4D EX VIVO CRISPR/CAS9 WHOLE-GENOME SCREEN TO IDENTIFY GENES REGULATING LUNG CANCER METASTASIS

Alexandria Plumer

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations

Part of the Cancer Biology Commons, Cell Biology Commons, Disease Modeling Commons, and the Laboratory and Basic Science Research Commons

Recommended Citation
Plumer, Alexandria, "4D EX VIVO CRISPR/CAS9 WHOLE-GENOME SCREEN TO IDENTIFY GENES REGULATING LUNG CANCER METASTASIS" (2021). The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access). 1146. https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1146

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.
4D EX VIVO CRISPR/CAS9 WHOLE-GENOME SCREEN TO IDENTIFY GENES
REGULATING LUNG CANCER METASTASIS
by
Alexandria Kathryn Plumer, M.S.

APPROVED:

__________________________________________
Prahlad Ram, Ph.D.
Advisory Professor

__________________________________________
Faye Johnson, M.D., Ph.D.

__________________________________________
Ju-Seog Lee, Ph.D.

__________________________________________
Rick Wetsel, Ph.D.

__________________________________________
Nidhi Sahni, Ph.D.

__________________________________________
Bill Mattox, Ph.D.

APPROVED:

__________________________________________
Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences
4D EX VIVO CRISPR/CAS9 WHOLE-GENOME SCREEN TO IDENTIFY GENES REGULATING LUNG CANCER METASTASIS

A DISSERTATION

Presented to the Faculty of

The University of Texas
MD Anderson Cancer Center UTHealth
Graduate School of Biomedical Sciences

in Partial Fulfillment
of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

by
Alexandria Kathryn Plumer, M.S.

Houston, Texas
December 2021
Metastatic lung cancer has a 5-year survival rate of 5%. Lung cancers tend to be asymptomatic until late stages, and almost 90% are not diagnosed until they are advanced. Metastases are very rare events, often initiated by a single cell from a primary tumor into a new niche at a distant location. Investigation of the early metastatic process is of urgent need for the development of early diagnostics and targeted therapeutics. We performed a proof-of-concept CRISPR/Cas9 whole genome knockout screen in the A549 lung adenocarcinoma cell line and utilized a novel ex vivo 4D lung metastasis model to find gene candidates to test in vitro. I prepared samples and performed next generation sequencing. I created a computational pipeline to demultiplex samples and align reads to the library. I recovered genes previously implicated in lung cancer and metastasis enriched in the Primary and Metastasis samples. SPI1 was identified as the central node in network analysis of the Primary and Metastases and present in every sample. SPI1 was validated in vitro as a novel transcription factor involved in lung cancer metastasis.
DEDICATION

For Mom, Dad, Jackie and Alex.
ACKNOWLEDGMENTS

I want to thank my parents, Kathryn Gardner, MD and Mark Plumer, who supported and encouraged me throughout the process of this work. My mom is a practicing ophthalmologist who has always upheld the highest ethical standards and encouraged me to do the same. My dad was an aeronautical engineer and inventor who could fix anything and had creative solutions to every problem. During the course of this project, my dad was diagnosed with lung adenocarcinoma. I am extraordinarily grateful to my dad for allowing me to be part of his treatment process and decision-making, which gave me an unusual scientific and personal understanding of lung cancer. My dad always used humor to help me through difficult situations. Shortly before his death in April of 2021, he told me to keep my shark fin sharp. His last words to me were "I see great things for you. Take it easy and enjoy your life."

I also want to thank my best friend, Jackie Jay, who I grew up with in California and has come to live in Houston since I have been here. She was indispensable to my ability to finish college, my Master's degree and my PhD project. Jackie helped me learn to stay organized and developed systems for me to use. I also wouldn’t have been able to survive without her cooking. I also want to thank my cat, Schmoo Plumer, who has been with me since my sophomore year of college.

I want to thank the friends I have made during my graduate school experience who made this work possible. I thank Alex Davis, PhD for his friendship, guidance, constructive criticism, Redcap lunches, and enlightening conversations. It is no overestimation to say that this work would not have been possible without him. I thank Anna Casasent, PhD, who was one of my
first friends in graduate school and has been one of the most instrumental people in my graduate work. When I needed something, Anna was the one who could find it. She introduced me to Keith Baggerly, PhD, who was able to help me articulate what my goals and interests were when I joined the Bioinformatics, Biostatistics and Systems Biology graduate program. Anna also introduced me to Beata Lerman, PhD, who gave me guidance and encouragement during my lab transition that I could not have done without. When I needed an electroporator and was unable to find one anywhere, Anna located the machine and helped me transfer it to my lab on Christmas Eve. I want to thank Matthew Cagley, who I met as a summer student and remained friends with until he returned to MD Anderson as a research coordinator. Matthew is willing to help anybody with anything, including hauling equipment during our lab move. I always admired and appreciated his understanding and concern about the culture and ethics of science, as well as the interesting technical parts. I want to thank Daniel McGrail, PhD, who took me under his wing during the final portion of my project when I really needed help. Daniel spent a lot of time helping me design and interpret my experiments. He also shared his lunch (first and second) and music with me, and I'm extremely grateful for his guidance and friendship.

I would like to thank those who collaborated with us on this project, Min Kim, MD and Dhuvra Mishra, PhD. Dhuvra’s work on the lung model and his explanations during our meetings were helpful in my understanding of the work and completion of this project. I also want to thank Ellie Seviour, PhD, who started this project prior to my arrival in Ram’s lab. After showing me the ropes in Dr. Ram's lab when I arrived, Ellie has remained a close friend who stuffs my Christmas stocking.
I want to thank the graduate school (GSBS), which is full of special people who are genuinely interested in the wellbeing of their students. The grad school went above and beyond to help me with my various difficulties, and I'm eternally grateful to everyone who participated in my education. In particular, I would like to thank Brenda Gaughan, who has given me encouraging words and great advice in our many meetings over the years. I also want to thank the Deans, in particular Bill Mattox, PhD, who took a special interest in my education, helped me through many hurdles over the years and became an important part of my thesis committee.

I thank my thesis committee, each of whom had a significant contribution to the development of my ideas and to this project. Nidhi Sahni, PhD and Ju-Seog Lee, PhD were sources of extremely interesting ideas on my project and welcome encouragement. I especially want to thank Faye Johnson, MD, PhD for her guidance on organization of a project and her help seeing the bigger picture. She was one of the first people who I spoke to about my dad's diagnosis, and her advice and experience was helpful to me and kind of her to share. I also want to give a special thanks to Rick Wetsel, PhD, the longest standing member of my thesis committee and the chair of my candidacy exam committee. I have enjoyed and appreciated Rick's presence for the long haul and feel that he has given me some of the most valuable advice I've ever received.

Finally, I would like to thank my mentor, Prahlad Ram, PhD. While we didn't always get along, I always felt that Ram supported and believed in me. Ram had an ability to understand what was going on in my head and helped me articulate it. He also had a unique ability to see the big picture of a project that involved multiple different disciplines, and understood the culture and interactions involved in communicating ideas to different people. I always admired
him for this. I do not feel that there was any other advisor who could have understood my skillset and interests in the way that Ram did (without my communicating it) and he encouraged me to pursue them in my future work.
# TABLE OF CONTENTS

Approval Page ....................................................................................................................... i

Title Page ............................................................................................................................... ii

Abstract ................................................................................................................................. iii

Dedication ............................................................................................................................... iv

Acknowledgments ................................................................................................................... v

Table of Contents .................................................................................................................... ix

Table of Figures ....................................................................................................................... xiv

List of Tables .......................................................................................................................... xvii

1. Introduction ........................................................................................................................ 1

1.1 The Steps of Metastasis .................................................................................................... 2

1.1.1 Proliferation of the primary tumor ............................................................................. 5

1.1.2 Escape of cells from the primary tumor ..................................................................... 5

1.1.3 Survival in the circulation .......................................................................................... 6

1.1.4 Arrest, extravasation, and metastasis formation ......................................................... 6

1.2 Modeling Metastasis ......................................................................................................... 7

1.2.1 4D ex vivo lung model of metastasis .......................................................................... 7

1.3 Whole-genome functional screening ................................................................................ 10

1.4 SPI1 Transcription factor ............................................................................................... 13

1.4.1 SPI1 is expressed in lung tissue ................................................................................ 14

ix
2. Development of a Computational Pipeline to Process and Demultiplex Pooled Screen Sequence Data

2.1 Introduction

2.2 Methods

2.2.1 GeCKOv2 CRISPR-Cas9 pooled library generation

2.2.2 4D ex vivo lung model of metastasis

2.2.3 Sample collection

2.2.4 Library Preparation & Sequencing

2.2.5 Input Library Sequencing

2.2.6 Sequence Processing and Alignment

2.3 Results

2.3.1 Library Preparation & Pooled Sequencing

2.3.2 Fragment analysis shows expected product size

2.3.3 Sequence Processing

2.3.4 Input Library Sequencing

2.4 Pooled screen sequence data summary

2.4.1 Sequence Analysis Demultiplexing & FASTQ Analysis

2.4.2 Processing of sequencing data

2.5 Discussion
3. Analysis of Screen Data to Identify Genes Regulating Metastasis ...........................................38

3.1 Introduction ......................................................................................................................................38

3.1.1 Library Dynamics and Expected Trends ..................................................................................38

3.2 Methods ..........................................................................................................................................40

3.2.1 Read Count Data Processing and Preparation .........................................................................41

3.2.2 Network Analysis ......................................................................................................................41

3.2.3 Database Analysis ......................................................................................................................41

3.3 Results ...........................................................................................................................................41

3.3.1 Input Cells Yielded High Library Coverage ...........................................................................41

3.3.2 Exploratory Data Analysis of Pooled Samples ........................................................................42

3.3.3 Selection of Gene Candidates ...................................................................................................52

3.3.4 Analysis Revealed Genes Previously Implicated in Lung Cancer .......................................54

3.3.5 SPI1 is a Novel Finding in Lung Cancer ...................................................................................58

3.4 Discussion .......................................................................................................................................61

4. Experimental Validation of SPI1 Screen Candidate in Vitro .......................................................65

4.1 Introduction .......................................................................................................................................65

4.2 Methods ..........................................................................................................................................66

4.2.1 Cell Culture ..............................................................................................................................66

4.2.2 siRNA Knockdown ....................................................................................................................66

4.2.3 Semi-quantitative RT-PCR .......................................................................................................67
4.2.4 Cell Growth & Viability .................................................................68
4.2.5 CyQUANT Cell Count Curves .......................................................69
4.2.6 Anchorage Independence Assay ..................................................69
4.2.7 Invasion Assay ............................................................................69

4.3 Results ..........................................................................................70
4.3.1 Expression of SPI1 in A549 Cells ....................................................70
4.3.2 SPI1 knockdown alters cell attachment but not overall cell number in A549 cells 71
4.3.3 SPI1 knockdown increases number of viable detached A549 cells in transfection plates ....................................................................................................................72
4.3.4 SPI1 knockdown decreases cell attachment ......................................75
4.3.5 Anchorage Independence Assays ...................................................76
4.3.6 SPI1 knockdown does not affect cellular migration or invasion ......80

4.4 Discussion .......................................................................................81

5. Conclusions and Future Directions .....................................................84
5.1 Project Summary .............................................................................84
5.2 Contributions of this thesis ...............................................................86
5.2.1 SPI1 may function as a metastasis suppressor in lung adenocarcinoma ....86
5.2.2 Screen Design ............................................................................89
5.3 Next Steps ......................................................................................92
5.4 Future Directions .............................................................................................................94

6. Bibliography .....................................................................................................................97

VITA ........................................................................................................................................110
TABLE OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Metastasis is a multistep process that allows cancer cells to spread from the primary site to a distant location in the body.</td>
</tr>
<tr>
<td>1.2</td>
<td>Metastasis formation in 4D system.</td>
</tr>
<tr>
<td>1.3</td>
<td><strong>CRISPR/Cas9 NHEJ schematic.</strong> The guide sequence guides the Cas9 enzyme to the desired sequence of double-stranded DNA.</td>
</tr>
<tr>
<td>1.4</td>
<td>The <strong>CRISPR/Cas9</strong> gene editing system allows for large-scale gene perturbations in cell lines.</td>
</tr>
<tr>
<td>1.5</td>
<td>Schematic overview of study design.</td>
</tr>
<tr>
<td>2.1</td>
<td>GeCKOv2 lentiviral vector containing guide sequences.</td>
</tr>
<tr>
<td>2.2</td>
<td>Sample collection for 4D model.</td>
</tr>
<tr>
<td>2.3</td>
<td>Harvest of cells from lung matrices.</td>
</tr>
<tr>
<td>2.4</td>
<td>PCR amplification and barcoding for pooled library sequencing.</td>
</tr>
<tr>
<td>2.5</td>
<td>Illumina sequencing schematic.</td>
</tr>
<tr>
<td>2.6</td>
<td><strong>Input gDNA library in A549 cells.</strong> gDNA was isolated and run through the fragment analyzer to reveal 2 peaks, one is close to the correct size and one is larger representing a primer dimer.</td>
</tr>
<tr>
<td>2.7</td>
<td>Submission of screening samples for fragment analysis.</td>
</tr>
<tr>
<td>2.8</td>
<td>Sequence processing and filtering steps.</td>
</tr>
<tr>
<td>2.9</td>
<td><strong>Code used to trim and align the Input library sequences.</strong></td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic overview of progressive loss of unique guides during screen.</td>
</tr>
<tr>
<td>3.2</td>
<td>Lorenz curve showing the inequality between the guides with the highest and lowest read. counts.</td>
</tr>
</tbody>
</table>
Figure 3.3. Number of genes per guide count for the input sequencing round using ion torrent (IPTC, left) and the pooled sequences (NSEQ, right). ..................................................46

Figure 3.4. Distribution of log2 read counts for the input sequencing (IPTC, purple) and the pooled input (NSEQ, yellow). .................................................................47

Figure 3.5. Comparing two methods for similarity analysis across all screening samples. ..........................................................................................................................49

Figure 3.6. Number of unique guides found in each sample........................................50

Figure 3.7. Euclidean distance heatmap of the primary and metastasis samples.. .....51

Figure 3.8. Venn diagrams created with R package eulerr ........................................52

Figure 3.9. Gene prioritization algorithm used to select gene hit candidates. ..........54

Figure 3.10. Network of genes with top average guide count for all samples in the screen. .......................................................................................................................55

Figure 3.11. Network generated by the gene selection algorithm for the Primary and Metastasis samples. ........................................................................................................56

Figure 3.12. All possible SPI1 network connections in the NetWalker database. ......57

Figure 3.13. Expression of SPI1 in human tissues at the levels of RNA (left) and protein (right). ......................................................................................................................59

Figure 3.14. RNA expression of SPI1 in clinical samples of normal and cancerous lung. .............................................................................................................................60

Figure 3.15. Kaplan-Meier survival curve of SPI1 low and high expression ..........61

Figure 4.1. Objective 3: In vitro validation of gene candidate SPI1. .......................65

Figure 4.2. Semi-quantitative Reverse Transcriptase PCR of SPI1 mRNA in A549 cells. ...........................................................................................................................71
Figure 4.3. Images of A549 cells 72h post siRNA transfection.................................72
Figure 4.4. PrestoBlue viability curve over 96h following SPI1 knockdown in A549s.
........................................................................................................................................73
Figure 4.5. PrestoBlue viability curve keeping original media.................................74
Figure 4.6. CyQuant growth curves show no significant difference in cells treated with
SPI1 siRNA compared to control groups...............................................................75
Figure 4.7. Number of viable cells suspended in media 48h post-transfection........76
Figure 4.8. Optimization of anchorage independence assay.....................................77
Figure 4.9. Colony count vs. number of cells seeded................................................78
Figure 4.10. Mean colony count and mean colony area per well.............................79
Figure 4.11. Colony size (right) and colony count (left) for anchorage independence
assay of siRNA knockdowns..................................................................................80
Figure 4.12. Representative figure of anchorage-independence assay following siRNA
transfection.................................................................................................................80
LIST OF TABLES

Table 2.1. Number of genes, miRNAs and nontargeting controls in the GeCKOv2 library. .......................................................... 20

Table 2.2. Primer sequences for PCR1 and PCR2.................................................. 24

Table 2.3. Thermocycler settings for PCR1 and PCR2 reactions. PCR1 used 18 cycles of amplification, while PCR2 used 24 cycles (not shown in table) ................................. 24

Table 2.4. Recovery of GeCKOv2 library guide sequences from sequencing of the input cells using Ion Torrent .......................................................... 34

Table 2.5. Specifications of LentiCRISPR primers used to amplify GeCKOv2 library. ................................................................................. 36

Table 3.1. Number of recovered guides for the genes, miRNAs and controls groups from the input library sequencing. ................................................................. 42

Table 3.2. Number of recovered guides for the genes, miRNAs and controls groups from the input library sequencing. ................................................................. 42

Table 3.3. Guide numbers. Read counts and library percentages covered by the IPTC (first round) and NSEQ (second round) sequencing, respectively ......................... 44

Table 4.1. Cell numbers and volumes of reagents used for siRNA knockdown experiments. ..................................................................................... 66

Table 4.2. Sequences of siRNA used to knock down SPI1. Sequences are in 5’ to 3’ direction .......................................................................................... 67

Table 4.3. PCR thermocycler settings for both PCRs performed ............................. 67
1. **INTRODUCTION**

Lung cancer is the leading cause of cancer death in the United States. While it consists of 14% of new cancer diagnoses, lung cancer is the most deadly, responsible for 25% of cancer deaths. Most patients with lung cancer present with metastatic disease at diagnosis. The 5-year survival rate for localized lung cancer is 55%, it drops to 4% once the disease has metastasized (Siegel et al. 2016). Seventy percent of lung cancer patients have detectable metastases at initial diagnosis and 90% of cancer deaths result from metastases (Tanaka et al. 2009).

Accumulation of alterations in genomes of somatic cells over time leads to cancer (Hanahan and Weinberg 2011). Tumors begin when a single cell acquires mutations, causing it to proceed through the cell cycle unchecked. The cell is unable to sense environmental cues about when to divide (Hanahan and Weinberg 2000). No single mutation causes cancer. Tumor cells divide without having to meet their own quality control standards (eg. the cell cycle checkpoints) and genomic integrity. Cancer has been referred to as a “wound that never heals” (Deyell et al. 2021). While the original understanding of cancer is that it comes from single cells, the long-standing definition of malignancy is changing. Malignancy refers to cancer that becomes deadly, which happens when cancer spreads (Welch and Hurst 2019).

The word “metastasis” is both a noun and a verb, referring to both the clinical outcome as well as the process by which this spread occurs. Metastasis is the process by which malignant cells travel and grow discontinuously from the primary tumor (Welch and Hurst 2019). This process is thought to occur in distinct stages, each of which require the cell to display a set of characteristics that allow survival of specific intercellular and microenvironmental threats (Gill and West 2014; Hecht et al. 2015; Icard et al. 2014).
Once metastasis occurs, the death rate of any given cancer increases precipitously (Vanharanta and Massagué 2013). The ability of a tumor cell to inhabit a distinct niche from the organ-of-origin is a clear delineation between survivable and deadly disease in the majority of cases. Primary tumors that have not invaded the surrounding tissue and are therefore isolated and localized are called “in situ.” Technological advances allowing for better resolution in imaging have allowed for the appreciation of how many tumors in situ actually exist in asymptomatic people. Prior to ability to detect earlier and smaller tumors, it was thought that the development of a tumor was coupled with subsequent metastatic progression and death. In reality, it looks as though many tumors are resolved without intervention, by natural defenses. It has been suggested that up to 60% of adenocarcinomas in situ of the breast might never progress to metastatic disease (Van Seijen et al. n.d.). Data on adenocarcinomas in situ of the lung are not available. Since non-invasive tumors are not deadly, the definition of malignancy should reflect this (Welch and Hurst 2019).

There is currently very little known about the early metastatic process, which will be very important knowledge in order to develop treatments, earlier diagnostics and prophylactics in order to improve the survival rate and quality of life in those who suffer from lung cancer.

1.1 THE STEPS OF METASTASIS

Metastasis is a multistep process that allows cancer cells to spread from the primary site to a distant location in the body and includes (but is not limited to) invasion, intravasation, circulation, extravasation and proliferation at a distant site (Arvelo et al. 2016; Kam et al. 2012; Martinez-Marti et al. 2017; Sahai 2007). Successful performance of all functions on this list is
required for a cell to metastasize, and therefore these functions are criteria for the definition of malignancy.

**Steps of metastasis** (Figure 1.1):

1. Proliferation of primary tumor.

2. Local invasion of surrounding extracellular matrix (ECM) and migration through stromal cell layers. Proteolytic destruction of the basement membrane and extracellular matrix.

3. Intravasation: the movement of cells through the basement membrane into the blood vessel.


5. Arrest at a new location.

6. Extravasation: movement of cells through the vessel wall into the parenchyma of tissue.

7. Formation of micrometastases.

8. Detectable metastasis. Tumor cells resume unchecked proliferative and locally invasive programs.
Figure 1.1. **Metastasis is a multistep process that allows cancer cells to spread from the primary site to a distant location in the body.** The primary tumor proliferates and begins to invade the surrounding tissue using matrix metalloproteases. Cells detach and intravasate into the bloodstream. Circulating tumor cells can survive detachment and circulation, a process that usually causes cell apoptosis (anoikis). Cells extravasate or enter distant tissue from the bloodstream through the vasculature. The secondary location where the cells can settle and begin growth is the final step in metastasis. Reproduced from Wirtz D, Konstantopoulos K, Searson PC. The physics of cancer: The role of physical interactions and mechanical forces in metastasis. Nat Rev Cancer. 2011;11(7):512-522. doi:10.1038/nrc3080 with permission from the Copyright Clearance Center (Wirtz et al. 2011).

In this dissertation, the steps of metastasis are described to understand the rationale behind development of a specific cell culture model of metastasis, which will aid in interpreting the results from the experiments and placement of those results within the larger conceptual framework of metastasis research as this study contributes to the gap in knowledge. For these purposes, the steps of metastasis can be grouped together into broader categories that reflect “compartments” or locations and functions of cancer cells as they progress from primary tumor to the end resulting metastatic lesions. The 4 resulting categories are the proliferation of the
primary tumor, escape of cells from the primary tumor, survival in the circulation, and formation of metastases.

1.1.1 Proliferation of the primary tumor

While discussed briefly in the introduction, basic oncogenesis and primary tumor characteristics are not the focus of this work. The primary tumor will only be discussed as the cells proliferating within it ultimately go on to seed metastatic tumors, and the primary tumor and surrounding microenvironment provide the conditions and selection pressures that are required for cells to begin the metastatic process.

The primary tumor grows as a result of uncontrolled cell cycle progression and loss of response to environmental growth or stop cues from the surrounding tissue. The primary tumor gets larger and takes on several new characteristics. There is an accumulation of genetic alterations in the primary tumor. Genetic instability means that the primary tumor is undergoing an evolutionary selection process. Cells that reproduce the most quickly are selected. As the primary tumor gets larger, it needs more nutrients and oxygen. The cells that can acquire necessary ingredients from the environment are selected (Gerlinger et al. 2014; Jamal-Hanjani et al. 2017).

1.1.2 Escape of cells from the primary tumor

The invasive compartment includes the acquisition of motility by cells at the periphery of the primary tumor, allowing them to leave. The process begins with migration of tumor cells to the tumor periphery and invasion into the stroma, which is the connective tissue surrounding the tumor. Invasive capability is a normal function that cells perform during embryogenesis.
Tumor cells develop invasive capacity in hypoxic conditions that occur as solid tumors grow rapidly and disrupt the normal tissue nutrient supply. The transcriptional programs for invasiveness are turned on in primary tumors as they become larger and the cells become low on nutrients and oxygen.

1.1.3 Survival in the circulation

Once tumor cells have invaded into the vasculature, they are confronted with an environment that is very different from the epithelial cell niche. When normal epithelial cells undergo loss of cell-cell adhesions, they undergo anoikis, or apoptotic cell death in response to loss of contact with other cells (Chambers et al. 2001; Sleeman et al. 2011). Normal human fibroblasts were thought to be incapable of anchorage-independent growth until it was shown that they possess this capability when provided the right conditions of high serum and hydrocortisone (Peehl and Stanbridge 1981).

Until recently, little was known about circulating tumor cells (CTCs). Methods for isolation and study are nascent and present many challenges. Quantification of circulating tumor cells in patients is discordant with origin or size of the primary tumor. Most tumor cells only survive 20-30 minutes in in the circulation. By 24 hours, only 1% of tumor cells are still alive in circulation (Chambers et al. 2001). Experimental evidence shows that tumor cells perform rolling adhesion like leukocytes along the vessel wall rather than float freely through the blood. This rolling adhesion form of movement throughout the vasculature explains the difficulty of studies to show consistency among each other related to tumor size and number of CTCs (Azevedo et al. 2015; Popper 2016; Vanharanta and Massagué 2013).

1.1.4 Arrest, extravasation, and metastasis formation
Once a cell has survived in the circulation, the next steps in metastasis involve settling in a distant location in a dormant state for a period of time. By the time a tumor is detectable, it consists of $10^{10}$-$10^{11}$ cells (Nolte and Singh 2014). Lineage tracing using DNA sequencing shows that most metastases arise clonally from a single cell (Blanpain 2013). We also know that millions of cells escape the primary tumor. The inefficiency of metastasis is a good thing for human health, because otherwise many more people would die of cancer and survival times would be much shorter. However, the inefficiency of metastasis also makes the process extremely difficult to study (Chambers et al. 2001; Sleeboom et al. 2018; Sleeman et al. 2012; Welch and Hurst 2019).

1.2 MODELING METASTASIS

Model systems are important tools for investigating biological phenomena. Historically metastasis is difficult to study, as little is known about early events in metastatic process. The process of metastasis itself creates limitations to its study. Metastasis starts out with single cells, which means that the limits of detection make collection and study of those cells difficult. Additionally, most of those cells do not make it to the end to meet the definition of metastasis. Because metastasis is so inefficient, collecting cells at the very beginning of the process (like circulating tumor cells) does not say anything about which of those cells will go on to form metastases (Chambers et al. 2001; Matikas et al. 2016; Tanaka et al. 2009).

1.2.1 4D ex vivo lung model of metastasis

The current methods of studying metastasis can broadly be separated into 2 groups: in vivo, which includes observational studies of human cancers, sample collection from patients, and animal studies, and in vitro, which encompasses the studies performed in cell culture labs using
ex vivo cells and cell lines cultivated in artificial growth conditions. The advantages to each broad category of study are also the reason for their limitations. The advantage to study of metastasis using in vivo systems is that the process occurs naturally. The limitation to in vivo studies is that there are several black holes in understanding the process of metastasis. For example, it is very difficult to isolate and study the early processes of metastasis, which involve single cells. In vitro model systems present an opposite set of limitations, as these assays tend to be limited to 2D growth (growth in a monolayer) as well as one single aspect of cancer cell growth (such as migration or invasion), which is not representative of the process of metastasis.

Tumor characteristics are highly dependent on tissue architecture, and many studies have shown that cells grown in traditional 2D culture to not adequately model cancer growth dynamics, cellular interactions, or drug sensitivity. It was found that the same cell line grown in 3D will express over 1,500 genes differently than when grown in a traditional 2D setup, with all other conditions like culture media and substrate being the same. While the 3D cell invasion assay overcomes some of the limitations of the 2D systems, they lack the ability to track and collect cells from one location to another (Nierode et al. 2015).

To overcome limitations of previous in vitro model systems, the ability to represent a broader process of metastasis is necessary. To overcome the limitations of study of metastasis in vivo, models must have the ability to be optimized and controlled. A model that represents a large swath of the in vivo invasion into the vasculature, survival in circulation and ability to colonize a distal location, while still allowing for the observability and control of an in vitro cell culture assay was the impetus for the development of a novel 4D ex vivo model of lung cancer metastasis (Mishra et al. 2014, 2015). Rat lungs are excised and decellularized to leave
an extracellular matrix lung for 3-dimensional growth of cells (Vishnoi et al. 2014). Tumor cells are then inoculated into the left lung through a one-way tracheal cannula, where they can adhere and form a primary tumor. For cells to reach the right lung, they first have to enter the vasculature, mimicking the in vivo system. Circulating tumor cells can be isolated from the vascular perfusion system within the first 2 weeks of inoculation. Metastases form in the 4D system in well-studied lung cancer cell lines A549, H1299 and H460 (Figure 1.2) (Mishra et al. 2014).

![Figure 1.2. Metastasis formation in 4D system.](image)

Metastases form in the right lung at varying rates, in a cell line-dependent manner. A549 cells (white bars) show metastases at 25 days. Reproduced from Mishra DK, Creighton CJ, Zhang Y, Gibbons DL, Kurie JM, Kim MP. Gene expression profile of A549 cells from tissue of 4D model predicts poor prognosis in lung cancer
1.3 WHOLE-GENOME FUNCTIONAL SCREENING

The era of pooled screening began with the forward genetics approach, which started with a phenotype and attempted to determine the genetic cause. These screens used random mutagenesis in the form of irradiation or chemical mutagens that covered the genome of the population being studied by knowing the size of the genome and number of mutations made by each unit or volume of mutagen. The phenotypes that appeared would be investigated by painstaking and time-consuming linkage studies or cloning of the mutated gene (Sanjana 2017).

What followed the random mutagenesis approach was a genome-wide manipulation at the transcriptional level by RNA interference (RNAi). RNAi screens were the transition from forward to reverse genetics (Ritchie et al. 2014). Pooled screening uses selection of phenotype to determine the genotype that is responsible for that phenotype. Pooled phenotypic screening allowed for unbiased analysis of the entire genome by making perturbations in each gene in individual cells of a population. While these screens were revolutionary in beginning a new era of genetics, pooled RNAi screens had issues associated with incomplete knockdown, problems involved in the use of the endogenous knockdown machinery, and significant off-target effects (Wang et al. 2018).

The next phase of the reverse genetics was allowed by harnessing of the clustered regularly interspaced short palindromic repeats (CRISPR) system, which allowed for programmable, highly targeted perturbation of large sets of genes in the genome (Chow and Chen 2018; Ledford 2015). The CRISPR system is an endogenous bacterial immune has been used to...
create pooled single guide RNA (sgRNA) libraries that target all of the known genes in the genome at several different locations of each gene. CRISPR sequences are bacterial genome sequences that protect against viral pathogens. CRISPR-associated protein Cas9 is an endonuclease that cleaves double-stranded DNA. Cas9 is directed to its target by a section of RNA that hybridizes to double-stranded DNA sequences (Jinek et al. 2012; Rath et al. 2015). A synthetic single guide RNA (sgRNA); the section of RNA which binds to the genomic DNA is 18–20 nucleotides. The protospacer adjacent motif (PAM) is a DNA sequence at the 3’ end of the guide RNA that is required to for cutting. Non-homologous end joining (NHEJ) when repaired incorrectly will cause an insertion/deletion of DNA (Leenay and Beisel 2017) (Figure 1.3).

![CRISPR/Cas9 NHEJ schematic](image)

**Figure 1.3. CRISPR/Cas9 NHEJ schematic.** The guide sequence guides the Cas9 enzyme to the desired sequence of double-stranded DNA.

Whole-genome knockout screening relies on imperfect DNA repair following double-stranded DNA breaks. When a double-stranded break is made and the cell repairs it using non-homologous end-joining (NHEJ), the process is imperfect a certain portion of the time. Because the Cas9 enzyme is integrated into the cell genome using lentiviral transduction, the enzyme is constantly transcribed with the guide RNA, and the Cas9 enzyme will cleave the
target sequence in the genome repeatedly until the genome sequence is mutated and no longer attracts the guide sequence (Pennisi 2013).

In order to synthesize a guide RNA library, oligonucleotides containing 20 base pair-long sgRNAs flanked by cloning site sequences are synthesized on a chip, which allows for parallel synthesis of the unique sequences in a large library. The flanking sites are used to clone the short segment containing the unique sgRNA sequence into the lentiviral backbone to create the library construct. The complete circular lentiviral construct contains the Cas9 gene, some enhancers, and lentiviral integration sequences. The lentiviral library is transduced into a cell line optimized for producing viral particles, which are then collected and used to infect the cells on which the phenotypic screening will be performed. The CRISPR/Cas9 system is transduced into cells using lentivirus, and next generation sequencing is used to determine enriched and depleted guides (Figure 1.4) (Carstens et al. 2018; Shalem et al. 2014; Zhou et al. 2014).

Figure 1.4. The CRISPR/Cas9 gene editing system allows for large-scale gene perturbations in cell lines. The oligonucleotides are synthesized on a chip and transduced
into cells using lentivirus. sgRNA pool is analyzed to determine enriched and depleted guides. Adapted from (Sanjana et al. 2014)

Reverse genetics screens are made possible by building mechanistic understanding and technological advancement to generate highly specific, targeted genetic perturbations to large numbers of elements in parallel. Reverse genetics screens identify genes that play a role in a specific phenotype. Pooled screens start with a theoretically equal mixture of genetic perturbations, which will change in distribution following phenotypic selection, with some elements influencing this phenotype being enriched and others depleted. With this system, determination of which genetic perturbations have influenced the phenotype is easily done by next generation sequencing of the region of interest, which is directly 3’ of the lentiviral backbone sequence that should be in every cell being screened. These pooled oligonucleotide synthesis technology and cloning techniques were used to create the first large Cas9-sgRNA lentiviral libraries (Koike-Yusa et al. 2014; Shalem et al. 2014; Wang et al. 2014).

1.4 SPI1 TRANSCRIPTION FACTOR

SPI1 is an ETS-domain transcription factor involved in myeloid and B-lymphoid differentiation and activation in normal development. Its full name is spleen focus forming virus (SFFV) proviral integration oncogene due to its discovery as a proto-oncogene. The SPI1 knockout mouse (PU.1) mounts an inadequate response in both the innate and acquired arms of immunity (Elton et al. 2013). Macrophage and monocyte cells do not mature appropriately, and B cell function is inhibited by inappropriate light chain response (McKercher et al. 1996; Zakrzewska et al. 2010) SPI1 also binds RNA and is thought to be involved in alternative splicing of mRNA transcripts (Guillouf et al. 2006). SPI1 encodes an ETS-domain transcription factor that activates gene expression during myeloid and B-lymphoid cell
development. SPI1 is expressed during differentiation. The SPI1/PU.1 transcription factor is a key regulator of many steps of hematopoiesis and limits self-renewal of hematopoietic stem cells (Mak et al. 2011). The deregulation of its expression or activity contributes to leukemia, in which SPI1 can be either an oncogene or a tumor suppressor. Cellular senescence is an antitumoral pathway that restrains cell proliferation, is a mechanism by which SPI1 limits hematopoietic cell expansion, and thus prevents the development of leukemia (Delestré et al. 2017).

SPI1 has been widely studied in hematopoietic cell lineages, particularly for its role in plasma cell, monocyte and macrophage differentiation and immune response (Homo 2017). There are also many studies linking SPI1 to leukemia and lymphoma, mainly originating from showing up on transcriptome analyses.

1.4.1 SPI1 is expressed in lung tissue

During development, lung tissue arises from hematopoietic progenitor cells. Multipotent cells from the bone marrow populate the lung, differentiating into lung resident macrophages and epithelia that lines the airway. While the transcriptional profile of lung epithelial cells clusters with other types of epithelia throughout the body, the epithelia of the lung is similar to hematopoietic cell types (Fagerberg et al. 2014). To understand this transcription factor’s role in lung cancer, we searched clinical data. We found that SPI1 was expressed in normal lung tissue and significantly decreased or eliminated in the histology specimens from lung cancer patients (Human Protein Atlas, Chapter 3). At the time of this writing, a PubMed search using terms “SPI1” and “lung cancer” yields 4 results (Bai and Hu 2012; Koch et al. 2017; Kopf et al. 2015; Kossenkov et al. 2011). Substitution of terms were attempted in a few combinations,
yielding alternative gene name “PU.1” or “lung adenocarcinoma.” Using Pubmed search terms “lung cancer” and “SPI1,” the result is 3 impressively diverse articles with little information applicable to our work. SPI1 in lung cancer and very little that discussed its presence or role in epithelial cells.

1.5 PROJECT DESCRIPTION AND OBJECTIVES

Metastatic lung cancer has a 5-year survival rate of 5%. Lung cancers tend to be asymptomatic until late stages, and almost 90% are not diagnosed until they are advanced. Metastases are very rare events, often initiated by a single cell from a primary tumor into a new niche at a distant location. There is currently very little known about how individual lung cancer cells to survive detachment from the primary tumor and travel to a distant location. In order to develop therapeutic, diagnostic and preventative approaches, elucidation of the genes that regulate early metastatic process is of urgent need.

We performed a whole-genome knockout screen using a CRISPR/Cas9 library knockout screen in and utilized the novel ex vivo 4D lung metastasis model. Analysis of the screen data revealed many known regulators of lung cancer metastasis as well as novel gene targets. Our screen identified several genes in the SPI1 transcription factor network. SPI1 is part of the ETS-domain transcription factor family whose expression is important for cellular differentiation in certain immune cell populations. Because SPI1 is a proto-oncogene implicated in some blood cancers and is expressed in normal lung epithelium, we hypothesize that knockout of this gene induces a circulating phenotype by decreasing cell-cell adhesion in lung epithelial cells. Experimental testing of SPI1 knockdown in A549 cells revealed increased
detachment and viability in phenotypic assays, consistent with the requirements of early metastasis.

The overall purpose of this project was to identify key regulator of early lung cancer metastasis. More specifically, this dissertation aimed to demonstrate the feasibility of a novel screening method for genetic regulators of lung cancer metastasis, design a relevant analysis pipeline to identify gene candidates, and experimentally investigate individual genes using phenotypic assays (Figure 1.5). The objectives of this project correspond to the chapters of this dissertation and are as follows:

1. Design a computational pipeline to demultiplex & process pooled sequencing data.
2. Analysis of screen data to identify genes regulating metastasis in 4D model.
3. Experimental testing of SPI1 gene target in vitro.
1. Generate knockout library
2. Screen for lung metastasis genes
3. Pooled NGS of samples
4. Analysis of screen data
5. Generate candidate gene list
6. Validate and assay individual genes
Figure 1.5. **Schematic overview of study design.** A whole-genome CRISPR/Cas9 knockout library is generated in A549s using a lentiviral delivery approach. A primary tumor consisting of 25 million cells is seeded in the left lung. The 4D lung model runs for 15 days with constant cell culture media circulation. On Day 15, the left and right lungs are harvested. Genomic DNA is extracted from the samples and PCR is performed to amplify and barcode samples for pooled sequencing. Sequence data are demultiplexed using the individual sample barcodes and the 5’ primer region is trimmed. The reads are aligned to the GeCKOv2 guide sequence library. Analysis is performed to determine the distribution of guides present in each sample and determine candidates for genes that may be involved in lung cancer metastasis. Network analysis provides increased stringency for narrowing down the list to genes that are connected in regulatory, interactional, or transcriptional databases. Experimental validation is performed on individual genes using *in vitro* cell culture phenotypic assays.
2. DEVELOPMENT OF A COMPUTATIONAL PIPELINE TO PROCESS AND DEMULTIPLEX POOLED SCREEN SEQUENCE DATA

2.1 INTRODUCTION

In this chapter, I describe the methods used for creation of a whole-genome CRISPR/Cas9 knockout library in A549 cells. I then go on to describe the 4D ex vivo lung cancer screen, collection of samples from the lung model and preparation of the samples for next generation sequencing. Finally, I describe how I developed a computational pipeline to process and demultiplex the sequencing data.

2.2 METHODS

Library generation, isolation of cells from the rat lung matrices and gDNA preparation was performed by Elena Seviour, PhD. The 4D ex vivo lung model was performed and harvested by Dhuvra Mishra, PhD.

2.2.1 GeCKOv2 CRISPR-Cas9 pooled library generation

The GeCKOv2 CRISPR-Cas9 knockout library generated by the Zhang lab at the Broad Institute was used (Shalem et al. 2014). The library consists of unique 20bp guide sequences that have been integrated into a universal plasmid vector (Figure 2.1). The GeCKOv2 library targets 19,050 genes and 1,864 microRNAs (miRs). Each gene is targeted by 6 sgRNAs and each microRNA is targeted by 4 sgRNAs (Table 2.1). The GeCKOv2 library also contains 1000 non-targeting controls. The GeCKOv2 library contains a total of 122,411 sgRNAs. Gene knockouts are generated using a lentiviral transduction approach for efficient delivery and stable gene editing (Shalem et al. 2014).
Figure 2.1. **GeCKOv2 lentiviral vector containing guide sequences.**

Table 2.1. **Number of genes, miRNAs and nontargeting controls in the GeCKOv2 library.**

<table>
<thead>
<tr>
<th>Group</th>
<th>Total</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>19050</td>
<td>6</td>
</tr>
<tr>
<td>miRNAs</td>
<td>1864</td>
<td>4</td>
</tr>
<tr>
<td>NT Controls</td>
<td>1000</td>
<td>1</td>
</tr>
</tbody>
</table>

2.2.2 **4D ex vivo lung model of metastasis**

The *ex vivo* 4D lung metastasis model was prepared as previously described (Mishra et al. 2014). Briefly, rat lungs are excised, tied off in specific places to maintain circulatory paths required for the desired flow. The lungs are placed in the bioreactor where they are first decellularized to leave only the extracellular matrix. The bioreactor setup consists of a motor and plastic tubing in order to circulate cell culture media and oxygen throughout the system for the duration of the screen. 25 million cells are seeded into the left lung on Day 0 and the circulation is started after 1 hour in order to allow the cells to attach. The bioreactor runs for 15 days after the primary tumor is seeded, with a brief pause for media change throughout the system every 24 hours.
2.2.3 Sample collection

Sample collection procedure was performed as described previously (Mishra et al. 2014). CTC samples were collected at 6 different time points (Days 9, 11, 12, 13, 14 and 15). All CTC samples were acquired from the used cell culture media that was removed from the system during the daily media changes or at the end of the screen. Cells were pelleted from the used media via centrifugation and live/dead counts were performed with Trypan Blue prior to storage at -20°C. Genomic DNA was extracted from frozen CTC pellets using the DNeasy Blood and Tissue Kit (Qiagen).

A different procedure was used to extract the cells from the lung matrices. Cells were isolated after 15 days from the left lung (Primary) and the right lung (Metastases) to isolate genomic DNA for PCR amplification of integrated genomic sgRNA and sequencing (Figure 2.2). 5mg Liberase TM (Sigma Aldrich) was reconstituted in 2ml non-enzymatic cell dissociation solution. Lung matrices were minced mechanically using scalpel and sterile forceps. Each lung was digested in 2ml of working solution for 2 hours at 37°C with shaking at 150 rpm. The lung matrix and cell slurry were incubated with 50µg of A549 surface antibody (Abcam) for 30 minutes (Figure 2.3).
Figure 2.2. **Sample collection for 4D model.** The tumor is seeded on day 0. Media is changed throughout the system every day. Starting on Day 9, the media collected is spun down and cells are collected. On Day 15, both left and right lungs are harvested.
Figure 2.3. **Harvest of cells from lung matrices.** Lung matrices are minced with scissors, digested in Liberase, and A549 cells are pulled down with an antibody that binds to A549 surface protein.

### 2.2.4 Library Preparation & Sequencing

Pre-sequencing quality control assays were performed following PCR amplification and gel extraction of the fragments to first confirm we had the fragment size expected from reaction and that our DNA preparation reagents and procedure were functional. Following these initial checks, performing fragment analysis prior to sequencing allows for determination of quality and purity of the desired band from extraction prior to sequencing and gives precise DNA concentration to calculate dilution and volume to submit to the core.

### 2.2.5 Input Library Sequencing

The lentiviral region was amplified from gDNA samples using Phusion High-Fidelity PCR polymerase. The PCR primers used are in Table 2.2 and the thermocycler settings are in Table 2.3. The PCR yielded a 288bp-long segment that was gel extracted using the Qiagen Gel Extraction kit. Prior to ion torrent sequencing, a quality control capillary electrophoresis step was carried out. DNA was submitted to MD Anderson’s Institute for Personalized Cancer Therapy (IPTC) Sequencing Core in a 50µL volume at a concentration of 25ng/µL and
sequenced using the Ion Torrent platform. Reads generated were 200-300 basepairs in length. Sequence data was generated in FASTQ file format.

Table 2.2. Primer sequences for PCR1 and PCR2.

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FWD1</td>
<td>AATGGACTATCATATGCTTACCCTTTGAAAGTATTTCG</td>
</tr>
<tr>
<td>REV1</td>
<td>CTTTGGTTTGTATGTCTGTGCTATTATGTCTATATTCTTTCC</td>
</tr>
<tr>
<td>FWD2</td>
<td>AATGATACGGCGACCACCGAGATCTACACTCTTTTCCATTACACGA CGCTCTTCCGATCT (8bp barcode) tcttgtggaaaggacgaacccg</td>
</tr>
<tr>
<td>REV2</td>
<td>CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT tctactattctttcccctgcactgt</td>
</tr>
</tbody>
</table>

Table 2.3. **Thermocycler settings for PCR1 and PCR2 reactions.** PCR1 used 18 cycles of amplification, while PCR2 used 24 cycles (not shown in table).

<table>
<thead>
<tr>
<th>TEMP (°C.)</th>
<th>TIME</th>
<th>CYCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>30 sec</td>
<td>1</td>
</tr>
<tr>
<td>98</td>
<td>10 sec</td>
<td>goto</td>
</tr>
<tr>
<td>60</td>
<td>10 sec</td>
<td>step2</td>
</tr>
<tr>
<td>72</td>
<td>15 sec</td>
<td>18X</td>
</tr>
<tr>
<td>72</td>
<td>5 min</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>forever</td>
<td>1</td>
</tr>
</tbody>
</table>

2.2.6 **Sequence Processing and Alignment**

Cutadapt ([Martin 2011](#)) was used for initial quantification of reads corresponding to the 5’ lentiviral sequence and trimming of the barcodes as well as separation of the sequences into different files. Input library sequences were analyzed by generating read counts for each guide sequence using MaGeCK alignment algorithm ([Li et al. 2014](#)). Input library quality was
assessed by visualizing the distribution of the library and determining the proportion of total guides and genes covered.

2.3 RESULTS

2.3.1 Library Preparation & Pooled Sequencing

The screen consists of 3 experimental replicates of 8 samples each (primary lung, cells isolated from circulating media on 6 different days, and a metastasis sample), which amounts to 24 samples total. In order to perform pooled sequencing, where all samples are sequenced together, two rounds of PCR were performed and designated PCR1 and PCR2 (schematic in Figure 2.4). In PCR1, all samples plus the input were amplified individually using the same primers against the lentiviral region in the first round of PCR amplification. PCR1 is performed with primers matching the lentiviral vector, which will amplify integrated fragments containing the sgRNAs from our library. This is expected to generate a 288bp fragment. In PCR2, the product of PCR1 was amplified using forward primers with unique 8bp barcodes assigned to each sample. The PCR1 and PCR2 rounds of amplification were performed using PCR was performed in two steps using Herculase II Fusion DNA Polymerase (Agilent). Amplification was carried out with 18 cycles for the first PCR and 24 cycles for the second PCR (Table 2.3). PCR2 products were gel extracted using the NEB gel extraction kit. Capillary electrophoresis was performed for fragment analysis of each individua PCR product. Sample concentrations were measured using Qubit and pooled based on these results. Samples were sequenced using MD Anderson Cancer Center’s Sequencing & Microarray Facility. All samples were pooled all samples to run on 1 lane on Illumina NextSeq 2500. Resulting amplicons from the second PCR were gel extracted, quantified, pooled and sequenced using
the NextSeq (Illumina). Figure 2.5 is a schematic describing the Illumina sequencing process and region of the amplicon.

Figure 2.4. **PCR amplification and barcoding for pooled library sequencing.** gDNA from cells containing the lentiviral insert (top). PCR with lentiviral primers flanking region with sgRNA. All samples use the same forward and reverse primer sequences for PCR1 (lentiCRISPR_FWD1 and lentiCRISPR_REV1). PCR2 uses FWD primers that are the same except for unique 8bp “barcode” sequence just 5’ to the fragment end in order to individually identify each sample. The same REV2 primer is used for all samples. Once PCR2 is done for
all samples, the fragment is purified by extracting it from a size gel, and the DNA concentration of each is measured precisely using a sensitive QuBit assay.

![Illumina sequencing schematic](image)

Figure 2.5. **Illumina sequencing schematic.** In order to be sequenced by the Illumina platform, the region must contain the Illumina 5’ index and 3’ ends.

### 2.3.2 Fragment analysis shows expected product size

Pre-sequencing quality control assays were performed following PCR amplification and gel extraction of the fragments to first confirm we had the fragment size expected from reaction and that our DNA preparation reagents and procedure were functional. Following these initial checks, performing fragment analysis prior to sequencing allows for determination of quality and purity of the desired band from extraction prior to sequencing and gives precise DNA concentration to calculate dilution and volume to submit to the core. The first submission of this fragment showed some high-mass DNA contamination, likely gDNA contamination (Figure 2.6). Second submission of a careful gel-extracted fragment showed the fragment without gDNA contamination that would eat reads. PCR1 is performed with primers matching the lentiviral vector, which will amplify integrated fragments containing the sgRNAs from our
library. This is expected to generate a 288 bp fragment (Figure 2.7). Fragment analysis shows expected product size and informs gel extraction procedure decisions.
Figure 2.6. Input gDNA library in A549 cells. gDNA was isolated and run through the fragment analyzer to reveal 2 peaks, one is close to the correct size and one is larger representing a primer dimer. Step 3 was performed with the Mageck algorithm.
2.3.3 Sequence Processing

Sequence data was generated in FASTQ file format and downloaded from the MD Anderson Cancer Center’s data server.

The raw sequencing data was processed using the following steps:

1. Sequences were filtered using the universal 5’ lentiviral region and a quality threshold.

2. Sequences were placed in separate files based on individual barcode.
3. Individual reads counted based on guide sequence.

The sequence processing steps for the input sequencing is shown in Figure 2.8. We found that about 40% of the raw reads passed filtering and quality control steps and had the lentiviral region.

**Figure 2.8. Sequence processing and filtering steps.** Steps 1-2 were performed with the standalone algorithm Cutadapt on Mac OS-X. Step 3 was performed with the Mageck algorithm.

2.3.4 Input Library Sequencing
In order to determine whether the library successfully integrated as expected into the A549s and optimize PCR and sequencing parameters, the Input library was PCR amplified. Input gDNA was sequenced using the Ion Torrent Ion Proton platform. The sequencing output yielded 3,609,713 reads. The sequence processing pipeline begins by using Cutadapt (Martin 2011) (Figure 2.9 top), which locates and trims read data based on the presence of a specified sequence. The lentiCRISPRrev1 primer in reverse complement order should be universal to all of the constructs. 1,965,260 (54.4%) of the reads met sequence and quality control criteria, and these were trimmed and output to a new file (output.fastq). The new file of trimmed reads is then aligned with the GeCKOv2 library using MaGeCK (Li et al. 2014) (Figure 2.9 bottom). The goal of this computational processing step is to end up with a new file containing a table of read counts.
Figure 2.9. **Code used to trim and align the Input library sequences.** Cutadapt was used to filter sequences based on quality and presence of 5’ region (top) and MaGeCK was used to align the trimmed reads to the GeCKOv2 library (bottom).

Sequencing with Ion Torrent showed that the input library contained about 80K unique guides and that the 98.7% of the GeCKOv2 library guide sequences were recovered from the input cells (Table 2.4). Most of the guide sequences in the library were present. These data show that the basic design and execution of this experiment was successful. The PCR product size and 5’ lentiviral sequence indicated that the lentiviral vector had successfully integrated into the cells. Guide distribution demonstrated enough unique species and gene representation
to meet quality control standards and continue with pooled sequencing of remaining samples using a second round of barcoding PCR.

Table 2.4. Recovery of GeCKOv2 library guide sequences from sequencing of the input cells using Ion Torrent.

<table>
<thead>
<tr>
<th>Target</th>
<th>In Library</th>
<th>Recovered</th>
<th>Percent Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>18992</td>
<td>18958</td>
<td>100</td>
</tr>
<tr>
<td>miRNA</td>
<td>1812</td>
<td>1700</td>
<td>94</td>
</tr>
<tr>
<td>Control</td>
<td>1000</td>
<td>884</td>
<td>88</td>
</tr>
<tr>
<td>Total</td>
<td>21804</td>
<td>21542</td>
<td>99</td>
</tr>
</tbody>
</table>

2.4 POOLED SCREEN SEQUENCE DATA SUMMARY

The pilot sequencing data indicated that preparation and analysis techniques could be used for pooled sequencing the remaining samples from the screen. Samples were PCR amplified using the original forward and reverse primers used for the test input step (now called PCR1). A second round of PCR was performed (PCR2) where all samples use the same forward primer and a unique reverse primer with an 8bp long barcode unique to each sample. Samples were pooled and sequenced using the Illumina NextSeq 550 platform to generate 75bp-long reads.

Sequencing data output revealed that the barcode sequence in the primers used to multiplex all samples was not in the location recognized by the Illumina equipment by default, so the assembled the binary files were re-demultiplexed manually. Raw output was 110M 75bp single end reads. The sequences were demultiplexed using Cutadapt and trimmed, then separated into
different files (one per sample). 77,541,167 reads aligned to guide sequences in the library (71%).

2.4.1 Sequence Analysis Demultiplexing & FASTQ Analysis

Raw output was 110 million 75bp single end reads. 77,541,167 reads aligned to one of the barcodes in the library (~71%). The barcode sequence in the primers used to multiplex all samples was not in the location recognized by the Illumina equipment by default, so the assembled the binary files were re-demultiplexed using Cutadapt.

2.4.2 Processing of sequencing data

Code to process sequencing data was written to perform the following steps in order.

1. Sequences that do not meet the following criteria are discarded: quality score >20 and contain all or part of the 5’ lentiviral sequence.

2. Removes (trims) the 5’ and 3’ primer sequences flanking the 20bp guide sequence.

3. Creates an output file containing only the 20bp guide sequences meeting criteria in 1.

2.5 DISCUSSION

In this chapter, I performed PCR amplification and sample preparation for next generation sequencing of the genomic DNA isolated from the input library cells.

Once I confirmed guide coverage and representation, I performed a second round of PCR where we barcoded individual samples with unique 5’ flanking ends.
Finally, I developed a computational pipeline to filter raw sequence output by quality and 5’ lentiviral sequence, trim adapters and demultiplex the samples into different files by sample barcode sequence.

Fragment size and purity was a big issue with this sequencing project, and it took several submissions in order to get the isolated band of the proper size from the gel extraction without incorrectly-sized fragments in the mixture. With the correctly-sized band, I often got a slightly larger band that was difficult to separate the target fragment from. Because the primers were such different melting temperatures and the in silico analysis predicted a primer dimer with a high free energy (Table 2.5), I hypothesized that there would be additional amplification at the end of the fragment. I had to use a permissive annealing temperature in the PCR, and the extra length might be from the primer dimer formation amplified at the end. This appeared to be the case once the fragment was sequenced.

Table 2.5. Specifications of LentiCRISPR primers used to amplify GeCKOv2 library.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Lenticrispr_fwd1</th>
<th>Lenticrispr_rev1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (bp)</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>GC Content</td>
<td>34.10%</td>
<td>53.33%</td>
</tr>
<tr>
<td>Melting Temp</td>
<td>72.8 ºC</td>
<td>86.8 ºC</td>
</tr>
</tbody>
</table>

Production of the appropriately sized fragment by PCR with lentiviral region-specific primer regions indicates that the transduction of the lentiviral vector into A549 cells was successful. The fragment was able to be isolated from all samples, indicating stability of this region in the A549 genome over a period of several weeks.

Design of the processing pipeline was not as straightforward as originally expected for the task of demultiplexing, applying a quality filter and trimming the sequences. The Cutadapt
algorithm had several potential settings, such as anchored or non-anchored ends, that changed the output and were difficult for me to differentiate between. Apparently with a low-diversity amplicon library (rather than shotgun genomic sequencing), either anchored or non-anchored ends can be used. There was also some issue with the amount of errors allowed in the guide sequences. The default for the program is 10%, which means that 2 bases per 20bp guide can be incorrect and still count as a read for a given guide. An important test to perform that I did not think of at the time would be to check whether the library was designed with more than 2 base differences between each guide.
3. ANALYSIS OF SCREEN DATA TO IDENTIFY GENES REGULATING METASTASIS

3.1 INTRODUCTION

The main purpose of this chapter was to perform analyses to understand how the statistically significant hits from our screen compared to available data and what is currently known about genes involved in lung cancer metastasis. We first needed to determine the utility and relevance of our novel methodology by comparing significant hits from the screen to previously implicated genetic regulators of lung cancer metastasis. We then wanted to choose clinically relevant targets to validate experimentally in the lab.

Dropout and enriched gene groups were defined by guide and read count thresholds and each group was analyzed for regulatory network connectivity. A portion of the genes from each group formed genetic subnetworks with several genes with well-established cancer roles and identified putative novel regulators of metastasis. Low-resolution data output due to inadequate coverage can be increased in power by supplementation with clinical data and network analysis. The main objective of this chapter is to propose a candidate list of gene knockouts that are significant in this screen.

3.1.1 Library Dynamics and Expected Trends

The purpose of this section is to discuss overall theory and dynamics of the model in order to understand the goal of these analyses and provide a framework for interpreting our results. This section provides the background I used to choose an appropriate readout of the screen in order to understand the earliest events in lung cancer.
The schematic overview of guide population dynamics (Figure 3.1) demonstrates the importance of understanding the conceptual framework in forming expectations useful in comparing data generated and interpretation of the results. We described methods for the library generation and performing the screen using the lung model in Chapter 2. To review briefly, the input cells were generated by transduction of lentiviral particles followed by selection with puromycin for 1 week in cell culture plates. At this time, the input cells were used for seeding of primary tumors in 3 replicates of the lung model. Cells seeded into the left lung grow in 3D for 15 days before they are collected with the metastasis sample upon completion of the screen.

Figure 3.1. **Schematic overview of progressive loss of unique guides during screen.** Guide diversity is highest in the input cells and should decrease as cells move from primary tumor to the metastases in the distant lung. Reproduced from Chen S, Sanjana NE, Zheng K, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell. 2015;160(6):1246-1260. doi:10.1016/j.cell.2015.02.038 with permission from the Copyright Clearance Center.

Summary of predictions based on theory of model

1. Library diversity should be as follows: input > primary > metastasis >> CTC
2. Library inequality, or skew, should increase as pools become less diverse and selected guides become enriched: input < primary < metastasis

3. The total number of guides should decrease faster than the total number of genes, due to redundancy created by each gene being covered by multiple guides.

Library diversity is the total number of unique guide sequences represented in a given sample. Input cells grown in 2D culture should contain the highest library diversity at this initial step. As cells move from primary tumor to the distant lung, the number of unique guides should decrease due to dropout (Figure 3.1).

Assuming similar cell counts collected from each lung, there would be increased skew of the guide distribution in the primary tumor and finally the metastasis. Previous studies using the GeCKOv2 library in a mouse model of metastasis have found that samples are most similar in guide distribution and total guide number among conditions. For example, primary tumors had more total guides, higher similarity in guide constitution and lower skew ratio than the metastases. Progression of tumor over time and distance was associated with guide dropout and increased population proportion of high representation guides (Chen et al. 2015).

3.2 METHODS

In Chapter 2, raw sequencing reads in FASTQ file formats were processed using Cutadapt (Martin 2011). Reads that did not meet the quality threshold and contain all or part of the 5’ lentiviral region were removed. Reads were then separated into individual files based on the 8bp barcode used to amplify each individual screen sample (25 total, including input). Reads were aligned in MAGeCK (Li et al. 2014, 2015). A table of read counts corresponding to rows
containing each guide ID (unique) and gene (repeated per number of guide sequences covered in the library) and columns containing the screen sample ID was generated.

### 3.2.1 Read Count Data Processing and Preparation

Tabular read count data were processed and analyzed using R5 and the environment R Studio (Haque et al. 2017). Count data were read into the software and prepare data for statistical analyses. Similarity matrices and heatmaps were generated for the screening samples.

### 3.2.2 Network Analysis

Network analysis was performed using NetWalker (https://netwalkersuite.org/) database assembly function. A list of genes was input into the software and “generate network.”

### 3.2.3 Database Analysis

The Human Protein Atlas was used to determine the normal tissue and cell line expression of SPI1. “SPI1” was used as the search term. Figures were generated using normal protein expression from different tissues and expression daof SPI1 at the mRNA level in normal and cancerous lung (Uhlen et al. 2017; Uhlén et al. 2015). The Cancer Genome Atlas (TCGA) was used to create a Kaplan-Meier survival curve in lung cancer of SPI1 high and low expression.

### 3.3 RESULTS

#### 3.3.1 Input Cells Yielded High Library Coverage

Next generation sequencing (NGS) of integrated lentiviral region isolated from the gDNA of the input cells was performed. Sequencing with Ion Torrent showed that the input library contained about 80K unique guides and that the 98.7% of the GeCKOv2 library guide
sequences were recovered from the input cells (Table 3.2). Most of the guide sequences in the library were present. These data show that the basic design and execution of this experiment was successful. The PCR product size and 5’ lentiviral sequence indicated that the lentiviral vector had successfully integrated into the cells. Guide distribution demonstrated enough unique species and gene representation to meet quality control standards. The results indicated that our preparation and analysis techniques were adequate to continue with pooled sequencing of remaining samples using a second round of barcoding PCR.

Table 3.1. **Number of recovered guides for the genes, miRNAs and controls groups from the input library sequencing.**

<table>
<thead>
<tr>
<th>Target</th>
<th>Library Total</th>
<th>Recovered</th>
<th>% Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>18,992</td>
<td>18,958</td>
<td>100%</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,812</td>
<td>1,700</td>
<td>94%</td>
</tr>
<tr>
<td>Control</td>
<td>1,000</td>
<td>884</td>
<td>88%</td>
</tr>
<tr>
<td>Total</td>
<td>21,804</td>
<td>21,542</td>
<td>99%</td>
</tr>
</tbody>
</table>

Table 3.2. **Number of recovered guides for the genes, miRNAs and controls groups from the input library sequencing.**

<table>
<thead>
<tr>
<th>Target</th>
<th>Library Total</th>
<th>Recovered</th>
<th>% Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>18,992</td>
<td>18,958</td>
<td>100%</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,812</td>
<td>1,700</td>
<td>94%</td>
</tr>
<tr>
<td>Control</td>
<td>1,000</td>
<td>884</td>
<td>88%</td>
</tr>
<tr>
<td>Total</td>
<td>21,804</td>
<td>21,542</td>
<td>99%</td>
</tr>
</tbody>
</table>

3.3.2  **Exploratory Data Analysis of Pooled Samples**
To determine how to analyze the pooled samples, I performed some exploratory analyses of the pooled sequencing dataset. I wanted to determine the overall numbers of reads, guides and genes in the screen and asked if they are in range for what I expected for the experiment and accepted standards for this type of screen. I also wanted to determine the distribution of reads across the screen. Finally, I wanted to determine how the samples compared to each other in order to determine how I could compare samples, conditions and replicates in order to learn about metastasis in lung cancer.

3.3.2.1 Comparison of Input Samples

In order to get a sense of how much of the actual information was captured, I compared the data from the two input samples sequenced. Input comparisons give a unique glimpse idea of how much was missed in the sequencing of the pooled samples. Comparing these inputs also provides valuable information about the effect of the sample preparation on the sequencing output. The original Ion Torrent sequencing pilot (IPTC) was compared to the pooled input (NSEQ). It is important to state that these two datasets were derived from the same tube of gDNA. Therefore, the differences between them are due to differences in preparation, amount of sample and proportion of sequencer, and sequencing platform.

First, I compared the numbers from the input libraries (Table 3.3). We can see that from the first to the second sequencing, half of the guides are lost. Both the IPTC and NSEQ samples had 43,650 guides in common. While nearly half of the guides were lost, however, most of genes in the library were retained. This is due to the redundancy in the library created by the fact that each gene is covered by multiple guides.
Table 3.3. Guide numbers. Read counts and library percentages covered by the IPTC (first round) and NSEQ (second round) sequencing, respectively.

<table>
<thead>
<tr>
<th>Sequencing Platform</th>
<th># Guides with &gt;0 reads</th>
<th>% Guides Covered</th>
<th>% Genes covered</th>
<th># Total Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPTC</td>
<td>87585</td>
<td>90%</td>
<td>99%</td>
<td>929609</td>
</tr>
<tr>
<td>NSEQ</td>
<td>46462</td>
<td>40%</td>
<td>90%</td>
<td>6967121</td>
</tr>
</tbody>
</table>

One interesting finding from the numbers in Table 1 is that despite the low coverage in the second sample coincides with nearly an order of magnitude increase in total read counts. This means that there is higher skew between the guides with the highest and lowest number of read counts. We can demonstrate this as an inequality using a Lorenz curve (Figure 3.2). This curve is a representation of the deviation from perfect equality (where all guides would have the same read count). The area between the straight diagonal line and the curve signifies the total inequality of the read distribution. The area of the curve between the diagonal line and the red line (IPTC) is smaller than the area under the blue line (NSEQ). Therefore, the first round of sequencing is less unequal in distribution than the second performed with the pooled samples. This result makes sense if we look at the total read count in comparison to the total number of genes and guides covered (Table 3.3).
Figure 3.2. **Lorenz curve showing the inequality between the guides with the highest and lowest read counts**

The red line represents the IPTC sequencing sample, while the blue line represents the NSEQ sample.

I compared the distribution of the number of guides per gene between the IPTC and NSEQ samples (Figure 3.3). This figure shows the number of genes for each guide count hit with at least 1 read count in each sample. In the IPTC sequencing, the average guide count per gene was greater than 4 guides per gene. In the NSEQ sample, the average guide count per gene was around 2. We can see that the whole distribution is shifted to the left from the IPTC to the NSEQ, and there are many more zeros in the NSEQ. What this means is that many guides were lost due to the sampling for the second round of PCR and the additional reads generated by the skew in the library.
Figure 3.3. Number of genes per guide count for the input sequencing round using ion torrent (IPTC, left) and the pooled sequences (NSEQ, right).

Next, I compared the distribution of read counts per guide in order to determine whether there was similarity in the distribution based on read count (Figure 3.4). In this figure, we used log2 normalization of read counts. We can see that the guides with the lowest read counts in the IPTC sample are more likely to be absent in the NSEQ sample.
Figure 3.4. Distribution of log2 read counts for the input sequencing (IPTC, purple) and the pooled input (NSEQ, yellow).

The purpose of the input analyses was to determine how the second round of PCR and change in sequencing platform affected the results we got from the whole screen pooled sequencing data in order to understand how we might be able to analyze, use and interpret the data. We found that over half of the guides were lost in the second round of sequencing. We also found that the second round of PCR generated higher inequality of sequencing reads and therefore higher skew in the count data. We were able to show what was lost and what was retained in the second pooled sequencing sample. Therefore, we have a better idea of how we look at all samples and see what we can and cannot find looking at this dataset.

Since we cannot distinguish between random loss due to low sequencing coverage and real positive selection, we will not analyze these data for dropout. The input data comparisons show
very clearly that just because a guide or gene is not present does not mean it’s not there. Instead, we will focus on what is present and discuss genes remaining as those of interest.

3.3.2.2 Comparison of Pooled Samples

To evaluate how we can compare and utilize the screening data, measurement of similarity index is made of all pairwise combinations of samples from the screen. Based on what we learned in comparing the input samples above, we expect low representation of the library in any given sample. We also expect low similarity between samples due to sample loss and lower representation. Euclidean distance is the linear distance between two samples and is calculated by taking the sum of all individual differences between the log2 read counts for each guide. Similarity of samples is a measure of how related two samples are to each other. The lower the distance value between two samples, the more similar the samples are and the darker the color. Higher distance values are considered more dissimilar and are lighter (Figure 3.5A). The Pearson correlation coefficient is a measure of strength of association between two samples. Guides are ranked in descending order by read count and the correlation coefficient is represented in a heatmap where lighter colors indicate correlation (Figure 3.5B).

I attempted several methods of comparing samples using various similarity indices and heatmaps and achieved markedly different sample groupings. Sample groupings are highly dependent on method used. Some measurement methods are more appropriate for these data than others. More broadly, it appears that experimental replicates (labelled A, B and C) grouped more closely with each other than samples from the same conditions. Similarity within replicates would be expected due to divergence in library distribution over time following the start of independent growth. Ultimately, we found that many of the samples were too low in
coverage and high in skew to assign groupings or conclusions with any confidence. There was no way to clearly rationalize use of one similarity measurement method over the others, and I ultimately did not find this comparison informative to my analysis and use of the samples.

Figure 3.5. **Comparing two methods for similarity analysis across all screening samples.**
A. Euclidean distance heatmap of the input library and all samples. B. Pearson correlation heatmap of log2 gene sum read counts >100.

The CTCs did not group with each other very well, nor did they show much consistency in similarity across the screen. The CTCs also tended to have on average much higher guide diversity (more total guides covered) than the primary or metastasis samples (Figure 3.6). This is a very curious finding, given the fact that the CTCs collected each day were in much smaller numbers than the final number of cells in the primary and metastasis samples. Therefore, there should be many more guides represented in the primary and metastasis samples.
Given the problem of guides integrated in dead cells likely contributing to the data retrieved from the CTCs and the totally different cell harvesting technique used for the primary and metastatic tumors from the lung matrices, I determined that the CTCs rendered data that had limited interpretability and could not be compared to the primary and metastasis samples. Therefore, we decided to use the primary and metastasis samples only, finding that replicates B and C were the most similar to each other (Figure 3.7).
Figure 3.7. **Euclidean distance heatmap of the primary and metastasis samples.** A distance matrix generated as in Figure 5 for all screening samples. Replicates are labeled A, B and C. P stands for Primary, and M stands for Metastasis.

Next, I was interested in seeing how many guides and genes the primary and metastasis samples had with each other (Figure 3.8). The top Venn diagrams (A and B) show the number of guides each sample has in common, while the bottom ones (C and D) are genes. Overall, we can see that replicates B and C have the most in common with each other, which is what we saw in Figure 3.7). Predictably, there are more genes in common between the samples than there are guides. Interestingly enough, the primary samples have more guides and genes in common with each other than the metastases do. This result is in line with what I predicted at
the beginning of this study, that metastasis tumors would be more divergent in their library representation than the primaries would be.

Figure 3.8. Venn diagrams created with R package eulerr (Larsson et al. 2020). A and B show guides, while C and D show genes. P stands for Primary, and M stands for Metastasis.

3.3.3 Selection of Gene Candidates

The overall goal of this chapter was to analyze the screen data for the genes most likely involved in metastasis. I made a gene candidate selection algorithm to choose a list of genes to
study furthers. In order to select a list of “hits” from our screen, it’s important to define a list of criteria. Gene candidates are a subset of genes in the original library that is selected with the following criteria: non-arbitrary, reproducible, minimizes the overall size of the list but selects a large enough list to generate a network in the NetWalker algorithm. I thought that maximizing the retention of true positives would be a higher priority than minimizing false positives.

A gene selection algorithm based on the concept of a “desirability function” using several criteria was devised (Lazic 2015). The algorithm chooses genes that were covered by more than one guide in all primary and metastasis samples, then puts this list into the NetWalker algorithm and the final list is what assembles into a network (Figure 3.9). The primary step of the algorithm creates a list of genes that are covered by 2 or more guides. A guide is considered to be present if it has 1 read count. The gene had to be present in all 3 replicates of each the primary and metastasis samples. The secondary step uses different criteria in order to decrease the list to desired size (network analysis). The selectivity of the numeric criteria can be toggled, which results in a network of a size that is large enough to include more than a few components, but small enough that it doesn’t have way too many items to use in subsequent analysis or too computationally cumbersome.
Initially, it was discussed that I should rank order the guides based on read count, and only use genes that had >100 reads. However, we found in the input analysis section that the library coverage was vastly decreased following the second round of PCR and sequencing. Not only was the guide count decreased by half, we also found that there were many more guides with 0 and 1 read count, and the distribution of genes per guide count was shifted to the left. We concluded from this decrease that 1 read count would be sufficient.

3.3.4 Analysis Revealed Genes Previously Implicated in Lung Cancer

We first wanted to see if there was a network that could sum up every sample in the screen. Using the list of 725 genes that were covered by at least 1 guide in all samples, we ranked the
list by number of average guides covered throughout all the samples. This gave us a list of 74
genes, which we put into the NetWalker algorithm. The network that was generated contained
several genes that have been previously implicated in lung cancer, which are highlighted in
orange (Figure 3.10). The network also had a central node, SPI1, which is a transcription factor
that has been implicated in blood cancers but if real, is a novel finding in lung cancer.

Figure 3.10. **Network of genes with top average guide count for all samples in the screen.**
The orange nodes denote genes previously implicated in lung cancer metastasis.

Next, we used our gene selection algorithm from the previous section to generate the
network that came from the genes that were covered by 2 or more guides in the primary and
metastasis samples in all 3 experimental replicates. The network that resulted was similar to
what we saw in Figure 3.10, with SPI1 again being the central node (Figure 3.11).
Figure 3.11. **Network generated by the gene selection algorithm for the Primary and Metastasis samples.**

Because SPI1 came up as the most highly connected node in the network, we wanted to know how many genes in the entire database connected to SPI1. We selected “all known connections” in the NetWalker database algorithm, finding more than 13,000 possible connections (Figure 3.12).
Figure 3.12. **All possible SPI1 network connections in the NetWalker database.** Different types of connections are denoted by different colors. Green lines are directional edges, indicating transcription factor to transcribed gene interactions. Black indicates physical interactions. Pink are miRNA interactions.

In our network analyses, we found that SPI1 was the most highly connected node in both networks generated from the genes present with the highest guide coverage throughout the whole screen and enriched in the primary and metastasis samples. In order to determine how many genes SPI1 theoretically connected to in our entire database of genes, we found about
13,000 connections, with most of them being directional transcription factor to transcribed gene interactions.

### 3.3.5 SPI1 is a Novel Finding in Lung Cancer

We found SPI1 sgRNAs consistently in every sample of our whole-genome screen. Presence of these sgRNAs is a significant finding because of sparsity due to low coverage in the original sequencing data. To determine what tissue SPI1 is expressed in and whether SPI1 is expressed in normal lung tissue, we looked at the Human Protein Atlas (www.proteinatlas.org). We found that SPI1 is expressed in normal lung tissue at the RNA and protein level (Figure 3.13).
Next, we wanted to know whether SPI1 expression was different between normal lung and cancerous lung tissue. I compared the mRNA expression of normal lung tissue to that of cancerous lung tissue and found that SPI1 mRNA expression was decreased more than 3-fold in lung cancer (Figure 3.14).
Finally, we wanted to see if SPI1 low and high expression corresponded to differences in lung cancer survival (Figure 3.15). Patients were separated into low and high SPI1 expression at 24.95 FPKM (units used in Figure 3.14). The resulting survival curve showed that there is a significant difference in survival between patients with low and high SPI1 expression ($p = 0.025$), with low expression having lower survival rates until year 10.
Figure 3.15. **Kaplan-Meier survival curve of SPI1 low and high expression.** Generated using the Human Protein Atlas (http://www.proteinatlas.org survival analysis function (Uhlén et al. 2015)).

3.4 DISCUSSION

In this chapter, I performed a computational analysis of the inputs and screening samples in order to determine the success of our novel library screening method and confirm appropriate guide representation in our input cells. We then examined the guide representation throughout the screen, assessed similarity across the screening samples for comparability and determined the resolution at which we could make conclusions about the presence or absence of guides. Finally, we designed a gene selection algorithm to narrow down a list of top hit gene candidates. We performed network analysis of our gene candidates to narrow down a list of gene candidates that were most likely to be real results based on their connection to each other in a regulatory network.
Our initial input sequencing showed that the majority of the library was covered with nearly 90,000 of the 123,000 guides and over 99% of the genes. This result spurred the decision to perform pooled sequencing of all of the samples in the screen. The pooled samples looked as though they had very few guides and genes in comparison to the original input sequencing. In order to determine how many guides were lost due to low coverage during PCR amplification and sequencing, which are technical artifacts, I compared the input read count data from the original ion torrent sequencing to that which was pooled with the rest of the samples. I found that about half of the guides were lost, while less than 10% of gene coverage was lost. From this, I concluded that the low coverage of the pooled samples was due in large part to technical artifact, low sequencing coverage, and the high skew was due to the second round of PCR amplification.

In order to determine how I might compare different conditions and decide on how to choose a list of gene candidates that I considered “hits” of the screen. I discovered that two different methods of comparison, Euclidean and Spearman rank correlation coefficient, showed two completely different groupings of samples. From this, I concluded that there was no way to rationally compare all of the samples, nor was there any way to decide that one method of comparison was better than the other.

Next, I looked at guide coverage in all samples across the screen, determining the number of unique guides in each sample. I discovered that the CTC samples covered on average larger numbers of unique guides. This was a curious finding, and I looked back at the cell collection methods used when harvesting cells from our model. I found that the CTCs were harvested during media change after spinning down the cells using the centrifuge. The primaries and
metastases, however, were harvested using an A549-specific cell surface antibody after the lung matrices were minced. The difference in total guide representation in these samples can be explained by the different harvesting techniques. Additionally, I found that a significant percentage (10-20%) of the CTCs harvested were not viable according to counts using Trypan Blue. The CTC harvesting technique was developed originally by Mishra et. al. in order to perform RNA sequencing and expression studies. Because RNA degrades so quickly, the dead cell would not contribute data in this situation. However, DNA is more stable, and would contribute to the data collected by DNA sequencing, and it would be impossible to distinguish which cells were alive in the circulation from those that had just died and fallen off the primary tumor or the metastasis and had not yet undergone membrane degradation at the time of collection. I also noticed that the DNA concentrations and total DNA content of the samples differed greatly between the CTCs and those harvested from the lungs, with an average concentration of 1.5ng/µL and 0.5ng/µL, respectively, with the same total volume. I concluded from this information that a large number of cells were lost from the primary and metastasis samples, and the low amount of guide and gene diversity in the primary and metastasis was due to the harvesting technique and not representative of the total number of cells in that compartment of the model. Given this information, I decided that there was no reasonable way to use the CTCs in my analysis and deemed them incomparable to the primary and metastasis samples.

After deciding to look at just the primary and metastasis samples, I determined that replicates B and C were more similar to each other than they were to A in a Spearman rank correlation coefficient analysis. By using a Venn diagram comparison tool to look at the number of guides and genes the primary and metastasis samples had with each other, I learned
that the primary samples had more in common with each other than the metastasis did with each other. This result is consistent with what we predicted based on our model described in the background section of this chapter, as well as previous studies that show that the farther away from the initial library integration the tumors get, the more the guide representation will diverge.

Overall, the dataset I have described is one that has low coverage due to both low DNA content of the samples, incomplete sequencing coverage, high skew due to PCR amplification, and incomparable sample conditions due to isolation techniques. Given the limitations introduced by these issues, the next problem we had to solve was how to choose a set of gene candidates from our screen. The gene selection algorithm that I described in this chapter uses 2 different methods of selection: the presence of 2 or more guides per gene, and a network analysis to choose genes that have a higher chance of interacting with each other (and therefore being real hits). Our method differs from the most commonly used method to determine which genes are “significant hits,” which is a gene set enrichment analysis using log2 fold change to choose genes that have been enriched from the primary to the metastasis. This type of analysis could not be done on our data given the low coverage, guides with lower read counts are lost often at random and the skew introduced by PCR amplification, which causes guides with a higher proportion of read counts to have inflated rank due to artifact. Most importantly, enrichment analysis would have been inappropriate for our analysis given the overall goal of this screen: to determine genes that when knocked out, regulate lung cancer metastasis.
4. EXPERIMENTAL VALIDATION OF SPI1 SCREEN CANDIDATE IN VITRO

4.1 INTRODUCTION

In Chapter 3, potential gene candidates were identified from the whole-genome screen in the 4D lung model of metastasis and pooled sequencing performed in Chapter 2. From a starting list of more than 21,000 genes, I created a gene enrichment algorithm and network analyses to narrow down a set of genes that could be reasonably validated in the lab. Gene selection criteria and network analyses independently implicated the transcription factor SPI1, which would be a novel finding in lung cancer. In this chapter, I experimentally validated SPI1 knockdown in the lab and investigated whether this perturbation resulted in a phenotype relevant to lung cancer metastasis (Figure 4.1).

Figure 4.1. **Objective 3: In vitro validation of gene candidate SPI1.** The gene of interest was knocked down using siRNA. Gene knockdown was then confirmed using semi-quantitative
RT-PCR. Phenotypic assays were performed on cells to assess effect of perturbation of gene candidate on cell functions related to metastasis.

4.2 METHODS

4.2.1 Cell Culture

A549 cells are maintained in RPMI with 10% Fetal Bovine Serum (FBS) (v/v) without antibiotics at 37°C with 5% CO2. Early passage cells (>p15) were used for experiments. Cells were passaged at ~80% confluence by detaching, counting and seeding at ~10% confluence for passage twice per week. Cells were detached with trypsin EDTA 0.25% by adding 1mL to cover the cells and then removed prior to a 5-minute incubation at 37°C. Trypsin was neutralized with RPMI media with 10% FBS, centrifuged at 400G for 5 minutes, resuspended in 10mL media and counted using a Countess automated cell counter.

4.2.2 siRNA Knockdown

Cells were detached and seeded at ~50% confluence in microplates of appropriate size for each chosen assay (Table 4.1). 18 hours after seeding, cells were transfected with 25nM siRNA (Santa Cruz Biotechnology, Table 4.2). using Lipofectamine3000 using the A549 siRNA knockdown protocol (Invitrogen). Cells were detached after 12h and seeded into one of the assays listed in methods described below.

Table 4.1. Cell numbers and volumes of reagents used for siRNA knockdown experiments.

<table>
<thead>
<tr>
<th>Culture Vessel</th>
<th>Growth Area (cm^2)</th>
<th>Number of Cells Seeded</th>
<th>Growth Medium</th>
<th>Opti-MEM Medium</th>
<th>siRNA (pmol)</th>
<th>Lipofectamine 3000 Reagent (μL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>96-well</td>
<td>0.33</td>
<td>1.98E+03</td>
<td>100 μL</td>
<td>2 x 5 μL</td>
<td>3 pmol</td>
<td>0.3</td>
</tr>
</tbody>
</table>


### 4.2.3 Semi-quantitative RT-PCR

24 hours after siRNA knockdown in 6-well plates (described above), RNA was harvested using TRIzol reagent (Thermo Fisher Scientific). RNA concentration was measured using a NanoDrop Lite (Thermo Fisher Scientific). cDNA synthesis was performed with SuperScript III (Thermo Fisher Scientific) using 5µg RNA. To demonstrate knockdown of SPI1 gene, a nested PCR was performed in 2 sequential steps using 2 sets of nested primers (PU.1 Cat. No. sc-36331 Santa Cruz Biotechnology) with proprietary sequences. We used Accuprime polymerase with the following settings for both rounds of PCR (Table 4.3). We optimized this reaction to yield a band for endogenous expression of SPI1 and not when it was knocked down, which was achieved with 20 cycles of each reaction. PCR reactions were performed on a Thermo Scientific SimpliAmp Thermocycler.

Table 4.3. **PCR thermocycler settings for both PCRs performed.**
<table>
<thead>
<tr>
<th>STEP</th>
<th>TEMP (°C)</th>
<th>TIME</th>
<th>CYCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>94</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>∞</td>
<td>1</td>
</tr>
</tbody>
</table>

### 4.2.4 Cell Growth & Viability

Cells were transfected for 12h and then detached with trypsin EDTA 0.25% and counted using a Countess automated cell counter. Cells are plated at 5000 cells/well in triplicate in a separate 96 well plate for each time point measured (3-5d). At the indicated time points, the media was removed, and the cells are washed with PBS containing calcium and magnesium, and 100µL PrestoBlue viability reagent (Thermo Fisher Scientific) was added to the wells with cells as well as empty wells for a background measurement. Plates were incubated at 37°C for 1 hour and fluorescence intensity was measured using a Tecan M1000 fluorescence plate reader and i-Control software (Tecan). Data were processed using the following steps:

1. Mean and standard deviation of the 6 fluorescence intensity measurements per condition for each day.

2. The mean background fluorescence on each day was subtracted from the mean for each condition.

3. Each condition was normalized to its own day 1 measurement, making the mean of all day 1 measurements equal to 1, and subsequent timepoint measurements represent fold change from their first measurement.

4. Statistical significance was determined using two-way analysis of variance (ANOVA).
4.2.5 CyQUANT Cell Count Curves

Cells were transfected for 12h and then detached with trypsin EDTA 0.25% and counted using a Countess automated cell counter. Cells were quantified using the CyQuant cell proliferation assay kit (C35006, Thermo Fisher Scientific). A standard curve was created by generating a dilution series with cell numbers ranging from 50 to 50,000 in 200µL volumes in 96-well plates. Plates were incubated at 37°C for 1 hour and fluorescence of the samples was measured at 480nm excitation and 520nm emission. Fluorescence intensity was measured using a Tecan M1000 fluorescence plate reader and i-Control software (Tecan).

4.2.6 Anchorage Independence Assay

Anchorage independence assays were performed in 96-well plates. Viable cells (1 × 10^3) were mixed in the RPMI medium in 0.5% agarose, plated on the top of 1% agarose base, and covered with RPMI containing 10% FBS. Cultures were maintained at 37°C in the incubator for 15 days. New medium was added every 3 days, and the number of anchorage-independent colonies was counted by choosing the z-axis with the most visible colonies in focus for each well at 10X magnification.

4.2.7 Invasion Assay

Cells were transfected as indicated above. Matrigel-coated invasion chambers (BD Biosciences) were used according to the manufacturer’s instructions with media containing 10% (v/v) fetal bovine serum as a chemoattractant. After 24h, the non-invading cells were removed with a sterile swab, and the remaining cells were fixed and stained in 100% methanol for 2 minutes and 10µM Propidium Iodide (in PBS) for 10 minutes. Inserts were air dried and visualized using fluorescence microscopy.
4.3 RESULTS

4.3.1 Expression of SPI1 in A549 Cells

In order to justify investigation of SPI1 in A549, we had to demonstrate endogenous expression of this gene. We performed Western blot (data not shown) to probe for ectopic expression of SPI1 following transient expression, but this technique was not sensitive enough to demonstrate endogenous expression despite several attempts with progressive increases in antibody concentration in incubation time.

While SPI1 was detectable in clinical samples of normal and cancerous lung at the RNA and protein level, there was no evidence in the protein atlas or other public databases of gene or protein expression of SPI1 in A549 cells. These data looked at whole-genome expression using microarray, RNA-seq and Western blots. Transcription factors are notorious for being present in the cell at low abundance, as relatively few molecules are required for the transcription of many genes under their control. A whole body of literature exists on methodology related to isolation and purification of transcription factors (Gadgil et al. 2001; Jarrett 1993).

To detect endogenously expressed SPI1, I isolated RNA following transfection of siRNA against SPI1. Using a semi-quantitative reverse transcriptase PCR (RT-PCR) with nested primers with 20 cycles per round, I was able to detect a baseline SPI1 expression in A549 cells (Figure 4.2).
4.3.2 SPI1 knockdown alters cell attachment but not overall cell number in A549 cells

We observed during early experiments that there appeared to be a higher number of A549s in multi-well plates following transfection with SPI1 siRNA (Figure 4.3). However, quantification of transfected cells presented the challenge of being extremely dependent on methodology. Following relatively minor procedural modifications, cell quantification yielded large discrepancies in results.
Figure 4.3. **Images of A549 cells 72h post siRNA transfection.** Image taken on a Nikon TE1000 inverted microscope at 10X magnification.

4.3.3 **SPI1 knockdown increases number of viable detached A549 cells in transfection plates**

4.3.3.1 **Paradoxical growth rate of A549 after SPI1 knockdown dependent on assay technique**

SPI1 knockdown increases the number of A549s in transfection plates. We observed during early experiments that there appeared to be a higher number of A549s in multi-well plates following transfection with SPI1 siRNA (Figure 4.4). However, quantification of transfected cells presented the challenge of being extremely dependent on methodology.
Given our observations in Figure 4.4 that A549 cell numbers increased or at the very least remained stable following SPI1 siRNA transfection, we decided to do a growth curve using PrestoBlue reagent (Thermo). PrestoBlue is a resazurin-based reagent that gives a linear readout for metabolic rate based on NADP+ conversion to NADPH and is used as a proxy for proliferation rate of cells in culture. In this first run, my procedure was as follows: I seeded the cells on Day 0 and on each measurement day, I removed the existing media, washed the cells using PBS (with Ca2+ and Mg2+), and added media with 1X PrestoBlue reagent. The growth curves based on metabolic rate showed that SPI1 had significantly lower proliferation rates on all days tested (Figure 4.5).
The result in Figure 4.5 did not match what I saw in the transfection plates, which had higher cell counts when I counted them straight out of the plates. Because the SPI1 transfection plates had higher numbers of cells and they were much more easily detached, I decided to modify the protocol to keep the old media and add 10X PrestoBlue reagent directly to the existing media without changing it. While the differences between SPI1 knockdown cells and the controls may not be as large, these results still implied a lower growth or viability rate (Figure 4.5).

Next I used a cell quantification reagent called CyQuant, which gives a fluorescence readout that is linear to cell number. This reagent requires cells be in PBS, as serum interferes with fluorescence readout. The protocol for adherent cells requires removal, washing and addition of PBS with reagent, and suspension cells can be centrifuged and responded. The problem with these procedures is that they cannot be done simultaneously, eg. cells must either be treated as suspension or adherent cells, and each protocol introduces error. I found that very careful removal of media manually using a pipette tip was the most reliable way to accurately represent cell numbers. Unlike the curves performed with PrestoBlue, there was no significant difference between numbers of SPI1 knockdown cells to the controls (Figure 4.6).
Figure 4.6. *CyQuant* growth curves show no significant difference in cells treated with SPI1 siRNA compared to control groups.

4.3.4 SPI1 knockdown decreases cell attachment

Following siRNA transfection, I noticed that SPI1 knockdown cells were less adherent than the control groups, with the number of viable floating cells increasing over 2-3 days post-transfection (Figure 4.7).
When detaching cells from the transfection plates for assays, I observed that SPI1 knockdown cells were extremely sensitive to detachment by Trypsin-EDTA. The SPI1 group detached from the plate in $\frac{1}{2}$ the time required for other conditions. The proportionality in timing was consistent at 2-4 days post-transfection and independent of plate well size. SPI1 appears to increase cell sensitivity to mechanical and chemical detachment.

### 4.3.5 Anchorage Independence Assays

#### 4.3.5.1 Assay Development & Optimization

The soft agar assay is used to assess anchorage independent growth capability in adherent cells. A pilot experiment to determine the cell number to plate was performed (Figure 4.8 and 4.9). I found that colony count after 16 days of growth is linear to number of cells originally plated (Figure 4.9 and 4.10). There was no significant difference between wells A-B and C-D. Well E is the negative control.
Figure 4.8. **Optimization of anchorage independence assay.** Serial dilution of cells used to determine appropriate seeding density. Wells are labelled A-E going from left to right as referenced in other figures.
Figure 4.9. **Colony count vs. number of cells seeded.** Optimization of anchorage independence assay.
Figure 4.10. **Mean colony count and mean colony area per well.** Wells go from A to E, with the most densely seeded cells in A, decreasing by half to well D, and E as a negative control well.

4.3.5.2 **Anchorage independence increased by SPI1 knockdown.**

SPI1 knockdown did not appear to increase the colony number in the anchorage-dependence assay (Figure 4.11). However, SPI1 knockdown cells form significantly larger colonies than controls. These colonies also appear to be more densely packed (Figure 4.12). Final colony size appears to be dependent on seeding density in the siSPI1 condition only. Control treatment cells grow to similar size in the tested seeding density range.
Figure 4.11. Colony size (right) and colony count (left) for anchorage independence assay of siRNA knockdowns.

Figure 4.12. Representative figure of anchorage-independence assay following siRNA transfection.

4.3.6 SPI1 knockdown does not affect cellular migration or invasion

Please note that these data are preliminary until completion of 3rd replicate.
This assay was performed using Matrigel-coated chambers and PES (uncoated) controls in a 24-well plate for 24 hours. There was no difference in number of invaded cells in the SPI1 knockdown as compared to the controls (data not shown).

4.4 DISCUSSION

In Chapter 3, I narrowed down the list of gene candidates to one gene of particular interest that was highly connected in the network and would be a novel find in lung cancer, SPI1. In this chapter, I performed experimental validation of SPI1.

Since there were no publicly available data on I started by confirming that SPI1 was expressed in A549s. I was unable to see SPI1 protein expression using a Western Blot. This is because SPI1 is likely expressed at a very low level, below the limit of detection for this assay. I was able to confirm SPI1 mRNA expression using a RT-PCR assay, which I also used to confirm that the SPI1 expression was knocked down following siRNA transfection.

In our computational analyses of the sequencing data of cells retrieved from the 4D model, we found that SPI1 guides were present in every single sample. If a guide is enriched in the distant lung (the metastasis samples), we can conclude that the knocked out gene was either highly proliferative and had many chances to metastasize, or was otherwise good at leaving the primary tumor and circulating. We were less interested in cells that were highly proliferative, given that this may not say very much about the actual process of metastasis.

Following SPI1 knockdown, growth curve analysis was used to determine if SPI1 had a role in cellular proliferation. I performed the first set of growth curve assays using a resazurin-based reagent, which is a dye that NADPH oxidase converts from a blue to a red color that fluoresces. In the first resazurin assay performed in 96-well plates, I removed the cell culture
media from the wells, washed the cells with PBS and then added new media with the reagent. The result was that the SPI1 knockdown condition had significantly less fluorescence than the control conditions. However, this result was not consistent with the number of cells I saw under the microscope when I checked the plates following transfection. Interestingly, I saw many detached and floating cells so I decided to count the cells that were both attached to the plate and those that were floating in suspension. What I found in a cell count using Trypan Blue exclusion dye was that there were many viable cells suspended in the media.

In vitro investigation of SPI1 knockdown in A549 cells showed that decreased expression of this gene did not alter proliferation, but it did decrease cell-cell attachments. We first saw decreased cellular attachments under the microscope, as there were many suspended cells and cells were easily detached with trypsin following SPI1 knockdown.

After observing that the cells looked detached and survive in suspension in cell culture, the next thing I looked at was anchorage independence using a soft agar assay. SPI1 knockdown increase the cells ability to form large colonies in the attachment-free environment. This result is consistent with what I saw under the microscope and found in the Trypan Blue assay, that the SPI1 knockdown cells had increased viability when detached.

Finally, I performed an invasion assay using Matrigel-coated transwell inserts. In this assay, cells must invade through the pores of the transwell inserts, which is an active process. There was no difference between SPI1 knockdown and the control in the invasive capability of the cells. This result is interesting because it demonstrates the discrete nature of the steps of metastasis. Just because a cell has heightened capacity for survival in circulation, does not mean that the cell also has an increased invasive capacity.
As discussed in the introductory chapter, tumor growth and metastasis occurs in discrete steps. As a tumor grows, cells are highly proliferative. At the edges of the tumor, cells begin to invade the surrounding tissue. Following invasion, cells detach from the primary tumor and enter the circulation. Circulating cells must survive attachment-free for a period of time prior to settling in the new metastatic niche. The results of these experiments have located the function of SPI1 to a specific step in metastasis.
5. CONCLUSIONS AND FUTURE DIRECTIONS

Most lung cancers are metastatic at diagnosis. Metastatic lung cancer has a 5-year survival rate of 5%. Little is known about the early steps of metastasis. My research question when I started this project was to whether a novel methodology could be implemented to investigate the broader gap in knowledge about the genetic regulators of the earliest events of metastasis. This question is dependent on two basic assumptions: 1) that metastasis is initiated by single cells and 2) monogenic events are responsible for metastatic phenotypes. There is a tripartite gap in knowledge, which is as follows:

I. Previous CRISPR screens haven’t been able to address metastasis, they only look at growth. A more realistic model is needed.

II. The statistical methodology that works for analyzing the cell culture screens is not realistic for analyzing data from these more realistic models. It does not answer the question.

III. We don’t know much about the genetics of the early stages of metastasis.

5.1 PROJECT SUMMARY

The first goal of this project was to process and demultiplex the sequencing data generated in the whole-genome CRISPR/Cas9 knockout screen performed using a novel 4D ex vivo lung model of metastasis. In Chapter 2, the methods used for creation of a whole-genome CRISPR/Cas9 knockout library in A549 cells are described. The first step was to prepare the DNA samples for sequencing by using PCR and DNA purification steps. Next, I describe the procedures used to run the 4D ex vivo lung cancer screen, collection of samples from the lung
model and preparation of the samples for next generation sequencing was given. The development of a computational pipeline to process and demultiplex the sequencing data is described. A pilot sequencing step of just the input cell DNA was performed, confirming that the basic assumptions of this project were true: that the lentiviral construct had integrated, and the library was present in the input cells. Finally, samples were barcoded, pooled and sequenced followed by demultiplexing using the computational pipeline.

In Chapter 3, I performed analysis on the sequencing data that was processed and demultiplexed in Chapter 2. The main purpose of the analyses was to understand how the statistically significant hits from the 4D ex vivo metastasis screen compared to available data and what is currently known about genes involved in lung cancer metastasis. The first goal was to determine the utility and relevance of the novel methodology by comparing significant hits from the screen to previously implicated genetic regulators of lung cancer metastasis. The initial analyses were aimed towards getting a broad picture of our read count data. To find common features among our replicates and samples, we first compared the data from primary and metastasis tumors in all 3 replicates. Using the gene selection algorithm, we decided that we would choose genes first by number of guide hits and then by read counts. SPI1 is the hub node of the gene network common among input, primary and metastasis samples. From a starting list of more than 21,000 genes, I used an in-house gene enrichment algorithm and network analyses to narrow down a list of genes that could be reasonably validated in the lab. Chapter 3 resulted in identification of SPI1 as a potential gene candidate from the whole-genome screen in the 4D lung model of metastasis.
The final goal of this project was to validate SPI1 knockdown in the lab and investigate whether decreased expression had a phenotype. In Chapter 4, assays to assess the proliferation, anchorage independence, and invasiveness of SPI1 knockdown cells were performed. It was found that SPI1 knockdown did not alter proliferation. However, SPI1 knockdown does decrease cell attachments and increase anoikis-resistance. We concluded that SPI1 acts in the process of metastasis that influences whether cells will survive in circulation.

5.2 CONTRIBUTIONS OF THIS THESIS

The contributions of this thesis include the processing and analysis of data from a novel screen design, and the proposal of a pipeline for experimental validation of genetic hits. We proposed an analysis methodology for screens where effects on cell viability and proliferation are confounding factors and obscure the desired information. To date, most whole genome-scale knockout screens have been performed using isolated cancer cell lines in 2D culture. Most of the rest are in vivo or using more simple cell culture growth assays. This thesis project contributed results from a novel system interpreted in a thoughtful and manner, addressing challenges presented by our methodology that are not addressed in the standard procedures and best practices recommended for whole-genome screens. Our in vitro validation pipeline can be scaled in the future to a high-throughput format for screening of candidates chosen by this screen.

5.2.1 SPI1 may function as a metastasis suppressor in lung adenocarcinoma

The analyses of sample sequencing data from the screen (Chapter 3) indicated that SPI1 was the top hit in our screen. To determine the role of SPI1 in lung cancer metastasis, we performed experimental validation using siRNA to knock down SPI1 gene expression in A549
cells (Chapter 4). It was found that SPI1 knockdown with siRNA did not alter proliferation or growth rate of the cells. We concluded that SPI1 decreases cellular attachments and increases the viability of A549s in suspension, which translates to lung adenocarcinoma cells being viable in circulation.

The definition of a metastasis driver or suppressor is different from that of an oncogene or tumor suppressor. Analogous to a tumor suppressor, a metastasis suppressor is a gene that inhibits metastasis when expressed (Khan and Steeg 2018). The definition of a metastasis suppressor also requires that the gene does not alter proliferative capacity of the cell. (Song 2004; Yoshida et al. 2000). Considering our results showing that SPI1 knockouts were enriched in the metastatic tumors in our screen and SPI1 knockdown decreased cellular attachment, this gene appears to be playing a role in lung tumor cells as a metastasis suppressor. In vitro investigation of SPI1 knockdown in A549 cells showed that decreased expression of this gene did not alter proliferation, but it did decrease cell-cell attachments. Our results fulfill the definition of a metastasis suppressor. More tests need to be performed to confirm this functionality.

In our screen using 4D ex vivo lung metastasis model, we found that metastatic tumor cell sequences were enriched for guide RNA sequences corresponding to the SPI1 gene and suggesting putative knockout of the gene. We hoped to investigate genes that were specifically involved in metastasis, which means that perturbation of the gene should not alter proliferation in an ideal situation. We found that SPI1 was expressed in lung tumor cells in cell culture lines (A549 and H1299) and in clinical tumor samples from patient tumor biopsies (see Chapter 3). In comparison to normal lung tissue, SPI1 expression was decreased in cancerous lung tissue.
SPI1 was decreased further in higher-grade cancers, when comparing Stage I to Stage IV for example. We also found that SPI1 expression corresponded to decreased survival during the first 10 years after initial diagnosis, with low expression having lower survival.

Speculation of the role of SPI1 in lung cancer metastasis can be informed by a discussion of development. During development of hematopoietic cells, SPI1 is very important in determination of lineage decisions. Because hematopoietic cell lineages have capacity for circulation, we can hypothesize that inhibition of SPI1 in A549 cells altered the cell’s capacity for circulation.

SPI1 is a novel finding, and its significance in this screen begs the question as to why it has not been implicated in lung cancer before. One reason for this may be that this screen looks at the earliest steps in metastasis, which due to the limitations of looking at single-cell events in this process, has not been probed many times before. Another reason may be that SPI1 is a transcription factor, which are notorious for low expression. Endogenous SPI1 expression in A549 cells is below the limit of detection for RNA-seq and Western blot, and therefore a semi-quantitative RT-PCR technique had to be employed in order to detect the presence of SPI1 mRNA. Finally, the analysis methods used for this screen were different from previous screens using this (or other) lung adenocarcinoma cell lines. Instead of using gene enrichment or fold change as criteria for choice of gene candidates, the top hits for this screen were chosen using a gene selection algorithm based on minimum read and guide number coverage followed by a network connectivity selection step. For all these possible reasons, SPI1 has the potential to be a real and relevant regulator of lung cancer metastasis that slipped under the radar until now.
Due to limitations in detection, collection and study, little is known about the early or initiating events of metastasis. Understanding the regulators of early metastatic events is important to development of prevention and treatment of metastatic disease. Development of methodology and appropriate analysis techniques is of urgent need to overcoming challenges associated with studying metastasis.

Chapter 1 introduced the idea that tumor growth and metastasis occurs in discrete steps. As a tumor grows, cells are highly proliferative. At the edges of the tumor, cells begin to invade the surrounding tissue. Following invasion, cells detach from the primary tumor and enter the circulation. Circulating cells must survive attachment-free for a period of time prior to settling in the new metastatic niche. The results of these experiments have located the function of SPI1 to a specific step in metastasis.

5.2.2 Screen Design

Because metastasis is a process that is made of multiple independent steps, where tumor cells must exhibit specific capabilities to overcome the barriers at each step, there are many different ways to model the metastatic process. Each method of modeling the process might be very informative and meaningful when taken in the context of the model system’s features, limitations, and comparisons to the conceptual framework and corresponding work done by others, and understanding the complicated process of metastasis in a clinically meaningful way will ultimately require the combined efforts and results from many different assays.

Studying metastasis using cancer cell lines presents certain limitations. These limitations must be considered in order to interpret and utilize any results from studies using cancer cell lines. Human cancer cell lines are made from immortalized cells from a patient’s tumor. The
modifications that must be made to patient cells from a biopsy make the cells immortal and capable of growing in cell culture. Cancer cell lines are not accurate representations of cancer cells for a number of reasons. Cell lines have abnormal genomes that are highly dynamic and variable and therefore create complications in comparing different aliquots and passages. Cell lines are grown in a monolayer on plastic addicted to growth hormone and therefore have very different environmental response and gene expression, making them difficult to compare to \textit{in vivo} cancer cells. While cell lines present the aforementioned limitations, they should not be discounted as a tool for investigating mechanisms to study cancer. Instead, \textit{in vitro} experiments using cell lines should be designed and results interpreted within the context of these limitations. Cell lines should be used for experimental exploration and validation of particular mechanisms involved in cancer. Efforts should be made to recapitulate the \textit{in vivo} phenomena of cancer and metastasis and to gain understanding of what aspects are effectively represented and what experimental artifacts exist.

The 4D model system used in this project has some unique advantages to both 2D cultures as well as tumor samples collected from in vivo animal studies or patient biopsies. Each method has different influence over the gene candidates that are significant, which is due to both technical artifact, expression potential and detectability. Combining the 4D lung model of metastasis with a whole-genome CRISPR/Cas9 knockout screen has potential for a powerful survey of all of the genes in the genome and their involvement in lung cancer metastasis.

The current published guidelines for pooled screens suggest that the screening procedures be optimized to achieve high cell-to-guide ratio at each sampling step (here referred to as “library coverage”). High library coverage is suggested in order to meet criteria for sequencing
data processing and analysis, and high statistical certainty is achieved by setting thresholds at each step of the screening process. These thresholds are based on amount of coverage of sgRNA sequences achieved at each step. Coverage is calculated using the total guide number in the library as a benchmark for how many cells are needed each step of the screen. To achieve 100X coverage for a library of 20 guides, for example, recommendations are that a minimum of 2000 cells are harvested at each step. The suggested coverage was not possible in our 4D ex vivo screen. Seeding of the primary tumor was the only step in our screen that achieved the suggested coverage, and it is unclear how much was remaining at the final harvest and in what proportion.

For phenotypic screens done using very well-defined and easily controlled model systems, the coverage recommendations are straightforward. However, model systems aimed at more accurately representing some aspects of physiology have added complexity and therefore become harder to anticipate and control. For example, quantification of viable cells at any given time in the 4D lung model following the initial seeding would require extensive optimization and study that would negate the advantages conferred by use of fewer animals, time and manpower. More importantly, the optimization would likely require automation of the system and simplification in a way that undermines the model’s ability to recapitulate the \textit{in vivo} phenomenon.

Model systems that have more in common with the in vivo process increase the complexity, technical need requirements, and, by definition, lack ease of control. Therefore, there is a trade-off between high statistical power and physiological similarity or significance. Systems where rigidly consistent cellular growth dynamics and guide sequence counts are the goal should be
re-thought and replaced with higher-complexity systems that require more thought and care in designing and interpreting.

The 4D model system used for this thesis project is of higher complexity, which did significantly complicate the analysis of the data and the conclusions drawn from the results. We addressed these complexities in our analysis by not using cell proliferation as a criterion for what we would consider a gene candidate. As a result of our use of a higher-complexity model system and thoughtful analysis, we ended up with a gene hit that we wouldn’t have seen in a cell culture model with traditional enrichment/depletion analysis.

5.3 NEXT STEPS

There were some unknowns about how things would turn out when this screen was conceptualized, and there was no way to do much optimization prior to just going in and seeing how it would work out. The 4D lung model used in this work had not been designed for or used in the context of DNA extraction previously. It was only known at that time that methods investigating cell-intrinsic expression patterns and morphology were feasible. The screen performed in this study can be optimized in future for better overall library coverage and predictability in yields that will be available at each step.

The contributions of this study include the processing and analysis of data from a novel screen design, and the proposal of a pipeline for experimental validation of genetic hits. An analysis methodology for screens was proposed where effects on cell viability and proliferation are confounding factors and obscure the desired information. To date, most whole genome-scale knockout screens have been performed using isolated cancer cell lines in 2D culture. Most of the rest are in vivo or using more simple cell culture growth assays. This project
contributed results from a novel system interpreted in a thoughtful and manner, addressing challenges presented by our methodology that are not addressed in the standard procedures and best practices recommended for whole-genome screens. Our *in vitro* validation pipeline can be scaled in the future to a high-throughput format for screening of candidates chosen by this screen.

The analysis methods were developed as a function of having low coverage in addition to a novel question about which genes might be relevant in this type of screen. During this process of analysis, it was realized that, to date, there has been no study of cell detachment, travel and attachment of this type. The vast majority of screens use proliferation and viability as a way to determine hits. However, it is clear that the process of metastasis does not involve changes in cell proliferation as part of its definition or required processes. Therefore, a metastasis screen that aims to study the cells that undergo detachment, circulation and reattachment while minimizing the confounding issue of proliferation (to directly compare metastatic ability of cells) is being proposed as a future step for this concept. The metastasis assay would be aimed at tracing causation for a cell detachment-reattachment phenotype that minimizes confounding effect of altered cell viability.

Tumor characteristics are highly dependent on tissue architecture, and many studies have shown that cells grown in traditional 2D culture to not adequately model cancer growth dynamics, cellular interactions, or drug sensitivity. An adhesion screen based on the 4D *ex vivo* lung model would be a useful tool to screen drugs in a high-throughput fashion.
The overall goal of the field of lung cancer research is to ultimately increase the quality and quantity of life in those who suffer from lung malignancy. There are many different potential avenues that research can take towards this overall goal. Initially, the long-term goals of this project were to find genetic drivers and suppressors of early events in metastasis, which could ultimately serve as targets for development of early molecular diagnostics, prognostic biomarkers, and targeted therapeutics. Based on what I learned in my thesis work, I am inspired to amend these long-term goals.

The long-term goals listed above share a basic assumption that genes are responsible for phenotype. While this is true, it is not complete. In order for this type of project to contribute to achieving of the original goals, the relationship between genotype and phenotype would need to be linear, and individual genes would need to contribute significantly or totally to a specific phenotype. In reality, the relationship between genotype and phenotype is not linear. First of all, genes are pleiotropic, with certain genes having many different functions. Genomes also have extraordinary functional redundancy, which is the ability for many genes to function in the same or similar capacity. Functional redundancy is the reason that cancers are so good at finding a way to escape pressure or inhibition on a specific function. Finally, the non-linear relationship between genotype and phenotype is further complicated by the fact that non-genetic factors like epigenetics, gene expression, post-transcriptional and post-translational modifications contribute significantly to phenotype.

Molecular diagnostics, prognostics and targeted therapeutics, collectively topics that fall under the category of personalized medicine (also called precision medicine), which aims to
use a patient’s genetic and biomarker data to prescribe a highly individualized plan to alter physiology. The precision medicine approach operates on the paradigm that the properties of the individual elements (genes) that contribute to a phenomenon (phenotype) can be understood at the individual level and can be modulated or finely-tuned.

Instead of focusing on the individual units, or genes, that make up the metastatic response, realizing that there are many different ways to achieve the same outcome may be a better philosophy. In Hanahan and Weinberg’s seminal review, “The Hallmarks of Cancer,” they attempt to distill down the vastly diverse array of distinct genotypes and phenotypes that are displayed by different cancers into a set of six characteristics, or hallmarks, that are common to all cancers because they are necessary and sufficient for their development (Hanahan and Weinberg 2000). In the quest for targeted therapeutics, certain drugs have been identified as having activity against cells with pre-metastatic traits, such as mesenchymal characteristics and anoikis-resistant cells (Anderson et al. 2019). These drugs have not been considered widely and tend to be dismissed relatively quickly following screens that identify them because they do not have activity against the most visible form of cancer (proliferating tumors) (Gandalovičová et al. 2017; Van Roosbroeck et al. 2017). However, anti-metastatic therapeutics have potential at times when treatment would decrease the likelihood of a primary tumor throwing a metastasis prior to, during, and in the period immediately following treatment.

Early diagnostics in the form of more sensitive and safer screening protocols is an important route of inquiry. Molecular diagnostics is a potential solution but should be viewed with extreme caution. False positives indicating further investigation, such as surveillance or
additional diagnostic testing are not without risk and a certain proportion of negative side effects and outcomes. Additionally, each molecular test detects only a very specific biomarker associated with often a specific cancer type or origin. Molecular diagnostics should be viewed as a supplementary tool to aid in early screening and diagnostics, rather than constitute a total solution and viable end-goal.

The most recent data in the United States has shown that cancer survival has made its largest increase in history. Lung cancer survival disproportionately contributed to this survival rate. The increase in lung cancer survival is due to the 2010 initiative to begin early screening in people who were asymptomatic but at high risk for lung cancer (Bernards 2010; Deffebach and Humphrey 2015). By calculating the rate of cancer in the population of people in the United States who meet the criteria to begin yearly screening, it is estimated that 12,000 deaths from lung cancer per year could be averted (Ma et al. 2013). The most recent data give hope for further effective implementations of early diagnostics in lung cancer, thereby acting as a tool for prevention, not of cancer itself, but of the steps involving invasion and metastasis that make lung cancer deadly.
6. BIBLIOGRAPHY


VITA

Alexandria Kathryn Plumer was born in Santa Monica, California, to Kathryn Gardner, MD and Mark Plumer. She attended University of California, Los Angeles (UCLA), completing Bachelor's and Master's of Science degrees in Microbiology, Immunology, and Molecular Genetics. Alexandria currently resides in Houston, Texas.