stands for metastasis associated lung adenocarcinoma transcript 1. This highly expressed lncRNA was discovered in primary LUAD tumours that had a high risk of metastasis. *MALAT1* interacts with the PRC2 complex to activate transcription of a number of cell growth related genes. Additionally, *MALAT1* has been proposed to be involved in the organization of subnuclear components that participate in activation of pro-cell growth pathways [81] [82].

Sustaining proliferative signaling: *PCAT-1* (prostate cancer associated transcript 1) is a transcriptional repressor for networks of genes associated with mitosis and cell cycle control.

Resisting cell-death: lncRNA *PANDA* interacts with a transcription factor called NF-YA and results in decreased expression of genes associated with apoptotic functions in multiple cancer cell types.

Genome instability: *NORAD* has been associated with breast cancer and loss of the lncRNA has been shown to promote chromosomal instability. Recent research has shown that *NORAD* is upregulated in breast cancer as part of the DNA damage response pathway and can promote nuclear assembly of the genome stability-associated NARC1 complex [83] [84].

Evading growth suppressors: *ANRIL* (antisense non-coding RNA in the *INK4* locus) interacts with polycolm repression complex 2 (PRC2) and recruits the complex to condense active chromatin and silence the genes around its transcriptional loci, including the well-known tumour suppressor INK4B (p15). This occurs in multiple cancer types including lung cancer.

Inducing angiogenesis: HIF1a is a well-known critical regulator of angiogenesis and has a natural antisense transcript (*aHIF*). *aHIF* expression is associated with HIF1a mRNA loss of stability and degradation in multiple cancer tissues [85].

Avoiding immune destruction*:* While researchers are just beginning to study the role that lncRNAs play in the immune system, the lncRNA *THRIL* is important for TNF-a and IL-6 secretion and is a vital part of innate immune response. *THRIL* has been associated with inflammatory disease and response to inflammation, which are major factors in multiple cancer types [86].

Enabling replicative immortality: The RNA component of telomerase holoenzyme called *TERC* (Telomerase RNA component) is vital for telomerase function and had been found to be upregulated in several human cancers [87] [88].

Deregulation of cellular energetics: Just as miRNAs have been known to modulate the expression of glucose transporters to favour glycolysis in cancer cells, so have lncRNAs. For example, the lncRNA *PCGEM1* promotes the expression of *GLUT1* in prostate tumours [89].

1.7.4 Long non-coding RNAs in the clinic

While not as well clinically studied as sncRNAs, lncRNAs have shown promise as disease biomarkers. In the context of blood based cancer diagnostic and prognostic biomarkers, lncRNA's unique chemistry may provide stability advantages over protein coding genes. The complex folding and tertiary structure of lncRNAs results in high stability when circulating in bodily fluids. Furthermore, lncRNAs have been shown to be able to resist ribonuclease-based degradation and to have even greater stability when packaged in exosomes or apoptotic bodies [90] [91]. As such, lncRNAs may make excellent candidates for early diagnosis, detection of different

subtypes, monitoring of metastasis and progression, sensitivity to treatment, and disease recurrence.

The attributes of a good biomarker are easily detectable (highly expressed), accessible (in the blood), as well as high sensitivity and specificity. An example of a lncRNA that fits these parameters is the lncRNA *H19*, which was one of the first lncRNAs discovered. *H19* is a promising biomarker for early stage gastric cancer patients, where its sensitivity and specificity is higher of than that of conventional protein biomarkers such as CEA and CA199 [92]. The lncRNA *MALAT1* has also been considered a biomarker in lung cancer [93]. Just as panels of sncRNAs can improve specificity over single genes, so can the addition of lncRNAs to current biomarker panels. For example, a panel including the lncRNA *GAS5* with CEA was more effective than CEA alone in diagnosing NSCLC [94].

As lncRNAs have been increasingly implicated in major cancer pathways and phenotypes, they have become intriguing clinical targets. Currently there are three main methods to target a lncRNA: transcriptional inhibition, steric-hindrance of RNA-protein interactions, and post-transcriptional RNA degradation. Methods using post transcriptional RNA degradation have been the most well-studied. One type of therapy is particularly intriguing with regards to targeting lncRNAs for degradation: antisense oligonucleotides (ASO's). ASO's are a nucleic acid based therapy designed to bind to complementary target RNA, causing either degradation, splicing alterations, or steric hindrance of protein interaction. ASO's are 15-20 nucleotides long and are quicker and cheaper to design than small molecule inhibitors for specific protein interaction domains [95]. Additionally, ASO-based targeting has higher specificity and fewer off-target effects [96]. These inhibitors have shown early promise in the clinic, with numerous ASO-based therapeutics entering clinical trials, and recent advances have added degradation resistance

phosporothioate backbone linkage modifications to improve stability [97]. Intriguingly, ASO's may be even better at targeting lncRNAs, as the RNA target is the final gene product rather than an intermediate product (Figure 1.10).

Due to its overexpression in lung cancer, and association with metastasis, *MALAT1* has become a potential target of interest. *MALAT1* shRNA-mediated knockdown resulted in a significant reduction in both cell migration, and invasive ability, indicating that it may be an attractive target [98]. ASO's developed against *MALAT1* have shown early promise in lung cancer xenografts models. Immune-deficient mice injected with human lung cancer cells show 70% less metastasis to the lungs when treated with *MALAT1* ASO's [99]. *MALAT1* ASO's are also being tested for aerosol delivery to the lungs for the treatment of lung cancer, and ongoing pre-clinical studies on patient-derived tumour organoids are moving these treatments closer to the clinic.

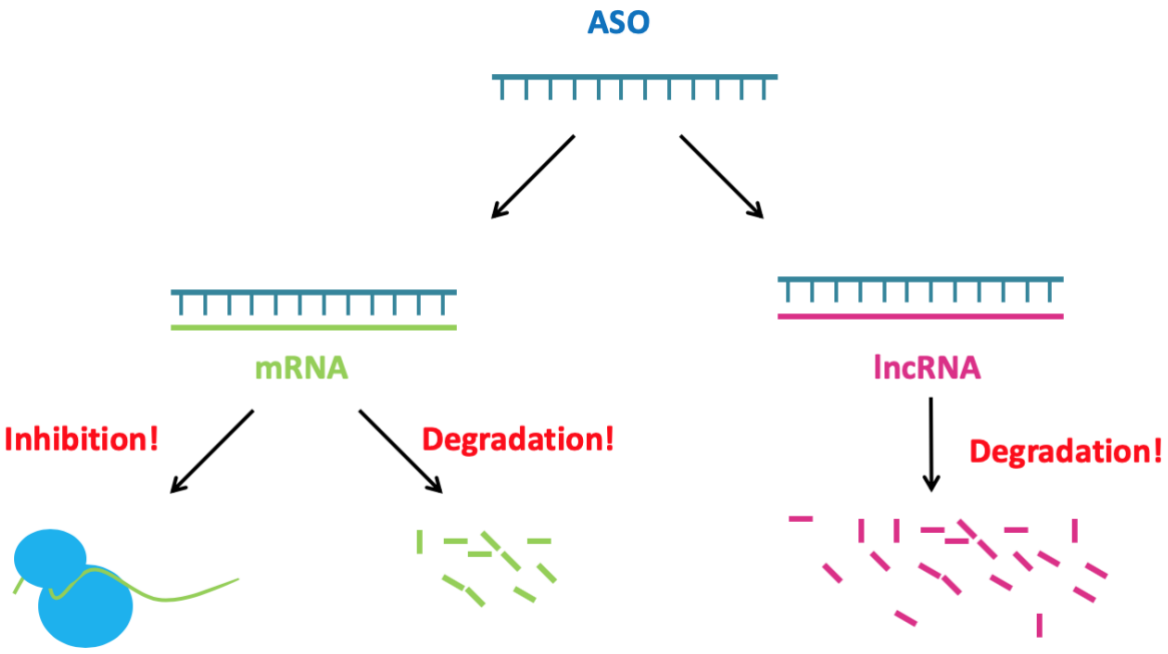


Figure 1.11 Function of Antisense Oligonucleotides (ASOs)

Antisense Oligonucleotides (ASO's) function by binding to a target RNA and causing either steric hindrance or degradation via RNAse H. LncRNAs make attractive ASO targets as their final states are an RNA product, rather than an intermediate product before translation.

1.7.5 Challenges in long non-coding RNA characterization

While lncRNAs have great clinical potential, there are significant challenges in the field. While there are several lncRNAs that have been well characterized, the vast majority remain undescribed. lncRNAs are notoriously difficult to characterize, and currently there are no reliable methodologies to predict lncRNA function. Unlike protein coding genes, lncRNAs do not have well characterized functional domains, and the techniques we use for protein coding gene functional characterization are not well suited to lncRNAs. For example, RNA-protein immunoprecipitation is commonly used to identify interaction partners. However, these assays have been found to be prone to nonspecific binding of RNAs, leading to false positive results [100]. An example of this is the chromatin modifying PRC2 complex; PRC2 has been observed to physically interact with 20% of lncRNAs, making identification of meaningful interactions difficult [101]. Additionally, the complex folding patterns and tertiary structure of lncRNAs complicates the identification of active sites. Furthermore, while highly conserved regions of a gene usually imply functional importance, lncRNAs evolved more recently in humans, so they are in general much less conserved. Lastly, unlike miRNAs, there are no "seed" sequences, and actives sites or important binding sequences remain unknown. Many studies have found lncRNAs to be associated with certain phenotypes, but it is more difficult to ascertain how these lncRNAs are functioning, and what their downstream targets may be. The number of lncRNAs discovered continues to be greatly outpaced by the number of lncRNAs characterized. There is a great need to better predict downstream targets of lncRNAs and this will be vital for this class of RNA to become viable clinical targets.

1.8 Thesis Rationale and objective

Lung cancer is the deadliest form of cancer, and lung adenocarcinoma is the most common subtype of this disease. While targeted therapies have been of great benefit to those who carry these targetable alterations, there is a need to identify new driver genes, as well as find new clinical action points for known pathways. Next generation sequencing and following studies have shown the clear importance of ncRNAs in cancer, making them attractive candidates for investigation as novel targets. However, the field is being held back by a number of complicating features endemic to lncRNAs: complex folding patterns, lack of conserved domains, and unknown binding sequences make these genes difficult to characterize and complicates the identification of target genes. Current prediction methods primarily assume lncRNA function through physical interaction with miRNAs, and while these potential interactions are plentiful, it is unknown if they occur in a biologically relevant manner. While the number of lncRNAs named has risen significantly, the number characterized remains low, emphasizing the need for better ways to predict lncRNA function.

Objective: To investigate the roles of non-coding RNAs in lung adenocarcinoma in order to identify alternate mechanisms of cancer gene deregulation.

Hypothesis: Non-coding RNAs are deregulated on a global scale in the LUAD genome, and this affects the protein coding genes that these ncRNAs interact with, including those that drive LUAD. This thesis describes the discovery and investigation of three different mechanisms of lung cancer gene deregulation by non-coding RNAs.

Chapter 2: Common methods

2.1 Next generation sequencing of lung adenocarcinoma patient samples

2.1.1 Patient samples

We performed Next Generation Illumina HiSeq RNA sequencing on a set of 36 micro-dissected LUAD tumours and matched adjacent non-malignant tissue (n=72). Our British Columbia Cancer Agency cohort (BCCA) was composed of fresh-frozen LUAD tumours and matched non-malignant lung parenchymal tissue collected from 36 patients at the Vancouver General Hospital with approval from the University of British Columbia-BCCA Research Ethics Board. Consent obtained from the tissue donors of this study was both informed and written. Matched non-malignant samples were collected from areas >2 cm away from the tumour. In order to reduce contaminating sequences derived from alternative cell types, tissue microdissection was guided by a pathologist. Samples used in this study contained >80% tumour cell or >80% non-malignant cell content. Total RNA was extracted using Trizol reagent and standard procedures.

2.1.2 Processing of RNA-sequencing data

Total RNA was used for library construction at the Genome Sciences Center (GSC, Vancouver, Canada). Briefly, samples were first analyzed using Agilent Bioanalyzer RNA nanochip, and samples that passed quality check were arrayed into a 96-well plate. PolyA+ RNA was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) from 2 µg total RNA with on-column DNaseI-treatment as per the manufacturer's instructions. Double-stranded cDNA was synthesized from the purified polyA+-RNA using the Superscript Double-Stranded cDNA Synthesis kit (Life Technologies, USA) and random hexamer primers at a concentration of 5µM. The paired-end sequencing library was

prepared following the GSC paired-end library preparation protocol, which is strand specific. Sequencing was performed using the Illumina HiSeq 2000 platform. Raw sequencing reads were subject to a quality control process. Reads with a length < 50nt (under two thirds of maximum read length of 75 nt) and quality level (Phred) < 20 were discarded. High quality reads (.fastq files) were aligned to the NCBI GRCh37 reference human genome build using the STAR aligner (v 2.4.1d) under default parameters [102]. Aligned reads (.bam files) were quantified using Ensembl Transcripts (Release 75) reference annotations[103]. Raw RNA sequencing reads from each patient (tumour and corresponding non-malignant tissue) were deposited at the at Bioproject http://www.ncbi.nlm.nih.gov/bioproject/516232 . Quantification was performed using the Partek Flow platform as reads per kilobase per million (RPKM). RNA sequencing (.bam files) and clinical data for a secondary set of LUAD tumours and matched non-malignant tissue (n=108) were downloaded from The Cancer Genome Atlas (TCGA) Data Portal for validation purposes (https://portal.gdc.cancer.gov/) (Table 2.1). Expression profiles from TCGA were processed as described above.

Small RNAs:

Small RNA sequencing data was downloaded for all LUAD tumours with available data from the TCGA (n= 586). Raw sequencing reads were then aligned to the human genome (hg 19), before being quantified using miRbase v22 (http://www.mirbase.org/ftp.shtml).

Table 2.1 Patient clinical characteristics of the discovery (BCCA), validation (TCGA), and survival analysis (KmPlotter) datasets.

Characteristic	BCCA	TCGA	KmPlotter
Samples (pairs)	72 (36)	108 (54)	866 (n/a)
Sex			
Male	10	24	344
Female	26	30	318
Average age	70	67	–
Stage			
I	20 (56%)	28 (52%)	370 (69%)
II	11 (31%)	14 (30%)	136 (25%)
III	3 (8%)	10 (19%)	24 (4%)
IV	1 (3%)	2 (4%)	4 (1%)
Ethnicity			
Caucasian	11 (31%)	51 (94%)	–
Asian	14 (39%)	–	–
Unknown	11 (31%)	–	–
Black	–	3 (6%)	–
Smoking			
Current	5 (14%)	7 (13%)	246 (Ever) [a]
Former	6 (17%)	36 (67%)	
Never	25 (69%)	5 (9%)	246

[a] Ever smokers is a term that includes both current and former smokers.

2.2 Basic laboratory techniques

Cell culture.

Lung cancer, and immortalized lung epithelial cell lines maintained in supplier recommended media and may be supplemented with 10% fetal bovine extract (FBS) or bovine pituitary extract (BPE). Cells were grown in a humidified incubator at 37^{o}C and 5% CO_2. Cell culture is performed in a ventilated type II biosafety hood. Trypsin is diluted and used to split cell lines prior to reaching confluence.

RT-qPCR based transcript quantification:

RNA was harvested from cell lines using published methods. In brief, the Trizol protocol was used, where the addition of Trizol disrupted cell structure, and precipitate cellular RNA while protecting from RNase activity. Isolated RNA was then resuspended in DEPC-treated water and converted into a cDNA template through use of reverse transcriptase. Appropriate cellular controls were used as appropriate depending on the RNA species being quantified (TaqMan - Applied Biosystems, Carlsbad, CA). Primers and controls are described in detail in each chapter.

Chapter 3: A novel *cis*-acting long non-coding RNA controls HMGA1 expression in lung adenocarcinoma

3.1 Introduction

Long non-coding RNAs (lncRNAs) are a class of transcripts that function in gene regulation and have since been implicated in the onset of many cancer-associated phenotypes, such as progression, tumourigenesis, and metastasis. Since their discovery, one of the key challenges is effective identification of downstream target genes. In order to harness their potential for disease-specific markers and potential therapeutic targets, a better understanding of lncRNA transcription and mechanisms of action in disease is required.

Recently, an emerging class of lncRNAs - *cis*-acting - has been shown to regulate the expression of neighbouring protein-coding genes, which frequently includes protein-coding genes with oncogenic or tumour-suppressive functions. Through a variety of mechanisms including the recruitment of protein complexes these non-coding transcripts can activate or repress transcription of neighbouring genes. For example, the lncRNA *Tarid* is able to recruit the protein complex GADD45A to actively de-methylate its neighbouring protein-coding gene *TCF21*, thereby increasing the expression of this tumour suppressive gene [78]. Thus, *cis*-acting lncRNAs may represent novel mechanisms of cancer-gene regulation as well as potentially actionable intervention points in known cancer-driving pathways. Despite the prevalence of genetic and epigenetic deregulation events in lung adenocarcinoma (LUAD), the extent and consequences of aberrant *cis*-acting lncRNA expression on known cancer-driving genes is unknown.

High mobility group A1 (HMGA1) chromatin remodeling protein is enriched in several aggressive cancer types, including non-small cell lung cancer (NSCLC), where its mRNA and protein expression are substantially increased [104]. HMGA1 is part of a family of proteins involved

in chromatin architectural maintenance, which have all been implicated in tumourigenesis, particularly in breast cancer where they have been shown to facilitate every hallmark of cancer [105]. Further, HMGA1 overexpression has been shown to be a key factor driving lung metastasis. The oncogenic role of HMGA proteins stems from their activation of cancer-driving genes such as *E2F1*, *AP1*, and *CCNA1*, as well as the repression of tumour suppressive genes such as *TP53* [106]. In LUAD high *HMGA1* gene expression has been associated with poor overall survival and chemotherapy resistance [104]. While *HMGA1* is deregulated in lung cancer, the mechanisms that mediate its expression are only beginning to emerge. Here I take a genome-wide large-scale approach to investigate putative cancer-relevant *cis*-acting lncRNAs and subsequently focus on an uncharacterized lncRNA that may represent an alternative mechanism of HMGA1 overexpression in LUAD patients.

3.2 Materials and Methods

3.2.1 Sample collection and processing

Two separate cohorts of raw RNA sequencing reads from LUAD tumours with matched adjacent non-malignant tissues were used in this study: an in-house microdissected (80% tumour purity) cohort collected at the BC Cancer Research Centre (BCCA; n=36 pairs) and a publicly available cohort obtained from The Cancer Genome Atlas (TCGA; n=54 pairs). Collection and sequencing of both cohorts were performed in congruent manners as described in Chapter 2 and in a previously published manuscript [107]. Briefly, after raw RNA sequencing reads were generated (Illumina HiSeq 2000), quality control analyses were performed (Phred>20; length>50nt), reads were aligned to the hg19 build of the human genome (STAR aligner), and expression was then

quantified as reads per kilobase per million (RPKM) [102,107] LncRNA genes were annotated according to release 75 of Ensembl.

3.2.2 Gene Expression Analyses

LncRNAs close enough to enact transcriptional or epigenetic changes (within 1.5 Kb) to protein-coding genes were considered as putative *cis*-acting lncRNAs. Here, we performed a Wilcoxon signed-rank test on both lncRNA and protein-coding gene expression between tumour and matched non-malignant tissues. Genes significantly deregulated in both cohorts (BH-$p<0.05$; $FC>1.5$) were considered for further analyses. To assess potential cancer-relevant *cis*-acting lncRNAs, those neighbouring protein-coding genes with experimental evidence of associations with tumour biology were assessed. Significant associations between lncRNA and neighbouring protein-coding gene expression were determined using a Mann-Whitney U-Test between the upper and lower tertiles of samples based on lncRNA expression.

3.2.3 *In vitro* analyses

The immortalized non-malignant epithelial lung cell line BEAS-2Bs was used to assess the effect of *HMGA1-lnc* inhibition on *HMGA1* expression *in vitro*. Cells were cultured in serum-free medium: K-SFM supplemented with 30µg/mL bovine pituitary extract (BPE) and 0.0002ng/µL epidermal growth factor (EGF); maintained in an incubator at 37$_o$C and 5% CO_2. Once confluent, 2 mL of cell solution was seeded into each well of a 6x2 cm-well plate at a concentration of 50,000 cells/mL. DharmaFECT siRNAs were prepared for transfection as per manufacturer's instructions in five conditions: i) untreated control; ii) a positive control siRNA targeting GAPDH (25 nM); iii) a non-targeting control siRNA (25 nM); iv) siRNA targeting *HMGA1-lnc* at a concentration of 12.5 nM; and v) siRNA targeting *HMGA1-lnc* at a concentration of 25 nM. Non-targetting control was designed to target no known human genes, and provide a baseline response for cellular

exposure to siRNAs (Dharmacon, D-001210-01-D001210-05). RNA was harvested after both 48 and 72 hours using the Quick-RNA™ MiniPrep Kit (Zymo Research, Catalog number R1055). Total RNA was converted to cDNA using the TaqMan Reverse Transcription kit. Gene expression was assessed using real-time quantitative PCR with custom primers specific to *HMGA1-lnc* generated by Thermo Fisher, as well as established primers for the *18S* ribosomal RNA (endogenous control), *GAPDH*, and *HMGA1*. RT-qPCR reactions were performed in triplicate and relative expression was determined using the $2^{-\Delta\Delta Ct}$ method.

Table 3.1 SiRNA sequences used to target *HMGA1-lnc*

Our ID	ID	Weight	Sequence
H1	ROWDG_000007	13400.9	GAGAGAAGACAGAGAGAAAUU
H2	ROWDG_000009	13370.9	CAACAAAGGCAUUAAGAAAUU
H3	ROWDG_000011	13370.9	GAGAAAUAUGUGAAGGAUAUU
H4	ROWDG_000013	13461.0	GGAAGGAGCAGGAGGAGGAUU

Table 3.2 qRT-PCR probes used in Chapter 3

Gene	ID
18S	Hs99999901_s1
GAPDH	Hs03929097_g1
HMGA1-lnc	Hs03921739_s1
HMGA1	Hs00852949_g1

3.3 Results

3.3.1 *cis*-acting long non-coding RNAs are deregulated in LUAD

We first sought to examine lncRNA expression in LUAD at a global level to identify potentially biologically relevant *cis*-acting lncRNAs. LncRNAs that were (i) expressed in both datasets, (ii) significantly deregulated in both datasets (Fold Change, FC: 1.5), and (iii) were overlapping or closely neighbouring (within 1.5 Kb) protein coding genes were considered for further analyses. Using these parameters we found 84 lncRNAs overlapping protein-coding genes that were significantly overexpressed in tumours relative to matched non-malignant samples. Similarly we found 324 lncRNAs overlapping protein-coding genes that were significantly downregulated at least 1.5-fold in tumours (Table A1, found in the Appendix). We were interested whether the neighbouring protein coding genes were also deregulated in LUAD, and if they were deregulated in the same direction. We observed that the substantial proportion of lncRNA:protein-coding gene expression relationships were deregulated in the same direction (concordant; 79%), while only 21% displayed inverse deregulation patterns (discordant).

A number of these deregulated lncRNAs overlap protein-coding genes that have been previously described to be involved in cancer biology (Table 3.3). For example, lncRNA *OIP5-AS1* is downregulated in both datasets, and overlaps the protein-coding gene *OIP5*. *OIP5* has been described to play a role in tumour progression and metastatic growth in multiple cancer types, and its elevated expression is associated with poor survival for lung cancer patients [108-110]. To identify whether the expression of the deregulated lncRNAs of interest was associated with the expression of their neighbouring protein-coding genes, we compared groups of tumours with high levels of lncRNA expression to groups of tumours with low lncRNA expression levels. As concordant

expression relationships may be more affected by transcriptional noise as a product of genetic "passenger effects", such as non-specific DNA copy number changes affecting multiple neighbouring genes, we focused on those with discordant relationships. Specifically, we decided to focus on a particular deregulated lncRNA, *RP11.513I15.6,* because of: (i) proximity to the known oncogene *HMGA1*, and (ii) the discordant expression relationship between the lncRNA and neighbouring protein-coding gene. For simplicity we refer to this lncRNA as *HMGA1-lnc*.

Table 3.3 LncRNAs deregulated in LUAD with cancer-associated neighbouring genes

lncRNA	Deregulation	Median FC (BCCA)	cis-gene	Deregulation	Median FC (BCCA)	cis-gene cancer literature	PMID ID
RN7SL5P	Upregulated	7.83	PTPRD	Downregulated	0.002	Lung Cancer	22245727
RP11-334E6.3	Upregulated	7.62	USP2	Downregulated	0.0291	Cancer	25687182
KRT18P12	Upregulated	86.68	PTPN14	Downregulated	0.208	Cancer	29017057
XXbac-BPG254F23.6	Upregulated	3.57E+04	HLA-DQB1	Downregulated	0.493	Lung Cancer	31114327
CTD-2033A16.3	Upregulated	3.09E+08	NQO1	Upregulated	2.54	Lung Cancer	30954648
RP11-276H19.1	Downregulated	0.051	GAS1	Downregulated	0.701	Cancer	26161998
AC012594.1	Downregulated	0.473	MYO3B	Upregulated	7.49	Cancer	25500906
RP11-122M14.1	Downregulated	0.127	NEK2	Upregulated	16.8	Lung Cancer	25202351
OIP5-AS1	Downregulated	0.217	OIP5	Upregulated	1.25	Lung Cancer	22129094
RP11-513I15.6	Downregulated	0.234	HMGA1	Upregulated	1.76	Lung Cancer	25344216

3.3.2 Expression of HMGA1-lnc and HMGA1 are deregulated in lung adenocarcinoma

We observed *HMGA1-lnc* to be significantly downregulated in tumours when compared to adjacent non-malignant tissues, which holds true in both datasets (Figure 3.1a). In contrast, the neighbouring protein-coding gene *HMGA1* was found to be significantly overexpressed in both tumour datasets relative to matched non-malignant tissue (Figure 3.1b). To highlight this difference in expression, we compared the levels of *HMGA1* between tertiles of tumours with the highest and lowest expression of *HMGA1-lnc*. Interestingly we found that *HMGA1* and *HMGA1-lnc* were negatively correlated (p=0.0153), and levels of *HMGA1* were significantly greater (p=0.0326) in the low lncRNA expressing tumours (Figure 3.1c,d). Thus, the expression of these two genes appears to be anti-correlated, possibly indicating that this lncRNA is involved in the inhibition of *HMGA1* expression in normal lung contexts.

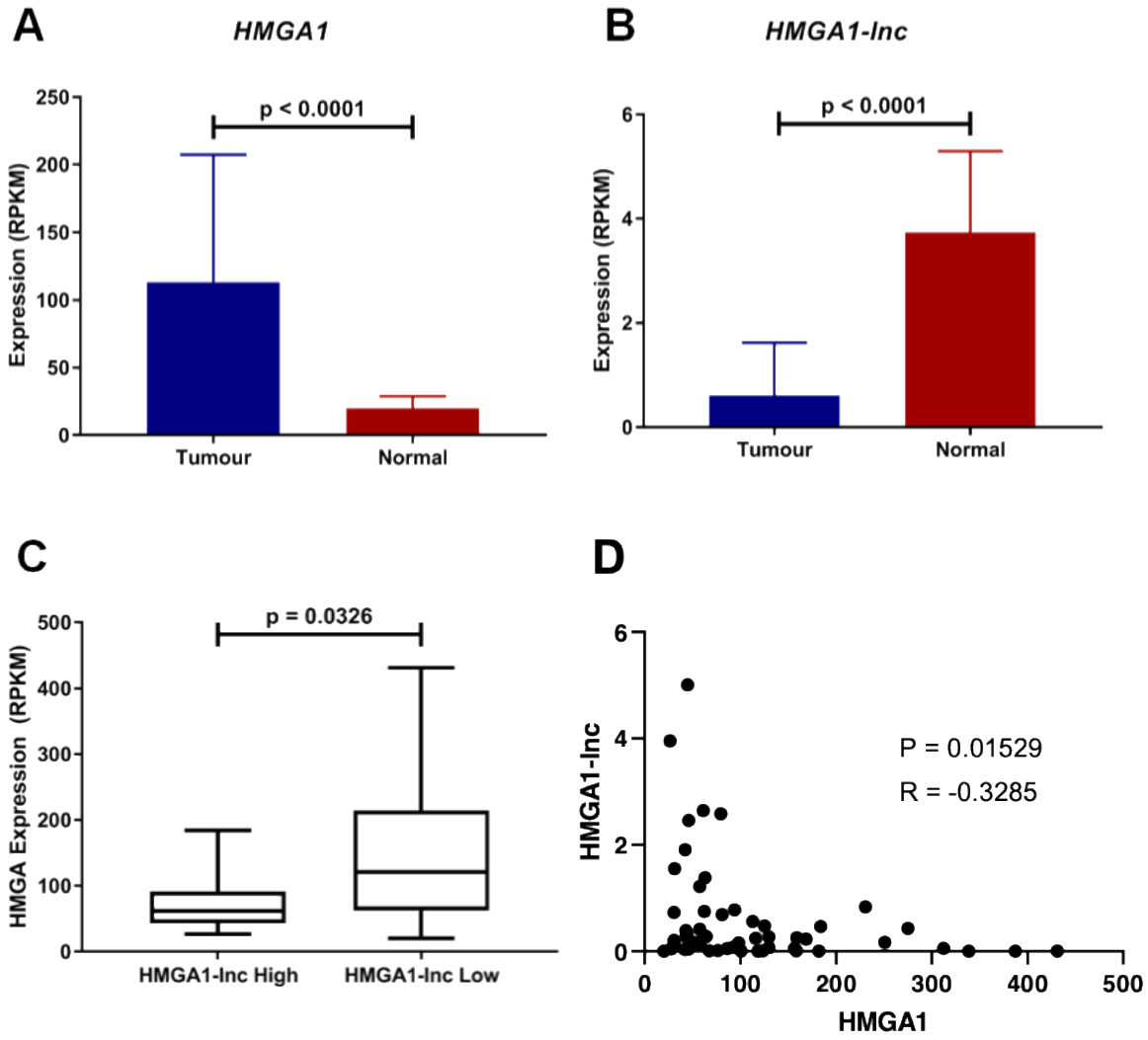


Figure 3.1 Expression of *HMGA1* and *HMGA1-lnc* in lung adenocarcinoma

Expression of *HMGA* is upregulated in LUAD compared to adjacent non-malignant tissue (p value) (A) while conversely, expression of *HMGA1-lnc* is downregulated in tumours (B). Additionally, tumours with high levels of *HMGA1-lnc*, have significantly lower levels of *HMGA1*, when compared to tumours with low levels of the lncRNA (C), and expression of HMGA1 and HMGA1-lnc are negatively correlated (D).

As expression of *HMGA1* has been previously described to increase with tumour stage and cancer aggressiveness, we examined whether expression of *HMGA1-lnc* was inversely associated with stage [104]. As the majority of our tumour samples were Stage I and II we performed a Mann-Whitney U-test between these two groups to identify significant associations (Figure 3.2). Interestingly, while *HMGA1* was associated with increased tumour stage (p=0.0011), the opposite was true for *HMGA1-lnc*, where expression of the lncRNA is significantly decreased in more advanced tumours (p=0.0125).

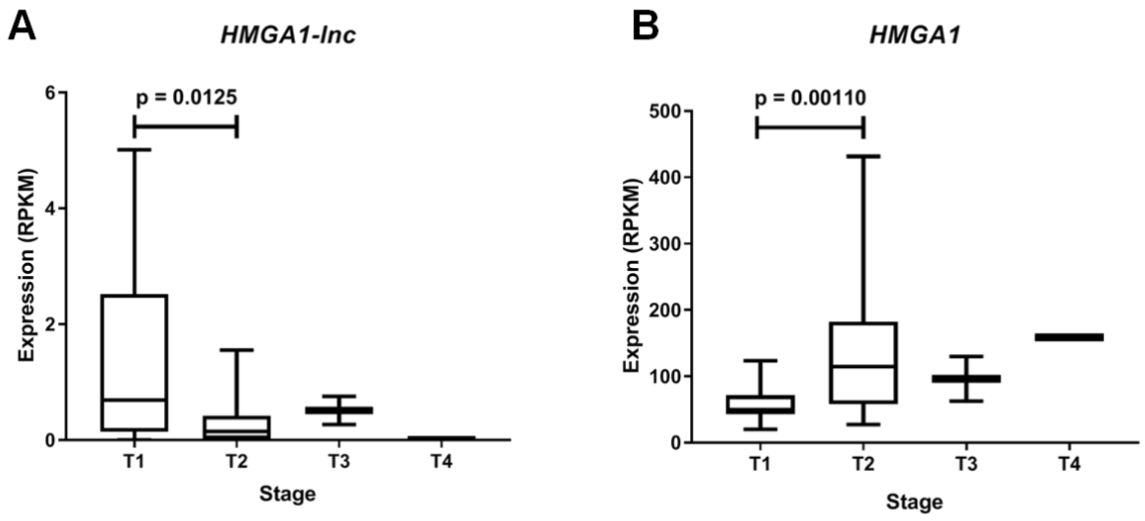


Figure 3.2 Expression of *HMGA1* and *HMGA1-lnc* is associated with tumour stage

Expression of *HMGA1-lnc* decreases with increasing tumour stage (A), whereas *HMGA1* expression increases with higher grade tumours (B).

3.3.3 *HMGA1-lnc* controls *HMGA1* expression

To determine whether lost *HMGA1-lnc* expression is a mechanism of *HMGA1* overexpression in the lung, we performed a siRNA-mediated knockdown of *HMGA1-lnc* in a non-malignant lung epithelial cell line (BEAS-2Bs) using a pool of siRNAs specific to *HMGA-lnc*. We then quantified expression changes using qRT-PCR as described in Methods, Chapter 2. From this, we observed a (3.42 fold reduction of the lncRNA) after (48 hours). Strikingly, in the cells with reduced *HMGA1-lnc* the mRNA expression levels of *HMGA1* were increased by 1.57 fold compared with cells transfected with non-targeting control siRNAs (Figure 3.3). This observed increase in HMGA1 levels in the lncRNA-inhibited cell lines, suggests that *HMGA1-lnc* acts to inhibit the expression of *HMGA1*, and that downregulation of this lncRNA in LUAD is a mechanism for the overexpression of this well known cancer-driving gene.

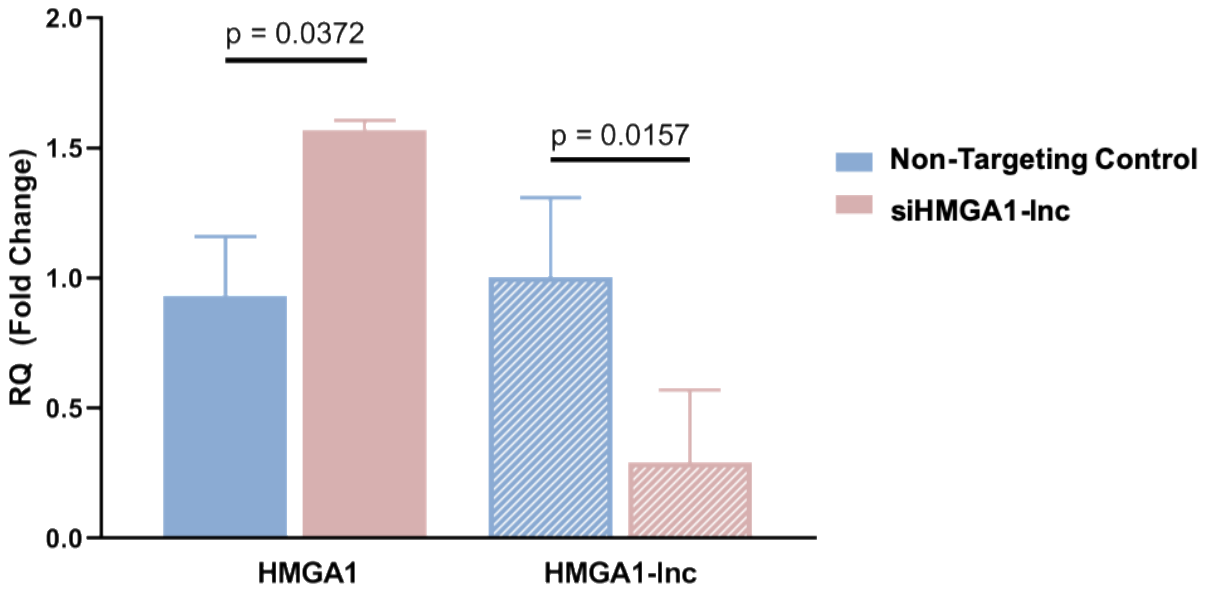


Figure 3.3 Inhibition of HMGA1-lnc results in increases of HMGA1 expression

SiRNA mediated inhibiton of *HMGA1-lnc* was performed in normal bronchial epithelial cells and resulted in significant reduction of the lncRNA conversely in cells where the lncRNA was inhibited significant increases in protein coding *HMGA1* expression was observed.

3.4 Discussion

The role of protein-coding genes in the onset and progression of LUAD is well-established; however, there remains a lack of treatment options for patients who do not harbor one of the few clinically-actionable driver-gene alterations. LncRNAs have been shown to have important roles in the regulation of cancer-associated genes, but complex folding patterns and unknown binding motifs make lncRNAs particularly difficult to functionally characterize. Here, we used an approach that considered the genomic location, as well as the known function of neighbouring protein coding genes, in order to identify candidate target genes of *cis*-acting lncRNAs deregulated in cancer.

Using this approach, we identified 408 lncRNAs that are deregulated in two datasets of LUAD, which closely neighbour or overlap protein-coding genes. Further, many of the protein-coding neighbours of the deregulated lncRNAs are similarly deregulated in both cohorts. Many of these neighbouring protein-coding genes have been previously described in cancer, including well-known lung-cancer-associated genes such as *NEK2*, suggesting that there may be a selective pressure for the deregulation of these lncRNAs in order to release these cancer-promoting genes from negative regulation. Thus, alterations in lncRNA expression may consequently disrupt coding-gene expression as a means of promoting tumour development.

The majority of the deregulated lncRNAs were positively correlated with the expression of their protein-coding neighbours. This may imply that this concordant regulation is favoured in LUAD, and previous studies have shown that many *cis*-acting lncRNAs have positive expression relationships with their neighbours in several tissue types [111]. However, genes in the same vicinity are often subject to regulation that can affect whole genomic regions, such as silencing through chromatin condensation. In particular, tumours often have significantly elevated levels of these

broad genomic alterations to the DNA, which enables tumour-suppressor-gene silencing or oncogene activation. Genes neighbouring these oncogenes and tumour suppressors are often caught in these regions of alteration, and display concordant expression with these genes, a phenomenon known as the passenger effect [112] [113]. For example, frequent DNA amplification of the *MET* oncogene occurs in 5-20% of LUAD, leading the surrounding genes to display significantly increased DNA copy number [114,115]. While it is difficult to separate *cis*-acting concordant regulatory relationships from oncogene passengers without further verifying direct interactions via *in vitro* expression modulation, genes displaying discordant expression relationships with their neighbours are less susceptible to this effect. Thus, we decided to focus on a deregulated lncRNA displaying a discordant expression pattern with its neighbouring oncogene, *HMGA1-lnc*.

We found *HMGA1-lnc* to be significantly downregulated in LUAD, where its expression levels is decreased 18 fold in tumours, compared to *HMGA1* which has expression levels 5 fold greater in tumours (TCGA). These observations in tandem with anti-correlated expression relationships within tumour samples led to our hypothesis that *HMGA1-lnc* acts to repress *HMGA1* expression in non-malignant samples. Consequently, the finding that *HMGA1-lnc* was downregulated with increasing stage, while *HMGA1* expression increased with more advanced stages strengthened this putative regulatory relationship. To verify that *HMGA1-lnc* was able to affect *HMGA1* expression levels, we performed siRNA-mediated knockdown of the lncRNA in cells derived from normal lung epithelium (BEAS-2B cells). When the lncRNA was inhibited we noted significant increases in *HMGA1* mRNA and protein levels, confirming that this lncRNA directly regulates the expression of *HMGA1 in vitro*. Previous studies modulating expression levels of *HMGA1* in these same normal lung epithelial cells (BEAS-2B) have shown that increased

HMGA1 expression leads to transformed phenotypes and increases in anchorage-independent cell growth [116]. These results suggest that downregulation of this previously-uncharacterized lncRNA may lead to *HMGA1* upregulation, potentially driving the onset of these same cancer phenotypes in normal human lung epithelial cells.

Chapter 4: Aberrant Expression of Pseudogene-derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma.

4.1 Introduction

While lncRNAs have been observed to be important in cancer biology, functional prediction of newly-discovered lncRNAs remains a major challenge. In Chapter 3 we focused on one approach to identify lncRNAs that may act *in cis* to drive the many aspects of LUAD. In this chapter we focus on lncRNAs that may be acting *in trans* to regulate cancer driving genes in LUAD. Specifically, how many lncRNAs expressed from pseudogene loci have been shown to regulate genes with which they have sequence homology.

Pseudogenes are DNA sequences that are defunct relatives of functional protein-coding genes (herein referred to as parent genes) and arise during either gene duplication events, or the reverse transcription of an mRNA transcript into a new genomic location. Through evolution these duplicated genes have acquired mutations such as premature stop codons and frameshifts, which results in the loss of protein coding ability, while still retaining a high degree of sequence homology with the original parent gene [117]. Recently, pseudogene-derived lncRNAs have been shown to regulate their parent genes and this novel mechanism has been observed in many tumour types, including lung cancer [118,119]. A prominent example is the tumour suppressor gene *PTEN* (chromosome 10), regulated both positively by *PTENP1* (chromosome 9), a lncRNA transcribed from the sense strand of the pseudogene locus, and negatively by the lncRNA *PTENP1-AS1*, which is transcribed from the strand antisense to the parent gene [76].
Pseudogenes have been continually omitted from large RNA-sequencing datasets due to the complexity of separating highly-similar pseudogene sequences from parent genes. However,

Millegan et al. recently generated an atlas of lncRNAs overlapping pseudogenes, which has provided a foundation for their analysis in RNA sequencing datasets [120]. We hypothesize that the functions of pseudogene-derived lncRNAs are an under-explored mechanism of gene regulation that occurs more broadly than previously realized, and that these events contribute to the tumourigenesis of LUAD. We performed next generation RNA-sequencing on microdissected LUAD tumours and matched non-malignant tissue to identify deregulated lncRNAs expressed from pseudogene loci (herein referred to as Ψ-lncs). We then explored the relationships of these Ψ-lncs with their parent genes, and explored their significance in relation to patient clinical features in our discovery dataset as well as a validation dataset.

4.2 Methods

4.2.1 Identification of long non-coding RNAs expressed from pseudogene loci and corresponding parent genes

Ψ-lnc annotation: Millegan et al. recently published a global atlas of lncRNAs that have exonic overlap with positionally non-redundant (unique) pseudogenes from 3 major pseudogene databases [120]. Using this resource we obtained a list of lncRNAs overlapping pseudogene loci (>1% intronic or exonic overlap) that we used as a foundation for our expression analysis (Supplemental table 1, Appendix A). Supplemental tables can be accessed in our published manuscript "Aberrant Expression of Pseudogene-Derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma," due to the large size of the supplementary tables in Chapter 4 (DOI: 10.3389/fgene.2019.00138) [107]. Descriptions of all supplemental tables can be found in Appendix A. As the degree of sequence overlap required for a pseudogene-derived lncRNA to regulate its parent gene is unknown, we did not restrict our

analysis to full-length, expressed pseudogenes, and included lncRNAs with any exonic overlap, including sense and antisense transcripts in order to annotate the most comprehensive list of Ψ-lncs (lncRNAs overlapping pseudogene loci).

Parent gene annotation: The parent gene information was also extracted for all pseudogenes overlapping our list of Ψ-lncs that have parent-gene annotations in the YaleHuman60 and Retroali5 databases (Supplemental Table 7, Appendix A). Manual literature search was performed for parent genes of deregulated Ψ-lncs that were not contained in these databases.

4.2.2 Statistical analysis

Identification of significantly deregulated Ψ-lncs in paired LUAD and non-malignant lung tissue: Gene expression for protein-coding and non-coding genes was compared between tumours and non-malignant tissue and significantly deregulated genes were identified using a Wilcoxon signed-rank test ($p<0.05$) and subjected to a Benjamini-Hochberg (BH) FDR correction. Ψ-lncs as identified previously were extracted, and those that were significantly deregulated between tumours and non-malignant tissue, in both our discovery (BCCA) and validation (TCGA) were selected for further analysis (n=104 deregulated lncRNAs) (Figure 4.1b).

Ψ-lncs and parent gene expression: Tumours were sorted by Ψ-lnc expression for each Ψ-lnc-parent gene pair, and grouped into top and bottom Ψ-lnc expressing tertiles. Parent gene expression was then compared between the two groups using the Mann Whitney U-test (p-value ≤ 0.05). We performed a global expression analysis to determine whether Ψ-lnc-parent pairs were more positive or negatively correlated than random chance. For all Ψ-lncs with expression data in both datasets, Spearman's correlation rho values were calculated for Ψ-lnc-parent gene pairs (n=390) and compared to Ψ-lnc-random gene pairs. Random genes in our expression

matrices were selected to pair with each Ψ-lnc. Each gene was assigned a number and pairs were chosen by using a random number generator (https://www.random.org/). Spearman's rho values were then plotted along a standard curve, and compared (Mann Whitney U-test, p-value ≤ 0.05). Rho value distribution was also compared between sense lnc-parent gene pairs (n=208) and antisense lnc-parent pairs (n=182).

Literature searches: To determine if each gene of interest (Ψ-lnc or parent gene) had been previously described in the context of tumours we searched Pubmed using the terms "gene + cancer" or "gene + lung cancer".

Hierarchical clustering and data visualization: Unsupervised hierarchical clustering was performed in order to visualize and examine the expression of the 104 deregulated Ψ-lncs in individual samples (Figure 4.2a). Average Linkage was used as a cluster distance metric, while Pearson Correlation was used as a point distance metric. To visualize the expression patterns of the most highly expressed Ψ-lncs, those with an average expression value of ≥ 10 RPKM in either tumour or non-malignant samples were included in the analysis.

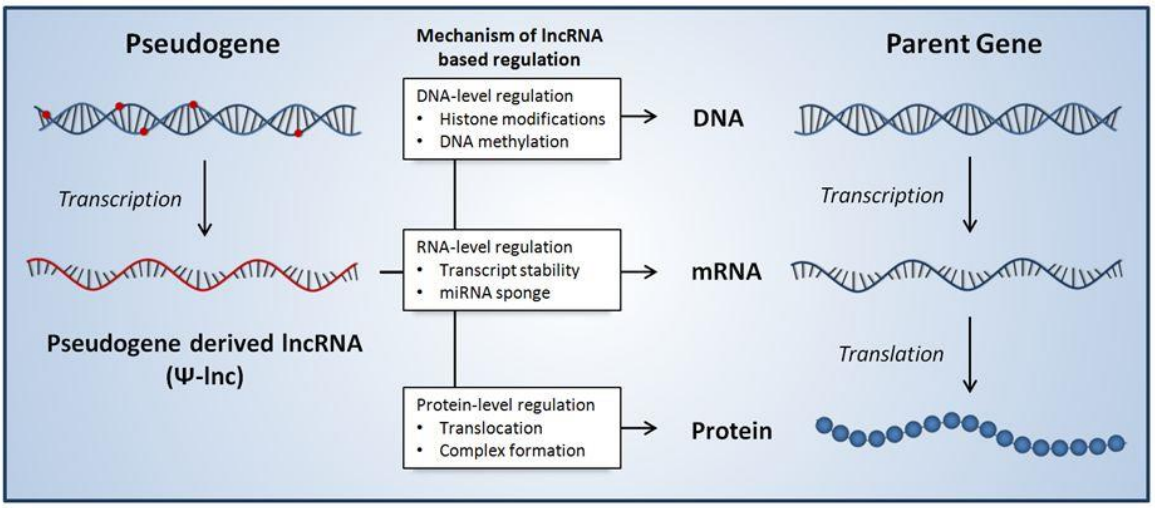
Distribution of deregulated Ψ-lncs across genome: Locations of Ψ-lncs and parent genes were compared to identify their genomic position (Supplemental Table 4, Appendix A). Circular plot visualization was performed using the R-package Circlize (Figure 4.3) [121]. LUAD-specific regions of significant recurrent somatic copy number alterations had been previously identified by TCGA, and were used in this study to determine if the deregulated Ψ-lncs overlapped with frequently altered regions [115]. All genomic coordinates correspond to the NCBI GRCh37 reference human genome build.

4.2.3 Clinical features

Survival analysis: A large public clinical database (Kaplan–Meier Plotter; http://kmplot.com/analysis/) comprised of 719 LUAD samples was used to determine the association between both protein-coding and non-coding gene expression with patient outcomes. Similar to the BCCA and TCGA cohorts, the patient samples in this 3rd dataset were mostly comprised of Stage I and Stage II tumours (Table 2.1). Of the 104 deregulated Ψ-lncs, 19 were represented in this database, while 70% of parent genes (72 out of 103) were present. Default settings were used and a log-rank (Mantel–Cox) test was applied to compare survival between groups of tumours with high and low expression of each gene tested, where $p < 0.05$ was considered statistically significant. The optimal expression cut-off was selected for each gene.

Association with tumour stage: The majority of tumours for BCCA and TCGA fell into the categories of Stage I and Stage II (Table 2.1). We compared expression of deregulated Ψ-lncs between Stage I tumours and tumours classified as Stage II and above using a Mann Whitney U-test (p-value ≤ 0.05).

A


Pseudogene
Mechanism of lncRNA based regulation
Parent Gene
DNA-level regulation
• Histone modifications
• DNA methylation
DNA
Transcription
Transcription
RNA-level regulation
• Transcript stability
• miRNA sponge
mRNA
Pseudogene derived lncRNA (Ψ-lnc)
Translation
Protein-level regulation
• Translocation
• Complex formation
Protein

B

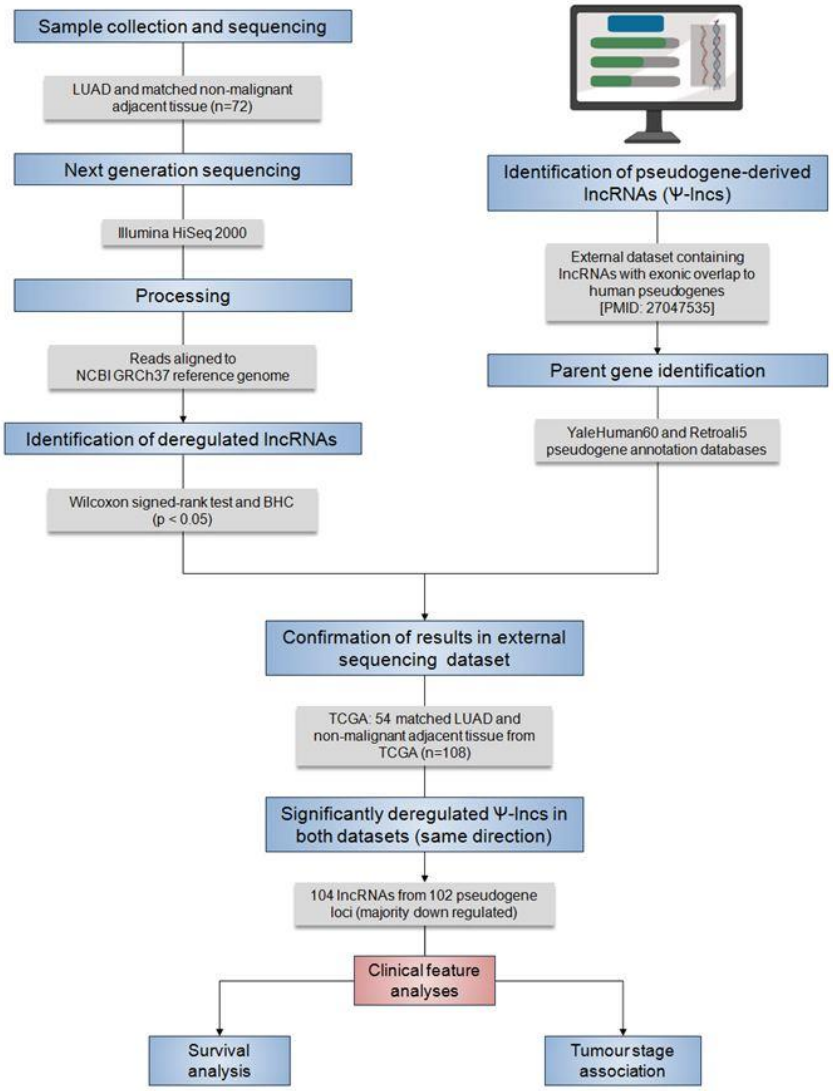

Sample collection and sequencing
LUAD and matched non-malignant adjacent tissue (n=72)
Next generation sequencing
Illumina HiSeq 2000
Processing
Reads aligned to NCBI GRCh37 reference genome
Identification of deregulated lncRNAs
Wilcoxon signed-rank test and BHC (p < 0.05)
Identification of pseudogene-derived lncRNAs (Ψ-lncs)
External dataset containing lncRNAs with exonic overlap to human pseudogenes [PMID: 27047535]
Parent gene identification
YaleHuman60 and Retroali5 pseudogene annotation databases
Confirmation of results in external sequencing dataset
TCGA: 54 matched LUAD and non-malignant adjacent tissue from TCGA (n=108)
Significantly deregulated Ψ-lncs in both datasets (same direction)
104 lncRNAs from 102 pseudogene loci (majority down regulated)
Clinical feature analyses
Survival analysis
Tumour stage association

Figure 4.1 Summary of the regulatory mechanisms of Ψ-lncRNAs and the analysis pipeline for the identification of their deregulation in lung adenocarcinoma.

Summary of the regulatory mechanisms of pseudogene-derived lncRNAs (Ψ-lncs) that retain sequence homology with the parent gene. Overall, lncRNAs have been shown to function through a variety of regulatory mechanisms, acting on the DNA, RNA, and protein levels **(A)**. Flow diagram description of the analysis pipeline applied for the identification of deregulated Ψ-lncs. Patient LUAD samples were collected and subjected to next generation sequencing to quantify RNA expression. Gene expression was then compared between tumours and matched non-malignant tissue to identify significantly deregulated transcripts. LncRNAs with exonic overlap to known pseudogenes were then identified and confirmed in a 2nd set of LUAD and matched non-malignant tissue. This lead to the identification of 104 deregulated Ψ-lncs, which were then assayed to determine associations with clinical features **(B)**.

4.3 Results

4.3.1 Ψ-lncRNA expression is deregulated in lung adenocarcinoma

Pseudogenes vary widely in terms of length, gene fraction, and identity to parent genes, and can be expressed as lncRNAs that are sense, antisense, partial overlapping, or internal to the parent. In light of this variation, Ψ-lncs are observed to have vastly different regulatory effects on downstream target genes (Figure 4.1). In our curation of Ψ-lncs in LUAD, we have included those that have exonic overlap with a pseudogene (partial or full length) and considered both sense and antisense transcripts (Supplemental Table 1, Appendix A). Ψ-lncs were analyzed in an in-house discovery (BCCA, n=72) and external validation (TCGA, n=108) cohort of LUAD and paired non-malignant lung tissues (Table 2.1).

We identified aberrantly expressed Ψ-lncs that are significantly deregulated in both the discovery and validation datasets with the same direction of expression alteration (Ψ-lncs upregulated or downregulated in tumours compared to matched non-malignant tissue).

We found 104 lncRNAs expressed from 102 pseudogene loci to be significantly deregulated in LUAD (Supplemental Table 2, Appendix A). To our surprise, we found that the majority of these deregulated Ψ-lncs were downregulated in tumours (Figure 4.2a and b). Most of these were unannotated lncRNAs, such as *RP11-1007O24.3*, which was downregulated in tumours, with only 24 of the total deregulated Ψ-lncs having been previously described in scientific literature annotated in PubMed, albeit none in the field of pseudogene-mediated deregulation (Supplemental Table 3, Appendix A / Figure 4.2b). Twenty of these 24 have been described in the context of cancer, with only four in lung cancer. This includes *DGCR5*, a lncRNA we found to be overexpressed in tumours. *DGCR5* has been reported to promote LUAD

progression by sequestering a variety of miRNAs involved in cell cycle regulation, although it has not been investigated with regard to its pseudogene-derived nature [122-124].

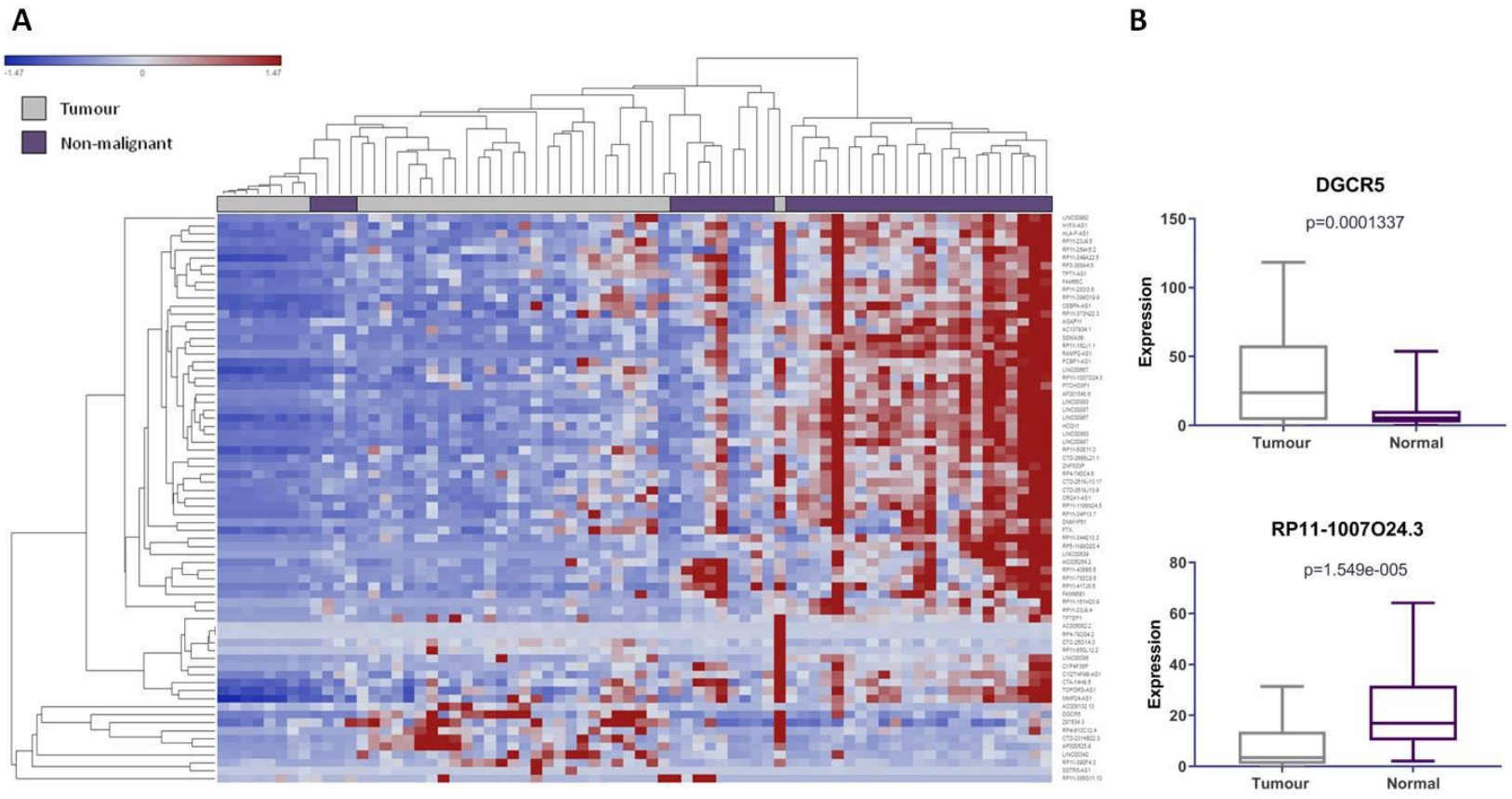

A
-1.47
0
1.47
Tumour
Non-malignant
B
DGCR5
p=0.0001337
Expression
150
100
50
0
Tumour
Normal
RP11-1007O24.3
p=1.549e-005
Expression
80
60
40
20
0
Tumour
Normal

Figure 4.2 LncRNAs derived from pseudogene loci are significantly differentially expressed in lung adenocarcinoma compared to matched non-malignant lung tissue. Unsupervised hierarchical clustering of pseudogene-derived lncRNAs differentially expressed between lung adenocarcinoma (grey) and matched non-malignant tissue (purple). Average linkage was used as the cluster distance metric and Pearson Correlation was used as the point correlation metric. Expression values are stratified from low (blue) to high (red). Only pseudogene-derived lncRNAs with average expression values of greater than or equal to 10 RPKM were included in the clustering analysis. Clustering of samples highlights relative similarity in pseudogene-derived lncRNA expression between the two sample groups, while clustering of gene expression reveals a trend towards the widespread underexpression of these transcripts in lung adenocarcinoma **(A)**. Highlighted examples of pseudogene-derived lncRNAs significantly deregulated between lung tumours and non-malignant tissues. Expression (RPKM) in tumours (grey) and normal tissues (purple) is represented on the Y-axis. Boxes represent the interquartile range and inner lines represent the median expression value **(B)**.

We were interested in examining the genetic events that could impact pseudogene loci, and thus affect Ψ-lnc expression. We mapped the chromosomal distribution of the deregulated Ψ-lncs, finding them to be distributed throughout the genome and detected on most chromosomes, except for chromosomes 4 and Y (Figure 4.3). The locations of each of the parent genes of deregulated Ψ-lncs are similarly distributed through the genome (Supplemental Table 4, Appendix A). We then determined the overlap of these genes with regions of recurrent chromosomal amplification and deletion as determined by The Cancer Genome Atlas (TCGA) for LUAD [115]. While some Ψ-lncs overlap with regions of recurrent deletion, the majority do not, indicating that they may be regulated by mechanisms other than copy number alteration (Figure 4.3).

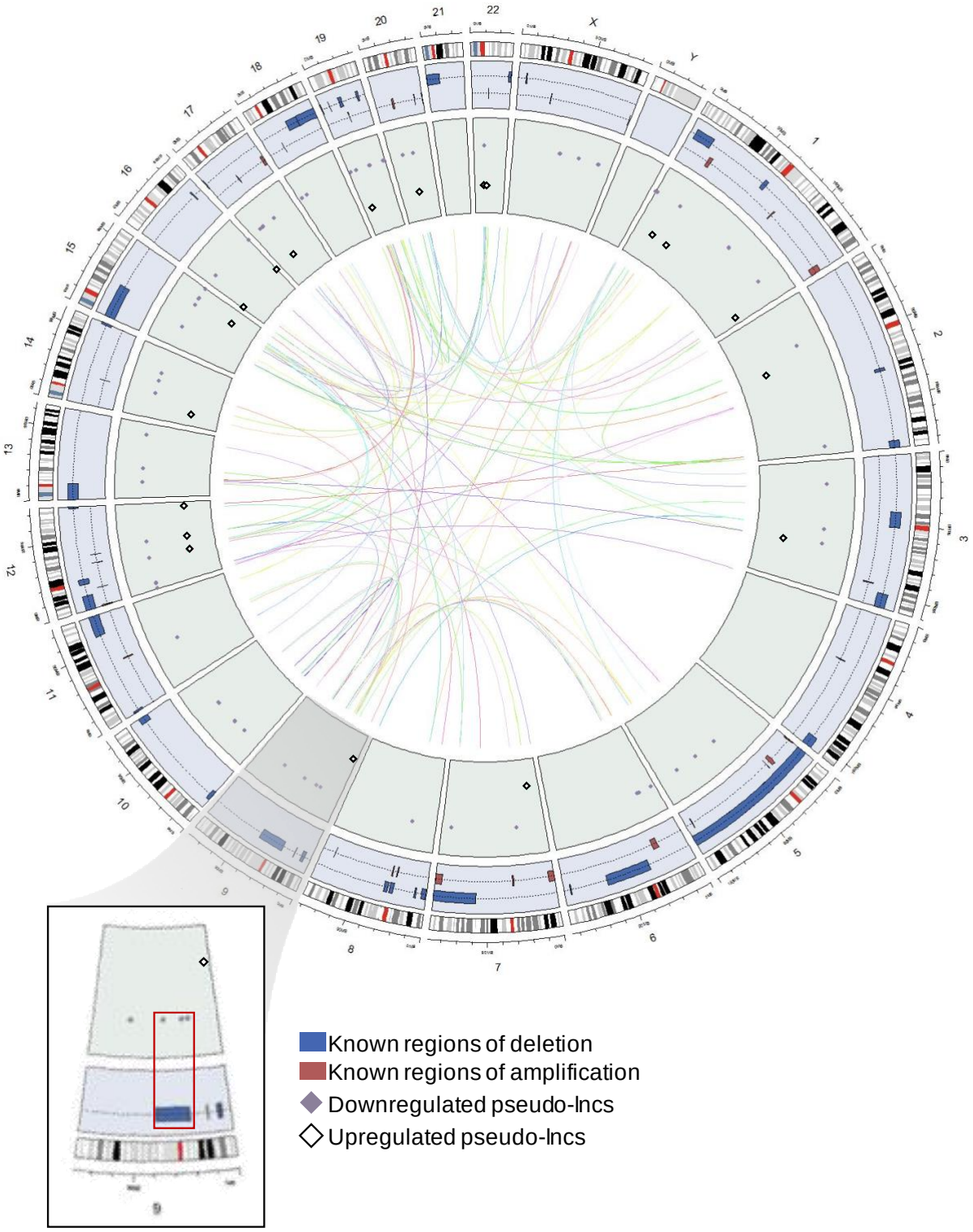

Known regions of deletion
Known regions of amplification
Downregulated pseudo-lncs
Upregulated pseudo-lncs

Figure 4.3 Genome-wide distribution of deregulated pseudogene-derived lncRNAs in lung adenocarcinoma.

Circular representation of the genomic distribution of the deregulated pseudogene-derived lncRNAs discovered in our study, as well as known regions of copy number alterations in lung adenocarcinoma as described by TCGA. The outer concentric circle represents the human karyotype from the genomic build hg19. The blue concentric circle contains known regions of copy number amplification (red boxes) and deletion (blue boxes) that have been previously published. The inner green circle represents the specific genomic location of our pseudogene-derived lncRNAs found to be either upregulated (green circles) or downregulated (purple circles) in lung adenocarcinoma. Finally, the inner connecting lines represent the interaction between the deregulated pseudogene-derived lncRNAs and the locations of their respective protein-coding parent genes. Chromosome 9 (magnified region) highlights that some of the downregulated pseudogene-derived lncRNAs overlap with genomic regions frequently deleted in lung adenocarcinoma.

4.3.2 Global patterns of Ψ-lncRNA and parental gene expression

As a first step to identify deregulated Ψ-lncs that may function through regulation of their respective parent genes, we explored whether Ψ-lncs with significantly deregulated expression were associated with altered parent gene transcript levels [76,119,125,126]. We obtained parent gene information for the 95 deregulated Ψ-lncs and determined that they shared 104 parent genes. Some pseudogenes contained multiple lncRNAs, and some lncRNAs overlapped multiple pseudogenes, constituting a total number of 116 Ψ-lnc-parent gene pairings. For each Ψ-lnc-parent pair we compared groups of tumours with high levels of Ψ-lnc expression to those with low levels of Ψ-lnc expression. We found that 33 Ψ-lncs have a significant expression relationship with their parent gene in at least 1 dataset (Supplemental Table 5, Appendix A). This included 21 sense Ψ-lnc-parent-gene pairs, and 13 antisense Ψ-lnc-parent gene pairs. Having identified Ψ-lncs with expression associated with parent gene expression, we investigated whether parent genes had known oncogenic or tumour suppressive roles. We performed a literature search to determine if any had been previously described in the context of cancer. Interestingly, we found that 65 of these parent genes had been previously described in cancer, and of those, 33 had been described in lung cancer (Table 4.1). Of the 34 significantly differentially expressed Ψ-lnc-parent gene pairs, 25 parent genes were described in cancer. This includes lung cancer associated genes like *CS*, which affects tumour drug response, as well as *RCN1*, which is associated with poor prognosis and tumour progression in lung cancer [127,128]. As the vast majority of the deregulated Ψ-lncs that were correlated with their parent gene had positive associations, we were interested whether this was a global phenomenon or exclusive to

Table 4.1. Parent genes of deregulated Ψ-lncRNA previously described in cancer literature

Parent Gene	Cancer [a]	Lung a Cancer	Ψ-lnc
ADC	8	-	*RP11-439L8.3*
AGAP1	12	-	*AGAP11*
ARIH1	4	-	*RP11-1007O24.3*
ATXN7L3	5	-	*RP11-56G10.2*
BCR	>1000	>1000	*AC008132.13*
BMS1	4	-	*AGAP11*
CDC42	1657	129	*RP11-390F4.3*
CECR7	3	-	*AP000525.9;*
			CTD-2314B22.3
CELSR1	11	2	*DGCR5*
CHRNB1	8	-	*RP11-650L12.2*
CIC	-	96	*RP11-34P13.7*
CS	-	10	*LINC00883;*
			RP11-446H18.5
CSPG4	62	5	*DNM1P51*
CTAGE1	49	-	*RP1-122P22.2*
CYP4F2	29	-	*CYP4F35P*
CYP4F3	10	1	*FAM95B1*
CYP4F31P	10	1	*CYP4F35P;*
			FAM95B1
DFFB	33	1	*TOPORS-AS1*
DRD2	121	10	*AP000438.2*
EGLN1	158	5	*RP11-182J1.1*
FAM103A1	1	-	*RP11-324H6.5*
GPR39	8	-	*RP11-399O19.9*
HMGB1	>1000	111	*RP11-349A22.5;*
			ZBED3-AS1
HMGN2P46	81	1	*RAMP2-AS1*
KRT8	94	7	*RP5-1198O20.4*
LINC00657	2	-	*CTA-14H9.5;*
			HCG11
LMNB2	3	1	*RP11-161H23.9*
MARK4	25	3	*CTD-2201G3.1*
MICE	1	-	*HLA-F-AS1*
MIPEPP3	4	1	*C1QTNF9B-AS1*

Parent Gene	Cancer [a]	Lung a Cancer	Ψ-lncRNA
MT1	>1000	95	*MMP24-AS1*
NEBL	17	-	*LINC00342*
NUP210	9	2	*H1FX-AS1*
PCBP2	26	-	*PCBP1-AS1*
PKD1	238	5	*RP11-1186N24.5*
POTEF	4	-	*RP11-193H5.1*
PPARGC1B	37	1	*RP11-527N22.1*
PPY	91	5	*CTD-2008P7.8*
PRKX	12	2	*RP11-526I2.1*
PTCHD3	2	-	*PTCHD3P1*
PTMA	34	-	*LINC00987*
PZP	20	-	*LINC00987*
RAB11FIP1	16	1	*RP3-368A4.5*
RAB40B	5	-	*LL0XNC01-250H12.3*
RCN1	17	1	*TPT1-AS1*
RNASEH1	7	-	*RP11-344E13.3*
RPL21	5	-	*AC005062.2*
RPL23A	8	-	*HLA-F-AS1*
RPSA	51	3	*FTX*
RPSAP58	-	3	*LINC00466*
SEMA3A	163	25	*SEMA3B*
SHQ1	10	1	*RP11-50E11.3*
SNAPC5	4	-	*LY86-AS1*
SNX18	2	-	*RP11-435B5.5*
SRSF9	5	-	*RP11-752G15.3*
TACC3	122	21	*LINC00667*
TOMM40	10	2	*CTD-2314B22.3*
TPTE	69	7	*TPTEP1*
TUBB4B	5	-	*RP11-386G11.10*
TULP3	5	-	*LINC00359*
VENTX	8	1	*RP11-81H3.2*
VWF	733	52	*TPTEP1*
ZNF14	2	-	*CTD-2666L21.1;*
			ZNF833P
ZNF44	1	-	*CTD-2666L21.1*
ZNF584	1	-	*CTD-2619J13.17*

[a] denotes number of entries in PubMed

deregulated genes. We performed a Spearman's correlation analysis on every Ψ-lnc-parent-gene pair with expression data in our dataset irrespective of deregulation status (n=390 gene pairs). We plotted the distribution of Spearman's rho (ρ) values for the Ψ-lnc-parent-gene pairs and compared them to the rho values for Ψ-lncs paired to randomly selected genes. We found that the Ψ-lnc-parent-gene pairs have significantly more positive relationships than the random gene pairs in both the BCCA (Mann Whitney U-test, $p<0.0001$) and TCGA datasets (Mann Whitney U-test, $p<0.0001$) (Figure 4.4a). Studies have shown that lncRNAs transcribed from opposite strands can have different regulatory effects on target genes [76,111]. To determine if transcriptional orientation has an effect on Ψ-lnc-parent relationships we compared the Spearman's rho values of sense Ψ-lnc-parent pairs to antisense Ψ-lnc-parent pairs. In both datasets we observed that the sense Ψ-lnc-parent pairs to have significantly more positive relationships than the antisense Ψ-lnc-parent pairs (Mann Whitney U-test, TCGA set ($p<0.0001$), and BCCA set ($p<0.0025$) (Figure 4.4b and Supplemental Table 6, Appendix A). Strongly positively correlated Ψ-lnc-parent pairs include *TPT1-AS1* / *RCN1* and *LINC00887* / *CS* (Figure 4.4c).

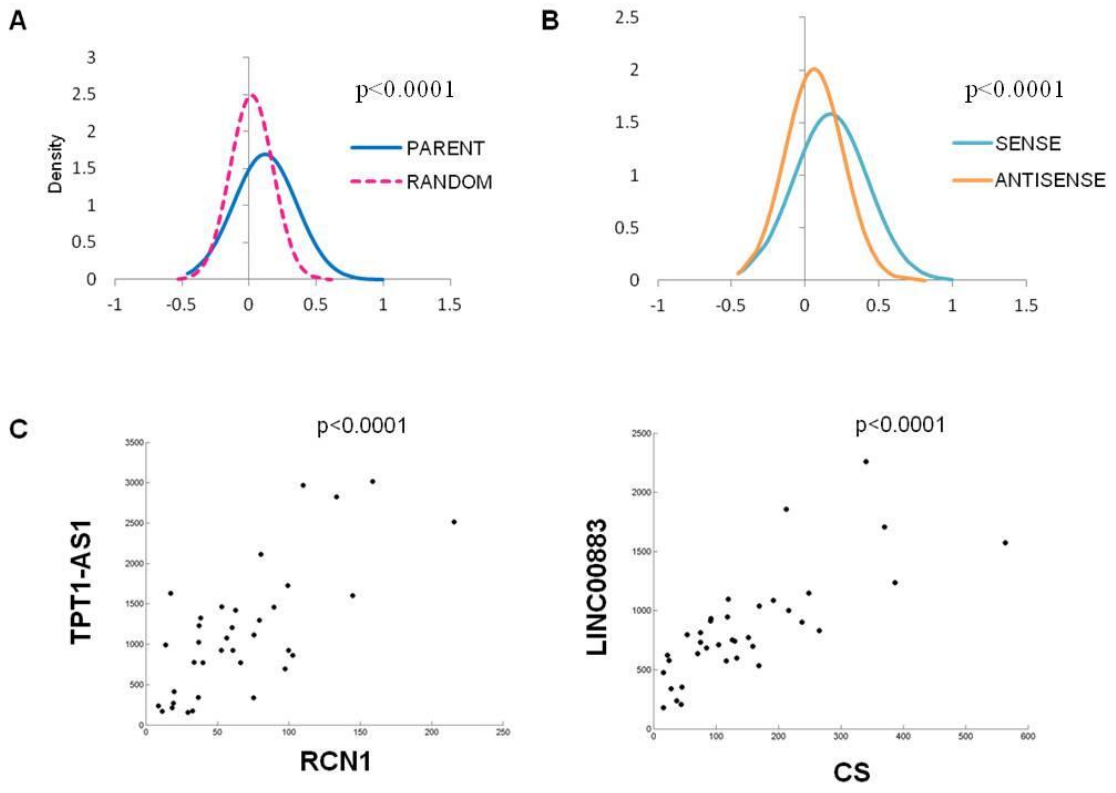


Figure 4.4 Distribution of Spearman's Correlation rho values for all Ψ-lnc-parent-gene pairs in the TCGA dataset (n=391).

Distribution of Spearman's correlation coefficients between Ψ-lnc-parent gene pairs (blue line, median R=0.088) and Ψ-lnc-random gene pairs (pink dashed line, median R=0.019). Rho values of the groups were compared by Mann Whitney U-test **(A)**. Distribution of Spearman's correlation coefficients between sense Ψ-lnc-parent-gene pairs (turquoise line, median R=0.140) and antisense Ψ-lnc-parent-gene pairs (orange line, median R=0.049) (**B)**. Correlation scatter plots of Ψ-lncs with positive expression correlations to cancer-associated parent genes (**C)**.

4.3.3 Ψ-lncRNAs and their parent genes are associated with patient survival

If the aberrant expression of Ψ-lncs is biologically relevant, it follows that they may be relevant in tumour aggressiveness, stage, and patient survival. We performed a two-group analysis using a Mann-Whitney U test between Stage I tumours and Stage II-IV tumours, as the majority of our tumours fell into these categories (Table 2.1). Of the deregulated Ψ-lncs we found *CTC-250I14.3* to be associated with Stage 1 disease and downregulated in both the BCCA and TCGA LUAD cohorts (Figure 4.5). Our discovery datasets were limited in sample size for survival association analysis; therefore, we examined a third cohort of 719 LUAD from the KM Plotter database (Kaplan–Meier Plotter; http://kmplot.com/analysis/) (Table 1.1). This cohort was limited to genes with probe coverage on microarray platforms. A total of 19 of the deregulated Ψ-lncs were represented on this platform, yet, the majority of these Ψ-lncs (16 of 19) were significantly associated with poor overall survival (log-rank $p<0.05$, Table 4.2).

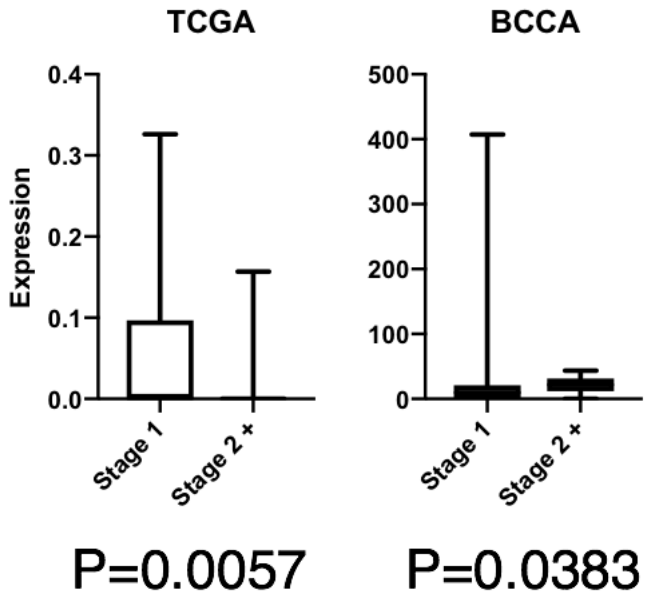


Figure 4.5 Comparison of deregulated Ψ-lnc *CTC-250I14.3* expression between Stage I tumors and tumors classified as Stage II and above using a Mann Whitney U-test (p-value ≤ 0.05).

While we were not able to investigate the survival associations of all deregulated Ψ-lncs as many were not covered by the microarray platforms, the majority (72 out of 103) of their parent genes were represented. We discovered that 67 of these parent genes were associated with patient survival. Twenty-eight of these survival associated parent genes were also significantly associated with the expression of their paired deregulated Ψ-lnc. Furthermore, we found 11 pairs where both Ψ-lnc and parent gene are associated with patient survival. For example *RP11-1007O24.3*, a Ψ-lnc downregulated in tumours, is positively associated with expression of survival-associated parent gene *ARIH1* in both the BCCA and TCGA datasets (Table 4.2, Figure 4.6a).

Table 4.2 Associations between Ψ-lncRNA, parent gene expression, and patient outcome

Ψ-lncRNA	Survival p (KmPlotter)	Parent Gene	Survival p (KmPlotter)	Association with Parent (p, BCCA)	Association with Parent (p, TCGA)
AC008132.13	-	*BCR*	0.00043 [a]	0.0173	-
AGAP11	0.0099 [b]	*AGAP1*	5.80E-10 [b]	0.0005	-
AGAP11	-	*BMS1*	0.00018 [a]	0.0121	-
AP000525.9	-	*CECR7*	3.20E-06 [a]	-	0.0371
CTA-14H9.5	-	*LINC00657*	6.00E-06 [b]	-	-
CTD-2314B22.3	-	*CECR7*	3.20E-06 [a]	-	0.0402
CTD-2619J13.17	-	*ZNF584*	1.70E-08 [a]	-	0.0205
CTD-2666L21.1	-	*ZNF44*	2.00E-15 [b]	-	0.0013
CTD-2666L21.1	-	*ZNF14*	1.90E-12 [a]	-	-
DGCR5	4.00E-04 [a]	*CELSR1*	2.70E-02 [a]	-	-
FAM66C	3.00E-04 [b]	*DEFB130*	-	-	-
FAM95B1	-	*CYP4F3*	2.00E-01 [b]	-	-
HCG11	1.40E-08 [b]	*LINC00657*	6.00E-06 [b]	-	0.0046
LINC00342	1.80E-06 [b]	*NEBL*	2.30E-05 [b]	-	-
LINC00466	4.60E-06 [a]	*RPSAP58*	-	-	-
LINC00639	0.0087 [b]	*ZFP41*	0.014 [b]	-	-
LINC00667	4.50E-08 [b]	*TACC3*	6.20E-09 [a]	-	-
LINC00883	2.90E-03 [b]	*CS*	0.00041 [a]	0.0449	-
LINC00957	0.042 [b]	*RASA4B*	0.035 [a]	-	-
LINC00982	1.30E-06 [b]	*n/a* [c]	-	-	-
LINC00987	-	*PZP*	0.02 [a]	-	<0.0001
LY86-AS1	0.00025 [b]	*SNAPC5*	5e-0.6 [b]	-	-
MMP24-AS1	0.023 [a]	*MT1*	1.20E-12 [a]	-	-
PCBP1-AS1	-	*PCBP2*	4.30E-13 [b]	-	-
RAMP2-AS1	0.014 [b]	*HMGN2P46*	-	-	-
RP11-1007O24.3	-	*ARIH1*	0.00093 [b]	<0.0001	0.0030
RP11-1186N24.5	-	*PKD1*	0.0013 [b]	-	<0.0001
RP11-182J1.1	-	*EGLN1*	1.20E-08 [b]	0.0011	-
RP11-344E13.3	-	*RNASEH1*	1.00E-07 [b]	0.0007	-
RP11-349A22.5	-	*HMGB1*	7.50E-10	-	-
RP11-34P13.7	-	*CIC*	1.20E-08 [a]	-	0.0205
RP11-390F4.3	-	*CDC42*	1.30E-05	0.0449	-
RP11-439L8.3	-	*ADC*	3.30E-09	0.0205	-
RP11-446H18.5	-	*CS*	0.00041 [a]	-	-
RP11-93K22.13	-	*FAM86B3P*	6.30E-05 [b]	-	0.0129
TOPORS-AS1	-	*DFFB*	9.10E-08 [b]	-	0.0033
TPT1-AS1	0.00019 [b]	*RCN1*	0.016 [a]	0.0242	-
ZBED3-AS1	1.10E-10 [b]	*HMGB1*	7.50E-10 [b]	0.0100	-
ZNF833P	-	*ZNF491*	0.016 [b]	-	-
ZNF833P	-	*ZNF14*	1.90E-12 [a]	-	0.0079

[a] Denotes poor overall survival associated with high gene expression

[b] Denotes poor overall survival associated with low gene expression

[c] No information for the parent gene

Further examples also include Ψ-lnc-parent pairs such as *ZBED3-AS1* and *HMGB1*, which are positively correlated at the expression level, and both significantly associated with survival (Figure 4.6b). We also observe Ψ-lnc-parent pairs that are associated with survival, but do not share an expression relationship such as *LINC00667* and parent gene *TACC3* (Figure 4.6c). Collectively, our discovery of the broad deregulation of Ψ-lncs, many of which are survival-associated and associated with parent gene expression, may indicate that Ψ-lncs impact LUAD biology through *trans* regulation of their cancer-associated parent genes.

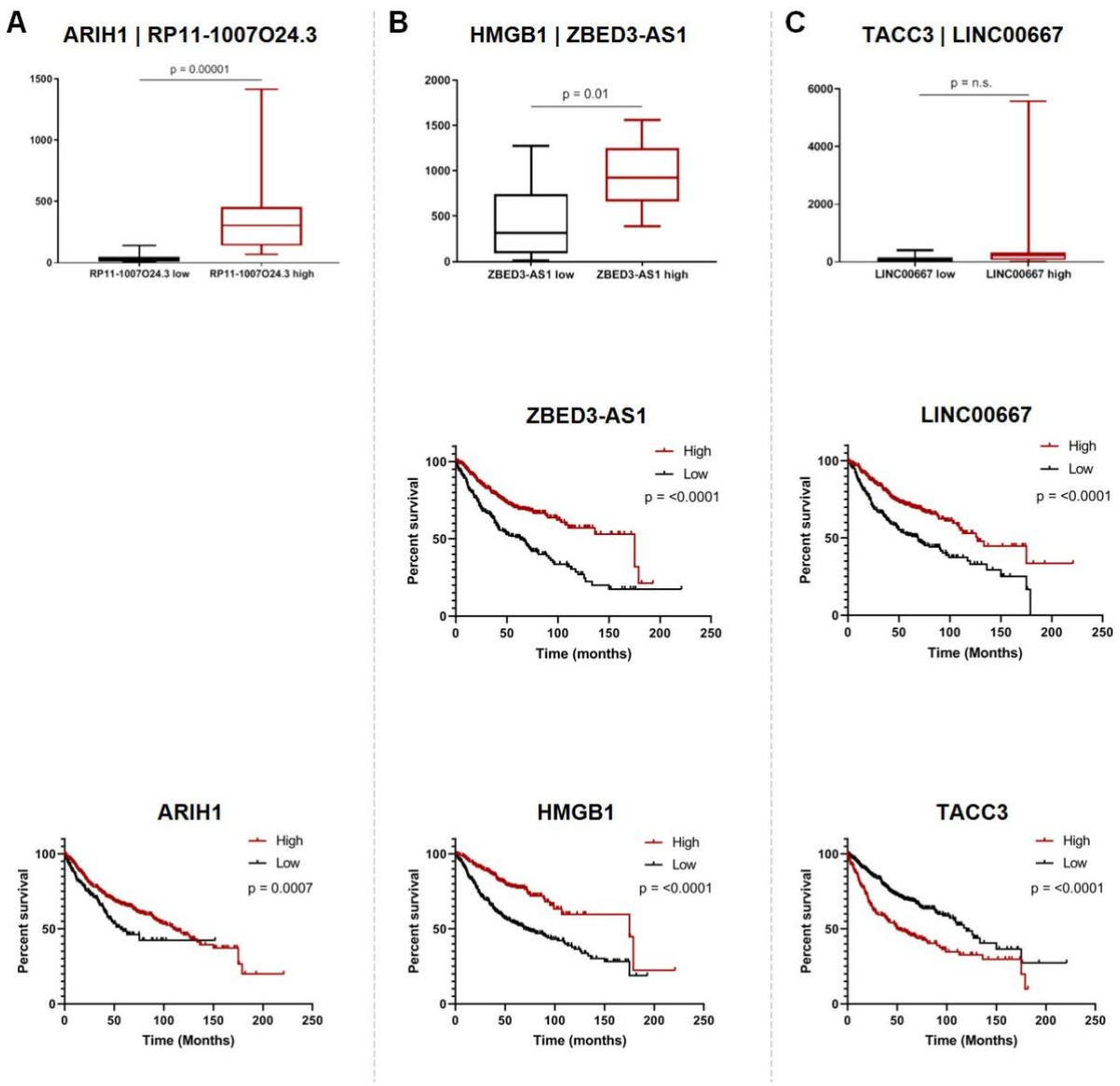

A
ARIH1 | RP11-1007O24.3
p = 0.00001
RP11-1007O24.3 low
RP11-1007O24.3 high
B
HMGB1 | ZBED3-AS1
p = 0.01
ZBED3-AS1 low
ZBED3-AS1 high
C
TACC3 | LINC00667
p = n.s.
LINC00667 low
LINC00667 high
ZBED3-AS1
High
Low
p = <0.0001
Percent survival
Time (months)
LINC00667
High
Low
p = <0.0001
Percent survival
Time (Months)
ARIH1
High
Low
p = 0.0007
Percent survival
Time (Months)
HMGB1
High
Low
p = <0.0001
Percent survival
Time (months)
TACC3
High
Low
p = <0.0001
Percent survival
Time (Months)

Figure 4.6 Associations of pseudogene-derived lncRNAs (upper row), their respective parent gene expression levels and their potential impact on patient outcome (middle row and bottom row).

Expression and survival associations as seen for: *RP11-1007O24.3* and parent gene *ARIH1* **(A)**, *ZBED3-AS1* and parent gene *HMGB1* **(B)**, and *LINC00667* and parent gene *TACC3* **(C)**. Expression associations (upper row) for each pseudogene-derived lncRNA and parent gene pair were found by stratifying samples into tertiles by high (red) and low (black) expression of the pseudogene-derived lncRNA, and plotting the expression of the parent-gene (RPKM) on the y-axis. Survival associations were found for pseudogene-derived lncRNAs (middle row) and their respective parent genes (bottom row). Samples were stratified into tertiles with high (red) and low (black) expression of the gene-of-interest, and the significance of the associations were assessed using the logRank method through GraphPad Prism 8 software on data obtained from kmPlot (n=673). Survival information was not available for RP11-1007O24.3 and it was thus omitted from this analysis.

4.4 Discussion

Here we expand upon the work done by Milligan et al [120], completing the first large-scale analysis of lncRNA expression from pseudogene loci in LUAD and paired non-malignant lung tissue. We discovered a broadly positive association between Ψ-lncs and parent-gene expression, suggestive of an alternative mechanism of cancer gene regulation. While there have been singular examples of deregulated pseudogene-derived lncRNAs in cancer, we show that this phenomenon is widespread in LUAD. In addition to being correlated with Ψ-lnc expression, we find that many of the parent genes of these deregulated lncRNAs are annotated cancer genes and are significantly associated with patient survival, highlighting how these previously-unappreciated non-coding genes may affect LUAD biology.

While the identification of lncRNAs associated with cancer phenotypes is increasing, a great challenge in the field remains the accurate downstream prediction of lncRNA function. Unlike protein-coding genes or small non-coding RNAs, features like complex folding patterns and unknown binding motifs have contributed to the challenging functional characterization of lncRNAs. We utilized the sequence similarity found between lncRNAs expressed from pseudogene loci, and their parent genes to predict the function of this subset of lncRNAs in LUAD. We identified a set of 104 Ψ-lncs deregulated in LUAD in two independent datasets. This greatly increases the number of deregulated lncRNAs known to be expressed from pseudogene loci in LUAD. Interestingly, the majority of these Ψ-lncs were under-expressed in tumours compared to non-malignant tissues, suggesting that they may have tumour-suppressive roles, and that their downregulation is advantageous to LUAD tumourigenesis.

Under-expression was not significantly associated with regions of recurrent copy number deletion in LUAD, although a subset of deregulated Ψ-lnc loci were localized to these regions

(Figure 4.3). These observations suggest that they may be regulated by alternative molecular mechanisms, including broad chromosomal aberrations that affect whole chromosome arms, or epigenetic mechanisms. For example, endogenous retroviruses and repetitive elements often become aberrantly expressed in cancer due to deregulated methylation patterns [129]. We did not observe enrichment of Ψ-lncs or their parent genes on any chromosomes, despite the fact that pseudogenes are known to be overabundant on the human X chromosome [130].

The direction of transcription often affects lncRNA function [111]. For example, the *PTEN* pseudogene (*PTENP1*) expressed in the sense direction can function as a decoy for inhibitory miRNAs that would otherwise cause translational inhibition of the *PTEN* parent mRNA. Conversely, when the antisense lncRNA is expressed from the *PTENP1* locus, the transcript is able to localize to the *PTEN* parent locus and recruit chromatin-remodeling machinery, which leads to the silencing of *PTEN* transcription [76,125]. Both mechanisms have been coopted by cancer cells for their respective tumour suppressive and oncogenic roles [119]. We found sense Ψ-lnc-parent pairs, (which account for 208 out of 391 Ψ-lncs examined) to be more positively correlated than antisense Ψ-lnc-parent pairs in both cohorts (Figure 4.4b). This may imply that Ψ-lncs are more likely to regulate their parent gene in a positive manner which may occur through mechanisms such as miRNA sponging or transcript stabilization when transcribed in the sense direction [119,131]. The distribution of Spearman's ρ-values for antisense-parent gene pairs suggests a more even split between positive and negative regulation. A limitation of this study is that we cannot discount the possibility that sequencing reads for sense overlapping Ψ-lncs that have sequence homology with their parent gene are being mapped to the parent gene instead of the Ψ-lnc. This potential issue warrants further investigation considering both the large number of annotated pseudogenes in the genome (n= 13,000) and the possibility of false interpretation of

sequencing data for both protein-coding and non-coding genes (https://www.genenames.org/cgi-bin/statistics). While these alignment errors could affect sense Ψ-lnc-parent gene pairs, antisense Ψ-lnc-parent gene pairs are not subjected to this same technical artefact. Recently, RNA-sequencing analysis strategies have emerged that begin to address this issue, and long read RNA-sequencing could be used to reduce errors in sequence alignment.

When looking at the parent genes of these deregulated Ψ-lncs we were interested to find that many had previously described roles in cancer (Table 4.1). This includes *EGLN1*, a well described cancer gene involved in regulation of tumour hypoxia, and *CDC42*, an oncogene involved in cell cycle control [132,133]. Many of these deregulated Ψ-lncs were also associated with clinical parameters such as patient survival and patient stage, in addition to the correlated expression between Ψ-lncs and their parent genes (Table 4.2). For example, *ZBED3-AS1* and *HMGB1* were positively correlated at the expression level, and low expression of both genes was associated with poor patient survival (Figure 4.6b). We also observed Ψ-lnc-parent pairs where both genes are associated with survival, but do not share an expression relationship. *LINC00667* and parent gene *TACC3*, for example, are both survival-associated, but not correlated at the expression level. *TACC3* is a component of the TACC3/ch-TOG/clathrin protein complex, and roles in complex assembly have been previously observed for lncRNAs [83] (Figure 4.6c). Thus, it is possible that *LINC00667* is involved in a form of regulation that would not affect transcript levels, including protein-complex assembly. While we were unable to assess survival associations for many of the deregulated Ψ-lncs, they may still impact patient survival through regulation of their parent genes. We found 28 out of 34 expression-associated Ψ-lnc-parent gene pairs to have parent genes associated with patient survival. *RP11-1007O24.3*, for example, was positively correlated with survival-associated parent gene *ARIH1* expression in both cohorts.

ARIH1 has been previously described to be a mediator of DNA-damage response and mitophagy in cancer cells (Figure 4.6a) [134,135]. The potential regulatory impact of Ψ-lncs on their clinically-relevant parent genes is considerable and may represent a novel avenue for targeted therapies. While this study focused on the broad effects of Ψ-lnc deregulation, future studies utilizing *in vitro* and *in vivo* experiments will be necessary to determine the specific mechanisms of parent gene regulation.

As Ψ-lncs may represent an unexplored area of cancer-associated parent gene regulation, their therapeutic relevance should be further explored. LncRNAs make ideal targets for therapies that target RNA products such as Antisense Oligonucleotide (ASO) therapies, since RNA is their final functional state, rather than the intermediate product for protein-coding genes [136]. In addition, ASOs are easier and less costly to develop than small molecule inhibitors, and are in development as aerosol sprays that may be ideal for lung cancer treatment [96,137]. However, as ASOs target through complementary sequence pairing, they would have to be designed in such a way as to not interfere with the parent gene, especially in the case of Ψ-lncs expressed from the sense strand.

This strategy of identifying lncRNAs aberrantly expressed from pseudogene loci may be useful when applied to other cancer types. Indeed, we see that several of our deregulated Ψ-lncs have been described in other tumour types, such as *TPT1-AS1* in cervical cancer, and *HGC11* in both prostate cancer and hepatocellular carcinoma [138-140]. Additionally, as lncRNA expression is highly tissue specific, the application of this approach to other cancer types may yield novel disease-specific Ψ-lnc-parent pairs, highlighting the clinical utility of examining these previously-underappreciated transcripts. Overall, Ψ-lnc-cancer-parent-gene axes represent

alternative mechanisms of cancer gene regulation, and their identification is a critical step towards the functional characterization of lncRNAs.

4.5 Chapter conclusions and contribution to the field

There is a growing need to functionally characterize lncRNAs. Pseudogene-derived lncRNAs have been shown to be involved in cancer and regulate the expression of their parent genes. We show here how pervasive this gene regulatory mechanism is in LUAD samples. We identify a large set of deregulated Ψ-lncs, with aberrant expression observed in RNA-sequencing data from two LUAD cohorts of paired tumour and non-malignant lung tissue samples. We show that these deregulated Ψ-lncs have clinical value and that the parent genes, many of which are correlated with Ψ-lnc expression, have been implicated in cancer phenotypes and are associated with clinical outcome. Together, our results highlight the important roles of the non-coding transcriptome in cancer cellular biology.

Chapter 5: An investigation of regulation of microRNA sponging, through the lens of *XIST*

5.1 Introduction

The previous chapters of this thesis have focused on understudied mechanisms used by lncRNAs to enact their function. In Chapter 5, we focused on the most widely cited mechanism of lncRNA function, which is when a lncRNA regulates another gene by acting as a miRNA sponge. As described earlier, miRNAs are small non-coding RNAs that enact their function by base-pairing to the 5'UTR of mRNAs that contain complementary target sequences. The mRNA is then either degraded or, more commonly, prevented from being translated into a functional protein product. Recently it was described that a lncRNA containing the same target "Seed" sequence as an mRNA can function as a sponge, or decoy for a particular miRNA. In this interaction, the lncRNA functions as a positive regulator of the target mRNA by decreasing the abundance of free miRNAs, thus preventing the inhibition of the mRNA by the shared targeting miRNA [141 125 77].

Since first being described, miRNA sponging by a lncRNA has become a hugely popular topic. In fact, over 600 publications described this mechanism last year, and many more of these types of manuscripts are published each year. However, although lncRNA mediated regulation through a shared miRNA is simple in concept, there are many factors that complicate this interaction within a biological system. As miRNA's target seed sequence can be as little as 6 nucleotides, the number of transcripts each miRNA is able to target can be in the hundreds. Furthermore, as lncRNAs are entirely untranslated, their whole sequence is theoretically available to be bound by miRNAs, thus these long transcripts may be bound by multiple miRNAs.

As such, there are a vast number of potential interactions for any lncRNA, miRNA, or mRNA target gene. As miRNA target prediction is based on complementary sequence homology, this complicates the identification of a biologically functional regulatory relationship. As the majority of recent manuscripts concerning cancer associated lncRNA now postulate they function as miRNA sponges, understanding the key features of this mechanism is increasingly important. This Chapter aims to evaluate the breadth of these predicted interactions and determine the likelihood that they hold true in a biologically relevant manner.

The lncRNA *XIST* is one of the most commonly described miRNA sponges, perhaps due to its notoriety and long length. *XIST* was one of the first lncRNAs to be characterized, which was an important milestone in proving that non-coding RNAs have cellular functions. XIST expression is often deregulated in cancer, and several studies postulate that it may have key functions in lung cancer [142] [143] [144]. Discovered in 1991 by Brown et al., this large (>27Kb) transcript was shown to be involved in the silencing of the inactive X chromosome in females. The second X chromosome is silenced through a mechanism similar to the cis-acting lncRNAs described in Chapter 3, thereby preventing drastic differences in gene expression in males and females [145] [146] [147].

Since the sequencing of the human transcriptome revealed the widespread expression of lncRNAs, many researchers have revisited *XIST* when new lncRNA functions are observed. Since the discovery of miRNA sponging, there has been a drastic increase in the number of publications concerning *XIST* have been concerning miRNA sponging, rather than *XIST*'s canonical function, and this translates to *XIST*'s potential role in cancer (Figure 5.1). Recently *XIST* has now been described to function as an oncogene through sponging tumour suppressive miRNAs in many cancer types, including lung cancer [148, 149]. However, these reports are

inconsistent, with *XIST* often associated with different phenotypes depending on the cancer being described. Furthermore, many of these studies describe *XIST* functioning through a single miRNA to protect a single mRNA, rather than a pool of miRNAs. Additionally, it is unknown whether certain regions of *XIST* are preferential or enriched for miRNA binding. Here we use the lncRNA *XIST* as a proof of principal to investigate the mechanism of miRNA sponging. We investigate the many shared miRNAs that target *XIST*, where and how they bind to the transcript, and how this affects the relationship of *XIST* with its putative sponge target genes.

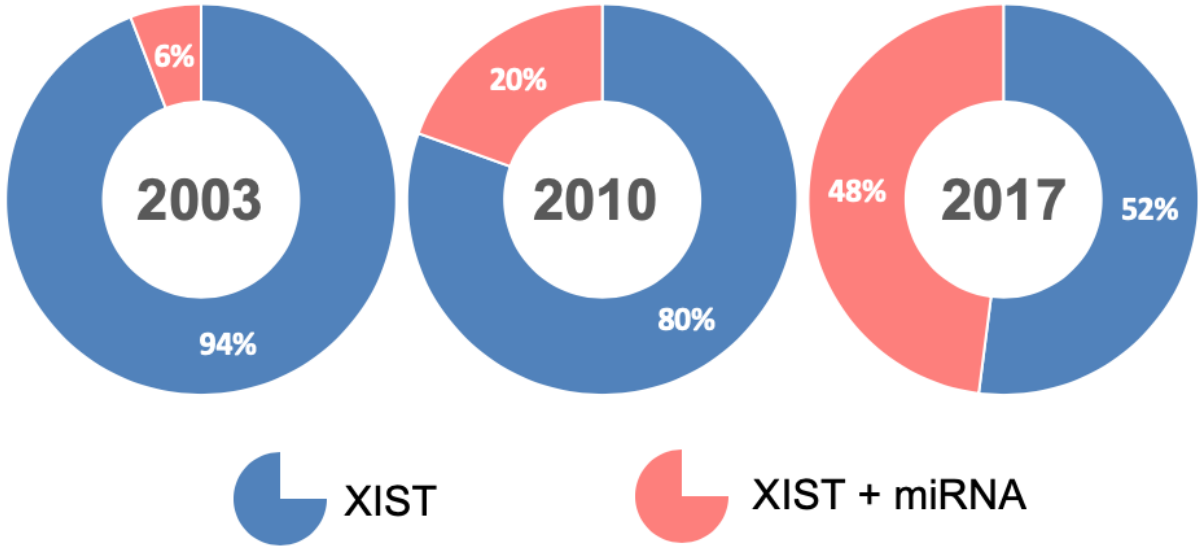


Figure 5.1 Number of published XIST-miRNA sponging manuscripts over time

Proportion of XIST and XIST-miRNA papers in 2003, 2010, and 2017. The proportion of functional manuscripts describing XIST as a miRNA sponge has drastically increased since the mechanism was first described.

5.2 Methods

5.2.1 Data processing

RNA sequencing data: Sequence data for all LUAD and non-malignant tissue samples available from the TCGA (n=304 female and 264 male) was obtained from CancerBrowser (Illumina HiSeq, https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/). The raw sequence reads were then aligned to the human genome (hg19) and the Ensemble gene reference (Release 75) was used to quantify gene expression.

Small RNA sequencing data: Sequence processing described as above in Chapter 2.2. miRNAs from the same LUAD sample set were considered for further analysis if they were expressed above 1 RPKM across 10% of all LUAD samples.

5.2.2 Data analysis

Identification of genes that may be positively regulated through miRNA sponging by XIST (***D****efended from* ***m****iRNA by* ***X****IST*, DMX genes):

Spearman's correlations with *XIST* were performed for every gene annotated in Ensembl (Release 75). As XIST is expressed in females, only female LUAD samples were used in this initial analysis (n=274). Candidate *DMX* genes considered for further analysis if they had a significant correlation coefficient (Rho) >0.4 (n=543 genes), and multiple testing correction was performed using the Benjimini-Hochberg method (BH $p<0.05$).

As miRNAs bind the 3'UTR of mRNAs, we obtained the sequence of 3'UTRs for all of the candidate DMX genes. The 3'UTRs were then analysed using the miRanda miRNA prediction algorithm. Predicted sequence homology and strength of binding energy was then determined for all miRNAs and the candidate DMX genes [150]. Candidate DMX-miRNA pairs were considered for further analysis if they contained a binding score of at least 150, and a

binding energy of at least -20kCal/mol. The full sequence of *XIST* was then input into miRanda and miRNAs that were predicted under the same parameters were identified. miRNAs that were predicted to target at least one candidate DMX gene as well as *XIST* are considered our candidate "shared miRNAs".

5.3 Identification of genes regulated by XIST through microRNA sponging

In order to identify potential sponge targets of XIST we must first define what a sponge candidate will look like. We surmise that a mRNA target gene that is regulated by XIST mediated sponging must:

1) Be expressed in LUAD
2) Be positively correlated with XIST expression
3) Share at least 1 miRNA target sequence with XIST

We define these candidate sponge target genes as target genes *defended from miRNA by XIST*, or DMX genes.

To identify candidate DMX genes we first wanted to identify mRNAs with positive correlations to XIST. LUAD is an ideal sample to study XIST's role as a miRNA sponge, as this cancer occurs in both males and females, and *XIST* has a wide range of expression in female lung samples [151]. As males should have negligible *XIST* expression, we excluded them from this initial analysis. In our female LUAD samples (n = 304) we identified all expressed Ensemble annotated genes. We then performed a Spearman's correlation to identify genes that were significantly positively correlated with *XIST* expression (Spearman's Rho>0.4, BH $p \leq 0.05$). This provided us with a list of 543 candidate DMX genes.

We then sought to identify miRNAs that targeted both the candidate DMX genes, as well as *XIST*. To perform this analysis, we retrieved the 3'UTRs of the 543 candidate genes from the UCSC Genome Browser. Next, we used the miRanda binding algorithm to identify miRNAs that would target these genes. All annotated miRNAs were tested for binding potential using a stringent threshold of binding ($\Delta G \geq$-20kCal/mol, score>150). This led to the identification of 10,654 potential DMX-miRNA interactions, involving 124 candidate DMX genes, and 2052 unique miRNAs.

Next, we wanted to determine which of these miRNAs also targeted *XIST*. We ran the full (unspliced) sequence of *XIST* in the miRanda binding algorithm using the same stringency thresholds ($\Delta G \geq$-20kCal/mol, score>150), which lead us to identify 864 unique miRNAs. Of these, 804 miRNAs were predicted to bind both XIST and at least one of the 124 candidate DMX genes. Lastly, as miRNAs have tissue specific expression, and are required to be expressed to function in a sponge-based regulation, we wanted to confirm the expression of these miRNAs in LUAD. All of the miRNAs identified were expressed in our LUAD samples, and therefore could potentially be sponged (Supplemental Table 3, Appendix B2).

5.4 microRNAs targeting XIST exonic regions display stronger DMX relationships

Many questions remain about what features make for an effective miRNA sponge. With regards to how efficient a sponge can function, we were interested in the distribution and enrichment of miRNAs across *XIST*, and how multiple shared miRNAs may affect *XIST*'s relationship with candidate DMX genes. *XIST* is a very long transcript, and as a result contains many miRNA binding sites. We were interested in whether or not it contained more miRNA binding sites per sequence length than other well-known lncRNAs that are similarly considered

to have secondary function through miRNA sponging. When compared to other lncRNAs, including *NEAT1* and *MALAT1*, *XIST* did not appear to be enriched for miRNA binding sites (Figure 5.2).

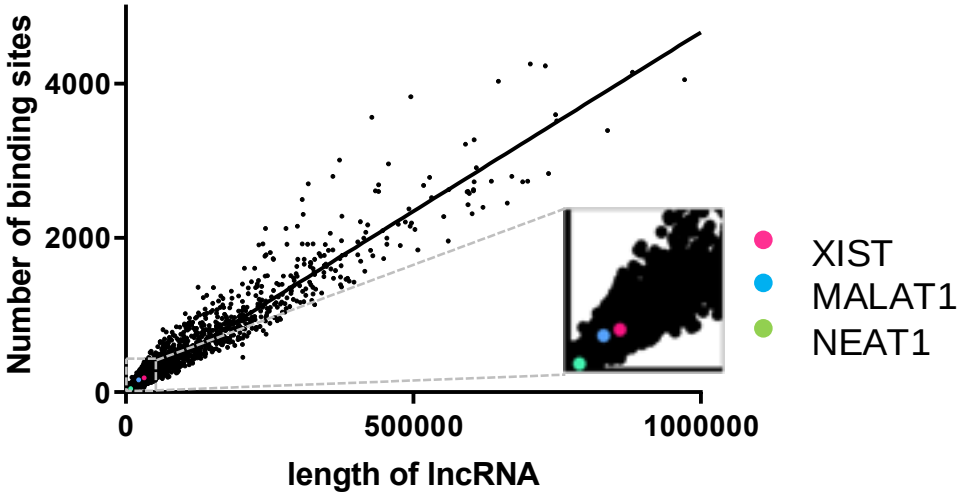


Figure 5.2 Number of binding sites per sequence length of lncRNAs.

Number of miRNA binding sites is displayed on the Y axis while lncRNA length is displayed on the X axis. Each dot represents a lncRNA. XIST is indicated in pink, while other well-known lncRNAs MALAT1 and NEAT1 are indicated by blue and green dots respectively.

Next, to investigate the distribution of miRNA binding sites across XIST we obtained and mapped the binding sites of the 804 miRNAs to the XIST transcript. Interestingly we found that one small region of XIST, exon 5, which is only 163 nucleotides was enriched for binding sites compared to the rest of the transcript with 15 binding sites (Figure 5.3). In terms of distribution of shared miRNA binding sites, we did not observe significant enrichment in either exonic or intronic regions (Figure 5.4).

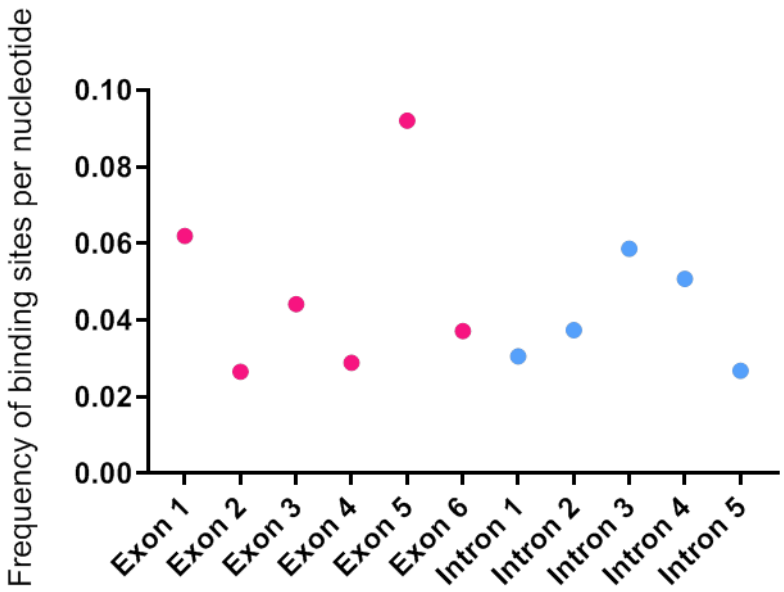


Figure 5.3 Frequency of miRNA binding sites to specific intronic and exonic regions of XIST.

The number of predicted miRNA binding sites were mapped to exonic and intronic regions of XIST sequence. The number of binding sites per sequence length was then calculated to determine if certain regions of XIST were enriched for miRNA binding. Pink circles indicate exonic regions, and intronic regions are denoted by baby blue.

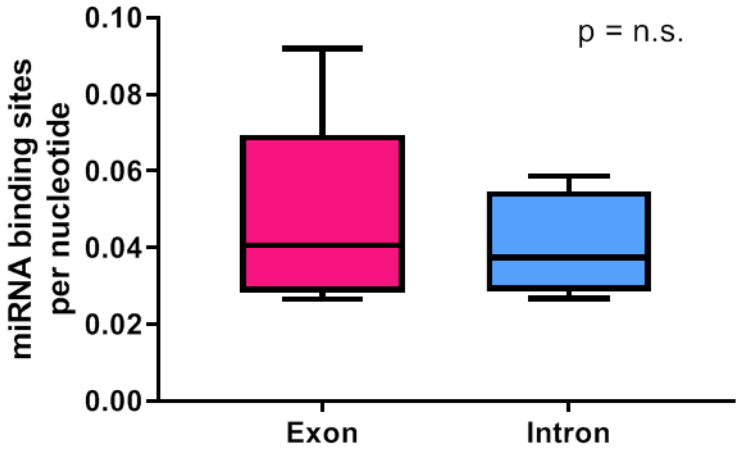


Figure 5.4 Comparison of miRNA exonic and intronic binding frequency.

Number of binding sites per nucleotide for all exons and introns was compared using a Student's t-test.

This led us to investigate whether the location of shared miRNA binding may affect *XIST*'s ability to act as a sponge. We then separated candidate DMX genes into two groups, those targeted exclusively by miRNAs that bind to *XIST*'s exonic regions, and those targeted by miRNAs that bind to *XIST*'s intronic regions. We then compared the expression association of genes in these groups with *XIST* expression. Interestingly, we found that DMX genes that shared miRNAs that bound to exonic regions of *XIST*, exhibit stronger positive correlations with *XIST*, indicating exonic regions may be better sponge targets (Figure 5.5a). Furthermore, when we investigated the correlations between miRNAs and the candidate DMX genes, we found that a greater number reached significance when the miRNAs were predicted to bind exonic regions of *XIST* (Figure 5.5b).

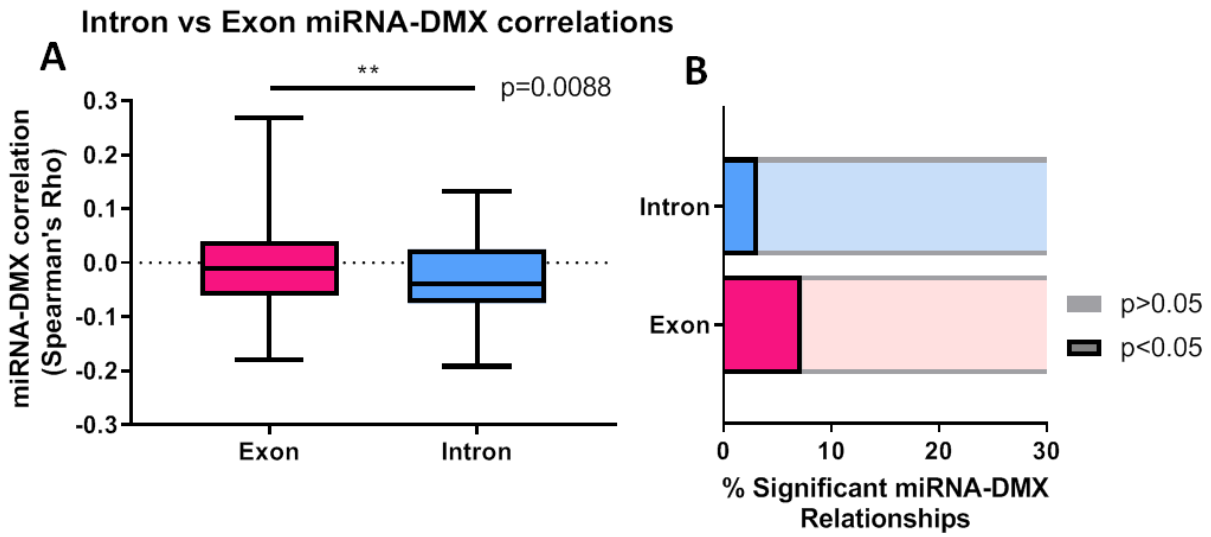


Figure 5.5 Location of miRNA binding affects XIST-DMX relationships.

MiRNAs bound to exonic regions of *XIST*, targeted DMX genes with significantly stronger positive expression associations with *XIST* (A). The number of significant correlations between miRNAs and the DMX genes also increased when miRNAs bound exonic regions of *XIST*.

In theory, an effective miRNA sponge would be able to quench more than a single miRNA at a time, and artificial miRNA sponges are optimized to contain multiple target sites [141 152]. We were interested in whether any of the shared miRNAs were able to bind XIST multiple times. Interestingly, we indeed found that many of the miRNAs targeted multiple regions of *XIST*'s sequence (Figure 5.6). Furthermore, when we compared DMX genes that shared multiple miRNAs with XIST, we found that they exhibited stronger correlations with *XIST* than DMX genes that shared single miRNAs. This suggests that efficient sponge regulation is mediated by a pool of shared miRNAs instead of a single shared miRNA (Figure 5.7a,b). Lastly, the expression level of miRNAs may be important for how readily a gene can be sponged. Interestingly, while

in general DMX-*XIST* correlations increase with shared miRNA expression, DMX genes that shared the highest expressed miRNAs show an unexpected decrease in correlation with *XIST*.

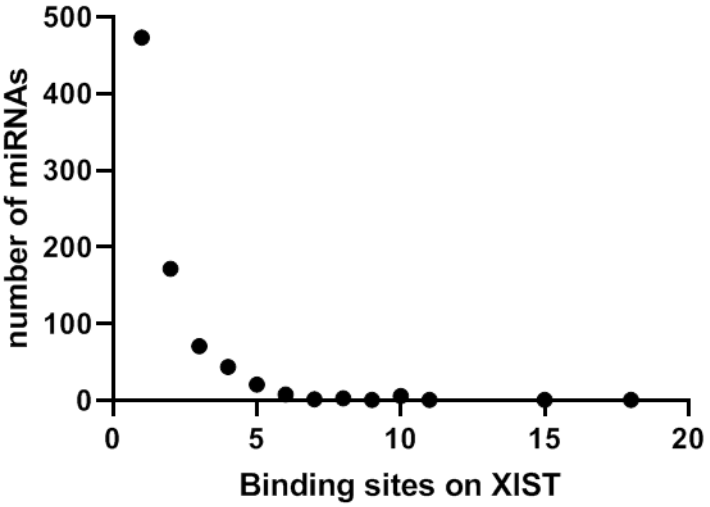


Figure 5.6 The number of target sequences each miRNA has on XIST.

While the majority of miRNAs target a single site on XIST, many target multiple sites along the XIST transcript.

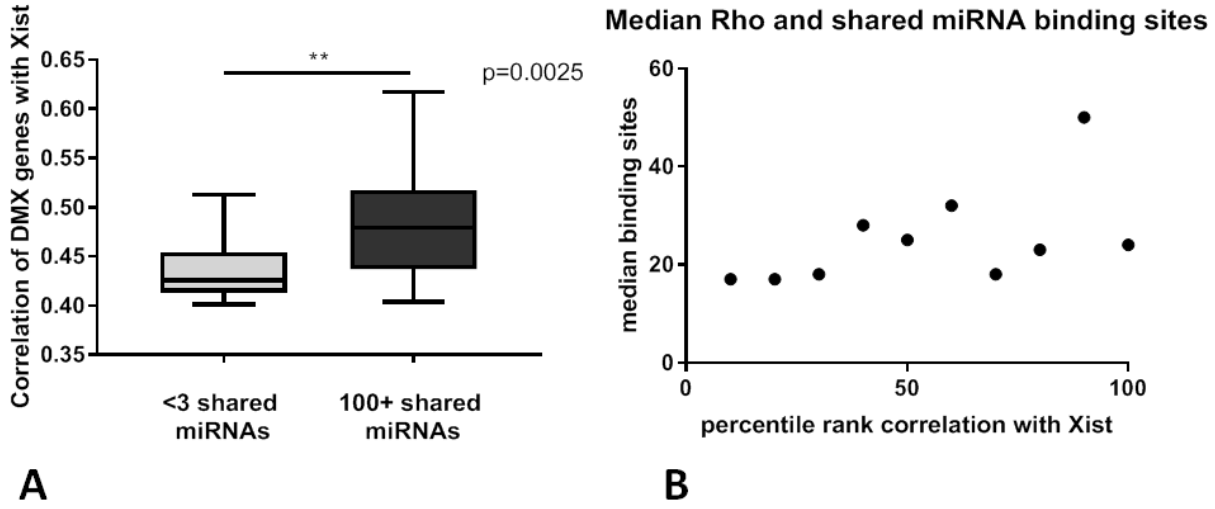


Figure 5.7 Number of shared miRNAs affects DMX and XIST expression correlations. DMX genes that shared less than 3 predicted miRNAs with XIST are significantly less correlated with XIST expression than DMX genes that share the top percental of shared miRNAs (100+) (A). Percent rank of DMX gene correlation with XIST plotted against the number of shared miRNAs (B).

5.5 Chapter discussion

This chapter aimed to identify the best candidates for XIST-mediated miRNA sponge regulation in LUAD, and to identify the factors that impact the efficiency of this mechanism. Importantly, we consider the biological sex of our samples, the full complement of miRNAs available to target *XIST*, and the expression of all mRNA targets and shared miRNAs. We identified a set of genes that are positively correlated with *XIST*, indicating a positive regulatory relationship consistent with miRNA sponging. We then used miRNA binding prediction

algorithms and discover that, of the miRNAs predicted to bind *XIST* (n=862), the vast majority (n=804) bind to these candidate DMX genes.

We investigated how and where these shared miRNAs bind to *XIST* to determine if this had an effect on the candidate DMX genes. *XIST* has many splice isoforms and this may affect which binding sites are available for a miRNA to bind. While we did not observe enrichment of miRNA binding sites in the exons, we found that DMX genes that shared exonically targeting miRNAs exhibited significantly stronger correlations with *XIST*. This is interesting as the fully-spliced *XIST* transcript is the most abundant in lung tissue, and as a result these increased DMX correlations may be a result of the abundance of *XISTs* exonic target sites available for miRNA sponging [151].

The number of miRNAs is also important for any sponge regulation. Studies on the optimization of synthetic miRNA sponges have determined that the optimum number of target sites is 4-10 per sponge to enact the greatest effect on target genes. We find that 10% of our shared miRNAs contain this many binding sites on *XIST*. We also observe that DMX genes that share multiple targeting miRNAs with *XIST* are more strongly positively correlated, indicating that this regulatory mechanism is more effective when a pool of shared miRNAs is involved. Lastly, when we tested the effect of miRNA expression upon DMX-*XIST* relationships, we observed that while general increased miRNA expression led to increased DMX-*XIST* correlations, the highest expressed miRNAs exhibited a reduction in association. While this may seem counter intuitive, it is possible that there is an upper limit to the number of free miRNAs XIST is able to bind; once this limit is reached, remaining miRNAs are free to target DMX genes once more, which reduces the association.

Overall, this chapter highlights the complexity of regulation by miRNA sponging. Any one lncRNA will have hundreds of genes that share targeting miRNAs, making identification of true regulatory targets difficult. We provide an in depth look at miRNA binding to *XIST*, and identify several features that may affect which target genes are preferentially protected from miRNA based degradation. These features will be critical to assist researchers studying this mechanism to identify better, more biologically relevant targets of this massively popular regulation type.

Chapter 6: Conclusions

6.1 Summary of thesis chapters

6.1.1 Overall summary of thesis findings

Up to this point, research on the role of lncRNAs in cancer has been hindered by a lack of functional prediction. Each chapter in this thesis work highlights a different lncRNA mediated mechanism that could be harnessed by tumours to regulate oncogenes and tumour suppressor genes. This work describes methodology's for identifying lncRNAs that function to deregulate cancer genes *in cis*, or *in trans* in LUAD, and additionally provides an in-depth analysis of the complex interactions involved in the mechanism of miRNA sponging.

6.1.2 Summary of thesis Chapter 3

With the goal of identifying mechanisms of lncRNA mediated deregulation of cancer genes, this chapter explored the role of lncRNAs that regulate protein coding genes *in cis*. In order to investigate the prevalence and landscape of *cis*-acting lncRNAs in lung adenocarcinoma, we harnessed two independent cohorts of RNA sequencing data from LUAD tumours, both with matched non malignant tissue. Combining sequence data with genomic location we were able to identify 408 deregulated lncRNAs that overlapped or closely neighboured protein coding genes. We performed literature review and discovered that many of these protein coding genes have been previously described in cancer. Additionally, several of these known cancer genes are associated with the expression of deregulated cis-acting lncRNAs, implying that these lncRNAs may be involved in regulating the expression of these genes. Furthermore, we surmise that these lncRNAs may be deregulated in order to modulate the expression of these cancer-associated genes.

To validate that this methodology can identify cis acting lncRNAs that function to regulate cancer driving genes, this chapter focused on a prospective cis-acting lncRNA that neighboured *HMGA1*, a known oncogene. We found *HMGA1-lnc* to be dowregulated, and inversely correlated with oncogene *HMGA1*. Furthermore, both genes were associated with tumour stage, with HMGA1 expression being associated with more aggressive disease, while *HMGA1-lnc* was associated with less aggressive disease. Lastly, we showed that *HMGA1-lnc* was able to control HMGA1 expression when we inhibited expression of *HMGA1-lnc in vitro* and we found significant increases in HMGA1 expression. This chapter demonstrates that deregulation of *cis*-acting lncRNAs is a frequent event in LUAD, and we observe that downregulation of *HMGA1-lnc* is an alternate mechanism of deregulation of the oncogene *HMGA1*.

6.1.3 Summary of thesis Chapter 4

Building upon work completed in Chapter 3, this chapter aimed to characterize the landscape of a type of lncRNA-based regulation previously never explored in lung cancer, the regulation of protein coding genes by pseudogene derived RNAs (Ψ-lncs). lncRNAs expressed from pseudogene loci have been previously shown to be able to regulate their related protein coding parent genes, and we hypothesized that this form of regulation could be harnessed by cancer cells to deregulate cancer driving genes.

The work here utilizes RNA-sequencing data in conjunction with sequence homology in two cohorts of lung adenocarcinoma, each of which contained matched non-malignant samples to identify 104 deregulated Ψ-lncs in LUAD. Interestingly, upon investigation, we find that many of these deregulated lncRNAs were expressed from the loci of pseudogenes related to

known cancer genes. Furthermore, we identify many deregulated pseudogene-derived lncRNAs whose expression significantly correlated with these cancer associated protein coding genes, suggesting a regulatory relationship. We harnessed a large public clinical dataset to then determine if Ψ-lncs, or their parent genes were associated with patient survival. While only 19 of the deregulated Ψ-lncs were represented on the microarrays compiled in this platform, 16 Ψ-lncs with deregulated expression were significantly associated with patient survival, implicating the clinical relevance of these genes. This work shows that deregulation of pseudogene-derived lncRNAs is a widespread phenomenon in LUAD and highlights trans-acting lncRNA regulation through sequence homology as an alternative mechanism of cancer gene deregulation in LUAD. A version of this chapter is published as "Aberrant Expression of Pseudogene-Derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma" in Frontiers in Genetics (2019) [107].

6.1.4 Summary of thesis Chapter 5

While lncRNAs have been increasingly implicated in cancer, functional characterization has remained a major challenge, and there is a growing gap between lncRNAs named, and those characterized. Due to the ease of use and widespread availability of miRNA prediction software, by far the most published and frequently studied lncRNA function is when a lncRNA acts as a miRNA sponge. However, questions remain over how robust these types of analysis are when applied to complex multi-gene interaction such as miRNA-sponge networks, and it is unknown what features make a more efficient sponge. This chapter explores the mechanism of lncRNA based miRNA sponging by analyzing the lncRNA *XIST* and potential targets of shared miRNAs in LUAD.

To investigate the role of *XIST* acting as a miRNA sponge, we designed a pipeline to identify mRNAs that may be "sponged" by *XIST* that we define as or Defended from miRNAs by *XIST* or DMX genes. These genes are expected to display an expression pattern that correlates positively with that of *XIST* and must have shared miRNA binding sites. To find candidate DMX genes we scanned both the sequence of *XIST* and the 3'UTRs of DMX transcripts to find common microRNA binding sites and utilized RNA sequencing and miRNA sequencing data from hundreds of LUAD samples. While XIST is targeted by over 800 miRNAs, and positively correlated with over 500 genes, using this approach allowed us to narrow down our discovery set to 124 genes potentially regulated by XIST through the sequestration of miRNAs (DMX). We then set out to further explore the features important to miRNA-mediated gene regulation. We evaluated whether multiple miRNA binding sites or certain regions of a gene are better at sequestering miRNAs. This led us to find that microRNAs targeting *XIST* at exonic regions and DMX genes that share multiple miRNA binding sites with *XIST* display a stronger *XIST*-DMX relationship. Another main finding of this chapter is the potential for false positive sponge interactions. While there are hundreds of genes that may be predicted to be targets of sponge-based regulation, many of these predictions may not be biologically relevant. lncRNAs can be targeted by hundreds of miRNAs (XIST alone is predicted to be targeted by over 800 miRNAs), and each of these miRNAs can in turn target tens to hundreds of genes. As miRNA predictions are used as the primary method to assign function to a deregulated lncRNA, this work emphasizes the need for higher stringency when studying this mechanism. Considering basic cell biology, expression, and localization will lead to increased confidence in identifying targets with more clinical relevance. An expanded version of this Chapter is published in PLoS One as

"Beyond sequence homology: Cellular biology limits the potential of XIST to act as a miRNA sponge" (2019) [153].

6.2 Strengths and limitations

6.2.1 Strengths and limitations of Chapter 3

6.2.1.1 Strengths

Identifying the function of lncRNAs has remained a major challenge in the field, and the mechanism of cis-acting lncRNAs is understudied. As over 30% of the genes in the human genome display overlap with other genes, this may be impacting a wide variety of genes globally [154] [111]. A major strength of this chapter is in its methodology to both generate and test hypotheses regarding lncRNA functions in cancer. This provides an alternate method to identify the candidate target genes of a lncRNA, as lab-based screens are both time consuming and expensive. Additionally, many of the screen based assays traditionally used to identify gene functions are intended for protein coding genes and are not well suited for lncRNAs. For example, pull downs on large protein complexes such as PRC2 reveal thousands of interacting RNAs, implying non-specific binding and complicating the identification of relevant gene interactions [155].

The use of two datasets of LUAD with matched non malignant tissue is also an advantage of this chapter, as it reduces the sample bias of using a single dataset, and allows for increased confidence that the lncRNAs identified as deregulated are biologically relevant to lung cancer. We highlight the utility of this approach with the identification of a deregulated lncRNA, *HMGA1-lnc*, which controls the expression of HMGA1. HMGA1 is an important cancer gene and the discovery of a lncRNA that regulates it has implications for many tumour types and demonstrates the utility of using RNA sequencing data to interrogate genomic loci. We

confirmed the ability of *HMGA1-lnc* to regulate HMGA1 in BEAS-2B cells, a cell model where previous studies have shown HMGA1 expression to drive oncogenic phenotypes. This study shows the potential of this type of lncRNA mediated regulation and that there may be many other undiscovered non-coding members of cancer pathways in LUAD and other malignancies.

6.2.1.1 Limitations

The many roles of lncRNAs in lung cancer have just begun to be uncovered, and while there remains considerable biology to discover, this methodology will be limited to capturing lncRNAs that are cis-acting. For example, this method will likely not identify trans-acting effects, as well as interactions with proteins and RNA that do not result in changes to neighbouring gene expression levels, including protein complex assembly, RNA splicing, and cellular localization. Additionally, while this methodology is hypothesis generating, there may be false positives and negatives as not all lncRNAs regulate their neighbouring genes, and these gene-pairs may be affected by other mechanisms that affect their transcript levels. This could include passenger effects such as DNA copy number alterations that by increasing or decreasing gene copies, can make *cis*-pairs appear to display concordant expression patterns. This highlights the importance of using cell models for additional experimental validation, and to confirm the regulatory ability of each prospective cis-acting lncRNA on their neighbouring genes of interest.

6.2.2 Strengths and limitations of Chapter 4

6.2.2.1 Strengths

Chapter 4 is the first global look at pseudogene derived lncRNAs, shedding light upon an entire class of gene previously unexplored in LUAD. The analysis of expression from

pseudogene loci has previously been complicated by the sequence homology of pseudogenes to protein coding genes, and inconsistencies between pseudogene databases. Until recently, even the location and name of many pseudogenes could vary greatly based on the database used. This chapter is the first to utilize a non-redundant database of the three largest pseudogene databases to analyze sequence data in conjunction with RNA sequencing profiles and protein coding gene homology.

We harnessed public databases containing hundreds of samples with clinical data to ascertain whether parent genes of these Ψ-lncs are associated with survival [107]. Similarly to the process used in Chapter 3, this methodology can be applicable to other cancer types, and may have a large impact on the cancer genome as there are estimated to be over 20, 000 pseudogenes in genome [156]. This chapter raises awareness of the broad spectrum of Ψ-lnc deregulation in LUAD, their expression relationships with known cancer genes, and associations with patient survival. Lastly, this chapter demonstrates the importance of this type of regulation and paves the way for further research on this understudied gene class, in LUAD, and other cancers.

6.2.2.2 Limitations

In this chapter our goal was to identify high confidence Ψ-lncs that may be involved in the regulation of their protein coding parent gene. To do this we focused on lncRNAs deregulated in 2 datasets of LUAD, but as this is a heterogenous disease we may have missed some Ψ-lncs that are less frequently deregulated or expressed at low levels. While the sequence homology of these lncRNAs allows them to perform their functions, it is a double-edged sword when it comes to expression analysis. Sequencing reads that are identical between the lncRNA and the parent gene will likely be mapped to the protein coding gene, potentially biasing expression relationships to appear concordant. While this is largely an issue for lncRNAs

expressed from the sense strand of the pseudogene, and is less of an issue for lncRNAs expressed from the opposite strand, future studies will need to address this issue, in addition to pseudogene-based reads contaminating protein coding expression analysis. Further, the exact amount of sequence homology needed for regulation of a parent gene is unknown, and it may be different for different types of regulation. For example, the sequence necessary for a Ψ-lnc to interact with DNA and recruit a protein complex may be much different in size than sequence homology to function as a decoy for shared miRNAs. Additionally, depending on said mechanism of regulation there may be certain locations where the sequence of the gene may be more important, for example 3' UTRs. While the purpose of this study was to observe the global deregulation of these lncRNAs, individual Ψ-lnc-parent interactions will need to be confirmed in cell models both to confirm their regulatory ability, as well as to assess their effect on cancer cell growth.

6.2.3 Strengths and limitations of Chapter 5

6.2.3.1 Strengths

MiRNA sponging is an immensely popular topic, and currently the most popular method of predicting the function of a lncRNA. As we study this mechanism in depth, a strength of this chapter is how applicable this research will be to those currently using this method to base their studies on. While our study focused on *XIST*, researchers studying other miRNA sponges may be able to identify better sponge targets by taking into account the binding features discussed in this chapter. Another strength of this chapter is the novel approach taken to explore the topic of the mechanism of miRNA sponging. We are able to perform this analysis by taking advantage of large public repositories of sequencing data, combining datasets of miRNA sequencing and RNA sequencing to probe for *XIST*-miRNA-DMX relationships.

6.2.3.2 Limitations

While this work has identified key features to consider while studying miRNA sponging, and finding several high confidence miRNA-*XIST*-DMX pairs in the process, confirming these complex interactions is still a complex and difficult process. Many things could interfere with any lncRNA-miRNA-mRNA interaction, including the number of total binding sites on the lncRNA, as well as the expression levels of a given miRNA. An even larger issue is that there are many other genes with varying numbers of miRNA binding sites that are competing for the binding of a miRNA. Additionally, the current standard method of confirming miRNA binding to a target mRNA in cell models usually involves the induced overexpression of a miRNA of interest. For example experimental validations such as luciferase assays or mutating binding sites do not mimic natural biological situations. This makes identification of biologically relevant sponging interactions difficult both *in silico* and *in vivo* as these conditions have artificially high miRNA levels and are not indicative of normal conditions within a cell.

6.3 Future directions

6.3.1 Chapter 3

The methodology used in Chapter 3 was able to identify candidate *cis*-acting lncRNAs that can regulate cancer driving genes like HMGA1. This approach may be useful in the study of other cancer types, where with different genetic backgrounds, other cis-acting lncRNAs may be regulating other known oncogenes or tumour suppressors specific to other cancer types. As HMGA1 is a known oncogene in other forms of malignancy, particularly breast cancer, it would be useful to determine if this lncRNA based mechanism is a common feature. If so this

interaction could represent a novel clinical intervention point in HMGA1-driven cancers. Additionally, future work into the exact nature of the HMGA-lnc mechanism may discover features that are applicable to a wide variety of cis-acting lncRNAs and could have clinical use. For example, lncRNAs that function through active methylation may be vulnerable to demethylating agents such as 5-azadeoxycytidine. Alternatively the discovery of a binding sequence or important secondary structure features used to recruit specific protein complexes would allow for a more targeted therapeutic approach.

6.3.2 Chapter 4

The Ψ-lncs identified in this chapter are deregulated, as well as associated with the expression of their cancer associated parental genes survival. However there is much variability in the homology and sequence overlap of each pseudogene and parent. As such, the regulatory ability of each Ψ-lnc-parent pair will need to be confirmed in cell models. Additionally, *in vitro* efforts concerning how the similarity of each Ψ-lnc to its parental gene, affects the mechanism of action would be very useful in predicting functional interactions. Similar to the *cis*-acting lncRNAs identified in Chapter 3, efforts to deconvolute the mechanisms of these *trans*-acting Ψ-lnc pairs will be important not only for understanding how this gene class functions, but for the potential therapeutic benefits of modulating expression of these genes.

Applying this methodology to other cancer types may reveal many other Ψ-lncs deregulated as an alternate mechanism of cancer gene deregulation. As these lncRNA-mRNA interactions are understudied there are many potential regulatory mechanisms at work here. For example, sense strand Ψ-lncs may make for ideal miRNA sponge candidates, as they can contain many of the same miRNA binding sites as their protein coding parent gene. However this same

sense strand homology can be problematic when it comes to differentiating reads from highly similar segments of Ψ-lncs and their parents. Current methods are better at differentiating regions of high dissimilarity, but have difficulty mapping similar reads. Future studies with higher depth sequencing will be able to better differentiate reads that map to regions with sparse, infrequent differences between each unique pseudogene and parent gene. Additionally, single cell sequencing will be a valuable asset to deconvolute Ψ-lnc parent pairs, and reduce the transcriptomic noise caused by the heterogeneity found in bulk tumours.

6.3.3 Chapter 5

While this chapter details the first in depth analysis of miRNA sponging through XIST, many questions remain in regard to this mechanism. The approach taken in this chapter was to determine whether the binding patterns of miRNAs to *XIST* affect the sponge regulated genes. In our manuscript published on this topic we dive further into *XIST*'s role as a miRNA sponge, by further investigating our candidate DMX genes. While a similar approach using gene expression profiles and miRNA prediction can be used on other lncRNAs to investigate and identify high confidence sponge targets, *XIST* has unique expression patterns that allowed us to further test our candidate DMX genes. As *XIST* is only expressed in females, we can compare expression patterns in biological systems with, and without *XIST* expression by analysing the male and female LUAD samples.

In our manuscript, we compare the correlation between miRNAs and the DMX genes in systems with (female n=304) and without (male n=264) *XIST*. Interestingly we find that a small subset of male samples have high XIST expression, and that the DMX genes are significantly more correlated with XIST than the males with no XIST expression. We expected that in the

systems with no sponge (males) the miRNA-DMX correlation of true sponged genes would be more negative as more free miRNAs would be available to perform their inhibitory action than those sequestered by the sponge. By comparing samples with, and without XIST, we discovered a high confidence set of 13 miRNA-DMX pairs with correlations that were significantly more negative in the LUAD samples without *XIST* (males). To further test the ability of these high confidence miRNAs to be sponged by *XIST* we then assessed the cellular location of the miRNAs. *XIST* is expressed exclusively in the nucleus, thus our candidate miRNAs must be present in the same compartment in order to interact with and be sponged by *XIST*. We tested five miRNAs for nuclear presence across 3 cells lines and found that all were present in the nucleus [153].

Further work on the high confidence miRNAs, and their target genes may reveal if these genes could impact cancer. *XIST* mediated miRNA sponging is frequently brought up in cancer, so it would be interesting to see if these miRNAs affect lung cancer phenotypes through DMX gene interactions in female LUAD. Additionally, future studies analyzing single cell sequencing may provide a more in depth system to analyze the interaction of *XIST* with these miRNAs and their target genes.

Currently, miRNA sponging is the most popular mechanism of lncRNA function, as *in silico* prediction of interactions with miRNAs are simpler to generate than many other possible interactions. However, future studies will require increased stringency to ensure that these interactions are important within the cell. Additionally, many questions remain and future studies will need to ascertain which binding sites are most relevant, and whether there are other features that determine which miRNAs are most likely to be used in sponging interactions. Further, certain lncRNAs may be better sponge candidates for certain genes. For example lncRNAs

expressed from pseudogene loci may share several of the same miRNA binding sites with their parent genes, and thus may function as superior sponges than lncRNAs sharing few miRNAs.

6.3.4 Future directions for the non-coding field

The immune system has become a major focus in lung cancer research, resulting in new therapeutics that have benefited patients with tumours that evade the immune system. The role of non-coding RNAs within immune cells remains largely unknown, especially with regards to cancer. Recent work from our group suggests that there are immune cell type specific lncRNAs, and that these lncRNAs can be found in bulk lung tumours, which may have important implications for detection of immune cells in tumours or suggest an active role of lncRNAs in the tumour-immune response. Future studies will reveal if these cell type specific lncRNAs have clinical benefit.

Lung cancer screening with low-dose computed tomography (CT), is currently in early studies to detect the disease early in high risk individuals, where the disease is less invasive and more treatable. It is estimated that by taking this approach lung cancer deaths can be reduced by 20% (www.bclungscreentrial.com). High risk individuals may be identified with blood based biomarkers and since miRNAs are stable in blood, these molecules could act as markers of disease. In particular, novel miRNAs have been shown to be extremely cell type specific and may have potential to better predict at-risk individuals [59] [157] [158]. Recent studies from our group have shown tumour specific novel miRNA expression as well as associations with patient survival, which may hint at their clinical utility [159] [160] [161]. Similar to how piRNAs were able to better separate survival curves for patients, adding these novel species of miRNAs to biomarker panels may also improve diagnostic and prognostic markers.

As focused on in Chapters 3 and 4, lncRNAs can act *in cis* and *in trans* as alternate mechanisms of cancer gene deregulation. As such, many of these non-coding genes may represent novel therapeutic intervention points. One promising method of targeting non-coding RNAs are Antisense Oligonucleotides (ASO's). As ASO's are cheaper to manufacture than small molecule inhibitors and specific to RNAs, they may be the most ideal way to target lncRNAs. Furthermore, they may make ideal drugs for lung cancer treatment, as they are currently being tested for dispersal by aerosol sprays to the lungs *in vivo*. [162]. Another potential benefit of targeting lncRNAs in this fashion is the prospect of re-activating silenced tumour suppressor genes by using ASO's targeting repressive cis-acting or trans acting lncRNAs. Future work will be needed to determine which deregulated lncRNAs would be ideal candidates for this type of treatment, but as more lncRNAs are identified as having important roles in cancer, ASO's are an exciting avenue for future treatment of lung cancer.

References

1 Thu, K. L. *et al.* Lung adenocarcinoma of never smokers and smokers harbor differential regions of genetic alteration and exhibit different levels of genomic instability. *PloS one* **7**, e33003, doi:10.1371/journal.pone.0033003 (2012).

2 Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers--a different disease. *Nature reviews. Cancer* **7**, 778-790, doi:10.1038/nrc2190 (2007).

3 Popper, H. H. Progression and metastasis of lung cancer. *Cancer metastasis reviews* **35**, 75-91, doi:10.1007/s10555-016-9618-0 (2016).

4 Alvarado-Luna, G. & Morales-Espinosa, D. Treatment for small cell lung cancer, where are we now?-a review. *Translational lung cancer research* **5**, 26-38, doi:10.3978/j.issn.2218-6751.2016.01.13 (2016).

5 Govindan, R. *et al.* Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: analysis of the surveillance, epidemiologic, and end results database. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **24**, 4539-4544, doi:10.1200/JCO.2005.04.4859 (2006).

6 Derman, B. A., Mileham, K. F., Bonomi, P. D., Batus, M. & Fidler, M. J. Treatment of advanced squamous cell carcinoma of the lung: a review. *Translational lung cancer research* **4**, 524-532, doi:10.3978/j.issn.2218-6751.2015.06.07 (2015).

7 Tang, E. R., Schreiner, A. M. & Pua, B. B. Advances in lung adenocarcinoma classification: a summary of the new international multidisciplinary classification system (IASLC/ATS/ERS). *Journal of thoracic disease* **6**, S489-501, doi:10.3978/j.issn.2072-1439.2014.09.12 (2014).

8 Ansari, J., Shackelford, R. E. & El-Osta, H. Epigenetics in non-small cell lung cancer: from basics to therapeutics. *Translational lung cancer research* **5**, 155-171, doi:10.21037/tlcr.2016.02.02 (2016).

9 Lu, F. & Zhang, H. T. DNA methylation and nonsmall cell lung cancer. *Anatomical record* **294**, 1787-1795, doi:10.1002/ar.21471 (2011).

10 Kullmann, L. & Krahn, M. P. Controlling the master-upstream regulation of the tumor suppressor LKB1. *Oncogene* **37**, 3045-3057, doi:10.1038/s41388-018-0145-z (2018).

11 Lockwood, W. W., Chandel, S. K., Stewart, G. L., Erdjument-Bromage, H. & Beverly, L. J. The novel ubiquitin ligase complex, SCF(Fbxw4), interacts with the COP9 signalosome in an F-box dependent manner, is mutated, lost and under-expressed in human cancers. *PloS one* **8**, e63610, doi:10.1371/journal.pone.0063610 (2013).

12 Lockwood, W. W. *et al.* Cyclin E1 is amplified and overexpressed in osteosarcoma. *The Journal of molecular diagnostics : JMD* **13**, 289-296, doi:10.1016/j.jmoldx.2010.11.020 (2011).

13 Rowbotham, D. A. *et al.* Multiple Components of the VHL Tumor Suppressor Complex Are Frequently Affected by DNA Copy Number Loss in Pheochromocytoma. *International journal of endocrinology* **2014**, 546347, doi:10.1155/2014/546347 (2014).

14 Cooper, W. A., Lam, D. C., O'Toole, S. A. & Minna, J. D. Molecular biology of lung cancer. *Journal of thoracic disease* **5 Suppl 5**, S479-490, doi:10.3978/j.issn.2072-1439.2013.08.03 (2013).

15 Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075, doi:10.1038/nature07423 (2008).
16 Hecht, S. S. Tobacco smoke carcinogens and lung cancer. *Journal of the National Cancer Institute* **91**, 1194-1210, doi:10.1093/jnci/91.14.1194 (1999).
17 Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121-1134, doi:10.1016/j.cell.2012.08.024 (2012).
18 Lovly C, Horn L & W., P. *2018 Molecular Profiling of Lung Cancer*, <https://www.mycancergenome.org/content/disease/lung-cancer> (2018).
19 Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Molecular and cellular biology* **10**, 28-36, doi:10.1128/mcb.10.1.28 (1990).
20 Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. *Annual review of genomics and human genetics* **6**, 69-92, doi:10.1146/annurev.genom.6.080604.162350 (2005).
21 Brown, C. J. & Willard, H. F. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature* **368**, 154-156, doi:10.1038/368154a0 (1994).
22 Jarroux, J., Morillon, A. & Pinskaya, M. History, Discovery, and Classification of lncRNAs. *Advances in experimental medicine and biology* **1008**, 1-46, doi:10.1007/978-981-10-5203-3_1 (2017).
23 Kufel, J. & Grzechnik, P. Small Nucleolar RNAs Tell a Different Tale. *Trends in genetics : TIG* **35**, 104-117, doi:10.1016/j.tig.2018.11.005 (2019).
24 Cao, T. *et al.* Biology and clinical relevance of noncoding sno/scaRNAs. *Trends in cardiovascular medicine* **28**, 81-90, doi:10.1016/j.tcm.2017.08.002 (2018).
25 Fu, Q. & Wang, P. J. Mammalian piRNAs: Biogenesis, function, and mysteries. *Spermatogenesis* **4**, e27889, doi:10.4161/spmg.27889 (2014).
26 Martinez, V. D. *et al.* Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology. *Scientific reports* **5**, 10423, doi:10.1038/srep10423 (2015).
27 Gebert, D., Ketting, R. F., Zischler, H. & Rosenkranz, D. piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals. *PloS one* **10**, e0124860, doi:10.1371/journal.pone.0124860 (2015).
28 Ku, H. Y. & Lin, H. PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *National science review* **1**, 205-218, doi:10.1093/nsr/nwu014 (2014).
29 Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320-324, doi:10.1126/science.1129333 (2006).
30 Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature* **450**, 304-308, doi:10.1038/nature06263 (2007).
31 Watanabe, T. & Lin, H. Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs. *Molecular cell* **56**, 18-27, doi:10.1016/j.molcel.2014.09.012 (2014).
32 Post, C., Clark, J. P., Sytnikova, Y. A., Chirn, G. W. & Lau, N. C. The capacity of target silencing by Drosophila PIWI and piRNAs. *Rna* **20**, 1977-1986, doi:10.1261/rna.046300.114 (2014).

33 Sage, A. P. *et al.* Oncogenomic disruptions in arsenic-induced carcinogenesis. *Oncotarget* **8**, 25736-25755, doi:10.18632/oncotarget.15106 (2017).
34 Orellana, E. A. & Kasinski, A. L. MicroRNAs in Cancer: A Historical Perspective on the Path from Discovery to Therapy. *Cancers* **7**, 1388-1405, doi:10.3390/cancers7030842 (2015).
35 Bhaskaran, M. & Mohan, M. MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. *Veterinary pathology* **51**, 759-774, doi:10.1177/0300985813502820 (2014).
36 Saini, H. K., Griffiths-Jones, S. & Enright, A. J. Genomic analysis of human microRNA transcripts. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17719-17724, doi:10.1073/pnas.0703890104 (2007).
37 Hammond, S. M. An overview of microRNAs. *Advanced drug delivery reviews* **87**, 3-14, doi:10.1016/j.addr.2015.05.001 (2015).
38 Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19**, 92-105, doi:10.1101/gr.082701.108 (2009).
39 Zou, Q., Mao, Y., Hu, L., Wu, Y. & Ji, Z. miRClassify: an advanced web server for miRNA family classification and annotation. *Computers in biology and medicine* **45**, 157-160, doi:10.1016/j.compbiomed.2013.12.007 (2014).
40 Enfield, K. S. *et al.* Deregulation of small non-coding RNAs at the DLK1-DIO3 imprinted locus predicts lung cancer patient outcome. *Oncotarget* **7**, 80957-80966, doi:10.18632/oncotarget.13133 (2016).
41 Krol, J., Loedige, I. & Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics* **11**, 597-610, doi:10.1038/nrg2843 (2010).
42 Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787-798, doi:10.1016/s0092-8674(03)01018-3 (2003).
43 Calin, G. A. *et al.* Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15524-15529, doi:10.1073/pnas.242606799 (2002).
44 He, L. *et al.* A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828-833, doi:10.1038/nature03552 (2005).
45 Chan, L. W., Wang, F. F. & Cho, W. C. Genomic sequence analysis of EGFR regulation by microRNAs in lung cancer. *Current topics in medicinal chemistry* **12**, 920-926, doi:10.2174/156802612800166747 (2012).
46 Qin, Q., Wei, F., Zhang, J., Wang, X. & Li, B. miR-134 inhibits non-small cell lung cancer growth by targeting the epidermal growth factor receptor. *Journal of cellular and molecular medicine* **20**, 1974-1983, doi:10.1111/jcmm.12889 (2016).
47 Chen, L. *et al.* Metastasis is regulated via microRNA-200/ZEB1 axis control of tumour cell PD-L1 expression and intratumoral immunosuppression. *Nature communications* **5**, 5241, doi:10.1038/ncomms6241 (2014).
48 Breunig, C. *et al.* MicroRNA-519a-3p mediates apoptosis resistance in breast cancer cells and their escape from recognition by natural killer cells. *Cell death & disease* **8**, e2973, doi:10.1038/cddis.2017.364 (2017).

49 Wang, T. *et al.* The expression of miRNAs is associated with tumour genome instability and predicts the outcome of ovarian cancer patients treated with platinum agents. *Scientific reports* **7**, 14736, doi:10.1038/s41598-017-12259-w (2017).
50 Kong, Q., Shu, N., Li, J. & Xu, N. miR-641 Functions as a Tumor Suppressor by Targeting MDM2 in Human Lung Cancer. *Oncology research* **26**, 735-741, doi:10.3727/096504017X15021536183490 (2018).
51 Fortunato, O. *et al.* Mir-660 is downregulated in lung cancer patients and its replacement inhibits lung tumorigenesis by targeting MDM2-p53 interaction. *Cell death & disease* **5**, e1564, doi:10.1038/cddis.2014.507 (2014).
52 Mao, G. *et al.* Tumor-derived microRNA-494 promotes angiogenesis in non-small cell lung cancer. *Angiogenesis* **18**, 373-382, doi:10.1007/s10456-015-9474-5 (2015).
53 Cortez, M. A. *et al.* PDL1 Regulation by p53 via miR-34. *Journal of the National Cancer Institute* **108**, doi:10.1093/jnci/djv303 (2016).
54 Li, J., Lei, H., Xu, Y. & Tao, Z. Z. miR-512-5p suppresses tumor growth by targeting hTERT in telomerase positive head and neck squamous cell carcinoma in vitro and in vivo. *PloS one* **10**, e0135265, doi:10.1371/journal.pone.0135265 (2015).
55 Xu, P., Li, Y., Zhang, H., Li, M. & Zhu, H. MicroRNA-340 Mediates Metabolic Shift in Oral Squamous Cell Carcinoma by Targeting Glucose Transporter-1. *Journal of oral and maxillofacial surgery : official journal of the American Association of Oral and Maxillofacial Surgeons* **74**, 844-850, doi:10.1016/j.joms.2015.09.038 (2016).
56 Ranade, A. R. *et al.* MicroRNA 92a-2*: a biomarker predictive for chemoresistance and prognostic for survival in patients with small cell lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **5**, 1273-1278, doi:10.1097/JTO.0b013e3181dea6be (2010).
57 Enfield, K. S. *et al.* MicroRNA gene dosage alterations and drug response in lung cancer. *Journal of biomedicine & biotechnology* **2011**, 474632, doi:10.1155/2011/474632 (2011).
58 Gilad, S. *et al.* Serum microRNAs are promising novel biomarkers. *PloS one* **3**, e3148, doi:10.1371/journal.pone.0003148 (2008).
59 Mitchell, P. S. *et al.* Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10513-10518, doi:10.1073/pnas.0804549105 (2008).
60 Xia, Y., Zhu, Y., Zhou, X. & Chen, Y. Low expression of let-7 predicts poor prognosis in patients with multiple cancers: a meta-analysis. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **35**, 5143-5148, doi:10.1007/s13277-014-1663-0 (2014).
61 Wang, H., Peng, R., Wang, J., Qin, Z. & Xue, L. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. *Clinical epigenetics* **10**, 59, doi:10.1186/s13148-018-0492-1 (2018).
62 Nadal, E. *et al.* A MicroRNA cluster at 14q32 drives aggressive lung adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* **20**, 3107-3117, doi:10.1158/1078-0432.CCR-13-3348 (2014).
63 Teplyuk, N. M. *et al.* Therapeutic potential of targeting microRNA-10b in established intracranial glioblastoma: first steps toward the clinic. *EMBO molecular medicine* **8**, 268-287, doi:10.15252/emmm.201505495 (2016).

64 Beg, M. S. *et al.* Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors. *Investigational new drugs* **35**, 180-188, doi:10.1007/s10637-016-0407-y (2017).
65 Yang, G., Zhang, W., Yu, C., Ren, J. & An, Z. MicroRNA let-7: Regulation, single nucleotide polymorphism, and therapy in lung cancer. *Journal of cancer research and therapeutics* **11 Suppl 1**, C1-6, doi:10.4103/0973-1482.163830 (2015).
66 Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA biology* **9**, 703-719, doi:10.4161/rna.20481 (2012).
67 Han, Q. *et al.* Long noncoding RNA CRCMSL suppresses tumor invasive and metastasis in colorectal carcinoma through nucleocytoplasmic shuttling of HMGB2. *Oncogene* **38**, 3019-3032, doi:10.1038/s41388-018-0614-4 (2019).
68 Kanduri, C., Thakur, N. & Pandey, R. R. The length of the transcript encoded from the Kcnq1ot1 antisense promoter determines the degree of silencing. *The EMBO journal* **25**, 2096-2106, doi:10.1038/sj.emboj.7601090 (2006).
69 Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338**, 1469-1472, doi:10.1126/science.1228110 (2012).
70 Yap, K. L. *et al.* Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular cell* **38**, 662-674, doi:10.1016/j.molcel.2010.03.021 (2010).
71 Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nature reviews. Genetics* **14**, 880-893, doi:10.1038/nrg3594 (2013).
72 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).
73 Cao, L., Zhang, P., Li, J. & Wu, M. LAST, a c-Myc-inducible long noncoding RNA, cooperates with CNBP to promote CCND1 mRNA stability in human cells. *eLife* **6**, doi:10.7554/eLife.30433 (2017).
74 Faghihi, M. A. *et al.* Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome biology* **11**, R56, doi:10.1186/gb-2010-11-5-r56 (2010).
75 Kang, M. J. *et al.* HuD regulates coding and noncoding RNA to induce APP-->Abeta processing. *Cell reports* **7**, 1401-1409, doi:10.1016/j.celrep.2014.04.050 (2014).
76 Johnsson, P. *et al.* A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature structural & molecular biology* **20**, 440-446, doi:10.1038/nsmb.2516 (2013).
77 Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353-358, doi:10.1016/j.cell.2011.07.014 (2011).
78 Arab, K. *et al.* Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. *Molecular cell* **55**, 604-614, doi:10.1016/j.molcel.2014.06.031 (2014).
79 Gibb, E. A. *et al.* Human cancer long non-coding RNA transcriptomes. *PloS one* **6**, e25915, doi:10.1371/journal.pone.0025915 (2011).
80 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

81 Gutschner, T., Hammerle, M. & Diederichs, S. MALAT1 -- a paradigm for long noncoding RNA function in cancer. *Journal of molecular medicine* **91**, 791-801, doi:10.1007/s00109-013-1028-y (2013).
82 Wang, D. *et al.* LncRNA MALAT1 enhances oncogenic activities of EZH2 in castration-resistant prostate cancer. *Oncotarget* **6**, 41045-41055, doi:10.18632/oncotarget.5728 (2015).
83 Munschauer, M. *et al.* The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **561**, 132-136, doi:10.1038/s41586-018-0453-z (2018).
84 Zhou, K. *et al.* High long non-coding RNA NORAD expression predicts poor prognosis and promotes breast cancer progression by regulating TGF-beta pathway. *Cancer cell international* **19**, 63, doi:10.1186/s12935-019-0781-6 (2019).
85 Rossignol, F., Vache, C. & Clottes, E. Natural antisense transcripts of hypoxia-inducible factor 1alpha are detected in different normal and tumour human tissues. *Gene* **299**, 135-140, doi:10.1016/s0378-1119(02)01049-1 (2002).
86 Li, Z. *et al.* The long noncoding RNA THRIL regulates TNFalpha expression through its interaction with hnRNPL. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 1002-1007, doi:10.1073/pnas.1313768111 (2014).
87 Andersson, S. *et al.* Frequent gain of the human telomerase gene TERC at 3q26 in cervical adenocarcinomas. *British journal of cancer* **95**, 331-338, doi:10.1038/sj.bjc.6603253 (2006).
88 Nowak, T. *et al.* Amplification of hTERT and hTERC genes in leukemic cells with high expression and activity of telomerase. *Oncology reports* **16**, 301-305 (2006).
89 Beltran-Anaya, F. O., Cedro-Tanda, A., Hidalgo-Miranda, A. & Romero-Cordoba, S. L. Insights into the Regulatory Role of Non-coding RNAs in Cancer Metabolism. *Frontiers in physiology* **7**, 342, doi:10.3389/fphys.2016.00342 (2016).
90 Akers, J. C., Gonda, D., Kim, R., Carter, B. S. & Chen, C. C. Biogenesis of extracellular vesicles (EV): exosomes, microvesicles, retrovirus-like vesicles, and apoptotic bodies. *Journal of neuro-oncology* **113**, 1-11, doi:10.1007/s11060-013-1084-8 (2013).
91 Shi, T., Gao, G. & Cao, Y. Long Noncoding RNAs as Novel Biomarkers Have a Promising Future in Cancer Diagnostics. *Disease markers* **2016**, 9085195, doi:10.1155/2016/9085195 (2016).
92 Zhou, X., Yin, C., Dang, Y., Ye, F. & Zhang, G. Identification of the long non-coding RNA H19 in plasma as a novel biomarker for diagnosis of gastric cancer. *Scientific reports* **5**, 11516, doi:10.1038/srep11516 (2015).
93 Weber, D. G. *et al.* Evaluation of long noncoding RNA MALAT1 as a candidate blood-based biomarker for the diagnosis of non-small cell lung cancer. *BMC research notes* **6**, 518, doi:10.1186/1756-0500-6-518 (2013).
94 Liang, W. *et al.* Circulating long noncoding RNA GAS5 is a novel biomarker for the diagnosis of nonsmall cell lung cancer. *Medicine* **95**, e4608, doi:10.1097/MD.0000000000004608 (2016).
95 Arun, G., Diermeier, S. D. & Spector, D. L. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends in molecular medicine* **24**, 257-277, doi:10.1016/j.molmed.2018.01.001 (2018).
96 Li, C. H. & Chen, Y. Targeting long non-coding RNAs in cancers: progress and prospects. *The international journal of biochemistry & cell biology* **45**, 1895-1910, doi:10.1016/j.biocel.2013.05.030 (2013).

97 Bennett, C. F., Baker, B. F., Pham, N., Swayze, E. & Geary, R. S. Pharmacology of Antisense Drugs. *Annual review of pharmacology and toxicology* **57**, 81-105, doi:10.1146/annurev-pharmtox-010716-104846 (2017).
98 Qi, P. & Du, X. The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **26**, 155-165, doi:10.1038/modpathol.2012.160 (2013).
99 Gutschner, T. *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research* **73**, 1180-1189, doi:10.1158/0008-5472.CAN-12-2850 (2013).
100 Leone, S. & Santoro, R. Challenges in the analysis of long noncoding RNA functionality. *FEBS letters* **590**, 2342-2353, doi:10.1002/1873-3468.12308 (2016).
101 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11667-11672, doi:10.1073/pnas.0904715106 (2009).
102 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
103 Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Research* **42**, D749-D755, doi:10.1093/nar/gkt1196 (2014).
104 Zhang, Z., Wang, Q., Chen, F. & Liu, J. Elevated expression of HMGA1 correlates with the malignant status and prognosis of non-small cell lung cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **36**, 1213-1219, doi:10.1007/s13277-014-2749-4 (2015).
105 Sgarra, R. *et al.* High Mobility Group A (HMGA) proteins: Molecular instigators of breast cancer onset and progression. *Biochimica et biophysica acta. Reviews on cancer* **1869**, 216-229, doi:10.1016/j.bbcan.2018.03.001 (2018).
106 Fusco, A. & Fedele, M. Roles of HMGA proteins in cancer. *Nature reviews. Cancer* **7**, 899-910, doi:10.1038/nrc2271 (2007).
107 Stewart, G. L. *et al.* Aberrant Expression of Pseudogene-Derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma. *Frontiers in genetics* **10**, 138, doi:10.3389/fgene.2019.00138 (2019).
108 Li, H. *et al.* OIP5, a target of miR-15b-5p, regulates hepatocellular carcinoma growth and metastasis through the AKT/mTORC1 and beta-catenin signaling pathways. *Oncotarget* **8**, 18129-18144, doi:10.18632/oncotarget.15185 (2017).
109 Chun, H. K. *et al.* OIP5 is a highly expressed potential therapeutic target for colorectal and gastric cancers. *BMB reports* **43**, 349-354, doi:10.5483/bmbrep.2010.43.5.349 (2010).
110 Koinuma, J. *et al.* Characterization of an Opa interacting protein 5 involved in lung and esophageal carcinogenesis. *Cancer science* **103**, 577-586, doi:10.1111/j.1349-7006.2011.02167.x (2012).
111 Balbin, O. A. *et al.* The landscape of antisense gene expression in human cancers. *Genome research* **25**, 1068-1079, doi:10.1101/gr.180596.114 (2015).
112 Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annual review of pathology* **10**, 25-50, doi:10.1146/annurev-pathol-012414-040312 (2015).
113 Sheltzer, J. M. & Amon, A. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends in genetics : TIG* **27**, 446-453, doi:10.1016/j.tig.2011.07.003 (2011).

114 Presutti, D. *et al.* MET Gene Amplification and MET Receptor Activation Are Not Sufficient to Predict Efficacy of Combined MET and EGFR Inhibitors in EGFR TKI-Resistant NSCLC Cells. *PloS one* **10**, e0143333, doi:10.1371/journal.pone.0143333 (2015).
115 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).
116 Hillion, J. *et al.* Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer. *Molecular cancer research : MCR* **7**, 1803-1812, doi:10.1158/1541-7786.MCR-08-0336 (2009).
117 Khachane, A. N. & Harrison, P. M. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC genomics* **10**, 435, doi:10.1186/1471-2164-10-435 (2009).
118 Sun, M. *et al.* The Pseudogene DUXAP8 Promotes Non-small-cell Lung Cancer Cell Proliferation and Invasion by Epigenetically Silencing EGR1 and RHOB. *Molecular therapy : the journal of the American Society of Gene Therapy* **25**, 739-751, doi:10.1016/j.ymthe.2016.12.018 (2017).
119 Huang, J. L. *et al.* The long non-coding RNA PTTG3P promotes cell growth and metastasis via up-regulating PTTG1 and activating PI3K/AKT signaling in hepatocellular carcinoma. *Molecular cancer* **17**, 93, doi:10.1186/s12943-018-0841-x (2018).
120 Milligan, M. J. *et al.* Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Frontiers in genetics* **7**, 26, doi:10.3389/fgene.2016.00026 (2016).
121 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812, doi:10.1093/bioinformatics/btu393 (2014).
122 Luo, J. *et al.* The effects of aberrant expression of LncRNA DGCR5/miR-873-5p/TUSC3 in lung cancer cell progression. *Cancer medicine*, doi:10.1002/cam4.1566 (2018).
123 Dong, H. X., Wang, R., Jin, X. Y., Zeng, J. & Pan, J. LncRNA DGCR5 promotes lung adenocarcinoma (LUAD) progression via inhibiting hsa-mir-22-3p. *Journal of cellular physiology* **233**, 4126-4136, doi:10.1002/jcp.26215 (2018).
124 Chen, E. G., Zhang, J. S., Xu, S., Zhu, X. J. & Hu, H. H. Long non-coding RNA DGCR5 is involved in the regulation of proliferation, migration and invasion of lung cancer by targeting miR-1180. *American journal of cancer research* **7**, 1463-1475 (2017).
125 Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-1038, doi:10.1038/nature09144 (2010).
126 Feng, F., Qiu, B., Zang, R., Song, P. & Gao, S. Pseudogene PHBP1 promotes esophageal squamous cell carcinoma proliferation by increasing its cognate gene PHB expression. *Oncotarget* **8**, 29091-29100, doi:10.18632/oncotarget.16196 (2017).
127 Chen, L. *et al.* Citrate synthase expression affects tumor phenotype and drug resistance in human ovarian carcinoma. *PloS one* **9**, e115708, doi:10.1371/journal.pone.0115708 (2014).
128 Chen, X. *et al.* Overexpression of RCN1 correlates with poor prognosis and progression in non-small cell lung cancer. *Human pathology*, doi:10.1016/j.humpath.2018.08.014 (2018).
129 Kassiotis, G. Endogenous retroviruses and the development of cancer. *Journal of immunology* **192**, 1343-1349, doi:10.4049/jimmunol.1302972 (2014).

130 Drouin, G. Processed pseudogenes are more abundant in human and mouse X chromosomes than in autosomes. *Molecular biology and evolution* **23**, 1652-1655, doi:10.1093/molbev/msl048 (2006).
131 Glenfield, C. & McLysaght, A. Pseudogenes Provide Evolutionary Evidence for the Competitive Endogenous RNA Hypothesis. *Molecular biology and evolution* **35**, 2886-2899, doi:10.1093/molbev/msy183 (2018).
132 Chan, D. A. & Giaccia, A. J. PHD2 in tumour angiogenesis. *British journal of cancer* **103**, 1-5, doi:10.1038/sj.bjc.6605682 (2010).
133 Maldonado, M. D. M. & Dharmawardhane, S. Targeting Rac and Cdc42 GTPases in Cancer. *Cancer research* **78**, 3101-3111, doi:10.1158/0008-5472.CAN-18-0619 (2018).
134 Perdomo, R. *et al.* Early diagnosis of hydatidosis by ultrasonography. *Lancet* **1**, 244 (1988).
135 von Stechow, L. *et al.* The E3 ubiquitin ligase ARIH1 protects against genotoxic stress by initiating a 4EHP-mediated mRNA translation arrest. *Molecular and cellular biology* **35**, 1254-1268, doi:10.1128/MCB.01152-14 (2015).
136 Crooke, S. T., Witztum, J. L., Bennett, C. F. & Baker, B. F. RNA-Targeted Therapeutics. *Cell metabolism* **27**, 714-739, doi:10.1016/j.cmet.2018.03.004 (2018).
137 Kjems, J. & Howard, K. A. Oligonucleotide delivery to the lung: waiting to inhale. *Molecular therapy. Nucleic acids* **1**, e1, doi:10.1038/mtna.2011.1 (2012).
138 Jiang, H. *et al.* Long non-coding RNA TPT1-AS1 promotes cell growth and metastasis in cervical cancer via acting AS a sponge for miR-324-5p. *Journal of experimental & clinical cancer research : CR* **37**, 169, doi:10.1186/s13046-018-0846-8 (2018).
139 Xu, Y., Zheng, Y., Liu, H. & Li, T. Modulation of IGF2BP1 by long non-coding RNA HCG11 suppresses apoptosis of hepatocellular carcinoma cells via MAPK signaling transduction. *International journal of oncology* **51**, 791-800, doi:10.3892/ijo.2017.4066 (2017).
140 Zhang, Y. *et al.* Downregulation of long non-coding RNA HCG11 predicts a poor prognosis in prostate cancer. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* **83**, 936-941, doi:10.1016/j.biopha.2016.08.013 (2016).
141 Bak, R. O. & Mikkelsen, J. G. miRNA sponges: soaking up miRNAs for regulation of gene expression. *Wiley interdisciplinary reviews. RNA* **5**, 317-333, doi:10.1002/wrna.1213 (2014).
142 Tantai, J., Hu, D., Yang, Y. & Geng, J. Combined identification of long non-coding RNA XIST and HIF1A-AS1 in serum as an effective screening for non-small cell lung cancer. *International journal of clinical and experimental pathology* **8**, 7887-7895 (2015).
143 Fang, J., Sun, C. C. & Gong, C. Long noncoding RNA XIST acts as an oncogene in non-small cell lung cancer by epigenetically repressing KLF2 expression. *Biochemical and biophysical research communications* **478**, 811-817, doi:10.1016/j.bbrc.2016.08.030 (2016).
144 Dinescu, S. *et al.* Epitranscriptomic Signatures in lncRNAs and Their Possible Roles in Cancer. *Genes* **10**, doi:10.3390/genes10010052 (2019).
145 Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527-542 (1992).
146 Disteche, C. M. & Berletch, J. B. X-chromosome inactivation and escape. *Journal of genetics* **94**, 591-599 (2015).

147 Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38-44, doi:10.1038/349038a0 (1991).
148 Detassis, S., Grasso, M., Del Vescovo, V. & Denti, M. A. microRNAs Make the Call in Cancer Personalized Medicine. *Frontiers in cell and developmental biology* **5**, 86, doi:10.3389/fcell.2017.00086 (2017).
149 Chow, J. C. *et al.* Inducible XIST-dependent X-chromosome inactivation in human somatic cells is reversible. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 10104-10109, doi:10.1073/pnas.0610946104 (2007).
150 John, B. *et al.* Human MicroRNA targets. *PLoS biology* **2**, e363, doi:10.1371/journal.pbio.0020363 (2004).
151 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).
152 Ebert, M. S., Neilson, J. R. & Sharp, P. A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature methods* **4**, 721-726, doi:10.1038/nmeth1079 (2007).
153 Marshall, E. A., Stewart, G. L., Sage, A. P., Lam, W. L. & Brown, C. J. Beyond sequence homology: Cellular biology limits the potential of XIST to act as a miRNA sponge. *PloS one* **14**, e0221371, doi:10.1371/journal.pone.0221371 (2019).
154 He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855-1857, doi:10.1126/science.1163853 (2008).
155 Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* **40**, 939-953, doi:10.1016/j.molcel.2010.12.011 (2010).
156 Tutar, Y. Pseudogenes. *Comparative and functional genomics* **2012**, 424526, doi:10.1155/2012/424526 (2012).
157 Kahraman, M. *et al.* Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase. *Clinical chemistry* **63**, 1476-1488, doi:10.1373/clinchem.2017.271619 (2017).
158 Glinge, C. *et al.* Stability of Circulating Blood-Based MicroRNAs - Pre-Analytic Methodological Considerations. *PloS one* **12**, e0167969, doi:10.1371/journal.pone.0167969 (2017).
159 Minatel, B. C. *et al.* Large-scale discovery of previously undetected microRNAs specific to human liver. *Human genomics* **12**, 16, doi:10.1186/s40246-018-0148-4 (2018).
160 Sage, A. P. *et al.* Expanding the miRNA Transcriptome of Human Kidney and Renal Cell Carcinoma. *International journal of genomics* **2018**, 6972397, doi:10.1155/2018/6972397 (2018).
161 Martinez, V. D. *et al.* Discovery of Previously Undetected MicroRNAs in Mesothelioma and Their Use as Tissue-of-Origin Markers. *American journal of respiratory cell and molecular biology* **61**, 266-268, doi:10.1165/rcmb.2018-0204LE (2019).
162 Hofman, V., Heeke, S., Marquette, C. H., Ilie, M. & Hofman, P. Circulating Tumor Cell Detection in Lung Cancer: But to What End? *Cancers* **11**, doi:10.3390/cancers11020262 (2019).

Appendices

Appendix A Supplementary material

A.1 Supplementary tables from Chapter 3

Table A.1 Deregulated prospective *cis*-acting lncRNAs

Tag	Direction	Median Fold Change (BCCA)	BCCA BHC pValue	*cis*-gene
A2M-AS1	Downregulated	4.99278313	9.31E-07	A2M
AC002066.1	Downregulated	8.4705882	4.39E-07	CAV1
AC002398.12	Downregulated	4.94318134	0.0000961	HSPB6
AC003102.3	Upregulated	6.72640629	0.00038918	RUNDC3A
AC004490.1	Downregulated	6.12145001	1.49E-07	DOT1L
AC004540.4	Downregulated	7.18341517	1.14E-07	SNX10
AC004540.5	Downregulated	5.74574039	0.00000033	SNX10
AC005264.2	Downregulated	4.24878293	0.00011162	GNA15
AC005740.6	Downregulated	5.04918033	4.87E-07	PCDH12
AC005789.11	Upregulated	2.375	0.00593123	SPRED3
AC006014.8	Downregulated	3.88757978	0.00062122	STAG3L1
AC007128.1	Upregulated	18282726.1	0.00036335	ICA1, NXPH1, GLCCI1
AC007277.3	Downregulated	4.67095866	0.00069463	MYO3B
AC007405.4	Downregulated	9.00000001	0.00010601	ERICH2
AC007405.6	Downregulated	3.89614598	0.0000444	ERICH2
AC007743.1	Downregulated	28.6394546	1.14E-07	CCDC85A
AC007750.5	Upregulated	2.91705041	0.00124182	FAP,GCG
AC007970.1	Downregulated	29.1722945	0	LANCL1
AC009005.2	Upregulated	2.95873543	0.02763744	BSG
AC010226.4	Downregulated	4.44841423	0.0000563	TMED7
AC010547.9	Downregulated	299.824543	0.0000321	ZNF19
AC010890.1	Upregulated	2	0.00298985	NCKAP5
AC011899.9	Downregulated	14.5490932	0	PTPRN2
AC012594.1	Downregulated	2.11530977	0.00303756	MYO3B
AC013264.2	Downregulated	6.99999999	0.000027	ANKRD44
AC016683.6	Downregulated	4.1058999	0.00072576	PAX8
AC026703.1	Downregulated	4.96545096	5.15E-07	NPR3

AC066593.1	Upregulated	3	0.01522016	DPP10
AC079210.1	Downregulated	3.36386868	0.0000264	FAM20A
AC083949.1	Downregulated	3.07382	0.0000647	EML4
AC090616.2	Downregulated	9.75046883	1.49E-07	RHOT1
AC093110.3	Downregulated	43.5289949	0	SPTBN1
AC093495.4	Downregulated	5.86308398	0.00000694	XPC
AC096670.3	Downregulated	12	0.00000033	ACOXL
AC097658.1	Downregulated	8.29400234	0.000002	GAB1
AC099850.1	Downregulated	9.73761196	0.00000345	SKA2
AC100830.3	Downregulated	3.5	0.0000621	OAZ2
AC124789.1	Downregulated	22.2039099	7.38E-08	ARHGAP23
ADAMTS9-AS1	Downregulated	42.3976211	0	ADAMTS9
ADAMTS9-AS2	Downregulated	21.9956398	0	ADAMTS10
ADIRF-AS1	Downregulated	14.5878435	0	ADIRF
ADORA2A-AS1	Downregulated	2.48939696	0.00073329	ADORA2A
AF131215.2	Downregulated	9.76271182	1.49E-07	XKR6
AFAP1-AS1	Upregulated	23.6712582	0	AFAP1, SORCS2
AGAP1-IT1	Downregulated	2.42857143	0.00320442	AGAP1
AL022476.2	Downregulated	2.87720175	0.00031687	TTLL1
AL049840.1	Downregulated	6.00663784	0.0001322	XRCC3
AL163636.6	Downregulated	8.25288088	0.0000013	ANG
AL356356.1	Downregulated	4.48040053	0.00025988	ADAMTSL4
AL358113.1	Upregulated	506282.799	0.0000159	TJP2
AL591684.1	Downregulated	16.5069495	0.0000295	ANXA8
AL603965.1	Downregulated	83.3838014	6.55E-07	ANXA8L2
ALG1L13P	Downregulated	7.23894278	3.07E-07	FAM86B3P
AP000322.54	Downregulated	5.45454546	0.00000243	SMIM11
AP000330.8	Upregulated	5.99836919	0.00000175	RUNX1, RCAN1, CLIC6
AP000525.9	Upregulated	3.62479848	0.00257119	DUXAP8
AP000662.4	Downregulated	8.07528954	0.0000176	YPEL4
AP000695.4	Upregulated	2	0.01408789	CLDN14
AP001065.2	Upregulated	13.6065886	0.04932941	TRPM2
AP001189.4	Downregulated	13	0	LRRC32
AP003026.1	Downregulated	2.49896563	0.00774901	DLG2
ATP13A4-AS1	Downregulated	3.85996	0.0000884	ATP13A4
ATP5F1P5	Downregulated	9.9	4.39E-07	ATP5F1P6

BCRP3	Downregulated	6.72922396	0.00092449	BCRP4
C1orf170	Upregulated	13.8841333	1.49E-07	PLEKHN1
C21orf128	Downregulated	5.12195123	0.0000828	UMODL1
C5orf56	Downregulated	4.00357782	0.0000227	IRF1
C6orf165	Downregulated	32.4549793	0.00973151	SLC35A1
CASC9	Upregulated	16.4383	0.00706918	CASC9
CCDC13-AS1	Downregulated	3.22517461	0.0000824	CCDC13
CCL15-CCL14	Downregulated	17.3095032	0.0000133	CCL15
CEBPA-AS1	Downregulated	4.90251126	0.00093249	CEBPA
CHIAP2	Downregulated	75.7752279	0	CHIAP2
CMAHP	Downregulated	4.35784012	0.0000891	CMAHP
COLCA1	Downregulated	15.3927404	0.00000281	COLCA2
CTA-134P22.2	Downregulated	16.4154455	0	CADM3
CTB-131B5.2	Downregulated	3.5	0.00016815	CYSTM1
CTB-134H23.3	Downregulated	41.7822538	0	RRN3P2
CTB-176F20.3	Downregulated	2.47368421	0.03855288	ZNF91
CTB-36H16.2	Downregulated	6.84415584	0.0000128	SNX2
CTB-50L17.14	Downregulated	4.46111796	0.00000163	LRG1
CTB-55O6.12	Downregulated	4.72625753	7.16E-07	LPHN1
CTB-55O6.4	Upregulated	2	0.00017365	RLN3
CTB-60B18.6	Downregulated	943396226	0.03046148	CGB
CTC-250I14.3	Upregulated	2.51718028	0.00642486	NACC1, STX10
CTC-255N20.1	Upregulated	2	0.03276568	STK32A
CTC-297N7.5	Downregulated	2.94728044	0.00000754	TMEM220
CTC-558O2.2	Downregulated	17.681967	1.85E-07	SLIT3
CTD-2033A16.3	Upregulated	309372122	0.0000766	NOB1, NFAT5, WWP2
CTD-2036P10.3	Downregulated	20746.888	0.00000425	TTBK2
CTD-2114J12.1	Upregulated	2.36189428	0.00623942	LRRC69
CTD-2207O23.10	Downregulated	218340.611	1.85E-07	PNPLA6
CTD-2207O23.3	Downregulated	310559006	1.85E-07	ARHGEF18
CTD-2314B22.3	Upregulated	2.06623421	0.00611609	DUXAP10

CTD-2319I12.2	Downregulated	2.11641332	0.00139631	HEATR6
CTD-2524L6.3	Downregulated	13.9993465	0.00012034	NR2E3
CTD-2527I21.4	Downregulated	74.9829957	0.00049933	FXYD1
CTD-2561B21.7	Upregulated	4929957469	0.00196435	CHMP6
CTD-2562J17.7	Downregulated	8.13401585	5.15E-07	ARRB1
CTD-2636A23.2	Downregulated	4.31053986	0.00012912	HMGCS1
CTD-3076O17.1	Upregulated	2.03031518	0.00986304	ADAMTS17
CTD-3105H18.14	Downregulated	2.53303703	0.00810136	ZNF799
CTD-3148I10.15	Downregulated	4524.82545	0.00165901	FLT3LG
CTD-3187F8.14	Downregulated	2.82514319	0.02196737	SIGLEC7
CTD-3214H19.16	Downregulated	5756.25705	0	TRAPPC5
CYP1B1-AS1	Downregulated	3.42857142	0.00012807	CYP1B1
CYP4F29P	Downregulated	2.38815909	0.00124182	CYP4F29P
DGCR5	Upregulated	3.10672322	0.00068241	DGCR5
DGCR9	Upregulated	5.31922816	0.00327422	DGCR9
DIO3OS	Downregulated	12.7827336	0	DIO3OS
DNAJC9-AS1	Downregulated	3.16173368	0.00346898	DNAJC9
DNM1P46	Downregulated	105485.232	0.00000268	DNM1P47
EPB41L4A-AS2	Downregulated	7.87368419	5.56E-07	EPB41L4A
EPHA1-AS1	Downregulated	5.57070425	0.00000473	EPHA1
ERICH2	Downregulated	2.18397626	0.00094987	ERICH2
FABP5P7	Downregulated	5.76451536	0.00000518	FABP5P7
FAM138A	Upregulated	8645.65772	0.02739795	FAM138F
FAM182A	Downregulated	4.2590818	0.00041186	FAM182A
FAM66C	Downregulated	7.06605283	0.0000121	FAM66C
FAM83A-AS1	Upregulated	5.96489641	0.00000425	FAM83A
FAM85A	Downregulated	7.74193546	7.38E-08	FAM85A
FAM86JP	Upregulated	3.74171985	0.01385637	ALG1L
FEZF1-AS1	Upregulated	9.62390064	0.00000184	FEZF1
FGF14-AS2	Downregulated	7.05527637	7.38E-08	FGF14
FLG-AS1	Downregulated	2.17470686	0.00012265	FLG

FLI1-AS1	Downregulated	7.61531087	0.00000251	FLI1
FLJ00388	Upregulated	3.07E+14	0.04194006	TRABD2B
FLJ20373	Downregulated	3.53830096	0.00038683	MAP4K4
FTO-IT1	Downregulated	7.46666669	5.27E-07	FTO
FZD10-AS1	Downregulated	29.3273004	0.00038683	FZD10
GAS6-AS2	Downregulated	6.14634145	0.0000312	GAS6
GATA6-AS1	Downregulated	6.31568645	0.00000694	GATA6
GGT3P	Downregulated	2.40649476	0.00888159	GGT3
GGTA1P	Downregulated	6.9976291	0.00000425	GGTA1
GRTP1-AS1	Downregulated	5.09986955	0.00000209	GRTP1
GTF2IRD1P1	Upregulated	16.8368972	0.00218852	RABGEF1
GVINP1	Downregulated	11.0946627	0.00000384	GVINP2
HBZP1	Upregulated	1620.50547	0.00551042	HBM
HERC2P3	Downregulated	231.857714	6.55E-07	HERC2P4
HHIP-AS1	Downregulated	11.6507031	0.00000694	HHIP
HM13-AS1	Downregulated	2.19808047	0.00611609	HM13
HNRNPA1P33	Downregulated	8.47330887	0.00000177	HNRNPA1P34
HSPB2-C11orf52	Downregulated	23.2927601	1.49E-07	CRYAB
IGHG4	Upregulated	3.08906608	0.00046879	IGHE
IMPDH1P10	Downregulated	176.054075	0.00046879	CFLAR
INMT-FAM188B	Downregulated	584795.322	1.14E-07	FAM188B
KM-PA-2	Upregulated	3.05904156	0.00014867	SCXB, MROH1
KRT18P12	Upregulated	86.684056	0.04164009	PTPN14
KRT8P9	Upregulated	3.33508045	0.0000304	LRRC49
LA16c-366D3.1	Downregulated	5.62915447	3.07E-07	LMF1
LBX1-AS1	Downregulated	3	0.00000215	LBX1
LIMD1-AS1	Downregulated	4.54678981	0.00000518	LIMD1
LINC00337	Upregulated	3.9994	0.00380086	ICMT
LL0XNC01-250H12.3	Downregulated	6.78185687	0.0000159	RAB40A
LL22NC03-86G7.1	Downregulated	4.18028034	6.55E-07	PPM1F
LY6G6E	Upregulated	2.78E+14	0.00011921	LY6G6E
MAGI2-AS3	Downregulated	12.6956306	0	MAGI2
MAMDC2-AS1	Downregulated	4.36130862	0.00103839	MAMDC2
MBL1P	Downregulated	4.21329572	0.00110557	MBL1P
MBNL1-AS1	Downregulated	7.07364952	0.00000103	MBNL1

MEGT1	Upregulated	12.8600327	0.00000455	LY6G6F, LY6G6E
MIR600HG	Downregulated	5.73430147	0.0000133	STRBP
NAV2-AS2	Downregulated	5.07240825	2.42E-07	NAV2
OIP5-AS1	Downregulated	4.61153189	0.000052	OIP5
OR2A20P	Downregulated	5.6551903	0.0000348	OR2A1
OR7E14P	Upregulated	2.12794267	0.00029623	OR7E14
PAN3-AS1	Downregulated	3.62068966	0.0000155	PAN3
PFN1P2	Downregulated	2.05551409	0.00698419	NBPF20
PGM5-AS1	Downregulated	6.59754999	0	PGM5
PPP1R14BP3	Upregulated	2.04527204	0.01942068	ELF2
PRKCQ-AS1	Downregulated	8.45562993	0.0000175	PRKCQ
PSMD6-AS2	Downregulated	3.67241283	0.0000101	PSMD6
RAMP2-AS1	Downregulated	51.0819202	0	RAMP2
RBM26-AS1	Downregulated	4.87339664	0.0007718	RBM26
RBPMS-AS1	Downregulated	5.12034542	0.00000083	RBPMS
RCC2P4	Downregulated	2	0.00326354	CHCHD6
RLIMP2	Downregulated	12422360.2	0.000041	LRIG2
RN7SL5P	Upregulated	7.83343797	0.00218852	PTPRD
RP1-170O19.22	Downregulated	688.471476	0.00033945	HOXA4
RP1-178F15.5	Downregulated	6.53739521	0.00000425	S100A1
RP1-18D14.7	Downregulated	14.4616734	0	TAL1
RP1-257A7.5	Downregulated	4.08391608	0.0000268	PHACTR1
RP1-310O13.7	Downregulated	9.81645516	0	CCM2L
RP1-78O14.1	Downregulated	19	0	SYT1
RP1-85F18.5	Downregulated	5.99999999	0.0000014	EP300
RP11-1002K11.1	Downregulated	13.3	4.39E-07	NRG1
RP11-1079K10.4	Upregulated	1400.70571	0.00353288	PHB
RP11-1090M7.1	Downregulated	6.202838	4.56E-07	ARHGAP44
RP11-122M14.1	Downregulated	7.875	0.00000083	NEK2
RP11-125B21.2	Downregulated	4.58149414	2.79E-07	VLDLR
RP11-127I20.5	Upregulated	3.06870667	0.00074099	SMIM22
RP11-1293J14.1	Downregulated	3	0.00000274	ADI1
RP11-129B22.1	Downregulated	7.55065924	6.55E-07	PRICKLE2

RP11-140H17.1	Downregulated	17953321.4	0.00000469	CYB5B
RP11-141J13.5	Downregulated	50.4950484	0	SMTNL2
RP11-146D12.2	Downregulated	8.66138157	0.00018456	ANKRD20A3
RP11-152P17.2	Upregulated	7.3165	0.0000368	ZFPM2
RP11-158M2.3	Downregulated	3.52149655	0.00000765	AKAP13
RP11-161H23.9	Downregulated	10.0206279	9.58E-07	PRPH
RP11-164J13.1	Downregulated	7.36517061	0.0000111	CAPN3
RP11-16K12.1	Downregulated	6.24320786	0.00022627	RASGRF1
RP11-170N16.3	Downregulated	4.87179487	0.00013007	FGF2
RP11-175K6.1	Downregulated	5.00108674	2.42E-07	EBF1
RP11-178L8.9	Downregulated	2	0.01173613	FBX031
RP11-182J1.1	Downregulated	13.0180573	1.85E-07	SCAND2P
RP11-192H23.6	Upregulated	58.6600343	0.03470539	SGK494
RP11-203J24.9	Downregulated	3.00952938	0.00041186	AK1
RP11-20I23.1	Downregulated	2.77490501	0.00012912	ATP6V0C
RP11-215G15.5	Downregulated	13.6187563	0.00000614	ANKRD33B
RP11-218M22.1	Downregulated	6.72450684	2.12E-07	NINJ2
RP11-23J9.4	Downregulated	8.08580141	4.39E-07	CCDC180
RP11-245J9.5	Downregulated	2.76	0.00049493	PSMD6
RP11-251M1.1	Downregulated	24.6262626	0	EGFL7
RP11-264F23.3	Downregulated	3	0.00020069	CCND2
RP11-275I14.4	Downregulated	5.74110583	0.000051	ACBD3
RP11-276H19.1	Downregulated	19.4871797	0	GAS1
RP11-283G6.4	Downregulated	2.22222222	0.00106582	SSPN
RP11-284F21.10	Upregulated	2.50373766	0.00308919	BCAN
RP11-284F21.7	Upregulated	2.11237062	0.00759476	BCAN

RP11-284F21.9	Upregulated	2	0.00552643	BCAN
RP11-284N8.3	Downregulated	3.86266178	0.00020872	KCNA3
RP11-286E11.1	Downregulated	3.36751757	0.00013845	SGMS2
RP11-286H15.1	Downregulated	15.173709	0	MYO7B
RP11-287D1.3	Downregulated	4.48080313	0.00053253	MTHFD2
RP11-290D2.6	Downregulated	333333.333	0.00000083	TPT1
RP11-295G20.2	Upregulated	2.00564593	0.00582219	DISC1, TSNAX
RP11-2B6.2	Downregulated	5.0000327	0.00000222	MIRLET7-DHG
RP11-302F12.1	Downregulated	2.625	0.03109525	SLC34A2
RP11-304L19.3	Upregulated	2.5	0.00567903	PKD1
RP11-304L19.8	Downregulated	1179.74751	0.00642486	PGP
RP11-307C18.1	Downregulated	4	0.0000168	BAIAP2L1
RP11-309N17.4	Downregulated	19	0	HID1
RP11-314B1.2	Downregulated	4.42105264	0.0000751	NYAP2
RP11-317J10.2	Downregulated	2	0.00076564	CA3
RP11-318A15.7	Downregulated	428.733183	0.00479286	MFSD11
RP11-324E6.9	Upregulated	4.15318	0.01382385	HCAR1
RP11-325F22.5	Downregulated	9.23006133	0.0000438	LHFPL3
RP11-326F20.5	Downregulated	3.59251192	0.00000311	B4GALT1
RP11-328C8.4	Downregulated	10.5336627	0.000048	PRICKLE1
RP11-332H18.4	Downregulated	11.9068133	0	TBX2
RP11-334E6.3	Upregulated	7.61950936	0.00051008	USP2, RNF26, MFRP, C1QTNF5, THY1
RP11-343N15.5	Downregulated	4.44755718	0.0000297	SRGAP2B
RP11-344E13.3	Downregulated	8.35397622	0.00000568	CCDC144NL
RP11-345P4.10	Upregulated	542.71028	0.04763161	SLC35E2B, CDK11B
RP11-350J20.12	Upregulated	2.06558949	0.01124402	LRP11

RP11-354E11.2	Downregulated	6.53850735	1.14E-07	C10ORF112
RP11-356K23.1	Downregulated	10.0038485	0.00000384	FOXN3
RP11-359E3.4	Downregulated	4.0261438	0.00016146	BMPR1A
RP11-360L9.7	Upregulated	2.69410912	0.01165937	GINS4
RP11-366L5.1	Downregulated	5	9.58E-07	SSFA2
RP11-373N22.3	Downregulated	3.88390648	0.00049819	SPINK13, FBXO38
RP11-382D8.3	Downregulated	2	0.00703491	GGPS1
RP11-388P9.2	Downregulated	3.66864289	0.00000818	ANK3
RP11-389C8.2	Downregulated	11.2087912	7.38E-08	ZNF366
RP11-38L15.3	Downregulated	9.39408058	0	SYT15
RP11-38P22.2	Downregulated	8.17791526	0.00000103	P2RY1
RP11-390F4.3	Upregulated	2.02978481	0.00084593	KDM4C
RP11-391L3.1	Downregulated	8.65180333	0.0299186	GAN
RP11-391L3.5	Downregulated	2	0.02878929	CMIP
RP11-391M1.4	Downregulated	3.40236687	0.0000593	RPL14
RP11-3K24.1	Downregulated	2	0.0180897	USP32
RP11-401P9.6	Downregulated	8	0.0000132	NKD1
RP11-403A21.1	Downregulated	8	1.14E-07	LAMA3
RP11-425I13.3	Downregulated	3.5	0.0000999	SCOC
RP11-426C22.5	Downregulated	4.55667048	0.00000922	RRN3P2
RP11-426L16.3	Downregulated	2.03279882	0.0000345	MOV10
RP11-435I10.3	Upregulated	14.0306036	0.0017061	EIF3C
RP11-449P15.1	Downregulated	4.83397474	0.0000827	GPR146
RP11-452C13.1	Downregulated	4	0.00000194	PTPRN2
RP11-455O6.2	Downregulated	9.00000001	1.14E-07	AZI1
RP11-465L10.10	Upregulated	3.17114259	0.00081961	MMP9
RP11-465N4.4	Upregulated	1.91719106	0.01942068	ELF3, RNPEP
RP11-473M20.9	Downregulated	7.45359457	0.0000121	ZSCAN10
RP11-480G7.1	Downregulated	3	0.0000347	MYLK3
RP11-482H16.1	Downregulated	9.66134372	0.0000321	CCDC85A

RP11-486B10.4	Downregulated	2.49044586	0.01141125	HIVEP3
RP11-488C13.6	Downregulated	2	0.00935694	VASH1
RP11-489D6.2	Upregulated	2.29155448	0.00216376	RYR3
RP11-505K9.4	Downregulated	2	0.00552883	MLYCD, OSGIN1
RP11-507K2.3	Downregulated	4	6.89E-07	PTPN21
RP11-509E16.1	Downregulated	6.99817184	0.0002222	EFR3B
RP11-510N19.5	Upregulated	2.04726226	0.00411536	ELF3
RP11-513I15.6	Downregulated	4.28406737	0.00000384	NUDT3, **HMGA1**
RP11-513M16.8	Downregulated	3.45476708	0.00027769	RPS6
RP11-529H20.5	Downregulated	2.61236238	0.00056669	ATXN3
RP11-531A24.5	Downregulated	4.20689655	0.00035946	SBSPON
RP11-541N10.3	Downregulated	6.31479736	1.14E-07	OBFC1
RP11-546B15.2	Downregulated	3.08878732	0.00027769	GANC
RP11-54F2.1	Downregulated	11	9.58E-07	ANKRD33B
RP11-54O7.14	Upregulated	1.11E+15	0.00353288	AGRN
RP11-558F24.4	Upregulated	2.30897436	0.00031986	PIK3CD
RP11-566K11.4	Upregulated	30.450495	0.000052	MC1R, TUBB3
RP11-566K19.5	Downregulated	3.5927	0.00053598	NIPA1
RP11-582J16.4	Downregulated	5.22931214	2.12E-07	PPP3CC
RP11-594N15.3	Downregulated	15.483871	7.38E-08	PKIA
RP11-598F7.1	Upregulated	1.8E+16	0.0120053	FAM138D
RP11-598F7.3	Downregulated	12.48	0	IQSEC3
RP11-598F7.4	Downregulated	4.73618215	5.56E-07	IQSEC3
RP11-613D13.5	Downregulated	4.02198091	0.00000226	HSD17B12
RP11-613D13.8	Downregulated	12.0031914	0	C11ORF96
RP11-624G17.3	Upregulated	58362344.7	0.0007718	RTN4RL2

RP11-627G18.1	Downregulated	9.45203938	0.00000117	GATA6
RP11-62H7.2	Downregulated	2.72727272	0.02408826	SNORA70
RP11-649E7.5	Upregulated	2.22321407	0.00231195	MGAT2
RP11-64D22.2	Downregulated	3.24368387	0.0000117	AADACL2
RP11-650L12.2	Upregulated	4.5876573	0.00168281	CHRNA5
RP11-666A8.9	Downregulated	8.41124784	1.85E-07	SNHG16
RP11-677M14.3	Downregulated	8.12769103	0.00000204	ESAM
RP11-680F20.12	Downregulated	3.45559314	0.0000159	CDON
RP11-680F8.3	Downregulated	6.38747552	0	TJP1
RP11-688G15.3	Upregulated	28.3487406	0.00000425	CCDC85C
RP11-690D19.3	Downregulated	6.404253	0.00000342	DCUN1D5
RP11-6O2.3	Downregulated	5.23015832	0.0000111	TTC23
RP11-6O2.4	Downregulated	54.4404014	5.15E-07	SYNM
RP11-70C1.1	Downregulated	3	4.56E-07	CCDC13
RP11-710C12.1	Downregulated	8	0	UNC5C
RP11-714G18.1	Downregulated	4.16000001	0.0000154	LRP2BP
RP11-71H17.7	Downregulated	5.58615003	0.0000348	KALRN
RP11-720L2.4	Downregulated	4.90129202	0	COLEC12
RP11-723O4.6	Downregulated	7.93709408	0.00000425	ACAD9
RP11-724O16.1	Downregulated	2.3575346	0.01644789	BBS5
RP11-736K20.5	Downregulated	11.8309858	7.38E-08	PRSS23
RP11-74E22.5	Downregulated	8307.64885	0.00231195	PAFAH1B1
RP11-750H9.5	Downregulated	13.3749201	7.38E-08	SLC39A13
RP11-75C9.1	Downregulated	14.9967827	0.00000033	PTPRD
RP11-77A13.1	Downregulated	30.1935485	0	MARCO
RP11-783K16.5	Upregulated	6.19957	7.51E-07	PPP1R14B, VEGFB,FKBP2
RP11-787I22.3	Downregulated	7.37729591	0.00000518	SESN1
RP11-78A19.3	Upregulated	3.624538	0.04763161	CHMP1B, GNAL
RP11-793H13.10	Downregulated	2.95661213	0.0000208	ATF7
RP11-793H13.3	Downregulated	8.16036801	0.000041	ATF7

RP11-797A18.6	Downregulated	4.16000001	0.0000517	TSPAN3
RP11-800A3.4	Downregulated	4.39125356	0.000029	P2RY2
RP11-800A3.7	Downregulated	7.13971409	6.55E-07	ARHGEF17
RP11-802O23.3	Downregulated	2.43498424	0.00335538	PDHB
RP11-82L18.4	Downregulated	13.125	2.42E-07	SHC3
RP11-82L2.1	Downregulated	6.99999999	2.42E-07	SMC2
RP11-830F9.6	Downregulated	10	7.38E-08	CBFA2T3
RP11-845C23.3	Downregulated	1876172608	0.00000083	NEDD4L
RP11-875O11.1	Downregulated	9.78884847	0	RHOBTB2
RP11-89B16.1	Downregulated	5.99999999	2.42E-07	FIP1L1
RP11-96C23.14	Downregulated	6.85892367	7.38E-08	ADIRF
RP11-96C23.5	Downregulated	19.2875202	0	ADIRF
RP11-986E7.7	Downregulated	15.000105	0.00000103	SERPINA3
RP11-98D18.15	Downregulated	4273504.27	0.00016013	MRPL9
RP13-514E23.2	Downregulated	3.02250968	0.00098133	ARHGAP24
RP13-580F15.2	Downregulated	1.9776915	0.02024785	SPNS3
RP13-638C3.4	Downregulated	424.390935	0.03738224	FOXK2
RP3-323P13.2	Downregulated	20.484355	2.12E-07	EYA4
RP3-340N1.5	Upregulated	2	0.04086601	UBXN10
RP3-395M20.8	Downregulated	2.35108159	0.03108627	TNFRSF14
RP3-497J21.1	Upregulated	2.021315	0.02020871	RPS6KA2
RP3-525N10.2	Downregulated	17.9908334	0	BAI3
RP4-539M6.21	Downregulated	2	0.00343067	SEC14L2
RP4-548D19.3	Upregulated	498.459208	0.02771379	SMARCD3, CHPF2
RP4-559A3.7	Downregulated	48780.4878	0.0039065	LEFTY1
RP4-569M23.2	Downregulated	2.16666666	0.04235529	ZMYND8
RP4-576H24.4	Downregulated	4.37016	0.00000346	SIRPB1
RP4-607J23.2	Downregulated	4.19756427	0.00000304	BAIAP2L1
RP4-639F20.1	Downregulated	6.40940696	0.0000175	CNN3
RP4-728D4.2	Downregulated	4.38729139	0.0000735	PSMB2
RP4-755D9.1	Downregulated	3.19476685	0.01165937	RHOXF1

RP5-1007M22.2	Downregulated	5.04377001	0.000013	LRRC8B
RP5-1024N4.4	Downregulated	2.40861445	0.00547407	SLC5A9
RP5-1050D4.2	Downregulated	2.01357227	0.00922503	CAMTA2
RP5-1052I5.1	Downregulated	11.4521213	0.00000922	HS2ST1
RP5-1198O20.4	Downregulated	23.7845319	0	KLF17
RP5-940J5.9	Upregulated	164.16984	0.00000163	GAPDH
RPL13AP17	Downregulated	39.9579786	0.00000204	MAGI2
RPL23AP1	Downregulated	10.1630053	0.00000033	HLA-F
SEMA3B-AS1	Downregulated	4.44444444	0.00000211	SEMA3B
SH3RF3-AS1	Downregulated	13.8333333	3.07E-07	SH3RF3
SLC2A1-AS1	Upregulated	3.64487871	0.00016968	SLC2A1
SLC2A3P1	Downregulated	1.8248E+13	0.00000117	PANK3
SNRK-AS1	Downregulated	9752.76735	0.00973151	SNRK
SRGAP3-AS2	Downregulated	19.8917325	1.14E-07	SRGAP3
ST3GAL1P1	Downregulated	2	0.00882936	UBA6
ST7-AS1	Downregulated	3.42026631	0.00000765	ST7
STARD13-AS	Downregulated	2	0.00000494	STARD13
TBX5-AS1	Downregulated	11.8753446	1.14E-07	TBX5
TCAM1P	Upregulated	4	0.00039809	TCAM1P
TFAP2A-AS1	Upregulated	2.66367	0.000345	TFAP2A
TM4SF1-AS1	Upregulated	2.87412	0.01232631	TM4SF1
TMX2-CTNND1	Downregulated	5.2356E+15	1.49E-07	TMX2-CTNND1
TRHDE-AS1	Downregulated	11.6864	0.0000025	TRHDE
TRIM53BP	Downregulated	199.23335	0.00066796	TRIM53AP
TSNAX-DISC1	Upregulated	18.8172043	0.0111517	TSNAX, SPRTN, EGLN1, EXOC8, GNPAT, C1orf131, TRIM67, FAM89A, ARV1, TTC13, C1orf198, CAPN9, COG2, AGT, PGBD5, GALNT2, SIPA1L2, MAP10, NTPCR, PCNXL2
TTLL10-AS1	Downregulated	11.8644001	0	TTLL10
TUBA3FP	Downregulated	5.51918276	0.00000345	TUBA3F
UBE2CP3	Downregulated	2	0.00110071	IGFBP7
UPK3BP1	Downregulated	46.6606555	0	UPK3B
USP30-AS1	Downregulated	2.94339622	0.0000913	USP30

VIPR1-AS1	Downregulated	9.80397126	0	VIPR1
VNN3	Downregulated	26.4842425	0.00000345	VNN3
VWFP1	Downregulated	31.8310044	7.38E-08	TPTEP1
WASIR2	Upregulated	3.06987	0.00972207	WASIR2
WDFY3-AS2	Downregulated	5.85003364	0.00000576	WDFY3
WDR11-AS1	Downregulated	3.69277987	0.0000351	WDR11
WI2-1896O14.1	Downregulated	53.2327787	0.03108627	NBPF9
WWC2-AS2	Downregulated	5.85365348	3.07E-07	WWC2
WWTR1-AS1	Downregulated	6.08674639	0.0000013	WWTR1
XXbac-BPG254F23.6	Upregulated	35707.1629	0.00739268	HLA-DQB1
XXyac-YX155B6.6	Upregulated	193800421	0.00295904	NBPF8, NBPF24, NBPF11
ZEB2-AS1	Downregulated	16.0666264	0.0000469	ZEB2
ZFYVE9P1	Downregulated	8.9150628	0.000011	ZFYVE9
ZNF300P1	Downregulated	6.56657969	0.00000165	ZNF300
ZNF582-AS1	Downregulated	15.8323997	0.0000208	ZNF582
ZRANB2-AS1	Downregulated	2.78395765	0.000028	ZRANB2

Appendix B Description of published supplementary tables

B.1 Description of published supplementary tables from Chapter 4

Supplemental tables can be accessed in our published manuscript "Aberrant Expression of Pseudogene-Derived lncRNAs as an Alternative Mechanism of Cancer Gene Regulation in Lung Adenocarcinoma" due to the large size of the supplementary tables in Chapter 4 (DOI: 10.3389/fgene.2019.00138) [107].

Supplemental Table 1: Detailed information of all pseudogene-derived lncRNAs.

This table includes a database of common gene names for the lncRNAs, stable ENSEMBLE IDs, pseudogene ID's, direction of transcription (relative to parent gene), and exon-exon overlap.

Supplemental Table 2: Statistical analysis of the deregulated lncRNAs in the BCCA and TCGA cohorts (BHC-corrected p-values).

This table includes the pseudogene overlapping lncRNAs that were deregulated in the same direction in both of our matched LUAD datasets. Direction of deregulation, and BHC-corrected p-values are included.

Supplemental Table 3: Number of PUBMED entries for lncRNA cancer-association.

This table includes the number of published manuscripts in "cancer" or "lung cancer" for our deregulated pseudogene overlapping lncRNAs.

Supplemental Table 4: Genomic locations of significant lncRNAs and their respective parent genes.

This table includes the genomic location, and relative size of each deregulated lncRNA and parent gene pair.

Supplementary Table 5: Expression correlations between the lncRNAs and their respective parent genes in the BCCA and TCGA cohorts.

This table includes the p-values for the expression association of the deregulated lncRNAs and their respective parent genes in both cohorts of LUAD.

Supplementary Table 6: Spearman's correlation coefficients for global expression analysis of Ψ-lnc-parent pairs.

This table includes the Spearman's correlation coefficients used to plot the global correlation patterns of Ψ-lnc-parent pairs, and Ψ-lnc-random gene pairs. Also included is the Spearman's correlation coefficients used to plot the global correlation patterns of sense and antisense Ψ-lnc-parent pairs.

Supplementary Table 7: Parent gene information for pseudogenes contained in Retrolali5 and Yale60 databases.

This table includes all matched parent genes for the pseudogenes in 2 major databases, Retrolali5 and Yale60.

B.2 Description of published supplementary tables from Chapter 5

Supplemental tables can be accessed in our published manuscript "Beyond sequence homology: Cellular biology limits the potential of XIST to act as a miRNA sponge" due to the large size of the supplementary tables in this Chapter (DOI: 10.1371/journal.pone.0221371) [153].

Supplementary Table 3: miRNA binding prediction to DMX genes and XIST

This table includes the binding energies for each of the miRNAs predicted to bind both the DMX genes, as well XIST. This table also includes the statistics for the as the Spearman's correlation of DMX genes to XIST.