Journal of Translational Medicine

RESEARCH

Open Access

Uncovering molecular features driving lung adenocarcinoma heterogeneity in patients who formerly smoked

Peiyao Wang[1,2], Raymond Ng[3], Stephen Lam[1] and William W. Lockwood[1,2,4*]

Abstract

Background An increasing proportion of lung adenocarcinoma (LUAD) occurs in patients even after they have stopped smoking. Here, we aimed to determine whether tobacco smoking induced changes across LUADs from patients who formerly smoked correspond to different biological and clinical factors.

Methods Random forest models (RFs) were trained utilizing a smoking associated signature developed from differentially expressed genes between LUAD patients who had never smoked (NS) or currently smoked (CS) from TCGA ($n=193$) and BCCA ($n=69$) cohorts. The RFs were subsequently applied to 299 and 131 formerly smoking patients from TCGA and MSKCC cohorts, respectively. FS were RF-classified as either CS-like or NS-like and associations with patient characteristics, biological features, and clinical outcomes were determined.

Results We elucidated a 123 gene signature that robustly classified NS and CS in both RNA-seq (AUC = 0.85) and microarray (AUC = 0.92) validation test sets. The RF classified 213 patients who had formerly smoked as CS-like and 86 as NS-like from the TCGA cohort. CS-like and NS-like status in formerly smoking patients correlated poorly with patient characteristics but had substantially different biological features including tumor mutational burden, number of mutations, mutagenic signatures and immune cell populations. NS-like formerly smoking patients had 17.5 months and 18.6 months longer overall survival than CS-like patients from the TCGA and MSKCC cohorts, respectively.

Conclusions Patients who had formerly smoked with LUAD harbor heterogeneous tumor biology. These patients can be divided by smoking induced gene expression to inform prognosis and underlying biological characteristics for treatment selection.

Keywords Lung adenocarcinoma, Smoking, Gene expression, Prognosis, Treatment selection

*Correspondence:
William W. Lockwood
wlockwood@bccrc.ca
[1]Department of Integrative Oncology, BC Cancer Research Institute, 675 West 10th Avenue, Vancouver, BC V5Z 1G1, Canada
[2]Interdisciplinary Oncology Program, Faculty of Medicine, 570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada
[3]Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, BC V6T 1Z4, Canada
[4]Department of Pathology and Laboratory Medicine, University of British Columbia, 899 West 12th Avenue, Vancouver, BC V5Z 4E6, Canada

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Lung cancer is the most lethal cancer in the world, with the majority of cases attributed to tobacco smoking. Successes in tobacco smoking control policies and smoking cessation programs have led to a decrease in the number of people who are actively smoking [1]. In countries such as the United States and Canada, over 50% of lung cancer deaths are now in people who had stopped smoking [2]. As a history of smoking is the biggest risk factor in terms of lung cancer development, people who have previously smoked are at an elevated risk. While lung cancer can occur even years after smoking cessation, [3] risk decreases gradually over time after smoking cessation at a rate that varies among individuals for reasons that remain unclear [4]. A recent meta-analysis showed the reducible relative risk after smoking cessation only marginally declines after 15 years from 26.7% (95% CI 20.2–34.3) to 19.7% (95% CI 13.3–26.4) at 20 years [3]. This motivates the study of tumor biology in formerly smoking lung cancer patients to determine underlying biological traits that may otherwise separate this population beyond clinical characteristics for the purposes of risk stratification.

In terms of clinical research, patients who have formerly smoked (FS) are often treated the same as those who are currently smoking, grouped together as 'ever smokers'. However, a recent study that subdivided its cohort into patients who had never smoked (NS), patients who currently smoked (CS), or FS showed that CS have significantly greater survival after PD-L1 inhibitor treatment than patients who previously smoked in refractory NSCLC, with NS experiencing significantly worse survival compared to both groups [5] In addition, another study showed that smoking exposure can be quantified using tumor mutational burden (TMB) and transversion/transition ratio, which can be applied to classify NS, CS, FS who quit in the last 15 years and those who quit over 15 years ago [6]. This supports the idea that FS are distinct from CS as well as among themselves, although the other molecular features that separate FS and how they can translate to clinical management and treatment strategies is currently unknown.

In this study, we aimed to further understand the diversity of FS with lung adenocarcinoma (LUAD) - the most common lung cancer subtype - by exploring their tumor biology and molecular features. We hypothesized that a subset of FS patients develop cancer due to the carcinogenic effects of previous tobacco smoking, while others may develop cancer through processes unrelated to smoking. To this end, we developed an active smoking associated gene expression signature to classify LUADs from FS, which revealed distinct subsets related to either CS or NS LUADs. Furthermore, we demonstrated that these subsets have unique underlying molecular features that influence heterogeneity in tumor biology across FS. This insight towards the mechanisms underlying tumor development in people who have stopped smoking have potential implications for treatment and clinical management of the largest LUAD patient group in the future.

Methods

Data sources

Gene expression data from LUAD tumor samples with information regarding patient smoking status were obtained from three sources. The Cancer Genome Atlas (TCGA) dataset contained 500 RNA-Seq samples (118 CS, 75 NS, 307 FS) (Illumina HiSeq RNA -Seq V2 RSEM) that were downloaded from Broad GDAC Firehose. Somatic copy number alterations (SCNAs), mutation frequency, and other genomic data including TMB and fraction of genome altered were also obtained from available samples.

The British Columbia Cancer Agency (BCCA) dataset comprised of 69 microarray samples (39 CS, 30 NS) profiled using the Illumina WG-6 v3.0 BeadChip and the Memorial Sloan Kettering Cancer Center (MSKCC) dataset involved 192 samples (25 CS, 36 NS, 131 FS) profiled using Affymetrix HG-U133A Arrays. Both microarray datasets were obtained from the GEO database (GSE75037 and GSE31547, respectively).

Differentially expressed Gene (DEG) analysis

Differentially expressed gene (DEG) analysis was conducted between NS and CS samples to develop a gene signature associated with active smoking. For both TCGA and BCCA datasets, genes expressed at low levels were removed and in instances of multiple probes corresponding to a single gene, only the probe with the highest mean expression was retained. Normalization was applied to each dataset using the *EdgeR* package in R and significantly up- and down-regulated genes were obtained using the *limma* package [7].

The overlapping DEGs from these two independent analyses that were also present within the MSKCC dataset constituted our active smoking gene signature. Principal component analysis (PCA) was performed using gene expression data from the genes within the signature derived from DEG analyses using the *ggfortify* package [8]. Receiver operating characteristic (ROC) curves were constructed with the principle component 1 values from each dataset's PCA and respective areas under the curve (AUCs) were calculated to determine the ability of the gene signature to separate samples based on their NS and CS status.

Functional analysis of DEGs

To understand the functions and pathways associated with the genes within the smoking associated gene

signature, Gene Ontology (GO), [9, 10] specifically Biological Process terms, and KEGG [11] databases were used. The DAVID tool [12] allowed integration of GO terms and pathways into clusters and ShinyGO [13] was utilized for confirmatory analysis and visualization purposes.

Random forest from gene signature

A random forest model (RF) utilizing genes of the derived gene signature and sex as features was trained to predict NS and CS status for future application to FS samples. RFs were created for both RNA-Seq and microarray data to account for inherent differences in the two data types. Each RF utilized the default settings from the *randomForest* package [14]. The RF built from RNA-Seq data was trained on 70% of the NS and CS within the TCGA dataset and tested on the remaining 30%, each of which was selected as described below. The RF built from microarray data was trained on the BCCA dataset and tested on the NS and CS of the MSKCC dataset. *YuGene* transformation was applied to both microarray datasets for cross-platform data consistency [15]. Performance metrics used to evaluate each RF included ROC curve AUC, overall accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, courtesy of the *caret* package [16]. The train-test sets used for the TCGA RF were resampled 10 times and both RFs were built on 10 distinct seeds. The model with the AUC closest to the average among each dataset was selected for classification of FS.

Random forest classification of patients who had formerly smoked

The RFs categorize samples as NS or CS based on the proportion of trees voting for either status. For FS, these classifications are interpreted as being "NS-like" or "CS-like", respectively.

Using RF classified NS-like or CS-like status given to lung tumors of FS, the relationships between FS and different clinical and biological characteristics could be investigated. RF defined FS classes were compared to variables that have been used to delineate higher risk of lung cancer, which include individuals between the ages of 50 and 80 who have previously smoked that have quit within the last 15 years and have over 20 pack years of smoking history according to the United States Preventive Services Task Force (USPSTF) [17].

Genomic traits including TMB, fraction of genomic altered, and number of mutations were compared across samples of all smoking statuses in available data. Frequency of oncogenic driver mutations and sex were also analyzed between RF defined FS classes.

When analyzing relationships between FS class and other traits, Fisher's exact test was used for categorical variables and Wilcoxon test was used for continuous variables. Correlations were assessed by Pearson correlation coefficient. In any comparisons that involved FS class as well as true NS and CS, Benjamini-Hochberg multiple testing correction was applied.

Copy number assessment and mutational analysis

GISTIC 2.0 [18] was used to identify frequent SCNAs in all smoking status groups within the TCGA dataset. The parameters of q-value, confidence, and focal length were set with 0.05, 0.95, and 0.5, respectively.

A total of 220 samples had mutation data for comparison of mutational signatures between all smoking statuses in the TCGA dataset. This was analyzed using the *mutSignatures* package and comparisons were made between smoking status groups by Wilcoxon test. From the mutation data of 144 FS patients from TCGA, driver mutation frequencies of each gene were compared between NS-like and CS-like samples using Fisher's exact test. Multiple testing correction was subsequently applied with the Benjamini-Hochberg method.

DEG analysis in patients who had formerly smoked

As with the DEG analysis between NS and CS samples, DEGs between NS-like and CS-like FS in the TCGA dataset were identified using the *edgeR* and *limma* packages. The thresholds for DEG selection were |log2 fold change| >1 and adjusted p value < 0.01. Functional analysis on these DEGs were performed using the DAVID tool and visualized with ShinyGO.

Immune cell content assessment

CIBERSORTx, [19] a deconvolution algorithm, was employed to estimate the relative proportion of 22 types of immune cells in the tumor tissue through the gene expression levels of 547 genes. The normalized gene expression data of the FS in the TCGA dataset were uploaded to the CIBERSORTx web interface with parameters set at 1000 permutations and relative proportions mode. Differences in each immune cell population between FS classes were compared by Wilcoxon test and adjusted by Bonferroni correction. The same process was repeated for true NS and CS as controls.

Assessment of clinical outcomes

The pathological stage of NS-like and CS-like FS samples was evaluated by Fisher's exact test and Benjamini-Hochberg multiple testing correction. In the TCGA dataset, FS were also assessed for their pathological T and N stages.

Univariate Cox regression analysis was conducted in the FS of the TCGA dataset using the *survivalAnalysis* package [20]. This was done to determine if RF-classified FS class could serve as an independent prognostic

factor alongside age, sex, and pack year history, years since quitting.

Kaplan-Meier survival curves were constructed with FS in both TCGA and MSKCC datasets to understand differences in overall survival between NS-like and CS-like FS. The *survival* and *survminer* packages were chosen for survival analysis by log rank test and for visualization, respectively, due to their robust capabilities and compatibility with one another [21, 22].

Statistical analysis

All statistical analyses were done using R (version 4.3.0) and p values <0.05 were considered statistically significant. All visualization of data was conducted with the *ggplot2* and *ggpubr* packages unless otherwise stated.

Results

Derivation of an active smoking gene expression signature to separate patients who had never smoked and currently smoke

Although all FS share a history of tobacco use, the range of their smoking history and susceptibility to tobacco smoke means that lung carcinogenesis in some FS is inevitably attributable to smoking, while in others, it may be due smoking unrelated factors. We aimed to stratify the FS lung adenocarcinoma population and thus better understand their heterogeneity by building a model to classify FS based on their smoking induced gene expression. In order to determine genes related to active smoking for later classification of FS, differentially expressed gene (DEG) analysis between NS and CS in both TCGA and BCCA cohorts was conducted, which yielded 4515 and 203 DEGs, respectively. The overlap between these genes and the ones available within the MSKCC dataset resulted in a 123-gene signature (Fig. 1a). Construction of PCAs using the expression levels from these 123 genes

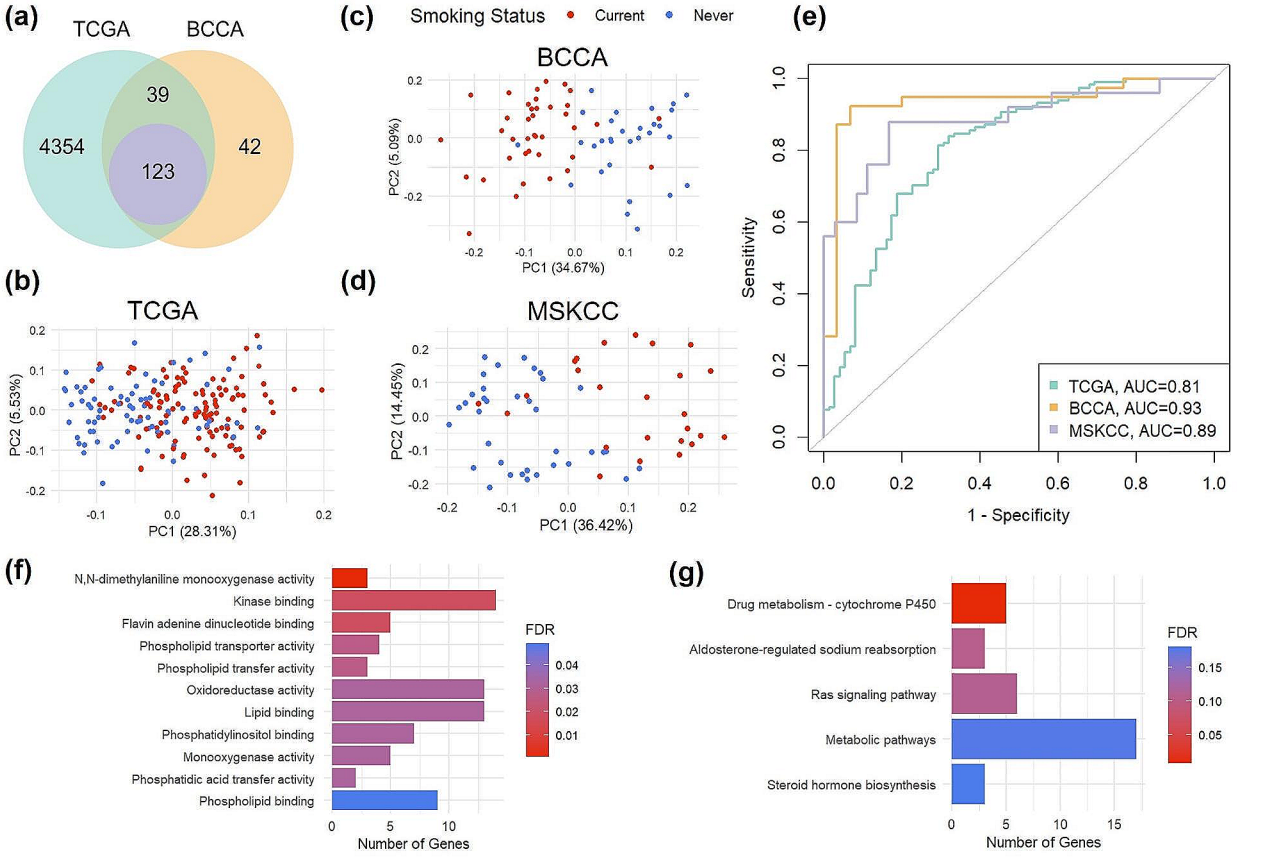


Fig. 1 Gene signature that discerns lung adenocarcinoma patients who had never smoked (NS) and currently smoked (CS) functionally relates to drug metabolism. **(a)** Overlap of DEGs between lung adenocarcinoma tumors of NS and CS from the TCGA and BCCA cohorts that are present within the MSKCC cohort (123 genes). **(b)** PCA of expression of 123 overlapped genes in NS and CS from TCGA ($n=193$), **(c)** BCCA ($n=69$), and **(d)** MSKCC ($n=61$). **(e)** ROC curves and AUCs generated from principle component 1 values for each sample, demonstrating strong separation of NS and CS by the selected 123 genes. **(f)** Functional analysis displaying significantly enriched Molecular Function Gene Ontology terms and **(g)** KEGG pathways from the 123 DEGs between NS and CS in TCGA and BCCA. FDR = false discovery rate; AUC = area under the curve

showed a visible separation between the NS and CS patients in all datasets, including within MSKCC, which was independent from the signature derivation process (Fig. 1b-d). A receiver operating characteristic area under the curve (AUC) comprised from each PCA's principal component 1 demonstrates that the 123-gene signature is robust in distinguishing NS and CS LUAD tumors. The AUCs for TCGA, BCCA, and MSKCC were 0.81, 0.93, and 0.89, respectively (Fig. 1e).

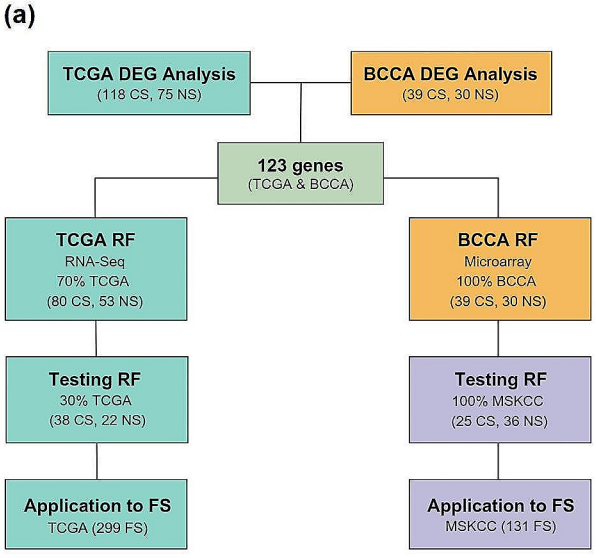
Functional analysis of the gene signature demonstrated that many of the 123 genes are related to regulation of monooxygenase activity (Fig. 1f). Molecular function GO terms that were most significantly enriched were "N, N-dimethylaniline monooxygenase activity", "kinase binding", and "flavin adenine dinucleotide binding". The genes associated with each term and whether they are more highly expressed in CS or NS are detailed in Supplementary Table 1.

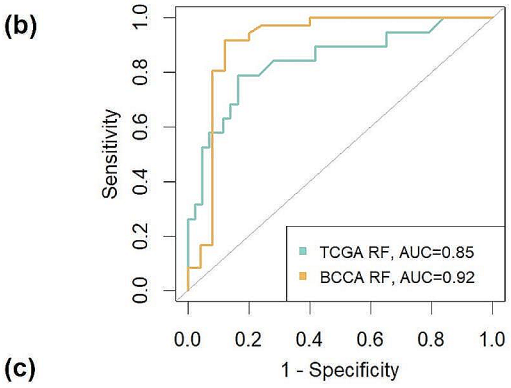
According to the KEGG pathway database, the 123 genes most commonly fell into drug metabolism by cytochrome P450, metabolic, aldosterone-regulated sodium reabsorption, and Ras signaling pathways (Fig. 1g). However, only the drug metabolism by cytochrome P450 pathway was significantly enriched (fold enrichment = 13.4, FDR = 0.007) and all the five genes that fall within this pathway (FMO3, FMO2, FMO4, MAOB, CYP3A5) are upregulated in NS tumor tissue.

Random forest models (RFs) were built to classify NS and CS LUAD patients with both RNA-seq data and microarray data using these 123 genes and sex as input features (Fig. 2a). These models were then validated by predicting NS and CS status from independent test data; the RNA-seq RF was trained on 70% of the TCGA dataset (n = 133) and tested on the remaining 30% (n = 60). The microarray RF was trained by the BCCA cohort (n = 69) and tested on the MSKCC data (n = 61). The AUCs from the validation these models were 0.85 and 0.92, respectively (Fig. 2b), with further performance metrics listed in the table of Fig. 2c.

Smoking induced gene expression correlates modestly with patient characteristics

The patients who had previously smoked from the TCGA cohort were defined as 72% (n = 213) CS-like and 28% (n = 86) NS-like according to our RF (Fig. 3a). The RF classifies a patient as NS-like or CS-like on a scale from





(c)

Metric	TCGA RF	BCCA RF
AUC	0.848	0.916
Overall Accuracy	0.806	0.869
Sensitivity	0.902	0.848
Specificity	0.644	0.883
PPV	0.817	0.834
NPV	0.793	0.894

Fig. 2 Pipeline of analysis and demonstration that random forest models (RFs) trained on gene signature and sex data can accurately distinguish between NS and CS with lung adenocarcinoma. **(a)** RF development and validation pipeline to differentiate NS and CS lung adenocarcinoma tumors. **(b)** ROC curves and AUCs generated from inputting previously unseen test data into random forest models trained on gene signature and sex data from TCGA (RNA-seq) and BCCA (microarray) datasets. **(c)** Table of random forest performance metrics. TCGA data was resampled 10 times for test and train sets and both models were built on 10 separate seeds and mean metrics are shown. AUC = area under the curve, PPV = positive predictive value, NPV = negative predictive value

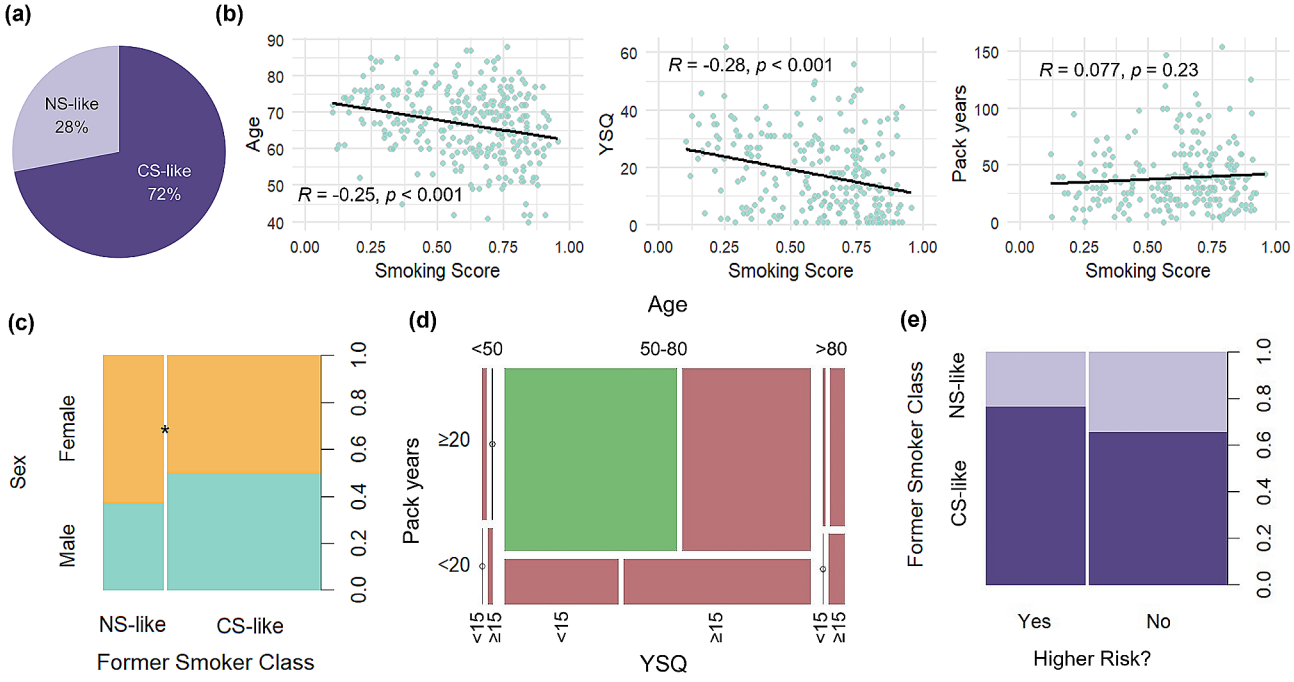


Fig. 3 Clinical characteristics correlate mildly with smoking induced gene expression and patients who formerly smoked (FS) with lung adenocarcinoma are a diverse demographic. **(a)** Percentage of FS categorized as either NS-like ($n = 86$) or CS-like ($n = 213$) by random forest model (RF). **(b)** Age, years since quitting (YSQ), and pack years depending on smoking score as predicted by RF. **(c)** Proportions of sex in FS relative to their RF classified CS- or NS-like status. **(d)** Mosaic plot of FS who would be higher risk and thus more likely recommended for screening (green) or lower risk (red). **(e)** Proportion of RF classified NS- or CS-like FS that are high or low risk. *$p < 0.05$

zero to one, with a score less than 0.5 being NS-like and a score greater than 0.5 being CS-like. Correlative analyses showed that age and years since quitting have a significant but weak negative correlation with smoking score (Fig. 3b). The former is likely due to the fact that with higher age, there is a greater amount of time for years since quitting to accrue; as such, age and years since quitting are correlated with one another (data not shown). FS in this cohort were defined as those who had quit for over a year; to assess the change in proportion of NS-like FS, this group was further parsed into those who had quit more than five, 10, or 15 years. The percentage of NS-like FS increased from 29.5 to 31.4%, 34.0%, and 39.1%, respectively, although none of these proportions are significantly different. This aligns with previous findings that some smoking related genes decrease in expression linearly as time since quitting increases, while other genes' expression remain resiliently expressed for years, potentially explaining why our NS-like status, which is defined by gene expression, is only modestly correlated with years since quitting [23]. Surprisingly, there was no correlation between smoking score in the FS and pack years (Fig. 3b). Finally, although sex is not currently part of lung cancer screening criteria, it was found that a slightly higher proportion of female FS patients were classified as NS-like compared to males (Fig. 3c).

The RF was also applied to the MSKCC dataset, which classified 57 FS to be CS-like and 74 to be NS-like. Although there was not enough information to assess the theoretical risk of these patients, available data indicated that pack years had a slight positive correlation with CS-like status and recapitulated that a greater proportion of female FS are NS-like compared to males (Figure S1a-b). Together, this suggests that clinical factors are not strongly correlated with active smoking gene expression levels in FS.

To understand how high risk traits relate to FS based on their active smoking gene expression levels, FS LUAD patients were presented to the 123-gene RF to be classified as never smoker-like (NS-like) or current smoker-like (CS-like). Only 41.7% of FS with LUAD would have been considered to be at relatively higher risk for lung cancer according to the USPSTF (Supplementary Table 2). Furthermore, of the FS who are at lower risk, 38.3% (79/206) were still categorized as CS-like (Fig. 3d-e). This demonstrates that, according to our model, a sizeable number of FS falling outside of high lung cancer risk attributes harbored tumors with high smoking induced gene expression.

NS-like and CS-like formerly smoking patient subgroups have significantly different genomic profiles

In order to determine whether other underlying biological differences are associated with active smoking gene expression levels, RF-classified FS were compared to different biological features. CS-like FS have markedly higher TMB than NS-like FS and their TMB is relatively similar to those of true CS. It is notable that relative to true NS, NS-like FS have significantly greater TMB. A similar stepwise trend from NS to NS-like FS to CS-like FS is also present in the number of mutations between groups. Regarding the fraction of the genome that is altered in each group, there is no significant difference between NS and NS-like tumors. However, both of these groups have distinguishable differences compared to both CS-like and CS tumors, which again demonstrate distinct genomic differences between NS-like and CS-like FS (Fig. 4a). Combining these three genomic traits through a composite score, having quit smoking for over 15 years was also able to separate FS by genomic features in addition to RF predicted class of FS (Figure S2). Taken together, these findings extend beyond the established knowledge that NS and CS possess pronounced genomic characteristics (Figure S3). Moreover, this reveals that FS occupy an intermediate position between these two groups that can be further delineated into two distinct groups based on our active smoking gene expression signature.

Mutational signature analysis revealed different mutational signature profiles between NS-like and CS-like FS

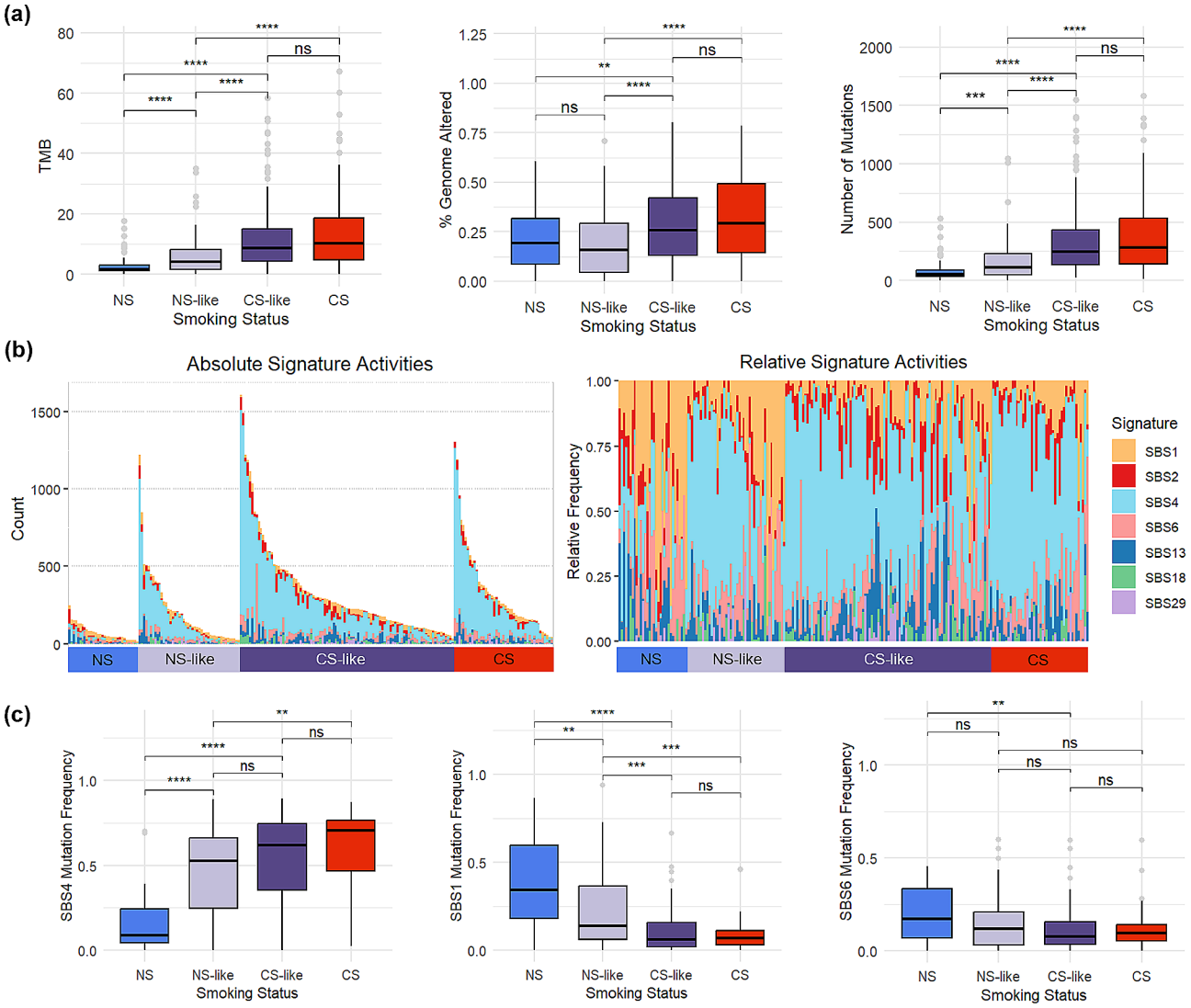


Fig. 4 Genomic profiles between NS- and CS-like FS with lung adenocarcinoma are significantly different. **(a)** Genome related measures and **(b)** absolute and relative frequencies of mutational signatures that have been previously detected in lung cancer between true NS, true CS, and RF classified NS- and CS-like FS. **(c)** Relative levels of SBS4 (tobacco), SBS1 (ageing) and SBS6 (DNA mismatch repair) mutational signatures between different smoking statuses. TMB = tumor mutational burden, ns = not significant, *$p < 0.05$, **$p < 0.001$, ***$p < 0.0001$, ****$p < 0.00001$

with LUAD (Fig. 4b). Specifically, relative level of SBS4, a well-defined COSMIC signature associated with tobacco mutagens, [24] is relatively higher in CS-like than NS-like FS (Figure S4) although the difference is only upon comparing absolute count of SBS4 between the two groups (Fig. 4c). This aids to verify the gene signature, which was developed to separate FS by smoking related gene expression. Furthermore, APOBEC signature SBS2 was relatively higher in NS than other groups as has been previously reported (Figure S4). True NS have the highest relative levels of SBS1 and SBS6, representing ageing and deficient mismatch repair, [24] which were significantly greater than CS-like FS but not NS-like FS, suggesting a greater impact of endogenous signatures in tumors of NS and NS-like FS (Fig. 4c).

The frequency of main oncogenic driver alterations was not significantly different between NS-like and CS-like tumors. The proportions of KRAS, EGFR, and ALK alterations were relatively similar among FS in TCGA and proportions of KRAS and TP53 alterations were also not significantly different between FS groups in MSKCC (Figure S5). Although the proportion of those with EGFR mutation was slightly higher in NS-like FS than that of CS-like FS, the significance of this difference does not hold after multiple testing correction.

SCNAs were reported in the TCGA dataset, which demonstrated that CS-like FS exhibit far more frequently altered regions of amplification and deletion compared to NS-like patients (Fig. 5a-b). In addition, compared to NS-like FS tumors, CS-like tumors demonstrated greater relative copy number alterations in multiple regions across

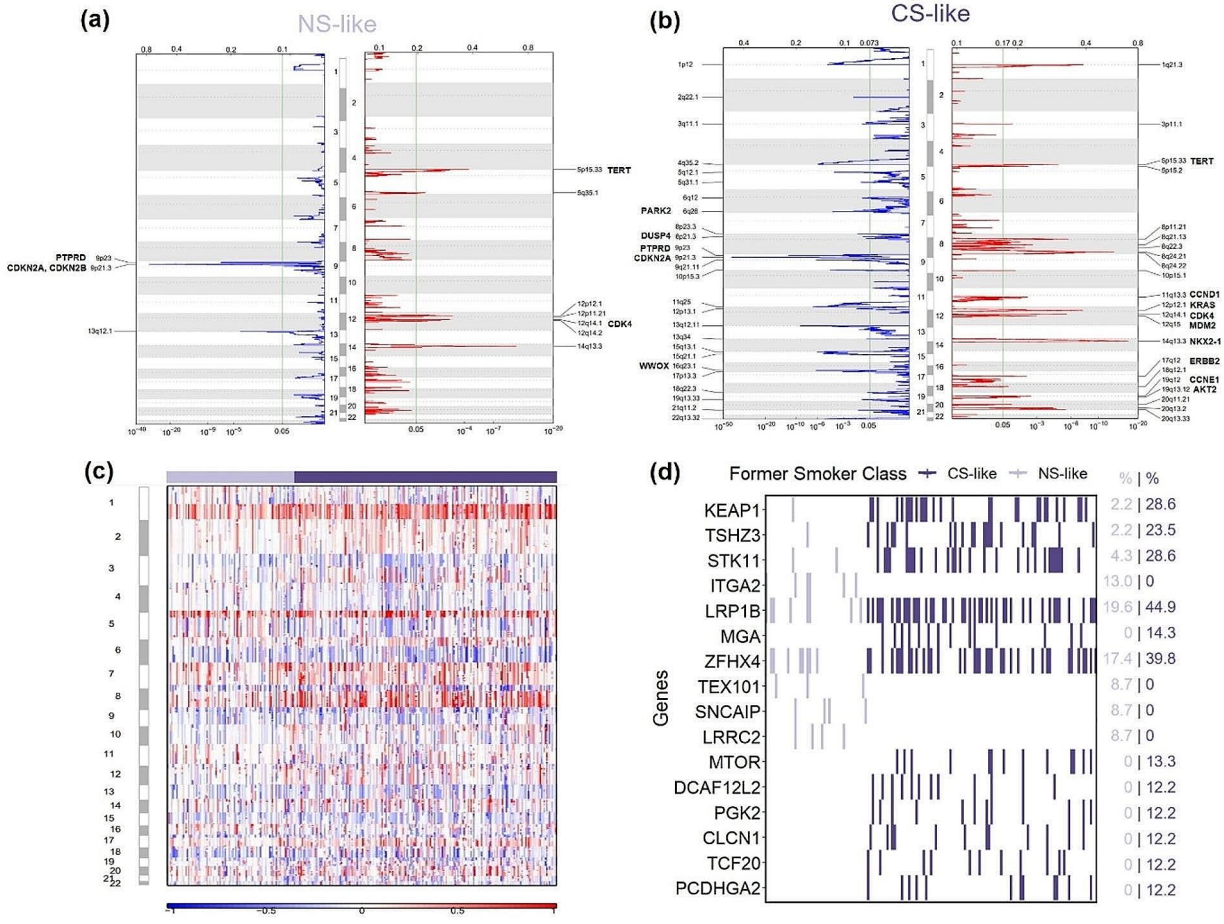


Fig. 5 Genomic analyses show greater somatic copy number alterations in CS-like than NS-like FS with lung adenocarcinoma, and genes with significantly different mutation frequencies between FS from TCGA are shown. GISTIC 2.0 analysis of copy number amplifications (red) and deletions (blue) of **(a)** NS-like ($n=85$) and **(b)** CS-like ($n=213$) FS with lung adenocarcinoma patients within TCGA. Significance is delineated by a green line at 0.05 representing the false discovery rate corrected p-value. The corresponding chromosome regions are labeled, and genes of interest are indicated. **(c)** Somatic copy number profiles of all FS in TCGA. Copy number amplifications (red) and losses (blue) are plotted as a heatmap with samples on the x-axis and chromosomes on the y-axis. **(d)** Genes with significantly different mutation frequencies between NS-like and CS-like FS in TCGA (Fisher's exact test, $n=144$, $p<0.01$)

the genome (Fig. 5c). This recapitulates the trend found in percentage of the genome altered between the two groups of FS. Regions that were significantly amplified in CS-like tumors held many genes known to be associated with cancer development, including KRAS, CDK4, and TERT. Comparing FS to true NS and CS, CS have comparable SCNAs to CS-like FS. However, NS-like FS have the least somatic copy number deletions and amplifications, even compared to true NS (Figure S6).

Significantly different mutation frequency between NS-like and CS-like FS was identified in 80 genes ($p<0.05$), with the top 16 genes ($p<0.01$) shown in Fig. 5d. CS-like tumors had significantly more highly mutated genes than NS-like samples, with the most significant genes being KEAP1, TSHZ3, and STK11. Genes that were significantly more frequently mutated in NS-like samples were ITGA2, TEX101, SNCAIP, and LRRC2, which had mutation frequencies ranging from 8.7 to 13.0% compared to 0% in CS-like samples (Fig. 5d). Although multiple testing adjustment did not retain any significant genes, the exploratory nature of this analysis highlights biological differences between subgroups of FS.

Tumors of CS-like patients NS-like tumors have different transcriptional and immune profiles

Upon investigating transcriptomic differences between NS-like and CS-like FS, the majority of significantly enriched GO terms relate to the cell cycle (Fig. 6a). This aligns with KEGG pathway analysis indicating the cell cycle as the most significantly enriched pathway, followed by drug metabolism, metabolism of xenobiotics, and ECM-receptor interaction (Fig. 6b).

Tumor immune microenvironment was investigated within FS using CIBERSORTx, which revealed that heterogeneity exists within NS-like and CS-like FS (Fig. 6c). In addition, differences between these subgroups of FS exist in that CS-like FS have greater proportions of activated CD4 memory T cells while NS-like FS have greater proportions of resting CD4 memory T cells, monocytes, dendritic cells, and mast cells after Bonferroni correction

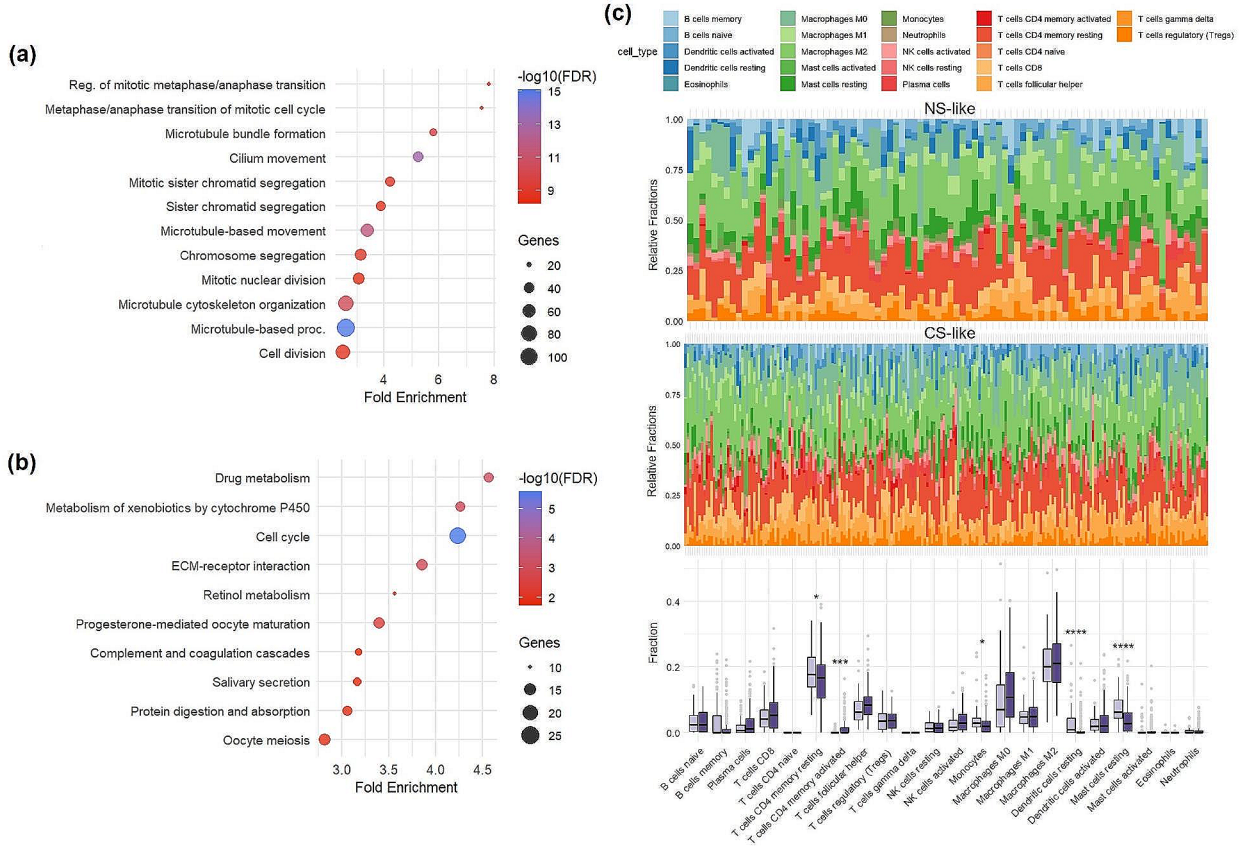


Fig. 6 Gene expression differences between NS-like and CS-like FS with lung adenocarcinoma revolve functionally around the cell cycle and drug metabolism and myeloid lineage immune cells are more abundant in NS-like FS. **(a)** Functional analysis displaying enriched Biological Process Gene Ontology terms and **(b)** enriched KEGG pathways from the 1050 DEGs between NS-like and CS-like FS in TCGA. **(c)** Relative fraction of 22 immune cell types in NS-like ($n=86$) and CS-like ($n=213$) FS in TCGA and comparisons of tumor-infiltrating immune cells between NS-like and CS-like FS (Wilcoxon test and Bonferroni correction, $n=193$, *$p<0.05$, ****$p<0.0001$). FDR = false discovery rate

(Fig. 6c). Comparing true CS to NS, it is observed that myeloid lineage cells, specifically dendritic cells and mast cells, are present in higher proportions among NS. Conversely, lymphoid lineage plasma cells and T follicular helper cells levels are proportionally greater in CS (Figure S7).

CS-like patients who had formerly smoked tend to have LUAD in later pathological stage and worse overall survival

To assess how active smoking gene expression is related to tumor characteristics and clinical outcomes, RF-classified FS were then associated with variables related to tumor stage and overall survival. In the TCGA dataset, CS-like patients more frequently present with tumors at advanced pathological stages compared to NS-like patients. After multiple testing adjustment, CS-like FS patients are significantly more likely to have stage III tumors compared to NS-like patients, whereas those who are NS-like have a higher likelihood of being in stage I than CS-like patients (Fig. 7a). The TNM staging data further illustrate that CS-like patients have a trend towards larger tumor size and increased lymph node involvement compared to NS-like patients. Notably, this difference in proportion is significant between T1 and T2 tumors as well as N0 and N2 tumors even after multiple testing correction. There were not enough metastatic events in the formerly smoking patient cohort to test for differences in proportion in metastasis (Fig. 7a).

Kaplan-Meier survival analysis of NS-like and CS-like FS followed by log rank test for statistical significance established that overall survival differs between the two subgroups of FS. Prognosis is 17.5 months longer in NS-like FS than CS-like FS, with a median survival time of 59.1 months compared to 41.6 months (Fig. 7b). This is consistent with CS-like patients demonstrating higher pathological stages of LUAD, which directly correlates to poorer survival outcomes (data not shown). In the MSKCC dataset, there is no significant difference in pathological stage between NS-like and CS-like groups (Fig. 7c). However, similarly to the TCGA dataset, NS-like FS have significantly longer overall survival than CS-like FS, with a median survival difference of 18.6 months (Fig. 7d). Importantly, other univariate survival analyses conducted demonstrated that NS-like FS are 49% less likely to die than CS-like FS (HR=0.49, CI=0.29, 0.81) and no other established clinical variables significantly affect overall survival in FS, including years since quitting, pack years, age, and sex (Fig. 7e). This suggests that our active smoking classifier is a more effective independent prognostic factor than any of these aforementioned clinical variables.

Discussion

A portion of the heterogeneity in the tumor biology of FS is due to the varied influence of smoking on lung carcinogenesis. We demonstrate that the FS population can be meaningfully segregated by their smoking related gene expression. Our study allowed FS to be classified as CS-like or NS-like based on the expression of a 123 gene signature that is associated with active smoking, defined through assessment of true NS and CS LUAD tumors.

Our work shows that despite lack of correlation with clinical characteristics, smoking related gene expression has relevance in predicting other aspects of LUAD tumor biology as well as overall survival in FS. The CS-like FS class had significantly greater genomic disturbances than NS-like FS even though it did not significantly correlate with smoking pack year history or years since quitting. RF-predicted FS classes also showed a divide in mutational signature profiles, where CS-like FS had relatively greater levels of tobacco mutagen signature SBS4 and NS-like FS had relatively higher levels of endogenous signatures SBS1, SBS6, and SBS2. These trends have been previously reported in true CS and NS patients with LUAD, respectively, supporting a clear and biologically relevant subdivision within FS [25].

Higher TMB is associated with greater sensitivity to immunotherapy in NSCLC [26]. Although FS have been shown to have significantly poorer response to immunotherapy than CS, [5] it is possible that a subset of FS identifiable as CS-like FS may confer great benefit since they have comparable TMB, fraction of genome altered, mutation counts, and copy number alterations to CS. This supports previous findings that FS are separable by TMB based on years since quitting, indicating distinct biological subgroups within the FS population [6]. In addition, tumors of CS-like FS have significantly higher proportions of activated CD4 memory T cells, and high levels of tumor infiltrating lymphocytes are well documented to predict good response to PD-1 blockade [27]. A caveat is that CS-like FS harbor significantly higher mutation frequencies in KEAP1 and STK11, both of which are associated with poor response to immunotherapy even with high TMB [28, 29]. This further refines the subgroup that may exist within FS who would benefit from immunotherapy and warrants further exploration in responses to this treatment specifically in patients who had previously smoked.

Aside from differing immune profiles and genomic characteristics, DEGs between FS classes were functionally related to the cell cycle, whose dysregulation is a hallmark feature of cancer and has been observed to be more highly disrupted in true CS than NS [30]. Another highly enriched pathway from the DEGs between FS groups is metabolism of xenobiotics; genes from this pathway are more highly upregulated in NS-like than CS-like FS.

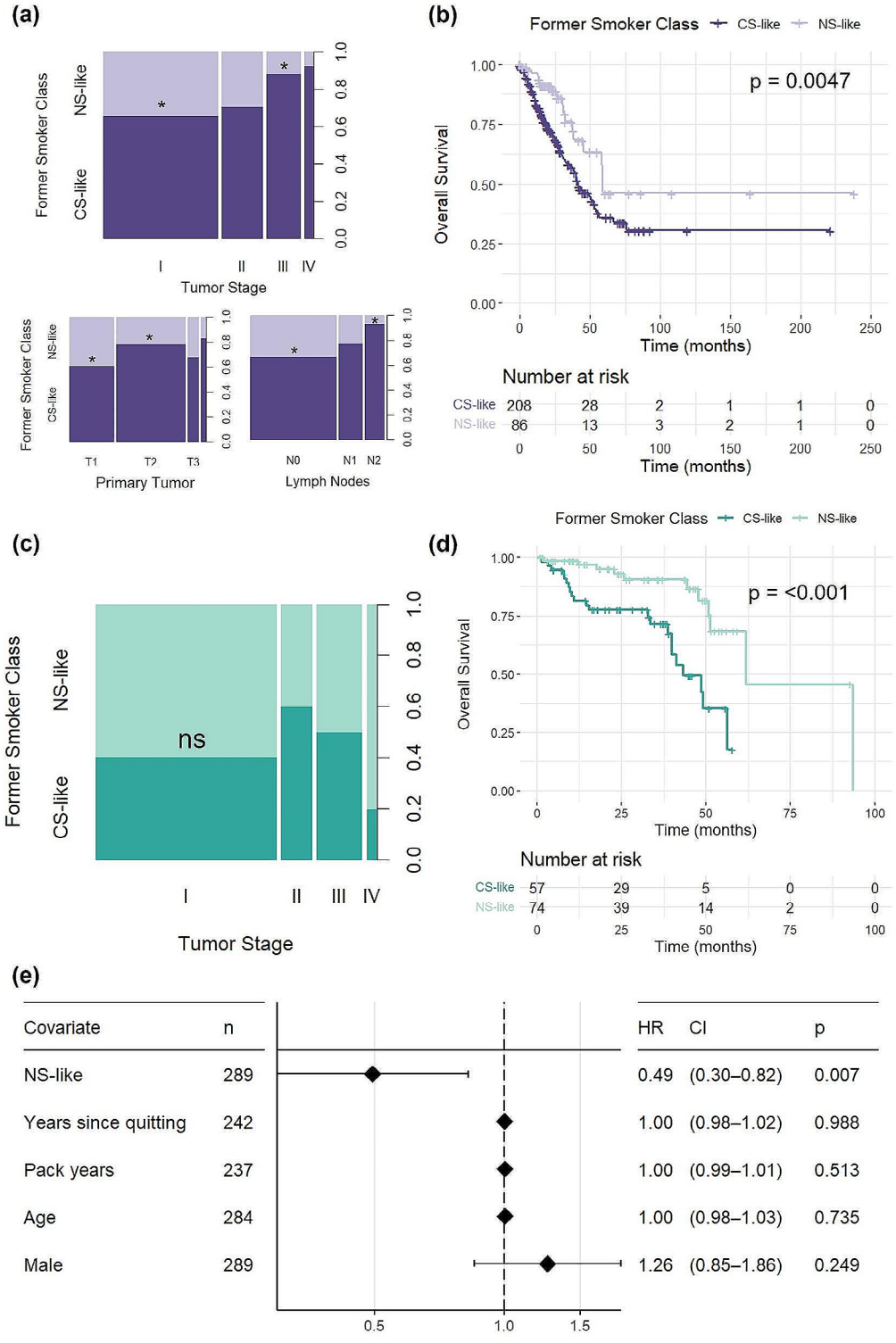


Fig. 7 CS-like FS with lung adenocarcinoma in TCGA have more advanced tumors and their overall survival is significantly worse than NS-like FS in both TCGA and MSKCC. Proportion of RF classified NS- and CS-like FS in different tumor stages and classifications in the **(a)** TCGA and **(c)** MSKCC cohort. **(b)** Kaplan Meier survival curve and number at risk table showing overall survival between NS- and CS-like FS. Median survival difference is 17.5 months in the TCGA cohort and **(d)** median survival difference is 18.6 months in the MSKCC cohort. **(e)** Univariate Cox regression of overall survival depending on RF-classified status in FS and other clinical characteristics in the TCGA cohort

Impairment in the metabolism of foreign compounds aligns with the idea that FS are more CS-like if their lung biology is less adept to process the compounds in tobacco smoke. Genes related to toxin removal from the metabolism of xenobiotics pathway including FMO3 [31] and CYP3A5 [32] are all down-regulated in CS-like tumors, potentially contributing to development of smoking induced tumors.

Although CS-like FS had significantly greater SCNAs as well as more chromosomal regions altered that belong to known tumor suppressors, some of these regions were also shared with NS-like FS, including CDKN2A. The overlap of these regions may suggest that these genes are consistently involved with LUAD of FS. Conversely, the abundance of other canonical cancer-related genes in CS-like FS may suggest distinct routes to tumorigenesis compared to NS-like tumors.

The subgroups of FS were also associated with tumor stage and overall survival. CS are more likely to be detected in advanced stage disease than FS, [33] and our findings follow this in that CS-like FS harbor a significantly higher proportion of late stage tumors than NS-like FS in the TCGA cohort. This may be a contributing factor to NS-like FS having a 17.5 month longer overall survival. However, it should be noted that there was no correlation between stage and FS classification in the MSKCC cohort, although it was also found that NS-like FS had significantly longer overall survival. In addition, no other clinical variable in univariate analyses was able to predict overall survival the way that the RF-classified FS class could. This may suggest that there are properties of certain tumors that are driven by smoking related gene expression that accelerate tumor progression, and that our gene signature may be able to identify these higher risk patients as candidates for more aggressive treatment regimens. The increased overall survival in NS-like FS is further supported by higher proportions of resting CD4 T cells, monocytes, resting dendritic cells, and resting mast cells, all of which have been previously significantly correlated with higher overall survival in LUAD patients of all smoking statuses [34]. Thus, this gene signature may be able to predict prognosis in LUAD FS patients and is a step towards personalized medicine for FS.

In our study probing tumors of LUAD FS by smoking related gene expression, there is no significant association between CS-like FS and patients who would have been considered at higher risk of developing lung cancer due to their age, pack year history, and years since quitting. Lung cancer risk over time is unique to the individual as some FS have been shown to remain at elevated risk despite quitting for more than 25 years [35]. Our work shows that CS-like tumors can present in FS who have quit for a long time, suggesting that persistently modified gene expression or genomic alterations may be a possible explanation for persistently high risk in some FS. This supports previous studies that have proposed a personalized approach to determining risk as being optimal for FS [4, 35] and research on FS should continue to investigate refinements towards early detection, including understanding genomic features of normal tissue in high risk populations.

Limitations

A limitation of this study is the lack of public databases separating current from FS with detailed smoking history, such as pack years and years since quitting. A previous retrospective study on lung cancer screening eligibility found that 36% of patients did not have smoking history documented in their medical records system among nearly 500 patients assessed [36]. This translated to a limited sample size in our study and restricted the ability to bridge tumor transcriptomics and genomics with clinical characteristics of FS from several public datasets. Considering the heterogeneity within a tumor and the small part extracted for transcriptomic and genomic sequencing, not only a larger sample size but a standardized protocol for extracting tumor samples could be established in the future for more generalizable results. Another future direction would be to integrate methylation data in the analyses to determine if it contributes to smoking related gene expression, but this data is not yet available. A further limitation is that all cohorts utilized in this study originated from North American centers and patients were primarily Caucasian. This calls for further investigation of FS with detailed smoking history in other geographical areas with more diverse racial backgrounds to understand if the results from this study are location- or race-specific.

There was also a lack of normal tissue and longitudinal datasets, which limit the direct applicability of these findings for screening and early detection purposes. Instead, this work is able to indirectly show that the use of clinical characteristics in screening may not adequately capture people who are at the highest risk of aggressive smoking related cancer. Our study serves as a proof of concept of the heterogeneity within tumor biology of FS that can be leveraged in patient care. Future directions include replicating this analysis in FS with more diverse geographical and racial backgrounds as well as in healthy patients longitudinally to determine the dynamics of smoking related gene expression over time.

Conclusions

FS are a diverse population not only due to variability of pack years and years since quitting, but due to differences in tumor biology. This study demonstrated the ability to stratify FS by smoking related gene expression that had weak correlations with clinical characteristics and

smoking history, but was associated with underlying factors including genomic alterations, immune infiltration and clinical factors including overall survival. This demonstrates the potential of considering gene expression in the clinical care of FS as well as motivates future research that focuses on FS with lung cancer to offer them personalized care as this population continues to grow.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05437-8.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
This work was supported by the Canadian Institutes of Health Research and the Terry Fox Research Institute. PW is supported by the Canadian Institute of Health Research Doctoral Award.

Author contributions
Peiyao Wang: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. Resources. Raymond Ng: Conceptualization, Methodology. Stephen Lam: Resources, Writing – Conceptualization, Review & Editing. William W. Lockwood: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding Acquisition, Project administration.

Funding
This work was funded by the Canadian Institutes of Health Research (CIHR, PJS-186324 and PJT – 169129) grants to William W. Lockwood and a Terry Fox Research Institute Program Project Grant to Stephen Lam and William W. Lockwood.

Data availability
The data that support the findings of this study are openly available in the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/gds/) at GSE75037 and GSE3147 as well as in Broad GDAC Firehose (https://gdac.broadinstitute.org/) under 'Lung adenocarcinoma'.

Declarations

Ethical approval
Not applicable.

Competing interests
The authors declare no potential conflicts of interest and no funding was received for this work.

Received: 11 January 2024 / Accepted: 26 June 2024
Published online: 08 July 2024

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. Cancer J Clin. 2021;71(3):209–49. https://doi.org/10.3322/caac.21660.
2. Jeon J, Holford TR, Levy DT, et al. Smoking and Lung Cancer Mortality in the United States from 2015 to 2065: a comparative modeling Approach. Ann Intern Med. 2018;169(10):684–93. https://doi.org/10.7326/M18-1250.
3. Reitsma M, Kendrick P, Anderson J, et al. Reexamining rates of decline in Lung Cancer Risk after Smoking Cessation. A Meta-analysis. Annals ATS. 2020;17(9):1126–32. https://doi.org/10.1513/AnnalsATS.201909-659OC.
4. Faselis C, Nations JA, Morgan CJ, et al. Assessment of Lung Cancer risk among smokers for whom Annual Screening is not recommended. JAMA Oncol. 2022;8(10):1428–37. https://doi.org/10.1001/jamaoncol.2022.2952.
5. Delasos L, Wei W, Hassan KA, Pennell NA, Patil P, Stevenson J. Clinical outcomes with pembrolizumab-based therapies in Recurrent/Refractory NSCLC after chemoradiation and Consolidative Durvalumab. Clin Lung Cancer. 2023;24(6):e205–13. https://doi.org/10.1016/j.cllc.2023.04.008.
6. Song K, Bi JH, Qiu ZW, et al. A quantitative method for assessing smoke associated molecular damage in lung cancers. Transl Lung Cancer Res. 2018;7(4):439–49. https://doi.org/10.21037/tlcr.2018.07.01.
7. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007.
8. Tang Y, Horikoshi M, Li W. Ggfortify: Unified Interface to visualize statistical results of Popular R packages. R J. 2016;8(2):474. https://doi.org/10.32614/RJ-2016-060.
9. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
10. The Gene Ontology Consortium, Aleksander SA, Balhoff J, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1). https://doi.org/10.1093/genetics/iyad031.
11. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
12. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–21. https://doi.org/10.1093/nar/gkac194.
13. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36(8):2628–9. https://doi.org/10.1093/bioinformatics/btz931.
14. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
15. Lê Cao KA, Rohart F, McHugh L, Korn O, Wells CA, YuGene. A simple approach to scale gene expression data derived from different platforms for integrated analyses. Genomics. 2014;103(4):239–51. https://doi.org/10.1016/j.ygeno.2014.03.001.
16. Kuhn M. Building Predictive models in R using the Caret Package. J Stat Softw. 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05.
17. US Preventive Services Task Force. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2021;325(10):962–70. https://doi.org/10.1001/jama.2021.1117.
18. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. https://doi.org/10.1186/gb-2011-12-4-r41.
19. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82. https://doi.org/10.1038/s41587-019-0114-2.
20. Wiesweg M, survivalAnalysis. High-Level Interface for Survival Analysis and Associated plots. Comprehensive R Archive Network. Published 2022. https://CRAN.R-project.org/package=survivalAnalysis.
21. Therneau T. A Package for Survival Analysis in R. Comprehensive R Archive Network. Published 2023. https://CRAN.R-project.org/package=survival.
22. Kassambara A, Kosinski M, Biecek P. Survminer: drawing Survival curves using ggplot2. Comprehensive R Archive Network. Published 2021. https://CRAN.R-project.org/package=survminer.
23. Bossé Y, Postma DS, Sin DD, et al. Molecular signature of smoking in human lung tissues. Cancer Res. 2012;72(15):3753–63. https://doi.org/10.1158/0008-5472.CAN-12-1160.
24. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101. https://doi.org/10.1038/s41586-020-1943-3.
25. Wang P, Sun S, Lam S, Lockwood WW. New insights into the biology and development of lung cancer in never smokers-implications for early detection and treatment. J Transl Med. 2023;21(1):585. https://doi.org/10.1186/s12967-023-04430-x.
26. Ricciuti B, Wang X, Alessi JV, et al. Association of High Tumor Mutation Burden in non–small cell lung cancers with increased Immune Infiltration and Improved Clinical outcomes of PD-L1 Blockade Across PD-L1 expression levels. JAMA Oncol. 2022;8(8):1160–8. https://doi.org/10.1001/jamaoncol.2022.1981.

27. Balança CC, Salvioni A, Scarlata CM, et al. PD-1 blockade restores helper activity of tumor-infiltrating, exhausted PD-1^{hi}CD39^{+} CD4 T cells. JCI Insight. 2021;6(2). https://doi.org/10.1172/jci.insight.142513.
28. Marinelli D, Mazzotta M, Scalera S, et al. KEAP1-driven co-mutations in lung adenocarcinoma unresponsive to immunotherapy despite high tumor mutational burden. Ann Oncol. 2020;31(12):1746–54. https://doi.org/10.1016/j.annonc.2020.08.2105.
29. Malhotra J, Ryan B, Patel M, et al. Clinical outcomes and immune phenotypes associated with STK11 co-occurring mutations in non-small cell lung cancer. J Thorac Dis. 2022;14(6):1772–83. https://doi.org/10.21037/jtd-21-1377.
30. Hu Y, Chen G. Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation. Oncol Lett. 2015;10(3):1350–70. https://doi.org/10.3892/ol.2015.3462.
31. Perez-Paramo YX, Chen G, Ashmore JH, et al. Nicotine-N'-oxidation by flavin monooxygenase enzymes. Cancer Epidemiol Biomarkers Prev. 2019;28(2):311–20. https://doi.org/10.1158/1055-9965.EPI-18-0669.
32. Ingelman-Sundberg M. Polymorphism of cytochrome P450 and xenobiotic toxicity. Toxicology. 2002;181–182:447–52. https://doi.org/10.1016/s0300-483x(02)00492-4.
33. Tindle HA, Stevenson Duncan M, Greevy RA, et al. Lifetime smoking history and risk of Lung Cancer: results from the Framingham Heart Study. JNCI: J Natl Cancer Inst. 2018;110(11):1201–7. https://doi.org/10.1093/jnci/djy041.
34. Guan M, Jiao Y, Zhou L. Immune Infiltration Analysis with the CIBERSORT Method in Lung Cancer. Dis Markers. 2022;2022:3186427. https://doi.org/10.1155/2022/3186427.
35. Pinsky PF, Zhu CS, Kramer BS. Lung cancer risk by years since quitting in 30 + pack year smokers. J Med Screen. 2015;22(3):151–7. https://doi.org/10.1177/0969141315579119.
36. Thuppal S, Hendren JR, Colle J, et al. Proactive recruitment strategy for patient identification for Lung Cancer Screening. Annals Family Med. 2023;21(2):119–24. https://doi.org/10.1370/afm.2905.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.