Title
The immune contexture and genomic landscape of lung adenomatous premalignancy

Author
Grimes, Brandon Scott

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The immune contexture and genomic landscape of lung adenomatous premalignancy

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Clinical Research

by

Brandon Scott Grimes

2017

ABSTRACT OF THE THESIS

The immune contexture and genomic landscape of lung adenomatous premalignancy

by

Brandon Scott Grimes

Master of Science in Clinical Research

University of California, Los Angeles, 2017

Professor Robert M. Elashoff, Chair

Rationale/Objective: A better understanding of genomic alterations and the tumor microenvironment along the spectrum of early disease could lead to identification of targetable neoantigens that may form the basis of future interceptive therapies.

Methods: From 41 lobectomy specimens for early stage lung adenocarcinoma, laser capture microdissection was utilized to obtain areas of atypical adenomatous hyperplasia (AAH), adenocarcinoma *in situ* (AIS), normal epithelium and the associated ADC for WES. Quantitative IHC assessed the immune microenvironment. Putative neoantigens were identified from somatic mutations by prediction of avid binding to patients' HLA.

Results: Non-synonymous (ns) somatic mutations were termed progression-associated mutations (PAMs), located in both premalignant lesions and ADC. Neoepitopes derived from PAMs, referred to as progression-associated neoepitopes (PAN), are associated with the highest levels of CD8+ T cell infiltration and PD-L1 expression. Immune-effector cell infiltration and accompanying adaptive immune suppression suggest specific immune antigen recognition.

The thesis of Brandon Scott Grimes is approved.

Steven M. Dubinett

David Elashoff

Robert M. Elashoff, Committee Chair

University of California, Los Angeles

2017

TABLE OF CONTENTS

The immune contexture and genomic landscape of lung adenomatous premalignancy

Brandon S. Grimes*[1], Linh Tran*[1,5], Kostyantyn Krysan*[1,5], Gregory Fishbein[2], Michael Fishbein[2], Dean Wallace[2], David Elashoff[1,5,6], Steven M. Dubinett[1,2,4,5,7]

[1]Division of Pulmonary and Critical Care Medicine, Departments of Medicine, [2]Pathology and Laboratory Medicine, [3]Urology and [4]Molecular and Medical Pharmacology, David Geffen School of Medicine at UCLA; [5]Jonsson Comprehensive Cancer Center, Los Angeles, California; [6]Department of Biostatistics at UCLA, Los Angeles, California; [7]VA Greater Los Angeles Health Care Center, Los Angeles, California.

*Equal Contribution

Introduction

Lung cancer is the world's leading cause of cancer death with adenocarcinoma (ADC) as the leading subtype[1]. One of the major driving forces of carcinogenesis is somatic mutagenesis. Over 75% of lung cancers bear driver mutations that are causally implicated in cancer development, while the remainder of lung cancers does not bear mutations in known oncogenes or tumor suppressors[2]. Despite advances in new targeted therapies, the overall 5 year survival for lung cancer has improved by less than 5% in the past thirty years. It has been suggested that a profound impact in survival could be achieved through early diagnosis and intervention before premalignant lesions advance to invasive lung cancer. Studies indicate that lung cancer develops through progressive pathologic changes evident as premalignant lesions. Understanding pulmonary premalignancy and the determinants of progression to invasive disease can facilitate those detection and prevention strategies that could reduce lung cancer mortality.

Atypical adenomatous hyperplasias (AAH), small focal proliferative lesions that can be found in the distal airways of patients with lung cancer as well as those at risk, are thought to be the earliest premalignant lesion in the progression from normal airway epithelium to ADC[3,4]. Despite knowledge regarding the histology of these lesions, highly effective lung cancer early detection and chemoprevention await definition of molecular targets and elucidation of disease pathogenesis.

Early attempts to evaluate somatic mutations in premalignant pulmonary lesions revealed mutations in known driver genes, such as *KRAS*[5], *EGFR*[6,7] and *TP53*[8]. AAH lesions are also known to harbor additional alterations seen in ADC including loss of heterozygosity at 9q and 16p[9]. In accord with the recent revisions in the histological classification for ADC, adenocarcinoma *in situ* (AIS), previously referred to as bronchiolooalveolar carcinoma, are non-

invasive, localized small (<3 cm) adenocarcinomas with growth restricted to neoplastic cells along preexisting alveolar structures[10]. Clonal analysis in previous investigations demonstrated identical monoclonal patterns in adjacent AAH and AIS lesions, strengthening the notion that AAH is a preneoplastic lesion rather than reactive hyperplasia[3]. Recent studies utilizing targeted sequencing of AAH lesions and associated cancers from the same surgical specimens, identified mutations in other cancer-related genes as well as clonality between premalignant lesions and cancer[11]. While the importance of the mutational landscape variations in progression from premalignancy to cancer have been highlighted, the genomic and microenvironmental determinants of progression have not yet been elucidated.

Next–generation sequencing of lung adenocarcinoma has provided insights in disease pathogenesis that informs new treatment strategies. While the majority of lung cancers bearing driver mutations that are causally implicated in lung cancer development[2], passenger mutations may also play roles in disease pathogenesis. Often carcinogen-induced, lung adenocarcinoma has among the highest mutational loads compared to other malignancies. Non-synonymous (n.s.) mutations can give rise to mutant proteins that, when processed, may result in immunogenic peptides, defined as neoepitopes, that avidly bind and are presented in the context of autologous MHC molecules. The recently documented clinical efficacy of checkpoint blockade immunotherapies in non-small cell lung cancer and other malignancies indicates that tumor-specific T cells can recognize neoantigens, resulting in tumor cell death. Could this also be consistent with the theory of immunosurveillance? The concept, as first proposed by Burnet[12], suggests that the host immune response is able to recognize and prevent the outgrowth of invasive cancer cells at the earliest point of development and thus before clinical recognition. While extensive data exists in laboratory models, the clinical evidence for the relevance of immunosurveillance in human lung cancer has not yet been defined; nor is it yet know when an individual's immune system begins to engage in the defense against the disease.

Here, to begin to understand the relationships among the genomic landscape of pulmonary adenomatous premalignancy and associated invasive ADC, we evaluated WES and immune cell infiltration in a cohort of patients following surgical resection. We report on evidence for mutational heterogeneity, pathway dysregulation and apparent immune recognition in pulmonary adenomatous premalignancy.

Results

Cell-mediated immunity and adaptive responses in pulmonary premalignancy.

To evaluate the presence of the early adaptive immune response against pulmonary premalignancy, we first assessed the degree of lymphocyte infiltration in premalignant (n = 328) and malignant lesions (n = 15 AIS and 50 ADC), along with adjacent histologically normal areas (n = 50) in lobectomy specimens from 41 patients who had undergone surgery for early stage ADC. The clinical features for these patients are summarized in **Table S1**. The median number of lesions evaluated per patient was six for AAH, and two for malignant lesions. The degree of lymphocyte infiltration was scaled by a 0-to-3 system (for details see Methods). We found that lymphocyte infiltration was significantly increased in AAH compared to adjacent normal areas (χ^2 test $p < 10^{-16}$), and became highest in AIS and ADC (χ^2 test $p < 10^{-14}$ compared to AAH) (**Figure 1A**). We further assessed the expression of regulators of cell-mediated immunity, including CD4, CD8, FOXP3, PD-1 and PD-L1 in premalignant lesions and ADC by immunohistochemistry (IHC) (**Figure 1B** illustrates immunostaining of the markers in an AAH lesion with a lymphocyte infiltration score of 2). Markers were quantified utilizing the Halo image analysis platform. We found both infiltration of T effector and cytotoxic cells as well as expression of the PD-L1 checkpoint in premalignancy. These findings indicate that cell-mediated immunity and possible recognition of neoepitopes occur within the cellular microenvironment of pulmonary premalignancy.

Immune-related gene signatures are associated with patient outcome in early stage disease in the TCGA lung adenocarcinoma (LUAD) cohort.

To determine if activation of immune pathways was associated with outcomes in early stage LUAD, the expression of immune-related genes in the TCGA LUAD cohort (444 tumors and 58 normal samples) was analyzed. We derived immune signatures based on expression of genes involved in 16 immune-related pathways from the Molecular Signature Database[13]. Gene Set Variation Analysis (GSVA)[14] was utilized to estimate the activities of immune-related pathways in individual patients, and these were then subjected to unsupervised hierarchical cluster analysis to stratify samples. Based on immune-related pathways activity, we identified three major groups (**Figure 1C**). Among them, group 1 (G1, annotated by black) had the highest levels of immune-related gene expression and included 51 tumors and the majority of normal samples (n = 52), whereas the other two groups included the remainder of the tumor samples (χ^2 test $p < 10^{-16}$): G2 (n = 198, blue) with intermediate and G3 (n = 201, red) with lowest expression of immune-related genes. The overall survival in G2 was better than in G3 (log-rank test (LRT) p = 0.063), however, and tumors in those groups were not significantly associated with either tumor stage (χ^2 test p = 0.14 for stage I vs. stage II and higher). Remarkably, the difference in survival between G2 and G3 was prominent for stage I (LRT p = 0.05, **Figure 1D**), but not evident for stage II and higher patients (LRT p = 0.44, **Figure 1E**). Together, these results suggest that the immune-related responses may play a role in patient outcomes, especially at the earliest stage of lung ADC.

Genetic heterogeneity between lesions from the same patients varies over wide range.

To determine if the immune infiltrates observed in pulmonary premalignancy were associated with expression of cognate neoantigens, we performed WES and identified putative neoantigens in 89 AAH, 15 AIS, and 55 ADC lesions from 41 lung cancer patients (**Table S2**). The cells of interest were dissected from the following regions of distal airways utilizing Laser Capture

Microdissection (LCM): **a**) normal airway epithelial cells (1-3 regions per patient), **b**) AAH lesions (2-4 regions per patient), **c**) AIS (1-3 regions per patient where present), and **d**) ADC (1-3 regions per patient). WES was conducted with at least 2 x 10^{10} bases sequenced per exome, which has been frequently achieved in WES studies[15]. The median number of mutations identified per individual region was 351. The median total of all mutations in all regions sequenced per patient was 1323. The mutational load per patient did not increase significantly by the addition of more sequenced regions (Kruskal-Wallis rank sum test $p = 0.20$) (**Figure S1)**. Because multiple regions including normal, premalignant and tumor were sequenced for each patient, we first characterized the heterogeneity and then the genomic relationship among regions sequenced. We utilized the Jaccard index, which measures the similarity in non-synonymous (n.s.) somatic mutations between a pair of lesions, and is inversely proportional to the level of heterogeneity. We found that lesions obtained from within individual patients had significantly higher Jaccard indices and, thus, lower heterogeneity than lesions compared between patients (Kruskal-Wallis rank sum test $p < 10^{-16}$) (**Figure 2A**). By further examining the heterogeneity between regions in individual patients, we found that their indices varied over a wide range (**Figure 2B)**. With the exception of the first four patients (P01 — P04, **Figure 2B)**, individual patients had higher indices (lower heterogeneity) among regions as compared to those from different patients. Thus, in this cohort, each patient most often demonstrated unique n.s. somatic mutations not shared among patients.

Phylogenetic trees were constructed to explore the relationship between sequenced regions for each individual patient (**Figures 2C and S2**). AAH lesions are pathologically classified as adenomatous premalignancy and AIS as non-invasive malignancy ( ). However, their somatic mutations have not been fully examined in the context of the associated ADC to determine the relationship of the histological classification to genomic profiles. AAH, AIS and ADC were all present in 10 out of 41 patients and their mutational profiles were compared to determine the

homology among regions. Phylogenetic trees for these 10 cases are illustrated in **Figure 2C**. Interestingly, in the majority of cases ADC (brown labels) shared common mutations with AIS (orange labels), but not with AAH (blue labels) (**Figure 2C**, top panel). This close relationship was not demonstrated in 3 of the 10 patients that revealed the following: **a**) one case in which one of two primary ADCs in the patient was closely related to the AAH lesions, while another ADC was clustered to AIS, **b**) one case in which ADC had its mutational landscape highly overlapped with that of AAH but not of AIS, and **c**) one case in which mutations in AAH and AIS lesions were closely related to each other, but not to the ADC (**Figure 2C**, bottom panel).

Neoantigens elicit immune responses in pulmonary premalignancy.

We next sought to determine if the immune infiltrates observed in adenomatous premalignancy were associated with the expression of putative cognate neoantigens. The pipeline outlined in **Figure S3** was applied to identify putative neoantigens, which were then classified into three distinguishing categories based on their location in various tissues including ADC, premalignancy or normal epithelium. To determine how n.s. somatic mutations affect tumor development in various stages, we first classified them into three different categories: **a**) premalignant mutations (PrMs) which were observed only in AAH lesions, **b**) progression-associated mutations (PAMs) which were located in both AAH and AIS/ADC lesions, and **c**) malignant-specific mutations (MSMs) which were only identified in AIS/ADC lesions. Recent studies that have focused on intra-tumoral heterogeneity or cancer evolution have classified mutations as trunk (or clonal), branch, and private (subclonal) mutations. [16,17] Here, the classification is also based on the type of the lesion in which the mutations are located. Therefore, PAMs are comprised of trunk and branch mutations, while MSMs are composed of branch and private mutations. The percentage of PAMs varied over a wide range (0.2% to 44%). The variation in PAM levels due to change in regions number was insignificant (Kruskal-

Wallis test p = 0.24). The high variation of the PAM percentage reflects the diversity in the mutational profiles among patients in our cohort.

Next, the association of neoantigens generated by PAMs with immune cell infiltration was investigated. Putative neoantigens were derived from n.s. somatic mutations to determine their association with apparent adaptive immune responses, as reflected by T cell infiltration and upregulation of checkpoints in premalignant and malignant tissues. Multiple algorithms were applied to predict binding affinity (IC_{50}) between mutant proteins and patient HLAs based on the Immune Epitope Database recommendations[18]. Mutant peptides with predicted IC_{50} < 500 nM were considered neoantigens. In accordance with our mutation classification, the neoantigens were also categorized into three groups as premalignant (PrNs), progression-associated (PANs) and malignant-specific (MSNs) neoantigens. As expected, the total number of putative neoantigens per patient was highly correlated with the corresponding mutational load (Kendall's $\tau = 0.90$). The distribution of neoantigen groups in 41 cases is summarized in **Figure 3A,** in which the cases are ordered based on the percentage of PANs. The percentage of PANs per patient varied from 0.2 to 40% with a median of 5%, while that of MSNs fluctuated from 5% to 92%, and 6% to 90% for PrNs. In addition to the patient level analysis, neoantigens were characterized in each specific region. For example, the percentages of PANs in the individual AAH lesions were similar to the associated cancer, whereas the percentage of PANs at the patient level demonstrated inter-region homogeneity. **Figure 3B** illustrates the variation in the percentage of PANs in individual AAHs for each case arranged based on the median observed level. The percentages of PANs in individual AAH lesions were comparable in most patients, with the exception of five cases in which the range exceeded 20%.

In order to evaluate the association of neoantigen load and immune cell infiltration, we assessed CD4, CD8, FOXP3, Granzyme B, PD1 and PDL1 by immunostaining a total of 55 regions from

nine cases (indicated by arrows in **Figure 3A**) with distinct levels of PANs (0.5% — 28.4%) and MSNs (10.4% — 78.5%). The relationship between neoantigens and observed lymphocyte infiltration was evaluated by lesion- and patient-wise comparisons. In the lesion-wise comparison, neoantigens and immune infiltrates were limited to the matched premalignant lesions, while in the patient-wise analysis these endpoints were aggregated for the corresponding patient. We first investigated if there was an association between immune response and neoantigen at the patient level. We found that the percentage of PANs detected in each patient significantly correlated with the average percentage of $CD8^+$ T-cells infiltrating AAH lesions (Kendall's $\tau = 0.61$, $p = 0.02$, **Figure 3C** top panel) but not to those infiltrating AIS/ADC (Kendall's $\tau = 0.14$, $p = 0.7$, **Figure 3C** bottom panel). At the lesion level, we found that the percentage of $CD8^+$ T-cells infiltrating AAH lesions correlated strongly with the percentage of PANs in the respective lesions (Kendall's $\tau = 0.56$, $p = 0.0003$) (**Figure 3D**). Furthermore, AAH lesions with greater neoantigen loads had significantly more infiltrating CD4+ T cells (Kendall's $\tau = 0.32$, $p = 0.05$) (**Figure 3E**) and PD-L1 expressing cells (Kendall's $\tau = 0.44$, $p = 0.01$) (**Figure 3F**). These results indicate that the high levels of PANs promote $CD8^+$ T cell infiltration, whereas the overall number of neoantigens is associated with $CD4^+$ T cell infiltration and PD-L1 expression in AAH regions. Together, these findings suggest the presence of an adaptive immune response to neoepitopes.

PAMs and MSMs lead to deregulation of distinct cancer-related pathways.

In addition to investigating the effect of PAMs in generating immune responses in premalignant lesions, the potential roles of these mutations in tumor development were explored. In this cohort of patients we evaluated the mutational status of 29 driver genes frequently mutated in lung ADC[2,19]. We found that these genes were frequently mutated in ADC, but rarely in AAH. These results suggest that the driver mutations were necessary for the progression from AAH to cancer. Although driver gene mutations are important for tumor development, they are absent in

24-36% of lung ADCs[2,20]. In the current cohort, n.s. mutations in the above mentioned 29 driver genes were not detected in 51% of patients. Therefore, we next investigated the mutations in the context of molecular pathways.

For the pathway analysis, enrichment scores (ES) of the mutated genes involved in each specific pathway were defined. Briefly, the ES is the ratio between the observed number of mutated genes involved in a specific pathway and its estimated value based on the total numbers of pathway genes, mutated genes and genes in the genome (for details see **Methods** and **Equation S1**). Next, it was postulated that a pathway was deregulated by a specific group of mutated genes if there were at least two genes overlapping between them and their ES was ≥ 2 (FDR = 0.03). We determined if the genes mutated in ADC in both the current cohort and TCGA caused deregulation of the 1341 well-defined hallmark gene sets and canonical pathways from the Molecular Signature Database[13]. We found that the frequently deregulated pathways were commonly shared between the two data sets. However, for a given pathway, the probability of being deregulated tended to be higher in our data set compared to TCGA (**Figure S4B**). For instance, a linear regression model comparing probabilities of deregulated pathways in the two cohorts suggested the recurrence threshold of 0.39 (16 out of 41 patients) for our data set that was equivalent to 0.34 (151 out of 444) for TCGA. Using these thresholds, we identified 58 and 24 regularly deregulated pathways for the current cohort and TCGA data sets, respectively (**Figure S4A**). Fourteen of these pathways were shared between the data sets (Fisher's exact test p =3.8e-17) and are involved in tumor proliferation and invasion. These results indicate that although patients in both cohorts had different demographic features (such as gender and smoking status), and in turn had different driver mutations, they demonstrated common affected pathways involved in carcinogenesis.

Mutated genes in ADC included those bearing both PAMs and MSMs; therefore, it was essential to determine the input of each of the gene groups in the pathway context. The ES of each gene group was calculated for all 1341 pathways. **Figure 4A** shows the recurrence rate of the top 27 pathways that are frequently deregulated by the genes bearing MSMs. These pathways were also affected by the PAM-bearing genes, but predominantly at a lower frequency than the MSM-bearing genes. The termination of O-glycan biosynthesis pathway was found to be affected by the PAM-bearing genes in 85% of patients. This glycoprotein sialylation pathway involves several mucin proteins, including *MUC4*, that have been found to bear PAMs in 90% of the patients in the current cohort. Among pathways de-regulated by MSM-bearing genes, the focal adhesion pathway was ranked highest based on the recurrence rate. In 30 of 41 patients, this pathway was affected by either PAM- (n = 1) or MSM- (n = 17) bearing genes or both (n = 12) (**Figure 4B**). These patients were divided into two groups: Group 1 demonstrated pathways that were affected by both PAM and MSM bearing genes, and Group 2 in which pathways were affected only by MSM bearing genes. We found that: **a**) the Group 1 patients had PAM-based ES higher than the MSM-based ES (Kruskal-Wallis p = 0.01), and **b**) in the Group 2 patients the MSM-based ES was generally higher than that in Group 1 (Kruskal-Wallis p = 0.0004). As the ES is proportional to the percentage of the mutated genes belonging to the pathway of interest, these results imply that if a high percentage of PAM-bearing genes is involved in the oncogenic pathway, few additional MSMs are required for its de-regulation (**Figure 4B**, Group 1). In contrast, other pathway activities are affected only by the MSM-bearing genes (**Figure 4B**, Group 2).

In the next step, we comprehensively evaluated deregulation of all 1341 pathways based on PAM- and MSM-bearing genes to gain insight into tumor initiation and development. Deregulation status of all pathways based on two mutation groups (PAM or MSM) was tabulated as a one-and-zero binary matrix for all patients. The unsupervised hierarchical cluster analysis

based on the pathway status identified three patient groups designated as high (H, n = 12), intermediate (I, n = 20), and low (L, n = 9) depending on the number of pathways deregulated by PAM- and MSM-bearing genes (**Figure 4C**). The intermediate group included the majority of study patients, in which MSMs (but not PAMs) were the main source of pathway deregulation (**Figure S4B**) and were frequently found in the driver genes. Overall, it appears that in this group MSMs in the driver genes were essential for malignant progression. Group L, the smallest group, had infrequent pathway deregulation by either PAM- or MSM-bearing genes. Group H had the highest number of deregulated pathways among the three groups (**Figure S4B**). The deregulated pathways in this group were frequently affected by both PAMs and MSMs, as well as driver genes (**Figure 4C and S4B**). Also, based on the deregulation pattern of the focal adhesion pathway, the majority of the patients in this group belong to Group 1 shown in **Figure 4B.** In this subgroup of patients, *PIK3CA*, *PIK3R3* (catalytic and regulatory subunits of PI3K, respectively) and *PPP2R1A* genes (regulatory subunit of phosphatase PP2A negatively regulating AKT kinase) had high frequency of somatic mutations in AAH lesions. However, in all patients with PAMs in either or both *PIK3CA* and *PPP2R1A* genes, these mutations were detected in only one AAH lesion, and thus appeared as branch mutations. Moreover, cluster analysis revealed that these patients clustered predominantly in group H, suggesting that the PI3K/AKT pathway deregulation required another pathway to be deregulated to induce malignant transformation. Similarly, higher numbers of deregulated pathways in group H suggest that deregulation of multiple non-critical pathways may synergize with those that are critical, leading to malignant transformation. Conversely, in group L there were very few pathways deregulated by both PAM- and MSM-bearing genes, suggesting that the transformation could be caused by events other than the somatic driver mutations that were not readily detectable by WES, such as gene rearrangements, copy number variation, epigenetic changes, deregulation of gene expression or alternative splicing. The fact that group L included both patients that only had AIS but no ADC, suggests that lesions in this group had a low

invasive potential. The majority of study patients had driver genes and pathways affected by MSMs, suggesting that the mutation profiles of their AAH lesions were less complex compared to the associated ADC.

Discussion

In this study, we sought to identify somatic mutations in adenomatous premalignancy and associated lung adenocarcinoma and to determine the extent of immune cell infiltration of premalignant lesions and the associated tumors. Here we find histologic evidence for immune recognition of AAH lesions which reveal an immune contexture characterized by lymphocyte infiltration and checkpoint molecule upregulation consistent with adaptive immune responses. Furthermore, WES reveals putative neoepitopes which correlate with this evidence of adaptive immunity. Consistent with the immunoediting concept of Schreiber[21], neoepitopes were frequently identified that were present only in the premalignant lesions (PrNs), suggesting active immune- editing in the progression of adenomatous premalignancy to invasive adenocarcinoma. PANs were detected in all patients with 37/41 patients expressing these epitopes at greater than 1% frequency. The presence of PANs associated with immune cell infiltration suggests ongoing immunoediting responses in premalignancy and the associated tumor that have not yet fully "edited" neoepitopes. This possibility is supported by the presence of $CD8^+$ T lymphocyte infiltration that was found to be significantly ($p=0.0004$) correlated with the percentage of PANs in individual AAH lesions. The correlation of $CD4^+$ T lymphocyte infiltration as well as $PD\text{-}L1^+$ cells with neoepitope load in AAH lesions further indicates the potential importance of immune recognition and adaptive responses in premalignancy.

Our findings indicate that premalignant AAH lesions from within an individual patient may have distinct mutational profiles (**Figure S3**) and bear a range of PAMs (**Figure 2**). We found that the mutational profiles of AIS are distinct from those of AAH and highly overlap with those of ADC in

the majority of cases (**Figure S3**). Previous studies suggest that passenger mutations can promote malignant progression in either an additive manner or by modulating the activity of oncogenic or tumor suppressor pathways[22,23]. Therefore, beyond the individual mutations, we assessed the effect of premalignant somatic mutations in the context of pathways. Among the pathways deregulated in both the UCLA and TCGA cohorts (**Figure S4A**, upper right quadrant) the highest frequency of deregulation was found in the termination of the O-glycan biosynthesis pathway that includes mucin proteins which protect epithelial cells from physical and chemical damage. Deregulated expression of mucins promotes tumor cell invasion and migration, and increases drug resistance in a variety of malignancies[24,25]. Zhang and colleagues have reported that genetic variation of *MUC4* increases lung cancer risk[26], and here we find that PAMs of *MUC4* were present in over 90% of our patients. Also, focal adhesion, extracellular matrix-receptor interaction and calcium signaling pathways were frequently deregulated (**Table S3**). These pathways are affected by KRAS and PI3K activation and have established roles in carcinogenesis, including proliferation, invasion and resistance to therapy[27,28].

Neoantigens, produced by PAMs, are potential immunotherapy targets, but these neoantigens do not necessarily correspond to known driver genes. Consistent with findings in melanoma[29] and colorectal cancer[30], our analysis of mutations in lung adenocarcinoma indicates that while there are many common driver mutations among tumors from different patients, mutations producing PANs are most often unique to individual patients. Due to high genomic plasticity, the established cancers have highly heterogeneous mutational landscapes in different areas of the tumor due to potential parallel evolution and subclonal expansion[31-33]. This has been postulated to be one of the reasons for tumor resistance to therapies targeting actionable somatic events. In contrast to intra-tumor heterogeneity, intra-premalignancy heterogeneity appears to be significantly lower[34]. Consistent with these findings, we found that heterogeneity between different AAH lesions from an individual patient is significantly lower than that among lesions

from different patients (**Figure 2A and B**). Thus, it appears that therapies that target PANs in cancer interception and prevention strategies will need to be tailored to individual patients.

Cancer interception is a strategy that seeks to block the progression of premalignancy to invasive cancer[35]. However, the scarcity of studies evaluating relevant mutational signatures in pulmonary premalignancy limits the development of novel interception and prevention strategies for lung cancer. Due to the apparent genomic simplicity of premalignant AAH lesions relative to invasive cancer, it has been suggested that these lesions may harbor decipherable interception targets that can block the progression to malignancy[34]. Establishing the Precancer Atlas, similar to The Cancer Genome Atlas, will reveal novel findings regarding clonal evolution, diversity, and the immunosuppressive microenvironment[36]. This, in turn, will help identify mechanisms, targets, and immunogenic neoepitopes which then will facilitate the design of novel therapies, such as vaccines, to prevent progression to invasive disease[37].

The concept that genes bearing somatic mutations often encode tumor specific neoantigens capable of eliciting immunity and tumor rejection was first described in murine models sixty years ago[38]. In accord with the cancer immunosurveillance theory, our current findings support the concept that the immune system is capable of recognizing cancer precursors[39,40]. Because evasion of immune surveillance has been implicated as an emerging hallmark of cancer development, future investigations will focus on stimulating specific immune responses[41]. Thus, it has been suggested that unleashing the immune response against pulmonary premalignancy may facilitate a blockade of the progression of premalignancy to invasive cancer at the earliest stages of disease (37). This will require a more complete understanding of the immune microenvironment of pulmonary premalignancy as well as the identification of premalignant markers that could be targeted in immunoprevention strategies.

Materials and Methods

Specimen identification and processing.

FFPE tissue blocks from 41 patients with premalignant lesions and lung adenocarcinoma were obtained from the VA Greater Los Angeles Healthcare Center Lung Cancer and UCLA Lung Cancer Tissue Repositories, and were subjected to pathology review to identify specific histologic areas for Laser Capture Microdissection (LCM). Tissues were first sectioned at 7 μm thickness onto specialized membrane slides, and serial sections were stained with hematoxylin and eosin. LCM was performed utilizing a Leica LMD7000 in the California NanoSystems Institute Advanced Light Microscopy/Spectroscopy (ALMS) Core at UCLA. The following regions were dissected from distal airways: **a**) at least one region of normal airway epithelial cells (type I and II pneumocytes) adjacent to but not contiguous with the tumor, **b**) a minimum of two premalignant AAH lesions, **c**) all AIS regions (if present), and **d**) at least one ADC region.

Genomic DNA isolation and library preparation for DNA sequencing.

DNA was extracted from microdissected cells utilizing the HiPure FFPE DNA isolation kit (Roche). Sequencing libraries were constructed using NuGen Ovation Ultralow V2 system, followed by exome capture using the Roche SeqCap EZ kit as recommended by the manufacturers. The quality of every library preparation and exome capture reaction was evaluated by utilizing a Bioanalyzer instrument (Agilent), Quant-iT assay and qPCR. Sequencing was then performed on an Illumina HiSeq2000 instrument as 100 bp paired-end runs with the aim of ~50X per base (based on the Illumina Sequencing Coverage Calculation with an assumption of 35% PCR duplication and a minimum of 85% target coverage). Samples with an estimated library size < $2x10^7$ based on Picard *MarkDuplicates* function were re-sequenced to achieve a higher depth of coverage.

Whole exosome sequencing (WES) analysis and variant calling

Sequencing Alignment. Sequence reads were aligned to the human genome based on the NCBI human genome reference build 37 (GRCh37) by following the pipeline suggested by Genome Analysis Toolkit (GATK)[42]. In brief, raw reads were first pre-processed to remove adapter contamination by *scythe* adapter trimmer (https://github.com/vsbuffalo/scythe) and low quality base calls (Phred score Q <15) and short reads (length < 20) by *sickle* (https://github.com/najoshi/ sickle). Reads were mapped to the reference human genome by Burrows-Wheeler Aligner (v 0.7.7)[43], and then marked for PCR and optical duplicates with the Picard (v 1.77) *MarkDuplicates* tool. The GATK 2.7 was used for local indel realignment and base recalibration. For cases with multiple normal samples, their bam files from the bases recalibration step were combined and re-aligned to local indels before being subjected to variant calling analysis. In case samples were re-sequenced by multiple runs, raw reads in each run were first aligned and base recalibrated independently. Their bam files were then combined and re-aligned for indel realignment. Default values were set for the parameters other than those mentioned.

Variant Calling and Annotation. Somatic variants between pairs of abnormal regions (i.e. AAH, AIS, and ADC) and matched normal tissue were determined by VarScan2[44]. Tumor and normal cells having exomes sequenced were obtained from LCM, so VarScan2 was performed with a) tumor purity set to 1, and b) minimum coverage for normal and abnormal exomes set to 4. Because multiple exomes from different areas were sequenced per patient, the p-value threshold was set to 0.1 in somatic variant calling of individual exomes, and would be adjusted further in the next step of mutation calling in which somatic variants from all regions were analyzed together to identify mutations for each patient. The remaining VarScan2 parameters were set at default values. The output single nucleotide variant (SNV) calls were filtered further to remove false positive calls due to sequencing- or alignment-related artifacts by utilizing VarScan2's associated *fpfilter.pl* script. The resulting somatic SNV and indel calls were then

annotated by ANNOVAR[45] to identify non-synonymous (n.s.) variants from silent variants and common SNPs.

Mutation calling. For each patient, a n.s. somatic mutation was defined if a n.s. variant was 1) supported by at least three reads, and 2) observed in either a) more than one region with p-value ≤ 0.1, or (b) a single region with p-value ≤ 0.01.

Genetic homogeneity analysis

The similarity in n.s. somatic mutations between any pair of regions was assessed by Jaccard index which was defined as the ratio between the number of shared mutations between the regions over the total number of mutations identified in the regions.

Phylogenetic analysis

Non-synonymous somatic mutations were first converted into the format with 1 being mutated and 0 otherwise. For each patient, the analysis only considered n.s. somatic mutations that were present in more than one region to determine resemblance among AAH, AIS and ADC regions based on their mutation profiles. The analysis was performed in R by using *ape* and *phangorn* packages[46,47]. In brief, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) approach was utilized to cluster regions based on their mutation-defined binary format matrix. Unrooted phylogenetic trees were then drawn with relative branch lengths disproportionate to the number of shared mutations among corresponding regions.

Mutational architecture analysis

For each individual patient, n.s. mutations in all regions were pooled together and categorized into three groups: premalignant mutations (PrMs), progression-associated mutations (PAMs) and malignant-specific mutations (MSMs) based on their presence in different regions. A PrM was defined as n.s. mutation was observed only in AAH lesion(s), while a MSM was only

identified in AIS/ADC lesion(s), and finally a PAM was present in both AAH and AIS/ADC lesions. For each patient, the number of mutations in each category was then normalized to the total number of n.s. mutations observed in the corresponding patient. For each individual region, its PAM was normalized to the total number of mutations identified in the respective region.

Identification of patient HLA typing

The OptiType algorithm[48] was applied to deduce a four-digit HLA genotype from WES data. Before applying the algorithm, raw reads were first pre-processed to a) remove adapter contamination by *scythe*, and b) remove low quality base calls (Phred score Q <20) by *sickle*, and c) keep reads that mapped on HLA reference regions by *bwa* and had a length of at least 50 bp by *fastqutils*[49]. For pair-end data, sequences from each end were pre-processed independently before subjecting them to the OptiType algorithm.

Identification of putative neoantigens

For every patient, each n.s. single nucleotide mutation was able to generate a maximum of ten 10-mer peptides having the mutated amino acid at different locations. Similarly, for each indel which did not cause early termination, ten 10-mer peptides were also created that had from 1-9 amino acids altered from the reference sequence. MHC-I binding prediction tools downloaded from Immune Epitope Database (IEDB)[18] were utilized to predict the binding affinity of 10-mer peptides to the patient's HLA germline alleles. IEDB protocol recommended using multiple algorithms including Artificial Neural Network[50,51], b) Stabilized Matrix Method[52], and c) NetMHCpan[53] for predicting binding strength to a given HLA allele due to the allele's available database and preferred algorithms previously proven to have outstanding performance for such allele. The smallest IC50 value derived from multiple algorithms was used as the predicted binding affinity of each peptide to each HLA allele. Approximately 60 peptide-MHC combinations (i.e. 10 peptides x 6 MHC-I) were derived from a single n.s. mutation. The peptide-MHC pair

with the lowest predicted IC50 was selected to represent the candidate mutant peptide and its binding MHC-I partner. Finally, candidate neoantigens were defined as those with the predicted binding strength IC50 < 500nM. Neoantigens were categorized as premalignant neoantigens (PrNs), progression-associated neoantigen (PANs) and malignant-specific neoantigen (MSNs) according to the groups that their corresponding mutations were classified.

Pathway analysis/Gene enrichment analysis

In pathway analysis, every affected gene should be counted once for each individual patient even though multiple n.s. mutation sites were identified on the same gene. Therefore, n.s. mutated sites were first consolidated to their corresponding gene identity. In our study, n.s. somatic mutations were categorized into three different groups based on their presence in various tissues. Thus, their affected genes should be assigned to the corresponding groups to evaluate their effects on molecular pathways, especially on tumor initiation and development. To achieve this, for each patient, eligible genes were first labeled based on PAMs, which were then removed from the available gene list for next steps labeling MSMs and PrMs. The labeling procedure was then repeated for MSMs, followed by PrMs. This meant that each patient had three mutually exclusive gene groups representing their PAMs, MSMs, and PrMs.

For each individual patient, the enrichment of mutated genes in the group *i* involved in a specific the pathway *j* is measured by an enrichment score, ES_{ij}, defined as:

$$ES_{ij} = \begin{cases} 0 & if\ H_{ij} < 2 \\ \frac{H_{ij}}{M_i * (S_j/P)} = \frac{H_{ij}/M_j}{S_i/P} & if\ H_{ij} \geq 2 \end{cases} \qquad \textbf{Equation S1}$$

where H_{ij} is the number of mutated genes in the group *i* (e.g. PAM-, MSM-, and PrM-bearing genes) involved in the pathway *j*. M_i , S_j and P are the numbers genes in group *i*, pathway *j,* and the genome. In other words, the enrichment score *ES* is the number of mutated genes involved in a pathway normalized by the estimated number based on the numbers of genes in the interested groups *i*, pathways *j* and the genome. Note that a non-zero *ES* requires a minimum of

two mutated genes associated with the pathway of interest. Furthermore, for a given pathway *j* (i.e. denominator is constant in the right most side of **Equation S1**), the discrepancy in *ES* between two groups of interest is proportional to the difference of the percentage of genes that are associated with the pathway in those groups.

The false discovery rate of *ES* was estimated by the permutation approach in which mutated genes in each patient were first randomly sampled from the genome, and then assigned to PAM- and MSM-bearing groups. *ES* were then calculated according to the above equation for a total of 123 mutated gene groups (41 patients x 3 groups: PAM-, MSM-, and union of PAM- and MSM-bearing genes) based on 1341 canonical and hallmark pathways downloaded from the Molecular Signature Database[13]. A total of 100 permutations were executed.
Finally, a pathway was defined to be deregulated by a certain mutated gene group if the corresponding *ES* was greater or equal to 2 (FDR = 0.03). In each patient, the deregulation states of all pathways based on PAM- and MSM-bearing genes were represented in binary format with 1 being deregulated and 0 for otherwise. The pathway-based binary data from all patients was then combined into the matrix form and subjected to unsupervised clustering analysis to stratify patients into subgroups. The cluster analysis was performed in R by utilizing Ward's clustering method (i.e. *ward.D2*).

Analyses using TCGA data sets (DNASeq, RNAseq and survival analysis)

Processed data sets from whole exome DNA and mRNA sequencing, as well as clinical information for lung adenocarcinoma (LUAD) samples were downloaded from the Cancer Genome Atlas (TCGA) data portal. The information of mutated genes in samples was extracted from somatic mutation calls (level 2 maf file), and organized into a table in which one was employed to indicate if the gene of interest had at least one non-silent mutation call located on

its coding regions, and zero for otherwise in the specific sample. The frequency of how often a gene was mutated in the cohort was then calculated from the table.

In gene expression analysis, RSEM normalized gene expression (level 3 text files) files were utilized to build a data matrix of all samples. The expression data was processed by removing a) genes with low abundance (i.e. < 1 CPM in >30% samples), and b) tumor samples without WES data. Pathway activities per individual sample were derived from its gene expression by using Gene Set Variation Analysis (GSVA)[14]. The information of gene sets involved in the immune regulated pathways was obtained from the Molecular Signature Database[13]. To eliminate the effect of genes commonly shared among pathways, the original gene sets were modified such that the overlapping genes were kept in the "child" and removed from the "parent" set. A "child" set was defined as the one having more than 90% of members overlapping with the parent set.

The GSVA scores of the interested pathways were then subjected to the unsupervised hierarchical cluster analysis to stratify samples into subgroups. The cluster analysis, which was performed in R, used Ward's clustering method (i.e. *ward.D2*) and Spearman correlation coefficient as the metric measuring similarity between sample pairs. Finally, patient survival among the subgroups was compared by log-rank test.

Evaluation of Lymphocytic infiltration

For each lesion, a section stained with hematoxylin and eosin underwent an initial qualitative evaluation by a board-certified pathologist to assess the overall degree of lymphocytic infiltration. This assessment utilized a simple graded scale: 0 (absent), 1 (focal with <3 clusters of 3 lymphocytes), 2 (multifocal with 3 or more clusters) and 3 (diffuse). χ^2 test was used to compared distributions of scores in different histological lesions (normal, AAH, AIS and ADC).

Immunohistochemistry analyses

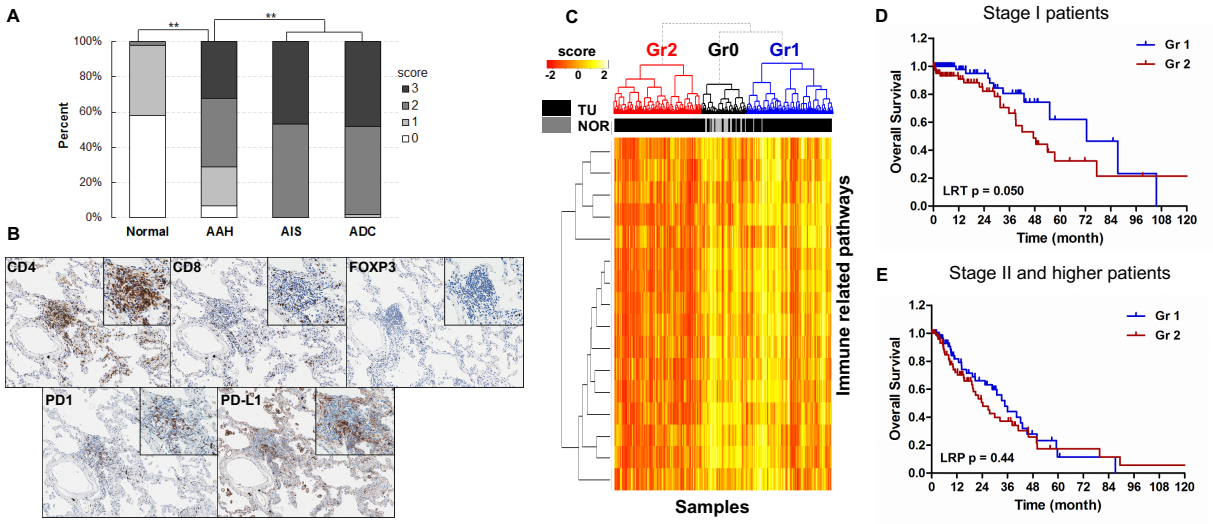
For nine cases, additional serial sections of 5 μm thickness were obtained from FFPE tissue blocks. Single-color immunostaining was performed on the Leica Bond III autostainer using Bond Low (H1) and High (H2) heat retrieval solutions, wash buffer, and Refine Polymer Detection system. Heat-induced epitope retrieval was performed in the autostainer, except for PD1 and PD-L1, which were treated in a pressure cooker. Antibodies used for detection of a single marker per slide included: CD8 (Dako #M7103), CD4 (Cell Marque #104R-16), Granzyme B (Dako #M7236), PD1 (Cell Marque #315M), PD-L1 (Spring Bio M4420), and FOXP3 (Bio SB #BSB676).

All slides were scanned at an absolute magnification of 3200 (resolution of 0.5 mm per pixel). Brightfield image analysis was performed using the Indica Labs Halo platform. With the assistance of a board-certified pathologist, each region of interest (AAH, AIS and ADC) was identified and outlined on the hematoxylin and eosin guide slide. The guide slide was aligned and synced with the corresponding serial sections immunostained for each marker. Existing Halo algorithms developed for detection of positive staining were accepted or modified based on the positive control slide for each marker. The final algorithm was then used to analyze the density (cells/mm^2) and percentage cellularity (% positive cells/all nucleated cells) for each marker on each region of interest. This raw data was then exported for statistical analysis.

Statistical analyses

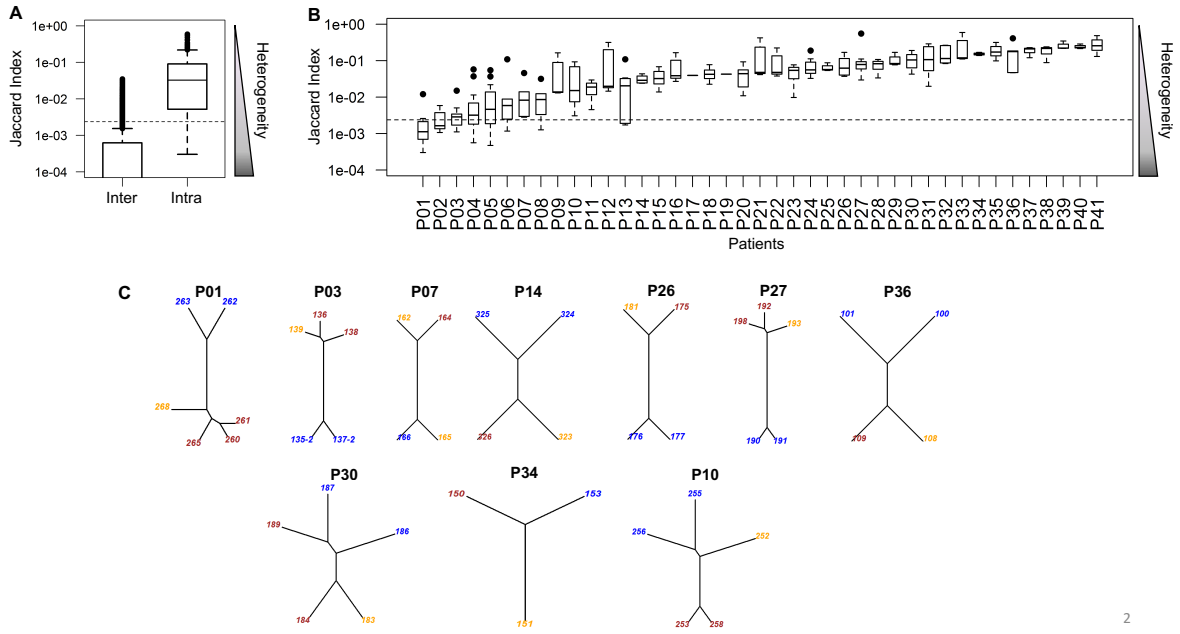
All analyses were performed utilizing R 3.2. Appropriate rank-based statistical tests were applied according to the nature of variables. For instance, Kendal's τ coefficient was used to assess association between the pairs of variables, such as percentage of PAMs, percentage of positively stained cells and log transformed neoantigen numbers, while the Kruskal-Wallis rank sum was applied to compare variables of interest between groups.

FIGURE 1



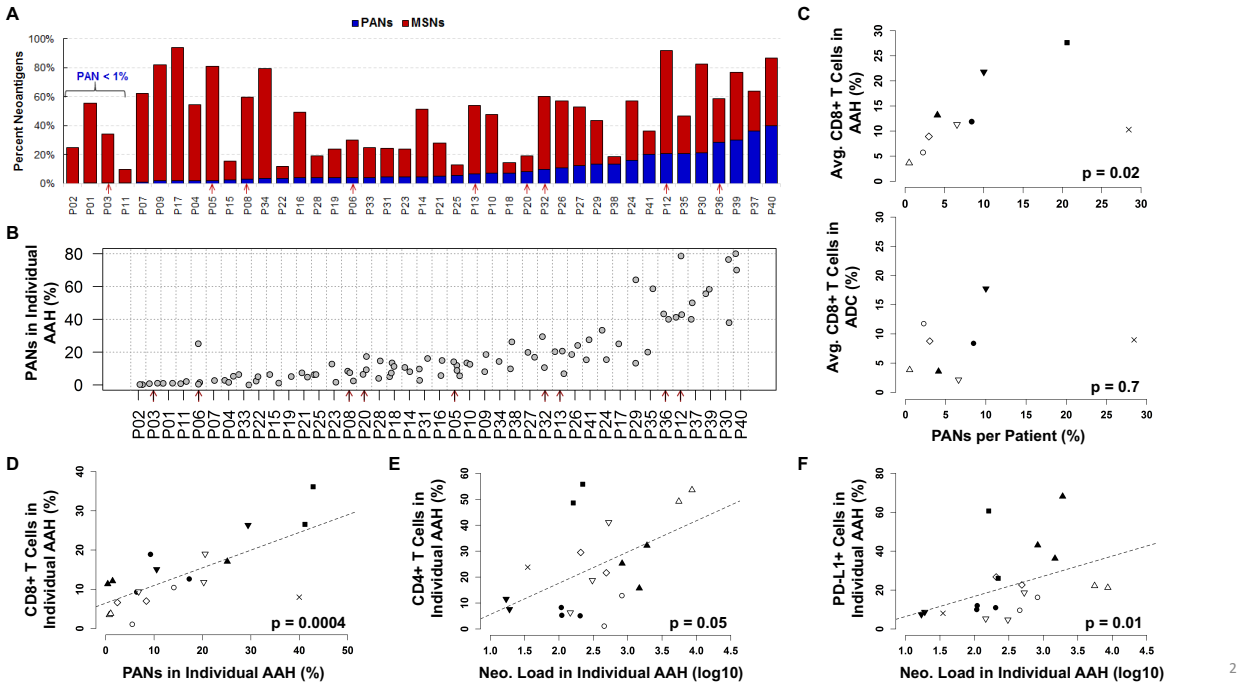
Cell-mediated immunity and adaptive responses in lung cancer continuum
A) Local lymphocyte infiltration index (0 — lowest, 3 — highest) in adjacent normal tissue, AAH, AIS and ADC (**χ^2 test $p < 10^{-10}$). **B**) A representative IHC staining of lymphocytic markers in an AAH lesion with local lymphocyte infiltration score = 2. **C**) Heatmap of gene expression scores of 16 immune-related pathways in TCGA LUAD and normal lung samples. Serial sections stained for the indicated markers shown at 10x and 20x magnification. **D**) Kaplan-Meier survival curves of stage I, and **E**) stage II and higher patients from the groups identified in **Figure 1C**.

FIGURE 2



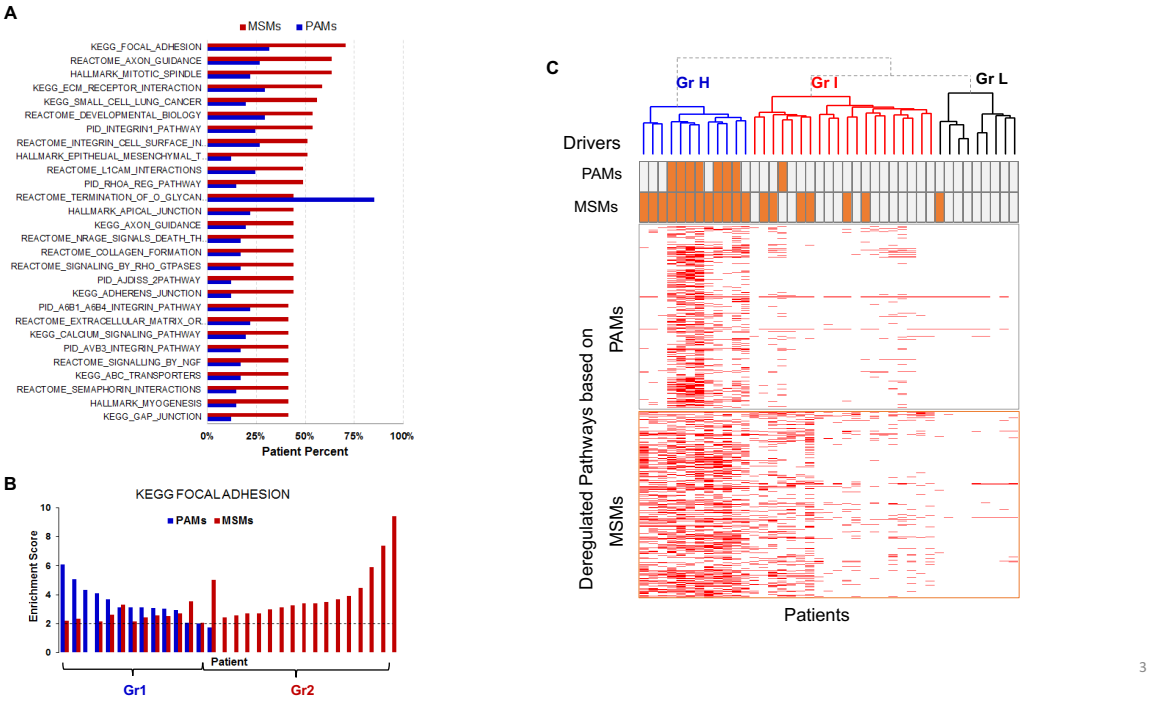
Intra- and inter-patient genetic heterogeneity of lung lesions.
A) Distribution of Jaccard indices comparing n.s. somatic mutation heterogeneity between AAH lesions from the same (intra-) or different (inter-) patients. Inter-patient Jaccard indices have the zero-median which cannot be displayed on the log-scale y-axis, so the distribution starts at 54 percentiles. **B**) Distribution of intra-patient Jaccard indices in 41 individual patients. The subjects are displayed in the low-to-high order based on their median values. In **A** and **B**, the side triangles represent the heterogeneity levels inversely proportional to Jaccard indices, and the dashed line marks the 90-percentile level of inter-subject Jaccard index. **C**) Phylogenic trees for 10 patients with AAH (blue), AIS (orange) and ADC (brown). Phylogenic trees for the entire cohort are shown in **Figure S2**.

FIGURE 3



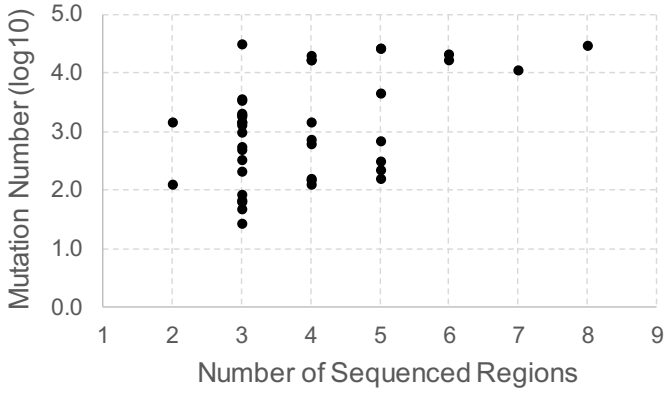
Neoantigens and the immune response in pulmonary premalignancy.
A) Distribution of PANs and MSMs in 41 study patients. The patients are displayed in the low-to-high order based on their percentages of PANs. Red arrows in **A** and in **B** indicate nine patients whose cellular immune response was evaluated. **B**) Percentage of PANs in individual AAH regions from 41 patients. The subjects are displayed in the low-to-high order based on their median levels, and not in the same order as those in **A**. **C**) Average percentages of infiltrating $CD8^+$ T cells observed in AAH (upper panel) and ADC (lower panel) plotted against percentage of patient-wise PANs. Each patient is represented by a data point indicated by a unique symbol. ADC in one patient was not evaluated. **D**) Correlation between the percentage of infiltrating $CD8^+$ T cells and the percentage of PANs in corresponding AAH lesions. **E-F**) Correlation between the percentage of infiltrating $CD4^+$ T cells (**E**) and PD-$L1^+$ cells (**F**) plotted against the corresponding log-transformed neoantigen number identified in AAHs. In **D-F** each region is represented by a point, while each patient is marked by the symbol identical to those in **B**. P-values in **B**, **D-F** are based on Kendall rank. The trend line (dashed line) indicates the linear association between variables. Other pair-wise comparisons between immune marker levels and neoantigen-related variables were insignificant, and were not shown.

FIGURE 4



Analysis of pathway deregulation by PAM and MSM.
A) The top 27 pathways frequently affected by MSM- (red) and PAM- (blue) bearing genes. **B**) Enrichment score of MSM- (red) and PAM- (blue) bearing genes involved in the KEGG-based focal adhesion pathway, which is the most deregulated pathway by MSM-bearing genes, plotted for each patient. Patients having less than two pathway genes (i.e. ES = 0) mutated in ADC are not included in the analysis. The gray dashed line indicates the significant ES threshold (= 2) to determine if the genes involved in pathway were significantly enriched by the mutated genes in the specific group. Two patient groups were defined based on their PAM-based ES ≥ 2. **C**) Heatmap of the pathways affected (red) by PAM- (top) and MSM- (bottom) bearing genes. The mutations in the 29 drive genes (listed in **Figure S3A**) observed in PAM and MSM are indicated by orange bars above the heatmap.

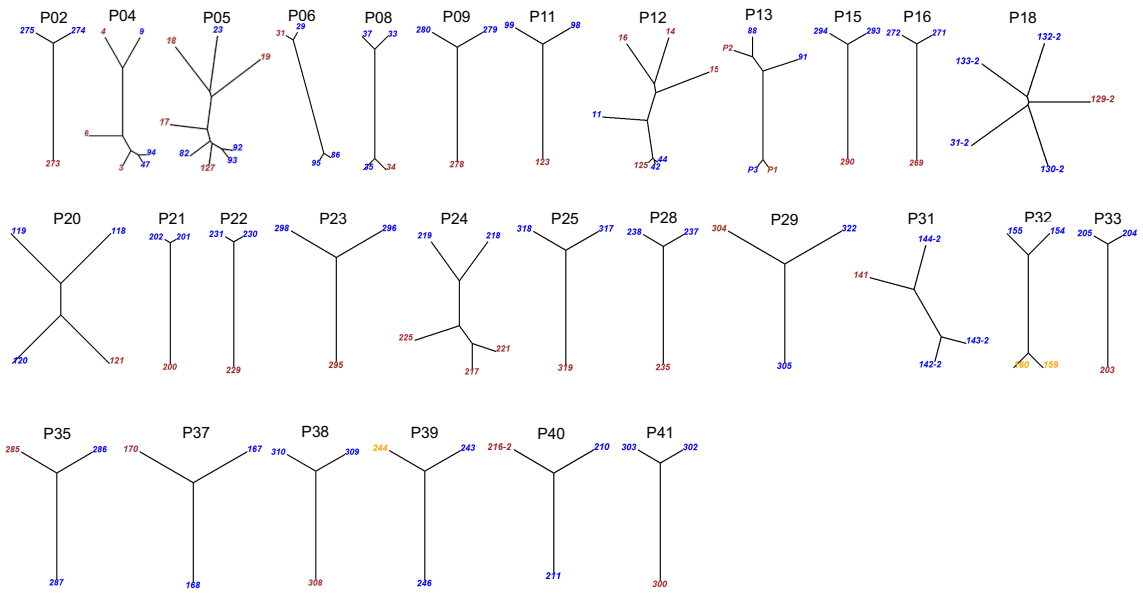
FIGURE S1



4

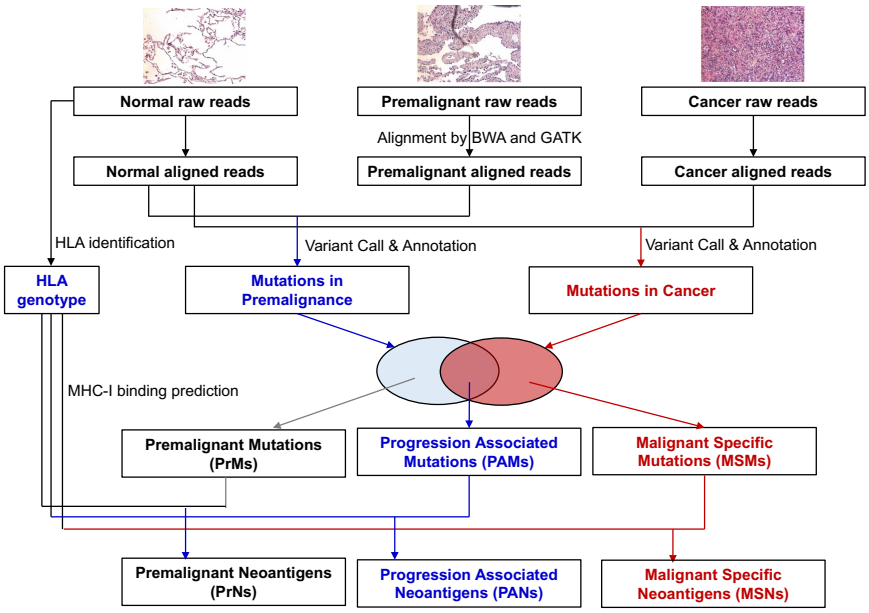
The of number of lesions sequenced per patient does not significantly alter the mutational load.

FIGURE S2



Phylogenetic trees.

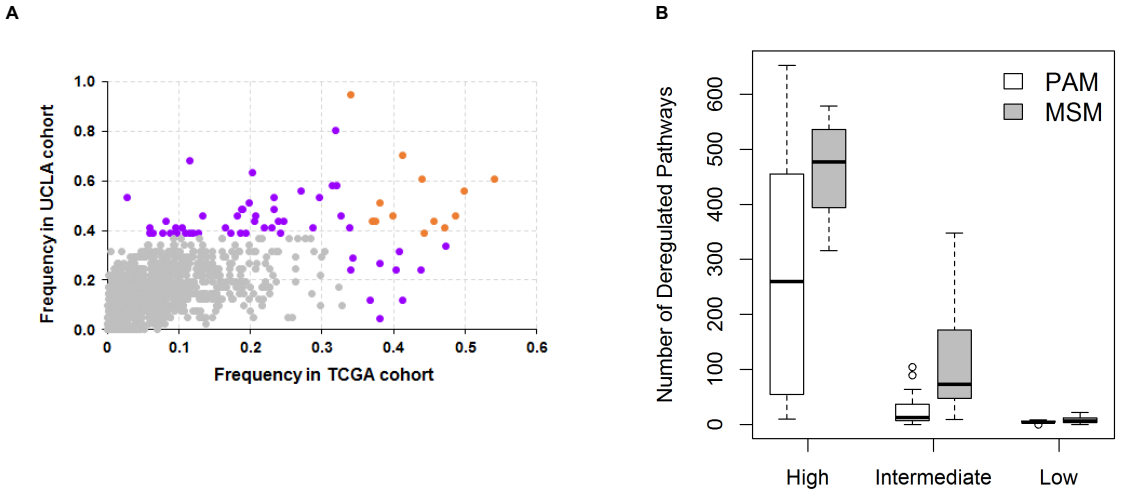
FIGURE S3



6

Outline of the experimental approach.

FIGURE S4



7

Pathways, frequently deregulated by the mutated genes in TCGA LUAD and UCLA cohorts.

TABLE S1

Characteristic	Study Population (N = 41)
Age – year (mean)	66.8
Female sex – no. (%)	32 (78.0)
Ethnicity – no. (%)	
Caucasian	33 (80.4)
East Asian	4 (9.8)
Other	4 (9.8)
Smoking Status – no. (%)	
Current	7 (17.1)
Former	28 (68.3)
Never	6 (14.6)
Pathologic Stage – no. (%)	
0	3 (7.3)
1	29 (70.7)
2	9 (22.0)

Demographics of UCLA cohort

FIGURE 2S

Patient ID	AAH	AIS	ADC
P01	2	1	3
P02	2	0	1
P03	2	1	2
P04	3	0	3
P05	4	0	4
P06	3	0	1
P07	1	2	1
P08	3	0	1
P09	2	0	1
P10	2	1	2
P11	2	0	1
P12	3	0	4
P13	3	0	2
P14	2	1	1
P15	2	0	1
P16	2	0	1
P17	1	1	0
P18	4	0	1
P19	1	0	1
P20	3	0	1
P21	2	0	1
P22	2	0	1

P23	**2**	**0**	**1**
P24	**2**	**0**	**3**
P25	**2**	**0**	**1**
P26	**2**	**1**	**1**
P27	**2**	**1**	**2**
P28	**2**	**0**	**1**
P29	**2**	**0**	**1**
P30	**2**	**1**	**2**
P31	**3**	**0**	**1**
P32	**2**	**2**	**0**
P33	**2**	**0**	**1**
P34	**1**	**1**	**1**
P35	**2**	**0**	**1**
P36	**2**	**1**	**1**
P37	**2**	**0**	**1**
P38	**2**	**0**	**1**
P39	**2**	**1**	**0**
P40	**2**	**0**	**1**
P41	**2**	**0**	**1**

Summary of regions sequenced in each patient

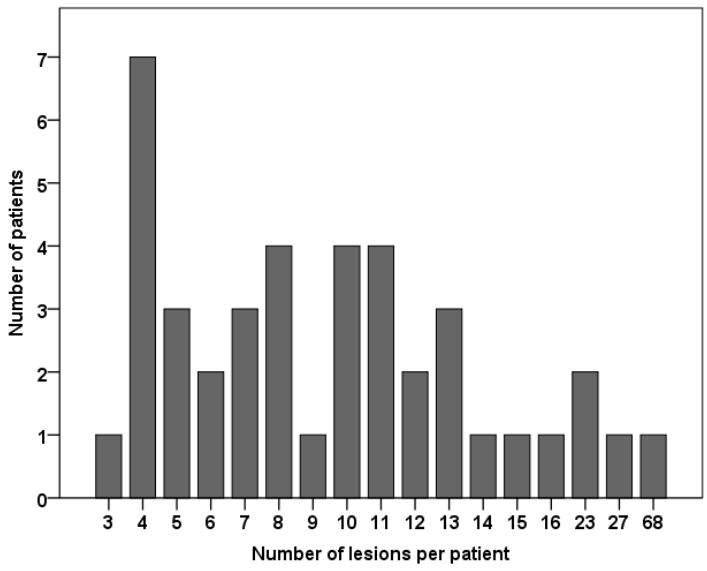
STATISTICAL APPENDIX

Section I: Extended Analysis of Lymphocyte Infiltration Score (LIS) Data

Cohort Summary Statistics

In total, 41 patients with a total of 453 regions were assigned an LIS. There were 3-68 regions per patient (Figure A1). AAH lesions totaled 337, making up 74% of all regions undergoing scoring. The primary adenocarcinoma tumor stage for each subject ranged from 0 (AIS only) to 3A. A total of 29 patients (70%) had stage IA or 1B disease at resection. 9 patients (21%) had stage IIA or higher disease. Interestingly, 3 patients had stage 0 disease, defined as adenocarcinoma in-situ (AIS) only.

Figure 1A



Effect of Staging and Patient Effects on LIS

Given that immunosurveillance is thought to weaken with tumor progression, we hypothesized that the LIS may be affected by the stage of the primary resected lung adenocarcinoma. We approached this question with 2 methods.

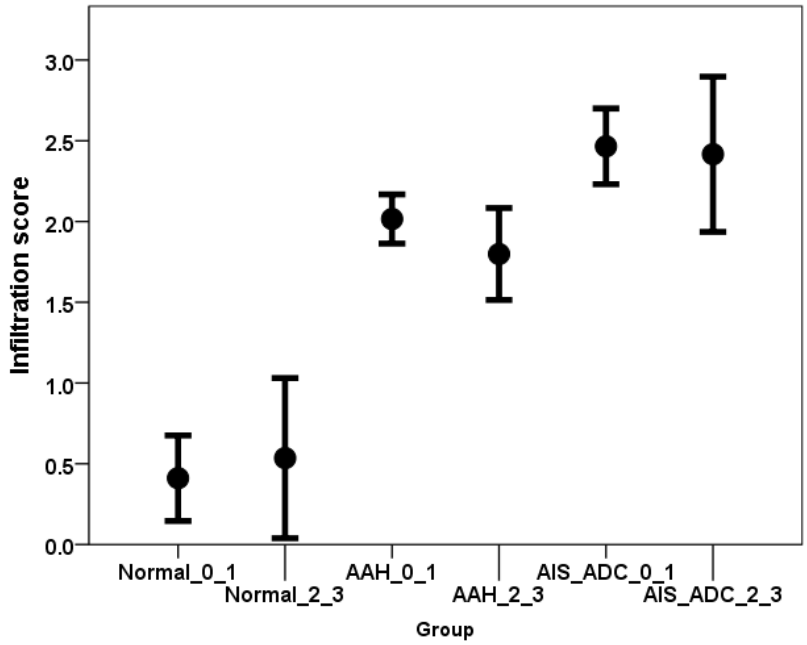
Method I: Linear mixed effects model

In the “naïve” model, we treated the lymphocyte infiltration score as continuous. Included in the model was tissue type (Normal, AAH or AIS/ADC), stage (0-1 vs. 2-3), an interaction term and the subject random effect. We found that while tissue type had a significant effect on LIS ($p < 0.001$), the tumor stage did not ($p = 0.78$). These values did not change significantly after removal of the non-significant interaction term.

Method II: Ordinal logistic mixed effects model

In this model, LIS was treated as an ordinal value. The model included the same terms as used in the linear mixed effects model. Not significantly different from the prior model, we again found that tissue type had a significant effect on LIS ($p<0.001$). The stage of the primary tumor did not significantly affect the LIS ($p=0.85$) These results are shown in Figure 2A.

Figure 2A



Section II: Extended Analysis of Quantitative Immunohistochemistry (IHC) Data

Summary Statistics

As noted previously, a total subset of 9 patients were selected to undergo quantitative IHC analysis using the HALO platform. In total, these patients 65 regions analyzed, which ranged from 5-12 per patient. Of the 9 patients, 6 had stage 0 or 1 disease. The remaining 3 patients had stage 2A or 2B disease.

Inter-correlation of IHC marker intensity

Given the known relationships between some of these markers from a biologic standpoint, we evaluated their inter-correlation using the Pearson correlation coefficient. We also calculated their average percentage of positive cells per marker. These findings are shown in Table A1.

Table A1

	CD4	CD8	FOXP3	Granzyme	PD1	PDL1
CD4	1.00					
CD8	0.27	1.00				
FOXP3	0.19	0.02	1.00			
Granzyme	0.25	0.32	0.14	1.00		
PD1	0.42	0.64	0.01	0.34	1.00	
PDL1	0.41	0.20	0.11	0.14	0.30	1.00
Average Percent Positive Cells per Area	24.34%	11.43%	4.04%	3.62%	3.59%	22.21%

Effect of Staging and Patient Effects on IHC marker intensity

Similar to our prior analysis of the LIS, we were interested in the effect of primary tumor stage upon the quantitative IHC scores for CD8+ and CD4+. We utilized a linear mixed effects model. Included in the model was tissue type (Normal, AAH or AIS/ADC), stage (0-1 vs. 2-3), an interaction term and the subject random effect. For CD8+ staining, we found that tissue type had a significant effect ($p < 0.02$), but staging did not ($p = 0.45$). Similarly, for CD4+ staining, tissue type had a significant effect ($p<0.04$), but tumor stage did not (0.34).

Discussion

In conclusion, both methods to quantifying intra-lesional lymphocyte infiltration, LIS and quantitative IHC, were significantly affected by tissue type. Both AAH and ADC lesions had significantly higher lymphocyte infiltration than normal regions. The stage of the primary tumor did not significantly affect the lymphocyte infiltration observed using either method. While this suggests that tumor stage may truly not influence the immune cell infiltrate seen across lesions, it is limited by the small sample size of patients with stage II or higher disease in the cohort. As expected, a strong correlation between CD8+ and PD-1 staining was observed.

References

1 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer Statistics, 2017. *CA: a cancer journal for clinicians* **67**, 7-30, doi:10.3322/caac.21387 (2017).

2 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

3 Niho, S. *et al.* Monoclonality of atypical adenomatous hyperplasia of the lung. *The American journal of pathology* **154**, 249-254, doi:10.1016/s0002-9440(10)65271-6 (1999).

4 Mori, M., Rao, S. K., Popper, H. H., Cagle, P. T. & Fraire, A. E. Atypical adenomatous hyperplasia of the lung: a probable forerunner in the development of adenocarcinoma of the lung. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **14**, 72-84, doi:10.1038/modpathol.3880259 (2001).

5 Westra, W. H. *et al.* K-ras oncogene activation in atypical alveolar hyperplasias of the human lung. *Cancer research* **56**, 2224-2228 (1996).

6 Yatabe, Y., Kosaka, T., Takahashi, T. & Mitsudomi, T. EGFR mutation is specific for terminal respiratory unit type adenocarcinoma. *The American journal of surgical pathology* **29**, 633-639 (2005).

7 Yoshida, Y. *et al.* Mutations of the epidermal growth factor receptor gene in atypical adenomatous hyperplasia and bronchioloalveolar carcinoma of the lung. *Lung cancer (Amsterdam, Netherlands)* **50**, 1-8, doi:10.1016/j.lungcan.2005.04.012 (2005).

8 Slebos, R. J. *et al.* p53 alterations in atypical alveolar hyperplasia of the human lung. *Human pathology* **29**, 801-808 (1998).

9 Takamochi, K. *et al.* Loss of heterozygosity on chromosomes 9q and 16p in atypical adenomatous hyperplasia concomitant with adenocarcinoma of the lung. *The American journal of pathology* **159**, 1941-1948, doi:10.1016/S0002-9440(10)63041-6 (2001).

10 Travis, W. D. *et al.* The IASLC Lung Cancer Staging Project: Proposals for Coding T Categories for Subsolid Nodules and Assessment of Tumor Size in Part-Solid Tumors in the Forthcoming Eighth Edition of the TNM Classification of Lung Cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **11**, 1204-1223, doi:10.1016/j.jtho.2016.03.025 (2016).

11 Izumchenko, E. *et al.* Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nature communications* **6**, 8258, doi:10.1038/ncomms9258 (2015).

12 Burnett, F. M. *Immunological surveillance*. (Pergamon Press, 1970).

13 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of*

Sciences of the United States of America **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

14 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **14**, 7, doi:10.1186/1471-2105-14-7 (2013).

15 Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics* **15**, 121-132, doi:10.1038/nrg3642 (2014).

16 McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science (New York, N.Y.)* **351**, 1463-1469, doi:10.1126/science.aaf1490 (2016).

17 Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science (New York, N.Y.)* **346**, 256-259, doi:10.1126/science.1256930 (2014).

18 Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic acids research* **43**, D405-412, doi:10.1093/nar/gku938 (2015).

19 Berger, A. H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer cell* **30**, 214-228, doi:10.1016/j.ccell.2016.06.022 (2016).

20 Sholl, L. M. *et al.* Multi-institutional Oncogenic Driver Mutation Analysis in Lung Adenocarcinoma: The Lung Cancer Mutation Consortium Experience. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **10**, 768-777, doi:10.1097/JTO.0000000000000516 (2015).

21 Mittal, D., Gubin, M. M., Schreiber, R. D. & Smyth, M. J. New insights into cancer immunoediting and its three component phases--elimination, equilibrium and escape. *Current opinion in immunology* **27**, 16-25, doi:10.1016/j.coi.2014.01.004 (2014).

22 Leedham, S. & Tomlinson, I. The continuum model of selection in human tumors: general paradigm or niche product? *Cancer research* **72**, 3131-3134, doi:10.1158/0008-5472.CAN-12-1052 (2012).

23 Muller, F. L. *et al.* Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **488**, 337-342, doi:10.1038/nature11331 (2012).

24 Brockhausen, I. Pathways of O-glycan biosynthesis in cancer cells. *Biochimica et biophysica acta* **1473**, 67-95 (1999).

25 Rao, C. V., Janakiram, N. B. & Mohammed, A. Molecular Pathways: Mucins and Drug Delivery in Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **23**, 1373-1378, doi:10.1158/1078-0432.CCR-16-0862 (2017).

26 Zhang, Z. *et al.* Genetic variants in MUC4 gene are associated with lung cancer risk in a Chinese population. *PloS one* **8**, e77723, doi:10.1371/journal.pone.0077723 (2013).

27 Stewart, T. A., Yapa, K. T. & Monteith, G. R. Altered calcium signaling in cancer cells. *Biochimica et biophysica acta* **1848**, 2502-2511, doi:10.1016/j.bbamem.2014.08.016 (2015).

28 Zhao, J. & Guan, J. L. Signal transduction by focal adhesion kinase in cancer. *Cancer metastasis reviews* **28**, 35-49, doi:10.1007/s10555-008-9165-4 (2009).

29 Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (New York, N.Y.)* **350**, 207-211, doi:10.1126/science.aad0095 (2015).

30 Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome biology* **16**, 64, doi:10.1186/s13059-015-0620-6 (2015).

31 De Sousa, E. M. F., Vermeulen, L., Fessler, E. & Medema, J. P. Cancer heterogeneity--a multifaceted view. *EMBO reports* **14**, 686-695, doi:10.1038/embor.2013.92 (2013).

32 McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell* **27**, 15-26, doi:10.1016/j.ccell.2014.12.001 (2015).

33 Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338-345, doi:10.1038/nature12625 (2013).

34 Hait, W. N. & Levine, A. J. Genomic complexity: a call to action. *Science translational medicine* **6**, 255cm210, doi:10.1126/scitranslmed.3009148 (2014).

35 Blackburn, E. H. Cancer interception. *Cancer prevention research (Philadelphia, Pa.)* **4**, 787-792, doi:10.1158/1940-6207.capr-11-0195 (2011).

36 Campbell, J. D. *et al.* The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer prevention research (Philadelphia, Pa.)* **9**, 119-124, doi:10.1158/1940-6207.CAPR-16-0024 (2016).

37 Spira, A. *et al.* Precancer Atlas to Drive Precision Prevention Trials. *Cancer research* **77**, 1510-1541, doi:10.1158/0008-5472.can-16-2346 (2017).

38 Prehn, R. T. & Main, J. M. Immunity to methylcholanthrene-induced sarcomas. *Journal of the National Cancer Institute* **18**, 769-778 (1957).

39 Zitvogel, L., Tesniere, A. & Kroemer, G. Cancer despite immunosurveillance: immunoselection and immunosubversion. *Nature reviews. Immunology* **6**, 715-727, doi:10.1038/nri1936 (2006).

40 Galon, J., Angell, H. K., Bedognetti, D. & Marincola, F. M. The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures. *Immunity* **39**, 11-26, doi:10.1016/j.immuni.2013.07.008 (2013).

41 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

42 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

43 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

44 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

45 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

46 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

47 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

48 Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310-3316, doi:10.1093/bioinformatics/btu548 (2014).

49 Breese, M. R. & Liu, Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494-496, doi:10.1093/bioinformatics/bts731 (2013).

50 Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic acids research* **36**, W509-512, doi:10.1093/nar/gkn202 (2008).

51 Lundegaard, C., Lund, O. & Nielsen, M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* **24**, 1397-1398, doi:10.1093/bioinformatics/btn128 (2008).

52 Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC bioinformatics* **6**, 132, doi:10.1186/1471-2105-6-132 (2005).

53 Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1-13, doi:10.1007/s00251-008-0341-z (2009).