Investigating the Genetic Influences of the Germline and Somatic Genomes in Three Subtypes of Lung Cancer

By

Timothy Daniel O'Brien

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May 31, 2017

Nashville, Tennessee

Approved:

Zhongming Zhao, Ph.D.

Melinda Aldrich, Ph.D.

Tony Capra, Ph.D.

Jirong Long, Ph.D.

Nancy Cox, Ph.D.

David Samuels, Ph.D.

To my parents, Dan and Bernie, for their unending support and who have always encouraged me to follow my dreams

and

To my wife, Barbara, who has always supported me in everything I do.

ACKNOWLEDGMENTS

The work in this dissertation would not have been possible without the support from the Human Genetics Training Grant provided by the National Institute of General Medical Sciences Training Grant (T32GM080178) and the National Institutes of Health grant (R01LM011177). Additionally, support was provided by the LUNGevity Foundation and Upstate Lung Cancer for the work done in Chapter IV. I would also like to acknowledge my outside collaborators for their contributions to this dissertation. William Pao and Hailing Jin from Vanderbilt and Uma Saxena, Martin J. Aryee, Mari Mino-Kenudson, Jeffrey A. Engelman, Long P. Le, A. John Iafrate, and Rebecca S. Heist from Massachusetts General Hospital for their work on Chapter IV. Also, my collaborator Maria T. Landi from the National Cancer Institute for her support on Chapter II. Finally, Pierre Massion from Vanderbilt for his support on Chapter III. Also, many members of the Zhao lab for all of their contributions to this dissertation.

I would also like to thank members of my dissertation committee: David Samuels (chair), Melinda Aldrich, Tony Capra, Jirong Long, Nancy Cox, and Zhongming Zhao. I have really been inspired by this group of great scientists. My thesis project, and my scientific thinking in general, have been greatly improved from each meeting. I thank David Samuels for being a great dissertation chair and always having an open door when I needed any help. Special thanks for Tony Capra and his lab for welcoming me over the last year. I really appreciate the friendship and support from the entire lab. Also, Melinda Aldrich, her lab, and members of the TREAT group at Vanderbilt for their feedback on my work and giving me the opportunity to see lung cancer from a different perspective.

Additional thanks to members of the Zhao lab past and present. Their help in technical skills and scientific thinking have helped me in my entire graduate career. Especially lab members: Junfeng Xia, Huy Vuong, Pora Kim, Qingguo Wang, Feixiong Cheng, Junfei Zhao, Ramkrishna (Santu) Mitra, Quan Wang, Mingyu Shao, and Peilin Jia for all their help with analyses in the lab. I especially would like to thank Peilin Jia. From the first day of my rotation she has helped me in many ways. From teaching me how to be a computational scientist to thinking critically and scientifically, she has greatly helped me as a scientist and has been a great mentor. I would also like to especially thank my mentor Zhongming Zhao. He has really helped me grow as a scientist and has been a great mentor. He has taught me about scientific writing, critical thinking, and many other skills required of a scientist. He has always had the time to help with any problems and was always available for discussion.

I would like to thank members of the CHGR/VGI and the students of the HGEN program. The student group has been a great resource of friendship and scientific support. Also, Roz Johnson and Dana Campbell for their help with everything related to the HGEN program.

I also thank David Miller and members of the Miller lab past and present. Cody Smith for all of the help and support for my early years in the lab. The work I did and friendships I made really inspired me to attend graduate school. Also, David Miller for his guidance and mentoring.

Additional thanks to my friends and family back in Washington and in Tennessee. I appreciate all the support and guidance over these past years. Finally, I would like to thank my wife and fellow grad student Barbara O'Brien. From the first day of graduate school, she has helped me in so many ways from qualifying exam prep to listening to way too many practice talks. At this point I think she knows as much about my project as I do. I really appreciate all of your support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Cancer is a disease of uncontrolled cellular growth. In the late stage, cancer cells may break the normal boundaries of their given cell type and invade surrounding tissue in a process called metastasis. Cancer can be classified into more than 100 distinct diseases that can affect nearly every cell and tissue in the human body and is ultimately a disease of genomic abnormalities (1). In 1982, strong evidence of the genetic component of cancer was discovered – Reddy *et al*. and Tabin *et al*. found a single mutation in the *HRAS* oncogene leading to cancer, as reported in two papers (2, 3). Since this breakthrough, hundreds of oncogenes and tumor suppressor genes have been discovered that may lead to cancer when their normal cellular mechanisms are disrupted through processes such as somatic mutations (4, 5). Although somatic mutations are important in cancer, the germline genome may also influence risk of cancer.

The germline genome is the genome of the germ cells and is inherited. This DNA is the same in every cell in the human body (with the exception of *de novo* mutations). The somatic genomes consist of the genomes of every cell in the body with mutations that have been acquired during the lifetime of the individual. While the germline genome is passed on to offspring for the next generation, somatic mutations do not pass on. Both of these genomes have been found to be important in the process leading to cancer. However, many cancer researchers focus on one or the other genome and rarely study them in combination. It is important to study both because variants

in the germline genome could act in combination with variants in the somatic genome to cause cancer in a process known as Knudson's two-hit hypothesis (6).

Most cancers arise from somatic mutations to a given cell type. Although there are some cancers attributable to germline genetic abnormalities, all of these disruptions to the germline genome confer a greater risk for cancer rather than causal tissue-specific somatic mutations that directly lead to disease. Somatic mutations in cancer cells can be classified into two main types: driver mutations and passenger mutations. Driver mutations typically refer to those somatic mutations that confer uncontrolled growth to the cell or allow the cell to survive in conditions where apoptosis should normally occur. These are the typical cancer genes that are known and are often implicated in more than one cancer type (1). There are now over 600 driver genes that have been discovered in cancer (7), and more novel mutations, including those with regulatory roles in noncoding sequences, have been reported with potential driving roles recently (8, 9). Passenger mutations are somatic mutations in the cancer cell that are not specific to uncontrolled growth. The majority of somatic mutations are passenger mutations in cancer cells (1). Many of these mutations are benign and were in the cell prior to the driver mutation event (1). However, the determination of passenger versus driver genes is still an active area of research (10, 11). Overall, cancer is a genetic disease where mutations in one of hundreds, or thousands, of different genes may lead to abnormal cellular growth and disease state from the germline, somatic, or both genomes. This dissertation uses genetic data from the germline and somatic genomes to investigate three subtypes of lung cancer.

Below, I give a brief overview of the history of lung cancer, environmental exposures associated with lung cancer, and genetic heritability. I also highlight the differences in lung cancer subtypes from a histological perspective. Finally, I summarize the genetic abnormalities in three

lung cancer subtypes using numerous examples from candidate gene studies, as well as genome-wide studies, from the germline and somatic genomes.

Overview and epidemiology of lung cancer

In 2012, there were ~8.2 million estimated deaths around the world attributable to cancer (12). The cancer types that were responsible for the most number of deaths were lung cancer in men and breast cancer in women. However, this trend is related to the income and status of the country – in developed countries, lung cancer is the leading cause of cancer related deaths for both sexes (13). Historically, this extremely high incidence of lung cancer around the world was not always true. Initially, lung cancer was a very rare cancer type with some institutions reporting it as comprising only ~1% of all tumors discovered (14). However, around the end of World War I (~1918), lung cancer cases began to skyrocket. Originally, there were several suspected links to lung cancer such as poison gas used in the trenches and the increase in pollution from the widespread use of the newly introduced automobiles (14). However, in Germany in the early 1940s, the first suspected links between cigarette smoke and lung cancer were reported. As reviewed in (14, 15), Müller's work (16), published in German, was one of the first studies to report a link between cigarette smoke and lung cancer. Additional studies in the UK and the US validated these claims in the 1950's (15). A preliminary report, by Doll and Hill, in 1950 demonstrated a link between smoking and lung cancer (17) while a much larger study (18) confirmed these findings in addition to the role of cigarette smoking in many other diseases. There are strong correlations with the prevalence of smoking and lung cancer for both sexes (19). Additionally, with an increase in cigarette smoking in developed nations, lung cancer has now

increased to become the number one diagnosed cancer type for men and the third most diagnosed cancer type for women around the world (13).

These worldwide trends are also prevalent in the United States (US). In the US, lung cancer is estimated to be the second leading cancer diagnosis for men (after prostate) and women (after breast) in 2017 (20). Although lung cancer is not estimated to be the number one diagnosed cancer type, it is predicted to be the leading cause of cancer related deaths in 2017 for both men and women. Current estimates place the total number of estimated deaths from lung cancer at ~85,000 for men and ~71,000 for women in the US for 2017 (20). One explanation for such a high mortality rate for lung cancer in comparison to other cancer types is due to its late-stage diagnosis. Over 50% of lung cancer diagnoses are made at the distant (metastasized to other organs) stage, while only ~15% of cases are discovered at a localized (constrained to the lung) level (20). There is a strong correlation between survival time and the stage at lung cancer diagnosis. At diagnosis, the 5 year survival rate for localized stage lung cancer is ~55%, but if diagnosed at a distant stage, the 5 year survival rate is <5% (21). In an attempt to increase the survival time of lung cancer sufferers, the US has rolled out an early screening program for lung cancer (22). Early results suggest that through the use of low-dose helical computed tomography (CT) scans, the mortality from lung cancer can be reduced (23). This program focuses on heavy smokers, but cigarette smoking is not the only known environmental agent associated with lung cancer.

Risk factors associated with lung cancer

Cigarette smoking is the number one cause of lung cancer and worldwide is estimated to be the cause of 85% of lung cancer cases in men and 47% of lung cancer cases in women (24).

However, many other environmental risks pose a threat. For example, smoking tobacco in non-cigarette forms such as pipes or cigars also increases a person's risk for lung cancer (25). Additionally, indoor and outdoor pollution poses a risk for lung cancer (25), like using specific cooking pots in many Asian countries (26). Many other occupational exposures exist such as arsenic and asbestos, as well as many heavy metals such as nickel and chromium. In addition, naturally occurring radon gas and radiation all have been associated with lung cancer risk (25). Although environmental hazards play a large role in lung cancer risk, there is also a genetic risk for lung cancer.

In 2001, Hemminki *et al*. (27) used the Swedish Family-Cancer Database to estimate the genetic heritability for colorectal cancer, melanoma, and lung cancer. This database was the largest that was used at the time of publication and contained over 6 million people and ~550,000 cancers (28). They estimated liability for cancer using genotype, shared environment, childhood environment and non-shared environments among family members. The genotype variable used corresponded to overall relationships. For example, the coefficient of relatedness for first-degree relatives is 0.5, such as siblings and parent-offspring, while half-siblings is 0.25. The results of this study concluded that the heritability for lung cancer was estimated to be approximately 14% in this registry. A follow-up study, published in 2002 (29), used the updated version of the same database and found that genetic effects accounted for approximately 8% of susceptibility of lung cancer. Recently, heritability estimates for lung cancer have been determined through genome-wide association studies (GWAS) (30). Sampson *et al*. (30) found that the heritability estimate for lung cancer is 0.206 in Europeans and 0.121 in Asians. Though these studies identified the percent of a genetic effect from lung cancer, earlier studies as far back as 1963 (31, 32) found familial aggregation of lung cancer. After 1963, many other studies have investigated the association

between family history and lung cancer including strong differences between ethnicity and lung cancer risk (33). For details about these studies, see the review this year by Musolf *et al*. (34). Importantly, these studies looked at the general category of lung cancer, and they did not separate lung cancer into its main subtypes.

Histological classifications of lung cancer

There are two main histological types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) (35). As its name implies, SCLC (or sometimes referred to as small cell carcinoma or its archaic name oat cell carcinoma) is defined by the World Health Organization (WHO) by the small size of the tumor cells that they define as smaller than the size of 3 small resting lymphocytes (24). SCLC cases comprise only ~13% of total number of new lung cancer cases worldwide (36). Additionally, SCLCs are neuroendocrine tumors unlike other tumors of the lung that are epithelial (bronchial or alveolar) or squamous in origin. Although SCLC can be further differentiated into a number of rare subtypes, most common classifications put SCLC into one main type (24). SCLC is very aggressive, has strong potential for early metastasis, and is difficult to treat (37).

Several classes of histologically different subtypes are classified under the larger lung cancer group NSCLC (24). Although NSCLC comprises many subtypes, the three most prevalent types are lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and large cell carcinoma (LCC). LCC comprises only ~15% of all lung cancer cases and most lung cancer cases are either LUAD (40%) or LUSC (30%) (19, 38). Additionally, LUAD and LUSC are the most

widely studied lung cancer subtypes and will be the NSCLC subtypes that are studied in this dissertation.

LUAD arises from epithelial cells and is most likely to occur on the periphery of the lung. The tumor cells range in size and usually resemble one of six macroscopic patterns. LUAD itself is comprised of many smaller subtypes, but like SCLC, it is often studied as a mixture of these histology types. LUAD is also the lung cancer subtype most likely to occur in never smokers and among women (24).

LUSC arises from cells in the bronchial epithelium, and although several histological variations of LUSC exist, they are usually grouped into one subtype. The tumor cells usually exhibit irregular nuclei and are of an abnormal shape. Additionally, the tumors are often a very large size (24). Among lung cancer subtypes, LUSC usually has the strongest association with smoking, although it is second to SCLC in some studies (24, 39).

Germline influence on lung cancer

Familial lung cancer

Linkage studies have been performed to identify regions of the genome that are linked to lung cancer in families. For example, the Genetic Epidemiology of Lung Cancer Consortium (GELCC) performed a genome-wide linkage analysis and identified a region of the genome at 6q23-25 associated with lung cancer (40). A follow-up study by You *et al*. (41) used microsatellites to fine-map this region to identify the gene responsible for the 6p23-25 peak from

the previous study. Their study concluded that the gene responsible for lung cancer in the linkage study was *RGS17.* An updated linkage analysis was performed by the GELCC in 2010. This new study replicated the earlier results on 6q using a larger number of families. Also, this study identified other regions on 6p, 1q, 8q, and 9p that may also be associated with lung cancer (42). In addition to the regions identified by linkage associated with lung cancer, there are a few cases of familial lung cancers discovered through gene-specific sequencing described below.

In 2005, Bell *et al*. (43) reported a potential inherited risk mutation for lung cancer in a European family. This study found that the T790M mutation in the Epidermal Growth Factor Receptor gene (*EGFR*) was mutated in multiple family members with lung cancer. This specific mutation has also been found in other families with lung cancer (44). Interestingly, this is a commonly observed mutation at the somatic level in lung tumor tissue that confers resistance to targeted EGFR inhibitors (45). Additionally, germline mutations in strong tumor suppressor genes such as p53 lead to an increased lung cancer risk (usually at a younger age) (46), along with many other cancer types.

Candidate gene studies

While there are many genetic variants that have been reported for an association with lung cancer, few have been replicated in large studies according to a review by Brennan *et al.* (47). However, two notable genes that contain germline variants have been replicated in more than one large study or meta-analysis. Variants within the glutathione S transferase M1 (*GSTM1*) gene have been implicated in risk for lung cancer. In a meta-analysis in 2008 that analyzed over 19,000 lung cancer cases and over 25,000 controls, the authors found an increased risk for lung cancer in

European and Asian individuals with germline variants in *GSTM1* (48). Another gene with germline variants associated with lung cancer is *CHEK2* (47). The missense mutation I157T, among other mutations, in *CHEK2* were found to lead to an increased risk of many cancer types in a Polish population (49). However, intriguingly, it was found that rare alleles of *CHEK2* decreased risk for lung cancer, while they increased risk of many other cancer types in a Polish population (50). However, one possibility for this observation is that the subjects died from the other cancers before they developed lung cancer and is therefore possible that risk for lung cancer may not have changed in this population. Many germline variants in lung cancer are thought to be of higher frequency in the population but of lower risk, so an ideal study design to discover these variants is through genome-wide association studies (GWAS).

Genome-wide association studies

The first successful GWAS was performed for age-related macular degeneration (AMD) in 2005 (51). Since then, this technique has been applied to an array of complex diseases or traits. As of February 13, 2017, the GWAS Catalog lists 31,394 unique SNP-trait associations for 2,000 traits across 2,742 studies (52). This large number of identified variants in peer-reviewed articles suggests strong interest in using this approach for many disease types, including lung cancer.

In 2008, there were four major GWA studies for lung cancer reported that identified several variants associated with lung cancer in European populations (53-56). Interestingly, all of the studies found strong genome-wide significant associated SNPs located in region 15q25. This region harbors a set of nicotinic cholinergic receptor (CHRNA) genes. This class of genes had previously been implicated in nicotine dependence (57). However, this region is associated with

lung cancer independent of smoking addiction (54). Additionally, other regions were found with associations in a subset of the studies such as 5p15 (55, 56) and 6p21 (56). Since 2008, there have been many more GWA studies (or meta-analyses) for lung cancer with larger samples sizes (58, 59) and in other racial/ethnic populations (60-62). A summary of these findings is shown in Table 1.1. Although large sample sizes and diverse populations provided additional regions of the genome to interrogate, there are several limitations to these studies. First, many of these studies did not analyze or stratify their results based upon all three major subtypes of lung cancer mentioned above. This made it difficult to investigate the level of association or similarity among the different subtypes. Secondly, since these are association studies, it is difficult to infer the correct causal SNP. It is also possible that the causal SNP may not have been genotyped and instead is detected through linkage disequilibrium (LD) (63). Finally, many of these detected variants are located in non-coding regions of the genome. This makes it difficult to infer a relationship between the non-coding SNP and its target gene(s). Therefore, many of the studies reported results with a set of "most likely" genes that are usually the closest ones in spatial proximity to the SNP based upon distance. Work has now demonstrated that the gene closest to the variant may not be its actual target (64). This lack of understanding for how non-coding variants are acting in lung cancer will be addressed in Chapter II where I will use two functional elements to identify regulatory SNPs and their target genes in three lung cancer subtypes: expression quantitative trait loci (eQTLs) and enhancers.

Table 1.1. Summary of GWA studies for lung cancer.

Lead author	Publication, year	Population/study characteristics	Initial discovery sample size	GWAS platform	Locus (or gene) reported	Genes reported	Outcome studied
Spinola	*Cancer Letters*, 2007	European	338 cases 335 controls	Affymetrix 100K	*KLF6*	*KLF6*	Risk of LUAD in discovery set and mixed NSCLC in replication set.
Amos	*Nature Genetics*, 2008	European	1,154 cases 1,137 controls	Illumina HumanHap300 v1.1	15q25.1	*PSMA4* and *CHRNA*	Risk of NSCLC.
Hung	*Nature*, 2008	European	1,989 cases 2,625 controls	Illumina Sentrix HumanHap 300	15q25	*CHRNA5, CHRNA3, CHRNB4*	Overall risk of lung cancer.
Liu	*JNCI*, 2008	Small family study GWAS	194 cases 219 controls	Affymetrix 500K Affymetrix SNP 6.0	15q24-25.1	*CHRNA3, CHRNA5, CHRNB4, PSMA4*	Risk of familial lung cancer.
Wang	*Nature Genetics*, 2008	European	1,952 cases 1,438 controls	Illumina HumanHap550	15q25.1, 6p21.33, 5p15.33	*BAT3, MSH5, CLPTM1L*	Overall risk of lung cancer in initial analysis, replication, and meta-analysis. Also identified subtype specific risk for top SNPs.
McKay	*Nature Genetics*, 2008	European	3,259 cases 4,159 controls	Illumina Sentrix HumanHap300	5p15.33	*TERT* and *CLPTM1L*	Overall risk of lung cancer.
Broderick	*Cancer Research*, 2009	European	1,952 cases 1,438 controls	Illumina HumanHap550	15q25.1, 5p15.33, 6p21.33	*CHNRA3, IREB2, PSMA4, TERT, CLPTM1L, BAT3, TNXB*	Overall risk of lung cancer in initial analysis and meta-analysis. The three most significant loci were analyzed by their histologies.

Landi	*AJHG*, 2009	European	5,739 cases 5,848 controls	Illumina HumanHap: 317K+240S, 550K, 610QUAD, 1M	15q25, 6p21, 5p15.33 (LUAD)	*CHRNA* region, *CLPTM1L* and *TERT* (LUAD)	Risk of lung cancer in LUAD, LUSC, and SCLC subtypes for initial analysis. Performed meta-analysis for all three subtypes in second study.
Li	*The Lancet Oncology*, 2010	Never smoking study. Mostly European	754 cases 377 controls	Illumina HumanHap: 370K, 610K	13q31.3	*GPC5*	Overall risk of lung cancer in never smokers.
Yoon	*Human Molecular Genetics*, 2010	Korean	621 cases 1,541 controls	Affymetrix 5.0	3q29, 5p15	*C3orf21, TERT, SLPTM1L*	Risk of NSCLC.
Hu	*Nature Genetics*, 2011	Han Chinese	2,383 cases 3,160 controls	Affymetrix 6.0	3q28, 5p15.33, 12q12.12, 22q12.2	*TP63, TERT, CLPTM1L, MIPEP, TNFRSF19, MTMR3, HORMAD2, LIF*	Overall risk of lung cancer. Stratified their six most significant SNPs by histology.
Ahn	*Human Genetics*, 2012	Korean never smokers	446 cases and 497 controls	Affymetrix 6.0	18p11.22	*APCDD1, NAPG, FAM38B*	Risk of NSCLC in never smokers.
Timofeeva	*Human Molecular Genetics*, 2012	European and Chinese meta-analysis	14,900 cases 29,485 controls	Illumina: 317K, 317K + 240S, 370Duo, 550K, 610QUAD, 1.2M	5p15, 6p21, 15q25, 12p13, 9p21	Several for each region	Meta-analysis for risk of lung cancer in LUAD, LUSC, SCLC, and large-cell lung cancer (LCLC).
Lan Q	*Nature Genetics*, 2012	Asian women never smokers	5,510 cases 4,544 controls	Illumina: 370K, 610Q, 660W	10q25.2, 6q22.2, 6p21.32, 5p15.33, 3q28, 17q24.3	*VTI1A, ROS1, DCBLD1, HLA class II* region, among other potential genes	Risk of lung cancer in never smoking women in initial study. Performed replication for 13 most significant SNPs in LUAD.
Dong	*PLoS Genetics*, 2013	Han Chinese	833 cases 3,094 controls	Affymetrix Genome-wide Human SNP Array 6.0	12q23.1	*SLC17A8, NR1H4*	Risk of LUSC.

Wang Y	*Nature Genetics*, 2014	European meta-analysis	11,348 cases 15,861 controls	Illumina: 317, 317 + 240S, 370Duo, 550, 610 1M	3q28 (LUAD)	*BRCA2, CHEK2* rare variants for *LUSC* and *TP63* for *LUAD*	Meta-analysis for risk of NSCLC. Stratified analysis by LUAD and LUSC.
Zanetti	*Lung Cancer*, 2016	African-American	1,737 cases 3,602 controls	Illumina HumanHap 1M Duo	5p15,15q25	*CHRNA5* and TERT	Risk of NSCLC. Stratified analysis by LUAD and LUSC.

These studies were located using the GWAS Catalog (52).

Functional elements

Non-coding functional variants can be classified into many categories, including transcription binding sites (TFBS), splice sites, methylation related CpG islands, eQTLs and enhancers. Among these non-coding variants, some are associated with a specific trait such as eQTLs and GWAS significant SNPs, while others, such as TFBS, are predicted to be functional based upon genomic context. Among these functional elements, eQTLs and enhancers have gained much attention due to their strong roles in gene regulation and disease association and the recent release of many high-quality data sets.

eQTLs are genetic variants in the genome that are correlated with variations in gene expression. These regulatory elements may be tissue specific, so it is necessary to use the correct disease related tissue for eQTL significance (65). Large collaborations such as the Genotype-Tissue Expression (GTEx) project (66) have generated sets of eQTLs in over 40 different human tissue types. Additionally, GTEx also generated sets of multi-tissue eQTLs that act in multiple tissues (67). Past work has also demonstrated that GWAS hits are enriched in eQTLs (68) and other regulatory regions (69) of the genome.

Enhancers are DNA sequences in the genome that can enhance the transcription of a gene or genes. They are mostly located in non-coding regions of the genome and can influence transcription of up-stream or down-stream genes. Additionally, enhancers may act on genes that are not their closest neighbors (70). Although identifying or predicting enhancers based on DNA sequence alone is difficult, many epigenetic marks associate with enhancers and their activity (70). This feature allows one to identify enhancers using experimental techniques such as Chip-Seq (70). Many methods are now available to predict enhancers in many different tissue types

integrating many different data types (71, 72). In addition to the epigenetic marks associated with enhancers, it was discovered that enhancers can also be transcribed into enhancer RNAs (eRNAs). The Functional ANnoTation of the Mammalian genome (FANTOM) project (73) used these eRNAs to identify enhancer regions across multiple tissue and cell-lines using the Cap Analysis of Gene Expression (CAGE) method (74). Recent work (75) has found that GWAS variants located in enhancers are important for many cancer types such as prostate cancer, breast cancer, and colorectal cancer.

Lung cancer from a somatic perspective

Although germline studies of lung cancer can provide insight into risks involved with lung cancer, studying somatic mutations in lung tumor tissue can help identify the genomic aberrations driving the tumor growth (1). Therefore, there has been much effort involved in the determination of the genetic aberrations in lung tumor tissue. These studies can be separated into smaller candidate gene studies and larger genome-wide tests to identify any locations in the genome that may be linked with lung cancer.

Candidate somatic studies

One approach to identify genetic alterations associated with lung cancer is to study gene sets that may have previous evidence for their involvement in cancer or cellular proliferation. Below, I summarize a single multi-gene study that has progressed the lung cancer research field for each lung cancer subtype.

LUAD

One of the first major multi-gene investigations to identify mutated genes associated with LUAD using sequencing was performed in 2008. Ding *et al*. (76) sequenced 623 known cancer related genes to identify novel mutations in lung cancer. In addition to replicating known gene associations with lung cancer such as *TP53, CDKN2A, STK11, KRAS, EGFR* and *NRAS*, this work also identified novel mutations in lung cancer including tumor suppressor genes (TSGs) *ATM,* and *RB1,* and the proto-oncogene tyrosine kinase *ERBB4*.

LUSC

In 2011, Hammerman *et al*. (77) sequenced 201 genes including all known kinase genes at that time to identify somatic mutations in LUSC using 20 samples. They found mutations in 25 genes (including *p53*) in this initial sample set. They performed a second screen with six kinase genes (*DDR2, FGFR2, NTKRK2, JAK2, FLT3*, and *CDK8*) that were mutated in the first phase. They identified many mutations in the *DDR2* gene and replicated it in an additional five samples from a validation cohort of 222 samples, confirming its role in LUSC.

SCLC

There are no large candidate-based sequencing studies for SCLC comparable to the studies mentioned above for LUAD and LUSC. However, smaller single gene studies have identified many genes with mutations that may be involved in the disease process such as *p53, PTEN, PIK3CA,* and *RB* (78).

Genome-wide somatic studies

In contrast to candidate gene studies, genome-wide studies for cancer interrogate the entire genome to identify genes that may influence lung cancer. Below, I highlight some of the most expansive genome-wide somatic studies for lung cancer in each subtype.

LUAD

Since the introduction of genome-wide genetic technologies, there have been several efforts to search for driver mutations in lung cancer. In 2007, Weir *et al*. (79) used SNP arrays to determine a set of copy number alterations across the genome for 371 LUAD tumor samples. This analysis found over 50 copy number changes across the genome for the LUAD samples. This genome-wide analysis was able to identify CNVs in LUAD, but was not able to obtain the level of mutational data generated using DNA sequencing. In 2012, one of the first efforts to use NGS to interrogate lung cancer on multiple samples was done. Govindan *et al*. (80) performed whole genome sequencing (WGS) and transcriptome sequencing (RNA-Seq) on 17 patients with NSCLC. Intriguingly, this study identified many chromatin modification genes that were significantly mutated. They also discovered novel fusion genes through their RNA-Seq analysis. In that same year, Imilinski *et al*. (81) performed a combination of WGS and WES on over 180 LUAD samples. Their work discovered several somatic mutations and insertions/deletions (indels) in LUAD. A few years later, The Cancer Genome Atlas (TCGA) working group published their results on LUAD (82). The TCGA group generated germline SNP data, somatic mutation data, mRNA sequencing data, microRNA sequencing data, methylation data, copy number alterations data, and protein expression levels using 230 LUAD tumor and matched normal samples. This

unprecedented level of data and methodological approaches discovered many new genomic alterations associated with LUAD (82). Work in Chapter III uses some of these results from somatic mutation calling and mRNA sequencing for the somatic analysis performed in that chapter.

LUSC

In 2009, Bass *et al*. (83) used SNP arrays to investigate the copy number alterations associated with LUSC and esophageal squamous cell carcinoma (ESCC) in 40 ESCC samples and 47 LUSC samples. This study identified amplifications and deletions in both cancer types. Interestingly, they found that genomic region 7p11.2 was amplified in both cancer types with *EGFR* as the target gene. *EGFR* is mutated in many NSCLC subtypes, but most are found in LUAD (35). Another region that was amplified in both cancer types was 8p12 that includes the candidate genes *FGFR1* and *WHSC1L1*. Their study also discovered *SOX2* as an amplified oncogene in both cancer types. The following year, Weiss *et al*. (84) used a much larger set of LUSC tumor samples (n = 155) to identify copy number alterations using a SNP array. Their study identified over 50 amplifications and deletions and confirmed the previous year's finding of *FGFR1* and *SOX2*. Interestingly, they looked at previously published results from LUAD (85) and observed that 8p12, which contains *FGFR1*, is not amplified in LUAD. In 2012, TCGA published their initial analyses on LUSC (86). For their analyses, they generated genome-wide data for mRNA sequencing, microRNA sequencing, copy number alterations, somatic mutations, and methylation levels. Their results indicated many genomic alterations in LUSC that were previously unknown. Intriguingly, they found somatic mutations in the *HLA-A* for the first time in lung cancer that suggests an immune role in this cancer subtype.

SCLC

In 1995, Levin *et al*. (87) examined copy number alterations in SCLC using comparative genomic hybridization (CGH) for ten SCLC samples. Their results indicated many regions of the genome that were increased (gain, or amplification) or decreased (loss) in copy number. Of note, they found copy number gains in regions that contained *MYC*, a known oncogene overexpressed in multiple cancer types (75). They also found decreased copy numbers in many genomic regions that harbor well-known TSGs *p53* and *RB*. In 2012, two large SCLC studies were published in the same issue of *Nature Genetics* (88, 89). Peifer *et al*. (88) generated WES and copy number alterations for ~30 samples, and they performed WGS on two of the samples and RNA-Seq on 15 of the samples. Their study found amplifications in 8p12, which contains the *FGFR1* gene, and a single sample with amplification of the *MYC* region. Their work also identified a set of significantly mutated genes in addition to sets of fusion genes. Rudin *et al*. (89) generated WES, RNA-Seq, and copy number alteration data on over 50 SCLC samples. Their analysis identified thousands of somatic mutations and many copy number alterations. They replicated the *MYC* amplification from the previous study and also found amplifications in the *SOX2* gene. They found many gene fusions, including four gene fusions involving kinase genes. Most recently, in 2015, George *et al*. (90) published the genomic profiles of WGS for 110 SCLC samples, RNA-Seq for 71 of these samples, and SNP copy number arrays for 103 of the samples. This study revealed several new genomic alterations in SCLC in genes such as *TP73* and many NOTCH signaling genes.

Overlap of genetic features in lung cancer subtypes

Results from the publication of the two large SCLC studies in 2012 (88, 89) highlighted above gave an indication of somatically mutated cancer driver genes, and oncogenic gene fusions, that are shared between LUAD, LUSC, and SCLC. These new results, in combination with previously discovered genomic alterations, reveled shared gene sets between subtypes (91). It was revealed that all three subtypes shared these somatic mutated driver genes: *TP53*, *CDKN2A*, *PIK3CA*, and *PTEN*. Only one gene, *KEAP1*, was uniquely shared between LUSC and LUAD, while two genes, *FGFR1*, and *SOX2*, were both driver genes shared between only LUSC and SCLC. This also revealed that many driver genes are unique to each subtype. These mutated driver genes and their comparison to each other are listed in Table 1.2.

A recent study (92) performed a deep comparison of genomic features in LUAD and LUSC. Campbell *et al*. (92) used WES and copy number profiles to identify CNVs and somatic mutations in over 1,000 tumor-normal matched pairs. Interestingly, they found that only six mutated genes (*TP53, RB1, ARID1A, CDKN2A, PIK3CA, and NF1)* were shared between the two subtypes. Three of these genes (*TP53*, *CDKN2A*, and *PIK3CA)* were previously found in all three lung cancer subtypes, *RB1* was previously only found in SCLC, and *ARID1A* and *NF1* were not previously found to be driver genes in lung cancer. This dissertation will be the first effort as a comprehensive comparison of LUAD, LUSC, and SCLC across the germline and somatic genomes.

Table 1.2. Summary of mutated driver genes in three lung cancer subtypes.

LUAD only	**LUSC only**	**SCLC only**	**LUAD-LUSC shared**	**LUAD-SCLC shared**	**LUSC-SCLC shared**	**Shared by three**
EGFR	*NFE2L2*	*RB1*	*KEAP1*	Nothing uniquely shared	*FGFR1*	*TP53*
KRAS	*TP63*	*RLF-MYCL1*			*SOX2*	*CDKN2A*
ERBB2	*NOTCH1*	*MYCL1*				*PIK3CA*
BRAF		*MYCN*				*PTEN*
ALK fusions		*MYC*				
ROS1 fusions						
RET1 fusions						
STK11						

Table is based off genes reported in Figure 1 in Pietanza and Ladanyi, *Nature Genetics*, 2012 (91).

Summary and overview of dissertation

Although cigarette use has strong correlations with lung cancer risk, genetic factors also play major roles. Lung cancer consists of several subtypes; the most common of these are LUAD, LUSC, and SCLC. Previous work has identified potential genetic risk for these subtypes from the germline level using small candidate-gene approaches and also GWA studies. Additionally, somatic studies have identified genes that have major roles in lung cancer tumor formation and growth. However, there is a lack of information on how genetically similar these three subtypes are across both genomes. This dissertation aims to perform a detailed interrogation of these three subtypes for the germline and somatic genomes.

In Chapter II, I investigate these three major subtypes from the germline perspective. I use common genetic variants discovered in a GWAS for LUAD, LUSC, and SCLC to investigate the common and distinct biology behind these three subtypes of lung cancer. Most GWA studies for lung cancer are not analyzed by each subtype separately, and most identified variants do not have a well-identified biological role or associated gene. Therefore, I identify a set of regulatory variants for each subtype. I also link these regulatory variants to their target genes using functional genomics data to provide insight into the biology behind each disease. Finally, I determine enriched biological pathways that contain these target genes.

In Chapter III, I use biological data from the somatic perspective to investigate these subtypes. I utilize gene expression levels identified from RNA sequencing (RNA-Seq) and DNA mutations identified from whole exome sequencing (WES) to interrogate the disease processes underlying each subtype. I generate a set of genes that are differentially expressed in the tumor versus normal tissue using the RNA-Seq expression data and use these genes to identify perturbed

biological pathways for each subtype. I also identify somatic mutational signatures for each subtype. These signatures can be used to help identify similarities behind specific cancer types by utilizing their unique pattern of somatic mutational processes. I use filtering and expression methods to obtain a set of potential driver genes for each subtype and investigate their overlap among the subtypes.

In Chapter IV, I explore the possibility of identifying somatic mutations using RNA-Seq in place of WES in lung cancer by performing a systematic comparison of variants identified in WES versus RNA-Seq. Specifically, I discover somatic mutations from lung cancer samples that have undergone WES and RNA-Seq and compare several technical and biological features of the mutations identified by each method.

In Chapter V, I apply the approach to identify GWAS regulatory variants and their target genes from Chapter II to other disease types. I utilize other lung diseases, population types, and cancer types to expand the usefulness of this approach. This chapter demonstrates the approaches from Chapter II can be expanded beyond lung cancer.

In Chapter VI, I summarize the major findings from Chapters II-IV. The main focus of the summary consists of ways that the germline findings from Chapter II and the somatic findings from Chapter III can be investigated together to identify unique new biological insights behind LUAD, LUSC, and SCLC. I finish with future directions that can be undertaken with these results to gain a clearer picture into the genetics and biology behind LUAD, LUSC, and SCLC.

CHAPTER II

EXPLORTATION OF THE GERMLINE GENOME IDENTIFIES WEAK SHARING OF GENETIC ASSOCIATION SIGNALS IN THREE LUNG CANCER SUBTYPES: EVIDENCE AT THE SNP, GENE, REGUALTION, AND PATHWAY LEVELS.

Introduction

Out of the three major subtypes of lung cancer, the non-small cell lung cancer (NSCLC) subtypes LUAD, and LUSC comprise ~60% of newly reported lung cancer cases, while SCLC comprises only a small subset (~15%) (93). One commonly used approach to identify variants associated with these lung cancer subtypes is to perform a GWAS. In 2007, Spinola *et al*. (94) performed a small GWAS in a European population of ~ 300 cases and ~ 300 controls and identified a variant near the KLF6 gene associated with lung cancer. The following year in 2008, four GWA studies were published (53-56) with all of the studies identifying variants in the 15q25 region that showed associations with lung cancer. Since 2008, several GWA studies and meta-analyses, (26, 58-62, 95-99) have discovered several common variants associated with lung cancer risk. However, only about half of these studies (54, 58-60, 95, 96) had data for all three lung cancer subtypes.

Additionally, most of these GWAS findings did not reach the stringent genome-wide significance in a single GWA study ($p < 5 \times 10^{-8}$), and most of the genome-level significant single nucleotide polymorphisms (SNPs) were located within non-coding regions of the genome, making it difficult to infer the underlying mechanism of the significant variants that could contribute to

disease. Since genome-wide significance level has been thought too stringent (i.e., by Bonferroni correction), those common SNPs with moderate association signals (e.g., $p < 1 \times 10^{-3}$) have been found informative in exploring the association evidence (100, 101). Moreover, recent studies have demonstrated that these marginally significant SNPs found from GWAS within non-coding regions of the genome may function in regulatory roles (68, 69). Therefore, one can use these results to obtain a set of regulated genes to investigate and compare the similarity of the three lung cancer subtypes at the germline gene level and at the regulation level.

In this study, we first identified a set of SNPs with moderate association signals ($p < 1 \times 10^{-3}$) from a prior GWAS (96) that covered three lung cancer subtypes, LUAD, LUSC, and SCLC. Then, we identified and compared regulatory variants associated with the three subtypes of lung cancer as well as their target genes. We used these results to investigate the similarity of the subtypes at the SNP, gene, regulatory, and pathway levels. We first remapped these SNPs to an updated genome reference (hg19) and expanded them using linkage disequilibrium (LD) patterns from a European population. We used this final set of SNPs to examine several lung tissue expression quantitative trait loci (eQTL) and enhancer datasets for evidence of regulatory function of each SNP and identified their target genes. When we compared the target genes of these regulatory SNPs, we observed that only five genes overlapped all three subtypes. Through this analysis, we have identified many genes that might have an important association with lung cancer for each specific subtype.

Methods

GWAS dataset

We used data from a previously performed multi-site GWAS for lung cancer in a European population that analyzed each sample by lung cancer subtype for the NCI GWAS for lung cancer (96). Briefly, this GWAS for lung cancer used cases and controls from four different studies, EAGLE, ATBC, PLCO, and CPS-II. After quality control (QC) of the genotyping results, there remained 5,739 cases and 5,848 controls of European ancestry and 515,922 SNPs. The analysis was stratified by lung cancer subtype with 1,730 LUAD cases, 1,400 LUSC cases, 678 SCLC cases, 5,848 shared controls, and was analyzed using unconditional logistic regression. We used the full set of significant lung cancer GWAS SNPs ($p < 1 \times 10^{-3}$) separated by subtype for this analysis.

Genomic annotation of GWAS SNPs

The online web tool SNP Nexus (102, 103) (http://snp-nexus.org/) was used to annotate the genomic location of the significant SNPs by lung cancer subtype using the NCBI36/hg18 genome assembly. We used the University of California Santa Cruz (UCSC) hg18 gene definitions for the genomic annotation of each region.

Converting hg18 SNPs to hg19 SNPs

The results from the lung cancer GWAS were originally generated using coordinates from the hg18 reference of the human genome. We converted these SNPs to hg19 coordinates using the online tool Remap from the National Center for Biotechnology Information (NCBI) with default settings (http://www.ncbi.nlm.nih.gov/genome/tools/remap). This conversion allowed us to map the SNPs to the regulatory annotation information, which are based on hg19 coordination.

We used these updated hg19 coordinates for the SNPs to obtain the updated SNP rsID numbers using dbSNP data for build 142 from the NCBI to account for any SNPs that may have been merged between assemblies. All the chromosomes with updated SNP IDs and coordinates for GRCh37.p13 (hg19) dbSNP b142 were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b142_GRCh37p13/chr_rpts/) on February 18, 2015 and matched with SNP results from Remap.

Identification of SNPs in LD with the genotyped SNPs

For each GWAS SNP, we retrieved all other SNPs in a 1Mb region both upstream and downstream from the SNP site using Tabix (104) (version 0.2.5). We obtained the SNP data from the European super population group from the 1000 Genomes Phase III data (v5.20120502). Vcftools (105) (version 0.1.12b) was used to convert the Tabix vcf files to the plink-tped file format. Then, we used the 1000 Genomes data for each GWAS SNP and applied PLINK (106) (version 1.07) to identify the final set of SNPs that were in LD with the tagging SNPs using an $r^2 > 0.8$ within the entire region 1Mb upstream and downstream of the SNP. The LD results from PLINK were combined for every SNP and any SNPs in LD that were duplicated across all SNP

sets were removed. These are the final set of LD SNPs for the analysis. Pipeline containing work flow is illustrated in Figure 2.1.

GTEx eQTLs

The full set of significant human lung tissue-specific eQTLs version 6 (V6) was downloaded from the GTEx website (https://www.gtexportal.org) on February 22, 2016. The eQTLs were identified using linear regression with the tool Matrix eQTL (107) with a +/- 1Mb region around the transcription start site in each individual tissue that had >70 samples. The significance of the eQTLs was determined by empirical p-values using permutations followed by a Storey false discovery rate (FDR). The eQTLs with a q-value ≤ 5% were considered significant.

We also downloaded the full set of all multi-tissue eQTLs for nine different tissue types from the pilot phase of the GTEx Project on June 11, 2015. This file contained eQTLs discovered using two different methods, the University of Chicago (UC) model (108) and the University of North Carolina at Chapel Hill (UNC) model (109). We used the results that contained the average between both methods including calculated posterior probabilities for every gene-snp pair titled "res_final_amean_com_genes_com_snps_all.txt." The whole SNP set (including LD SNPs) was used to detect eQTLs in this dataset. We plotted the distribution of posterior probabilities of all the eQTLs found using the SNPs and defined an eQTL as "significant" if its posterior probability was >80%. We removed all duplicated genes in each subtype to obtain the final GTEx set of genes.

Lung tissue eQTLs from Hao *et al*. study

Hao *et al*. (110) investigated how genetic variation affects gene expression levels in human lung tissues. The authors used lung tissue and blood from more than 1,000 patients across three cohorts to identify a set of eQTLs in lung tissue and used their results to interrogate SNPs associated with asthma. We downloaded the entire set of cis-eQTLs in lung tissue identified from this study with FDR at 10%. We removed the target genes without annotated gene names in order to combine the genes with the results from our other analyses. If more than one SNP-gene pair were identified as eQTLs, but differed in their probes used, we considered them distinct eQTLs. We removed duplicated genes in each subtype to define the final Hao *et al*. set of genes.

FANTOM5 transcribed enhancers

The FANTOM consortium aims to identify and assign regulatory function to the mammalian genome. Part of this comprehensive project is to identify all transcribed enhancers and promoters in multiple human cell lines and tissue types. The entire set of permissive enhancers found in the FANTOM5 data was downloaded from http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed on August 26, 2015. The gene-report function was used in PLINK v1.07 to search for any SNPs that were located within permissive enhancer regions defined by FANTOM. To identify the possible target genes of these enhancers, these enhancer regions were then matched with the set of FANTOM5 enhancer transcription start site's significant associations downloaded from http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed on August 25, 2015.

IM-PET predicted enhancers

He *et al*. (72) developed a novel approach to identify the targets of histone-derived enhancers using a random forest classifier. The authors used this approach to define a set of enhancer target genes for 12 different cell types. We used the results for two lung cell types, IMR90 and NHLF, from the supplemental tables of their publication (72) for our analysis. We used Bedtools (111) version 2.17.0 to identify lung cancer GWAS SNPs located within the enhancer regions that had an associated target gene. To remove non-expressed genes, we filtered the results to remove target genes with Reads Per Kilobase per Million mapped reads (RPKM) = 0. The enhancer targets were originally formatted as Ensembl defined transcripts, so we converted them to gene symbols using the BioMart tool from Ensembl using the archived site pertaining to genome assembly CRCh37.p13 (112).

Locus level analysis

biomaRt (113) was used to annotate the genomic locations for the germline-regulated genes discovered from each dataset for each subtype using gene start and stop coordinates from Ensembl gene definitions using genome build GChR37.3. Genomic locations that were not defined from Ensembl, were manually annotated using NCBI's Gene online web resource https://www.ncbi.nlm.nih.gov/gene. The function "cluster" from Bedtools (111) was used to cluster the genes into independent 1Mb regions.

Pathway enrichment analysis

The final set of germline-regulated genes was uploaded to the WebGestalt online resource (114). The hypergeometric test was used for enrichment with specific pathways followed by Benjamini & Hochberg multiple test correction (115).

GWAS Catalog SNPs

We downloaded all SNPs from the GWAS Catalog using the search term "lung cancer" on January 13, 2016. We removed the SNPs where the initial or replication population was other than European. We also removed the SNPs that were reported in Landi's original lung cancer GWAS report (96) because we used them for our analyses, so we could not use them for any replication purposes.

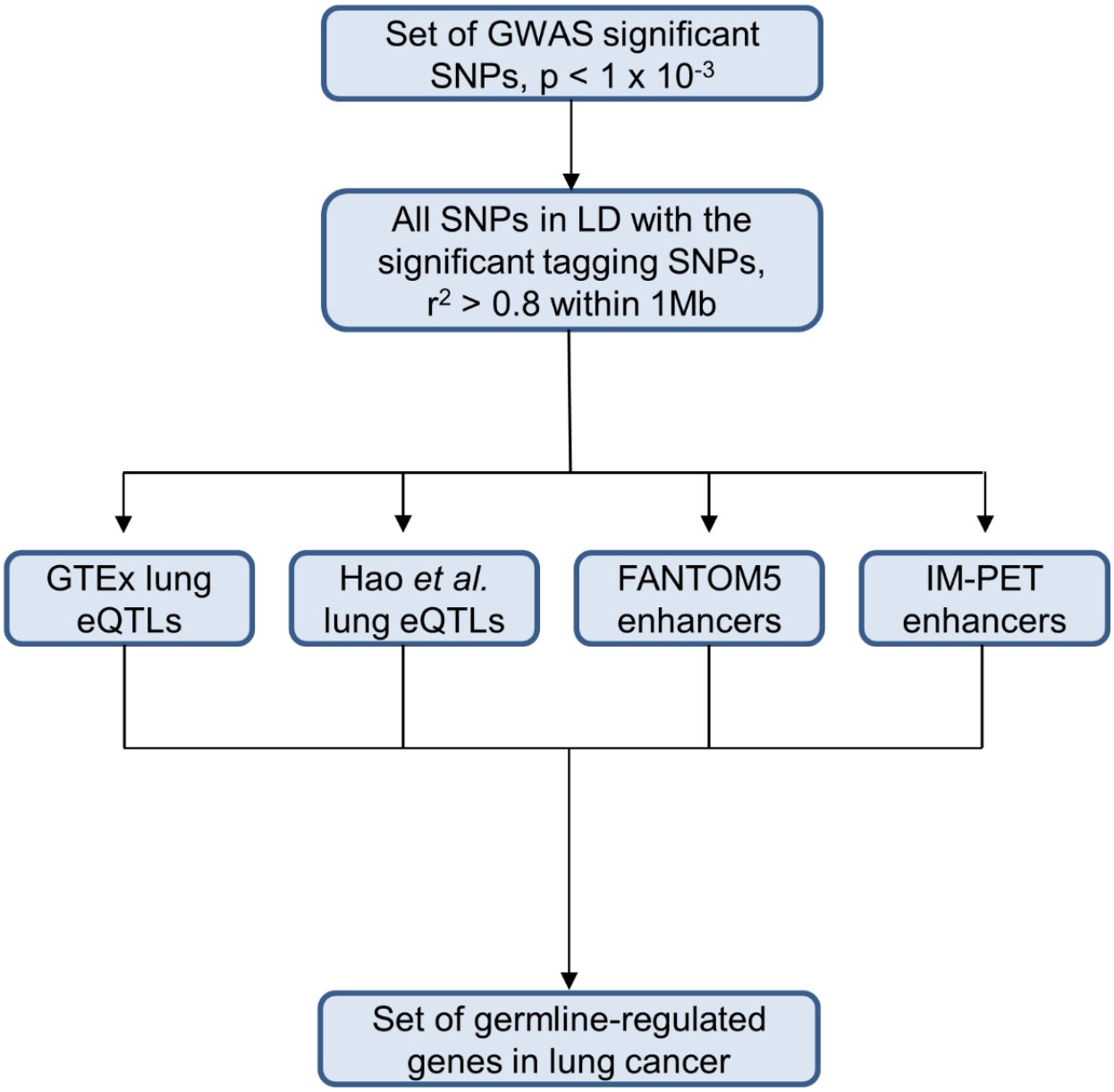


Figure 2.1. Pipeline to identify a set of germline genes for SNPs that were moderately associated with three subtypes of lung cancer from the genome-wide association studies (GWAS) (96). SNPs were run through the pipeline separately for each lung cancer subtype: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and small cell lung cancer (SCLC). After LD expansion, two eQTL and two enhancer datasets were used to identify expanded or original SNPs that were within regulatory regions with an identified, or predicted, target gene.

Results

Description of data and SNP expansion

We obtained SNPs ($p < 1 \times 10^{-3}$) for three lung cancer subtypes, LUAD, LUSC, and SCLC, from a National Cancer Institute (NCI) GWAS for lung cancer (96). This GWAS utilized cases and controls from four smaller studies: Environment and Genetics in Lung Cancer Etiology (EAGLE), Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC), Prostate, Lung, Colon, Ovary (PLCO) screening trial, and the Cancer Prevention Study II (CPS-II) nutrition cohort. Subjects from these four smaller studies were genotyped at two institutions. The EAGLE samples and part of the PLCO samples were genotyped at the Center for Inherited Disease Research (CIDR). The ATBC, CSP-II, and part of the PLCO samples were genotyped at the Core Genotyping Facility (GCF) at the NCI. Table 2.1 shows the characteristics of these samples and their study origin.

Table 2.1. Summary of data used from GWAS for lung cancer.

Study	Cases	Controls	Population	Illumina HumanHap platform
EAGLE*	1917	1978	European ancestry	550K, 610QUAD
ATBC**	1732	1270	European ancestry	550K, 610QUAD
PLCO*/**	1355	1896	European ancestry	317K, 240S, 550K, 610QUAD
CSP-II**	695	674	European ancestry	550K, 610QUAD, 1M
Total	**5699**	**5818**		

Number of cases for the three subtypes: lung adenocarcinoma (LUAD) = 1730, lung squamous cell carcinoma (LUSC) = 1400, small cell lung cancer (SCLC) = 678.
* Genotyped at the Center for Inherited Disease Research (CIDR).
** Genotyped at the Core Genotyping Facility (GCF).

Figures 2.2-2.4 are Manhattan plots for each lung cancer subtype. These plots illustrate the significance and location of variants we used for our analysis. This stratified GWAS by subtype confirmed previous lung cancer associations at the 15q25 locus (53, 54) for each subtype. Table 2.2 shows the total number of cases genotyped for each subtype, the total number of SNPs discovered by selection criterion ($p < 10^{-3}$), and distribution of their locations within the genome. Interestingly, only 10 SNPs (<1%) overlapped all three subtypes (Figure 2.5A). We found that, similar to many GWA studies for various disease types, only 2-3% of variants were located within coding regions of the genome.

Table 2.2: Summary of SNP results from lung cancer GWAS.

Subtype	Sample size	SNPs ($p < 1x10^{-3}$)				
		Total SNPs	Coding	Intron	UTR	Intergenic
LUAD	1730	544	13	228	7	296
LUSC	1400	598	18	299	16	265
SCLC	678	558	14	247	10	287

LUAD: lung adenocarcinoma. LUSC: lung squamous cell carcinoma. SCLC: small cell lung cancer.

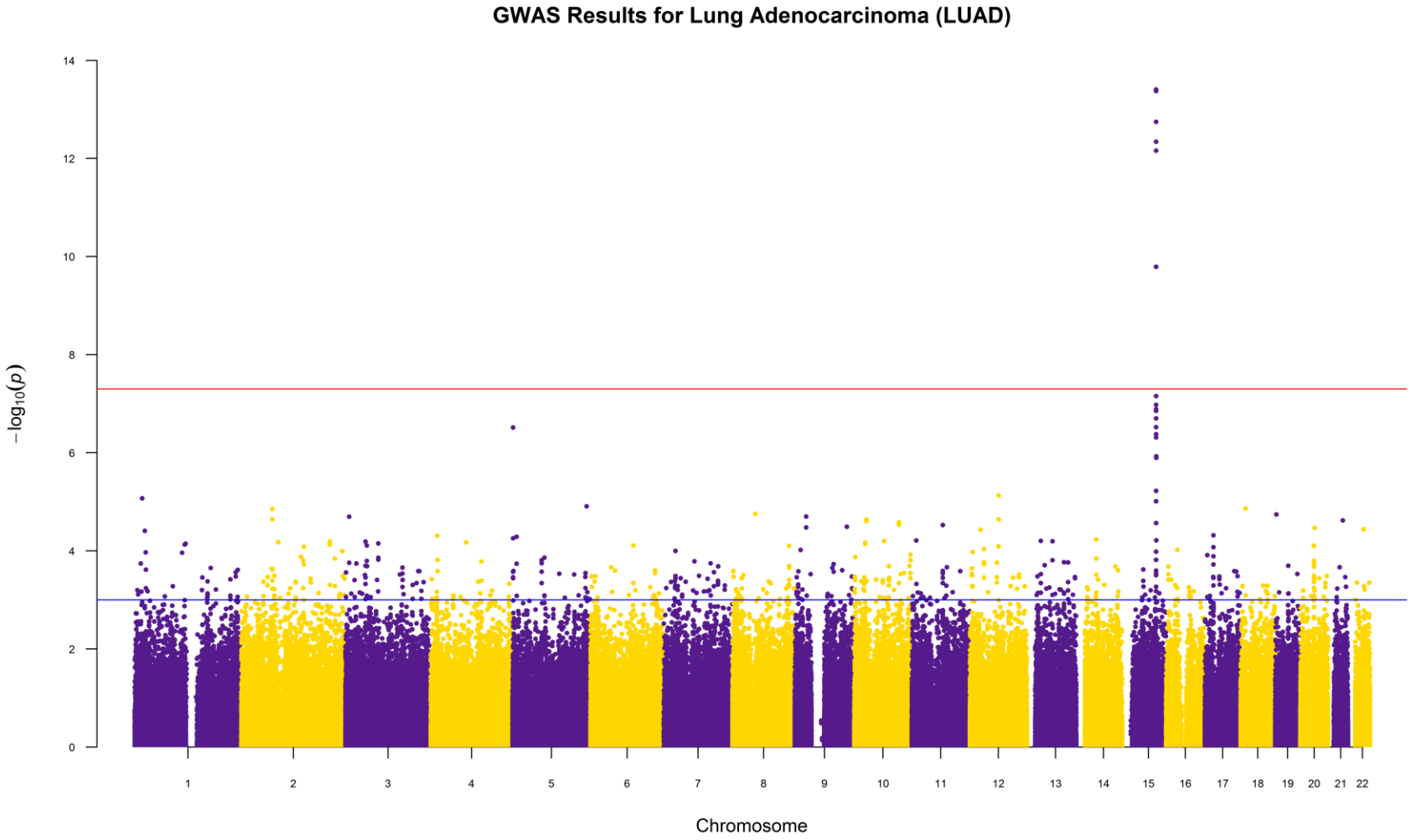


Figure 2.2. Manhattan plot of GWAS results for LUAD. Red line represents genome-wide significance for GWAS ($p < 5 \times 10^{-8}$). Blue line represents significance level of the SNPs ($p < 1 \times 10^{-3}$) used in this study.

Figure 2.3. Manhattan plot of GWAS results for LUSC. Red line represents genome-wide significance for a GWAS ($p < 5 \times 10^{-8}$). Blue line represents significance level of the SNPs ($p < 1 \times 10^{-3}$) used in this study.

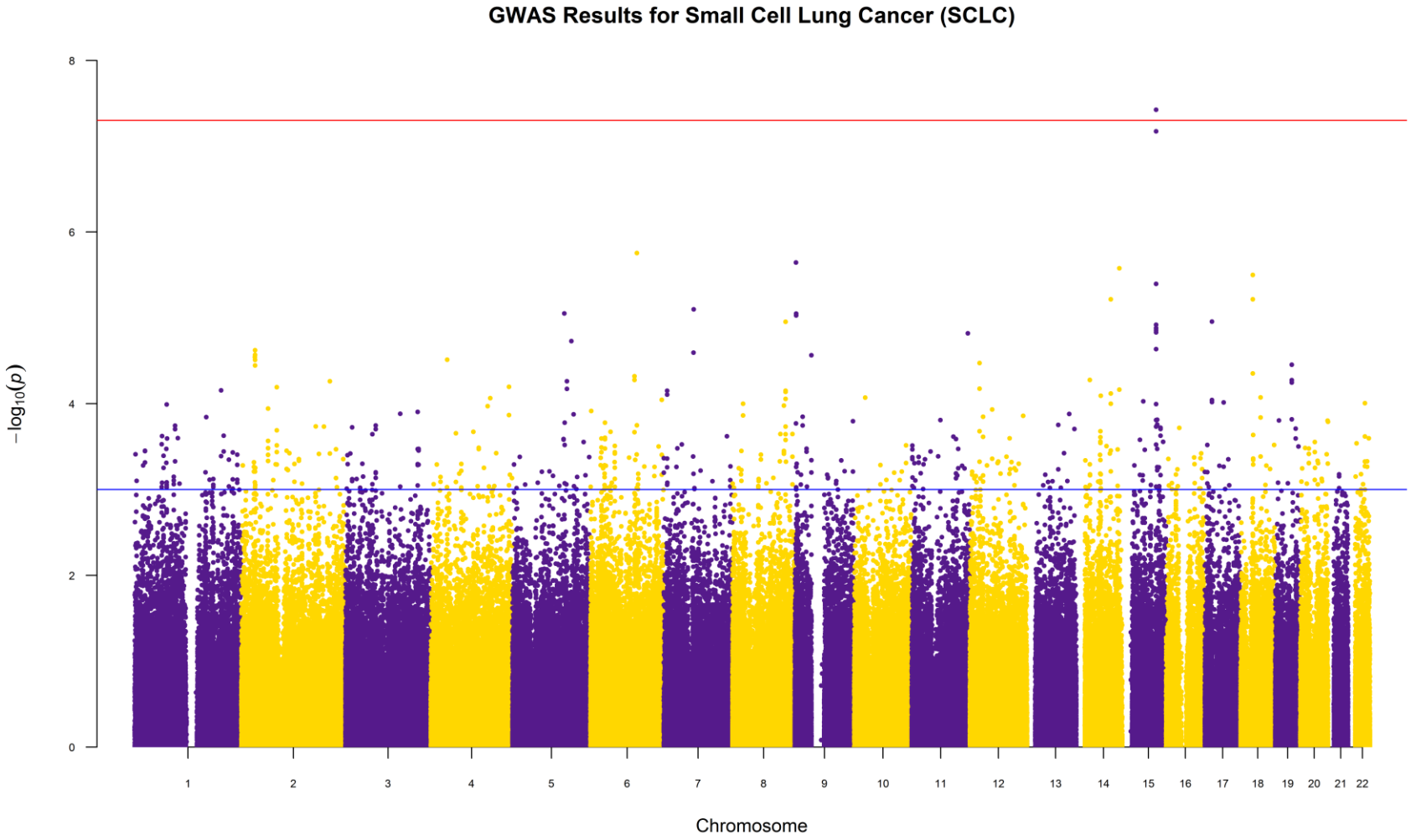


Figure 2.4. Manhattan plot of GWAS results for SCLC. Red line represents genome-wide significance for a GWAS ($p < 5 \times 10^{-8}$). Blue line represents significance of the SNPs ($p < 1 \times 10^{-3}$) used in this study.

We generated a set of SNPs in LD with these genotyped SNPs for each subtype (see Methods), and summarize the LD expansion in Table 2.3. After we removed duplicated SNPs within each subtype, we found 8295 SNPs associated with LUAD, 8734 with LUSC, and 8361 with SCLC (Figure 2.5B). As illustrated in Figure 2.5C, we found very little correlation (Pearson Correlation Coefficient (PCC) < 0.03) between the p-values of the subtypes.

Table 2.3. Sample results and LD expansion.

	LUAD	LUSC	SCLC
SNPs (GWAS, $p < 10^{-3}$)	544	598	558
SNPs (LD, $r^2 > 0.8$, within 1Mb)	14312	16021	13104
duplicated SNPs	6561	7885	5301
Final SNPs	8295	8734	8361

The details of SNP data processes are provided in main text. LD: linkage disequilibrium. LUAD: lung adenocarcinoma. LUSC: lung squamous cell carcinoma. SCLC: small cell lung cancer.

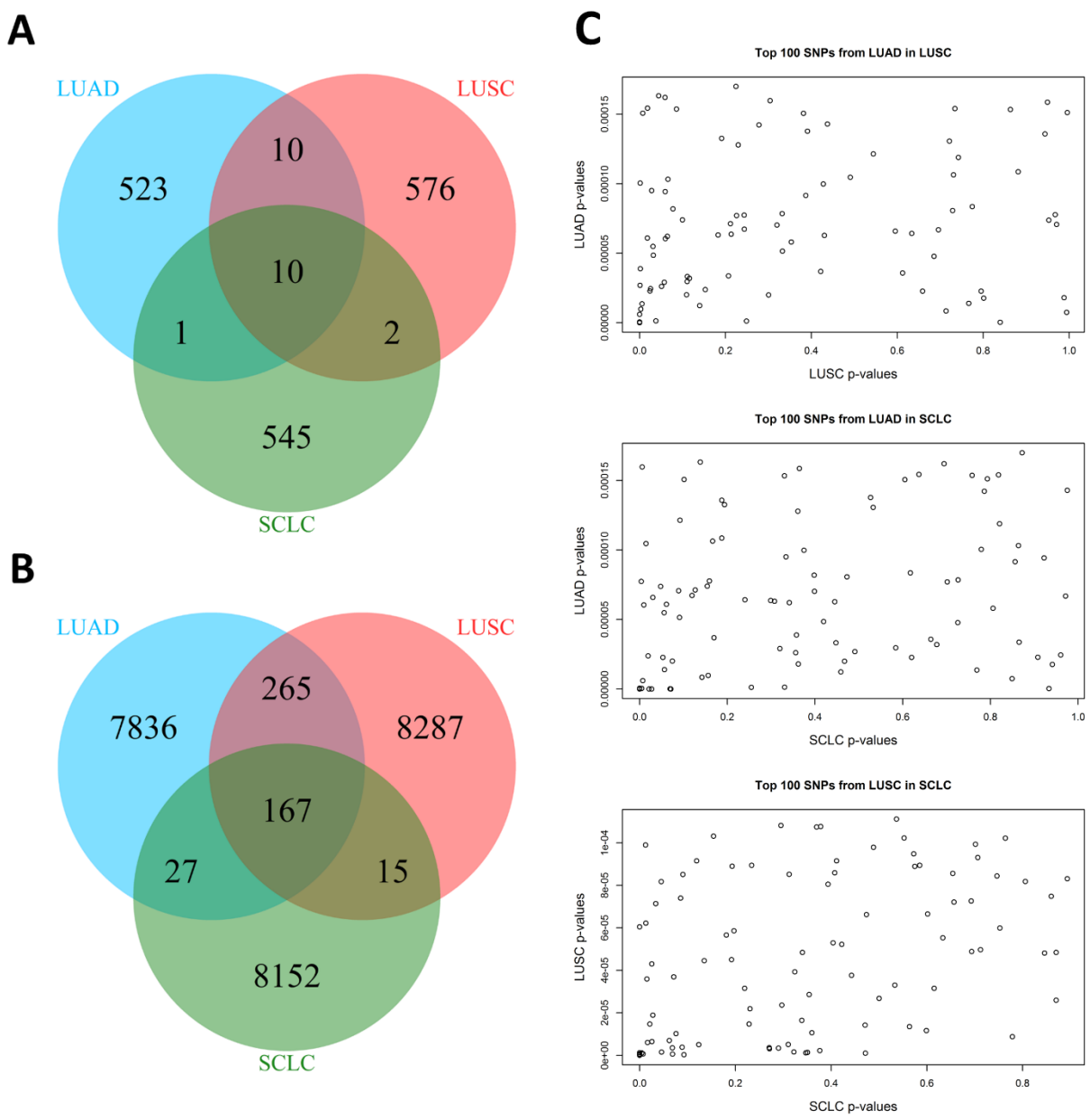


Figure 2.5. Comparison of SNPs from GWAS for lung cancer. Venn diagrams show overlap of SNPs found in GWAS ($p < 10^{-3}$) for lung cancer by subtype (A) and after LD expansion (B). (C) Plots show the top 100 variants for each subtype plotted against each subtype. All evidence shows very little overlap or correlation between GWAS significant SNPs per subtype.

Lung tissue eQTLs

We first utilized three sets of lung eQTLs to annotate the SNPs. The first lung eQTL dataset was retrieved from the Genotype-Tissue Expression (GTEx) project (67). Using this dataset, we found 1,297 SNPs for LUAD, 1,429 for LUSC, and 1,171 for SCLC (Figure 2.6A) that acted as eQTLs using a set of pre-compiled significant lung tissue-specific eQTLs in GTEx. To explore all eQTLs for lung, including non-tissue-specific eQTLs, we used a second set of eQTLs identified using a multi-tissue model from GTEx. These eQTL results were generated using a statistical model different from the single tissue eQTLs (see Methods). We gauged significance based upon the distribution of multi-tissue eQTLs in each subtype (Figure 2.7). We combined the single and multi-tissue eQTLs represented by the SNPs to form the final set of GTEx eQTLs. Many of these eQTL SNPs were within strong LD of each other and controlled the expression of the same target gene, so we collapsed all eQTLs to the specific genes they control. As illustrated in Figure 2.6B, we found a total of 71 genes for LUAD, 108 for LUSC, and 67 for SCLC. Three genes overlapped from one unique signal in all three subtypes (*CHRNA5, PSMA4,* and *RP11-650L12.2*). *CHRNA5* is in the nicotinic acetylcholine region that has well-known associations with lung cancer (53, 54, 116), while *PSMA4* has also been reported to be associated with lung cancer (117).

We examined a third set of lung tissue eQTLs generated from a meta-analysis that used lung tissue samples from three different recruitment sites (not including GTEx data) (110). We refer to this set of eQTLs as the Hao *et al.* eQTLs. We found 25 SNPs for LUAD, 34 for LUSC, and 16 for SCLC that acted as eQTLs (Figure 2.6C). We reduced the number of eQTLs to unique target genes (see Methods) and found no genes that overlapped all three subtypes, no genes that overlapped LUAD and SCLC, two genes that overlapped LUSC and SCLC in one genomic region (*MYL4 and RPRML*), and one gene (*IREB2*) that overlapped the two NSCLC subtypes (Figure

2.6D). *IREB2* has previous associations with both chronic obstructive pulmonary disease (COPD) and lung cancer, but recent work suggests that *IREB2* has a stronger association for lung cancer than COPD (118).

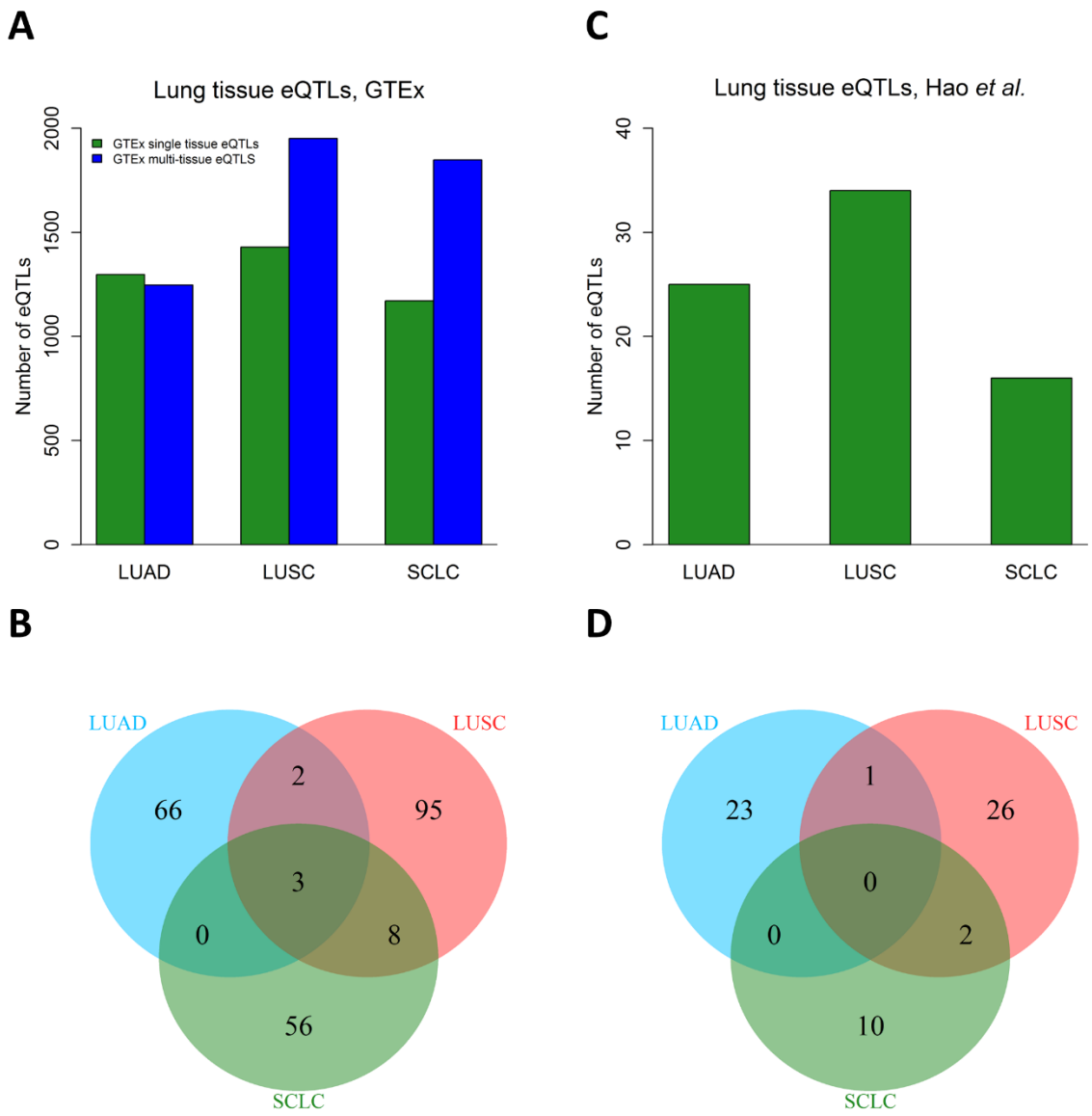


Figure 2.6. Lung tissue eQTLs in three lung cancer subtypes. (A) Total number of significant eQTLs found in each lung cancer subtype using lung tissue specific data (q-value ≤ 5%) and multi-tissue data (posterior probability > 0.8) from GTEx. (B) Venn diagram shows the overlap of genes discovered from the GTEx eQTLs. For each lung cancer subtype, we obtained the final gene set by collapsing all SNPs from (A) into genes. (C) Total number of eQTLs (false discovery rate, FDR < 10%) found in the lung tissue specific dataset from Hao *et al. (110)*. (D) Venn diagram shows the overlap of genes based on Hao *et al*. eQTLs. Duplicate genes were removed from (C) for this comparison.

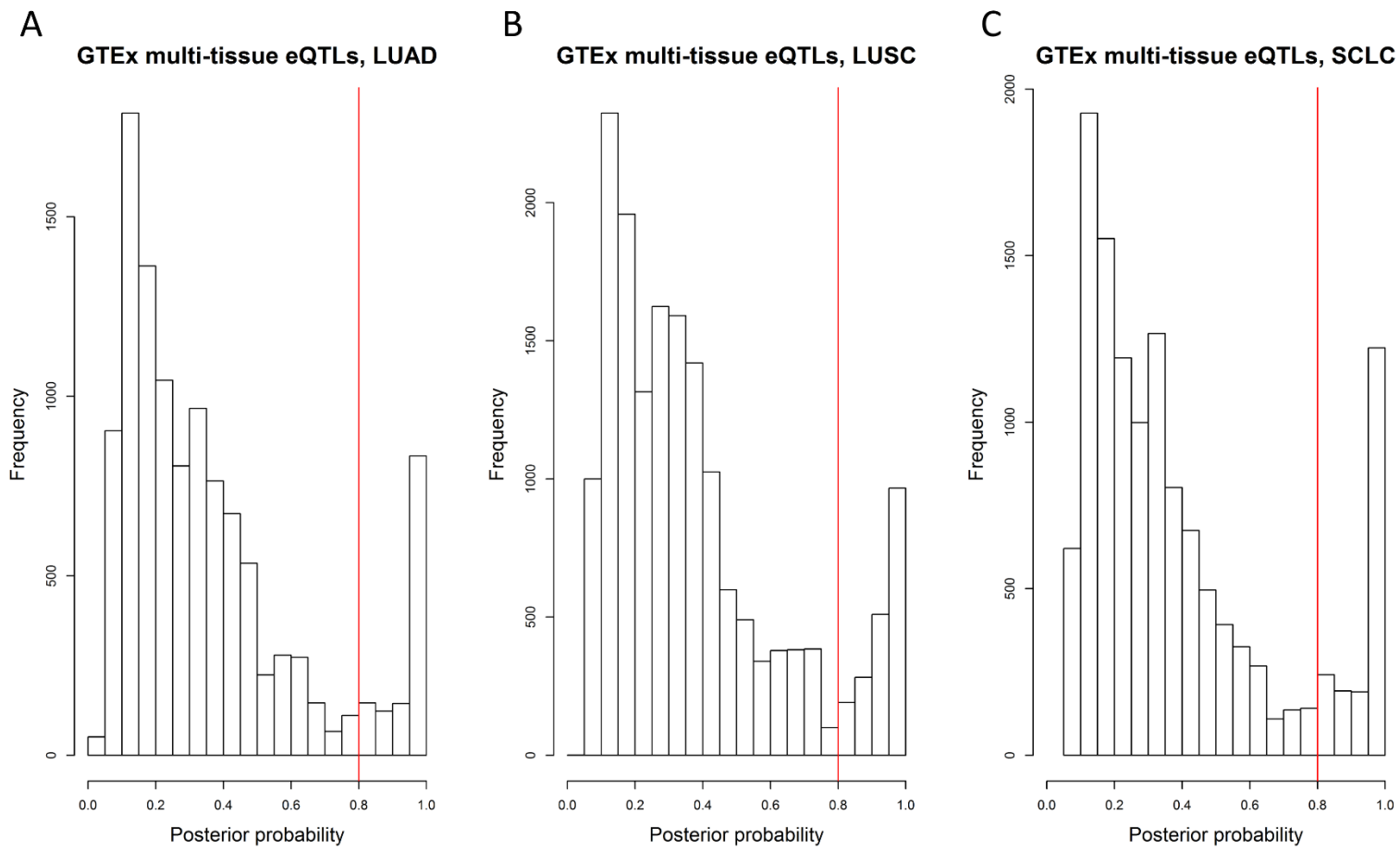


Figure 2.7. Determination of significance for GTEx multi-tissue eQTLs. Posterior probabilities in lung tissue for all multi-tissue eQTLs are plotted for each subtype. The posterior probabilities of the eQTLs for each subtype, LUAD (A), LUSC (B) and SCLC (C), resemble a bimodal distribution. We chose a significance threshold to capture the second distribution of values. Red line indicates the cutoff used of a posterior probability of 0.8.

Finding transcribed enhancers and their target genes

We next examined SNPs located within enhancer regions of the genome that had associated target genes. We used data from The Functional ANnoTation Of the Mammalian genome (FANTOM) (73) collaborative project that identified transcribed enhancer regions of the genome known as "eRNAs" using the Cap Analysis of Gene Expression (CAGE) method (74). We used this permissive set of enhancers and their corresponding transcribed target genes from the Promoter Enhancer Slider Selector Tool (PrESSTo) website (73, 119). We found the number of genes that were targeted by the enhancers were 45 for LUAD, 104 for LUSC, and 43 for SCLC (Figure 2.8A). We removed duplicated genes in each subtype and found no overlap for these enhancer target genes among all three subtypes (Figure 2.8B). We also observed no overlap among LUAD and SCLC or SCLC and LUSC. However, we did find five target genes from two genomic loci that overlapped LUAD and LUSC (*EPB49*, *LGI3*, *LPCAT1*, *NPM2* and *PHYHIP*).

Finding epigenetically defined enhancers and their predicted target genes

To find SNPs located within epigenetically defined enhancers, we used a dataset that defined enhancers using histone modifications such as H3K4me1 (120) and H3K27ac (121). Specifically, we used the results from a newly developed software tool, Integrated Methods for Predicting Enhancer Targets (IM-PET), that uses specific histone marks to identify enhancers and other data types to predict their targets using a sophisticated random forest classifier (72). We found more than 100 enhancer targets in all subtypes across two lung related cell lines (IMR90 and NHLF) (Figure 2.8C). These enhancer targets are reported as mRNA transcripts. Therefore, to perform a comparison similar to the previous datasets, we collapsed all transcripts into single

genes (see Methods). We merged the genes found across both cell lines and removed the duplicated genes within subtypes. There were only two genes from one unique signal that overlapped all subtypes (*ID3HA and TBC1D2B)* (Figure 2.8D). *IDH3A* is an enzyme in the metabolic tricarboxylic acid (TCA) cycle that is frequently altered in cancer cells (122).

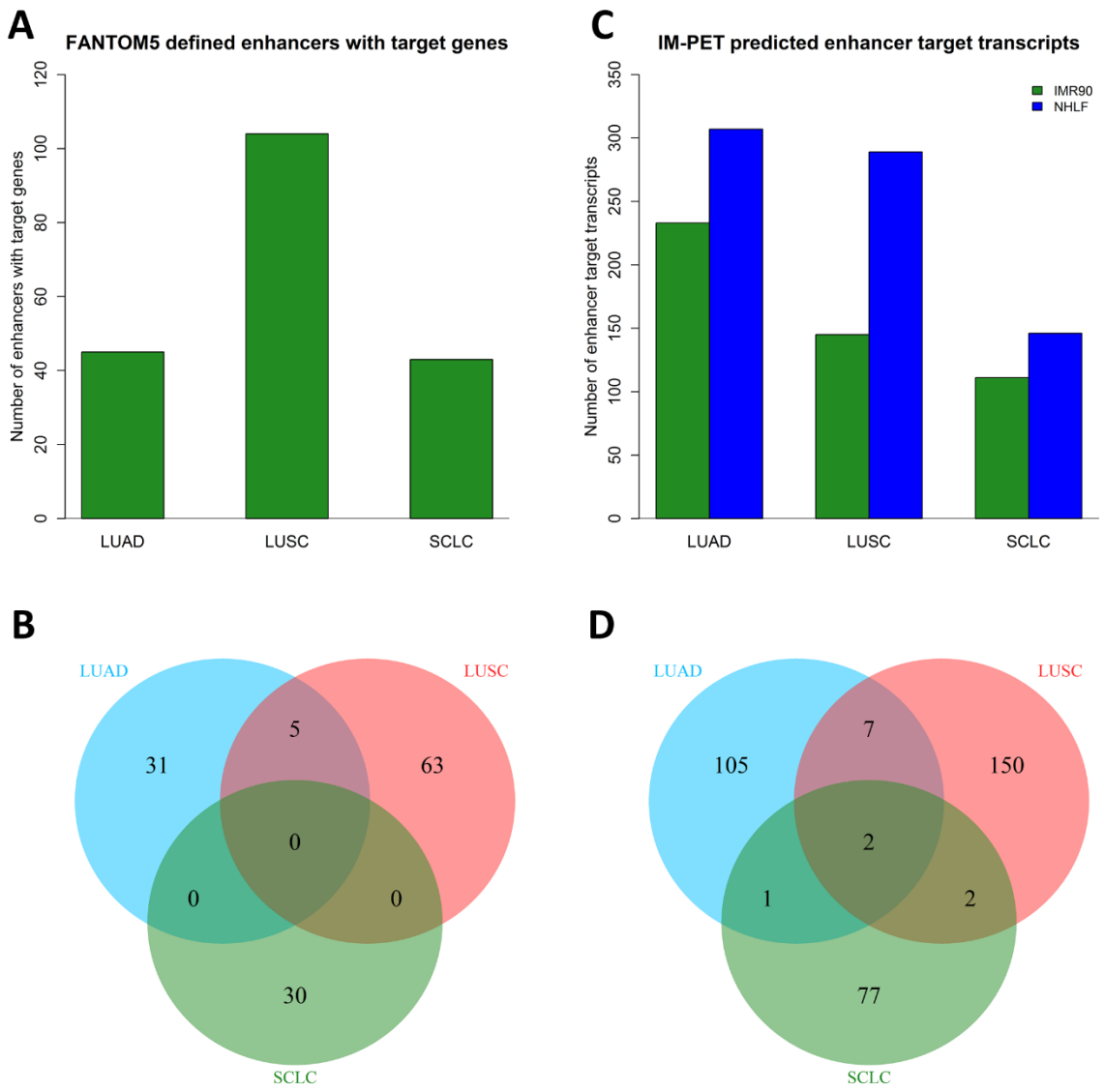


Figure 2.8. Comparison of the SNPs located within the enhancer regions and their target genes among three lung cancer subtypes. (A) Total number of enhancer target genes identified by FANTOM5. (B) Venn diagram shows the overlap of FANTOM5 enhancer target genes by subtype. (C) Total number of enhancer target transcripts identified by IM-PET for two lung-related cell lines. (D) Venn diagram shows the overlap of the lung cancer predicted enhancer target genes for IMR90 and NHLF identified by IM-PET.

Final set of germline-regulated genes and comparison to the original study

We collected all of the genes identified by all of the above methods, removed duplicated genes within subtypes, and refer to this final collection of genes as germline-regulated genes. There were only five genes shared by all of the subtypes: *CHRNA5, IDH3A, PSMA4, RP11-650L12.2,* and *TBC1D2B* (Figure 2.9A). Although we found five unique genes, these genes are all located together in one unique genomic region on 15q25 and probably represent only one unique signal. We also compared the genes found across all of the different methods per subtype. Surprisingly, we observed very little overlap between the different methods across all lung cancer subtypes (Figure 2.10).

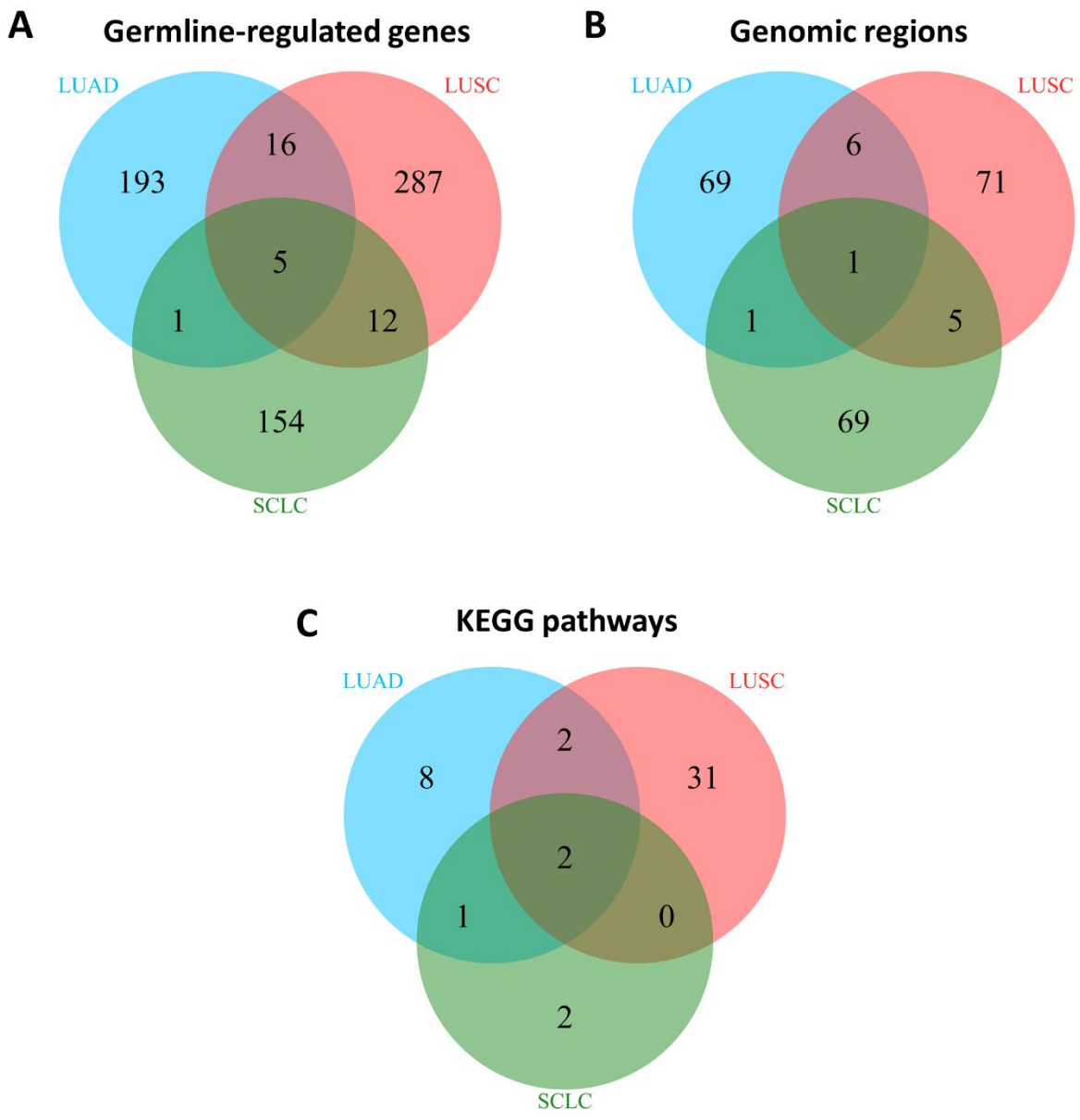


Figure 2.9. Comparison of the germline genes and their enriched biological pathways by subtype. (A) Venn diagram shows the overlap of germline-regulated genes identified in the present study for the three lung cancer subtypes. (B) Venn diagram shows the overlap of the germline-regulated genes from (A) represented as unique genomic loci. (C) Venn diagram shows the overlap of KEGG pathways enriched with the germline-regulated genes.

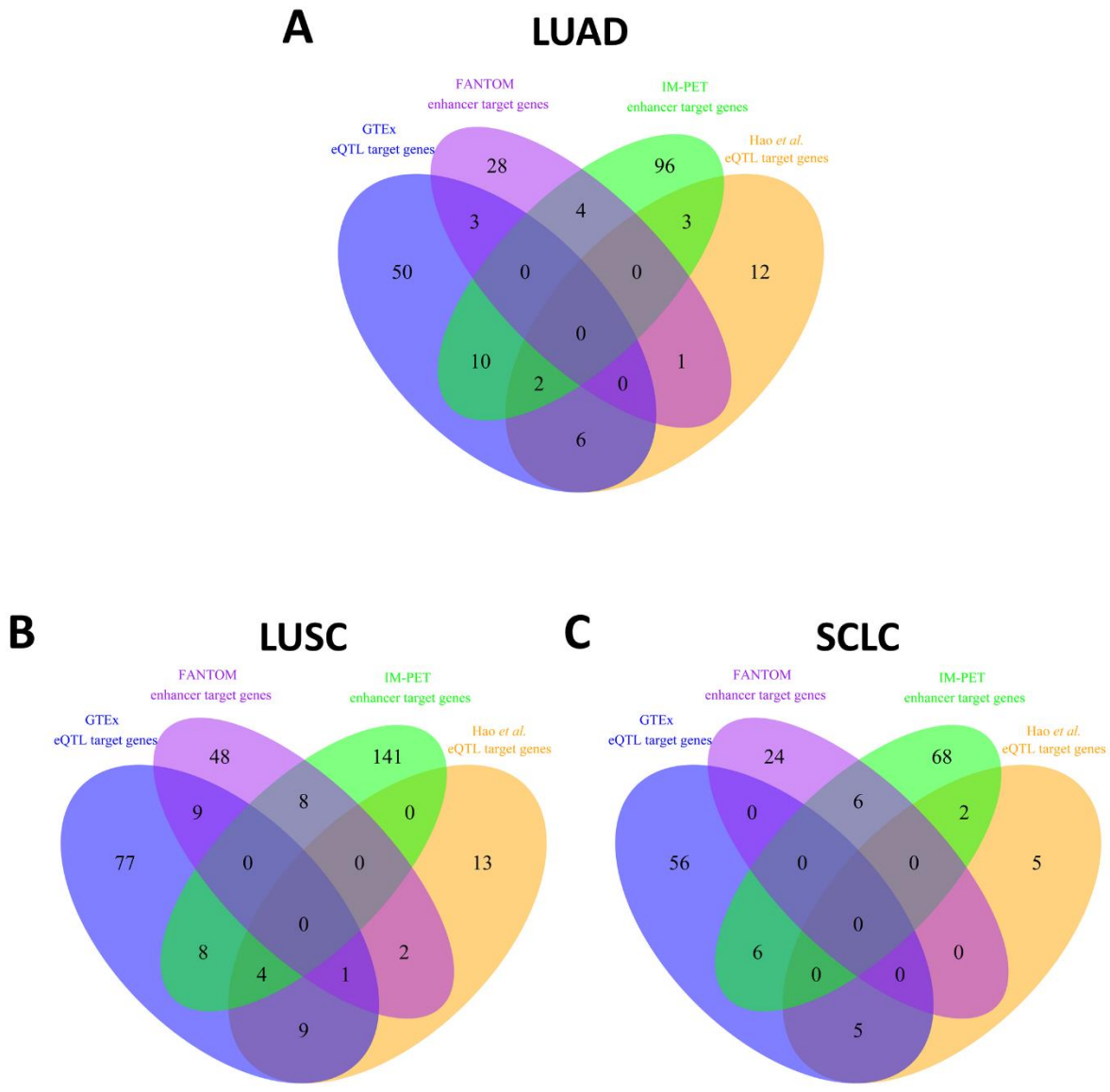


Figure 2.10. Comparison of the final germline-regulated genes discovered in each subtype separated by the different data sources. Venn diagrams show the overlap between genes found from each data source for (A) lung adenocarcinoma (LUAD), (B) lung squamous cell carcinoma (LUSC), and (C) small cell lung cancer (SCLC).

A common approach used to report genes that may be associated with SNPs found from GWA studies is to report genes that are the closest in proximity up-stream or down-stream of the genotyped SNP. Therefore, we next verified that our approach to identifying target genes from GWAS SNPs identified a different set of genes than the genes reported using the "closest gene" approach in the original Landi *et al*. study (96). To perform this comparison, we ran the same pipeline described above, and used the same set of SNPs reported in Landi's original paper's (96) supplemental tables with a defined significance $p < 1 \times 10^{-4}$. We found that only ~25% of the germline-regulated genes that we found using our approach were reported in the original GWAS publication (Figure 2.11A).

We further applied our approach to analyze the data from the GWAS Catalog and obtained a set of SNPs for matched European population type from the GWAS Catalog (52) using the search term "lung cancer" (see Methods). After removing the SNPs from the original study, we identified 17 SNPs to run through our pipeline. We ran the SNPs through the pipeline and identified six germline-regulated genes from the GWAS Catalog SNPs: *CHRNA5, CLPTM1L, PSMA4, RP11-650L12.2, TP63* and *ZSCAN29*. Three of these genes, *CHRNA5*, *PSMA4*, and *RP11-650L12.2* are located in the 15q25 locus, while the other three genes are in three independent genomic locations. We examined the overlap between these genes and our defined germline-regulated genes by lung cancer subtype. There was a strong overlap (67%) between the genes in at least one subtype from our analysis and the target genes associated with lung cancer from the GWAS Catalog (Figure 2.11B).

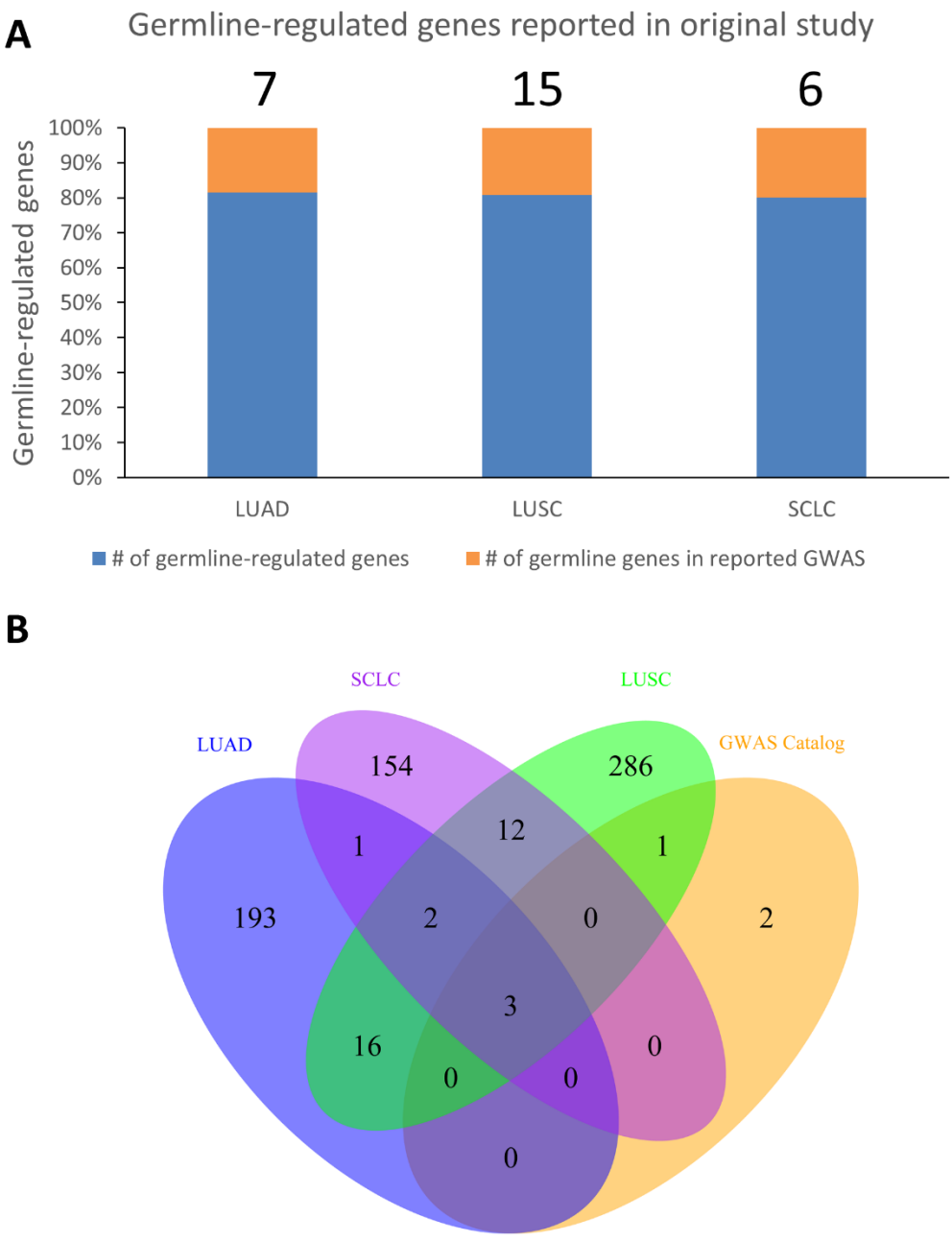


Figure 2.11. Comparison of germline-regulated genes to original study by Landi *et al*. and the GWAS Catalog. In panel A, we show the proportion of germline-regulated genes we discovered that were reported in the original GWAS publication by Landi *et al*. (96). The genes originally reported were discovered at significance level $p < 1 \times 10^{-4}$ and were based upon their physical location to the significant SNPs. The numbers above the bars are the total number of germline-regulated genes found in this study that were originally reported. The majority of germline-regulated genes discovered in this chapter were initially missed in the original report because the authors reported them based only upon physical location. Panel B shows the overlap of germline-regulated genes found using SNPs from the GWAS Catalog with the three lung cancer subtypes.

Pathway enrichment analysis of germline-regulated genes

To gain a deeper understanding of the biology driven by these germline-regulated genes, we performed biological pathway enrichment analysis of the genes in each subtype. We used the web-based tool, WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) (114, 123), to identify significantly enriched pathways with the set of germline-regulated genes for each subtype using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (124). We list the full sets of pathways enriched in each subtype for the KEGG pathways in Tables 2.4-2.6. We found that all three subtypes had genes enriched in the Metabolic pathways and Proteasome pathways (Figure 2.9B). We note that many of the pathways found for LUSC represent only one genomic locus (HLA region, chromosome 6p21) that contains the same sets of genes (Table 2.5).

Table 2.4. KEGG pathway enrichment results of germline-regulated genes for LUAD.

KEGG pathway name	# enriched genes in pathway	Gene symbols	Unique genomic regions	# genes in pathway	Raw p	BH Adjusted p
Metabolic pathways	17	*HIBADH, NT5C2, PIGN, GAPDH, AMT, IDH3A, IMPDH2, MDH2, ACOX1, QARS, POLR3D, UGCG, UGT2B4, LPCAT1, COX4I2, GGPS1, CD38*	15	1130	1.49E-05	0.0006
Tight junction	4	*MYH9, MYH4, CLDN23, LLGL2*	4	132	0.003	0.024
Viral myocarditis	3	*MYH9, MYH4, CAV1*	3	70	0.0039	0.024
Endocytosis	5	*NEDD4L, CAV1, ASAP1, CAV2, CHMP6*	4	201	0.0022	0.024
Bacterial invasion of epithelial cells	3	*CAV1, SHC4, CAV2*	2	70	0.0039	0.024
Insulin signaling pathway	4	*TSC1, SHC4, FOXO1, PRKAR1A*	4	138	0.0035	0.024
Nicotinate and nicotinamide metabolism	2	*NT5C2, CD38*	2	24	0.0052	0.0275
Apoptosis	3	*ENDOD1, BCL2L1, PRKAR1A*	3	87	0.0071	0.0328
Citrate cycle (TCA cycle)	2	*IDH3A, MDH2*	2	30	0.008	0.0329
Focal adhesion	4	*LAMB2, CAV1, SHC4, CAV2*	3	200	0.0127	0.047

Proteasome	2	*PSMA4, PSMD14*	2	44	0.0167	0.0492
Lysosome	3	*GGA3, GGA1, CTSH*	3	121	0.0173	0.0492
Aldosterone-regulated sodium reabsorption	2	*NEDD4L, SFN*	2	42	0.0153	0.0492

Table 2.5. KEGG pathway enrichment results of germline-regulated genes for LUSC.

KEGG pathway name	# enriched genes in pathway	Gene symbols	Unique genomic regions	# genes in pathway	Raw p	BH Adjusted p
Staphylococcus aureus infection	15	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, C4A, HLA-DMB, C5, CFB, HLA-DOA, HLA-DOB, HLA-DPB1, HLA-DRB5*	2	55	1.44E-20	1.15E-18
Asthma	12	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	1	30	4.65E-19	1.86E-17
Type I diabetes mellitus	13	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, HSPD1, HLA-DOB, HLA-DRB5, HLA-DPB1*	2	43	1.17E-18	3.12E-17

Antigen processing and presentation	15	*HLA-DRB1, HLA-DRA, HSPA1L, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, TAP1, TAP2, HLA-DOA, HLA-DOB, HLA-DPB1, HLA-DRB5*	1	76	3.04E-18	6.08E-17
Allograft rejection	12	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	1	37	9.56E-18	1.53E-16
Graft-versus-host disease	12	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	1	41	3.98E-17	5.31E-16
Intestinal immune network for IgA production	12	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1,*	1	48	3.37E-16	3.85E-15

		HLA-DQB1, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1				
Autoimmune thyroid disease	12	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	1	52	9.75E-16	9.75E-15
Viral myocarditis	13	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, EIF4G3, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	2	70	1.29E-15	1.15E-14
Leishmaniasis	13	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, NFKBIA, HLA-DMB, HLA-DOA, HLA-*	2	72	1.91E-15	1.53E-14

		DOB, HLA-DRB5, HLA-DPB1				
Rheumatoid arthritis	13	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, ATP6V1G2, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	1	91	4.61E-14	3.35E-13
Phagosome	15	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, ATP6V1G2, HLA-DMB, TAP1, TAP2, HLA-DOA, HLA-DOB, HLA-DPB1, HLA-DRB5*	1	153	1.50E-13	1.00E-12
Toxoplasmosis	14	*HLA-DRB1, HLA-DRA, HSPA1L, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, NFKBIA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRB5, HLA-DPB1*	2	132	3.30E-13	2.03E-12

Systemic lupus erythematosus	14	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, C4A, HLA-DMB, C5, HLA-DOA, HLA-DOB, HLA-DPB1, HLA-DRB5*	2	136	5.00E-13	2.86E-12
Cell adhesion molecules (CAMs)	13	*HLA-DRB1, HLA-DRA, HLA-DMA, HLA-DPA1, HLA-DQA2, HLA-DQA1, HLA-DQB1, HLA-DMB, HLA-DOA, ALCAM, HLA-DOB, HLA-DRB5, HLA-DPB1*	2	133	6.70E-12	3.57E-11
Spliceosome	6	*HSPA1L, PPIL1, RBM25, ZMAT2, SF3A1, DDX39B*	5	127	0.0002	0.0009
Metabolic pathways	19	*PPT2, EARS2, ALDH6A1, GNPDA1, IDH3A, UQCR10, PMM2, GAL3ST1, LCLAT1, PON1, SDHA, NT5C2, PON2, POLR3D, NDUFA2, PON3,*	14	1130	0.0002	0.0009

		ATP6V1G2, LPCAT1, CYP17A1				
Wnt signaling pathway	6	*TCF7, CSNK2B, NFATC3, PSEN1, BTRC, DAAM2*	6	150	0.0005	0.0022
Aminoacyl-tRNA biosynthesis	4	*HARS2, VARS2, EARS2, HARS*	3	63	0.0008	0.0034
Huntington's disease	6	*GNAQ, DNAL1, NDUFA2, SDHA, UQCR10, TBPL1*	6	183	0.0015	0.006
Proteasome	3	*PSMB8, PSMB9, PSMA4*	2	44	0.0032	0.0122
Jak-STAT signaling pathway	5	*LIF, SPRED1, PIM1, CBLB, OSM*	4	155	0.0039	0.0142
Alzheimer's disease	5	*GNAQ, NDUFA2, PSEN1, SDHA, UQCR10*	5	167	0.0054	0.0188
Pathways in cancer	7	*TCF7, FGF17, SUFU, NFKB2, NFKBIA, E2F3, CBLB*	6	326	0.0065	0.0217
NOD-like receptor signaling pathway	3	*CARD8, NFKBIA, NLRC4*	3	58	0.0069	0.0221
Shigellosis	3	*DIAPH1, BTRC, NFKBIA*	3	61	0.0079	0.0243
Biosynthesis of unsaturated fatty acids	2	*ACOT2, ACOT1*	1	21	0.0085	0.0252
RNA degradation	3	*SKIV2L, HSPD1, PATL1*	3	71	0.0119	0.0317
Complement and coagulation cascades	3	*C5, CFB, C4A*	2	69	0.0111	0.0317

Oxidative phosphorylation	4	*NDUFA2, ATP6V1G2, SDHA, UQCR10*	4	132	0.0119	0.0317
Chronic myeloid leukemia	3	*NFKBIA, E2F3, CBLB*	3	73	0.0129	0.0333
Collecting duct acid secretion	2	*SLC12A7, ATP6V1G2*	2	27	0.0139	0.0347
Citrate cycle (TCA cycle)	2	*IDH3A, SDHA*	2	30	0.017	0.0412
Hematopoietic cell lineage	3	*HLA-DRB1, HLA-DRA, HLA-DRB5*	1	88	0.0211	0.0496
Prostate cancer	3	*TCF7, NFKBIA, E2F3*	3	89	0.0218	0.0498

Table 2.6. KEGG pathway enrichment results of germline-regulated genes for SCLC.

KEGG pathway name	# enriched genes in pathway	Gene symbols	Unique genomic regions	# genes in pathway	Raw p	BH adjusted p
Metabolic pathways	13	*MGAT3, EPRS, REV3L, DHRS4L2, PLA2G7, DPM1, GNPDA1, IDH3A, AMD1, OXSM, BPNT1, ATP6V1E1, DHRS4*	10	1130	0.0002	0.003
Retinol metabolism	3	*DHRS4L2, DHRS4, CYP26A1*	2	64	0.0015	0.0112
Focal adhesion	4	*ACTN4, TNN, ILK, PAK4*	3	200	0.0055	0.0275
Proteasome	2	*PSMD8, PSMA4*	2	44	0.0105	0.0387
N-Glycan biosynthesis	2	*MGAT3, DPM1*	2	49	0.0129	0.0387

Discussion

Understanding the regulatory roles that common genetic variants play in the development of many disease types, including lung cancer, is an important research question because the majority of common variants that increase risk for a diverse set of diseases are located within non-coding regions and most likely act as regulators of gene expression. These results can also be used to interrogate the differences between different subtypes of cancer. To address these questions, we performed a detailed analysis of common genetic variants (SNPs) associated with three subtypes of lung cancer (LUAD, LUSC, and SCLC).

We used marginally significant GWAS results ($p < 1 \times 10^{-3}$) to search for regulatory roles for common variants associated with LUAD, LUSC, and SCLC. We expanded this set of results to include all SNPs in LD with the genotyped SNPs using data from the 1000 Genomes Phase III project. This expansion resulted in ~15,000 more SNPs to test per subtype that may be acting as the actual causal variant (63). We used a diverse set of regulatory data to identify SNPs that were within regulatory regions of the genome that had an identified target gene. It is important to use data that contain the target genes for regulatory SNPs, because most regulatory looping interactions influence distant genes rather than the closest gene(s) (125). We first examined lung tissue eQTL data from the GTEx project and the Hao *et al*. study. Interestingly, our results indicated there was little overlap in the eQTL genes identified from these separate datasets. This is not surprising because they were analyzed using different methods and with different sample sizes, but a more thorough examination into the details of this small overlap would be interesting. Although the methods differ, they are from the same tissue and a detailed analysis of the

differences may reveal further insight into the heterogeneity of whole tissue eQTL analyses. This issue is rarely discussed in the literature, but is important for phenotypes where the specific type of cell within the mix of cells in the entire tissue directly affects disease. Through the search of the FANTOM5 set of permissive enhancers and their target genes, we observed that several SNPs from each subtype were within enhancer regions of the genome. This finding may be important because recently one SNP (rs6983267) within a gene desert at 8q24, which harbors many SNPs associated with several cancers, was found to disrupt an enhancer region that controls expression of the oncogene MYC, an important gene involved in many cellular growth pathways. Also, MYC is tightly regulated due to its essential role in cell proliferation where it is implicated in the genesis of many cancer types (126). This regulatory mechanism highlights the importance of enhancers in the maintenance of cell division and growth (75). For our final data source, we used IM-PET, a machine classifier method that has high predictive power to detect the target genes of enhancers (72). Specifically, we used the results from two lung related cell lines, NHLF and IMR90. We found the largest number of regulatory target genes for all three subtypes using the combined results from both cell lines for IM-PET (LUAD = 115, LUSC = 161 and SCLC = 82). Overall, our results indicated very small overlap between all three subtypes at the SNP, gene, pathway, and regulatory levels. Of note, we found similar lack of overlap between all subtypes when we used SNPs with $p < 1 \times 10^{-4}$ (Appendix B). Importantly, the weak overlap we observed at the gene level between all subtypes was from five separate genes, but was only representative of one genomic region. Therefore, there is likely only one independent region on 15q25 that overlaps all three subtypes of lung cancer and is likely driven by the lead peak in the GWAS Manhattan plot.

It is worth highlighting that three (*CHRNA5, IDH3A,* and *PSMA4)* out these five genes shared in all three subtypes of lung cancer have been previously reported to be associated with

lung cancer. *CHRNA5* has strong implications in its association with lung cancer (53, 54, 116). *CHRNA5* encodes a nicotinic acetylcholine receptor (nAChR). nAChRs are a class of ligand-gated ion channels that are activated by the neurotransmitter acetylcholine to allow the flow of ions across the cell membrane (127). There is still an ongoing debate about *CHRNA5*'s role in lung cancer risk versus its risk for lung cancer through nicotine addiction (128), but finding this gene in all three subtypes of lung cancer that have biological and environmental differences suggests it may be playing a direct role in lung cancer risk. *IDH3A* encodes an isocitrate dehydrogenase (IDH). IDHs are important enzymes in the regulation of the TCA cycle (129). Recently, *IDH3A* was shown to promote tumor growth by activating hypoxia-inducible factor 1 (HIF-1) alpha and promoting the stability of HIF-1 to participate in angiogenesis and was also associated with poor survival in lung cancer (130). Additionally, *IDH3A* acts in the conversion of metabolism that occurs with cancer fibroblasts (131). *PSMA4* encodes a subunit of the proteasome. Experimental studies have shown that *PSMA4* mRNA is increased in lung tumor versus normal samples and also plays a major role in cell proliferation using data from lung carcinoma cell lines (132). Another gene, *RP11-650L12.2*, has not been characterized. It warrants future experimental studies due to its association with all three subtypes of lung cancer. The final gene shared by all subtypes, *TBC1D2B*, is a protein coding gene that may have GTPase activity and may play a role in autophagy (133).

In addition to the five overlapping genes above, our pathway enrichment analysis revealed two biological pathways shared by the three subtypes. Among them, all three subtypes shared Metabolic pathways. Metabolic pathways are frequently modified in cancer to provide the over proliferating cells with required nutrients (134). We also observed that the oxidative phosphorylation pathway was significantly enriched in LUSC (adj. $p < 0.05$). It is interesting to

find this pathway dysregulated in the germline genome, because it has strong associations in the transition from oxidative phosphorylation to the less efficient aerobic glycolysis, known as the Warburg effect, that takes place in cancer cell proliferation (135). Although the Warburg effect may be attributable to glycolysis inhibiting a still active oxidative phosphorylation pathway, this result still suggests that commonly occurring variants in LUSC may lead to some disruption in the oxidative phosphorylation pathway that makes this process easier to arrest or inhibit and enhance cell proliferation after some somatic disruption in lung tissue. We also found several cancer related pathways in LUSC such as Pathways in cancer, Prostate cancer and many signaling pathways associated with cancer. We discovered that the focal adhesion (adj. $p < 0.01$) pathway was significantly enriched with genes from SCLC. This is an intriguing finding because genes in this pathway are involved in the epithelial-mesenchymal transition (EMT), which is important in cancer metastasis (136). Although this pathway is also found in LUAD (adj. $p = 0.047$), it is more significant in SCLC (adj. $p = 0.028$) and may help explain the much higher rate of metastases seen in SCLC compared to NSCLC (137). In summary, this pathway-based evidence suggests both shared subtype and unique subtype associations.

To ensure that our approach to identify target genes from GWAS SNPs was not just identifying genes found in the original study, or reporting the closest gene to each SNP, we ran the SNPs found in Landi's original paper's supplemental tables through our pipeline. These SNPs were reported at a more stringent threshold ($p < 1 \times 10^{-4}$) than we used in our analysis ($p < 1 \times 10^{-3}$). We compared the germline-regulated genes identified using these SNPs to the genes reported in the original study. We found little overlap, ~25%, suggesting the reported genes (closest to the SNPs) may not be the correct target genes. We also looked at this overlap using the set of germline-regulated genes that we discovered with our pipeline using SNPs at $p < 1 \times 10^{-4}$ and observed

overlap of ~15%. These findings agree with the studies that have reported that the gene closest to a non-coding variant is oftentimes not the target gene (125, 138). One well-known example is the *FTO* locus and obesity. A recent study (64) discovered that a SNP within an intron in *FTO* was not controlling the expression levels of the gene *FTO,* but instead the more distant gene *IRX3*. We observed similar findings when we investigated lung cancer genes reported from the GWAS Catalog. We compared the reported genes for lung cancer from the GWAS Catalog to our germline-regulated genes and found only *CHRNA5* and *PSMA4* overlapped all subtypes. In total, out of ~50 reported genes in the Catalog, we observed overlap of two, six, and three genes for LUAD, LUSC, and SCLC, respectively. To determine if this weak overlap occurred because the Catalog links SNPs to the closest gene, we ran the SNPs from the GWAS Catalog through our pipeline. The results showed a strong overlap (67%) between the germline-regulated genes obtained using the GWAS Catalog SNPs and our set of germline-regulated genes in any subtype. This finding suggests that the target gene of a non-coding regulatory SNP may not be the closest gene.

In addition to the analysis reported in this chapter for SNPs ($p < 1 \times 10^{-3}$), we also performed the same analysis using a more stringent $p < 1 \times 10^{-4}$. Our results from that analysis agreed with our discoveries in this chapter that indicated very small overlap between the lung cancer subtypes. At $p < 1 \times 10^{-4}$, we only observed three genes (*CHRNA5, IDH3A, and RP11-650L12.2*) that overlapped between all subtypes in comparison to five genes (*CHRNA5, IDH3A, PSMA4, RP11-650L12.2,* and *TBC1D2B)* that overlapped at $p < 1 \times 10^{-3}$. Both sets of genes are located within the same single genomic locus.

There are several limitations to this study. First, we utilized a set of marginally significant SNPs. Although previous studies (139, 140) have shown it is a practical approach, this may have

resulted in some false positive SNPs in our study. Second, we did not impute the GWAS data to obtain a larger set of SNPs for the analysis. This would have resulted in more SNPs that could have been tested for significance. We will integrate such SNPs in future analyses. Third, for validation of our results, we were limited to a small set of SNPs reported in the GWAS Catalog because we only focused on SNPs specifically found in one population. Though we saw a strong overlap (67%), it would have been better to include a larger set of SNPs for better power of confirming the validity of our pipeline. Whether similar patterns are in other populations remains for further investigation when such data becomes available. Another limitation of our study is that we may have discovered several different genes that may represent only one unique signal because we used SNPs in LD for our analysis. For example, if we found five genes that were shared by all subtypes, but these genes were clustered in one genomic location, it may represent a single unique signal. To account for this potential bias, we separated the gene sets into unique signals to give a better idea of the true overlap of subtypes while still including all discovered germline-regulated genes.

In summary, we used common genetic variants found in three lung cancer subtypes to interrogate the similarity between them at four biological levels. We found that there is very little overlap between the three subtypes at the SNP, gene, regulatory and pathway levels. At the most basic level (SNPs), we observed less than a 1% overlap between the subtypes. Similarly, we found only five genes (< 1%) (all five from one unique genomic locus) overlap that were discovered in all three subtypes, but three of the five (*CHRNA5, IDH3A,* and *PSMA4*) are well-known lung cancer genes. We observed the same trend at the pathway level and found only two KEGG pathways (~4%) overlapped all three subtypes. At the regulatory level, we found many differences in how the genetic variants in non-coding regions control their target genes. Not much work has

been done comparing all three subtypes at the somatic level, but recent work interrogating the differences between LUAD and LUSC concluded similar findings of little overlap between these two subtypes at the molecular level in somatic lung tumor tissue (92). Overall, this study provides some important insight into the genetic architecture of three subtypes of lung cancer.

CHAPTER III

EXPLORATION OF SOMATIC MUTATION AND GENE EXPRESSION FEATURES IN THREE LUNG CANCER SUBTYPES

Introduction

The interrogation of genetic variants detected in the germline genome of lung cancer subjects through approaches such as GWA studies may reveal risks associated with this disease. However, many agents used in lung cancer therapies target somatic alterations in tumor tissue. For example, treatments such as erlotinib (141) and gefitinib (142) target actionable somatic mutations in EGFR in NSCLC. These molecularly targeted therapies have better success rates than the traditional non-targeted platinum-based chemotherapy (143, 144). Additionally, many new driver genes have been discovered in somatic lung tumor tissue that are effectively targeted by pharmaceutical compounds (35). These studies support current efforts to identify genes at the somatic level that are critical for lung cancer treatment.

The TCGA is one of the largest collaborative consortia with its main goal to uncover the landscapes of the genetic alterations at the genome level in major tumor types. TCGA has characterized genomic alterations for both LUAD (82) and LUSC (86). In addition to TCGA, several other groups have studied mutations in LUAD (76, 80) and LUSC (77, 84). SCLC was not studied as part of TCGA, likely due to its unavailability of tumor samples. However, two recent

studies (88, 89) investigated this tumor type. Collectively, these two SCLC studies found several new driver genes and genomic alterations in SCLC. Importantly, most of the aforementioned studies included the investigation of genetic aberrations at both DNA and RNA levels. Studying gene expression at the RNA level is important because it can identify over-expressed or under-expressed genes in tissues that may not be mutated but can still cause disease (70, 145). Combined, both data types may reveal a more complete genomic picture of the tumor in each lung cancer subtype.

Although recent somatic work (92) has compared the two most common NSCLC subtypes, LUAD and LUSC, much less is known how these subtypes compare to SCLC. Investigating all three subtypes together at the DNA and RNA somatic levels is important because it may lead to the discoveries for SCLC treatment options that currently work with NSCLC. This is vital because there are currently few treatments for SCLC, which is the most aggressive lung cancer subtype (137).

In this study, we used somatic data from lung tumor tissue to compare three lung cancer subtypes at the DNA and RNA levels. First, we identified a set of differentially expressed genes (DEGs) in each subtype and used these gene sets to perform a pathway enrichment analysis. We next examined the proportion of the substitution types, transitions and transversions, among the somatic non-synonymous mutations. We also generated mutational signatures for each subtype. We further used the somatic mutations to identify a set of potential driver genes for each subtype. Overall, this work suggests both shared and distinct genes that are altered among three subtypes of lung cancer at the somatic level. Deeper interrogation of these genes may give greater insight into the biology and potential therapeutic targets of each subtype.

Methods

Summary of somatic mutations

We downloaded the full set of somatic mutations called at the Broad Institute for LUAD and LUSC from tumor samples with matched normal controls from the TCGA Web Portal (updated link https://gdc-portal.nci.nih.gov/) on September 23, 2015. Rudin *et al.* (89) performed a study that identified genomic features in SCLC. They performed WES, RNA-Seq, and copy number analysis (CNA) for multiple SCLC samples. We extracted the full set of somatic mutations identified in this study from the publication's Supplemental Table 3.

Extracting mutational information

We used the R package "maftools" (146) version 1.0.40 to extract mutational information. We used the "titv" function to generate a list of transitions and transversions in each subtype. We generated a matrix of each single nucleotide variant (SNV) and its preceding and proceeding base using human genome reference 19 (hg19). We downloaded the hg19 file in 2bit format from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit on January 31, 2017. We used the TwoBitToFa script, downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/, to convert the 2bit file to a FASTA file. We used the FASTA file for the location of the SNVs and then used non-negative matrix factorization (NMF) (147) to generate the mutational signatures in each subtype. We used up to 6 signatures to search for the top 3 signatures for each subtype separately, based upon the parameters set up in the R package. This tool then used cosine similarity (148) to determine the closest

matching signature from Catalogue of Somatic Mutations in Cancer (COSMIC) to the top three mutational signatures generated from the NMF analysis.

Generating the final set of somatic mutated genes

We used several filtering steps to define a list of possible driver genes to compare the lung cancer subtypes (Figure 3.1). We first filtered our lists to exclude any genes that are not part of the Cancer Genome Census (CGC) list from COSMIC since these have evidence of driving tumor growth (149). We acknowledge that this approach may miss novel driver genes, but since lung cancer is overwhelmed by passenger mutations (150), this approach helps to identify the actual driver genes. The current build we used was downloaded on October 10, 2016, and contained 602 cancer genes. Our primary filter only kept mutations from each cancer type that were within any of the 602 cancer genes. We then removed duplicated somatic mutations per gene per sample. For example, if one subject had 20 TP53 mutations, we only kept one of the mutations so that we could filter by total mutations in more than one subject. We note this may be a limitation, but for our purpose we only needed information if a gene had at least one mutation in a subject. We further filtered by genes mutated in more than one subject. We made two lists at different thresholds defined below that we refer to as "strict" and "lenient." To find a suitable cutoff for these thresholds, we plotted the histograms (Figure 3.2) of the somatic mutations filtered by COSMIC per subtype and manually defined the thresholds as follows. The histograms showed one large peak to the left of the plot, followed by a trailing set of genes mutated in many samples. For our lenient threshold, we chose a cutoff that allowed most of the large bar to be represented, while for our strict threshold, we removed all genes from the primary bar on the histogram that represented the genes that were mutated in only a small number of samples. However, for SCLC, we relaxed

this threshold due to the much smaller set of somatic mutations so we could have a large enough set to use for our analysis. After obtaining a set of possible somatic driver genes for each subtype, we further filtered the list of genes by their expression levels. Rudin *et al*. (89) previously filtered their somatic mutation gene list using expression levels from RNA-Seq, so we only filtered the somatic genes from TCGA for LUAD and LUSC, which had not been previously filtered. We removed genes that had read counts < 10 per gene in the RNA-Seq expression data.

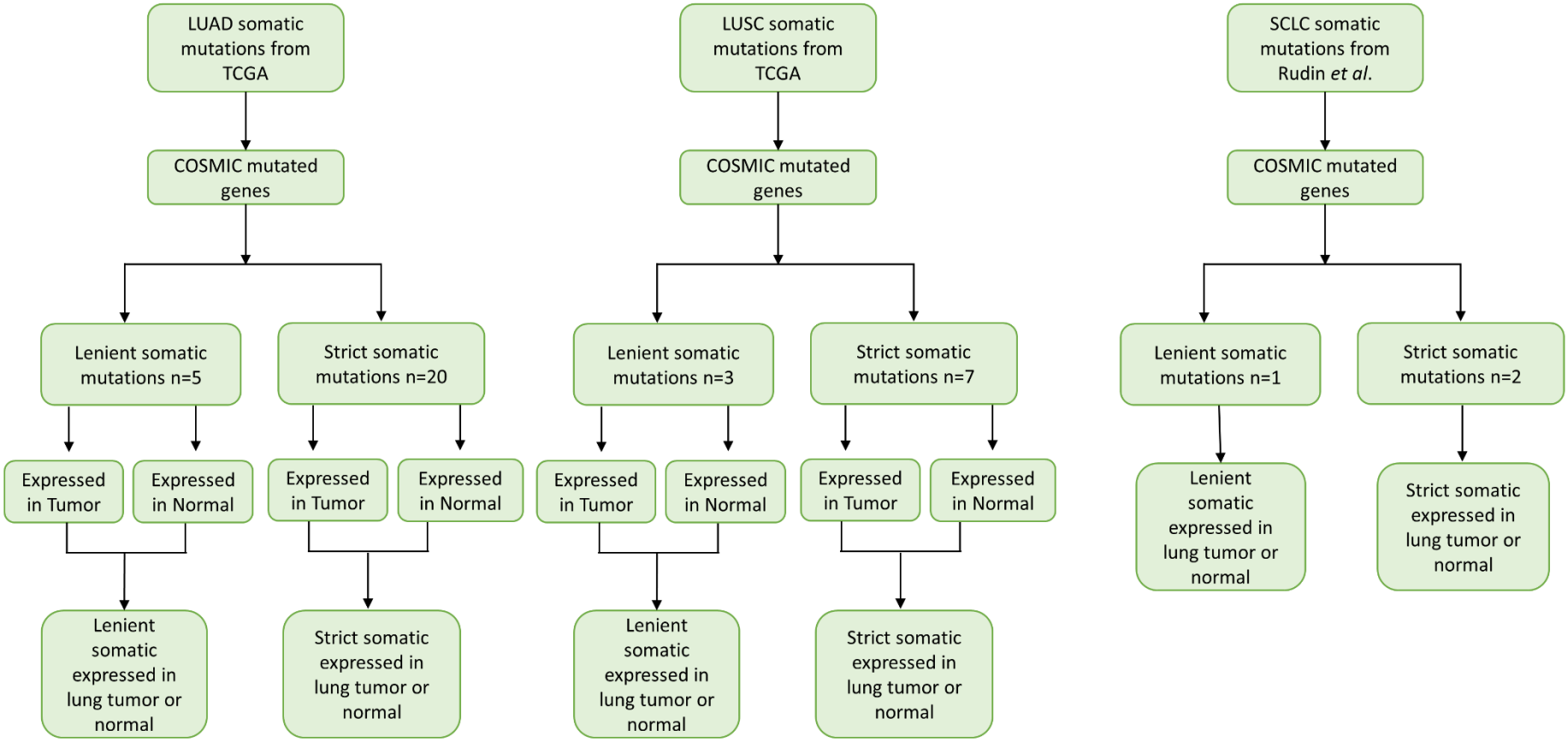


Figure 3.1. Pipeline to obtain somatic mutated genes for three lung cancer subtypes. SCLC somatic mutations were previously filtered by expression levels (89).
Expressed in tumor: ≥ 10 reads
Expressed in normal: ≥ 10 reads
LUAD = lung adenocarcinoma, LUSC = lung squamous cell carcinoma, SCLC = small cell lung cancer

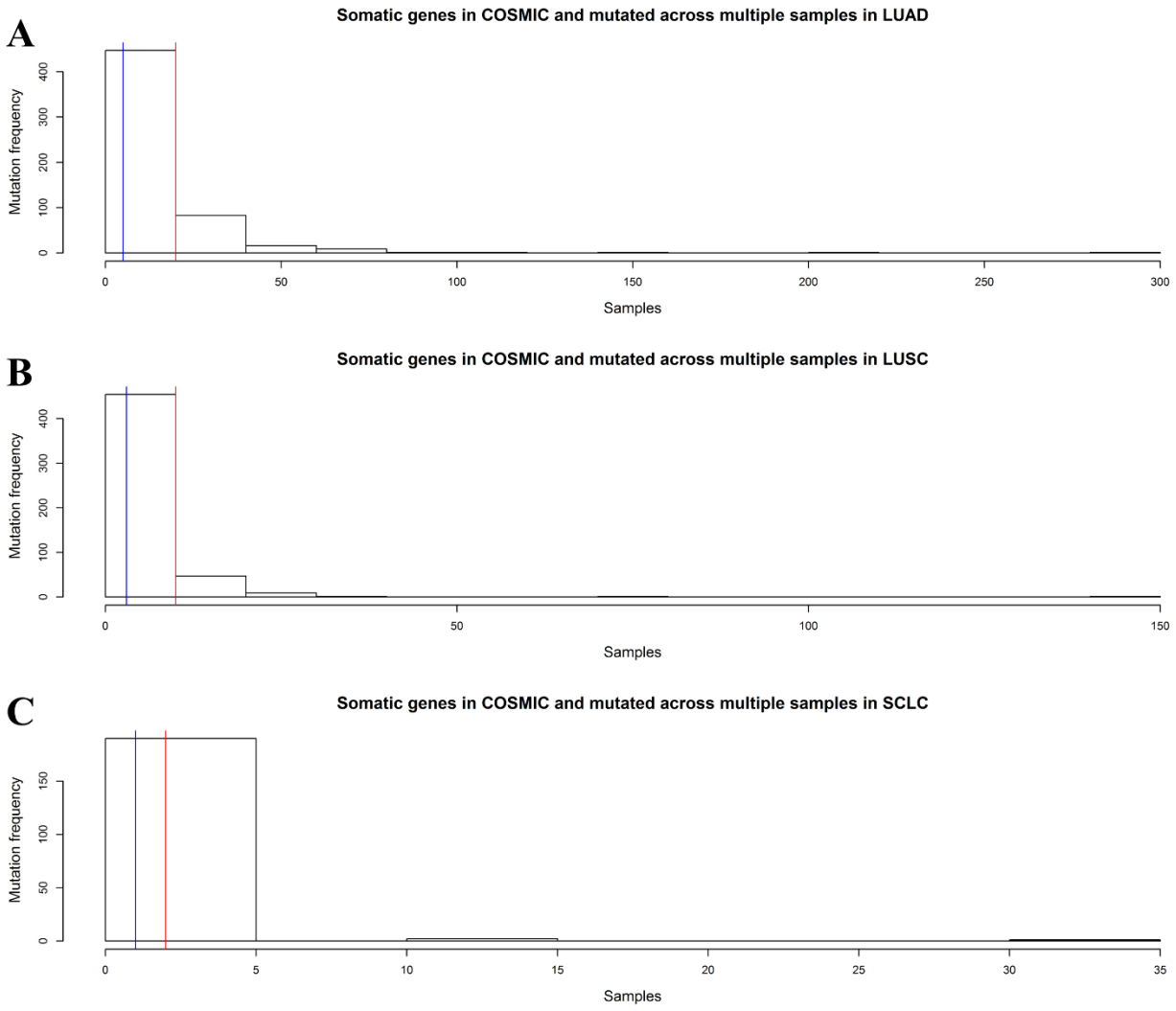


Figure 3.2. Histogram of somatic mutated genes across multiple samples. We show the distribution of somatic mutated genes across multiple samples for LUAD (A), LUSC (B), and SCLC (C). The blue line represents our cutoff for lenient, and the red line represents our cutoff for strict.

Extracting mRNA-Seq raw count values for LUAD and LUSC

Rahman *et al*. (151) reprocessed all of the RNA Sequencing data that was available in TCGA. Briefly, the authors obtained the raw FASTQ formatted RNA-Seq files from the NCI's Cancer Genomics Hub. They also obtained all of the clinical records for the RNA-Seq files. The data were run through their pipeline that relied heavily on the Rsubread package (152). In addition to raw read counts, the authors also normalized the results using two standards in the field: transcripts per million (TPM) (153) and fragments per kilobase of exon per million reads mapped (FPKM) (154). The authors made all of the results freely available on the GEO website: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62944. To extract the raw counts files for LUAD and LUSC, we downloaded the file GSE62944_RAW.tar that contained several data matrices. We unzipped the file and used two data frames for the extraction. For tumor samples, we used: GSM1536837_06_01_15_TCGA_24.tumor_Rsubread_FeatureCounts.txt; and for normal samples, we used: GSM1697009_06_01_15_TCGA_24.normal_Rsubread_FeatureCounts.txt. We also downloaded other files for annotation from the website: GSE62944_06_01_15_TCGA_24_CancerType_Samples.txt.gz for the tumor samples and GSE62944_06_01_15_TCGA_24_Normal_CancerType_Samples.txt.gz for the normal samples. We observed that many sample IDs were duplicated in the tumor sample files. We found this resulted from some RNA-Seq data being generated for more than one part of a tumor tissue for each sample. To keep one sample ID for our analysis, we removed duplicated RNA-Seq results from the same patient by keeping the first mention of a tumor sample and removing additional results from the same tumor sample but having a secondary portion or RNA-Seq run. We matched the LUAD tumor and normal samples and the LUSC tumor and normal samples with these sample names and imported them into DESeq2 (155).

Differential expression analysis using DESeq2

All tumor and normal RNA-Seq samples above were used for differential expression analysis. We imported the raw read counts for all three subtypes into R version 3.2.3 and used the R package DESeq2 version 1.10.1 (155) for the analysis. DESeq2 uses raw non-normalized read count data for its estimation of differential expression. We first filtered our RNA-Seq data to remove low expressed genes (counts ≤ 1). The data were normalized following a negative binomial distribution. We then determined differential expression using a generalized linear model (GLM).

Identification of TSGs and oncogenes

Zhao *et al*. (4) catalogued a set of human tumor suppressor genes (TSGs) through a comprehensive literature review. We downloaded the set of human TSGs from https://bioinfo.uth.edu/TSGene/download.cgi on January 31, 2017.

Liu *et al*. (5) completed an exhaustive literature review to obtain a high-confidence set of oncogenes. We downloaded the entire set of human oncogenes from http://ongene.bioinfo-minzhao.org/download.html on January 31, 2017.

Pathway enrichment analysis

We used the online WebGestalt resource (114) to identify pathways enriched with the DEGs. We separated our analyses into up-regulated and down-regulated gene sets. We used the KEGG pathways (124) for the analysis and filtered out pathways that contained less than five genes.

Results

RNA-Seq data used for DEG analysis

We collected RNA-Seq raw read count tumor and normal samples for LUAD and LUSC from reprocessed TCGA data (151). TCGA did not study SCLC, so we obtained tumor and normal RNA-Seq raw read counts for SCLC from a previous study (89). We removed duplicated samples per subtype to generate our final set of samples. The total number of tumor samples we studied included 515, 501, and 54 for LUAD, LUSC, and SCLC, respectively. We had 59, 51, and 25 matched normal samples for LUAD, LUSC, and SCLC, respectively. The numbers of samples by data source are summarized in Table 3.1.

Table 3.1. Summary of RNA-Seq data.

Subtype	# tumor samples	# normal samples	Data source
LUAD	515	59	TCGA
LUSC	501	51	TCGA
SCLC	54	25	Rudin *et al.* (89)
Total	**1,070**	**135**	

Differentially expressed genes for three lung cancer subtypes

We used DESeq2 to generate DEGs for all three subtypes (Figure 3.3). To define significant DEGs, we used $|\log_2 FC| > 2$ and Benjamini-Hochberg (BH) adj. $p < 0.05$. By first using the adjusted $p < 0.05$, we found a total of 3,710, 5,623, and 3,888 DEGs for LUAD, LUSC, and SCLC, respectively. We further filtered the BH adjusted DEGs by $\log_2 FC > 2$ and detected 1,818,

2,377, and 1,470 up-regulated DEGs for LUAD, LUSC, and SCLC, respectively (Figure 3.4A). We also generated a set of down-regulated DEGs by requiring $\log_2FC < -2$. This revealed 604, 1,328, and 1,798 DEGs for LUAD, LUSC, and SCLC, respectively (Figure 3.4B). The SCLC results were originally reported using Ensembl gene ID's. In our analysis, we removed 45 Ensembl Gene IDs that did not map to Gene Symbols. Accordingly, we found 554 up-regulated genes and 325 down-regulated genes that were shared by all three subtypes (Figure 3.4C, Figure 3.4D). Our results indicated that the number of up-regulated DEGs shared between LUAD and LUSC was significantly higher than that between LUSC and SCLC ($p = 2.2 \times 10^{-16}$, binomial test). Surprisingly, we observed the opposite trend in overlap for the down-regulated genes. We found that more genes were shared between SCLC and LUSC than between LUAD and LUSC ($p = 0.01$, binomial test). To verify that this lack of overlap between all subtypes was not due to a single threshold of DEGs, we generated overlap for three separate expression thresholds ($\log_2FC < -1$, $\log_2FC < -2$, $\log_2FC < -3$, $\log_2FC > 1$, $\log_2FC > 2$, and $\log_2FC > 3$) and observed the same trend in the overlap between subtypes (Figure 3.5).

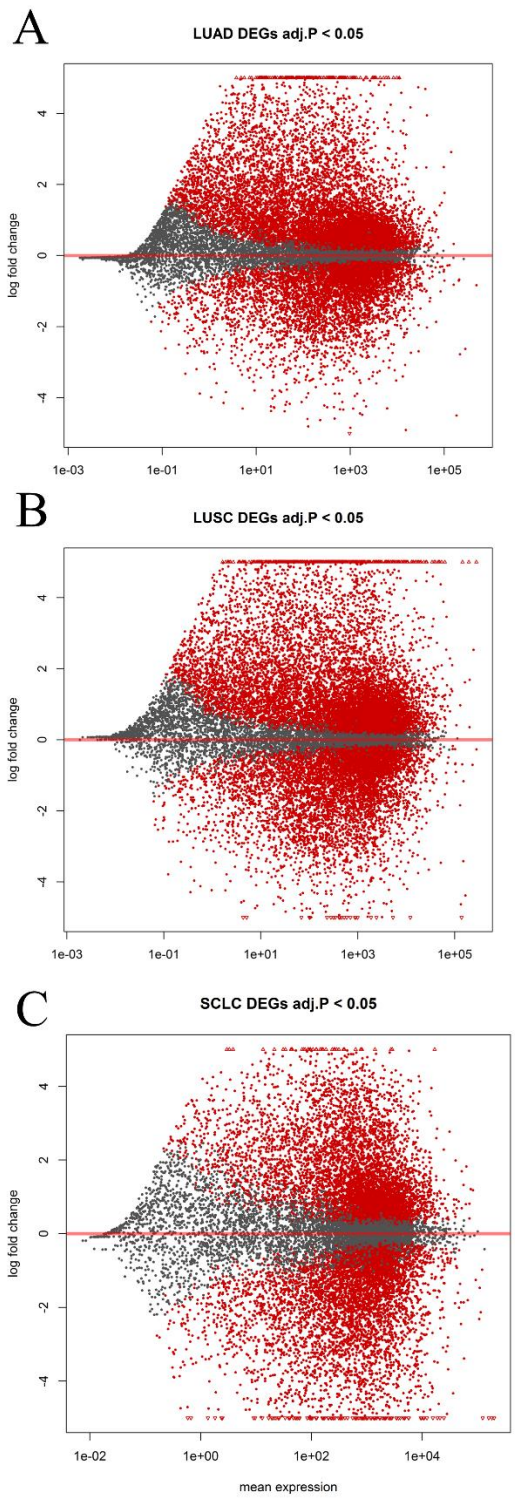


Figure 3.3. MA (log ratio over mean) plots for DEGs. Log fold change is plotted on the y-axis versus mean expression on the x-axis. LUAD, LUSC, and SCLC are plotted in A, B, and C, respectively. Red colored dots indicate significant DEGs at Benjamini-Hochberg (BH) adj.p < 0.05.

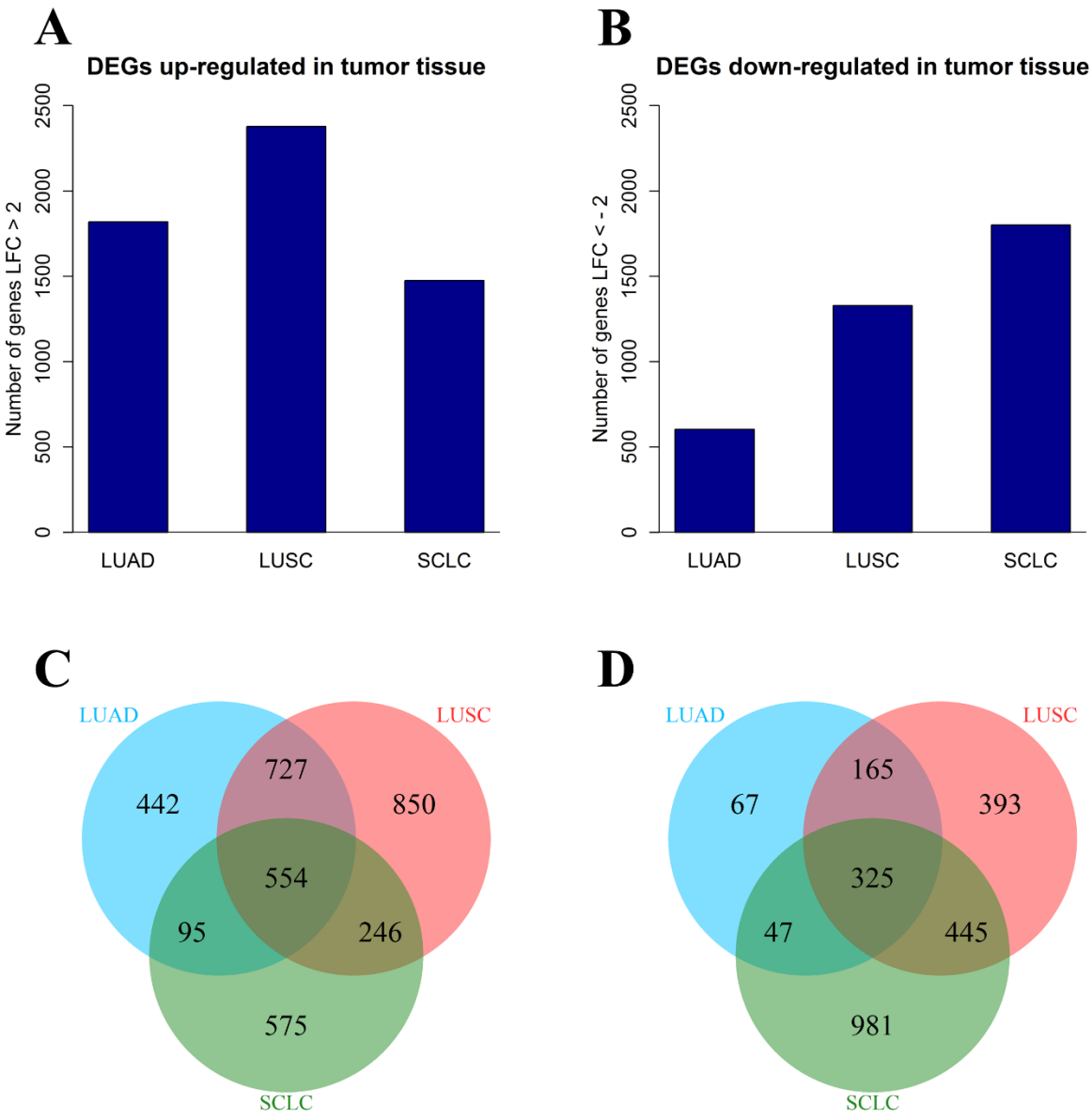


Figure 3.4. DEGs found in three lung cancer subtypes. Panel A shows the total number of DEGs up-regulated in lung tumor tissue versus normal controls. Panel B shows the total number of down-regulated DEGs in lung tumor versus normal controls. Panel C illustrates the overlap between up-regulated DEGs per subtype. Panel D illustrates the overlap between down-regulated genes per subtype. LFC = $\log_2$fold change.

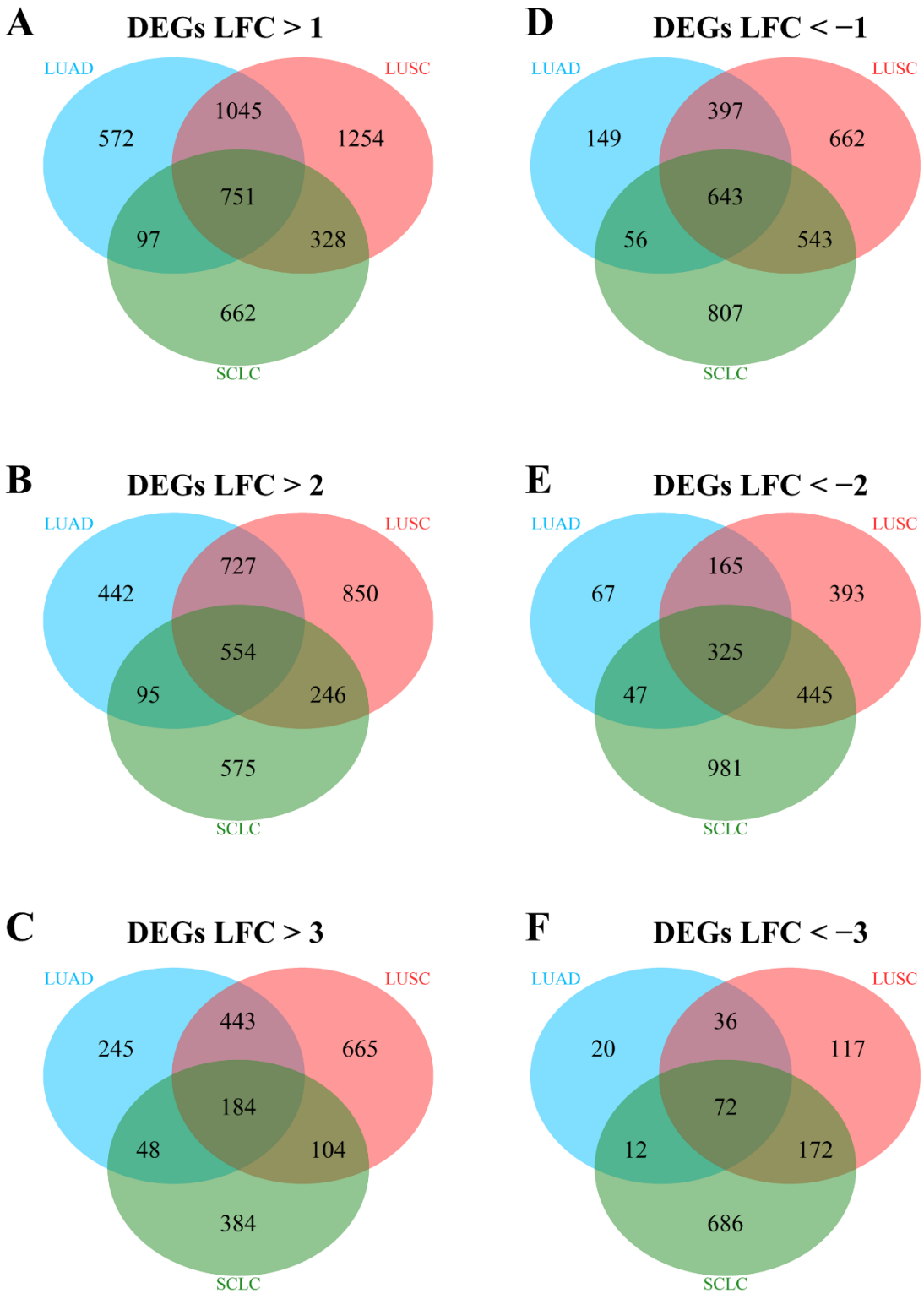


Figure 3.5. Overlap of DEGs for each subtype at multiple differential expression thresholds. Venn diagrams show the overlap in DEGs between subtypes. Panels A-C show up-regulated DEGs and D-F show down-regulated DEGs. LFC = $\log_2$fold change.

Tumor suppressor genes

We used the Tumor Suppressor Gene 2.0 (TSGene 2.0) database (156) to investigate the sets of down-regulated DEGs. The tumor suppressor genes (TSGs) found in all uniquely down-regulated genes per subtype and among overlapping gene sets are summarized in Appendix C. We calculated the percentage of TSGs per unique subtype and overlap and found that the largest percentage of TSGs were in the genes that overlapped SCLC and LUAD (Table 3.2). However, there was no significant difference between the sets ($p = 0.064$, test of proportions).

Table 3.2. Summary of TSGs found in down-regulated DEG sets.

Gene set	# genes in set	# TSGs	% overlap
LUAD unique	67	5	7.4
LUSC unique	393	34	8.7
SCLC unique	981	93	9.5
LUAD overlap LUSC	165	9	5.5
LUAD overlap SCLC	47	8	17.0
LUSC overlap SCLC	445	50	11.2
All overlap	325	21	6.5

Oncogenes

We used the oncogene database ONGene (5) to investigate the up-regulated DEGs found in all subtypes (Appendix D). We calculated the percentage of oncogenes found in each subtype and the overlapping gene sets. The results are listed in Table 3.3. There was a significant difference in the oncogenes found between all sets of genes ($p = 2.802 \times 10^{-10}$, test of proportions). The largest percentage of oncogenes were in the overlapping set of DEGs.

Table 3.3. Summary of oncogenes found in up-regulated DEG sets.

Gene set	# genes in set	# oncogenes	% overlap
LUAD unique	442	8	1.8
LUSC unique	850	26	3.1
SCLC unique	575	15	2.6
LUAD overlap LUSC	727	28	3.9
LUAD overlap SCLC	95	3	3.2
LUSC overlap SCLC	246	11	4.5
All overlap	554	54	9.7

Pathway enrichment of DEGs

To determine the biological activity driven by these DEGs, we performed biological pathway enrichment of the DEGs in each set of genes. We used WebGestalt (123) for the enrichment analysis with KEGG (124) as the source of the pathway definitions. Table 3.4 shows the 20 pathways that overlapped all three subtypes for the up-regulated genes and the 35 pathways for the down-regulated genes. We show the overlap of these pathways for all three subtypes in Figure 3.6.

Table 3.4. Enriched KEGG pathways from overlapping DEGs.

KEGG pathways up-regulated DEGs	KEGG pathways down-regulated DEGs
Alanine, aspartate and glutamate metabolism	African trypanosomiasis
Axon guidance*	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Calcium signaling pathway*	Axon guidance*
Cell cycle	Bile secretion
ECM-receptor interaction*	Calcium signaling pathway*
Gastric acid secretion	Cardiac muscle contraction
Homologous recombination	Cell adhesion molecules (CAMs)
Long-term potentiation	Chemokine signaling pathway
Maturity onset diabetes of the young	Complement and coagulation cascades
Melanoma	Cytokine-cytokine receptor interaction
Metabolic pathways	Dilated cardiomyopathy
Neuroactive ligand-receptor interaction*	Drug metabolism - cytochrome P450
Nitrogen metabolism	ECM-receptor interaction*
Oocyte meiosis	Endocytosis
p53 signaling pathway	Focal adhesion
Pathways in cancer*	Hematopoietic cell lineage
Progesterone-mediated oocyte maturation	Hypertrophic cardiomyopathy (HCM)
Protein digestion and absorption*	Jak-STAT signaling pathway
Salivary secretion*	Leukocyte transendothelial migration
Systemic lupus erythematosus	Long-term depression
	Malaria
	MAPK signaling pathway
	Metabolism of xenobiotics by cytochrome P450
	Neuroactive ligand-receptor interaction*
	Pancreatic secretion
	Pathways in cancer*
	Phagosome
	PPAR signaling pathway
	Protein digestion and absorption*
	Regulation of actin cytoskeleton
	Retinol metabolism
	Salivary secretion*
	Staphylococcus aureus infection
	Tight junction
	Vascular smooth muscle contraction

* Enriched with both up-regulated and down-regulated genes.

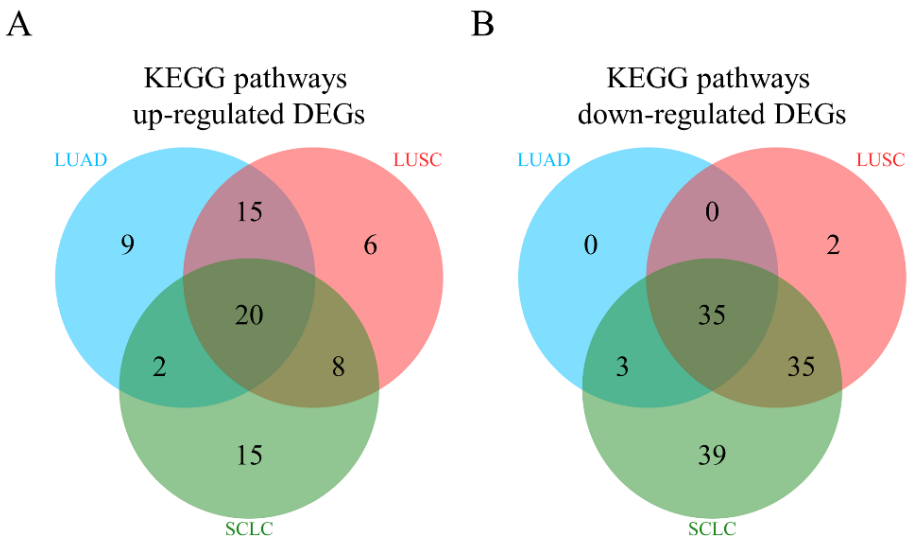


Figure 3.6. KEGG pathways that overlap between all three subtypes. Panel A shows the up-regulated DEGs shared between all three subtypes. Panel B shows the down-regulated genes shared by subtype.

Somatic mutations in three lung cancer subtypes

We selected all somatic mutations for LUAD and LUSC generated from TCGA. There were 234,278 mutations for LUAD and 65,305 for LUSC. SCLC was not studied as part of the TCGA, so we obtained a set of somatic mutations for SCLC from the same study that generated the RNA-Seq results (89). There were 7,945 somatic mutations for SCLC. For the following explorations of the specific somatic mutations per subtype, we removed synonymous SNVs. First, we determined the transition (Ti) to transversion (Tv) (Ti/Tv) ratio per subtype. We found that all three subtypes shared similar ratios of Ti (~35%) to Tv (~65%) (Figures 3.7A – 3.9A) based upon the average values across all samples per subtype. Specifically, the mean percentages of Tv mutations were 61.95, 63.87, and 65.20 for LUAD, LUSC, and SCLC, respectively. The mean

percentages of Ti mutations were 38.05 for LUAD, 36.13 for LUSC, and 34.80 for SCLC. To obtain greater insight behind the mutational profiles of each subtype, we generated their mutational signatures. We used non-negative matrix factorization (NMF) to generate the mutational signatures for each subtype. The three most significant signatures are plotted in Figures 3.7B-3.9B. Next, we compared the mutational signatures to a compiled set of known signatures in cancer from COSMIC (7). The top 3 signatures for each subtype and their relationships (see Methods) to COSMIC signatures are listed in Table 3.5. Interestingly, all three subtypes share the 4th COSMIC signature in common. This signature has recently been shown to be increased in cancers derived from smokers versus non-smokers (157). Also, LUAD and SCLC both share the 5th signature, while LUAD and LUSC share the 13th signature.

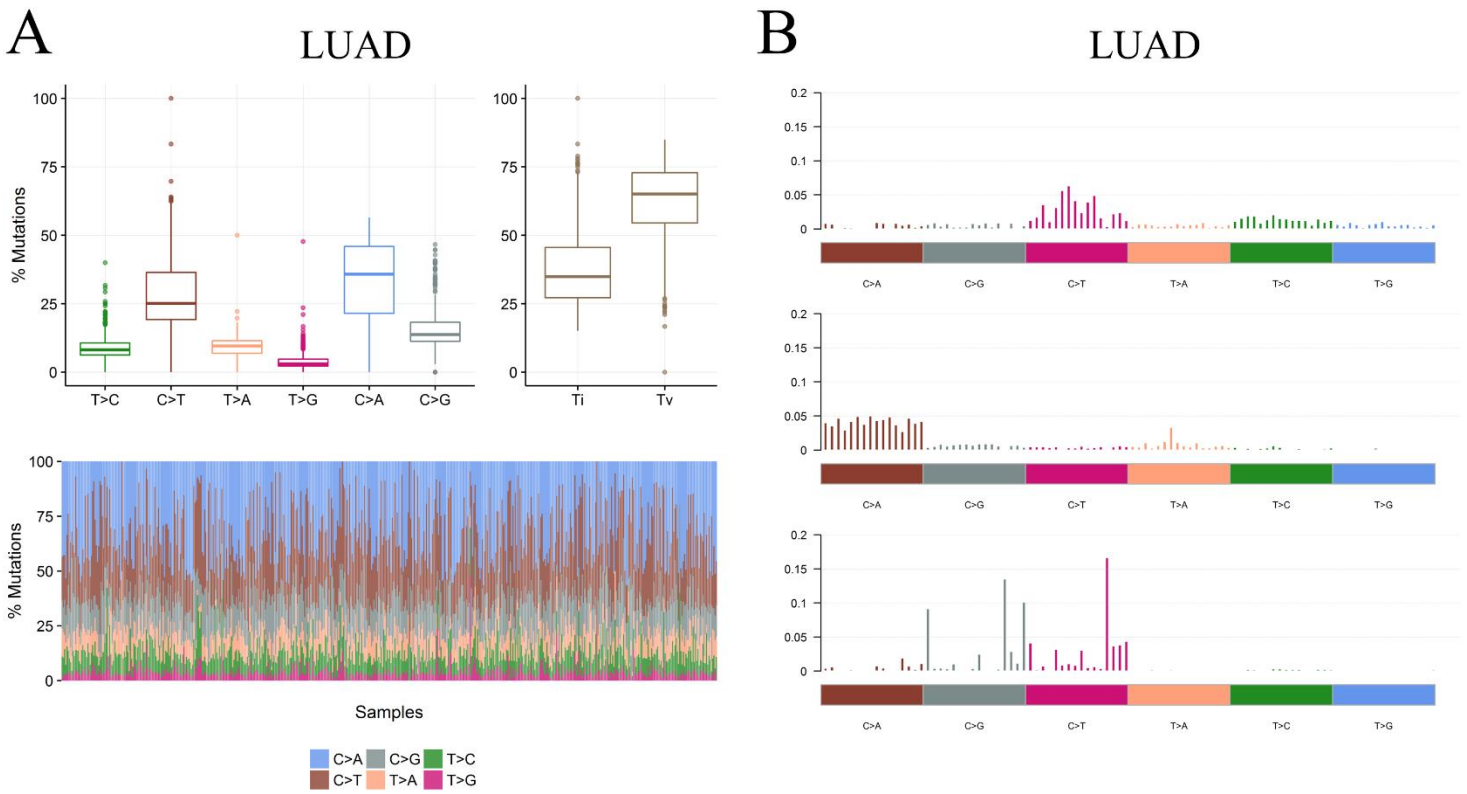


Figure 3.7. Summary of Ti/Tv ratios and mutational signatures for LUAD. Panel A shows the Ti/Tv ratio in LUAD. Panel B shows the frequency of mutations in the top three mutational signatures in LUAD.

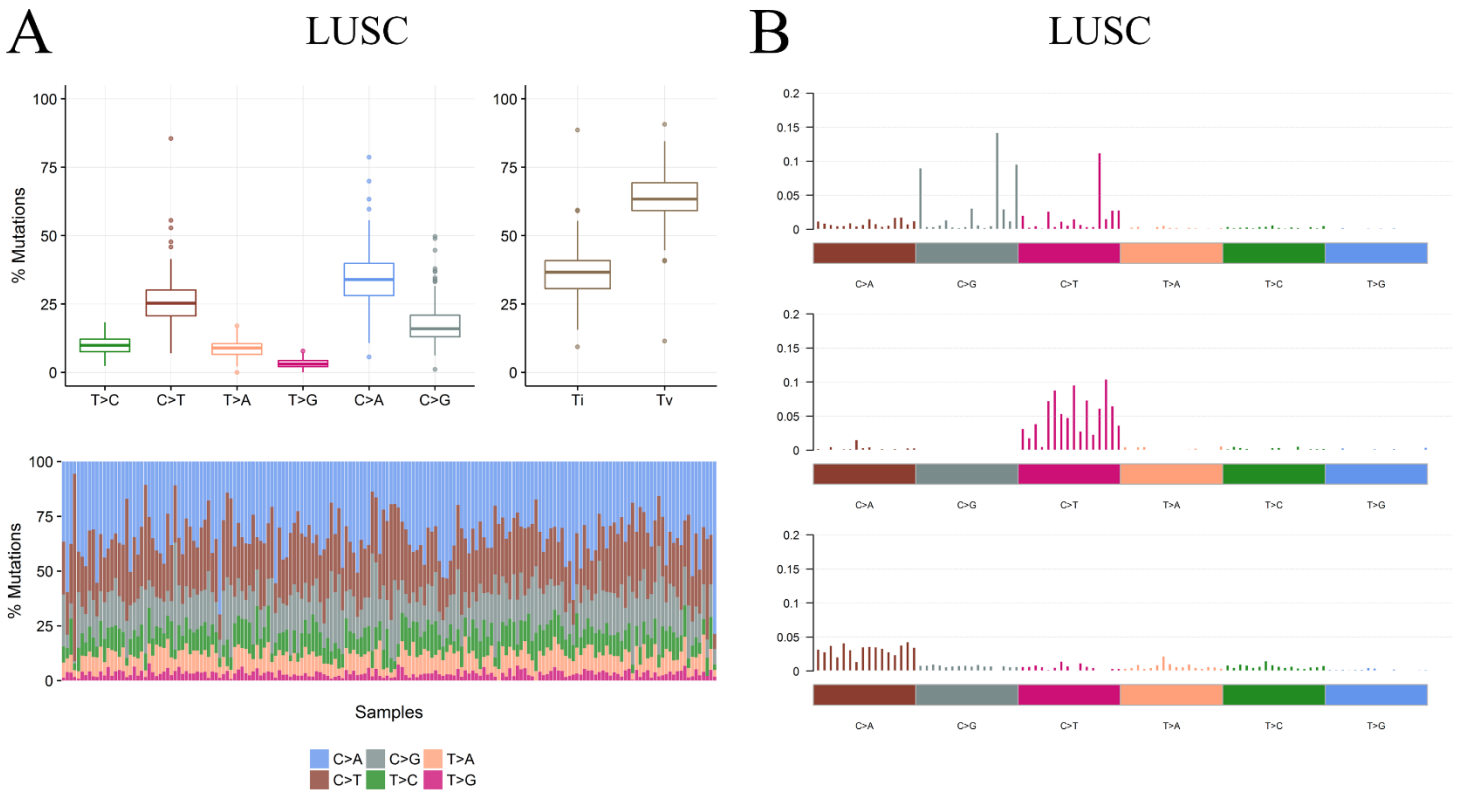


Figure 3.8. Summary of Ti/Tv ratios and mutational signatures for LUSC. Panel A shows the Ti/Tv ratio in LUSC. Panel B shows the frequency of mutations in the top three mutational signatures in LUSC.

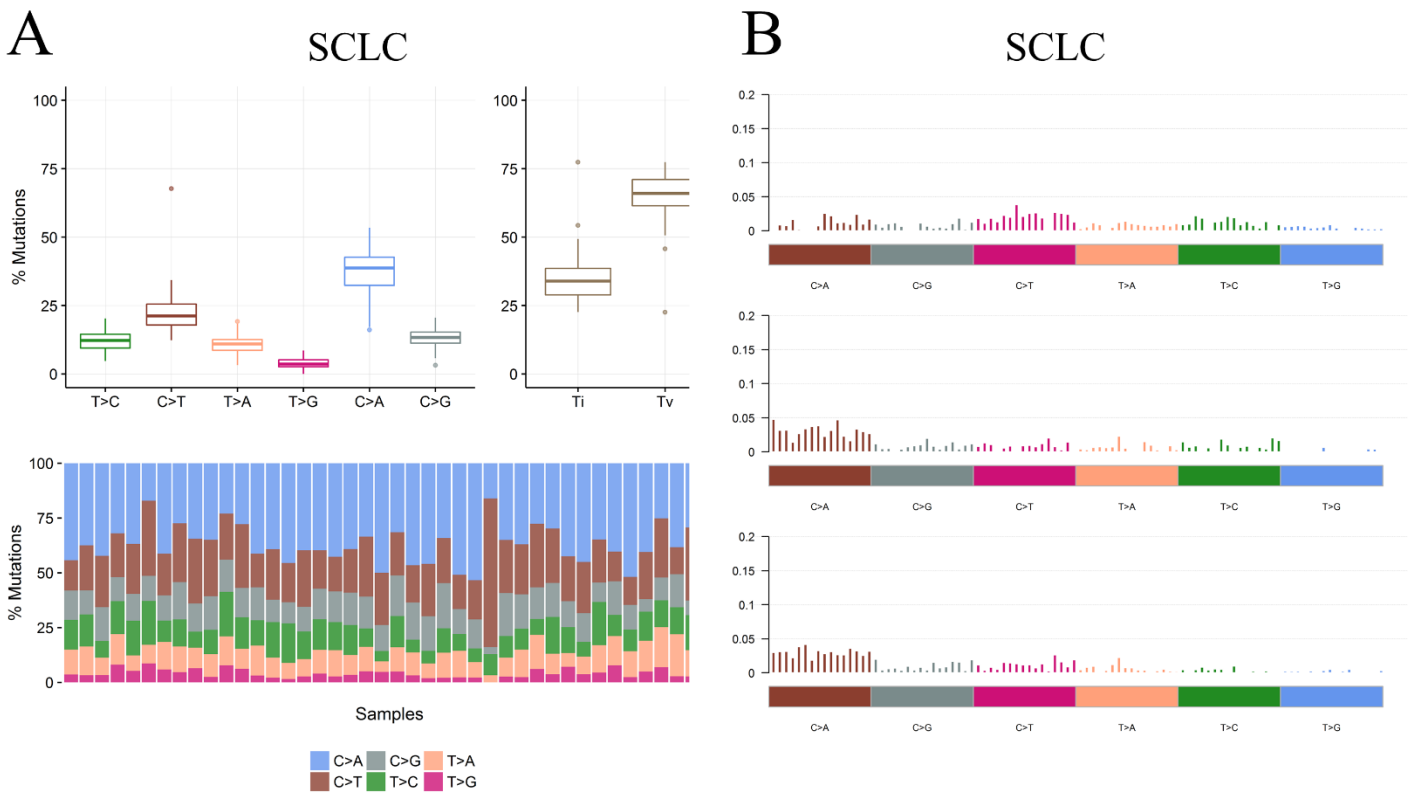


Figure 3.9. Summary of Ti/Tv ratios and mutational signatures for SCLC. Panel A shows the Ti/Tv ratio in SCLC. Panel B shows the frequency of mutations in the top three mutational signatures in SCLC.

Table 3.5. Summary of mutational signatures.

Mutational signature	Closest match in COSMIC	Cosine-Similarity
LUAD sig 1	Sig 5	0.77
LUAD sig 2	Sig 4	0.91
LUAD sig 3	Sig 13	0.77
LUSC sig 1	Sig 13	0.81
LUSC sig 2	Sig 30	0.85
LUSC sig 3	Sig 4	0.91
SCLC sig 1	Sig 5	0.84
SCLC sig 2	Sig 4	0.85
SCLC sig 3	Sig 4	0.92

The mutation rate for lung cancer is much higher than many cancer types (158), and the tumor genomes may harbor many passenger mutations, so we used several filtering steps to attempt to remove many of the genes that carried passenger mutations. (see Methods). Importantly, for our filtering process we did not exclude synonymous variants in LUAD or LUSC as was done in the previous analysis for mutational signatures (for details, see Discussion). The original number of genes that harbored at least one somatic mutation were 18,068, 14,789 and 5,180 for LUAD, LUSC, and SCLC, respectively. We applied our filtering steps to determine a list of possible lung cancer driver genes for each subtype at two different thresholds. The final genes are listed in Table 3.6. We refer to these sets of genes as the lenient and strict set of somatic mutated genes. Using our lenient threshold, we found 382, 247, and 67 genes for LUAD, LUSC, and SCLC, respectively (Figure 3.10A). Using our strict threshold, we found 106, 57, and 26 genes for LUAD, LUSC, and SCLC, respectively (Figure 3.10B). We illustrate the overlap of the somatic mutated genes in Figures 3.10C-D,

Table 3.6. Summary of filtered somatic mutated genes.

	LUAD	LUSC	SCLC
# samples	538	178	42
# initial mutations	234,278	65,305	7,945
# mutations in COSMIC	10,604	3,080	364
# genes mutated in COSMIC	560	513	193
# lenient genes	411	260	67
# strict genes	113	59	26
# genes expressed in normal tissue	410 (lenient), 112 (strict)	260 (lenient), 59 (strict)	N/A
# genes expressed in tumor tissue	410 (lenient), 112 (strict)	260 (lenient), 59 (strict)	N/A
# genes > 10 counts in normal tissue	379 (lenient), 106 (strict)	245 (lenient), 57 (strict)	N/A
# genes > 10 counts in tumor tissue	377 (lenient), 104 (strict)	238 (lenient), 56 (strict)	N/A
# final genes - lenient	**382**	**247**	**67**
# final genes - strict	**106**	**57**	**26**

Lenient thresholds: LUAD = mutated in at least 20 different subjects, LUSC = mutated in at least 7 different subjects, SCLC = mutated in at least 2 different subjects

Strict thresholds: LUAD = mutated in at least 5 different subjects, LUSC = mutated in at least 3 different subjects, SCLC = mutated in at least 2 subjects

N/A = genes for SCLC were previously filtered for expression, so did not include filtering step for this subtype.

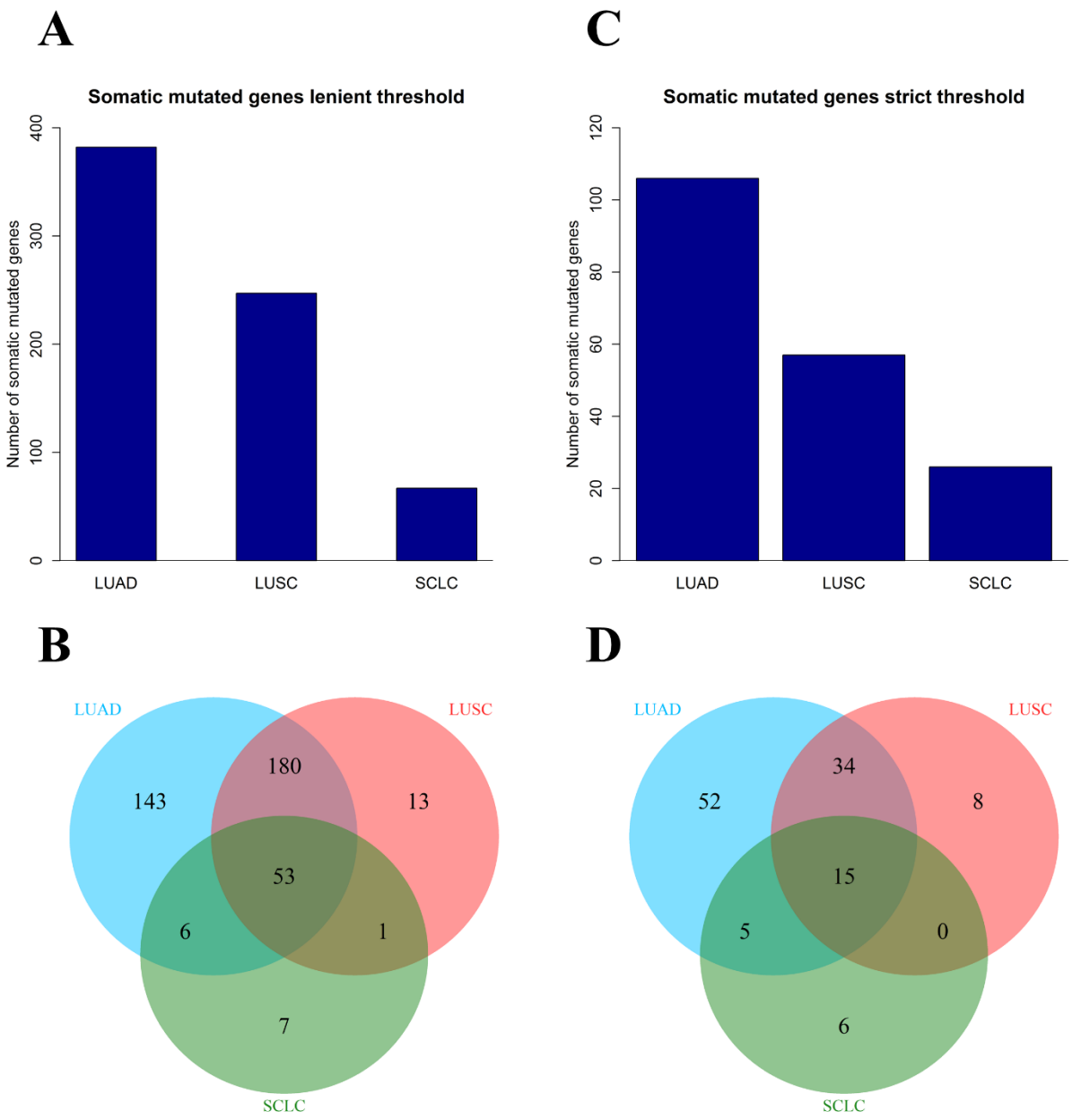


Figure 3.10. Final somatic mutated genes. Panel A shows the total number of lenient defined somatic mutated genes for each subtype. We show their overlap between subtypes in panel B. Panels C and D show the total number of strict defined somatic mutated genes and their overlap, respectively.

Discussion

The identification of the similarities and differences in three lung cancer subtypes at the somatic level is an important question to help ascertain and differentiate the molecular characteristics of the subtypes. This level of data is crucial because previous studies (159, 160) have demonstrated that typical chemotherapy-based options do not perform as well as targeted therapies in lung cancer. In this study, we used DNA mutations and RNA expression levels to characterize and compare three subtypes of lung cancer. We used RNA expression levels to generate a set of DEGs that are dysregulated in lung tumor tissue. We then used the DEGs to perform a pathway enrichment analysis. Additionally, we used somatic mutations to generate a Ti/Tv ratio and mutational signatures between subtypes. We then filtered a set of broad somatic mutations to identify a set of potential driver genes for each subtype.

RNA level analyses

We first generated a set of DEGs for all subtypes. We found that there not strong overlap between all of the three subtypes. We identified KEGG biological pathways that were enriched with the sets of DEGs. We also found that LUSC and SCLC shared many immune-related pathways that were not shared with LUAD, such as Antigen processing and presentation, Intestinal immune network for IgA production, and Antigen processing and presentation. Additionally, immune-related genes are affected in many cancer types such as breast cancer (161), so this warrants future investigation.

DNA level analyses

We generated the Ti/Tv ratio for all three subtypes. Our results showed that all subtypes had much higher transversions (~65%) than transitions (~35%). This finding agrees with previous work that suggested high transversions in NSCLC due to cigarette smoking (92). Specifically, we see a very high percentage of C>A transversions (~30% of mutations) in all subtypes. This pattern has been well-established in LUAD and LUSC (162). The observation of this same pattern in SCLC suggests it shares somatic mutation features with the NSCLC subtypes attributable to smoking. However, we would have expected to see much higher C>A mutations in this subtype because studies have found the strongest smoking association with SCLC (39). To expand upon this mutational investigation, we generated mutational signatures in each subtype. We identified the top three signatures and compared them to a set of COSMIC signatures. The COSMIC signatures are based upon a set of over 10,000 exomes and ~1,000 genomes over 40 different cancer types (7). Our results indicated that all subtypes shared the Signature 4 from COSMIC. This signature has been found in all three subtypes previously and appears to be associated with tobacco smoke (7). Additionally, recent work has found that this signature is more common in cancers that are found in smokers compared to non-smokers (157). The mutational signatures for SCLC were overwhelmed by smoking, and therefore, out of the three top signatures for SCLC, two of them matched to COSMIC signature 4. This suggests that there are not even three distinct mutational signatures and that a lot of the mutation pattern may be attributed to cigarette smoke in SCLC. This result indicates that the mutational signature for SCLC is overwhelmed by tobacco smoke related mutations. We also identified that COSMIC signature 5 matched signatures in LUAD and SCLC but not LUSC. The etiology of this signal is unknown. Overall, we found that the three subtypes shared most of their mutational signatures with at least one other subtype. However, one

signature for LUSC, signature 30 from COSMIC, was unique to LUSC. This pattern deserves attention because it has only been previously found in a small set of breast cancers (7). We also filtered a set of previously identified (82, 86, 89) somatic mutations to generate a set of possible driver genes in each subtype. We filtered out the majority of the mutated genes by initially removing genes not in the CGC list from COSMIC (7).

Study limitations and summary

There are several limitations to this study. First, the sample size for SCLC was much smaller than the NSCLC sample size. More comparable sample sizes between SCLC and NSCLC may have revealed more significant associations, but we were limited by available data. Second, the somatic mutations were identified using different approaches for NSCLC versus SCLC. Previous work (163, 164) has compared mutation callers and found differences, so there may be some issues with comparing the called variants. We also included silent mutations for LUAD and LUSC for filtering the somatic mutations to identify potential driver genes, but we did not include synonymous mutations for SCLC because they were already filtered out of the results. We did not want to remove the silent mutations in the NSCLC subtypes because previous work (165, 166) has demonstrated that synonymous variants may contribute to cancer (167). For future work, we can also exclude all silent mutations from LUAD and LUSC to compare with our results in this study. Finally, if clinical data is sufficient and sample size is large enough, we will consider smoking and other clinical factors like drug treatment as co-variates in our future analyses.

In summary, we used DNA somatic mutations and RNA expression data to compare three lung cancer subtypes. We calculated DEGs and a set of potential driver genes. We determined

mutational signatures for each cancer subtype. We identified the overlap between all three subtypes for somatic mutations and DEGs. Overall, this study provided strong insight into the biological similarities and differences at the somatic level for three subtypes of lung cancer.

CHAPTER IV

INVESTIGATION INTO THE CHALLENGES OF IDENTIFYING SOMATIC MUTATIONS IN LUNG CANCER USING RNA SEQUENCING VERSUS WHOLE EXOME SEQUENCING*

Introduction

As demonstrated in Chapter III, investigations into the somatic alterations in lung cancer are best approached at both the DNA and RNA levels. However, unless the analyses are done in large consortia such as TCGA (86), it may be cost prohibitive to generate DNA sequencing to call somatic variants and RNA sequencing (RNA-Seq) to identify expression levels. It would be ideal if one could use RNA-Seq as a tool to call somatic variants in addition to its role of determining RNA expression levels. In order to determine the effectiveness of using RNA-Seq to call variants compared to traditional based DNA whole exome sequencing (WES), we performed an integrative analysis for both techniques on the same set of lung cancer samples. Due to the restrictions mentioned above, we limited our analysis to the non-small cell lung cancer (NSCLC) subtype. We specifically focus on RNA-Seq's role to identify a single type of mutation, the single nucleotide variant (SNV). SNVs are the most abundant form of genetic variation in genome sequences and somatic SNVs play critical roles in disease including lung cancer (150). The discovery of many driver SNVs has led to new targets for therapeutic treatments and preventive measures.

---

* Adapted from O'Brien *et al*. *Methods* 2015. 83:118-127 (168)

Examples include vemurafenib specifically targeting BRAF V600 mutations in melanoma (169, 170) and gefitinib, erlotonib, and afatinib for EGFR mutations in lung cancer (142). The recent advances in next-generation sequencing (NGS) technologies, especially WES and whole transcriptome sequencing (RNA-Seq), have helped investigators generate a massive amount of NGS data, from which genetic variants, including SNVs, are detected. Many tools are now available for the detection of somatic SNVs from NGS data (163).

Both whole genome sequencing (WGS) and WES have been applied to detect SNVs in large scale cancer studies. While WGS can detect the full spectrum of variants, including SNVs, insertions/deletions (indels), copy number variations (CNVs), and structural variants (SVs), across the whole cancer genome, WES is more cost-effective in detecting SNVs and indels located in the 1-2% of the genome that encodes for functional proteins (171). There is good evidence that SNVs within the exome are responsible for many diseases, so WES has been applied extensively in research and clinically (171-173). RNA-Seq is commonly used for the measurement of gene expression levels, detection of gene fusions, and identification of splicing events. Because RNA-Seq is based on direct sequencing of cDNA, the product of the mRNA through reverse transcription, it may be feasible to detect SNVs from RNA-Seq data (174, 175). This is a unique feature that is different from the traditional microarray-based gene expression. RNA-Seq also has the ability to detect RNA editing, which is a post-transcriptional process that modifies RNA transcripts. One of the most common mechanisms of RNA editing is the deamination of adenosine to inosine by the protein Adenosine Deaminase Acting on RNA (ADAR). The inosine is interpreted in a similar way to guanosine and, thus, results in an adenosine to guanine (A → G) change (176).

RNA-Seq has been extensively applied to genomic and transcriptomic studies, including cancer. For example, a large-scale RNA-Seq study of lung adenocarcinoma identified several cancer driver genes (177), indicating its utility in a transcriptome analysis of cancer samples. This study demonstrated that in addition to identifying fusion genes and differential gene expression, RNA-Seq could detect well-known cancer driver genes. RNA-Seq has also been combined with WGS to better understand the mutational landscape of lung cancer (80, 178). These studies, in addition to showing the standard applications of RNA-Seq in gene expression analysis, highlight its usefulness as a technology platform for SNV detection, though challenges remain (179). As was demonstrated in the previous chapter, large consortia such as TCGA have applied both WES and RNA-Seq, as well as other platforms, to comprehensively catalog the cancer genome landscape (82, 86). The combination of WES and RNA-Seq data from the same tumor samples allows for large-scale examinations of somatic mutations in both the DNA and RNA. By applying these two types of technology together, one can improve the detection of various mutations, including those in the expressed genes with different splicing and expression levels, and those in non-transcribed regions. However, sequencing the same tumor using both platforms is rarely used due to cost and analysis issues.

A detailed comparison of SNVs called from WES and RNA-Seq data using the same lung cancer samples can not only reveal the technical differences of these two technologies, but also help us better understand the underlying biological processes that lead to the ambiguous observations of SNVs at the DNA and RNA levels, respectively. Such a comparison can provide guidance on the utility of WES and RNA-Seq in SNV detection. So far, there have been only a few attempts to unveil the advantages and disadvantages of WES and RNA-Seq in SNV detection. For example, Cirulli et al. (180) recently compared WGS with RNA-Seq in detecting SNVs using

peripheral blood mononuclear cells from the same subjects. They highlighted many important aspects for SNV detection such as expression levels and read depth, but its conclusions are yet to be validated due to the limited sample size. Another recent review compared WES and RNA-Seq (181), but it only discussed several global features without a systematic comparison of detailed features.

In this study, we compared the features of SNVs from WES and RNA-Seq using a collection of 27 NSCLC tumor and matched normal samples from the same patients. Through our systematic analyses, we attempted to unveil the unique features of SNVs from each platform and determined why variants are missed between these platforms. Because of the high false calling rate of indels, we only focused on SNVs. We observed only a small overlap of SNVs between WES and RNA-Seq, and identified multiple technological and biological reasons leading to discrepancies in SNV calling.

Methods

Samples and sequencing

Twenty-seven paired tumor and normal NSCLC samples from patients undergoing lung cancer surgery at Massachusetts General Hospital were used for this analysis. For all 27 paired tumor and normal lung cancer samples, we performed both WES and RNA-Seq experiments. All participants provided written informed consent. Tumor content was assessed with an average of 60% across samples. The exome regions were captured using the Agilent SureSelect Human All

Exon kit and then sequenced on an Illumina HiSeq 2000 platform (paired end, 100 bp) in a Massachusetts General Hospital (MGH) core. We obtained a total of 3,677,811,274 paired-end reads with an average sequencing depth of 121×. For RNA-Seq, Illumina Tru-Seq v2 RNA-Seq kit was used for enrichment of mRNA, cDNA synthesis, and library construction. Then, RNA sequencing was performed on an Illumina HiSeq 2000 platform in the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core (paired end, 100 bp). We obtained a total of 4,778,766,598 paired end reads with an average of 88,495,678 paired end reads per sample. We used FASTQC to check the quality of reads of all samples (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

WES data analysis

We mapped the WES reads to the human reference genome hg19 (GRCh37) using Burrows-Wheeler Aligner (BWA) (version 0.5.9c) (182). In order to further process the data, we used Picard (version 1.95) (183) to mark duplicate reads and used GATK (version 1.0.3825) to perform local realignment and recalibration (184, 185). After post-alignment processing of the data, we called SNVs with MuTect (version 1.1.4). To generate mpileup files for each tumor and normal sample, we used the "mpileup" function in Samtools (version 0.1.19) (186). Read count values were obtained from the mpileup files using VarScan2 (version 2.3.5) (187) with the "readcounts" function. Read count values were split up into categories of values: not covered (NA), single read (1), low coverage (2-7) and high coverage (≥ 8).

RNA-Seq data analysis

We used TopHat2 (version 2.0.0) (188) to map RNA-Seq reads to the human reference transcriptome and genome (hg19). TopHat2 first attempts to map reads to the reference transcriptome and then for the unmapped reads, it attempts to map them to the human genome reference. As we did for WES data, we called SNVs using MuTect (version 1.1.4). Specifically, we generated mpileup files using Samtools and obtained read count values using VarScan2. We used Cufflinks (version 2.1.1) (189) to obtain gene-based FPKM (Fragments Per Kilobase of exon per Million fragments Mapped) values for all samples. FPKM values corresponding to degrees of expression were as follows: not covered (NA), no expression (FPKM < 1), very low expression (FPKM 1-5), low to moderate expression (FPKM 5-20), and high expression (FPKM > 20).

Read counting for the RNA-Seq SNVs covered by the WES capture kit

We used Bedtools (version 2.17.0) to determine whether the SNVs identified from RNA-Seq were covered by the WES capture kit using the "-intersectBed" function. SNVs were categorized into four groups by read count values as was done for the aforementioned read count analysis: not covered (NA), single read (1), low coverage (2-7) and high coverage (≥ 8).

Mutation pattern categorization for all SNVs

We categorized SNVs into six groups according to their nucleotide changes: A:T→C:G, A:T→T:A, A:T→G:C, C:G→A:T, C:G→G:C, and C:G→T:A.

The computational tools that we used for all analyses are summarized in Table 4.1.

Table 4.1. Tools used for comparing WES versus RNA-Seq data.

Method/tool	Purpose	URL
FASTQC	Check quality of WES and RNA-Seq reads	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BWA	Map WES reads to the reference genome	http://bio-bwa.sourceforge.net/
Picard	Mark duplicate WES reads	http://broadinstitute.github.io/picard/
GATK	Perform local realignment and recalibration of WES reads	https://www.broadinstitute.org/gatk/
MuTect	Detect SNVs in WES and RNA-Seq	http://www.broadinstitute.org/cancer/cga/mutect
Samtools	Generate mpileup files for WES and RNA-Seq	http://samtools.sourceforge.net/
VarScan2	Generate read counts for WES and RNA-Seq	http://varscan.sourceforge.net/
TopHat2	Map RNA-Seq reads to the human reference transcriptome and genome	http://ccb.jhu.edu/software/tophat/index.shtml
Cufflinks	Calculate FPKM gene expression levels for RNA-Seq	https://github.com/cole-trapnell-lab/cufflinks
Bedtools	Intersect RNA-Seq SNVs with WES capture kit	https://github.com/arq5x/bedtools2
R	Perform the analysis for SNV comparisons	http://www.r-project.org/

This table summarizes the computational tools used in our WES versus RNA-Seq comparative analysis. We include each tool used in our analysis, our use for the tool, and the URL link to the website. Further details including citations for all tools listed above are in the main methods section of the text

Results

Figure 4.1 illustrates the concept of our SNV comparison from WES and RNA-Seq data. There are several factors that may cause a difference in detecting SNVs from WES and RNA-Seq data, even from the same samples. First, the two sequencing technologies and their sequencing strategy will have variation in the enrichment of sequence regions. Second, at the biological level, SNVs detected from DNA-Seq (i.e., WES) may not be detectable by RNA-Seq due to low coverage, or tissue-specific expression and alternative splicing. In contrast, SNVs in the transcriptome may not be detected in WES because of low coverage, RNA editing, or their location outside of the WES capture regions. With these factors, we performed an in-depth comparison between SNVs detected by the two sequencing techniques.

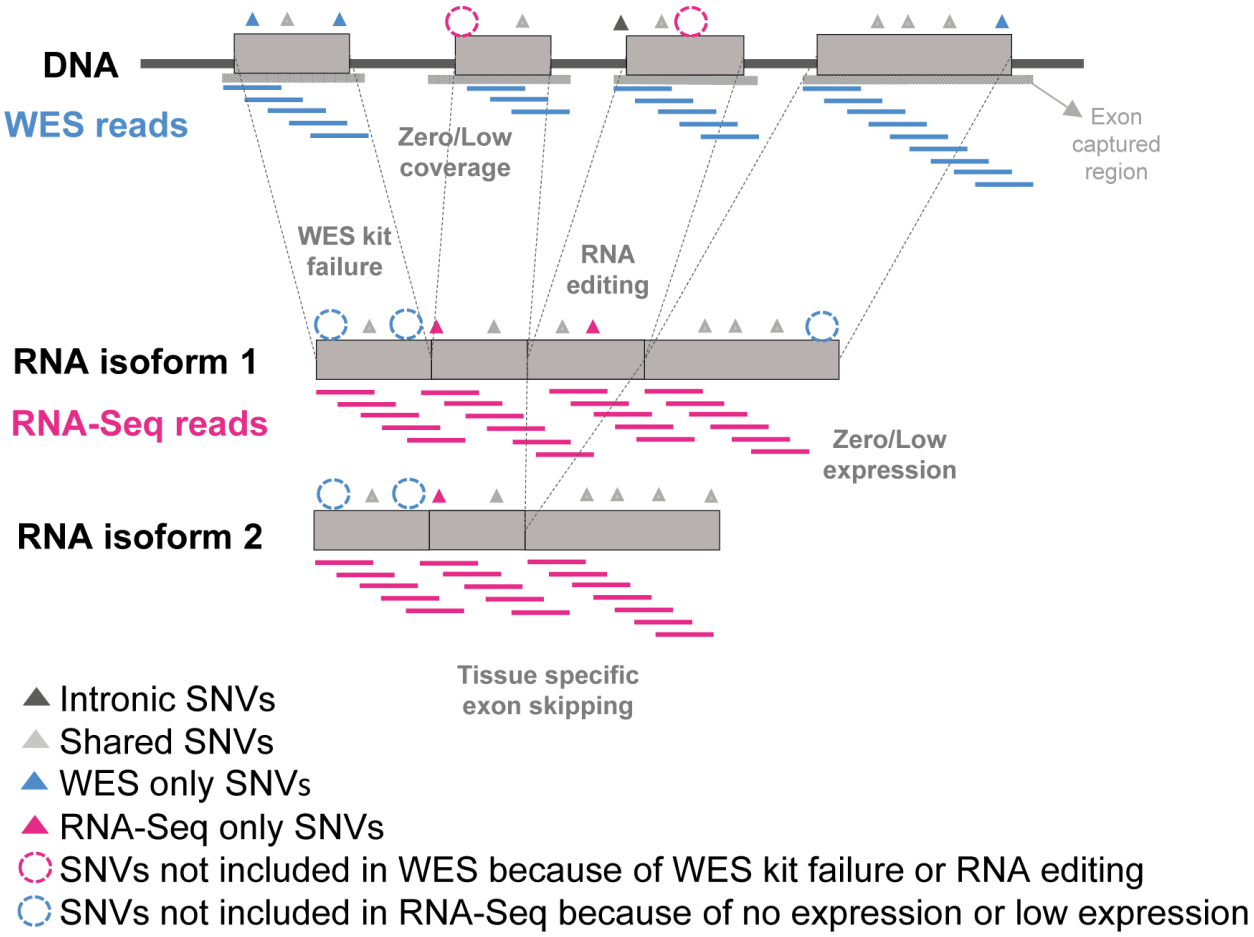


Figure 4.1. Comparison between WES data and RNA-Seq data. This figure shows the motivation and the concept behind our study. WES reads are generated on the exon captured regions. RNA-Seq reads are generated on the content of gene expression conditions. SNVs may exist in various locations of the genome including introns adjacent to exons in the DNA, and for locations within the transcriptome. SNVs for the intronic, WES and RNA-Seq shared, WES only, RNA-Seq only are colored with dark grey, light grey, blue and pink, respectively. SNVs not included in WES by the low coverage or WES kit failure or RNA editing are represented with pink dotted circles. SNVs not included in RNA-Seq by the low expression or coverage are represented with blue dotted circles.

Poor concordance for SNVs called in WES and RNA-Seq data

We obtained WES and RNA-Seq data for 27 lung cancer tumor samples and their matched normal samples. We applied a standard pipeline to analyze the samples and detect somatic SNVs (Figure 4.2). We refer to the SNVs that were uniquely detected in WES but not in RNA-Seq data as "WES unique SNVs," the SNVs that were uniquely detected in RNA-Seq but not in WES data as "RNA-Seq unique SNVs," and those observed in both WES and RNA-Seq as "WES shared SNVs" or "RNA-Seq shared SNVs." Note that although the WES shared SNVs and the RNA-Seq shared SNVs have the same genomic coordinates, they may have different alternative allele frequencies, or even different alternative alleles, in the WES data and in the RNA-Seq data. Thus, we referred to them separately as WES shared SNVs and RNA-Seq shared SNVs. Overall, we identified 15,662 SNVs from the WES data, with an average of 580 ± 517 SNVs per sample, and 15,473 SNVs from the RNA-Seq data, with an average of 573 ± 332 SNVs per sample. Surprisingly, only ~14% (2,150) of these SNVs were detected by both WES and RNA-Seq (Table 4.2).

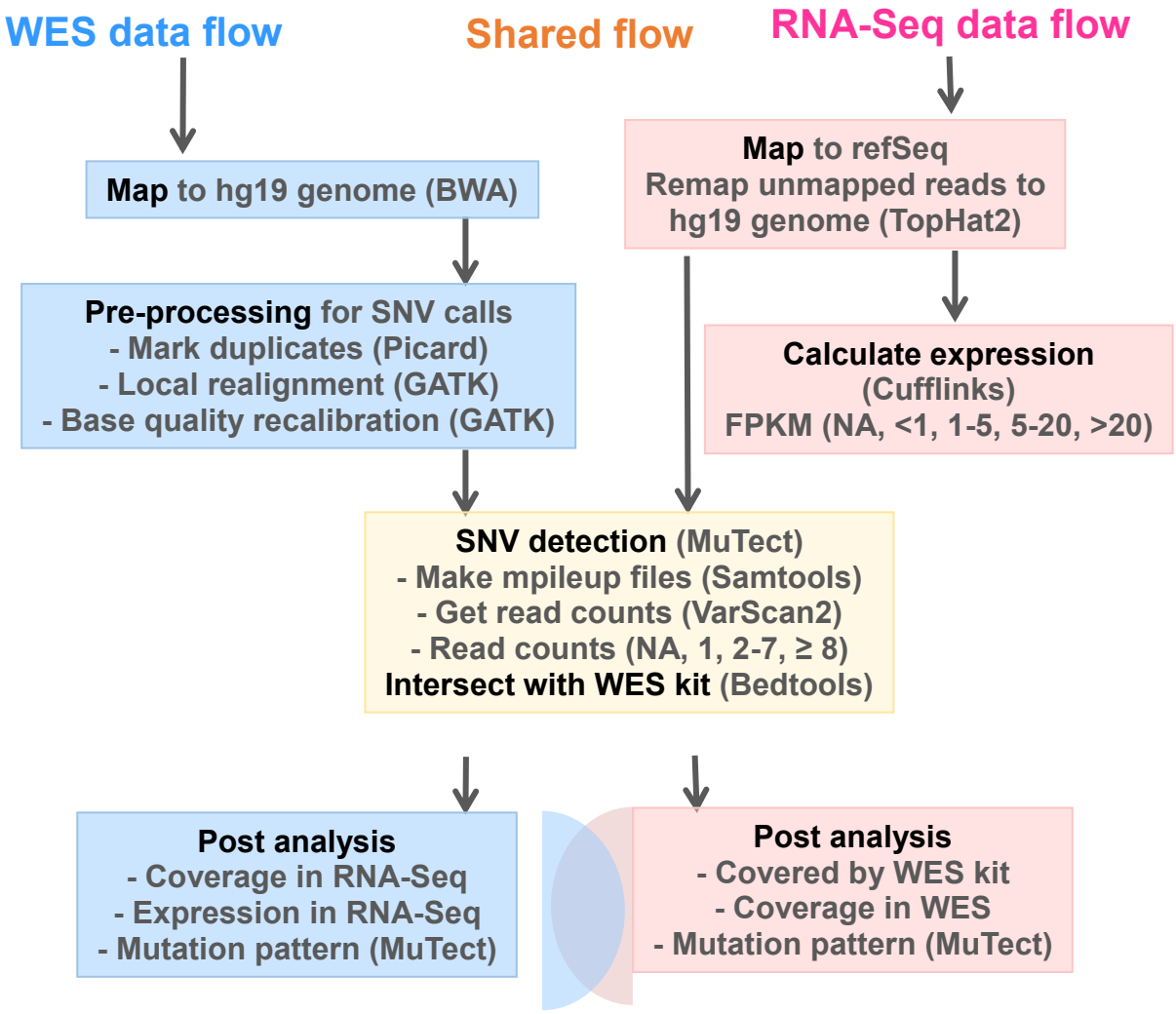


Figure 4.2. Work flow for the overall analysis. The blue color scheme is for the WES work flow. The pink color scheme is for the RNA-Seq work flow. Typical pipelines for WES data and for RNA-Seq data were used. From the SNVs called using MuTect, we performed several comparisons. For the WES data sets, we compared SNV lists from RNA-Seq, and analyzed gene expression patterns. For the RNA-Seq data sets, we compared SNV lists from WES and compared the mutation patterns.

Table 4.2. Summary of all SNVs detected in RNA-Seq and WES by MuTect.

Sample ID	RNA-Seq	WES	Overlap	Overlap with RNA-Seq (%)	Overlap with WES (%)
1	452	388	52	11.5	13.4
2	1,082	1,206	263	24.3	21.8
3	731	902	175	23.9	19.4
4	531	62	9	1.7	14.5
5	572	83	4	0.7	4.8
6	640	619	92	14.4	14.9
7	317	94	8	2.5	8.5
8	220	85	5	2.3	5.9
9	659	168	23	3.5	13.7
10	524	78	8	1.5	10.3
11	529	447	36	6.8	8.1
12	597	773	112	18.8	14.5
13	432	335	54	12.5	16.1
14	533	1,360	124	23.3	9.1
15	403	540	30	7.4	5.6
16	768	892	143	18.6	16.0
17	590	296	56	9.5	18.9
18	753	172	26	3.5	15.1
19	313	1,060	42	13.4	4.0
20	422	1,017	125	29.6	12.3
21	188	716	9	4.8	1.3
22	1,425	901	158	11.1	17.5
23	348	309	19	5.5	6.1
24	310	227	45	14.5	19.8
25	97	66	1	1.0	1.5
26	508	577	58	11.4	10.1
27	1,529	2,289	473	30.9	20.7
Mean ± SD	573 ± 332	580 ± 517	80 ± 102	13.9 ± 9.0	13.7 ± 6.0
Total	15,473	15,662	2,150		

We explored the reasons why such a small portion of WES SNVs was detected in the RNA-Seq data. One possibility is that the positions of the WES SNVs are not well covered in RNA-Seq. A large proportion of the WES unique SNVs (41.0%) are not covered in RNA-Seq. However, the majority (96.9%) of the WES shared SNVs have at least eight RNA-Seq reads mapped to their position (Figure 4.3). There is a small proportion of WES unique and WES shared SNVs moderately covered in RNA-Seq (2-7 reads), 8.8 – 24.2% and 0 – 33.3% respectively. Interestingly, 11.2 – 58.8% of the WES unique SNVs have a high number (≥8) of RNA-Seq reads aligned to their position. However, these are still undetected in RNA-Seq. Only one WES shared SNV was not covered (NA) in RNA-Seq, and this is likely a false positive detected from the MuTect analysis. We hypothesized that some of the WES unique SNVs may be located in genes which are not expressed, or have very low expression levels, and therefore are undetected by RNA-Seq.

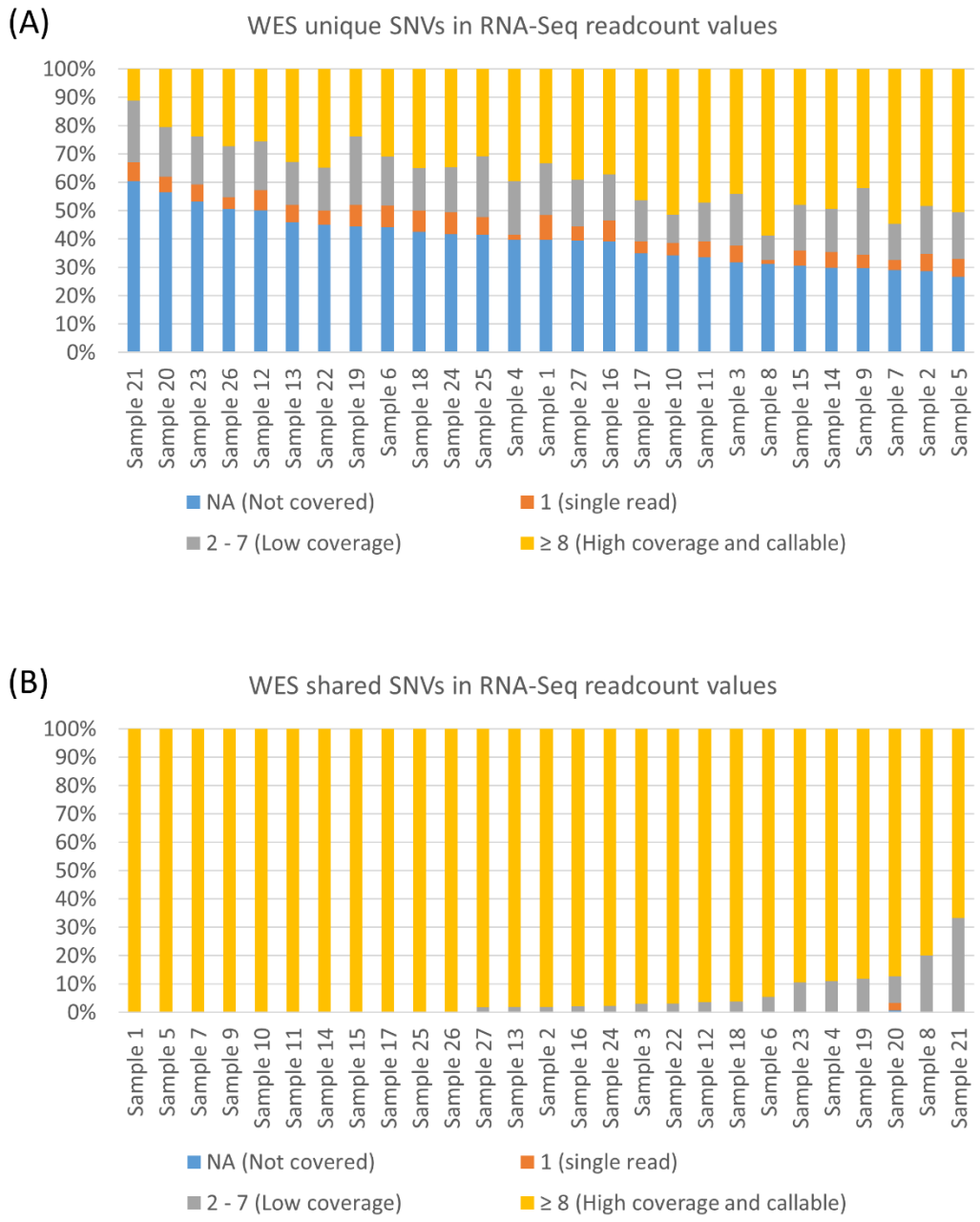


Figure 4.3. VarScan2 read count values determine why WES unique SNVs are not called by RNA-Seq. (A) Stacked column graph showing read counts results in RNA-Seq for WES unique SNVs. (B) Bar plot showing read counts results in RNA-Seq for WES shared SNVs. Blue represents read counts NA (not covered), orange represents read counts 1, grey represents read counts 2-7, and yellow represents read counts ≥ 8. Around 50% of WES unique SNVs are not covered in RNA-Seq. Samples ordered by decreasing NA (in A) and decreasing ≥ 8 (in B).

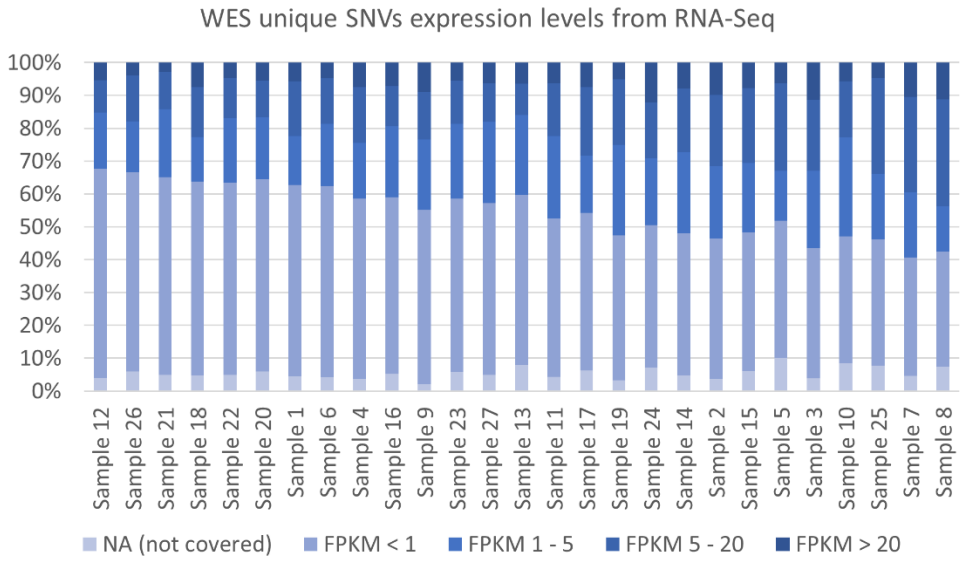
We further explored the features of WES unique SNVs regarding their gene expression levels. We used the software Cufflinks to generate FPKM values from RNA-Seq data for the chromosomal loci of WES SNVs (Table 4.3). We categorized FPKM values as not covered (NA), not expressed (< 1 FPKM), low expression (1-5 FPKM), low to moderate expression (5 – 20 FPKM) and high expression (> 20 FPKM) (Figure 4.4). Many of the WES unique SNVs are located in genes that are not expressed (51.0%). In contrast, 77.7% of WES shared SNVs are located in genes with FPKM > 5, including 0 – 66.7% of WES shared SNVs located in genes with low to moderate expression (FPKM 5-20), and 11.1 - 100% WES shared SNVs located in genes with high expression levels (> 20 FPKM).

Table 4.3. Summary of FPKM[a] levels from RNA-Seq for SNVs detected by WES.

	NA[b]	**< 1**	**1 − 5**	**5 – 20**	**> 20**	**Total**
WES unique SNVs[c]						
Mean ± SD	24 ± 20	255 ± 220	109 ± 106	79 ± 68	33 ± 31	500 ± 429
Range	2 - 90	25 - 948	9 - 449	9 - 240	3 - 114	53 - 1816
Range of %	2.1 - 10.1%	35.0 - 63.5%	13.7 - 30.0%	9.6 - 32.5%	2.8 - 12.1%	
WES shared SNVs[d]						
Mean ± SD	2 ± 3	1 ± 2	15 ± 23	38 ± 47	24 ± 31	80 ± 102
Range	0 - 10	0 - 7	0 - 103	0 - 224	1 - 132	1 - 473
Range of %	0 - 11.5%	0 - 5.3%	0 - 27.2%	0 - 66.7%	11.1 - 100%	

[a] FPKM: Fragments Per Kilobase of transcript per Million mapped reads.
FPKM gene expression values were generated by Cufflinks.
[b] NA: SNV positions from WES that are not covered at the gene level in RNA-Seq.
[c] WES unique SNVs: SNVs detected only in WES.
[d] WES shared SNVs: SNVs detected in both WES and RNA-Seq.

(A)



(B)

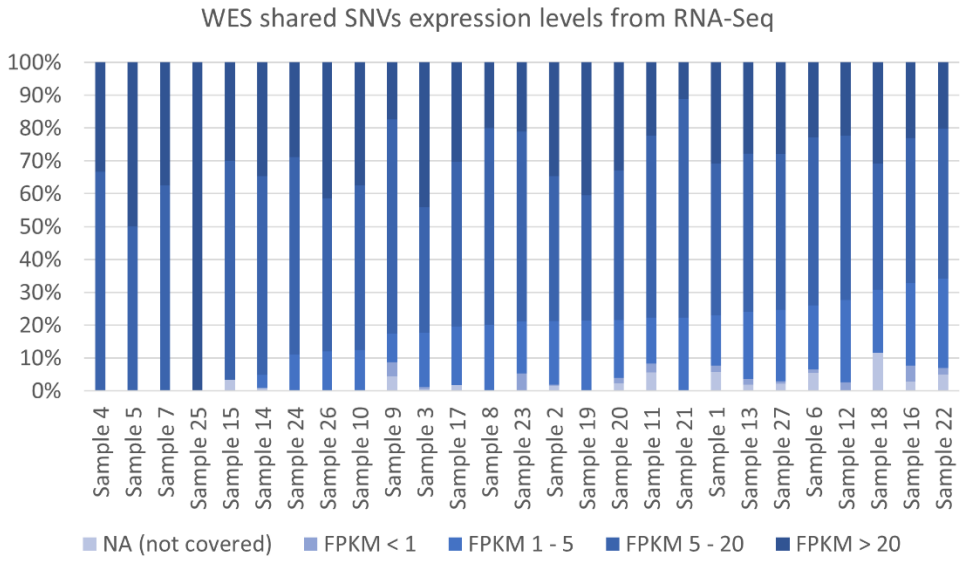


Figure 4.4. Cufflinks analysis to determine gene expression levels of WES unique SNVs in RNA-Seq. (A) FPKM values for WES unique SNVs. (B) FPKM values for SNVs shared between WES and RNA-Seq. Most WES unique SNVs are located within genes which are not expressed in RNA-Seq. FPKM NA: not covered, FPKM < 1: not detected; FPKM 1-5: not expressed; FPKM 5 -20: low to moderate expression; and FPKM > 20: high expression. Samples ordered by decreasing percentage of SNVs FPKM < 1 (in A) and decreasing percentage of FPKM > 5 (in B).

Feature analysis of RNA-Seq unique SNVs

We then examined the features of RNA-Seq unique variants. We first explored RNA-Seq unique SNVs that may be located outside of the WES capture regions. RNA-Seq does not contain a specific exome capture step, so the variants detected are not constrained to the specific 1-2% of the genome sequenced by WES, and are only limited to the genomic regions that are being transcribed. We first explored the proportion of RNA-Seq unique SNVs that lie outside of the WES capture region. We used the "-intersectBed" command in Bedtools to identify RNA-Seq unique SNVs that are not covered by the WES capture region. For the 13,323 RNA-Seq unique SNVs, 9,513 (71.4%) are located outside of the WES capture regions (Figure 4.5). We used VarScan2 to identify the read count values for the positions that are covered by the WES capture kit. We discovered that for the RNA-Seq unique SNVs covered by the kit, an average of ~93% (82.2 – 98.3%) are in locations that are highly covered ($\geq$ 8 reads) (Table 4.4). This is an interesting observation - it means that only approximately 7.0% of the SNVs uniquely called in RNA-Seq are potentially missed in WES due to low coverage of sequencing. Thus, the remaining SNVs are not missed due to technical issues, but due to biological issues.

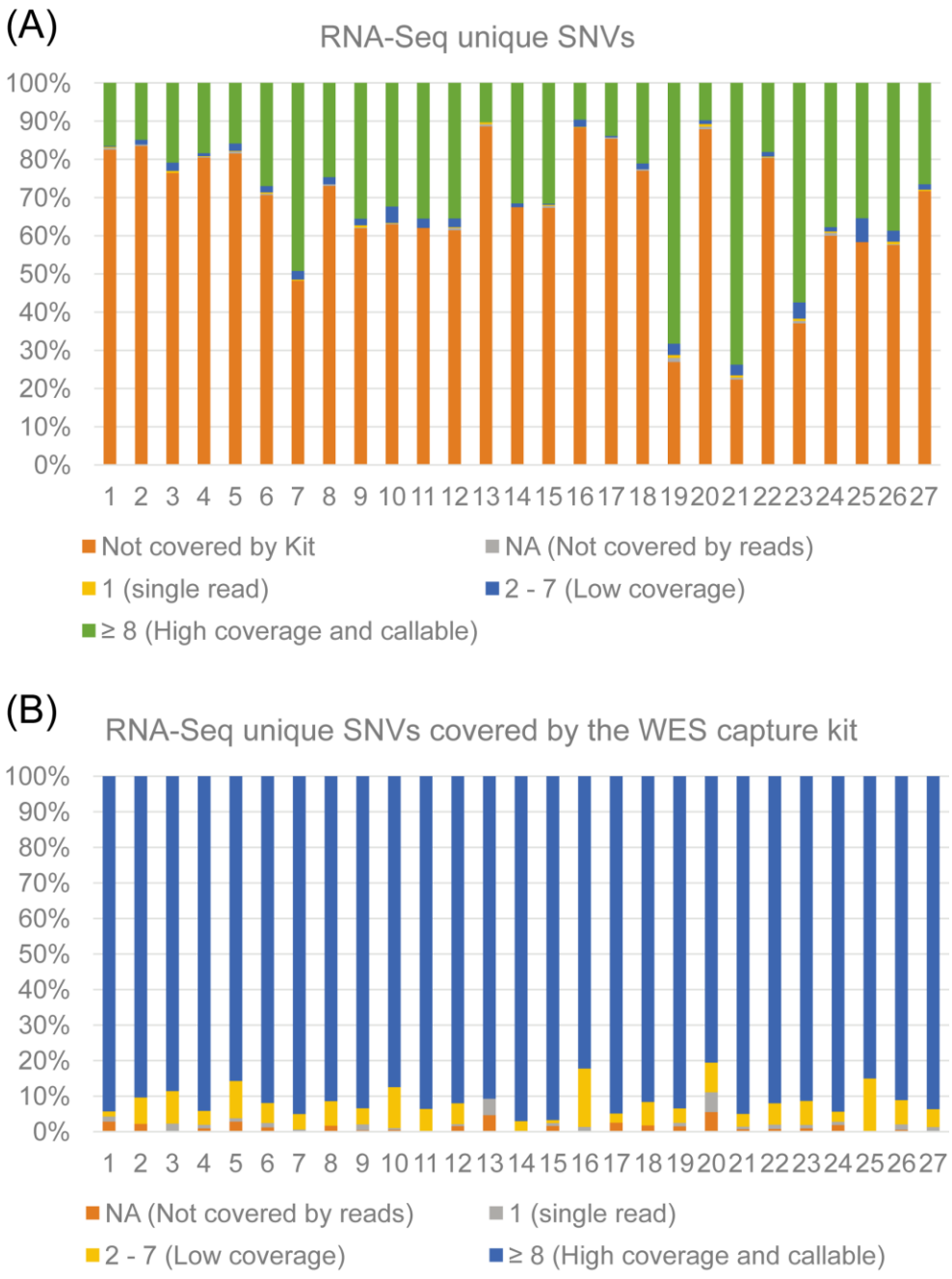


Figure 4.5. RNA-Seq unique SNVs not covered by the WES kit and coverage levels. (A) Bar plot shows the percentage of RNA-Seq unique SNVs within each sample that are not covered by the WES capture kit. Also included are VarScan2 read count values for covered positions. Figure 4.5A shows that most SNVs are not covered by the WES kit. Here, 'not covered by kit' represents RNA-Seq SNVs outside of the capture kit region; read counts values represented by NA, 1, 2 – 7, and ≥ 8. (B) Bar plot containing VarScan2 read counts values for only the positions covered by the WES kit. Most SNVs covered by the WES kit have high coverage. Read counts values represented by NA, 1, 2 – 7, and ≥ 8.

Table 4.4. Summary of WES coverage for RNA-Seq SNVs that are covered by the WES capture kit.

	NA	**1**	**2 - 7**	**≥ 8**	**Total in kit**
Mean	1 ± 1	1 ± 1	8 ± 5	130 ± 64	140 ± 68
Range	0 - 3	0 - 4	0 - 22	29 - 280	34 - 299
Range of %	0 - 5.9%	0 - 5.9%	0 - 16.4%	82.2 - 98.3%	

We hypothesized that RNA editing is another factor leading to the RNA-Seq SNVs being undetected in WES. Although there are known difficulties detecting RNA editing in NGS data (190-192), we explored this mechanism as a potential reason for inconsistencies in mutation calling between WES and RNA-Seq. We used the results from MuTect to analyze the base-pair mutation pattern across all SNVs for signatures of RNA editing. Interestingly, the most common mutation pattern for the RNA-Seq unique SNVs was the A:T→G:C mutation pattern, occurring in 55.3% of SNVs (Figure 4.6). Another interesting finding was that 21.4% of the RNA-Seq unique SNVs that were covered by the WES capture kit (but not detected in WES) also shared this same mutation pattern. In comparison, only 6.7% of the total number of overlapping SNVs called in both WES and RNA-Seq had this mutation pattern. The A→G mutation is a common RNA-editing mechanism arising from A→I editing acted upon by Adenosine Deaminase Acting on RNA (176). We summarize the list of factors that may lead to inconsistencies in detecting SNVs in RNA-Seq versus WES in Table 4.5.

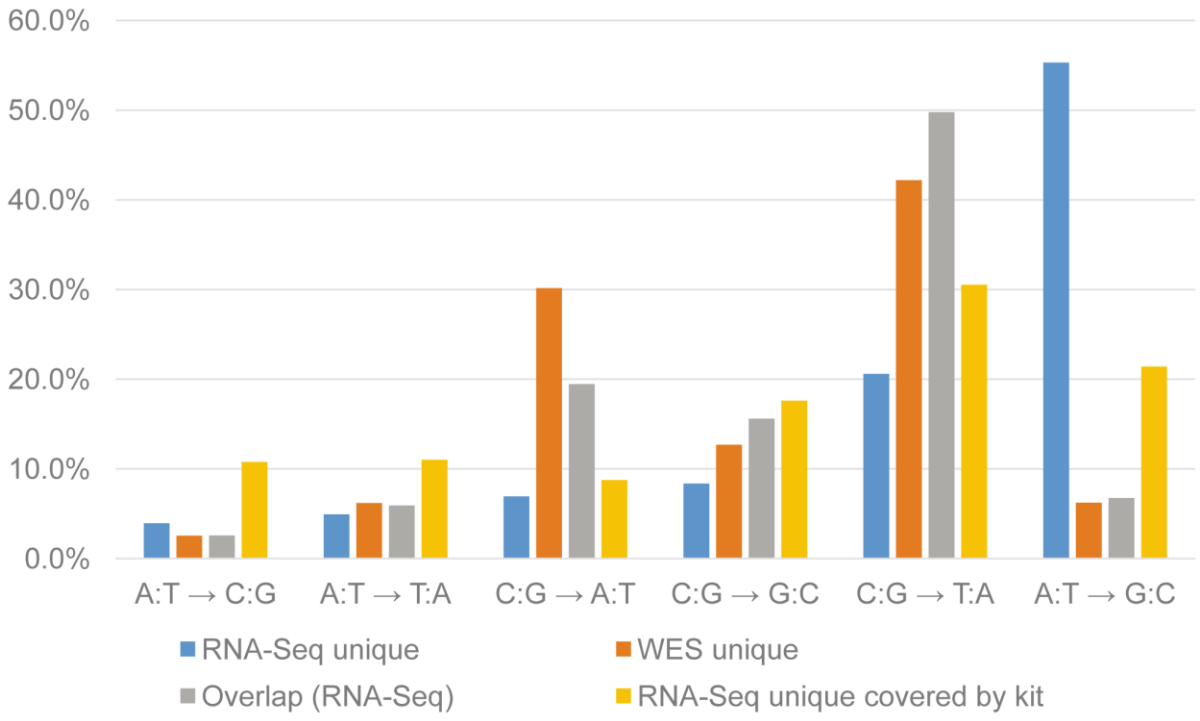


Figure 4.6. Mutation pattern for all SNVs. Mutation pattern was determined for all categories of SNVs and percentages plotted. Several patterns are more highly enriched than others, such as the A:T→G:C mutation in RNA-Seq.

Table 4.5. Summary of factors that may lead to inconsistencies in detecting SNVs in WES versus RNA-Seq.

Factors causing RNA-Seq unique SNVs	Observation	Factors causing WES unique SNVs	Observation
SNVs outside of the WES capture regions	71.4% of RNA-Seq unique SNVs	Low coverage of SNVs in RNA-Seq	41.0% of WES-unique SNVs have no RNA-Seq coverage
Low coverage of SNVs in WES	8.0% of RNA-Seq unique SNVs that are within the WES regions, have low or no WES coverage	SNVs located in non-expressed genes (< 1 FPKM)	51.0% of WES-unique SNVs
RNA-editing	55.3% of RNA-Seq unique SNVs were A:T→G:C mutations	SNVs potentially edited in RNA-Seq	55.3% of RNA-Seq unique SNVs were A:T→G:C mutations

Discussion

Few studies have examined mutation detection from both WES and RNA-Seq data of the same samples. However, such information is critical in assessing the mutations at different biological stages as well as their effects on disease. In this study, our comparison of WES and RNA-Seq data from the 27 pairs of NSCLC tumor and matched normal samples revealed that on average only ~14% of SNVs overlap. This value is quite low considering that the samples are identical. Thus, we explored possible reasons that cause this small overlap. We found that many of the WES unique SNVs are not called in RNA-Seq because they are poorly covered in RNA-Seq with many SNVs mapping with less than eight reads. This information is important for using a

SNV-calling software tool like MuTect, where the coverage limitations allowed to call an SNV is at least 14 reads in the tumor and at least 8 in the normal.

We noticed that although low coverage levels explained why most WES unique SNVs were not detected in RNA-Seq, many other SNVs had high read counts values but were still missed. We decided to interrogate gene expression levels to determine if this may explain why some SNVs are not detected in RNA-Seq. We used FPKM values, and found that the majority of WES unique SNVs are located in genes which are not expressed. In contrast, the SNVs that were shared between sequencing methods were found to have moderate to high expression levels. This is an important finding, because many studies use WES as the single method for somatic mutation detection in cancer, and this analysis demonstrates that it is important to measure expression levels when trying to determine deleterious variants. Many SNVs may be called in WES, but may not have an impact at the biological level because the variant is located within a non-expressed gene.

After determining these potential causes for the WES unique SNVs not being called in RNA-Seq, we next focused on the reasons why RNA-Seq unique SNVs were missed by WES.

We first thought that many of the RNA-Seq unique SNVs may be missed by WES because they fall outside of the WES capture regions. This is an important aspect to consider, because while RNA-Seq covers the whole transcriptome, WES is limited to detecting variants in the exons and their flanking regions. Currently, many exon capture kits are designed to have their probes covering well-annotated coding genes using representative gene models like Consensus CDS (CCDS) and RefSeq. And the capture method using target-probe hybridization has the limitation of GC-content bias. To compare the regions covered by the kit with the RNA-Seq, we used Bedtools and found that 71.4% of the RNA-Seq unique SNVs are not covered in WES. This suggests that many potentially important SNVs not located in exome regions would be missed if

WES were applied. This is becoming more important as ENCODE data has determined that many non-exonic regions in the genome are expressed, and that they may be playing important roles in gene regulation (193). It also implies that only performing WES on a tumor sample may miss potential variants that may be of important function.

Another biological reason why RNA-Seq unique SNVs may be missed in WES is due to RNA editing. We used the output from MuTect to generate the mutation pattern for all samples. An interesting mutation pattern in RNA-Seq unique SNVs was A:T→G:C. This pattern is indicative of RNA editing occurring by deamination of the adenosine to inosine, which gets interpreted as a guanine, editing in RNA achieved by the Adenosine Deaminase Acting on RNA proteins (176). This result has two implications for tumor sequencing. First, there may be a defect in the RNA editing machinery that leads to over-editing occurring in loci that normally do not get edited. Studies have shown that increased and decreased levels of RNA-editing may occur in different types of cancer (194, 195). This editing may give rise to new functions, or lose functions of important proteins in the tissue of interest. These mutations would be completely missed if sequencing were only focused on the whole genome or whole exome. Second, these mutations edited at RNA level are not expected to be detected by WES or WGS; therefore, their potential causative or deleterious effects will remain hidden.

Although we discovered many important differences between variants detected in WES versus RNA-Seq, there are some limitations to the interpretation of the results. Our samples were exclusively from tumor material, and we focused on an important category – somatic SNVs. It will be interesting to see if these results are similar for non-tumor tissue and germline mutations. We only used a total of 27 pairs of samples, and while this is large number and adequate for this analysis, it may miss some important conclusions. Furthermore, while the number of reads per

sample in our RNA-Seq is large, it is not sufficient enough for RNA splicing analysis. Experimental validation of RNA editing variants is also required. Finally, the SNVs called in each tumor type and sequencing type vary widely, so a pan-cancer study may identify additional reasons for the small overlap of variants detected in WES versus RNA-Seq.

In conclusion, our systematic comparison of SNVs from WES and RNA-Seq NSCLC data revealed a low overlap. We pinpointed multiple reasons for the inconsistencies in SNV detection with RNA-Seq and WES. It was discovered that most WES SNVs were undetected by RNA-Seq because of low coverage or low expression levels. We found that most SNVs detected by RNA-Seq were missed in WES because they are located outside the boundary of the WES capture regions. Lastly, we found that many SNVs detected by RNA-Seq had a mutational signature of RNA editing. This analysis has provided answers to our original posed question above about the feasibility to detect WES level SNVs using only RNA-Seq. Although we found that many variants would be missed using RNA-Seq alone, many of them may be of less importance because they are not expressed at high enough levels to cause damage. However, the SNVs detected by RNA-Seq may have potentially undergone RNA-editing and therefore would be difficult to target in DNA due to the converted base change at the RNA level.

CHAPTER V

APPLICATION OF THE GWAS-BASED REGULATORY PIPELINE AND APPROACH TO OTHER DISEASE TYPES

Introduction

Chapter II illustrated that it is possible to identify a set of regulatory variants and their target genes in lung cancer using several tissue-specific data sources and methods. However, that approach can be extended to other disease types beyond lung cancer and other populations besides European. In this chapter, we applied the approach developed in Chapter II to another disease of the lung, other cancer types, and other instances of lung cancer in non-European populations using SNPs with $p < 1 \times 10^{-4}$ from three GWA studies.

We selected a lung related disease that sometimes co-occurs with lung cancer: chronic obstructive pulmonary disease (COPD). COPD was originally defined as a disease that encompassed emphysema and chronic bronchitis, but recent efforts have determined it is a much more complex disease. COPD patients usually present with shortness of breath and other symptoms of lung dysfunction (196). There are several known risk factors for COPD, but like lung cancer, the most well-known risk factor is cigarette smoking (197). COPD is currently diagnosed when patients have a post bronchodilator Forced Expiratory Volume in one second (FEV-1) to Forced Vital Capacity (FVC) ratio < 70% (198). We use this lung disease to demonstrate our pipeline is applicable to lung diseases other than cancer. We also illustrate the applicability of the

approach to lung cancer in an Asian population of never smoker women. Interestingly, ~50% of lung cancer cases in women occurred in never smokers worldwide (199). The risk of never smoking lung cancer is especially high in Asian countries where many women cook with traditional stoves that emit toxic fumes. Overall, there are many differences between risk factors and the biology for smoking versus never smoker lung cancer (200). For an extensive discussion of this subject, see the review from Sun *et al*. in Nature Reviews Genetics 2007 (200). These data allow us to extend our pipeline to lung cancer identified in non-European populations to identify regulatory mechanisms of disease. Finally, we extend our approach to two histologically different, but anatomically close, cancer types: gastric cancer (GC) and esophageal squamous cell carcinoma (ESCC). These cancer types are usually combined and categorized as upper gastrointestinal cancers. The highest incidence of these cancer types occur in areas of China, although rates are decreasing (201). Interestingly, although alcohol consumption and cigarette smoking are major risk factors for these cancers in the west, they have decreased influence for Chinese populations (202). This suggests a different mechanism of disease for these cancers between different populations. Additionally, the different environmental exposures between these populations may influence risk of these cancer types. For example, dietary factors, such as the consumption of moldy bread (203) that is consumed in Linxian China may be a stronger constituent of disease in China versus the west.

In this study, we applied our approach from Chapter II on three different GWA studies to highlight the pipeline's generalizability to other disease types. The three studies consisted of five different disease types. Three of the diseases are lung related: COPD, never smoking LUAD (N.S. LUAD) and never smoking LUSC (N.S. LUSC). The other two diseases, ESCC and GC, expand the usefulness of our approach to non-lung related cancer subtypes. We also looked at the overlap

between many of the diseases and cancer subtypes. We give a brief overview of each study below and our combined results.

Methods

Datasets

COPD GWAS dataset

Pillai *et al.* (204) performed the first GWAS for COPD using subjects from a case-control study for the Bergen cohort in Norway. Cases and controls were required to have at least 2.5 pack-years of smoking history. COPD cases were defined by a post-bronchodilator FEV_1 of < 90% and $FEV_1/FVC < 0.7$. Controls were defined by $FEV_1 > 80\%$ and $FEV_1/FVC > 0.7$. After quality control (QC) of the samples, 823 cases and 810 controls of European ancestry remained. The subjects were genotyped with Illumina's HumanHap550 chip. After QC of the SNPs, there were 538,030 SNPs left for the association analysis. The authors analyzed the data using a logistic regression model including age, sex, smoking status, pack-years, and 12 principal components. We used the set of significant SNPs ($p < 1 \times 10^{-4}$) from the discovery phase of the GWAS in Supplementary Table 1 of the Pillai publication for our analysis.

Lung cancer in never smoking women GWAS dataset

Lan *et al.* (26) performed a GWAS for never-smoking women of Asian ancestry. This GWAS used subjects from 14 smaller studies that were scanned at six different centers. The genotypes were combined using a clustering approach described in their publication (26). The authors performed QC of the genotyped data to remove samples and SNPs that did not reach their QC criteria. After QC, 5,510 cases and 4,544 controls for the discovery analysis with 512,226 SNPs remained. The authors analyzed the data using logistic regression including age, study group, and eigenvectors. The authors also stratified their analysis based upon histology. We used all SNPs ($p < 1 \times 10^{-4}$) from the study's discovery phase for LUAD and LUSC. We obtained the SNP results from dbGaP study ID phs000716.v.1.p1. To download the SNP results separated by subtype, we utilized the Analyses tab on the webpage. This tab gave results in the online browser for LUAD http://www.ncbi.nlm.nih.gov/projects/SNP/gViewer/gView.cgi?aid=3852&pvf=0 and for LUSC http://www.ncbi.nlm.nih.gov/projects/SNP/gViewer/gView.cgi?aid=3853. To download the data, we chose no log(P-value_filter) and clicked to download the displayed data. We did this same procedure for each subtype to download all SNPs associated with each subtype. We further filtered this list to include SNPs $p < 1 \times 10^{-4}$ using the statistics analysis software R (205).

GWAS datasets for gastric cancer and esophageal cancer

Abnet *et al*. (206) performed a GWAS for GC and ESCC in an ethnic Chinese population. For the discovery phase, the authors used participants from two studies: the Shanxi Upper Gastrointestinal Cancer Genetics Project (Shanxi) and the Linxian Nutrition Intervention Trial (NIT). After QC of the genotyping and subjects, 3,523 cases and 2,100 controls for the association

analysis and 551,152 SNPs remained. The SNPs were analyzed using logistic regression including age, study, and sex. We obtained all significant SNPs from dbGaP study ID phs000361.v1.p1. To download the SNP results for both cancer types, we clicked on the Analyses tab, and under the Analyses folder, we chose both cancer types. The results were displayed in the online browser, and we downloaded all unfiltered results. We then filtered the list to include SNPs $p < 1 \times 10^{-4}$ using R.

Methods to obtain final germline-regulated genes

The same approach and datasets used in this chapter were extensively explained in the methods for Chapter II. Below, we briefly discussed data used and any modifications to the approach used in Chapter II. We do not go into depth about every method, as was done in Chapter II. We point the reader to the methods section of Chapter II for full details and explanation of the data sets used in this current chapter.

Remapping SNPs between genome builds and updating SNP rs ID numbers

The online tool Remap from NCBI (http://www.ncbi.nlm.nih.gov/genome/tools/remap) was used to remap SNPs between genome builds. For COPD, we remapped SNPs from hg18 to hg19 using default settings. The updated SNP positions were used to extract updated SNP rsID values using build 142 of dbSNP from NCBI. We used dbSNP files downloaded from ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b142_GRCh37p13/chr_rpts/. The SNPs for the other two disease types were downloaded from dbGaP (207), so we had to preprocess them first before we remapped. For example, data on dbGaP is structured differently than our data from

dbSNP and their positions are shifted by -1 bp relative to dbSNP. For N.S. LUAD and N.S. LUSC, we added +1 bp to each position before we remapped. The new positions were then remapped from hg38 to hg19. These new positions were updated to new rsID values as explained previously. For GC and ESCC, we also added +1 bp to their position. However, these SNPs were already in hg19, so we just matched the new positions directly to the files from dbSNP.

Generation of SNPs in LD for all diseases

We used each SNP for all disease types to obtain a set of all SNPs in a 1Mb region upstream and downstream from each SNP using Tabix (104) version 0.2.5. For COPD, we obtained the SNP data from the European Super Population Group, and for all other disease types we used the Asian Super Population Group. All data were obtained from the 1000 Genomes Phase III data v5.20120502 (208). We used Vcftools (105) version 0.1.12b to convert the Tabix vcf files to plink-tped file format. PLINK version 1.07 (106) was used in combination with the 1000 Genomes population genotype data to extract all SNPs in LD using $r2 > 0.8$ within 1Mb of each genotyped SNP. We combined all SNPs in LD from PLINK for each disease type and removed any duplicated SNPs from the LD expansion.

GTEx eQTLs

We used the full set of human tissue-specific eQTLs version 6 (V6) that was downloaded from the GTEx website (www.gtexportal.org) on February 22, 2016 (67). For full details of how the eQTLs were generated, see methods in Chapter II. We used lung tissue for COPD and the never smoking lung cancer samples. For ESCC, we combined the eQTLs found in two esophageal

tissues: Esophagus Mucosa and Esophagus Muscularis. For GC, we used stomach tissue. We also used multi-tissue eQTLs generated by GTEx in lung tissue for COPD, N.S. LUAD, and N.S. LUSC. There were no appropriate tissues to use for GA or ESCC in the multi-tissue analysis. We plotted the distribution of posterior probabilities for lung tissue eQTLs to determine the threshold for significance.

Hao et al. lung eQTLs (110)

The full set of cis-eQTLs in lung tissue with FDR at 10% identified from this study was used to identify eQTLs in COPD, N.S. LUAD, and N.S. LUSC.

FANTOM5 transcribed enhancers

We used the entire set of permissive (all identified) enhancers downloaded from http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed on August 26, 2015. We used PLINK v1.07 (106) to find any SNPs within enhancer regions using the gene-report function. To find the target gene of the enhancer regions with SNPs, we used the FANTOM5 enhancer transcription start site's associations downloaded from http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed on August 25, 2015. Enhancer target genes were determined for every disease studied.

IM-PET predicted enhancer target genes

Data from a study by He *et al*. (72) was used to find enhancer target genes for COPD, N.S. LUAD, and N.S. LUSC. Two lung related cell lines, IMR90 and NHLF, were used. Target genes with Reads Per Kilobase per Million mapped reads (RPKM) = 0 were removed, and Ensembl transcript IDs were mapped to gene symbols using BioMart (112).

Results

Description of data

We obtained SNPs ($p < 1 \times 10^{-4}$) from the discovery phases for three GWA studies. The first dataset was obtained from the first GWAS for COPD (204). This GWAS for COPD was performed in Norway using the Bergen Cohort. All cases and controls were current or former smokers. This GWAS contained 823 cases and 810 controls of European ancestry. A summary of the participants is listed in Table 5.1. The second GWAS dataset was obtained for lung cancer cases in never smoking women of Asian descent (26). This GWAS was performed using data from 14 different studies. There were 5,458 cases and 7,457 controls of Asian ancestry. A summary of the participants is listed in Table 5.2. The third GWAS dataset was obtained for GC and ESCC in an ethnic Chinese population (206). This GWAS was performed using data from two studies in China and contained 3,523 cases and 2,100 controls of Asian ancestry. A summary of the participants is listed in Table 5.3.

Table 5.1. Summary of GWAS for COPD.

	# Cases	# Controls
Participants	823	810
Post-FEV_1 in liters (±SD)	1.59 (±0.71)	3.25 (±0.74)
Post-FEV_1 % pred (±SD)	50.26 (±17.33)	93.91 (±9.22)
Post-FEV_1/FVC ratio (±SD)	0.52 (±0.13)	0.79 (±0.04)
Population	European ancestry	European ancestry

Adapted from Pillai *et al*. *PLoS Genetics* 2009 (204).

Table 5.2. Summary of GWAS for never smoking women in Asia.

Study	# LUAD samples	# LUSC samples	# Controls
CAMSH	555	32	334
FLCS	212	49	386
GDS	535	7	123
GEL-S	120	8	296
GELAC	1,059	75	1,095
HKS	226	0	666
JLCS	407	10	549
SKLCS	419	28	1,082
SLCS	378	98	1,024
CNULCS	498	51	480
SWHS	78	9	200
TLCS	49	32	237
WLCS	0	14	343
YLCS	179	330	642
Total	**4,715**	**743**	**7,457**

Adapted from Supplementary Table 1. Lan *et al*. *Nature Genetics* 2012 (26).

Table 5.3. Summary of GWAS for GC and ESCC in ethnic Chinese.

Study	# GC	# ESCC	# controls
Shanxi	1,368	1,399	1,650
NIT	257	499	450
Total	**1,625**	**1,898**	**2,100**

Adapted from Supplementary Table 1. Abnet *et al*. Nature Genetics 2010 (206).

Remapping SNPs to an updated genome and LD expansion

We remapped the SNPs for each disease to hg19 (see Methods) and updated the SNP rs ID numbers in the new genome build using dbSNP b142. The datasets are from two different population types, so first we used these hg19 SNPs, and data from the 1000 Genomes Phase III European Super Population, to expand our COPD SNP list to include all SNPs in LD ($r^2 > 0.8$) within 1Mb of each updated SNP. We used the 1000 Genomes Phase III Asian Super Population data for N.S. LUAD, N.S. LUSC, GC, and ESCC to obtain all SNPs in LD ($r^2 > 0.8$) within 1Mb of each SNP. The results from the LD expansion are listed in Table 5.4. We used the same pipeline as in Chapter II to obtain a set of functional SNPs and their target genes for the lung related diseases. We modified the pipeline for the gastric related diseases by removing the SNP mapping steps for the Hao *et al*. eQTLs, IM-PET enhancers, and the GTEx multi-tissue eQTLs.

Table 5.4. Summary of LD SNP expansion for all SNPs.

SNP category	COPD	N.S. LUAD	N.S. LUSC	GA	ESCC
SNPs genotyped	538,030	512,226	512,226	551,152	551,152
SNPs from GWAS $p < 1 \times 10^{-4}$	58	95	69	61	98
SNPs in LD r2 0.8 within 1 Mb	1,341	3,195	2,278	2,084	2,594
Duplicated SNPs	494	1,276	945	519	1,251
Final SNPs	**847**	**1,919**	**1,333**	**1,565**	**1,343**

Regulatory variants for all disease types

GTEx single tissue eQTLs

We first determined regulatory variants and their target genes from the data sources that contained tissue types other than lung and could be broadly used in the other disease types. Initially, we used single tissue eQTLs generated from the GTEx project (66) to identify regulatory SNPs. We used the lung tissue eQTL results and identified a set of SNPs that acted as eQTLs for COPD (n = 8), N.S. LUAD (n = 2,400), and N.S. LUSC (n = 52). There are two ESCC related tissues with eQTLs: esophagus mucosa (EMC) and esophagus muscularis (EMS). We used these two tissue types and discovered 438 eQTLs and 560 eQTLs for EMC and EMS, respectively. For GC, we used stomach tissue and found 129 eQTLs (Figure 5.1A). The eQTLs found above may be acting to control the same gene, so we collapsed all eQTLs to the genes they controlled for the final gene sets.

FANTOM transcribed enhancers and their target genes

We next determined the SNPs from each disease type that were located within enhancer regions of the genome with associated target genes. We used the FANTOM data (73) for the enhancer definitions. We used the set of permissive enhancers and their correlated transcribed target genes from the Promoter Enhancer Slider Selector Tool (PrESSTo) website (119). We discovered the number of enhancers with an associated target gene for each disease type. For the lung diseases, we found 1, 104, and 6 enhancers with target genes, for COPD, N.S. LUAD, and

N.S. LUSC, respectively. For the non-lung diseases, we found 2 enhancers with target genes for GC and 8 enhancers with target genes for ESCC (Figure 5.1B).

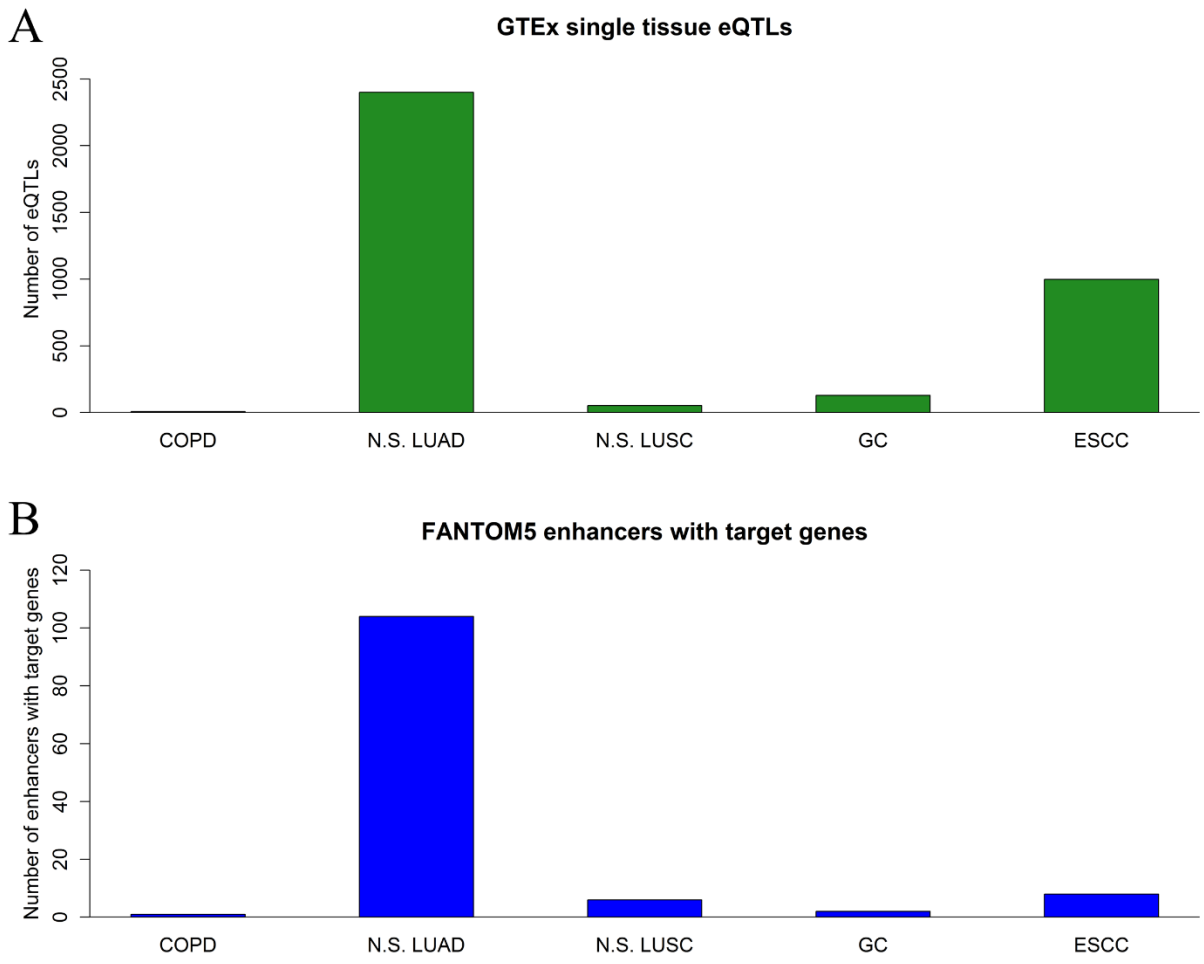


Figure 5.1. Regulatory elements discovered in all diseases. Panel A shows the total number of single tissue eQTLs discovered in each disease type. We combined the eQTLs for ESCC found in two esophageal tissues: EMC and EMS. Panel B shows the number of enhancers with target genes that contained SNPs from each disease type.

Regulatory variants for lung diseases

GTEx multi-tissue eQTLs

Next, we determined the additional regulatory information that could be obtained using the lung related diseases. Although, as we demonstrated above, our approach can be used across different cancer types, it was designed for a lung disease, and therefore additional data can be applied to the lung diseases. The GTEx project identified a set of multi-tissue eQTLs in addition to the single tissue eQTLs (for full details see Chapter II). There are no GC or ESCC related tissues used in the multi-tissue analysis; therefore, we only used the lung tissue results for COPD, N.S. LUAD, and N.S. LUSC. The eQTLs identified in the multi-tissue approach differed from the single tissue approach (see Chapter II, Methods), so we plotted the distribution of each set of eQTLs to determine a significance threshold. Based upon the distributions, we selected posterior probabilities of 0.7 for COPD and N.S. LUAD and 0.8 for N.S. LUSC. As illustrated in Figure 5.2A, we found 40, 1310, and 101 multi-tissue eQTLs for COPD, N.S. LUAD, and N.S. LUSC, respectively. Similar to the single tissue eQTLs, many of these eQTLs controlled the same target gene, so we combined all results to their unique target genes for the final gene sets.

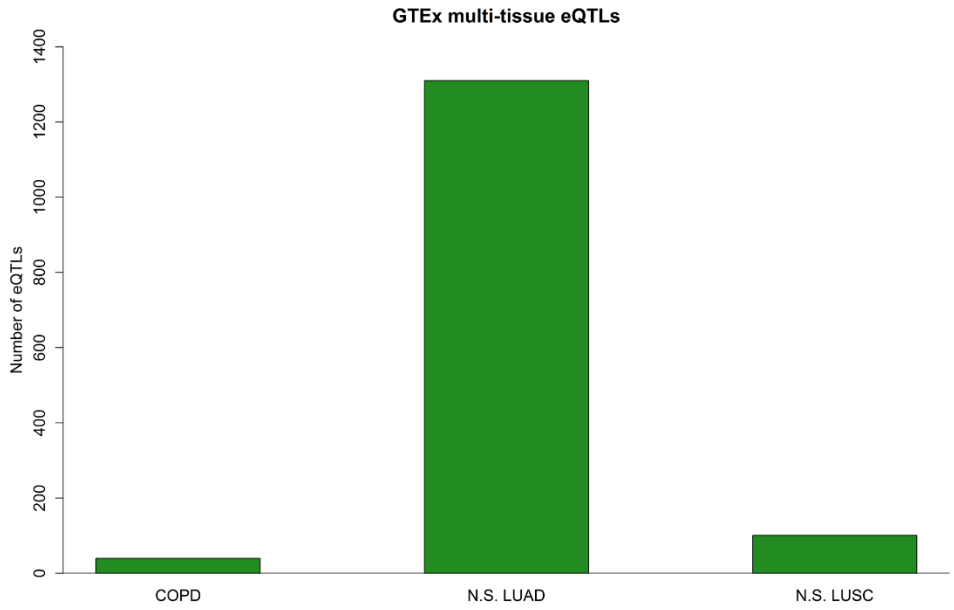
Hao et al. lung tissue eQTLs

We used a third source of lung eQTLs for our final eQTL dataset for the lung diseases. We used the lung tissue eQTLs derived in the study by Hao *et al*. (110) that is described in Chapter II. We only discovered two eQTLs for COPD and N.S. LUSC and only one eQTL for N.S. LUAD.

Epigenetically defined enhancers and their predicted target genes

We used the data generated using the software IM-PET (72) for two lung related cell lines: IMR90 and NHLF. We first used the IMR90 cell line data and discovered 5, 6, and 4 target transcripts for COPD, N.S. LUAD, and N.S. LUSC, respectively. We found a greater number of target transcripts using the NHLF results for each disease type. Specifically, we found 8 for COPD, 13 for N.S. LUAD, and 8 for N.S. LUSC. We plotted these results in Figure 5.2B.

A



B

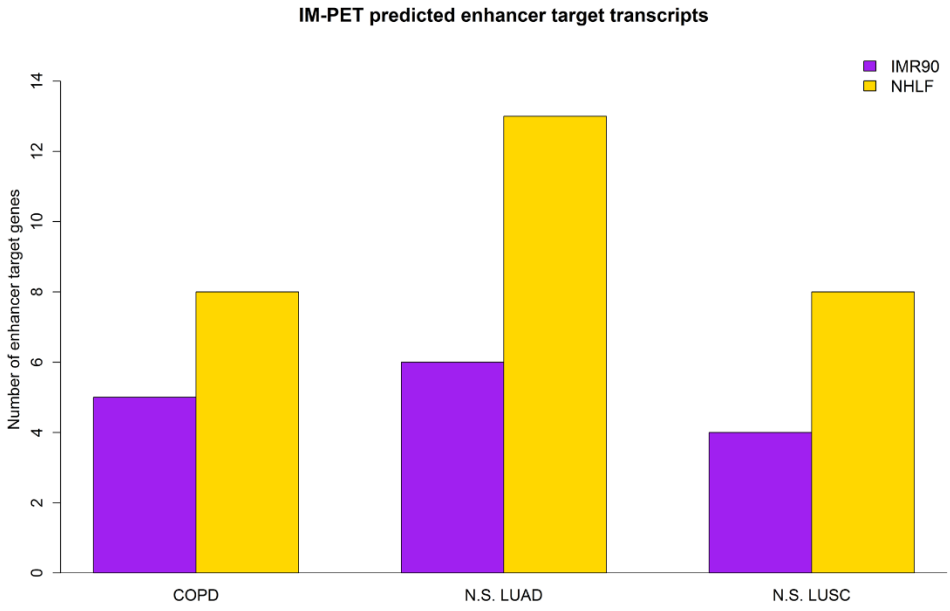


Figure 5.2. Total number of regulatory elements for lung related diseases. In panel A, we show the number of lung tissue eQTLs found from the multi-tissue analysis. For panel B, we show the number of enhancer target genes for each lung related disease colored by cell type.

Little overlap between the different histological cancer types

We next determined the overlap between diseases by utilizing the target genes discovered in the regulatory data sets. We first determined the overlap between all cancer types using the GTEx single tissue and FANTOM5 data. We used all four cancer types because the two non-lung related cancers are the same histological class as the never smoking lung cancer subtypes. Histologically, the GCs are adenocarcinomas like LUAD, and the ESCCs are squamous cell carcinomas like LUSC. Surprisingly, we discovered there were no single tissue eQTLs or FANTOM enhancer targets that were shared among any of the four cancers. We originally hypothesized we would see sets of eQTLs and enhancers shared in histologically similar cancer types based upon previous evidence that showed some cancers are more similar by histological subtype than by tissue-specific cancer type (209). Solely for the lung related diseases, we also examined their overlap using the multi-tissue eQTL target genes (Figure 5.3A), and the IM-PET combined enhancer target genes (Figure 5.3B). Using the GTEx multi-tissue eQTL target genes, we found only one gene that overlapped between COPD and N.S. LUAD and no other genes that overlapped between the lung diseases. Our results also indicated that there were no genes from IM-PET that overlapped between any of the lung diseases. We did not illustrate the Hao *et al*. eQTL overlap because there are only one or two genes identified for each disease and they do not overlap.

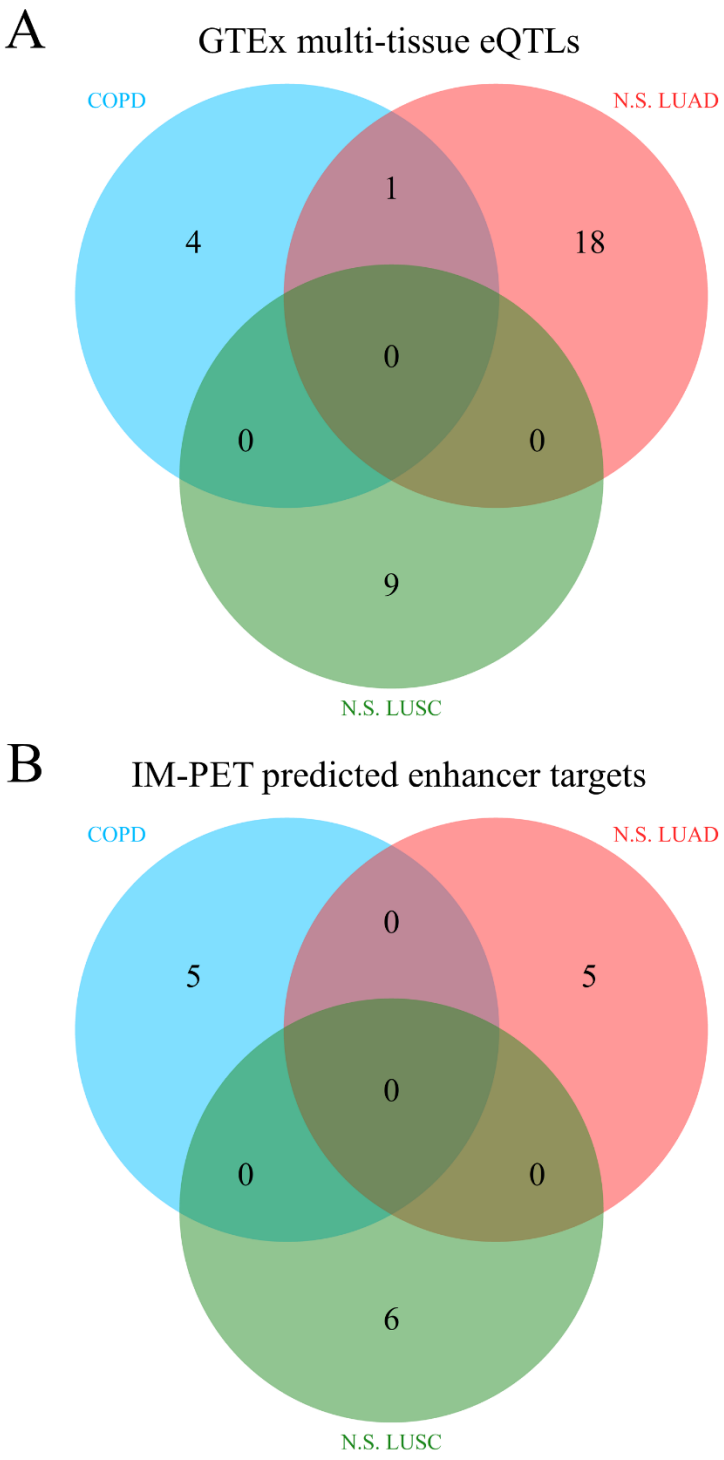


Figure 5.3. Overlap between lung-related diseases for multi-tissue eQTLs and predicted enhancer targets. Panel A shows the overlap between lung tissue eQTLs from GTEx that are active in multiple tissue types. Panel B shows the overlap between IM-PET predicted enhancer targets that were combined for two lung cell lines: IMR90 and NHLF.

Generation of final germline regulated genes for each disease

We combined all results from the datasets for each disease type, and removed duplicated genes that were discovered using more than one dataset, to obtain the final germline-regulated genes. The summary of the final germline-regulated genes per data source is listed in Table 5.5. We also determined the overlap between all diseases using these germline-regulated genes (Figure 5.4). Overall, our final germline gene results indicated very little overlap between disease types.

Table 5.5. Summary of final germline-regulated genes.

Disease	GTEx S.T. eQTLs	GTEx M.T. eQTLs	Hao *et al*. eQTLs	FANTOM enhancer targets	IM-PET combined	**Germline-regulated genes**
COPD	1	5	2	1	5	**14**
N.S. LUAD	16	19	1	29	5	**48**
N.S. LUSC	4	9	2	4	6	**23**
GC	8	NA	NA	1	NA	**9**
ESCC	28	NA	NA	6	NA	**34**

S.T. = single tissue. M.T = multi-tissue.

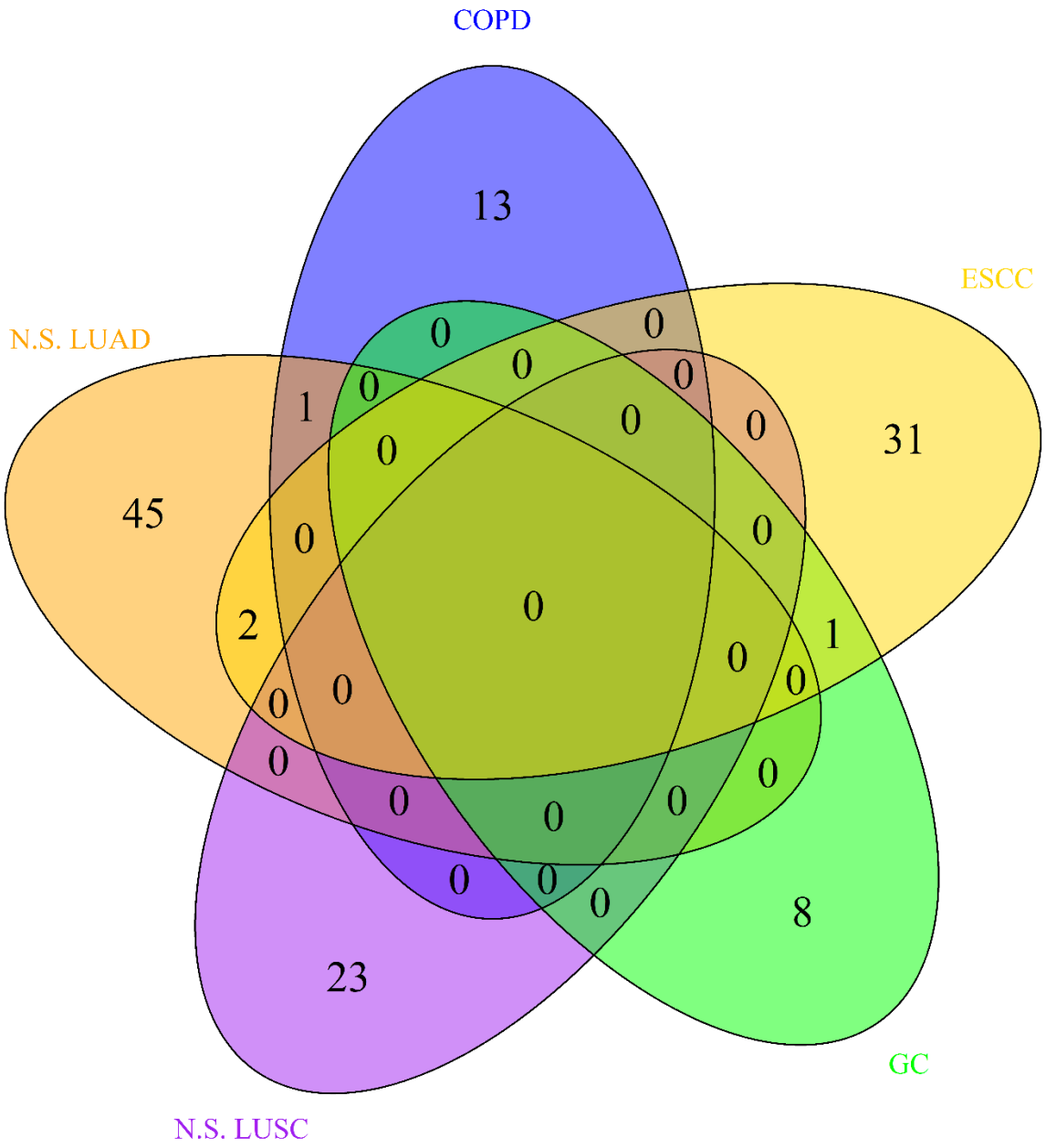


Figure 5.4. There is little overlap between all germline-regulated genes for each disease.

Discussion

In this chapter, we demonstrated that we can apply the approach developed in Chapter II for lung cancer to other diseases and ethnicities. Although we could not use all of the datasets on the non-lung related diseases, we still identified a large set of germline-regulated genes (Table 5.5). For the lung related diseases, we discovered a set of germline-regulated genes using all of the datasets from Chapter II.

We first identified a set of target genes using the single tissue eQTL results from GTEx. Surprisingly, we observed no overlap between the eQTLs identified in each cancer type with any other cancer type. Also, we did not observe any single tissue eQTLs that overlapped between COPD and the two never smoking lung cancer subtypes. These results are intriguing because it suggests that even though these subtypes are of the same histology (GC and N.S. LUAD, ESCC and N.S. LUSC), or the same cancer type (N.S. LUAD and N.S. LUSC), they may not share common regulatory mechanisms at the single tissue eQTL level. Next, we determined the overlap between all cancer subtypes and their FANTOM5 defined enhancer target genes. Similar to the eQTL results, we did not observe any overlap between any subtypes. This lack of overlap between regulatory elements is surprising because in Chapter II, we observed at least some overlap between lung cancer subtypes. However, in Chapter II, we had many more eQTLs and target genes to compare.

Next, we focused on the lung specific diseases where we could obtain more results using our lung-specific regulatory datasets. First, we used the GTEx multi-tissue eQTL data to find lung tissue eQTLs that are active in multiple tissues. We determined the overlap between the three lung diseases, and we found that one gene overlapped between COPD and N.S. LUAD: *LINC01137*.

Second, we identified predicted enhancer targets from two lung related cell lines using IM-PET. We discovered that no target genes were shared between any of the lung diseases. This is an interesting result and, combined with the lack of overlap of enhancers from FANTOM, may indicate separate regulatory programs for each lung disease.

Finally, we determined the overlap between all diseases using the final sets of germline-regulated genes. This comparison revealed some interesting overlapping genes. For example, we found that ESCC had two genes, from a single genomic region, *HLA-DQA1* and *HLA-DQB2*, which overlapped N.S. LUAD and none that overlapped N.S. LUSC. This is surprising because these two subtypes are of different histologies. We hypothesized that we would see overlap between different cancer types of the same histology, but we did not observe that with these results. Intriguingly, we found one gene that overlapped between GC and ESCC: *NOC3L*. The original publication by Abnet *et al*. (206) reported one gene that was shared between both subtypes: *PLCE1*. This finding was reported because of SNPs within the genic region. However, our analysis also discovered several SNPs within *PLCE1*, but they are eQTLs for a more distant gene *NOC3L*. This finding suggests that the original study may have reported the wrong affected gene that was shared between subtypes. This is one example of how we used the approach from Chapter II to discover new biology that may contribute to two non-lung cancer types.

Although we found several germline-regulated genes using our approach, there are several limitations. First, we did not identify enough genes to perform any more meaningful analyses, such as pathway analyses, as we did in Chapter II. We hypothesize this is because we used a more stringent p-value here ($p < 1 \times 10^{-4}$) instead of the cutoff used in Chapter II ($p < 1 \times 10^{-3}$). In future studies using this approach, it is advised to use the second p-value threshold for more meaningful results. Second, it is difficult to obtain a set of GWAS values at the above threshold. Although

many times the data can be requested from repositories such as dbGaP (207), it is more difficult to obtain than thresholds such as $p < 1 \times 10^{-4}$ that may be reported in supplemental tables (96). Third, we relied on datasets that have their own limitations. Although GTEx uses > 70 samples per tissue for their eQTL analyses, more results could be obtained with greater sample sizes. Therefore, we may have used incomplete sets of eQTLs. Additionally, we rely on regulatory features that are generated for either tissue-wide data or for cell-lines. Since many different cell types make up each tissue, there may be noise in the datasets from this heterogeneity of cells types. Fourth, we are limited in the application of our approach to non-lung related disease types. Although we found an interesting result with the GC and ESCC subtypes, we did not have the same sensitivity as the lung-related diseases because of the limitation with the datasets. However, if GTEx expands their multi-tissue analysis to more tissue types, future work can be completed outside of lung tissue.

In conclusion, we utilized the pipeline and approach from Chapter II on non-lung cancer GWAS results. We demonstrated that the approach identified regulatory variants and their corresponding germline-regulated in non-lung cancer samples. We identified a shared germline-regulated gene between GC and ESCC that may have been missed in the original study. We also discussed several limitations to our results. Overall, we discovered that our approach to the determination of regulatory variants from GWA studies and the identification of their target genes is applicable across disease types and population types, but caution remains. Approaches like this are needed to help unravel the unknown nature of many non-coding variants found in GWAS for many phenotypes, which currently is a challenging but important research topic.

CHAPTER VI

CONCLUSION

In this dissertation, I discussed my work studying the germline and somatic genomes in three subtypes of lung cancer, as well as some extended work for other cancer types and lung diseases. Although previous studies have investigated single lung cancer subtypes on a germline or somatic level (see Chapter I, Introduction), it is not well understood how to integrate these findings. Additionally, due to different study designs and methodologies, it has been difficult to systematically compare the different subtypes of lung cancer amongst each other. My aims in this thesis work were to perform exhaustive interrogations of both genomes across lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and small cell lung cancer (SCLC). In addition to studying each subtype alone and in one genome (somatic or germline), I determined overlaps that exist across genomes and subtypes. This integrative analysis approach is important because outside of environmental hazards, the causal factors behind the genetic basis of lung cancer remain largely unknown. Therefore, I used a combination of genetics and functional genomics to study both genomes.

In Chapter II, I developed novel approaches to interrogate non-coding variants associated with lung cancer. Specifically, I used a functional genomics approach to identify regulatory variants in each subtype. Interestingly, I found that on a single nucleotide polymorphism (SNP) level, the subtypes are quite distinct from each other. Through the identification of regulatory

variants and their target genes, I found that this lack of overlap extended beyond the SNP level and this observation will be discussed below.

In Chapter III, I used DNA sequencing (WES) and RNA-Sequencing (RNA-Seq) to explore the differences in the subtypes at the somatic level. This analysis also revealed an overall lack of overlap between subtypes, but not as severe as at the germline level. I used RNA-Seq data to generate a set of genes that were differentially expressed in the tumor-versus-normal tissue for each subtype. These differentially expressed genes showed different patterns of overlap in the subtypes based upon up-regulation or down-regulation. I also identified a set of mutational signatures in each subtype and sets of potential driver genes.

In Chapter IV, I explored the accessibility of calling somatic variants from RNA-Seq data in lung cancer and identified the biological and technical reasons for inconsistencies in calling these variants in WES versus RNA-Seq. In this chapter, I performed a deep investigation into the feasibility of generating a set of somatic single nucleotide variants (SNVs) in RNA-Seq that is equal to SNVs called in whole exome sequencing (WES). The results in Chapter IV provided a greater understanding of the limitations to using this approach to call variants. I found that RNA-Seq can be used to call variants, but many caveats exist that need to be carefully considered.

In Chapter V, I applied the approaches from Chapter II on GWA studies for other disease types and populations. In this chapter, I performed the same analysis that was done in Chapter II in order to demonstrate the transferability of the approach. I illustrated that the same approach can be applied, with some small changes for non-lung related diseases, to GWAS for many different diseases and ethnicities.

Below, I highlight some intriguing discoveries found in the overall analyses and their contribution to the research field of lung cancer.

Filling in the knowledge gap for GWAS variants

Many of the GWA studies introduced in Chapter I identified common variants located within non-coding regions of the genome. The interpretation of these variants remains challenging because it is not easy to identify the genes that may be affected by these variants. Additionally, past work (68) has demonstrated that non-coding GWAS variants are enriched for regulatory function. Lung cancer is one of these diseases where more significantly associated variants from GWA studies were in non-coding regions rather than in coding regions of the genome (52). The lack of insight into the biological mechanisms behind these variants has left many questions about the biology of lung cancer from the germline perspective unanswered. However, in Chapter II of this work, I used several lung-related eQTL and enhancer datasets to investigate the role of GWAS results for lung cancer and their LD SNPs in regulatory regions of the genome. I also determined the target genes for these regulatory SNPs (see Chapter II). These results indicated that some of these variants were acting in regulatory roles in the genome that control expression on one or several target genes. Specifically, I found that these variants act as SNPs in the eQTLs that control the gene's expression, known as expression SNPs (eSNPs), and are also located within enhancer regions of the genome. Additionally, the identification of the target genes of these regulatory SNPs generated a set of targets that could be explored in future experimental studies. These results suggested that many common variants associated with all three lung cancer subtypes are in control

of genes with already established roles in lung cancer, such as *CHRNA5*, *IDH3A*, and *PSMA4* which were identified as target genes of regulatory variants in every subtype.

The weak overlap between all three subtypes at the germline and somatic genomes

One discovery from this work was that across both genomes, there was a lack of overlap between all three lung cancer subtypes at the gene-level. Although histological analysis and cell type studies group these three lung cancers into distinct subtypes, they are all derived from lung tissue (210). Therefore, I originally hypothesized that there would be moderate overlap of the three subtypes. However, I discovered that across genomes and biological factors, many of the genes discovered in each subtype were not shared by the other two subtypes (see Chapter II and Chapter III).

I first observed the weak overlap between LUAD, LUSC, and SCLC in the germline genome using lung cancer GWAS results at $p < 1 \times 10^{-3}$. I found that only 10 SNPs (< 1%) overlapped all three subtypes even though each subtype had over 500 significant SNPs (see Chapter II). I further identified sets of germline genes that were the targets of the regulatory SNPs for each subtype and again found weak overall overlap and only observed one independent region on 15q25 that overlapped all subtypes. This region contained five germline-regulated genes that may be contributing risk for lung cancer independent of subtype.

I next observed the overlap between subtypes at the somatic level. First, I identified sets of DEGs that were down-regulated and up-regulated in somatic lung tumor tissue compared to normal lung tissue in each subtype. Although 554 up-regulated DEGs overlapped all three subtypes, there

were still 442, 850, and 575 unique up-regulated DEGs for LUAD, LUSC, and SCLC, respectively. Fewer down-regulated genes overlapped all three subtypes (325), but there were also fewer DEGs detected overall. For both sets of DEGs, only ~ 14% of all DEGs were shared between all three subtypes. At the somatic mutation level, I first identified the three most significant mutational signatures for each subtype. Only one mutational signature (COSMIC signature 4) was shared between the subtypes. This mutational signature is most associated with tobacco mutagenesis (149). However, LUSC had Signature 13 in common with LUAD. This signature has been found in many cervical and bladder cancer types (7), suggesting that these subtypes may share functions or regulation with other cancer types. Additionally, signature 5 was shared between LUAD and SCLC. This signature has been found in all cancer types and exhibits T>C substitutions (7). I also generated a set of potential driver genes, and found that at two different thresholds, I found many potential driver genes unique to each subtype and only ~13% of all potential driver genes overlapped all three subtypes (see Chapter III). Overall, although there was a lack of strong overlap in the somatic genome for all three subtypes, it was not nearly as strong as the overlap between all three subtypes at the germline level.

Linking acetylcholine receptors from the germline to somatic genomes

Several genomic studies (53, 116) have identified the region of 15q25 to be associated with multiple subtypes of lung cancer (for full discussion, see Chapter I). However, most of the variants were located within non-coding regions of the genome and in the vicinity of a closely related set of genes. These genes comprise a set of nicotinic cholinergic receptor (CHRNA) genes. Debate has ensued about this region's role in lung cancer risk from a genetic versus environmental

perspective (128). However, using the combined results of Chapters II and III, one can gain insight into this region's role in lung cancer. For example, I found several regulatory SNPs within this genomic region for all three subtypes. However, through my regulatory functional genomics approach, I identified the affected gene from these regulatory SNPs as *CHRNA5*. The results from the somatic DEG analysis further validated this finding. I discovered that *CHRNA5* is up-regulated 3.64, 3.22, and 2.52 log fold in LUAD, LUSC, and SCLC tumor tissue compared to normal lung tissue. This finding expands the previous knowledge about this gene's role at the germline level and extends it to the somatic level. To further validate *CHRNA5* as the probable target of common variants, I used the DEG results to investigate another CHRNA gene that has also been implicated at the germline level, *CHRNA3*. I searched the DEG results and found that although *CHRNA3* is up-regulated in SCLC, it is not significantly up-regulated in LUSC or LUAD. In this situation, it was crucial to compare across the subtypes rather than just studying SCLC because identifying *CHNRA3* only in SCLC suggests that *CHRNA5* is probably the correct target. However, the possibility still remains that *CHRNA3* is also acting solely in SCLC. Although *CHRNA5* has not garnered much attention at the somatic level, this finding suggests it may be beneficial to explore the biology behind acetylcholine receptors (ACRs) in cancer in more detail.

Shared pathways across germline and somatic genomes (using gene sets)

I combined the pathways between the germline and somatic genomes in each subtype to identify biological pathways perturbed in both genomes. Ignoring any overlap between other subtypes, I looked at the number of pathways identified in each subtype from the germline genome and from the somatic genome. For the germline genome, I used the germline-regulated genes in

each subtype, and for the somatic genome, I used the combined up-regulated and down-regulated DEGs. I did not filter the pathways for the germline-regulated genes because I did not have very large gene sets. However, for the somatic DEGs, I filtered out pathways that contained less than five DEGs. I found that four pathways are enriched with genes from both genomes for LUAD. For LUSC, I found that 18 pathways overlap the germline genome and somatic genome. Finally, for SCLC, I observed that three pathways overlapped both genomes. These final pathways are listed in Table 6.1. Interestingly, many of these pathways do not have immediate relations with cancer. Therefore, future studies may look to these pathways as guidance for the exploration of biological pathways associated with each cancer type. The evidence from both genomes gives moderate expectations that these pathways may be important in each cancer subtype.

Table 6.1. Final overlap in enriched KEGG biological pathways shared in the germline and somatic genomes.

LUAD	LUSC	SCLC
Metabolic pathways	Staphylococcus aureus infection	Metabolic pathways
Tight junction	Asthma	Retinol metabolism
Endocytosis	Antigen processing and presentation	Focal adhesion
Focal adhesion	Graft-versus-host disease	
	Intestinal immune network for IgA production	
	Viral myocarditis	
	Leishmaniasis	
	Rheumatoid arthritis	
	Phagosome	
	Toxoplasmosis	
	Systemic lupus erythematosus	
	Cell adhesion molecules (CAMs)	
	Metabolic pathways	
	Jak-STAT signaling pathway	
	Pathways in cancer	
	NOD-like receptor signaling pathway	
	Complement and coagulation cascades	
	Hematopoietic cell lineage	

Future directions

For each part of this work there are several follow up analyses that could be explored. However, I think that the most interesting future studies would involve the possible biological interactions and relationships between the germline and somatic genomes in each lung cancer subtype. Although this work only highlighted a few aspects of how the germline and somatic genomes may work together in lung cancer, there are specific follow up studies that may reveal closer interactions.

One particular method that can be used is EW_dmGWAS (139). This is a network-based method that can be used to integrate both genomes. EW_dmGWAS uses the germline and somatic genomes with a protein-protein interaction (PPI) network to help identify sets of closely related genes from both genomes. This method uses GWAS genome-wide gene-level p-values to weigh each node in the network. It uses differential expression values at the somatic level to weigh the edges in the network. The algorithm identifies closely related modules in the network. This is one way that a network can be used to combine both genomes. Several iterations can be used for this analysis because many tools exist to generate gene-level p-values. For example, instead of using proximity to the SNP to generate a gene-level p-value that is done in tools such as Versatile Gene-based Association Study (VEGAS) (211), newer approaches such as MetaXcan (212) can be used. MetaXcan generates a gene-level p-value based upon predicted gene expression using quantitative trait loci (eQTLs) derived from the GWAS SNPs. Using these values in the network may generate more biologically accurate results since I hypothesized, and confirmed, that many of the lung cancer GWAS variants are regulatory (Chapter II).

Finally, future studies can expand upon the biological pathways discovered in this work. The pathways I discovered were identified using a set of genes that were already filtered due to their potential role in disease. However, this can be expanded upon to generate pathways identified from all SNPs in the GWAS (213) and all genes in the somatic genome (214). This larger unbiased approach may identify additional genes that are closely associated, and may function together, in both genomes.

Concluding remarks

This dissertation's deep interrogation into the molecular differences between three histologically distinct lung cancers suggested several potential shared and distinct mechanisms of disease. Although this dissertation focused on comparing the subtypes of lung cancer, these methods can easily be extended to other cancer types. This insight into biological differences between cancers that arise in the same organ, but of different cell types (210), will be of greater importance as sequencing technologies become more accessible and affordable to all patients (http://www.businesswire.com/news/home/20170109006363/en/), as we are entering to the era of the $100 per genome sequencing. New initiatives such as the Precision Medicine Initiative (PMI) (https://www.nih.gov/research-training/allofus-research-program) will generate large diverse genomic datasets that will help to unravel the mysteries and inner workings of the genome.

APPENDIX

Appendix A. Locus level analyses for germline-regulated genes discovered in Chapter II.

A.1. Independent locus level analysis for LUAD unique.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	41	14	25
GTEx multi-tissue eQTL in lung > 0.80	43	18	28
GTEx combined	66	18	37
Hao single tissue lung eQTL	23	12	18
FANTOM5 enhancer target gene	31	12	15
IM-PET IMR90	43	12	20
IM-PET NHLF	80	15	28
IM-PET combined	105	16	34
All genes	193	21	69

A.2. Independent locus level analysis for LUSC unique.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	47	14	23
GTEx multi-tissue eQTL in lung > 0.80	72	18	36
GTEx combined	95	18	42
Hao single tissue lung eQTL	26	9	17
FANTOM5 enhancer target gene	63	13	16
IM-PET IMR90	59	16	27
IM-PET NHLF	110	16	28
IM-PET combined	150	18	37
All genes	287	19	71

A.3. Independent locus level analysis for SCLC unique.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	29	13	25
GTEx multi-tissue eQTL in lung > 0.80	45	14	30
GTEx combined	56	16	37
Hao single tissue lung eQTL	10	6	10
FANTOM5 enhancer target gene	30	12	15
IM-PET IMR90	32	13	21
IM-PET NHLF	56	13	29
IM-PET combined	77	16	38
All genes	154	20	69

A.4. Independent locus level analysis for LUAD overlap LUSC.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	0	0	0
GTEx multi-tissue eQTL in lung > 0.80	2	2	2
GTEx combined	2	2	2
Hao single tissue lung eQTL	1	1	1
FANTOM5 enhancer target gene	5	2	2
IM-PET IMR90	2	2	2
IM-PET NHLF	7	3	3
IM-PET combined	7	3	3
All genes	16	6	6

A.5. Independent locus level analysis for LUAD overlap SCLC.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	0	0	0
GTEx multi-tissue eQTL in lung > 0.80	0	0	0
GTEx combined	0	0	0
Hao single tissue lung eQTL	0	0	0
FANTOM5 enhancer target gene	0	0	0
IM-PET IMR90	0	0	0
IM-PET NHLF	1	1	1
IM-PET combined	1	1	1
All genes	1	1	1

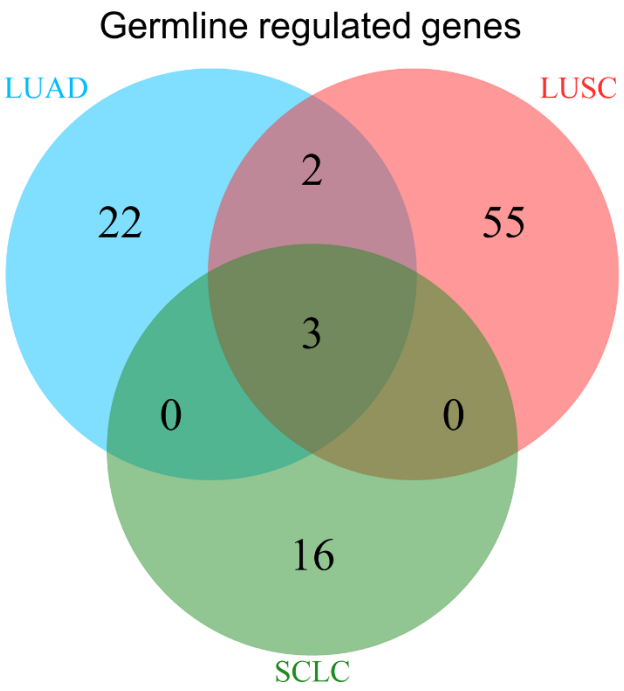
A.6. Independent locus level analysis for LUSC overlap SCLC.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	4	2	2
GTEx multi-tissue eQTL in lung > 0.80	4	2	2
GTEx combined	8	3	3
Hao single tissue lung eQTL	2	1	1
FANTOM5 enhancer target gene	0	0	0
IM-PET IMR90	1	1	1
IM-PET NHLF	2	1	1
IM-PET combined	2	2	2
All genes	12	5	5

A.7. Independent locus level analysis for ALL OVERLAP.

Regulatory category	# unique genes	# chromosomes	# total unique regions
GTEx single tissue V6 Lung eQTL	2	1	1
GTEx multi-tissue eQTL in lung > 0.80	3	1	1
GTEx combined	3	1	1
Hao single tissue lung eQTL	0	0	0
FANTOM5 enhancer target gene	0	0	0
IM-PET IMR90	0	0	0
IM-PET NHLF	2	1	1
IM-PET combined	2	1	1
All genes	5	1	1

Appendix B. Final germline-regulated genes and overlap at GWAS SNP $p < 1 \times 10^{-4}$. Below illustrates the set of final germline-regulated genes discovered using the more stringent $p < 1 \times 10^{-4}$. The overall lack of overlap between subtypes is consistent with results at the more lenient threshold used in Chapter 2.



Appendix C. Summary of the down-regulated somatic DEGs discovered in Chapter 3 that are TSGs. TSGs were defined according to TSGene Database (156).

C.1. LUAD unique TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
DACH1	dachshund family transcription factor 1	protein-coding
DCC	DCC netrin 1 receptor	protein-coding
GPC3	glypican 3	protein-coding
MME	membrane metallo-endopeptidase	protein-coding
PACRG	PARK2 co-regulated	protein-coding

C.2. LUSC unique TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
ADARB1	adenosine deaminase, RNA-specific, B1	protein-coding
CYB5A	cytochrome b5 type A (microsomal)	protein-coding
ERBB4	erb-b2 receptor tyrosine kinase 4	protein-coding
FOXA2	forkhead box A2	protein-coding
LIFR	leukemia inhibitory factor receptor alpha	protein-coding
NFATC2	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2	protein-coding
NTRK3	neurotrophic tyrosine kinase, receptor, type 3	protein-coding
PGR	progesterone receptor	protein-coding
SEPT4	septin 4	protein-coding
SPTBN1	spectrin, beta, non-erythrocytic 1	protein-coding
LEFTY2	left-right determination factor 2	protein-coding
SEMA3B	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B	protein-coding
NR0B2	nuclear receptor subfamily 0, group B, member 2	protein-coding
RECK	reversion-inducing-cysteine-rich protein with kazal motifs	protein-coding
AKAP12	A kinase (PRKA) anchor protein 12	protein-coding
RASSF2	Ras association (RalGDS/AF-6) domain family member 2	protein-coding
CADM1	cell adhesion molecule 1	protein-coding
PCDH17	protocadherin 17	protein-coding
ZMYND10	zinc finger, MYND-type containing 10	protein-coding
DCDC2	doublecortin domain containing 2	protein-coding
CASC1	cancer susceptibility candidate 1	protein-coding
FAT4	FAT atypical cadherin 4	protein-coding
MFSD2A	major facilitator superfamily domain containing 2A	protein-coding
CABLES1	Cdk5 and Abl enzyme substrate 1	protein-coding
MIA2	melanoma inhibitory activity 2	protein-coding

SLC5A8	solute carrier family 5 (sodium/monocarboxylate cotransporter), member 8	protein-coding
MIRLET7F1	microRNA let-7f-1	ncRNA
MIR126	microRNA 126	ncRNA
MIR135A2	microRNA 135a-2	ncRNA
MIR142	microRNA 142	ncRNA
MIR26A1	microRNA 26a-1	ncRNA
MIR326	microRNA 326	ncRNA
VTRNA2-1	vault RNA 2-1	ncRNA
ADAMTS9-AS2	ADAMTS9 antisense RNA 2	ncRNA

C.3. SCLC unique TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
ADPRH	ADP-ribosylarginine hydrolase	protein-coding
AHR	aryl hydrocarbon receptor	protein-coding
AIF1	allograft inflammatory factor 1	protein-coding
FAS	Fas cell surface death receptor	protein-coding
ARG1	arginase 1	protein-coding
RHOB	ras homolog family member B	protein-coding
ATF3	activating transcription factor 3	protein-coding
ZFP36L2	ZFP36 ring finger protein-like 2	protein-coding
CASP5	caspase 5, apoptosis-related cysteine peptidase	protein-coding
CD4	CD4 molecule	protein-coding
CD44	CD44 molecule (Indian blood group)	protein-coding
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	protein-coding
CNN1	calponin 1, basic, smooth muscle	protein-coding
MAP3K8	mitogen-activated protein kinase kinase kinase 8	protein-coding
CSF2	colony stimulating factor 2 (granulocyte-macrophage)	protein-coding
CST6	cystatin E/M	protein-coding
CTGF	connective tissue growth factor	protein-coding
DAB2	Dab, mitogen-responsive phosphoprotein, homolog 2 (Drosophila)	protein-coding
DCN	decorin	protein-coding
DPP4	dipeptidyl-peptidase 4	protein-coding
DUSP6	dual specificity phosphatase 6	protein-coding
EGR2	early growth response 2	protein-coding
EPHA2	EPH receptor A2	protein-coding
EMP1	epithelial membrane protein 1	protein-coding
ESR1	estrogen receptor 1	protein-coding
HIC1	hypermethylated in cancer 1	protein-coding
IGF1	insulin-like growth factor 1 (somatomedin C)	protein-coding
IGFBP4	insulin-like growth factor binding protein 4	protein-coding
IRF1	interferon regulatory factor 1	protein-coding
IRF5	interferon regulatory factor 5	protein-coding
ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	protein-coding

MSMB	microseminoprotein, beta-	protein-coding
MT2A	metallothionein 2A	protein-coding
PLCD1	phospholipase C, delta 1	protein-coding
PPARG	peroxisome proliferator-activated receptor gamma	protein-coding
PRKCD	protein kinase C, delta	protein-coding
PRODH	proline dehydrogenase (oxidase) 1	protein-coding
KLK10	kallikrein-related peptidase 10	protein-coding
PTGDR	prostaglandin D2 receptor (DP)	protein-coding
PTPN13	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)	protein-coding
PTPRC	protein tyrosine phosphatase, receptor type, C	protein-coding
S100A11	S100 calcium binding protein A11	protein-coding
CXCL12	chemokine (C-X-C motif) ligand 12	protein-coding
SOD2	superoxide dismutase 2, mitochondrial	protein-coding
SP100	SP100 nuclear antigen	protein-coding
TAGLN	transgelin	protein-coding
TGFB1	transforming growth factor, beta 1	protein-coding
THBS1	thrombospondin 1	protein-coding
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	protein-coding
VIM	vimentin	protein-coding
ZNF185	zinc finger protein 185 (LIM domain)	protein-coding
ZYX	zyxin	protein-coding
SRPX	sushi-repeat containing protein, X-linked	protein-coding
TNFSF9	tumor necrosis factor (ligand) superfamily, member 9	protein-coding
TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b	protein-coding
IER3	immediate early response 3	protein-coding
LIMD1	LIM domains containing 1	protein-coding
SOCS3	suppressor of cytokine signaling 3	protein-coding
DLEC1	deleted in lung and esophageal cancer 1	protein-coding

CTDSPL	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like	protein-coding
SPRY2	sprouty homolog 2 (Drosophila)	protein-coding
YAP1	Yes-associated protein 1	protein-coding
ARL6IP5	ADP-ribosylation factor-like 6 interacting protein 5	protein-coding
TXNIP	thioredoxin interacting protein	protein-coding
PLA2G16	phospholipase A2, group XVI	protein-coding
RASSF8	Ras association (RalGDS/AF-6) domain family (N-terminal) member 8	protein-coding
PHLDA3	pleckstrin homology-like domain, family A, member 3	protein-coding
LATS2	large tumor suppressor kinase 2	protein-coding
DKK3	dickkopf WNT signaling pathway inhibitor 3	protein-coding
PYCARD	PYD and CARD domain containing	protein-coding
G0S2	G0/G1 switch 2	protein-coding
TNFRSF12A	tumor necrosis factor receptor superfamily, member 12A	protein-coding
ERRFI1	ERBB receptor feedback inhibitor 1	protein-coding
HRASLS2	HRAS-like suppressor 2	protein-coding
LXN	latexin	protein-coding
ADAMTS9	ADAM metallopeptidase with thrombospondin type 1 motif, 9	protein-coding
NDRG2	NDRG family member 2	protein-coding
MTUS1	microtubule associated tumor suppressor 1	protein-coding
ZBTB4	zinc finger and BTB domain containing 4	protein-coding
EDA2R	ectodysplasin A2 receptor	protein-coding
LRRC4	leucine rich repeat containing 4	protein-coding
BHLHE41	basic helix-loop-helix family, member e41	protein-coding
TNFAIP8L2	tumor necrosis factor, alpha-induced protein 8-like 2	protein-coding
CREB3L1	cAMP responsive element binding protein 3-like 1	protein-coding
CYGB	cytoglobin	protein-coding
JDP2	Jun dimerization protein 2	protein-coding
SIK1	salt-inducible kinase 1	protein-coding
SYNPO2	synaptopodin 2	protein-coding

SAMD9L	sterile alpha motif domain containing 9-like	protein-coding
SGMS1	sphingomyelin synthase 1	protein-coding
HCAR2	hydroxycarboxylic acid receptor 2	protein-coding
RASL11A	RAS-like, family 11, member A	protein-coding
PTPLAD2	protein tyrosine phosphatase-like A domain containing 2	protein-coding

C.4. LUAD overlap LUSC TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
CDO1	cysteine dioxygenase type 1	protein-coding
PCDH9	protocadherin 9	protein-coding
SFRP5	secreted frizzled-related protein 5	protein-coding
SLIT2	slit homolog 2 (Drosophila)	protein-coding
CMTM5	CKLF-like MARVEL transmembrane domain containing 5	protein-coding
MIR223	microRNA 223	ncRNA
MIR23A	microRNA 23a	ncRNA
MIR27A	microRNA 27a	ncRNA
MIR34C	microRNA 34c	ncRNA

C.5. LUAD overlap SCLC TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
ALOX15	arachidonate 15-lipoxygenase	protein-coding
THBD	thrombomodulin	protein-coding
WNT7A	wingless-type MMTV integration site family, member 7A	protein-coding
KLF4	Kruppel-like factor 4 (gut)	protein-coding
ABCG2	ATP-binding cassette, sub-family G (WHITE), member 2 (Junior blood group)	protein-coding
THSD1	thrombospondin, type I, domain containing 1	protein-coding
AHNAK	AHNAK nucleoprotein	protein-coding
SOX7	SRY (sex determining region Y)-box 7	protein-coding

C.6. LUSC overlap SCLC TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
ALOX15B	arachidonate 15-lipoxygenase, type B	protein-coding
ALPL	alkaline phosphatase, liver/bone/kidney	protein-coding
BMP2	bone morphogenetic protein 2	protein-coding
BTK	Bruton agammaglobulinemia tyrosine kinase	protein-coding
CAT	catalase	protein-coding
CFTR	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)	protein-coding
KLF6	Kruppel-like factor 6	protein-coding
CST5	cystatin D	protein-coding
DAPK1	death-associated protein kinase 1	protein-coding
DMBT1	deleted in malignant brain tumors 1	protein-coding
DUSP1	dual specificity phosphatase 1	protein-coding
EGR1	early growth response 1	protein-coding
FABP3	fatty acid binding protein 3, muscle and heart	protein-coding
FBP1	fructose-1,6-bisphosphatase 1	protein-coding
GPC5	glypican 5	protein-coding
NR4A1	nuclear receptor subfamily 4, group A, member 1	protein-coding
HPGD	hydroxyprostaglandin dehydrogenase 15-(NAD)	protein-coding
IGFALS	insulin-like growth factor binding protein, acid labile subunit	protein-coding
GADD45B	growth arrest and DNA-damage-inducible, beta	protein-coding
PF4	platelet factor 4	protein-coding
PLA2G2A	phospholipase A2, group IIA (platelets, synovial fluid)	protein-coding
PRKCE	protein kinase C, epsilon	protein-coding
RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2	protein-coding
SPI1	Spi-1 proto-oncogene	protein-coding
TBX5	T-box 5	protein-coding
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)	protein-coding

TIMP3	TIMP metallopeptidase inhibitor 3	protein-coding
ZFP36	ZFP36 ring finger protein	protein-coding
SPARCL1	SPARC-like 1 (hevin)	protein-coding
TNFSF12	tumor necrosis factor (ligand) superfamily, member 12	protein-coding
ALDH1A2	aldehyde dehydrogenase 1 family, member A2	protein-coding
SELENBP1	selenium binding protein 1	protein-coding
DOK2	docking protein 2, 56kDa	protein-coding
GPRC5A	G protein-coupled receptor, class C, group 5, member A	protein-coding
ARHGAP29	Rho GTPase activating protein 29	protein-coding
TSPAN32	tetraspanin 32	protein-coding
CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	protein-coding
RHOBTB2	Rho-related BTB domain containing 2	protein-coding
SASH1	SAM and SH3 domain containing 1	protein-coding
HSPB7	heat shock 27kDa protein family, member 7 (cardiovascular)	protein-coding
RBMS3	RNA binding motif, single stranded interacting protein 3	protein-coding
WFDC1	WAP four-disulfide core domain 1	protein-coding
SPRY4	sprouty homolog 4 (Drosophila)	protein-coding
HOPX	HOP homeobox	protein-coding
SCGB3A1	secretoglobin, family 3A, member 1	protein-coding
GATA5	GATA binding protein 5	protein-coding
SHISA3	shisa family member 3	protein-coding
ZNF366	zinc finger protein 366	protein-coding
GKN2	gastrokine 2	protein-coding
BCL6B	B-cell CLL/lymphoma 6, member B	protein-coding

C.7. All overlap TSGs from somatic down-regulated DEGs.

Gene Symbol	Description	Gene class
AGTR1	angiotensin II receptor, type 1	protein-coding
CAV1	caveolin 1, caveolae protein, 22kDa	protein-coding
CDH5	cadherin 5, type 2 (vascular endothelium)	protein-coding
EDNRB	endothelin receptor type B	protein-coding
EMP2	epithelial membrane protein 2	protein-coding
EPAS1	endothelial PAS domain protein 1	protein-coding
FHL1	four and a half LIM domains 1	protein-coding
GPX3	glutathione peroxidase 3 (plasma)	protein-coding
CXCR2	chemokine (C-X-C motif) receptor 2	protein-coding
MT1M	metallothionein 1M	protein-coding
TGFBR3	transforming growth factor, beta receptor III	protein-coding
ZBTB16	zinc finger and BTB domain containing 16	protein-coding
NR4A3	nuclear receptor subfamily 4, group A, member 3	protein-coding
KL	klotho	protein-coding
DLC1	DLC1 Rho GTPase activating protein	protein-coding
ADAMTS8	ADAM metallopeptidase with thrombospondin type 1 motif, 8	protein-coding
WIF1	WNT inhibitory factor 1	protein-coding
DAPK2	death-associated protein kinase 2	protein-coding
CSRNP1	cysteine-serine-rich nuclear protein 1	protein-coding
C2orf40	chromosome 2 open reading frame 40	protein-coding
STARD13	StAR-related lipid transfer (START) domain containing 13	protein-coding

Appendix D. Summary of the up-regulated somatic DEGs discovered in Chapter 3 that are oncogenes. Oncogenes were defined based upon the ONGene database (5).

D.1. LUAD unique oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
WISP1	WNT1 inducible signaling pathway protein 1	protein-coding
WNT3	wingless-type MMTV integration site family member 3	protein-coding
MUC4	mucin 4, cell surface associated	protein-coding
MIR135B	microRNA 135b	ncRNA
FHL2	four and a half LIM domains 2	protein-coding
CDH17	cadherin 17	protein-coding
LCN2	lipocalin 2	protein-coding
ADAM28	ADAM metallopeptidase domain 28	protein-coding

D.2. LUSC unique oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
SMO	smoothened, frizzled class receptor	protein-coding
MOS	v-mos Moloney murine sarcoma viral oncogene homolog	protein-coding
TP63	tumor protein p63	protein-coding
WNT10A	wingless-type MMTV integration site family member 10A	protein-coding
ZNF703	zinc finger protein 703	protein-coding
CT45A1	cancer/testis antigen family 45, member A1	protein-coding
LMX1B	LIM homeobox transcription factor 1 beta	protein-coding
MAFA	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog A	protein-coding
JUP	junction plakoglobin	protein-coding
HOXA1	homeobox A1	protein-coding
H19	H19, imprinted maternally expressed transcript (non-protein coding)	ncRNA
FGF4	fibroblast growth factor 4	protein-coding
CKS1B	CDC28 protein kinase regulatory subunit 1B	protein-coding
PTTG2	pituitary tumor-transforming 2	protein-coding
WNT10B	wingless-type MMTV integration site family member 10B	protein-coding
WNT5A	wingless-type MMTV integration site family member 5A	protein-coding
TSPY1	testis specific protein, Y-linked 1	protein-coding
MIR663A	microRNA 663a	ncRNA
BMP7	bone morphogenetic protein 7	protein-coding
S100A8	S100 calcium binding protein A8	protein-coding
BCL11A	B-cell CLL/lymphoma 11A	protein-coding
CENPW	centromere protein W	protein-coding
HSPB1	heat shock protein family B (small) member 1	protein-coding
GRM1	glutamate receptor, metabotropic 1	protein-coding

FGF8	fibroblast growth factor 8	protein-coding
TGM3	transglutaminase 3	protein-coding

D.3. SCLC unique oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
BCL2	B-cell CLL/lymphoma 2	protein-coding
ALK	anaplastic lymphoma receptor tyrosine kinase	protein-coding
MYB	MYB proto-oncogene, transcription factor	protein-coding
HOXA9	homeobox A9	protein-coding
WHSC1	Wolf-Hirschhorn syndrome candidate 1	protein-coding
TAL2	T-cell acute lymphocytic leukemia 2	protein-coding
RFC3	replication factor C subunit 3	protein-coding
MYCL	v-myc avian myelocytomatosis viral oncogene lung carcinoma derived homolog	protein-coding
HOXD9	homeobox D9	protein-coding
E2F1	E2F transcription factor 1	protein-coding
DUSP26	dual specificity phosphatase 26 (putative)	protein-coding
SOX4	SRY-box 4	protein-coding
PRDM8	PR domain 8	protein-coding
FEV	FEV, ETS transcription factor	protein-coding
E2F3	E2F transcription factor 3	protein-coding

D.4. LUAD overlap LUSC oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
MYCN	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog	protein-coding
MCF2	MCF.2 cell line derived transforming sequence	protein-coding
TNS4	tensin 4	protein-coding
S100A7	S100 calcium binding protein A7	protein-coding
PVT1	Pvt1 oncogene (non-protein coding)	ncRNA
STYK1	serine/threonine/tyrosine kinase 1	protein-coding
NME1	NME/NM23 nucleoside diphosphate kinase 1	protein-coding
MMP12	matrix metallopeptidase 12	protein-coding
MAGEA11	MAGE family member A11	protein-coding
MIR196A1	microRNA 196a-1	ncRNA
LHX1	LIM homeobox 1	protein-coding
SBSN	suprabasin	protein-coding
UHRF1	ubiquitin like with PHD and ring finger domains 1	protein-coding
FGF3	fibroblast growth factor 3	protein-coding
DSG3	desmoglein 3	protein-coding
CYP24A1	cytochrome P450 family 24 subfamily A member 1	protein-coding
DPPA2	developmental pluripotency associated 2	protein-coding
CDX2	caudal type homeobox 2	protein-coding
KIAA0101	KIAA0101	protein-coding
UCA1	urothelial cancer associated 1 (non-protein coding)	ncRNA
PRDM9	PR domain 9	protein-coding
MIR130B	microRNA 130b	ncRNA
BCAR4	breast cancer anti-estrogen resistance 4 (non-protein coding)	ncRNA
TCL6	T-cell leukemia/lymphoma 6 (non-protein coding)	ncRNA
ETV4	ETS variant 4	protein-coding
HOTTIP	HOXA distal transcript antisense RNA	ncRNA
HOTAIR	HOX transcript antisense RNA	ncRNA
GREM1	gremlin 1, DAN family BMP antagonist	protein-coding

D.5. LUAD overlap SCLC oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
RET	ret proto-oncogene	protein-coding
PAX4	paired box 4	protein-coding
LIN28A	lin-28 homolog A	protein-coding

D.6. LUSC overlap SCLC oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
TP73	tumor protein p73	protein-coding
SOX2	SRY-box 2	protein-coding
SKP2	S-phase kinase-associated protein 2, E3 ubiquitin protein ligase	protein-coding
KIAA1524	KIAA1524	protein-coding
GMNN	geminin, DNA replication inhibitor	protein-coding
PAX2	paired box 2	protein-coding
LMO1	LIM domain only 1	protein-coding
GALR2	galanin receptor 2	protein-coding
CNTN2	contactin 2	protein-coding
SYT1	synaptotagmin 1	protein-coding
PAK7	p21 protein (Cdc42/Rac)-activated kinase 7	protein-coding

D.7. All overlap oncogenes from somatic up-regulated DEGs.

Gene symbol	Full name	Gene type
PTTG1	pituitary tumor-transforming 1	protein-coding
HMGA2	high mobility group AT-hook 2	protein-coding
AURKA	aurora kinase A	protein-coding
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit	protein-coding
CDC6	cell division cycle 6	protein-coding
PIWIL1	piwi-like RNA-mediated gene silencing 1	protein-coding
CCNB2	cyclin B2	protein-coding
CCNE1	cyclin E1	protein-coding
FAM83D	family with sequence similarity 83 member D	protein-coding
TYMS	thymidylate synthetase	protein-coding
TWIST1	twist family bHLH transcription factor 1	protein-coding
SSX1	synovial sarcoma, X breakpoint 1	protein-coding
STIL	SCL/TAL1 interrupting locus	protein-coding
STRA6	stimulated by retinoic acid 6	protein-coding
SALL4	spalt-like transcription factor 4	protein-coding
HES6	hes family bHLH transcription factor 6	protein-coding
PAX3	paired box 3	protein-coding
LIN28B	lin-28 homolog B	protein-coding
HOXD13	homeobox D13	protein-coding
TLX1	T-cell leukemia homeobox 1	protein-coding
HMGA1	high mobility group AT-hook 1	protein-coding
FGF5	fibroblast growth factor 5	protein-coding
EEF1A2	eukaryotic translation elongation factor 1 alpha 2	protein-coding
ECT2	epithelial cell transforming 2	protein-coding
DLX5	distal-less homeobox 5	protein-coding
UBE2C	ubiquitin conjugating enzyme E2C	protein-coding
MLLT11	myeloid/lymphoid or mixed-lineage leukemia; translocated to, 11	protein-coding
IGF2BP1	insulin like growth factor 2 mRNA binding protein 1	protein-coding
CDKN3	cyclin-dependent kinase inhibitor 3	protein-coding
CDC25C	cell division cycle 25C	protein-coding
CDC25A	cell division cycle 25A	protein-coding

KIF14	kinesin family member 14	protein-coding
CDK1	cyclin-dependent kinase 1	protein-coding
ESPL1	extra spindle pole bodies like 1, separase	protein-coding
CDK5R2	cyclin-dependent kinase 5, regulatory subunit 2 (p39)	protein-coding
CCNB1	cyclin B1	protein-coding
ZIC2	Zic family member 2	protein-coding
UCHL1	ubiquitin C-terminal hydrolase L1	protein-coding
FAM72A	family with sequence similarity 72 member A	protein-coding
SIX1	SIX homeobox 1	protein-coding
PRDM13	PR domain 13	protein-coding
PRDM12	PR domain 12	protein-coding
PBK	PDZ binding kinase	protein-coding
PLK1	polo like kinase 1	protein-coding
PITX2	paired like homeodomain 2	protein-coding
OTX2	orthodenticle homeobox 2	protein-coding
MSI1	musashi RNA binding protein 1	protein-coding
ASCL1	achaete-scute family bHLH transcription factor 1	protein-coding
FEZF1	FEZ family zinc finger 1	protein-coding
PRAME	preferentially expressed antigen in melanoma	protein-coding
FOXM1	forkhead box M1	protein-coding
FOXG1	forkhead box G1	protein-coding
EN2	engrailed homeobox 2	protein-coding
BIRC5	baculoviral IAP repeat containing 5	protein-coding

REFERENCES

1. Stratton MR, Campbell PJ, & Futreal PA (2009) The cancer genome. *Nature* 458(7239):719-724.

2. Reddy EP, Reynolds RK, Santos E, & Barbacid M (1982) A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* 300(5888):149-152.

3. Tabin CJ*, et al.* (1982) Mechanism of activation of a human oncogene. *Nature* 300(5888):143-149.

4. Zhao M, Sun J, & Zhao Z (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic acids research* 41(Database issue):D970-976.

5. Liu Y, Sun J, & Zhao M (2017) ONGene: A literature-based database for human oncogenes. *J Genet Genomics* 44(2):119-121.

6. Knudson AG (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* 68(4):820-823.

7. Forbes SA*, et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* 43(Database issue):D805-811.

8. Weinhold N, Jacobsen A, Schultz N, Sander C, & Lee W (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics* 46(11):1160-1165.

9. Fredriksson NJ, Ny L, Nilsson JA, & Larsson E (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics* 46(12):1258-1263.

10. Jia P & Zhao Z (2014) VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS computational biology* 10(2):e1003460.

11. Gonzalez-Perez A*, et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods* 10(8):723-729.

12. Torre LA, Siegel RL, Ward EM, & Jemal A (2016) Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol Biomarkers Prev* 25(1):16-27.

13. Torre LA*, et al.* (2015) Global cancer statistics, 2012. *CA Cancer J Clin* 65(2):87-108.

14. Witschi H (2001) A short history of lung cancer. *Toxicol Sci* 64(1):4-6.

15. Proctor RN (2012) The history of the discovery of the cigarette-lung cancer link: evidentiary traditions, corporate denial, global toll. *Tob Control* 21(2):87-91.

16. Müller FH (1940) Tabakmissbrauch und lungencarcinom. *Journal of Cancer Research and Clinical Oncology* 49(1):57-85.

17. Doll R & Hill AB (1950) Smoking and carcinoma of the lung; preliminary report. *Br Med J* 2(4682):739-748.

18. Hammond EC & Horn D (1954) The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. *J Am Med Assoc* 155(15):1316-1328.

19. Cheng TY*, et al.* (2016) The International Epidemiology of Lung Cancer: Latest Trends, Disparities, and Tumor Characteristics. *J Thorac Oncol* 11(10):1653-1671.

20. Siegel RL, Miller KD, & Jemal A (2017) Cancer Statistics, 2017. *CA Cancer J Clin* 67(1):7-30.

21. Howlader N NA, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (2013) SEER Cancer Statistics Review, 1975-2013. (National Cancer Institute).

22. Commitee ALALCS (2015) Providing Guidance on Lung Cancer Screening to Patients and Physicians. (American Lung Association).

23. National Lung Screening Trial Research T*, et al.* (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine* 365(5):395-409.

24. Travis WD, World Health Organization., & International Agency for Research on Cancer. (2004) *Pathology and genetics of tumours of the lung, pleura, thymus and heart* p 344 p.

25. Alberg AJ, Wallace K, Silvestri GA, & Brock MV (2013) Invited commentary: the etiology of lung cancer in men compared with women. *Am J Epidemiol* 177(7):613-616.

26. Lan Q*, et al.* (2012) Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nature genetics* 44(12):1330-1335.

27. Hemminki K, Lonnstedt I, Vaittinen P, & Lichtenstein P (2001) Estimation of genetic and environmental components in colorectal and lung cancer and melanoma. *Genet Epidemiol* 20(1):107-116.

28. Hemminki K & Vaittinen P (1998) National database of familial cancer in Sweden. *Genet Epidemiol* 15(3):225-236.

29. Czene K, Lichtenstein P, & Hemminki K (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *International journal of cancer. Journal international du cancer* 99(2):260-266.

30. Sampson JN*, et al.* (2015) Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *Journal of the National Cancer Institute* 107(12):djv279.

31. Tokuhata GK & Lilienfeld AM (1963) Familial aggregation of lung cancer in humans. *Journal of the National Cancer Institute* 30:289-312.

32. Tokuhata GK & Lilienfeld AM (1963) Familial aggregation of lung cancer among hospital patients. *Public Health Rep* 78:277-283.

33. Coté ML*, et al.* (2012) Increased risk of lung cancer in individuals with a family history of the disease: A pooled analysis from the International Lung Cancer Consortium. *European Journal of Cancer* 48(13):1957-1968.

34. Musolf MA*, et al.* (2017) Familial Lung Cancer: A Brief History from the Earliest Work to the Most Recent Studies. *Genes* 8(1).

35. Pao W & Girard N (2011) New driver mutations in non-small-cell lung cancer. *Lancet Oncol* 12(2):175-180.

36. van Meerbeeck JP, Fennell DA, & De Ruysscher DK (2011) Small-cell lung cancer. *Lancet* 378(9804):1741-1755.

37. Hann CL & Rudin CM (2007) Fast, hungry and unstable: finding the Achilles' heel of small-cell lung cancer. *Trends Mol Med* 13(4):150-157.

38. Bender E (2014) Epidemiology: The dominant malignancy. *Nature* 513(7517):S2-3.

39. Khuder SA (2001) Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. *Lung cancer* 31(2-3):139-148.

40. Bailey-Wilson JE*, et al.* (2004) A Major Lung Cancer Susceptibility Locus Maps to Chromosome 6q23–25. *The American Journal of Human Genetics* 75(3):460-474.

41. You M*, et al.* (2009) Fine Mapping of Chromosome 6q23-25 Region in Familial Lung Cancer Families Reveals RGS17 as a Likely Candidate Gene. *Clinical Cancer Research* 15(8):2666.

42. Amos CI*, et al.* (2010) A Susceptibility Locus on Chromosome 6q Greatly Increases Lung Cancer Risk among Light and Never Smokers. *Cancer research* 70(6):2359.

43. Bell DW*, et al.* (2005) Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. *Nature genetics* 37(12):1315-1316.

44. Gazdar A*, et al.* (2014) Hereditary lung cancer syndrome targets never smokers with germline EGFR gene T790M mutations. *J Thorac Oncol* 9(4):456-463.

45. Pao W*, et al.* (2005) Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLOS Medicine* 2(3):e73.

46. Birch JM*, et al.* (2001) Relative frequency and morphology of cancers in carriers of germline TP53 mutations. *Oncogene* 20(34):4621-4628.

47. Brennan P, Hainaut P, & Boffetta P (2011) Genetics of lung-cancer susceptibility. *Lancet Oncol* 12(4):399-408.

48. Benhamou S*, et al.* (2002) Meta- and pooled analyses of the effects of glutathione S-transferase M1 polymorphisms and smoking on lung cancer risk. *Carcinogenesis* 23(8):1343-1350.

49. Cybulski C*, et al.* (2004) CHEK2 Is a Multiorgan Cancer Susceptibility Gene. *The American Journal of Human Genetics* 75(6):1131-1135.

50. Cybulski C*, et al.* (2008) Constitutional CHEK2 mutations are associated with a decreased risk of lung and laryngeal cancers. *Carcinogenesis* 29(4):762-765.

51. Klein RJ*, et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720):385-389.

52. Welter D*, et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42(Database issue):D1001-1006.

53. Amos CI*, et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature genetics* 40(5):616-622.

54. Hung RJ*, et al.* (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 452(7187):633-637.

55. McKay JD*, et al.* (2008) Lung cancer susceptibility locus at 5p15.33. *Nature genetics* 40(12):1404-1406.

56. Wang Y*, et al.* (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature genetics* 40(12):1407-1409.

57. Saccone SF*, et al.* (2007) Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* 16(1):36-49.

58. Timofeeva MN*, et al.* (2012) Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* 21(22):4980-4995.

59. Wang Y*, et al.* (2014) Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics* 46(7):736-741.

60. Hu Z*, et al.* (2011) A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nature genetics* 43(8):792-796.

61. Dong J*, et al.* (2013) Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese. *PLoS Genet* 9(1):e1003190.

62. Zanetti KA*, et al.* (2016) Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung cancer* 98:33-42.

63. Bush WS & Moore JH (2012) Chapter 11: Genome-wide association studies. *PLoS computational biology* 8(12):e1002822.

64. Smemo S*, et al.* (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507(7492):371-375.

65. Albert FW & Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics* 16(4):197-212.

66. Consortium GT (2013) The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 45(6):580-585.

67. Consortium GT (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348(6235):648-660.

68. Nicolae DL*, et al.* (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6(4):e1000888.

69. Jiang J, Jia P, Shen B, & Zhao Z (2014) Top associated SNPs in prostate cancer are significantly enriched in cis-expression quantitative trait loci and at transcription factor binding sites. *Oncotarget* 5(15):6168-6177.

70. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, & Bejerano G (2013) Enhancers: five essential questions. *Nature reviews. Genetics* 14(4):288-295.

71. Erwin GD*, et al.* (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS computational biology* 10(6):e1003677.

72. He B, Chen C, Teng L, & Tan K (2014) Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 111(21):E2191-2199.

73. Consortium F*, et al.* (2014) A promoter-level mammalian expression atlas. *Nature* 507(7493):462-470.

74. Kodzius R*, et al.* (2006) CAGE: cap analysis of gene expression. *Nature methods* 3(3):211-222.

75. Sur I, Tuupanen S, Whitington T, Aaltonen LA, & Taipale J (2013) Lessons from functional analysis of genome-wide association studies. *Cancer research* 73(14):4180-4184.

76. Ding L*, et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455(7216):1069-1075.

77. Hammerman PS*, et al.* (2011) Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. *Cancer Discov* 1(1):78-89.

78. Arriola E*, et al.* (2008) Genetic changes in small cell lung carcinoma. *Clinical and Translational Oncology* 10(4):189-197.

79. Weir BA*, et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature* 450(7171):893-898.

80. Govindan R*, et al.* (2012) Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 150(6):1121-1134.

81. Imielinski M*, et al.* (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150(6):1107-1120.

82. Cancer Genome Atlas Research N (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511(7511):543-550.

83. Bass AJ*, et al.* (2009) SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature genetics* 41(11):1238-1242.

84. Weiss J*, et al.* (2010) Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci Transl Med* 2(62):62ra93.

85. Beroukhim R*, et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences* 104(50):20007-20012.

86. Cancer Genome Atlas Research N (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417):519-525.

87. Levin NA, Brzoska PM, Warnock ML, Gray JW, & Christman MF (1995) Identification of novel regions of altered DNA copy number in small cell lung tumors. *Genes Chromosomes Cancer* 13(3):175-185.

88. Peifer M*, et al.* (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature genetics* 44(10):1104-1110.

89. Rudin CM*, et al.* (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nature genetics* 44(10):1111-1116.

90. George J*, et al.* (2015) Comprehensive genomic profiles of small cell lung cancer. *Nature* 524(7563):47-53.

91. Pietanza MC & Ladanyi M (2012) Bringing the genomic landscape of small-cell lung cancer into focus. *Nature genetics* 44(10):1074-1075.

92. Campbell JD*, et al.* (2016) Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature genetics* 48(6):607-616.

93. Houston KA, Henley SJ, Li J, White MC, & Richards TB (2014) Patterns in lung cancer incidence rates and trends by histologic type in the United States, 2004-2009. *Lung cancer* 86(1):22-28.

94. Spinola M*, et al.* (2007) Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. *Cancer Lett* 251(2):311-316.

95. Broderick P*, et al.* (2009) Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer research* 69(16):6633-6641.

96. Landi MT*, et al.* (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *American journal of human genetics* 85(5):679-691.

97. Yoon KA*, et al.* (2010) A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. *Hum Mol Genet* 19(24):4948-4954.

98. Li Y*, et al.* (2010) Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol* 11(4):321-330.

99. Ahn MJ*, et al.* (2012) The 18p11.22 locus is associated with never smoker non-small cell lung cancer susceptibility in Korean populations. *Human genetics* 131(3):365-372.

100. Jia P*, et al.* (2012) Network-Assisted Investigation of Combined Causal Signals from Genome-Wide Association Studies in Schizophrenia. *PLoS computational biology* 8(7):e1002587.

101. Jia P, Wang L, Meltzer HY, & Zhao Z (2010) Common variants conferring risk of schizophrenia: A pathway analysis of GWAS data. *Schizophrenia Research* 122(1–3):38-42.

102. Dayem Ullah AZ, Lemoine NR, & Chelala C (2012) SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research* 40(Web Server issue):W65-70.

103. Chelala C, Khan A, & Lemoine NR (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* 25(5):655-661.

104. Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27(5):718-719.

105. Danecek P*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156-2158.

106. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559-575.

107. Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353-1358.

108. Flutre T, Wen X, Pritchard J, & Stephens M (2013) A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* 9(5):e1003486.

109. Li G, Shabalin AA, Rusyn I, Wright FA, & Nobel AB (2013) An Empirical Bayes Approach for Multiple Tissue eQTL Analysis. *ArXiv e-prints.* 1311.

110. Hao K*, et al.* (2012) Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 8(11):e1003029.

111. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841-842.

112. Flicek P*, et al.* (2014) Ensembl 2014. *Nucleic acids research* 42(Database issue):D749-755.

113. Durinck S*, et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21(16):3439-3440.

114. Wang J, Duncan D, Shi Z, & Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* 41(Web Server issue):W77-83.

115. Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 57(1):289-300.

116. Liu P*, et al.* (2008) Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *Journal of the National Cancer Institute* 100(18):1326-1330.

117. Wang T*, et al.* (2015) Association of PSMA4 polymorphisms with lung cancer susceptibility and response to cisplatin-based chemotherapy in a Chinese Han population. *Clin Transl Oncol* 17(7):564-569.

118. Ziolkowska-Suchanek I*, et al.* (2015) Susceptibility loci in lung cancer and COPD: association of IREB2 and FAM13A with pulmonary diseases. *Sci Rep* 5:13502.

119. Andersson R*, et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455-461.

120. Heintzman ND*, et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* 39(3):311-318.

121. Creyghton MP*, et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* 107(50):21931-21936.

122. Levine AJ & Puzio-Kuter AM (2010) The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* 330(6009):1340-1344.

123. Zhang B, Kirov S, & Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research* 33(Web Server issue):W741-748.

124. Kanehisa M & Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* 28(1):27-30.

125. Sanyal A, Lajoie BR, Jain G, & Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489(7414):109-113.

126. Dang Chi V (2012) MYC on the Path to Cancer. *Cell* 149(1):22-35.

127. Taly A, Corringer PJ, Guedin D, Lestage P, & Changeux JP (2009) Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system. *Nat Rev Drug Discov* 8(9):733-750.

128. Improgo MR, Scofield MD, Tapper AR, & Gardner PD (2010) From smoking to lung cancer: the CHRNA5/A3/B4 connection. *Oncogene* 29(35):4874-4884.

129. Huh TL, Kim YO, Oh IU, Song BJ, & Inazawa J (1996) Assignment of the human mitochondrial NAD+ -specific isocitrate dehydrogenase alpha subunit (IDH3A) gene to 15q25.1-->q25.2by in situ hybridization. *Genomics* 32(2):295-296.

130. Zeng L*, et al.* (2015) Aberrant IDH3alpha expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. *Oncogene* 34(36):4758-4766.

131. Zhang D*, et al.* (2015) Metabolic reprogramming of cancer-associated fibroblasts by IDH3alpha downregulation. *Cell Rep* 10(8):1335-1348.

132. Liu Y*, et al.* (2009) Haplotype and cell proliferation analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. *Cancer research* 69(19):7844-7850.

133. Popovic D*, et al.* (2012) Rab GTPase-activating proteins in autophagy: regulation of endocytic and autophagy pathways by direct binding to human ATG8 modifiers. *Mol Cell Biol* 32(9):1733-1744.

134. Dang CV (2012) Links between metabolism and cancer. *Genes Dev* 26(9):877-890.

135. Zheng J (2012) Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). *Oncology letters* 4(6):1151-1157.

136. Thiery JP (2002) Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2(6):442-454.

137. Karachaliou N*, et al.* (2016) Cellular and molecular biology of small cell lung cancer: an overview. *Transl Lung Cancer Res* 5(1):2-15.

138. Musunuru K*, et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466(7307):714-719.

139. Wang Q, Yu H, Zhao Z, & Jia P (2015) EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics* 31(15):2591-2594.

140. Jia P, Wang L, Meltzer HY, & Zhao Z (2011) Pathway-based analysis of GWAS datasets: effective but caution required. *Int J Neuropsychopharmacol* 14(4):567-572.

141. Pao W*, et al.* (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America* 101(36):13306-13311.

142. Paez JG*, et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304(5676):1497-1500.

143. Maemondo M*, et al.* (2010) Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *The New England journal of medicine* 362(25):2380-2388.

144. Mitsudomi T*, et al.* (2010) Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *Lancet Oncol* 11(2):121-128.

145. Lettice LA*, et al.* (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12(14):1725-1735.

146. Mayakonda A & Koeffler HP (2016) Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *bioRxiv*.

147. Gaujoux R & Seoighe C (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 11:367.

148. Dhillon IS & Modha DS (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning* 42(1):143-175.

149. Bamford S*, et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer* 91(2):355-358.

150. Vogelstein B*, et al.* (2013) Cancer genome landscapes. *Science* 339(6127):1546-1558.

151. Rahman M*, et al.* (2015) Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* 31(22):3666-3672.

152. Liao Y, Smyth GK, & Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research* 41(10):e108.

153. Wagner GP, Kin K, & Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131(4):281-285.

154. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5(7):621-628.

155. Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15(12):550.

156. Zhao M, Kim P, Mitra R, Zhao J, & Zhao Z (2016) TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic acids research* 44(D1):D1023-1031.

157. Alexandrov LB*, et al.* (2016) Mutational signatures associated with tobacco smoking in human cancer. *Science* 354(6312):618-622.

158. Lawrence MS*, et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214-218.

159. Rosell R*, et al.* (2012) Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol* 13(3):239-246.

160. Shaw AT*, et al.* (2013) Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *The New England journal of medicine* 368(25):2385-2394.

161. Choi J*, et al.* (2014) The associations between immunity-related genes and breast cancer prognosis in Korean women. *PloS one* 9(7):e103593.

162. Kandoth C*, et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502(7471):333-339.

163. Wang Q*, et al.* (2013) Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome medicine* 5(10):91.

164. Xu H, DiCarlo J, Satya RV, Peng Q, & Wang Y (2014) Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC genomics* 15:244.

165. Gartner JJ*, et al.* (2013) Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proceedings of the National Academy of Sciences of the United States of America* 110(33):13481-13486.

166. Supek F, Minana B, Valcarcel J, Gabaldon T, & Lehner B (2014) Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156(6):1324-1335.

167. Gotea V, Gartner JJ, Qutob N, Elnitski L, & Samuels Y (2015) The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment Cell Melanoma Res* 28(6):673-684.

168. O'Brien TD*, et al.* (2015) Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods* 83:118-127.

169. Chapman PB*, et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England journal of medicine* 364(26):2507-2516.

170. Davies H*, et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature* 417(6892):949-954.

171. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, & Jabado N (2011) What can exome sequencing do for you? *Journal of medical genetics* 48(9):580-589.

172. Rabbani B, Tekin M, & Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *Journal of human genetics* 59(1):5-15.

173. Jia P*, et al.* (2013) Next-generation sequencing of paired tyrosine kinase inhibitor-sensitive and -resistant EGFR mutant lung cancer cell lines identifies spectrum of DNA changes associated with drug resistance. *Genome research* 23(9):1434-1445.

174. Chepelev I, Wei G, Tang Q, & Zhao K (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic acids research* 37(16):e106.

175. Greif PA*, et al.* (2011) Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia* 25(5):821-827.

176. Maas S (2012) Posttranscriptional recoding by RNA editing. *Advances in protein chemistry and structural biology* 86:193-224.

177. Seo JS*, et al.* (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome research* 22(11):2109-2119.

178. Liu J*, et al.* (2012) Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome research* 22(12):2315-2327.

179. Wilkerson MD*, et al.* (2014) Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic acids research* 42(13):e107.

180. Cirulli ET*, et al.* (2010) Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome biology* 11(5):R57.

181. Ku CS*, et al.* (2012) Exome versus transcriptome sequencing in identifying coding region variants. *Expert review of molecular diagnostics* 12(3):241-251.

182. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.

183. Picard Web Site (http://picard.sourceforge.net/.

184. DePristo MA*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43(5):491-498.

185. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.

186. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

187. Koboldt DC*, et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22(3):568-576.

188. Kim D*, et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14(4):R36.

189. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28(5):511-515.

190. Kleinman CL & Majewski J (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335(6074):1302; author reply 1302.

191. Pickrell JK, Gilad Y, & Pritchard JK (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335(6074):1302; author reply 1302.

192. Lin W, Piskol R, Tan MH, & Li JB (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 335(6074):1302; author reply 1302.

193. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

194. Paz N*, et al.* (2007) Altered adenosine-to-inosine RNA editing in human cancer. *Genome research* 17(11):1586-1595.

195. Chen L*, et al.* (2013) Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nature medicine* 19(2):209-216.

196. Balkissoon R, Lommatzsch S, Carolan B, & Make B (2011) Chronic obstructive pulmonary disease: a concise review. *Med Clin North Am* 95(6):1125-1141.

197. Marsh S, Aldington S, Shirtcliffe P, Weatherall M, & Beasley R (2006) Smoking and COPD: what really are the risks? *European Respiratory Journal* 28(4):883.

198. Global Initiative for Chronic Obstructive Lung Disease I (2017) Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease (2017 Report).

199. Jemal A*, et al.* (2011) Global cancer statistics. *CA Cancer J Clin* 61(2):69-90.

200. Sun S, Schiller JH, & Gazdar AF (2007) Lung cancer in never smokers [mdash] a different disease. *Nat Rev Cancer* 7(10):778-790.

201. Ke L (2002) Mortality and incidence trends from esophagus cancer in selected geographic areas of China circa 1970–90. *International Journal of Cancer* 102(3):271-274.

202. Gao Y*, et al.* (2011) Risk factors for esophageal and gastric cancers in Shanxi Province, China: A case–control study. *Cancer Epidemiology* 35(6):e91-e99.

203. Tran GD*, et al.* (2005) Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. *International journal of cancer. Journal international du cancer* 113(3):456-463.

204. Pillai SG*, et al.* (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 5(3):e1000421.

205. R Core Team (2013) R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria).

206. Abnet CC*, et al.* (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics* 42(9):764-767.

207. Mailman MD*, et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics* 39(10):1181-1186.

208. Genomes Project C*, et al.* (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

209. Hoadley KA*, et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158(4):929-944.

210. Lemjabbar-Alaoui H, Hassan OU, Yang YW, & Buchanan P (2015) Lung cancer: Biology and treatment options. *Biochim Biophys Acta* 1856(2):189-210.

211. Liu JZ*, et al.* (2010) A versatile gene-based test for genome-wide association studies. *American journal of human genetics* 87(1):139-145.

212. Barbeira A*, et al.* (2016) MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results. *bioRxiv*.

213. Lamparter D, Marbach D, Rueedi R, Kutalik Z, & Bergmann S (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS computational biology* 12(1):e1004714.

214. Mootha VK*, et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* 34(3):267-273.