5-2019

COMPUTATIONAL GENOMIC MODELS FOR SPATIO-TEMPORAL INVESTIGATION OF EARLY LUNG CANCER PATHOLOGY

Smruthy Sivakumar

COMPUTATIONAL GENOMIC MODELS FOR SPATIO-TEMPORAL INVESTIGATION OF EARLY LUNG CANCER PATHOLOGY

by

Smruthy Sivakumar, M.S.

APPROVED:

Paul Scheet, Ph.D.
Advisory Professor

Humam Kadara, Ph.D.
Secondary Advisory Professor

Eduardo Vilar Sanchez, M.D. Ph.D.

Sadhan Majumder, Ph.D.

Yin Liu, Ph.D.

APPROVED:

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

COMPUTATIONAL GENOMIC MODELS FOR SPATIO-TEMPORAL INVESTIGATION OF EARLY LUNG CANCER PATHOLOGY

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Smruthy Sivakumar, M.S.

Houston, Texas

May 2019

To Amma and Appa, for their eternal love and support.

To my friends, who have constantly been by my side.

To my husband, Varun, for his support in every imaginable way and for all his encouragement and love over the years.

ACKNOWLEDGEMENTS

I thank my advisor, Paul Scheet, for being a very supportive and motivating mentor. He always encouraged me to think independently, ask questions no matter how silly, and patiently explained any statistical concepts that I have had trouble understanding. He has filled many shoes including that of a teacher, a friend and a parent, and I am very thankful for everything he has done for me. I would also like to thank my co-advisor, Humam Kadara, for all the time he dedicated to teaching me about lung cancer and helping me interpret my computational analyses for translational applications. I want to thank him for all the conversations, both professional and personal, during the long sessions of manuscript preparation. I am very fortunate to have trained under their joint mentorship. My work is very collaborative and they taught me how to thrive in such a space. Their constant guidance helped me in planning, executing and publishing my research. I would like to thank them for always looking out for me, both professionally and personally.

I would also like to thank my lab peers, from the past and present, for making it a very fruitful and nourishing lab environment, where I could lean on anyone for their support in terms of my work and otherwise. I would especially like to thank Anthony San Lucas, for serving as a mentor and friend. I learned a lot by observing him work and enjoyed the projects we co-led in the lab. His hardworking nature and humility has constantly encouraged me to become a better person. I also want to thank Jerry Fowler, for developing SyQADA, a framework which helped me run all my analyses. He also trained me to develop more automated frameworks for cancer genomic applications. I extend my gratitude to my long-time office mate, Yihua Liu. Since the time I joined the lab, she made me feel very comfortable, offered a helping hand and always watched out for my well being. I would also like to thank Sasha Jakubek for her extensive help throughout my work. I have enjoyed my conversations with her and have traveled in her shadow at many conferences, where she has encouraged me to socialize and approach people. I would like to thank each of them for taking the time to patiently review all my submissions including conference abstracts, posters and presentations, research articles, scholarship applications as well as my dissertation. I thank Margie Stembridge for her administrative help, for pampering

me with brownies and other gifts over the years. I thank all lab members for making me enjoy my time in the lab. I also thank the Department of Epidemiology and the Division of Cancer Prevention at MD Anderson.

I thank my advisory committee members for their valuable insights that have helped me sculpt my dissertation into what it is now and for providing me their letters of support over the years. I thank Eduardo Vilar Sanchez, for his valuable feedback on developing the translational concepts of my work. I thank Yin Liu for her comments and insights on computational improvement of my methods. I thank Sadhan Majumder, for serving as external member who has always shown encouragement and excitement for my work, and for helping me better explain my research. I also thank him for serving as a chair on my candidacy examination committee. I also extend my thanks to the other members of my candidacy exam committee - Eduardo Vilar, James Hixson, Arvind Rao and Chad Huff. I am grateful to Sanjay Shete for his constant advice and help. His help extends to days before I joined the school, and I thank him for recommending courses when I started here and providing me with letters of support throughout my training here. I also want to thank the program directors of the BBSB program for their constant support and my peers in the program for their friendship. I extend my gratitude to Ken Chen, for introducing me to the research work at MD Anderson, and for providing me an opportunity to rotate in his lab.

A significant portion of the research presented in this dissertation was supported by the Pauline Altman-Goldstein Foundation Discovery Fellowship. I want to thank the organization and the MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for providing me with this honor. I thank Dean Michelle Barton and Dean Michael Blackburn, for actively taking an interest in my work and for being very approachable. I also thank my advisors at school, Bill Mattox and Brenda Gaughan for all their advice, encouragement and assistance. I extend my appreciation to all the other school staff members who have helped me with various deeds such as reserving rooms for my committee meetings, providing technical support, assisting with course registration, among other things. I am also thankful to all my course instructors for providing me with a strong foundation in bioinformatics and cancer genomics.

My interest to pursue a doctoral degree came from a lot of encouragement I received

before I started at GSBS. I would like to thank my mentors and colleagues at Dow AgroSciences, for their advice, support and assistance in pursuing a doctoral degree. I would like to extend my gratitude to Uma Ranjan, for introducing me to computational research and for mentoring me during my undergraduate training. I am also grateful to Georgia Tech, for providing me with a strong foundation in bioinformatics that helped me pursue cancer genomics research in my doctoral training.

I believe that the people you meet are more important than the places you go. I feel very blessed to have been surrounded by a tight knit group of friends who, like family, have always been my side through all the ups and downs. Some friends I have known for more than 20 years and some others for less than a year. I am grateful to have them as a part of my life and want to thank each of them for the countless Whatsapp/Skype/Hangouts texts, phone calls and video chats, that were a huge part of my doctoral journey as well.

I sincerely thank my parents, Anuradha and Sivakumar for their unwavering support and love these 28 years of my life. I feel very lucky to have such supportive parents, who have constantly encouraged me to pursue my dreams. I am grateful to them for everything they have done for me. I would not be in this position today if not for their sacrifices. I would also like to thank my extended family and my cousins for their love.

Lastly, I want to thank my partner through it all, Varun Rao. I thank him for patiently helping me these past few years, even if it meant something as small as reviewing an email. I thank him for always being with me during the tough times, and always cracking me up with his silly jokes whenever I needed it the most. Even though we were never in the same city physically, I never felt a distance between us. And that was the most comforting feeling. His growing love and support over the years has been a constant driving force for me.

Every single person I have met in this journey has impacted me in one way or the other. I thank each and every one for helping me build memories that I will cherish for the years to come.

COMPUTATIONAL GENOMIC MODELS FOR SPATIO-TEMPORAL INVESTIGATION OF EARLY LUNG CANCER PATHOLOGY

Smruthy Sivakumar, M.S.

Advisory Professor: Paul Scheet, Ph.D.

Secondary Advisor: Humam Kadara, Ph.D.

Lung cancer, of which non-small cell lung cancer (NSCLC) is the most common form, is the second most prevalent cancer and the leading cause of cancer-related deaths. NSCLCs primarily comprise adenocarcinomas (LUAD) and squamous cell carcinomas (LUSC). Advances in early detection and prevention have been limited by the lack of early-stage biomarkers and targets. A comprehensive molecular characterization of premalignant lesions and tumor-adjacent normal tissue can aid in better understanding NSCLC pathogenesis. However, these investigations are further challenged by limited tissue availability and low cellular fractions of detectable somatic mutations.

Therefore, there is a dearth of knowledge about the pathogenesis of premalignant lung lesions, especially for atypical adenomatous hyperplasia (AAH), the only known precursor to LUADs. We performed a cross-platform integrative analysis comprising targeted DNA sequencing, genotype array profiling and transcriptome sequencing of matched AAHs, LUADs and normal tissues from 23 early-stage patients. The study revealed potentially divergent pathways based on the mutation status of AAH (*BRAF* vs *KRAS*), recurrent chromosomal aberrations (17p loss) and the presence of immune deregulation early in the pathogenesis of AAHs.

Molecular changes, characteristic of NSCLCs, might also occur in normal tissues, preceding identifiable premalignancy-associated morphological changes. We sought to comprehensively survey the somatic mutational architecture of the normal airway in early-stage NSCLCs. Targeted DNA sequencing allowed us to capture driver mutations at low cellular fractions, typical of these non-malignant tissues. Additionally, genotype array profiling helped characterize subtle chromosomal aberrations in these tissues. This multi-region

study included tumor-adjacent and -distant airways, nasal epithelia and uninvolved normal lung (collectively cancerized field) along with matched multi-region NSCLCs and blood cells from 48 patients. Integrative computational analysis revealed genomic airway field carcinogenesis in 52% of cases. The airway field exhibited mutations in known drivers, that were present at lower frequencies compared to NSCLCs, suggestive of selection-driven clonal expansion. These driver events also comprised somatic two-hit alterations in matched airway field and NSCLCs.

Our study design offers spatiotemporal insights into NSCLC development and suggests potential targets for early detection and treatment, in possibly less hostile environments of premalignancy. To validate and enhance the utility of the bioinformatic techniques devised and implemented for these investigations, I also provide methods to expand such analyses across multiple tumor sites.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABBREVIATIONS

AAH	Atypical adenomatous hyperplasia
ACC	Adrenocortical carcinoma
AI	Allelic imbalance
AIS	Adenocarcinoma *in situ*
BAF	B-allele frequency
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CCP	Comprehensive Cancer Panel
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
CIS	Carcinoma *in situ*
COAD	Colon adenocarcinoma
CN	Copy number
CNB	Core needle biopsy (CNB1-CNB8)
cnLOH	Copy neutral loss of heterozygosity
CT	Computerized tomography
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic acid
ESCA	Esophageal carcinoma
ExAC	Exome Aggregation Consortium
FCAUC	Field cancerization area under the curve
FFPE	Formalin-fixed and paraffin-embedded
FDR	False discovery rate
GBM	Glioblastoma multiforme
H&E	Hematoxylin and eosin
HNSC	Head and Neck squamous cell carcinoma
ITH	intra-tumor heterogeneity
LOH	Loss of heterozygosity
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
L	Large airway
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LRR	log R ratio
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
MIA	Minimally invasive adenocarcinoma
N	Normal tissue
Na	Nasal epithelium

NDR	Non-reference discordance rate
NGS	Next-generation sequencing
NL	Normal lung parenchyma
NSCLC	Non-small cell lung cancer
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
qPCR	Quantitative polymerase chain reaction
READ	Rectum adenocarcinoma
RNA	Ribonucleic acid
S	Small airway (S1-S5)
SARC	Sarcoma
SCLC	Small cell lung cancer
SCNA	Somatic copy number alterations
SKCM	Skin Cutaneous Melanoma
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SRA	Sequence read archive
STAD	Stomach adenocarcinoma
T	Tumor
TCGA	The cancer genome atlas
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
TRU	Terminal respiratory unit
TVC	Torrent Variant Caller
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UTR	Untranslated regions
UV	Ultraviolet light
UVM	Uveal Melanoma
VAF	Variant allele frequency
VCF	Variant call format
WT	Wild type

CHAPTER 1
INTRODUCTION

1.1 Background

Cancer is a multi-step complex genetic disease that is characterized by several hallmarks of development [1]. Cancer is thought to arise from a single normal cell. Each cell in our body contains genetic information encoded within the deoxyribonucleic acid (DNA) that is packaged into chromosomes. Human DNA has 3 billion bases consisting of a sequence of four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T), that are together organized in a double-stranded structure. Humans have 23 pairs of chromosomes consisting of 22 pairs of autosomes (1-22) and 1 pair of sex chromosomes (XX in females and XY in males). DNA has the property of self-replication, where in each strand serves as a template for the synthesis of two new identical strands. This process of DNA replication is crucial for cells to divide, grow and replenish. While this process has many checkpoints to assure the correctness of the replicated DNA, the machinery is not perfect. This might result in a mistake, such as of a single base change, that we term a mutation. Mutations may also develop due to lifestyle habits and environmental factors such as smoking and ultraviolet light (UV) exposure. Mutations in certain genes called proto-oncogenes that normally help cell growth, can result in an activation and overproduction of oncogenes that drive cancer. Mutations might also occur in genes termed as tumor suppressors, that are meant to safeguard normal cellular functions such as DNA repair and apoptosis; these may become inactivate resulting in abnormal cell growth that might further drive cancer development. The presence and accumulation of such mutational processes may result in genomic instability, an emerging hallmark of cancer and known to play a critical role in tumor initiation and progression [2].

Genomic instability is thought to occur in all stages of cancer development, from a normal cell acquiring a mutation, to precancerous lesions and all the way until the formation of more advanced cancers. Mutations as small as a single nucleotide variation (SNV) or

those that span larger regions such as whole-chromosomes or chromosome arms as well as genomic rearrangements may occur in genomically unstable cells. Characterizing these changes, early on, such as in precancerous lesions or normal cells in the vicinity of the tumor can aid in understanding the molecular mechanisms preceding tumor formation.

1.1.1 Genetic characterization of lung cancer

Lung cancer is the second more frequent form of cancer in the United States with an estimate of 228,150 new cases in 2019 [3]. It is also the leading cause of cancer-related deaths worldwide, accounting for one quarter of all cancer deaths [3]. Among the different types of lung cancer, the predominant form ($\sim 85\%$) is non-small cell lung cancer (NSCLC). Other forms include small cell lung cancer (10-15%) and lung carcinoid tumors($<5\%$).

NSCLCs consist of three major subtypes: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and large-cell lung cancer. LUAD and LUSC together account for about 65-70% of all NSCLCs [4]. Disparate molecular pathways and cells of origin have been implicated in these different subtypes. Smoking has a huge influence on lung carcinogenesis, particularly in LUSC, with cells of origin often located in the central airway. NSCLCs are also seen in non-smokers and often are LUADs that develop in the peripheral airways. Some of the earliest mutations implicated in NSCLCs are within oncogenes such as *EGFR* and *KRAS*, as well as mutations and loss of heterozygosity (LOH) of chromosome arms such as 3p, 9p and 17p spanning tumor suppressor genes *CDKN2A* and *TP53* [4]. These mutations show preferential abundance based on smoking patterns and histology. For example, *KRAS* mutations have found be predominant among smokers and 3p LOH is more prevalent in LUSCs [4]. Large studies such as those by The Cancer Genome Atlas (TCGA) have identified several other mutational patterns and gene expression changes specific to these different subtypes of NSCLCs [5, 6].

In spite of the progress made in understanding the genetic and molecular aberrations in NSCLCs, these tumors continue to account for high rates of mortality. This is largely attributed to their late diagnosis [4]. The National Lung Cancer Screening Trial has evaluated the importance of low-dose computerized tomography (CT) in screening for lung cancers; the study identified that LUADs and LUSCs were detected more frequently at

the earliest stage by low-dose helical CT compared to standard chest X-rays [7]. However, advances (e.g., molecular-based) in early detection and prevention of NSCLC have been limited by a poor understanding of early changes in the pathogenesis of this tumor [8]. This necessitates studies that examine precancerous stages as well normal tissues that precede the development of tumors.

1.1.2 Field cancerization in lung cancer development

The phenomenon of field cancerization was first proposed by Slaughter *et al.* following the observation that the epithelium adjacent to oral tumors exhibited histological abnormalities, it was proposed that these patches of abnormal tissues could lead to multiple primary tumors as well as local recurrence [9]. Although first observed in oral cancers, the concept of field cancerization was later extended to cancers of other organs, such as esophagus, skin, stomach, colon, cervix and lung [9, 10].

Exposure to mutagens and age-related random mutations from errors in the natural process of DNA replication are thought to initiate a mutant lineage that might ultimately lead to cancerized fields [10]. Under the principles of field cancerization, this mutant lineage may acquire additional mutations and become phenotypically strong to survive and expand in that microenvironment. This may in turn give rise to patches of genetically distinct clones that continue developing, especially in the presence of continued carcinogen exposure. Some of these patches may eventually progress to a neoplasm, thus giving rise to a cancerized field. The cancerized field consists of cells that are further along on the evolutionary path of cancer development, that may precede observable histopathological abnormalities. Through the continuous expansion of these cancerized fields, premalignant and malignant phenotypes may result [10].

In lung cancer, previous studies have identified mutations and gene expression changes in normal airway epithelia. For example, a study of histologically normal epithelia and mildly abnormal premalignant lesions of LUSCs identified multiple sequentially occurring chromosomal loss of heterozygosity (LOH) events, such as of chromosome arms 3p and 9q, that increased in frequency based on the severity of histological changes in the multistage progression to LUSCs [11]. Another study identified additional deletions at chromosomal

regions 2q35-q36 and 12p12-p13 in matched histological normal bronchial epithelium and tumor tissue of NSCLC in long-term smokers [12]. A recent multi-region study by our group provided a genome-wide landscape of chromosomal aberrations, including frequent and recurrent events on chromosome 9 in the airway epithelium of early-stage NSCLC patients [13]. Apart from chromosomal changes, studies have also observed gene expression changes in the cancerization field of normal-appearing airway including those shared with adjacent NSCLCs [14, 15, 16], thus suggesting the utility of airway gene expression in early lung cancer diagnosis among intermediate-risk patients. In comparison to chromosomal aberrations and gene expression, studies of gene mutations in the normal bronchial epithelium have been limited. One of the first few observations of point mutations in non-malignant tissue, involved the *KRAS* oncogene, in non-malignant and matched NSCLC tumors of smokers [17]. Around the same time, another group reported widespread *TP53* mutations in the bronchial epithelia of a cancer-free individual [18]. Later, mutations in *EGFR* were detected in normal respiratory epithelium of 43% of patients with *EGFR*-mutant adenocarcinomas [19]. More recent studies have identified epigenetic alterations in the normal bronchial epithelium of lung cancer cases and cancer-free smokers such as methylation of *CDKN2A*, *DAPK*, *RAR-2* and *GSTP1* [20, 21, 22]. Although often limited to specific known lung cancer drivers and targeted regions, these studies shed light on the molecular alterations in histologically normal tissues.

1.1.3 Premalignant lesions of lung cancers

Lung cancers are thought to result from a series of progressive pathological changes, comprising several precursor lesions in the respiratory mucosa. Although many studies have identified molecular abnormalities in NSCLCs, much fewer studies have focused on exploring the molecular characteristics of early precancerous lesions that precede overt lung tumors. Among the different subtypes of lung cancer, most is known about the sequential progression of centrally arising LUSCs, while premalignant lesions of other subtypes such as LUADs, large cell carcinomas or SCLCs still remain poorly documented.

Among NSCLCs, due to the central origin of LUSCs, serially sampled large bronchi have helped identify a series of changes from hyperplasia (either basal cell or mucous),

squamous metaplasia, different degrees of dysplasia to carcinoma *in situ* (CIS) and invasive squamous oriented cancer [23]. Since then, studies have described molecular aberrations in these different lesions along the development of LUSCs, starting with early loss of 3p and 9p, followed by LOH events spanning 8p and 17p [11, 24]. Methylation of *CDKN2A* has also been observed in LUSC premalignancy, with a frequency that increased with histopathologic progression [25]. More recent studies have identified differentially expressed genes in premalignant lesions of lung squamous tumors, including the increased expression of *SLC2A1*, *CEACAM5*, and *PTBP3*, activation of *PI3K* and *MYC* [26, 27], as well as depletion of interferon signaling, antigen presentation and immune cells in these lesions [28].

In comparison to the well characterized multi-step progression in the development of LUSC, little is known about the premalignant stages of LUAD, the other major subtype of NSCLCs. Atypical adenomatous hyperplasia (AAH) is the only known premalignant lesion, with the pathogenesis of many adenocarcinomas being largely unknown [8]. Other low-grade lesions such as adenocarcinoma *in situ* (AIS) and minimally invasive adenocarcinomas (MIA) have been shown to occur prior to development of invasive LUADs. There is a suggested continuum of these lesions in the development of LUADs from AAHs; this was later shown to not be the case for all LUADs, thereby allowing for non-linear progression mechanisms [29]. Very few studies have characterized these premalignant lesions. Mutations in *KRAS* and *EGFR* have been identified in premalignant, preinvasive and invasive lesions of LUADs [30, 31, 32]. It was further shown that *KRAS* mutations were more frequent in premalignant lesions than LUADs, while *EGFR* mutations showed similar frequencies across the different stages of LUAD pathogenesis [32]. Some AAH lesions have demonstrated LOH of chromosomes arms, including 3p, 9p, 9q, 17p, and 17q; these aberrations have also been commonly found in invasive adenocarcinomas [33, 34]. Small, targeted studies of gene expression have also identified changes occurring early in the development of AAH and LUADs. For example, the loss of *STK11* was associated with dysplasia, thereby suggestive of its role in the malignant transformation of AAHs to LUADs [35]. Other expression changes observed in AAH lesions include the activation of *NKX2-1*, cyclin D1, survivin and an RNA-binding protein called *hnRNP B1*, as well as loss of p16 (*CDKN2A*) [36, 37, 38]. Another study investigated the role of epigenetic modifications in LUAD de-

velopment; the authors found elevated DNA methylation at *CDKN2A* and *PTPRN2* in AAHs, while also describing epigenetic changes that occur at later stages, such as in AIS and invasive LUADs [39]. A recent targeted next-generation sequencing (NGS) study of multi-focal AAH, AIS and MIA lesions identified mutations in *KRAS*, *TP53* and *EGFR* as indicators of malignant transition to invasive adenocarcinomas; the study further detected these early mutations in paired circulating DNA [40]. These studies not only point to the molecular complexity of AAH, but highlight the challenges in studying their role as precursors to LUAD, which still remains largely unknown.

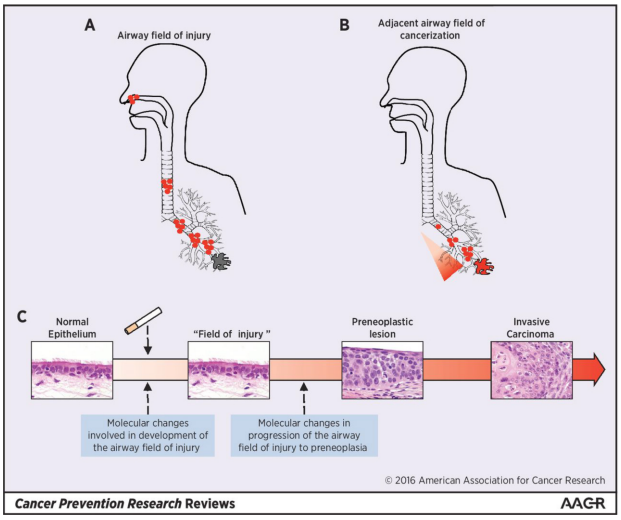


Figure 1.1: **Investigations of airway field cancerization and premalignancy in lung cancer pathogenesis.** Kadara and colleagues highlight the importance of understanding molecular changes involved in the development of the cancerized airway field as well as their progression to premalignant and malignant phenotypes that would aid the early detection and treatment of lung cancers. (*H Kadara, P Scheet, I. I. Wistuba, and A. E. Spira, Early events in the molecular pathogenesis of lung cancer, Cancer Prevention Research, vol. 9, no. 7, pp. 518 - 527, 2016. Permission obtained through the Copyright Clearance Center*).

1.2 Objectives

Our long term goal is to better characterize the evolution of tumors by studying apparent normal and premalignant tissues in early-stage cancer patients to identify potential biomarkers for the early detection of these tumors in intermediate-risk individuals (e.g., smokers) or predict targets for the prevention of these fatal tumor by better tackling preneoplastic

tissues, as illustrated in Figure 1.1.

This not only necessitates better profiling technologies, but also demands improved computational methods to push the limits of mutation detection to normal and premalignant lesions that exhibit low overall fraction of mutant cells.

A comprehensive annotation of the genomic landscape of mutational processes occurring in normal and premalignant tissues using a combination of microarray and next generation sequencing (NGS) methods is critical for enabling future therapies or preventive measures. With decreasing costs for NGS technologies, several tumor genomes, including those originating in the lung, have been assessed to infer comprehensive landscapes of genetic alterations and gene expression patterns, such as in The Cancer Genome Atlas (TCGA) consortium. We utilize the results from these large-scale tumor studies, particularly in NSCLCs, to compare our findings in normal or premalignant tissues and NSCLCs, that might elucidate insights into early events in the development of NSCLCs. We aim to achieve this through a cross-platform analyses, comprising multiple DNA and RNA-based technologies, and validating our findings *in silico* across these platforms. As such, our objectives are the following:

1. **To characterize atypical adenomatous hyperplasia and study its neoplastic progression to early-stage invasive lung adenocarcinomas**, by performing a multi-platform assessment comprising deep targeted DNA sequencing, broad-range SNP genotyping arrays and transcriptome sequencing of matched normal, AAH and LUAD tissues.

2. **To identify the mutational landscape of the cancerized field in the normal airway epithelium of early-stage NSCLCs**, by performing a genome-wide analysis of single nucleotide mutations (SNVs) as well as large chromosomal aberrations crucial to the evolution of NSCLCs from a cancerized field comprising multi-region samples of tumor-adjacent and distant airway epithelia.

3. **To extend our analysis to identify the pan-cancer landscape of chromosomal aberrations in the TCGA consortium**, by developing and implementing sensitive computational methods that can identify, compare and contrast our findings across

platforms as well as highlight potentially discordant results.

Mutational patterns in the normal airway epithelium and AAHs can help answer fundamental questions involving the pathogenesis of these tumors. Our data-driven genomics approach, with significant potential in predicting outcomes in high-risk patients, can lead to novel biomarker discovery and personalized chemo-preventive strategies that may delay or halt the tumorigenesis process in the earliest stages.

This dissertation explains the research study design, bioinformatic analyses devised and implemented, and biological interpretation of premalignant and normal-tissues in lung cancer patients as well as computational methods developed to compare our findings with the tumors in TCGA consortium. Give the collaborative nature of the study, my work focused on designing novel computational strategies for molecular characterization as well as interpreting the analyses for biological insights. In chapter 2, I describe our work on assessing the mutational and transcriptional landscape of matched normal, premalignant and tumor tissues from early-stage LUAD patients. In chapter 3, I outline our work on inferring tumor evolution from a cancerized field in early-stage NSCLCs using mutation profiles from a multi-region sampling of tumors and matched airway epithelia. In chapter 4, I report the computational methods we developed and implemented to compare and contrast our findings in the TCGA consortium as well as to extend our analysis to identify patterns of chromosomal aberrations across multiple tumor types. In chapter 5, I outline the significance of our findings, describe future research directions and discuss the utility of the computational methods that I developed and implemented over the course of this project.

CHAPTER 2
INVESTIGATION OF PREMALIGNANT LESIONS OF LUNG ADENOCARCINOMAS

Atypical adenomatous hyperplasia (AAH) is the only known precursor lesion to lung adenocarcinomas. AAHs are challenging to identify and are often captured incidentally, thus making molecular interrogations of these premalignant lesions limited. Due to the low cellular fractions of detectable somatic mutational processes within these premalignant tissues, little is known about its pathogenesis and progression to LUAD. Previously, mutations in known lung cancer drivers such as *KRAS*, *EGFR* and *TP53* as well loss-of-heterozygosity in targeted chromosomal arms such as 9p, 9q, 16p and 17q have been shown to occur in AAHs. There is also evidence for gene expression and epigenetic modifications early in AAH development. However, most studies of AAHs have been limited to specific targeted regions known to be important in LUAD pathogenesis, and therefore, the landscape of genomic and transcriptomic changes in AAH developments remains unexplored. Our goal was to perform a multi-platform assessment of matched normal, premalignant AAH and invasive LUAD specimens from early-stage patients to infer and integrate the landscape of somatic mutations and gene expression changes that inform the molecular pathogenesis of AAH.

In this chapter, I describe my work on identifying single nucleotide mutations from deep targeted DNA sequencing and genome-wide patterns of large chromosomal aberrations identified from broad-scale SNP genotyping arrays in AAHs. I also describe investigations of the gene expression changes occurring in initiation of AAH from normal tissues as well as those occurring later in their progression to LUADs. I conclude by proposing a list of candidate genes based on an integrative analysis that might suggest important roles in the initiation or progression of AAHs. The contents of this chapter are based on the following publications:

Sivakumar S, Lucas FAS, McDowell TL, Lang W, Xu L, Fujimoto J, Zhang J, Futreal PA, Fukuoka J, Yatabe Y, Dubinett SM, Spira AE, Fowler J, Hawk ET, Wistuba II, Scheet

P, Kadara H. Genomic landscape of atypical adenomatous hyperplasia reveals divergent modes to lung adenocarcinoma. Cancer Research 2017;77(22):611930.
Copyright permissions are not required, since AACR states Authors of articles published in AACR journals are permitted to use their article or parts of their article in the following ways without requesting permission from the AACR - Submit a copy of the article to a doctoral candidate's university in support of a doctoral thesis or dissertation.

Sivakumar S*, Lucas FAS*, Jakubek Y, McDowell TL, Lang W, Kallsen N, Peyton S, Davies GE, Fukuoka J, Yatabe Y, Zhang J, Futreal PA, Fowler J, Fujimoto J, Ehli EA, Hawk ET, Wistuba II, Kadara H, Scheet P. Genomic landscape of allelic imbalance in premalignant atypical adenomatous hyperplasias of the lung. *In Press.* Ebiomedicine 2019.
Ebiomedicine is part of the Lancet group and states that all requests to reproduce or make available anything in the journalin whole or in part, in electronic or in any other form, including translationshould be made through Elsevier. Elsevier states that "Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes."

2.1 Study design

I present a characterization of single nucleotide mutations, gene expression and chromosomal allelic imbalance profiles in the pathogenesis of AAHs by studying these premalignant lesions along with normal lung parenchyma tissues (NL) and primary LUADs from a cohort of 23 patients with early-stage LUAD (Figure 2.1). Recurrently mutated genes and chromosomal aberrations that encompass known lung cancer driver genes were further assessed for shared (AAH and LUAD) and tissue-specific (AAH or LUAD only) patterns. Gene expression changes were also interrogated to assess their role early in AAH development as well as for those occurring later in progression of AAH to LUAD.

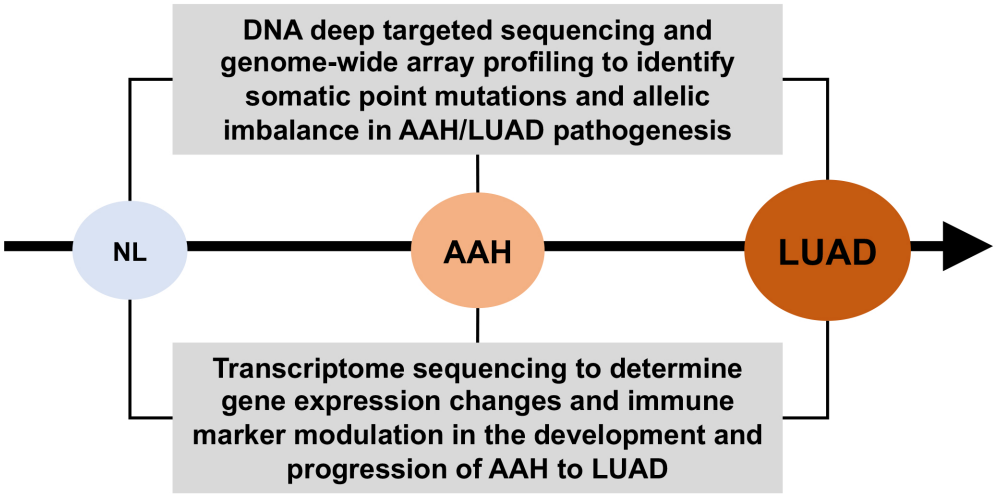


Figure 2.1: **Study design to understand the development and progression of adenomatous atypical hyperplasia.** A two-pronged approach, consisting of DNA-based profiling and transcriptome sequencing were used to study the pathogenesis of AAH.

2.2 Methods

2.2.1 Cohort

Normal lung parenchyma tissues (NL), AAHs, and LUADs were acquired from 23 patients with early-stage LUAD who were evaluated at the Aichi Cancer Center (Nagoya, Japan) and Nagasaki University (Nagasaki, Japan). Specimens were approved for study by Institutional Review Boards and according to the international ethical guidelines for biomedical research involving human subjects (Council for International Organizations of Medical Sciences). Informed written consents were received from all subjects wherever necessary. Clinicopathologic features of these patients are summarized in Table 2.1. The diagnosis, specimen collection and slide preparation were carried out between 2011 and 2015 for all patients. The AAH lesions were incidental and identified by radiological imaging. Specimens were obtained formalin-fixed and paraffin-embedded (FFPE) and stained by hematoxylin and eosin (H&E). Assessment of histopathology of AAHs and LUADs was performed by analy-

sis of H&E stained alternating slides (from 5 micron sections) with sections in between (10 micron) preserved for RNA isolation. Normal lung was taken from resected areas and was confirmed histopathologically following H&E staining to be consistent with normal tissue devoid of preneoplastic or neoplastic cells. Tissues were pathologically examined following the World Health Organization on the classification of lung tumors in the report by Travis and colleagues [41]. Images of the lesions were scanned using the Aperio platform (Leica Biosystems). A tabulation of cases and samples analyzed by DNA-sequencing (22 cases), transcriptome sequencing (17 cases) and SNP genotyping arrays (16 cases) is provided in Table 2.2.

Case	**Age**	**Gender**	**Tobacco history**	**Stage**
1	70	Male	Ever	IA
2	21	Female	Ever	IB
3	67	Male	Ever	IA
4	46	Female	Ever	IA
5	79	Female	Never	IA
6	40	Male	Ever	IA
7	72	Male	Never	IA
8	48	Female	Never	IA
9	51	Female	Never	IB
10	81	Female	Never	IA
11	67	Male	Ever	IA
12	63	Female	Never	IA
13	79	Female	Ever	IA
14	54	Female	Never	IA
15	62	Male	Ever	IA
16	64	Female	Never	IA
17	67	Male	Ever	IA
18	57	Female	Never	IA
19	71	Male	Ever	IB
20	63	Male	Ever	IIIA
21	74	Female	Never	IA
22	60	Male	Ever	IA
23	65	Male	Ever	IB

Table 2.1: Clinicopathological features of the cohort comprising matched normal lung parenchyma, AAH and LUAD

	DNA targeted sequencing			Transcriptome sequencing			SNP genotyping arrays		
Case	**Normal**	**AAH**	**LUAD**	**Normal**	**AAH**	**LUAD**	**Normal**	**AAH**	**LUAD**
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	0	0	0
4	1	1	1	1	1	1	0	0	0
5	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1
8	1	1	1	0	1	1	0	0	0
9	1	1	1	1	1	1	0	0	0
10	1	1	1	1	1	1	1	1	1
11	1	1	1	0	0	0	1	1	1
12	1	1	1	0	1	0	0	0	0
13	1	1	1	0	0	0	1	1	1
14	1	1	1	1	1	1	0	0	0
15	1	1	1	0	0	0	1	1	1
16	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1	1
19	1	1	1	0	0	0	1	1	1
20	1	1	2	1	1	1	0	0	0
21	1	1	1	0	0	0	1	1	1
22	1	1	1	1	1	1	1	1	1
23	0	0	0	0	0	0	1	1	1
Total:	22	22	23	15	17	16	16	16	16

Table 2.2: Samples analyzed by DNA sequencing, transcriptome sequencing and SNP genotyping arrays.

2.2.2 DNA and RNA isolation

DNA/RNA was extracted from all samples using the AllPrep DNA/RNA FFPE kit from Qiagen and suspended in nuclease free water (RNA) or AE buffer (DNA). 5 to 15 sections/slides per specimen were deparaffinized prior to isolation of normal tissues and lesions by scraping with 25-gauge needles under a stereomicroscope. Tissue fragments were then collected in 1.5 ml self-lock tubes containing 100 to 200 μl of lysis buffer PKD (Qiagen). Sample concentrations were measured on a NanoDrop 1000 (Thermo Fisher Scientific), and RNA integrity numbers indicative of overall quality were obtained on the 2100 Bioanalyzer (Agilent Technologies) using the RNA 6000 Nano or Pico kit according to the manufacturers protocol. DNA was quantified using the Quant-iT PicoGreen double stranded DNA (dsDNA) kit (Thermo Fisher Scientific) according to the manufacturers instructions.

2.2.3 DNA targeted sequencing

The Ion AmpliSeq Comprehensive Cancer Panel (Thermo Fisher Scientific) comprising primers for 409 canonical cancer-associated genes and the AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific) were used to prepare barcoded libraries from the FFPE DNA samples. Target amplification was carried out in 5 μl reactions with 17 cycles of amplification. The pools were then combined for digestion and ligation using Ion Xpress barcode adapters (Thermo Fisher Scientific) according to the manufacturers protocol. The libraries were then quantified with quantitative PCR (qPCR) using the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific). Template reactions were prepared using the Ion PI Hi-Q OT2 200 Kit and the Ion PI Hi-Q Chef kit (Thermo Fisher Scientific) based on the commercial protocol. Templates were then assessed on a Qubit 2.0 fluorometer before loading onto Ion PI chip v3 (Thermo Fisher Scientific). Sequencing was performed on the Ion Torrent Proton platform according to the manufacturer's instructions. Specimens from two cases were processed together in one chip and sequenced on an Ion Proton sequencer. Sequencing reports generated in the Ion Torrent Suite 5.0 were used to assess the quality of the libraries and sequencing runs. Base calling results were aligned to the reference `hg19_ampliseq_transcriptome_ercc_v1.fasta` provided by the manufacturer. The

aligned BAM files were then used to run Torrent Variant Caller 5.0 (TVC) using manufacturers targeted region BED file to generate VCF files. The targeted DNA sequencing data files have been deposited in the sequence read archive (SRA) under Bioproject accession PRJNA398260. This is a subsection in Chapter 2.

2.2.4 Transcriptome sequencing

A subset of the cases (15 NLs, 17 AAHs, and 16 LUADs from 17 different cases) was selected for transcriptome sequencing based on specimen availability as well as transcriptome sequencing quality indicated by percentage of mapped reads and valid on-target reads. For library preparations, approximately 30 ng of RNA was used based on sample concentrations obtained from the Qubit HS RNA assay (Thermo Fisher Scientific). All samples were heat shocked at 80C for 10 minutes and cooled to room temperature for 5 minutes and then reverse-transcribed to generate cDNA libraries using the Ion AmpliSeq Transcriptome Human Gene Expression Kit (Thermo Fisher Scientific) adhering to the manufacturers protocol for FFPE samples with 16 cycles of target amplification. Library concentrations were determined by qPCR using an absolute quantitation method and the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific) following the manufacturers protocol. Template reactions were carried out using the Ion PI Hi-Q OT2 200 Kit (Thermo Fisher Scientific) according to the manufacturers instructions and then loaded onto Ion PI chips v3 using the Ion PI Hi-Q Sequencing 200 Kit based on the manufacturers protocol (Thermo Fisher Scientific). The Ion Torrent Suite 5.0 was used to assess the quality of the libraries and sequencing runs. For sequencing, specimens from two cases were processed together in one chip. All samples were sequenced on an Ion Proton sequencer. Raw transcriptome sequence data files have been deposited in the gene expression omnibus under data series GSE102511.

2.2.5 Genome-wide high-density array profiling

A subset of 16 cases (16 NLs, 16 AAHs, and 16 LUADs) were selected for genotype array profiling based on specimen availability. The extracted DNA was processed through the Infinium HD FFPE DNA Restoration protocol (Illumina Inc., San Diego, CA.) followed

by SNP genotyping using the Illumina Infinium Global Screening Array-24 v1.0 BeadChip array. Raw intensity files were analyzed with GenomeStudio Genotyping Module v2.0 (Illumina Inc., San Diego, CA.) to call genotypes, normalize and cluster data in order to obtain SNP metrics such as B-allele frequency (BAF) and log R ratio (LRR).

2.2.6 Strategy to identify somatic single nucleotide mutations

2.2.6.1 Using multiple variant calling algorithms

In the Ion Torrent server, results from base calling were aligned to a reference file `hg19_ampliseq_transcriptome_ercc_v1.fasta`, which produced aligned BAM files. An automated analysis was performed with Ion Torrent proprietary software Ion Reporter to call somatic variants in LUADs and AAHs by contrasting events in NLs. The two available tumor specimens from case 20, were pooled together for analysis. The following two programs were used to augment variant calls in AAHs and LUADs: MuTect, using the default settings with the exception of retaining those annotated with the tag `nearby_gap_events`; and Varscan2, using a minimum variant allele frequency (VAF) threshold of 0.01987 instead of its default of 0.2 and filtering variants with a P value <10-6. The VAF threshold for Varscan2 was reduced to identify a larger number of low frequency mutations given the high average depth of sequencing; the threshold of 0.01987 was derived based on the minimum VAF detected in Ion Reporter, the caller natively calibrated for this platform. Finally as a fourth caller, the VCF files generated marginally from Torrent Variant caller (TVC) were subjected to a simple subtraction of variants, removing those observed in the matched normal sample for each case. Instead of using all mutations detected by any of the four callers, for each sample, the stringency was increased to include mutations that were detected by at least two different callers, with the exception of those identified by Ion Reporter software and TVC since they are produced from the same source and are expected to share more mutations. Mutations in exonic, splicing, and untranslated regions (UTR) were assessed, focusing on exonic single-nucleotide variants within the targeted 409 cancer gene panel. Mutations from all four callers/methods were annotated using ANNOVAR and Oncotator.

A post-processing protocol was then applied to remove potentially-overlooked germline

variants based on their presence in the Exome Aggregation Consortium (ExAC). The ExAC data was used as it encompasses commonly used databases 1000 genomes project and NHLBI-GO Exome Sequencing Project. In addition to the matched normal, AAH and LUAD, mutations identified in other normal tissues were also excluded as potentially germline variants. Mutations within known lung adenocarcinoma cancer drivers [5] and other cancer associated genes [42] were further assessed for specific patterns in AAH and LUADs of our cohort.

2.2.6.2 PolyAna

An additional concern with Ion semiconductor sequencing technology is the potential false positives mutations due to reduced accuracy at loci with homopolymer repeats of the same nucleotide. Briefly, the sequencing chemistry works by releasing a hydrogen ion with the incorporation of every nucleotide in the DNA strand being sequenced. The released hydrogen ion results in a change in pH of the solution that is then converted to a voltage signal in the ion sensors. However, in regions of homopolymers, multiple identical bases often result in inaccurate measurement of the magnitude of voltage pulse, thus causing difficulty in accurately estimating the length of the homopolymer. Preliminary approaches have been implemented to better correct and align the sequencing data to avoid homopolymer derived mutations [43]. In order to be more stringent, in our study, we decided to exclude mutations that are in regions of homopolymers.

I developed **PolyAna**, a homopolymer filter that identifies homopolymers in the vicinity of the mutation based on its genomic location. The pseudocode for the program is shown in Figure 2.2. The script takes the following as arguments: the variant file to be annotated, reference fasta file, homopolymer length cut off, a window threshold and an output file name. Defaults for minimum homopolymer length and window size are set to 6bp and 10bp respectively.

For every mutation, it first computes a repeat length to identify if the mutation is in a repeat segment. Based on the repeat length, it runs the appropriate segment of the code. First, if the repeat length is between four and the homopolymer length cut off (default = 6bp), it annotates the variant as being in a homopolymer region if the the

```
For each mutation
{
        polyx=number of consecutive bases at position

        Part 1: Check ALT base
        if (polyx is between 4 and min_length)
        {
                If ALT matches adjacent base  { Result=Homopolymer }
        }

        Part 2: Check for homopolymer within window
        if (polyx < min_length)
        {
                A: Is there an adjacent Homopolymer?
                B: Is there a homopolymer in the vicinity?
                C: Are there small stretches of consecutive homopolymers?
                If A | B | C { Result=Homopolymer } Else { Result=Potentially_Valid }
        }

        Part 3 : If in a homopolymer
        else
        {
                if (middle of homopolymer) { Result = Homopolymer}
                else if (at end of a homopolymer)
                {
                                if (ALT matches adjacent base) {Result = Homopolymer }
                                else { Result = Potentially_Valid }
                }
        }
}
```

Figure 2.2: **Pseudocode for PolyAna.** A quality control step in the processing of Ion Torrent sequencing data to identify and remove potential homopolymer-derived false positive somatic mutations.

alternate allele matches the adjacent nucleotide base. Then, if the variant has a short repeat length, lesser than the homopolymer length cut off, and the alternate allele doesn't match the adjacent base, it looks for homopolymers in the window. This section of the code is divided into three components: (A) a test to check if there is a homopolymer immediately adjacent to the mutation, (B) a test to check if there is a homopolymer in the vicinity (i.e., within the window specified), and (C) a test to check for short stretches of homopolymers (e.g., AAACCC) in the vicinity as specified by the window. If either of these conditions are satisfied, the variant is annotated as being in the vicinity of the variant while the remaining mutations are treated as potentially valid hits. Among the remaining variants that have a repeat length greater than the homopolymer length cut off, if the mutation is in within a homopolymer segment or if the alternate allele matches the adjacent nucleotide, the variant is annotated as a homopolymer, leaving the remaining as potentially valid mutations. In this way, the mutations arising from a homopolymer region, or with a homopolymer or short stretches of homopolymers in the vicinity can be filtered out during quality control to retain only the potentially valid mutation calls. PolyAna is available at `http://scheet.org/software.html` and was published as a part of a recent publication [44].

2.2.7 Validation of somatic mutations using digital PCR

Somatic mutations identified in the driver genes *BRAF* and *KRAS* in AAHs as well as the samples (AAH and LUAD) carrying the *EGFR* p.L858R mutation as detected in DNA targeted sequencing were verified by digital PCR using the QuantStudio 3D system (Thermofisher, A26317) following the manufacturers protocol for use with a chip loader. TaqMan and Custom TaqMan SNP genotyping assays were used as probes for the mutations (Thermofisher 4351379 and 4332077). Samples with less than 30 ng of input DNA were given 7 cycles of pre-amplification using the Platinum PCR SuperMix High Fidelity (Thermofisher 12532016) and SNP genotyping assays as primers. Allele frequencies were obtained from analyzing the chip files in the QuantStudio 3D software available on the Thermofisher cloud.

2.2.8 Gene expression analysis

Transcriptomes were quantified from BAM alignment files generated in the Ion Torrent server using a expectation-maximization (E/M) algorithm based procedure [45]. Resultant gene-based counts were normalized, log (base 2) transformed, and corrected for batch effects using the R limma package [46]. Differentially expressed genes were identified by ANOVA based on a P <0.001, false discovery rate (FDR) of 0.01 and minimum of 2-fold change in at least one of three comparisons (AAH-NL, LUAD-NL and LUAD-AAH). The model incorporated specimen type as a fixed effect and with different patients considered random effects. Hierarchical clustering of the samples was performed in R using Pearson correlation. To understand immune signaling in the pathogenesis of AAH, an *a priori* list of 730 markers of immune response and function (nCounter PanCancer Immune Profiling Panel from nanoString technologies) were used. Housekeeping genes in this panel were excluded from analysis. The expression of these 730 immune markers was analyzed similar to the global gene expression analysis, but with a minimum 1.5-fold change required in at least one of the three comparisons mentioned above. For identifying genes differentially expressed among different groups of AAHs based on select mutation status, a P <0.01 threshold and a 1.5-fold change cut-off was used. Pathways, gene-network identification and gene set enrichment analyses were performed using the commercially available software Ingenuity Pathways Analysis (IPA).

In order to better visualize and assess the gene expression changes from NL to AAH to LUAD, I developed a trivial classifier composed of two one-sided t-tests (P <0.05) to identify eight disparate patterns. The two tests consisted of comparing NL and AAH, followed by comparing AAH and LUAD. The eight disparate patterns included gene expression changes that (i) progressively decrease from NL to AAH to LUAD, (ii) progressively increase from NL to AAH to LUAD, (iii) decrease from NL to AAH only with no change from AAH to LUAD, (iv) increase from NL to AAH only with no change from AAH to LUAD, (v) decrease from AAH to LUAD only with no change from NL to AAH, (vi) increase from AAH to LUAD only with no change from NL to AAH, (vii) increase from NL to AAH with a simultaneous decrease from AAH to LUAD, (viii) increase from NL to AAH

with a simultaneous decrease from AAH to LUAD. The patterns (vii) and (viii) are expected to be rare among these different patterns of expression changes. This classifier was applied to assess patterns in the global gene expression profiles as well as in the assessment of expression changes in a subset of immune markers.

Finally, TIMER [47] was used to reanalyze the gene expression data to estimate the abundance of six tumor-infiltrating immune cell subsets (B cells, CD4 T cells, CD8 T cells, macrophages, neutrophils, and dendritic cells) among all three tissue types (NL, AAH and LUAD).

2.2.9 Identification of subtle genome-wide allelic imbalance

Allelic imbalance (AI) was inferred using hapLOH, a method developed in our laboratory, to detect subtle patterns of BAFs at heterozygous markers consistent with a relative haplotype imbalance [48]. Using regions that exhibit deviations in their BAFs along with LRR intensities for markers within regions of AI, events were classified as gain, loss and copy-neutral loss of heterozygosity (cnLOH) as described previously [13]. Briefly, the event regions with LRR $\geq$ 0.05 were classified as gains while those with LRR $\leq$ -0.05 were classified as losses. Among the remaining calls, regions with BAF deviation of 0.1 or greater were classified as cnLOH. The event calls that were not classified into these three event types were marked as subtle AI. Subsequently, detected AI events were specifically tested for statistical evidence of existence in other samples from the same individual, using a binomial test of similarities between the two sample-specific haplotypes in putative excess (derived from the sample BAFs) within each event region.

For each patient, genomic regions under AI were compared between AAH and LUAD samples to identify regions that were either aberrant in both samples or only in one of the tissues. To describe heterogeneity between the samples, I then quantified proportions of the genome exhibiting AI in both samples and proportions of the genome harboring AAH or LUAD-specific AI. For each patient, markers profiled in the SNP genotyping array were annotated as either being in an event in AAH, LUAD or both tissues. Based on this, the proportion of markers within AI events in matched AAH and LUAD specimens were determined as shared events, while those specific to only one of the tissues were determined as

the proportion of AAH-specific and LUAD-specific events. In addition, for shared AI events between matched AAH and LUAD, the regions exhibiting over-representation of opposite haplotypes were identified using RECUR [49] and excluded since they might be suggestive of independent events or secondary events. Phylogenetic trees were then constructed using these shared and tissue-specific AI events between the LUAD and AAH for each patient using the *ape* package in R [50].

All patients profiled using SNP genotyping arrays (with the exception of patient 23) were also profiled for single nucleotide mutations (SNVs) using ultra-deep DNA targeted sequencing of a panel of 409 known tumor associated genes (Sections 2.2.3,2.2.6). Given the prevalence of *EGFR* mutations in this cohort, the samples exhibiting mutations in this oncogene were assessed for patterns in their overall genomic AI burden. The presence of *EGFR* p.L858R in the tumor sample of patient 23 was confirmed by digital PCR as described previously (Section 2.2.7). The identified AI events and SNVs in known oncogenes and tumor suppressors [5, 42] were assessed for patterns of two-hit mutations (AI and SNV) in AAH and LUAD as well as shared-first hit (AI or SNV) with LUAD-specific second hit (AI and SNV).

2.3 Comprehensive genomic and transcriptomic characterization of AAHs

2.3.1 Mutation profiling

Exonic single nucleotide mutations (SNVs) within a targeted cancer gene panel consisting of 409 cancer associated genes were identified in AAHs and LUADs of our cohort of 22 patients. A total of 67 samples were analyzed, with one case having two tumor tissues profiled. The mutations in AAH and LUAD were identified using the matched normal lung tissue as a comparator. All 45 AAHs and LUADs exhibited at least one somatic mutation (exonic, splicing or in UTRs) with a mean of 6.1 variants (min = 1, max = 19) in AAHs and 10.6 (min = 1, max = 60) in LUADs (Figure 2.3).

Non-smokers displayed an expected lower somatic mutation burden than ever-smokers in the LUADs (5.4 vs 14.5); however, they exhibited a similar burden in AAHs (6.4 vs 5.9; Figure 2.4).

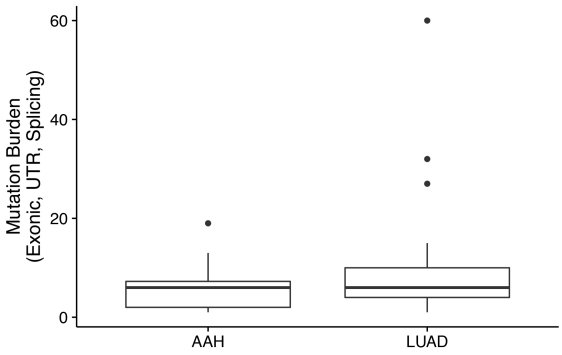


Figure 2.3: **Mutation burden in AAH and LUAD.** Somatic point mutations in exonic, splicing and UTR regions within the 409 genes sequenced in the panel were identified for the 45 specimens (22 AAH and 23 LUAD). Point mutation burdens for AAH and LUAD were plotted as boxplots.

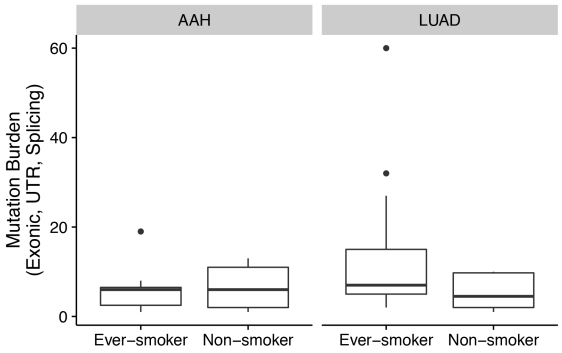


Figure 2.4: **Mutation burden in AAH and LUAD based on tobacco history.** Specimens (AAH and LUAD) were classified based on tobacco history (non-smoker and ever-smoker) in all 22 patients. Point mutation burdens for the tissues from non-smokers and smokers were plotted as boxplots.

To gain further insights into point mutations in the pathogenesis of AAH, nonsynonymous mutations in genes considered to be bona fide drivers of cancer were interrogated [42]. Mutations in genes previously determined by TCGA to be significantly recurrently altered in LUAD were also examined [42, 5]. Figure 2.5A shows 17 cases that exhibited a mutation in either driver gene set within their AAH specimens. The paired LUADs were also plotted depicting mutations in genes previously established by the TCGA to be significantly mutated in LUAD. Figure 2.5B describes a tissue level analysis of mutations in AAH and LUAD samples to identify mutated genes, from the same set of driver genes, that were common or disparate between AAH and LUAD. Figure 2.5C shows a lollipop plot for mutations in the *BRAF* oncogene and their prevalence in AAHs of our cohort.

AAHs from five patients (23%) exhibited somatic activating mutations in the *BRAF* oncogene (Figure 2.5A). Interestingly, the *BRAF* mutations were not detected in the paired LUAD specimens (Figure 2.5A; Figure 2.5B). Four of the five AAHs exhibited a *BRAF* p.K601E mutation, the other AAH contained a *BRAF* p.N581S variant (Figure 2.5C). *KRAS* was the second most recurrently mutated gene in AAHs (four cases, 18%; Figure 2.5A). All four *KRAS*-mutant premalignant tissues were from ever-smokers, in contrast to *BRAF*-mutant AAHs (from three non-smokers and two ever-smokers; Figure 2.5A). Also, AAH *KRAS* and *BRAF* mutations showed mutual exclusivity (Figure 2.5A). An interesting observation was that for four of the five cases (80%) with *BRAF*-mutant AAHs, their paired LUADs harbored activating mutations (three p.L858R in exon 21 and one p.S752F in exon 19) of the *EGFR* oncogene (Figure 2.5A). The other LUAD exhibited inactivating mutations in *KEAP1* and *STK11* tumor suppressors. LUADs of cases with the *KRAS*-mutant AAHs exhibited mutations in other drivers besides *KRAS* such as *TP53* (Figure 2.5A).

We further validated by digital PCR, the presence of all sequencing derived *BRAF* and *KRAS* mutations in AAHs, and the *EGFR* p.L858R mutation in both AAHs and LUADs that exhibited these patterns (Figure 2.5). Variant allele frequencies (VAFs) based on digital PCR were consistent with sequencing-based VAFs for these loci (Table 2.3).

TSC1 was the most frequently mutated tumor suppressor in AAHs (13.6%; two nonsense and one missense mutation; Figure 2.5A). We also noted other mutated oncogenes (*EGFR* and *JAK3*) and tumor suppressors (*CDKN2A* and *TP53*) in AAHs (Figure 2.5A).

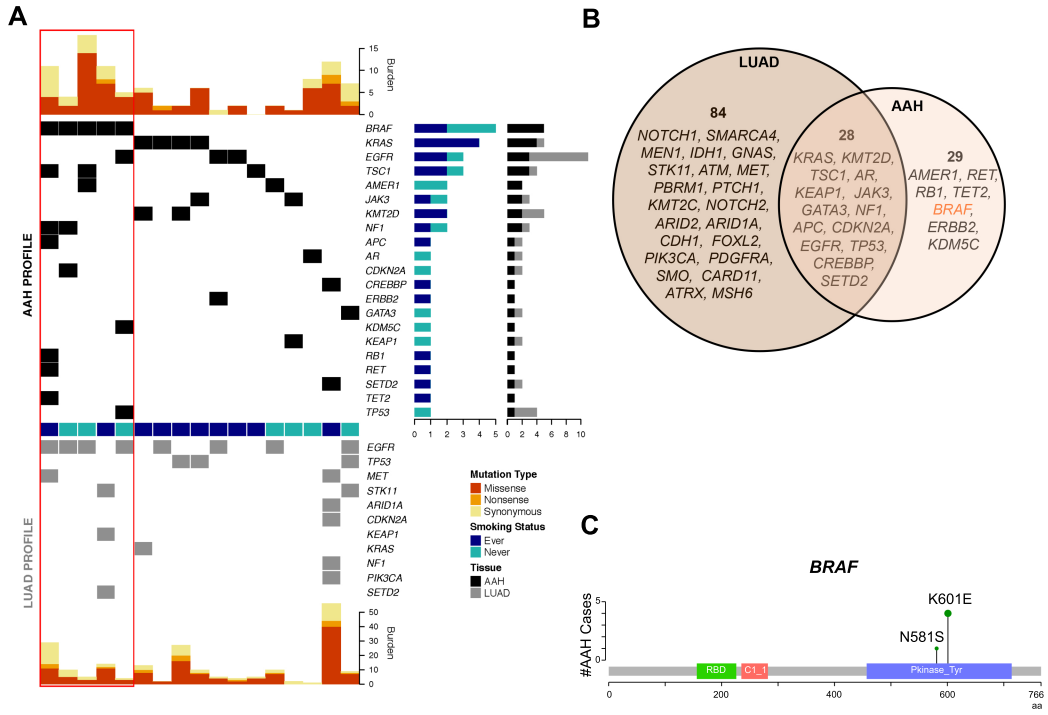


Figure 2.5: **Driver mutation profiles in AAH.** Somatic nonsynonymous mutations in AAHs and LUADs were identified. (A)Mutations in previously established cancer driver genes were examined. AAH specimens that exhibited a mutation in either driver gene set were plotted. The paired LUADs were also plotted depicting mutations in known LUAD driver genes. Shown within the red panel is the enrichment of *EGFR* mutations in LUAD paired to *BRAF*-mutant AAH. (B) A tissue level analysis of mutations in AAH and LUAD specimens was performed to identify mutated genes, from the same set of driver genes surveyed in panel A, that were common or disparate between AAH and LUAD. (C) Lollipop plot for mutations (p.K601E; n = 4 and p.N581S; n = 1) in the *BRAF* gene and their prevalence in AAH specimens.

Gene	**Mutation**	**Protein Change**	**Case**	**Tissue**	**Digital PCR assay**	**DNAseq derived VAF**	**Digital PCR derived VAF**
BRAF	chr7:140453134T>C	p.K601E	11	AAH	BRAF_478	0.3656	0.3853
BRAF	chr7:140453134T>C	p.K601E	12	AAH	BRAF_478	0.1609	0.1882
BRAF	chr7:140453193T>C	p.N581S	14	AAH	BRAF_462	0.0966	0.1208
BRAF	chr7:140453134T>C	p.K601E	15	AAH	BRAF_478	0.2197	0.2116
BRAF	chr7:140453134T>C	p.K601E	16	AAH	BRAF_478	0.2517	0.2251
EGFR	chr7:55259515T>G	p.L858R	2	LUAD	EGFR_6224	0.2139	0.1977
EGFR	chr7:55259515T>G	p.L858R	7	LUAD	EGFR_6224	0.1519	0.1753
EGFR	chr7:55259515T>G	p.L858R	12	LUAD	EGFR_6224	0.3157	0.2728
EGFR	chr7:55259515T>G	p.L858R	13	LUAD	EGFR_6224	0.0773	0.0873
EGFR	chr7:55259515T>G	p.L858R	14	LUAD	EGFR_6224	0.2602	0.2558
EGFR	chr7:55259515T>G	p.L858R	16	AAH	EGFR_6224	0.2087	0.2208
EGFR	chr7:55259515T>G	p.L858R	16	LUAD	EGFR_6224	0.2681	0.2845
EGFR	chr7:55259515T>G	p.L858R	18	LUAD	EGFR_6224	0.2379	0.2456
EGFR	chr7:55259515T>G	p.L858R	22	AAH	EGFR_6224	0.1863	0.1979
KRAS	chr12:25398284C>G	p.G12A	1	AAH	KRAS_522	0.3499	0.3293
KRAS	chr12:25398285C>A	p.G12C	4	AAH	KRAS_516	0.1492	0.132
KRAS	chr12:25398285C>A	p.G12C	19	AAH	KRAS_516	0.1078	0.121
KRAS	chr12:25380275T>A	p.Q61H	20	AAH	KRAS_555	0.1496	0.1571

Table 2.3: Validation of specific *BRAF*, *KRAS* and *EGFR* mutations using digital PCR. Variant allele frequencies (VAF) in both platforms are shown. DNAseq derived VAF corresponds to the VAF derived from deep targeted DNA sequencing.

	Mutated Genes
Common	*APC, AR, CDKN2A, CREBBP, CSMD3, DST, EGFR, EPHA3, ESR1, GATA3, GUCY1A2, JAK3, KAT6B, KEAP1, KMT2D, KRAS, LPHN3, LRP1B, NF1, NLRP1, NUP214, PBX1, PIK3CG, PTPRD, SETD2, SYNE1, TP53, TSC1*
AAH-specific	*AMER1, AURKC, BRAF, CASC5, CDKN2B, CRTC1, CYP2C19, ERBB2, ERBB3, FLI1, GPR124, IKBKB, IRS2, ITGB3, KDM5C, MDM4, MTOR, MYH11, NIN, PKHD1, RB1, RET, SH2D1A, SOX11, TAF1L, TCF3, TCF7L1, TET2, TFE3*
LUAD-specific	*ABL2, ADAMTS20, AFF1, AKAP9, ARID1A, ARID2, ARNT, ATM, ATR, ATRX, BCL11A, BCL11B, BTK, BUB1B, CARD11, CDH1, CDH2, CDH20, CMPK1, COL1A1, CRKL, CTNNA1, DCC, DICER1, EP400, ERBB4, ERCC1, FH, FLT1, FOXL2, FOXP1, GNAS, GRM8, HOOK3, IDH1, IL7R, ITGA9, KMT2A, KMT2C, LCK, MAGI1, MARK1, MEN1, MET, MN1, MSH6, MTR, MTRR, MUTYH, MYC, NBN, NCOA2, NOTCH1, NOTCH2, NOTCH4, NTRK3, PAK3, PARP1, PAX3, PBRM1, PDGFRA, PDGFRB, PIK3C2B, PIK3CA, PMS1, POT1, PTCH1, PTPRT, RECQL4, RUNX1T1, SAMD9, SEPT9, SMARCA4, SMO, STK11, TAF1, TET1, THBS1, TLR4, TNK2, TPR, USP9X, WAS, ZNF521*

Table 2.4: Genes mutated in AAHs and LUADs.

Additionally in this cohort, 28 genes were mutated in both AAHs and LUADs (e.g., *KRAS*, *TP53*, *KEAP1*, *CDKN2A*), 84 were found only in LUADs (*STK11*, *PIK3CA*) and 29 were found only in the preneoplastic lesions (*AMER1*, *BRAF*, *KDM5C*, *ERBB2*) (Figure 2.5B; Table 2.4). Even for genes that were shared between AAH and LUAD tissues, there were notable examples of differential frequency (e.g., *KRAS* more common in AAH; *EGFR* and *TP53* more common in LUAD; Figure 2.5A). On further examination of these 28 shared genes, we found that *EGFR* and *KAT6B* exhibited the same mutations in both tissues. There was also an enrichment of different mutations in codon 12 of *KRAS* in both the tissues. Our findings underscore subgroups of AAH with different mutated driver genes (*BRAF* vs *KRAS*) suggestive of potentially different mechanisms in the pathogenesis of these premalignant lesions.

2.3.2 Expression profiles in the development and progression of AAH

Next I sought to characterize expression profiles signifying the development of AAH from normal lung tissue (NL), and its progression to LUAD. Transcriptome sequencing of a subset of the cases and samples (Table 2.2) using a capture method targeting over 20,000 Refseq genes.

2.3.2.1 Global gene expression patterns

I identified 1,008 genes differentially expressed in at least one of three tissue types (Figure 2.6). Using one-sided t-tests to interrogate the two-step (NL to AAH and AAH to LUAD) modes of differential expression, I identified eight patterns or clusters of expression among the global gene expression profiles (Figure 2.6). These consisted of the following: decrease (n = 214) from NL to AAH and from AAH to LUAD; increase (n = 204) from NL to AAH and from AAH to LUAD; decrease (n = 116) and increase (n = 146) from NL to AAH alone with no change from AAH to LUAD; decrease (n = 85) and increase (n = 126) from AAH to LUAD alone with no change from NL to AAH and, less prevalent forms with no net change in expression such as an increase (n = 33) or decrease (n = 84) in AAH alone relative to other tissues. A pathway based enrichment analysis for genes in each cluster further identified potentially altered signaling pathways (Figure 2.6). This analysis pinpointed decreased anti-tumor T-helper (Th1) immunity, and conversely, increased pro-tumor Th 2-based immune response and signaling in both phases, the development of AAH from NL and their progression to LUAD. Inhibition of IFN-γ and TGFB1 signaling occurred early in AAH, when compared to NLs, and reduced surfactant protein signaling occurred thereafter in LUAD only. Pathway and gene set enrichment analysis also revealed an activation of B-cell receptor, *CSF2* (indicative of pro-tumor immune response), *MYC* and *ERBB2* signaling in AAH and LUAD (Figure 2.6). Activation of WNT and β-catenin signaling as well as modulation of gene sets associated with increased immune cell (phagocytes) migration were activated in AAH relative to NL (Figure 2.6). Gene sets associated with enhanced cell cycle and proliferation as well as reduced apoptosis were modulated in LUAD relative to AAH or NL.

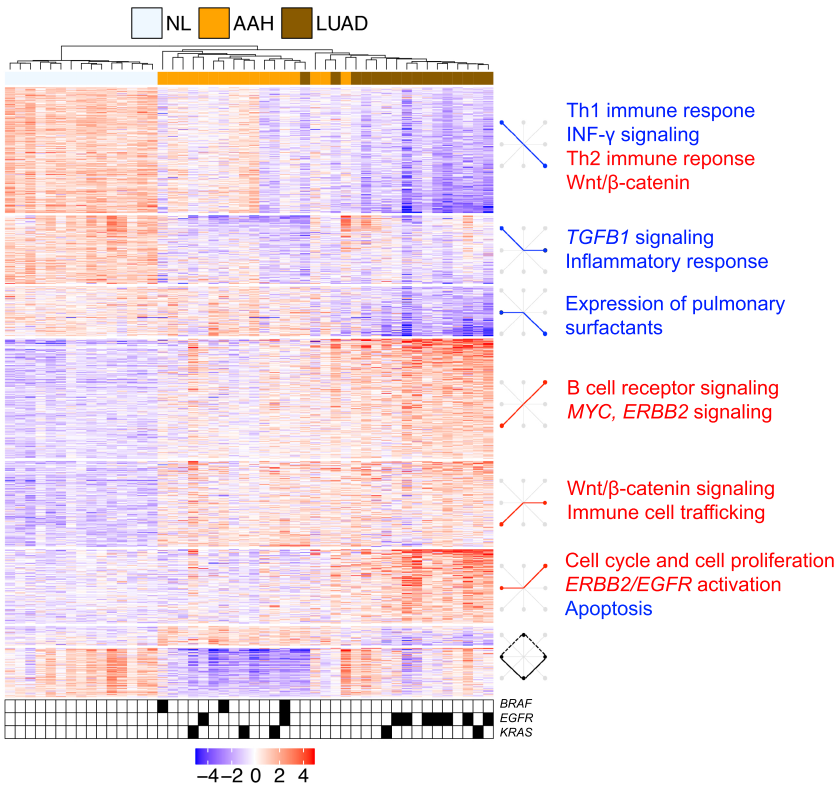


Figure 2.6: **Expression profiles differentially modulated in development of AAH and LUAD.** Genes (n = 1008) differentially expressed between the three tissues (AAH vs. NL, LUAD vs. NL, or LUAD vs. AAH) were analyzed by hierarchical clustering (red, upregulated relative to median sample; blue, downregulated relative to median sample). Genes were grouped into eight different patterns, with patterns of differential expression in each gene cluster schematically depicted on the right. Pathways and gene set enrichment analysis were performed and pathways deregulated in each cluster of genes are depicted in red (activation) and blue (inhibition) alongside the heatmap. Mutations status of *EGFR*, *KRAS*, and *BRAF* for AAH and LUAD specimens is depicted below.

Then, I compared and contrasted gene expression among three groups of AAHs based on driver gene mutation status identified above: *BRAF* mutant, *KRAS* mutant and *BRAF*/*KRAS* wild type (WT). 327 differentially modulated genes were observed between the three different groups of AAHs (Figure 2.7). Accordingly, these gene features indeed clustered the three groups separately based on the driver mutation status but with *BRAF*- and *KRAS*-mutant AAHs grouped closer together than with *BRAF*/ *KRAS* WT AAHs (Figure 2.7). Among the genes that were enriched in the *BRAF*-mutant AAHs were the cytokinesis promoting gene *KIF5C* and the cell proliferation promoting transcription factor (*MYC* Associated Factor X) *MAX*, typically associated with *MYC* oncoprotein[51](Figure 2.7). On the other hand, *KRAS*-mutant AAHs displayed up-regulated expression of tumor necrosis factor receptor superfamily members 9 and 10B (*TNFRSF9* and *TNFRSF10B*), the *NF-κB* subunit *RELB* and the proliferation promoting ubiquitin ligase *UBE2C* (Figure 2.7). Of note, both *BRAF*-mutant and *KRAS*-mutant AAHs exhibited suppressed expression of the epithelial mesenchymal transition-promoting tyrosine kinase receptor *AXL* relative to *BRAF*/*KRAS* WT AAHs (Figure 2.7). These findings suggest shared and disparate expression programs among AAHs with activating mutations in the oncogenic GTPases *BRAF* and *KRAS*.

2.3.2.2 Immune marker gene expression profiling

Accumulating evidence suggests a pivotal role for the host immune response in the evolution of cancer as well as dynamic interplay between emerging tumor cells and immune-based expression programs [52, 53]. We sought to begin to understand contextual immune marker profiles in the development and progression of AAHs. Among the global gene expression profiles, genes with known roles in immune signaling based on an annotated and *a priori* list were assessed. I identified 131 markers of immune response that were differentially modulated among NLs, AAHs and LUADs (Figure 2.8).

Overall, the immune markers followed similar patterns or clusters of expression described above. This analysis revealed that *IL12A*, a cytokine most notably associated with an anti-tumor immune response [54], was decreased in AAHs and LUADs relative to NLs (Figure 2.8). Conversely, the cytokines *CXCL13* and *CXCL14*, indicative of activated

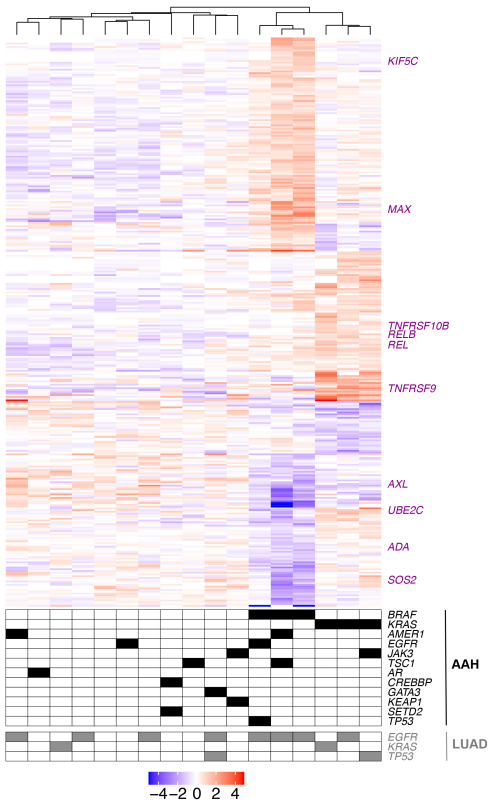


Figure 2.7: **Differential gene expression based on driver mutation status in AAH.** AAHs were subgrouped based on *BRAF* and *KRAS* mutation status: *BRAF*-mutant, *KRAS*-mutant, and *BRAF*/*KRAS* wild-type. Genes (n = 327) differentially expressed between the three AAH subgroups were identified and analyzed by hierarchical clustering (red, upregulated relative to median sample; blue, downregulated relative to median sample).

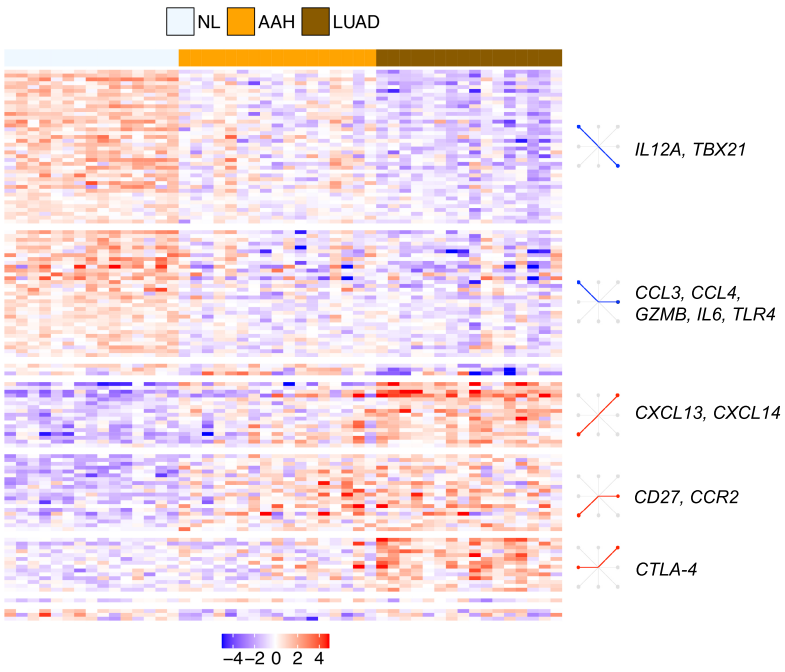


Figure 2.8: **Deregulation of immune signaling in the molecular pathogenesis of AAH.** Expression profiles for an *a priori* list of immune markers was compiled and studied to identify differentially expressed immune genes (n = 131). The genes were divided into different clusters based on patterns of differential expression between NL, AAH, and LUAD. Patterns of differential expression in each gene cluster are schematically depicted on the right. Select immune markers present in the major clusters are also depicted on the right.

B-cell chemotaxis and signaling [55, 56], were up-regulated in AAHs and LUADs (Figure 2.8). Moreover, aberrant immune marker expression occurred early in AAHs, relative to NLs (Figure 2.8). I found early and significantly decreased expression of prototypical markers of the anti-tumor immune response (e.g., *GZMB*) in AAHs relative to NLs (Figure 2.8). On the other hand, AAHs exhibited increased expression of the tumor-supportive chemokine receptor *CCR2* (Figure 2.8) [57]. Of note, I found that the major immune checkpoint cytotoxic T-lymphocyte-associated antigen 4 (*CTLA-4*) [58] was significantly up-regulated in LUADs relative to AAHs but not in the premalignant lesions relative to NLs (Figure 2.8) suggesting that aberrant immune checkpoint function by *CTLA-4* may be implicated in progression of AAH to LUAD. These findings accentuate the role of aberrant immune function and signaling early on in the development and progression of AAH.

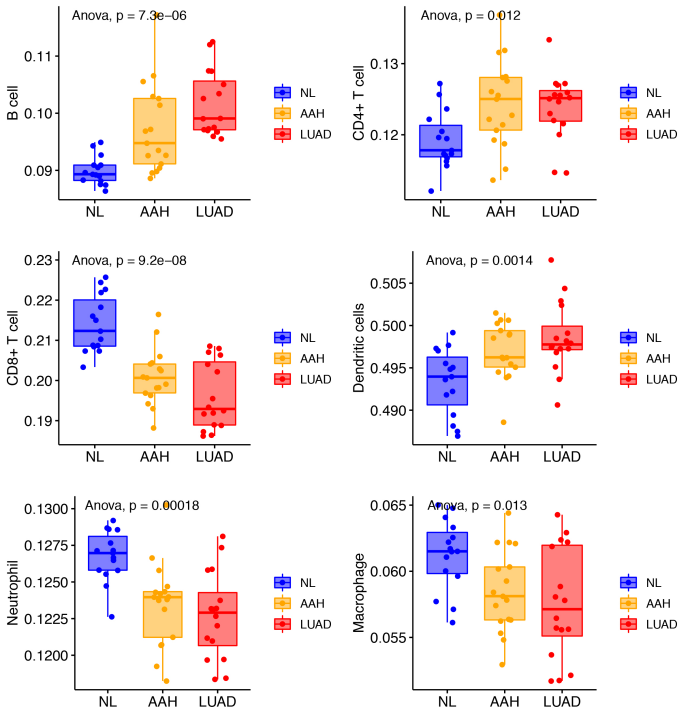


Figure 2.9: **Abundance of tumor-infiltrating immune cells in the different tissues.** Expression profiles were used to estimate the abundance of six tumor-infiltrating immune cells. Patterns of their abundance across the three tissue types (NL, AAH and LUAD) are depicted as boxplots.

The gene expression data was also analyzed to estimate the abundance of tumor-infiltrating immune cells using TIMER [47]. The results are summarized in Figure 2.9.

Particularly, our analysis revealed progressively and significantly increasing levels of B cells and CD4+ T cells from NL to AAHs and finally LUADs. In contrast, we identified significantly decreasing levels of CD8+ T cells from NL to AAH to LUAD. These results point to the relevance of immune activation early in the development of AAHs, perhaps even prior to the occurrence of mutations in these preinvasive lesions.

2.3.3 Chromosomal instability in AAH and LUAD

A haplotype-based computational framework, hapLOH [48], was used to infer subtle genome wide AI, including alterations present at lower cellular fractions. Among the 48 samples in our cohort profiled using SNP genotyping arrays, AI was detected in nine AAHs (56%), 15 LUADs (94%) and four normal lung parenchyma tissues (25%) (Figure 2.10).

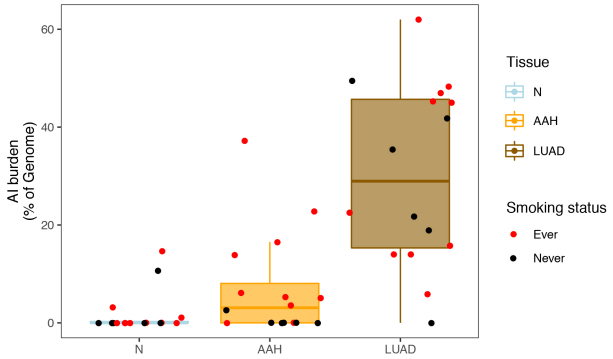


Figure 2.10: **Chromosomal allelic imbalance burden in normal, AAH and LUAD tissues.** Regions with subtle chromosomal allelic imbalance (AI) were identified in the normal (N), AAH and matched LUAD tissues using genome-wide genotype array profiling. AI burdens, defined as a percent of the genome, are represented by box plots for each tissue type (N, AAH and LUAD). The burden for each patient is shown as a point overlaid on the boxplots. The points are colored red if the patient had a smoking history and black if the patient was a non-smoker.

I identified 53 chromosome-arm AI events ($\geq$ 50% of chromosomal arm) and 19 focal AI events (<50% of chromosomal arm) in AAHs; and 210 arm-level AI events and 97 focal events in LUADs. Overall, the detectable AI burden (defined as a percent of genome exhibiting AI) in AAHs was significantly lower than LUAD (Wilcoxon, P value=0.0002; Figure 2.11). While AAHs showed significantly higher AI burden in lifetime smokers compared to non-smokers (Wilcoxon, P value = 0.005), their matched LUADs showed similar

distributions of AI burden between non-smokers and smokers (Figure 2.10, Figure 2.11). Of note, the AI burdens of *EGFR*-mutant non-smoker LUADs were larger than those of smokers as well as non-smoker LUADs without the mutation (Figure 2.11).

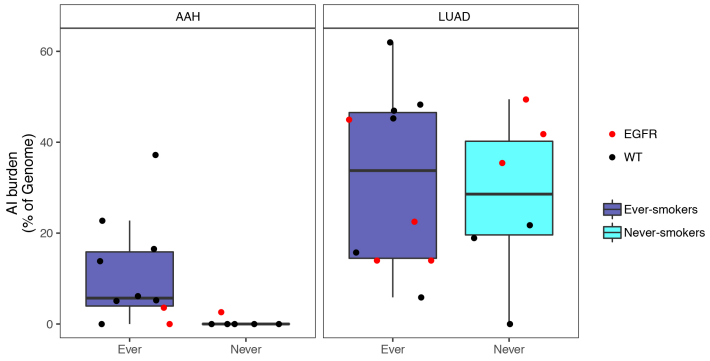


Figure 2.11: **Chromosomal allelic imbalance burden in AAH and LUAD based on tobacco history.** AI burdens, defined as a percent of the genome, are represented by box plots for each tissue type (AAH and LUAD) based on tobacco history (never-smoker and ever-smoker). The burdens for each individual case are overlaid as points on the boxplot. Specifically, samples exhibiting *EGFR* point mutations are shown as red dots.

2.3.3.1 Genomic landscape of chromosome-arm and focal allelic imbalance events

We then assessed large AI events that spanned chromosome arms. Recurrent allelic loss events in 17p harboring tumor suppressors *TP53* (17p13) and *PER1* (17p13), were the most frequently detected chromosomal change in AAHs of our cohort (n = 6; Figure 2.12). Additionally, five of these six cases with 17p loss events in AAHs were identified in patients with a history of tobacco use. Other recurrent AI events in AAHs included the following: gain of 1q, harboring oncogene *ABL2* (1q25) and cell proliferation genes *PARP1* (1q42) and *PBX1* (1q23); gain of 18q harboring *BCL2* (18q21); loss of 8p harboring tumor suppressor *MTUS1* (8p22); loss of 16q encompassing *CYLD* (16q12), *CDH1* (16q22); loss of 19p harboring *KEAP1* (19p13), *STK11* (19p13), *SMARCA4* (19p13); and loss of 19q as well as mixed events on 13q (n=3) (Figure 2.12). The matched LUADs exhibited more complex patterns of allelic imbalance across the entire genome (Figure 2.12). These tissues also showed frequent gains spanning known oncogenes including those on 8q (*MYC*: 8q24), 7p (*EGFR*: 7p11) and 2p (*DNMT3A*, *ALK*: 2p23); they showed loss or cnLOH events harboring

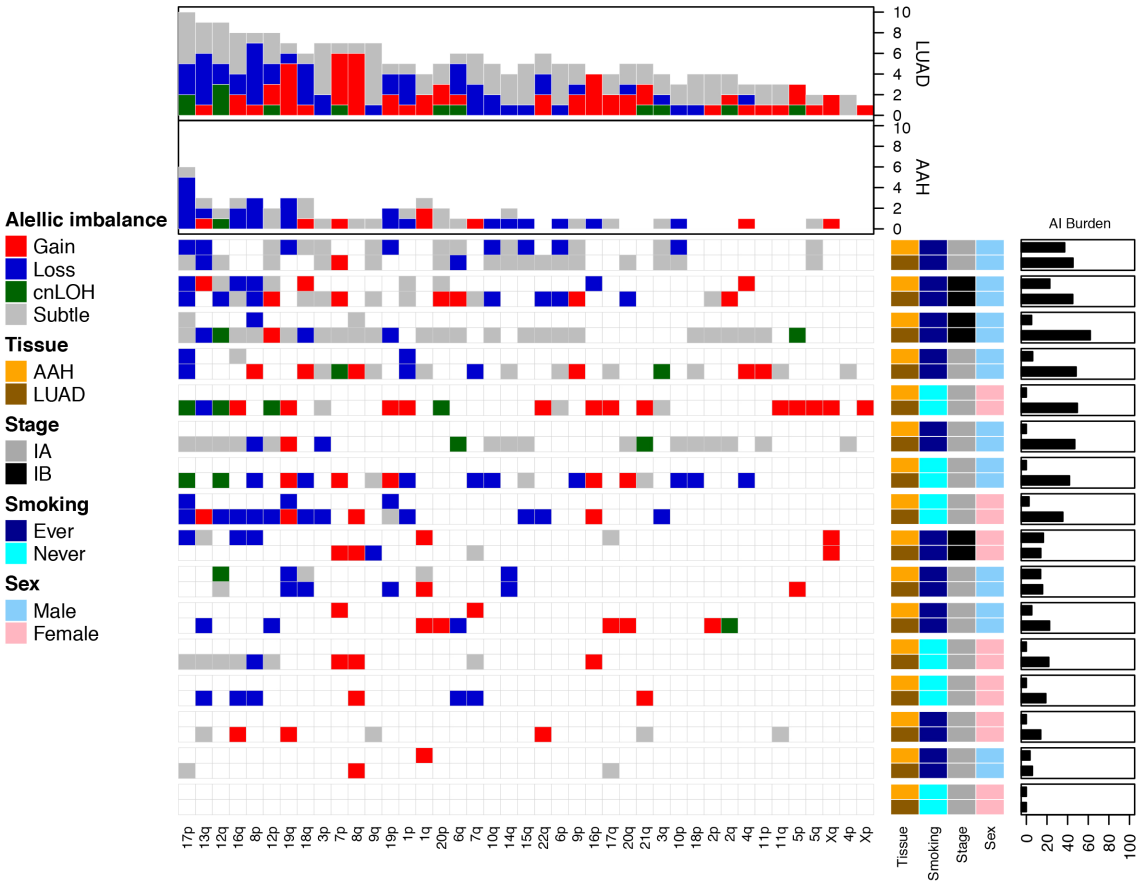


Figure 2.12: **Genome-wide chromosome-arm allelic imbalance events in matched AAH and LUAD.** The distribution of chromosomal arm events in AAHs and LUADs are shown, with rows representing individual patients and columns representing chromosome arms. Each individual row is further divided to show profiles of matched AAH and LUAD from that individual. The events are annotated as gain (red), loss (blue) or copy-neutral loss of heterozygosity (green) while unclassifiable events are annotated as subtle (gray). The overall burden across all chromosomal arms is shown in the bar plots at the top, while allelic imbalance burdens in each sample are shown on the right. Patients are also annotated to denote their clinicopathological features.

Chr	Start	End	Sample1	Sample2	Haplotype Correlation	P value
6	150279	57127760	1-LUAD	1-AAH	-0.725958882	0.00E+00
6	69077820	170823119	1-LUAD	1-AAH	-0.336894721	3.38E-90
20	371781	27500000	1-LUAD	1-AAH	-0.467812513	1.50E-61
6	205213	9189039	1-AAH	1-LUAD	-0.81494625	2.33E-119
6	16751926	57169715	1-AAH	1-LUAD	-0.792269377	2.47E-245
6	67420037	166042705	1-AAH	1-LUAD	-0.340701478	7.82E-86
19	39014184	59083268	1-AAH	1-LUAD	-0.350611209	3.13E-28
20	61098	10834078	1-AAH	1-LUAD	-0.520616776	1.69E-40
11	18742043	45419113	15-LUAD	15-N	-0.345607357	7.23E-26
18	19830977	77938901	15-LUAD	15-N	-0.425760462	1.54E-95
11	18742043	45419113	15-LUAD	15-AAH	-0.344892749	9.24E-26
18	19830977	77938901	15-LUAD	15-AAH	-0.405048942	8.39E-86
17	72487	24000000	16-LUAD	16-AAH	-0.763214541	1.23E-192
19	1095581	7205240	16-LUAD	16-AAH	-0.844794475	9.81E-84
19	50751361	59083268	16-LUAD	16-AAH	-0.854232301	1.12E-154
17	72487	16819909	16-AAH	16-LUAD	-0.803706227	1.41E-205
19	367313	26500000	16-AAH	16-LUAD	-0.313266578	3.80E-22
19	51136844	58819113	16-AAH	16-LUAD	-0.866329216	1.42E-155
8	68295959	101670653	19-LUAD	19-AAH	-0.308031765	1.46E-21
12	1456930	18459387	23-AAH	23-LUAD	-0.438251952	1.03E-39

Table 2.5: Mirrored events between matched tissues exhibiting opposite haplotypes in excess

tumor suppressors such as those on 12q (*ARID2*: 12q12, *MLL2*: 12q13), 3p (*SETD2*: 3p21, *VHL*: 3p25, *FOXP1*: 3p13), 9q (*KLF4*: 9q31, *PTCH1*:9q22, *GNAQ*: 9q21, *TSC1*: 9q34, *ABL1*, *NOTCH1*: 9q34), 18q (*SMAD4*: 18q21), and 6q (*FOXO3*: 6q21). Although the overall AI burden in AAHs was seemingly lower than in LUADs, four cases exhibited similar burdens across these matched tissues (Figure 2.12). These cases also exhibited high sharing of specific AI events in AAHs and matched LUADs, including loss of chromosomal arms 17p (*TP53*, *PER1*: 17p13), 13q (*RB1*: 13q14), 19p (*KEAP1*, *STK11*, *SMARCA4*: 19p13), 19q and 9q (*KLF4*: 9q31, *PTCH1*:9q22, *GNAQ*: 9q21, *TSC1*: 9q34, *ABL1*, *NOTCH1*: 9q34). Of note, we identified AI events that exhibited patterns of mirroring (opposite haplotypes in excess across the same event) between matched AAH and LUAD using RECUR [49]. The mirrored events are summarized in Table 2.5.

In addition to chromosomal-arm AI events, we also identified subtle focal events in AAH of six patients that included 11p gain encompassing *HRAS* and *IGF2* (11p15), 5q gain

spanning *RAD50* (5q31), *FGFR4* and *NSD1* (5q35), 19p loss comprising *STK11* (19p13), 3p amplification at *FOXP1* (3p13), 11p gain encompassing the oncogene *WT1* (11p13), 17q loss harboring NF1 (17q11) and 4q gain covering *KIT* (4q12) (Figure 2.13). Finally, AI detected in the four normal lung parenchyma tissues included three patients with smoking history and exhibited large chromosomal loss events on 19p and 19q, gain of 18q as well as several subtle, yet, large events on 1q, 6q, 7q, 8q, and 20q (Figure 2.13). Three of these cases exhibited events that were shared with matched LUAD specimens and the remaining one case showed events shared with its matched AAH and LUAD tissues (Figure 2.13).

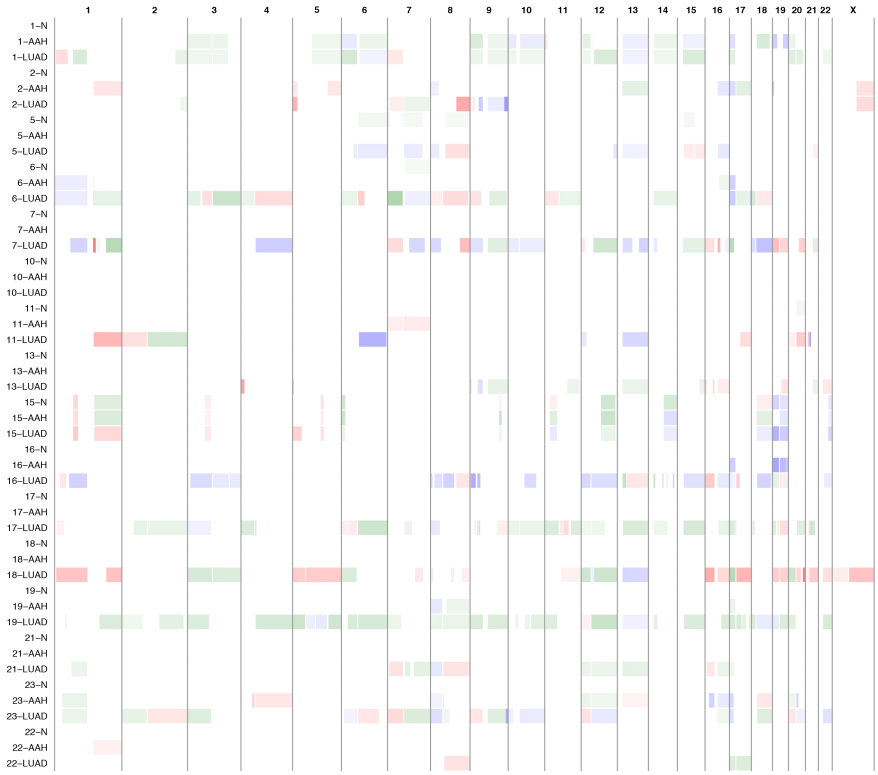


Figure 2.13: **Chromosomal arm and focal allelic imbalance events in matched normal lung parenchyma, AAH and LUAD.** The genomic locations of the identified chromosomal allelic imbalance events were plotted for all 48 samples of matched normal lung parenchyma, AAH and LUAD from 16 patients. Allelic imbalance regions are first classified as gains (red) or losses (blue), the intensity of which is based on the log R ratio of the event. The remaining events are annotated in green as subtle and copy-neutral loss of heterozygosity (cnLOH) events, intensity of which is based on B-allele frequency deviation for the event region, with darker shaded regions representing increased evidence for cnLOH.

2.3.3.2 Genomic evolution processes in AAH and LUAD.

I used the identified chromosomal-arm and focal AI events of matched AAH and LUAD tissues for all patients to construct phylogenetic trees depicting the genomic evolution of these tissues. Mirrored events (2.5) were excluded from this analysis. Seven patients exhibited regions of shared AI events between matched AAH and LUAD forming trunks of phylogenetic trees (Figure 2.14(A)).

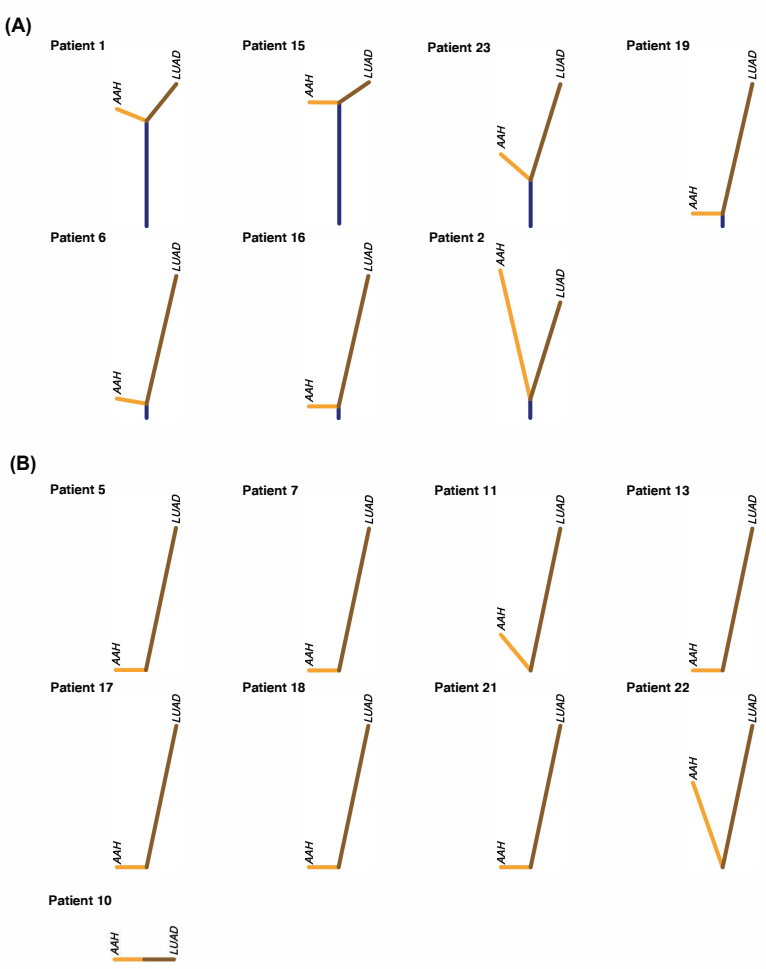


Figure 2.14: **Phylogenetic reconstruction of truncal, AAH-specific and LUAD-specific chromosomal aberrations.** Matched AAH and LUAD specimens from individual patients were assessed for patterns of shared as well as tissue-specific allelic imbalance events and phylogenetic rooted trees were constructed. Cases exhibiting any evidence for shared events are shown in (A) and remaining cases are shown in (B). Vertical distances in each tree are scaled to the proportion of shared as well as tissue-specific events. Shared events, thereby trunks of the trees, are shown in dark blue; tissue-specific events are shown separately for AAH (orange) and LUAD (brown).

The length of the trunk, and therefore the extent of shared events between matched AAH and LUAD varied across patients. Truncal events included chromosomal arms that spanned known lung cancer associated genes such as loss or cnLOH events harboring tumor suppressors on 17p (*TP53*, *PER1*: 17p13) , 8p (*MTUS1*: 8p22), 9p (*CDKN2A*: 9p21), 9q (*KLF4*: 9q31, *PTCH1*: 9q22, *GNAQ*: 9q21, and *ABL1*, *NOTCH1*, *TSC1*: 9q34), 19p (*KEAP1*, *STK11*, *SMARCA4*: 19p13), as well as gains of chromosomal arms encompassing oncogenes on 8q (*MYC*: 8q24), 12p (*KRAS*: 12p12), 1q (*ABL2*: 1q25). Patient 1 showed the largest percentage of shared AI events (36.9%) that included subtle events on chromosomal arms 3p, 5q, 6p, 6q, 9p, 9q, 12p and 17p; LUAD-specific events such as 1p, 7p and 12q and AAH-specific events such as 18q, 19p and 19q. Patients 15 and 23 also exhibited shared AI events between AAH and LUAD (15.6% and 16.0% respectively) that included chromosomal arms 1q, 11p, 18q and 19p in patient 15 and 1p, 12p, 12q, 16q, 17p, 18q, 20p and 20q in patient 23. While patient 15 showed similar overall AI burdens in both AAH and LUAD tissues, patient 23 showed an overall higher AI burden in LUAD compared to its AAH with LUAD-specific AI events including 1p, 2q, 3p, 7p, 9p, 9q and AAH-specific AI events on 4q and 13q. Patient 6 and 19 exhibited shared AI events in a small proportion of the genome (5.2% and 6.1% respectively) followed by patients 2 (2.3%), and 16 (3.2%) showing much lesser sharing between matched AAH and LUADs. Further, among all seven cases with evidence for shared AI between matched AAH and LUAD, a majority were identified as smokers (6 of 7) with only one case identified as a non-smoker (patient 16). In the remaining cases, the AAH and LUAD showed distinct and independent AI profiles (Figure 2.14 (B). These cases exhibited private somatic AI events unique to AAH or LUAD such as those on 1q, 7p, 7q, 13q and 16q. The distribution of shared AI events as well as AAH-specific and LUAD-specific AI events across the genome is shown in Figure 2.15.

2.3.3.3 Somatic multi-hit progression of AAH to LUAD

Finally, I also integrated my previous analysis of single nucleotide mutations (SNVs) within this cohort to identify cancer driver genes exhibiting somatic multi-hit mutational processes (i.e. mutation and a chromosomal-arm or focal AI events encompassing the mutated gene). While the LUADs exhibited somatic-two hit events in known cancer associated genes such

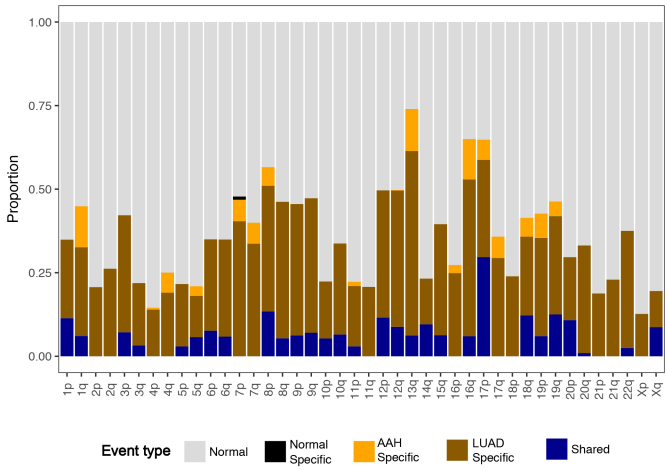


Figure 2.15: **Distribution of shared and tissue-specific allelic imbalance events across chromosomal arms.** The identified allelic imbalance events across all patients were averaged and assessed for chromosomal patterns of shared as well as tissue-specific events. A stacked bar plot representing the proportion of normal region (grey), shared AI (blue), AAH-specific AI (orange), LUAD-specific AI (brown) and normal tissue-specific AI (black) for each chromosome arm is shown.

as *EGFR, TP53, KRAS, CDH1, JAK3, ARID1A, ARID2, CDKN2A, GNAS* and *MSH6*, I identified only two AAH cases with such patterns. One case exhibited a *KRAS* mutation and a 12p gain, that was shared between its AAH and LUAD tissues and another case with an AAH-specific *BRAF*/7p gain event (Figure 2.16). I also identified an additional two cases that exhibited a single shared AI event (i.e. present in AAH and LUAD) with a LUAD-specific second mutation hit (SNV) such as subtle AI on 9q/*NOTCH1* and 17p/*TP53* (Figure 2.16). Table 2.6 summarizes these multi-hit patterns observed in this cohort.

2.4 Discussion

There is a lack of understanding of the molecular aberrations leading to the initiation as well as the progression of AAH, the only known precursor lesion to LUAD. The challenges in physical acquisition of incidental AAHs, in addition to the low mutant cell fractions that are typical for these samples, have limited the molecular characterization of these premalignant lesions to date. In this chapter, I described our findings of the mutation and gene expression landscapes of AAHs in comparison with normal tissues and early-stage LUADs (matched) from the same patients.

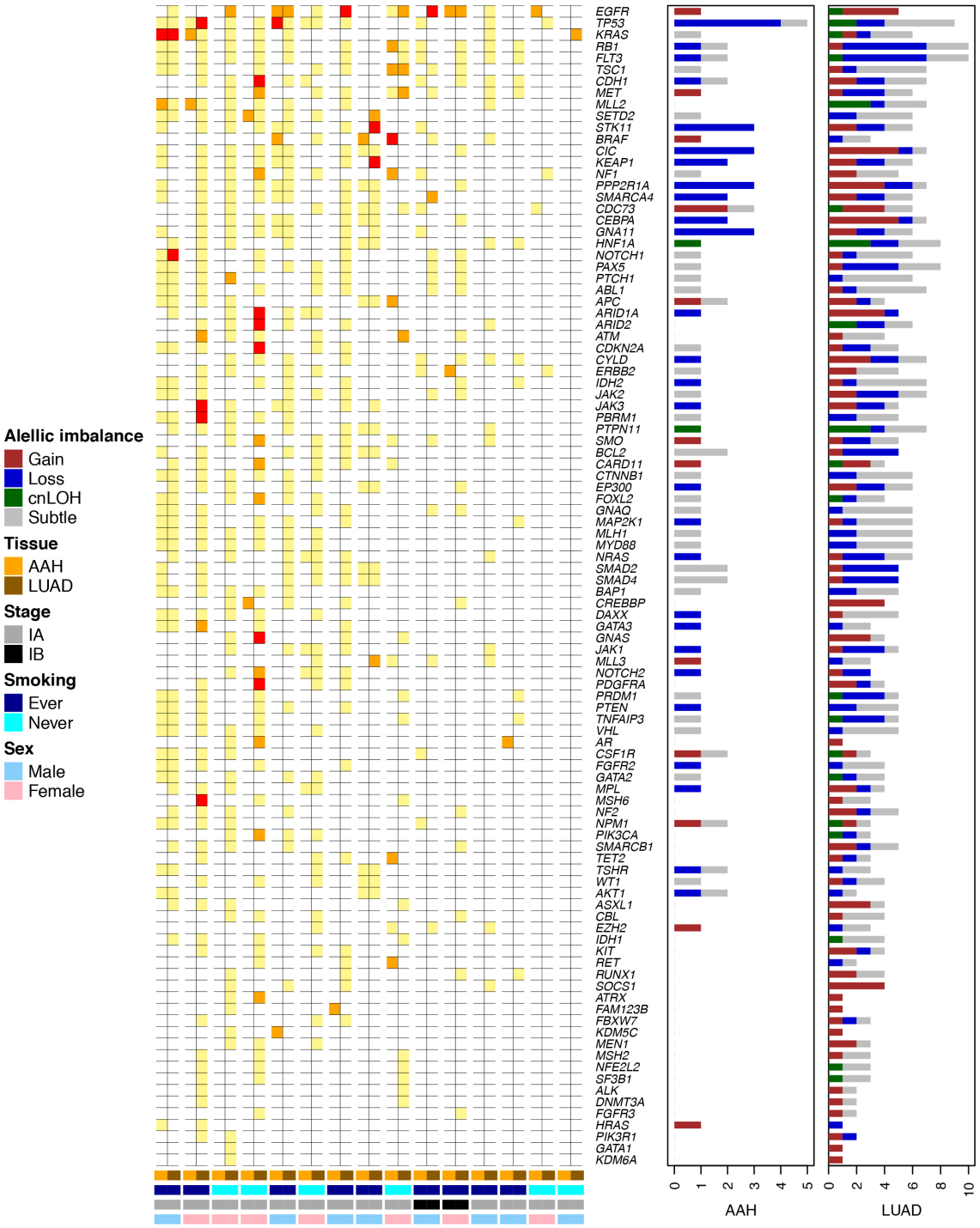


Figure 2.16: **Progressive and somatic two-hit processes in matched AAH and LUADs.** Known cancer driver genes within regions of allelic imbalance or with single nucleotide mutations (SNVs) were examined. The figure depicts genes exhibiting somatic two-hits (both SNVs and AI; red) in AAHs and LUADs as well as those exhibiting a first shared hit (either SNV: orange or AI: yellow) in the AAH accompanied by a second tumor-specific hit in the matched LUAD. For samples with allelic imbalance, the event types are shown as bar plots on the right, accompanying each gene, in both the AAH and LUAD specimens.

Case	AAH	LUAD
1	*KRAS*; 12p subtle	*KRAS*; 12p subtle
1	9q subtle	*NOTCH1*; 9q subtle
2	-	*EGFR*; 7p gain
7	-	*EGFR*; 7p gain
11	*BRAF*; 7q gain	-
15	-	*KEAP1*; 19p loss
15	-	*STK11*; 19p loss
16	*TP53*; 17p loss	17p loss
17	-	*CDH1*; 16q subtle
17	-	*ARID1A*; 1p gain
17	-	*ARID2*; 12q subtle
17	-	*CDKN2A*; 9p loss
17	-	*GNAS*; 20q subtle
17	-	*PDGFRA*; 4q gain
19	17p subtle	*TP53*; 17p subtle
19	-	*JAK3*; 19p loss
19	-	*PBRM1*; 3p subtle
19	-	*MSH6*; 2p subtle

Table 2.6: Multi-hit somatic mutational events in matched AAH and LUAD

2.4.1 Significance of findings

We delineated subgroups of AAHs with mutually exclusive and distinct driver gene point mutations; namely *BRAF*-mutant (both nonsmokers and ever-smokers), *KRAS*-mutant (ever-smokers only) and *KRAS/BRAF* WT AAHs. We also identified the enrichment of chromosomal aberrations such as the loss of 17p in AAHs of this cohort, particularly in ever-smokers. By agnostic transcriptome sequencing analysis, we also presented various patterns of expression profiles and pathways in the molecular pathogenesis of AAH. Further analysis underscored markers of immune function that are significantly differentially modulated, early on, in AAH (downregulation of *GZMB*) relative to NL as well as those deregulated in LUADs relative to AAHs (e.g., *CTLA-4*). Our findings highlight early recurrent driver mutations, chromosomal aberrations, gene expression profiles, and markers of immune response in AAH that offer a better understanding of the molecular pathogenesis of these premalignant lesions.

Our study also underscores previously uncharacterized properties of these AAH

BRAF mutations, namely mutual exclusivity with *KRAS* and correlation with smoking patterns. Whereas *KRAS*-mutant AAHs were from ever-smokers, *BRAF* mutations in AAHs occurred in both non-smokers and ever-smokers. The *BRAF* p.K601E variant has been previously noted in preneoplastic melanocytic lesions and melanomas in situ as well as in thyroid adenomas [59, 60, 61], thus pointing to the probable role of *BRAF* in early stages of oncogenesis (i.e., development of preneoplasia such as AAH). The *BRAF* p.K601E mutation was also found in small proportions of cancers of the thyroid, colon and skin [62]. This may suggest that an enrichment for this hotspot driver mutation highlights a crucial mechanism for AAH and LUAD pathogenesis. Indeed, studies by the TCGA [5] and our group [63] showed relatively infrequent ($\sim 3\%$) *BRAF* mutations in LUADs. Yet, its absence in our sample set of LUADs, including in tumor specimens from patients with *BRAF*-mutant AAH is intriguing. It cannot be neglected that this may, in part, be due to our relatively modest set of samples.

Statistical approaches to discover large chromosomal alterations in AAH samples may be limited to mutant cell fractions of 15% with standard SNP array technology. Here we applied hapLOH, a sensitive, haplotype-based method [48] that offers resolution perhaps down to 5% mutant fraction. From our results, the frequency of detectable allelic imbalance events in these samples may at first appear high. However, aspects of our analysis, study design, and findings in other nonmalignant tissues serve to contextualize these findings. First, the increased resolution from using our statistical approach probably captures a critical region of the within-sample mutation frequency spectrum, specifically, mutations in a small proportion of cells, consistent with their involvement in early stage of development and the heterogeneous nature of the tissues. Second, the use of SNP arrays, instead of exome sequencing, allows for more power to detect copy number changes, particularly those leading to AI. Third, we applied additional statistical testing; when we detected an event in a tissue, we specifically looked at that same region in matched tissues. Finally, rates of half for premalignant lesions or field cancerization samples demonstrating detectable AI have been observed in the lung [13] and colon [64].

Chromosomal aberrations that we identified not only corroborate previously described LOH events but also provide better resolution of genome-wide gain, loss and cn-

LOH, including previously undocumented aberrations such as those on chromosomes 1,7, 8, 12 and 19. Chromosomal aberrations such as loss and cnLOH of arms 9p, 12q, 17p, 19p and 19q and gain of 1q, 8q, 18q, 7p and 7q in AAHs of our cohort have been shown in previous studies of chromosomal changes in early-stage non-small cell lung cancer (NSCLC), including *EGFR*-mutant LUADs, of Asian patients, that form a major subset of our cohort [65, 66, 67, 68, 69, 70]. Evidence for shared chromosomal aberrations between matched AAHs and LUADs was primarily observed among smokers in our cohort, alluding to the role of field cancerization in the development of these preneoplastic lesions. Further that these changes are not only shared with NSCLCs but exhibit reduced overall proportions in AAHs compared to matched LUADs are consistent with the morphological changes in these lesions and might suggest their role in the malignant transformation of these premalignant lesions. Chromosomal aberrations identified in our study have also been previously described in premalignant lesions of other tumor sites [71, 72]. For example, the most common event in AAHs of our cohort, 17p loss, has been previously described as an early event, preceding mutations in TP53, and a predictor of neoplastic progression in Barretts esophagus, a premalignant lesion which predisposes to esophageal adenocarcinoma [73, 74, 71, 72]. Another study described the importance of loss events such as on 17p, 8p and 13q in addition to early LOH events of 3p and 9p in conferring increased relative risk of malignant transformation in oral premalignant lesions [75]. Further, the higher incidence of 17p loss events observed here compared to previous studies [23], particularly in smokers, might be attributable to the East Asian origin of this cohort. These findings implicate a role of chromosomal imbalances early in the development and progression of these preneoplastic lesions.

Complementing our DNA analysis is our agnostic transcriptome sequencing analysis that revealed differential gene expression programs that occur in different stages of LUAD pathogenesis – early in development of AAH from normal tissue, in LUADs, or in both lesion types. Several altered gene expression programs and pathways that we identified in AAHs of our cohort were also independently identified in other premalignant and invasive tumor tissues. WNT/β-catenin signaling has been previously shown to be activated in progression of oral leukoplakia, a precancerous lesion of head and neck squamous cell carcinoma

(HNSCC) [76]. Increased *EGFR* was shown to promote cellular proliferation, inhibit apoptosis and drive development and progression of bronchial dysplasia [77]. Similarly, *MYC* overexpression has been previously reported in colorectal polyps with a level of expression proportional to the polyp size as well as dysplastic histology [78]. Taken together, these data point to early changes in the development and progression of AAH and that would thus comprise ideal targets for chemoprevention of LUAD.

Our immune marker profiling overall suggested an activation of pro-tumor immune pathways (i.e., Th2) and B-cell receptor signaling as well as an inhibition of anti-tumor immune response (e.g., Th1-derived *IFN-*γ signaling). Similar findings have been reported in previous studies of Barretts esophageal tissues, a premalignant condition with a high risk of progression to esophageal adenocarcinomas [79]. *IL12A*, known for its proinflammatory anti-tumor response, along with anti-tumor immune chemokines (e.g. *CCL3, CCL4, TLR4*) and apoptosis-inducing proteases (*GZMB*) were decreased in AAH relative to normal lung. On the other hand, we found elevated expression of the *CCL2* chemokine receptor *CCR2* in AAH relative to normal lung. *CCR2* has been shown to enhance tumor growth, angiogenesis and tumor progression and was demonstrated to be over-expressed in several tumor tissues [57]. Recently, *CCL2/CCR2*-based immune prevention models were shown to attenuate tumor development and metastasis [80, 57]. Of note, we identified an increasing expression in chemokines *CXCL13* and *CXCL14*, both known for their role in inflammatory processes and immune response [55], and *SPP1*, previously shown to be overexpressed in premalignant lesions of the oral epithelium as well as actinic keratosis, the premalignant lesion to skin squamous cell carcinomas [81]. We also found that CD27, which in combination with its ligand CD70 is known to generate a potent co-stimulatory signal, was increased in AAH relative to normal lung. Notably, our analysis pointed to significantly increased expression of the major immune checkpoint *CTLA-4* in LUAD relative to AAH [58]. Our brief analysis of the abundance of tumor-infiltrating immune cells corroborate previous findings [82], further suggesting the presence and importance of immune activation, early in premalignancy, potentially prior to other mutational processes. Progression of these preinvasive lesions may be characterized by additional genomic and transcriptomic aberrations, that may eventually lead to the acquisition of immune tolerance or evasion mechanisms in more

malignant phenotypes.

Based on our findings on mutual exclusivity of *BRAF* and *KRAS* in AAHs along with the disparate patterns of mutations in the paired LUADs, it is plausible to suggest that there are divergent pathways in pathogenesis of these preneoplastic lesions. Disparate and shared gene expression and immune marker deregulation between *BRAF*- and *KRAS*-mutant AAHs further point to differential aberrant immune signaling among AAH based on driver mutation status. A schematic of this paradigm is represented in Figure 2.17. A similar divergent model to malignancy has also been recently described in the evolution of different melanoma subtypes from their precursor lesions [61]. We posit here that aberrant immune signaling (e.g., attenuated anti-tumor immune response) is a common, perhaps critical, feature of AAH and LUAD development, as illustrated in Figure 2.17. Given that our analysis of chromosomal allelic imbalance events came from a smaller subset of cases, I did not include it in this proposed paradigm. However, a natural extension to this paradigm would be the preferential enrichment of 17p loss events in ever-smokers and the important role of cancerized fields in the development of AAH and LUAD, particularly in smokers, as described by the prevalence of shared DNA mutations and chromosomal allelic imbalance identified in our cohort of AAH and LUAD.

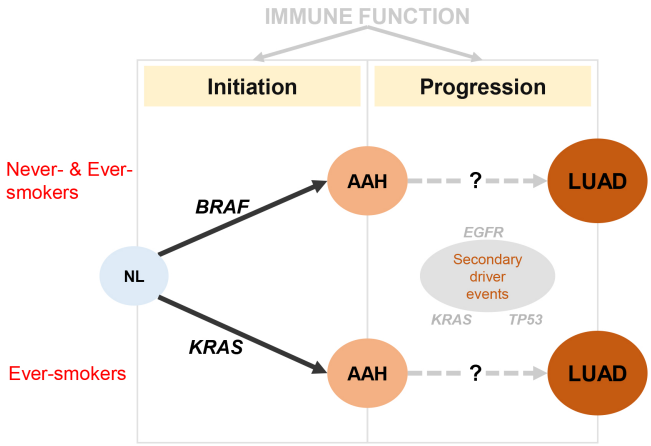


Figure 2.17: **Proposed models for the pathogenesis of AAH.** Two potentially divergent modes in the pathogenesis of these preneoplastic lesions are proposed based on the mutual exclusivity of point mutations and disparate expression profiles.

2.4.2 Limitations

While we comprehensively studied paired AAHs and LUADs, our cross sectional study design is not best positioned to thoroughly characterize the progression of AAH to LUAD. Naturally, the AAHs that already progressed to LUADs are no longer available for analysis. In addition, our cohort consisted of only one AAH specimen for each individual, and therefore, limits investigations of heterogeneity that might be crucial to identify lesions that progress to malignancies.

Further, based on our findings on the absence of *BRAF* mutations in the LUADs studied in our cohort, one cannot rule out the possibility that the *BRAF*-mutant AAHs are benign and may not be the preneoplastic lesions that eventually progress to LUADs. Earlier studies have insinuated that *BRAF* mutations are important for initiation of premalignancy rather than their expansion or progression [83, 40]. Lungs of mice genetically engineered to express a mutant form of Braf (p.V600E) were shown to develop hyperplasias that progressed to adenoma [84]. Of note, only after mutations in other genes (e.g. *Tp53*) did the *Braf* mutant lesions progress to LUAD [84]. Yet, the strong pairing of *BRAF*-mutant AAHs with *EGFR*-mutant LUADs is nonetheless an interesting observation that is worth investigating in future studies comprising a larger number of patients with both AAHs and LUADs. Further, that these patterns hold across lesions arising independently, although potentially from the same cell lineage, reflect the patient-specific nature of their development and highlight the potential for personalized prevention strategies.

It is also worthwhile to mention that our cohort was mainly comprised of East Asian patients. Earlier studies have demonstrated that LUADs of East Asians exhibit disparate mutational spectra (e.g. more prevalent activating mutations in *EGFR*) relative to LUADs from Western (or Caucasian) patients [85, 57]. It is reasonable to surmise that mutational differences in AAHs, across patients of different ethnicities, to roughly reflect those we observe in LUADs. In this context, our results and proposed paradigm could be more relevant to the East Asian population based on our cohort.

Lastly, recent pathological classification guidelines for LUAD have underscored subgroups with pure lepidic growth (adenocarcinoma *in situ*, AIS) and those that exhibit

predominant lepidic growth and with less than 5 mm invasion (minimally invasive adenocarcinoma, MIA). Earlier work in East Asian LUAD patients suggested that LUADs of the terminal respiratory unit (TRU) progress from AAH to AIS and then to invasive lesion. It is plausible that AIS may have distinct profiles that suggest an intermediate stage [61, 40] in the progression of AAH to LUAD. However, our cohort largely comprised LUADs with very few AIS or MIA, too limited in size to further delineate profiles along this progression.

CHAPTER 3
INVESTIGATION OF FIELD CANCERIZATION IN EARLY STAGE NON-SMALL CELL LUNG CANCERS

Field cancerization in non-small cell lung cancer (NSCLC) was originally coined based on observations of histological abnormalities in normal tissues adjacent to the tumor. Work thus far has identified gene expression modifications, chromosomal aberrations and, to a limited extent, single-gene mutations in tumor-adjacent and distant airway epithelium. We sought to comprehensively characterize the largely unexplored, somatic mutation landscape of normal-appearing airway epithelia in early-stage NSCLC patients. Somatic DNA alterations in these normal tissues often exist at low mutant cell fraction and thereby demand better computational approaches. In this chapter, I describe our approach to identify somatic driver mutational processes in airways of early-stage NSCLC patients that might suggest the presence and importance of field cancerization in NSCLC pathogenesis. I discuss the bioinformatics tools and methods that I implemented to integrate point mutations and large chromosomal aberrations as well as to compare mutations in the airway epithelia to those of the matched NSCLC tumor. I conclude by proposing a list of driver mutations that might inform of the temporal and spatial events in the initiation and development of NSCLCs from the airway epithelium. The contents of this chapter is based on the following publications:

Kadara H*, **Sivakumar S***, Jakubek Y, Lucas FAS, Lang W, McDowell TL, Weber Z, Behrens C, Davies GE, Kalhor N, Moran C, El-Zein R, Mehran R, Swisher SG, Wang J, Zhang J, Fujimoto J, Fowler J, Heymach JV, Dubinett S, Spira AE, Ehli EA, Wistuba II, Scheet P. Driver mutations in normal airway epithelium elucidate spatiotemporal resolution of lung cancer. *In press*, American Journal of Respiratory and Critical Care Medicine 2019.

3.1 Study design

We surveyed mutational processes in the cancerized field of early-stage NSCLC patients using a combination of deep targeted DNA sequencing of a panel of 409 cancer-associated genes and high-resolution single nucleotide polymorphism (SNP) array profiling (Figure 3.1). Multi-region normal airways, comprising tumor-adjacent small airways, tumor-distant large airways, nasal epithelium and uninvolved normal lung (collectively field), matched NSCLCs as well as blood cells were interrogated for somatic driver point mutations and genome-wide chromosomal allelic imbalance events (Figure 3.1). Point mutations and chromosomal allelic imbalance events were integrated to study patient-specific patterns as well as interrogated for specific driver gene-associated multi-hit patterns of pathogenesis and progression to NSCLCs. I also present computational methods to identify and assess somatic DNA alterations in these normal tissues that exist at low mutant cell fractions, such as a novel measure to quantify the extent of field carcinogenesis. I then relate the molecular changes detected in these normal-appearing tissues to early alterations in the transition to the malignant phenotype of NSCLC; these may in turn serve as potential targets for early detection and treatment.

3.2 Methods

3.2.1 Cinical cohort

Multi-region samples comprising tumor-adjacent small airways, tumor-distant ipsilateral large airways, nasal epithelium, normal lung tissue, peripheral blood cells as well as NSCLC tumors were obtained from 48 early-stage patients (stages IA-IIIA; 37 LUADs and 11 LUSCs; 42 ever-smokers and six non-smokers) who were evaluated at The University of Texas MD Anderson Cancer Center. All 48 patients did not receive neoadjuvant therapy or any therapy for at least a year prior to surgery. Demographic and clinicopathological data for all cases are summarized in Table 3.1. The study was approved by the Institutional Review Board and all patients provided written informed consents. A breakdown of samples obtained from each patient is summarized in Table 3.1.

Case	Histology	Gender	Vital status	Recurrence	Stage	Tobacco History
AIR_001	LUAD	M	A	Y	IIA	Never
AIR_002	LUAD	F	A	N	IIIA	Ever
AIR_003	LUAD	M	D	N	IIIA	Ever
AIR_004	LUAD	M	A	Y	IIB	Never
AIR_005	LUAD	F	D	N	IIB	Ever
AIR_007	LUAD	M	A	Y	IIIA	Ever
AIR_008	LUAD	F	D	Y	IIA	Ever
AIR_009	LUAD	M	A	Y	IB	Ever
AIR_010	LUAD	M	A	N	IIA	Ever
AIR_011	LUAD	M	A	N	IB	Ever
AIR_012	LUSC	M	A	Y	IIA	Ever
AIR_013	LUAD	F	A	N	IA	Ever
AIR_014	LUAD	F	A	N	IA	Ever
AIR_015	LUSC	F	A	Y	IIB	Ever
AIR_016	LUAD	M	D	Y	IB	Ever
AIR_017	LUSC	M	D	Y	IIB	Ever
AIR_018	LUSC	F	A	N	IB	Ever
AIR_019	LUAD	M	A	N	IA	Ever
AIR_020	LUAD	M	A	N	IB	Ever
AIR_022	LUSC	F	A	N	IA	Ever
AIR_023	LUSC	M	A	N	IB	Ever
AIR_024	LUSC	F	A	N	IIA	Ever
AIR_026	LUAD	F	A	N	IIB	Ever
AIR_027	LUAD	F	D	Y	IIA	Never
AIR_028	LUAD	F	A	Y	IA	Never
AIR_031	LUSC	M	D	Y	IB	Ever
AIR_032	LUSC	M	A	N	IIB	Ever
AIR_033	LUAD	F	A	Y	IA	Ever
AIR_034	LUAD	F	D	N	IB	Ever
AIR_035	LUAD	M	A	N	IIB	Ever
AIR_036	LUAD	M	A	N	IA	Ever
AIR_037	LUAD	F	A	N	IIIA	Ever
AIR_039	LUSC	F	A	N	IIA	Ever
AIR_040	LUAD	F	A	N	IB	Never
AIR_041	LUAD	F	A	N	IA	Ever
AIR_042	LUAD	M	A	N	IIA	Ever
AIR_043	LUAD	F	A	N	IA	Ever
AIR_044	LUSC	F	A	N	IB	Ever
AIR_045	LUAD	F	D	N	IIB	Never
AIR_047	LUAD	M	D	Y	IIA	Ever
AIR_048	LUAD	F	A	N	IB	Ever
AIR_049	LUAD	F	A	Y	IIA	Ever
AIR_050	LUAD	F	A	Y	IIIA	Ever
AIR_052	LUAD	F	A	N	IB	Ever
AIR_053	LUAD	M	A	Y	IIA	Ever
AIR_054	LUAD	M	D	N	IIIA	Ever
AIR_055	LUAD	M	A	N	IB	Ever
AIR_056	LUAD	F	A	N	IIIA	Ever

Table 3.1: Clinicopathological features of patients studied for airway field cancerization. (Histology - LUAD: Lung adenocarcinoma, LUSC: Lung squamous cell carcinoma; Sex - M: Male F: Female; Vital status - A: Alive, D: Dead; Recurrence - Y: Yes; N: No)

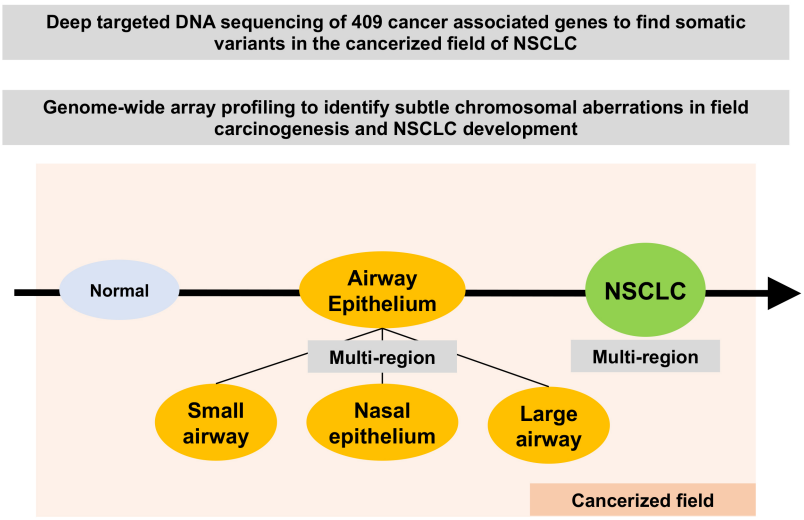


Figure 3.1: **Study design to understand the field cancerization mechanism in NSCLCs.** A two-pronged approach, consisting of deep targeted DNA sequencing and a broad-scale SNP genotype array based profiling, was used to study point mutations and chromosomal aberrations in the normal-appearing cancerized field of early-stage NSCLCs.

3.2.2 Multi-region sample collection

The different types of samples used in this study were collected in the manner we performed previously [14]. The small airway epithelia adjacent to NSCLCs from the resected specimens. Small normal-appearing airways adjacent to the tumors (S1-S5; S1, relatively closest from NSCLC and S5, relatively farthest from the tumor) were collected using Cytosoft brushes from the resected specimens in all 48 patients in a manner previously described [14]. Briefly, the spatial distance between two consecutive airway brushings was similar (approximately 2 cm). Airways were denoted by numbers 1 (relatively closest from tumor) to 5 (relatively farthest). The relative distance of an airway brushing (e.g., airway 1) from the adjacent NSCLC tumor was similar across all case patients. Brushings of the nasal epithelium and large airway were collected during endoscopic bronchoscopy prior to resective surgery. Nasal brushings were acquired using sterile Cytosoft cytology brushes (Medical Packaging Corporation) and large airway epithelia were obtained endoscopically using ConMed disposable bronchial cytology brushes (ConMed Corporation). Brushings from the nasal epithelium (Na) and from large airways (L; mainstem bronchi) were collected from

a subset of patients (21 and 24 patients respectively). Lung parenchyma and NSCLC tissues were immediately snap-frozen after resection. Additionally, 169 multi-region NSCLC biopsies were obtained from 28 patients, ranging from three to eight biopsies per patient (CNBs 1-8). Histopathological evaluation of brushings was performed in a manner reported previously [14]. Epithelial content was confirmed by immunohistochemical analysis of pan-cytokeratin and absence of preneoplastic or neoplastic cells was assessed by hematoylin eosin staining. Briefly, two touch preparation cytology slides were prepared from brushings by pressing the samples against glass slides. Cytology slides were then fixed, stained with Hematoxylin Eosin and analyzed histopathologically using experienced pathologists. Preneoplastic cells, if identified, were excluded because of the minute amount of cells and for being outside the scope of the present study. The normal appearing tissues including the tumor-adjacent small and distant large airways, nasal epithelium and uninvolved normal lung are collectively referred as field for the purpose of further analysis. A breakdown of samples obtained from each patient is summarized in Table 3.2.

3.2.3 DNA targeted sequencing

DNA was extracted from all fresh-frozen samples using the QIAamp DNA kit from Qiagen according to manufacturers instructions. DNA was quantified using the RNase P assay (Life Technologies) according to the manufacturers protocol. Sequencing of a panel of 409 canonical cancer-associated genes was performed in the manner I previously reported in Chapter 2. The Ion AmpliSeq Comprehensive Cancer Panel (CCP; Thermo Fisher Scientific) comprising primers for 409 canonical cancer-associated genes and the AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific) was used to prepare barcoded libraries from the DNA samples (40 ng). Target amplification steps were carried out in 5 μl reactions with 13 cycles of amplification. The pools were then combined for digestion and ligation steps. Libraries were quantitated with qPCR using the Ion Library TaqMan Quantitation Kit. Sequencing was performed using the Ion Torrent Proton platform. The Ion Torrent Suite was used to assess overall quality of the libraries, chips, and reagents. Raw DNA sequencing data files have been deposited in the sequence read archive under Bioproject accession PRJNA453609.

Case	Sample Availability							
	T	**CNB**	**S**	**L**	**N**	**Na**	**BL**	**Total**
AIR_001	1	0	2	1	1	1	1	7
AIR_002	1	0	2	1	1	1	1	7
AIR_003	1	0	2	1	1	1	1	7
AIR_004	1	0	2	1	1	1	1	7
AIR_005	1	0	2	1	1	1	0	6
AIR_007	1	6	2	1	1	1	1	13
AIR_008	1	8	2	1	1	1	1	15
AIR_009	1	0	2	1	1	1	1	7
AIR_010	1	0	1	1	1	1	1	6
AIR_011	1	0	1	1	1	1	1	6
AIR_012	1	0	2	1	1	1	1	7
AIR_013	1	0	2	1	1	1	1	7
AIR_014	1	0	2	1	1	1	1	7
AIR_015	1	6	2	1	1	1	1	13
AIR_016	1	6	2	1	1	1	1	13
AIR_017	1	6	1	1	1	1	1	12
AIR_018	1	0	2	1	1	1	1	7
AIR_019	1	0	1	1	1	1	1	6
AIR_020	1	0	2	1	1	1	1	7
AIR_022	1	0	2	1	1	1	1	7
AIR_023	1	0	1	1	1	1	1	6
AIR_024	1	8	2	0	1	1	1	14
AIR_026	1	8	2	0	1	1	0	13
AIR_027	1	8	2	0	1	1	1	14
AIR_028	1	7	2	0	1	0	1	12
AIR_031	1	6	2	0	1	0	1	11
AIR_032	1	6	2	0	1	0	1	11
AIR_033	1	4	2	0	1	0	0	8
AIR_034	1	8	5	0	1	0	0	15
AIR_035	1	0	5	0	1	0	1	8
AIR_036	1	8	5	0	1	0	1	16
AIR_037	1	8	5	0	1	0	1	16
AIR_039	1	8	5	0	1	0	0	15
AIR_040	1	8	5	0	1	0	1	16
AIR_041	1	0	4	0	1	0	1	7
AIR_042	1	8	5	0	1	0	1	16
AIR_043	1	0	5	0	1	0	1	8
AIR_044	1	0	4	0	1	0	0	6
AIR_045	1	8	5	0	1	0	1	16
AIR_047	1	4	4	0	1	0	1	11
AIR_048	1	4	5	0	1	0	1	12
AIR_049	1	3	5	0	1	0	1	11
AIR_050	1	4	5	0	1	0	1	12
AIR_052	1	3	5	0	1	0	1	11
AIR_053	0	4	5	0	0	0	1	10
AIR_054	1	4	5	0	1	0	1	12
AIR_055	1	4	5	0	1	0	1	12
AIR_056	1	4	5	0	1	0	1	12
								498

Table 3.2: Samples analyzed for DNA-based profiling. Tissues are labeled as T: Tumor, CNB: Core-needle biopsy, S: Tumor-adjacent small airways, L: Tumor-distant large airway, Na: Nasal epithelium, N: Normal uninvolved lung, and BL: Blood

3.2.4 Strategy to identify somatic point mutations

In a manner described in Section 2.2.6, I used a multiple mutation callers to identify somatic SNVs in all samples profiled in this cohort. When available, white blood cells were used as germline control (n = 42); for the remaining cases, their matched normal lung tissues were used as control (n = 6).

3.2.4.1 Within-patient sample quality control

Given the large size of our study cohort, prior to running all the mutation callers, I wanted to inspect the association of samples from the same individual based on sample labels by performing a quantitative assessment of within-patient and between-patient samples. One approach to do this verification is to test for the concordance of genotypes at germline variant sites. Our expectation would be to observe higher concordance between samples from the same individual than those from different individuals. I used the *vcf-compare* feature in *vcftools* to perform this genotype-level correlation between the marginal VCF files generated by Torrent variant caller. To make it more accurate, variants can be restricted to those observed in the 1000 genomes project. The reported non-reference discordance rates (NDR) were then used as measure to infer sample relationships. A low NDR is expected between samples from the same individual, while samples from different individuals show high NDR values.

3.2.4.2 Multiple mutation callers

Somatic mutations were rigorously identified based on four different methods: the Ion Torrent proprietary software Ion Reporter, marginal variant files (VCFs) generated from Torrent Variant Caller (TVC), MuTect and Varscan2. The stringency of mutation calling was increased to only include mutations that were identified by at least two different callers in at least one sample for every patient. This approach was implemented to account for lower variant allele frequency (VAF) mutations, particularly in the airway field, that might not be identified by all mutation callers. Somatic SNVs in exonic, splicing and untranslated regions (UTRs) within the targeted 409 cancer gene panel were assessed. Nonsynonymous

mutations in bona fide NSCLC and cancer driver genes [5, 6, 42] were also studied *a priori* in both the NSCLCs tumors and biopsies (T, CNB1-8, respectively) and the field (S1-5, L, Na, N) samples.

3.2.4.3 Inference of mutation signatures

Although, traditionally, larger sequencing platforms such as whole-genome sequencing and whole-exome sequencing derived mutations are used to study mutation signatures, I inspected the mutations identified from known mutation signatures, under the assumption of high mutation burden in NSCLCs. Therefore, SNVs detected in the NSCLC and field samples were assessed for patterns in base substitutions and potential mutational signatures using the R package *deconstructSigs* [86]. First, the NSCLC samples were dichotomized based on their smoking status and then analyzed for specific mutation substitution patterns and signatures (e.g., smoking-associated C>A signature). Using signatures detected in NSCLCs of smokers, the matched normal field samples were also assessed for enrichment of these specific signatures. Normal airway samples from non-smokers were not tested for mutational signatures due to the limited number of non-smokers samples and their relatively low mutation burden.

3.2.5 Identification of subtle genome-wide allelic imbalance

AI profiles for 39 of the 48 NSCLC cases in this cohort were previously assessed and reported by our group [13]. AI profiles in the additional nine patients were analyzed using hapLOH [48] and chromosomal aberrations were characterized in a manner previously reported [13] and as described in Section 2.2.9.

3.2.6 Quantitative analysis of the cancerized field in the normal-appearing airway.

The SNVs as well as the large chromosomal aberrations were used to perform quantitative tests to better study the extent of shared events between the tumor and other tissues within each patient as well as across all patients. In this section, I describe the approaches

implemented to infer the spatial and temporal ordering of events in the normal-appearing airway epithelium and in its transition to NSCLCs.

3.2.6.1 Statistical testing of spatial airway field of cancerization

In each patient, the presence of a spatial gradient of variant allele frequencies (VAFs) was tested. For every individual, at each mutated locus in the NSCLC, *samtools* was used to obtain a VAF for the matched airway field tissues from allelic depths at those loci. This was performed solely to increase the data points of mutations for comparison and to account for variants that might not be identified by any mutational caller, but show subtle signal of presence in their raw allelic depths. This set of mutation VAFs was termed forced VAFs. Then, for each patient, linear regression was performed using the forced VAFs for the airway field samples against numeric constants assigned to each airway field sample based on the proximity to NSCLC in the order (S(1-5), L, N, Na as 1.1,1.2,1.3,1.4,1.5,8,11 and 14 respectively). The distances were assigned based on a close approximation of physical distance. Other arbitrarily assigned distances were also tested. The total read depths at the mutated loci were used as weights in the linear regression model. The regression slope was then calculated, with a negative slope indicating a spatial field effect i.e. tissues closer to the tumor (e.g., S1) exhibit higher VAFs than those farther away from the tumor (e.g., Na). Tumors were excluded from this slope estimation, since it is expected that tumors will always force a negative slope by virtue of being more aberrant. The mean slope across all 48 patients was calculated. This mean was then tested for statistical significance using permutation analysis, consisting of 1000 iterations, by altering the field tissue order in each iteration. Due to the lack of a natural null distribution for our test statistic, a conservative randomization approach robust to model misspecification was used. Each iteration involved altering the tissue order/distance (e.g., iteration 1: S3, S2, N, Na, L, S4, S1, S5, iteration 2: L, S3, S1, N, Na, S4, S3, S5) and re-estimating the mean slope across all 48 patients. Z-scores and P values were then computed from these iterated values against the previously obtained mean slope from field effect derived tissue orders.

3.2.6.2 Quantification of airway field of cancerization

We then wanted to better quantify the extent of field of cancerization effect in our cohort. To do this, I developed a measure based on the proportion of shared events between matched NSCLCs and field specimens. For each individual, SNVs and AI events identified in the NSCLCs were examined in the matched field samples. A similarity measure was then computed between tumors and each matched airway field sample based on the extent of shared somatic events, derived from the presence (or absence) of events. Airway field samples were sorted based on their similarity measure from the corresponding NSCLC and were assigned equidistant numeric values, with all measures scaled from 0 to 1 (0, corresponding to NSCLC; and 1, corresponding to the farthest field tissue). For each patient, the ordered field tissues and their proportion of sharing with matched NSCLCs were used to compute a field cancerization area under the curve (FCAUC) using the *DescTools* package in R, ranging from 0 if there is no evidence of field cancerization to 1 if field tissues shared all their mutations with the matched NSCLC.

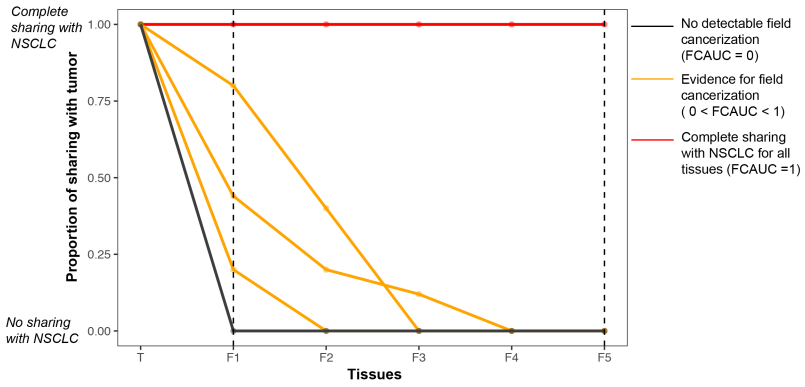


Figure 3.2: **Pictorial representation of the calculation of field cancerization area under the curve (FCAUC).** Genomic airway field cancerization was quantified based on shared SNV and AI profiles and summarized as FCAUC (between 0: lack of airway cancerization evidence and no sharing of alterations in the airway field with the tumor (black line) and 1: complete sharing of alterations between all airway field samples and matched NSCLCs (red)). Shown here are three representative cases with relatively varied FCAUCs (orange). The x-axis denotes an ordinal distance of airway field tissues from its matched NSCLC (0 to 1), and y-axis denotes the proportion of shared aberrations with the matched NSCLC (0 indicates no shared events, to 1 for complete sharing). The FCAUC is computed between the dashed lines by excluding the primary tumor specimen from the calculation

Figure 3.2 shows a pictorial representation of the computation of the FCAUC measure. Suppose that an individual had one tumor sample (T) and five field samples (F1-F5). The x-axis shows the ordinal scale of all the tissues available for an individual, ordered based on the extent of sharing with the primary NSCLC. The y-axis shows the proportion of shared events with the NSCLC. The profile for the NSCLCs of each individual are derived from both the tumor sections as well as core needle biopsy specimens, where available. FCAUC is computed between the dashed lines shown in Figure 3.2, i.e., using only the field tissues (F1-F5). When there is no detectable field of cancerization, the field tissues exhibit no shared events with the primary NSCLC and therefore will appear like the black line plot in Figure 3.2, which will result in a FCAUC of zero. In contrast, although rare, if the field tissues showed all the mutations observed in the matched NSCLCs, the FCAUC value will be one. In other individuals, we might observe different proportions of shared events in the field tissues, probably based on the proximity to the tumor, that might result in orange line plots shown in Figure 3.2. These cases will show a FCAUC between zero and one. Our primary interest was identifying these patients with non-zero FCAUC, as potential exhibitors of field effects. Of these, cases with a visually pronounced FCAUC were further interrogated for specific mutation patterns.

3.2.6.3 Phylogenetic analysis of the cancerization field

Phylogenetic trees were then constructed from the identified somatic SNVs and AI events across all samples for each individual to study the specific shared and disparate mutation events between matched NSCLCs and field tissues. Distance matrices were generated using *dist.gene* from the *ape* package in R. The difference in this section compared to the previous section of FCAUC computation is that mutations private to the field tissues were also used in this analysis to generate the distance measures. For each individual, the distance matrix derived from all tissues profiled for that individual was then used to construct unrooted neighbor joining trees. While the two sources of somatic alterations were merged for the combined analysis, I also constructed individual level trees for SNVs and AI events independently and then tested the concordance of the two tree topologies (SNV tree and AI tree) for each patient. The patients exhibiting an evidence for field cancerization were

then identified based on their distinct topologies from those lacking evidence for field effects (straight line).

3.3 Comprehensive genomic characterization of the cancerized field of early-stage NSCLCs

In this section, I present the results from our comprehensive characterization of point mutations and large chromosomal aberrations in the normal-appearing cancerized field of early-stage NSCLC patients.

3.3.1 Sample quality control and testing

Targeted DNA sequencing resulted in an average depth of 1278.4X across all samples profiled in this study. The study originally consisted of 500 samples, however based on the methodology described in Section 3.2.4.1, I tested for individual-level sample genotype matching and identified two problematic samples. Figure 3.3 shows the distribution of NDR values for samples that were labeled to be from the same individual versus those labeled to be from different individuals. Overall, within-patient NDR values were much lower then between-patient NDR values. However, a few points stool out as outliers in this boxplot for each category of samples.

A closer inspection identified that the tumor and uninvolved normal lung tissue of one individual (AIR_053) failed the test. The tumor and normal tissue for this individual were in fact concordant with those derived from patient AIR_052, suggesting a possible technical error while sequencing. Figure 3.4 shows a heatmap generated from correlation values computed using the presence (or absence) of a subset of variants, derived from the 1000 genomes project, in the samples from individuals AIR_052 and AIR_053. The tumor section and uninvolved normal lung parenchyma from individual AIR_053 clustered with samples from patient AIR_052. However, since this individual had profiles from core-needle biopsies to serve as a tumor comparator and blood to serve as a germline control, the individual was still included in the study after removing the two problematic samples (AIR_053 T and N).

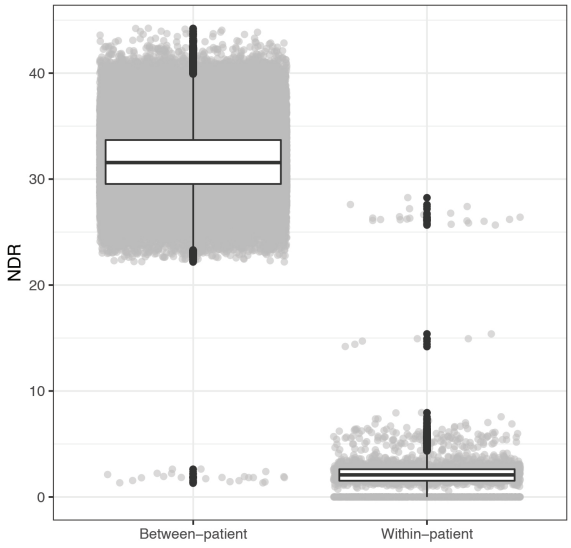


Figure 3.3: **Pairwise comparisons to test for individual-level concordance of samples.** NDR values were computed for paired samples when the two samples were labeled to be from the same individual versus those labeled to be from different individuals. Boxplots with overlaid points of each NDR value computed are shown.

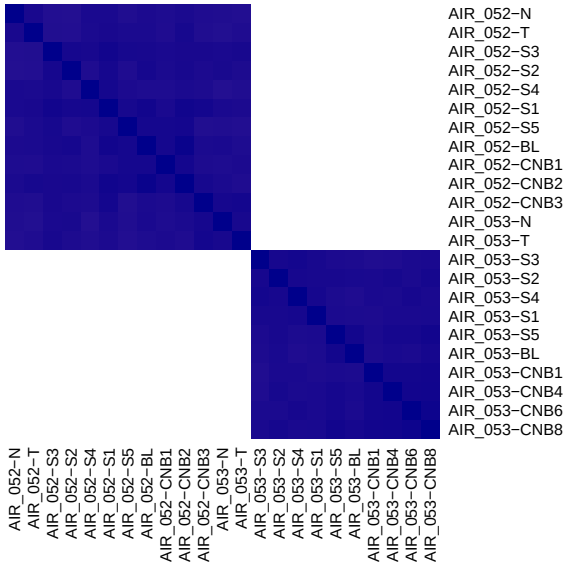


Figure 3.4: **Heatmap of correlation between samples from individuals AIR_052 and AIR_053.** A correlation based heatmap depicts the relationship between samples from individuals AIR_052 and AIR_053 in this study cohort.

This highlights the importance of sanity checking the cohort, especially when profiling large number of samples within an individual, to identify and exclude possibly problematic and error prone samples such as from contamination, sample swap and other technical errors. To enhance our focus on normal-appearing samples, we pooled the T and CNB samples collectively as a NSCLC set and S, L, N and Na (airway) samples were denoted as the cancerized field.

3.3.2 Somatic point mutation processes in the uninvolved normal-appearing field

3.3.2.1 Mutation burden

I identified 3,286 somatic mutations in 285 samples, mostly in NSCLCs (T or CNB; 3,017 mutations in 209 samples from 48 patients). I identified 269 somatic mutations in 76 airway field samples from 36 patients, the vast majority of which were observed in airways adjacent to the tumor (226 out of 269 field mutations). The overall airway field mutation burden decreased as distance from tumors increased (S: 226; L: 19; N: 16; Na: 8 mutations), a classical field phenotype (Figure 3.5).

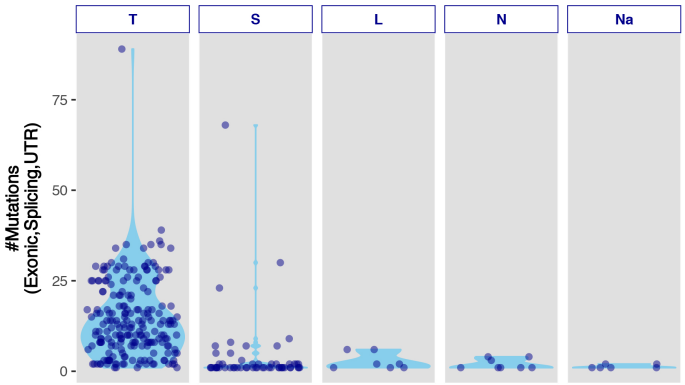


Figure 3.5: **Mutation burden in multi-region samples of normal-appearing airway epithelia and matched NSCLCs.** The total number of somatic SNVs across the airway field comprising multi-region samples from tumor-adjacent small airways (S), distant large airways (L), nasal epithelium (Na) and uninvolved normal lung tissue (N) as well as their matched NSCLCs (T) are depicted. Each point represents a single sample and plots within each sample type show somatic SNV burden distributions.

The mean somatic mutational burden was significantly higher in NSCLCs (13.9 mutations per sample) compared to the airway field (1.2 mutations per sample) (Figure 3.6). Although mutation burdens in lifetime smoker NSCLCs were significantly higher than in non-smokers (P <10-15), we observed marginal evidence for this pattern in the airway field (Figure 3.6).

Further interrogation of 27 cases with both multi-region CNBs and tumor sections identified 266 mutations unique to CNBs that were missed in whole tumor sections. Of these, 70 mutations from five cases were shared with the airway field. This attests to the importance of profiling multiple tumor samples to better capture heterogeneity and detect low frequency subclonal events.

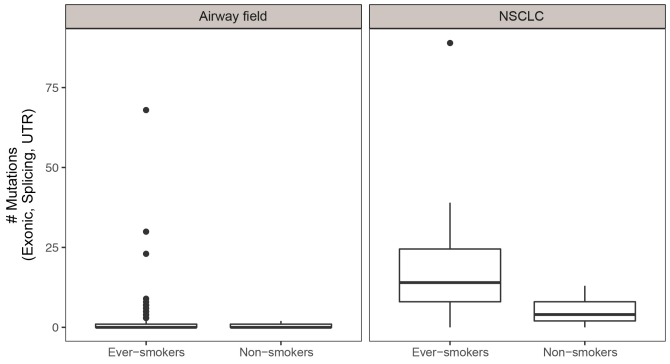


Figure 3.6: **Mutation burden differences between normal airway epithelia and NSCLCs based on tobacco history.** Somatic point mutations were identified in the airway field and matched NSCLCs (including multi-region tumor biopsies) from deep targeted DNA sequencing. Mutation burdens were plotted for each sample type (airway field and NSCLC) separately for non-smokers and smokers.

3.3.2.2 Mutation signatures in airway field of cancerization

The mutations identified in NSCLCs were then used to estimate mutation signatures. Despite the lower overall number of mutations captured in targeted panel sequencing, in comparison to whole genome and whole exome sequencing, I used the more aberrant NSCLC specimens to identify mutation signatures. For example, smoker NSCLCs exhibited more tobacco-associated [signature 4] C >A base substitutions compared to non-smoker tumors (Figure 3.7). Enrichment of this signature was also observed in smoker airway field sam-

ples albeit to a lesser extent (Figure 3.7), while using the NSCLC derived signatures as a background. Figure 3.7 shows the enrichment of signatures in NSCLCs (smokers and non-smokers) as well as field tissues of non-smokers.

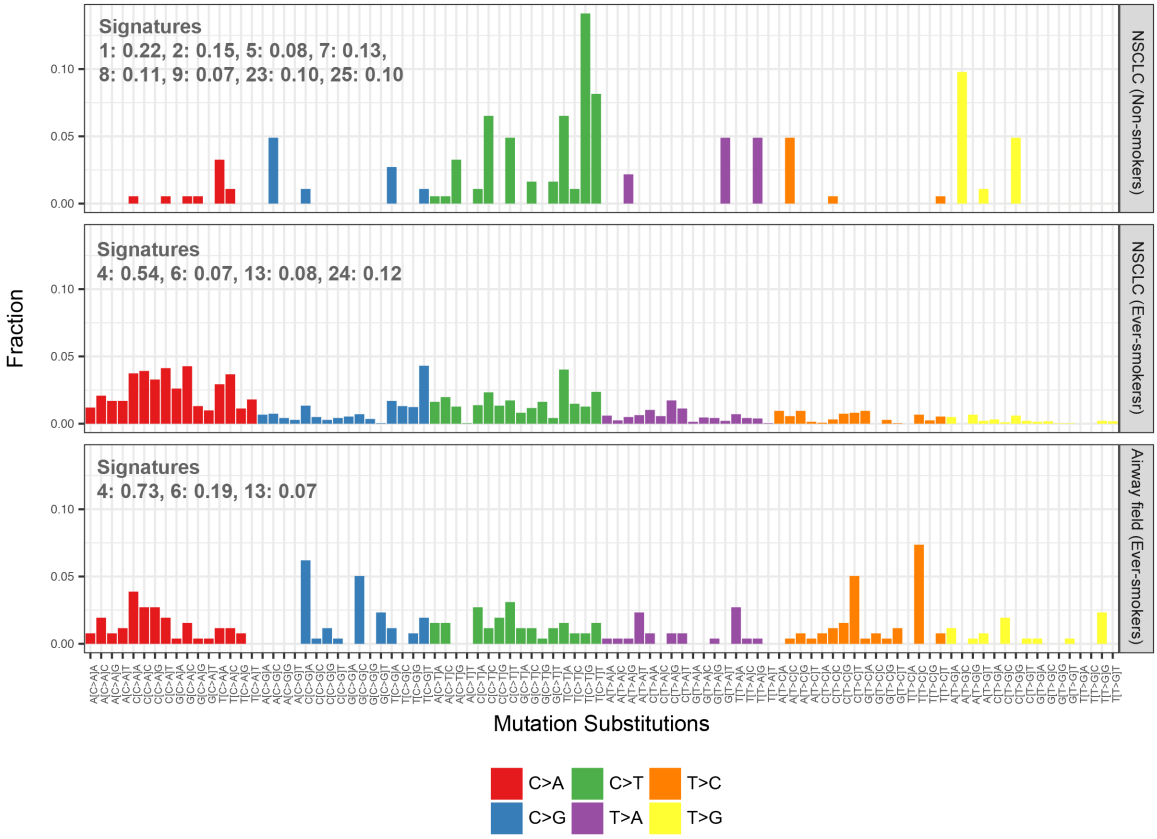


Figure 3.7: **Spectrum of base substitutions and mutational signatures in the normal-appearing airway field of smoker NSCLC patients.** Mutation substitution patterns in airway field and NSCLC were annotated and plotted based on tobacco history. Airway field of non-smokers was excluded due to low sample availability and lower mutation counts. The airway field samples from smokers were tested for enrichment of specific canonical mutational signatures using those identified in smoker NSCLCs as background.

3.3.2.3 Mutation frequencies and spatial field effects

I then tested for the presence of a spatial gradient in mutation VAFs based on the proximity of the field tissues to the tumor. Overall, VAFs of mutations in field samples decreased as distances from matched NSCLCs increased, as shown in Figure 3.8.

While Figure 3.8 shows the VAFs for all mutations observed within each tissue profiled and across all patients, it doesn't account for spatial distribution in VAFs of the

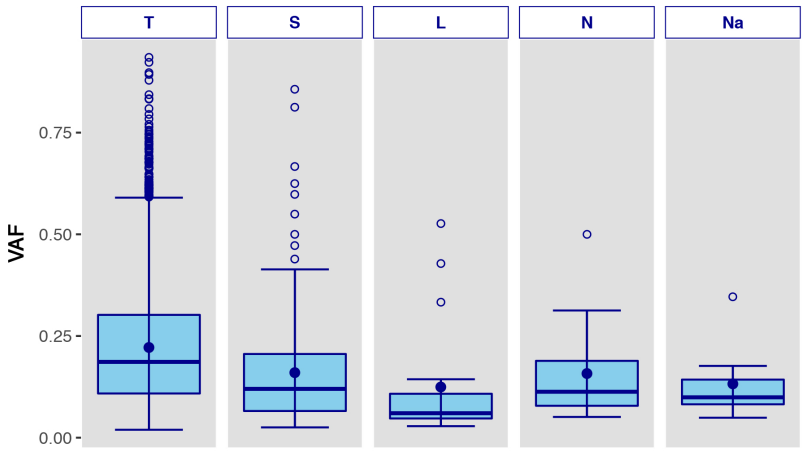


Figure 3.8: **Variant allele frequency distribution in airway field and NSCLCs.** Box plots demonstrating the VAF distributions of the identified SNVs across the multi-region samples: tumor-adjacent small airways (S), distant large airways (L), nasal epithelium (Na) and uninvolved normal lung tissue (N) as well as their matched NSCLCs (T) are shown.

same mutation in all tissues of the individual. To test this, I performed a statistical analysis of the airway field VAFs. Mutations in the tumor were tested for presence and variant allele frequency in each of the field tissues, using raw allelic depth intensities. Using the approach explained in Section 3.2.6.1 of this chapter, I obtained a net negative mean slope across all patients (-0.003). Figure 3.9 shows the weighted regression based slope for all 48 cases profiled in this study. These mutations were then used to perform a permutation analysis, comprising 1000 iterations of differentially ordered tissues, to test the significance of this negative slope. The result indicated that this mean negative slope was also statistically significant ($P = 0.03$; Figure 3.10) suggesting the presence of an overall spatial field effect, with clonal mosaicism enriched in tissues closer to NSCLC.

3.3.2.4 Driver mutation landscape of the cancerized field

I found 28 mutated bona fide drivers [5, 6, 42], among other genes, displaying protein damaging mutations in the airway field samples (Figure 3.11; Table 3.3). *TP53, KRAS, KEAP1, KMT2D, KMT2C, STK11, ATRX, IDH1* and *JAK1* were mutated in normal-appearing airway samples from more than one case with *KMT2C* and *TP53* being most

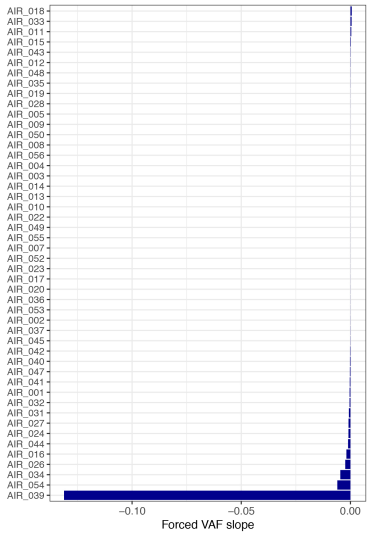


Figure 3.9: **Spatial analysis of variant allele frequencies in the normal airway field with respect to proximity to NSCLC.** Within patient variant allele frequencies (VAFs) were obtained for airway field tissues at sites that exhibited mutations in the matched NSCLCs. A weighted linear regression slope between the VAFs and ordered airway distances (based on their relative proximity to the NSCLC) was computed. A barplot of the derived slopes for all 48 cases is shown.

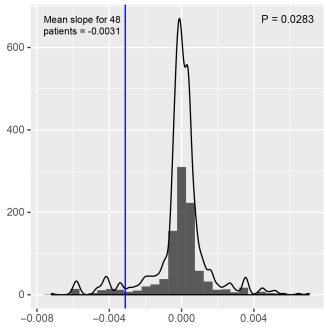


Figure 3.10: **Statistical testing of the spatial field effect using variant allele frequencies in the normal airway field with respect to proximity to NSCLC.** Permutation testing was performed to test the significance of the mean negative slope (blue line) obtained across all patients. A histogram of all permutation-inferred slope values along with the mean slope (blue line) are plotted.

recurrently mutated in the airway field (n = three cases; Figure 3.11). The majority (19 out of 28) of these genes exhibited the same mutation in matched NSCLCs, some of which were shared within multiple field samples for the patient (Figure 3.11; Table 3.3). NSCLC-adjacent small airways (21 samples from 13 cases) exhibited mutations in 22 driver genes. The majority (17 out of 22) of these genes were also mutated in the matched NSCLCs (Table 3.3). Relatively more distant large airways (n = 2) showed mutations in five genes - *CDKN2A, PIK3CA, SETD2, TP53, TSHR*; all were also mutated in matched NSCLCs (Table 3.3). Uninvolved normal lung tissue (n = 4) and nasal epithelium (n = 4) showed mutations in *RB1, RET, TSHR*; and *ATK1* respectively, none of which were shared with matched NSCLCs (Table 3.3). Airway field samples comprised mutations consistent with previously characterized variants specific to LUADs (*KRAS, STK11* and *KEAP1*) and LUSCs (*CDKN2A, PIK3CA* and *KMT2D*) or to both (*TP53*) (Figure 3.11).

Tissue	**Cancer associated mutated genes**
Small airway	*AMER1, ARID2, ATM, ATRX, CDC73, CDKN2A, CTNNB1, CYLD, DNMT3A, HRAS, IDH1, JAK1, JAK2, JAK3, KEAP1, KIT, KMT2C, KMT2D, KRAS, NOTCH1, STK11, TP53*
Large airway	*CDKN2A, PIK3CA, SETD2, TP53, TSHR*
Normal lung	*RB1, RET, TSHR*
Nasal epithelium	*AKT1*

Table 3.3: Cancer driver genes exhibiting mutations in different airway field tissues. Genes shown in red exhibited mutations that were shared with their matched NSCLC

Figure 3.12 shows cases exhibiting shared driver genes mutations [5, 6, 42] in matched airway field and NSCLC specimens. For genes mutated in both NSCLCs and airway field samples, the observed VAFs in the tumors were often higher than in field samples (Figure 3.12) - suggestive of selection-driven clonal expansion of the airway field. As an exception, case AIR_054 however showed a higher VAF in its field tissues compared to the matched NSCLCs. Cases AIR_039 and AIR_043 showed similar VAFs of driver mutations in matched airway field and NSCLCs.

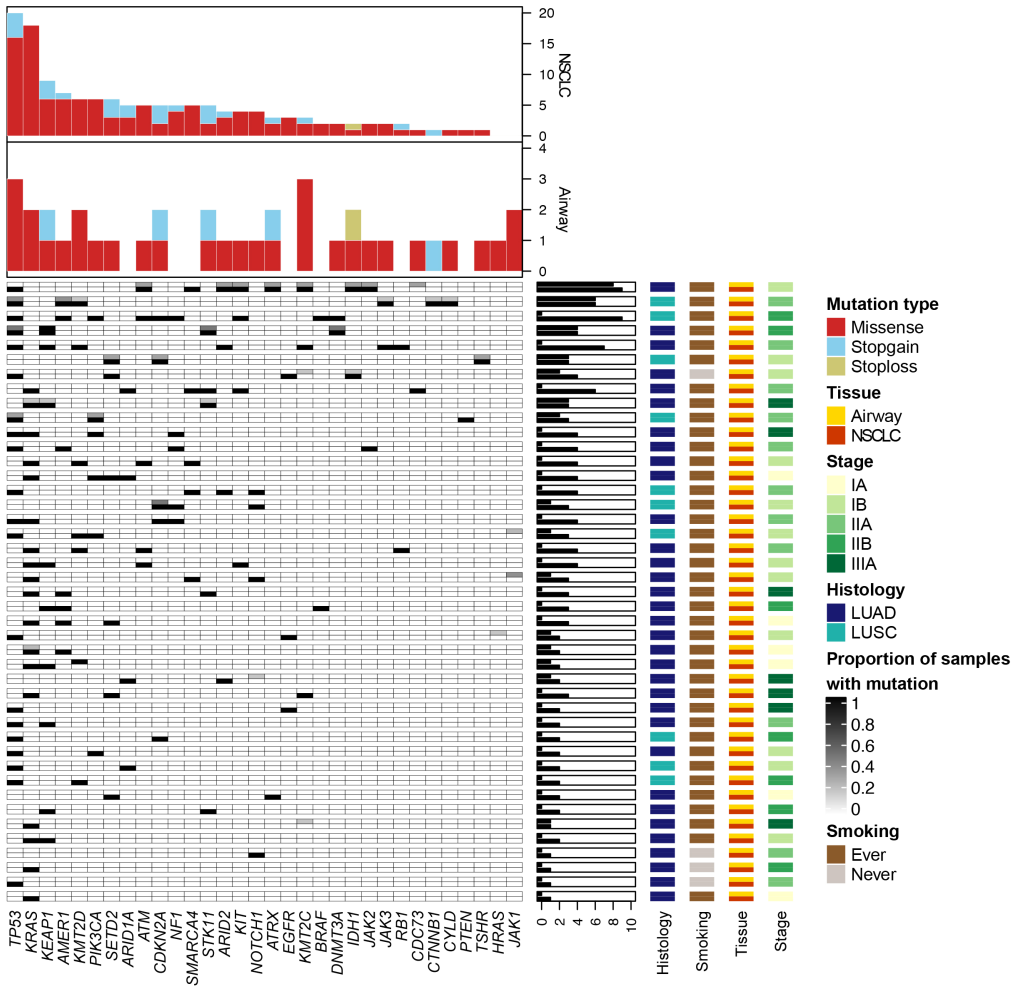


Figure 3.11: **Landscape of somatic driver mutations in the NSCLC-adjacent and -distant normal airway epithelium.** Somatic nonsynonymous (e.g., missense, nonsense and stoploss) variants in all airway field and matched NSCLCs were identified from targeted sequencing of a panel of 409 genes. Mutated genes previously implicated as drivers in NSCLC or other malignancies are shown for the airway field and tumor samples. Columns denote genes and rows represent individual patients. Each patient, denoted by a row, has the airway field presented on top half of the cell and the matched NSCLC in the bottom half. Mutated genes are color coded based on the proportion of airway samples carrying a variant within the gene (proportion range 0 to 1; white to black; right panel) and presence in the matched NSCLC (white: absent, black: present; right panel). The number of patients with the indicated driver mutated genes in the airway field and NSCLC are shown as bar plots (top panels). Annotations for stage, histology, smoking and tissue type (airway and NSCLCs) for all patients are also shown. Patients were ordered, top to bottom, based on airway field and NSCLC somatic mutation burdens (middle horizontal barplots).

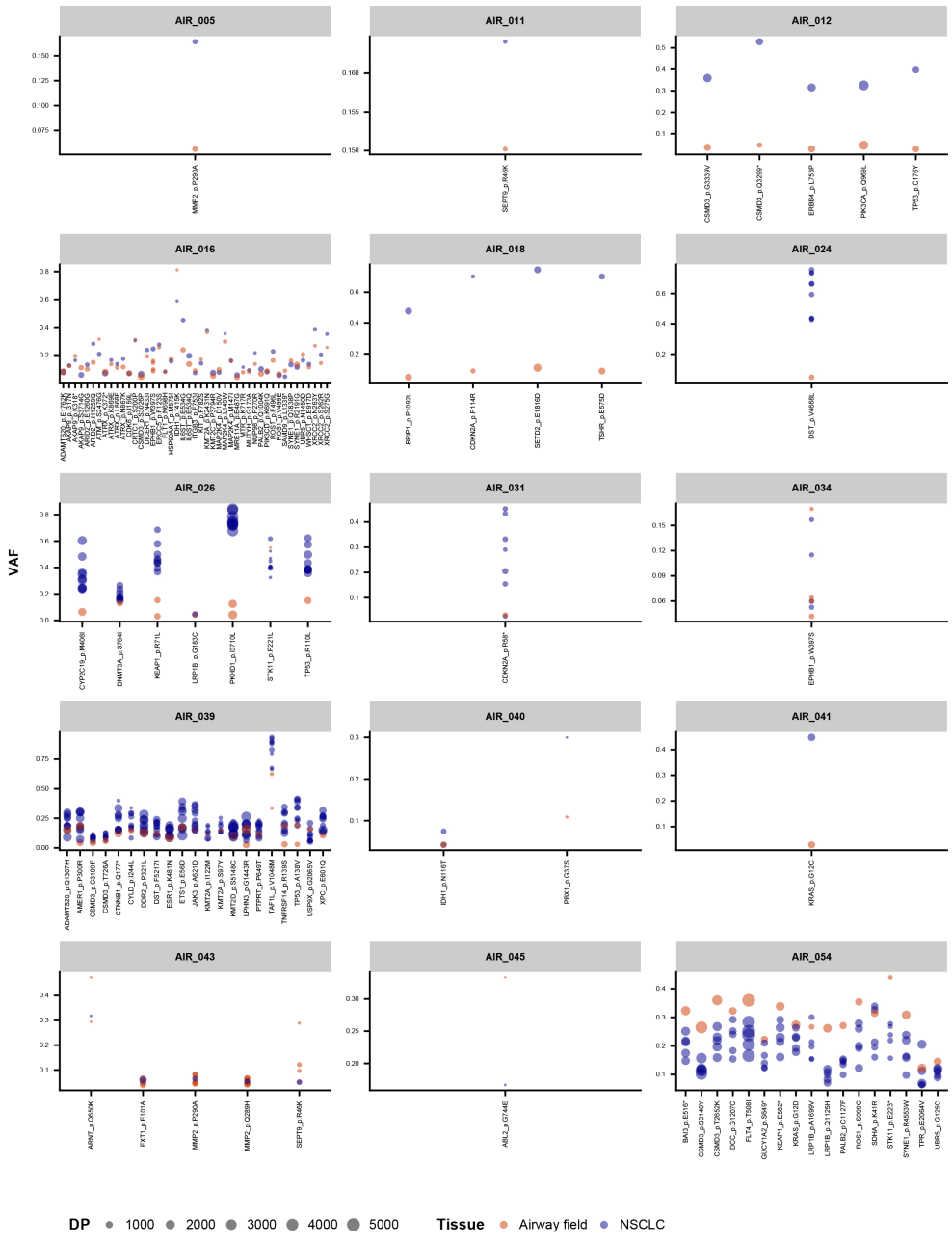


Figure 3.12: **Variant allele frequencies of shared driver mutations in NSCLCs relative to their normal airway cancerization field.** Allele frequencies of somatic variants identified to be shared between the airway field and NSCLC within patient were plotted. Each dot represents a mutation in a given sample within a patient. The size of the dot represents the total sequencing depth available at the locus and the color of the dot corresponds to the type of sample (normal airway field, red; NSCLC, purple).

3.3.3 Integrative mutational mechanisms in airway field carcinogenesis

I next identified and integrated previous findings of chromosomal mutations (events) leading to AI [13], to infer regions of allelic imbalance (AI) at a whole-genome scale, and integrated those with the identified somatic SNVs. NSCLCs exhibited a relatively high abundance of somatic SNVs and AI with an overall concordance in burden of these mutation types (rho = 0.43; P <10-10, Spearman rank test; Figure 3.13, right). Matched airway field samples also exhibited a positive relationship between AI and SNVs, albeit to a lesser extent (rho = 0.22; P <10-3; Figure 3.13, left).

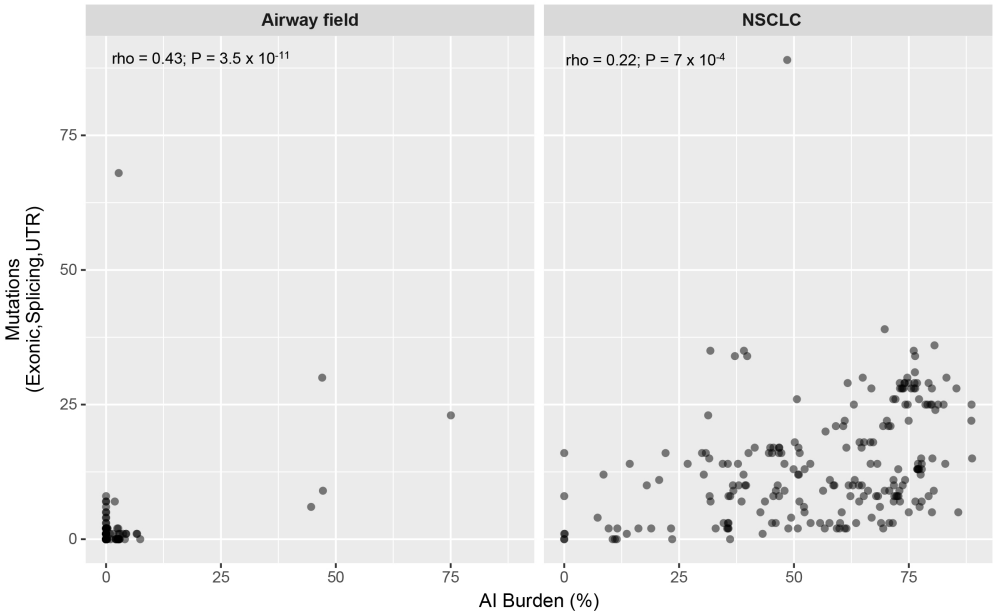


Figure 3.13: **Association of SNV and allelic imbalance somatic mutation burdens in the normal-appearing airway field and matched NSCLC.** The correlation between somatic SNV and AI burdens (percentage of aberrant genome) was tested in the normal-appearing airway field and NSCLC. A scatter plot of the two types of somatic mutation burdens as well as their correlation is shown for the normal airway field (left) and NSCLC (right). Each point represents an airway field or NSCLC sample profiled. Correlations between SNV and AI burdens were statistically evaluated using the Spearman rank test.

3.3.3.1 Field of cancerization area under the curve (FCAUC)

I then attempted to construct a genomic airway field phenotype using the field of cancerization area under the curve (FCAUC) measure. In Figure 3.14, I show the plots for three

cases that exhibit varying degrees of sharing with the matched primary, thereby resulting in different FCAUC values. When applied to the entire cohort of 48 individuals, I identified 25

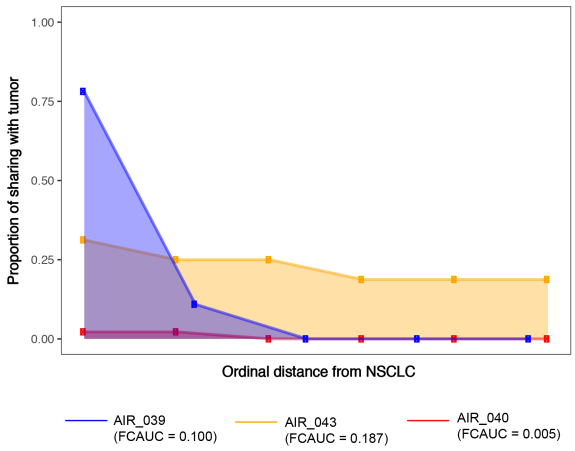


Figure 3.14: **Genomic field of cancerization quantification for three representative cases.** The ordinal field tissue distances and proportion of shared events in each field tissue with its matched NSCLC is shown for three representative cases. Similarly, FCAUC values were computed for all cases in the cohort.

cases that exhibited a non-zero FCAUC, indicating evidence for shared somatic aberrations between normal-appearing airway samples and matched NSCLCs. FCAUC values for all 48 cases is shown as a barplot in Figure 3.15. Specifically, a few individuals with high FCAUC values were identified; these are suggestive of higher degrees of field cancerization effects.

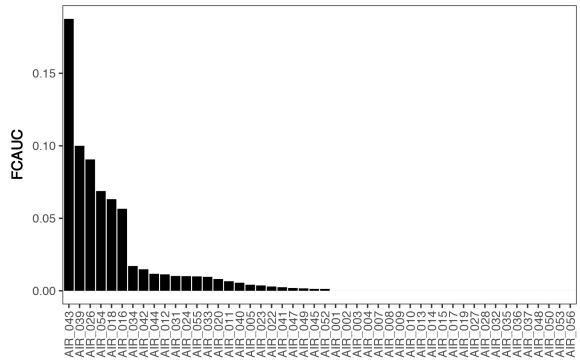


Figure 3.15: **Distribution of FCAUCs across all patients profiled.** A barplot with the FCAUC values for each individual is shown. Cases are ordered by their FCAUC value.

3.3.3.2 Phylogenetic assessment of field carcinogenesis

I performed phylogenetic analysis to further interrogate relationships between the field samples and matched NSCLCs within each individual. The combined mutation profile, consisting of both SNVs and large chromosomal arm allelic imbalance events was used. I also generated trees separately for SNVs and AIs detected in each patient. The tree topologies for cases with evidence for genomic field cancerization differ from a typically straight line that would be expected if only NSCLCs presented mutations. For a majority of patients (30/48), I observed a concordance of tree topologies between their separately-constructed SNV and AI profiles. Among the 48 patients, six patients (four LUADs and two LUSCs) exhibited phylogenetic tree structures with relatively pronounced airway field carcinogenesis phenotypes (Figure 3.16) – the same patients statistically determined above to markedly display field effects (Figure 3.15). These phylogenetic trees represent a serial inclusion of mutations, especially within airway epithelia, that might imply a temporal order of events in airway field cancerization and NSCLC pathogenesis. For example, cases AIR_039, AIR_016 and AIR_026 exhibited multiple airway specimens with differential mutation loads. Case AIR_039 likely presented an early TP53 mutation (S3) that was sequentially followed by additional hits comprising a *KMT2D* mutation, and chromosomal aberrations such as 3p loss and 17p loss in the relatively closer airway epithelium (S2) that eventually acquired additional mutations to progress to NSCLCs (e.g., 12p gain). Similarly, AIR_026 exhibited an early mutation in *KEAP1*, with subsequent *STK11* and *TP53* point mutations and 9q loss; the matched NSCLC exhibited additional events including *SMARCA4* mutation, 17p loss and a subtle 12p AI event. These trees also encompass driver genes with two-hit alterations; such as a driver with a shared mutation in both the airway field and NSCLC (e.g., *TP53* in AIR_026) but with an additional NSCLC-specific hit (e.g., 17p loss in AIR_026), therefore alluding to the two-hit model of progression of normal airway epithelia to NSCLC development, described in more detail in the next section. Overall, this analysis highlighted potential sequential patterns of mutations and key drivers in the progression of airway field to NSCLC. The phylogentic trees for all the cases, as well as trees generated separately for SNVs and AIs for each patient are provided in the Appendix.

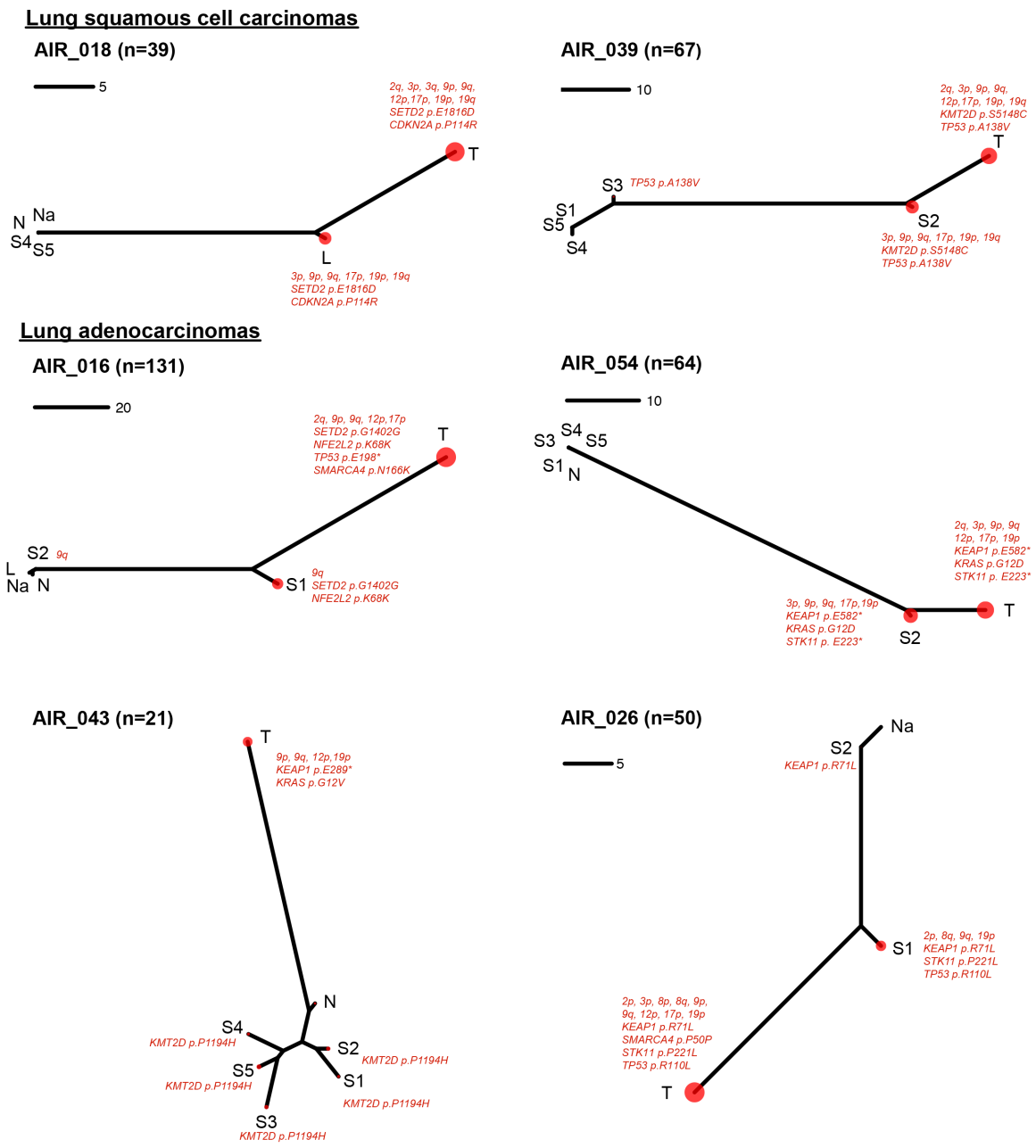


Figure 3.16: **Molecular spatial and temporal relationships between the normal airway cancerization field and early-stage NSCLC.** For every patient, the SNVs and AIs detected (n) across airway field and NSCLC tissues were integrated to generate unrooted neighbor-joining phylogenetic trees to study intra-patient multi-region samples. Six cases (two LUSC and four LUAD) with pronounced field effects are shown. The phylogenetic trees were annotated with mutations in known cancer associated genes as well as large chromosomal aberrations previously implicated in NSCLC pathogenesis. Each tree is accompanied by a scale to denote the number of mutations. The relative somatic burden for each tissue in a tree is denoted by a correspondingly sized red circle. The distances between the multiple points of a tree correspond to the extent of shared as well as as disparate mutational events among samples of a patient.

3.3.3.3 Somatic multi-hit oncophenotypic alterations in the cancerized field

In addition to looking at the presence (or absence) of SNVs and AI events in matched field and NSCLC tissues within an individual, I also probed for somatic two-hit alterations (genes with both SNVs and within AI events). Figure 3.17 shows the mutation patterns at known cancer driver genes [42], including NSCLC driver genes [5, 6]. Genes with two-hit aberrations are depicted in red and genes comprising either single SNVs or AI events are depicted in orange and yellow, respectively. Airway field of four cases exhibited two hits in known NSCLC drivers such as *TP53*/17p focal or whole-arm loss, *KRAS*/12p focal gain, *KEAP1*/19p arm loss, *STK11*/19p arm loss, *CDKN2A*/9p arm loss and *SETD2*/3p arm loss and that were also shared with matched NSCLCs (Figure 3.17). These hits are summarized in Table 3.4.

Sample	**Two-hit genes**
AIR_018-L	*CDKN2A, SETD2, TSHR*
AIR_026-S1	*DNMT3A, KEAP1, STK11, TP53*
AIR_039-S2	*CTNNB1, CYLD, JAK3, TP53*
AIR_054-S2	*KEAP1, KRAS, STK11*
AIR_055-S2	*JAK1*
AIR_055-S3	*JAK1*

Table 3.4: Airway field samples exhibiting somatic two-hit mutations in known cancer associated genes.

I expanded the analysis to include other known cancer-associated genes [42] and identified additional two-hit genes in the airway field such as *CYLD*/16q loss and *DNMT3A*/2p gain (Table 3.4, Figure 3.17). I also noted 12 cases whose airway field samples exhibited single hits (AI or SNV) for driver genes, and where the matched NSCLCs exhibited two-hits for the same genes resulting in first shared hit/second tumor hit pairs (Table 3.5). These included *CDKN2A*/9p cnLOH, *PIK3CA*/3q gain, *TP53*/17p cnLOH, *KRAS*/12p focal gain and *IDH1*/2q gain where the first (shared) hit is a SNV with a presumably subsequent AI event observed in matched NSCLCs; and pairs such as 9q cnLOH/*NOTCH1*, 9q cnLOH/*PTCH1*, 9q cnLOH/*ABL1*, 16q loss/*CDH1* and 2p gain/*MSH2*, where the first (shared) hit is an AI event with a presumably subsequent SNV in matched NSCLCs (Fig-

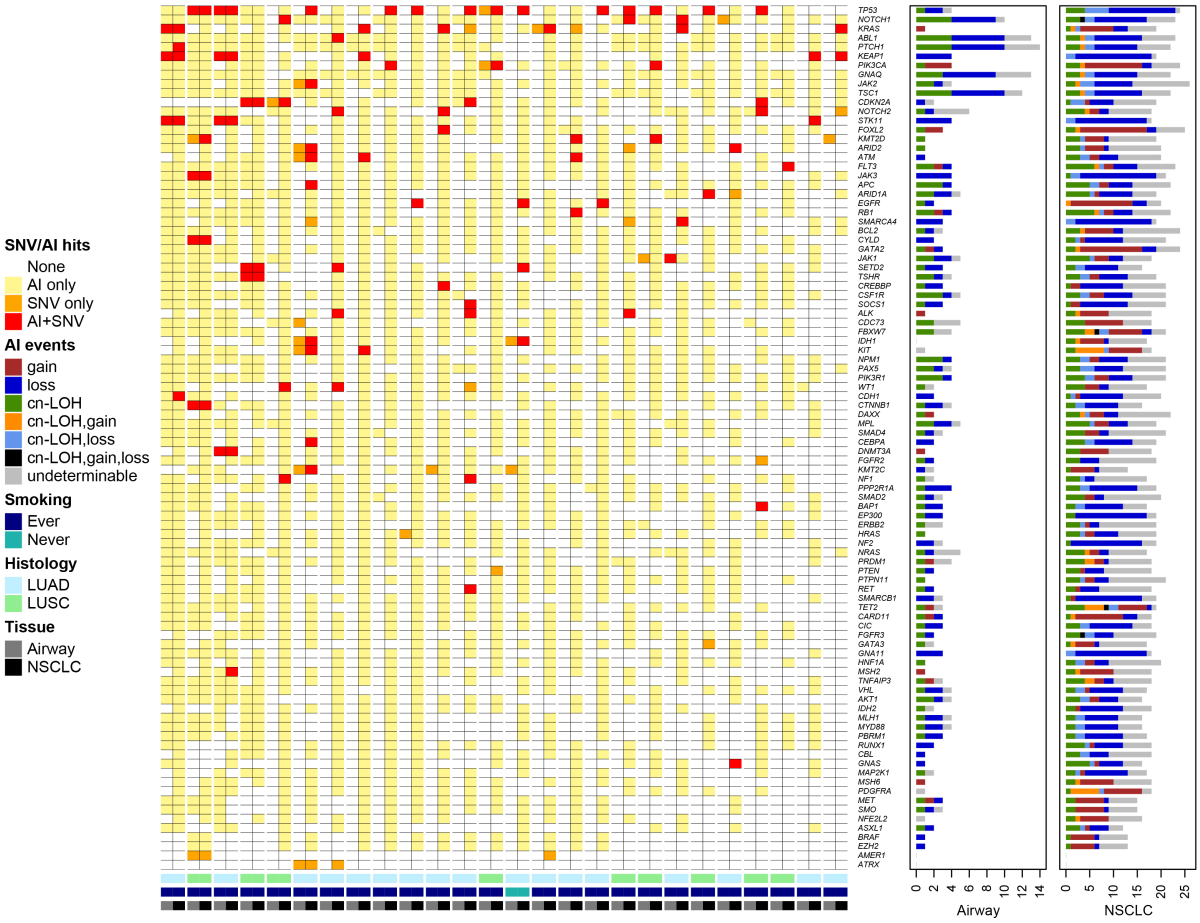


Figure 3.17: **Somatic two-hit aberrations in the adjacent and distant normal airway epithelium of early-stage NSCLC patients.** Data from deep DNA sequencing and SNP array profiling were integrated to identify NSCLC-associated drivers that comprised either somatic SNVs or AI as well as genes with two-hit aberrations (both SNVs and AI) in the airway field and NSCLC samples. Columns and rows represent patients and NSCLC-associated driver genes, respectively. Each column denotes a patient with the left half of the cell corresponding to the airway field (grey) and right half (black) to its matched NSCLC. NSCLC-associated driver genes are ordered top to bottom based on overall two-hit and single-hit patterns in the airway field and NSCLC; the cases (columns) are ordered left to right based on overall burden of somatic hits across these genes. The detected AI events were annotated as gain (brown), loss (blue), cnLOH (green) and undeterminable (grey). Events exhibiting intra-tumor heterogeneity within multi-region tumor samples (e.g., one CNB with a cnLOH and another biopsy from the same tumor showing a copy gain for the same chromosomal region: cnLOH,gain) are annotated separately.

ure 3.17, Table 3.5). I identified other examples exhibiting this pattern involving subtle (i.e, from a low fraction of cells exhibiting the alteration) AI events in the genome that spanned mutated genes such as *ARID2, ATM, CDKN2A, IDH1, KIT, KMT2D, KMT2C, JAK2* and *NOTCH2* (Table 3.5). Therefore, these somatic AIs and SNVs can offer insights into NSCLC pathogenesis by suggesting temporal ordering of events in the development of NSCLCs from a cancerized field.

Sample	**Shared first hit**	**Second NSCLC hit**
AIR_012-L	*PIK3CA*	3q gain
AIR_012-L	*TP53*	17p cnLOH
AIR_016-S1	*ARID2*	12q undeterminable
AIR_016-S1	*ATM*	11q undeterminable
AIR_016-S1	*IDH1*	2q undeterminable
AIR_016-S1	*JAK2*	9p undeterminable
AIR_016-S1	*KIT*	4q undeterminable
AIR_016-S1	*KMT2C*	7q undeterminable
AIR_017-S4	1p undeterminable (focal)	*NOTCH2*
AIR_024-S1	9q cnLOH	*NOTCH1*
AIR_026-S1	2p gain	*MSH2*
AIR_031-S2	*CDKN2A*	9p cnLOH
AIR_033-S1	9q cnLOH	*ABL1*
AIR_033-S2	9q cnLOH	*ABL1*
AIR_039-S2	*KMT2D*	12q undeterminable
AIR_040-S2	*IDH1*	2q gain
AIR_041-S1	*KRAS*	12p gain (focal)
AIR_042-S2	9p undeterminable	*CDKN2A*
AIR_054-S2	9q cnLOH	*PTCH1*
AIR_054-S2	16q loss	*CDH1*

Table 3.5: Airway field samples with shared first somatic hit and matched NSCLC-specific second somatic hit.

3.4 Discussion

Previously, somatic mutations in the *EGFR* oncogene have been identified in normal epithelium of *EGFR*-mutant LUADs [19] and *KRAS* mutations have been detected in lung tissue adjacent to resected tumors [17]. Yet, the spectrum of somatic driver mutations and genes in the normal-appearing airway cancerization field is not known. In this chapter, I presented the most comprehensive analysis of genomic aberrations in normal-appearing air-

way epithelium to this date. Our study consisted of a rich cohort of 498 samples comprising multi-region and spatially distributed airway and NSCLC specimens, along with germline samples, from 48 patients to identify the landscape of somatic point mutations and allelic imbalance (AI) in the normal-appearing airway. The adjacent and distant-to-tumor uninvolved normal-appearing airway field comprised somatic mutations in key drivers that were present at higher allele frequencies in the matched NSCLCs. We also identified key driver genes with shared two-hit alterations (both SNVs and AI) in the airway field as well as those with single hits in the field coupled with NSCLC-specific mutations. Findings from our study offer insights into a continuum of alterations with plausible spatiotemporal properties in the normal-appearing airway epithelium and nearby NSCLC.

3.4.1 Significance of findings

Here, I found that genomic airway field cancerization phenotypes were identified in over 50% of the cases suggesting that airway field carcinogenesis is not uncommon in early-stage NSCLC. Given our focus on somatic genomic changes, I contrasted genomic profiles in airways and tumors to peripheral blood cells or distant normal lung parenchyma in each patient, thus enabling us to focus on likely pathogenic changes in the lungs of NSCLC patients. Additionally, we not only pinpointed driver alterations (e.g. *KRAS* and *PIK3CA* mutations) in the normal airway epithelium (both adjacent and distant) but also demonstrated that these changes were shared with the NSCLC with many occurring sequentially – lending further confidence to the probable role of these genomic changes in pathogenesis of this malignancy from the epithelial cancerization field. These changes may hint at mechanisms that underlie tumor recurrence or development of second primary tumors in the remaining lung following surgical treatment with curative intent.

In this study, I present complementary findings on genomic airway cancerization that allude to potential clonally selected changes in the transition of normal airway field to NSCLC: overall increased somatic VAFs in NSCLCs relative to the airway field; shared mutated driver genes between the airway field and NSCLCs; acquisition of additional driver events in the NSCLCs themselves along with overall increased driver gene VAF in the tumors. We surmise that these genomic airway field cancerization changes provide insights

into spatial and temporal development of NSCLCs. If so, studies that include longitudinally profiled airway field samples from lung cancer patients and/or smokers who are free of lung cancer would bear this out. Mutations in key drivers that we find here to be shared between the airway field and NSCLC were previously described as truncal mutations in intra-tumor heterogeneity (ITH) studies of NSCLC [70, 87]. It is noteworthy that our present analysis of multi-region NSCLC biopsies allowed us to capture more truncal tumoral mutations that are shared with the field.

It is noteworthy that several driver genes I found to be mutated in the airway field have been previously implicated in lung preneoplasia. For instance, our previous study and that of Izumchenko and colleagues identified mutations in *TP53* and *KRAS* in atypical adenomatous hyperplasia (AAH), the precursor lesion to LUAD [40, 44]. Specific to my survey of AAHs (Chapter 2), I found that *KEAP1* and *KMT2D* that were previously identified to be mutated in AAHs, were also mutated in the airway field and matched NSCLCs in this cohort, indicating that, under certain selective pressures, the cancerized field may evolve to preneoplastic and ultimately to malignant lesions [10, 88].

In conclusion, our analysis of a rich set of spatially distributed multi-region normal airway epithelia and early-stage NSCLCs illuminated somatic variants of key driver mutated genes in the airway field that were mostly shared with matched NSCLCs. Overall, these somatic variants were positively selected in the tumors suggestive of clonal expansion in NSCLC. Our study not only points to early mutational processes that likely demarcate key events in the emergence of NSCLC from the normal-appearing cancerization field but can also pave the way for similar interrogations in other malignancies. These airway field changes may comprise potential targets for early treatment (e.g., adjuvant therapy to prevent tumor recurrence) of NSCLCs.

3.4.2 Limitations

While our study provides a comprehensive characterization of field cancerization in early-stage NSCLCs, it is unable to decompose these effects further due to the lack of control subjects. An additional question that remains unanswered is the clinical implications of identifying mutations in normal-appearing cancerized fields.

For the purpose of this analysis that was focused on characterizing the normal-appearing airway field, a combined tumor profile consisting of the tumor section as well as multiple CNB specimens, where available, was used. However, the significance of CNBs in the field cancerization, perhaps to test for the presence of truncal tumor events, is not leveraged here.

While we do hypothesize that the cancerized field gives rise to preneoplastic and malignant phenotypes, it is however plausible that airway field mutations may not be clonally selected in progression to lung premalignant and malignant phenotypes. For instance, previous studies have described driver mutations (e.g., *BRAF, KRAS*) that are relatively more frequent in AAHs compared with LUADs [32, 44].

Nonetheless, our study of multiple tissues from early stage NSCLC patients points toward the future of genomic investigations in medicine, where evolutionary trajectories can be inferred from diverse spatial or temporal sampling. The molecular changes we detect in these normal tissues present as early alterations in the transition to the malignant phenotype of NSCLC.

CHAPTER 4

INVESTIGATION OF PAN-CANCER PATTERNS OF CHROMOSOMAL ALLELIC IMBALANCE IN THE CANCER GENOME ATLAS

The Cancer Genome Atlas (TCGA) provides a large repository of tumor specimens across multiple tissue types and varied clinicopathological features, thus facilitating several pan-cancer studies of cancer aneuploidy and tumor-specific copy-number signatures, such as those derived from single nucleotide polymorphism (SNP) genotyping array platforms [89, 90, 91]. However, unique challenges compound the automated detection of chromosomal somatic copy number alterations (SCNAs) from SNP arrays. First, the tumor samples are often contaminated with normal cells and thereby necessitate more sensitive algorithms to identify subtle events in possibly low cellularity samples. Second, most SCNA detection methods report genomic regions of copy number alterations and their segment mean copy number (CN) estimates, the characterization of which heavily relies on accurate identification of non-aberrant regions of the genome to establish a baseline signal intensity representative of neutral CN. However, tumor samples exhibiting high levels of genomic instability pose a challenge for such analyses.

Output from SNP genotyping arrays include the following two measurements per site: the B-allele frequency (BAF), representing the ratio of the alleles at a locus; and log R ratio (LRR), the total intensity of both allelic probes at the locus. The former is used to identify regions of allelic imbalance (AI) while the latter is used for the characterization of the identified events. Detection of AI can lead to the identification of SCNAs as well as provide a largely unexplored, yet valuable class of chromosomal aberrations called copy-neutral loss of heterozygosity (cnLOH). cnLOH events represent regions of zero net copy number change, but present more severely altered ratio of alleles (e.g., change of germline heterozygous loci AB to AA or BB, resulting in a 2:0 or 0:2 ratio). The landscape of cnLOH regions in TCGA remains largely unknown due to the lack of algorithms for its automated detection. Therefore, it is important to identify the landscape of these cnLOH events across

tumor sites to better understand their role in the complex mechanisms of tumorigenesis such as in two-hit modes of pathogenesis of tumors.

In this chapter, I describe my approach to identify the pan-cancer genomic landscape of chromosomal allelic imbalance in the TCGA. I highlight the previously uncharacterized set of cnLOH events and interrogate these for additional diagnostic or clinical implications. I end the chapter by comparing our findings with previously reported SCNAs in this dataset, and propose an automated method for the identification and adjustment of putative problematic cases in TCGA.

4.1 Study Design

I present a comprehensive characterization of the pan-cancer atlas of allelic imbalance derived copy number changes (e.g., gain, loss) as well as regions of cnLOH (Figure 4.1). I utilized a sensitive haplotype-based framework, hapLOH, to identify regions in the genome that exhibit AI, a deviation from the expected 1:1 ratio at germline heterozygous loci [48]. SNP genotype profiling of 11,074 paired tumor-normal tissues across 33 tumor types in TCGA were studied. AI derived events, particularly cnLOH events, were assessed for global pan-cancer patterns as well as for tissue site-specificity. An automated method to identify and adjust cases with putative problematic calls as compared to previously reported SCNAs, was applied to the entire cohort, as described in Figure 4.1.

4.2 Methods

4.2.1 Dataset

The Level 1 raw CEL files from Affymetrix Genome-Wide Human SNP Array 6.0 profiling of 11,074 paired tumor-normal samples across 33 cancer sites in The Cancer Genome Atlas (TCGA) were downloaded from the Genomic Data Commons data portal, along with available clinical annotations. The cohort consisted of a majority of primary tumors (n=10,680), with a few metastatic specimens (n = 394; SKCM accounting for 368 of these samples). SNP metrics such as genotypes, BAF and LRR were used to reanalyze the dataset. The cohort is summarized in Table 4.1.

Tumor	**TCGA abbreviation**	**Number of tumor samples**
Adrenocortical carcinoma	ACC	90
Bladder Urothelial Carcinoma	BLCA	411
Breast invasive carcinoma	BRCA	1101
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	303
Cholangiocarcinoma	CHOL	36
Colon adenocarcinoma	COAD	462
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	48
Esophageal carcinoma	ESCA	186
Glioblastoma multiforme	GBM	545
Head and Neck squamous cell carcinoma	HNSC	527
Kidney Chromophobe	KICH	66
Kidney renal clear cell carcinoma	KIRC	529
Kidney renal papillary cell carcinoma	KIRP	288
Acute Myeloid Leukemia	LAML	200
Brain Lower Grade Glioma	LGG	527
Liver hepatocellular carcinoma	LIHC	377
Lung adenocarcinoma	LUAD	517
Lung squamous cell carcinoma	LUSC	502
Mesothelioma	MESO	87
Ovarian serous cystadenocarcinoma	OV	595
Pancreatic adenocarcinoma	PAAD	185
Pheochromocytoma and Paraganglioma	PCPG	183
Prostate adenocarcinoma	PRAD	496
Rectum adenocarcinoma	READ	168
Sarcoma	SARC	261
Skin Cutaneous Melanoma	SKCM	472
Stomach adenocarcinoma	STAD	442
Testicular Germ Cell Tumors	TGCT	156
Thyroid carcinoma	THCA	509
Thymoma	THYM	124
Uterine Corpus Endometrial Carcinoma	UCEC	546
Uterine Carcinosarcoma	UCS	55
Uveal Melanoma	UVM	80
		11074

Table 4.1: Summary of the tumor sites and samples analyzed across TCGA

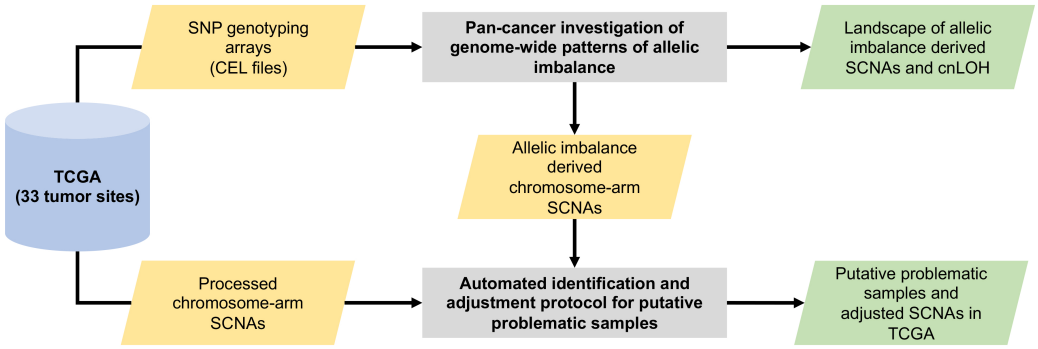


Figure 4.1: **Study design to identify, compare and contrast chromosomal aberrations in TCGA.** A comprehensive characterization of allelic imbalance derived landscape of somatic copy number alterations (SCNAs) as well as cnLOH in TCGA was carried out. An automated approach to compare findings from this study with previously reported events in the TCGA database was developed. Putative problematic samples were highlighted and an automated adjustment procedure was applied to rectify these calls.

4.2.2 Pan-cancer allelic imbalance profiles using hapLOH

For each tumor sample in TCGA, the corresponding control sample (blood or tumor adjacent normal tissue) was assigned and phasing was performed using MaCH for the statistical reconstruction of haplotypes. The phased genotypes as well as the BAFs were supplied as inputs to run hapLOH using default parameters. The resulting regions of allelic imbalance as inferred from hapLOH were then characterized based on the extent of BAF and LRR deviations for each event region. Regions with LRR deviation ≥ 0.05 were classified as gains while those with LRR deviations of $\leq$ -0.05 were classified as losses. The remaining events were characterized as copy-neutral loss of heterozygosity if their BAF deviation was >0.1. The events that were unable to be characterized into these different types were deemed to be too subtle, due to low mutant cell fraction, for annotation. Of note, AI events with LRR $\geq$ 0.08 and lesser than 2Mb in length were excluded as likely inherited duplications. The allelic imbalance events spanning more than 70% of the genome were considered chromosomal-arm level events, while the remaining were annotated as focal events. For each tumor sample, the total number of events identified was used as a measure of its count burden, while the percent of its genome under allelic imbalance was used as a measure of its genomic burden.

4.2.3 TCGA pan-cancer copy number profiles

The processed chromosome arm-level SCNA files were downloaded from Broad GDAC Firehose. The latest analyses version at the time of download was dated `2016_01_28`. The `broad_values_by_arm.txt` file for each TCGA study was then processed into `bed` formatted files using the specified amplification/deletion threshold of 0.1. These results were compared with allelic imbalance derived chromosome-arm level SCNAs identified by hapLOH for the same individuals.

4.2.4 Identification of putative problematic calls in TCGA

Allelic imbalance events identified in each tumor sample were reassessed at a chromosomal-arm level, by concatenating events on each arm to identify specific chromosome-arms that exhibited events at a fraction ≥ 0.7 of the arm, using a custom script. For the purpose of SCNA comparisons, cnLOH events and subtle unclassified events were excluded from this analysis. For every marker genotyped in the array, the presence (or absence) of an event spanning the marker in both TCGA and hapLOH derived event calls were annotated as 1 (or 0) respectively. A Pearson correlation coefficient was computed from all markers. Samples with a negative correlation were identified as potentially discordant and hence putative problematic samples.

4.2.5 Automated adjustment of potentially problematic calls in TCGA

For each of the negatively correlated tumor samples identified through the procedure describe above, the normal region, as determined by hapLOH was identified. Events reported by TCGA within these normal regions, as well as those that were identified as normal by both methods were identified. A new weighted median copy number was calculated from these events, weighed by the length of the event. The original calls made by TCGA were recalibrated using this newly determined normal copy number. Using the same specifications as before, i.e. a amplification/deletion threshold of 0.1, the new set of chromosome-arm event calls were reclassified. A correlation between these adjusted TCGA SCNA calls and hapLOH derived SCNAs was calculated in a manner explained in the previous section.

4.3 Results

Tumor genomes often exhibit high genomic instability, making the identification of copy number alterations challenging due to limited normal region in their genomes. Here, a sensitive haplotype-based technique was applied to identify the landscape of chromosomal copy number changes (e.g.; gain, loss) as well as previously uncharacterized landscape of cnLOH events from a survey of 11,074 paired tumor-normal specimens across 33 tumor types in the TCGA. The cohort is summarized in Table 4.1. I also present the performance of the automated procedure I developed to identify and adjust putative problematic cases.

Shown in Figure 4.2, are two pancreatic adenocarcinoma (PAAD) tumor samples that motivate the utility of B-allele frequency (BAF) in the identification of chromosomal aberrations. In Figure 4.2A, the tumor sample showed high concordance between hapLOH-derived SCNAs and those reported in TCGA. In such cases, our approach of a BAF-derived AI estimator supplements the database with additional, potentially impactful, chromosomal aberrations. In this particular example, our method identified additional chromosome-arm level cnLOH on chromosome 22 and arm 20p. Through this investigation, we aim to supplement the SCNAs in TCGA using BAF patterns, a complementary data element to the LRR-derived signal estimates, to identify chromosomal aberrations leading to allelic imbalance. On the contrary, the PAAD tumor sample shown in Figure 4.2B showed a complete discordance between the two methods. Here, the incorporation of deviations in BAF suggested an incorrect estimation of the normal region. The SCNAs reported in TCGA did not align with BAF deviations. However, the SCNAs, after adjustment, align with deviations in BAF and thereby are concordant with hapLOH event calls. Through this approach, we address differences between the call sets and suggest an automated method to adjust specific cases with potentially problematic calls to aid the database with more accurate SCNAs.

4.3.1 Pan-cancer allelic imbalance burden

Our method identified at least one AI event in 10,411 cases (94%), with a median count burden of 20 events per case, spanning on average 31.88% of the genome (genomic burden).

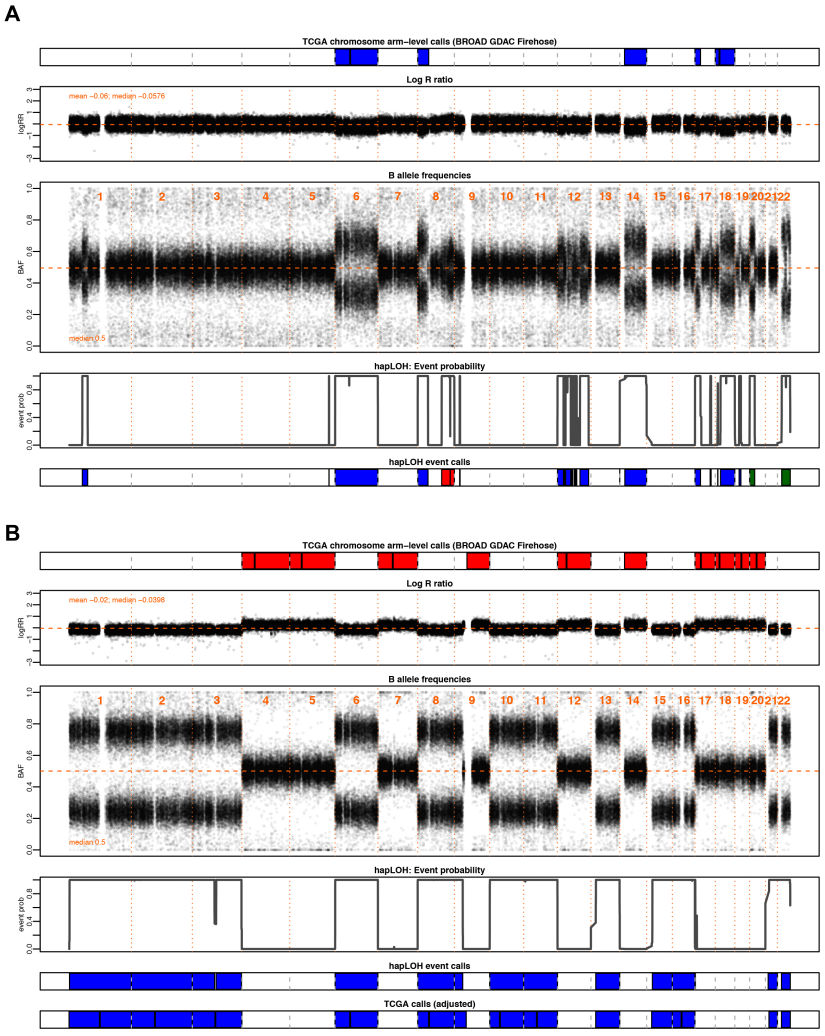


Figure 4.2: **Examples highlighting the utility of a B-allele frequency in the identification of chromosomal aberrations.** Two examples of pancreatic adenocarcinoma (PAAD) tumor samples that motivate the implementation of a BAF-based approach are shown. The tumor samples are annotated with chromosomal arm-level events downloaded from BROAD GDAC Firehose along with the BAF and LRR distributions of the markers profiled across the genome for that individual. Below these panels, are shown the event probabilities inferred from hapLOH using the shifts in BAF, as well as classified event calls from hapLOH using a threshold-based approach from BAF and LRR deviations, for the identified event boundaries. Although all hapLOH events are shown, only chromosomal-arm level events were used for the comparison to SCNAs identified in TCGA. (A) An example of a PAAD tumor sample that exhibited overall concordance between the two call sets, with additional cnLOH events identified by hapLOH. (B) An example of a PAAD tumor sample with discordant calls between the two approaches. An automated adjustment approach was applied, the result of which is shown in the at the bottom of this panel.

But with high variation and cancer site specificity in this value, the patterns of genomic burden and count burden were further assessed for each tumor site independently, to account for the variability in number of samples and in molecular complexities across tumor types (Figure 4.3, Table 4.2).

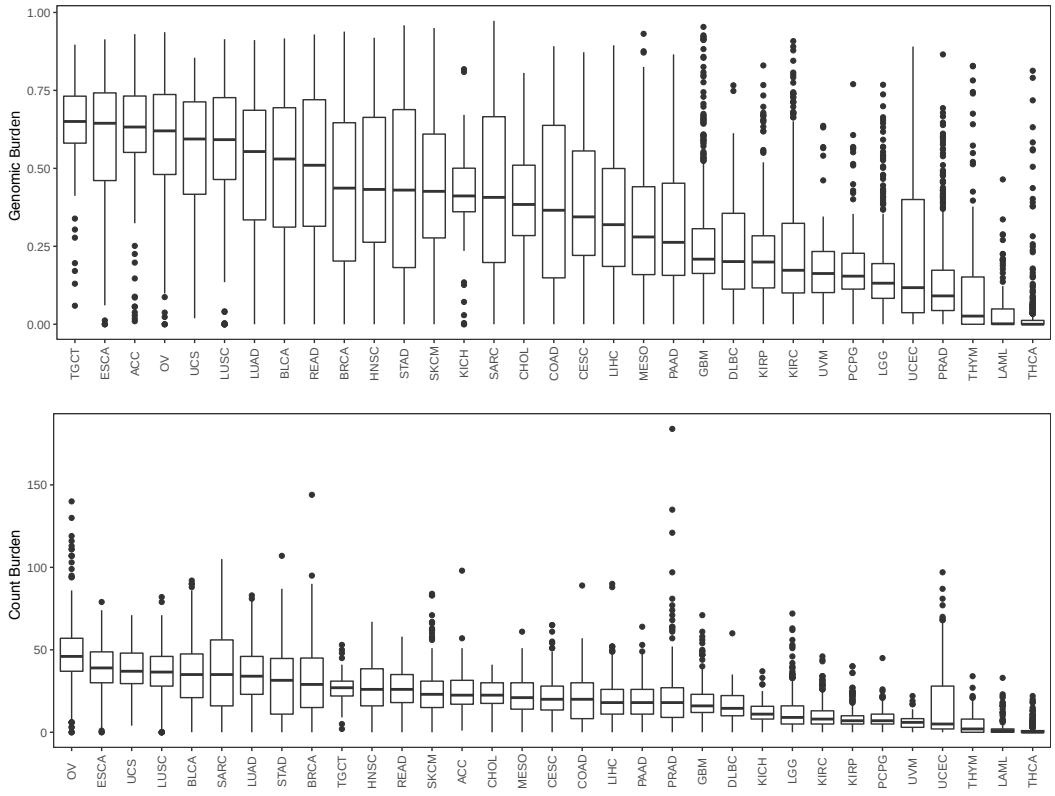


Figure 4.3: **Distribution of genomic burden and count burden across 33 tumor sites.** Boxplots of the overall genomic burden and count burden for the tumors studied across 33 sites in TCGA are shown. The tumor sites are ordered by their median burden for each plot.

Ovarian carcinoma (OV) showed the highest median count burden of 46, followed by esophageal carcinoma (ESCA) with median count burden of 39 (Figure 4.3, Table 4.2). Lung squamous cell carcinoma (LUSC), sarcomas (SARC), uterine carcinosarcoma (UCS), lung adenocarcinomas (LUAD) and bladder cancer (BLCA) showed median count burdens between 34 and 37 events (Figure 4.3, Table 4.2). At the other end of the spectrum were thyroid carcinoma (THCA), acute myeloid leukemia (LAML), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uveal melanoma (UVM) that exhibited a median count burden of less than five events (Figure 4.3, Table 4.2). Alternatively, tumors were also

Tumor	Total samples	Samples with AI	Samples with AI (%)	Count burden (Mean)	Count burden (Median)	Genomic burden (Mean)	Genomic burden (Median)
ACC	90	90	100.00	25.08	22.5	0.5802	0.6326
BLCA	411	408	99.27	35.22	35	0.4924	0.5300
BRCA	1101	1093	99.27	31.21	29	0.4292	0.4366
CESC	303	301	99.34	21.74	20	0.3885	0.3444
CHOL	36	34	94.44	22.78	22.5	0.3928	0.3842
COAD	462	455	98.48	20.41	20	0.3927	0.3656
DLBC	48	47	97.92	16.65	14.5	0.2527	0.2011
ESCA	186	184	98.92	38.87	39	0.5954	0.6446
GBM	545	539	98.90	18.12	16	0.2656	0.2088
HNSC	527	526	99.81	27.67	26	0.4499	0.4324
KICH	66	64	96.97	12.32	11	0.4116	0.4115
KIRC	529	521	98.49	9.82	8	0.2387	0.1731
KIRP	288	281	97.57	8.59	7	0.2229	0.1996
LAML	200	106	53.00	2.37	1	0.0368	0.0014
LGG	527	517	98.10	11.92	9	0.1617	0.1317
LIHC	377	369	97.88	19.37	18	0.3509	0.3194
LUAD	517	514	99.42	33.98	34	0.5105	0.5539
LUSC	502	495	98.61	36.62	36.5	0.5858	0.5921
MESO	87	85	97.70	21.49	21	0.3221	0.2800
OV	595	592	99.50	47.23	46	0.6010	0.6202
PAAD	185	170	91.89	18.82	18	0.2994	0.2627
PCPG	183	179	97.81	8.30	7	0.1839	0.1543
PRAD	496	465	93.75	20.15	18	0.1369	0.0910
READ	168	167	99.40	26.53	26	0.5003	0.5100
SARC	261	253	96.93	38.21	35	0.4327	0.4069
SKCM	472	469	99.36	24.48	23	0.4441	0.4263
STAD	442	436	98.64	30.03	31.5	0.4420	0.4303
TGCT	156	156	100.00	26.69	27	0.6425	0.6504
THCA	509	195	38.31	1.22	0	0.0280	0.0000
THYM	124	85	68.55	5.24	2	0.1238	0.0263
UCEC	546	481	88.10	15.92	5	0.2350	0.1174
UCS	55	55	100.00	38.16	37	0.5599	0.5941
UVM	80	79	98.75	6.54	6	0.1827	0.1628

Table 4.2: Count burden and genomic burden for tumors across the 33 sites in TCGA.

assessed for patterns in the genomic burden from the identified AI events. Testicular germ cell tumors (TCGT), ESCA, adrenocortical carcinoma (ACC) and OV had high overall genomic burdens, with a median greater than 60 percent of the genome (Figure 4.3, Table 4.2). UCS, LUSC and LUAD also exhibited high genomic burden, with medians over 50% of the genome (Figure 4.3, Table 4.2). Of note, as shown in Figure 4.4, the metastatic SKCM samples (n=368) exhibited significantly higher genomic burden of AI events than the primary tumor samples (n=104) within this dataset (Wilcoxon rank sum test, P value = 0.001).

Each event type showed varied patterns for count and genomic burdens across tumor types (Figure 4.6, Figure 4.5). Overall, among all event types, gains and loss events were frequent while cnLOH events were less common. In particular, OV tumors showed the highest median counts for loss and gain events (Figure 4.5). In terms of overall genomic burden, TGCT, UCS, ESCA and OV showed the highest median genomic burdens for gains, while ACC, KICH, OV, LUSC and UCS showed high genomic burdens for losses (Figure 4.6). The relative abundance of cnLOH was overall lesser than gains and losses, and spanned

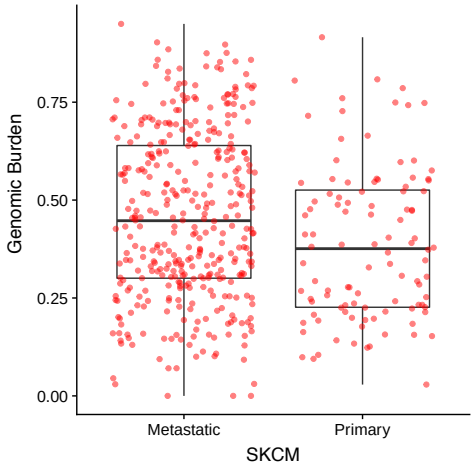


Figure 4.4: **Allelic imbalance burden in primary and metastatic melanoma samples.** Boxplots of the overall genomic burden for the primary SKCM tumors and metastatic SKCM tumors are shown. The individual sample-level burden is overlaid as red points on the boxplot.

smaller proportions of the genome (Figure 4.6); the highest rates of cnLOH genomic burdens were observed in TGCT, ESCA, LUSC, UCS and OV (Figure 4.6).

The different tumor sites also seems to show different patterns of enrichment of the three event types (Figure 4.7). While some cancers, such as KICH and ACC, showed pronounced and preferential enrichment of loss events, tumors such as KIRP and TGCT showed high gain burdens (Figure 4.7).

4.3.2 Landscape of chromosome arm-level copy number changes across tumor sites

Since our method is better designed to identify large chromosomal changes such as those that span an entire chromosome or chromosome arm, I examined in greater depth chromosome-arm level events across the 33 tumor types (Figure 4.8). Our method identified 121,645 events in 10,004 tumor samples, of which 32,925 were gains and 57,161 were loss events. Among these, the most common pan-cancer chromosome arm event occurred on 17p (Figure 4.8). Although 17p events were common among multiple tumor sites including ACC, KICH, COAD, LUAD, LUSC, PAAD, ESCA and BRCA, some tumor sites did not show an enrichment for 17p allelic imbalance, such as GBM, KIRC, THCA, UVM an PRAD (Figure

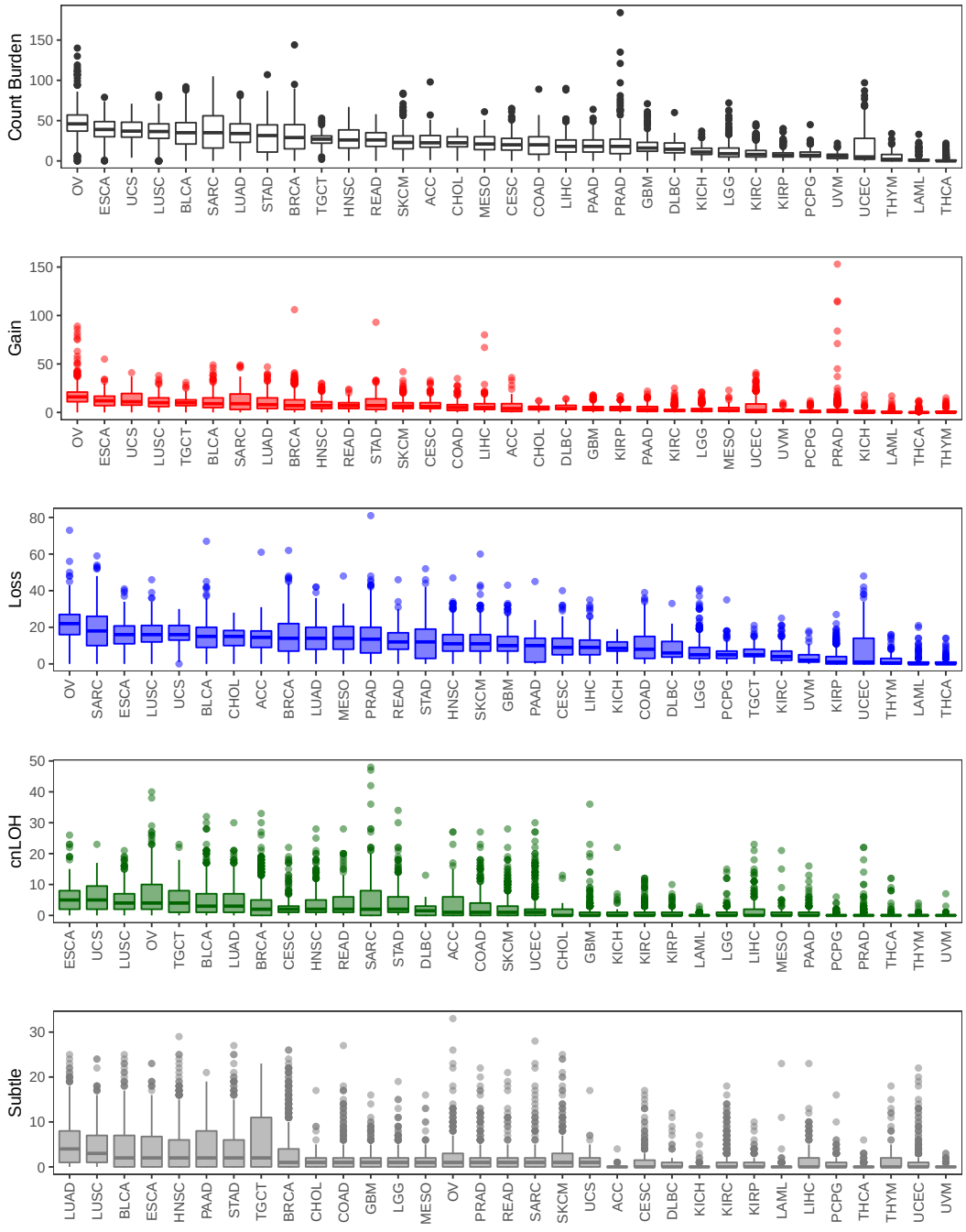


Figure 4.5: **Distribution of count burden for each event type across 33 tumor sites.** Boxplots of the overall count burden and burdens for each event type are shown across 33 tumor types in TCGA. The tumor sites are ordered by their median burden for each plot.

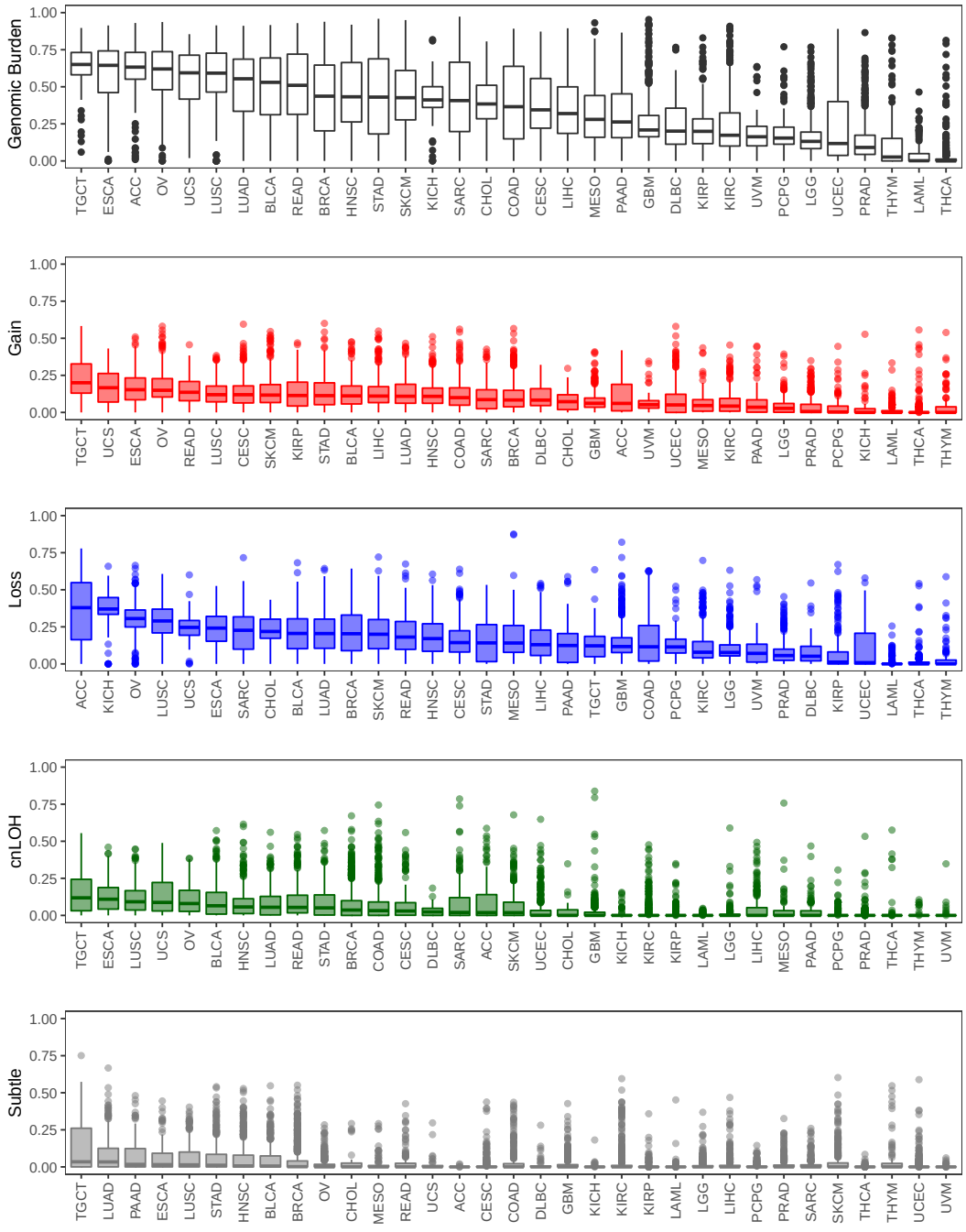


Figure 4.6: **Distribution of genomic burden for each event type across 33 tumor sites.** Boxplots of the overall genomic burden and burdens for each event type are shown across 33 tumor types in TCGA. The tumor sites are ordered by their median burden for each plot.

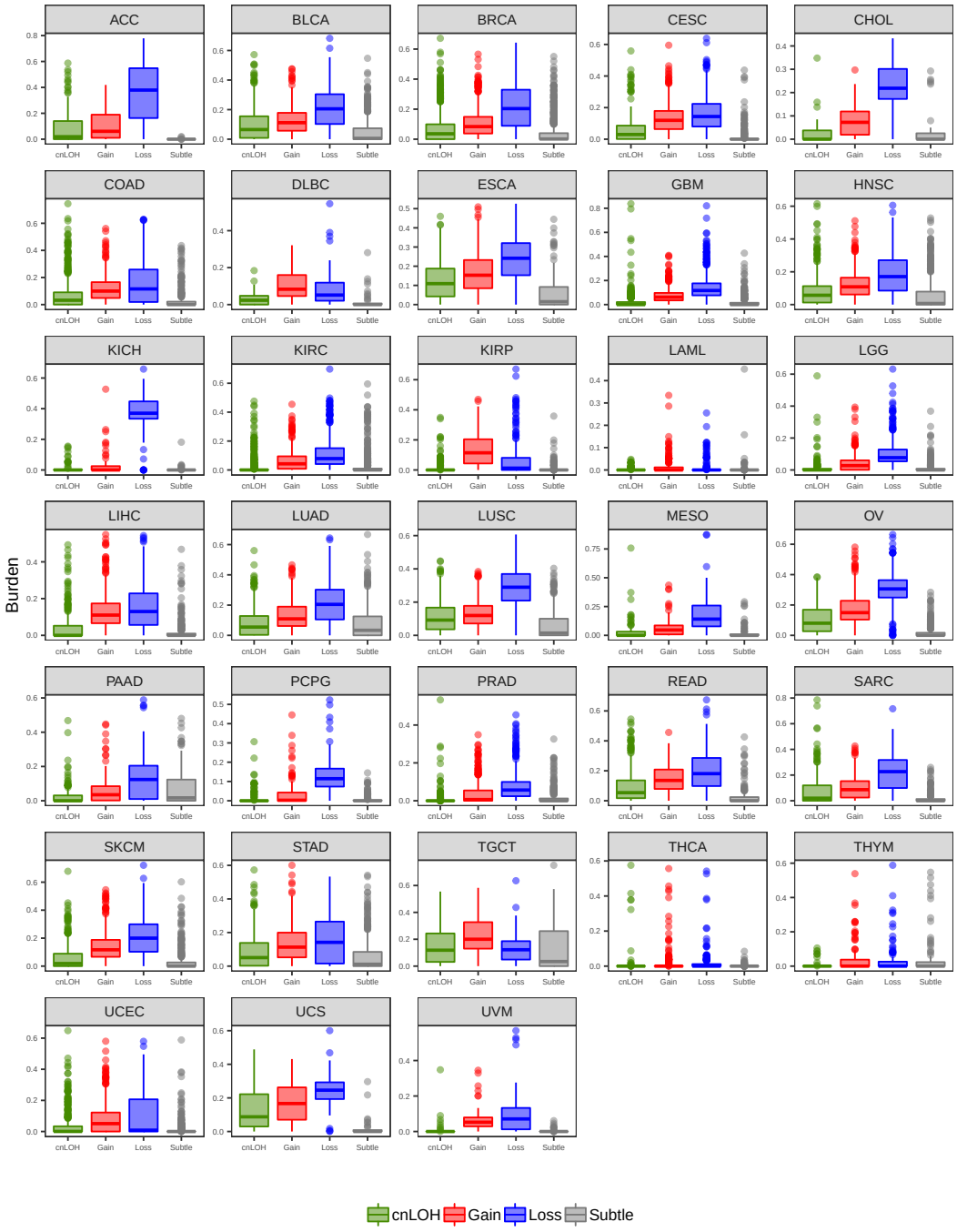


Figure 4.7: **Genomic burden for each event type compared within each of the 33 tumor sites.** Boxplots of the overall genomic burden for each event type in all 33 tumor sites in TCGA are shown.

4.8). Among cases that showed copy number changes on 17p, most comprised loss events; KIRP was the only tumor site that showed abundance of 17p gain events. Events on 8p and 3p were also prevalent across multiple tumor sites (Figure 4.8). While most cancer types showed a loss of 8p, LAML and UVM exhibited predominantly gain events; STAD, UCEC, and COAD showed mixed event types on 8p (Figure 4.8). Particularly in PRAD, 8p loss events were predominant with rest of genome being relatively stable, showing limited events in the rest of the genome such as 8q gain and 18q loss. As with 17p events, loss of 3p occurred across many tumor sites (Figure 4.8). Particularly in KIRC, 3p loss events seemed to the driver event, with rest of the genome showing very limited evidence for chromosomal instability (Figure 4.8). 3p loss events were also prevalent in UVM, LUAD, LUSC, HNSC, CHOL and CESC (Figure 4.8). 8q amplification was the most frequent pan-cancer gain event, showing high occurrence in multiple tumor sites including UVM, LAML, COAD, HNSC, STAD, UCEC, SKCM and LIHC (Figure 4.8). The second most prevalent amplification was identified on chromosome arm 7p, particularly in KIRP, DLBC, COAD, GBM, SKCM and STAD. Amplification of 1q was also observed across many tumor sites; LUAD, LIHC, CESC, UCEC, SKCM and THYM showed relatively high occurrences of 1q gain (Figure 4.8). Similarly, amplification of 20q seemed to be prevalent in gastrointestinal tumors COAD, READ and STAD. (Figure 4.8). In contrast, chromosome arms 2p and 2q were the least altered across tumor types; a predominant loss of 2p and 2q was observed only in ACC and KICH, both of which consisted of very few cases.

Tumor sites could broadly be classified into two categories, based on the distribution of chromosomal arm events across the genome. Some tumor sites such as UVM, THCA, KIRC, LGG, LAML and PRAD showed a significant enrichment of at most a single or very few chromosome arm events with the remaining parts of the genome being stable (Figure 4.8). For example, three tumor sites showed single arm events that dominated the allelic imbalance profile in those tumors such as 22q loss in THCA, 3p loss in KIRC and 8p loss in PRAD tumors. LGG tumors showed a significant enrichment of 1p and 19q loss events, consistent with the known phenotype of 1p/19q codeletion in LGGs. In contrast to LGGs, the other class of brain tumors, GBMs, exhibited a different allelic imbalance profile. In GBM tumors, the frequent events included the loss of chromosome 10 and gain

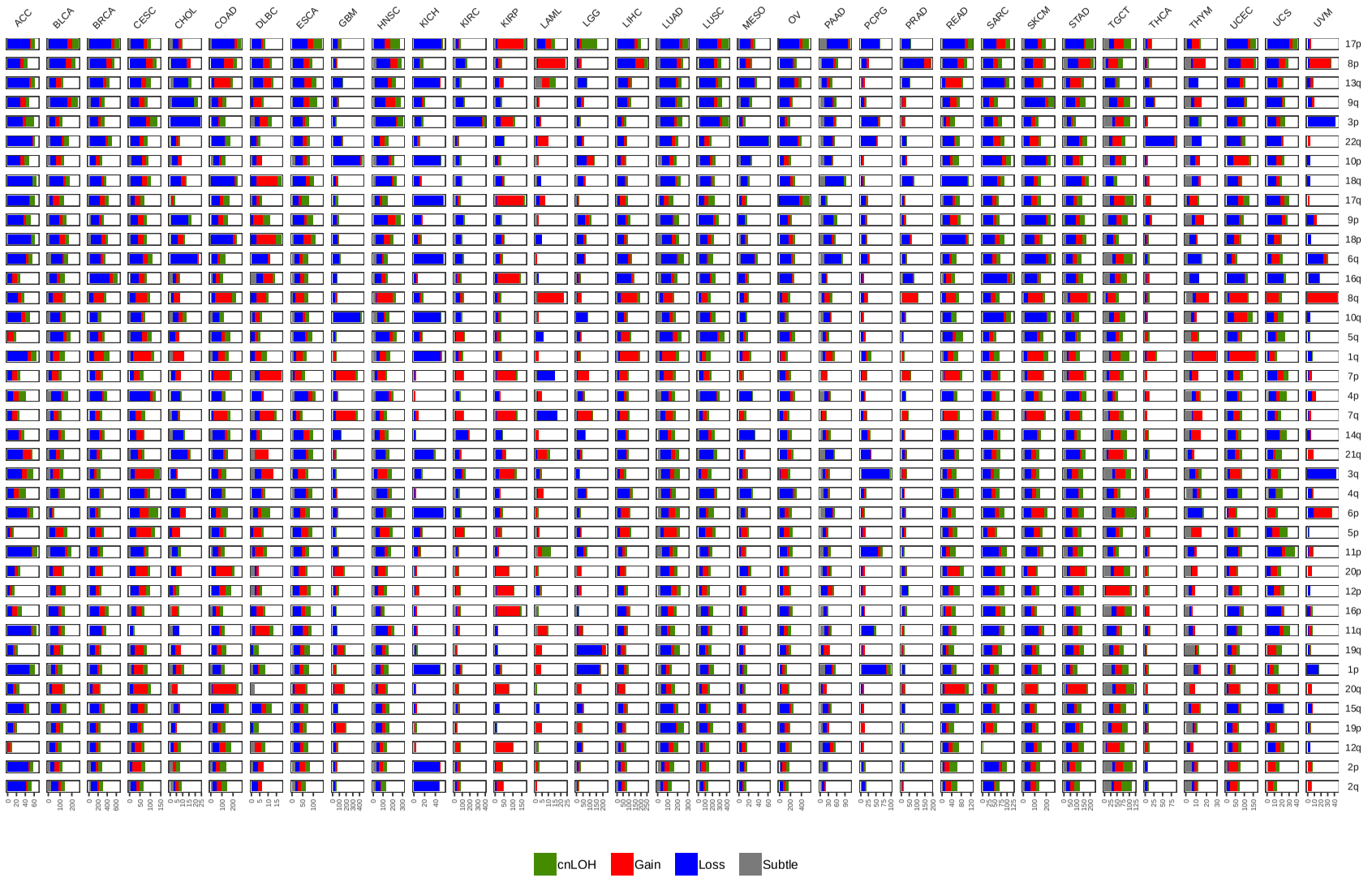


Figure 4.8: **Landscape of chromosome arm-level allelic imbalance events across the 33 tumor sites.** The distribution of different types of chromosome-arm level allelic imbalance events in each tumor site profiled in TCGA is shown. The scale of each tumor site is different, corresponding to number of samples that exhibited allelic imbalance events within that dataset. The chromosome arms are sorted by overall burdens across all tumors profiled.

of chromosome 7 (Figure 4.8). LAML tumors, although from a small dataset, showed high prevalence of chromosome 8 gains and to a lesser extent chromosome 7 loss events (Figure 4.8). UVM tumors also seemed to exhibit relatively few chromosomal arm events that spanned losses of chromosome 3 and chromosome arm 6q, as well as gains of chromosome 8 and chromosome arm 6p 4.8). PCPG also showed few chromosome arms under allelic imbalance, the most frequent being loss of 1p and 3q 4.8). Such tumor sites that were often accompanied by the enrichment of a single or few chromosomal aberrations, might suggest the role of these chromosome-arm events in driving tumorigenesis. In contrast to these tumor sites that showed lower overall allelic imbalance burdens, tumor sites such as LUAD, LUSC, BRCA, ESCA, ACC, TGCT, STAD, SKCM and SARC showed genome-wide allelic imbalance patterns with multiple chromosomal aberrations 4.8).

Tumor sites could also be classified based on the enrichment of a specific event type among the allelic imbalance events detected. For example, KIRP tumors seemed to be primarily driven by gain events across multiple chromosomal arms. In contrast, tumor sites such as KICH, PCPG, STAD and PAAD showed an enrichment of loss events across their genomes (Figure 4.8, Figure 4.7). These results aid in understanding the site-specific origin of different chromosomal changes that might be drive the development of different tumor types.

4.3.3 Copy-neutral loss of heterozygosity patterns across tumor sites

Our method relies on identifying deviations in the expected 1:1 allelic ratios at germline heterozygous loci, and therefore can accurately detect regions of cnLOH that result in allelic ratios of 2:0 or 0:2 depending on the haplotype under cnLOH. Given that standard methods that rely on identifying copy number changes from LRR intensities, this particular class of chromosomal aberrations will be missed in such procedures. Therefore, I sought to characterize the previously unknown landscape of cnLOH events across the 33 tumor sites in TCGA. To date, this has not been comprehensively characterized in the TCGA.

Our method identified 20,454 cnLOH arm-level events across 5,222 cases in TCGA. Among the different tumor sites profiles, TGCT and ESCA showed the highest rates of genomic cnLOH burden as well as chromosome arm-level cnLOH events (Figure 4.6, Figure

4.8). Figure 4.9 shows the distribution of chromosome arm-level cnLOH events across the genome for each tumor site. The rates and patterns of cnLOH varied among the tumor types and a few visually pronounced chromosome-arm level cnLOH events were identified. Chromosome 17 showed the highest rates of cnLOH across tumor sites (Figure 4.9). Chromosome arm 3p, 6p and 2p also showed high rates of cnLOH events across tumor sites (Figure 4.9).

I interrogated, in greater depth, the chromosome arm-level cnLOH that showed enrichment within specific tumor sites. For example, LGG tumors showed prevalence of cnLOH events on 17p (Figure 4.9), an observation that would be completely missed in current copy number detection approaches that typically only use LRR intensities. In addition, cases that exhibited 17p cnLOH in LGGs (n=114) were found to be mutually exclusive of cases that exhibited the characteristic LGG event, 1p/19q co-deletion (n=160). 17p cnLOH were also identified as the predominant cnLOH event in ESCA, OV, LUSC and HNSC tumors (Figure 4.9). These tumor sites also showed evidence for cnLOH events on the q arm of chromosome 17, albeit to a lesser extent than 17p. 6p also showed high rates of cnLOH, particularly in CESC tumors (Figure 4.9). In UCS tumors, despite high rates of cnLOH across the genome, 11p cnLOH occurred more frequently than other cnLOH events (Figure 4.9). I also observed tumor sites with multiple cnLOH events across the genomes, i.e., without having any visually observable specificity for particular chromosome arm events. For example, TGCT showed high rates of cnLOH arm-level events across the genome, with events on chromosomes 6 and 2 being slightly more common than other parts of the genome (Figure 4.9). Similarly, READ, LUAD, SKCM and STAD also showed genome-wide cnLOH events. These results suggest the important role of investigating cnLOH events that might be crucial drivers of oncogenesis across tumor sites.

4.3.4 Copy neutral loss of heterozygosity events and survival trends

Next, I investigated the correlations between the presence of specific chromosome-arm cnLOH events with survival within multiple tumor sites. I first looked for patterns in the most common cnLOH in our dataset, 17p, across multiple tumor sites. I did not identify a significant association between 17p cnLOH events and survival in LGG (P value = 0.41),

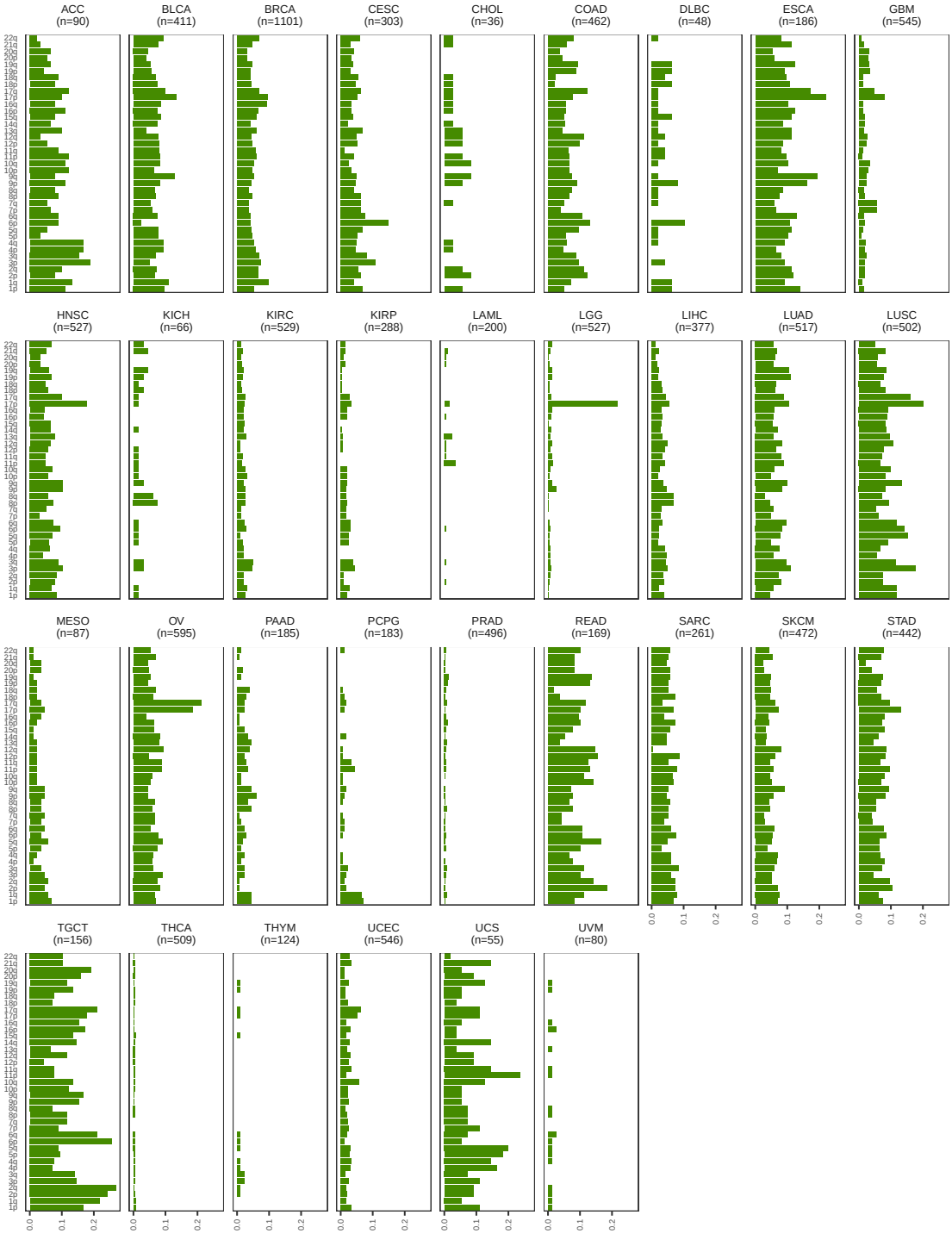


Figure 4.9: **Distribution of copy-neutral loss of heterozygosity chromosome-arm events across 33 tumor sites.** For each tumor site, the proportion of cases exhibiting chromosome-arm level cnLOH events are shown as bar plots. The total number of samples profiled for each tumor site is listed above each plot.

HNSC (P value = 0.18), ESCA (P value = 0.43) or LUSC (P value = 0.77). LUSC patients with 3p cnLOH also did not show significant association with survival (P value = 0.33). The presence of 11p cnLOH in UCS patients showed a mild evidence of association with survival (P value = 0.07), with tumors that exhibited this cnLOH event showing better survival. Given the mutual exclusivity of 17p cnLOH and 1p/19q co-deletion in LGGs, I also investigated for potential differences in survival trends among the two subtypes based on their chromosomal aberration status. I found marginal evidence for a poor overall survival in cases that exhibited 17p cnLOH compared to those with the 1p/19q codeletion (P value = 0.06). Overall, I was unable to find significant associations between the presence of cnLOH events and survival in this cohort; however a few events exhibited trends that would support the association, pending future studies. Further that these cnLOH may be associated with additional gain or loss events, or point mutations, may allude to the lack of strong trends in survival estimation for cnLOH events.

4.3.5 Putative problematic copy number profiles

Given the differences in the detection approaches, i.e. to identify allelic imbalance events using BAF deviation based measurements, in comparison to the commonly used LRR measure for SCNA detection, I was interested in testing for consistency of our event calls with previously identified chromosome-arm level SCNAs, using the approach described in Section 4.2.4. I used 10,680 tumor samples for this comparison, among which 8,893 samples showed overall consistency (positive correlation) between the two sets. A total of 1,787 cases showed negative correlation between our calls with those previously identified. A closer inspection of these cases revealed a strong negative correlation (Pearson correlation, -0.7417) between the overall genomic burden, derived from regions of allelic imbalance, and the concordance of the two call sets, i.e., samples that exhibited high overall allelic imbalance burden tended to show patterns of discordance between the two calls sets, consistent with the hypothesis of incorrectly estimating the true normal region in an aberrant tumor genome. This trend was consistent across all tumor sites (Figure 4.10).

A subset of 1653 tumor samples that exhibited a negative correlation between the two SCNA cell sets as well as a high overall AI burden ($\geq$ 50% of the genome) was identi-

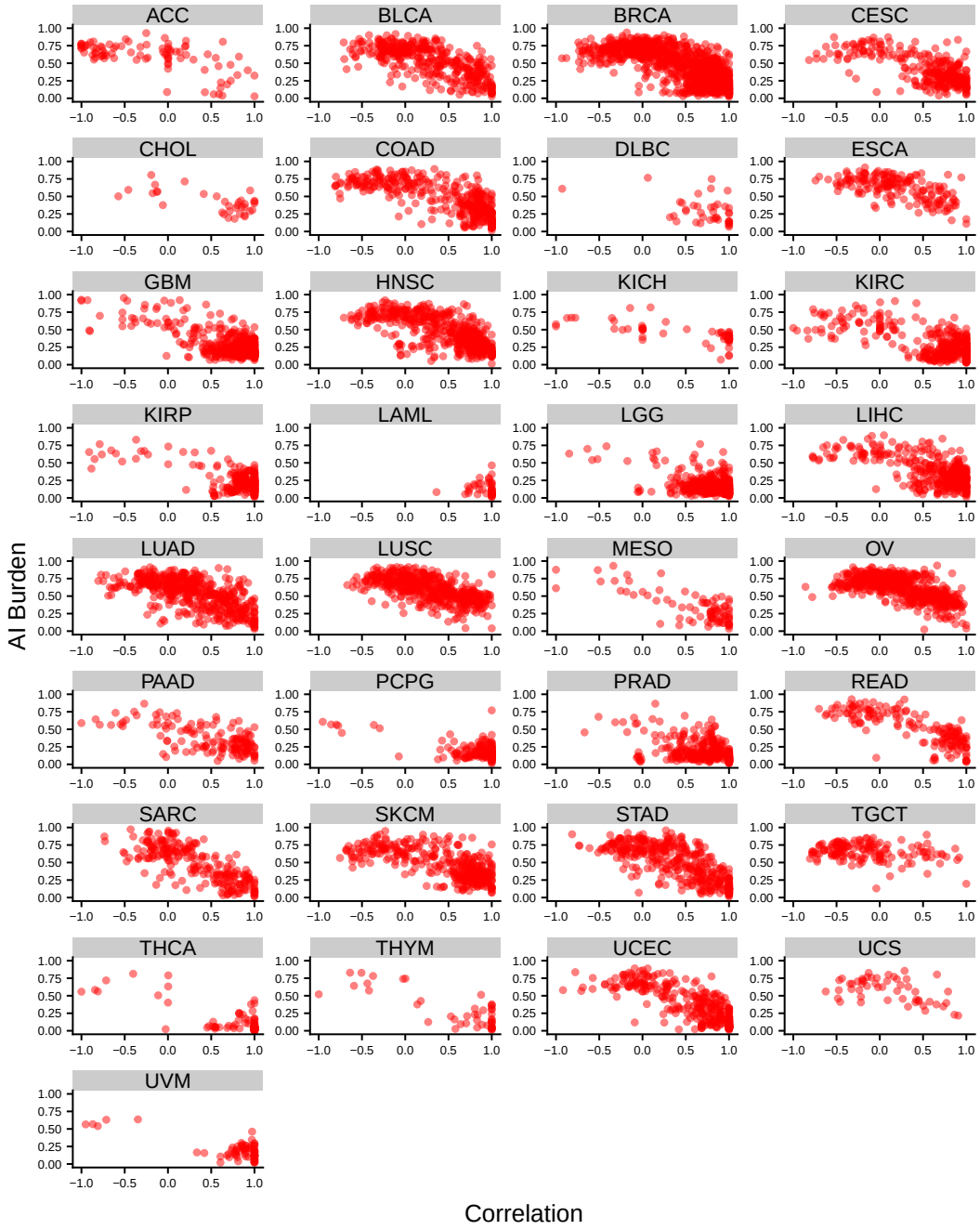


Figure 4.10: **Relationship between overall allelic imbalance burden and correlation between the SCNA calls.** For each tumor site, a scatter plot of the overall allelic imbalance (AI) genomic burden (y axis) and the correlation value (x axis), signifying concordance of calls, for all samples profiled, is shown. Overall, samples that showed poor correlation exhibited a higher genomic burden of allelic imbalance.

fied. The rates of these putative problematic samples varied across tumor sites (Figure 4.11). Table 4.3 summarizes the identified putative problematic samples by tumor site. The highest proportion of negatively correlated samples was observed in TGCT (64.7%) and ACC (54.4%) consisting of a total of 150 and 90 profiled cases respectively. UCS also showed a high rate of discordant calls (37%) from a total of 50 cases profiled. All three datasets were relatively small, therefore sampling variation might account for this higher percentage of discordant samples. However, relatively high rates of discordance, greater than 25% of the samples, were also observed in larger studies such as ESCA (30.4%), READ (27.8%), OV (27.6%) and BLCA (26.2%) (Figure 4.11, Table 4.3). The lung tumors (LUAD and LUSC) also showed rates of 24% discordance (Figure 4.11, Table 4.3). In contrast, LAML dataset (n=191) contained no discordant samples. (Figure 4.11, Table 4.3). Other tumor sites that exhibited very low percent of discordant calls were THCA (1.2%), LGG (1.2%), PRAD (1.2%) and DLBC (2.1%) (Figure 4.11, Table 4.3). The low discordant cases were often observed in tumor sites that showed lower allelic imbalance burden, further supporting the notion of incorrectly estimating the normal region in aberrant tumor genomes that may result in potentially erroneous calls.

I next attempted to automatically adjust the copy number calls in these cases using the approach described in Section 4.2.5. Among the identified problematic cases, the adjusted SCNAs in 1224 samples resulted in a positive correlation. However, the rates of performance of our automated adjustment protocol also varied by tumor sites, as shown in Figure 4.12 and summarized in Table 4.3. For example, in tumor sites such as LGG, THCA, PCPG and THYM, our automated approach successfully adjusted all discordant cases to achieve a positive correlation of calls between all samples (Table 4.3). ACC and MESO also achieved high rates of adjustment in over 90% of the potentially problematic cases (Table 4.3). Across tumor sites, after performing the automated adjustment, the percent of negatively correlated samples that remained was drastically lower. For example, the percent of negatively correlated samples reduced from 54.4% to 4.4% in ACC, 23.1% to 6.6% in SARC, 21.2% to 5.1% in SKCM, 20.9% to 4.1% in COAD. However, some tumor sites continued to exhibit relatively high numbers of discordant cases, even after adjustment. For example, TGCT showed negative correlation in 14% cases after adjustment; similarly ESCA,

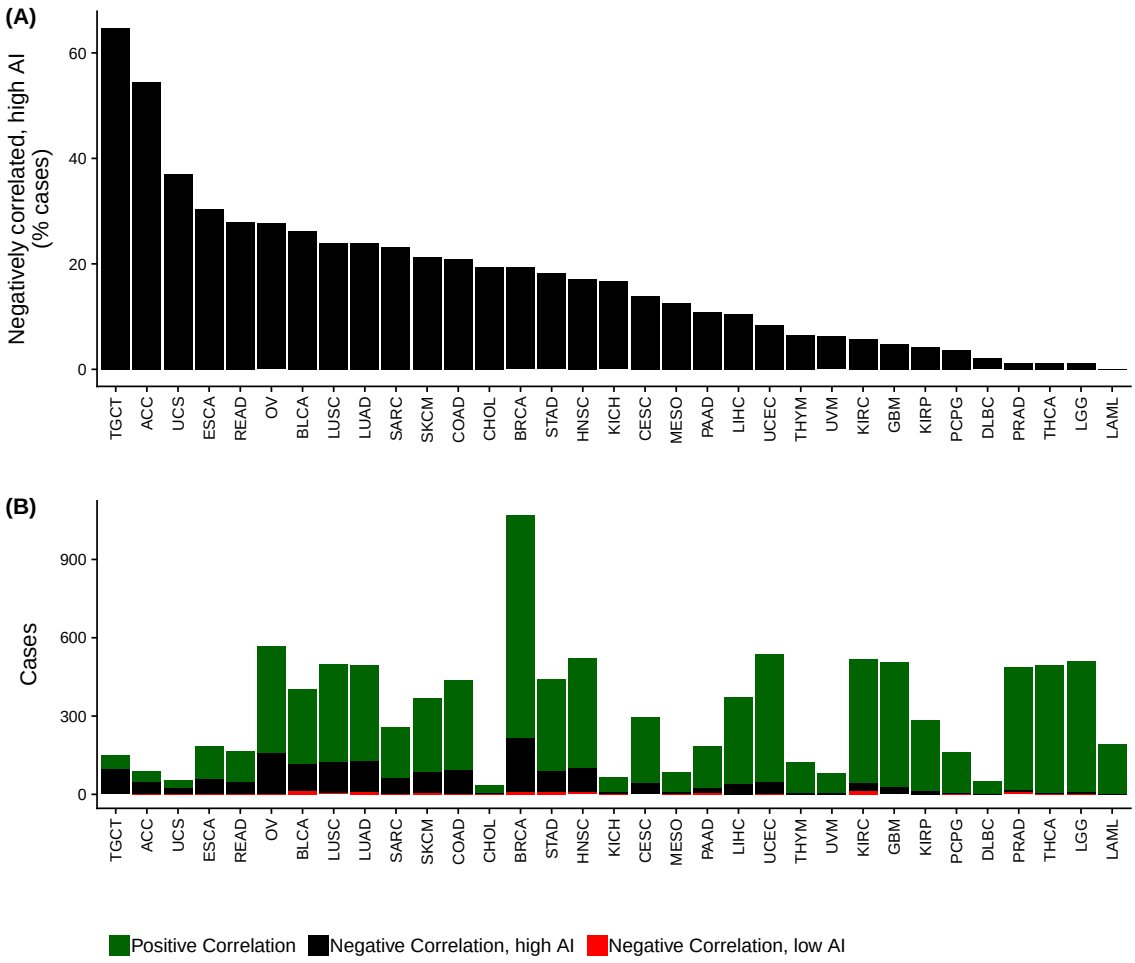


Figure 4.11: **Distribution of concordant and discordant samples across TCGA.** (A) For each tumor site, a bar plot of the percent of putative problematic calls is shown. The tumor sites are sorted by the percent of negatively correlated, high AI samples identified. (B) For each tumor site, a stacked bar plot showing the number of cases that were positively correlated and negatively correlated (high AI for those with $\geq 50\%$ AI burden, low AI for those with $<50\%$ AI burden) are shown.

TCGA Study	Total cases	Cases with Correlation	Positive Correlation	Negative Correlation (Before)	Negative Correlation, high AI (Before)	Negative Correlation, high AI (Adjusted)	Negative Correlation, high AI (Before) %	Negative Correlation, high AI (After) %	Negative Correlation, high AI (Adjusted) %
ACC	90	88	38	50	49	45	54.44	4.44	91.84
BLCA	404	378	260	118	106	74	26.24	7.92	69.81
BRCA	1071	1035	818	217	207	148	19.33	5.51	71.50
CESC	294	286	242	44	41	30	13.95	3.74	73.17
CHOL	36	34	26	8	7	7	19.44	0.00	100.00
COAD	439	389	293	96	92	74	20.96	4.10	80.43
DLBC	48	39	38	1	1	1	2.08	0.00	100.00
ESCA	184	177	119	58	56	38	30.43	9.78	67.86
GBM	505	497	470	27	24	21	4.75	0.59	87.50
HNSC	521	499	399	100	89	64	17.08	4.80	71.91
KICH	66	62	50	12	11	11	16.67	0.00	100.00
KIRC	516	481	438	43	30	22	5.81	1.55	73.33
KIRP	285	269	256	13	12	11	4.21	0.35	91.67
LAML	191	56	56	0	0	0	0.00	0.00	-
LGG	511	469	459	10	6	6	1.17	0.00	100.00
LIHC	370	356	316	40	39	31	10.54	2.16	79.49
LUAD	496	482	353	129	119	85	23.99	6.85	71.43
LUSC	500	492	365	127	120	81	24.00	7.80	67.50
MESO	87	83	72	11	11	10	12.64	1.15	90.91
OV	568	565	406	159	157	106	27.64	8.98	67.52
PAAD	183	144	119	25	20	16	10.93	2.19	80.00
PCPG	162	153	145	8	6	6	3.70	0.00	100.00
PRAD	488	331	314	17	6	4	1.23	0.41	66.67
READ	165	161	113	48	46	33	27.88	7.88	71.74
SARC	255	235	172	63	59	42	23.14	6.67	71.19
SKCM	367	355	271	84	78	59	21.25	5.18	75.64
STAD	440	396	308	88	80	55	18.18	5.68	68.75
TGCT	150	148	48	100	97	76	64.67	14.00	78.35
THCA	496	143	136	7	6	6	1.21	0.00	100.00
THYM	123	46	38	8	8	8	6.50	0.00	100.00
UCEC	535	389	342	47	45	34	8.41	2.06	75.56
UCS	54	52	28	24	20	15	37.04	9.26	75.00
UVM	80	74	69	5	5	5	6.25	0.00	100.00

Table 4.3: Summary statistics for the automated procedure for the identification and adjustment of putative problematic samples in TCGA

UCS and OV also showed rates of 9.8%, 9.2% and 8.9% respectively, after adjustment (Table 4.3). Nonetheless, our approach was able to significantly reduce the rates of discordance, compared to the trends observed before applying the adjustment protocol across all tumor sites.

4.4 Discussion

Acquired chromosomal alterations such as deletions, duplications and copy-neutral loss-of-heterozygosity serve as hallmarks of tumorigenesis. Large alterations span multiple heterozygous markers and thus result in deviations from the expected one-to-one allelic ratio, thereby leading to allelic imbalance (AI). The SCNA pipeline of the TCGA consortium reports genomic regions and their segment mean copy number estimates from SNP genotyping arrays. Calibration of this process relies on the accurate identification of non-aberrant regions of the genome to establish a baseline signal intensity representative of neutral copy number. However, tumor samples exhibiting high levels of genomic instability pose a challenge for such analyses. The resulting poor calibration may lead to erroneous SCNA calls, with obvious examples of highly aberrant chromosomes (by visual inspection) being missed by the automated calling procedure and instead SCNAs called across normal-appearing chromosome arms. Here we attempt to address this problem by integrating an orthogonal data source to triangulate regions more likely to be copy-neutral and then apply a systematic adjustment procedure to rescue such cases.

4.4.1 Significance of findings

A sensitive allelic imbalance detector was used to reveal allele frequency patterns consistent with acquired SCNAs as well as the landscape of previously unknown regions of cnLOH across all 33 tumor sites in TCGA. The results presented here supplement the TCGA repository with additional allelic imbalance derived chromosomal aberrations and suggest tissue-site specific patterns of AI signatures.

A recent study summarized pan-cancer chromosomal aberrations and correlated chromosomal arm aneuploidy to somatic point mutations and expression of immune sig-

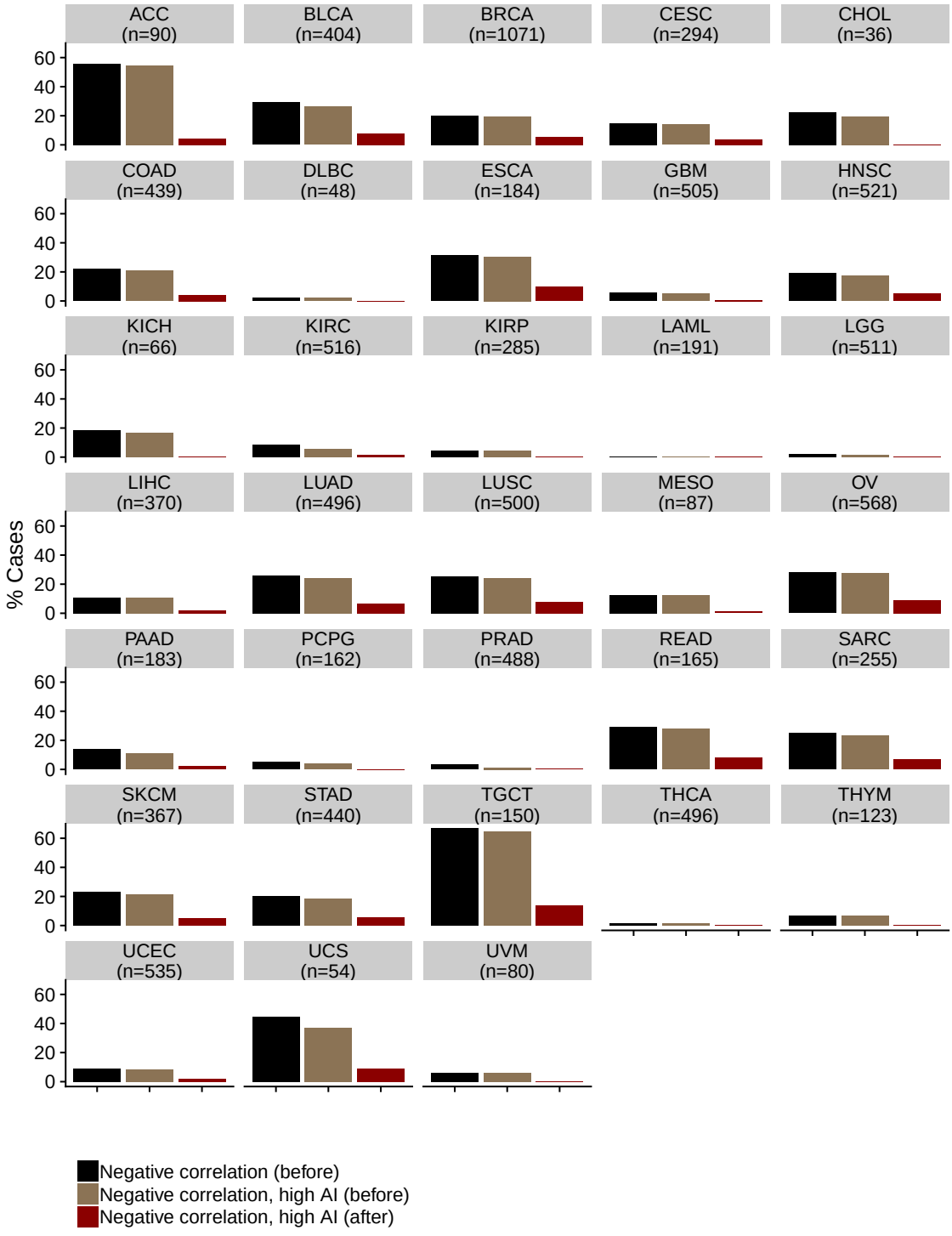


Figure 4.12: **Trends of negatively correlated samples before and after applying an automated adjustment protocol, across all tumor sites in TCGA.** For each tumor site, barplots of the percentage of samples that were negatively correlated before and after the adjustment procedure are shown.

naling genes [89]. Chromosomal copy number events identified in our present study corroborated with their findings, such as the high prevalence of loss events on 17p and 8p, as well as gains of 8q across tumor sites. Similarly, both studies identified that chromosome 2 was the least aberrant across tumor sites. Our allelic imbalance burden derived estimates were also similar to the aneuploidy scores they presented across TCGA types [89]; however I report higher burdens in ESCA and OV. Another pan-cancer atlas of the TCGA tumor sites identified specific clusters of tumors based on the extent of aneuploidy and the type of events [90]. I identified similar patterns from our allelic imbalance derived chromosome arm-level aberrations. Both studies identified a subset of low burden sites that included PRAD, THYM, LAML and THCA. Our results also identified the enrichment of 13q gain and chromosome 18 loss in gastrointestinal tumors (COAD, READ, and STAD); in addition to these events, I also observed high rates of chromosome 20 gains in these tumors. Similarly, the previously identified enrichment of chromosome 7 gains and chromosome 10 losses in GBM tumors was also observed in our dataset. These evidences for shared chromosomal SCNA patterns between studies, validates our approach of identifying SCNAs from allelic imbalance patterns.

In comparison to both these previous studies [89, 90], we also provide an additional complementary landscape of pan-cancer cnLOH events that opens a window of opportunities to investigate the importance of these subtle cnLOH events in tumorigenesis. A striking result observed in this investigation was the recurrent cnLOH of 17p across multiple tumor sites. This suggest that loss as well as cnLOH of 17p might be wide-spread and recurrent across tumor sites. Based on recent evidences [92], it is possible that these pan-cancer 17p loss and cnLOH affect a combination of genes, extending beyond the effects on the *TP53* tumor suppressor gene.

Of particular interest were tumor samples that displayed conflicting SCNA events as compared to previous reports. A closer examination of these cases, especially, when overlaid with BAF distributions across the genome, suggested an incorrectly estimated normal-region within these cases. These striking observations as well as the wide user base of the TCGA repository motivated me to develop an automated procedure to identify and adjust these putative problematic cases. By developing an automated identification and

adjustment procedure, I not only provide a list of these putative problematic call sets, but also renormalize the calls for improved SCNA detection.

This study also showcased high rates of previously unknown chromosomal aberrations across multiple tumor sites. For example, in KIRP tumors that were predominantly driven by gain events, our findings aligned with previously identified high frequency gains on chromosome 7 and 17, but I also identified a higher proportion of chromosome 16 gains than previously reported [93]. Similarly in UCS tumors, I identified recurrent loss and cnLOH events of 17p as well as novel recurrent cnLOH of 11p. Given the prevalence of *TP53* mutations in UCS [94], loss and cnLOH of 17p might confer somatic two-hit mechanisms of mutagenesis in UCS tumors. In HNSC tumors, previously, highest rates of chromosomal changes were reported on 8p and 3p [95]. Our method not only identified these events, but also detected high rates of 17p loss and cnLOH events in these tumors. In this way, findings from our study supplement the current database of chromosome-arm SCNAs across multiple tumor sites, through the detection of additional chromosomal aberrations.

It is noteworthy that cnLOH events we identified to be enriched in our present investigation across multiple tumor sites have been shown to be prognostically relevant in independent studies. For example, I observed recurrent 17p cnLOH in LGGs in addition to the well known 1p/19q codeletion; cnLOH of 17p, as well as its mutual exclusivity with 1p/19q codeletion, has been previously shown to be a potential marker in independent cohorts of gliomas [96, 97, 98]. Similarly, 6p cnLOH were enriched in CESC tumors; this corroborates previous studies that have identified LOH on 6p21.2 that was suggestive of recurrence of cervical carcinoma after radiotherapy [99]. Therefore, cnLOH events presented in this study may have the potential to serve as prognostic markers for the detection or prediction of recurrence across tumor sites.

4.4.2 Limitations

Although our methods to identify as well as adjust previous SCNAs provide as a useful resource for pan-cancer investigations of chromosomal aberrations, it doesn't come without limitations.

First, our approach relies on deviations in allelic ratios at germline heterozygous

sites to identify regions of allelic imbalance. As a result of it, our method is currently incapable of identifying balanced duplications since these events do not perturb the allelic ratios.

Second, in line with the requirement of altered allelic ratios, our approach is better at detecting loss or cnLOH events that results in a more severe change in allelic ratios (i.e. 1:0 or 2:0), in comparison to gains (e.g., one copy gains that results in a ratio of 1:2).

Third, although the statistical model is accurate and sensitive in identifying regions of allelic imbalance, I currently implement a naive thresholding based classification to annotate events as loss, gain and cnLOH. However, this results in subtle events that show a signal for allelic imbalance, but are not classifiable using our current LRR threshold based approach. Nonetheless, these additional events supplement the current repository of copy number alterations.

Lastly, our methodology to systematically adjust potential differences in the identified copy number alterations might suffer from overcorrection since this procedure is applied across the entire genome, or an undercorrection in cancer genomes with complex events (mixed gains and losses) in the expected normal region, as inferred by hapLOH, that will not alter the estimated normal CN and therefore fails to adjust those event calls.

However, our overall goal, through this study, is to highlight the importance of integrating multiple data types (B-allele frequency and Log R ratio) for more robust automated inference procedures, which can run amok with single sources of data. These results have the potential to support exploration of more rigorous methods across all the cancer types in TCGA in order to improve downstream analyses and empirical discoveries, including clinical evaluations and copy number-derived signatures. The landscape of cnLOH events presented in this study will provide opportunities for future investigations of chromosomal instability in tumorigenesis.

CHAPTER 5
CONCLUSIONS AND FUTURE DIRECTIONS

The evolutionary trajectory of cancer development from a normal cell or tissue involves a natural process of initiation and different stages of progression, that may comprise one or more non-invasive premalignant or intermediate lesions. These stages prior to the development of overt malignancies provide a window of opportunities for studying and designing tumor intervention and prevention strategies.

A series of progressive pathological changes involving certain precursor lesions, with corresponding genetic and epigenetic aberrations have been shown to be critical in tumor development across organ sites, an example of which is lung [100]. This multi-step tumorigenesis of precursor lesions is often accompanied by an accumulation of molecular changes, and therefore genomic instability, that might be crucial for the development of malignant phenotypes. A deeper understanding of the evolutionary dynamics of the initiation of premalignant lesions suggests that molecular mechanisms leading to tumorigenesis begin even before the growth of these clinically detectable premalignant lesions [10]. Therefore, molecular changes might occur in normal-appearing tissues, prior to any morphological changes associated with premalignancy are identifiable. A cancerized field is described as the preconditioning of an area of the epithelium to a cancer-primed cell population, with no apparent morphological changes [10]. These theories allude to the important role of *cancerized fields* in initiation of premalignant and tumor phenotypes.

In lung cancers, a major challenge for early detection and intervention is the lack of molecular characterization of the earliest stages of development, and thereby the paucity of markers indicative of early stage disease. Studying stages preceding invasive lung tumors such as preneoplastic and normal tissues may help address this gap in knowledge, however investigations of these tissues are often compounded by additional challenges. First, there is often very limited tissue available for molecular characterization of premalignant and normal tissues, perhaps due to the relatively small size of these premalignant lesions. Second,

genomic changes are present at low cellular fractions in preneoplastic and normal tissues, making the mutant cell fraction low, i.e., few cells exhibit the change of interest. Third, premalignant tissues may include a high contamination from DNA from normal cells. These challenges not only demand more precise profiling technologies, but also require computational innovations to better detect and characterize molecular changes in these tissues.

Therefore, we attempted to comprehensively investigate the normal-appearing respiratory epithelia and premalignant lesions of early-stage non-small cell lung cancers, by performing multi-platform integrative and innovative computational analyses to overcome the challenges faced with profiling these premalignant and malignant tissues. In this chapter, I summarize our conclusions and outline the contributions of this thesis. I discuss possible future research directions and conclude by speculating on the impact of the bioinformatics approaches that I developed here to other cancer genomics studies, particularly for investigations of premalignant fields of cancerization.

5.1 Contributions of this thesis

In this thesis, I describe our efforts to understand the molecular changes of premalignant lesions and normal tissues using early-stage non-small lung cancers as a model. Given the high prevalence and mortality of lung cancers worldwide, there is an urgent need to devise better detection and treatment strategies to intercept or prevent the development of overt lung malignancies at the earliest stages. Here, we characterize some of the earliest stages preceding non-small cell lung cancers while attempting to capture the spatial and temporal changes that might provide a better trajectory of normal tissues as well as premalignant lesions transitioning to malignancy. I provide studies that address complementary and independent questions, towards the long term goal of better characterizing these stages preceding tumor development.

In Chapter 2, I describe our cross-platform integrative approach to comprehensively characterize atypical adenomatous hyperplasia, the only known premalignant lesion to lung adenocarcinomas. AAH were interrogated along with matched normal lung parenchyma and tumor specimens. Such a study design allowed us to make comparisons between AAH

and LUAD within and between patients, as well as better delineate the initiation and progression phases of AAH and LUAD development. I proposed different mechanisms in the pathogenesis of AAH, possibly driven from tobacco exposure; the study underscores some of earliest mutation processes as well as gene expression changes, such as those leading to altered immune signaling, that may play a critical role in the pathogenesis of AAH and their progression to lung adenocarcinomas.

In Chapter 3, I report the mutational landscape of airway field of cancerization from an independent and large collection of multi-region samples consisting of tumor-adjacent and distant airway epithelia, nasal epithelia, uninvolved normal lung parenchyma, blood and matched NSCLC specimens. I developed a quantitative measure (FCAUC) to better identify and quantify the extent of field effects in these early-stage NSCLC patients. These morphologically normal-appearing airway epithelia not only exhibited mutations in known lung cancer drivers, but their overall mutational burden and variant allele frequencies suggested a spatial field of cancerization effect, suggestive of clonal selection and expansion in the transition to malignant phenotypes. Lastly, the identified multi-hit somatic progression models suggest a temporal ordering of events leading to mutation accumulation and transition of normal-appearing airway lesions to NSCLC.

In Chapter 4, I expand our investigations of genomic instability to other tumor types in The Cancer Genome Atlas Project. I provide a pan-cancer landscape of chromosomal allelic imbalance events, comprising copy number changes such as gains and losses, as well as the previously undocumented landscape of copy-neutral loss of heterozygosity in this data set of tumor genomes. I also propose an automated approach to identify and adjust potentially problematic chromosomal copy number events previously reported in the TCGA consortium. This work not only provides the cancer research community with a complementary source of large chromosomal copy number alterations, but also provide a new and important class of mutational events, cnLOH, that may play a crucial role in the classical two-hit hypothesis for tumorigenesis, that may further involve an interplay with germline variations.

This thesis also outlines several bioinformatic approaches to aid data analysis in low mutant samples such as premalignant and airway field samples. For example, I developed a

filter to remove potential false positive point mutation calls within homopolymer regions in the genome; this tool is now part of our standard pipeline for quality control and processing of point mutation calls in Ion Torrent technology. In addition, I utilized a combination of mutational callers to increase our confidence in the identified mutation calls, particularly in these pre-tumor stages. In Chapter 2, I developed a classifier to better visualize patterns of gene expression changes between the normal lung and AAH as well as between AAH and LUAD; this method serves as a novel extension to standard gene expression analysis by aligning these changes temporally along the different stages of tumor development (e.g., normal, AAH, LUAD). Similarly, in Chapter 3, I developed statistical approaches to test for spatial field effects using the mutation frequencies and proposed a new measure to quantify the extent of field cancerization using the proportion of shared aberrations with matched NSCLCs. These methods can further aid discoveries and interpretation of field cancerization effects across other tumor types. In Chapter 4, using the TCGA cohort as an example, I also developed methods to compare and contrast large chromosomal changes identified using different software and platforms. Overall, the methods developed in this dissertation highlight the need for integrative approaches to investigate these aberrations not only in cancer genomes, but also in studies of precancerous lesions and somatic mosaicism in normal tissues and cancerized fields.

In summary, the findings in this thesis highlight the utility of normal and premalignant lesions as way to improve our knowledge of tumor development, using non-small cell lung cancer as an example. This research may contribute towards developing strategies for improved screening and detection, such as profiling airway epithelium, as well as towards early treatment and possible prevention regimens of premalignant disease, particularly in smokers or other individuals at elevated risk for lung cancer.

5.2 Future directions

The bioinformatic methods developed as a part of this thesis provide a proof-of-concept for the utility and importance of applying quantitative metrics to understand pre-tumor and tumor biology. However, there is potential to extend these approaches to incorporate more

sophisticated statistical models to better characterize these molecular changes. For example, the test for spatial field effects uses a linear regression model. However, this is limited by fluctuations based on the distance measures supplied to the model. A framework that treats these field tissues as an ordinal variable, or uses precise distances available through histopathological aids, may improve its performance or aid interpretation and characterization of these spatial effects. Similarly, the automated procedure to identify and adjust potentially problematic chromosomal alterations in the TCGA repository, can be improved by applying separate adjustments for gain and loss events.

Our efforts to characterize AAH lesions in early stage LUADs of 23 individuals, lends to the need for future longitudinal studies with a larger number of AAH samples and in larger cohorts, wherein mutations, gene expression and markers of the immune response can be aligned with time and space. Our study was restricted to an East-Asian cohort, further suggesting that investigations in different cohorts, perhaps comprising individuals from different populations, may help identify population-specific molecular patterns and trends in AAHs. Additionally, tracking AAHs in patients without overt lung malignancy may also help distinguish drivers of AAH progression from benign lesions. Such a study design may further delineate the effects of field cancerization into those involved in progression to premalignant and malignant phenotypes. Another natural extension to the current study involves the investigation of other minimally invasive lesions such as AIS and MIA implicated in LUAD pathogenesis along with matched AAH lesions, to better contextualize the role of AAHs in development of these less invasive lesions. Recent and ongoing efforts have begun to distinguish potentially distinct profiles in the multi-step progression of AAH to AIS and subsequently to LUAD [101, 8, 40].

Immune-based therapy has come to the forefront of targeted therapeutic strategies for various malignancies, including lung cancers [102]. In this context, there is a growing demand to leverage premalignant lesions in cancer immunotherapeutic strategies [103, 104]. Therefore, additional work to extend the immune marker deregulation identified in AAHs of our cohort, might suggest measures for targeting immune responses and signaling (e.g., immune checkpoint blockade) as a viable strategy to prevent progression of preneoplasia such as AAH. For example, ongoing studies have begun to investigate neoantigen profiles

in AAH [101], analogous to patterns seen in LUADs [105]. Hence, future studies examining protein levels of markers of various immune cell infiltrates (e.g., immunohistochemistry methods) will shed more light on the role of the immune response in AAH pathogenesis.

Interestingly, recent studies suggest that driver mutations are likely to be found in preconditioned epithelial fields of phenotypically healthy individuals [106, 107]. In this regard, future studies, perhaps integrating our efforts of characterizing the field cancerization effect in early-stage NSCLCs with similar studies in control subjects (e.g., lung cancer-free smokers), may improve our understanding of field cancerization in lung cancer pathogenesis. Such a study will help differentiate and better understand the multi-step path to NSCLC pathogenesis from the airway field of injury. Additionally, given the young age of this cohort collection, a longitudinal follow-up, including a continued sampling of airway epithelia as well as an up-to-date record of clinical features (e.g., treatment regimen, smoking status, recurrence status, survival status), may help provide insights into the potential role of profiling airway epithelia to predict recurrence, relapse, response to treatment as well as survival. Another computational extension to the study involves probing intra-field and intra-tumor heterogeneity patterns in early-stage NSCLCs, especially in a subset of cases with multiple core needle biopsies of the tumor. Such an analysis might provide evidence for the presence of truncal tumor mutations in the airway field, further offering insights into tumor development or recurrence.

I presented independent studies to investigate molecular aberrations in premalignant and cancerized fields of NSCLCs. However, future studies, concurrently studying matched airway field, premalignant and malignant tumor tissues may help create a more comprehensive genomic roadmap to NSCLC pathogenesis.

The work described in Chapter 4 involving a pan-cancer investigation of TCGA, provides a repository of adjusted chromosome-arm copy number alterations in TCGA as well as additional allelic imbalance derived chromosomal aberrations, such as regions of copy-neutral loss of heterozygosity. Future work, perhaps integrating these sources of aberrations along with focal copy number alterations and point mutation profiles, can aid in discoveries of compounded mutation signatures across tumor types, that may also present prognostic potential. Our findings of widespread loss and cnLOH on 17p warrants studies

that investigate the differences between these two event types on 17p, as well as their role with respect to the presence of important driver gene mutations (e.g., *TP53*) can aid in better understanding the dynamics of tumors that exhibit these multi-hit aberrations. In addition, investigations of allelic imbalance in tumor-adjacent normal tissues in TCGA can also inform of field cancerization mechanisms across tumor sites. Computationally, future methods to jointly model BAF and LRR in the statistical estimation of allelic imbalance might offer more sensitivity and help identify events such as balanced duplications, that are currently unidentifiable using our approach.

In this thesis, I focused on identifying the somatic mutation landscape of these premalignant and normal tissues, however, a crucial and complementary form of data derives from the germline landscape of inherited variants. Therefore, the somatic mutations and chromosomal aberrations identified in these studies can be integrated with matched germline variants to infer other complex mechanisms of NSCLC pathogenesis.

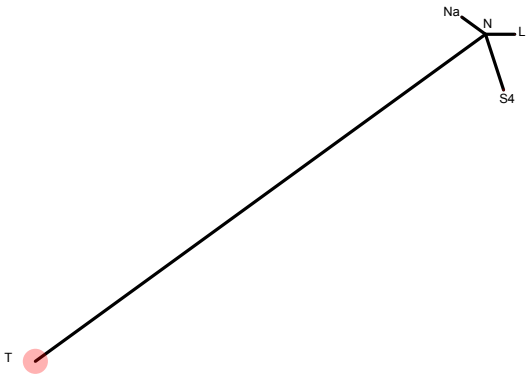
5.3 Summary

This dissertation presented a data-driven genomics approach to investigate normal and premalignant tissues, using early-stage lung cancers as a model of study. The study design, profiling technologies and bioinformatic approaches developed and implemented in this thesis provide great potential for utility in ongoing large-scale investigations, such as in the development of the PreCancer Atlas [108]. The findings also carry significant potential in predicting outcomes in high-risk patients, and may lead to novel biomarker discovery and personalized chemo-preventive strategies in these tumors.

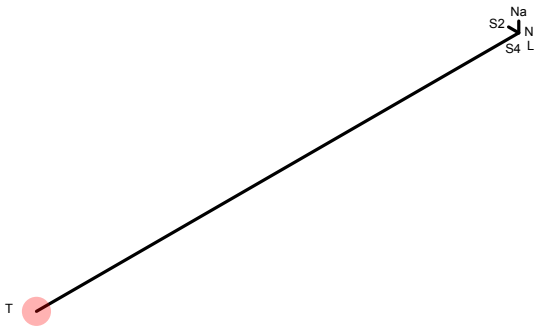
APPENDIX A
PHYLOGENETIC ANALYSIS OF FIELD CANCERIZATION

The methods described in Section 3.2.6.3 were used to infer within-patient inter-sample relationships for all 48 individuals in the cohort. Phylogenetic trees were constructed based on the extent of shared events, comprising both single nucleotide mutations (SNVs) and allelic imbalance derived chromosome-arm level alterations (AI). A scale is attached to each tree, describing the number of aberrations. Each node in the tree, corresponding to the samples profiled for that individual, are denoted by their tissue annotation and a red circle, scaled to the total number of observed aberrations in that tissue. Where available, the profiles of tumor core-needle biopsy samples were combined with the matched tissue biopsy (collectively denoted as T), for the construction of these trees. The tissues are denoted as follows: Tumor (T), tumor-adjacent small airways (S1-S5), distant large airway (L), nasal epithelium (Na) and uninvolved normal lung parenchyma (N). Trees exhibiting non-linear structures, thereby indicating shared events between matched NSCLCs and airway field, were identified as potential exhibitors of genomic airway field carcinogenesis. The trees for all 48 individuals are shown below.

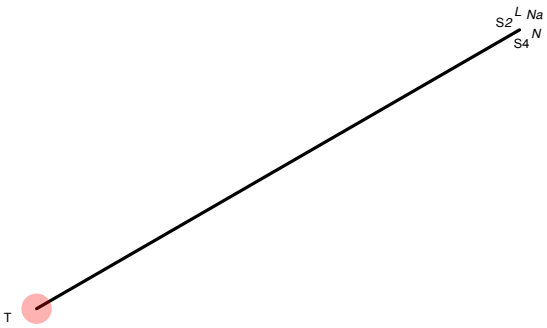
AIR_001 (n=23)



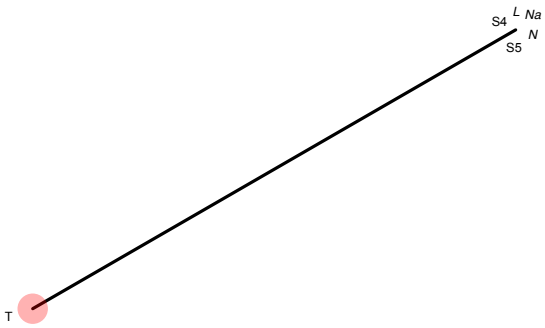
AIR_002 (n=49)



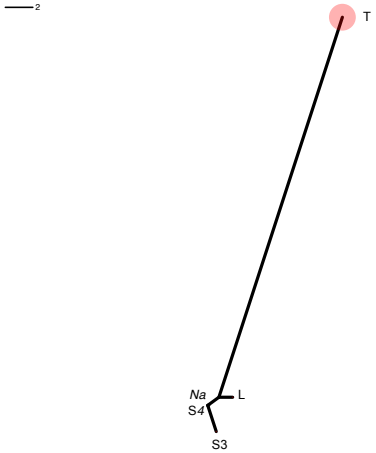
AIR_003 (n=24)



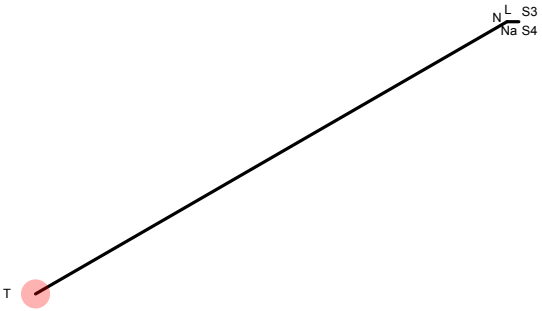
AIR_004 (n=7)



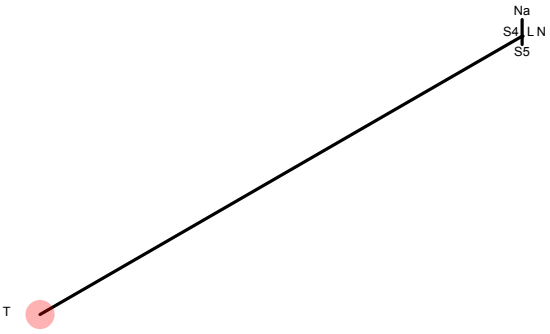
AIR_005 (n=33)



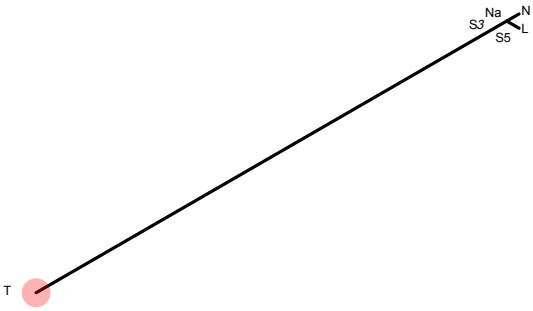
AIR_007 (n=47)



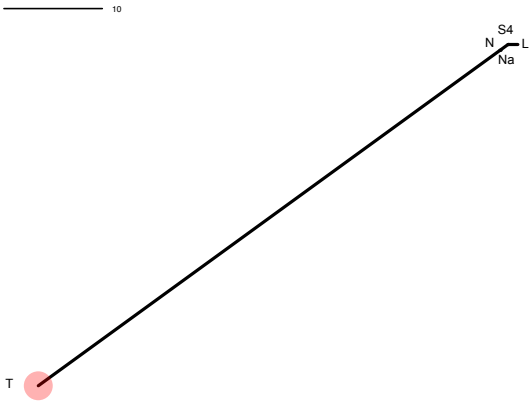
AIR_008 (n=70)



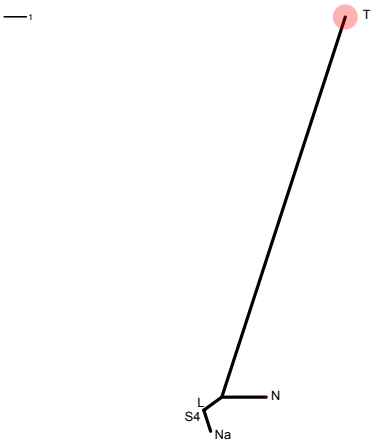
AIR_009 (n=40)



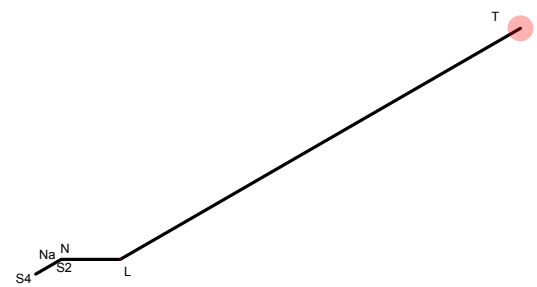
AIR_010 (n=60)



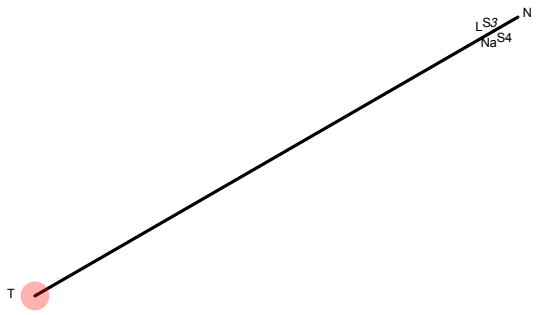
AIR_011 (n=22)



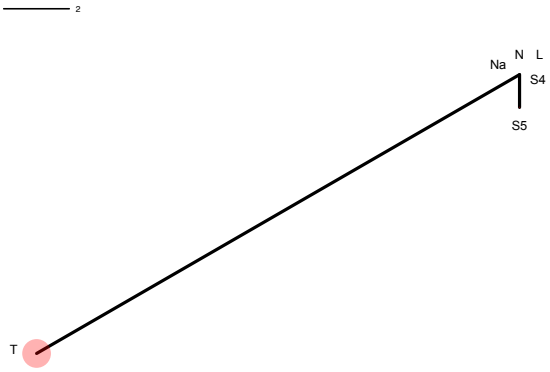
AIR_012 (n=56)



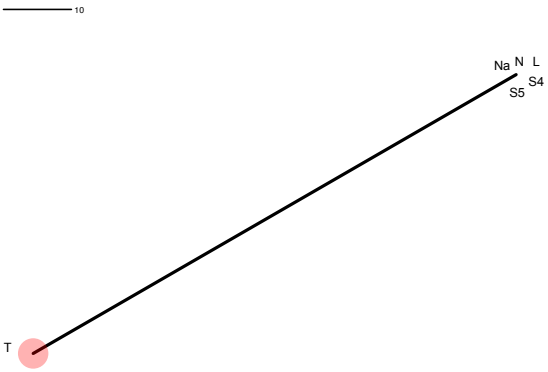
AIR_013 (n=20)



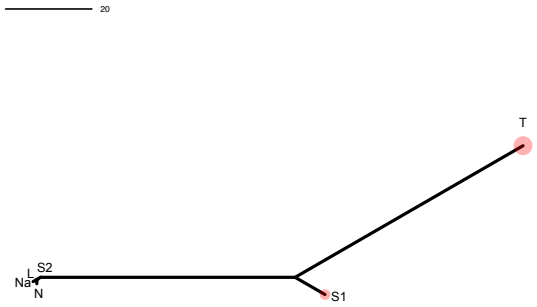
AIR_014 (n=18)



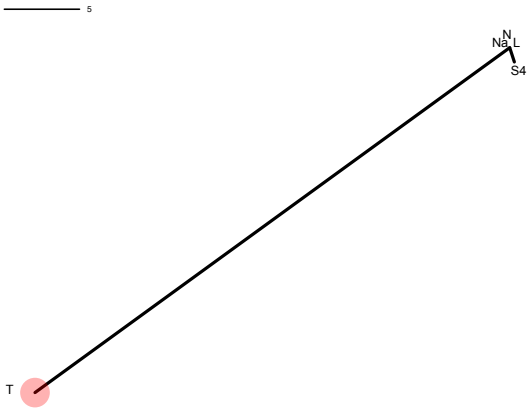
AIR_015 (n=82)



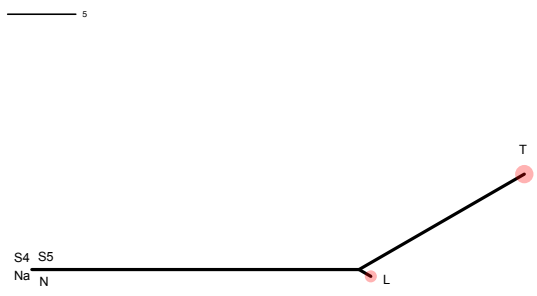
AIR_016 (n=131)



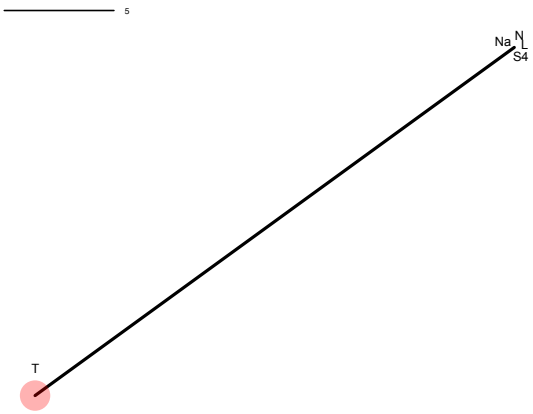
AIR_017 (n=40)



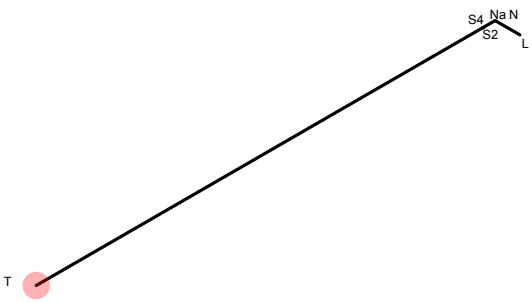
AIR_018 (n=39)



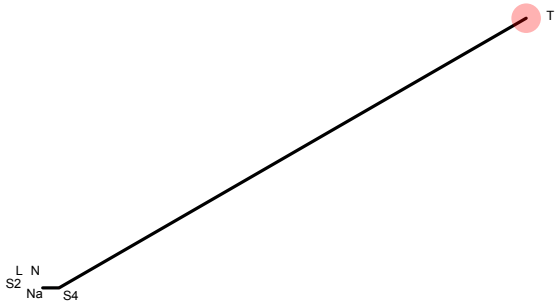
AIR_019 (n=27)



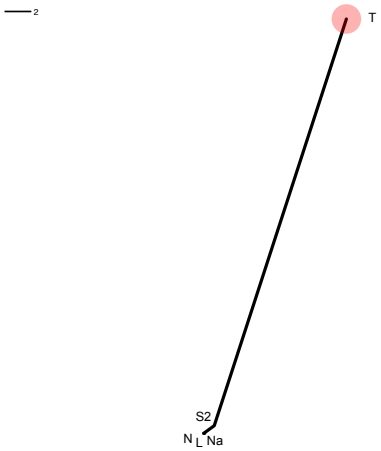
AIR_020 (n=39)



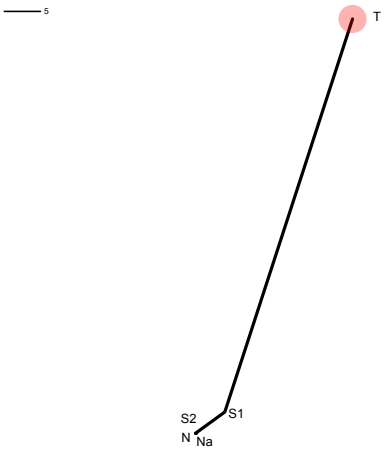
AIR_022 (n=34)



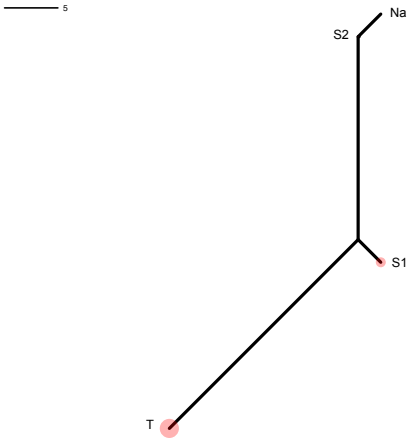
AIR_023 (n=34)



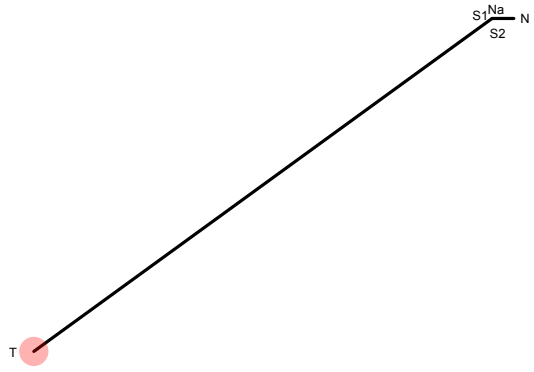
AIR_024 (n=62)



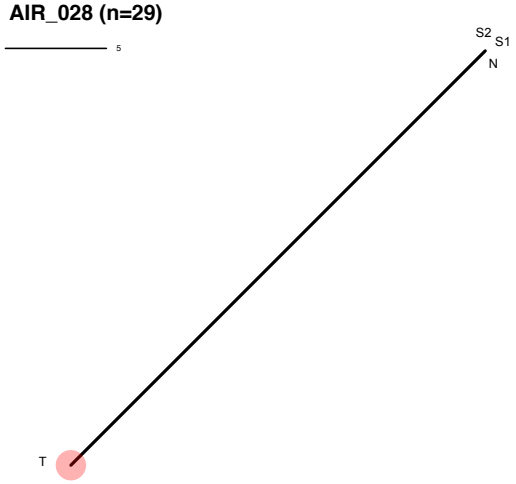
AIR_026 (n=50)



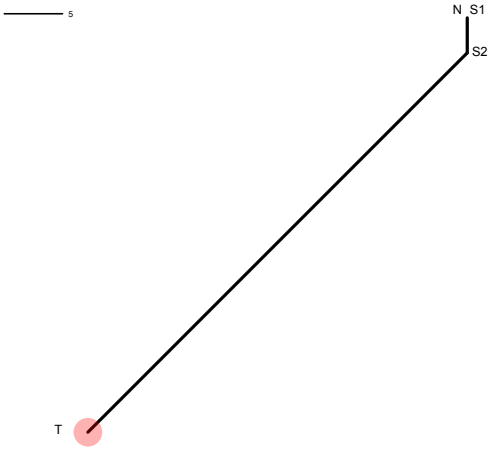
AIR_027 (n=27)



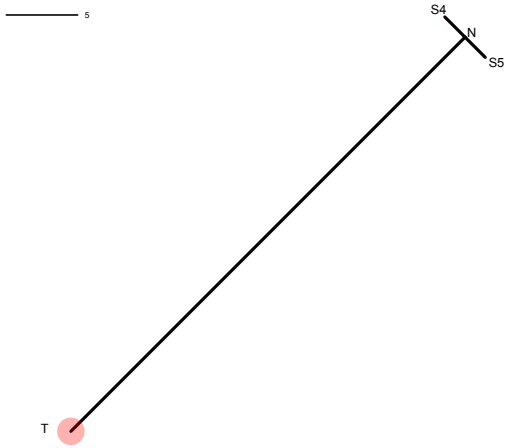
AIR_028 (n=29)



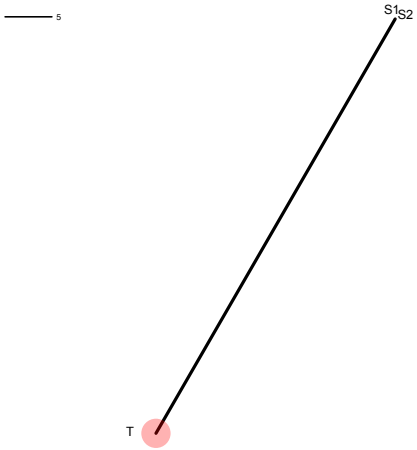
AIR_031 (n=49)



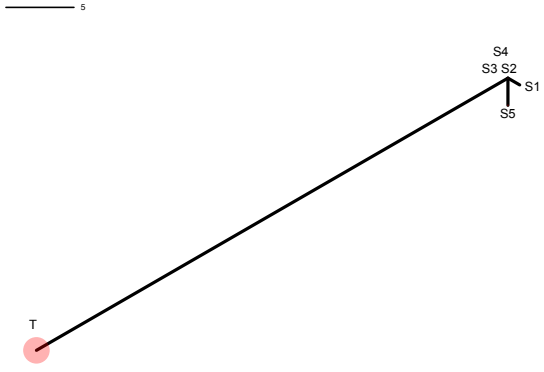
AIR_032 (n=43)



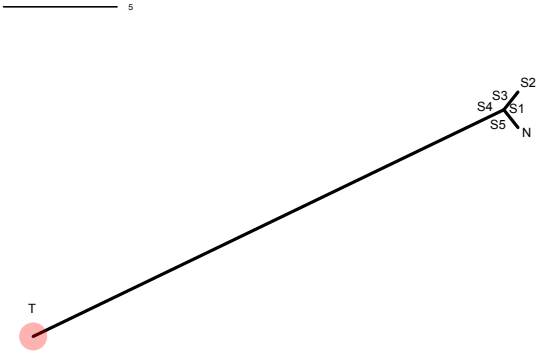
AIR_033 (n=52)



AIR_034 (n=43)

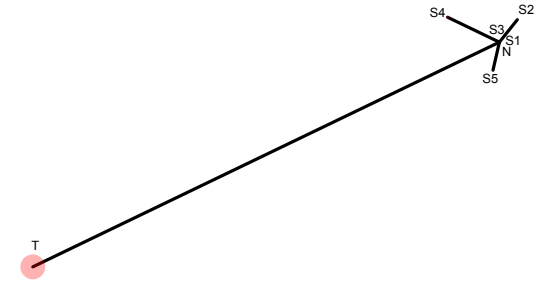


AIR_035 (n=25)



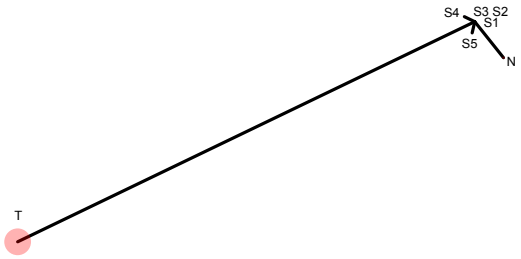
AIR_036 (n=22)

5



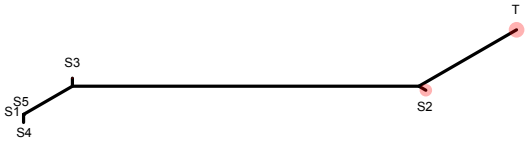
AIR_037 (n=50)

10



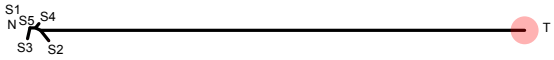
AIR_039 (n=67)

10



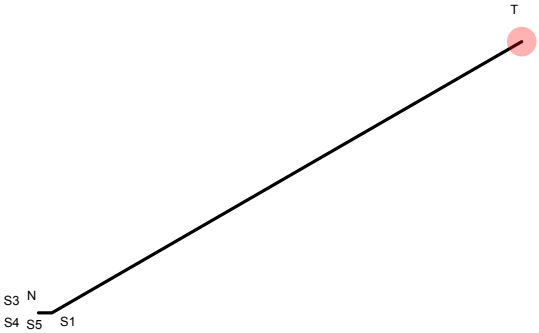
AIR_040 (n=47)

10



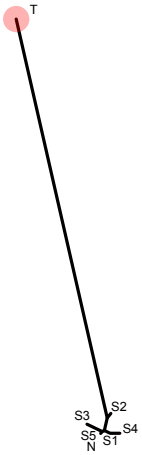
AIR_041 (n=41)

5

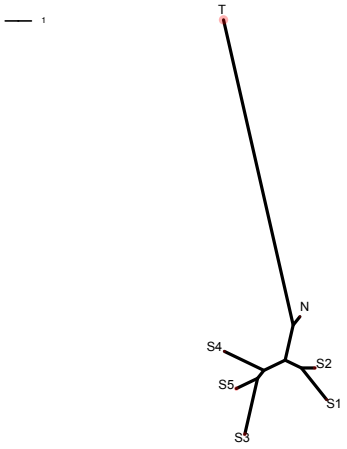


AIR_042 (n=47)

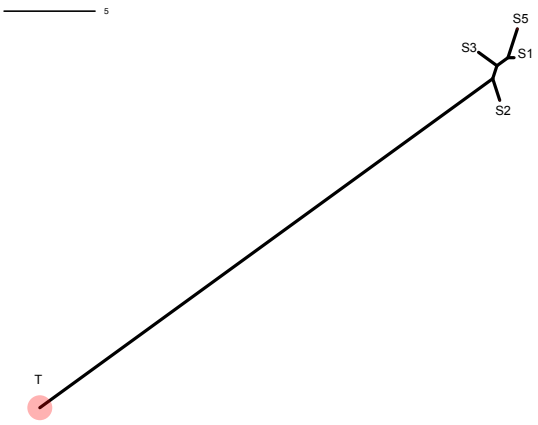
2



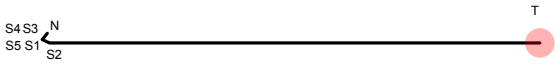
AIR_043 (n=21)



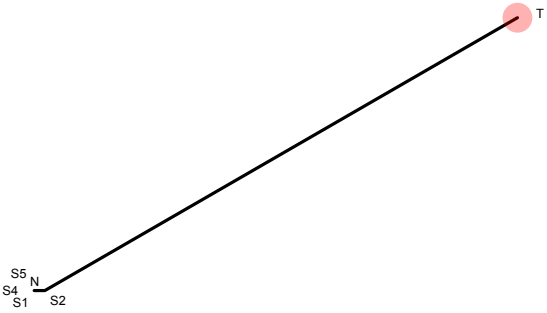
AIR_044 (n=36)



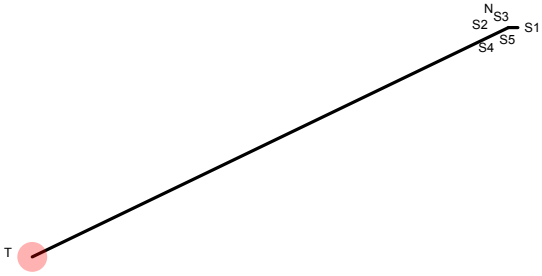
AIR_045 (n=64)



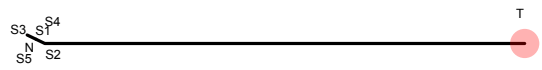
AIR_047 (n=52)



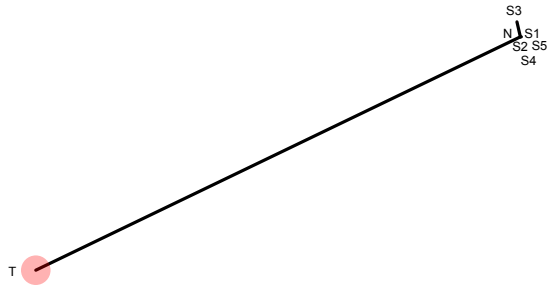
AIR_048 (n=57)



AIR_049 (n=51)

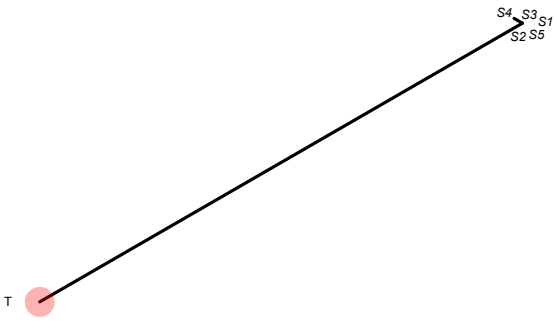


AIR_050 (n=36)

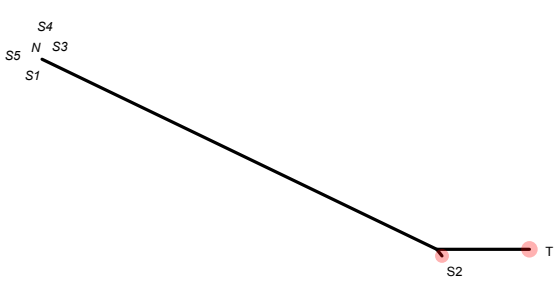


AIR_052 (n=64)

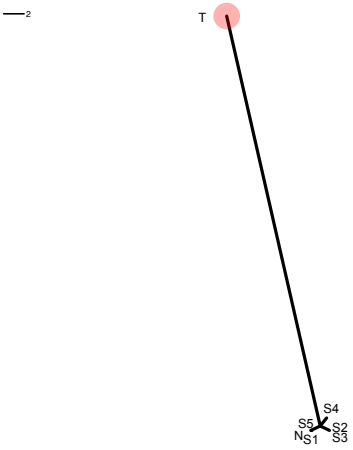
AIR_053 (n=59)



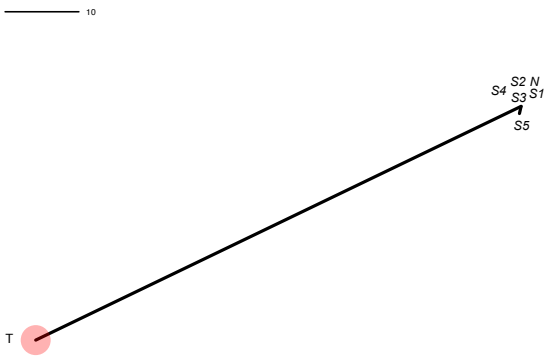
AIR_054 (n=64)



AIR_055 (n=44)

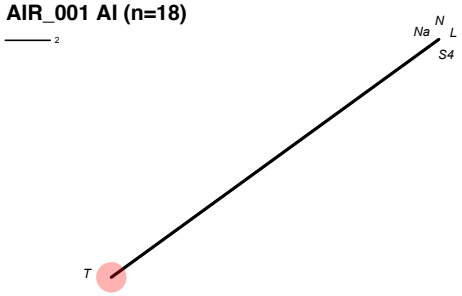


AIR_056 (n=74)

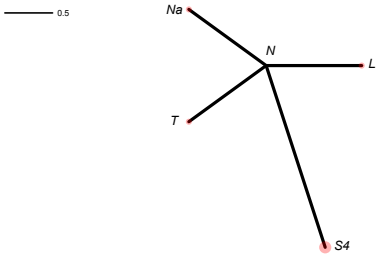


Phylogenetic trees were also constructed independently for the two different types of aberrations studied: single nucleotide mutations (SNVs) and allelic imbalance derived chromosome-arm level alterations (AI). The concordance of tree structures was tested using `all.equal.phylo` function in the `ape` R package, while setting `use.tip.label=TRUE`. The AI-derived trees and SNV-derived trees for all 48 individuals are shown below. A scale is attached to each tree, describing the number of aberrations. Each node in the tree, corresponding to the samples profiled for that individual, are denoted by their tissue annotation and a red circle, scaled to the total number of observed aberrations in that tissue. Where available, the profiles of tumor core-needle biopsy samples were combined with the matched tissue biopsy (collectively denoted as T), for the construction of these trees. The tissues are denoted as follows: Tumor (T), tumor-adjacent small airways (S1-S5), distant large airway (L), nasal epithelium (Na) and uninvolved normal lung parenchyma (N).

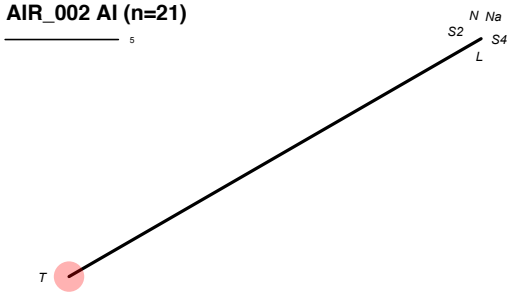
AIR_001 AI (n=18)



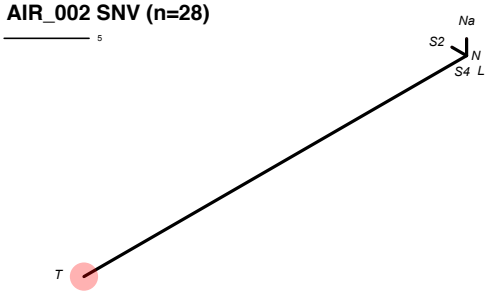
AIR_001 SNV (n=5)



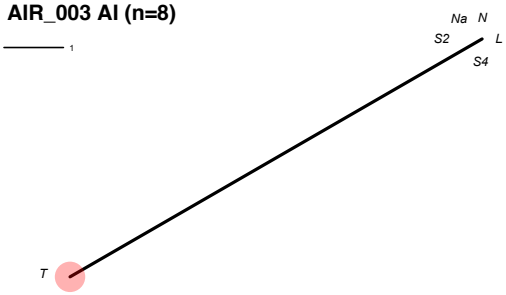
AIR_002 AI (n=21)



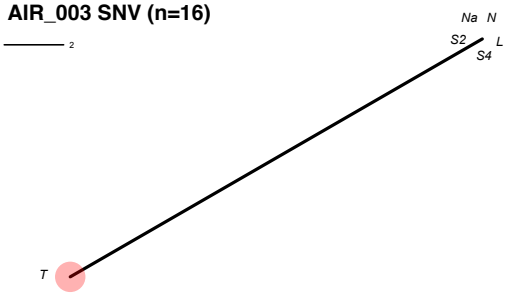
AIR_002 SNV (n=28)



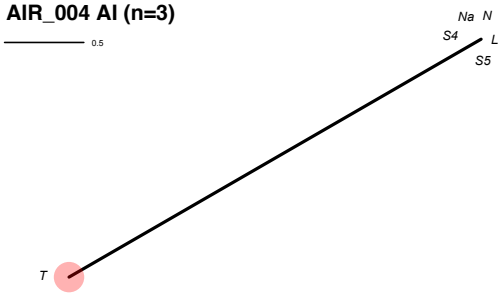
AIR_003 AI (n=8)



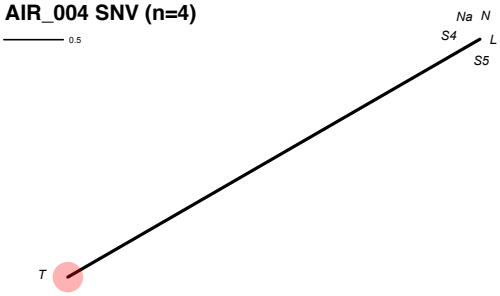
AIR_003 SNV (n=16)



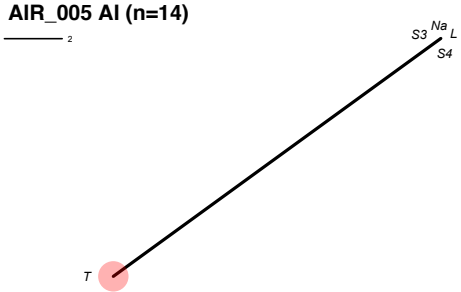
AIR_004 AI (n=3)



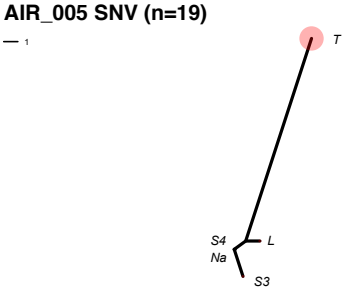
AIR_004 SNV (n=4)



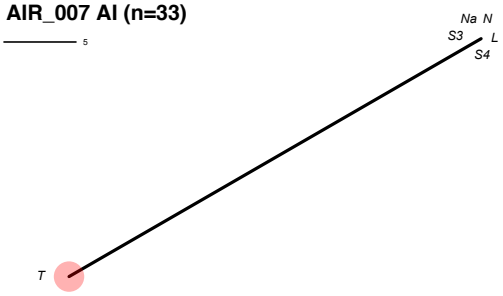
AIR_005 AI (n=14)



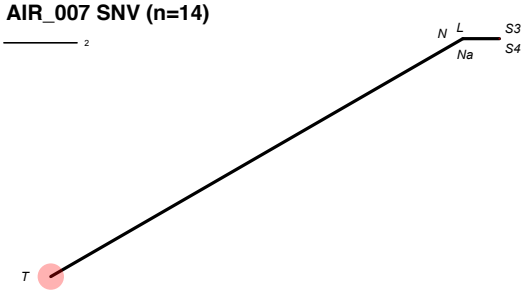
AIR_005 SNV (n=19)



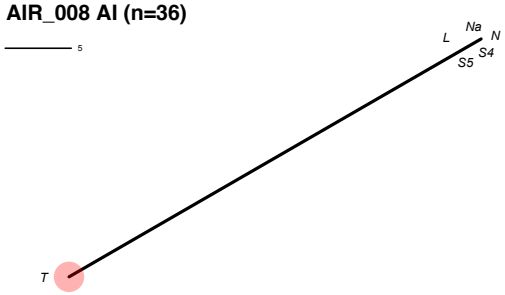
AIR_007 AI (n=33)



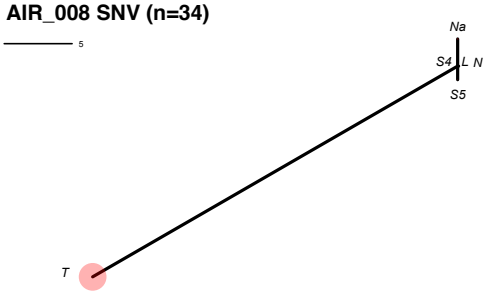
AIR_007 SNV (n=14)



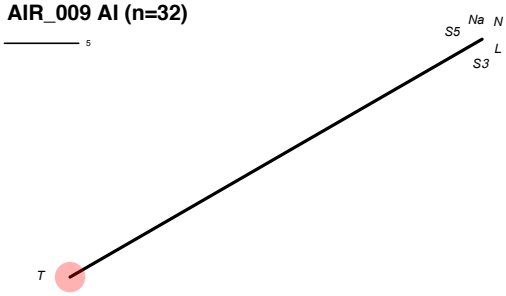
AIR_008 AI (n=36)



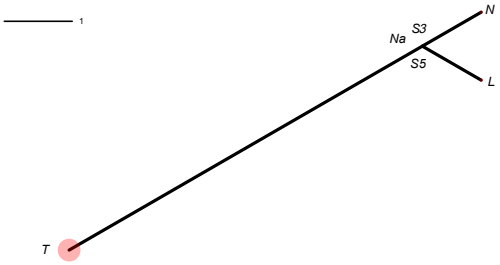
AIR_008 SNV (n=34)



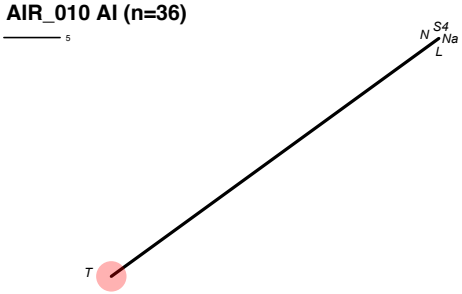
AIR_009 AI (n=32)



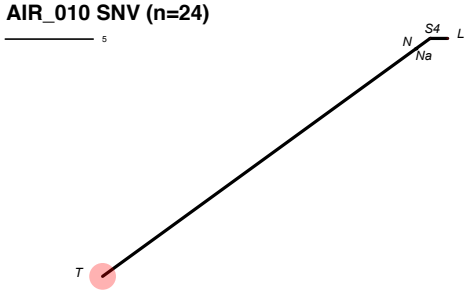
AIR_009 SNV (n=8)



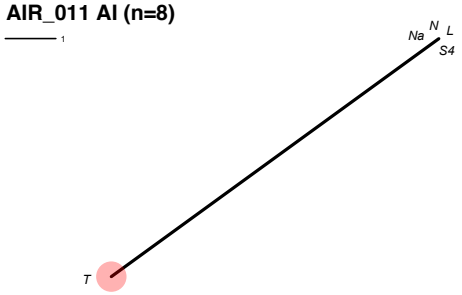
AIR_010 AI (n=36)



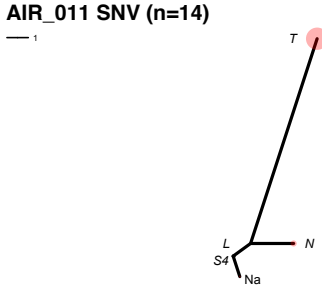
AIR_010 SNV (n=24)



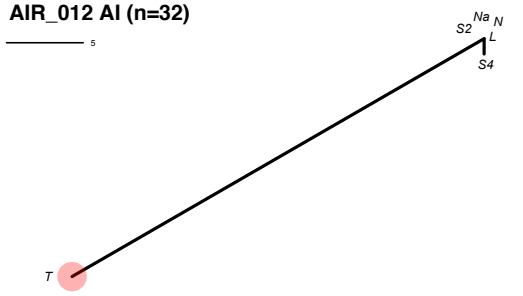
AIR_011 AI (n=8)



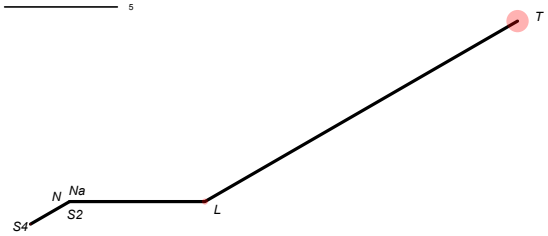
AIR_011 SNV (n=14)



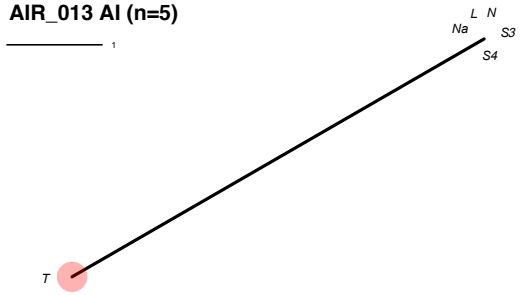
AIR_012 AI (n=32)



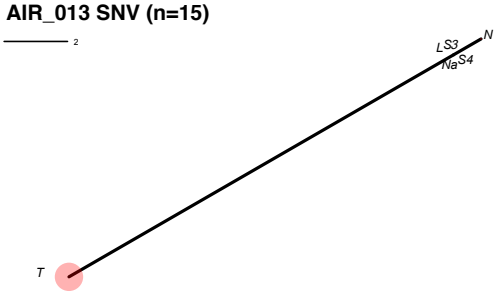
AIR_012 SNV (n=24)



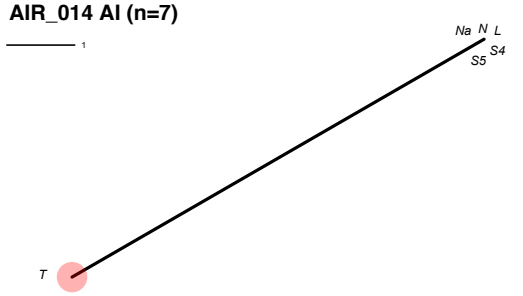
AIR_013 AI (n=5)



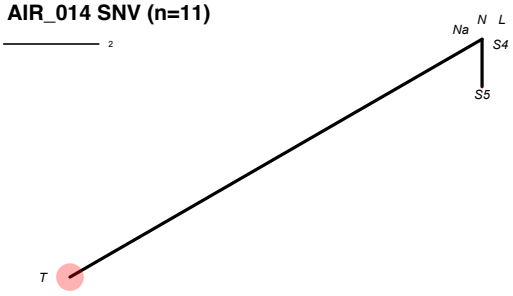
AIR_013 SNV (n=15)



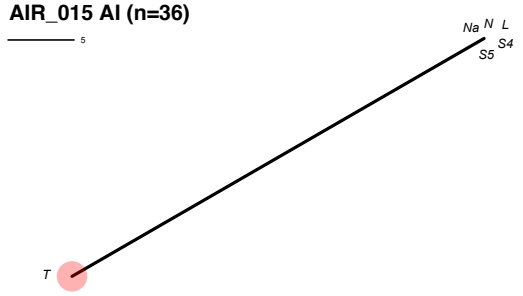
AIR_014 AI (n=7)



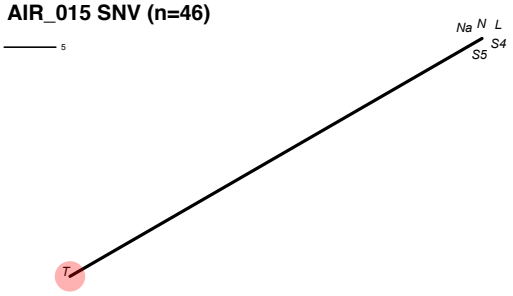
AIR_014 SNV (n=11)



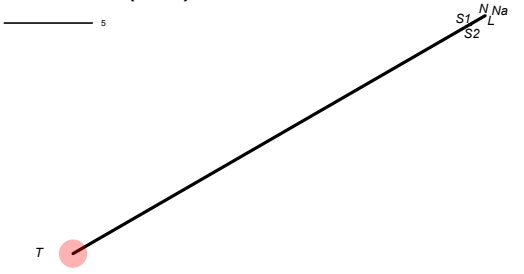
AIR_015 AI (n=36)



AIR_015 SNV (n=46)

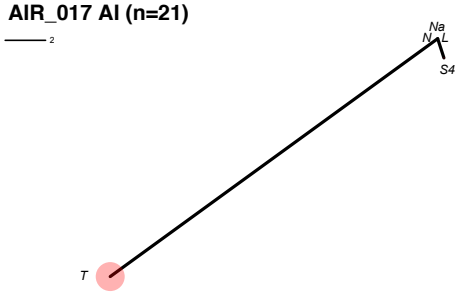


AIR_016 AI (n=27)

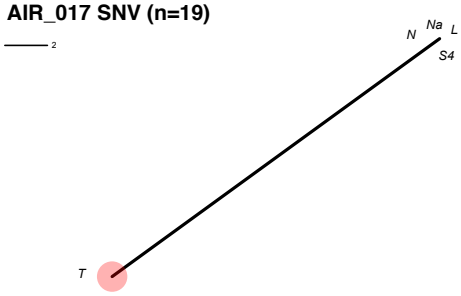


AIR_016 SNV (n=104)

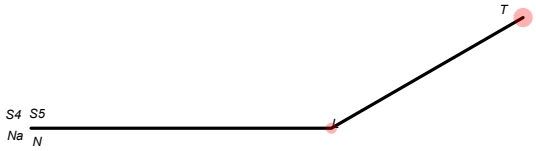
AIR_017 AI (n=21)



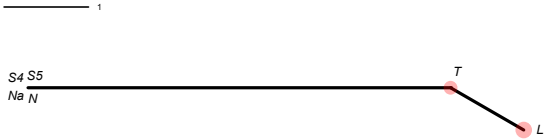
AIR_017 SNV (n=19)



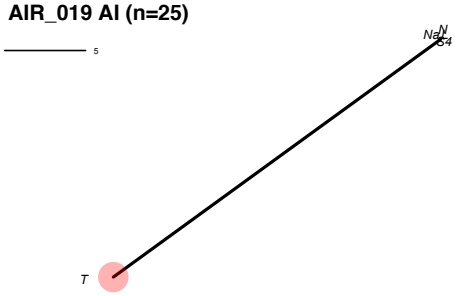
AIR_018 AI (n=33)



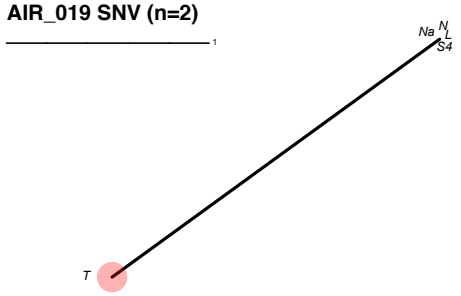
AIR_018 SNV (n=6)



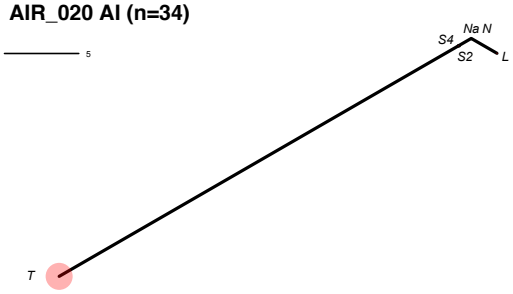
AIR_019 AI (n=25)



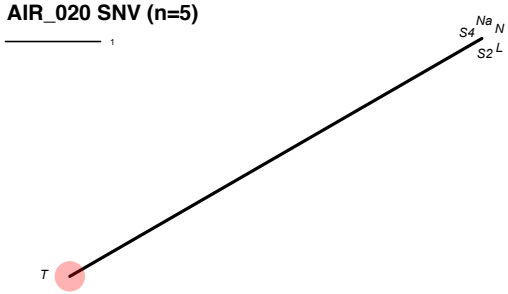
AIR_019 SNV (n=2)



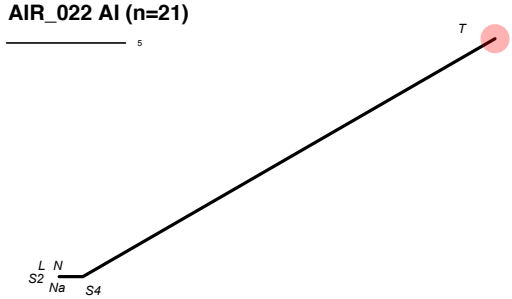
AIR_020 AI (n=34)



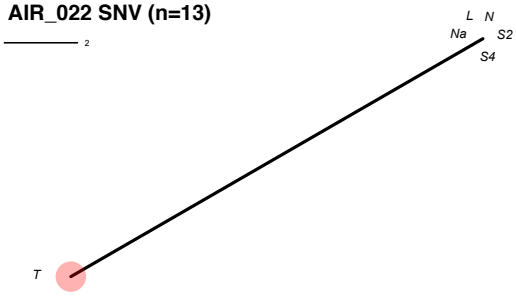
AIR_020 SNV (n=5)



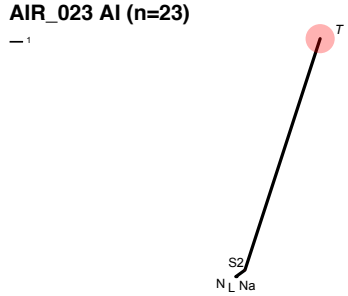
AIR_022 AI (n=21)



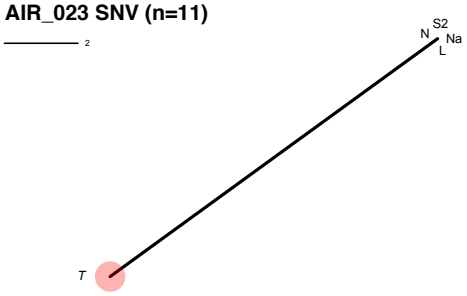
AIR_022 SNV (n=13)



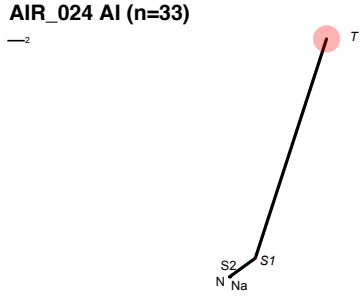
AIR_023 AI (n=23)



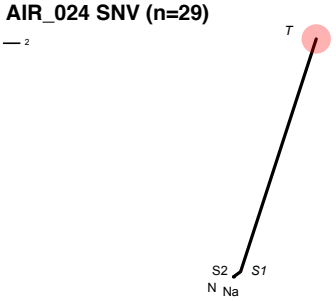
AIR_023 SNV (n=11)



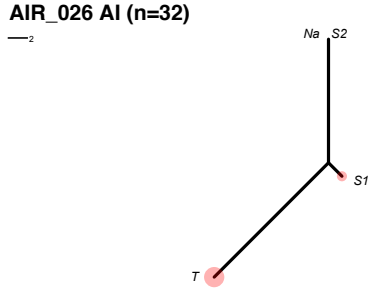
AIR_024 AI (n=33)



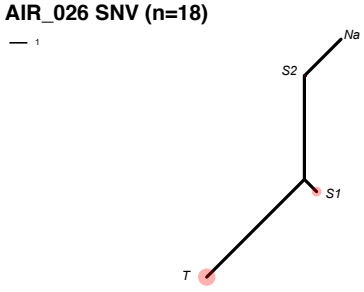
AIR_024 SNV (n=29)



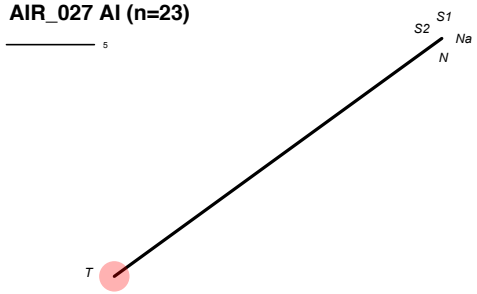
AIR_026 AI (n=32)



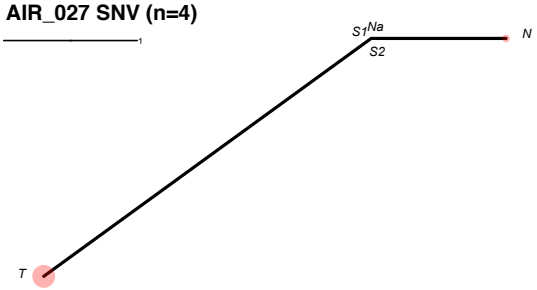
AIR_026 SNV (n=18)



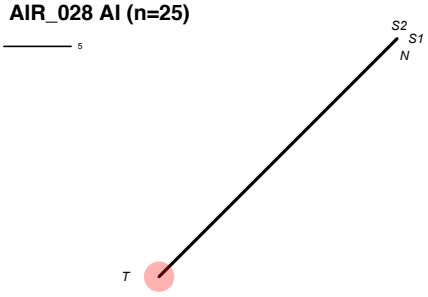
AIR_027 AI (n=23)



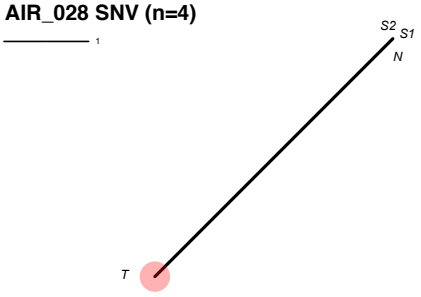
AIR_027 SNV (n=4)



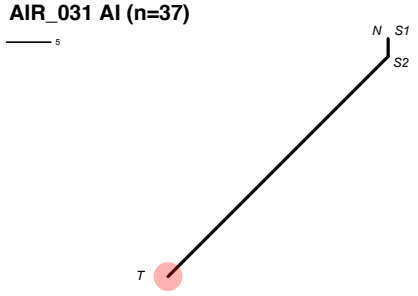
AIR_028 AI (n=25)



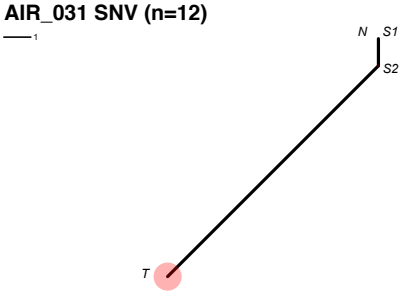
AIR_028 SNV (n=4)



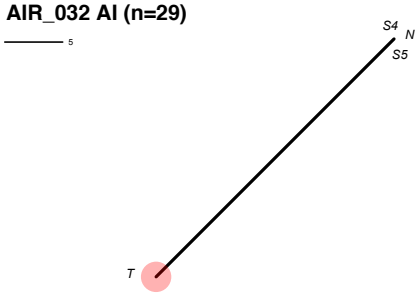
AIR_031 AI (n=37)



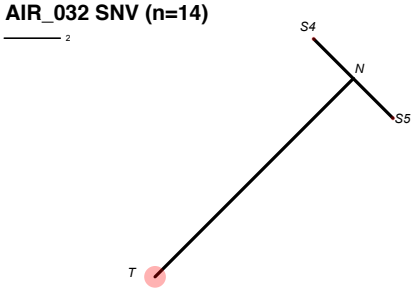
AIR_031 SNV (n=12)



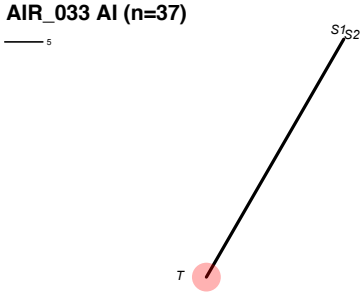
AIR_032 AI (n=29)



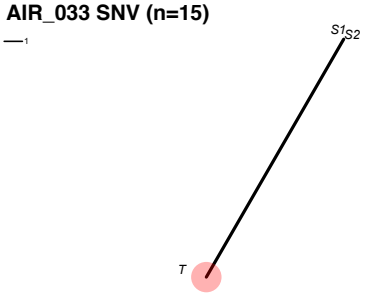
AIR_032 SNV (n=14)



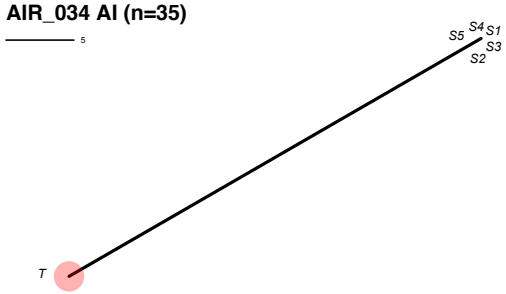
AIR_033 AI (n=37)



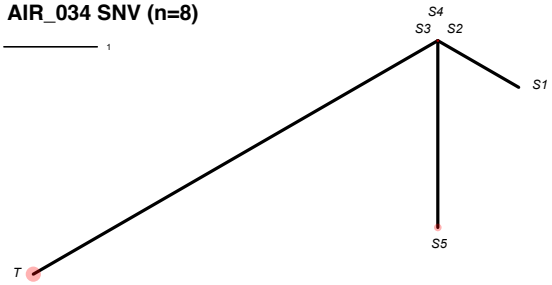
AIR_033 SNV (n=15)



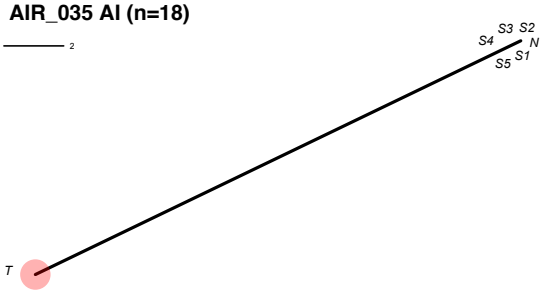
AIR_034 AI (n=35)



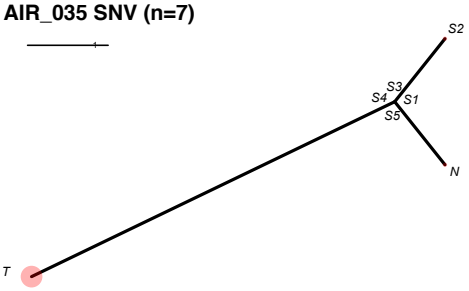
AIR_034 SNV (n=8)



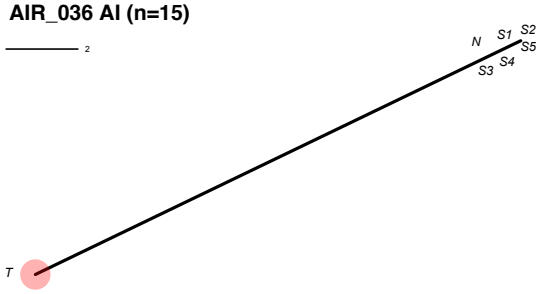
AIR_035 AI (n=18)



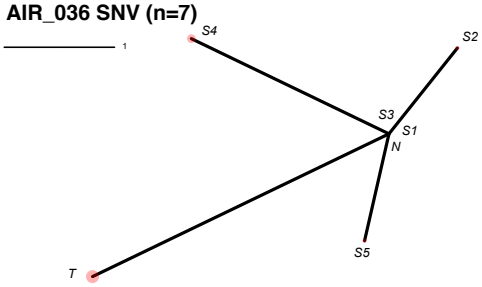
AIR_035 SNV (n=7)



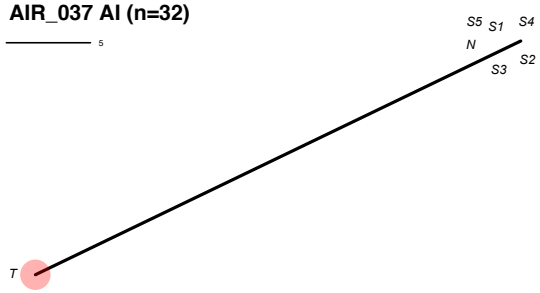
AIR_036 AI (n=15)



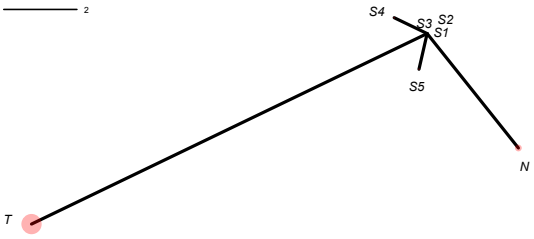
AIR_036 SNV (n=7)



AIR_037 AI (n=32)

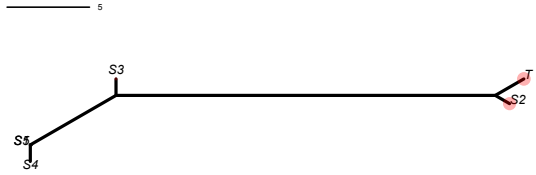


AIR_037 SNV (n=18)

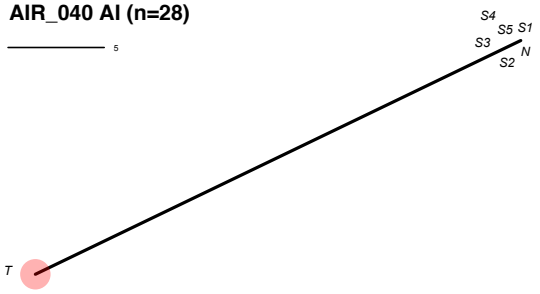


AIR_039 AI (n=33)

AIR_039 SNV (n=34)

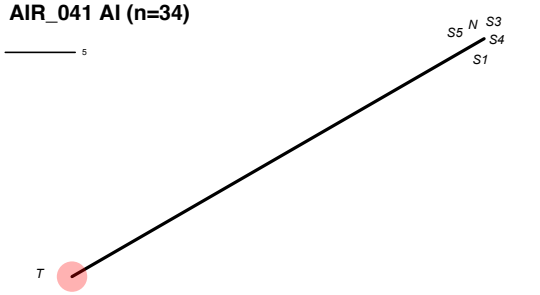


AIR_040 AI (n=28)

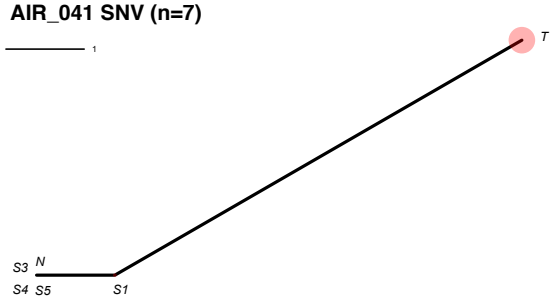


AIR_040 SNV (n=19)

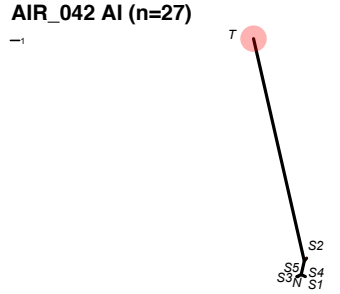
AIR_041 AI (n=34)



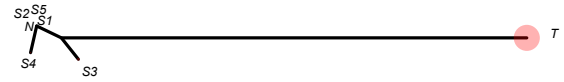
AIR_041 SNV (n=7)



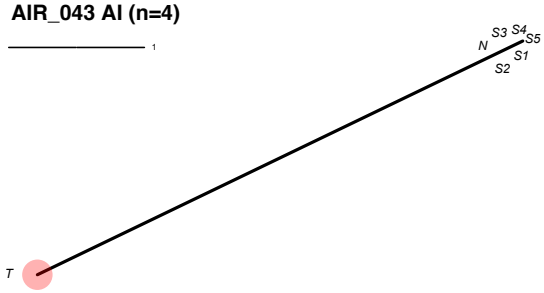
AIR_042 AI (n=27)



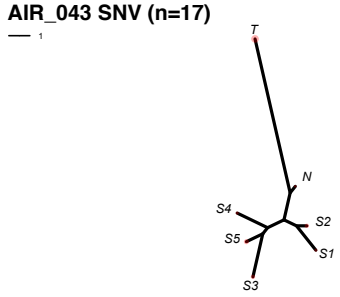
AIR_042 SNV (n=20)



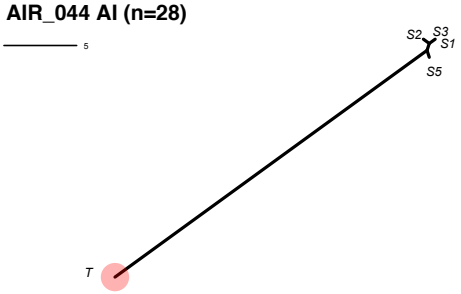
AIR_043 AI (n=4)



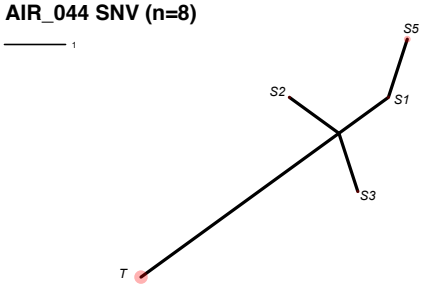
AIR_043 SNV (n=17)



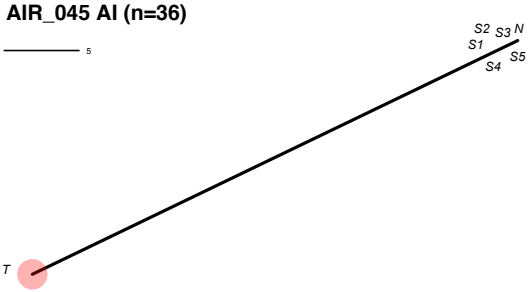
AIR_044 AI (n=28)



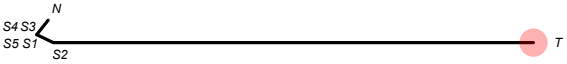
AIR_044 SNV (n=8)



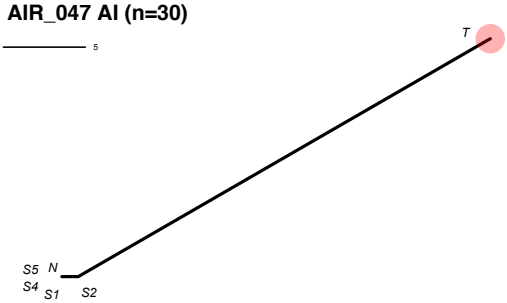
AIR_045 AI (n=36)



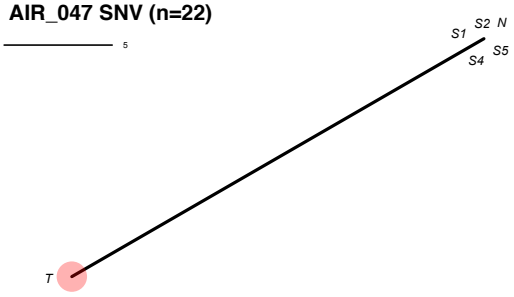
AIR_045 SNV (n=28)



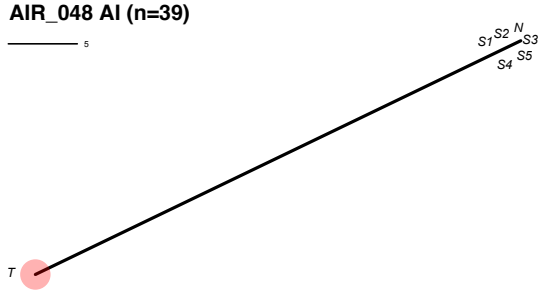
AIR_047 AI (n=30)



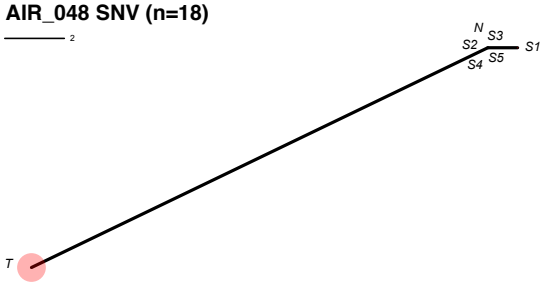
AIR_047 SNV (n=22)



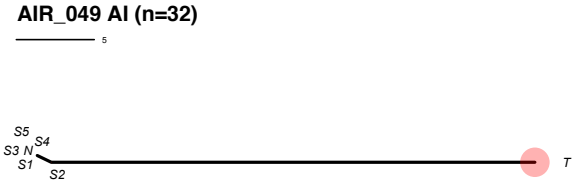
AIR_048 AI (n=39)



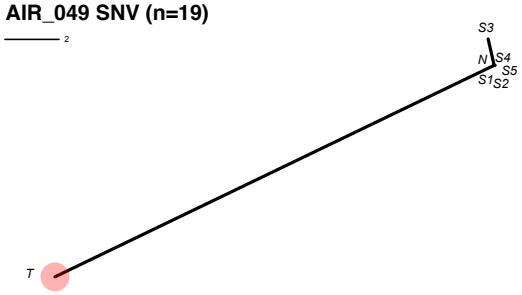
AIR_048 SNV (n=18)



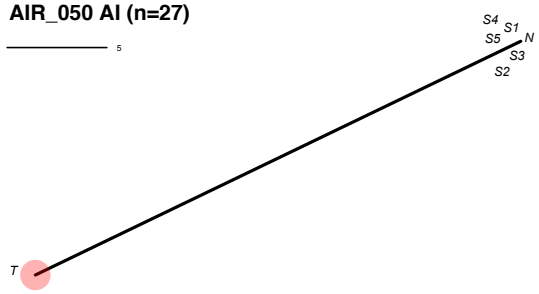
AIR_049 AI (n=32)



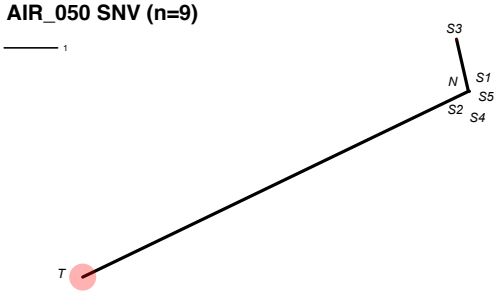
AIR_049 SNV (n=19)



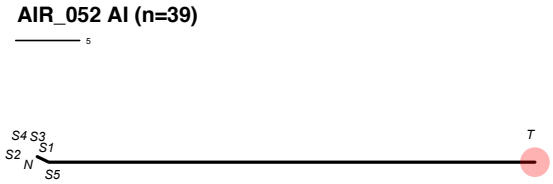
AIR_050 AI (n=27)



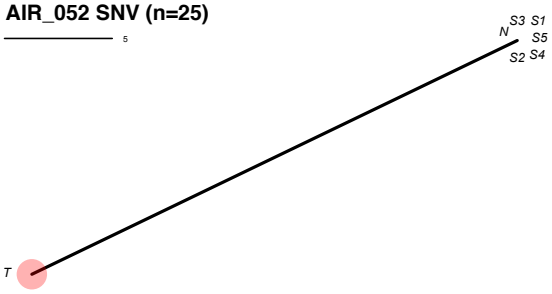
AIR_050 SNV (n=9)



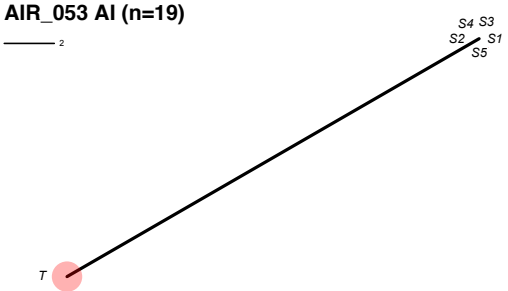
AIR_052 AI (n=39)



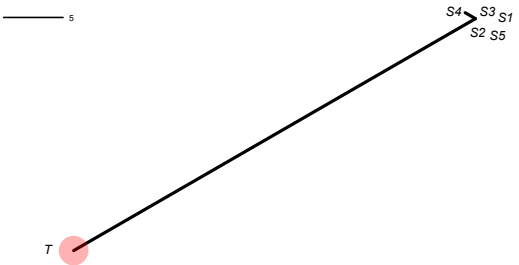
AIR_052 SNV (n=25)



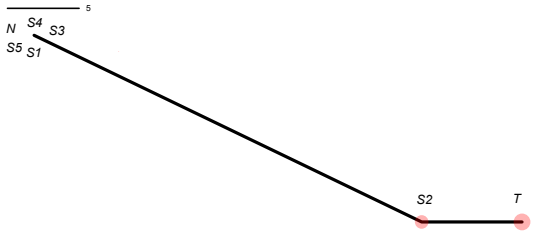
AIR_053 AI (n=19)



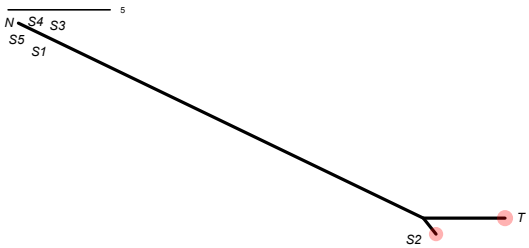
AIR_053 SNV (n=40)



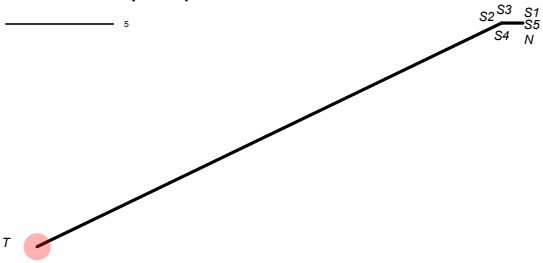
AIR_054 AI (n=37)



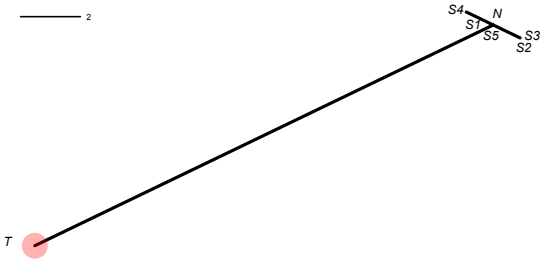
AIR_054 SNV (n=27)



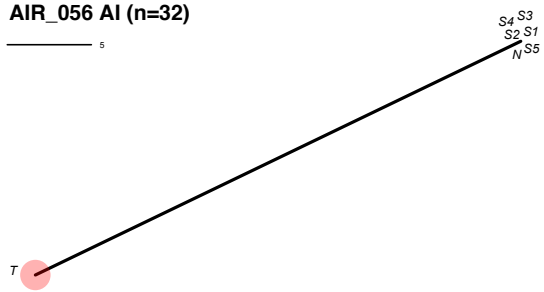
AIR_055 AI (n=25)



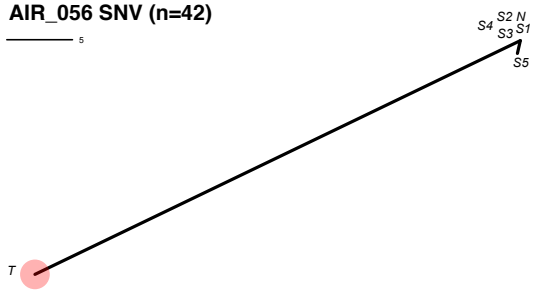
AIR_055 SNV (n=19)



AIR_056 AI (n=32)



AIR_056 SNV (n=42)



BIBLIOGRAPHY

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[2] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis, "Genomic instability — an evolving hallmark of cancer," *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 3, p. 220, Mar. 2010.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, 2019.

[4] R. S. Herbst, J. V. Heymach, and S. M. Lippman, "Lung cancer," *N. Engl. J. Med.*, vol. 359, no. 13, pp. 1367–1380, Sep. 2008.

[5] Cancer Genome Atlas Research Network, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014.

[6] The Cancer Genome Atlas Research Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, p. 519, Sep. 2012.

[7] National Lung Screening Trial Research Team, D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *N. Engl. J. Med.*, vol. 365, no. 5, pp. 395–409, Aug. 2011.

[8] H Kadara, P Scheet, I. I. Wistuba, and A. E. Spira, "Early events in the molecular pathogenesis of lung cancer," *Cancer Prev. Res.*, vol. 9, no. 7, pp. 518–527, 2016.

[9] R Kaufmann, "The concept of field cancerization," *Melanoma Res.*, vol. 20, e13–e14, 2010.

[10] K. Curtius, N. A. Wright, and T. A. Graham, "An evolutionary perspective on field cancerization," *Nat. Rev. Cancer*, vol. 18, no. 1, p. 19, Dec. 2017.

[11] I. I. Wistuba, C Behrens, S Milchgrub, D Bryant, J Hung, J. D. Minna, and A. F. Gazdar, "Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma," *Oncogene*, vol. 18, no. 3, pp. 643–650, Jan. 1999.

[12] U. Grepmeier, W. Dietmaier, J. Merk, P. J. Wild, E. C. Obermann, M. Pfeifer, F. Hofstaedter, A. Hartmann, and M. Woenckhaus, "Deletions at chromosome 2q and 12p are early and frequent molecular alterations in bronchial epithelium and NSCLC of long-term smokers," *Int. J. Oncol.*, vol. 27, no. 2, pp. 481–488, Aug. 2005.

[13] Y. Jakubek, W. Lang, S. Vattathil, M. Garcia, L. Xu, L. Huang, S.-Y. Yoo, L. Shen, W. Lu, C.-W. Chow, Z. Weber, G. Davies, J. Huang, C. Behrens, N. Kalhor, C. Moran, J. Fujimoto, R. Mehran, R. El-Zein, S. G. Swisher, J. Wang, J. Fowler, A. E. Spira, E. A. Ehli, I. I. Wistuba, P. Scheet, and H. Kadara, "Genomic landscape established by allelic imbalance in the cancerization field of a normal appearing airway," *Cancer Res.*, vol. 76, no. 13, pp. 3676–3683, Jul. 2016.

[14] H. Kadara, J. Fujimoto, S.-Y. Yoo, Y. Maki, A. C. Gower, M. Kabbout, M. M. Garcia, C.-W. Chow, Z. Chu, G. Mendoza, L. Shen, N. Kalhor, W. K. Hong, C. Moran, J. Wang, A. Spira, K. R. Coombes, and I. I. Wistuba, "Transcriptomic architecture of the adjacent airway field cancerization in Non–Small cell lung cancer," *JNCI: Journal of the National Cancer Institute*, vol. 106, no. 3, 2014.

[15] J. Beane, J. Vick, F. Schembri, C. Anderlind, A. Gower, J. Campbell, L. Luo, X. H. Zhang, J. Xiao, Y. O. Alekseyev, S. Wang, S. Levy, P. P. Massion, M. Lenburg, and A. Spira, "Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq," *Cancer Prev. Res.*, vol. 4, no. 6, pp. 803–817, Jun. 2011.

[16] E. Billatos, J. L. Vick, M. E. Lenburg, and A. E. Spira, "The airway transcriptome as a biomarker for early lung cancer detection," *Clin. Cancer Res.*, vol. 24, no. 13, pp. 2984–2992, Jul. 2018.

[17] M. A. Nelson, J Wymer, and N Clements Jr, "Detection of k-ras gene mutations in non-neoplastic lung tissue and lung cancers," *Cancer Lett.*, vol. 103, no. 1, pp. 115–121, May 1996.

[18] W. A. Franklin, A. F. Gazdar, J Haney, I. I. Wistuba, F. G. La Rosa, T Kennedy, D. M. Ritchey, and Y. E. Miller, "Widely dispersed p53 mutation in respiratory epithelium. a novel mechanism for field carcinogenesis," *J. Clin. Invest.*, vol. 100, no. 8, pp. 2133–2137, Oct. 1997.

[19] X. Tang, H. Shigematsu, B. N. Bekele, J. A. Roth, J. D. Minna, W. K. Hong, A. F. Gazdar, and I. I. Wistuba, "EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients," *Cancer Res.*, vol. 65, no. 17, pp. 7568–7572, Sep. 2005.

[20] S. A. Belinsky, W. A. Palmisano, F. D. Gilliland, L. A. Crooks, K. K. Divine, S. A. Winters, M. J. Grimes, H. J. Harms, C. S. Tellez, T. M. Smith, P. P. Moots, J. F. Lechner, C. A. Stidley, and R. E. Crowell, "Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers," *Cancer Res.*, vol. 62, no. 8, pp. 2370–2377, Apr. 2002.

[21] J.-C. Soria, M. Rodriguez, D. D. Liu, J. J. Lee, W. K. Hong, and L. Mao, "Aberrant promoter methylation of multiple genes in bronchial brush samples from former cigarette smokers," *Cancer Res.*, vol. 62, no. 2, pp. 351–355, Jan. 2002.

[22] S. Zöchbauer-Müller, S. Lam, S. Toyooka, A. K. Virmani, K. O. Toyooka, S. Seidl, J. D. Minna, and A. F. Gazdar, "Aberrant methylation of multiple genes in the upper aerodigestive tract epithelium of heavy smokers," *Int. J. Cancer*, vol. 107, no. 4, pp. 612–616, Nov. 2003.

[23] I. I. Wistuba and A. F. Gazdar, "Lung cancer preneoplasia," *Annu. Rev. Pathol.*, vol. 1, pp. 331–348, 2006.

[24] I. I. Wistuba, C Behrens, A. K. Virmani, S Milchgrub, S Syed, S Lam, B Mackay, J. D. Minna, and A. F. Gazdar, "Allelic losses at chromosome 8p21-23 are early and frequent events in the pathogenesis of lung cancer," *Cancer Res.*, vol. 59, no. 8, pp. 1973–1979, Apr. 1999.

[25] S. A. Belinsky, K. J. Nikula, W. A. Palmisano, R Michels, G Saccomanno, E Gabrielson, S. B. Baylin, and J. G. Herman, "Aberrant methylation of p16INK4a is an early event in lung cancer and a potential biomarker for early diagnosis," *Proceedings of the National Academy of Sciences*, vol. 95, no. 20, pp. 11 891–11 896, 1998.

[26] A. T. Ooi, A. C. Gower, K. X. Zhang, J. L. Vick, L. Hong, B. Nagao, W. D. Wallace, D. A. Elashoff, T. C. Walser, S. M. Dubinett, M. Pellegrini, M. E. Lenburg, A. Spira, and B. N. Gomperts, "Molecular profiling of premalignant lesions in lung squamous cell carcinomas identifies mechanisms involved in stepwise carcinogenesis," *Cancer Prev. Res.*, vol. 7, no. 5, pp. 487–495, May 2014.

[27] D. Xiong, J. Pan, Q. Zhang, E. Szabo, M. S. Miller, R. A. Lubet, M. You, and Y. Wang, "Bronchial airway gene expression signatures in mouse lung squamous cell carcinoma and their modulation by cancer chemopreventive agents," *Oncotarget*, vol. 8, no. 12, pp. 18 885–18 900, Mar. 2017.

[28] J. Beane, S. A. Mazzilli, J. D. Campbell, G. Duclos, K. Krysan, C. Moy, C. Perdomo, M. Schaffer, G. Liu, S. Zhang, H. Liu, J. Vick, S. S. Dhillon, S. J. Platero, S. M. Dubinett, C. Stevenson, M. E. Reid, M. E. Lenburg, and A. E. Spira, "Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions," Sep. 2018.

[29] Y. Yatabe, A. C. Borczuk, and C. A. Powell, "Do all lung adenocarcinomas follow a stepwise progression?" *Lung Cancer*, vol. 74, no. 1, pp. 7–11, Oct. 2011.

[30] W. H. Westra, I. O. Baas, R. H. Hruban, F. B. Askin, K Wilson, G. J. Offerhaus, and R. J. Slebos, "K-ras oncogene activation in atypical alveolar hyperplasias of the human lung," *Cancer Res.*, vol. 56, no. 9, pp. 2224–2228, May 1996.

[31] W. H. Westra, "Early glandular neoplasia of the lung," *Respir. Res.*, vol. 1, no. 3, 2000.

[32] H Sakamoto, J Shimizu, Y Horio, R Ueda, T Takahashi, T Mitsudomi, and Y Yatabe, "Disproportionate representation of KRAS gene mutation in atypical adenomatous

hyperplasia, but even distribution of EGFR gene mutation from preinvasive to invasive adenocarcinomas," *J. Pathol.*, vol. 212, no. 3, pp. 287–294, Jul. 2007.

[33] K. Takamochi, T. Ogura, K. Suzuki, H. Kawasaki, Y. Kurashima, T. Yokose, A. Ochiai, K. Nagai, Y. Nishiwaki, and H. Esumi, "Loss of heterozygosity on chromosomes 9q and 16p in atypical adenomatous hyperplasia concomitant with adenocarcinoma of the lung," *Am. J. Pathol.*, vol. 159, no. 5, pp. 1941–1948, 2001.

[34] E. a. Kitaguchi S, *Proliferative activity, p53 expression and loss of heterozygosity on 3p, 9p and 17p in atypical adenomatous hyperplasia of the lung. - PubMed - NCBI*, `https://www.ncbi.nlm.nih.gov/pubmed/9583279`, Accessed: 2019-1-27.

[35] R. M. S. Amin, K. Hiroshima, A. Iyoda, K. Hoshi, K. Honma, M. Kuroki, T. Kokubo, T. Fujisawa, Y. Miyagi, and Y. Nakatani, "LKB1 protein expression in neuroendocrine tumors of the lung," *Pathol. Int.*, vol. 58, no. 2, pp. 84–88, 2008.

[36] K. Nakanishi, T. Kawai, F. Kumaki, S. Hiroi, M. Mukai, and E. Ikeda, "Survivin expression in atypical adenomatous hyperplasia of the lung," *Am. J. Clin. Pathol.*, vol. 120, no. 5, pp. 712–719, Nov. 2003.

[37] M. Tominaga, N. Sueoka, K. Irie, K. Iwanaga, O. Tokunaga, S.-I. Hayashi, K. Nakachi, and E. Sueoka, "Detection and discrimination of preneoplastic and early stages of lung adenocarcinoma using hnRNP B1 combined with the cell cycle-related markers p16, cyclin d1, and ki-67," *Lung Cancer*, vol. 40, no. 1, pp. 45–53, 2003.

[38] Y. Yatabe, T. Mitsudomi, and T. Takahashi, "TTF-1 expression in pulmonary adenocarcinomas," *Am. J. Surg. Pathol.*, vol. 26, no. 6, pp. 767–773, 2002.

[39] S. A. Selamat, J. S. Galler, A. D. Joshi, M Nicky Fyfe, M. Campan, K. D. Siegmund, K. M. Kerr, and I. A. Laird-Offringa, "DNA methylation changes in atypical adenomatous hyperplasia, adenocarcinoma in situ, and lung adenocarcinoma," *PLoS One*, vol. 6, no. 6, e21443, 2011.

[40] E. Izumchenko, X. Chang, M. Brait, E. Fertig, L. T. Kagohara, A. Bedi, L. Marchionni, N. Agrawal, R. Ravi, S. Jones, M. O. Hoque, W. H. Westra, and D. Sidransky, "Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA," *Nat. Commun.*, vol. 6, no. 1, 2015.

[41] W. D. Travis, E. Brambilla, A. G. Nicholson, Y. Yatabe, J. H. M. Austin, M. B. Beasley, L. R. Chirieac, S. Dacic, E. Duhig, D. B. Flieder, K. Geisinger, F. R. Hirsch, Y. Ishikawa, K. M. Kerr, M. Noguchi, G. Pelosi, C. A. Powell, M. S. Tsao, I. Wistuba, and WHO Panel, "The 2015 world health organization classification of lung tumors: Impact of genetic, clinical and radiologic advances since the 2004 classification," *J. Thorac. Oncol.*, vol. 10, no. 9, pp. 1243–1260, Sep. 2015.

[42] B Vogelstein, N Papadopoulos, V. E. Velculescu, S Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[43] W. Feng, S. Zhao, D. Xue, F. Song, Z. Li, D. Chen, B. He, Y. Hao, Y. Wang, and Y. Liu, "Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies," *BMC Genomics*, vol. 17 Suppl 7, p. 521, Aug. 2016.

[44] S. Sivakumar, F Anthony San Lucas, T. L. McDowell, W. Lang, L. Xu, J. Fujimoto, J. Zhang, P Andrew Futreal, J. Fukuoka, Y. Yatabe, S. M. Dubinett, A. E. Spira, J. Fowler, E. T. Hawk, I. I. Wistuba, P. Scheet, and H. Kadara, "Genomic landscape of atypical adenomatous hyperplasia reveals divergent modes to lung adenocarcinoma," *Cancer Res.*, vol. 77, no. 22, pp. 6119–6130, 2017.

[45] Y. Xing, T. Yu, Y. N. Wu, M. Roy, J. Kim, and C. Lee, "An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs," *Nucleic Acids Res.*, vol. 34, no. 10, pp. 3150–3160, Jun. 2006.

[46] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, e47, Apr. 2015.

[47] T. Li, J. Fan, B. Wang, N. Traugh, Q. Chen, J. S. Liu, B. Li, and X. S. Liu, "TIMER: A web server for comprehensive analysis of Tumor-Infiltrating immune cells," *Cancer Res.*, vol. 77, no. 21, e108–e110, Nov. 2017.

[48] S. Vattathil and P. Scheet, "Haplotype-based profiling of subtle allelic imbalance with SNP arrays," *Genome Res.*, vol. 23, no. 1, pp. 152–158, Jan. 2013.

[49] Y. A. Jakubek, F. A. San Lucas, and P. Scheet, "Directional allelic imbalance profiling and visualization from multi-sample data with RECUR," *Bioinformatics*, Nov. 2018.

[50] E. Paradis and K. Schliep, "Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R," *Bioinformatics*, vol. 35, no. 3, pp. 526–528, Feb. 2019.

[51] C. V. Dang, "MYC on the path to cancer," *Cell*, vol. 149, no. 1, pp. 22–35, Mar. 2012.

[52] T. F. Gajewski, H. Schreiber, and Y.-X. Fu, "Innate and adaptive immune cells in the tumor microenvironment," *Nat. Immunol.*, vol. 14, no. 10, pp. 1014–1022, 2013.

[53] M. L. Disis, "Immune regulation of cancer," *J. Clin. Orthod.*, vol. 28, no. 29, pp. 4531–4538, Oct. 2010.

[54] S Tugues, S. H. Burkhard, I Ohs, M Vrohlings, K Nussbaum, J vom Berg, P Kulig, and B Becher, "New insights into IL-12-mediated tumor suppression," *Cell Death Differ.*, vol. 22, no. 2, pp. 237–246, 2014.

[55] J. Lu, M. Chatterjee, H. Schmid, S. Beck, and M. Gawaz, "CXCL14 as an emerging immune and inflammatory modulator," *J. Inflamm.*, vol. 13, p. 1, Jan. 2016.

[56] J Panse, K Friedrichs, A Marx, Y Hildebrandt, T Luetkens, K Barrels, C Horn, T Stahl, Y Cao, K Milde-Langosch, A Niendorf, N Kröger, S Wenzel, R Leuwer, C Bokemeyer, S Hegewisch-Becker, and D Atanackovic, "Chemokine CXCL13 is overexpressed in the tumour tissue and in the peripheral blood of breast cancer patients," *Br. J. Cancer*, vol. 99, no. 6, pp. 930–938, Sep. 2008.

[57] S. Y. Lim, A. E. Yuzhalin, A. N. Gordon-Weeks, and R. J. Muschel, "Targeting the CCL2-CCR2 signaling axis in cancer metastasis," *Oncotarget*, vol. 7, no. 19, pp. 28 697–28 710, May 2016.

[58] J. F. Grosso and M. N. Jure-Kunkel, "CTLA-4 blockade in tumor models: An overview of preclinical and translational research," *Cancer Immun.*, vol. 13, p. 5, Jan. 2013.

[59] M. Afkhami, A. Karunamurthy, S. Chiosea, M. N. Nikiforova, R. Seethala, Y. E. Nikiforov, and C. Coyne, "Histopathologic and clinical characterization of thyroid tumors carrying the BRAF(K601E) mutation," *Thyroid*, vol. 26, no. 2, pp. 242–247, Feb. 2016.

[60] E. Macerola, L. Torregrossa, C. Ugolini, S. Bakkar, P. Vitti, G. Fadda, and F. Basolo, "BRAFK601E mutation in a follicular thyroid adenoma," *Int. J. Surg. Pathol.*, p. 106 689 691 668 808, 2017.

[61] A. H. Shain, I. Yeh, I. Kovalyshyn, A. Sriharan, E. Talevich, A. Gagnon, R. Dummer, J. North, L. Pincus, B. Ruben, W. Rickaby, C. D'Arrigo, A. Robson, and B. C. Bastian, "The genetic evolution of melanoma from precursor lesions," *N. Engl. J. Med.*, vol. 373, no. 20, pp. 1926–1936, Nov. 2015.

[62] G. Zheng, L.-H. Tseng, G. Chen, L. Haley, P. Illei, C. D. Gocke, J. R. Eshleman, and M.-T. Lin, "Clinical detection and categorization of uncommon and concomitant mutations involving BRAF," *BMC Cancer*, vol. 15, p. 779, Oct. 2015.

[63] H Kadara, M Choi, J Zhang, E. R. Parra, J Rodriguez-Canales, S. G. Gaffney, Z Zhao, C Behrens, J Fujimoto, C Chow, Y Yoo, N Kalhor, C Moran, D Rimm, S Swisher, D. L. Gibbons, J Heymach, E Kaftan, J. P. Townsend, T. J. Lynch, J Schlessinger, J Lee, R. P. Lifton, I. I. Wistuba, and R. S. Herbst, "Whole-exome sequencing and immune profiling of early-stage lung adenocarcinoma with fully annotated clinical follow-up," *Ann. Oncol.*, Mar. 2017.

[64] E. Borras, F. A. San Lucas, K. Chang, R. Zhou, G. Masand, J. Fowler, M. E. Mork, Y. N. You, M. W. Taggart, F. McAllister, D. A. Jones, G. E. Davies, W. Edelmann, E. A. Ehli, P. M. Lynch, E. T. Hawk, G. Capella, P. Scheet, and E. Vilar, "Genomic landscape of colorectal mucosa and adenomas," *Cancer Prev. Res.*, vol. 9, no. 6, pp. 417–427, Jun. 2016.

[65] B. R. Balsara and J. R. Testa, "Chromosomal imbalances in human lung cancer," *Oncogene*, vol. 21, no. 45, p. 6877, Oct. 2002.

[66] R. Nahar, W. Zhai, T. Zhang, A. Takano, A. J. Khng, Y. Y. Lee, X. Liu, C. H. Lim, T. P. T. Koh, Z. W. Aung, T. K. H. Lim, L. Veeravalli, J. Yuan, A. S. M. Teo, C. X. Chan, H. M. Poh, I. M. L. Chua, A. A. Liew, D. P. X. Lau, X. L. Kwang, C. K. Toh, W.-T. Lim, B. Lim, W. L. Tam, E.-H. Tan, A. M. Hillmer, and D. S. W. Tan, "Elucidating the genomic architecture of asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing," *Nat. Commun.*, vol. 9, no. 1, p. 216, Jan. 2018.

[67] F.-Y. Lo, J.-W. Chang, I.-S. Chang, Y.-J. Chen, H.-S. Hsu, S.-F. K. Huang, F.-Y. Tsai, S. S. Jiang, R. Kanteti, S. Nandi, R. Salgia, and Y.-C. Wang, "The database of chromosome imbalance regions and genes resided in lung cancer from asian and caucasian identified by array-comparative genomic hybridization," *BMC Cancer*, vol. 12, p. 235, Jun. 2012.

[68] C. Vinayanuwattikun, F. Le Calvez-Kelm, B. Abedi-Ardekani, D. Zaridze, A. Mukeria, C. Voegele, M. Vallée, D. Purnomosari, N. Forey, G. Durand, G. Byrnes, J. Mckay, P. Brennan, and G. Scelo, "Elucidating genomic characteristics of lung cancer progression from in situ to invasive adenocarcinoma," *Sci. Rep.*, vol. 6, p. 31 628, Aug. 2016.

[69] X.-J. Ding, M.-X. Liu, L. Ao, Y.-R. Liang, and Y. Cao, "Frequent loss of heterozygosity on chromosome 12q in non-small-cell lung carcinomas," *Virchows Arch.*, vol. 458, no. 5, pp. 561–569, May 2011.

[70] E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Grönroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S.-M. Lee, M. Forster, T. Ahmad, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefe, S.-C. Chen, S. Begum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell, and C. Swanton, "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution," *Science*, vol. 346, no. 6206, pp. 251–256, Oct. 2014.

[71] M. T. Barrett, C. A. Sanchez, L. J. Prevo, D. J. Wong, P. C. Galipeau, T. G. Paulson, P. S. Rabinovitch, and B. J. Reid, "Evolution of neoplastic cell lineages in barrett oesophagus," *Nat. Genet.*, vol. 22, no. 1, pp. 106–109, May 1999.

[72] X Li, P. C. Galipeau, C. A. Sanchez, P. L. Blount, C. C. Maley, J Arnaudo, D. A. Peiffer, D Pokholok, K. L. Gunderson, and B. J. Reid, "Single nucleotide Polymorphism-Based Genome-Wide chromosome copy change, loss of heterozygosity, and aneuploidy in barrett's esophagus neoplastic progression," *Cancer Prev. Res.*, vol. 1, no. 6, pp. 413–423, 2008.

[73] B. J. Reid, L. J. Prevo, P. C. Galipeau, C. A. Sanchez, G. Longton, D. S. Levine, P. L. Blount, and P. S. Rabinovitch, "Predictors of progression in barrett's esophagus II: Baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased

risk for neoplastic progression," *Am. J. Gastroenterol.*, vol. 96, no. 10, pp. 2839–2848, 2001.

[74] K. Dolan, A. I. Morris, J. R. Gosney, J. K. Field, and R. Sutton, "Loss of heterozygosity on chromosome 17p predicts neoplastic progression in barrett's esophagus," *J. Gastroenterol. Hepatol.*, vol. 18, no. 6, pp. 683–689, 2003.

[75] M. P. Rosin, X Cheng, C Poh, W. L. Lam, Y Huang, J Lovas, K Berean, J. B. Epstein, R Priddy, N. D. Le, and L Zhang, "Use of allelic loss to predict malignant risk for low-grade oral epithelial dysplasia," *Clin. Cancer Res.*, vol. 6, no. 2, pp. 357–362, Feb. 2000.

[76] K. Ishida, S. Ito, N. Wada, H. Deguchi, T. Hata, M. Hosoda, and T. Nohno, "Nuclear localization of beta-catenin involved in precancerous change in oral leukoplakia," *Mol. Cancer*, vol. 6, p. 62, Oct. 2007.

[77] D. T. Merrick, J. Kittelson, R. Winterhalder, G. Kotantoulas, S. Ingeberg, R. L. Keith, T. C. Kennedy, Y. E. Miller, W. A. Franklin, and F. R. Hirsch, "Analysis of c-ErbB1/Epidermal growth factor receptor and c-ErbB2/HER-2 expression in bronchial dysplasia: Evaluation of potential targets for chemoprevention of lung cancer," *Clin. Cancer Res.*, vol. 12, no. 7, pp. 2281–2288, Apr. 2006.

[78] H Imaseki, H Hayashi, M Taira, Y Ito, Y Tabata, S Onoda, K Isono, and M Tatibana, "Expression of c-myc oncogene in colorectal polyps as a biological marker for monitoring malignant potential," *Cancer*, vol. 64, no. 3, pp. 704–709, Aug. 1989.

[79] M. E. Kavanagh, M. J. Conroy, N. E. Clarke, N. T. Gilmartin, K. E. O'Sullivan, R. Feighery, F. MacCarthy, D. O'Toole, N. Ravi, J. V. Reynolds, J. O'Sullivan, and J. Lysaght, "Impact of the inflammatory microenvironment on t-cell phenotype in the progression from reflux oesophagitis to barrett oesophagus and oesophageal adenocarcinoma," *Cancer Lett.*, vol. 370, no. 1, pp. 117–124, Jan. 2016.

[80] H. Jung, L. Ertl, C. Janson, T. Schall, and I. Charo, "Abstract a107: Inhibition of CCR2 potentiates the checkpoint inhibitor immunotherapy in pancreatic cancer," *Cancer Immunology Research*, vol. 4, no. 11 Supplement, A107–A107, 2016.

[81] P.-L. Chang, L. Harkins, Y.-H. Hsieh, P. Hicks, K. Sappayatosok, S. Yodsanga, S. Swasdison, A. F. Chambers, C. A. Elmets, and K.-J. Ho, "Osteopontin expression in normal skin and non-melanoma skin tumors," *J. Histochem. Cytochem.*, vol. 56, no. 1, p. 57, Jan. 2008.

[82] K. Chang, M. W. Taggart, L. Reyes-Uribe, E. Borras, E. Riquelme, R. M. Barnett, G. Leoni, F. A. San Lucas, M. T. Catanese, F. Mori, M. G. Diodoro, Y. N. You, E. T. Hawk, J. Roszik, P. Scheet, S. Kopetz, A. Nicosia, E. Scarselli, P. M. Lynch, F. McAllister, and E. Vilar, "Immune profiling of premalignant lesions in patients with lynch syndrome," *JAMA Oncol*, vol. 4, no. 8, pp. 1085–1092, Aug. 2018.

[83] W. McCaskill-Stevens, D. C. Pearson, B. S. Kramer, L. G. Ford, and S. M. Lippman, "Identifying and creating the next generation of Community-Based cancer prevention studies: Summary of a national cancer institute think tank," *Cancer Prev. Res.*, vol. 10, no. 2, pp. 99–107, Feb. 2017.

[84] D. Dankort, E. Filenova, M. Collado, M. Serrano, K. Jones, and M. McMahon, "A new mouse model to explore the initiation, progression, and therapy of BRAFV600E-induced lung tumors," *Genes Dev.*, vol. 21, no. 4, pp. 379–384, Feb. 2007.

[85] V. G. Krishnan, P. J. Ebert, J. C. Ting, E. Lim, S.-S. Wong, A. S. M. Teo, Y. G. Yue, H.-H. Chua, X. Ma, G. S. L. Loh, Y. Lin, J. H. J. Tan, K. Yu, S. Zhang, C. Reinhard, D. S. W. Tan, B. A. Peters, S. E. Lincoln, D. G. Ballinger, J. M. Laramie, G. B. Nilsen, T. D. Barber, P. Tan, A. M. Hillmer, and P. C. Ng, "Whole-genome sequencing of asian lung cancers: Second-hand smoke unlikely to be responsible for higher incidence of lung cancer among asian never-smokers," *Cancer Res.*, vol. 74, no. 21, pp. 6071–6081, Nov. 2014.

[86] R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, and C. Swanton, "DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution," *Genome Biol.*, vol. 17, p. 31, Feb. 2016.

[87] J. Zhang, J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C.-W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, I. I. Wistuba, and P. A. Futreal, "Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing," *Science*, vol. 346, no. 6206, pp. 256–259, Oct. 2014.

[88] "Evolutionary dynamics in pre-invasive neoplasia," *Current Opinion in Systems Biology*, vol. 2, pp. 1–8, Apr. 2017.

[89] A. M. Taylor, J. Shih, G. Ha, G. F. Gao, X. Zhang, A. C. Berger, S. E. Schumacher, C. Wang, H. Hu, J. Liu, A. J. Lazar, Cancer Genome Atlas Research Network, A. D. Cherniack, R. Beroukhim, and M. Meyerson, "Genomic and functional approaches to understanding cancer aneuploidy," *Cancer Cell*, vol. 33, no. 4, 676–689.e3, Apr. 2018.

[90] K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, Cancer Genome Atlas Network, J. M. Stuart, C. C. Benz, and P. W. Laird, "Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, 291–304.e6, Apr. 2018.

[91] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B.

Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim, "Pan-cancer patterns of somatic copy number alteration," *Nat. Genet.*, vol. 45, no. 10, pp. 1134–1140, Oct. 2013.

[92] Y. Liu, C. Chen, Z. Xu, C. Scuoppo, C. D. Rillahan, J. Gao, B. Spitzer, B. Bosbach, E. R. Kastenhuber, T. Baslan, S. Ackermann, L. Cheng, Q. Wang, T. Niu, N. Schultz, R. L. Levine, A. A. Mills, and S. W. Lowe, "Deletions linked to TP53 loss drive cancer through p53-independent mechanisms," *Nature*, vol. 531, no. 7595, pp. 471–475, Mar. 2016.

[93] Cancer Genome Atlas Research Network, W. M. Linehan, P. T. Spellman, C. J. Ricketts, C. J. Creighton, S. S. Fei, C. Davis, D. A. Wheeler, B. A. Murray, L. Schmidt, C. D. Vocke, M. Peto, A. A. M. Al Mamun, E. Shinbrot, A. Sethi, S. Brooks, W. K. Rathmell, A. N. Brooks, K. A. Hoadley, A. G. Robertson, D. Brooks, R. Bowlby, S. Sadeghi, H. Shen, D. J. Weisenberger, M. Bootwalla, S. B. Baylin, P. W. Laird, A. D. Cherniack, G. Saksena, S. Haake, J. Li, H. Liang, Y. Lu, G. B. Mills, R. Akbani, M. D. M. Leiserson, B. J. Raphael, P. Anur, D. Bottaro, L. Albiges, N. Barnabas, T. K. Choueiri, B. Czerniak, A. K. Godwin, A. A. Hakimi, T. H. Ho, J. Hsieh, M. Ittmann, W. Y. Kim, B. Krishnan, M. J. Merino, K. R. Mills Shaw, V. E. Reuter, E. Reznik, C. S. Shelley, B. Shuch, S. Signoretti, R. Srinivasan, P. Tamboli, G. Thomas, S. Tickoo, K. Burnett, D. Crain, J. Gardner, K. Lau, D. Mallery, S. Morris, J. D. Paulauskis, R. J. Penny, C. Shelton, W. T. Shelton, M. Sherman, E. Thompson, P. Yena, M. T. Avedon, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, T. Santos, L. Wise, E. Zmuda, J. A. Demchok, I. Felau, C. M. Hutter, M. Sheth, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, B. Ayala, J. Baboud, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, A. Ally, M. Balasundaram, S. Balu, R. Beroukhim, T. Bodenheimer, C. Buhay, Y. S. N. Butterfield, R. Carlsen, S. L. Carter, H. Chao, E. Chuah, A. Clarke, K. R. Covington, M. Dahdouli, N. Dewal, N. Dhalla, H. V. Doddapaneni, J. A. Drummond, S. B. Gabriel, R. A. Gibbs, R. Guin, W. Hale, A. Hawes, D. N. Hayes, R. A. Holt, A. P. Hoyle, S. R. Jefferys, S. J. M. Jones, C. D. Jones, D. Kalra, C. Kovar, L. Lewis, J. Li, Y. Ma, M. A. Marra, M. Mayo, S. Meng, M. Meyerson, P. A. Mieczkowski, R. A. Moore, D. Morton, L. E. Mose, A. J. Mungall, D. Muzny, J. S. Parker, C. M. Perou, J. Roach, J. E. Schein, S. E. Schumacher, Y. Shi, J. V. Simons, P. Sipahimalani, T. Skelly, M. G. Soloway, C. Sougnez, A. Tam, D. Tan, N. Thiessen, U. Veluvolu, M. Wang, M. D. Wilkerson, T. Wong, J. Wu, L. Xi, J. Zhou, J. Bedford, F. Chen, Y. Fu, M. Gerstein, D. Haussler, K. Kasaian, P. Lai, S. Ling, A. Radenbaugh, D. Van Den Berg, J. N. Weinstein, J. Zhu, M. Albert, I. Alexopoulou, J. J. Andersen, J. T. Auman, J. Bartlett, S. Bastacky, J. Bergsten, M. L. Blute, L. Boice, R. J. Bollag, J. Boyd, E. Castle, Y.-B. Chen, J. C. Cheville, E. Curley, B. Davies, A. DeVolk, R. Dhir, L. Dike, J. Eckman, J. Engel, J. Harr, R. Hrebinko, M. Huang, L. Huelsenbeck-Dill, M. Iacocca, B. Jacobs, M. Lobis, J. K. Maranchie, S. McMeekin, J. Myers, J. Nelson, J. Parfitt, A. Parwani, N. Petrelli, B. Rabeno, S. Roy, A. L. Salner, J. Slaton, M. Stanton, R. H. Thompson, L. Thorne, K. Tucker, P. M. Weinberger, C. Winemiller, L. A. Zach, and R. Zuna, "Comprehensive molecular characterization of papillary Renal-Cell carcinoma," *N. Engl. J. Med.*, vol. 374, no. 2, pp. 135–145, Jan. 2016.

[94] A. D. Cherniack, H. Shen, V. Walter, C. Stewart, B. A. Murray, R. Bowlby, X. Hu, S. Ling, R. A. Soslow, R. R. Broaddus, R. E. Zuna, G. Robertson, P. W. Laird, R. Kucherlapati, G. B. Mills, Cancer Genome Atlas Research Network, J. N. Weinstein, J. Zhang, R. Akbani, and D. A. Levine, "Integrated molecular characterization of uterine carcinosarcoma," *Cancer Cell*, vol. 31, no. 3, pp. 411–423, Mar. 2017.

[95] Cancer Genome Atlas Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, no. 7536, pp. 576–582, Jan. 2015.

[96] M. Labussière, A. Rahimian, M. Giry, B. Boisselier, Y. Schmitt, M. Polivka, K. Mokhtari, J.-Y. Delattre, A. Idbaih, K. Labreche, A. Alentorn, and M. Sanson, "Chromosome 17p homodisomy is associated with better outcome in 1p19q Non-Codeleted and IDH-Mutated gliomas," *Oncologist*, vol. 21, no. 9, pp. 1131–1135, Sep. 2016.

[97] A. Idbaih, F. Ducray, C. Dehais, C. Courdy, C. Carpentier, S. de Bernard, E. Uro-Coste, K. Mokhtari, A. Jouvet, J. Honnorat, O. Chinot, C. Ramirez, P. Beauchesne, A. Benouaich-Amiel, J. Godard, S. Eimer, F. Parker, E. Lechapt-Zalcman, P. Colin, D. Loussouarn, T. Faillot, P. Dam-Hieu, S. Elouadhani-Hamdi, L. Bauchet, O. Langlois, C. Le Guerinel, D. Fontaine, E. Vauleon, P. Menei, M. J. M. Fotso, C. Desenclos, P. Verrelle, F. Ghiringhelli, G. Noel, F. Labrousse, A. Carpentier, F. Dhermain, J.-Y. Delattre, D. Figarella-Branger, and POLA Network, "SNP array analysis reveals novel genomic abnormalities including copy neutral loss of heterozygosity in anaplastic oligodendrogliomas," *PLoS One*, vol. 7, no. 10, e45950, Oct. 2012.

[98] H. Suzuki, K. Aoki, K. Chiba, Y. Sato, Y. Shiozawa, Y. Shiraishi, T. Shimamura, A. Niida, K. Motomura, F. Ohka, T. Yamamoto, K. Tanahashi, M. Ranjit, T. Wakabayashi, T. Yoshizato, K. Kataoka, K. Yoshida, Y. Nagata, A. Sato-Otsubo, H. Tanaka, M. Sanada, Y. Kondo, H. Nakamura, M. Mizoguchi, T. Abe, Y. Muragaki, R. Watanabe, I. Ito, S. Miyano, A. Natsume, and S. Ogawa, "Mutational landscape and clonal architecture in grade II and III gliomas," *Nat. Genet.*, vol. 47, no. 5, p. 458, Apr. 2015.

[99] Y Harima, K Harima, S Sawada, Y Tanaka, S Arita, and T Ohnishi, "Loss of heterozygosity on chromosome 6p21.2 as a potential marker for recurrence after radiotherapy of human cervical cancer," *Clin. Cancer Res.*, vol. 6, no. 3, pp. 1079–1085, Mar. 2000.

[100] B Sadikovic, K Al-Romaih, J. A. Squire, and M Zielenska, "Cause and consequences of genetic and epigenetic alterations in human cancer," *Curr. Genomics*, vol. 9, no. 6, pp. 394–408, Sep. 2008.

[101] K. Krysan, L. M. Tran, B. S. Grimes, T. C. Walser, W. D. Wallace, and S. M. Dubinett, "Abstract 1016: Evaluation of progression associated neoepitopes and immune contexture in pulmonary premalignancy," *Cancer Res.*, vol. 77, no. 13 Supplement, pp. 1016–1016, Jul. 2017.

[102] K. L. Knutson and M. L. Disis, "Tumor antigen-specific T helper cells in cancer immunity and immunotherapy," *Cancer Immunol. Immunother.*, vol. 54, no. 8, pp. 721–728, 2005.

[103] A. Spira, M. L. Disis, J. T. Schiller, E. Vilar, T. R. Rebbeck, R. Bejar, T. Ideker, J. Arts, M. B. Yurgelun, J. P. Mesirov, A. Rao, J. Garber, E. M. Jaffee, and S. M. Lippman, "Leveraging premalignant biology for immune-based cancer prevention," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 39, pp. 10 750–10 758, Sep. 2016.

[104] M. R. I. Young, "Redirecting the focus of cancer immunotherapy to premalignant conditions," *Cancer Lett.*, vol. 391, pp. 83–88, Apr. 2017.

[105] N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, and T. A. Chan, "Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer," *Science*, vol. 348, no. 6230, pp. 124–128, Apr. 2015.

[106] I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M. R. Stratton, P. H. Jones, and P. J. Campbell, "Tumor evolution. high burden and pervasive positive selection of somatic mutations in normal human skin," *Science*, vol. 348, no. 6237, pp. 880–886, May 2015.

[107] I. Martincorena, J. C. Fowler, A. Wabik, A. R. J. Lawson, F. Abascal, M. W. J. Hall, A. Cagan, K. Murai, K. Mahbubani, M. R. Stratton, R. C. Fitzgerald, P. A. Handford, P. J. Campbell, K. Saeb-Parsy, and P. H. Jones, "Somatic mutant clones colonize the human esophagus with age," *Science*, Oct. 2018.

[108] S. Srivastava, S. Ghosh, J. Kagan, R. Mazurchuk, and National Cancer Institute's HTAN Implementation, "The making of a PreCancer atlas: Promises, challenges, and opportunities," *Trends Cancer Res.*, vol. 4, no. 8, pp. 523–536, Aug. 2018.

VITA

Smruthy Sivakumar is the daughter of Anuradha and K V Sivakumar and the wife of Varun Rao. She was born in Chennai, Tamil Nadu, India. She grew up in Bangalore, Karnataka, India and graduated from National Public School (Rajajinagar) in 2008. She received a Bachelors in Engineering degree with a major in Biotechnology from the PES Institute of Technology, Bangalore in 2012. During her undergraduate training, she pursued internship opportunities, including brief appointments at Philips Healthcare and GVN Institute of Oncology. She moved to the United States in 2012, and received a Master of Science degree in Bioinformatics at the Georgia Institute of Technology, Atlanta, Georgia in 2013. She briefly worked as a Bioinformatics Scientist at Dow AgroSciences before moving to Houston, TX. In August 2014, she joined the MD Anderson UTHealth Graduate School of Biomedical Sciences to pursue her doctoral degree. Smruthy's research interests are in applying bioinformatic approaches to aid clinical and translational cancer research. She expects to receive her Doctor of Philosophy Degree in Bioinformatics from the MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.