NON-CODING RNA EXPRESSION IN LUNG ADENOCARCINOMA: THE LONG AND THE SHORT OF IT

by

Adam Patrick Sage

B.Sc.H., Queen's University, 2016

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

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Interdisciplinary Oncology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2019

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

Non-coding RNA expression in lung adenocarcinoma: The long and the short of it

submitted by Adam Sage in partial fulfillment of the requirements for

the degree of Master of Science

in Interdisciplinary Oncology

Examining Committee:

Dr. Wan Lam
Supervisor

Dr. Carolyn Brown
Supervisory Committee Member

Dr. Miriam Rosin
Additional Examiner

Dr. Cathie Garnis
Oral Defence Chair

Additional Supervisory Committee Members:

Dr. Decheng Yang
Supervisory Committee Member

Supervisory Committee Member

Abstract

Lung cancer genomic profiling has led to the development of therapies targeting aberrant protein expression. However, the majority of patients present tumours with undruggable or unidentified driver mutations, highlighting the need for a new approach to discover lung cancer genes. Non-coding RNAs (ncRNAs) have emerged as critical regulators of cellular processes, such as proliferation, apoptosis, and the immune response; functions that can be perturbed in cancer. I take a global approach to explore the broad role of ncRNAs in lung tumours to uncover alternative regulatory mechanisms and potential therapeutic targets.

Non-coding RNAs are divided into two main categories based on their size: small (sncRNAs; <200nt) and long (lncRNAs; >200nt). I analyzed small RNA sequencing data from two cohorts of paired lung tumour and non-malignant tissue samples to identify previously-unannotated microRNAs (miRNAs). Using *in silico* algorithms and subsequent curation, I discovered 141 novel miRNA sequences, representing a substantial increase in the lung miRNA transcriptome. Not only were these transcripts specifically expressed in lung tissues, but they also displayed deregulated expression patterns in tumours and potential prognostic value. This strategy has been instrumental for miRNA discovery in tumours from other organs.

Immunotherapy has provided a promising treatment option for lung cancer, but the ability to predict treatment response is an emerging challenge. LncRNAs are known to have a significant role in the immune system. Several lncRNAs have described functions in lung tumours, but lncRNA expression in tumour-infiltrating lymphocytes remains uncharacterized. I determined the lncRNA expression patterns from purified human immune cell subsets, and delineated their cell-type specificity, which may contribute to their particular immune-regulatory

roles. These expression patterns were recapitulated in both bulk lung and in single-cell RNA sequencing data. My observations highlight the contribution of infiltrating immune cells to sequencing analyses and also the relevance of lncRNAs to the biology of the tumour microenvironment.

Together, my results emphasize the importance of high-throughput deep-sequencing efforts and lay a foundation for the discovery of novel genes involved in lung cancer biology. Assessment of ncRNAs represents the next frontier of cancer biology research and new opportunities for therapeutic target and biomarker discovery.

Lay Summary

Lung cancer is the leading cause of cancer-related death, but it is not solely a smoker's disease. Uncovering genetic alterations causing tumours has helped design new therapies, but for many patients these are ineffective. Thus, we need to find new genes driving tumours. Non-coding RNAs (ncRNAs) are largely unexplored as they do not encode proteins, but now have demonstrated important regulatory functions. However, the extent of their expression in human cells is poorly understood. I comprehensively describe ncRNA expression in lung tumours. I discovered 141 novel lung-specific genes that were altered in tumours and predict patient outcome. I further define ncRNAs specifically expressed in immune cells, which can also be detected in lung tumours. These analyses may be used to identify immune cells within tumours and new functional roles for ncRNAs. Collectively, my results underscore the necessity of in-depth molecular profiling and the undeniable roles of ncRNAs in cancer.

Preface

This thesis contains work of which I was principally responsible for both the design and execution. However, this would not have been possible without guidance from Dr. Wan Lam, as well as our collaborators, namely Dr. Ninan Abraham, Dr. Igor Jurisica, Dr. Carolyn Brown, and Dr. Decheng Yang. The members of the Wan Lam Lab were also integral to this thesis, through both supportive and physical roles. Particularly, Brenda Minatel worked in concert with me in our development of the platform for novel miRNA discovery and performed the cell culture of the RT-qPCR experiments (Chapter 3). Erin Marshall led the initial study on the sncRNA transcriptome of the NCI-60 cell line panel, which provided the rationale for Chapter 3, as well as provided helpful insight and guidance throughout my work. Kevin Ng aided in the design of my research question approached in Chapter 4, assisted with sequence analysis and identification of potentially relevant transcripts, and has provided invaluable support throughout. Finally, Dr. Victor Martinez provided assistance with the various statistical analyses and interpretation of raw sequencing data.

Work in this thesis was performed with ethics approval from the UBC Research Ethics Board, Certificate Numbers: EDRN H09-00008, CCSRI H09-00934, W81XW-10-1-0634.

In Chapter 1 and Chapter 3.1, I discuss concepts that are included in manuscripts that I have been published throughout my degree, listed here. In each of these 6 publications, I was part of the study design, data collection and analysis, as well as manuscript preparation. Co-first authorship is demarcated by asterisks (*).

- Guisier F*, Barros-Filho MC*, Rock LD*, Constantino FB, Minatel BC, **Sage AP**, Marshall EA, Martinez VD, Lam WL (2019). Small non-coding RNA expression in cancer. In *Gene Expression Profiling in Cancer*. Dimitrios Vlachakis ed., InTechOpen, London. In press.
- Martinez VD*, **Sage AP***, Marshall EA, Suzuki M, Goodarzi AA, Dellaire G, Lam WL (2018) Oncogenetics of lung cancer induced by environmental carcinogens. In *Oncogenes and Carcinogenesis*. Pinar Erkekoğlu ed., InTechOpen, London. 1-24.
- **Sage AP***, Martinez VD*, Minatel BC, Pewarchuk ME, Marshall EA, MacAulay GM, Hubaux R, Pearson DD, Goodarzi AA, Dellaire G, Lam WL (2018). Genomics and epigenetics of malignant mesothelioma. *High Throughput* 7:20, 1-23.
- Minatel BC*, **Sage AP***, Anderson C, Hubaux R, Marshall EA, Lam WL, Martinez VD (2018) Environmental arsenic exposure: genetic and epigenetic contributions to carcinogenesis. *Environment International* 112:183-97.
- **Sage AP***, Minatel BC*, Ng K, Stewart GL, Dummer TJB, Lam WL, Martinez VD (2017) Oncogenomic disruptions in arsenic-induced carcinogenesis. *Oncotarget* 8:25736-55.
- Marshall EA, **Sage AP**, Ng KW, Martinez VD, Firmino NS, Bennewith KL, Lam WL (2017) Small non-coding RNA transcriptome of the NCI-60 cell line panel. *Scientific Data* 4:170157, 1-8.

Chapter 3 has been presented at local, national, and international conferences and contains reference to data published in the following manuscript.

- **Sage AP**, Minatel BC, Marshall EA, Martinez VD, Stewart GL, Enfield KSS, Lam WL (2018) Expanding the miRNA transcriptome of human kidney and renal cell carcinoma. *International Journal of Genomics* 2018:1-10.

I designed the study and led data generation, analysis, and manuscript preparation. Brenda Minatel assisted in data analysis and optimization of the bioinformatics platform, as well as cell culture experiments. Erin Marshall, Dr. Victor Martinez, Greg Stewart, and Dr. Katey Enfield assisted with data analysis and manuscript preparation.

The data pertaining to the discovery of novel miRNAs in lung tissues in Chapter 3 is formatted as a manuscript to be submitted. I co-led the design of the study and was responsible for the majority of the data generation and analysis along with Brenda Minatel, with assistance from Dr. Victor Martinez and Erin Marshall.

- Minatel BC*, **Sage AP***, Martinez VD, Marshall EA, Reis PP, Lam WL (2019) Discovery of novel miRNAs in lung (Manuscript in preparation).

I have presented the work in Chapter 4 at local and international conferences, in both poster and oral presentations. The work was selected for the poster award at the BC Cancer Summit, in November of 2018. Additionally, it is formatted as a manuscript to be submitted.

- **Sage AP**, Ng KW, Marshall EA, Enfield KSS, Stewart GL, Martin SD, Minatel BC, Brown CJ, Abraham N, Lam WL (2018). Long non-coding RNA expression patterns delineate infiltrating immune cells in the lung tumour microenvironment (Manuscript in preparation).

Kevin Ng helped me to formulate the questions asked in this study, as well as with initial data generation. I performed the majority of the data analysis and have prepared the manuscript. Erin Marshall, Dr. Katey Enfield, Greg Stewart, Dr. Spencer Martin, and Brenda Minatel aided in data interpretation and provided day-to-day support.

Table of Contents

Abstract ..... **iii**
Lay Summary ..... **v**
Preface ..... **vi**
Table of Contents ..... **ix**
List of Tables ..... **xii**
List of Figures ..... **xiii**
List of Symbols and Abbreviations ..... **xv**
Acknowledgements ..... **xvii**
Dedication ..... **xviii**
Chapter 1: Introduction **1**
1.1 Lung cancer background 1
1.1.1 Health burden 1
1.2 Histological and cellular heterogeneity in lung cancer 2
1.2.1 Classification of lung cancer 2
1.2.2 The lung tumour microenvironment 3
1.3 Molecular heterogeneity in lung cancer and effects on treatment 4
1.3.1 Molecular alterations in the development of lung cancer 4
1.3.2 Current treatment options for lung cancer patients 7
1.4 Non-coding RNAs as emerging players in lung cancer 8
1.4.1 Next generation sequencing and non-coding RNAs 8
1.4.2 microRNAs 9
1.4.3 Long non-coding RNAs 12
1.5 Challenges in ncRNA research 15
1.6 Thesis themes 16
1.6.1 Rationale 16
1.6.2 Objectives, hypotheses, and specific aims 17
Chapter 2: Methods **20**
2.1 Patient tumour cohorts and collection of RNA sequencing data 20
2.1.1 BC Cancer Lung Adenocarcinoma Samples (BCCA-LUAD) 20
2.1.2 TCGA Lung Adenocarcinoma Samples (TCGA-LUAD) 21
2.1.3 TCGA Clear Cell Renal Cell Carcinoma Samples (TCGA-KIRC) 21
2.2 Processing of RNA sequencing data 23

2.2.1 Processing of small RNA sequencing data 23
2.2.2 Processing of long-read RNA sequencing data 23
2.3 Discovery of novel miRNA sequences 24
2.3.1 Description of predictive algorithms 24
2.3.2 Sequence detection and candidate curation 25
2.3.3 Statistical analyses of gene expression 26
Chapter 3: Discovery of previously-unannotated miRNAs specific to normal and malignant human tissues 28
3.1 Introduction 28
3.1.1 miRNAs in lung cancer 28
3.1.2 Previously-unannotated miRNAs 29
3.1.3 Small non-coding transcriptome of the NCI-60 cell line panel 30
3.2 Methods 31
3.2.1 Patient Samples and small RNA sequencing 31
3.2.2 Discovery and analysis of previously-unannotated miRNAs 31
3.2.3 Analysis of previously-unannotated miRNA expression 33
3.2.4 Validation in different contexts: human ccRCC samples and cell lines 33
3.2.5 Cell culture and real-time quantitative PCR (RT-qPCR) 34
3.3 Results 35
3.3.1 Expression of previously-unannotated miRNAs in human lung samples 35
3.3.2 Dysregulation of novel miRNAs in lung tumours and their clinical potential 40
3.3.3 Novel miRNAs target important protein interaction networks in lung tumours 44
3.3.4 Technical validation of the novel miRNA discovery platform 46
3.3.5 Novel miRNAs are similarly deregulated and associated with survival in ccRCC tumours as in lung adenocarcinoma samples 47
3.3.6 Confirmation of novel miRNA expression patterns *in vitro* 50
3.4 Discussion 52
Chapter 4: Examining the expression of long non-coding RNAs in infiltrating immune cells in the lung tumour microenvironment 55
4.1 Introduction 55
4.1.1 The lung cancer immune microenvironment and immunotherapy 55
4.1.2 Non-coding RNAs in immune cells and cancer immunology 56
4.2 Methods 59
4.2.1 Analysis of lncRNA expression in RNA sequencing data from sorted healthy human immune cells 59
4.2.2 Assessment of lncRNA expression patterns in bulk tumour data 61
4.2.3 Analysis of lncRNA expression in single-cell RNA sequencing data 62
4.3 Results 63

4.3.1 LncRNAs are specifically expressed in healthy human immune cells 63
4.3.2 Known and novel lncRNA expression patterns are indicative of function in immune cells 66
4.3.3 LncRNAs are deregulated in tumours but may result from tumour impurity 70
4.3.4 Immune-lncRNAs are co-expressed with immune cell markers in single cell RNA sequencing data isolated from bulk lung tumour samples 76
4.4 Discussion 81
Chapter 5: Conclusions 85
5.1.1 Summary and significance 85
5.1.2 Limitations and future directions 88
References 91
Appendix A: Publications 102

List of Tables

Table 2.1. Cohort sample characteristics 22

List of Figures

Figure 1.1. Epidemiological, histological, and molecular features of lung cancer. 6

Figure 1.2. General mechanisms of transcription and function for miRNAs and lncRNAs. 14

Figure 3.1. Sequence analysis flow chart for novel miRNA sequences discovered in human lung samples from the TCGA-LUAD and BCCA-LUAD cohorts, and subsequent validation in TCGA-KIRC patient samples and ccRCC cell lines. 32

Figure 3.2. Detection parameters and molecular features of novel miRNA candidates. 36

Figure 3.3. Lung-specific expression of previously-unannotated miRNAs revealed by principle components analysis. 39

Figure 3.4. Previously-unannotated miRNAs are differentially expressed between lung tumours and matched adjacent non-malignant tissues in the BCCA-LUAD cohort. 41

Figure 3.5. Prognostic utility of previously-unannotated miRNAs. 43

Figure 3.6. Pathways significantly enriched in genes targeted by novel miRNAs discovered in human lung samples. 45

Figure 3.7. Differential expression of novel miRNA sequences in ccRCC samples relative to non-malignant tissues. 49

Figure 3.8. Fold change values for novel miRNA candidates in ccRCC (TK-10) relative to non-malignant kidney (HEK-293T) cell lines. 51

Figure 4.1. Long non-coding RNAs involved in immune cell differentiation. 58

Figure 4.2. Analysis pipeline. 60

Figure 4.3. Distribution of lncRNAs expressed in healthy human immune cells. 64

Figure 4.4. Expression of lncRNAs in healthy human immune cells. 65

Figure 4.5. Expression patterns of long non-coding RNAs and protein-coding genes in healthy human immune cell subsets. 68

Figure 4.6. Genomic region (Chr19, q13.41) highlighting the location and orientation of *NKG7*, *SIGLEC10*, *AC008750.1*, and *AC008750.2* genes. 69

Figure 4.7. General dysregulation of lncRNAs between lung adenocarcinoma samples and matched adjacent non-malignant tissues. 72

Figure 4.8. Dysregulation of immune-associated lncRNAs in bulk tumour data in TCGA-LUAD (A) and BCCA-LUAD (B) cohorts. 73

Figure 4.9. LncRNA expression is confounded by infiltrating immune cells in bulk tumour data. 75

Figure 4.10. linc00649 expression across immune cell subsets..... 77

Figure 4.11. linc00861 expression in healthy immune cells and association with tumour-infiltrating lymphocytes. 80

List of Symbols and Abbreviations

3'UTR	3-prime untranslated region
Ago2	Argonaute 2
ASO	Antisense oligonucleotide
ATCC	American Type Culture Collection
BCCA	BC Cancer Agency
BCCA-LUAD	BC Cancer Lung adenocarcinoma sample cohort
BH	Benjamini-Hochberg
BLASTn	Standard nucleotide blast
ccRCC	Clear cell renal cell carcinoma
cDNA	Complementary DNA
cgHUB	Cancer Genomics Hub data repository
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR inactivation
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
E	Expect value
EGFR	Epidermal growth factor receptor
FBS	Fetal bovine serum
FFPE	Formalin-fixed paraffin-embedded
FPKM	Fragments per kilobase of transcript per million mapped reads
HOTAIR	HOX antisense intergenic RNA
HSC	Hematopoietic stem cell
IFN-γ	Interferon gamma
IHC	Immunohistochemistry
kb	Kilobase
Knm	Kidney novel miRNA
KRAS	Kirsten Rat Sarcoma virus protein
LCC	Large cell carcinoma
lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
LUMP	Leukocytes Unmethylation for Purity
LUSC	Lung squamous cell carcinoma
MALAT1	Metastasis-associated lung adenocarcinoma transcript 1
miRNA	microRNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA

NGS	Next generation sequencing
NIH	National Institute of Health
NK	Natural killer cells
NM	Non-malignant
NSCLC	Non-small cell lung carcinoma
ng	Nanogram
nt	Nucleotide
PD-1	Programmed cell death receptor 1
PD-L1	Programmed cell death ligand 1
piRNA	PIWI-interacting RNA
Pol II	RNA polymerase II
PRC2	Polycomb repressive complex 2
pri-miRNA	Primary-miRNA
Rfam	RNA Families database
RISC	RNA-induced silencing complex
RNA	Riobnucleic acid
RPM	Reads per million normalization method
RPMI	Roswell Park Memorial Institute medium
rRNA	Ribosomal RNA
RT-qPCR	Real-time quantitative Polymerase Chain Reaction
SCLC	Small cell lung carcinoma
scRNAseq	Single-cell RNA Sequencing
siRNA	Short-interfering RNA
sncRNA	Small non-coding RNA
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
SRA	Sequence Read Archive
STAR	Spliced Transcripts Alignment to a Reference algorithm
TAA	Tumour-associated antigen
TCGA	The Cancer Genome Atlas
TCGA-KIRC	TCGA clear cell renal cell carcinoma sample cohort
TCGA-LUAD	TCGA Lung adenocarcinoma sample cohort
TKI	Tyrosine kinase inhibitor
TMM	Weighted trimmed mean of log expression ratios normalization method
TP	Tumour
tRNA	Transfer RNA
UCSC	University of California Santa Cruz
v.	Version number
µL	Microlitre
Δ	Delta (change in)

Acknowledgements

Firstly, I would like to thank Dr. Wan Lam for his continuous guidance and fostering an environment where students feel encouraged, supported, and happy. Beyond instilling a sense of scientific rigour and broadening my expertise, he has taught me the importance of collaboration. He has given me the tools and confidence to work towards achieving my goals. Without his support I would not be where I am today. I am also sincerely appreciative of the guidance provided by my committee members, Drs. Carolyn Brown and Decheng Yang, whose valuable feedback helped this thesis come to fruition.

I would like to recognize the incredible collaborative environment facilitated by the past and present members of the Wan Lam Lab, and the other students at the BCCRC. Particularly those that were integral to all aspects of this thesis, particularly Erin, Brenda, Cassia, Kevin, Katey, Greg, Victor, and Matt, from the data analysis to their less-tangible support that kept me going throughout.

I have had the privilege to obtain support from various funding sources over the course of my degree. I sincerely appreciate the assistance provided by the Interdisciplinary Oncology Program, the University of British Columbia Faculty of Medicine Graduate Award and Graduate Student Travel Award, and the Canadian Institutes of Health Research Frederick Banting and Charles Best Canada Graduate Scholarship Masters Award. The work in this thesis was funded by the Canadian Institutes of Health Research.

Finally, I am very grateful for the enduring support offered by my family and Olivia throughout all aspects of this endeavour.

Dedication

To my family, and friends who are family.

Chapter 1: Introduction

1.1 Lung cancer background

1.1.1 Health burden

Despite major advancements in diagnostics, prognostics, and therapeutics, cancer represents an enormous burden on global health. A broad term that represents the unregulated growth of cells, cancer can arise in any bodily tissue, from the blood to the brain. However, each individual cancer presents a unique clinical challenge and requires different treatment approaches.

Known to be associated with exposure to tobacco smoke, lung cancer is the deadliest form of the disease to date, with a substantial impact on developing countries[1]. In Canada, lung cancer represented 14% of all new cancer cases, was marked by a dismal five-year survival of 14% in males and 20% in females, and accounted for 26% of all cancer-related deaths in 2017 (Figure 1.1A)[2]. However, lung cancer can also arise in individuals that have never smoked, a sub-group that represents nearly 25% of cases[3]. While facilitating the critical function of gas exchange, the lungs are continuously exposed to compounds in the environment. Arsenic, radon gas, asbestos fibres, and the fine particulate matter in air pollution have been shown to contribute to the development of lung cancer (Appendix A; Items 6, 12)[4-8]. Additionally, heritable factors may also affect individual susceptibility (Appendix A; Item 9)[9,10]. Thus, there remains an urgent need to understand the intricacies of this multifaceted disease, which will enable the design of strategies that will lead to better outcomes and quality of life for patients.

1.2 Histological and cellular heterogeneity in lung cancer

1.2.1 Classification of lung cancer

Lung cancer is not one, but many diseases as the lung is made up of a complex milieu of cells with varying but critical functions[11]. Broadly, lung cancer is separated by histology into small cell lung carcinoma (SCLC) which accounts for roughly 15% of cases, and non-small cell lung carcinoma (NSCLC) representing an overwhelming 85% of cases (Figure 1.1B)[12]. NSCLC can be further broken down into its most frequently observed histological categories: the most common lung adenocarcinoma (LUAD; ~50% of lung cancer cases), followed by lung squamous cell carcinoma (LUSC), and the relatively infrequent large cell carcinoma (LCC)[13]. These classifications are representative of the cell-of-origin of the tumours, in that the gland-forming alveolar cells in the distal airways are more commonly associated with LUAD, while the cells lining the upper airways are typically associated with LUSC and LCC[14]. Due to the deep tissue residency of alveolar cells, LUAD tumours are the most common type observed in individuals who have never smoked, the incidence rates for which continue to increase despite smoking cessation programs[15-18]. Thus, there is a need to understand the differences of these tumours at the cellular and molecular level in order to address the contribution of heterogeneity to the poor treatment efficacy faced by many patients.

1.2.2 The lung tumour microenvironment

Beyond the differences between tumours arising from various cells and spatial locations in the lung, these tumours are incredibly heterogeneous in terms of cellular organization. Due in part to the strong influence of exposure to environmental carcinogens on the development of lung cancer, it is characterized as having a strikingly high number of non-synonymous mutations and inflammatory responses compared to other types of cancer[19,20]. Among others, these features contribute to a relatively high degree of immune and other stromal cells that are found within lung tumours. This community of cells is recognized as the tumour microenvironment, of which the various cell populations, such as fibroblasts, immune cells, and epithelial cells can contribute to both tumour growth and regression[21]. Tumour mutation profiles may lead to the presentation of tumour-specific antigens that can be recognized by cytotoxic immune cells to attack tumour cells and halt progression[22]. Conversely, tumour cells can also employ mechanisms to either avoid immune detection, or alter the phenotypes of neighbouring cells to promote malignant growth and survival[23-25]. These biological features are under intense scrutiny and have prompted the development of therapeutic strategies looking to promote immune-cell recognition of and subsequent action against tumour cells[26]. However, accurate identification of cell populations and their activation status within the tumour microenvironment will be critical to improving the efficacy of these immunotherapeutic regimes.

1.3 Molecular heterogeneity in lung cancer and effects on treatment

1.3.1 Molecular alterations in the development of lung cancer

Heterogeneity at the histological, cellular, and genetic levels is a major contributing factor to the poor prognosis of lung cancer patients, which is compounded by frequent late-stage diagnosis (Figure 1.1C). Thus, understanding and characterizing the specific molecular events that lead to lung cancer is critical for directing its treatment. Large-scale genome sequencing efforts of hundreds of lung tumour samples have been invaluable to this endeavour, such as The Cancer Genome Atlas (TCGA) consortium, which has next generation sequencing data on over one thousand lung tumours and has since uncovered many important genetic events (Figure 1.1D)[27,28].

In LUAD, one of the most commonly observed mutations constitutively activates the *epidermal growth factor receptor* (*EGFR*) gene, which is part of signaling pathways that promote many cellular phenotypes including proliferation and DNA synthesis[29]. Interestingly, mutations in *EGFR* are more prevalent in never-smokers, and thus LUAD, while mutations in the gene encoding Kirsten Rat Sarcoma virus protein (KRAS) are more common in individuals with a history of smoking[30]. This in-depth profiling has led to the development of inhibitors targeting the molecular drivers of lung cancer. However, many patients present with acquired resistance, which occurs through the activation of alternative pathways or secondary mutations[31]. Interestingly, LUSC tumours are not as frequently driven by *EGFR* mutations as in LUAD, but rather approximately 47% of cases may be the result of alterations in the PI3K pathway, which maintains cell survival and proliferation[32]. Together, these observations highlight the heterogeneity amongst individual lung tumours at the genomic level.

Despite the benefits of sequencing efforts, many genetic aberrations are still considered "undruggable", such as *KRAS* alterations, and upwards of 40% of lung adenocarcinoma cases remain characterized as driven by unknown molecular events[33]. Thus, the lack of well-defined molecular drivers further impedes access to effective treatment for patients, underscoring the need to identify novel genes relevant to the development of lung cancer.

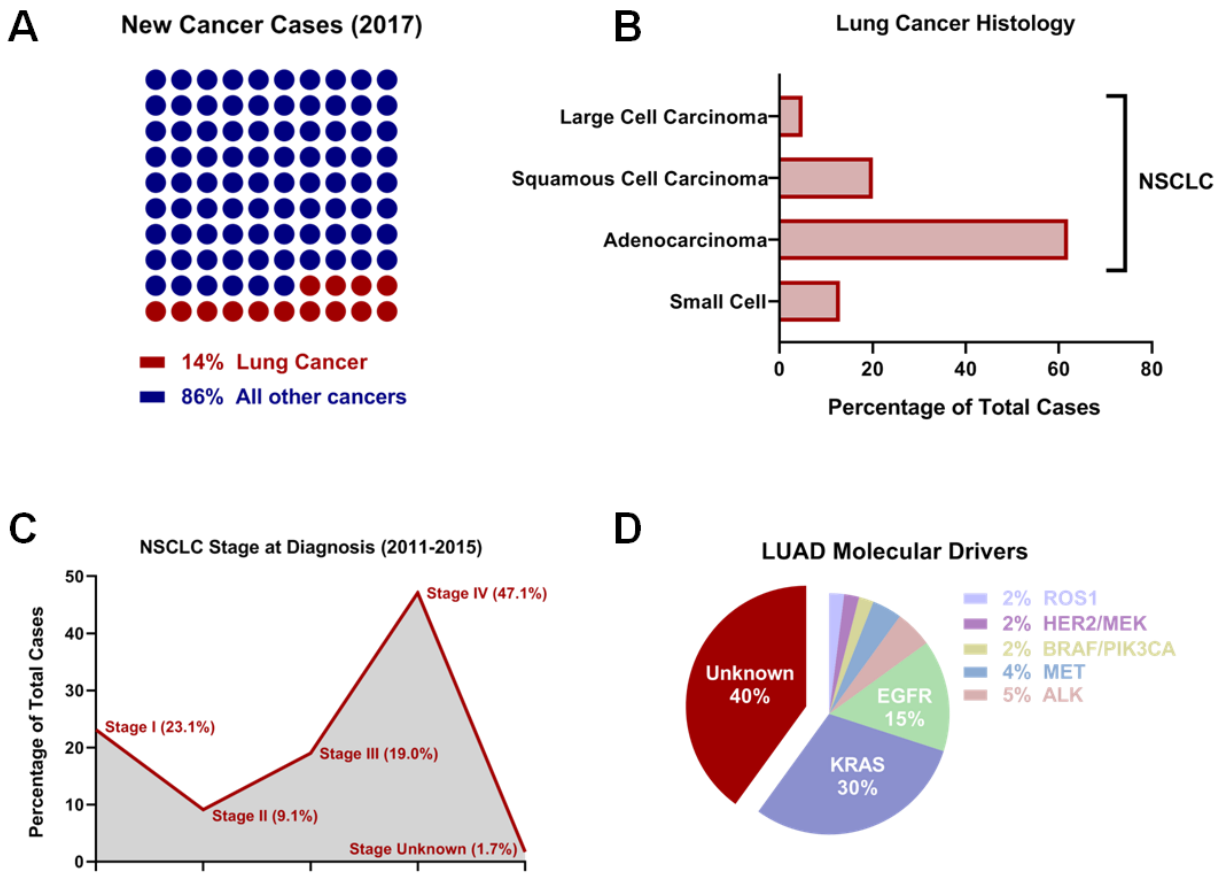

Figure 1.1. Epidemiological, histological, and molecular features of lung cancer.
A) Proportion of new lung cancer cases (red) relative to all other cancer types (blue) in Canada in 2017[2]. B) Histological breakdown of lung cancer as a percentage of total cases (x-axis)[34]. C) Tumour stage for patients newly diagnosed with non-small cell lung cancer (NSCLC)[2]. D) Molecular landscape of lung adenocarcinoma tumours. The genes mentioned here represent some of the well-documented examples of genetic alterations driving lung tumour development[33].

1.3.2 Current treatment options for lung cancer patients

Patients with early stage lung cancer can often be effectively treated with surgical intervention, but the vast majority of lung tumours are diagnosed at much more advanced stages because of the lack of symptom presentation and effective detection programs[11]. Stage III (locally advanced) lung cancer patients are subjected to combined chemotherapy and radiotherapy, which has tangible but only slight benefits in outcome[31,35]. However, as previously mentioned, advancements in genomic sequencing and analysis technologies have enabled not only the determination of key molecular aberrations driving lung cancer, but also the development of therapeutic agents specifically targeting these alterations[36]. For example, there are now three generations of targeted EGFR inhibitors, the most recent of which is osimertinib[31]. Each new compound is developed to improve upon resistance and lack of durable response in previous generations[37]. Despite the promise of these tyrosine kinase inhibitors (TKIs), observations of new resistance mutations as well as patients presenting with lung tumours driven by non TKI-related events have necessitated research into alternative methods of treatment. Recent breakthroughs in the understanding of the immune microenvironment of lung tumours and its dichotomous role in pro- and anti-tumour effects have made immunotherapy an exciting potential treatment option for many cancer patients[31]. Namely, checkpoint blockade inhibitors such as Nivolumab are able to promote cytotoxic activity against tumour cells and have shown promising responses in patients[38]. However, response rates to immunotherapy remain fairly low, and thus further work is required in order to identify markers of treatment efficacy[39]. Together, broadening our understanding of the unique genomic events in different lung tumours may stimulate the development of novel biomarkers and therapies to combat persistently poor prognoses.

1.4 Non-coding RNAs as emerging players in lung cancer

1.4.1 Next generation sequencing and non-coding RNAs

Although the relevance of molecular aberrations to human pathology has been widely recognized for decades, effective profiling of the human genome has only been made possible in recent years. The completion of the Human Genome Project in 2003 laid the groundwork for the emergence of large-scale genome sequencing initiatives that have been invaluable to the characterization and treatment of human disease[40]. Initial sequencing techniques such as Sanger sequencing have now been improved upon from technical, practical, and economic perspectives[41,42]. Since the first large-scale sequencing technology emerged in 2005, many techniques – collectively termed next generation sequencing (NGS) – are now available that can provide high-coverage parallel reactions and accurate sequencing at reduced costs in a timely manner[43-46]. As technology has developed there has also been an increased demand for the accurate analysis of the data that are produced, which has resulted in the development of numerous computational algorithms[43]. These algorithms decode the extensive information provided by NGS experiments and represent a necessary step that continues to be refined. Together, sequencing and subsequent analyses enable the discovery of novel genetic events important to normal and disease biology, especially cancer[47].

The analyses of the thousands of samples in public repositories such as TCGA are largely focused on open reading frames (or coding sequences), which only account for approximately 2% of the genome, while transcribed genes represent about one-third of the genome[48]. However, numerous studies have described widespread gene transcription throughout the genome, of which 80% of these transcripts may have functional roles[49]. While originally discarded in early genomic explorations due to their inability to code for protein, non-coding RNAs (ncRNAs) are

now recognized as functional RNA units and important regulators of gene expression. Broadly characterized by the basis of size into long (lncRNA; >200nt) and small (sncRNA; <200nt) non-coding RNAs, ncRNAs include many classes of transcripts that serve key functions in the cell[50]. Beyond ncRNAs with critical functions in cellular housekeeping, such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), ncRNAs have also been shown to act in the regulation of gene expression at many levels. For instance, PIWI-interacting RNAs (piRNAs) are sncRNAs that largely function in the regulation of histone modifications at the DNA level, while microRNAs (miRNAs) act to regulate messenger RNA (mRNA) transcripts[51]. Alternatively, lncRNAs, which include those derived from pseudogene loci and natural antisense transcripts, can regulate genes at the DNA, RNA, and protein levels[52]. Together, this class of RNA transcripts represents an immensely vast repertoire of previously unexplored transcripts in the biology of normal and malignant cells. This thesis will explore the role, transcription, and deregulation of miRNAs and lncRNAs in further detail with regards to lung cancer.

1.4.2 microRNAs[1]

Small non-coding RNAs – particularly microRNAs – are perhaps the better characterized of the two broad classes of ncRNAs. These 18-25 nucleotide transcripts typically act to negatively regulate protein-coding genes post-transcriptionally but pre-translationally, through RNA-RNA base-pairing (Figure 1.2A)[53]. Since their discovery in 1993, the evolutionarily-conserved and single-stranded RNA transcripts have been demonstrated to be associated with nearly every aspect of cancer biology[54]. Many miRNAs also show promise as predictors of

[1] Chapter 1.4.2 alludes to a book chapter that has been accepted for publication (Appendix A; Item 5): Guisier F*, Barros-Filho MC*, Rock LD*, Constantino FB, Minatel BC, [Sage AP], Marshall EA, Martinez VD, Lam WL (2019). Small non-coding RNA expression in cancer. In *Gene Expression Profiling in Cancer*, Dimitrios Vlachakis ed., InTechOpen, London. In press.

patient outcome and disease recurrence[55]. Thus, miRNA-based genomic profiling has become a prevalent focus in cancer research particularly because of their translational utility as not only markers of disease and outcome, but also their potential use as therapeutic agents and targets[56].

The miRNA biogenesis pathway is a tightly regulated process, of which its disruption is a common mechanism of miRNA deregulation in cancer[57]. Following transcription by RNA polymerase II (Pol II), the newly-formed primary-miRNA (pri-miRNA) transcripts are cleaved and processed into mature miRNAs[50]. Interestingly, pri-miRNAs can contain the sequences of multiple mature miRNA transcripts as well as protein-coding exons[58]. Next, pri-miRNAs are cleaved by Drosha, producing pre-miRNAs[59]. These ~70 nucleotide sequences are transported from the nucleus through Exportin-5, where their processing and cleavage into mature double-stranded miRNA sequences is carried out by the Microprocessor complex, containing Dicer and DGCR8[60]. Mature miRNA duplexes are unwound, a single-stranded miRNA (17-22 nt) is loaded into the argonaute 2 (Ago2) protein, which together with other proteins form the RNA-induced silencing complex (RISC)[61]. At this point, the seed sequence of the miRNA (nucleotides 2-7) binds with target RNA transcripts through complementary sequences primarily at their 3'untranslated region (3'UTR)[62]. If the interaction is perfectly complementary in the middle region of the miRNA (nucleotides 9-11), Ago2 will cleave the target transcript, imperfect binding directs expression repression through the blockage of translation[63].

As miRNA target recognition relies on a relatively short sequence of nucleotides, a single miRNA can have a variety of RNA targets, highlighting the consequences of miRNA dysregulation. Thus, the targets of miRNAs can include key genes in cellular biology, as well as both tumour suppressors and oncogenes (Appendix A; Item 4). These observations have led to the designation of oncogenic (oncomiRs) and tumour suppressive roles for numerous miRNAs[64].

In light of this, aberrant expression of a miRNA (or a pool of miRNAs) which normally targets a known tumour suppressor would thus have an oncogenic effect on the cells. For example *let-7*, is proposed to be tumour suppressive through its ability to target *KRAS* and *HMGA2*, both related to the proliferation of lung cancer cells[65]. Reduced expression of this miRNA is commonly observed in lung cancer samples, and correlates with worsened patient prognosis[66]. Together, this highlights an alternative mechanism for the regulation of well-established cancer-associated genes. Similar associations have also been noted for the aberrant expression of numerous other miRNAs, including miR-200 (promotes epithelial-to-mesenchymal transition), miR-494 (decreases proliferation), and miR-135b (inflammatory response)[67-70]. Excitingly, beyond the development of miRNA-based diagnostic and prognostic signatures, there is a growing body of work looking into the use of synthetic miRNA-mimics for inappropriately lost miRNAs with tumour-suppressive functions and conversely, anti-miRs targeting upregulated oncomiRs[71].

The proportion of the total pool of miRNAs currently-annotated in public repositories (~2400 miRNAs) with established roles in cancer is marginal relative to their widespread regulation of the genome[72]. While high-throughput sequencing efforts have enabled the characterization of miRNA function, the landscape of their expression has been proposed to be underestimated in human tissue samples as many large-scale sequencing projects are enriched for larger, protein-coding transcripts[73]. Thus, miRNAs present exciting opportunities from both biological and therapeutic perspectives and further exploration of their expression patterns may uncover alternative regulatory programs in difficult-to-treat cancers.

1.4.3 Long non-coding RNAs

Unlike small non-coding RNAs, long non-coding RNAs are frequently discovered in whole-genome sequencing projects due to their shared features with mRNAs. Despite the frequent description of pervasive transcription of long RNAs across the genome beyond coding genes, lncRNAs as a class of transcripts remain relatively poorly characterized. The difficulty in assigning function to lncRNAs is due in part to their widespread recognition as transcriptional noise[74]. However, the renewed interest in the non-coding transcriptome that has accompanied sequencing improvements has resulted in the identification of a multitude of functional roles for lncRNAs, particularly the modulation of gene expression in both *cis* and *trans*[75]. The biogenesis of lncRNAs follows strikingly similar pathways to that of mRNAs. They are transcribed by Pol II and often similarly processed post-transcriptionally (5' capped, spliced, 3'poly-adenylated), although there is growing evidence of exceptions to these observations[76]. Beyond the major difference in their ability to encode proteins, lncRNAs diverge from mRNAs in that they have relatively shorter length and fewer exons, as well as expression levels that are lower in magnitude but more specific than their coding counterparts[77]. While lncRNAs do not exhibit a high-degree of conservation between different species, lncRNA transcription occurs to a larger extent and may have better correlations to organism complexity than mRNA transcription[78-80]. These observations emphasize the importance and broad relevance of lncRNAs to cellular biology as well as their necessary inclusion in future genomic investigations.

LncRNA action occurs through a multitude of mechanisms at all levels of gene expression including transcriptional enhancement, decoys for miRNAs, as well as protein and/or RNA scaffolding (Figure 1.2B)[78]. LncRNAs are commonly found in the nucleus, where they impact chromosomal organization as well as transcription. They are also frequently transported to the

cytoplasm and other sub-cellular compartments, where they can regulate processes such as translation and mRNA/protein stability[81]. Similar to recent sncRNA explorations, a number of lncRNAs have been described to mediate key cancer-related phenotypes and are often shown to have deregulated expression levels[82]. In fact, lncRNAs have garnered attention as sites of frequent genomic alteration in many cancer types, such as mutation events, copy-number alterations, and epigenetic disruptions[83]. Some of the most well-characterized cancer-associated lncRNAs are described in lung cancer, including *Metastasis-Associated Lung Adenocarcinoma Transcript 1* (*MALAT1*) and *HOX antisense intergenic RNA* (*HOTAIR*)[65,84,85]. The overexpression of these transcripts is associated with poor survival for lung cancer patients[86]. These lncRNAs are involved in the regulation of numerous metastasis-associated genes. In the case of *HOTAIR*, this occurs via polycomb repressive complex 2 (PRC2) recruitment and epigenetic remodelling[87]. Excitingly, treatment of lung tumours *in vivo* with antisense oligonucleotides (ASOs) targeting *MALAT1* have been successfully shown to prevent metastasis[88]. Together, these studies highlight the utility of lncRNA-based analyses from both biological and clinical perspectives and emphasize the need for their further exploration.

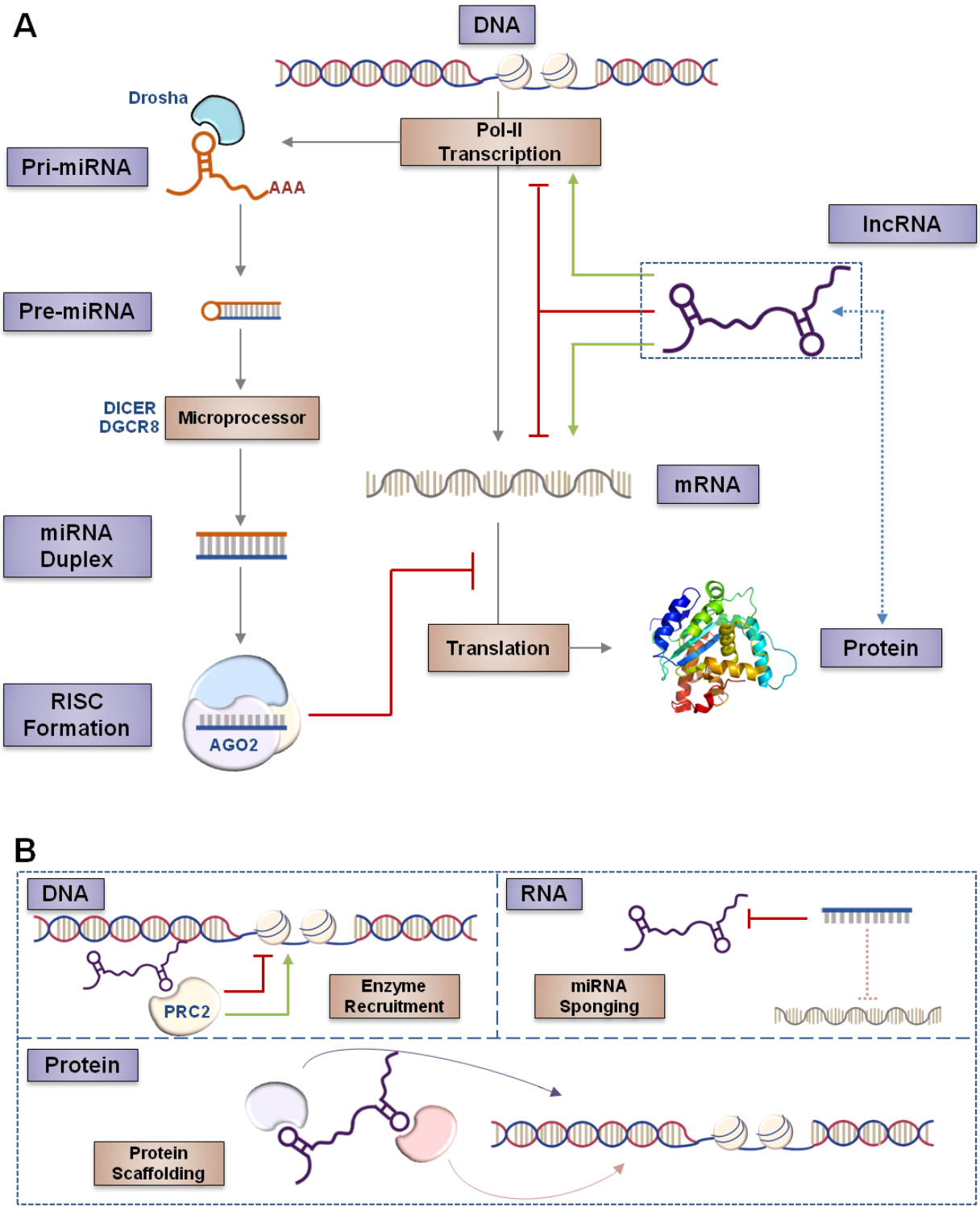

Figure 1.2. General mechanisms of transcription and function for miRNAs and lncRNAs. A) General transcription pathways for miRNAs (left), mRNAs (centre), and lncRNAs (right). Inhibitory relationships are represented by red lines, while positive regulatory relationships are denoted by green lines. B) Common mechanisms of action for lncRNAs at the DNA, RNA, and protein levels.

1.5 Challenges in ncRNA research

The recognition of the immense molecular heterogeneity of lung cancer subtypes has been made possible with large-scale genome sequencing efforts, yet, many cases present with undefined or unactionable molecular drivers[89]. Thus, non-coding RNAs present a unique opportunity to explore largely-uncharted areas of the genome for alternative mechanisms of gene regulation and potentially novel therapeutic intervention points.

Despite their relatively strong functional characterization recent studies differ in their descriptions of the landscape of human small non-coding RNA expression. A possible explanation for this may be the requirement of small RNA enrichment steps in RNA extraction and genomic analyses[90]. Even those studies that do focus beyond mRNAs to ncRNAs often examine a single subset of transcripts, resulting in the underdevelopment of general ncRNA detection tools[91]. Further, initial work that sought to characterize the human miRNA transcriptome was largely based on hypotheses that miRNAs were highly expressed and conserved between genera[92]. Since, evidence has emerged showing miRNA transcripts that are organism specific[93]. These observations have led to the suggestion that the human genome encodes a substantially greater number of small non-coding RNAs than currently anticipated[94]. Thus, further characterization of the miRNA transcriptome in human tissues is necessary prior to elucidating their widespread translational value.

Alternatively, while lncRNA research does not face barriers in sequence detection, the identification of the broad relevance and categorization of these transcripts limits our appreciation of their utility from the cellular to clinical level. Indeed, the rapid attribution of potential functions to an increasing number of annotated transcripts outpaces our ability to identify trends within non-coding RNA interaction networks, thereby hampering the

development of platforms aimed at functional lncRNA identification[95]. Specifically, a complex secondary structure, poor target-prediction and analysis tools, and difficulty in the design of effective knock-out/knock-in experiments impede our ability to generate phenotypically-relevant candidates[76,82,95]. Further, most discovery and functional characterization of lncRNAs have been performed in mice but lncRNAs show complex mechanisms of conservation between species, which impedes reproducible results between organisms[79,96,97]. Because of these features, although there are many functional studies, they are largely focused on single transcripts in specific contexts. Therefore, biochemical analyses will be critical to addressing challenges associated with structure and function, such as studies examining the functional domains and motifs within the dynamic lncRNA structure[98]. Similarly, it has become increasingly important to identify the broad expression patterns of lncRNAs in human diseases and disease phenotypes, in order to elucidate trends that aid in the downstream delineation of function[82]. Findings from these types of studies will not only strengthen our understanding of this complex class of transcripts, but also uncover potential avenues for their implementation in future cancer clinics.

1.6 Thesis themes

1.6.1 Rationale

There are a large proportion of lung cancer cases that result from unknown molecular events or exposures, which is true despite both a renewed global focus on smoking cessation and large-scale genetic profiling of these tumours. It remains critical to identify novel genes and mechanisms relevant to lung cancer development in order to direct specific and effective treatment strategies. Thus, the comprehensive analysis of non-coding RNA expression patterns and regulatory relationships represents the next frontier of lung cancer biology and treatment.

However, our understanding of the implications of ncRNA expression in human tissues remains limited. Profiling these transcripts at a broad level in human lung tumours and non-malignant samples is necessary to identify their lung-specific mechanisms of gene regulation and subsequent roles in tumour development and progression.

1.6.2 Objectives, hypotheses, and specific aims

The main ***objective*** of this thesis is to generate a clearer picture of the landscape of ncRNA expression in human lung cancer tissues, which will enable the discovery of novel genes that are central to lung cancer biology. Particularly, I endeavour to comprehensively characterize the miRNA transcriptome of normal and malignant lung samples, to supplement further analysis of tissue- and disease-specific transcripts. From the perspective of lncRNAs, while easily detectable in NGS data, their broad functional relevance remains veiled. Thus, I will assess their specificity to the cells within the lung tumour microenvironment, an area of recent intense exploration in light of its role in the response to immunotherapy treatment. I have formed ***two hypotheses***:

(1) The landscape of small non-coding RNA expression, particularly miRNAs, while largely functionally characterized, is underestimated in human lung tissues; and

(2) LncRNAs are specifically expressed and mediate regulatory networks in the cells of the tumour microenvironment.

For the first hypothesis, I aim to demonstrate the existence of previously-unannotated miRNAs in human tissue samples and subsequently analyze their relevance to tissue- and cancer-specific biology. SncRNA detection requires a targeted approach that results in the more widespread use of low coverage analysis techniques, such as microarrays, as well as a reliance

on current annotations of miRNAs. However, early miRNA discoveries based on sequence analyses relied on observable sequence conservation between species, which consequently eliminated a large proportion of specifically expressed miRNAs from further consideration. I thus re-explore small-RNA sequencing data with the goal of discovering and paneling the expression of previously-unannotated miRNA sequences that are specifically expressed in non-malignant and cancerous lung tissues. To assess whether this phenomenon is also true for other human tissues, I apply my approach to the same type of data from clear cell renal cell carcinoma samples. The ***specific aims*** are:

Aim 1: Assess the existence of previously-unannotated miRNAs in human lung samples

Aim 2: Explore the potential roles and utility of these transcripts in LUAD biology

Aim 3: Determine whether this phenomenon holds true in other tissue contexts

Collectively, I show that current analysis pipelines may result in the misrepresentation of a large proportion of biologically important sequences, and describe a platform to provide a more global representation of their expression.

For the second hypothesis, I aim to delineate the expression and potential roles of long non-coding RNAs in infiltrating immune cells in the lung tumour microenvironment. Lung cancer is characterized by a relatively high mutational burden and subsequent high degree of immunogenicity. Immunotherapy has become a viable and effective treatment option for patients with difficult-to-treat lung cancer, and is being explored as first-line therapy for select cases[99]. However, it is important to identify which patients will respond to treatment, which may be facilitated through the identification of specific markers of immune infiltration. Long non-coding RNAs are known to have specific expression patterns and despite numerous studies describing

lncRNAs that function in the regulation of the immune response, the landscape of lncRNA expression in human immune cells remains poorly characterized. Thus, I aim to describe the expression of lncRNAs in healthy human immune cells and further apply these observations to tumour infiltrating lymphocytes in human lung cancer samples. The ***specific aims*** are:

Aim 1: Delineate the expression of lncRNAs in healthy human immune cells

Aim 2: Panel these immune-related lncRNAs in tumour-infiltrating lymphocytes

My work aims to identify novel patterns of ncRNA expression that are important to lung cancer biology. Together, my results will lay the groundwork for the identification of unique markers of disease and potential therapeutic intervention points.

Chapter 2: Methods

2.1 Patient tumour cohorts and collection of RNA sequencing data

Numerous sample cohorts and datasets were used for the analyses described in this thesis. All datasets are publicly available, with the exception of the BC Cancer Agency (BCCA) lung adenocarcinoma dataset, which is a cohort specifically obtained for use by the BC Cancer Research Centre. Data were downloaded from public servers using the Sequence Read Archive (SRA) toolkit command line application[100]. Table 2.1 provides a summary and description of the datasets used.

2.1.1 BC Cancer Lung Adenocarcinoma Samples (BCCA-LUAD)

Lung tumour samples with matched adjacent non-malignant tissues were acquired during surgery at the Vancouver General Hospital. Patients had not undergone any previous treatments. Tissue samples were obtained with informed patient consent as well as the approval of the Research Ethics Board of the University of British Columbia and the BCCA. After collection, tumour samples were microdissected to 80% tumour cell-content (tumour purity) and subsequently analyzed using whole genome sequencing. A total of 236 samples were processed using the Illumina HiSeq 2000 small RNA sequencing platform for the analysis of small RNAs described in Chapter 3, and the 36 paired samples sequenced using the Illumina HiSeq platform were used to assess long RNA expression levels.

2.1.2 TCGA Lung Adenocarcinoma Samples (TCGA-LUAD)

A cohort of lung adenocarcinoma tumours obtained from TCGA Research Network under the TCGA-LUAD heading was obtained for analysis of whole genome sequencing data. Small RNA sequencing data were available for 118 tumour samples with matched non-malignant tissues, processed using the Illumina Hi Seq 2000 small RNA sequencing platform. The TCGA-LUAD contains long RNA sequencing data from 515 LUAD tumours generated using the Illimuna Hi Seq 2000 platform, 54 of which were used with their matched non-malignant samples in paired analyses. The data from small RNA sequencing experiments were used in the analysis of novel miRNAs in lung tissue in Chapter 3, while the long RNA data were used in multiple analyses of lncRNA expression in lung tumours described in Chapter 4.

2.1.3 TCGA Clear Cell Renal Cell Carcinoma Samples (TCGA-KIRC)

Clear cell renal cell carcinoma (ccRCC) tumours with matched non-malignant samples (n=71 pairs), as well as unpaired tumour samples (n=431) were collected and processed by TCGA Research Network (TCGA-KIRC; http://cancergenome.nih.gov/). The Illumina HiSeq2000 platform was used to generate small RNA sequencing reads, which were obtained from the Cancer Genomics Hub (cgHUB) Data Repository (dbgap Project ID: 6208). These data were used in the exploration and analysis of previously-unannotated miRNA sequences in Chapter 3.

Table 2.1. Cohort sample characteristics

		TCGA-LUAD	BCCA-LUAD	TCGA-KIRC
Non-malignant samples		54 (long-read) 91 (short-read)	36 (long-read) 118 (short-read)	71 (short-read)
Malignant samples		515 (long-read) 91 (short-read)	36 (long-read) 118 (short-read)	502 (short-read)
Tumour cell content		>60%	>80%	>60%
Clinical Information				
Mean Age		67	70	61
Gender	*Male*	24	10	331
	Female	30	26	171
Ethnicity	*Caucasian*	51	11	-
	Asian	-	14	-
	Hispanic/Latino	-	-	24
	Not Hispanic/Latino	-	-	332
	Other	3	11	-
	Not Reported	-	-	146
Stage	*I*	28	20	241
	II	14	11	55
	III	10	3	122
	IV	2	1	81
	Not Reported	-	-	3
Smoking	*Current*	7	5	64
	Former	36	6	9
	Never	5	25	431

2.2 Processing of RNA sequencing data

2.2.1 Processing of small RNA sequencing data

Raw small RNA sequencing data generated from the BCCA-LUAD, TCGA-LUAD, and TCGA-KIRC cohorts were processed according to a custom pipeline for the detection and quantification of small RNA sequences[101]. First, the raw sequencing reads (.bam files) were converted into unaligned reads (.fastq files) using the PartekFlow software, build 7.0.18.0724[102]. The unaligned reads were then trimmed according to their sequencing quality scores (Phred scores ≥ 20). The Spliced Transcripts Alignment to a Reference (STAR; v2.4.1d) algorithm[103] was used to align the reads that met the quality threshold were to the most recent annotation of the human genome (hg38 build). Quantification of reads was then performed, using various algorithms and normalization methods described in more detail in Chapter 2.3.2.

2.2.2 Processing of long-read RNA sequencing data

Raw RNA sequencing data were processed similarly to the small RNA sequencing reads as described previously. Reads were unaligned to .fastq files and realigned to the current build of the human genome using the STAR aligner. Quantification of long RNA sequencing reads was performed using the Cufflinks algorithm[104]. Long non-coding RNAs were defined according to the Ensembl v89 annotation, wherein transcripts labeled as antisense, bidirectional promoter lncRNA, lincRNA, and macro lncRNA were considered for further analysis[105].

2.3 Discovery of novel miRNA sequences

2.3.1 Description of predictive algorithms

The discovery of previously-unannotated miRNA sequences described in Chapter 3 used two publicly available platforms that are based on the miRDeep2 prediction algorithm[106]. The miRDeep2 algorithm combines sequencing read numbers with secondary structure information. The relative free energy of the secondary structure predicted by the sequence is combined with the significance associated with random folding to generate a score representing the likelihood of a miRNA-like secondary structure. This measurement is known as the miRDeep2 score and higher miRDeep2 scores are reflective of more reliable predictions.

Recently, studies have sought to expand upon the *in silico* insight provided by algorithms like miRDeep2. In 2017, Fehlmann and colleagues described a new predictive tool coined miRMaster that could integrate the base algorithm for the prediction of novel miRNAs with direct quantification of miRNA expression and identification of isomiRs/polymorphisms in miRNAs[107]. Furthermore, while this algorithm does not curate any of the predicted novel miRNA candidates, it provides the user with comparisons to miRBase and RefSeq databases, which enables the identification of sequences overlapping currently annotated miRNAs, and potential viral or bacterial contaminants[72,107,108]. Thus, miRMaster provides not only a high degree of reliability in predicted candidates through its extensive set of features, but also provides fewer requirements for manual curation, lessening the possibility of false discovery through user error.

Algorithms similar to miRMaster based upon the miRDeep2 prediction algorithm also provide comparable results, while operating with slightly different specifications. Another prominent prediction tool is OASIS, which generates grouped predictions of novel miRNA

candidates from raw sequencing reads that map to regions of the genome that are not currently annotated for miRNAs[109]. While quantification of miRNA expression is not directly available in this algorithm, read counts are described, which separates the analysis of novel miRNA detection from the evaluation of their expression. Together, the development of these algorithms represents the growing need for and continual improvement of data analysis and interpretation tools.

2.3.2 Sequence detection and candidate curation

To determine a robust set of novel miRNA candidates for further analysis of their expression in human tissues, curation based on the molecular features of the predictions was performed. In the analyses using miRMaster, the raw prediction output files were merged with normalized expression quantification using the perl command line application. The predictions were then filtered to include only candidates that had: (1) completely novel annotation; (2) no complete sequence overlap with sequences annotated in miRBase; and (3) GC content within 2 standard deviations from the mean of all sequences. Candidates with duplicate 5p and 3p sequences were consolidated and their expression levels summed. Finally, an expression threshold of $\geq$ 1 read per million (RPM) in 10% of samples was set to exclude candidate sequences with little or no expression.

The OASIS platform provides raw sequence information for the predicted novel miRNA sequences[109]. Thus, for the analyses using this platform, sequences were first filtered to select for candidates that had: (1) no overlap with rRNA/tRNA reads in the RNA families (Rfam) database[110]; (2) significant ($p < 0.05$) miRNA-like secondary structure predictions; (3) loci with an acceptable number of sequencing reads ($\geq$ 10 reads); and (4) a GC content within 2 standard deviations from the mean. Duplicate candidates were consolidated and the remaining sequences

were assessed for their similarity to miRNAs currently annotated in miRBase v21[72], which was performed using standard nucleotide blast (BLASTn) through the BLAST+ command line application[111]. These analyses yielded expect (*E*) values for each sequence, and candidates with $E < 0.1$ were determined to have strong overlap with previously-annotated sequences and were subsequently discarded. As OASIS does not use expression information, the sequencing reads were quantified using the featureCounts v1.4.6 algorithm[112] on a per-sample basis and normalized using the weighted trimmed mean of log expression ratios method (TMM). Candidates with summed expression levels ≥ 10 as well as ≥ 0.1 in at least 10% of samples were considered expressed and used for further analysis.

2.3.3 Statistical analyses of gene expression

Several statistical tests were used to analyze and interpret information from next generation sequencing data. All analyses of differential expression were performed using a two-tailed Mann-Whitney U-test, performed using MATLAB R2013a, with the Benjamini-Hochberg (BH) method performed in RStudio v3.3.3 used to correct for multiple testing and ensure adequate control of the false-discovery rate[113]. All unsupervised hierarchical clustering analyses were performed using Partek Flow software (build 7.0.18.0724[102]), with a cluster distance metric of Average Linkage and the point distance metric as Pearson Correlation. Patient clinical information was obtained for all publicly available samples through the University of California Santa Cruz (UCSC) Xena browser (http://xena.ucsc.edu/). Survival analyses were performed by stratifying samples into tertiles of high to low expression of the gene of interest, categorizing vital status and days to death/last follow up. The significance of how well patient outcome was

predicted by gene expression was assessed by the logrank method[114], using both GraphPad Prism v8 and MATLAB R2013a software.

Protein-coding target genes of miRNAs were assessed using the miRanda v3.3a algorithm[115]. This algorithm compares the sequences of candidate miRNAs against the 3'UTR of all human protein-coding genes (obtained from the Ensembl BioMart tool; https://www.ensembl.org). Predicted targets are based on: (1) the sequence complementarity of the miRNA to the 3'UTR; (2) the free energy of the duplex formed; and (3) the conservation of the target sites. Predicted targets generated by the algorithm were first curated by analyzing five scrambled miRNA sequences, wherein protein-coding targets predicted by both candidate miRNAs and scrambled sequences were discarded. Targets were further filtered to only include those with an alignment score of ≥ 140 and a free energy threshold of ≤ -20 kcal/mol. Transcripts targeted by at least 10% of the miRNAs-of-interest were analyzed for their membership in key signaling pathways via the pathDIP v.2.5.21.6 algorithm[116], which comprehensively examines pathways enriched in the predicted target genes in 15 distinct public repositories.

Chapter 3: Discovery of previously-unannotated miRNAs specific to normal and malignant human tissues[2]

3.1 Introduction

3.1.1 miRNAs in lung cancer

Beyond their roles in the regulation of close to 60% of protein-coding genes, miRNAs are now also regarded as important players in cancer biology[117]. As the recognition (seed) sequence for miRNA-mRNA interactions can be as little as only 6 nucleotides, a single miRNA can have an array of protein-coding targets, including many well-established tumour-suppressors and oncogenes[71,118]. Every step of the miRNA biogenesis pathways is tightly regulated[119]. However, dysregulation of miRNA expression in cancer both specifically (amplification/deletion events) and indirectly (disruption of miRNA biogenesis enzymes) is frequently observed and may promote pro-tumour phenotypes and poor outcome[71,120,121].

In lung cancer, the deregulation of many miRNAs is implicated to be pathogenic[122]. For example, the decreased expression of miR-101 through DNA copy-number losses has been shown to result in the upregulation of EZH2, a key epigenetic modifying enzyme, which subsequently leads to the inhibition of apoptosis and disruption of proliferation, among other effects[123,124]. Further, there has been immense investigation into the ability of miRNAs to supplement lung cancer early detection and subtype identification, in light of their expression patterns in cancer and their stability in biofluids and formalin-fixed paraffin-embedded (FFPE)

[2] The work in Chapter 3 describing the detection of novel miRNA sequences in renal cell carcinoma has been published. [Sage AP], Minatel BC, Marshall EA, Martinez VD, Stewart GL, Enfield KSS, Lam WL. (2018) Expanding the miRNA transcriptome of human kidney and renal cell carcinoma. *International Journal of Genomics* 2018:1-10. (Appendix A; Item 8).

tissues[125]. However, despite the rapidly increasing number of miRNA-based signatures in lung cancer none have yet to be translated into the clinic, necessitating further exploration of sensitive and specific markers of disease[126-128].

3.1.2 Previously-unannotated miRNAs

The continuously developing body of evidence suggesting diverse roles for miRNAs in the regulation of numerous important cellular processes underscores the renewed emphasis on the accurate characterization of the sncRNA transcriptome. However, many early genomic explorations focused on miRNA transcripts that were abundantly expressed and conserved between organisms, features that may exclude important specifically-expressed miRNAs[92,93]. Thus, it follows that with next generation sequencing efforts it is possible to uncover the transcription of miRNAs that have yet to be described in any capacity, genes that may have more specialized expression and function than their well-characterized predecessors.

Indeed, algorithms have now been developed to search for miRNA transcripts that are not annotated in public repositories, namely miRDeep2, Oasis, and miRMaster[107,129,130]. As described in Chapter 2.3.1, these algorithms are designed to integrate NGS reads with molecular features of miRNAs (e.g. DICER processing, secondary folding structure). These platforms enable the re-evaluation of RNA-sequencing data to better represent the landscape of transcription in these samples. Further, separate curation and bioinformatic analyses can be performed to estimate the probability of a true-positive prediction, such as novomiRank which compares candidate sequences with those annotated based on features such as free energy, folding, length, and sequence composition[131]. While their existence in human tissue and disease

contexts still requires comprehensive analysis, these novel sequences represent exciting discoveries that may present specific markers and intervention points.

3.1.3 Small non-coding transcriptome of the NCI-60 cell line panel

Cell lines provide invaluable genetic and phenotypic resources for the study of cancer biology[132]. The NCI-60 cell line panel is perhaps the most widely used in basic cancer research, which contains 59 cell lines from nine different cancer types[133]. However, cell lines are often limited by their ability to truly recapitulate the complex nature of human tumours. Additionally, while many novel cancer-related ncRNAs have been described, adequate expression profiles of the non-coding transcriptome in tissues-of-interest are necessary for the further characterization of these important transcripts. Thus, to provide a resource to facilitate genome-level research of sncRNA expression in cancer and cancer cell lines, our group recently described the small non-coding transcriptome of the NCI-60 cell line panel (Appendix A; Item 11)[134].

Through high-throughput small RNA sequencing, we were able to characterize sncRNAs in each cell line and found the widespread distribution, proportion, and unique expression patterns for piRNAs, miRNAs, small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs). Additionally, we discovered that novel miRNA transcripts were not only detectable in these cell lines, but represented striking increases of roughly 10% from those currently-annotated. This previous work provides a foundation for the exploration of sncRNA-mediated gene regulation in cancer cell lines and emphasizes the existence of many previously-unannotated miRNA transcripts. Together, these observations highlight the necessity of determining whether these findings hold true in human tissue samples, from both malignant and non-malignant perspectives.

In light of these findings, I hypothesized that human lung samples not only expressed a significant number of previously-unannotated miRNA sequences, but also that their expression patterns and molecular targets were indicative of roles in the maintenance of important lung cancer pathways. I thus set out to develop a platform for the discovery of these sequences in human cancer tissues, and then assess both the potential functions of these transcripts as well as whether this phenomenon remains true in other tissue contexts.

3.2 Methods

3.2.1 Patient Samples and small RNA sequencing

The BCCA-LUAD and TCGA-LUAD cohorts were used in this analysis (Chapter 2.1.1 and Chapter 2.1.2). Raw unaligned sequencing reads from these samples were filtered for read quality (Phred $\geq$ 20) and aligned to the hg38 build of the human genome (STAR aligner) (Figure 3.1). Aligned sequencing read files for both LUAD cohorts were quantified using the miRMaster platform. Gene expression levels ascertained by these algorithms were then normalized to the total number of reads in each sample, using the RPM method.

3.2.2 Discovery and analysis of previously-unannotated miRNAs

LUAD small RNA sequencing files were submitted to the miRMaster online sequence analysis platform. The miRMaster algorithm allows for the integration of per-sample and per-group features to directly quantify transcript expression levels and examine novel miRNA candidates. Thus, miRMaster provided an extensive set of novel miRNA candidates that were then curated based on molecular and cellular features to generate a robust set of predictions.

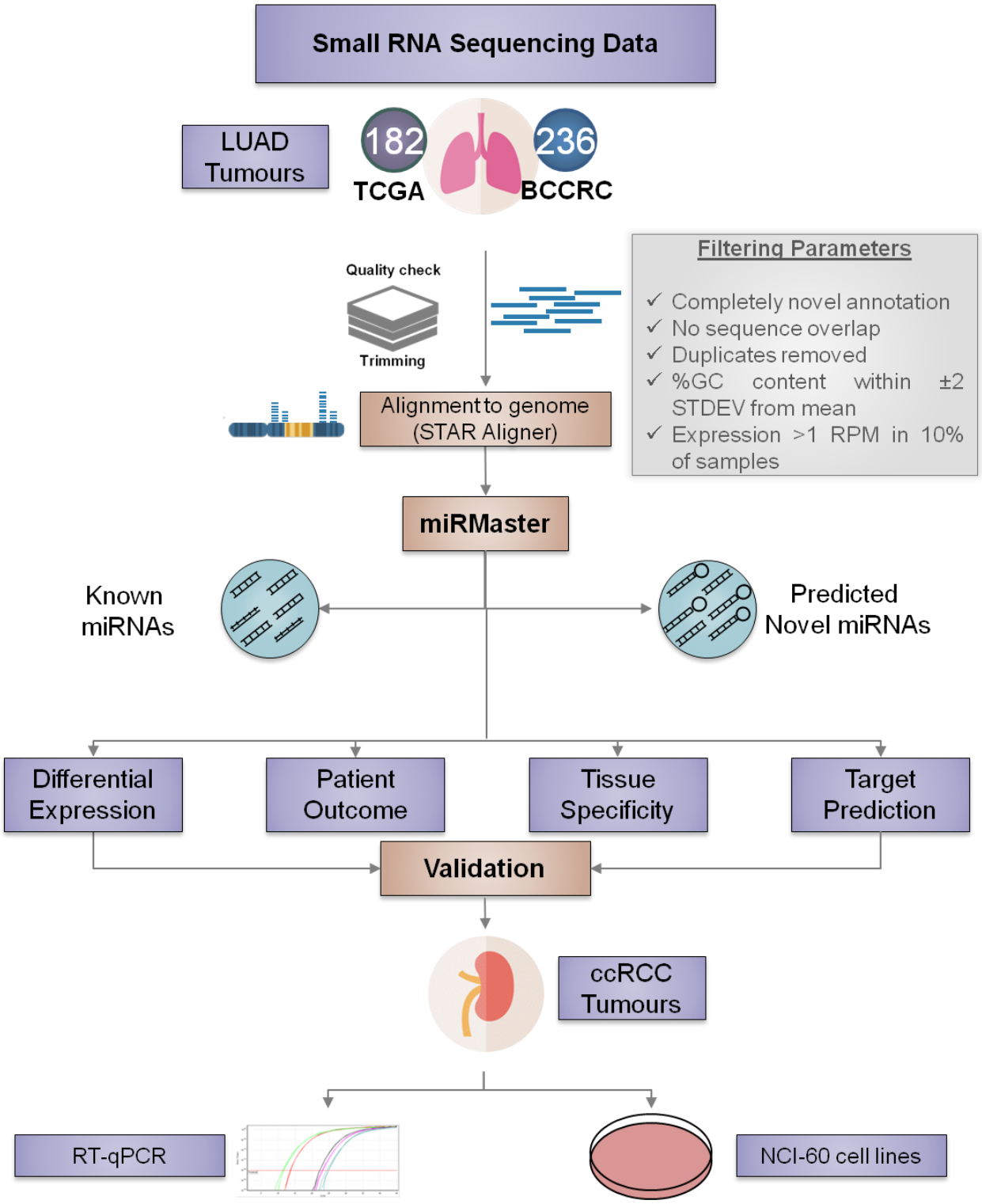

Figure 3.1. Sequence analysis flow chart for novel miRNA sequences discovered in human lung samples from the TCGA-LUAD and BCCA-LUAD cohorts, and subsequent validation in TCGA-KIRC patient samples and ccRCC cell lines.

3.2.3 Analysis of previously-unannotated miRNA expression

Following the generation of this set of miRNAs newly-discovered in human lung samples, I set out to assess their expression patterns and potential relevance to cancer biology. The tissue specificity of these transcripts was analyzed through principal components analysis using the PartekFlow software (build 7.0.18.0724)[102]. We also examined their expression patterns in matched tumour and non-malignant samples in both the TCGA-LUAD and BCCA-LUAD cohorts using MATLAB R2013a and RStudio v3.3.3. The associations of key miRNAs with patient outcome were assessed in GraphPad Prism v8, using the log rang method. Finally, we examined the potential molecular targets for these miRNAs using the miRanda v3.3a algorithm and the associated pathways that were disrupted, through PathDip v2.5.21.6 (Chapter 2.3.3).

3.2.4 Validation in different contexts: human ccRCC samples and cell lines

To assess whether the phenomenon of novel miRNA expression held true in different human tissue contexts, we performed the discovery analysis on human kidney samples from the TCGA-KIRC cohort (Chapter 2.1.3). Novel miRNA discovery was performed using the OASIS platform (Chapter 2.3.1 and 2.3.2), and transcript expression values were quantified using the featureCounts v1.4.6 algorithm through PartekFlow (build 7.0.18.0724)[102]and normalized using the TMM method. To validate the novel miRNAs detected in kidney samples, we assessed their expression in renal cancer cell lines from the NCI-60 cell line panel. Following the same procedure, the detection of previously-unannotated miRNAs was confirmed in eight renal cell lines (A498, CAKI-1, 786-0, TK-10, UO-31, ACHN, RXF393, and SN12C). Sequences with expression levels of TMM > 0.1 were considered to be expressed in the cell lines and thus validated. Differential expression, association with patient outcome, as well as target prediction

and subsequent pathway analysis were performed for these transcripts, in the same manner as those discovered in lung samples (Chapter 2.3.3, Chapter 3.2.3).

3.2.5 Cell culture and real-time quantitative PCR (RT-qPCR)

To further confirm not only the existence, but also the deregulated expression of novel miRNA sequences in human tissues, two sequences with the greatest differential expression levels in tumour and non-malignant tissues (Knm17_1130, Knm3_1968) were assessed *in vitro*. The ccRCC tumour cell line TK10 and immortalized non-malignant embryonic kidney cell line HEK-293T were used in this analysis. Cultures of both lines were maintained according to guidelines provided by the American Tissue Culture Collection (ATCC) and the National Institute of Health (NIH) (TK10 – Roswell Park Memorial Institute medium (RPMI) 1640 + 10% Fetal Bovine Serum (FBS); HEK-293T – Dulbecco's Modified Eagle Medium (DMEM) + 10% FBS) and stored in a 37 ^{o}C incubator with 5% CO_2. Total cellular RNA was extracted from plates of confluent cells using the Quick-RNA™ MiniPrep Kit Zymo Research, Catalog number R1055). To specifically convert the predicted novel miRNAs to complementary DNA (cDNA) products, custom primers were designed through the Custom TaqMan® Small RNA Assay Design Tool from Thermo Fisher (Knm3_1968 - GCAGAUUCCCAGAGUGGGACAG; Knm17_1130 - UGAGGUGGAGGGUUGUGGGA). The TaqMan miRNA Reverse Transcription Kit was used to perform the cDNA conversion, with an input RNA concentration of 2 ng/µL. The resulting cDNA was then analyzed by real-time quantitative PCR (RT-qPCR) using the Applied Biosystems® 7500 Real-Time PCR System. Finally, the $2^{-\Delta\Delta Ct}$ method was used to quantify the relative expression of the novel miRNA candidates, after normalization to the expression of the widely-expressed snRNA U6.

3.3 Results

3.3.1 Expression of previously-unannotated miRNAs in human lung samples

We have previously shown the widespread expression of not only small non-coding RNAs in cancer cell lines, but also the presence of a striking number of miRNAs that have yet to be annotated in public databases (Appendix A; Items 8, 10)[134,135]. Thus, I aimed to assess whether this phenomenon holds true in human lung samples, in both normal and disease contexts. After obtaining and processing small RNA sequencing data from the TCGA-LUAD and BCCA-LUAD cohorts (Chapter 2.1, 3.2.1), we analyzed these data using our custom analysis pipeline for the discovery of novel miRNAs, which makes use of the miRMaster online tool (Chapter 2.3.1, 3.2.2)[107]. Initial assessments of quality and the removal of sequences with strong similarity to currently-annotated sequences generated a list of 379 previously-unannotated miRNA candidates discovered in both cohorts. Further curation by expression and sequence composition resulted in a set of 141 previously-unannotated miRNAs expressed in human lung tissues. Together, this represents an approximately 14.6% increase in the known miRNA transcriptome of the lung (Figure 3.2A). We then examined the similarity of these novel sequences to those currently-annotated in miRBase, using features including: GC content, seed sequence composition, and genomic location (Figure 3.2B-D). These comparisons emphasize the miRNA-like character of these sequences, their potential roles in the regulation of protein-coding target genes, and subsequently tissue- and cancer-specific pathways.

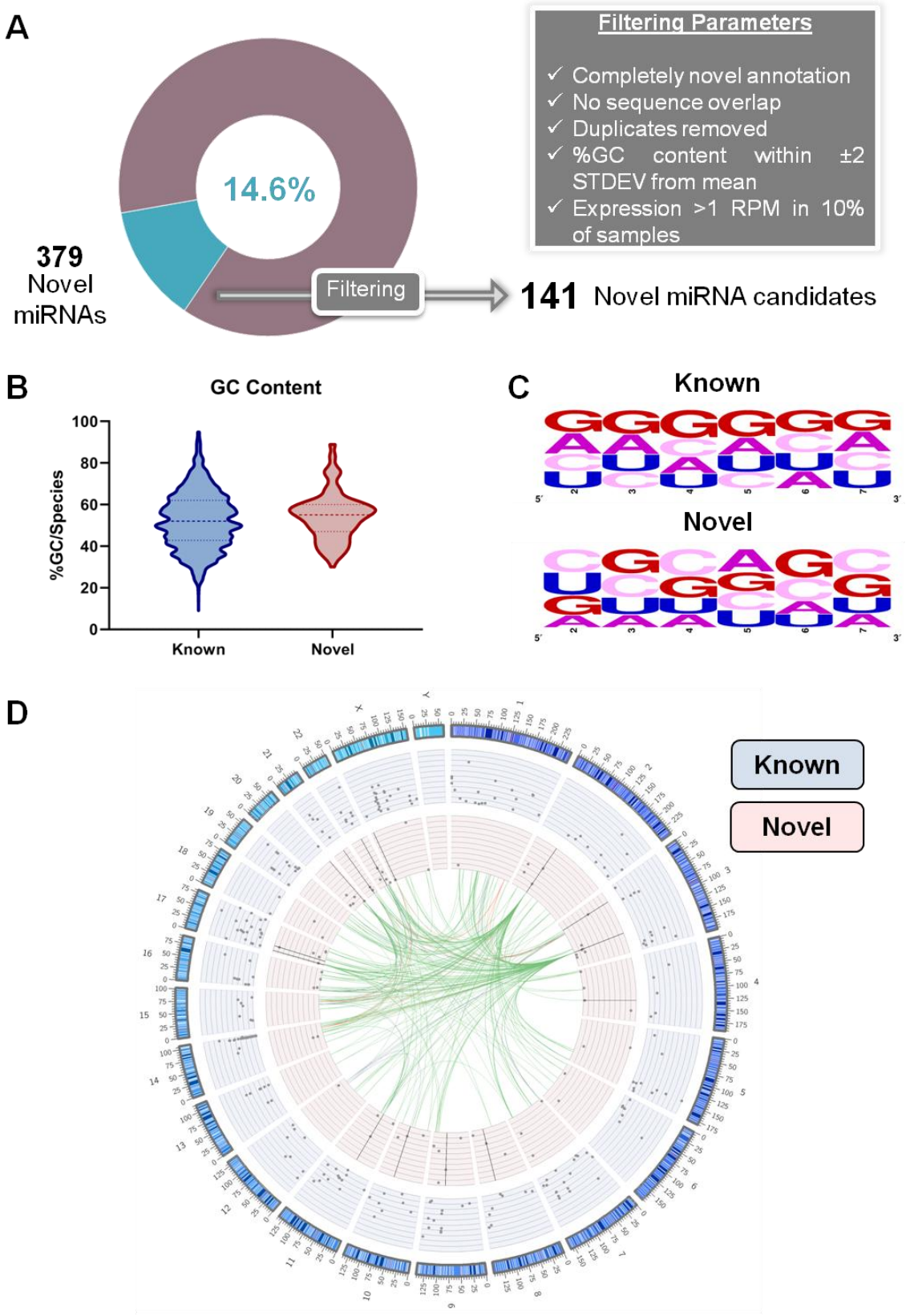

Figure 3.2. Detection parameters and molecular features of novel miRNA candidates.

A) Relative proportion of the known lung miRNA transcriptome represented by newly-discovered novel miRNA candidates (blue). B) Violin plot comparing the distribution of G/C content per sequence between known (blue) and novel (red) miRNAs. C) Sequence logo comparison between known and novel miRNAs in the nucleotide composition of the seed region[136]. D) Genomic localization of known (grey points on blue concentric circle) and novel (grey points on red concentric circle)[137]. Human chromosomes are represented on the outer concentric circle. Grey lines highlight novel miRNAs differentially expressed between tumour and non-malignant samples. Links between regions represent the predicted interaction pairings between novel miRNAs and protein-coding targets, where genes targeted by at least 2 miRNAs are coloured with green lines, at least 3 miRNAs are purple lines, and at least 4 miRNAs are red lines.

A prominent hypothesis explaining why these newly-discovered miRNAs have been overlooked is the reliance of previous genomic explorations on sequence conservation between species and tissues, and their relatively low expression escaping detection by the relatively low-coverage techniques[94]. In light of this observation, I aimed to assess the specificity of the 141 newly-detected sequences to lung tissue. Principle component analysis revealed that the combined expression pattern of the newly-detected miRNA candidates was able to distinguish lung samples from the other 12 non-malignant tissues that were analyzed (Figure 3.3). Further, while samples from other tissue types are not completely separate from one another, they can be seen to be organized within their tissue type, highlighting the potential tissue- and context-dependence of previously-unannotated miRNA sequences. Together, these results suggest the specificity of the newly-detected sequences to non-malignant lung tissue, as well as the potential translational utility of these sequences as tissue-of-origin and disease-specific markers.

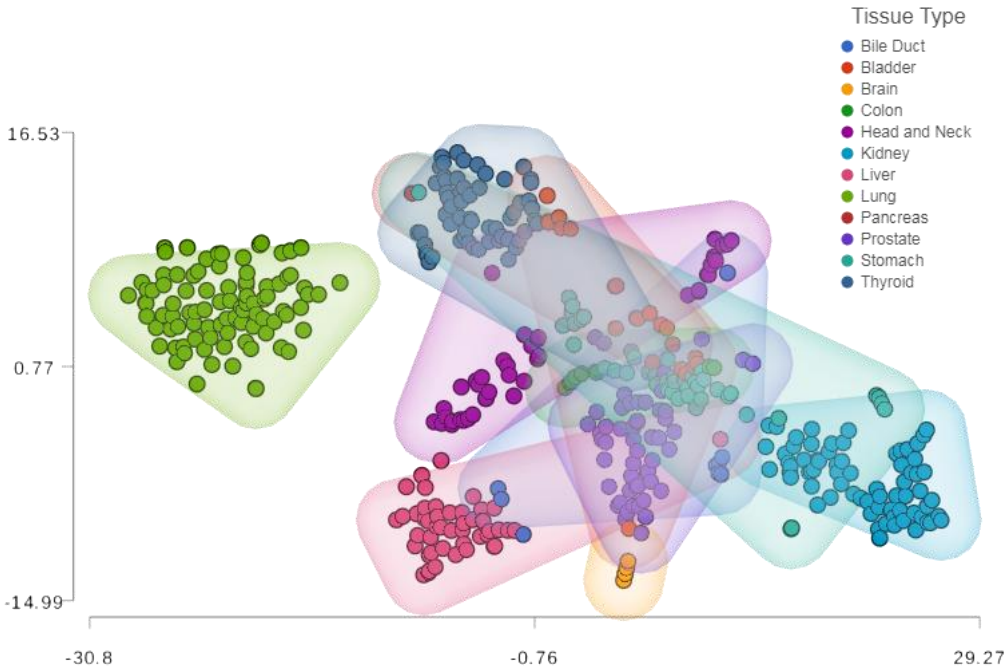

Figure 3.3. Lung-specific expression of previously-unannotated miRNAs revealed by principle components analysis.

Non-malignant tissue samples are represented by circles on the graph, and samples from the same tissues are grouped within their respective coloured and shaded regions (e.g. lung samples are represented by green circles, surrounded by light green shading).

3.3.2 Dysregulation of novel miRNAs in lung tumours and their clinical potential

After observing that the newly-detected sequences were specifically expressed in non-malignant lung tissue, I was curious as to whether this trend of specific expression would apply to tumour contexts. Initially, it was striking to see that of the 141 novel miRNA candidates, 49 were discovered in non-malignant samples and displayed no detectable expression in tumours, while 74 were only expressed in tumour samples (Figure 3.4A). The context-dependence of previously-unannotated sequences suggests the clinical utility of these transcripts as markers of disease onset and progression, as well as targets for therapy that could be specific to tumour cells.

We also examined the 18 previously-unannotated miRNA loci that were discovered and had detectable expression in both non-malignant and tumour samples. Analysis of their expression revealed that of these 18, 14 were differentially expressed (BH-$p < 0.05$), the majority of which were over-expressed in tumours (Figure 3.4B). The genomic locations of these differentially-expressed previously-unannotated miRNA transcripts are also highlighted by grey bars on the circos plot in Figure 3.2D. Collectively, these results lay the groundwork for future studies examining the impact of previously-unannotated miRNA dysregulation on lung tumour onset and progression, as well as their translational potential, which I aimed to explore further.

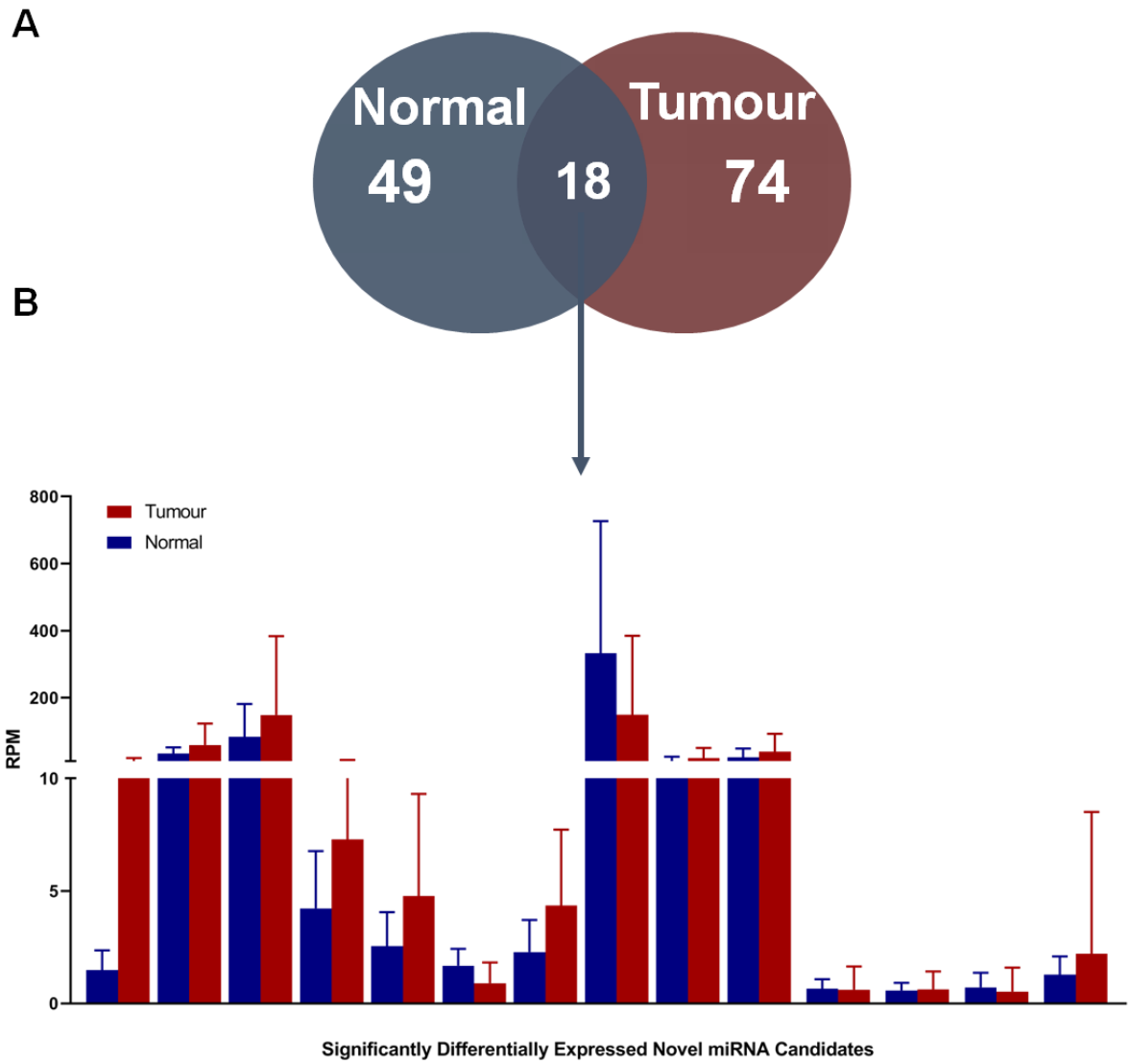

Figure 3.4. Previously-unannotated miRNAs are differentially expressed between lung tumours and matched adjacent non-malignant tissues in the BCCA-LUAD cohort.
A) Venn diagram representing the number of sequences discovered and displaying detectable expression in either normal (blue) or tumour (red) samples. B) Histogram of the 14 previously-unannotated miRNAs differentially expressed between tumour (red) and normal (blue) samples.

In order to gain insight into the potential use of these novel miRNA sequences in clinics, I first wanted to assess whether these specific candidates could be predictive of patient outcome in individuals with lung adenocarcinoma. We examined de-identified patient phenotype and clinical characteristics data for the samples from the BCCA-LUAD cohort. Indeed, we observed that the expression of several previously-unannotated miRNAs was able to stratify patients by outcome in a logrank survival analysis, revealing striking examples of pred-nov-miR-10972_TP and pred-nov-miR-5274_TP (Figure 3.5). It is particularly interesting that the expression of singular novel miRNAs alone is observed to have prognostic value. Thus, it can be suggested that these highly specific sequences may be able to strengthen current molecular prognostic panels, allowing for earlier detection of more aggressive tumour characteristics and subtypes.

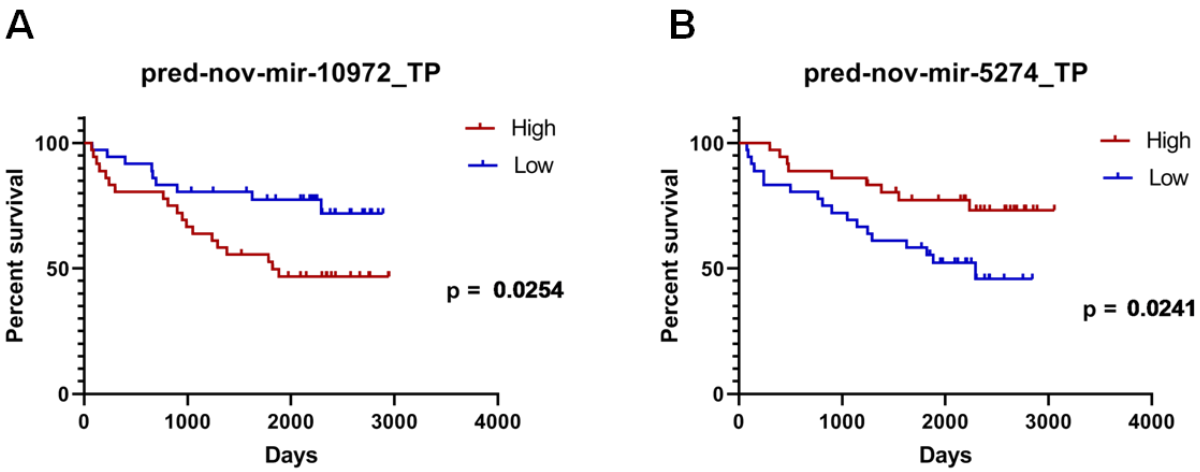

Figure 3.5. Prognostic utility of previously-unannotated miRNAs.
Red lines represent the tertile of samples with relatively high transcript expression, while blue lines represent the lower tertile of expression. Overall survival was assessed using the log-rank test ($p < 0.05$) for pred-nov-miR-10972_TP (A) and pred-nov-miR-5274_TP (B).

3.3.3 Novel miRNAs target important protein interaction networks in lung tumours

Gene regulation enacted by miRNAs relies on sequence complementarity of the miRNA with the 3'UTR of a given protein-coding mRNA transcript, specifically, the six-nucleotide seed sequence (nucleotides 2-7 of the miRNA)[138]. Because of this feature, *in silico* algorithms have been developed that assess sequence complementarity and the thermodynamic stability of the corresponding RNA duplex to predict potential protein-coding targets of a given set of miRNAs[115]. We used the algorithm known as miRanda v3.3a to predict the genes targeted by the newly-discovered miRNA sequences. Predicted miRNA-coding-gene pairs were filtered to have an alignment score ≥ 140 and an energy threshold of ≤ -20 kcal/mol, which represents stronger potential interactions. To find the most likely protein-coding targets, we subjected only the genes predicted to be targeted by at least 10% of the novel miRNA sequences to pathway enrichment analysis using the pathDIP algorithm[116]. These analyses revealed the enrichment of pathways curated from 15 literature sources that included *EGFR* and *ERK* signaling pathways, as well as genes that mediate mRNA stabilization (Figure 3.6). Additionally, the genome-level regulation enacted by these miRNAs can be seen by the linking lines in the center of the circos plot in Figure 3.2D, which further illustrates the functional impact of previously-unannotated miRNA expression and deregulation.

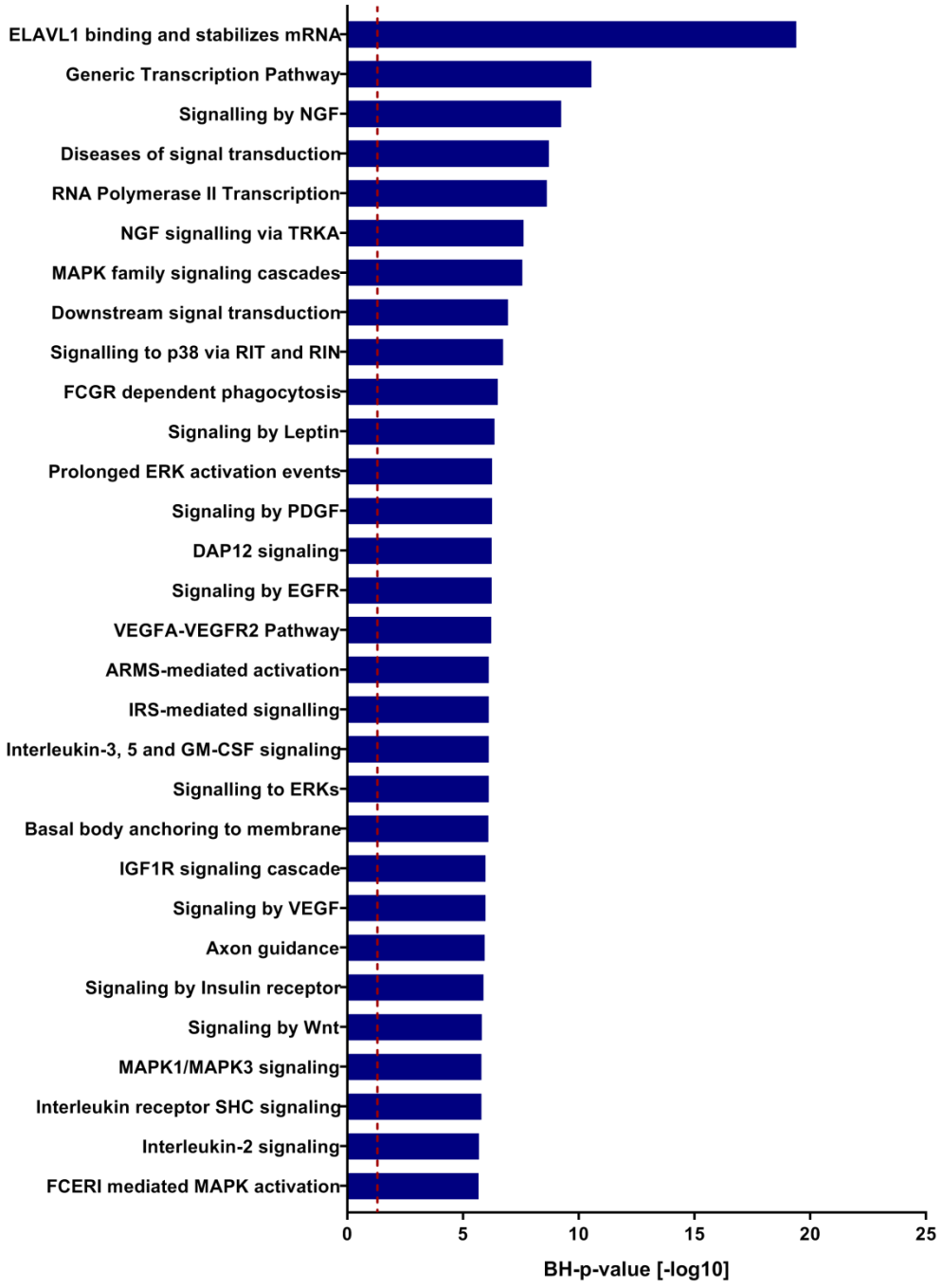

Figure 3.6. Pathways significantly enriched in genes targeted by novel miRNAs discovered in human lung samples.
Pathways were determined using the PathDip algorithm and plotted according to the logged value of their statistical significance in enrichment for genes predicted to be targeted by novel miRNA transcripts. Individual genes in each pathway are predicted to be targeted by at least 10% of previously-unannotated miRNAs. The red line represents the threshold of significance (BH-corrected p-value < 0.05).

3.3.4 Technical validation of the novel miRNA discovery platform

With the discovery of 141 previously-unannotated miRNAs expressed in human lung tissues, I was interested in examining if this phenomenon could be extended to other tissue contexts. Similar to lung adenocarcinoma, while specific molecular events including miRNA disruption have been implicated in ccRCC disease progression and prognosis, patient outcome remains impeded by the limited availability and efficacy of relevant diagnostic and prognostic markers. Following our previous work with the NCI-60 cell lines, I decided to assess clear cell renal cell carcinoma tissue samples to expand upon the landscape of novel miRNA expression in human tissues and diseases.

The OASIS/miRDeep2 algorithm was used to discover previously-unannotated miRNAs in the TCGA-KIRC cohort (Chapter 2.1.3). Similar to the initial landscape of novel miRNA expression in human lung samples, we discovered a total of 376 miRNA candidates in human ccRCC samples. Manual curation as described in Chapter 2.3.2 (BLASTn, GC content) resulted in our final set of 35 and 134 previously-unannotated miRNA candidates detected in non-malignant kidney and ccRCC samples, respectively (Kidney novel miRNAs, Knm). To further confirm the existence of these novel sequences in ccRCC samples, we assessed the expression of the newly-detected miRNA loci in the 8 renal cancer cell lines in the NCI-60 cell line panel. As the renal cell lines are derived from different patients and tumour subtypes, we considered novel miRNA sequences with normalized expression greater than 0.1 in at least one cell line to be detected. Excitingly, we found that 65% (26) of the sequences in non-malignant human kidney samples and 71% (102) in ccRCC patient tumours to also be present in the NCI-60 renal cell lines. Further, there is nearly no sequence overlap between the novel miRNA transcripts detected in renal samples with those discovered in lung tissues. Together, these results first confirm the

existence of novel miRNAs in human tissues and patient-derived cell lines beyond lung samples, and also highlight the efficacy of our platform for novel miRNA discovery.

3.3.5 Novel miRNAs are similarly deregulated and associated with survival in ccRCC tumours as in lung adenocarcinoma samples

To ensure that there were no systematic biases in the discovery of previously-unannotated miRNAs specific to lung tissue we examined the potential roles of novel miRNAs in human ccRCC samples in a similar fashion to the assessment in lung samples. This also allowed me to explore the relevance of previously-unannotated miRNA sequences to human tissue and disease biology. Of the 65 previously-unannotated miRNA loci detected in ccRCC and non-malignant samples, 31 were significantly differentially expressed (Figure 3.7A**;** 15 over-expressed, 16 under-expressed; BH-p < 0.05). Further, the combined expression of these miRNA sequences was sufficient to stratify samples into non-malignant and tumour groups (Figure 3.7A), which has implications for the use of these newly-detected sequences as diagnostic markers.

In line with the exploration of the potential diagnostic and therapeutic utility of these candidates, specific sequences are of particular interest because of the striking magnitude of their deregulation. Knm22_2209 has little to no detectable expression in tumours, representing a 28-fold decrease in expression, while Knm3_1968 and Knm17_1130 show strongly increased expression in tumour tissues, 100- and 13-fold, respectively (Figure 3.7B). We also examined the associations between the expression of these candidate miRNAs and patient outcome. In fact, we found the increased expression of Knm6_2419 to be significantly associated with worsened patient survival, which agreed with its significant over-expression in tumours (Figure 3.7C). Finally, as observed with the previously-unannotated miRNA sequences discovered in human lung samples, the miRNAs newly-detected in ccRCC were predicted to target genes significantly

enriched in relevant kidney cancer pathways. These pathways include the *VEGF* pathway and pathways associated with cellular metabolism, which are important to ccRCC biology.

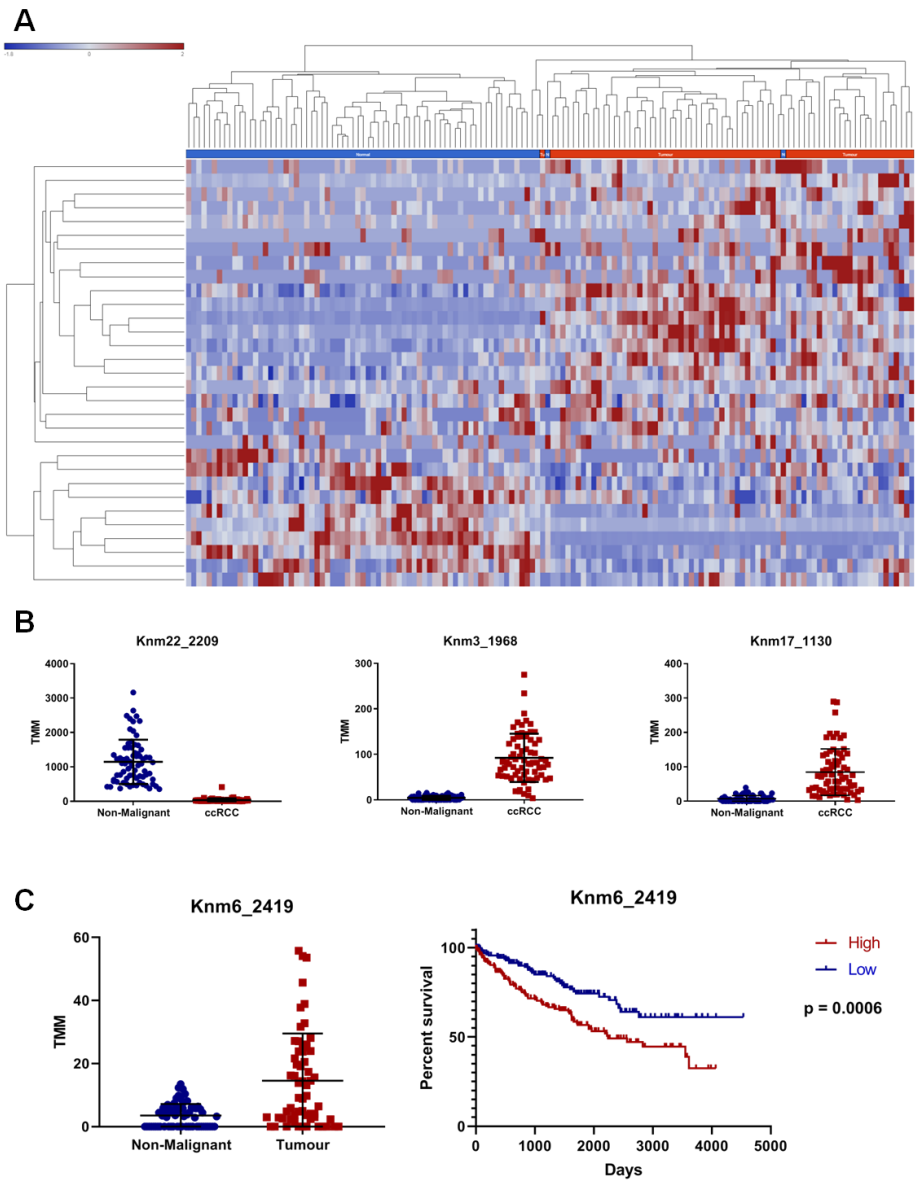

Figure 3.7. Differential expression of novel miRNA sequences in ccRCC samples relative to non-malignant tissues.[3]

A) Unsupervised hierarchical clustering of tumour (red) and non-malignant (blue) ccRCC samples (columns), and kidney novel miRNAs (rows). Relative expression values are represented from low (blue) to high (red). B) Specific examples of deregulated novel miRNA expression between non-malignant (blue) and ccRCC samples (red). All corrected p values are below 0.0001. C) Deregulated expression and associations with patient outcome for Knm6_2419. Patients were stratified into tertiles based on the expression of these novel miRNA transcripts.

[3] Figure 3.7 is adapted from the following publication: [Sage AP], Minatel BC, Marshall EA, Martinez VD, Stewart GL, Enfield KSS, Lam WL. (2018) Expanding the miRNA transcriptome of human kidney and renal cell carcinoma. *International Journal of Genomics* 2018:1-10. (Appendix A; Item 8).

3.3.6 Confirmation of novel miRNA expression patterns *in vitro*

With the confirmation of the widespread but tissue- and context-specific expression of previously-unannotated miRNA sequences in multiple human tissues, I finally aimed to experimentally validate the expression of key miRNA candidates (Knm3_1968 and Knm17_1130). To this end, we cultured TK-10 renal cancer cells as well as HEK-293T embryonic kidney cells and extracted total RNA. After specifically converting our miRNA candidates-of-interest to cDNA, we assessed their relative expression levels by RT-qPCR. Both miRNAs, which were over-expressed in ccRCC samples in the analysis of small RNA sequencing data, had detectable expression and displayed similar patterns of deregulation in these cell lines (fold-change = 8.56 and 16.54, for Knm3_1968 and Knm17_1130, respectively; Figure 3.8). Additionally, we also assessed Knm22_2209 expression, which was not detected in the renal cancer cell line, but it was also not detected in the non-malignant cells. However, the embryonic nature of HEK-293T cells may account for these observations. Collectively, these results confirm the existence of previously-unannotated miRNA sequences *in vitro* and further emphasize the impact of the *in silico* findings.

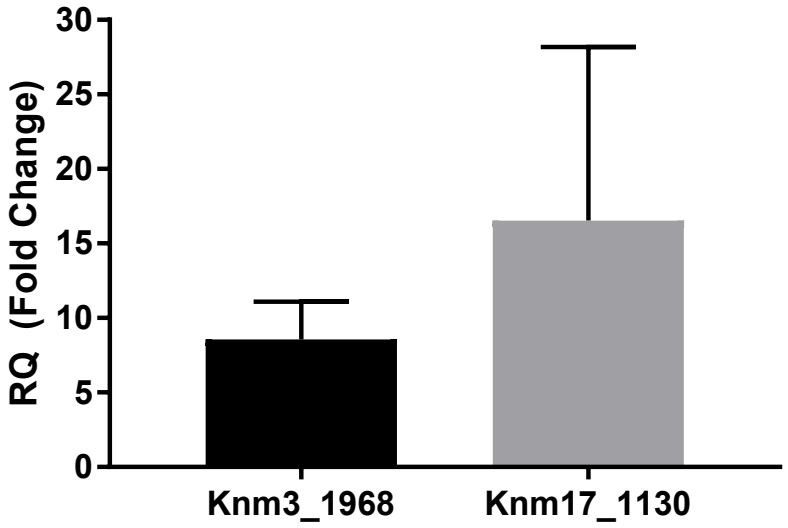

Figure 3.8. Fold change values for novel miRNA candidates in ccRCC (TK-10) relative to non-malignant kidney (HEK-293T) cell lines.[4]
RQ was calculated by analyzing RT-qPCR data, using the $2^{-\Delta\Delta Ct}$ method.

[4] Figure 3.8 is taken from the following publication: [Sage AP], Minatel BC, Marshall EA, Martinez VD, Stewart GL, Enfield KSS, Lam WL. (2018) Expanding the miRNA transcriptome of human kidney and renal cell carcinoma. *International Journal of Genomics* 2018:1-10. (Appendix A; Item 8).

3.4 Discussion

In this study, I assessed high-throughput small RNA sequencing data from TCGA to build a more comprehensive picture of the miRNA transcriptome of human lung tissues, from both non-malignant and malignant perspectives. Leveraging *in silico* tools for novel-miRNA analysis, we discovered 141 previously-unannotated miRNAs in human lung samples from two separate cohorts of RNA-sequencing data, representing a substantial 14.6% increase in the known lung miRNA transcriptome. Together my results highlight the extent of miRNA transcription that is missed in many current analyses, which may provide novel opportunities in future lung cancer research.

In line with the reliance of previous analyses of the human miRNA transcriptome on sequence conservation, I found that the newly-discovered miRNAs displayed remarkably specific expression in lung samples. Comparing non-malignant tissues with lung adenocarcinoma samples, many novel miRNAs were exclusively expressed in only one context. These results are particularly encouraging in light of the need for specific and easily-detectable markers of tumour onset and progression, which is further supported by the associations between individual novel miRNA transcripts and LUAD patient outcome. In fact, these expression patterns combined with the stability of miRNAs in biofluids and FFPE specimens makes these attractive candidates for the development of sensitive diagnostic and prognostic markers, or potential therapeutic targets with limited off-target effects[139-141]. Further, there is a potential for the utility of the specific of novel miRNA sequences to be extended to other cancers, which we have begun to explore in their ability to distinguish malignant from benign nodules in papillary thyroid carcinoma as well as differentiate mesothelioma cases from lung tumours that have metastasized to the pleura (Appendix A; Items 1, 2).

Many of the novel miRNAs expressed in both malignant and non-malignant lung samples were observed to show significant differential expression between these two contexts. These observations suggest the potential for the dysregulation of novel miRNAs to represent an alternative mechanism whereby cancer-associated genes become aberrantly expressed in tumours. While the exact mechanisms of dysregulation for many miRNAs are relatively poorly characterized, common causes of genomic alterations such as biogenesis defects, changes in DNA and histone methylation status, or even infection – as has been described for miR21 in gastric cancer – may be relevant[142,143]. Characterizing the impact and mechanisms of miRNA dysregulation will be integral to uncovering novel miRNAs with oncogenic or tumour-suppressive roles that may be clinically actionable.

Similarly, as miRNAs act through the regulation of protein-coding targets, it was interesting to observe that many of the genes predicted to be targeted by the newly-discovered transcripts were significantly enriched for their involvement in lung cancer pathways, including *EGFR* signaling. While these interactions and associations require validation *in vitro*, my preliminary observations can be used to guide the selection and design of necessary miRNA-inhibition and over-expression experiments. Another area warranting exploration may be the effects of inter-individual genetic variations, due to the heavy reliance on sequence complementarity for miRNA action[144]. Additionally, these features of miRNA gene-targeting in tandem with the discovery of specific novel miRNAs may be able to be used to guide the development of future miRNA-mimics targeting key oncogenic mRNAs[145].

To assess whether the phenomenon of previously-unannotated miRNA expression was widespread throughout human samples from both normal and cancer biology perspectives, I described and validated the expression of these transcripts in human kidney samples. Not only

did I observe similar context-specific expression and deregulation of these transcripts from RNA sequencing data, but I was also able to validate their expression in cell lines by RT-qPCR as well as NGS experiments. Moreover, these transcripts displayed patterns of deregulation in line with sequencing data between malignant and non-malignant cell lines. The results of my analysis into ccRCC serve to extend the impact of novel miRNA discovery beyond lung tissues, and present implications for all cancer types.

Beyond the biological and therapeutic information uncovered by the discovery of novel miRNA sequences, my results highlight the usefulness of whole-genome sequencing and subsequent *in silico* analysis in basic cancer research. While *in silico* algorithms are continuously being optimized and refined, they provide unique opportunities for large-scale analyses of gene expression and regulatory patterns. In this case, I combined the use of established miRNA-discovery algorithms with our custom curation pipeline. Filtering candidate sequences by their miRNA-like characteristics such as GC content, homology with known sequences, and their detectable expression in samples allowed me to ensure a much lower degree of false-positive predictions. These observations were confirmed by my validation of sequence expression in additional cohorts, and *in vitro.*

Collectively, this work highlights the misrepresentation of a significant portion of the transcriptome, both in quantity and in relevance to cellular and disease biology. My results provide a foundation for further *in vivo* validation of these sequences, their targets, and their subsequent functions. With this information, the specific expression and gene regulation of novel miRNA transcripts may be exploited in the clinic in the management of lung cancer and extended to any type of cancer.

Chapter 4: Examining the expression of long non-coding RNAs in infiltrating immune cells in the lung tumour microenvironment

4.1 Introduction

4.1.1 The lung cancer immune microenvironment and immunotherapy

Immune cells able to infiltrate bulk tumours can act as a natural defense system against malignant growth. Paradoxically, they can also actively contribute to tumour proliferation, growth, and survival. In fact, features such as chronic inflammation, cytokine release, and hypoxia have been associated with tumour development[146-148]. Thus, comprehensive characterization of the cells that make up the tumour microenvironment is required to adequately understand the individual contributions of various cell types to tumourigenesis.

Uncovering the mechanisms used by tumours to evade infiltrating immune cells has catalyzed an immense body of work towards therapeutic agents that can help the cells of the immune system to recognize and attack tumour cells[149]. The most prominent immunotherapeutic strategy is the use of immune-checkpoint inhibitors, which have shown particular efficacy in cancers resulting from high mutagen exposure, such as melanoma and non-small cell lung cancer[150]. These agents (e.g. Pembrolizumab) are antibodies that target and block the interactions between inhibitory receptors (e.g. programmed cell death receptor 1; PD-1) and their ligands (e.g. PD-L1), which are frequently upregulated by tumour cells as a mechanism of immune evasion[26,151].

While immunotherapy regimes have been shown to be effective, an outstanding clinical challenge is the wide variation in patient-response to treatment[152,153]. Currently, numerous features of the tumour microenvironment have demonstrated potential prognostic value, such as

the expression of the genes encoding PD-L1 and interferon gamma (IFN-γ), the mutational load of the tumour, as well as the relative proportion of infiltrating immune cells[154-157]. For example, a higher proportion of cytotoxic T cells is associated with improved outcome for NSCLC patients[158]. However, accurate immunophenotyping from bulk tumour samples can be difficult and can require a number of different analytical techniques[159]. These factors highlight the importance of identifying detectable microenvironment-specific markers for the efficacy of immunotherapy regimes.

4.1.2 Non-coding RNAs in immune cells and cancer immunology

The increasing accessibility of NGS technologies have corresponded to an increase in the search of disease and treatment markers at the genomic level, as well as uncovering new molecular features of the tumour microenvironment. This is particularly applicable to immunotherapy, wherein gene expression profiling approaches such as microarrays have been used to explore mRNA expression of immune-related genes that may be indicative of distinct immune cell types within tumours[160]. Indeed, immune-gene signatures, namely those representing an active immune response such as *NKG7*, *IDO1*, and *IFNG*, have been described to be differentially expressed between patients that are responsive to treatment versus those that are not[161]. However, many of the genes identified in these panels can be expressed at varying levels in cells and in other cell types, which convolutes accurate analysis[162].

Demarcating immune cell infiltrate and response to immunotherapy through high-throughput sequencing of tumour samples would provide advantages over other techniques such as immunohistochemistry (IHC), as the same data could also be assessed for other molecular features of the tumour samples (e.g. mutation status or prognosis). In fact, there are a number of

algorithms that have been built to integrate gene expression information to generate estimations of relative proportions immune cell infiltrate, such as CIBERSORT, and TIMER[163,164]. However, these estimates are unable to provide accurate determination of cellular proportions and specific composition within and between tumours, necessitating the use of genome-wide high-throughput techniques to improve these analyzes[165].

As specifically-expressed transcripts representing alternative mechanisms of gene regulation, the consideration of ncRNAs may present a unique opportunity to address these issues. Unlike sncRNAs, lncRNAs are readily detectable in common next generation sequencing experiments, but have broad functional roles and patterns of expression that remain poorly characterized. Long non-coding RNAs have been described to be specifically expressed in cancer tissues and also have demonstrated gene-regulatory roles in immune-related phenotypes, including development, homeostasis, and response to infection[166-168]. The specific regulation enacted by lncRNAs enables the fine-tuning required by the immune system to balance pro- and anti-inflammatory phenotypes (Figure 4.1). For example, the lncRNA *Morrbid* negatively regulates the expression of its neighbouring gene *Bcl2l11,* promoting differentiation of myeloid progenitor cells[169]. Despite their recognized roles in immune cell function, immune-related lncRNA research has been focused on particular transcripts relevant to a single phenotype.

Leveraging the potential of lncRNA expression patterns to strengthen our knowledge of tumour immunology and mechanisms of the immunotherapy response, requires a high-level understanding of their expression in immune cells and tumour infiltrating lymphocytes. In light of their roles in the immune system and the tumour microenvironment, I hypothesized that lncRNAs are specifically expressed in human immune cells, which can be detected in bulk tumour samples.

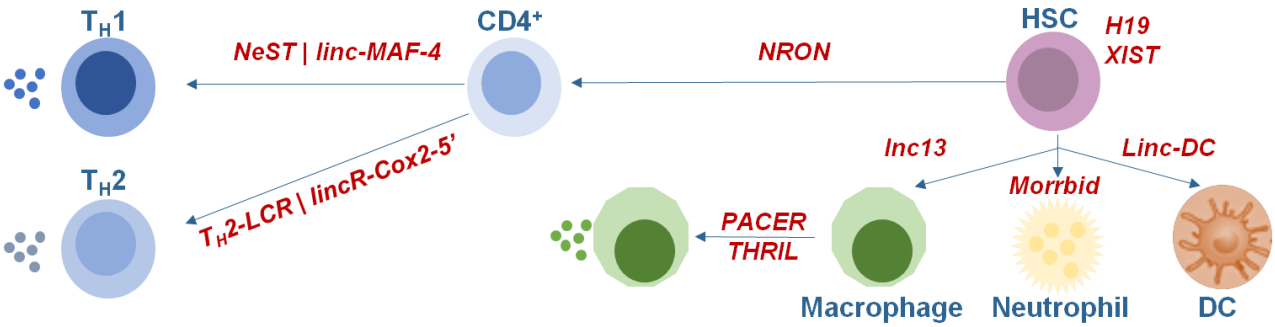

Figure 4.1. Long non-coding RNAs involved in immune cell differentiation.
The differentiation of immune cells from hematopoietic stem cells (HSCs) relies on the gene regulation exerted by numerous lncRNAs (red text). Figure is adapted from Chen et al, with permission from the publisher[169].

To this end, I first sought to characterize the landscape of lncRNA expression in healthy human immune cells, which then allowed me to assess lncRNA expression from immune cell populations in the lung tumour microenvironment. Together, these observations can be used to facilitate further exploration of lncRNAs as both mediators of the tumour-immune response and as markers of immunotherapy effectiveness, particularly in immunogenic malignancies such as lung cancer.

4.2 Methods

4.2.1 Analysis of lncRNA expression in RNA sequencing data from sorted healthy human immune cells

To probe lncRNA expression in tumour infiltrating lymphocytes, their expression was first assessed in whole genome sequencing data from two datasets of purified, flow-sorted cells from six immune-cell subsets ($CD8^+$ T, $CD4^+$ T, B, Monocytes, Neutrophils, and Natural Killer) obtained from a healthy donor (GSE62408[170]) as well as the same immune-cell subsets from patients with various auto-immune diseases, which included 4 healthy donors (GSE60424[171]). As described in Chapter 2.2.2, following quality filtering and sequence alignment to the human reference genome (STAR v2.4.1d), reads were quantified using the Cufflinks algorithm v2.2.1[104] and normalized using the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) method[172]. Loci encoding lncRNAs were extracted based on the Ensembl v89 annotation. Transcripts were filtered based on their expression in samples, wherein lncRNAs with summed expression levels $\geq$ 1 FPKM across all samples were considered in further analyses. Gene expression patterns were analyzed through hierarchical clustering and differential expression tests described in Chapter 2.3.3. Figure 4.2 summarizes the analysis pipeline.

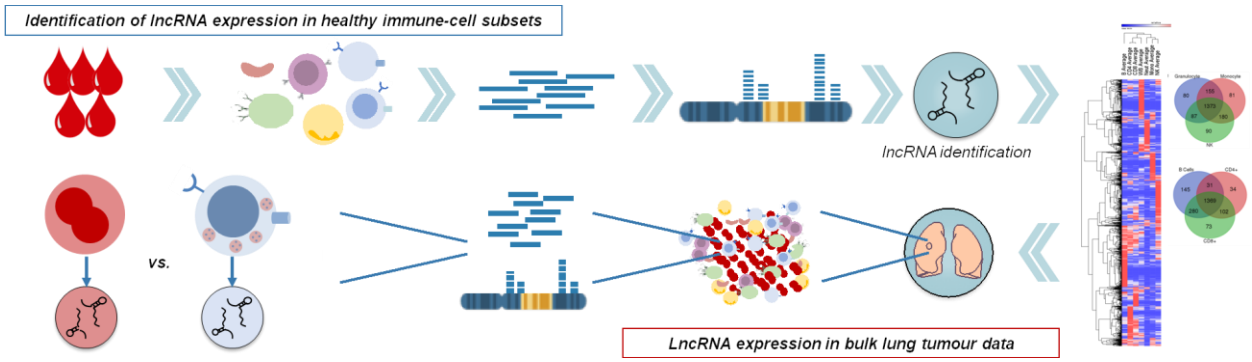

Figure 4.2. Analysis pipeline.
LncRNAs were first assessed for their expression in healthy human immune cells (upper portion), which was applied to examine the contribution of infiltrating immune cells to bulk lung tumour data (lower portion).

4.2.2 Assessment of lncRNA expression patterns in bulk tumour data

In order to assess the expression of immune-associated lncRNAs in bulk lung tumours, the TCGA-LUAD cohort (n = 54 pairs; Chapter 2.1.2) was used to first identify lncRNAs significantly deregulated (two-tailed Mann-Whitney U-test, BH-p < 0.05) between lung tumours and matched adjacent non-malignant samples. In a similar manner, immune-associated lncRNAs identified in healthy human immune cells were then assessed for their deregulated expression patterns in both the TCGA-LUAD and BCCA-LUAD cohorts. Hierarchical clustering analyses (Chapter 2.3.3) were used to identify whether the expression of immune-associated lncRNAs was able to stratify samples by disease status.

Three separate markers of tumour immune-cell infiltrate were used in the identification of lncRNAs associated with tumour-infiltrating lymphocytes. First, tumours were stratified by their average expression of 200 known tumour-associated antigens (TAAs), as the presence of these proteins is known to trigger an adaptive immune response to the tumour[154,173]. From this subset of genes encoding tumour-associated antigens, tertiles were defined, and lncRNAs were assessed for differential expression between tumours with high average expression of tumour-associated antigens versus those with lower average expression. Secondly, immune-associated lncRNAs expressed in the tumour samples of the TCGA-LUAD cohort were assessed for their correlation (Spearman's Correlation; r > 0.4; p < 0.05) with the expression of *PTPRC* (encodes CD45), the canonical marker of immune cells. Finally, associations between lncRNA expression and Leukocytes Unmethylation for Purity (LUMP) scores were examined. Briefly, LUMP scores are an established marker of immune cell infiltrate in tumours, which are a measure of the methylation at sites unmethylated in immune cells but frequently methylated in cancer cells[174]. Thus, lncRNAs with significantly negative associations with LUMP scores (Spearman's

Correlation; $r < -0.4$; $p < 0.05$) were considered associated with immune cell content. Additionally, as samples from the BCCA-LUAD cohort are microdissected to 80% tumour-cell content, only TCGA-LUAD tumours were considered for infiltrating immune cell analyses.

4.2.3 Analysis of lncRNA expression in single-cell RNA sequencing data

Following the assessment of lncRNA expression in bulk tumour data, I sought to examine whether these lncRNAs were similarly expressed in tumour-infiltrating lymphocytes. To accomplish this, we examined RNA sequencing data at the single cell resolution (scRNAseq) obtained from a study that sought to evaluate the subpopulations of cells in the lung tumour microenvironment[175]. By analyzing coding-gene expression programs, this study was able to recapitulate the tumour microenvironment of five lung tumour samples into 52 sub-clusters of stromal cells, organized into seven overarching cell types (alveolar, fibroblast, endothelial, T cells, B cells, myeloid, epithelial). The transcriptomic data generated from these samples were extensive, but subsequent analyses were largely focused on cell-type identification and marker-gene expression. Thus, these data provided me with an opportunity to assess the expression of the immune-related lncRNAs from single cell populations within a lung tumour microenvironment. As such, we subjected these scRNAseq data to our lncRNA-analysis pipeline and examined their expression in the clusters of infiltrating stromal cells, as well as their co-expression with known protein-coding markers of gene expression. We also assessed associations with the expression of these immune lncRNAs and the relative proportions of given cell types, focusing on clusters of T cells. Finally, to assess the potential translational value of the immune-related lncRNAs, selected transcripts with cytotoxic expression patterns were assessed for their associations with patient outcome, as described in Chapter 2.3.3.

4.3 Results

4.3.1 LncRNAs are specifically expressed in healthy human immune cells

Despite the emerging importance of lncRNAs to both immune biology and cancer development, the landscape and relevance of lncRNA expression in tumour infiltrating lymphocytes is poorly understood. To explore the landscape of lncRNA expression in tumour-infiltrating lymphocytes, gene expression data from flow-cytometry sorted and purified healthy human immune cell subsets were first explored. High throughput RNA sequencing data from granulocytes, monocytes, $CD8^+$ T cells, $CD4^+$ T cells, and B cells from healthy donors were comprehensively analyzed for lncRNA expression patterns. Of the 13,006 lncRNA genes annotated in Ensembl v89, 4,919 had detectable expression (summed FPKM > 1 across all samples). Further, just over one quarter of the lncRNAs identified in immune cells were expressed in all cell types, which indicates their potentially conserved functions and highlights the ubiquitous expression of this class of transcripts in human immune cells (Figure 4.3A). As lncRNAs are known to have relatively tissue- and context-specific expression patterns, it was interesting that 15% of lncRNAs were found to be exclusively expressed in only one cell type, compared to just 3% of protein-coding genes (Figure 4.3B). The cell-type specific expression of lncRNAs is further emphasized by examining unsupervised hierarchical clustering analyses, wherein lncRNA expression alone was able to recapitulate immune cell differentiation and developmental lineage patterns (Figure 4.4). Together, these data highlight not only the widespread transcription of lncRNAs in human immune cells in various contexts, but also suggest the importance of these lncRNAs to the differentiation and function of the immune system.

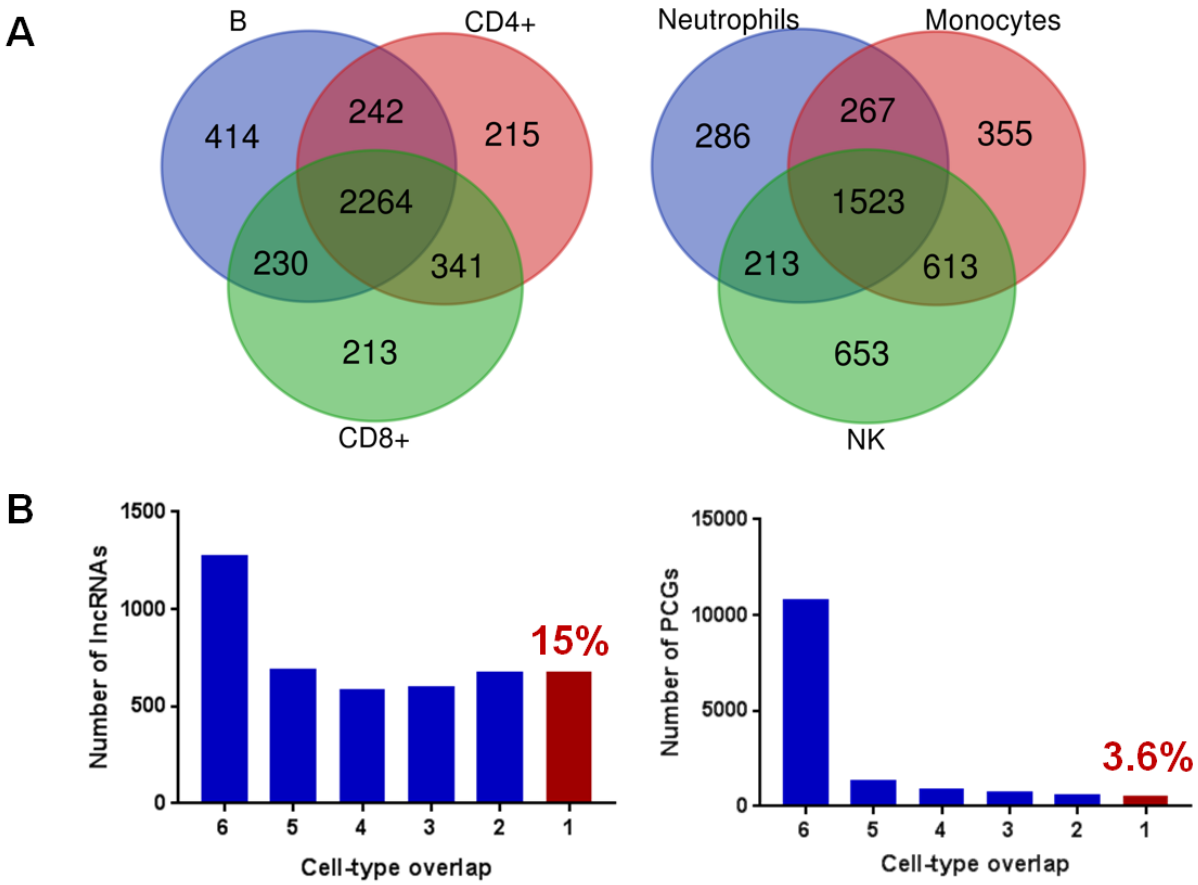

Figure 4.3. Distribution of lncRNAs expressed in healthy human immune cells.
A) Venn diagram representing the number of lncRNAs expressed in adaptive (left) and innate (right) immune cell subsets. Overlapping regions denote sequences expressed in multiple cell subsets. B) Histogram of the cell-type specificity of lncRNA (left) and protein-coding (right) transcripts in human immune cells. Cell-type overlap (x-axis) denotes the number of cell subsets with detectable expression of a given lncRNA (quantified as number of lncRNAs; y axis). The red bar represents the number of transcripts specific to only one of the cell subsets analyzed (quantified as percent of total; red text).

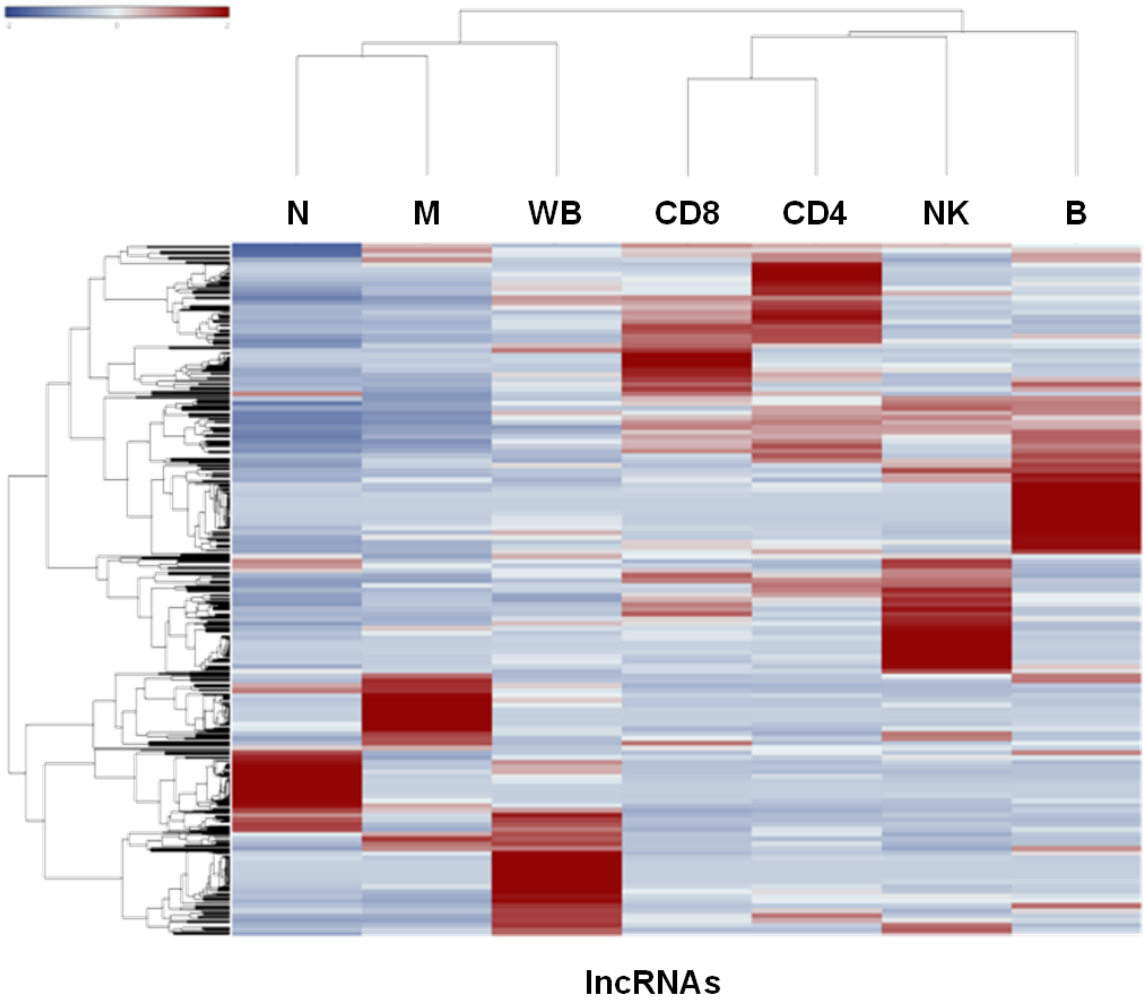

Figure 4.4. Expression of lncRNAs in healthy human immune cells.
Unsupervised hierarchical clustering of lncRNAs expressed in purified immune cell subsets (N: neutrophils; M: monocytes; WB: whole blood; CD8: $CD8^+$ T cells; CD4: $CD4^+$ T cells, NK: natural killer cells). LncRNA expression is represented from high (red) to low (blue).

4.3.2 Known and novel lncRNA expression patterns are indicative of function in immune cells

Functional classification of lncRNAs remains impeded by complex secondary structure, veiled functional motifs, and poor conservation between species. Although hundreds of lncRNAs have been functionally characterized to date, no broad functional categories have emerged and many lncRNAs are labeled as artefacts of transcription[78]. Together, this necessitates consideration of alternative factors for assessing potential functional roles for lncRNAs, such as genomic location and orientation, sequence homology, as well as patterns of expression. Thus, we first examined the expression patterns of specific lncRNAs to identify whether the observed patterns were congruent with ascribed functions of previously-characterized lncRNAs. The tumour suppressive lncRNA *MEG3,* recently shown to regulate IL-1β abundance in alveolar macrophages[176], is observed to be highly expressed in healthy human monocytes with little to no detectable expression in other cell types (Figure 4.5A). Similarly, *IFNG-AS1* (also known as *NeST*) displays an expression pattern characteristic of a recently uncovered function in the regulation of IFN-γ production in $CD8^+$ T cells[177] and the Th1-lineage specificity of IFN-γ[178], although its relatively high expression in B cells has not been described (Figure 4.5B).

The idea of using the expression patterns and genomic location of lncRNAs to elucidate potential biological function can also be applied to lncRNAs that have yet to be functionally characterized. As a representative example, *AC008750.1* (and transcript variant *AC008750.2*; located on chromosome 19q13.41) display markedly high expression levels in natural killer cells, as well as detectable but less pronounced expression in $CD8^+$ T cells; cell subsets which share analogous cytotoxic function (Figure 4.5C and D). Interestingly, 20 kilobases (kb) upstream of *AC008750* sits the gene encoding the NKG7/GMP-17 protein (*NKG7*; Figure 4.5E). NKG7 is

involved in the granulation response in both natural killer and cytotoxic T cells, and has been shown to demarcate cytotoxic effector function in $CD8^+$ T cells[179]. Conversely, the expression of *SIGLEC10*, another immune-related gene that is transcribed from a region overlapping *AC008750*, is most prevalent in neutrophils, monocytes, and B cells (Figure 4.6). These observations suggest that the expression of *AC008750* is not merely a passenger of transcription in cytotoxic immune cells. Specifically, these results suggest a potential role for *AC008750* in the regulation of cytotoxic function, and more broadly, highlight the relevance of examining widespread lncRNA transcription programs to immune biology in healthy and disease contexts.

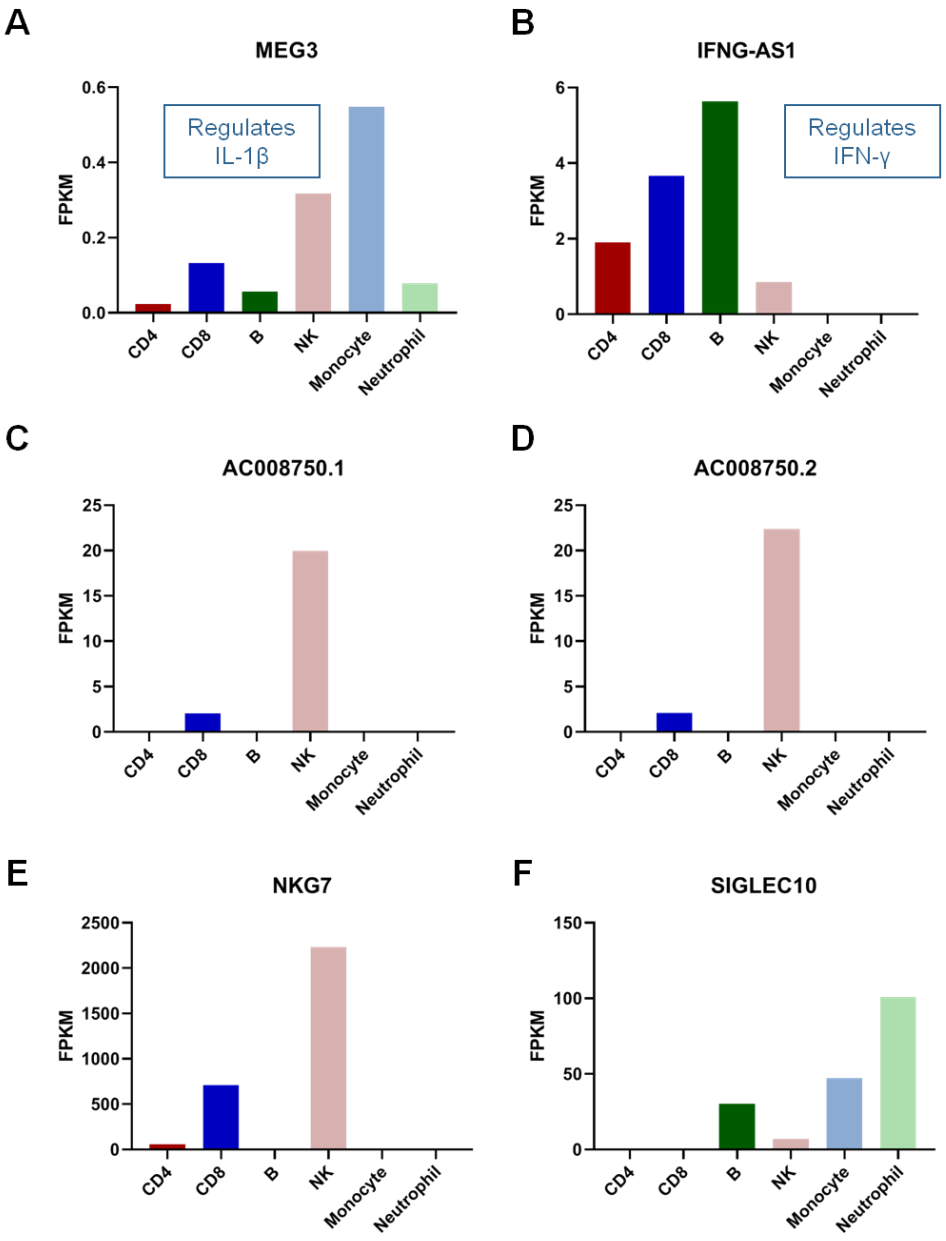

Figure 4.5. Expression patterns of long non-coding RNAs and protein-coding genes in healthy human immune cell subsets.
The expression patterns of lncRNAs with known immune function, *MEG3* (A) and *IFNG-AS1* (B); previously-uncharacterized lncRNAs, *AC008750.1* (C) and *AC00850.2* (D); and the protein-coding genes at this genomic region, also with known immune function, *NKG7* (E) and *SIGLEC10* (F), were assessed in healthy immune cell subsets.

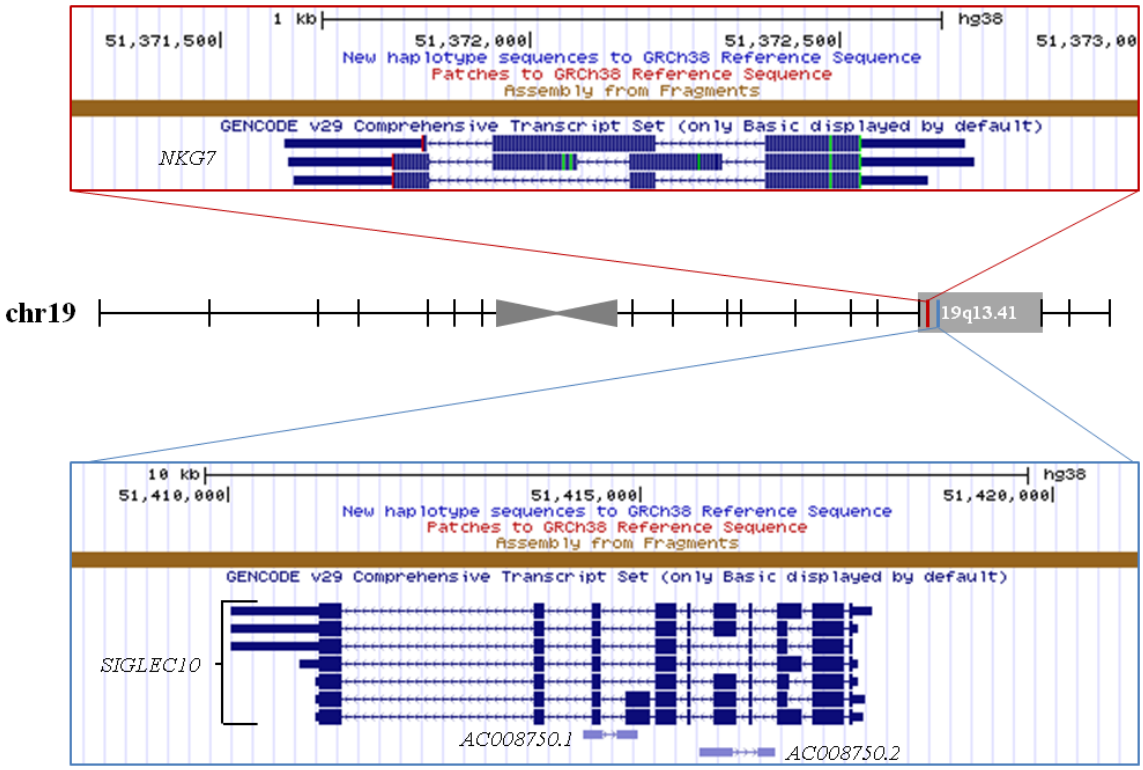

Figure 4.6. Genomic region (Chr19, q13.41) highlighting the location and orientation of *NKG7*, *SIGLEC10*, *AC008750.1*, and *AC008750.2* genes.
This figure was designed using images obtained from the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway), where filled boxes represent exons, lines represent introns, and chevrons represent the direction of transcription.

4.3.3 LncRNAs are deregulated in tumours but may result from tumour impurity

LncRNAs have frequently been described to be differentially expressed in many solid tumours, compared to non-malignant tissues[180,181]. Additionally, the importance of stromal cells in bulk tumours and their relevance to disease progression and treatment has become a major focus of basic cancer research. However, as lncRNA expression profiles in stromal cells are only superficially described, the contribution of these cells to the analysis of lncRNA expression and potential function in tumours is poorly understood. Thus, I aimed to explore the contribution of tumour-infiltrating immune cells to the analysis of lncRNA expression data from bulk tumours.

Confirming the observations of previous studies, the general dysregulation of lncRNAs in lung tumours from the TCGA-LUAD cohort was observed to a relatively high degree (Figure 4.7)[182]. In fact, 679 lncRNAs displayed significantly deregulated expression in tumours relative to matched non-malignant tissues (BH-p $\leq$ 0.05). These include lncRNAs with established roles in lung cancer, such as *PVT1* (Average fold-change = 6.70, BH-p = $8.94*10^{-7}$)[183]. When examining only the 815 immune-associated lncRNAs expressed in the TCGA-LUAD cohort, 194 were significantly over-expressed in tumours while 156 were significantly under-expressed (Figure 4.8A), which is suggestive of their potential oncogenic or tumour-suppressive functions. However, given the observed expression of these lncRNAs in immune cells and the relatively low requirement for tumour cell content of TCGA samples (>60%)[174], I expected these results to be altered when examining the same immune-associated lncRNAs in a cohort of microdissected tumours. Indeed when the same analyses were performed in the BCCA-LUAD cohort, only 61 lncRNAs are significantly over-expressed, with 321 significantly under-expressed (Figure 4.8B). Further, 209 of the differentially expressed lncRNAs are significantly deregulated in both TCGA and BCCA cohorts, yet the cohorts show different directionality for 68 transcripts. As a prime

example, *TUG1* has been shown in multiple studies to be associated with NSCLC proliferation, with some citing its under-expression as a mechanism of tumour development, while others describe its increased expression in tumours[184,185]. In these analyses, *TUG1* is observed to be over-expressed in tumours in the TCGA-LUAD cohort, while it is under-expressed in the BCCA-LUAD cohort. These inconsistent results may be in part due to the presence of non-tumour cells depending on the samples used and suggest that the consideration of tumour purity may help to elucidate the true effects of *TUG1* deregulation. The stark contrast in immune-associated lncRNA expression observed depending on the purity of the samples warrants caution when examining the potential dysregulation of lncRNAs in tumours.

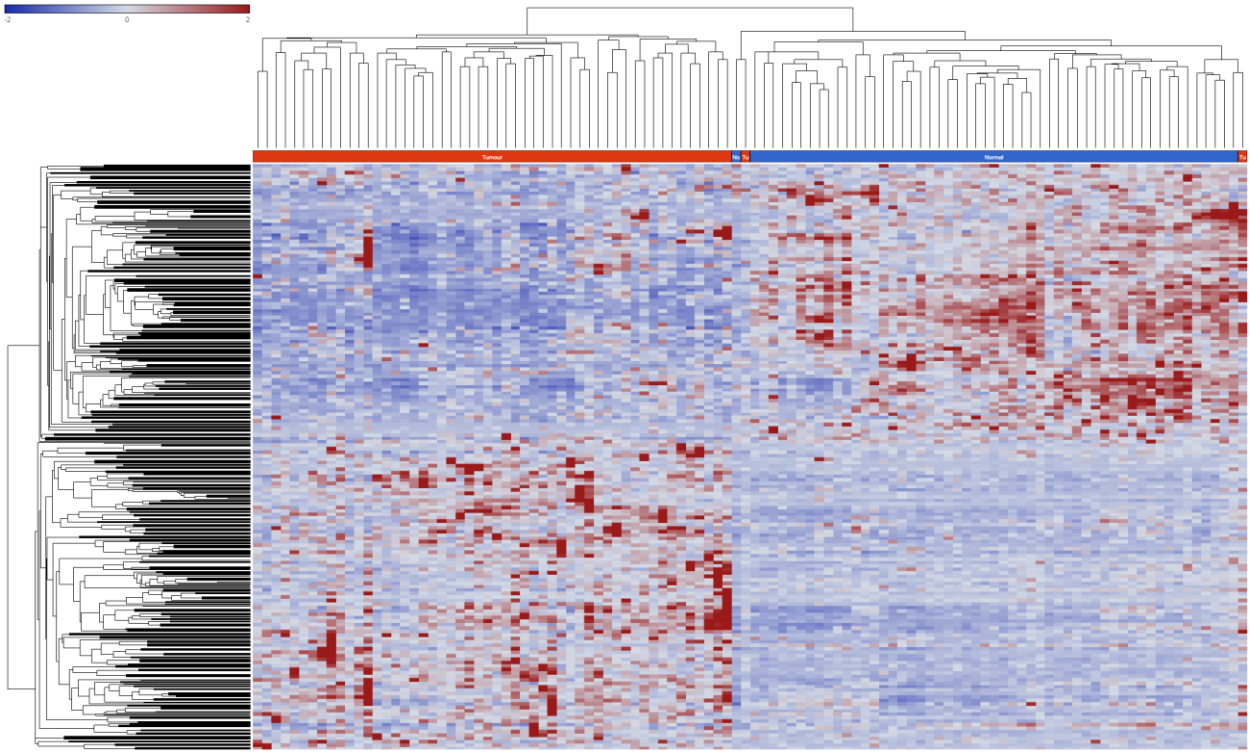

Figure 4.7. General dysregulation of lncRNAs between lung adenocarcinoma samples and matched adjacent non-malignant tissues.
Samples are organized along columns, where tumours are represented in red and non-malignant samples in blue. LncRNAs are organized along rows, with low to high expression represented by dark blue to dark red, respectively. Both samples and transcripts were clustered in an unsupervised fashion as described in Chapter 2.3.3. Only significantly differentially expressed lncRNAs are shown.

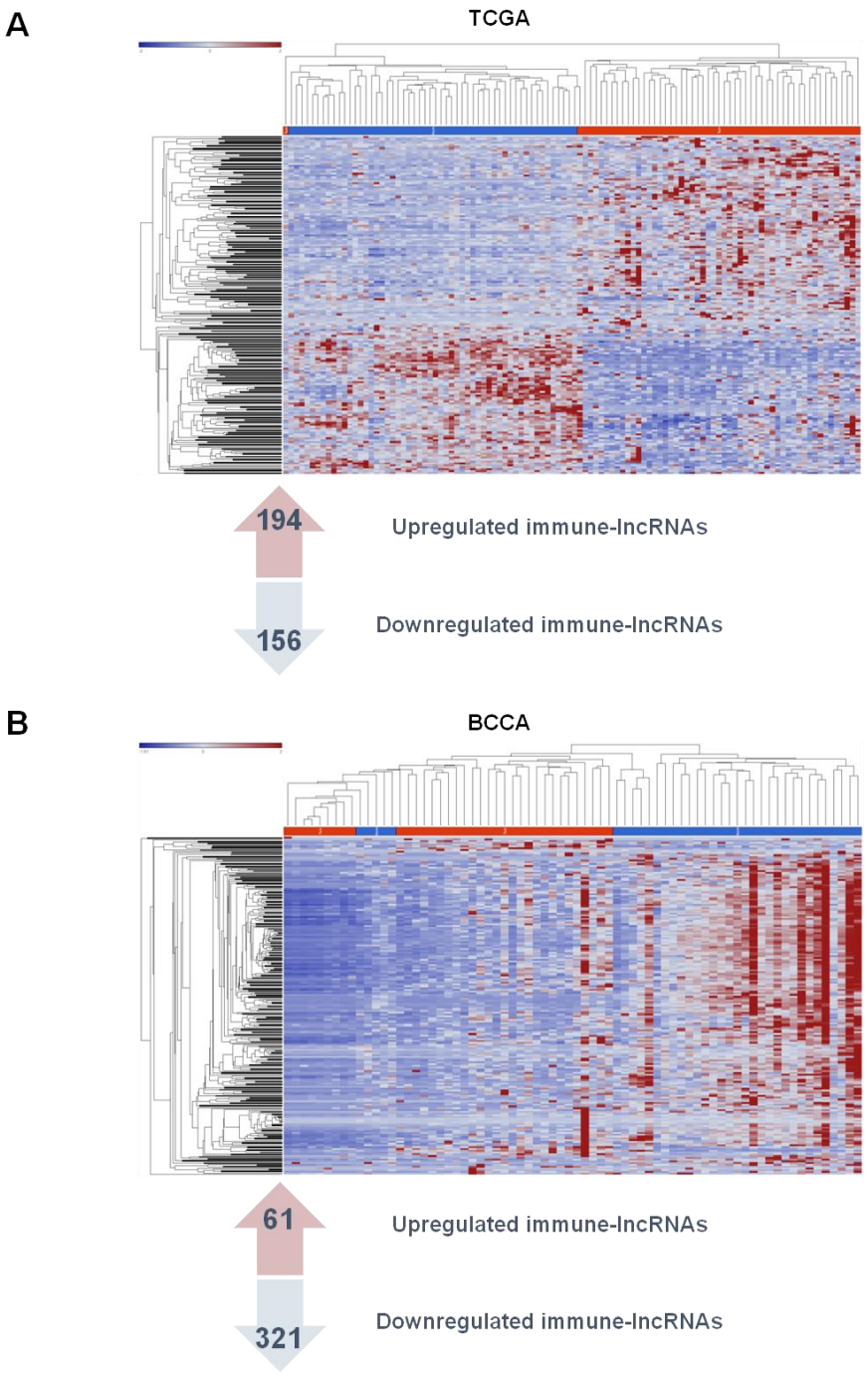

Figure 4.8. Dysregulation of immune-associated lncRNAs in bulk tumour data in TCGA-LUAD (A) and BCCA-LUAD (B) cohorts.
Heat map of immune-associated lncRNA expression from low (blue) to high (red). Samples (columns) and lncRNAs (rows) are organized based on unsupervised hierarchical clustering analysis, where blue bars represent non-malignant tissues and red bars are tumour samples. Only significantly differentially expressed lncRNAs are shown. Both samples and transcripts were clustered in an unsupervised fashion as described in Chapter 2.3.3.

To further confirm whether the observed differential expression of immune-associated lncRNAs in bulk lung tumour data from the TCGA-LUAD cohort may be confounded by the presence of infiltrating lymphocytes, we assessed the associations between the expression of these lncRNAs and established immune-cell markers. First, as the expression of TAAs is known to be immunogenic, we stratified samples into tertiles by their average TAA expression[186]. We observed 152 differentially-expressed immune-associated lncRNAs to be significantly differentially expressed between tumours with relatively high TAA expression and those with low, including a number of lncRNAs specifically expressed in one of the cell subsets analyzed such as *AC027763.2* (Neutrophils), *LPP-AS2* (Monocytes), and *THAP7-AS1* (Neutrophils) (Figure 4.9A). Next, of these lncRNAs we found 11 that were strongly and significantly correlated ($r > 0.4$, $p < 0.05$) with the expression of the gene encoding CD45 (*PTPRC*), a canonical marker of immune cells. For example, *linc00426* is significantly overexpressed in tumour samples (average fold-change = 2.25, BH-p = 0.0107), but strongly correlated with *PTPRC* expression ($r = 0.725$, $p < 0.0001$) (Figure 4.9B). Finally, we also assessed the association of lncRNA expression with an established marker of tumour-infiltrating lymphocytes, the LUMP score. This score represents the relative methylation values of 44 CpG sites frequently unmethylated (<5%) in immune cells, and averagely methylated (>30%) in all tumour types[174]. Similarly, we found 7 differentially-expressed immune-associated lncRNAs strongly negatively correlated with LUMP score ($r < -0.4$, $p < 0.05$; e.g. *linc00426*, Figure 4.9C). Together, not only do these results emphasize the necessary consideration of tumour purity and infiltrating immune cells when examining bulk tumour data, but also highlight potential novel functions for lncRNAs from the perspective of tumour-infiltrating lymphocyte activity.

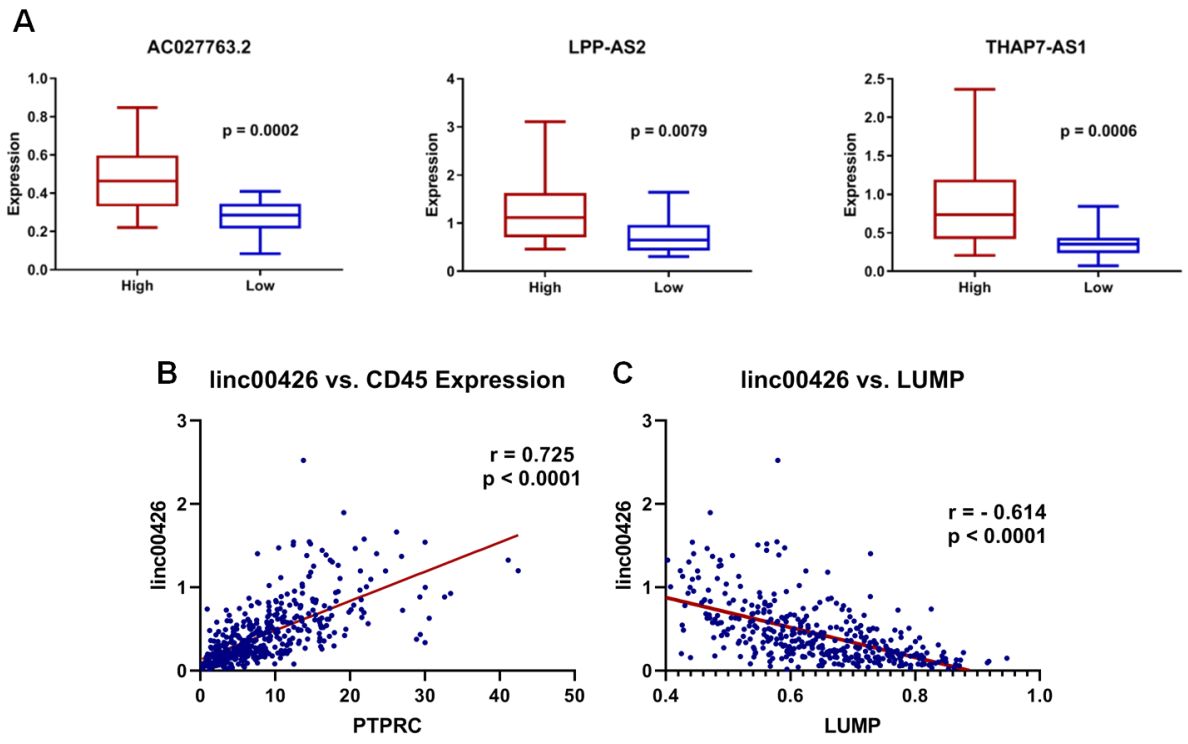

Figure 4.9. LncRNA expression is confounded by infiltrating immune cells in bulk tumour data.
A) Differential expression of selected cell-type specific immune-associated lncRNAs in tumours with high (red) versus low (blue) tumour-associated antigen expression. B) Spearman's correlation of *linc00426* with CD45 ($r = 0.725$, $p < 0.0001$), where blue dots represent tumour samples. C) Spearman's correlation of *linc00426* with LUMP scores ($r = -0.614$, $p < 0.0001$), where blue dots represent tumour samples.

4.3.4 Immune-lncRNAs are co-expressed with immune cell markers in single cell RNA sequencing data isolated from bulk lung tumour samples

The observation that lncRNAs expressed in healthy human immune cell subsets were also associated with markers of immune infiltration in bulk tumours led me to ask whether it was possible to identify similar trends in tumour sequencing data at the single cell resolution. To accomplish this, we examined the expression of the immune-related lncRNAs in scRNAseq data of stromal cells isolated from five lung tumour samples[175]. The original analysis of these data yielded 52 stromal cell subsets, which included 9 subsets each of T and B cells.

We first assessed whether the lncRNAs associated with immune cells in bulk tumour data were co-expressed with canonical immune cell markers in the T and B cell subsets. Indeed, we observed lncRNAs with specific or strongly elevated expression in healthy $CD4^+$ and/or $CD8^+$ T cells to display congruent expression patterns in clusters of stromal cells isolated from bulk lung tumours (e.g. *linc00649*; Figure 4.10), as well as an expression association with the three subunit genes of CD3, an established marker of T cells. These observations serve to provide further evidence suggesting the detectable and specific expression of numerous lncRNAs from tumour infiltrating lymphocytes.

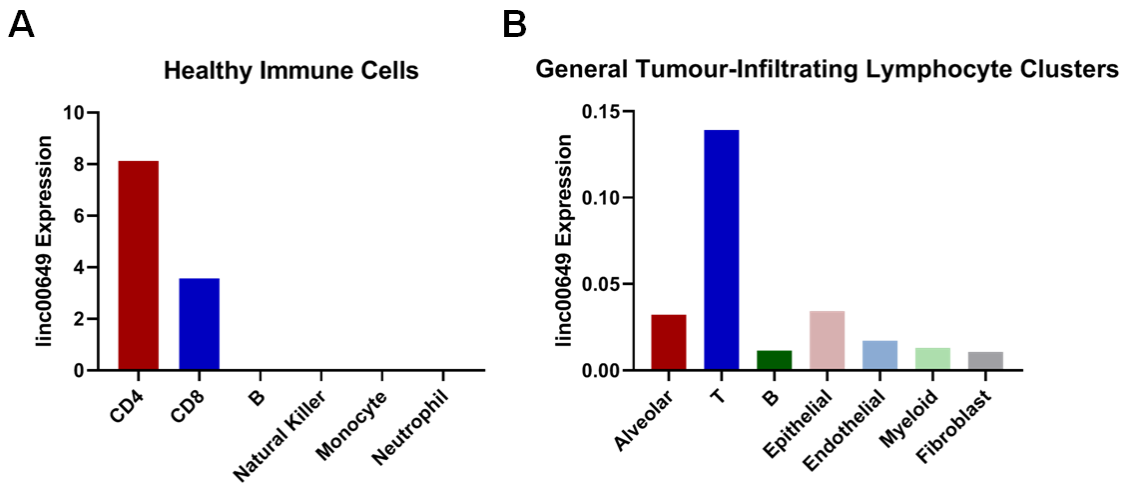

Figure 4.10. linc00649 expression across immune cell subsets.
A) *linc00649* expression in healthy human immune cell subsets. B) *linc00649* expression in subsets of stromal cells isolated from bulk lung tumour samples.

Interestingly, when the broad cell categories are further stratified into their specific cell subsets, such as the separation of T cells into $CD4^+$, $CD8^+$, and natural killer cells, lncRNA expression patterns that confirm our earlier observations in healthy human immune cells can be seen. For instance, *linc00861* was observed to be relatively highly expressed in healthy $CD4^+$, $CD8^+$, and natural killer cells from the initial dataset that was examined (Figure 4.11A). We also observed the expression of this lncRNA to be associated with decreased tumour purity in bulk tumour samples, in that it was strongly negatively correlated with LUMP scores in the TCGA-LUAD cohort (Figure 4.11B, $r = -0.556$, $p < 0.0001$). Analysis of *linc00861* expression in scRNAseq data from stromal cells in the lung microenvironment revealed a similarly high expression in the cluster representing T cells, relative to all other stromal cell clusters (Figure 4.11C). Further examination of the specific sub populations within the T cell cluster found that much like in healthy immune cells, *linc00861* was highly expressed in natural killer cells with detectable expression in $CD4^+$, $CD8^+$ cells, but relatively lower expression in all B cell subsets assessed (Figure 4.11D). Together, this suggests not only the expression of this and other lncRNAs from immune cells in the lung tumour microenvironment, but also the potential functional relevance of lncRNAs such as *linc00861* to tumour biology through key immune pathways.

Finally, as demonstrated in Chapter 3, tissue- and cell-type-specific expression patterns of non-coding RNAs may have insightful implications for their clinical translation as markers of disease progression and prognosis. Although the proportion and identity of infiltrating immune cells in tumours have been observed to display significant associations with patient outcome, it remains a challenge to easily identify the specific cells in a three-dimensional tumour microenvironment with current techniques[159]. Thus, I explored the association of *linc00861*

(highly expressed in cytotoxic T cells) with patient survival and found that its low expression is significantly associated with a worsened prognosis for patients ($p = 0.0273$, Figure 4.11E). Although these observations can be attributed to a potentially lower proportion of anti-tumour immune cells in patients with low expression of this lncRNA, they support the expression of *linc00861* from immune cells and highlight translational utility of ncRNA analysis.

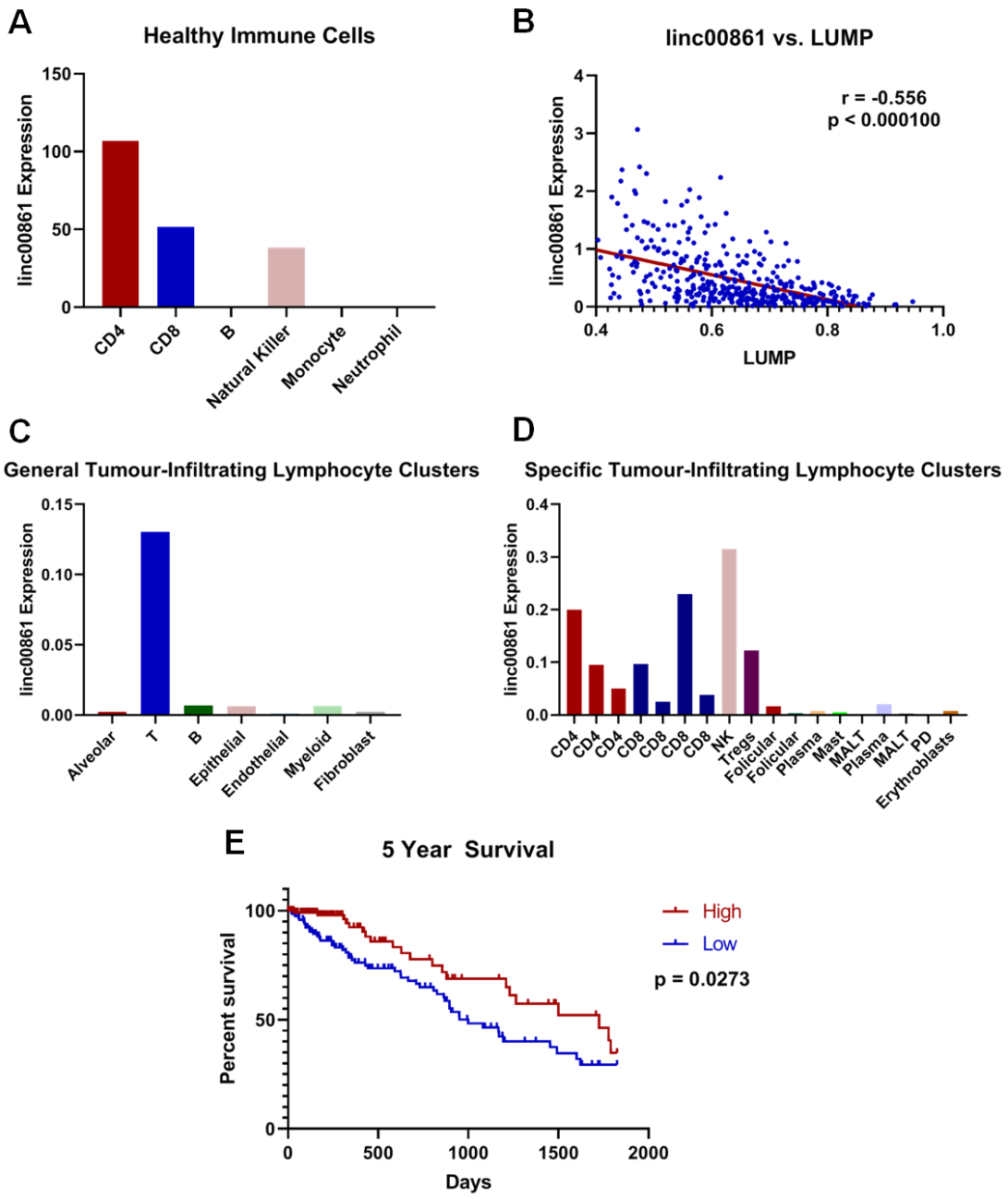

Figure 4.11. linc00861 expression in healthy immune cells and association with tumour-infiltrating lymphocytes.

A) *linc00861* expression in healthy human immune cell subsets. B) Spearman's correlation of *linc00861* expression and tumour purity as estimated by LUMP scores, where higher LUMP scores indicate higher tumour-cell content. C) *linc00861* expression in clusters of stromal cells isolated from bulk lung tumour samples. D) *linc00861* expression in specific sub-clusters of T (n=9) and B (n=9) cells isolated from bulk lung tumour samples. E) Five-year survival of lung cancer patients stratified into tertiles of high (red) versus low (blue) expression of *linc00861*. The log-rank test was used to calculate the statistical significance of this association.

4.4 Discussion

Contrary to miRNAs, of which specific gene targets can be predicted by sequence analysis, elucidating lncRNA mechanisms of action presents a unique challenge in light of their complex secondary structure, targeting mechanisms, and wide-reaching interaction partners. Large-scale analyses of lncRNA expression in specific contexts are necessary to provide a foundation from which functional studies can be built. Here, I describe an atlas of lncRNAs expressed in human immune cells, which I use to assess their expression and potential roles in the lung tumour microenvironment. My approach is based on the consideration of: (i) the emerging roles for lncRNAs in homeostatic and disease biology; (ii) the recognized importance of infiltrating lymphocytes to tumour development; and (iii) the ability of lncRNAs to be detected in NGS data.

I uncovered the widespread expression of lncRNAs in human immune cells, with transcripts expressed at incredibly varying levels. Approximately one quarter of the lncRNAs expressed in this cohort were expressed in all cell types (Figure 4.3). Conversely, I observed numerous lncRNA transcripts to be expressed exclusively in one cell type, which aligns with the observation that lncRNAs typically display tissue- and context-specific expression. In my assessment, I found approximately 15% of the lncRNAs expressed in the immune cell subsets analyzed to be expressed in only one cell type, whereas this proportion was only 3% for mRNAs. Further, unsupervised hierarchical clustering analysis showed that lncRNA expression was capable of recapitulating known patterns of immune cell differentiation (Figure 4.4). From this, it can be suggested that lncRNA expression may be of value in supplementing current deconvolution panels, such as CIBERSORT. As such, future studies may seek to examine the ability of lncRNA expression in tandem with established markers to better identify proportions of cells in bulk samples, such as tumours.

Broadly, the expression pattern and genomic location of a given lncRNA may be a means whereby its potential functional relevance can be inferred, particularly for transcripts that act *in cis*. This idea is supported by the observation of lncRNAs with previously-characterized functions in immune cells, such as high *MEG3* expression in monocytes aligning with its observed function in the regulation of IL-1β release (Figure 4.5)[176]. As a prominent example, I observed transcript variants of *AC008750* to be expressed only in cytotoxic cells ($CD8^+$-T and NK) and transcribed from a genomic locus that is neighbouring *NKG7*. *NKG7* not only displays a congruent expression pattern but is also an important gene in the granulation response and cytotoxic effector function, suggesting a potential regulatory relationship between this gene and its neighbouring lncRNA. Further, the gene overlapping *AC008750*, *SIGLEC10*, while also involved in the immune response displays a clearly distinct expression pattern, making the transcription of this lncRNA unlikely to be the result of a passenger effect. While biochemical assays are required to elucidate whether there is indeed a regulatory relationship between *AC008750* and *NKG7*, such as a CRISPR-induced knockout or inactivation of *AC008750*, my results highlight the utility of broad panelling of lncRNA transcription to identify potential functionally-relevant candidates.

Generating an atlas of lncRNA expression in healthy human immune cells allowed me to ask whether these expression patterns could be detected in bulk tumour data, potentially indicating the presence of infiltrating lymphocytes. To explore this, I first examined RNA sequencing data from LUAD tumours from TCGA, where I found many immune-related lncRNAs to be differentially expressed between tumours with matched non-malignant samples. Given their detection in immune cells, I expected differences between tumours with high versus low potential for infiltrating stromal cells. To this end, examining the dysregulation of these

same transcripts in our microdissected BCCA-LUAD cohort revealed the marked under-expression of immune-related lncRNAs in these tumours. Additionally, I found a number of these lncRNAs to significantly correlate with three separate markers of immune cell infiltration: CD45 gene expression, purity (LUMP) scores, and tumour associated antigen expression. Finally, to confirm the expression of immune-related lncRNAs from immune cells in the lung tumour microenvironment, I assessed single cell RNA sequencing data from lung tumour samples. Here, I showed that immune-related lncRNAs were not only expressed in similar patterns in these cell populations as initially observed in purified healthy cells, but also that this expression was specific to their respective cell type.

Together, these results suggest the contribution of infiltrating immune cells to gene expression data from bulk samples. Thus, my repertoire of immune-related lncRNAs may be used to identify the presence of specific immune cell populations within tumour samples from NGS data, as well as provide novel explanations for observed deregulated expression patterns. Although these results cast doubt on the widely-held idea that differential expression may indicate a functional role in tumour development and warrant further analysis, they may be used to identify potential roles in the maintenance of tumour-associated phenotypes, such as immune-infiltration and inflammation. My analysis of immune cells can also be extended to the landscape of RNA expression and gene regulation in the broader tumour microenvironment. The importance of lncRNA-based analysis shown here provides both the rationale and the methodology to more widely explore these types of transcripts in the larger context of bulk tumour data. In fact, the expression of immune-related lncRNAs could be assessed in immunogenically “hot” tumours versus those that are immunogenically “cold”, such as colorectal cancer with and without microsatellite instability[187]. Additionally, to identify potential

tumour-associated roles for these lncRNAs, their expression and functions could be assessed in immunological malignancies, such as acute myeloid leukemia. Similarly, their functions in cases of hyperactive immune systems, such as amyotrophic lateral sclerosis could be surveyed.

Finally, in light of the increasing relevance of the tumour microenvironment to lung cancer development and treatment, as well as the functional importance and specific expression of lncRNAs make them attractive transcripts for clinical intervention. Although studies to confirm their specific expression from immune cells in bulk tumours are required, such as fluorescence *in situ* hybridization, the lncRNAs described here could have translational utility as prognostic markers of immune cell infiltration or response to immunotherapy. I observed a subset of these immune-related lncRNAs to be associated with patient outcome in LUAD, which aligns with the poor outcome for patients lacking the presence of specific lymphocytes. Further, the specific expression of lncRNAs may provide the advantage of limited off-target effects for novel therapeutic agents. Additionally, the fact that the RNA is the functional unit rather than an intermediate may make them amenable to RNA-based inhibitors such as antisense oligonucleotides. Collectively, the widespread consideration of lncRNAs reveals broad features of this class of transcripts, which can be used to further elucidate their functions as novel mediators of gene expression and disease.

Chapter 5: Conclusions

5.1.1 Summary and significance

Non-coding RNAs, while initially discarded for not encoding proteins, have since become recognized as transcripts with specific expression patterns and important functions in gene regulation. Here I take a large-scale, high-throughput approach to study the non-coding transcriptome of lung adenocarcinoma cases, a disease with a markedly poor prognosis and a persisting lack of well-established molecular drivers. I sought to address some of the broad challenges in ncRNA research to identify novel genes in lung cancer biology. First, I addressed the difficulty in accurately describing the landscape of sncRNA expression in human tissues, especially those with specific expression patterns that have largely been underestimated. From this, I shifted focus to the expression of long non-coding RNAs which, while specific examples are well-documented, a lack of broad characterization imposes barriers to uncovering their roles in human tissue and disease contexts.

In Chapter 3, I re-evaluated whole-genome small RNA sequencing data from lung tumours in both the TCGA-LUAD and BCCA-LUAD cohorts. Through this, I not only described a substantial proportion of previously-unannotated miRNAs expressed between the individual cohorts, but also observed their remarkable specificity to lung samples. These results serve to significantly expand the miRNA transcriptome of human lung tissues, which can be of value in finding new candidates for important players in lung biology and disease. In fact, I found several of these miRNAs to display significantly deregulated expression patterns between non-malignant and tumour samples, many of which were specifically expressed in only one context. These features combined with their observed associations with survival emphasize the biological

importance and translational utility of deep sequencing initiatives. To further explore the extent of previously-unannotated miRNA expression in human samples, I applied a similar pipeline to the analysis of kidney and clear cell renal cell carcinoma samples. Here, I observed a congruent phenomenon of a significant number of novel miRNAs expressed in and deregulated between normal and malignant human kidney samples, as well as in widely-used renal cell lines. Furthermore, I validated the expression of novel miRNA candidates *in vitro*, wherein they showed detectable expression as well as similar patterns of dysregulation between malignant and non-malignant cell lines. The discovery of the widespread transcription of previously-unannotated miRNAs highlights the information that is missed by relying on abundance and conservation. Similarly, I illustrate the extensive data that can be obtained from previous RNA-sequencing initiatives and the ability of *in silico* analyses to identify robust novel gene candidates for further investigation. Collectively, my results serve to address part of the difficulty surrounding sncRNA identification in human samples, and provide a resource for the deeper characterization of their specific functional roles.

Conversely, while sncRNAs have relatively defined mechanisms of gene regulation, lncRNA-based research remains focused on specific examples. In Chapter 4, I sought to highlight the application of a broad approach to studying lncRNA expression from the perspective of a key feature of lung cancer, the tumour microenvironment. Given that lung cancer is one of the most immunogenic malignancies and that the presence of specific infiltrating lymphocyte populations has prognostic value, I sought to identify whether there were lncRNAs relevant to this phenomenon. Rather than narrowing my focus to specific transcripts, I first panelled lncRNA expression in purified healthy human immune cell populations, wherein I observed the widespread and specific expression of close to 5000 lncRNAs. Their striking degree

of cell-type specificity – to a greater extent than protein-coding genes – and their ability to recapitulate immune differentiation pathways is suggestive of their potential applications to immune-cell deconvolution of bulk sequencing studies. With this atlas of immune-related lncRNAs, I explored their expression in bulk-tumour data. I first observed significant differential expression between tumour and normal samples, which is suggestive of a potential relevance of lncRNAs to tumour biology. However, the same analyses performed in the BCCA-LUAD cohort, which is microdissected to 80% tumour-cell content, revealed the widespread under-expression of lncRNAs in tumours. This suggested that infiltrating immune cells in bulk tumour samples may confound RNA-sequencing data. To further explore this, I found several immune-related lncRNAs that were strongly associated with immune cell markers, namely purity estimates (LUMP scores), correlation with CD45 expression, and tumour-associated antigen exposure. Finally, I aimed to assess this specifically in sequencing data from lung tumours at the single cell resolution. Here, I observed many of the lncRNAs identified in healthy human immune cells to co-express with cell-type markers from stromal cells found within tumours. Further, the specific expression patterns observed for a number of lncRNAs in purified cell subsets was maintained in the scRNAseq data. Collectively, these results highlight the expression of a subset of lncRNAs from immune cells in the tumour microenvironment. Firstly, this emphasizes the necessary consideration of tumour-cell content when examining bulk sequencing data, in order to avoid conclusions and follow-up experiments on the basis of results confounded by infiltrating lymphocytes. Further, my approach highlights the utility of genome-wide approaches to studying lncRNA-related function. While specific functional lncRNAs in the tumour microenvironment require further study, I provide the groundwork to elucidate genomic

links between malignant and immune cells, which in the future may uncover truly novel therapeutic intervention points.

5.1.2 Limitations and future directions

Although my results underscore the value of deep sequencing efforts and *in silico* analyses, they will need to be confirmed via *in vitro* and *in vivo* experiments. Beyond validating the expression patterns described here, these experiments may seek to address the dysregulation of the various functional ncRNA candidates. The most pertinent experiments would involve targeted knockdown or knock-out of the ncRNA candidate of interest, using a system such as CRISPR. A CRISPR-mediated interference (CRISPRi) design may be particularly effective for the study of the phenotypic consequence of lncRNA dysregulation. CRISPRi captures the potential action of a lncRNA *in cis* or *in trans*, as well as the potential effects of the act of lncRNA transcription or transcriptional enhancement[188]. Further, this could be coupled with CRISPR activation (CRISPRa) experiments, to assess for functional/phenotypic rescue with the re-expression of a deleted lncRNA[189]. Alternatively, miRNA analysis may be more realistically performed by targeted oligonucleotides to block or degrade over-expressed functional miRNAs, such as short-interfering RNAs (siRNAs).

The prediction of miRNA gene targets through algorithms based on free energy and sequence homology such as miRanda are subject to a relatively high degree of false-positive results[190]. Thus, the predicted targets and subsequent pathways for the newly-discovered miRNA transcripts described in Chapter 3, while encouraging, should be taken with caution. For lncRNAs, target prediction is almost wholly unreliable due to unknown binding motifs and dynamic secondary structures. To combat this, gene-editing experiments coupled with gene

expression analysis could be used to assess miRNA targets, as well as interaction assays such as RNA immunoprecipitation or luciferase-based assessments. These experiments could also be used to examine the functional outcomes of mutations or polymorphisms in ncRNAs, an understudied but widespread phenomenon[191]. Similarly, my results describe the dysregulation of many ncRNAs in the context of lung tumours, but do not address the mechanisms of their deregulation. Future studies may look to examine whether the most frequently deregulated ncRNAs are found within commonly amplified or deleted genomic regions, display aberrant methylation, or even the product of deregulated miRNA targeting.

I have shown the usefulness of examining broad expression levels of ncRNAs, to provide a glimpse of their potential functions in cellular biology. Analyzing ncRNAs from this perspective allows the elucidation of alternative regulatory mechanisms, but also focuses on the functional unit of the genes. Conversely, protein-coding mRNAs are an intermediate step to functional products, meaning that mRNA expression alterations do not necessarily indicate effects at the protein-level. In light of this, I did not assess expression correlations between ncRNAs and potential target genes; however, these analyses could provide rationale for the further confirmation of the protein-level consequences of ncRNA:mRNA regulatory relationships.

I also suggest the broad translational utility of many of the aforementioned ncRNA transcripts, both in theory and through the development of prognostic signatures in both Chapters 3 and 4. Although I performed these analyses on a relatively extensive dataset, the validation of these signatures in individual cohorts and markedly higher sample sizes is required before these results could be used to direct clinical intervention. Additionally, while lung cancer in smokers is known to display different molecular signatures, I did not focus my analyses on tumours arising in individuals that have never smoked. Smoking status could be a contributing factor to the

results described here, such as in the differential expression of lncRNAs between tumours with high versus low tumour-associated antigen expression. Thus, future studies could assess ncRNA expression and specificity while controlling for additional factors such as smoking, histology, or stage. With this information in tandem with the confirmation of their functional impact, certain ncRNAs may prove to be valuable therapeutic intervention points, particularly those that are cell-type specific. Therefore, the design of anti-miRs, miRNA mimics, as well as the potential of ASO-based lncRNA inactivation (e.g. targeting *MALAT1*), may be seen more frequently in future clinical trials.

Finally, while my results are performed under the lens of lung cancer, the platforms described here can be applied to all cancer types, and other disease contexts, as evidenced by the findings of novel miRNAs expressed in ccRCC in Chapter 3. In a similar fashion, lncRNAs with immune-specific expression and functions could be panelled in other malignancies that can evade the immune system, or on the other end of the spectrum, in autoimmune diseases that represent aberrant hyperactivity of specific lymphocytes. Regardless of the context in which they are assessed, I show that the high-level analysis of ncRNA expression can generate extensive hypotheses and is necessary for the downstream identification of their functions.

Overall, I show the plethora of functions and genomic information that is encoded within non-coding RNAs, despite their absence of protein-coding propensity. I identified novel patterns of non-coding RNA expression and used these to broaden our understanding of the lung tumour genome. While further experiments to confirm and extend these results are required, my results set the stage for the identification of alternative mechanics of gene regulation, impacting cellular phenotypes and cancer development.

References

1 Wong, M. C. S., Lao, X. Q., Ho, K. F., Goggins, W. B. & Tse, S. L. A. (2017). Incidence and mortality of lung cancer: global trends and association with socioeconomic status. *Scientific reports* **7**, 14300

2 Canadian Cancer Statistics Advisory Committee. *Canadian Cancer Statistics 2018*,Toronto ON: Canadian Cancer Society. Available at: <http://www.cancer.ca/en/cancer-information/cancer-101/canadian-cancer-statistics-publication/?region=pe> (2018).

3 Pallis, A. G. & Syrigos, K. N. (2013). Lung cancer in never smokers: disease characteristics and risk factors. *Critical reviews in oncology/hematology* **88**, 494-503

4 Hubaux, R., Becker-Santos, D. D., Enfield, K. S., Lam, S., Lam, W. L. & Martinez, V. D. (2012). Arsenic, asbestos and radon: emerging players in lung tumorigenesis. *Environmental health : a global access science source* **11**, 89

5 Martinez, V. D., Sage, A.P., Marshall, E.A., Suzuki, M., Goodarzi, A.A., Dellaire, G., & Lam, W.L. *Oncogenetics of Lung Cancer Induced by Environmental Carcinogens*, in *Oncogenes and Carcinogenesis* (IntechOpen, 2018).

6 Sage, A. P., Minatel, B. C., Ng, K. W., Stewart, G. L., Dummer, T. J. B., Lam, W. L. & Martinez, V. D. (2017). Oncogenomic disruptions in arsenic-induced carcinogenesis. *Oncotarget* **8**, 25736-25755

7 Torres-Duran, M., Barros-Dios, J. M., Fernandez-Villar, A. & Ruano-Ravina, A. (2014). Residential radon and lung cancer in never smokers. A systematic review. *Cancer letters* **345**, 21-26

8 Global Burden of Disease Cancer Collaboration, Fitzmaurice, C., Dicker, D., Pain, A., Hamavid, H., Moradi-Lakeh, M., MacIntyre, M. F., Allen, C., Hansen, G., Woodbrook, R., Wolfe, C., Hamadeh, R. R., Moore, A., Werdecker, A., Gessner, B. D., Te Ao, B., McMahon, B., Karimkhani, C., Yu, C., Cooke, G. S., Schwebel, D. C., Carpenter, D. O., Pereira, D. M., Nash, D., Kazi, D. S., De Leo, D., Plass, D., Ukwaja, K. N., Thurston, G. D., Yun Jin, K., Simard, E. P., Mills, E., Park, E. K., Catala-Lopez, F., deVeber, G., Gotay, C., Khan, G., Hosgood, H. D., Santos, I. S., Leasher, J. L., Singh, J., Leigh, J., Jonas, J. B., Sanabria, J., Beardsley, J., Jacobsen, K. H., Takahashi, K., Franklin, R. C., Ronfani, L., Montico, M., Naldi, L., Tonelli, M., Geleijnse, J., Petzold, M., Shrime, M. G., Younis, M., Yonemoto, N., Breitborde, N., Yip, P., Pourmalek, F., Lotufo, P. A., Esteghamati, A., Hankey, G. J., Ali, R., Lunevicius, R., Malekzadeh, R., Dellavalle, R., Weintraub, R., Lucas, R., Hay, R., Rojas-Rueda, D., Westerman, R., Sepanlou, S. G., Nolte, S., Patten, S., Weichenthal, S., Abera, S. F., Fereshtehnejad, S. M., Shiue, I., Driscoll, T., Vasankari, T., Alsharif, U., Rahimi-Movaghar, V., Vlassov, V. V., Marcenes, W. S., Mekonnen, W., Melaku, Y. A., Yano, Y., Artaman, A., Campos, I., MacLachlan, J., Mueller, U., Kim, D., Trillini, M., Eshrati, B., Williams, H. C., Shibuya, K., Dandona, R., Murthy, K., Cowie, B., Amare, A. T., Antonio, C. A., Castaneda-Orjuela, C., van Gool, C. H., Violante, F., Oh, I. H., Deribe, K., Soreide, K., Knibbs, L., Kereselidze, M., Green, M., Cardenas, R., Roy, N., Tillmann, T., Li, Y., Krueger, H., Monasta, L., Dey, S., Sheikhbahaei, S., Hafezi-Nejad, N., Kumar, G. A., Sreeramareddy, C. T., Dandona, L., Wang, H., Vollset, S. E., Mokdad, A., Salomon, J. A., Lozano, R., Vos, T., Forouzanfar, M., Lopez, A., Murray, C. & Naghavi, M. (2015). The Global Burden of Cancer 2013. *JAMA oncology* **1**, 505-527

9 Minatel, B. C., Sage, A. P., Anderson, C., Hubaux, R., Marshall, E. A., Lam, W. L. & Martinez, V. D. (2018). Environmental arsenic exposure: From genetic susceptibility to pathogenesis. *Environment international* **112**, 183-197

10 Kanwal, M., Ding, X. J. & Cao, Y. (2017). Familial risk for lung cancer. *Oncology letters* **13**, 535-542

11 Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, L. & Rosell, R. (2015). Non-small-cell lung cancer. *Nature Reviews Disease Primers* **1**, 15009

12 Inamura, K. (2017). Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. *Frontiers in oncology* **7**, 193

13 Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., Chirieac, L. R., Dacic, S., Duhig, E., Flieder, D. B., Geisinger, K., Hirsch, F. R., Ishikawa, Y., Kerr, K. M., Noguchi, M., Pelosi, G., Powell, C. A., Tsao, M. S., Wistuba, I. & WHO Panel. (2015). The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **10**, 1243-1260

14 Hanna, J. M. & Onaitis, M. W. (2013). Cell of origin of lung cancer. *Journal of carcinogenesis* **12**, 6

15 Sun, S., Schiller, J. H. & Gazdar, A. F. (2007). Lung cancer in never smokers-a different disease. *Nature reviews. Cancer* **7**, 778-790
16 Herbst, R. S., Morgensztern, D. & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* **553**, 446-454
17 Meza, R., Meernik, C., Jeon, J. & Cote, M. L. (2015). Lung cancer incidence trends by gender, race and histology in the United States, 1973-2010. *PloS one* **10**, e0121323
18 Sholl, L. M. (2016). The Molecular Pathology of Lung Cancer. *Surgical pathology clinics* **9**, 353-378
19 Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. (2013). Cancer genome landscapes. *Science* **339**, 1546-1558
20 Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. (2008). Cancer-related inflammation. *Nature* **454**, 436-444
21 Hanahan, D. & Coussens, L. M. (2012). Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer cell* **21**, 309-322
22 Li, X., Song, W., Shao, C., Shi, Y. & Han, W. (2019). Emerging predictors of the response to the blockade of immune checkpoints in cancer therapy. *Cellular & molecular immunology* **16**, 28-39
23 Mittal, V., El Rayes, T., Narula, N., McGraw, T. E., Altorki, N. K. & Barcellos-Hoff, M. H. (2016). The Microenvironment of Lung Cancer and Therapeutic Implications. *Advances in experimental medicine and biology* **890**, 75-110
24 Albini, A. & Sporn, M. B. (2007). The tumour microenvironment as a target for chemoprevention. *Nature reviews. Cancer* **7**, 139-147
25 Gajewski, T. F., Schreiber, H. & Fu, Y. X. (2013). Innate and adaptive immune cells in the tumor microenvironment. *Nature immunology* **14**, 1014-1022
26 Nishino, M., Ramaiya, N. H., Hatabu, H. & Hodi, F. S. (2017). Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nature reviews. Clinical oncology* **14**, 655-668
27 Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550
28 Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A. & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *The New England journal of medicine* **375**, 1109-1112
29 Oda, K., Matsuoka, Y., Funahashi, A. & Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology* **1**, 2005 0010
30 Rudin, C. M., Avila-Tang, E., Harris, C. C., Herman, J. G., Hirsch, F. R., Pao, W., Schwartz, A. G., Vahakangas, K. H. & Samet, J. M. (2009). Lung cancer in never smokers: molecular profiles and therapeutic implications. *Clinical cancer research : an official journal of the American Association for Cancer Research* **15**, 5646-5661
31 Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Jr., Wu, Y. L. & Paz-Ares, L. (2017). Lung cancer: current therapies and new targeted treatments. *Lancet* **389**, 299-311
32 Morgensztern, D., Devarakonda, S. & Govindan, R. (2013). Genomic landscape of squamous cell carcinoma of the lung. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, 348-353
33 Chan, B. A. & Hughes, B. G. (2015). Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. *Translational lung cancer research* **4**, 36-54
34 Pikor, L. A., Ramnarine, V. R., Lam, S. & Lam, W. L. (2013). Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer* **82**, 179-189
35 Senan, S., Brade, A., Wang, L. H., Vansteenkiste, J., Dakhil, S., Biesma, B., Martinez Aguillo, M., Aerts, J., Govindan, R., Rubio-Viqueira, B., Lewanski, C., Gandara, D., Choy, H., Mok, T., Hossain, A., Iscoe, N., Treat, J., Koustenis, A., San Antonio, B., Chouaki, N. & Vokes, E. (2016). PROCLAIM: Randomized Phase III Trial of Pemetrexed-Cisplatin or Etoposide-Cisplatin Plus Thoracic Radiation Therapy Followed by Consolidation Chemotherapy in Locally Advanced Nonsquamous Non-Small-Cell Lung Cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **34**, 953-962
36 Zappa, C. & Mousa, S. A. (2016). Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research* **5**, 288-300
37 Mayekar, M. K. & Bivona, T. G. (2017). Current Landscape of Targeted Therapy in Lung Cancer. *Clinical pharmacology and therapeutics* **102**, 757-764

38 Bustamante Alvarez, J. G., Gonzalez-Cao, M., Karachaliou, N., Santarpia, M., Viteri, S., Teixido, C. & Rosell, R. (2015). Advances in immunotherapy for treatment of lung cancer. *Cancer biology & medicine* **12**, 209-222

39 Raju, S., Joseph, R. & Sehgal, S. (2018). Review of checkpoint immunotherapy for the management of non-small cell lung cancer. *ImmunoTargets and therapy* **7**, 63-75

40 International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945

41 Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467

42 Yohe, S. & Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. *Archives of pathology & laboratory medicine* **141**, 1544-1557

43 Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27-38

44 van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics : TIG* **30**, 418-426

45 Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46

46 Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology* **2012**, 251364

47 Le Gallo, M., Lozy, F. & Bell, D. W. (2017). Next-Generation Sequencing. *Advances in experimental medicine and biology* **943**, 119-148

48 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J. & International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921

49 Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337**, 1159, 1161

50 Cech, T. R. & Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94
51 Romano, G., Veneziano, D., Acunzo, M. & Croce, C. M. (2017). Small non-coding RNA and cancer. *Carcinogenesis* **38**, 485-491
52 Sun, W., Yang, Y., Xu, C. & Guo, J. (2017). Regulatory mechanisms of long noncoding RNAs on gene expression in cancers. *Cancer genetics* **216-217**, 105-110
53 He, L. & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics* **5**, 522-531
54 Bracken, C. P., Scott, H. S. & Goodall, G. J. (2016). A network-biology perspective of microRNA function and dysfunction in cancer. *Nature reviews. Genetics* **17**, 719-732
55 Calin, G. A. & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nature reviews. Cancer* **6**, 857-866
56 Li, Z. & Rana, T. M. (2014). Therapeutic targeting of microRNAs: current status and future challenges. *Nature reviews. Drug discovery* **13**, 622-638
57 Gurtner, A., Falcone, E., Garibaldi, F. & Piaggio, G. (2016). Dysregulation of microRNA biogenesis in cancer: the impact of mutant p53 on Drosha complex activity. *Journal of experimental & clinical cancer research : CR* **35**, 45
58 Macfarlane, L. A. & Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. *Current genomics* **11**, 537-561
59 Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. & Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415-419
60 Michlewski, G. & Caceres, J. F. (2019). Post-transcriptional control of miRNA biogenesis. *Rna* **25**, 1-16
61 Saliminejad, K., Khorram Khorshid, H. R., Soleymani Fard, S. & Ghaffari, S. H. (2019). An overview of microRNAs: Biology, functions, therapeutics, and analysis methods. *Journal of cellular physiology* **234**, 5451-5465
62 Kim, D., Sung, Y. M., Park, J., Kim, S., Kim, J., Park, J., Ha, H., Bae, J. Y., Kim, S. & Baek, D. (2016). General rules for functional microRNA targeting. *Nature genetics* **48**, 1517-1526
63 Hammond, S. M. (2015). An overview of microRNAs. *Advanced drug delivery reviews* **87**, 3-14
64 Svoronos, A. A., Engelman, D. M. & Slack, F. J. (2016). OncomiR or Tumor Suppressor? The Duplicity of MicroRNAs in Cancer. *Cancer research* **76**, 3666-3670
65 Enfield, K. S., Pikor, L. A., Martinez, V. D. & Lam, W. L. (2012). Mechanistic Roles of Noncoding RNAs in Lung Cancer Biology and Their Clinical Implications. *Genetics research international* **2012**, 737416
66 Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R. M., Okamoto, A., Yokota, J., Tanaka, T., Calin, G. A., Liu, C. G., Croce, C. M. & Harris, C. C. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell* **9**, 189-198
67 Castro, D., Moreira, M., Gouveia, A. M., Pozza, D. H. & De Mello, R. A. (2017). MicroRNAs in lung cancer. *Oncotarget* **8**, 81679-81685
68 Izumchenko, E., Chang, X., Michailidi, C., Kagohara, L., Ravi, R., Paz, K., Brait, M., Hoque, M. O., Ling, S., Bedi, A. & Sidransky, D. (2014). The TGFbeta-miR200-MIG6 pathway orchestrates the EMT-associated kinase switch that induces resistance to EGFR inhibitors. *Cancer research* **74**, 3995-4005
69 Romano, G., Acunzo, M., Garofalo, M., Di Leva, G., Cascione, L., Zanca, C., Bolon, B., Condorelli, G. & Croce, C. M. (2012). MiR-494 is regulated by ERK1/2 and modulates TRAIL-induced apoptosis in non-small-cell lung cancer through BIM down-regulation. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16570-16575
70 Halappanavar, S., Nikota, J., Wu, D., Williams, A., Yauk, C. L. & Stampfli, M. (2013). IL-1 receptor regulates microRNA-135b expression in a negative feedback mechanism during cigarette smoke-induced inflammation. *Journal of immunology* **190**, 3679-3686
71 Rupaimoole, R. & Slack, F. J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature reviews. Drug discovery* **16**, 203-222
72 Kozomara, A. & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**, D68-73
73 Meiri, E., Levy, A., Benjamin, H., Ben-David, M., Cohen, L., Dov, A., Dromi, N., Elyakim, E., Yerushalmi, N., Zion, O., Lithwick-Yanai, G. & Sitbon, E. (2010). Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic acids research* **38**, 6234-6246
74 Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A.

R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F., Group, R. G. E. R. & Genome Science, G. (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563

75 Morris, K. V. & Mattick, J. S. (2014). The rise of regulatory RNA. *Nature reviews. Genetics* **15**, 423-437

76 Quinn, J. J. & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature reviews. Genetics* **17**, 47-62

77 Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J. & Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22**, 1775-1789

78 Kung, J. T., Colognori, D. & Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics* **193**, 651-669

79 Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in genetics : TIG* **30**, 121-123

80 Mattick, J. S., Taft, R. J. & Faulkner, G. J. (2010). A global view of genomic information-moving beyond the gene and the master regulator. *Trends in genetics : TIG* **26**, 21-28

81 Marchese, F. P., Raimondi, I. & Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. *Genome biology* **18**, 206

82 Gutschner, T. & Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. *RNA biology* **9**, 703-719

83 Schmitt, A. M. & Chang, H. Y. (2016). Long Noncoding RNAs in Cancer Pathways. *Cancer cell* **29**, 452-463

84 Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W. E., Serve, H. & Muller-Tidow, C. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031-8041

85 Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S. & Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076

86 Wei, M. M. & Zhou, G. B. (2016). Long Non-coding RNAs and Their Roles in Non-small-cell Lung Cancer. *Genomics, proteomics & bioinformatics* **14**, 280-288

87 Zhao, W., An, Y., Liang, Y. & Xie, X. W. (2014). Role of HOTAIR long noncoding RNA in metastatic progression of lung cancer. *European review for medical and pharmacological sciences* **18**, 1930-1936
88 Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., Zornig, M., MacLeod, A. R., Spector, D. L. & Diederichs, S. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research* **73**, 1180-1189
89 Devarakonda, S., Morgensztern, D. & Govindan, R. (2015). Genomic alterations in lung adenocarcinoma. *The Lancet. Oncology* **16**, e342-351
90 Dalmay, T. (2010). Detection of small non-coding RNAs. *Methods in molecular biology* **655**, 265-274
91 Raasch, P., Schmitz, U., Patenge, N., Vera, J., Kreikemeyer, B. & Wolkenhauer, O. (2010). Non-coding RNA detection methods combined to improve usability, reproducibility and precision. *BMC bioinformatics* **11**, 491
92 Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G. & Tuschl, T. (2003). A uniform system for microRNA annotation. *Rna* **9**, 277-279
93 Friedlander, M. R., Lizano, E., Houben, A. J., Bezdan, D., Banez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., Gonzalez, J., Chen, K. C., LeProust, E. M., Marti, E. & Estivill, X. (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome biology* **15**, R57
94 Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., Hatzimichael, E., Kirino, Y., Honda, S., Lally, M., Ramratnam, B., Comstock, C. E., Knudsen, K. E., Gomella, L., Spaeth, G. L., Hark, L., Katz, L. J., Witkiewicz, A., Rostami, A., Jimenez, S. A., Hollingsworth, M. A., Yeh, J. J., Shaw, C. A., McKenzie, S. E., Bray, P., Nelson, P. T., Zupo, S., Van Roosbroeck, K., Keating, M. J., Calin, G. A., Yeo, C., Jimbo, M., Cozzitorto, J., Brody, J. R., Delgrosso, K., Mattick, J. S., Fortina, P. & Rigoutsos, I. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E1106-1115
95 Freedman, J. E., Miano, J. M., National Heart, L. & Blood Institute Workshop, P. (2017). Challenges and Opportunities in Linking Long Noncoding RNAs to Cardiovascular, Lung, and Blood Diseases. *Arteriosclerosis, thrombosis, and vascular biology* **37**, 21-25
96 Johnsson, P., Lipovich, L., Grander, D. & Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta* **1840**, 1063-1071
97 Roux, B. T., Heward, J. A., Donnelly, L. E., Jones, S. W. & Lindsay, M. A. (2017). Catalog of Differentially Expressed Long Non-Coding RNA following Activation of Human and Mouse Innate Immune Response. *Frontiers in immunology* **8**, 1038
98 Xue, Z., Hennelly, S., Doyle, B., Gulati, A. A., Novikova, I. V., Sanbonmatsu, K. Y. & Boyer, L. A. (2016). A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage. *Molecular cell* **64**, 37-50
99 Schulze, A. B. & Schmidt, L. H. (2017). PD-1 targeted Immunotherapy as first-line therapy for advanced non-small-cell lung cancer patients. *Journal of thoracic disease* **9**, E384-E386
100 Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic acids research* **39**, D19-21
101 Martinez, V. D., Vucic, E. A., Thu, K. L., Hubaux, R., Enfield, K. S., Pikor, L. A., Becker-Santos, D. D., Brown, C. J., Lam, S. & Lam, W. L. (2015). Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology. *Scientific reports* **5**, 10423
102 Partek Genomics Suite software v. 7.0.18.0724 (St. Louis, MO, USA, 2018).
103 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
104 Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515
105 Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D. N., Nag, R., Newman, V., Nuhn, M., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S. P., Zadissa, A., Kostadima, M., Martin, F. J.,

Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Cunningham, F., Yates, A., Zerbino, D. R. & Flicek, P. (2017). Ensembl 2017. *Nucleic acids research* **45**, D635-D642

106 An, J., Lai, J., Lehman, M. L. & Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic acids research* **41**, 727-737

107 Fehlmann, T., Backes, C., Kahraman, M., Haas, J., Ludwig, N., Posch, A. E., Wurstle, M. L., Hubenthal, M., Franke, A., Meder, B., Meese, E. & Keller, A. (2017). Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic acids research* **45**, 8731-8744

108 O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-745

109 Capece, V., Garcia Vizcaino, J. C., Vidal, R., Rahman, R. U., Pena Centeno, T., Shomroni, O., Suberviola, I., Fischer, A. & Bonn, S. (2015). Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics* **31**, 2205-2207

110 Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R. & Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic acids research* **37**, D136-140

111 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421

112 Liao, Y., Smyth, G. K. & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930

113 Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300

114 Bland, J. M. & Altman, D. G. (1998). Survival probabilities (the Kaplan-Meier method). *Bmj* **317**, 1572

115 Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome biology* **5**, R1

116 Rahmati, S., Abovsky, M., Pastrello, C. & Jurisica, I. (2017). pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic acids research* **45**, D419-D426

117 Di Leva, G., Garofalo, M. & Croce, C. M. (2014). MicroRNAs in cancer. *Annual review of pathology* **9**, 287-314

118 Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. (2005). Principles of microRNA-target recognition. *PLoS biology* **3**, e85

119 Ha, M. & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology* **15**, 509-524

120 Karube, Y., Tanaka, H., Osada, H., Tomida, S., Tatematsu, Y., Yanagisawa, K., Yatabe, Y., Takamizawa, J., Miyoshi, S., Mitsudomi, T. & Takahashi, T. (2005). Reduced expression of Dicer associated with poor prognosis in lung cancer patients. *Cancer science* **96**, 111-115

121 Su, X., Chakravarti, D., Cho, M. S., Liu, L., Gi, Y. J., Lin, Y. L., Leung, M. L., El-Naggar, A., Creighton, C. J., Suraokar, M. B., Wistuba, I. & Flores, E. R. (2010). TAp63 suppresses metastasis through coordinate regulation of Dicer and miRNAs. *Nature* **467**, 986-990

122 Wu, X., Piper-Hunter, M. G., Crawford, M., Nuovo, G. J., Marsh, C. B., Otterson, G. A. & Nana-Sinkam, S. P. (2009). MicroRNAs in the pathogenesis of Lung Cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **4**, 1028-1034

123 Thu, K. L., Chari, R., Lockwood, W. W., Lam, S. & Lam, W. L. (2011). miR-101 DNA copy loss is a prominent subtype specific event in lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **6**, 1594-1598

124 Zhang, J. G., Guo, J. F., Liu, D. L., Liu, Q. & Wang, J. J. (2011). MicroRNA-101 exerts tumor-suppressive functions in non-small cell lung cancer through directly targeting enhancer of zeste homolog 2. *Journal of*

thoracic oncology : official publication of the International Association for the Study of Lung Cancer **6**, 671-678
125 Uddin, A. & Chakraborty, S. (2018). Role of miRNAs in lung cancer. *Journal of cellular physiology*
126 Florczuk, M., Szpechcinski, A. & Chorostowska-Wynimko, J. (2017). miRNAs as Biomarkers and Therapeutic Targets in Non-Small Cell Lung Cancer: Current Perspectives. *Targeted oncology* **12**, 179-200
127 Brown, D., Rahman, M. & Nana-Sinkam, S. P. (2014). MicroRNAs in respiratory disease. A clinician's overview. *Annals of the American Thoracic Society* **11**, 1277-1285
128 Inamura, K. & Ishikawa, Y. (2016). MicroRNA In Lung Cancer: Novel Biomarkers and Potential Tools for Treatment. *Journal of clinical medicine* **5**
129 Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W. & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **40**, 37-52
130 Rahman, R. U., Gautam, A., Bethune, J., Sattar, A., Fiosins, M., Magruder, D. S., Capece, V., Shomroni, O. & Bonn, S. (2018). Oasis 2: improved online analysis of small RNA-seq data. *BMC bioinformatics* **19**, 54
131 Backes, C., Meder, B., Hart, M., Ludwig, N., Leidinger, P., Vogel, B., Galata, V., Roth, P., Menegatti, J., Grasser, F., Ruprecht, K., Kahraman, M., Grossmann, T., Haas, J., Meese, E. & Keller, A. (2016). Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic acids research* **44**, e53
132 Masters, J. R. (2000). Human cancer cell lines: fact and fantasy. *Nature reviews. Molecular cell biology* **1**, 233-236
133 Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer* **6**, 813-823
134 Marshall, E. A., Sage, A. P., Ng, K. W., Martinez, V. D., Firmino, N. S., Bennewith, K. L. & Lam, W. L. (2017). Small non-coding RNA transcriptome of the NCI-60 cell line panel. *Scientific data* **4**, 170157
135 Minatel, B. C., Martinez, V. D., Ng, K. W., Sage, A. P., Tokar, T., Marshall, E. A., Anderson, C., Enfield, K. S. S., Stewart, G. L., Reis, P. P., Jurisica, I. & Lam, W. L. (2018). Large-scale discovery of previously undetected microRNAs specific to human liver. *Human genomics* **12**, 16
136 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190
137 Yap, S.-J., Cheong, W.-H., Tan, Y.-C. & Ng, K.-P. (2015). ClicO FS: an interactive web-based service of Circos. *Bioinformatics* **31**, 3685-3687
138 Kehl, T., Backes, C., Kern, F., Fehlmann, T., Ludwig, N., Meese, E., Lenhof, H. P. & Keller, A. (2017). About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget* **8**, 107167-107175
139 Mall, C., Rocke, D. M., Durbin-Johnson, B. & Weiss, R. H. (2013). Stability of miRNA in human urine supports its biomarker potential. *Biomarkers in medicine* **7**, 623-631
140 Max, K. E. A., Bertram, K., Akat, K. M., Bogardus, K. A., Li, J., Morozov, P., Ben-Dov, I. Z., Li, X., Weiss, Z. R., Azizian, A., Sopeyin, A., Diacovo, T. G., Adamidi, C., Williams, Z. & Tuschl, T. (2018). Human plasma and serum extracellular small RNA reference profiles and their clinical utility. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E5334-E5343
141 Hall, J. S., Taylor, J., Valentine, H. R., Irlam, J. J., Eustace, A., Hoskin, P. J., Miller, C. J. & West, C. M. (2012). Enhanced stability of microRNA expression facilitates classification of FFPE tumour samples exhibiting near total mRNA degradation. *British journal of cancer* **107**, 684-694
142 Palmero, E. I., de Campos, S. G., Campos, M., de Souza, N. C., Guerreiro, I. D., Carvalho, A. L. & Marques, M. M. (2011). Mechanisms and role of microRNA deregulation in cancer onset and progression. *Genetics and molecular biology* **34**, 363-370
143 Irmak-Yazicioglu, M. B. (2016). Mechanisms of MicroRNA Deregulation and MicroRNA Targets in Gastric Cancer. *Oncology research and treatment* **39**, 136-139
144 Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., Sommer, S. S. & Rossi, J. J. (2009). SNPs in human miRNA genes affect biogenesis and function. *Rna* **15**, 1640-1651
145 Christopher, A. F., Kaur, R. P., Kaur, G., Kaur, A., Gupta, V. & Bansal, P. (2016). MicroRNA therapeutics: Discovering novel targets and developing specific therapy. *Perspectives in clinical research* **7**, 68-74
146 Multhoff, G., Molls, M. & Radons, J. (2011). Chronic inflammation in cancer development. *Frontiers in immunology* **2**, 98
147 Muz, B., de la Puente, P., Azab, F. & Azab, A. K. (2015). The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* **3**, 83-92

148 Landskron, G., De la Fuente, M., Thuwajit, P., Thuwajit, C. & Hermoso, M. A. (2014). Chronic inflammation and cytokines in the tumor microenvironment. *Journal of immunology research* **2014**, 149185

149 Postow, M. A., Callahan, M. K. & Wolchok, J. D. (2015). Immune Checkpoint Blockade in Cancer Therapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **33**, 1974-1982

150 Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., Miller, M. L., Rekhtman, N., Moreira, A. L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E. B., Merghoub, T., Wolchok, J. D., Schumacher, T. N. & Chan, T. A. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124-128

151 Chen, L. & Han, X. (2015). Anti-PD-1/PD-L1 therapy of human cancer: past, present, and future. *The Journal of clinical investigation* **125**, 3384-3391

152 Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E., Barlesi, F., Kohlhaufl, M., Arrieta, O., Burgio, M. A., Fayette, J., Lena, H., Poddubskaya, E., Gerber, D. E., Gettinger, S. N., Rudin, C. M., Rizvi, N., Crino, L., Blumenschein, G. R., Jr., Antonia, S. J., Dorange, C., Harbison, C. T., Graf Finckenstein, F. & Brahmer, J. R. (2015). Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *The New England journal of medicine* **373**, 1627-1639

153 Reck, M. & Rabe, K. F. (2017). Precision Diagnosis and Treatment for Advanced Non-Small-Cell Lung Cancer. *The New England journal of medicine* **377**, 849-861

154 Goodman, A. M., Kato, S., Bazhenova, L., Patel, S. P., Frampton, G. M., Miller, V., Stephens, P. J., Daniels, G. A. & Kurzrock, R. (2017). Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Molecular cancer therapeutics* **16**, 2598-2608

155 Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., Bajorin, D. F., Al-Ahmadie, H. A., Chaft, J. E., Rudin, C. M., Riely, G. J., Baxi, S., Ho, A. L., Wong, R. J., Pfister, D. G., Wolchok, J. D., Barker, C. A., Gutin, P. H., Brennan, C. W., Tabar, V., Mellinghoff, I. K., DeAngelis, L. M., Ariyan, C. E., Lee, N., Tap, W. D., Gounder, M. M., D'Angelo, S. P., Saltz, L., Stadler, Z. K., Scher, H. I., Baselga, J., Razavi, P., Klebanoff, C. A., Yaeger, R., Segal, N. H., Ku, G. Y., DeMatteo, R. P., Ladanyi, M., Rizvi, N. A., Berger, M. F., Riaz, N., Solit, D. B., Chan, T. A. & Morris, L. G. T. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics*

156 Shimizu, K., Okita, R. & Nakata, M. (2013). Clinical significance of the tumor microenvironment in non-small cell lung cancer. *Annals of translational medicine* **1**, 20

157 Al-Shibli, K. I., Donnem, T., Al-Saad, S., Persson, M., Bremnes, R. M. & Busund, L. T. (2008). Prognostic effect of epithelial and stromal lymphocyte infiltration in non-small cell lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **14**, 5220-5227

158 Dai, F., Liu, L., Che, G., Yu, N., Pu, Q., Zhang, S., Ma, J., Ma, L. & You, Z. (2010). The number and microlocalization of tumor-associated immune cells are associated with patient's survival time in non-small cell lung cancer. *BMC cancer* **10**, 220

159 Barnes, T. A. & Amir, E. (2017). HYPE or HOPE: the prognostic value of infiltrating immune cells in cancer. *British journal of cancer* **117**, 451-460

160 Stoll, G., Bindea, G., Mlecnik, B., Galon, J., Zitvogel, L. & Kroemer, G. (2015). Meta-analysis of organ-specific differences in the structure of the immune infiltrate in major malignancies. *Oncotarget* **6**, 11894-11909

161 Ji, R. R., Chasalow, S. D., Wang, L., Hamid, O., Schmidt, H., Cogswell, J., Alaparthy, S., Berman, D., Jure-Kunkel, M., Siemers, N. O., Jackson, J. R. & Shahabi, V. (2012). An immune-active tumor microenvironment favors clinical response to ipilimumab. *Cancer immunology, immunotherapy : CII* **61**, 1019-1031

162 Gibney, G. T., Weiner, L. M. & Atkins, M. B. (2016). Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *The Lancet. Oncology* **17**, e542-e551

163 Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457

164 Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., Li, B. & Liu, X. S. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer research* **77**, e108-e110
165 Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., Coussens, L. M., Gabrilovich, D. I., Ostrand-Rosenberg, S., Hedrick, C. C., Vonderheide, R. H., Pittet, M. J., Jain, R. K., Zou, W., Howcroft, T. K., Woodhouse, E. C., Weinberg, R. A. & Krummel, M. F. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature medicine* **24**, 541-550
166 Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nature medicine* **21**, 1253-1261
167 Gibb, E. A., Brown, C. J. & Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Molecular cancer* **10**, 38
168 Batista, P. J. & Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298-1307
169 Chen, Y. G., Satpathy, A. T. & Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nature immunology* **18**, 962-972
170 Hrdlickova, B., Kumar, V., Kanduri, K., Zhernakova, D. V., Tripathi, S., Karjalainen, J., Lund, R. J., Li, Y., Ullah, U., Modderman, R., Abdulahad, W., Lahdesmaki, H., Franke, L., Lahesmaa, R., Wijmenga, C. & Withoff, S. (2014). Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. *Genome medicine* **6**, 88
171 Linsley, P. S., Speake, C., Whalen, E. & Chaussabel, D. (2014). Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PloS one* **9**, e109760
172 Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* **12**, R22
173 Ng, K. W., Marshall, E. A., Enfield, K. S., Martin, S. D., Milne, K., Pewarchuk, M. E., Abraham, N. & Lam, W. L. (2018). Somatic mutation-associated T follicular helper cell elevation in lung adenocarcinoma. *Oncoimmunology* **7**, e1504728
174 Aran, D., Sirota, M. & Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nature communications* **6**, 8971
175 Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwe, H., Pircher, A., Van den Eynde, K., Weynand, B., Verbeken, E., De Leyn, P., Liston, A., Vansteenkiste, J., Carmeliet, P., Aerts, S. & Thienpont, B. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature medicine* **24**, 1277-1289
176 Li, R., Fang, L., Pu, Q., Bu, H., Zhu, P., Chen, Z., Yu, M., Li, X., Weiland, T., Bansal, A., Ye, S. Q., Wei, Y., Jiang, J. & Wu, M. (2018). MEG3-4 is a miRNA decoy that regulates IL-1beta abundance to initiate and then limit inflammation to prevent sepsis during lung infection. *Science signaling* **11**
177 Gomez, J. A., Wapinski, O. L., Yang, Y. W., Bureau, J. F., Gopinath, S., Monack, D. M., Chang, H. Y., Brahic, M. & Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* **152**, 743-754
178 Collier, S. P., Henderson, M. A., Tossberg, J. T. & Aune, T. M. (2014). Regulation of the Th1 genomic locus from Ifng through Tmevpg1 by T-bet. *Journal of immunology* **193**, 3959-3965
179 Turman, M. A., Yabe, T., McSherry, C., Bach, F. H. & Houchins, J. P. (1993). Characterization of a novel gene (NKG7) on human chromosome 19 that is expressed in natural killer cells and T cells. *Human immunology* **36**, 34-40
180 Chiu, H. S., Somvanshi, S., Patel, E., Chen, T. W., Singh, V. P., Zorman, B., Patil, S. L., Pan, Y., Chatterjee, S. S., Cancer Genome Atlas Research, N., Sood, A. K., Gunaratne, P. H. & Sumazin, P. (2018). Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell reports* **23**, 297-312 e212
181 Fatica, A. & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews. Genetics* **15**, 7-21
182 Peng, Z., Wang, J., Shan, B., Yuan, F., Li, B., Dong, Y., Peng, W., Shi, W., Cheng, Y., Gao, Y., Zhang, C. & Duan, C. (2017). Genome-wide analyses of long noncoding RNA expression profiles in lung adenocarcinoma. *Scientific reports* **7**, 15331
183 Cui, D., Yu, C. H., Liu, M., Xia, Q. Q., Zhang, Y. F. & Jiang, W. L. (2016). Long non-coding RNA PVT1 as a novel biomarker for diagnosis and prognosis of non-small cell lung cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **37**, 4127-4134
184 Lin, P. C., Huang, H. D., Chang, C. C., Chang, Y. S., Yen, J. C., Lee, C. C., Chang, W. H., Liu, T. C. & Chang, J. G. (2016). Long noncoding RNA TUG1 is downregulated in non-small cell lung cancer and can regulate CELF1 on binding to PRC2. *BMC cancer* **16**, 583

185 Liu, H., Zhou, G., Fu, X., Cui, H., Pu, G., Xiao, Y., Sun, W., Dong, X., Zhang, L., Cao, S., Li, G., Wu, X. & Yang, X. (2017). Long noncoding RNA TUG1 is a diagnostic factor in lung adenocarcinoma and suppresses apoptosis via epigenetic silencing of BAX. *Oncotarget* **8**, 101899-101910
186 Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P. & Boon, T. (2014). Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nature reviews. Cancer* **14**, 135-146
187 Leman, J. K., Sandford, S. K., Rhodes, J. L. & Kemp, R. A. (2018). Multiparametric analysis of colorectal cancer immune responses. *World journal of gastroenterology* **24**, 2995-3005
188 Liu, S. J., Horlbeck, M. A., Cho, S. W., Birk, H. S., Malatesta, M., He, D., Attenello, F. J., Villalta, J. E., Cho, M. Y., Chen, Y., Mandegar, M. A., Olvera, M. P., Gilbert, L. A., Conklin, B. R., Chang, H. Y., Weissman, J. S. & Lim, D. A. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**
189 Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H. & Joung, J. K. (2013). CRISPR RNA-guided activation of endogenous human genes. *Nature methods* **10**, 977-979
190 Pinzon, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F. & Seitz, H. (2017). microRNA target prediction programs predict many false positives. *Genome research* **27**, 234-245
191 Hrdlickova, B., de Almeida, R. C., Borek, Z. & Withoff, S. (2014). Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et biophysica acta* **1842**, 1910-1922

Appendix A: Publications

1. Martinez VD*, Marshall EA*, Anderson C, Ng KW, Minatel BC, **Sage AP**, Enfield KSS, Xu Z, Lam WL (2019). Discovery of previously-undetected miRNAs in mesothelioma and their use as tissue-of-origin markers. *American Journal of Respiratory Cell and Molecular Biology*. In Press. [*Co-first Authorship].
2. Barros-Filho MC*, Pewarchuk ME*, Minatel BC*, **Sage AP**, Marshall EA, Martinez VD, Rock LD, MacAulay G, Kowalski LP, Rogatto SR, Garnis C, Lam WL (2019). Previously-undescribed thyroid-specific miRNA sequences in papillary thyroid carcinoma. *Journal of Human Genetics*. In press. [*Co-first Authorship].
3. Stewart GL, Enfield KSS, **Sage AP**, Martinez VD, Minatel BC, Pewarchuk ME, Marshall EA, Lam WL (2019). Aberrant expression of long non-coding RNAs from pseudogene loci highlights alternative mechanisms of cancer gene regulation in lung adenocarcinoma. *Frontiers in Genetics*. 10:138.
4. Barros-Filho MC*, Guisier F*, Rock LD*, Becker-Santos DD, **Sage AP**, Marshall EA, Lam WL (2019). Tumour suppressor genes with oncogenic roles in lung cancer. In *Tumour Suppressor Genes*. Guy-Joseph Lemamy ed., InTechOpen, London. In press. [*Co-first Authorship].
5. Guisier F*, Barros-Filho MC*, Rock LD*, Constantino FB, Minatel BC, **Sage AP**, Marshall EA, Martinez VD, Lam WL (2019). Small non-coding RNA expression in cancer. In *Gene Expression Profiling in Cancer* Dimitrios Vlachakis ed., InTechOpen, London. In press. [*Co-first Authorship].
6. **Martinez VD*, Sage AP*,** Marshall EA, Suzuki M, Goodarzi AA, Dellaire G, Lam WL (2018) Oncogenetics of lung cancer induced by environmental carcinogens. In *Oncogene and Carcinogenesis.* Pinar Erkekoğlu ed., InTechOpen, London. 1-24. [*Co-first Authorship].
7. **Sage AP*, Martinez VD***, Minatel BC, Pewarchuk ME, Marshall EA, MacAulay GM, Hubaux R, Pearson DD, Goodarzi AA, Dellaire G, Lam WL (2018). Genomics and epigenetics of malignant mesothelioma. *High Throughput* 7:20, 1-23. [*Co-first Authorship].
8. **Sage AP**, Minatel BC, Marshall EA, Martinez VD, Stewart GL, Enfield KSS, Lam WL (2018) Expanding the miRNA transcriptome of human kidney and renal cell carcinoma. *International Journal of Genomics* 2018:1-10.
9. **Minatel BC*, Sage AP***, Anderson C, Hubaux R, Marshall EA, Lam WL, Martinez VD (2018) Environmental arsenic exposure: genetic and epigenetic contributions to carcinogenesis. *Environment International* 112:183-97. [*Co-first Authorship]
10. Minatel BC, Martinez VD, Ng KW, **Sage AP**, Tokar T, Marshall EA, Anderson C, Enfield KSS, Stewart GL, Reis PP, Jurisica I, Lam WL (2018) Large-scale discovery of previously-undetected microRNAs specific to human liver. *Human Genomics* 12:16, 1-6.
11. Marshall EA, **Sage AP**, Ng KW, Martinez VD, Firmino NS, Bennewith KL, Lam WL (2017) Small non-coding RNA transcriptome of the NCI-60 cell line panel. *Scientific Data* 4:170157, 1-8.
12. **Sage AP*, Minatel BC***, Ng K, Stewart GL, Dummer TJB, Lam WL, Martinez VD (2017) Oncogenomic disruptions in arsenic-induced carcinogenesis. *Oncotarget* 8:25736-55. [*Co-first Authorship].