2-1-2016

Regulatory networks in non-small cell lung cancer: Connecting differentially expressed genes, miRNAs, and lncRNAs

Jasmine Omran

THE AMERICAN UNIVERSITY IN CAIRO
الجامعة الأمريكية بالقاهرة

School of Sciences and Engineering

Regulatory Networks in Non-Small Cell Lung Cancer: Connecting Differentially Expressed Genes, miRNAs, and lncRNAs

A Thesis Submitted to the:

The Biotechnology Graduate Program

In partial fulfillment of the requirements for

The degree of Master of Science in Biotechnology

By: Jasmine Kamal Omran

Bachelor of Science in Human Biology, University of California San Diego

Under the supervision of Dr. **Hassan Azzazy**

Professor, Department of Chemistry, The American University in Cairo

December/2016

The American University in Cairo

School of Science and Engineering

Regulatory Networks in Non- Small Cell Lung Cancer: Connecting Differentially Expressed Genes, miRNAs, and lncRNAs

A Thesis Submitted by: Jasmine Kamal Omran

To the Biotechnology Graduate Program

December 2016

In partial fulfillment of the requirements for the degree of

Master of Science in Biotechnology

Has been approved by:

Thesis Committee Supervisor/Chair _____

Affiliation_____

Thesis Committee Supervisor _____

Affiliation_____

Thesis Committee Reader/Examiner _____

Affiliation_____

Thesis Committee Reader/Examiner _____

Affiliation_____

_____	_____	_____	_____
Dept. Chair/Director	Date	Dean	Date

DEDICATION

I dedicate this thesis to my mother and father. My mother provides me with unconditional love, support, and motivation. She is my queen. She is the power behind the throne of the person I am today. And my father taught me early on that being compassionate was not enough to achieve my dream of becoming a doctor; he taught me that it requires diligence and perseverance. He instilled in me the desire to contribute to society. And despite my losses I still view life with enchanted optimism.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisor, Dr. Hassan Azzazy for all of his support. I would like to thank him as he helped me grow not only as a student but also as a scientist and a teacher. He continues to inspire me with his devotion to students and immense scientific knowledge. He is a role model for hard work and scientific innovations. Next, I would like to thank all my professors in the AUC Biotechnology program for contributing to the honing of my scientific knowledge. Also, I would like to thank Ahmed Elhosseiny for his support, bioinformatics help, and late hours. He is an icon of hard work and resilience. Also, I would like to extend many thanks to Marwa Zahra as she is supportive and provided her time, guidance, and intelligence. Last but not least, I would like to extend my gratitude to my mother and my friends for believing in me and for their continuous support.

ABSTRACT

The American University in Cairo

Regulatory Networks in Non-Small Cell Lung Cancer: Connecting Differentially Expressed Genes, miRNAs, and lncRNAs

By: Jasmine Omran

Under the Supervision of: **Dr. Hassan Azzazy**

Non-small cell lung cancer (NSCLC) is the most prevalent class of lung cancer and the most common cancer worldwide. NSCLC accounts for 85% of total lung cancer cases and leads to the most cancer-related deaths worldwide. Long non-coding RNAs (lncRNAs) and micro RNAs (miRNAs) are gene regulatory elements that play crucial roles in cancer biology such as cancer cell growth and metastasis. Understanding the gene regulatory elements that influence cancer biology is critical for diagnostic and therapeutic purposes. A systems approach can help simulate interactions between these elements. In this study 110 microarray samples from NSCLC patients were analyzed by computational methods to identify differentially expressed genes in two tissue types: NSCLC and normal lung tissue. Identified differentially expressed genes were functionally clustered and annotated with their miRNA and lncRNA targets using miRTarBase and starBase, respectively. Regulatory networks were created to suggest an interplay between these miRNAs, lncRNAs, and differentially expressed genes. This approach led to the identification of 108 differentially expressed genes. Innumerable miRNAs target the differentially expressed genes but 66 miRNAs were identified by literature mining and strong evidence validation methods to identify miRNA and differentially expressed gene targets. The filtered miRNAs were also paired with seven of the most common NSCLC-associated lncRNAs. Based on the findings of this computational study and other studies in literature, connections of differentially expressed genes, miRNAs, and lncRNAs were suggested. TGFBR3 and HHIP, tumor suppressor genes, and CAV1, an oncogene, were functionally related to carcinogenesis and cancer cell metastasis, respectively and were related to cell signaling and extracellular matrix genes. This study suggests that MALAT1, PVT1, and GAS5 are lncRNAs that regulate gene expression via miRNA targeting. Since miRNAs, and lncRNAs are instrumental gene regulatory factors in determining NSCLC diagnosis and prognosis, these regulatory pathways can lead to novel approaches in cancer therapy. Therefore, these networks propose mechanisms of actions to further study miRNAs and lncRNAs suggesting a crosstalk between miRNAs, lncRNAs, and differentially expressed genes.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CDC: Center for Disease Control and Prevention

COPD: Chronic Obstructive Pulmonary Disease

CSV: Comma Separated Values

DAVID: Database for Annotation, Visualization and Integrated Discovery

EMT: Epithelial Mesenchymal Transition

GEO: Gene Expression Omnibus

IARC: International Agency for Research on Cancer

JNCI: Journal of the National Cancer Institute

LATS2: Large Tumor Suppressor Kinase 2

lncRNAs: long non-coding RNAs

miR: micro RNA

miRNAs: micro RNAs

MMP: Membrane Metalloprotease

ncRNAs: non-coding RNAs

NCRP: National Cancer Registry Program

NSCLC: Non-Small Cell Lung Cancer

Pre-miRNA: precursor miRNA

pri-miRNA: primary miRNA

RISC: RNA-induced silencing complex

RMA: Robust Multi-Array Average

SCLC: Small Cell Lung Cancer

TXT: Text format

CHAPTER 1: Introduction

1.1. NSCLC

1.1.1. Epidemiology

Lung cancer encompasses two main categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC is the more prevalent of the lung cancers responsible for 85% of those diagnosed with lung cancer. NSCLC encompass adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Lung cancer is the prevailing cause of cancer-related deaths worldwide for both males and females and is more common in the older generation (65 years or older). More specifically, lung cancer is the most common cause of cancer related-deaths in males while it is the second most common cause of cancer related-deaths in females. In both sexes, 19% of all cancer deaths are attributed to lung cancer alone totaling to 1.6 million deaths (CDC). According to the American Cancer Society, 1 out of 4 cancer-related deaths are due to lung cancer. More deaths result from lung cancer every year than of colon, breast, and prostate combined (American Cancer Society).

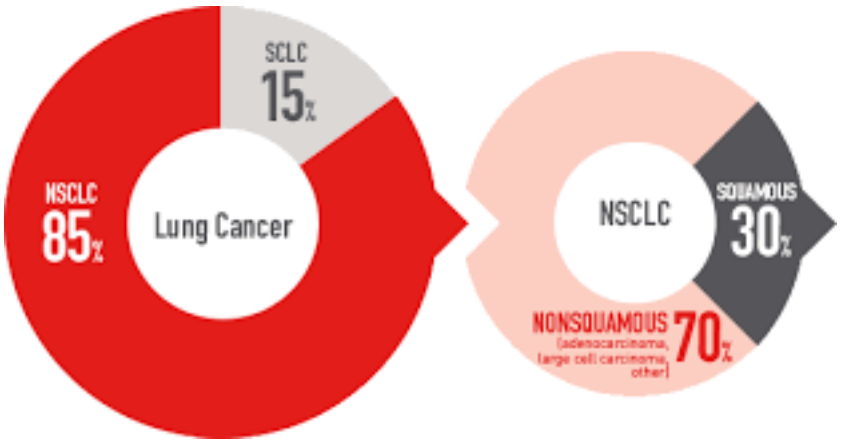


Figure 1. Sub-groups of Lung Cancer (Lilly Oncology).
Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are two dominant classifications within lung cancer. NSCLC accounts for 85% of lung cancer diagnoses and encompasses adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Figure was reproduced with permission from reference 30; see appendix.

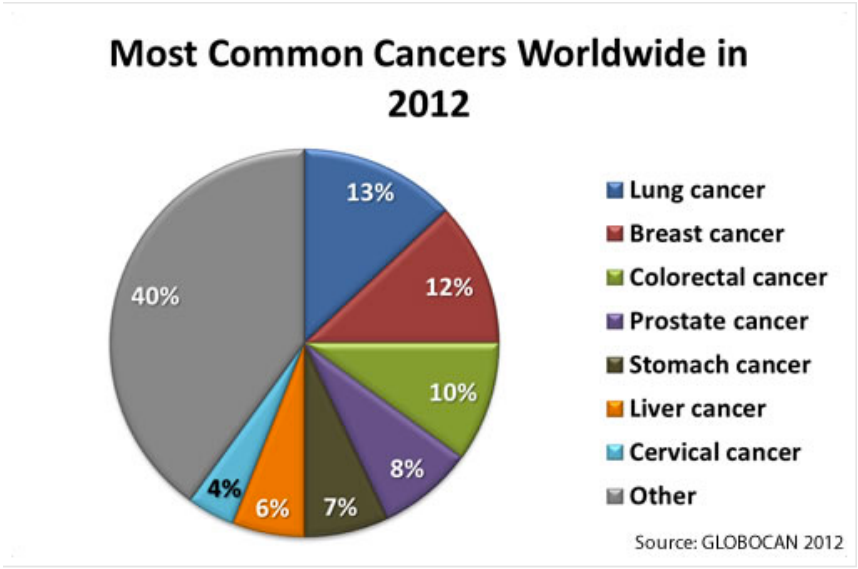


Figure 2. Most Common Cancers Worldwide in 2012 (CDC).
In both males and females, lung cancer is the most common classified cancer globally (CDC). Figure was reproduced with permission from reference 7; see appendix.

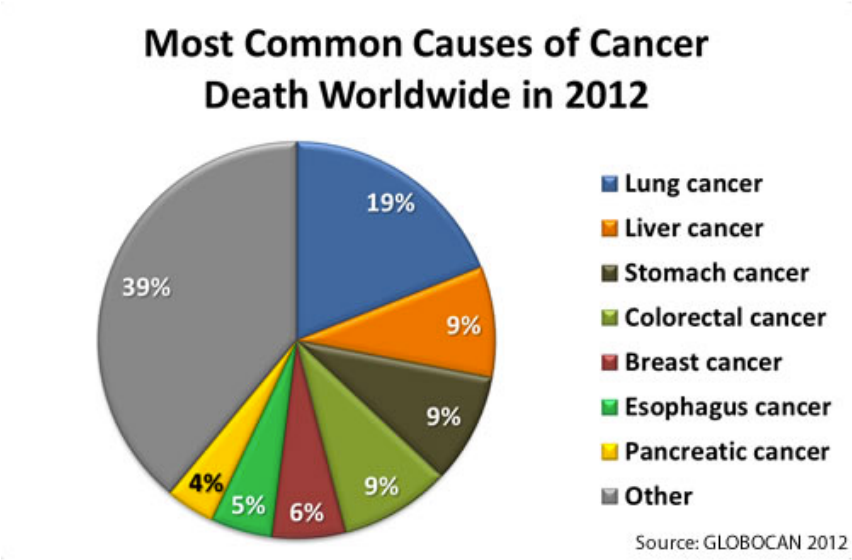


Figure 3. Most Common Causes of Cancer Death Worldwide in 2012 (CDC).
In both females and males, 19% of all cancer-related deaths are due to lung cancer accounting for 1.6 million deaths (CDC). Figure was reproduced with permission from reference 7; see appendix.

1.1.2. Statistics

According to the American Cancer Society, 158,080 Americans are supposed to die in 2016 which will be responsible for 27% of all cancer deaths. Also, as indicated by the American Lung Association and the American Cancer Society, an expected 224,390 Americans are expected to be diagnosed with lung cancer in 2016 accounting for 14% of all cancer diagnoses. Despite these statistics, the incidence rate for both men and women has been declining. One in 14 men and one in 17 women are likely to develop lung cancer during their lifetime (American Cancer Society).

1.1.3. NSCLC in Egypt

Lung cancer accounts for 5-7% of cancers in Egypt (National Cancer Registry Program). The incidence rate in Egypt is expected to drastically increase by a 3-time fold in both males and

females until the year 2050. The expected number of cases are expected to reach over 140,000 and 160,000 for women and men, respectively until the year 2050 according to the National Cancer Registry Program of Egypt (NCRP) (Ibrahim, Khaled et al. 2014).

	Lower Egypt 2009–2011				Middle Egypt 2009				Upper Egypt 2008			
	Site	%	Crude rate	ASR	Site	%	Crude rate	ASR	Site	%	Crude rate	ASR
Males	Liver	41.7	57.8	81.0	Liver	20.4	22.4	37.6	Bladder	12.6	12.2	19.3
	Bladder	8.8	12.2	19.0	Bladder	14.2	15.6	26.4	Liver	11.8	11.5	17.5
	NHL	6.0	8.3	10.3	Brain#	7.3	8.0	12.5	Lung##	7.6	7.4	11.5
	Lung##	5.5	7.6	10.1	Lung##	5.8	6.3	10.8	Leukemia	6.1	6.0	6.7
	Prostate	4.8	6.7	11.7	Leukemia	4.9	5.4	6.7	Prostate	5.9	5.7	9.2
Females	Breast	33.2	43.8	53.0	Breast	26.8	25.8	35.6	Breast	38.7	45.3	64.5
	Liver	16.4	21.6	32.6	Liver	8.9	8.6	13.7	Ovary	6.1	7.1	10.2
	Brain#	4.4	5.8	7.4	Brain#	7.7	7.4	11.1	Liver	5.1	6.0	8.7
	NHL	4.1	5.4	6.7	Leukemia	4.7	4.5	5.6	Leukemia	4.8	5.6	7.2
	Thyroid	3.9	5.1	5.4	NHL	4.4	4.2	5.8	Uterus	3.5	4.1	6.7
Both Sexes	Liver	29.6	40.1	56.8	Liver	15.2	15.6	25.7	Breast	21.6	23.1	33.2
	Breast	16.1	21.7	26.9	Breast	12.4	12.8	18.1	Liver	8.2	8.8	13.1
	Bladder	5.9	8.0	12.5	Bladder	9.2	9.5	15.7	Bladder	7.4	7.9	12.5
	NHL	5.1	6.9	8.5	Brain#	7.5	7.7	11.8	Leukemia	5.4	5.7	7.0
	Brain#	4.5	6.0	7.8	Leukemia	4.8	4.9	6.2	Lung##	4.6	4.9	7.7

#Includes brain and nervous system tumors.

##Includes trachea, bronchus and lung tumors.

Figure 4. Incidence rates of the most frequently observed cancers in Egypt.
In Egypt, lung cancer is ranked third and fourth amongst males and fifth amongst both genders (NCRP) (Ibrahim, Khaled et al. 2014). Figure was reproduced with permission from reference 21; see appendix.

	2013			2015			2020			2025			2050		
	Males	Females	Total	Males	Females	Total	Males	Females	Total	Males	Females	Total	Males	Females	Total
Lip	135	126	262	146	135	281	178	164	342	206	202	408	438	427	866
Tongue	155	175	330	164	185	349	186	231	417	219	296	515	417	763	1180
Mouth	228	163	391	243	178	420	275	216	491	314	261	575	544	528	1071
Salivary glands	147	108	255	158	117	275	189	130	319	222	153	375	495	265	760
Tonsil	6	31	37	6	33	40	9	41	50	11	54	65	13	129	142
Other oropharynx	42	25	67	45	28	73	54	30	85	63	37	100	132	65	198
Nasopharynx	145	32	178	154	34	188	171	40	211	207	42	249	428	68	496
Hypopharynx	72	80	152	76	85	161	88	96	184	102	111	213	200	173	372
Pharynx unspec.	35	7	42	36	7	43	42	7	48	55	7	62	94	8	102
Oesophagus	699	485	1184	746	525	1271	897	644	1542	1065	762	1827	2249	1504	3752
Stomach	726	969	1695	772	1045	1816	922	1249	2171	1080	1484	2565	2185	2877	5062
Small intestine	98	179	277	106	194	300	120	229	349	134	274	408	223	507	730
Colon	1522	1339	2862	1618	1437	3055	1893	1715	3608	2225	2063	4287	4465	4120	8585
Rectum	464	406	871	490	432	922	568	502	1070	645	584	1230	1097	1052	2149
Anus	133	50	183	142	53	195	162	65	227	178	71	249	291	127	418
Liver	19646	8345	27991	20932	9043	29975	24420	10900	35320	28580	12933	41513	59047	26425	85471
Gallbladder and so forth	235	324	559	248	350	598	297	413	710	348	488	835	715	967	1682
Pancreas	1350	876	2226	1440	957	2397	1676	1160	2836	1961	1405	3366	3912	2971	6883
Nose, sinuses and so forth	98	136	234	104	144	247	124	170	294	154	186	340	340	322	661
Larynx	933	134	1067	993	142	1136	1194	173	1367	1428	201	1629	3094	395	3489
Trachea, Bronchus, Lung	3304	1586	4890	3530	1703	5233	4168	2031	6198	4889	2404	7293	10176	4895	15071

Figure 5. Estimated number of cancer cases in Egypt 2013-2015.
According to the National Cancer Registry Program (NCRP) in Egypt, the incidence rate is expected to drastically increase by 3-time fold in both males and females until the year 2050 genders. The expected number of cases are expected to reach over 140,000 and 160,000 for women and men, respectively until the year 2050 (Ibrahim, Khaled et al. 2014). Figure was reproduced with permission from reference 21; see appendix.

1.1.4. Risks and Protective Factors of NSCLC

Some of the risks that lead to NSCLC are cigarette smoke, second hand smoking, alcohol consumption, and occupational exposure to carcinogens such as radon and asbestos (American Cancer Society). Long-term air pollution exposure, previous radiation therapy to the lungs, and family history of lung cancer are irreversible risk factors (American Cancer Society). The heredity inheritance of the TP53 gene and other chromosome markers may also lead to lung cancer (American Cancer Society). Any combination of these risk factors increases the likelihood of lung cancer. Some protective factors of lung cancer include exercise, specifically

leisure-time physical activity and fruits and vegetables that are rich in antioxidants, even among heavy smokers (Molina, Yang, Cassivi, Schild, & Adjei, 2008).

1.1.5. Smoking and NSCLC

NSCLC is more common amongst former smokers (60%) than current smokers (25%) (American Lung Association). Male and female smokers are 23 and 13 times, respectively, more probably to develop lung cancer than those who never smoked (American Lung Association). Nonsmokers have a 20 to 30% chance of developing lung cancer if disclosed to second-hand smoke (American Lung Association). Smoking leads to 90% and 80% of male and female lung cancer related-deaths, respectively (American Lung Association).

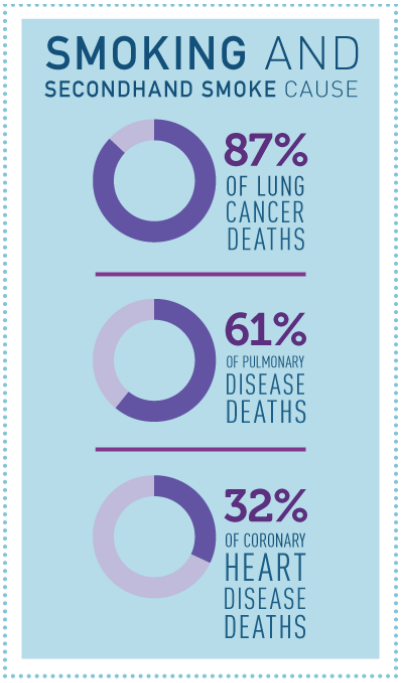


Figure 6. Smoking, a cause of cancer-related deaths.
Smoking is the dominant element of cancer related deaths followed by other pulmonary diseases and heart disease (National Cancer Institute). Figure was reproduced with permission from reference 38; see appendix.

1.1.6. Types of NSCLC

Adenocarcinoma, squamous cell carcinoma, and large cell carcinoma are three primary divisions in NSCLC. Adenocarcinoma effects the outer area of lung and accounts for 40% of lung cancers; it is slower to spread and is considered the most common form of cancer in smokers or former smokers but it is also the most popular form of cancer in nonsmokers. Squamous cell carcinoma, flat cells that line the airways, effects the center of the lung and bronchi and accounts for 25-30% of lung cancers and is connected to a smoking history. Large cell (undifferentiated) carcinoma effects any part of the lung and accounts for 10-15% of lung cancers. Large cell carcinoma grows and spreads fast (American Cancer Society).

1.1.7. NSCLC Signs and Symptoms

A worsening cough, a new onset of wheezing, coughing up blood or rust-colored phlegm, shortness of breath, chest pain, weight loss, loss of appetite, recurring respiratory conditions are among the most common symptoms that appear with lung cancer patients (American Cancer Society). Usually lung cancer is pronounced at a later disease phase or when the disease spreads to distant body parts. As the lung malignancy metastasizes to the brain or spinal cord, bone pain and nervous system changes may occur. If the lung cancer spreads to the liver this may result in jaundice. Symptoms increase and appear as the cancer progresses and proliferates. Some cancers may even lead to syndromes like: Horner syndrome, superior vena cava syndrome, and paraneoplastic syndromes. Lung cancer screening is recommended for patients that are 55-74 years old, in fairly good health condition, and are currently smoking or formerly smoked in the past 15 years (American Cancer Society). Lung cancer screening is also recommended for patients with a "30 pack-year smoking history" meaning the patient smoked one cigarette pack per day for 30 years (American Cancer Society).

1.1.8. NSCLC Diagnosis and Treatments

The current diagnostic procedure includes a medical history and physical exam, chest x-ray, chest CT scan with infusion or contrast material, and biopsy (American Cancer Society). Lung cancer is diagnosed microscopically by observing a collection of lung cells (American Cancer Society). Current treatments include surgery, radiation therapy, combination chemotherapy such as Cisplatin and Paclitaxel (Taxol), targeted therapy such as Bevacizumab and (Avastin) and Ramucirumab (Cyramza), and immunotherapy like Nivolumab (Opdivo) (American Cancer Society). Treatment of NSCLC depends on the stage of the cancer (American Cancer Society).

1.1.9. NSCLC Stages

The stage of the disease is according to the size of the tumor and progression of the spreading of the cancer. Stage 0 is considered the non-invasive stage and the cancer cells are found in the airway lining and have not spread any further. Stage 1A is the beginning of invasive stages; the tumor size is ≤3 cm and the tumor has grown through the airway lining into deep lung tissue. Stage 1B, the abnormal mass size is ≥3cm and the tumor has reached the primary airway. Stage 2A, the tumor is ≤3 cm and the tumor has proliferated to nearby lymph glands on the same side of the chest. Two types of stage 2B progression are present. Stage 2B1, the tumor is ≥5 cm and the tumor has advanced to the lymph nodes. Stage 2B2, the tumor is ≥7 cm and the tumor has not penetrated the lymphatic structure or distant structures. Stage 3A, the abnormal mass varies in size and the cancer spreads to the lymph nodes on the same side of the lung tumor and to nearby structures such as the chest wall and other parts of the thoracic cavity. Stage 3B, the tumor can be any size and the cancer spreads to the lymph glands above the chest or to the other side of the chest and nearby structures. Stage 4, the tumor can be any size and the malignancy spreads to other body parts such as the brain, bone, liver, and adrenal glands (American Cancer Society).

1.2. Gene Regulation

1.2.1. miRNAs

miRNAs are 22 nucleotides in length; these small non-coding pieces of RNA were overlooked until their discovery in 1993. miRNAs play different roles in various cancers including post transcriptional gene regulation, cell-to-cell communication, and cell cycle regulation. miRNAs are known gene regulatory factors that can function as tumor suppressors

and oncogenes. The abnormal miRNA expression levels signify an anomalous state and can be used to target cancer regulatory pathways and provide insight into disease progression. This can lead to the uncovering of original biomarkers and curative breakthroughs. Hence, the study of miRNAs and the fluctuation of miRNA levels is a growing field for disease progression, prognosis, diagnosis, and therapy.

1.2.1a. miRNA Biogenesis

Diseased states of miRNA targets may result from alterations in the processing of miRNAs; hence, it is important to understand the biogenesis of miRNAs. miRNA biogenesis begins in the nucleus where primary miRNAs (pri-miRNA) are processed by Drosha into pre-miRNAs before leaving the nucleus. In the nucleus, Drosha, an RNase III enzyme removes the tails of the pri-miRNA, consisting of 1- 4 kilobases, which leads to the formation of a stem loop called precursor miRNA (Pre-miRNA) to construct the 22 nucleotide non-coding sequence (Bartel 2004). Pre-miRNAs are transported from the nucleus via Exportin 5. Once out of the nucleus, the pre-miRNA is further prepared by Dicer in the cytoplasm. Dicer removes the stem loop from the pre-miRNA producing a mature miRNA. The mature miRNA can be primed into the Argonaute protein in the RNA-induced Silencing Complex also known as RISC (Bartel 2004). This RISC complex in turn degrades mRNA and or represses gene expression. This further proves the ability of miRNAs to regulate gene expression post transcription.

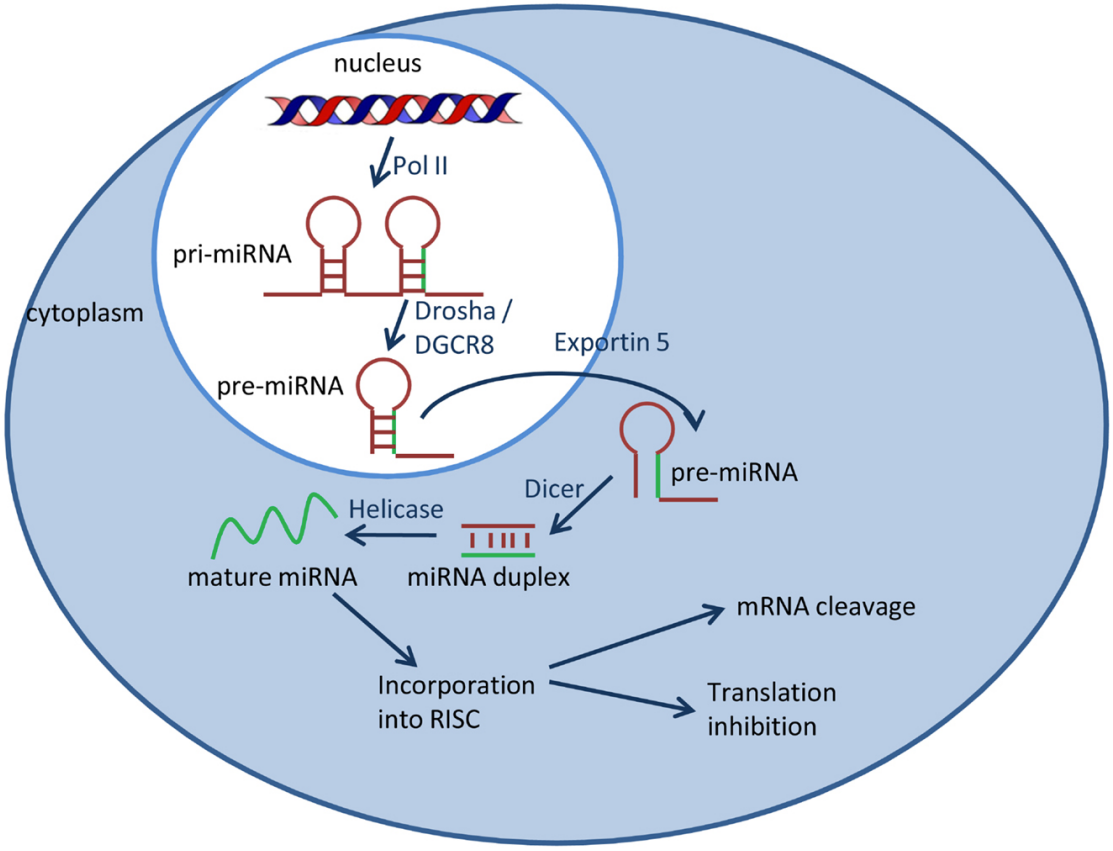


Figure 7. miRNA Biogenesis overview.
Primary miRNA (pri-miRNA) transcribed by pol II or pol III must be processed by Drosha before it is exported from the nucleus. After processing takes place pri-miRNA becomes Precursor miRNA (Pre-miRNA). Pre-miRNA is transported from the cytoplasm to nucleus via Exportin 5 and is further refined by Dicer to produce a mature micro RNA (miRNA). Mature miRNA can be armed into the Argonaute protein containing RNA-induced silencing complex (RISC). This RISC complex consecutively degrades mRNA and represses gene expression (Rothschild, 2013). Figure was reproduced with permission from reference 42; see appendix.

1.2.1b. miRNA Functions

miRNAs have various functions including post-transcriptional gene regulation, cell-to-cell communication, and cell cycle regulation all of which are important to understand disease progression. miRNAs create related cell networks and loop mechanisms to regulate many target cells simultaneously (Gurtan and Sharp 2013). This proves the ability of miRNAs to reach multiple targets which further confirms the gene regulatory capabilities of miRNAs. A recent

study found that "miRNA transfer can fine-tune gene expression during generation of the immune response and increase the complexity of communication between immune cells" (Chen, Liang et al. 2012). Cell communication is a key factor which can manipulate the cell environment and produce analogous cell responses that can alter gene expression and lead to the activation or repression of corresponding molecular pathways. miRNAs can also have hormone like affects in terms of reaching and manipulating close and far away neighbors which further promotes the capability of miRNAs to regulate genes. Thus, the lack of cell-to-cell communication can lead to diseased states and can promote cancer development. Most often miRNAs are studied by observing the expression levels of miRNAs in altered states. This approach can pose difficulties due to the genetic diversity of tumors and constant mutations that arise from cancer (Chen, Liang et al. 2012). Since miRNAs can mediate the cell-to-cell communication further studies should focus on targeting the altered cellular mechanisms to be help discover biomarkers and develop therapeutics.

miRNAs not only a critical factor in cell communication but they also take part in cell growth and reproduction. Shifts in cell cycle checkpoints can lead to alterations in cell cycle progression, cell differentiation, and programmed cell death. Diseased states can lead to over-expression or under-expression of miRNAs. Downregulation of miRNAs can result from genomic loss, alterations of genomic histone acetylation, variations in methylation, and repression of oncogenic and/or tumor suppressor transcription factors (Jansson and Lund 2012). Upregulation of miRNAs can result from loss of epigenetic markers (Jansson and Lund 2012). Upregulated miRNAs are functionally classified as oncogenes while downregulated miRNAs are functionally classified as tumor suppressors (Jansson and Lund 2012). For instance, in non-small cell lung cancer miRNA-17 and miRNA-200b are upregulated while miRNA-181 is

downregulated (Nadal, Truini et al. 2015) (Markou 2015). Hence, miRNA-17 and miRNA-200b act as oncogenes while miRNA-181 acts as a tumor suppressor.

1.2.1c. Examples of miRNAs in NSCLC

miRNAs from tissue and blood samples are signature factors for diagnosis, prognosis, and therapy. There are many examples of miRNAs that are associated with NSCLC. Novel biomarkers of NSCLC including downregulated miRNAs have been cited such as: miR-520h, miR-34b, and miR-448. Novel biomarkers of NSCLC including upregulated miRNAs have also been cited such as: miR-22 and miR-654-3p. According to *Xu et al*. (2015), miR-34b and miR-520h take part in the coordination of NSCLC, miR-22 is an oncogene biomarker, and miR-654-3p prohibits NSCLC progression while miR-448 promotes NSCLC progression. More upregulated miRNAs present in NSCLC include: miR-141, miR-193b, miR-200b, miR-301, let-7g, miR-331, miR-331, miR-758, miR-744, miR-106a, miR-19a, miR-17, miR-19b, miR-93, miR-20b, miR-106b, miR-215, miR-25, miR-200c, and miR-24. According to Nadal *et al.* (2015), miR-141, miR-200b, miR-193b, and miR-301 all of which were upregulated in NSCLC and were distinguished as novel serum biological markers for lung malignancy detection.

miRNAs are also prognostic markers for lung cancer. High-expression of miR-155, miR-21, miR-106a, miR-93 and low-expression of let-7a-2, let-7b, miR-145 are associated with unfavorable outcomes of adenocarcinoma patients (Shen & Jiang, 2012). Low levels of miR-1 and miR-499 and high levels of miR-486 and miR-30d are associated with unfavorable prognosis. miR-145 is a predictive biomarker for lung adenocarcinoma. miR-200b, miR-30c-1, miR-510, miR-630, miR-657 predict lung cancer recurrence. miR-221 and miR-222 are related

to aggressive NSCLC while miR-374a is associated with poor patient survival in early-stage NSCLC (Shen & Jiang, 2012).

There are experimentally targeted miRNAs that regulate many pathways responsible for repression of E- cadherin, cytoskeleton rearrangement, and focal adhesion and stress fibers all leading to the spread of lung cancer (Figure 8) (L. Jiang & Qiu, 2013). Involved pathways include JAK/STAT, MAPK pathway, Wnt signaling pathway, and Notch Signaling pathway. miRNAs that are known to regulate invasion and metastasis in lung cancer (Figure 9) (L. Jiang & Qiu, 2013). miR-125a-5p, miR-21, and miR-378 lead to the spread of lung cancer to the brain, liver, and bone (L. Jiang & Qiu, 2013). miR-126, miR-30a, miR-206, miR-200, miR-200c, and miR125a-3p suppress the spread of lung cancer (L. Jiang & Qiu, 2013). The invasive lung cancer phenotype may be addressed by the regulation of the following miRNAs: miR-125b, miR-210, miR-103, miR-194, and miR-500 (L. Jiang & Qiu, 2013).

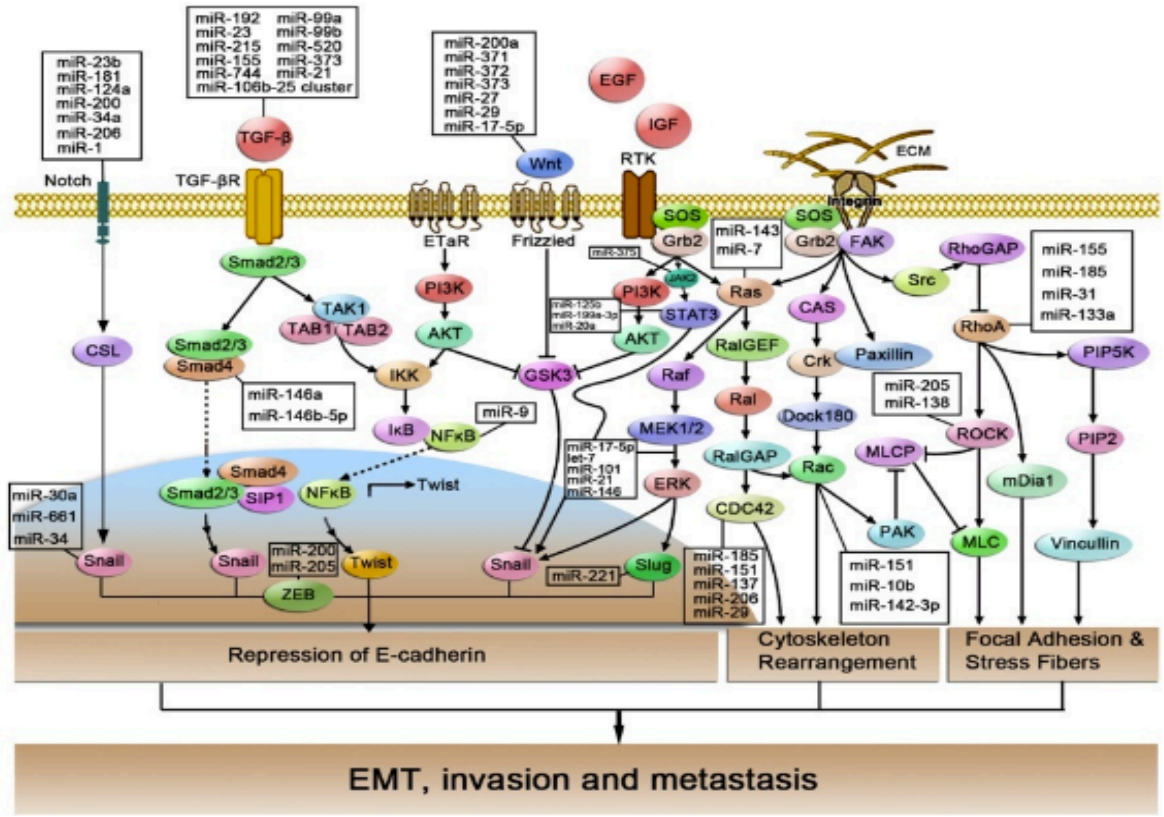


Figure 8. Experimentally validated miRNAs and their target pathways involved in invasion and metastasis.

The micro RNAs (miRNAs) target many pathways responsible for repression of E- cadherin, cytoskeleton rearrangement, and focal adhesion and stress fibers all leading invasion and metastasis of lung cancer. Involved pathways include JAK/STAT, MAPK pathway, Wnt signaling pathway, and Notch Signaling pathway (L. Jiang & Qiu, 2013). Figure was reproduced with permission from reference 23; see appendix. EMT: Epithelial- Mesenchymal Transition PI3K: Phosphoionositide-3-Kinase, EGF: Epidermal Growth Factor, IGF: Insulin-like Growth Factor, TGF: Transforming Growth Factor, ECM: Extracellular Matrix, RTK: Receptor Tyrosine Kinase

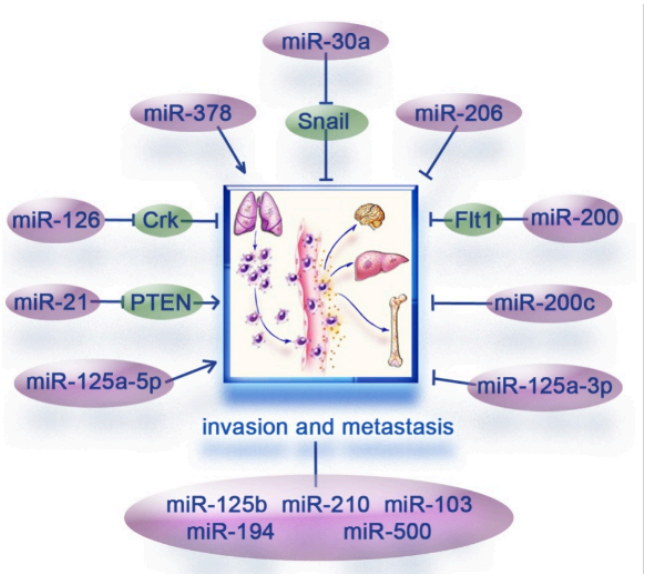


Figure 9. miRNAs known to regulate invasion and metastasis in lung cancer.
miR-125a-5p, miR-21, and miR-378 lead to the invasion and metastasis of lung cancer to the brain, liver, and bone. miR-126, miR-30a, miR-206, miR-200, miR-200c, and miR125a-3p inhibit the invasion and metastasis of lung cancer. The invasive lung cancer phenotype may be addressed by the regulation of the following miRNAs: miR-125b, miR-210, miR-103, miR-194, and miR-500 (L. Jiang & Qiu, 2013). Figure was reproduced with permission from reference 23; see appendix.

1.2.2. lncRNAs

The study of lncRNAs in the molecular biology field is relatively new. lncRNAs have become more important in understanding the biology of cancer and the paradigm of gene regulation. lncRNAs are functional RNA molecules that bypass translation and regulate gene function. lncRNAs are composed of 200 base pairs or greater and were discovered using histone modifications; lncRNAs are a class of non-coding RNAs (ncRNAs) which were once thought to be "junk" in human genome. Many lncRNAs have been identified but only a few have been functionally annotated. More studies on lncRNAs are being carried out to define lncRNA regulatory pathways. It has also been studied that lncRNA expression differs between tissues types and disease conditions. For example, lncRNA expression differ between normal tissue and

NSCLC tissue; lncRNA expression also differs between adenocarcinoma and squamous cell carcinoma which are subtypes of NSCLC. This regulatory factor is promising for cancer diagnosis and prognosis and introduces RNA-based therapy (J. Yang et al., 2014).

1.2.2a. lncRNA categorization based on target location

lncRNAs can be categorized into two categories based on the location of the genomic target: cis and trans acting lncRNAs. If the genomic target is located close to the site of synthesis and regulates nearby genes, this is known as cis-acting lncRNAs. On the other hand, trans-acting lncRNA has a genomic target located far from the site of synthesis and acts to regulate gene expressions that are farther away, on different chromosomes, or on homologous chromosomes (Vance & Ponting, 2014).

1.2.2b. lncRNA categorization based on transcription site

lncRNAs can be differentiated according to the location of transcription. lncRNAs can be transcribed from different regions: intronic, intergenic, sense, antisense, and bidirectional regions (Figure 10). lncRNAs transcribed from the introns of coding genes are called intronic lncRNAs also known as lintronic RNAs (J. Chen, Wang, Zhang, & Chen, 2014). lncRNAs induced from in between two coding genes are called intergenic lncRNAs (J. Chen, Wang, Zhang, & Chen, 2014). Sense lncRNAs are transcripts from the same sequence as the coding gene and lncRNAs from the opposite gene coding sequence are called antisense lncRNAs (J. Chen, Wang, Zhang, & Chen, 2014). Bidirectional lncRNAs or divergent lncRNAs are transcripts that begin at the start of another coding gene (Zhao & Lin, 2015).

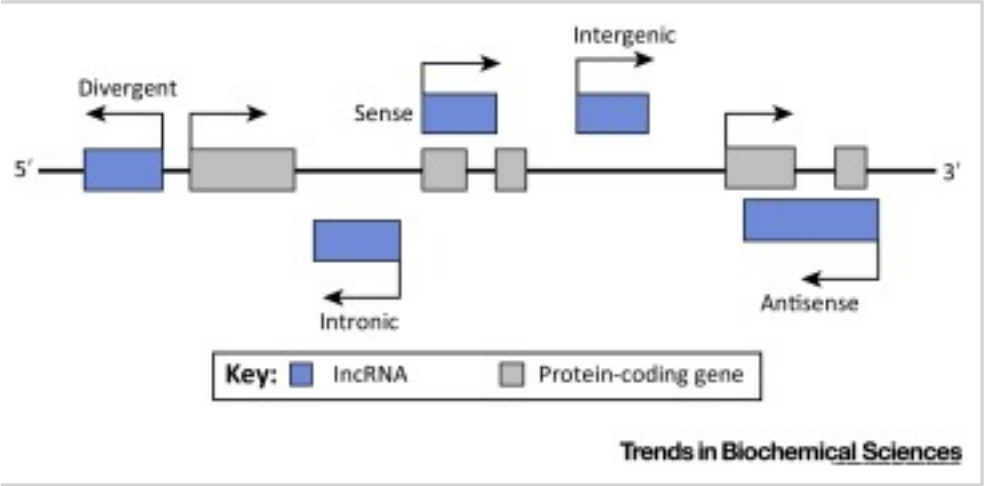


Figure 10. lncRNAs transcribed from different regions.
Long non-coding RNAs (lncRNAs) can be transcribed from in between two coding genes, from introns of coding genes, from the same gene coding sequence, from the opposite gene coding sequence, from the beginning of another coding gene known as intergenic lncRNA, intronic lncRNA, sense lncRNA, antisense lncRNA, and divergent lncRNA respectively (Zhao & Lin, 2015). Figure was reproduced with permission from reference 58; see appendix.

1.2.2b. lncRNA Functions

lncRNAs play a role in chromatin regulation, transcription regulation, and post-transcriptional regulation; they also play an important role in tumorigenesis. lncRNAs functions range from a molecular level to a cellular level. lncRNAs serve many functions such as a transcriptional activator, a transcriptional repressor, a transcriptional guide, and a scaffold for chromatin modification complex. In addition to transcriptional and epigenetic control, lncRNAs are involved in post-transcriptional regulation, like miRNAs. Studies reveal interactions between miRNAs and lncRNAs suggesting a role play in cancer diagnosis and prognosis (G. Yang et al., 2014). Much like miRNAs, lncRNAs were thought to be information deficient. Like miRNAs, lncRNAs regulate transcription and therefore regulate gene expression but the impact of gene regulation by lncRNAs is not yet understood. The similarities between these regulatory factors can be important in understanding disease progression and developing cancer therapies.

lncRNAs can also act as decoys and signals. A lncRNA that functions as a decoy binds to the protein and removes them from chromatin. A lncRNA that functions as a scaffold allows for subunits to assemble and work together. A lncRNA that functions as a guide binds to the protein and targets the gene to regulate gene expression. A lncRNA that function as signals send signals to distinguish gene expression (J. Chen, Wang, Zhang, & Chen, 2014). Furthermore, lncRNAs can act as flexible scaffold for chromatin- modifying complexes, enhancer RNAs, tumor suppressing signalers, RNA processors, RNA-RNA inter-actors, and miRNA sequesters. For example, MEG3 is noted as a tumor suppressor as it is downregulated in cancer cell lines and when overexposed prevents proliferation of the cancer cells in hepatocellular carcinoma (Prensner & Chinnaiyan, 2011). Further understanding of how lncRNAs function is important in understanding molecular disease progression thus leading to advanced mechanistic targeting of the disease.

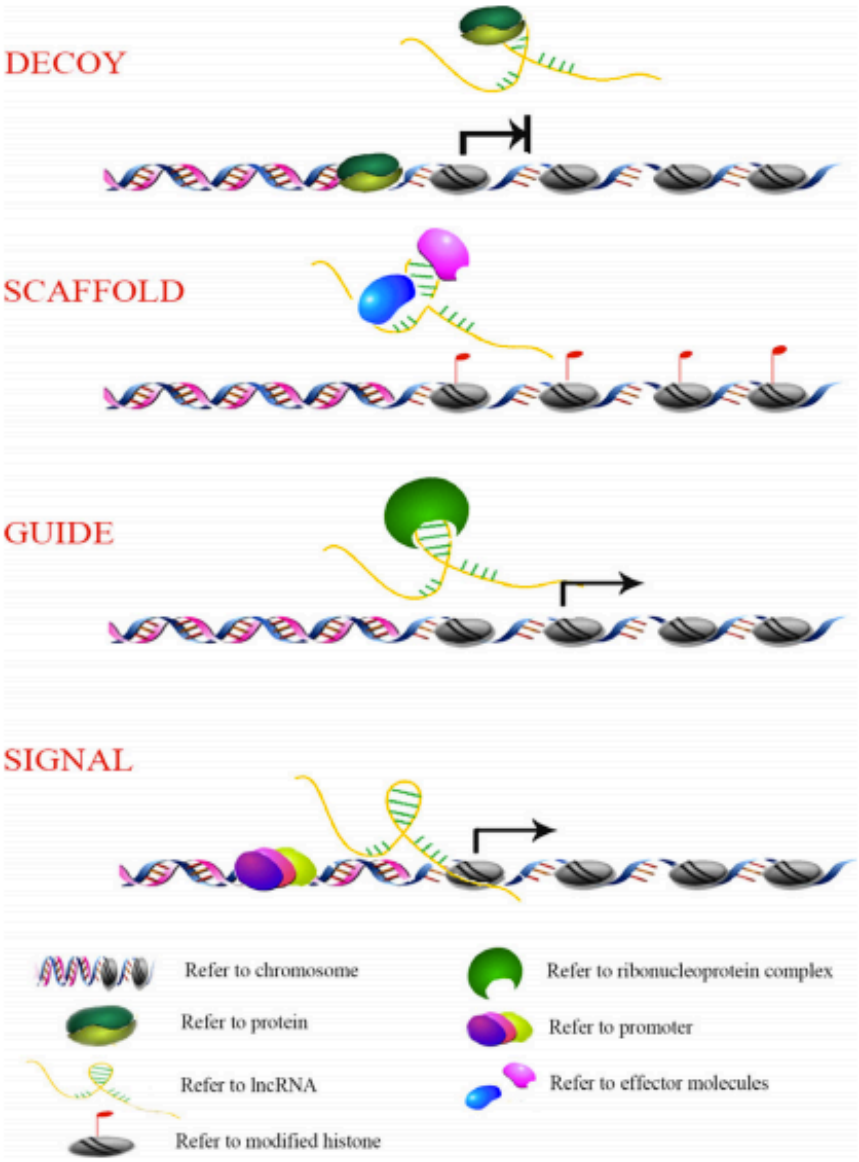


Figure 11. lncRNAs have various functions.

Long non-coding RNAs (lncRNAs) could act as decoys, scaffolds, guides, or signals. A lncRNA that functions as a decoy binds to the protein and removes them from chromatin. A lncRNA that functions as a scaffold allows for subunits to assemble and work together. A lncRNA that functions as a guide binds to the protein and targets the gene to regulate gene expression. A lncRNA that function as signals send signals to distinguish gene expression (J. Chen, Wang, Zhang, & Chen, 2014). Figure was reproduced with permission from reference 8; see appendix.

1.2.2c. Examples of lncRNAs in NSCLC

Much like miRNAs, lncRNAs have also been found in cancer. MALAT1, H19, GAS5, PVT1, and HOTAIR are among the few lncRNAs found in lung cancer. Like miRNAs, lncRNAs have been associated with oncogenic and tumor suppressor activities. H19 is highly expressed in lung carcinomas, upregulated by Myc and hypoxic conditions, and functions as an oncogene in tumorigenesis. MALAT1 is also an oncogene and is highly expressed in lung cancers. H19 function as oncogenes while GAS5 acts a tumor suppressor. BANCR is a known lncRNA and is under-expressed in NSCLC in tumor tissues and takes part in melanoma cell migration (G. Yang et al., 2014). Many other lncRNAs have been found such as CAR10, RGMBAS1, GHSROS, NKX2-AS1, BCYRN1, DLX6-AS1, SOX2-OT, CARLo-5, Lnc_bc060912, MVIH, HNF1A-AS1, CCAT2, LUADT1, ZXF1, ANRIL, SCALI, NRG1, GAS6-AS1, LOC788228, DQ786227, PANDAR, MEG3, SPRY4-IT1, and AK126698. Many of these lncRNAs function to promote cell proliferation, migration, and metastasis. Some of these lncRNAs function to induce apoptosis and suppress cell proliferation (Wei & Zhou, 2016).

There are known lncRNAs in NSCLC that are correlated to specific functions. CCAT2, HOTAIR, BANCR, AK126698, MALAT1, GAS6-AS1, and MEG3 are NSCLC-associated lncRNAs (Figure 12). CCAT2 and AK126698 are involved in pathways that result in invasion and metastasis. AK12669 may also be applied as a potential target for reversing NSCLC cisplatin resistance (J. Chen, Wang, Zhang, & Chen, 2014). MALAT1, GAS6-AS1, and MEG3 are involved in different pathways that lead to invasion and metastasis. MALAT1 acts as a potential diagnostic and prognostic marker for NSCLC. MEG3 not only acts a tumor suppressor but acts as potential therapeutic target for NSCLC. HOTAIR is involved in pathway leading to apoptosis and G0/G1 cell cycle regulations. HOTAIR also act as a potential chemotherapy target.

BANCR is involved in pathways leading to epithelial mesenchymal transition (EMT) (J. Chen, Wang, Zhang, & Chen, 2014).

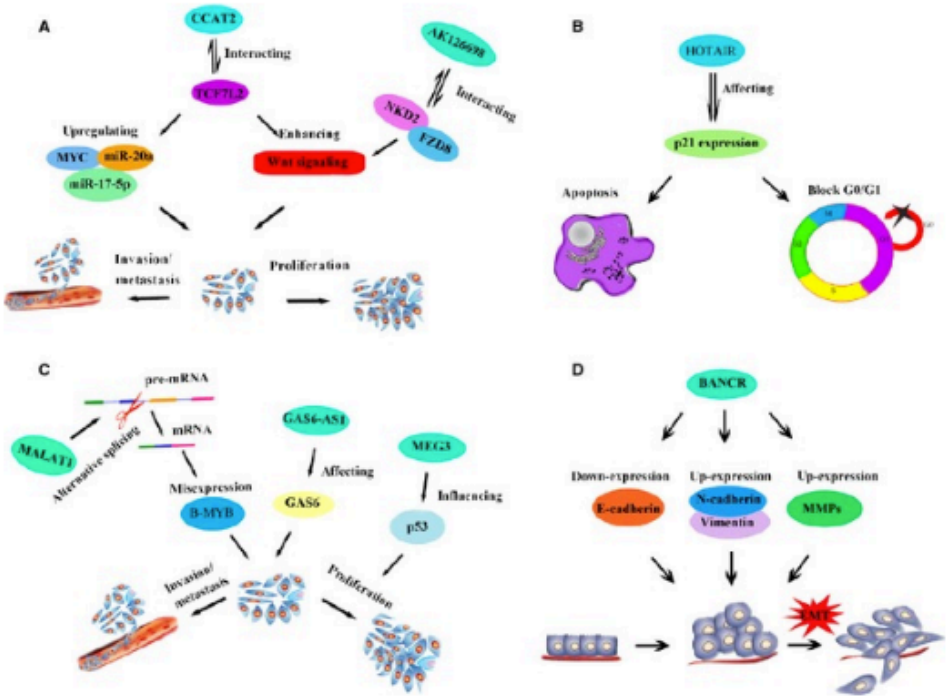


Figure 12. Known lncRNAs in NSCLC and their functions.
CCAT2, HOTAIR, BANCR, AK126698, MALAT1, GAS6-AS1, and MEG3 are NSCLC-associated lncRNAs. CCAT2 and AK126698 are involved in pathways that result in invasion and metastasis. MALAT1, GAS6-AS1, and MEG3 are involved in different pathways that lead to invasion and metastasis. HOTAIR is involved in pathway leading to apoptosis and G0/G1 cell cycle regulations. BANCR is involved in pathways leading to epithelial mesenchymal transition (J. Chen, Wang, Zhang, & Chen, 2014). Figure was reproduced with permission from reference 8; see appendix.

1.3. Microarray Analysis

In biology, genes are studied in many different ways including experimental and computational approaches. In this study, a microarray was used for gene expression analysis. A microarray is a 2D chip that harbors thousands of bound DNA sequences to its surface. Target

DNA sequences hybridize to the surface that is also probed with complementary DNA sequences. There are three types of microarray: spotted arrays on glass, self-assembled arrays, and *in situ* synthesized arrays (Bumgarner, 2013). Microarrays are assembled in a way to allow for gene expression analysis. First, RNA is extracted from isolated cells of interest and enriched for mRNA. The enriched mRNA is reverse transcribed into cDNA and the samples are fluorescently labeled. These labeled samples hybridize to the microarray plate and the plate is scanned to measure hybridization and in turn gene expression. Microarrays are not only used for gene expression analysis but they are also used for transcription factor binding analysis, and genotyping; they have been widely used to study single nucleotide polymorphisms. Given the vast array of functions a microarray contributes to, microarrays comprise of some limitations. Microarrays are not a direct measurement of DNA or RNA at high or low concentration, hence the signal is not linearly proportional. Microarrays can only detect nucleic acid sequences that are probed on the microarray plate therefore not contributing to the discovery of novel genes. Microarrays remain advantageous in allowing for gene regulation analysis (Bumgarner, 2013).

Microarray analysis of NSCLC can lead to the understanding of regulatory pathways. Many regulatory pathways are involved in NSCLC including the STAT3 signaling pathway, Hedgehog signaling pathway, Ras pathway, and TGF-B pathway (Brambilla & Gazdar, 2009). These regulatory pathways function in cell proliferation, invasion, angiogenesis, metastasis, and resistance to apoptosis (Brambilla & Gazdar, 2009). Therefore, studying the genes involved in these pathways as well as the regulatory factors of these genes like miRNAs and lncRNAs is important for targeting tumor pathogenesis and honing therapeutic strategies.

CHAPTER 2: Hypothesis and Objectives

Hypothesis:

Variable gene expression was observed due to the different sample characteristics of NSCLC. NSCLC in this study is further examined according to regulatory factors. Molecular factors such as lncRNAs and miRNAs are fundamental in the interplay of gene regulation in order to ensure cellular homeostasis.

In this study, it is hypothesized that: miRNAs, lncRNAs, and differentially expressed genes in NSCLC may have a biological relationship which further recommends molecular target experiments.

Objective 1:

To identify biologically and statistically significant differentially expressed genes in non-small cell lung cancer patient samples and non-small cell lung cancer patient samples with normal lungs

Objective 2:

To identify miRNAs and lncRNAs that target the aforementioned identified differentially expressed genes

Objective 3:

To propose regulatory networks that include the found miRNAs and lncRNAs that target the differentially expressed genes

CHAPTER 3: Materials and Methods

3.1. Dataset

The dataset selection criteria included sample size, sample type, recent studies, and median centered values. The dataset chosen corresponds to the molecular pathology article authored by *Kadara et al.,* 2013, from the Journal of the National Cancer Institute (JNCI) titled "Transcriptomic Architecture of the Adjacent Airway Field Cancerization in Non-Small Cell Lung Cancer" . The dataset selected contains 226 total samples consisting of NSCLC tumor lung tissue samples, NSCLC normal lung tissue samples, and NSCLC airway samples. Two groups from the chosen dataset were obtained: NSCLC tumor lung tissue samples vs. NSCLC normal lung tissue samples.

3.2. Software

The microarray samples were obtained from the Gene Expression Omnibus (GEO) dataset database and the gene expression profiles were downloaded. GEO is a user-friendly database allowing users to search for datasets in published sources as well as allowing users to access and download the corresponding data. Once the dataset was downloaded, RStudio was used to process the microarray data samples, an open source used for statistical computations and graphics; it is a front-end interface allowing for easy data input, analysis, and visualization. Once the differentially expressed genes were manually paired with miRNAs and lncRNAs, each list was filtered, and uploaded into Cytoscape v3.4.0. Cytoscape is also an open source used for network visualization. The differentially expressed genes' list was uploaded into Database for Annotation, Visualization and Integrated Discovery (DAVID 6.8) for functional annotation.

3.3. Databases

GeneCards was used to perform functional annotation of the differentially expressed genes as well as to find the miRNAs that target the respective gene. miRTarBase is an experimentally verified miRNA target interaction database. miRTarBase was used to pair miRNAs that target the differentially expressed genes. NONCODE is a database that annotates lncRNAs. NONCODE was used to manually pair the differentially expressed genes with lncRNAs. starBase v2.0 is a lncRNA database used to make miRNA-lncRNA target interactions. starBase v2.0 was used to create miRNA-lncRNA target interactions. Target Explorer powered by Ingenuity is an online database used to convey biological content through providing interaction networks and biological pathways and relationships. Target Explorer was used to suggest relationships between miRNAs, lncRNAs, and differentially expressed genes.

Table 1. Table of open source databases used in this study.
This table contains a list of open source databases used in this study as well as their respective references.

Database	**Reference**
Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/
GeneCards	http://www.genecards.org/
miRTarBase	http://mirtarbase.mbc.nctu.edu.tw/
NONCODE2016	http://noncode.org/index.php
starBase v2.0	http://starbase.sysu.edu.cn/
Target Explorer-Ingenuity	https://targetexplorer.ingenuity.com/index.htm
DAVID 6.8	https://david.ncifcrf.gov/

.

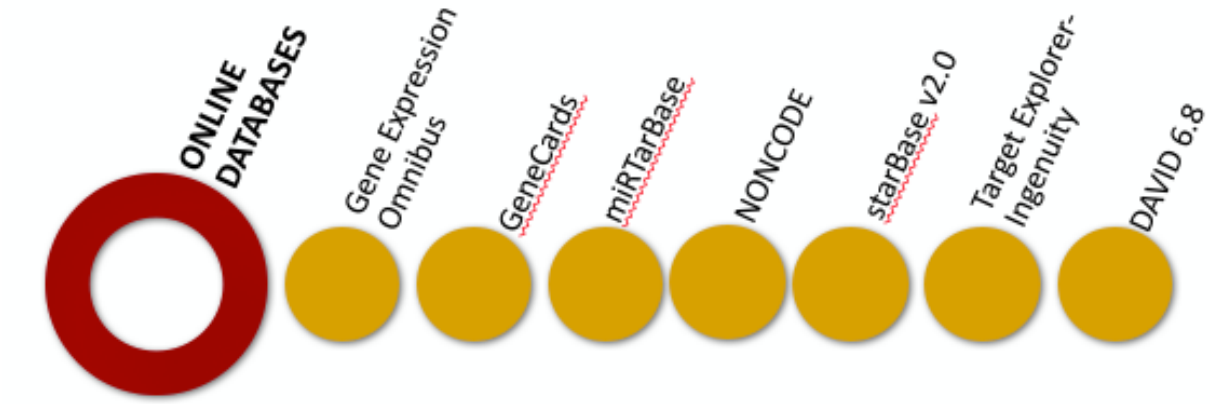


Figure 13. Schematic overview of online databases used in this study.
This overview shows the open source and validated online databases used in this computational study.

3.4. Samples

Samples were retrieved and downloaded from GEO using the following accession number: GSE44077. This dataset included 226 total samples comprising a 986.1 Mb file. This dataset included 110 NSCLC samples and 116 NSCLC airway samples. Of the 226 samples, 110 NSCLC samples were chosen for this study. Selection criteria for the samples was based on the location profile of the sample; different location profiles of this dataset included matched NSCLC tumors and airway epithelia with varying distances from the tumor. In this study, only matched NSCLC tumors were studied. The raw expression data were in .CEL format. Further processing was done for the downloaded samples. The samples were corrected and normalized using Robust Multi-Array Average (RMA) and transformed using log base 2 in order to reveal the values of the expression data to allow for further analysis of the differentially expressed genes.

3.5. Sample Characteristics

This study contained matched samples from NSCLC patients. The two groups studied include: 55 samples from NSCLC patients and 55 samples from NSCLC patients with normal

lungs. NSCLC patients may have adjacent lung tissue to the tumor that appears normal at an early diagnosis stage or the adjacent normal lung tissue may not have been effected by radiation or other treatments (American Cancer Society). The remaining samples from the dataset include: lung cancer patients' samples and airway samples which are irrelevant to this study.

3.6. Sample Processing

The samples were processed using R version 3.3.2. The data was uploaded in .txt format and named DATA. Csv. This file contains all the samples including the NSCLC samples (n=55) and NSCLC with normal lungs' (n=55) samples. The NSCLC samples correspond to GSM1077844,GSM1077846,GSM1077848,GSM1077853,GSM1077855,GSM1077857,GSM1077864,GSM1077866,GSM1077868,GSM1077873,GSM1077875,GSM1077877,GSM1077892,GSM1077895,GSM1077896,GSM1077902,GSM1077904,GSM1077906,GSM1077911,GSM1077913,GSM1077915,GSM1077922,GSM1077924,GSM1077926,GSM1077933,GSM1077935,GSM1077937,GSM1077944,GSM1077946,GSM1077948,GSM1077952,GSM1077954,GSM1077956,GSM1077963,GSM1077965,GSM1077967,GSM1077973,GSM1077981,GSM1077984,GSM1077985,GSM1077991,GSM1077993,GSM1077995,GSM1078001,GSM1078003,GSM1078005,GSM1078010,GSM1078016,GSM1078018,GSM1078020,GSM1078027,GSM1078029,GSM1078031,GSM1078061,GSM1078063. The NSCLC with normal lungs' samples correspond to GSM1077845,GSM1077847,GSM1077849,GSM1077854,GSM1077856,GSM1077858,GSM1077865,GSM1077867,GSM1077869,GSM1077874,GSM1077876,GSM1077878,GSM1077893,GSM1077894,GSM1077903,GSM1077905,GSM1077907,GSM1077912,GSM1077914,GSM1077916,GSM1077923,GSM1077925,GSM1077927,GSM1077934,GSM1077936,GSM1077938,GSM1077945,GSM1077947,GSM1077949,GSM1077953,GSM1077955,GSM1077957,GSM1077964,GSM1077966,GSM1077968,GSM1077974,GSM1077975,GSM1077976,GSM1077982,GSM

1077983,GSM1077986,GSM1077992,GSM1077994,GSM1077996,GSM1078002,GSM1078004,GSM1078006,GSM1078011,GSM1078012,GSM1078017,GSM1078019,GSM1078021,GSM1078028,GSM1078030,GSM1078032. This data was saved for further processing.

3.7. Sample Processing: Differential Expression of Genes

A log2 histogram was created to visualize the distribution of the expression value means of the samples. The values on the extremes of the plot represent the differentially expressed genes. A box plot was created to determine if the NSCLC samples and the normal samples are comparable. Box plots are used to compare different variables of a group of data samples in order to show a trend among the dataset. This means that a boxplot shows distribution values for selected samples for a dataset; a box plot specifically shows if the values selected are median centered. A centered configuration of the median line indicates samples are fit for comparison. A cluster dendogram was created to group the samples according to similar characteristics. Cluster dendograms group samples based on correlation coefficients of the expression values. This illustrates the samples in nodes and branches. A scatter plot was created to visualize the relationship between the two groups of samples: NSCLC patients and NSCLC patients with normal lungs. Points that lay farther from the red line represent the differentially expressed genes. Biological and statistical significance were defined using a fold cut-off value = 2 and a p-value = 0.01, respectively. A volcano plot was produced to present the biological (fold cut-off value) and statistical significance (p-value) in one graph. The fold cut-off correlates to the difference between the means of the conditions, the differences between the differentiated genes and the undifferentiated genes. The larger the fold cut-off the more significant the data. The p-value correlates to the likelihood of finding significance by chance. In other words, defining a low p-value means that the significance is not by chance. A heatmap was also produced to

visualize the level of expression of the hierarchically clustered differentially expressed genes; red represents the over-expressed genes and blue represents the under-expressed genes.

The differentially expressed genes were manually inputted into GeneCards to perform functional annotation. The differentially expressed genes were also manually paired with miRNAs using miRTarBase. The miRNAs were filtered using literature associated with NSCLC as well as choosing miRNAs that have a strong evidence validation method (i.e. reporter assay, western blot, and qPCR). The differentially expressed genes were also manually paired with lncRNAs using NONCODE. The lncRNAs were filtered using literature associated with NSCLC.

3.8. Systems Approach

Filtered miRNAs and their matched differentially expressed genes were organized into a table and uploaded into Cytoscape version 3.4.0. The table was created into a text file and used to link the filtered miRNAs to the differentially expressed genes to create a gene regulatory network. A network was created to visualize the manually paired miRNAs with their respective differentially expressed genes.

lncRNA associated with NSCLC were paired with the filtered miRNAs using starBase v2.0, organized into a table, and uploaded into Cytoscape. The table was also created into a text file used to link the filtered miRNAs to already NSCLC-associated lncRNAs to create a regulatory network. Another regulatory network was created to connect the filtered miRNAs and the differentially expressed genes from this dataset to the NSCLC-associated lncRNAs. Another table was created that included the filtered miRNAs, their associated differentially expressed genes, and their associated NSCLC-associated lncRNAs, converted into a text file, and uploaded

into Cytoscape. This network was created to visualize connections between the filtered miRNAs, NSCLC-associated lncRNAs, and differentially expressed genes.

All of the differentially expressed genes were uploaded with the official gene symbol into DAVID 6.8 in a list format and were functionally annotated. DAVID clustered the genes based on similarly related genes into functionally related groups. This was used for further interpretation.

Target Explorer was used to provide more comprehensive results. Selected differentially expressed gene symbols, miRNAs, and lncRNAs were entered into this online biomedical toolkit to determine biological relationships. These biological relationships were used to suggest regulatory relationships between miRNAs, lncRNAs, and differentially expressed genes. It helped confirm regulatory targets and suggest other biological functions. This will also help suggest further molecular target experiments.

Figure 14. Schematic overview of study design.
This study design shows the steps taken to create regulatory networks that connect differentially expressed genes, miRNAs, and NSCLC-associated lncRNAs in NSCLC.

CHAPTER 4: Results

4.1. Data Validation and Quality Check for Samples in this Study

The 110 NSCLC samples in this study were obtained from GEO with the following accession number GSE44077 (Figure 15). A log2 histogram, box plot, and cluster dendogram were created for sample analysis and quality assurance of results. In addition, a scatter plot, volcano plot, and heatmap were created in order to allow for further analysis. Differentially expressed genes were identified and used to create gene regulatory networks to their respective miRNAs and lncRNAs.

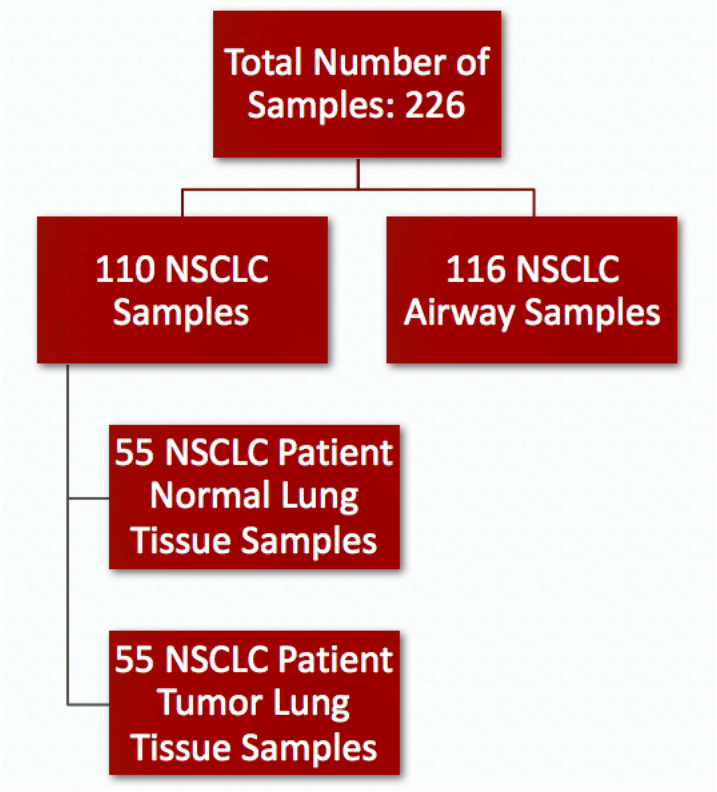


Figure 15. Flowchart of samples chosen from GEO (GSE44077).
A total of 226 samples were acquired from GSE44077. 55 NSCLC patient samples and 55 NSCLC patients with normal lung tissue samples were chosen for this study and further analyzed. The remaining samples from the dataset include: lung cancer patients' samples and airway samples which are irrelevant to this study.

4.1.1. Testing for Data Distribution of Samples

Expression data from different data sets were collected. The data were first pre-processed to allow for appropriate comparison. First, the expression values were log-2 transformed then the distribution of the log-2 transformed data was examined. The log2 transformed histogram of the data was produced to check the behavior of the data and to visualize the distribution of the expression value means. The advantage of taking the log2 allows the user to compare the expression value means and take the ratio between the means. The distribution of the expression values in the log2 histogram is not evenly distributed nor symmetric (Figure 16). The log2 transformation of this dataset skews the bell curve formation of the histogram to the left. The left tail of the histogram suggests differentially expressed genes that are under-expressed and the right tail of the histogram suggest differentially expressed genes that are over-expressed. In this study, most of the differentially expressed genes appear to be over-expressed. This was later confirmed with the volcano plot (Figure 20).

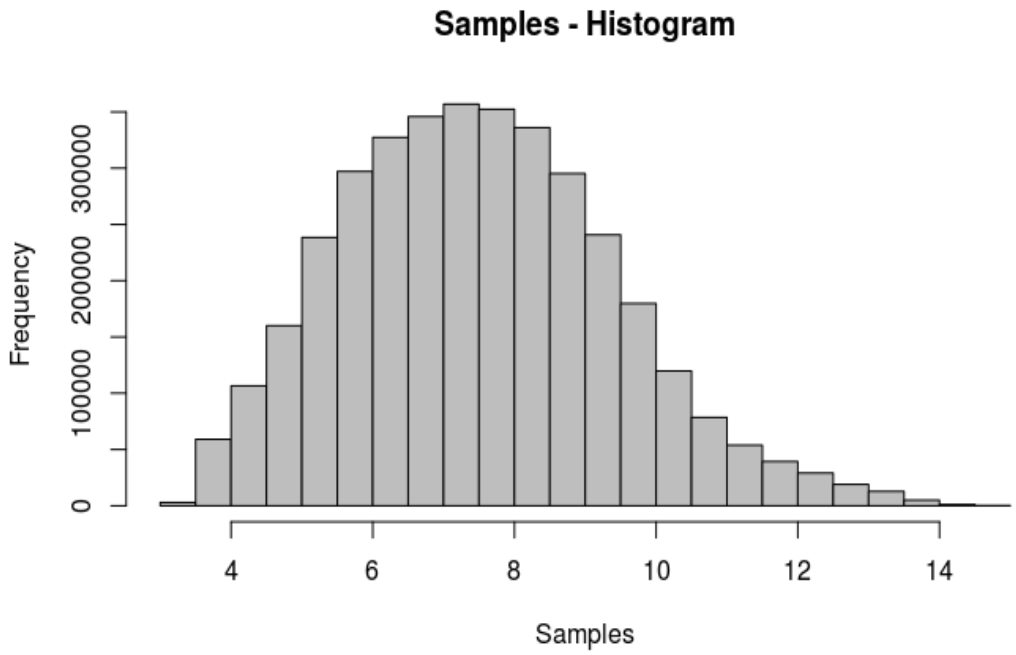


Figure 16. Log2 histogram of the NSCLC and normal lung tissue samples.
The distribution of the expression values is not evenly distributed nor is it symmetric. The histogram is bell-shaped curved and skewed to the left. The left tail of the histogram suggests differentially expressed genes that are under-expressed and the right tail of the histogram suggest differentially expressed gene that are over-expressed in NSCLC and normal lung tissue samples. This Figure was produced by RStudio using DATA.csv.

4.1.2. Testing for Data Uniformity of Samples

All the green box plots represent the normal lung tissue samples from this dataset and all the red box plots represent the NSCLC samples from this dataset. This box plot shows an almost even distribution of values which suggests that these samples are fit for comparison (Figure 17).

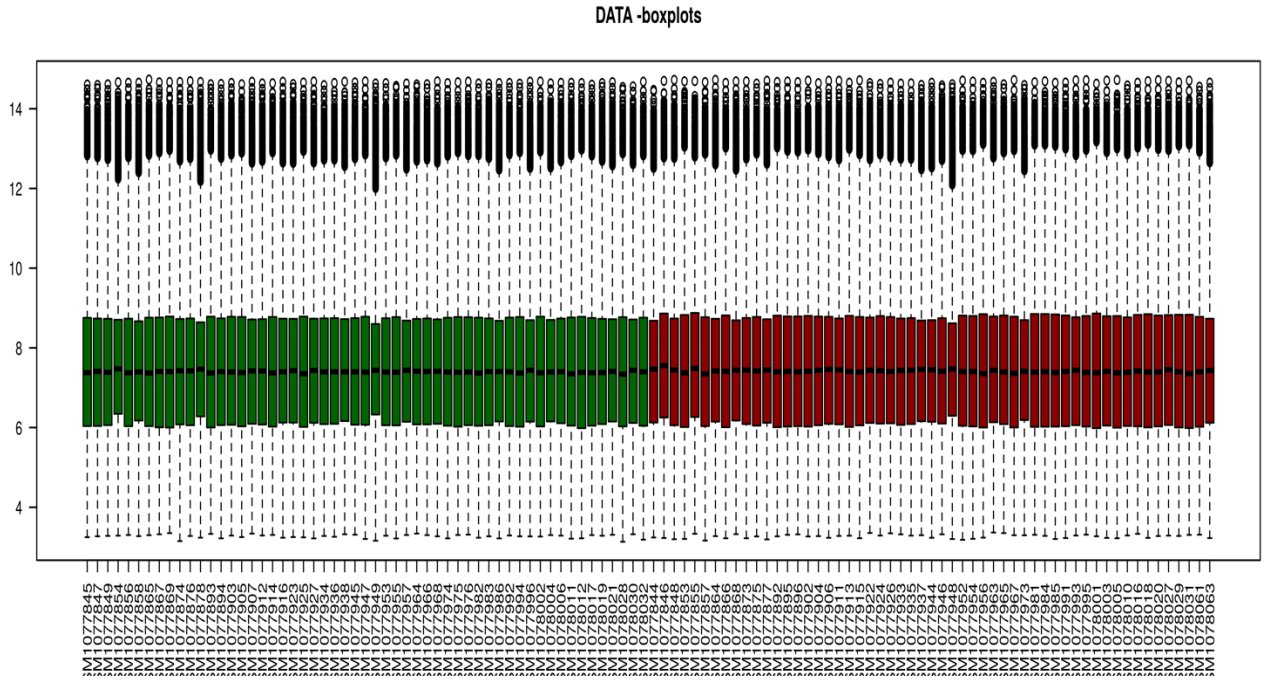


Figure 17. Box plot of the NSCLC and normal lung tissue samples.
The green box plots represent the normal lung tissue samples from the GSE44077 dataset and all the red box plots represent the NSCLC samples from this dataset. An almost even distribution of median values is observed among the samples deeming the samples fit for comparison. This Figure was produced by RStudio using DATA.csv.

4.1.3. Classifying Samples by Cluster Analysis

The hierarchical clustering shows that the characteristically related samples were grouped together. All the NSCLC and normal lung tissue samples were grouped together except for three NSCLC samples; these three NSCLC samples were also grouped together (Figure 18). There were two major clusters present: NSCLC and NSCLC normal lung samples. A third cluster was also present that represented three NSCLC samples; this remote cluster may be a result of a misdiagnosis or mislabeling.

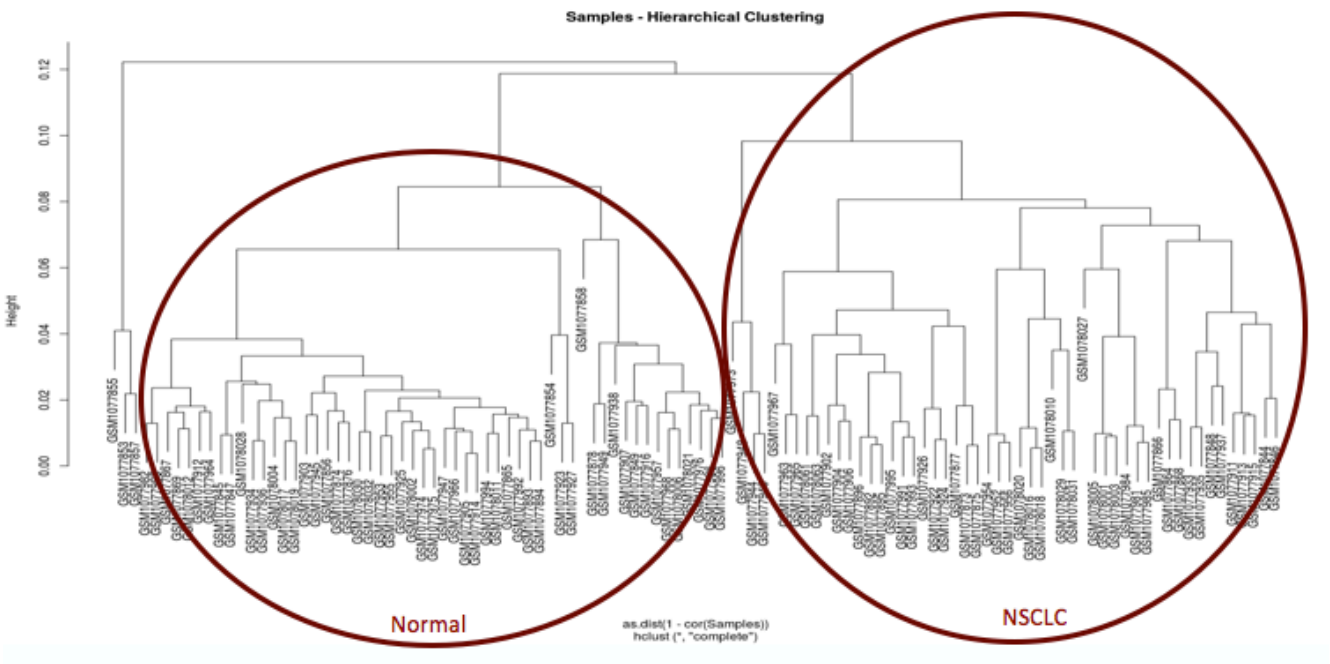


Figure 18. Cluster Dendogram of the NSCLC and normal lung tissue samples.
All the NSCLC and normal lung tissue samples were grouped together except for three NSCLC samples; these three NSCLC samples were also grouped together. All characteristically similar samples were grouped together. This Figure was produced by RStudio using DATA.csv.

4.2. Identification of Biologically and Statistically Significant Differentially Expressed Genes

4.2.1. Genomic Expression Levels Between NSCLC and Normal Lung Samples

The scatter plot produced shows the relationship between two variables (Figure 19). The relationship between the NSCLC samples and the NSCLC normal lung samples is roughly linear (Figure 19). The majority of the data samples have similar genomic correlations which is indicated by the overlap of data points. Each gene has a specific expression value denoted by an individual data point. Data points closer to the line indicate similar gene expression. Data points above the line are over-expressed and data points below the line are under-expressed. The outliers on the scatter plot, lie further from the line and indicate variable expression differences

in the data. In this study these data points represented the differentially expressed genes and the differentially expressed genes were further analyzed.

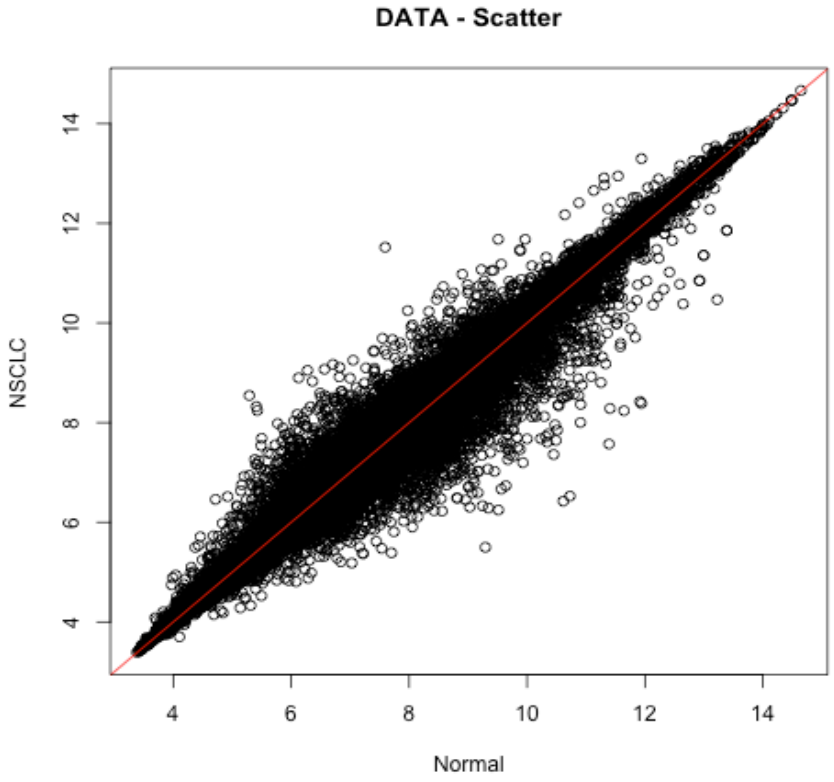


Figure 19. Scatter Plot showing gene expression between NSCLC patients and NSCLC patients with normal lungs.
Scatter plot of the NSCLC cancerous lung samples and the NSCLC normal lung samples was illustrated by RStudio using DATA.csv. The relationship is roughly linear. Overlap of data points indicate similar gene expression. Data points above the line are over-expressed and data points below the line are under-expressed.

4.2.2. Determining Biologically and Statistically Significant Differentially Expressed Genes

A common approach in bioinformatics analysis of microarray is to combine statistical analysis with negative-log expression ratios. A typical visualization tool for this two-dimensional analysis is the volcano plot (Figure 20). The volcano plot measures the fold cut-off, biological

significance, on the x-axis and the p-value, statistical significance, on the y-axis. The more statistically significant data appear higher in the graph and the upregulated data and the downregulated data appear to the left and right side of the graph, respectively. More specifically, data points to the right of the red line and above the green line represent statistically and biologically significant upregulated differentially expressed genes. Data points to the left of the blue line and above the green line represent statistically and biologically downregulated differentially expressed genes. Applying such analysis to our dataset resulted in 108 differentially expressed genes, of which 105 are known genes of biological and statistical significance in this dataset; two genes lack significant found similarities and one gene is of unknown function. Most of the differentially expressed genes in this study are upregulated (Figure 20). Figure 20 also reiterates the findings in Figure 16.

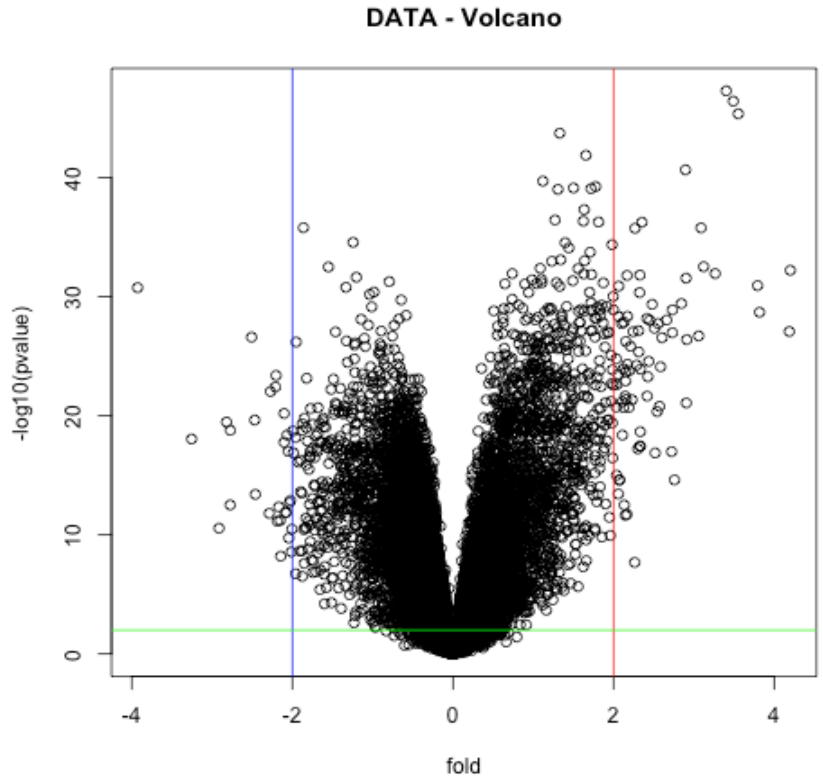


Figure 20. Volcano Plot demonstrating the relationship between NSCLC patients and NSCLC patients with normal lungs.

The volcano plot shows a relationship between the biological (fold cut-off =2) and statistical (p-value=0.01) significance in one graph. Genes below the horizontal green line are considered not significant. Genes to the right of red line are upregulated and genes to the left of the blue line are downregulated. Genes of interest lie in the upper right sector and the upper left sector. 108 differentially expressed genes were identified and most of these genes are upregulated. This Figure was illustrated by RStudio using DATA.csv.

4.2.3. Visualizing Expression Patterns of Differentially Expressed Genes

Hierarchical clustering of the differentially expressed genes (n=108) resulted into 2 clusters and the samples (n=110) into 3 clusters as previously determined from Figure 18 (Figure 21). This heat map showed expression values of the differentially expressed genes in the different NSCLC sample type. The over-expressed genes are depicted in red and the under-expressed genes are depicted in blue. The identified differentially expressed genes that are over-expressed in cancer samples are under-expressed in normal samples.

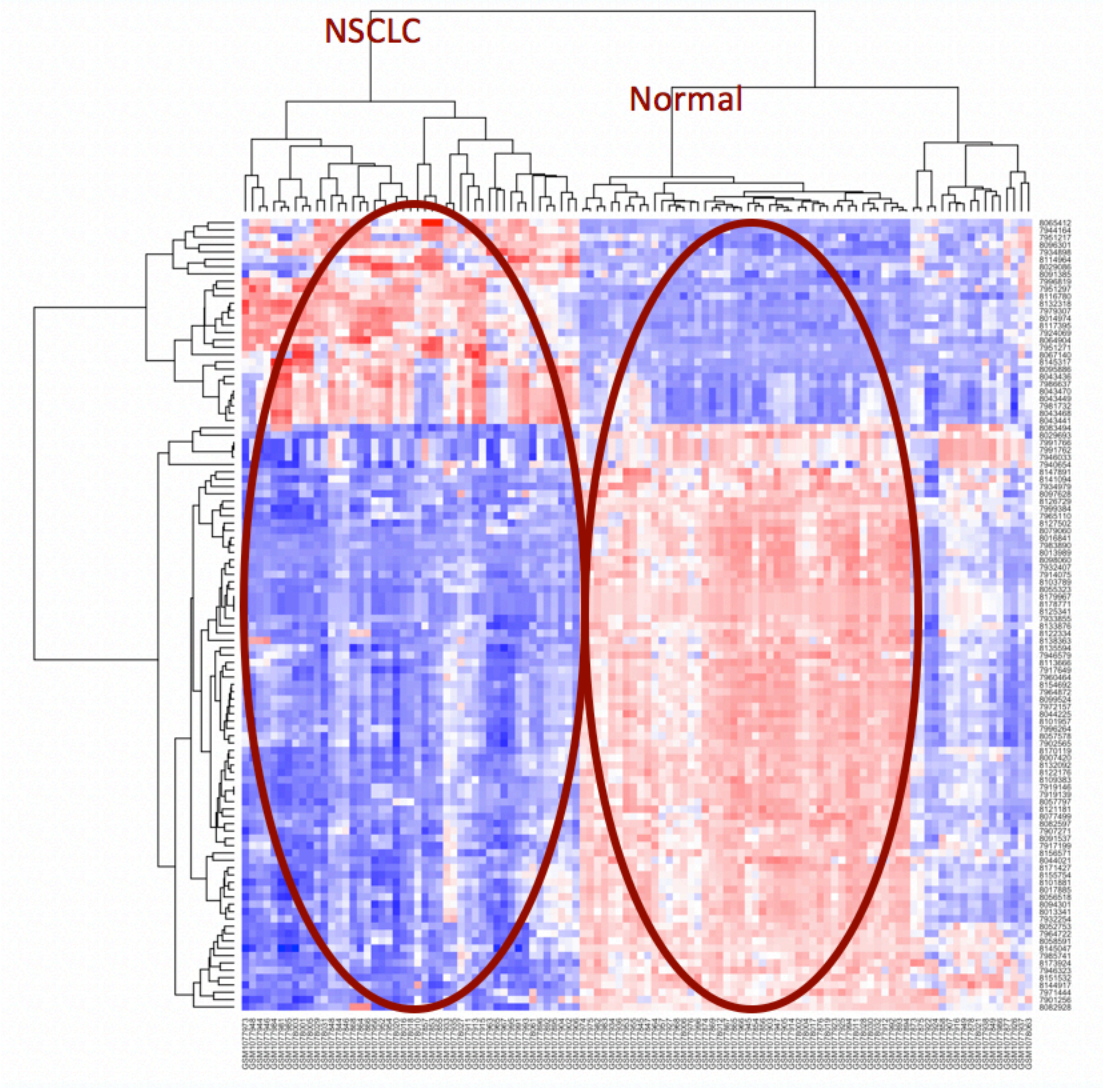


Figure 21. Heatmap exhibiting expression pattern of identified differentially expressed genes. *The heatmap hierarchically clusters the differentially expressed genes (n=108) and the samples (n=110). Red represents the over-expressed genes and blue represents the under-expressed genes. The differentially expressed genes are labeled on the y-axis and divided into 2 clusters and the samples are labeled on the x-axis and divided into 3 clusters. Samples are outlined in columns and the differentially expressed genes are depicted in rows. This Figure was illustrated by RStudio using DATA.csv.*

4.2.4. Functional Annotation of Differentially Expressed Genes of Biological and Statistical Significance

There were a total of 33,252 genes in this dataset, but only 108 differentially expressed genes of biological and statistical significance were identified (Table 2). They were further

analyzed using online databases to create regulatory networks to connect miRNAs, NSCLC-associated lncRNAs, and differentially expressed genes.

Table 2. Table of differentially expressed genes identified.

The 108 differentially expressed genes (DEGs) are listed and annotated with their respective Affymetrix gene ID and their respective official gene symbol. There are 105 known differentially expressed genes, 2 genes that lack significant hits., 1 unknown gene.

AFFYMETRIX_ID	Name	Gene_Symbol
8055323	NCK associated protein 5	NCKAP5
8099524	LIM domain binding 2	LDB2
7917649	Transforming growth factor beta receptor 3	TGFBR3
8095886	C-X-C motif chemokine ligand 13	CXCL13
8056518	Sodium voltage-gated channel alpha subunit 7	SCN7A
8133876	CD36 molecule	CD36
8145047	Surfactant protein C	SFTPC
8094301	Slit guidance ligand 2	SLIT2
7932407	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 6	ST8SIA6
7971444	Carboxypeptidase B2	CPB2
7972157	Endothelin receptor type B	EDNRB
8064904	Fermitin family member 1	FERMT1
7932254	Integrin subunit alpha 8	ITGA8
8091385	Ceruloplasmin	CP
8082928	Claudin 18	CLDN18
7907271	Flavin containing monooxygenase 2	FMO2

8156571	Chromosome 9 open reading frame 3	C9orf3
8044225	Sulfotransferase family 1C member 4	SULT1C4
7979307	DLG associated protein 5	DLGAP5
8113666	Semaphorin 6A	SEMA6A
8122176	Transcription factor 21	TCF21
8147891	Polycystic kidney and hepatic disease 1 autosomal recessive-like 1	PKHD1L1
7985741	ATP/GTP binding protein like 1	AGBL1
7934898	Ankyrin repeat domain 22	ANKRD22
8103789	Glycoprotein M6A	GPM6A
8127502	Long intergenic non-protein coding RNA 472	LINC00472
8058591	Acyl-CoA dehydrogenase, long chain	ACADL
7933855	Rhotekin 2	RTKN2
8013341	Microfibrillar associated protein 4	MFAP4
7944164	Transmembrane protease, serine 4	TMPRSS4
8044021	Interleukin 1 receptor like 1	IL1RL1
8013989	Solute carrier family 6 member 4	SLC6A4
8029693	FosB proto-oncogene, AP-1 transcription factor subunit	FOSB
7914075	Ficolin 3	FCN3
8016841	Transmembrane protein 100	TMEM100
8114964	Serine peptidase inhibitor, Kazal type 1	SPINK1
8082597	Collagen type VI alpha 6 chain	COL6A6

8126729	Chloride intracellular channel 5	CLIC5
8057797	Serum deprivation response	SDPR
8179967	Advanced glycosylation end-product specific receptor	AGER
8125341	Advanced glycosylation end-product specific receptor	AGER
8178771	Advanced glycosylation end-product specific receptor	AGER
8017885	ATP binding cassette subfamily A member 8	ABCA8
8096301	Secreted phosphoprotein 1	SPP1
7901256	Cytochrome P450 family 4 subfamily B member 1	CYP4B1
8141094	Pyruvate dehydrogenase kinase 4	PDK4
8132318	Anillin actin binding protein	ANLN
7946033	Hemoglobin subunit beta	HBB
8077499	Long intergenic non-protein coding RNA 312	LINC00312
8097628	Hedgehog interacting protein	HHIP
7940654	Secretoglobin family 1A member 1	SCGB1A1
7951297	Matrix metallopeptidase 12	MMP12
7996819	Cadherin 3	CDH3
8138363	Sclerostin domain containing 1	SOSTDC1
8029086	Carcinoembryonic antigen related cell adhesion molecule 5	CEACAM5
7996264	Cadherin 5	CDH5
8144917	Lipoprotein lipase	LPL

7964722	WNT inhibitory factor 1	WIF1
8132092	INMT-FAM188B readthrough	NMD candidate
8116780	Desmoplakin	DSP
8067140	Cytochrome P450 family 24 subfamily A member 1	CYP24A1
8171427	PIR-FIGF readthrough	PIR-FIGF
7981732	Immunoglobulin heavy variable 4-31	IGHV4-31
8145317	ADAM like decysin 1	ADAMDEC1
8117395	Histone cluster 1 H2B family member f	HIST1H2BF
8091537	Immunoglobulin superfamily member 10	IGSF10
7964872	Protein tyrosine phosphatase, receptor type B	PTPRB
8151532	Fatty acid binding protein 4	FABP4
8079060	Vasoactive intestinal peptide receptor 1	VIPR1
8154692	TEK receptor tyrosine kinase	TEK
7983890	Myocardial zonula adherens protein	MYZAP
8007420	Amine oxidase, copper containing 3	AOC3
8052753	Gastrokine 2	GKN2
7946323	Olfactory receptor family 5 subfamily P member 2	OR5P2
8109383	Glutamate ionotropic receptor AMPA type subunit 1	GRIA1
8135594	Caveolin 1	CAV1
8098060	Relaxin/insulin like family peptide receptor 1	RXFP1
8122334	Atypical chemokine receptor 4	ACKR4

8057578	Calcitonin receptor like receptor	CALCRL
8065412	Cystatin SN	CST1
7960464	Von Willebrand factor	VWF
7917199	Tubulin tyrosine ligase like 7	TTLL7
8155754	MAM domain containing 2	MAMDC2
7946579	Lymphatic vessel endothelial hyaluronan receptor 1	LYVE1
7951217	Matrix metallopeptidase 7	MMP7
8121181	Four and a half LIM domains 5	FHL5
7951271	Matrix metallopeptidase 1	MMP1
7991762	Hemoglobin subunit alpha 1	HBA1
7991766	Hemoglobin subunit alpha 1	HBA1
7919139	Ankyrin repeat domain 20 family member A12, pseudogene	ANKRD20A12P
7919146	Ankyrin repeat domain 20 family member A12, pseudogene	ANKRD20A12P
8014974	Topoisomerase DNA II alpha	TOP2A
7934979	Ankyrin repeat domain 1	ANKRD1
8101957	Endomucin	EMCN
8043468	Immunoglobulin kappa constant	IGKC
8043436	Immunoglobulin kappa constant	IGKC
8043449	Immunoglobulin kappa constant	IGKC
8083494	Membrane metalloendopeptidase	MME
8101881	Alcohol dehydrogenase 1B class 1, beta	ADH1B

	polypeptide	
7902565	Adhesion G protein-coupled receptor L2	ADGRL2
8170119	Four and a half LIM domains 1	FHL1
7965110	Noncoding transcript identified by NONCODE	NONHSAT029647
7924069	RNA, U5A small nuclear 8, pseudogene	RNU5A-8P
8043441	Immunoglobulin Kappa Variable 1D-27, pseudogene	IGKV1D-27
7999384	Noncoding transcript identified by NONCODE	NONHSAT140505
7986637	Immunoglobulin Heavy Variable 1/OR15-1, nonfunctional	IGHV1OR15-1
8043470	Immunoglobulin Kappa Variable 3D-11	IGKV3D-11
8173924	Unknown noncoding sequence	Unknown

4.3. Systems Approach: Creating Regulatory Networks

4.3.1. Connecting Differentially Expressed Genes and miRNAs via Interaction Networks

Six hundred eighty-one miRNAs were matched to the differentially expressed genes. The miRNAs were filtered according to literature and miRTarBase's strong evidence validation methods (reporter assay, western blot, and qPCR). The filtration narrowed the original list to 66 miRNAs (Table 3). The regulatory network created in Cytoscape shows many overlapping connections between the miRNAs and the differentially expressed genes and indicates a many-to-many systemic analysis relationship (Figure 22). This network was constructed on Cytoscape and exhibits the filtered miRNAs that are known to target these differentially expressed genes.

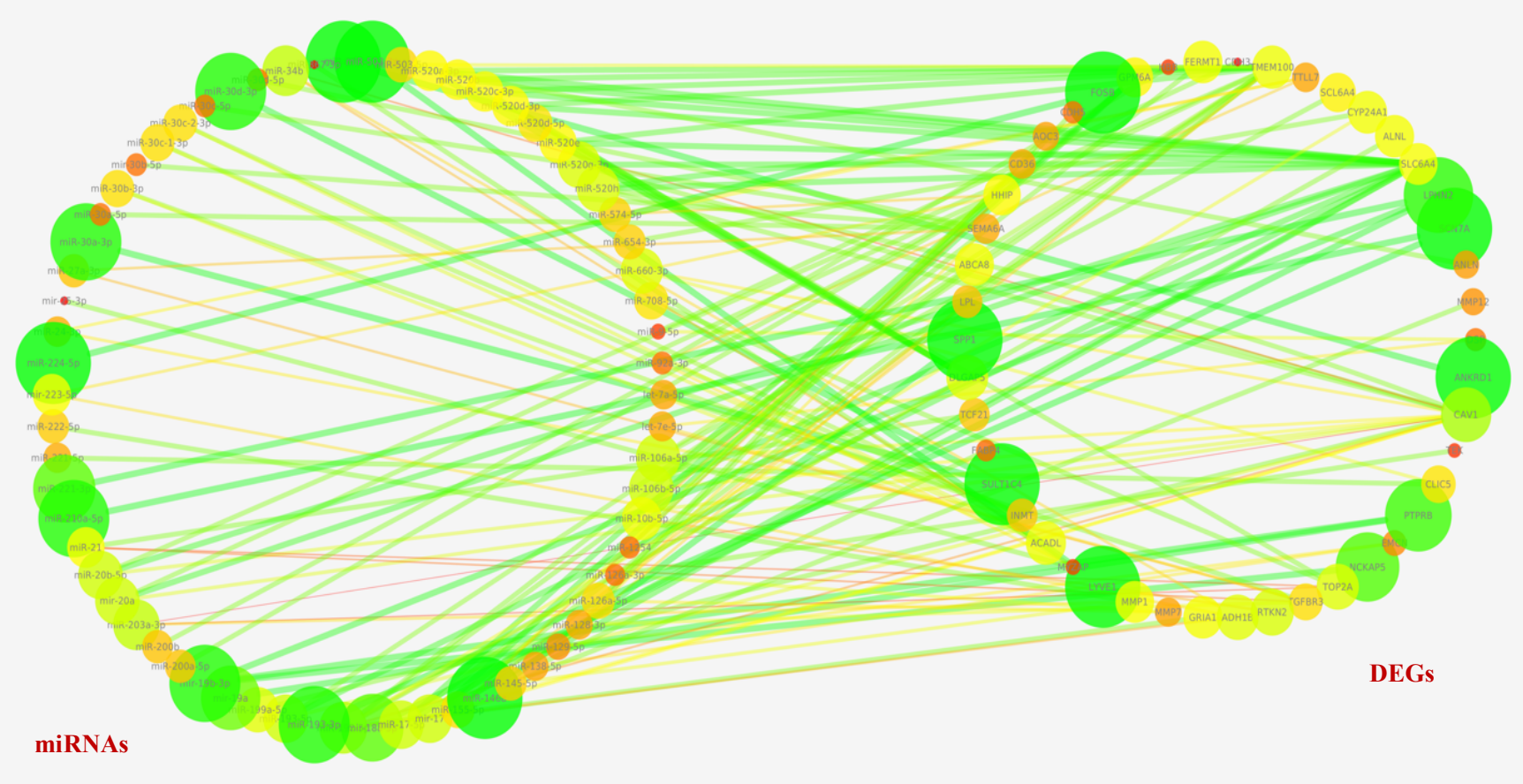


Figure 22. Regulatory network of miRNAs and differentially expressed genes.

This network shows many overlapping and many-to-many relationships, specifically, it shows the miRNAs that target the differentially expressed genes (DEGs). Table 3 represents the annotations in this interaction network. Table 3 below was uploaded into Cytoscape in .csv format and Cytoscape was used to create this network. The colors and sizes of the circles and the lines do not have significant meaning.

Table 3. Table of filtered regulatory miRNAs paired with DEGs.

This table annotates 66 of the 681 miRNAs and their respective differentially expressed gene target from this dataset. These miRNAs were filtered based on literature and miRTarBase's strong validation methods (i.e. reporter assay, western blot, qPCR).

miRNA	**Gene Symbol**
mir-223-5p	HHIP, ADH1B
mir-25-3p	MYZAP
miR-654-3p	CYP24A1
miR-520h	CAV1, SLC6A4
miR-34b	ADH1B, CAV1, ACADL
miR-200b	HHIP
miR-138-5p	LPL, FABP4, EMCN
miR-503-5p	ANLN, TMEM100
miR-503-3p	LYVE1
miR-145-5p	MMP1, MMP12
miR-222-5p	MMP1
miR-367-3p	MYZAP
mir-18a-3p	LPHN2
mir-19a	PTPRB
mir-20a	SLC6A4, CAV1, TMEM100
mir-19b-3p	NCKAP5, LPHN2, PTPRB
miR-21	TGFBR3, TOP2A, TCF21
miR-106a-5p	SCL6A4, TMEM100, CAV1
miR-146a	SPP1
miR-155-5p	HHIP, CD36, LPL
miR-192-5p	CAV1, RTKN2, DLGAP5, FERMT1, CYP24A1, GRIA1, ALNL, ABCA8
miR-203a-3p	CAV1, MMP1, TOP2A
miR-210a-5p	SCN7A
miR-9-5p	DSP, CDH3
miR-708-5p	TOP2A
miR-375	ANKRD1
miR-126a-3p	TEK, MMP7
miR-126a-5p	MMP7, AOC3, TMEM100

miR-30d-3p	SULT1C4
miR-30d-5p	SEMA6A
miR-129-5p	EMCN, SEMA6A
miR-128-3p	TGFBR3, CDH5
miR-30b-3p	ACADL, INMT
mir-30b-5p	SEMA6A
miR-30c-5p	SEMA6A
miR-30c-1-3p	ACADL, INMT
miR-30c-2-3p	INMT, ACADL
miR-520c-3p	DLGAP5, SLC6A4, TMEM100
miR-520e	DLGAP5, SLC6A4, TMEM100
miR-520b	DLGAP5, SLC6A4, TMEM100
miR-520d-3p	DLGAP5, SLC6A4, TMEM100
miR-520a-3p	DLGAP5, SLC6A4, TMEM100
miR-520h	CAV1, SLC6A4
miR-520g-3p	CAV1, SLC6A4
miR-520d-5p	TOP2A
miR-17-5p	TMEM100, GPM6A, CAV1, SLC6A4
miR-200a-5p	HHIP
miR-106b-5p	SLC6A4, TMEM100, GPM6A, CAV1
miR-193-5p	SLC6A4, TMEM100, CAV1
miR-193-3p	SCN7A
miR-20b-5p	SLC6A4, TMEM100, CAV1
miR-224-5p	FOSB
let-7e-5p	DSP, TGFBR3
miR-221-5p	CLIC5
miR-221-3p	LPHN2
let-7a-5p	TGFBR3
miR-27a-3p	SEMA6A, TGFBR3
miR-10b-5p	HHIP, RTKN2
miR-1254	CD36
miR-574-5p	HHIP, TTLL7
miR-24-3p	INMT, TTLL7

miR-199a-5p	CAV1
miR-660-3p	CAV1, CLIC5
miR-92a-3p	TTLL7, HBB, MYZAP
miR-30a-3p	SULT1C4
miR-30a-5p	SEMA6A

4.3.2. Connecting NSCLC-associated lncRNAs and miRNAs via Interaction Networks

The NSCLC-associate lncRNAs include: MALAT1, PVT1, HOTAIR, H19, TUG1, GAS5, and DLX6-AS1. These 7 lncRNAs were chosen from literature (Table 4) (Wei & Zhou, 2016). Most of these NSCLC-associated lncRNAs are oncogenes (Wei & Zhou, 2016). MALAT1, PVT1, HOTAIR, H19, and DLX6-AS1 are associated with oncogene functions. GAS5 and TUG1 are associated with tumor suppressor functions. This network shows many overlapping and many-to-many relationships between the filtered miRNAs and NSCLC-associated lncRNAs (Figure 23).

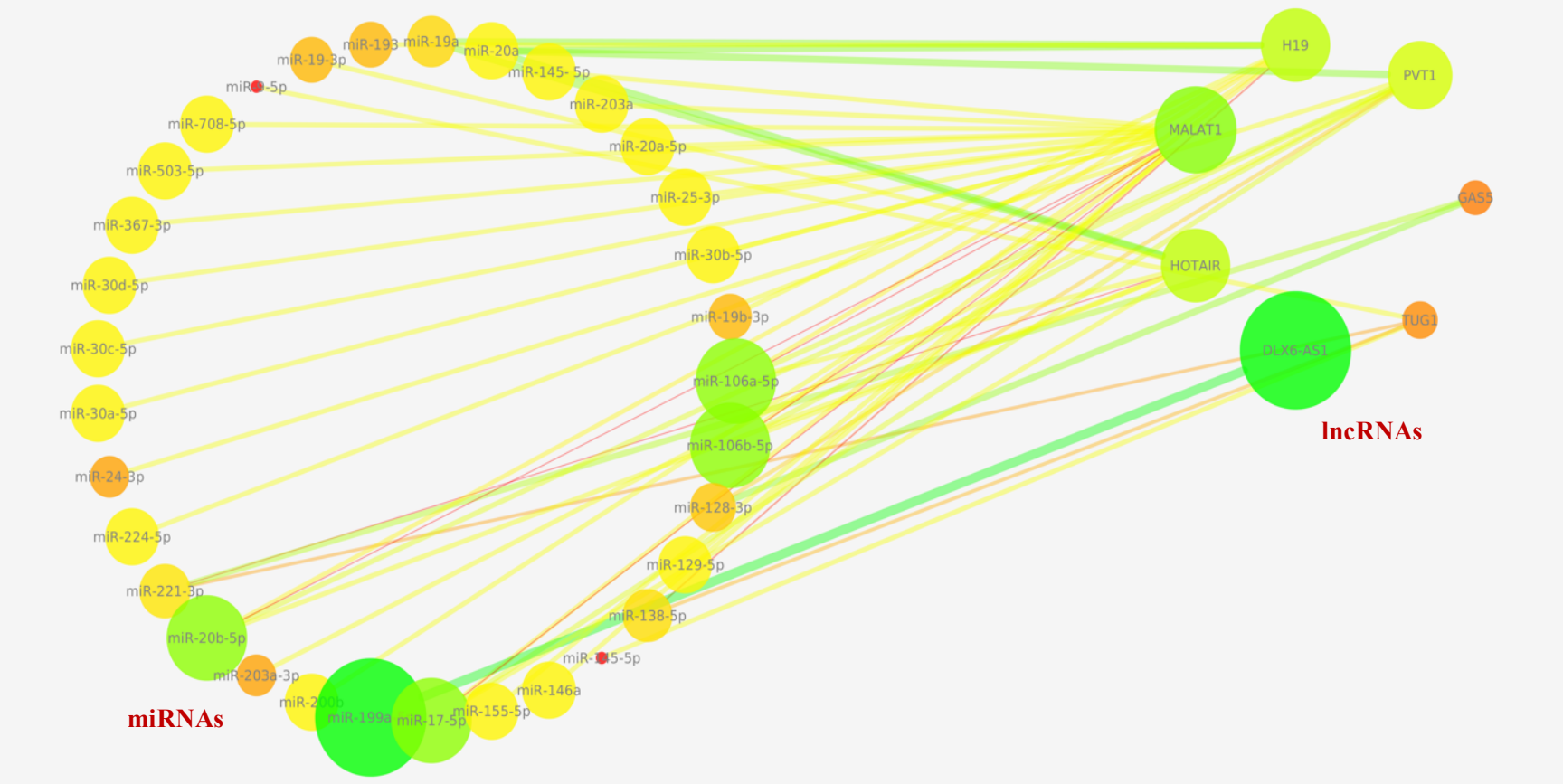


Figure 23. Regulatory network of NSCLC-associated lncRNAs and miRNAs.

This network shows many overlapping and many-to-many relationships specifically it shows the NSCLC- associated long non-coding RNAs (lncRNAs) that are known to target miRNAs. Table 4 represents the annotations in this interaction network. Table 4 below was uploaded into Cytoscape in .csv format and Cytoscape was used to create this network. The colors and sizes of the circles and the lines do not have significant meaning.

Table 4. Table of miRNA and lncRNA targets.

These seven long non-coding RNAs (lncRNAs) are associated with NSCLC according to literature (Wei & Zhou, 2016). These already NSCLC associated-lncRNAs were paired with miRNA targets already found in this dataset using the online starBase database.

lncRNA associated with NSCLC	**miRNA**
MALAT1	miR-200b, miR-30c-5p, miR-129-5p, miR-708-5p, miR-17-5p, miR-20a-5p, miR-203a, miR-155-5p, miR-367-3p, miR-145- 5p, miR-146a, miR-30a-5p, miR-25-3p, miR-106b-5p, miR-30b-5p, miR-30d-5p, miR-20b-5p, miR-106a-5p, miR-503-5p, miR-224-5p
H19	miR-17-5p, miR-19a, miR-20a, miR-19b-3p, miR-138-5p, miR-193, miR-106b-5p, miR-20b-5p, miR-106a-5p
TUG1	miR-9-5p, miR-138-5p, miR-145-5p, miR-221-3p
HOTAIR	miR-17-5p, miR-19a, miR-20a, miR-19-3p, miR-106b-5p, miR-221-3p, miR-20b-5p, miR-106a-5p
GAS5	miR-128-3p, miR-221-3p
PVT1	miR-17-5p, miR-20a, miR-203a-3p, miR-24-3p, miR-128-3p, miR-106b-5p, miR-20b-5p, miR-106a-5p
DLX6-AS1	miR-199a-5p

4.3.3. Connecting Differentially Expressed Genes, miRNAs, and NSCLC-associated lncRNAs via Interaction Networks

The regulatory network created shows many overlapping interactive connections between the miRNAs, NSCLC-associated lncRNAs, and differentially expressed genes (Figure 24). Figure 24, also indicates many-to-many relationship which combines are reiterates findings from Figure 22 and Figure 23. These many-to-many systemic analysis relationships observed in Figure 22, Figure 23, and Figure 24 means that these differentially expressed genes, miRNAs, and lncRNAs are intertwined and may be involved in NSCLC regulation. These overlapping interactions were further analyzed using Target Explorer.

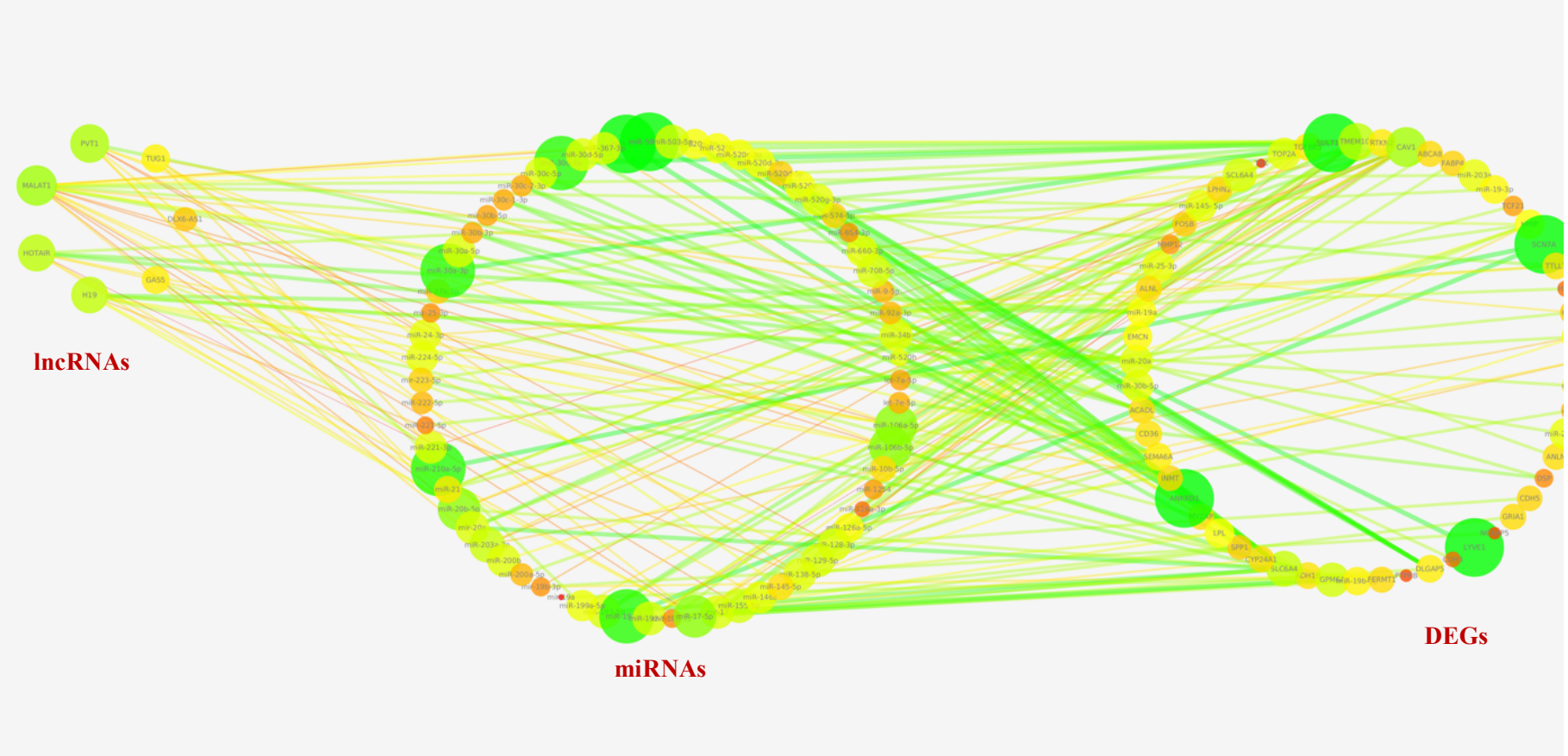


Figure 24. Regulatory network of miRNAs, NSCLC-associated lncRNAs, and differentially expressed genes.
The program used to create the network of NSCLC-associated lncRNAs, miRNAs, and differentially expressed genes (DEGs). This network shows many overlapping and many-to-many relationships. Table 5 represents the annotations in this interaction network. Table 5 below was uploaded into Cytoscape in .csv format and Cytoscape was used to create this network. The colors and sizes of the circles and the lines do not have significant meaning.

Table 5. Table of lncRNA, miRNA, and DEG targets.

This table annotates the 66 miRNAs with their respective identified differentially expressed gene (DEG) target from this dataset. The 66 miRNAs are also annotated with their respective NSCLC-associated lncRNAs targets.

lncRNA	**miRNA**	**Gene Symbol**
no matches found	mir-223-5p	HHIP, ADH1B
MALAT1	mir-25-3p	MYZAP
no matches found	miR-654-3p	CYP24A1
no matches found	miR-520h	CAV1, SLC6A4
no matches found	miR-34b	ADH1B, CAV1, ACADL
MALAT1	miR-200b	HHIP
H19, TUG1	miR-138-5p	LPL, FABP4, EMCN
MALAT1	miR-503-5p	ANLN, TMEM100
no matches found	miR-503-3p	LYVE1
TUG1, MALAT1	miR-145-5p	MMP1, MMP12
no matches found	miR-222-5p	MMP1
MALAT1	miR-367-3p	MYZAP
no matches found	mir-18a-3p	LPHN2
HOTAIR, H19	mir-19a	PTPRB
HOTAIR, H19, MALAT1, PVT1	mir-20a	SLC6A4, CAV1, TMEM100
H19, HOTAIR	mir-19b-3p	NCKAP5, LPHN2, PTPRB
TUG1	miR-21	TGFBR3, TOP2A, TCF21
H19, HOTAIR, MALAT1, PVT1	miR-106a-5p	SCL6A4, TMEM100, CAV1
MALAT1	miR-146a	SPP1
MALAT1	miR-155-5p	HHIP, CD36, LPL
other lncRNAs found	miR-192-5p	CAV1, RTKN2, DLGAP5, FERMT1, CYP24A1, GRIA1, ALNL, ABCA8
MALAT1, PVT1	miR-203a-3p	CAV1, MMP1, TOP2A
LINC00473	miR-210a-5p	SCN7A
TUG1	miR-9-5p	DSP, CDH3
MALAT1	miR-708-5p	TOP2A
no matches found	miR-375	ANKRD1
other lncRNAs found	miR-126a-3p	TEK, MMP7
no matches found	miR-126a-5p	MMP7, AOC3, TMEM100

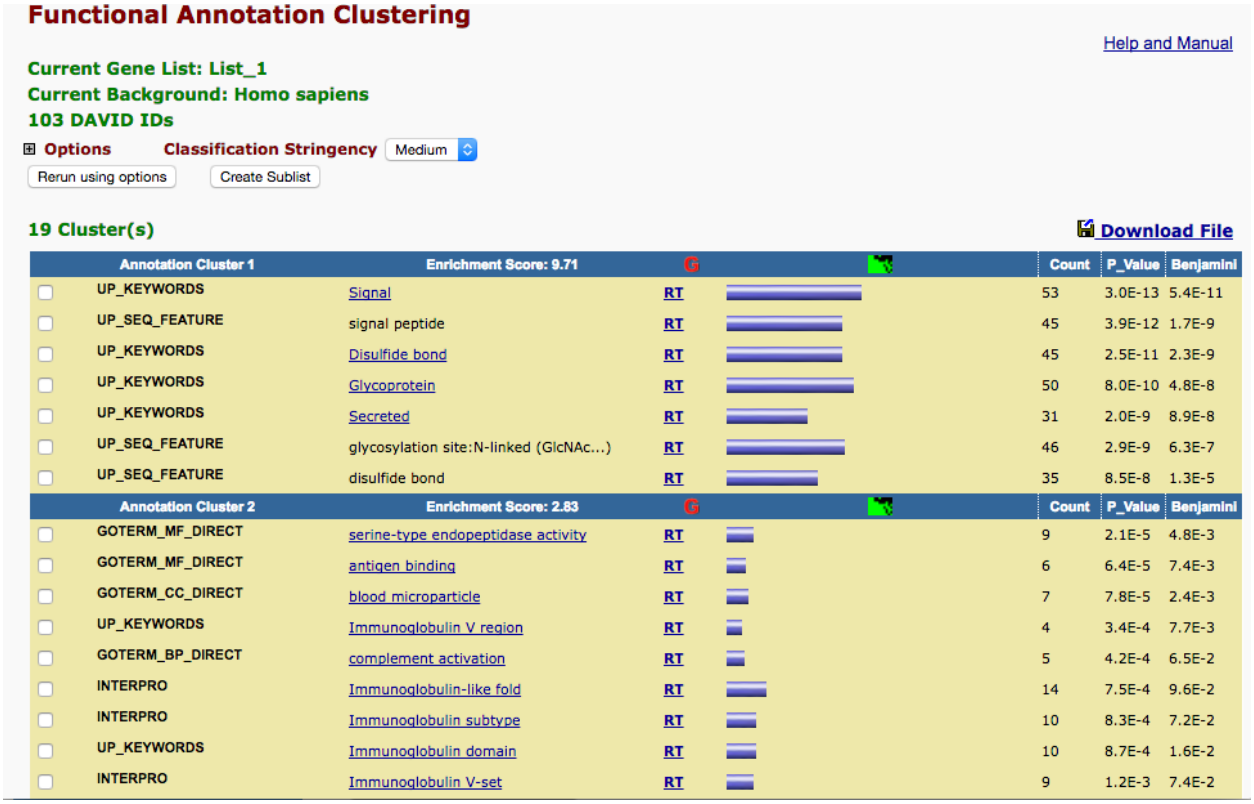
no matches found	miR-30d-3p	SULT1C4
MALAT1	miR-30d-5p	SEMA6A
MALAT1	miR-129-5p	EMCN, SEMA6A
GAS5, PVT1	miR-128-3p	TGFBR3, CDH5
no matches found	miR-30b-3p	ACADL, INMT
MALAT1	mir-30b-5p	SEMA6A
MALAT1	miR-30c-5p	SEMA6A
no matches found	miR-30c-1-3p	ACADL, INMT
no matches found	miR-30c-2-3p	INMT, ACADL
MALAT1	miR-520c-3p	DLGAP5, SLC6A4, TMEM100
MALAT1	miR-520e	DLGAP5, SLC6A4, TMEM100
MALAT1	miR-520b	DLGAP5, SLC6A4, TMEM100
MALAT1	miR-520d-3p	DLGAP5, SLC6A4, TMEM100
MALAT1	miR-520a-3p	DLGAP5, SLC6A4, TMEM100
no matches found	miR-520h	CAV1, SLC6A4
no matches found	miR-520g-3p	CAV1, SLC6A4
no matches found	miR-520d-5p	TOP2A
H19, MALAT1, HOTAIR, PVT1	miR-17-5p	TMEM100, GPM6A, CAV1, SLC6A4
no matches found	miR-200a-5p	HHIP
H19, HOTAIR, MALAT1, PVT1	miR-106b-5p	SLC6A4, TMEM100, GPM6A, CAV1
no matches found	miR-193-5p	SLC6A4, TMEM100, CAV1
H19	miR-193-3p	SCN7A
H19, HOTAIR, MALAT1, PVT1	miR-20b-5p	SLC6A4, TMEM100, CAV1
MALAT1	miR-224-5p	FOSB
other lncRNAs found	let-7e-5p	DSP, TGFBR3
no matches found	miR-221-5p	CLIC5
GAS5, HOTAIR, TUG1	miR-221-3p	LPHN2
other lncRNAs found	let-7a-5p	TGFBR3
no matches found	miR-27a-3p	SEMA6A, TGFBR3
other lncRNAs found	miR-10b-5p	HHIP, RTKN2
no matches found	miR-1254	CD36
no matches found	miR-574-5p	HHIP, TTLL7

PVT1	miR-24-3p	INMT, TTLL7
DLX6-AS1	miR-199a-5p	CAV1
no matches found	miR-660-3p	CAV1, CLIC5
MALAT1	miR-92a-3p	TTLL7, HBB, MYZAP
no matches found	miR-30a-3p	SULT1C4
MALAT1	miR-30a-5p	SEMA6A

4.4. Differentially Expressed Genes: DAVID Functional Annotation

According to DAVID, the differentially expressed genes were functionally annotated. DAVID functionally annotated the differentially expressed genes into 19 clusters in order of highest to lowest enrichment scores. This means that DAVID classified the genes into 19 gene functional groups. The enrichment score ranks the gene clusters in order of biological significance. The primary cluster has an enrichment score of 9.71 and the genes are related to signaling and secretion (Figure 25). The highly common annotations of the first gene cluster includes: signal, glycoprotein, glycosylation site N-linked, signal peptide, and disulfide bond annotations. The second cluster has an enrichment score of 2.83 and the genes are related to immune response and immunoglobulins (Figure 25). The highly common annotations of the second gene cluster includes: immunoglobulin-fold, immunoglobulin-domain, immunoglobulin subtype, antigen binding, serine-type endopeptidase activity, gamma receptor pathway signaling pathway involved in phagocytosis, and complement activation classical pathway annotations.

The third cluster has an enrichment score of 2.36 and the genes are related to the cell membrane (Figure 26). The highly common annotations of the third gene cluster includes: membrane, transmembrane, transmembrane helix, transmembrane region, integral component of membrane, cell membrane, topological domain cytoplasmic, plasma membrane, and topological domain extracellular annotations.



Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
103 DAVID IDs

Options Classification Stringency: Medium

Rerun using options Create Sublist

19 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 9.71			Count	P_Value	Benjamini
UP_KEYWORDS	Signal	RT		53	3.0E-13	5.4E-11
UP_SEQ_FEATURE	signal peptide	RT		45	3.9E-12	1.7E-9
UP_KEYWORDS	Disulfide bond	RT		45	2.5E-11	2.3E-9
UP_KEYWORDS	Glycoprotein	RT		50	8.0E-10	4.8E-8
UP_KEYWORDS	Secreted	RT		31	2.0E-9	8.9E-8
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		46	2.9E-9	6.3E-7
UP_SEQ_FEATURE	disulfide bond	RT		35	8.5E-8	1.3E-5
Annotation Cluster 2	**Enrichment Score: 2.83**			**Count**	**P_Value**	**Benjamini**
GOTERM_MF_DIRECT	serine-type endopeptidase activity	RT		9	2.1E-5	4.8E-3
GOTERM_MF_DIRECT	antigen binding	RT		6	6.4E-5	7.4E-3
GOTERM_CC_DIRECT	blood microparticle	RT		7	7.8E-5	2.4E-3
UP_KEYWORDS	Immunoglobulin V region	RT		4	3.4E-4	7.7E-3
GOTERM_BP_DIRECT	complement activation	RT		5	4.2E-4	6.5E-2
INTERPRO	Immunoglobulin-like fold	RT		14	7.5E-4	9.6E-2
INTERPRO	Immunoglobulin subtype	RT		10	8.3E-4	7.2E-2
UP_KEYWORDS	Immunoglobulin domain	RT		10	8.7E-4	1.6E-2
INTERPRO	Immunoglobulin V-set	RT		9	1.2E-3	7.4E-2

Figure 25. DAVID Gene Functional Annotation Clustering: Cluster 1 &2 of 19.

DAVID functionally clustered the uploaded list of 108 differentially expressed genes (DEGs) into 19 clusters. The most biologically significant cluster is associated with signaling and secretion and has an enrichment score of 9.71. The second most biologically significant cluster is associated with immune response and immunoglobulins and has an enrichment score of 2.83.

Annotation Cluster 3	Enrichment Score: 2.36			Count	P_Value	Benjamini
UP_KEYWORDS	Cell membrane	RT		34	2.9E-6	1.0E-4
UP_SEQ_FEATURE	topological domain:Extracellular	RT		26	5.9E-4	5.1E-2
GOTERM_CC_DIRECT	plasma membrane	RT		34	1.3E-3	2.8E-2
UP_SEQ_FEATURE	topological domain:Cytoplasmic	RT		28	2.8E-3	9.8E-2
GOTERM_CC_DIRECT	integral component of plasma membrane	RT		16	3.5E-3	5.7E-2
UP_KEYWORDS	Transmembrane	RT		38	1.4E-2	1.8E-1
UP_KEYWORDS	Membrane	RT		47	1.4E-2	1.7E-1
UP_KEYWORDS	Transmembrane helix	RT		37	2.2E-2	2.1E-1
UP_KEYWORDS	Receptor	RT		15	2.3E-2	2.0E-1
UP_SEQ_FEATURE	transmembrane region	RT		32	4.2E-2	6.2E-1
GOTERM_CC_DIRECT	integral component of membrane	RT		32	1.1E-1	5.8E-1

Figure 26. DAVID Gene Functioning Annotation Clustering: Cluster 3 of 19.

The third most biologically significant cluster DAVID identified is associated with the cell membrane and has an enrichment score of 2.36.

DAVID also allowed for the differentially expressed genes to be categorized. Categories included cancer, lung disease or lung cancer, lung function, and NSCLC. These categories and the corresponding associated differentially expressed genes are presented in Table 6. Some of the differentially expressed genes overlap between the categories.

Table 6. Table of categorized DEGs.

This table presents the categorized differentially expressed genes (DEGs). AGER, LPL, SCGB1A, CAV1, GKN2, CYP24A1, MMP1, MMP7 and MMP12 are overlapping between the categories.

Category	**Genes**
Cancer	CD36, ST8SAI6, WIF1, AGER, ADH1B, ANKRD1, CDH3, CPB2, CAV1, CYP24A1, CYP4B1, EDNRB, FMO2, FHL5, GKN2, GRIA1, IL1RL1, LPL, MMP1, MMP12, MMP7, MME, MFAP4, SPP1, SCGB1A1, SPINK1, SLIT2, SLC6A4, TOP2A, TGFBR3, VWF
Lung Disease or Lung Cancer	AGER, CAV1, CYP24A1, HHIP, GKN2, LPL, MMP1, MMP12, MMP7, SCGB1A1, SFTPC
Lung Function	CP, MMP1, MMP12
NSCLC	AGER, MMP1, MMP12, MMP7

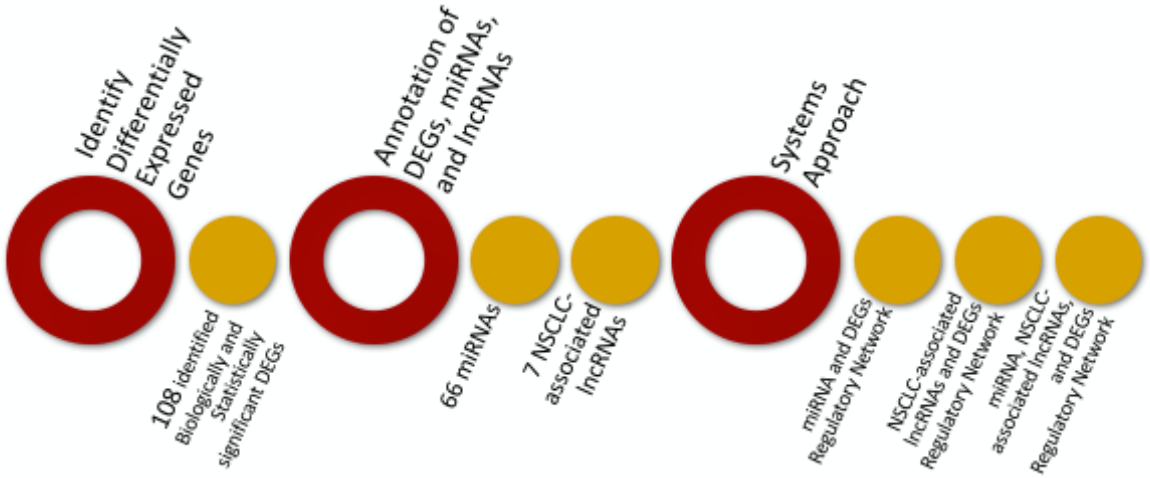


Figure 27. Schematic overview of study results.

This overview summarizes this study's results of connecting differentially expressed genes, miRNAs, and NSCLC-associated lncRNAs in NSCLC.

CHAPTER 5: Discussion and Conclusion

5.1. Regulatory Networks: Connecting Differentially Expressed Genes, miRNAs, and NSCLC-associated lncRNAs

TGFBR3, CAV1, and HHIP were further analyzed. All of these differentially expressed genes were extracted from our samples and are related to cancer or lung cancer. Using miRNAs from this dataset and NSCLC associated-lncRNAs potential regulatory networks were created to show an interplay between the regulatory factors and differentially expressed genes.

5.1.1. Regulatory Network: Connecting TGFBR3, GAS5, miR-21, miR-128

TGFBR3 is a differentially expressed gene in this dataset in which miR-21 targets TGFBR3 (miRTarBase). miR-128 targets TGFBR3 but it also targets GAS5 (miRTarBase)(starBase). According to Wei *et al.*,, GAS5 is a known lncRNA that is associated with NSCLC; GAS5 was not found in this dataset when DEGs were manually paired to lncRNAs. GAS5 is also known to target miR-21(starBase). GAS5 and TGFBR3 function as tumor suppressors. In cancer conditions, miR-21 is upregulated and miR-128 is downregulated *(Jiang et al., 2010) (Liang et al., 2016)*. TGFBR3 and GAS5 are also downregulated in cancer (Jiang et al., 2010; Liang et al., 2016). Thus, it appears that miR-21 suppresses TGFBR3 expression and miR-128 also appears to induce TGFBR3 expression (Figure 28). Hence, miR-128 appears to induce GAS5 expression and in turn GAS5 appears to suppress miR-21 expression (Figure 28).

miR-21 and miR-128 are from the filtered miRNAs in this dataset. miR-128 targeting is supported by next generation sequencing (less strong evidence). According to Hu et al., miR-128 plays a role in NSCLC tumorigenesis, angiogenesis, and lymph-angiogenesis (Hu et al., 2014). Hu *et al.* also demonstrates that miR-128 is downregulated in non-small cell lung cancer tissues

and acts a tumor suppressor. Recent studies also show, miR-128 also targets TGFBR3 confirming that there is crosstalk between miRNAs and TGF-B signaling *(Hu et al., 2014)*. According to miRTarBase, miR-21 targeting is supported by more strong evidence (reporter assay, qPCR). Previous studies show that miR-21 targets TGFBR3, a player in a key tumor suppressor pathway, TGF-B, and that TGFBR3, a membrane proteoglycan, is under-expressed in NSCLC (Papagiannakopoulos, Shapiro, & Kosik, 2008). TGF-B pathway is known to promote tumor suppressor effects (Butz, Rácz, Hunyady, & Patócs, 2012). Previous studies have confirmed with strong evidence (RT-qPCR) that GAS5 represses miR-21 expression but the study also shows the negative relationship between miR-21 and GAS5 meaning that mir-21 also suppresses the expression of GAS5. Both miR-21 and GAS5 regulate each other (Zhang et al., 2013). This double-sided relationship strengthens the fact that miRNAs and lncRNAs regulate cellular processes. In the case of NSCLC, this study suggests that the under-expression of GAS5 suppresses miR-21 which suppresses expression of TGFBR3 leading to tumor growth and metastasis. Finger *et al.* further validates this point by showing that the under-expression of TGFBR3 leads to increase in cell migration, invasion, and anchorage independent growth of lung cancer cells (Finger et al., 2008). Furthermore, GAS5 has been determined as a novel biomarker for the diagnosis of non-small cell lung cancer (Liang et al., 2016). Recently, miR-21 has also been coined a novel therapeutic target in lung cancer (Markou, 2016). This further proves the potential of miRNAs and lncRNA to regulate cancer development which emphasizes the importance of creating these regulatory network relationships. These regulatory network relationships can lead to novel biomarkers and therapeutic targets for cancer diagnosis and progression.

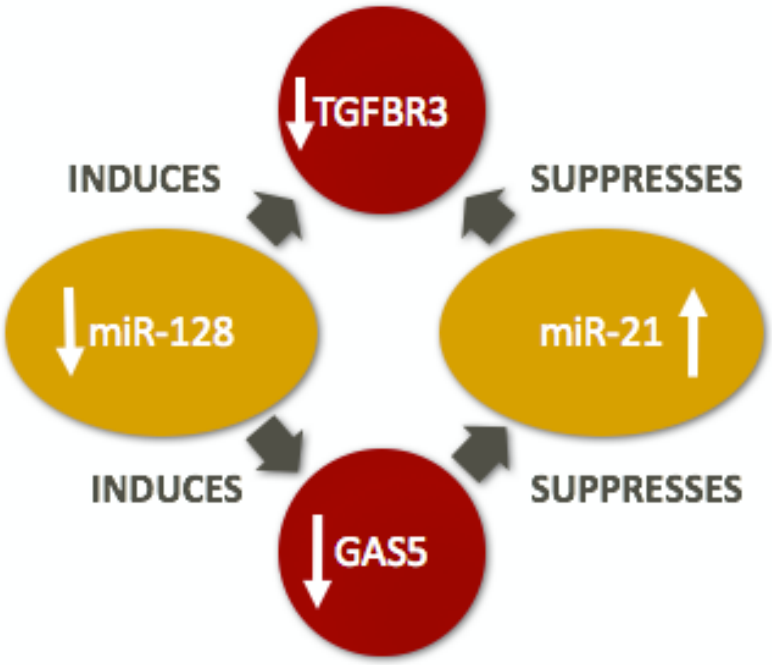


Figure 28. Regulatory Network Interplay Between TGFBR3, miR-21, miR-128, and GAS5. *In cancer, TGFBR3, miR-128, and GAS5 are downregulated while miR-21 is upregulated NSCLC (Jiang et al., 2010) (Liang et al., 2016). miR-21 and miR-128 target TGFBR3 while GAS5 targets miR-21 and miR-128 targets GAS5 (Papagiannakopoulos, Shapiro, & Kosik, 2008) (Hu et al., 2014) (Zhang et al., 2013). Thus, this suggests that miR-128 induces GAS5 expression. It also appears that GAS5 suppresses miR-21 expression and in turn it appears that miR-21 suppresses TGFBR3 expression. It also appears that miR-128 induces TGFBR3 expression.*

5.1.2. Regulatory Network: Connecting HHIP, MALAT1, miR-200b, miR-155-5p

HHIP is a differentially expressed gene in this dataset in which miR-200b targets HHIP. miR-155-5p targets HHIP but it also targets MALAT1 (miRTarBase)(starBase). According to Wei *et al.*, MALAT1 is a known lncRNA that is associated with NSCLC; MALAT1 was also not found in this dataset when DEGs were manually paired to lncRNAs. MALAT1 is also known to target miR-200b (starBase). MALAT1 functions as and oncogene while HHIP functions as a tumor suppressor. In cancer conditions, miR-200b and miR-155-5p are upregulated *(Zhou et al., 2013)*. HHIP is downregulated in lung cancer (Huang et al., 2011). MALAT1 is also upregulated

in cancer (Wei & Zhou, 2016). Thus, it appears that miR-200b suppresses HHIP expression and miR-155-5p also appears to suppress HHIP expression (Figure 29). Hence, miR-200b appears to induce MALAT1 expression and in turn MALAT1 appears to induce miR-155-5p expression (Figure 29).

MALAT1 functions by regulating gene expression and not by regulating alternative splicing in turn influencing lung cancer metastasis (Gutschner et al., 2013). In the case of NSCLC, this study suggests that the over-expression of MALAT1 induces the expression of miR-200b which suppresses the expression of HHIP leading to lung cancer cell metastasis. HHIP has been associated with Chronic Obstructive Pulmonary Disease (COPD) but the pathogenesis of COPD by HHIP remains unclear. Studies suggest that HHIP influences COPD via the hedgehog signaling pathway, a pathway important for carcinogenesis. Zhou *et al.*, performed functional annotation analysis of significant gene expression values between COPD and normal lung tissues; this analysis demonstrated a relationship with extracellular matrix and cell growth genes (Zhou et al., 2013).

DAVID was used for further analysis of extracellular matrix genes. MMP1, MMP7, and MMP12 are extracellular matrix genes found in this dataset. According to DAVID, MMP1 is functionally associated with a rate of decline of lung function in Chronic Obstructive Pulmonary Disease and is associated with an increased risk for lung cancer. Polymorphisms in MMP1 and MMP12 are related to smoking-related lung injury (R. Zhang, He, Yang, Lu, & Liu, 2005). Over-expression of MMP1 is associated with NSCLC and lymphatic metastasis of NSCLC (UN et al., 2004). MMP1 and MMP7 are over-expressed in lung microenvironment and distinguish pulmonary fibrosis from lung diseases (Rosas et al., 2008). MMP1 may also be associated with lung cancer development (Gouyer et al., 2005). MMP1 is upregulated by substance P to promote

collagen degradation in lung fibroblasts (Ramos et al., 2007). MMP1 and MMP12 play a role in lung structural changes leading to development of emphysema (Joos et al., 2002). MMP1 promoter is demonstrated to be a direct target of cigarette smoke in lung epithelial cells (Mercer, Wallace, Brinckerhoff, & D'Armiento, 2009). This reiterates the findings of Zhou *et al.* and further proves the relationship of extracellular genes with COPD or development and regulation of lung cancer.

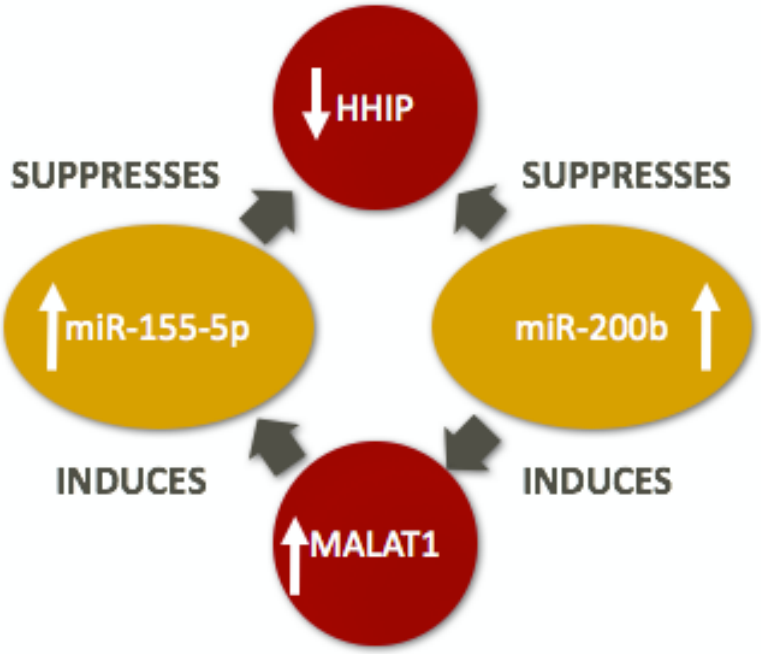


Figure 29. Regulatory Network Interplay Between HHIP, miR-155-5p, miR-200b, and MALAT1.
In cancer, miR-155-5p, miR-200b, and MALAT1 are over-expressed and HHIP is under-expressed (Huang et al., 2011) (Wei & Zhou, 2016). miR-200b and miR-155-5p target HHIP (Zhou et al., 2013). According to starBase, miR-200b targets MALAT1 and MALAT1 targets miR-155-5p. Thus, this suggests that miR-200b induces MALAT1 expression. In turn it appears that MALAT1 induces miR-155-5p expression which appears to result in miR-155-5p suppressing HHIP expression. It also appears that miR-200b suppresses HHIP expression.

5.1.3. Regulatory Network: Connecting CAV1, PVT1, miR-20b-5p, miR-17-5p

CAV1 is a differentially expressed gene in this dataset in which miR-20b-5p targets

CAV1 (miRTarBase). miR-17-5p targets CAV1 but it also targets PVT1 (miRTarBase)(starBase). According to Wei *et al.*, PVT1 is a known lncRNA that is associated with NSCLC; PVT1 was not found in this dataset when DEGs were manually paired to lncRNAs. PVT1 is also known to target miR-20b-5p (miRTarBase). PVT1 and CAV1 function as oncogenes (Wei & Zhou, 2016). The tumor suppressor function of CAV1 depends on the tissue type and the stage of the tumor; CAV1 is known to act as both a tumor suppressor and an oncogene (Sunaga et al., 2004). In cancer conditions, miR-20b-5p is downregulated and miR-17-5p is upregulated (Mogilyansky & Rigoutsos, 2013). CAV1 and PVT1 are also upregulated in NSCLC cancer. Thus, it appears that miR-20b-5p induces CAV1 expression and miR-17-5p also appears to induce CAV1 expression (Figure 30). Hence, miR-17-5p appears to induce PVT1 expression and in turn PVT1 appears to induce miR-20b-5p expression (Figure 30).

PVT1 has been defined as a biomarker and therapeutic target for NSCLC (Y.-R. Yang et al., 2014). PVT1, a 1716 nucleotide lncRNA, is a known oncogene and works to promote NSCLC cell proliferation through large tumor suppressor kinase 2 (LATS2) expression (Wan et al., 2016). In the case of NSCLC, this study suggests that the over-expression of PVT1 suppresses miR-20b-5p which suppresses CAV1 expression. CAV1, an oncogene, plays a role in tumor spreading, growth factor signaling, matrix remodeling, cell to cell adhesion, and angiogenesis. CAV1 is also known to interact with membrane metalloproteases (MMP) like MMP1, MMP7, and MMP12. MMPs function to control tumor invasion and metastasis by degrading extracellular matrix proteins. CAV1 remodels the extracellular membrane by interacting with MMPs leading to cancer cell migration and metastasis (Senetta et al., 2013).

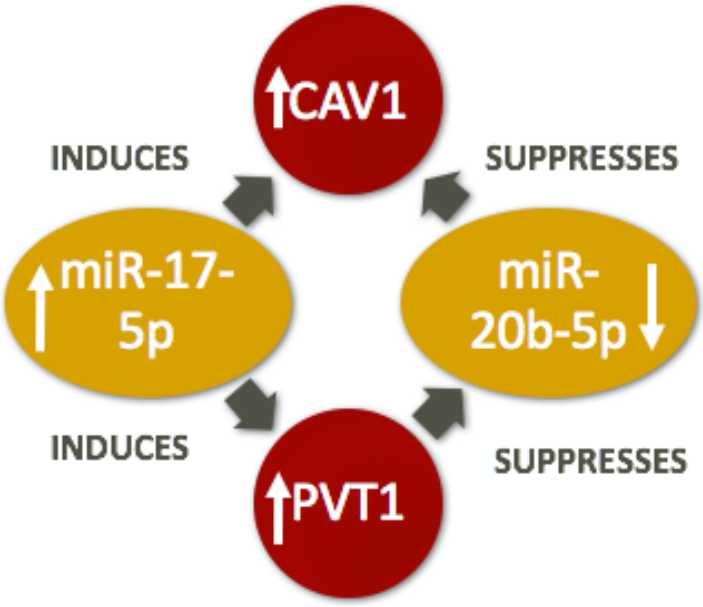


Figure 30. Regulatory Network Interplay Between CAV1, miR-17-5p, miR-20b-5p, and PVT1. *In NSCLC, CAV1, miR-17-5p, and PVT1 are over-expressed and miR-20b-5p is under-expressed (Wei & Zhou, 2016) (Mogilyansky & Rigoutsos, 2013). According to miRTarBase, miR-17-5p and miR-20b-5p target CAV1. PVT1 targets miR-20b-5p and miR-17-5p targets PVT1. Thus, this suggests that miR-17-5p induces PVT1 expression and in turn it appears that PVT1 suppresses miR-20b-5p expression which in turn appears to result in miR-20b-5p suppressing CAV1 expression. On the other hand, it also appears that miR-17-5p induces CAV1 expression.*

Not only is miR-20b-5p downregulated but so is miR-106a-5p and miR-203a-3p which are suppressed by PVT1 and in turn suppresses CAV1 expression *(Wei & Zhou, 2016)*. This suggests that miR-17-5p induces PVT1 expression and in turn PVT1 suppresses miR-20b-5p, miR-106a-5p, and miR-203a-3p expression which in turn results in the aforementioned miRNAs suppressing CAV1 expression (Figure 31).

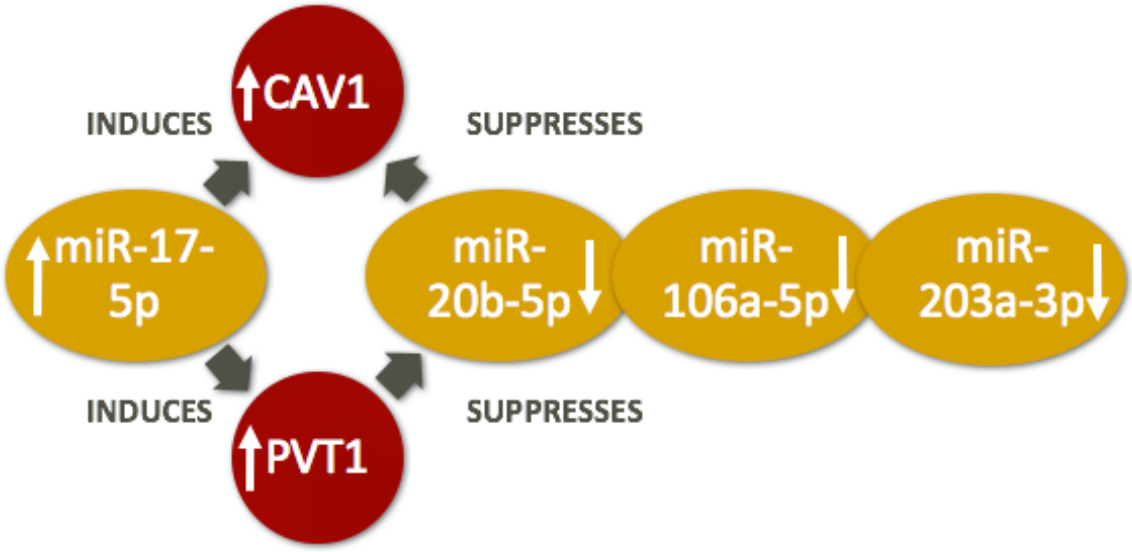


Figure 31. Regulatory Network Interplay Between CAV1, miR-17-5p, miR-20b-5p, miR-203a-3p, miR-106a-5p, and PVT1.
According to miRTarBase, miR-20b-5p, miR-106a-5p, and miR-203a-3p target CAV1. In NSCLC, not only is miR-20b-5p downregulated but so is miR-106a-5p and miR-203a-3p which appear to be suppressed by PVT1. In turn, it also appears that miR-106a-5p and miR-203a-3p suppress CAV1 expression (Shen & Jiang, 2012). This suggests similar biological relationships as in Figure 30.

In NSCLC, miR-17-5p, miR-106b-5p, and miR-20a-5p target CAV1 (miRTarBase). Not only is miR-17-5p upregulated but so is miR-106b-5p and miR-20a-5p which induces CAV1 expression (*Shen & Jiang, 2012)*. This suggests that miR-17-5p, miR-106b-5p, miR-20a-5p induce CAV1 expression but also induce PVT1 expression and in turn PVT1 suppresses miR-20b-5p, miR-106a-5p, and miR-203a-3p expression which in turn results in the aforementioned miRNAs suppressing CAV1 expression (Figure 32).

Experimental validation of these proposed computational interactions will contribute to the roadmap for lung cancer diagnosis and therapy.

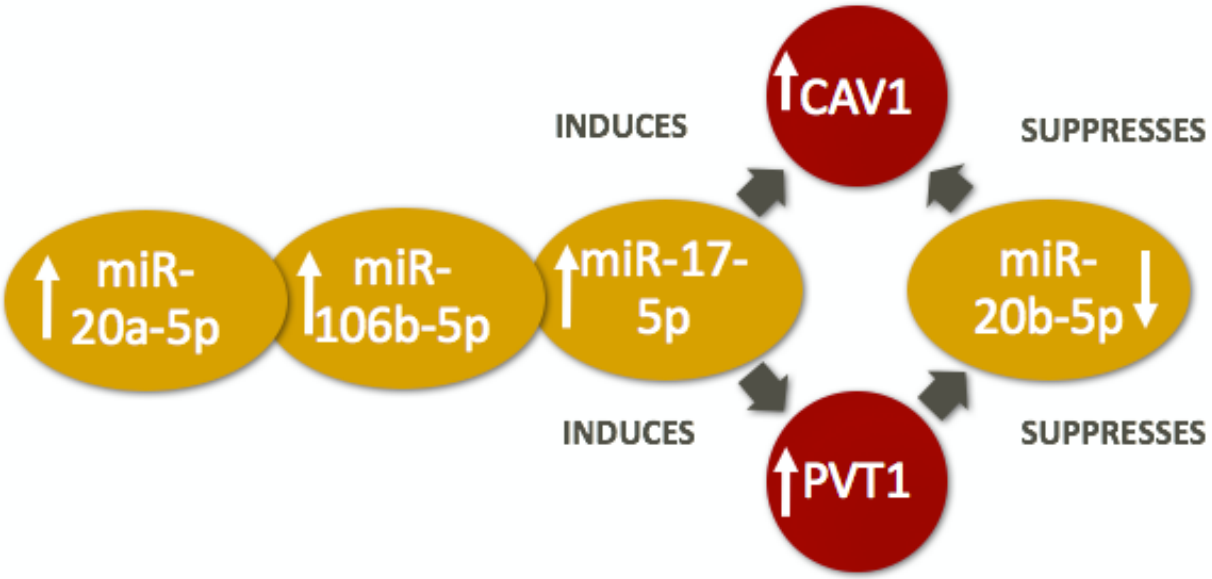


Figure 32. Regulatory Network Interplay Between CAV1, miR-17-5p, miR-106b-5p, miR-20a-5p, miR-20b-5p, and PVT1.

According to miRTarBase, miR-17-5p, miR-106b-5p, and miR-20a-5p target CAV1. In NSCLC, not only is miR-17-5p upregulated but so is miR-106b-5p and miR-20a-5p which appear to induce CAV1 expression (Shen & Jiang, 2012). This suggests similar biological relationships as in Figure 30.

The fold cut-off and p-value are harsh filters and this may present a problem with extracting the intended samples from this dataset. An unfiltered heatmap may be produced to validate the number of extracted samples from the volcano plot. Also, the number of samples in this study may be a limitation. In the future, computational studies patterns can be validated with other datasets and compared with RNA-sequencing datasets.

5.2. Conclusion

miRNAs and lncRNAs are fundamental molecular factors that regulate gene expression to ensure cellular homeostasis in NSCLC and in normal lungs. In this study the aim was to suggest a biological relationship and to further recommend molecular target experiments to further connect differentially expressed genes, miRNAs, and lncRNAs. It is clear that miRNAs and lncRNAs play regulatory mechanisms in controlling NSCLC and it is also clear that there

are a plethora of miRNA and lncRNA interactions that lead to this regulation. It is also clear that there are connections between differentially expressed genes, miRNAs, and NSCLC-associated lncRNAs. The suggested regulatory networks in this study reinforce this notion. However, what is not clear is how the miRNAs, lncRNAs, and differentially expressed genes work together to regulate cellular homeostasis.

Tumor suppressor genes like TGFBR3 and HHIP show a connection between extracellular matrix genes which are functionally known and related to carcinogenesis (Finger et al., 2008 and Zhou et al., 2013). CAV1, an oncogene, displays a relationship between growth factor signaling and matrix remodeling all leading to tumor growth and cancer cell metastasis (Senetta et al., 2013). MALAT1, PVT1, and GAS5 are lncRNAs that regulate gene expression via miRNA targeting. These networks propose mechanisms of actions to further study miRNAs and lncRNAs suggesting a crosstalk between miRNAs, lncRNAs, and differentially expressed genes. Because miRNAs and lncRNAs have proven to play important roles in cellular regulation as well as cancer progression the regulation of these regulatory pathways can lead to novel approaches in cancer therapy.

CHAPTER 6: Future Directions

A better knowledge of the regulatory pathway is important for understanding tumor pathogenesis. In the present study, the aim was to analyze regulatory gene factors in order to suggest connections between differentially expressed genes, miRNAs, and lncRNAs. These factors were computationally studied and suggested. Future studies are needed to validate or revoke these connections with complementary experimental research. Confirming these suggested relationships will not only provide insight to the pathogenesis of cancer, but will also hone therapeutic strategies. Future studies should accomplish the following objectives.

Objective 1

Experimentally validate the suggested regulatory networks. Experimentally determine whether miRNAs and lncRNAs together can influence differentially expressed genes. This will lead to a confirmed mechanism of action for NSCLC regulation. This can give insight into manipulating lung cancer progression and novel therapeutic methods.

Objective 2

Identify more lncRNAs associated with NSCLC and determine their respective functions. This can contribute to clinical usefulness in cancer diagnosis, prognosis, and therapy. This crucial especially since lncRNAs are potential biomarker for early diagnosis of NSCLC.

Future studies should also expand on lncRNAs, as it is a new field in molecular biology. lncRNAs are a novel emerging field and require in-depth analysis including experimental identification and functional annotation of lncRNAs. This can contribute to online lncRNA databases for future computational analysis and will lead to a better understanding of cancer pathogenesis and treatment.

Objective 3

Future studies should also include RNA sequencing data from NSCLC patients to not only identify the same patterns, but also validate the connections between differentially expressed genes, miRNAs, and lncRNAs. This will also lead to possible diagnostic targets and targeted treatment options for NSCLC patients which could lead to halting the progression of the disease. This will lead to a better mechanistic understanding of lung cancer regulation.

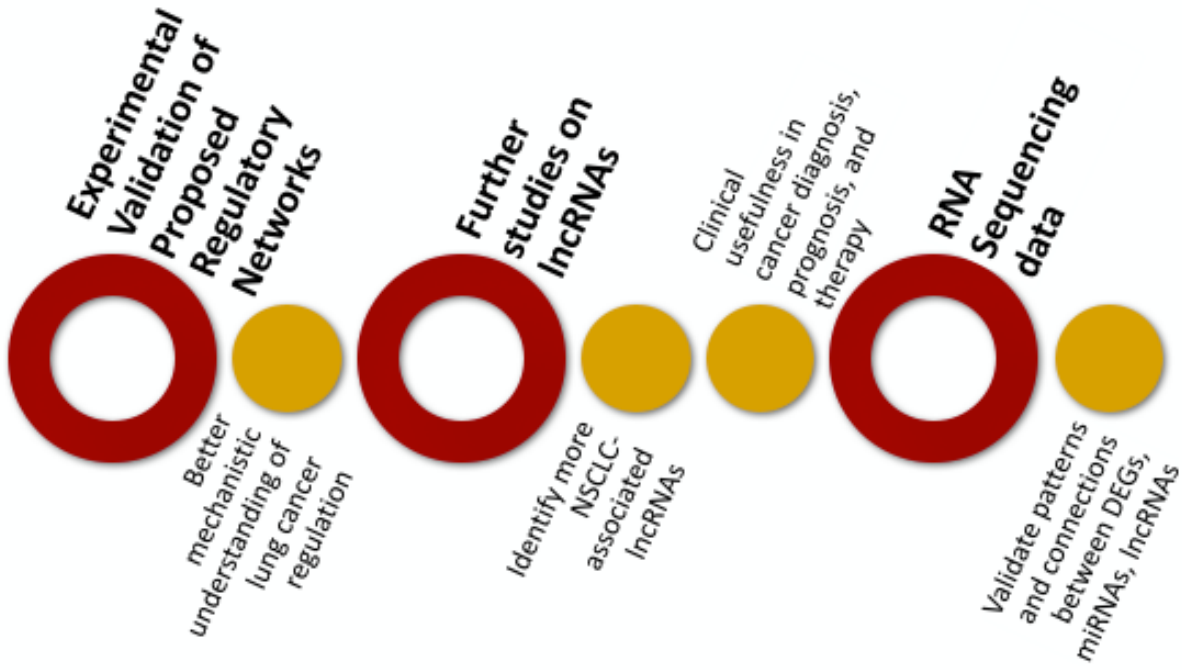


Figure 33. Schematic overview of future directions.
This overview summarizes the future directions for NSCLC diagnosis, prognosis, and therapy.

References

1.AmericanCancerSociety. (5/16/2016). Lung Cancer (Non-Small Cell). Retrieved from http://www.cancer.org/cancer/lungcancer/

2.AmericanLungAssociation. (Nov 3, 2016). Lung Cancer Fact Sheet. Retrieved from http://www.lung.org/

3.Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *cell, 116*(2), 281-297.

4.Brambilla, E., & Gazdar, A. (2009). Pathogenesis of lung cancer signaling pathways: roadmap for therapies. *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*, *33*(6), 1485–1497. http://doi.org/10.1183/09031936.00014009

5.Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, 22.21. 21-22.21. 11.

6.Butz, H., Rácz, K., Hunyady, L., & Patócs, A. (2012). Crosstalk between TGF-β signaling and the microRNA machinery. *Trends in pharmacological sciences, 33*(7), 382-393.

7.CDC. (Jan 4, 2016). Global Cancer Statistics. Retrieved from http://www.cdc.gov/cancer/international/statistics.htm

8.Chen, J., Wang, R., Zhang, K., & Chen, L. B. (2014). Long non-coding RNAs in non-small cell lung cancer as biomarkers and therapeutic targets. *Journal of cellular and molecular medicine, 18*(12), 2425-2436.

9.Chen, X., Liang, H., Zhang, J., Zen, K., & Zhang, C.-Y. (2012). Secreted microRNAs: a new form of intercellular communication. *Trends in cell biology, 22*(3), 125-132.

10.Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.H., Yang, C.D., Hong, H.C., Wei, T.Y., Tu, S.J. and Tsai, T.R. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research, 44*(D1), D239-D247. W.H., Yang, C.D., Hong, H.C., Wei, T.Y., Tu, S.J. and Tsai, T.R

11.Cure. (April 23, 2016). Medical Illustration: Non-Small Cell Lung Cancer. Retrieved from http://www.curetoday.com/publications/cure/2016/lung-2016/medical-illustration-nonsmall-cell-lung-cancer

12.Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research, 30*(1), 207-210.

13.Finger, E. C., Turley, R. S., Dong, M., How, T., Fields, T. A., & Blobe, G. C. (2008). TβRIII suppresses non-small cell lung cancer invasiveness and tumorigenicity. *Carcinogenesis, 29*(3), 528-535.

14.Gouyer, V., Conti, M., Devos, P., Zerimech, F., Copin, M. C., Créme, E., Wurtz, A., Porte, H. and Huet, G. (2005). Tissue inhibitor of metalloproteinase 1 is an independent predictor of prognosis in patients with non-small cell lung carcinoma who undergo resection with curative intent. *Cancer, 103*(8), 1676-1684.

15.Gurtan, A. M., & Sharp, P. A. (2013). The role of miRNAs in regulating gene expression networks. *Journal of molecular biology, 425*(19), 3582-3600.

16.Gutschner, T., Hämmerle, M., Eißmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Groß, M., Zörnig, M. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research, 73*(3), 1180-1189.

17.Hu, J., Cheng, Y., Li, Y., Jin, Z., Pan, Y., Liu, G., Fu, S., Zhang, Y., Feng, K. and Feng, Y. (2014). microRNA-128 plays a critical role in human non-small cell lung cancer tumorigenesis, angiogenesis and lymphangiogenesis by directly targeting vascular endothelial growth factor-C. *European journal of cancer, 50*(13), 2336-2350.

18. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research, 37*(1), 1-13.

19. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols, 4*(1), 44-57.

20. Huang, S., Yang, L., An, Y., Ma, X., Zhang, C., Xie, G., Chen, Z.Y., Xie, J. and Zhang, H. (2011). Expression of hedgehog signaling molecules in lung cancer. *Acta histochemica, 113*(5), 564-569.

21.Ibrahim, A. S., Khaled, H. M., Mikhail, N. N., Baraka, H., & Kamel, H. (2014). Cancer incidence in Egypt: results of the national population-based cancer registry program. *Journal of cancer epidemiology, 2014*.

22.Jansson, M. D., & Lund, A. H. (2012). MicroRNA and cancer. *Molecular oncology, 6*(6), 590-610.

23.Jiang, L., & Qiu, X. (2013). MicroRNAs in Invasion and Metastasis in Lung Cancer: INTECH Open Access Publisher.

24.Jiang, X., Liu, R., Lei, Z., You, J., Zhou, Q., & Zhang, H. (2010). [Defective expression of TGFBR3 gene and its molecular mechanisms in non-small cell lung cancer cell lines]. *Chinese journal of lung cancer, 13*(5), 451-457.

25.Joos, L., He, J.-Q., Shepherdson, M. B., Connett, J. E., Anthonisen, N. R., Paré, P. D., & Sandford, A. J. (2002). The role of matrix metalloproteinase polymorphisms in the rate of decline in lung function. *Human molecular genetics, 11*(5), 569-576.

26.Kadara, H., Fujimoto, J., Yoo, S.-Y., Maki, Y., Gower, A. C., Kabbout, M., Garcia, M.M., Chow, C.W., Chu, Z., Mendoza, G. and Shen, L. (2014). Transcriptomic architecture of the adjacent airway field cancerization in non–small cell lung cancer. *Journal of the National Cancer Institute, 106*(3).

27.Lancet, D., Safran, M., Olender, T., Dalah, I., Iny-Stein, T., Inger, A., Harel, A. and Stelzer, G. (2008, April). *GeneCards tools for combinatorial annotation and dissemination of human genome information.* Paper presented at the GIACS Conf. Data Complex Syst.

28.Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., & Yang, J.-H. (2013). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, gkt1248.

29.Liang, W., Lv, T., Shi, X., Liu, H., Zhu, Q., Zeng, J., Yang, W., Yin, J. and Song, Y. (2016). Circulating long noncoding RNA GAS5 is a novel biomarker for the diagnosis of non-small cell lung cancer. *Medicine, 95*(37), e4608.

30.LillyOncology. Squamous NSCLC. Retrieved from http://www.lillyoncology.com/education/squamous-nsclc.html

31.MacFarlane, L.-A., & R Murphy, P. (2010). MicroRNA: biogenesis, function and role in cancer. *Current genomics, 11*(7), 537-561.

32.Markou, A. (2015). MicroRNA signatures as clinical biomarkers in lung cancer.

33.Markou, A. (2016). miRNA-21 as a novel therapeutic target in lung cancer.

34.Mercer, B. A., Wallace, A. M., Brinckerhoff, C. E., & D'Armiento, J. M. (2009). Identification of a cigarette smoke–responsive region in the distal MMP-1 promoter. *American journal of respiratory cell and molecular biology, 40*(1), 4-12.

35.Mogilyansky, E., & Rigoutsos, I. (2013). The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death & Differentiation, 20*(12), 1603-1614.

36.Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., & Adjei, A. A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Paper presented at the Mayo Clinic Proceedings.

37.Nadal, E., Truini, A., Nakata, A., Lin, J., Reddy, R. M., Chang, A. C., Ramnath, N., Gotoh, N., Beer, D.G. and Chen, G. (2015). A novel serum 4-microRNA signature for lung cancer detection. *Scientific reports, 5*.

38.National Cancer Institute. (n.d.). Retrieved January 10, 2017, from http://www.cancer.gov/

39.Papagiannakopoulos, T., Shapiro, A., & Kosik, K. S. (2008). MicroRNA-21 targets a network of key tumor-suppressive pathways in glioblastoma cells. *Cancer research, 68*(19), 8164-8172.

40.Prensner, J. R., & Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer discovery, 1*(5), 391-407.

41.Ramos, C., Montaño, M., Cisneros, J., Sommer, B., Delgado, J., & Gonzalez-Avila, G. (2007). Substance P up-regulates matrix metalloproteinase-1 and down-regulates collagen in human lung fibroblast. *Experimental lung research, 33*(3-4), 151-167.

42.Rothschild, S. I. (2013). "Epigenetic therapy in lung cancer–role of microRNAs." Frontiers in oncology **3**: 158.

43.Rosas, I. O., Richards, T. J., Konishi, K., Zhang, Y., Gibson, K., Lokshin, A. E., Lindell, K.O., Cisneros, J., MacDonald, S.D., Pardo, A and Sciurba, F. (2008). MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med, 5*(4), e93.

44.Senetta, R., Stella, G., Pozzi, E., Sturli, N., Massi, D., & Cassoni, P. (2013). Caveolin-1 as a promoter of tumor spreading: when, how, where and why. *Journal of cellular and molecular medicine, 17*(3), 325-336.

45.Shen, J., & Jiang, F. (2012). Applications of microRNAs in the diagnosis and prognosis of lung cancer. *Expert opinion on medical diagnostics, 6*(3), 197-207.

46.Sunaga, N., Miyajima, K., Suzuki, M., Sato, M., White, M. A., Ramirez, R. D., Shay, J.W., Gazdar, A.F. and Minna, J. D. (2004). Different roles for caveolin-1 in the development of non-small cell lung cancer versus small cell lung cancer. *Cancer research, 64*(12), 4277-4285.

47.UN, T.-S., Chiou, S.-H., Wang, L.-S., Huang, H.-H., CHIANG, S., Shih, A. Y., Chen, Y., Chen, C.Y., Hsu, N.Y. and Ming-Chihchou, S.J. (2004). Expression spectra of matrix metalloproteinases in metastatic non-small cell lung cancer. *Oncology reports, 12*, 717-723.

48.Vance, K. W., & Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in Genetics, 30*(8), 348-355.

49.Wan, L., Sun, M., Liu, G.-J., Wei, C.-C., Zhang, E.-B., Kong, R., Xu, T.P., Huang, M.D. and Wang, Z.-X. (2016). Long Noncoding RNA PVT1 Promotes Non–Small Cell Lung Cancer Cell Proliferation through Epigenetically Regulating LATS2 Expression. *Molecular cancer therapeutics, 15*(5), 1082-1094.

50.Wei, M.-M., & Zhou, G.-B. (2016). Long Non-coding RNAs and Their Roles in Non-small-cell Lung Cancer. *Genomics, Proteomics & Bioinformatics, 14*(5), 280-288.

51.Xu, C., Zheng, Y., Lian, D., Ye, S., Yang, J., & Zeng, Z. (2015). Analysis of microRNA expression profile identifies novel biomarkers for non-small cell lung cancer. *Tumori Journal, 101*(1), 104-110.

52.Yang, G., Lu, X., & Yuan, L. (2014). LncRNA: a link between RNA and cancer. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1839*(11), 1097-1109.

53.Yang, J., Lin, J., Liu, T., Chen, T., Pan, S., Huang, W., & Li, S. (2014). Analysis of lncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes. *Lung Cancer, 85*(2), 110-115.

54.Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., & Qu, L.-H. (2011). starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic acids research, 39*(suppl 1), D202-D209.

55.Yang, Y.-R., Zang, S.-Z., Zhong, C.-L., Li, Y.-X., Zhao, S.-S., & Feng, X.-J. (2014). Increased expression of the lncRNA PVT1 promotes tumorigenesis in non-small cell lung cancer. *Int J Clin Exp Pathol, 7*(10), 6929-6935.

56.Zhang, R., He, Q., Yang, R., Lu, B., & Liu, Y. (2005). [Study on matrix metalloproteinase 1, 9, 12 polymorphisms and susceptibility to chronic obstructive pulmonary disease among Han nationality in northern China]. *Chinese Journal of Epidemiology*, 907-910.

57.Zhang, Z., Zhu, Z., Watabe, K., Zhang, X., Bai, C., Xu, M., Wu, F. and Mo, Y.Y. (2013). Negative regulation of lncRNA GAS5 by miR-21. *Cell Death & Differentiation, 20*(11), 1558-1568.

58.Zhao, X.-Y., & Lin, J. D. (2015). Long noncoding RNAs: a new regulatory code in metabolic control. *Trends in Biochemical Sciences, 40*(10), 586-596.

59.Zhao, Y., Li, H., Fang, S., Kang, Y., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q. and Chen, R. (2015). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*.

60.Zhou, X., Qiu, W., Sathirapongsasuti, J. F., Cho, M. H., Mancini, J. D., Lao, T., Thibault, D.M., Litonjua, A.A., Bakke, P.S., Gulsvik, A. and Lomas, D.A. (2013). Gene expression analysis uncovers novel hedgehog interacting protein (HHIP) effects in human bronchial epithelial cells. *Genomics, 101*(5), 263-272.

APPENDIX

Permission to use Figures in Thesis

Copyright ©1994-2016 Eli Lilly and Company. All rights reserved.

Lilly USA, LLC (Lilly) is a wholly owned subsidiary of Eli Lilly and Company. This Copyright Policy applies to this website and any promotional e-mail, text message or other electronic content received by you in response to an opt-in registration from Lilly (the Content). Lilly hereby authorizes you to print individual pages from this site or the Content, unless otherwise expressly noted, solely for your own personal, non-commercial use in learning about the services or products offered by Lilly or for your non-commercial use in connection with healthcare or education. If you are a healthcare professional or provider, you may print individual pages from this site or the Content, unless otherwise expressly noted, and share the information and materials with others. No other permission is granted to you to print, copy, reproduce, distribute, license, transfer, sale, transmit, upload, download, store, display in public, alter, modify or create derivative works of the Content. You may not remove any copyright, trademark or other proprietary notations from this site or the Content.

Figure A1. Permission to use Figure 1

Figure A2. International Agency for Research on Cancer (IARC) grants permission to use Figure 2 and Figure 3 "as is" for research purposes.

Figure A3. Permission to use Figure 4 and Figure 5

Title:	Smoking and Health Infographic
Description:	This infographic is from the National Cancer Institute's Building on Opportunities in Cancer Research: An Annual Plan and Budget Proposal for Fiscal Year 2016. See also http://www.cancer.gov/about-nci/budget/annual-plan/nci-plan-2016.pdf.
Topics/Categories:	Cancer Types -- Lung Cancer Risk Factors and Causes
Type:	Color, Illustration
Source:	National Cancer Institute (NCI)
Date Created:	November 2014
Date Added:	December 2, 2014
Access:	Public
Reuse Restrictions:	**None** - This image is in the public domain and can be freely reused. Please credit the source and, where possible, the creator listed above.

Figure A4. Permission to use Figure 6

Figure A5. Permission to use Figure 7

Figure A6. Permission to use Figure 8 and Figure 9

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jan 09, 2017

This Agreement between Jasmine K Omran ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4010160440125
License date	Dec 15, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Biochemical Sciences
Licensed Content Title	Long Noncoding RNAs: A New Regulatory Code in Metabolic Control
Licensed Content Author	Xu-Yun Zhao,Jiandie D. Lin
Licensed Content Date	October 2015
Licensed Content Volume Number	40
Licensed Content Issue Number	10
Licensed Content Pages	11
Start Page	586
End Page	596
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations

Figure A7. Permission to use Figure 10

Figure A8. Permission to use Figure 11 and Figure 12