ORIGINAL ARTICLE

Tobacco Smoking-Related Mutational Signatures in Classifying Smoking-Associated and Nonsmoking-Associated NSCLC

Sophie M. Ernst, MD,[a] Joanne M. Mankor, MD,[a] Job van Riet, PhD,[b]
Jan H. von der Thüsen, MD, PhD,[c] Hendrikus J. Dubbink, PhD,[c]
Joachim G. J. V. Aerts, MD, PhD,[a] Adrianus J. de Langen, MD, PhD,[d]
Egbert F. Smit, MD, PhD,[f] Anne-Marie C. Dingemans, MD, PhD,[a,*]
Kim Monkhorst, MD, PhD[e]

[a]Department of Respiratory Medicine, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[b]Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[c]Department of Pathology, Erasmus University Medical Center, Rotterdam, The Netherlands
[d]Department of Thoracic Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[e]Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[f]Department of Pulmonology, Leiden University Medical Center, Leiden, The Netherlands

Received 19 July 2022; revised 9 November 2022; accepted 29 November 2022
Available online - XXX

ABSTRACT

Introduction: Patient-reported smoking history is frequently used as a stratification factor in NSCLC-directed clinical research. Nevertheless, this classification does not fully reflect the mutational processes in a tumor. Next-generation sequencing can identify mutational signatures associated with tobacco smoking, such as single-base signature 4 and indel-based signature 3. This provides an opportunity to redefine the classification of smoking- and nonsmoking-associated NSCLC on the basis of individual genomic tumor characteristics and could contribute to reducing the lung cancer stigma.

Methods: Whole genome sequencing data and clinical records were obtained from three prospective cohorts of metastatic NSCLC ($N = 316$). Relative contributions and absolute counts of single-base signature 4 and indel-based signature 3 were combined with relative contributions of age-related signatures to divide the cohort into smoking-associated ("smoking high") and nonsmoking-associated ("smoking low") clusters.

Results: The smoking high ($n = 169$) and smoking low ($n = 147$) clusters differed considerably in tumor mutational burden, signature contribution, and mutational landscape. This signature-based classification overlapped considerably with smoking history. Yet, 26% of patients with an active smoking history were included in the smoking low cluster, of which 52% harbored an *EGFR*/*ALK*/*RET*/*ROS1* alteration, and 4% of patients without smoking history were included in the smoking high cluster. These discordant samples had similar genomic contexts to the rest of their respective cluster.

*Corresponding author.

Dr. Ernst and Dr. Mankor contributed equally to this work.

Disclosure: Dr. von der Thüsen reports receiving research grants and institutional fees from AstraZeneca, and Roche Diagnostics; personal fees from Eli Lilly, Johnson&Johnson, Merck Sharp & Dohme, Pfizer, and Amgen, outside of the submitted work. Dr. Dubbink reports receiving research funding and support from AstraZeneca, Merck Sharp & Dohme, and Illumina; personal fees from AbbVie, AstraZeneca, Bayer, Janssen, Eli Lilly, Merck Sharp & Dohme, Pfizer, and Bayer, outside of the submitted work. Dr. Aerts reports receiving institutional fees from Eli Lilly, Amphera, Merck Sharp & Dohme, Takeda, BIOCAD, Pfizer, and Bristol-Myers Squibb; having stock/ownership interest in Amphera and PamGene, outside of the submitted work. Dr. de Langen reports receiving grants from Bristol-Myers Squibb, Merck Sharp & Dohme, AstraZeneca and Boehringer Ingelheim; non-financial support from Merck Serono and Roche, outside the submitted work. Dr. Dingemans reports receiving institutional fees from Roche, Takeda, Sanofi, Amgen, Bayer, Eli Lilly, Boehringer Ingelheim and AstraZeneca, outside of the submitted work. Dr. Monkhorst reports receiving institutional fees from AstraZeneca, Merck Sharp & Dohme, Roche, Benecke, Pfizer, Takeda, Bristol-Myers Squibb, AbbVie, Diaceutics, Eli Lilly, Bayer, and Boehringer Ingelheim, outside of the submitted work. The remaining authors declare no conflict of interest.

Address for correspondence: Anne-Marie C. Dingemans, MD, PhD, Department of Respiratory Medicine, Erasmus MC Cancer Institute, Dr. Molewaterplein 40, Rotterdam, 3015 GD, The Netherlands. E-mail: a.dingemans@erasmusmc.nl

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2022.11.030

Conclusions: A substantial subset of metastatic NSCLC is differently classified into smoking- and nonsmoking-associated tumors on the basis of smoking-related mutational signatures than on the basis of smoking history. This signature-based classification more accurately classifies patients on the basis of genome-wide context and should therefore be considered as a stratification factor in clinical research.

Keywords: Non–small cell lung cancer; Tobacco smoking; Whole genome sequencing; Mutational signature; Biomarker

Introduction

Lung cancer is the leading cause of global cancer-related mortality.[1] Approximately 85% of lung cancer is NSCLC, which is a notoriously heterogeneous disease.[2] It has become clear that within this heterogeneity, specific subgroups of NSCLC may be defined, which potentially derive greater benefit from certain treatments. Some of these subgroups, for example, squamous cell carcinomas or tumors with *KRAS* transversion mutations such as *KRAS* G12C or G12V mutations, are more prevalent in patients who smoke or have previously smoked, whereas tumors harboring an *EGFR* mutation or *ALK* translocation are more prevalent in patients who have never smoked.[3–5] Therefore, NSCLC is often divided into smoking-associated and nonsmoking-associated tumors on the basis of patient-reported smoking history. Nevertheless, this division falls short because tumors with nonsmoking-associated carcinogenesis may also occur in patients who smoke. In addition, clinical smoking history can be subject to recall bias and does not account for possible passive smoke exposure.

Fortunately, more precise tools than clinical smoking history are available to select individual patients for specific treatments, such as targeted next-generation sequencing and programmed death-ligand 1 (PD-L1) tumor proportion score. Clinical smoking history might still help guide molecular testing as some targets that are much more common in patients who have never smoked, such as gene fusions, might require additional testing to confirm. Nevertheless, the current guidelines recommend testing all patients with adenocarcinoma for molecular drivers, regardless of clinical smoking history.[6] Therefore, in the era of personalized treatment and precision medicine, clinical smoking history has limited diagnostic or therapeutic consequences in daily clinical practice. In contrast, in clinical research, the classification of patients in “smokers” and “never smokers” on the basis of clinical smoking history is still frequently used as a stratification factor and as a basis for subgroup analyses. This highlights a gap between clinical practice and clinical research that could possibly come at the expense of the external validity of clinical trials. There is a need to bridge this gap by implementing a more precise classification method than patient-reported clinical smoking history.

Several techniques enabling this classification method are already in practice. Next-generation sequencing, including targeted panels, whole exome sequencing, and whole genome sequencing (WGS), allow for an in-depth analysis of the lung cancer genome. Several genome-based studies highlighted major differences in oncogenic events between lung cancer in patients who smoke and patients who have never smoked, including different types of single-base substitutions (SBS), doublet base substitutions (DBS), and small insertions and deletions (indels), which can group together to derive distinct biologically relevant mutational signatures.[7–9] For instance, the SBS signature 4 (SBS4) is characterized by transcriptional strand bias for C>A mutations. This signature was found to be strongly associated with tobacco smoking and to correlate with the extent of tobacco smoke exposure. Similar to SBS4, the indel-based signature 3 (ID3) is associated with tobacco smoking.[10] Therefore, these signatures seem to provide an accurate way of classifying smoking- and nonsmoking-associated tumors. Nevertheless, the tobacco smoking mutational signatures have not yet found their way to randomized controlled trials.

In this study, we aim to provide a genomic classification of smoking- and nonsmoking- associated NSCLC on the basis of the observed frequencies of the smoking-related signatures SBS4 and ID3. This could allow for a more accurate subgrouping of NSCLC for future clinical research. To this end, we leveraged high-quality WGS data obtained from three uniform prospective cohorts of metastatic NSCLC.

Material and Methods

Patient Cohort and Study Procedures

We selected patients with metastatic NSCLC who were included under the protocol of the Center for Personalized Cancer Treatment consortium (CPCT-02 Biopsy Protocol, ClinicalTrial.gov number NCT01855477), the Whole Genome Sequencing Implementation in standard Diagnostics for Every Cancer Patient study (Samsom et al.[11]), and the Drug Rediscovery Protocol study (ClinicalTrial.gov number NCT02925234). All three trials were approved by the local institutional review board and were conducted in accordance with good clinical practice guidelines and the Declaration of Helsinki's ethical principles for medical research. All patients provided written informed consent before any study procedure. Core needle biopsies were taken following local institutional guidelines, aiming to take two to four biopsies with 18 G needles. A minimum tumor percentage of 20% and a minimum DNA yield of 50 ng were needed. Matched whole blood samples were collected to discriminate somatic mutations from germline DNA background variations. The handling, processing, and sequencing of the samples have previously been described in detail for these cohorts.[11–13] The WGS data were made available by the Hartwig Medical Foundation.

Here, we present the in-depth analysis of patients with metastatic NSCLC who were included between July 2012 and October 2020 in five different hospitals in the Netherlands, and of whom clinical records were available. We collected demographic and clinical information including age, sex, disease stage at diagnosis, date of diagnosis of metastatic disease, smoking history, treatment(s) before and after study biopsy, and pathologic information from the local pathology reports including histopathological diagnosis and PD-L1 expression. Smoking history was abstracted from the electronic patient charts. Patients who had previously smoked or were currently smoking were defined as having an active smoking history. Patients with less than 1 pack-year were considered to have never smoked.

Supervised Clustering of Samples Based on Smoking-Related Mutational Signatures

The processing and analysis of the WGS data are described in detail in Supplementary Data 1.[7,12,14–22] Mutational signature contribution was determined by the number of somatic mutations falling into the 96 SBS, 78 DBS, and 83 ID contexts (as described in the COSMIC catalog; https://cancer.sanger.ac.uk/signatures/). These contexts are defined by the substitution class and the sequence context immediately 3′ and 5′ to the mutated base. Each mutational signature is therefore characterized by the predominant substitutional class(es) and the predominant sequences in those classes.[23] The relative mutational signature contribution was determined relative to the total tumor mutational burden (TMB).

To classify the samples as smoking or nonsmoking associated, we calculated the proportion of the relative contribution of single-base mutational signature SBS4 (tobacco smoking) compared with the summed relative contributions of SBS1 (age), SBS4, and SBS5 (age). In addition, we calculated the proportion of the indel mutational signature ID3 (tobacco smoking) compared with the summed relative contributions of ID1 (mismatch repair deficiency [MMRd]/age), ID2 (MMRd/age), and ID3. As inspired by Lee et al.,[7] we used these relative proportions of SBS4 and ID3 together with the absolute counts of SBS4 and ID3 to form two distinct clusters (k-means; $k = 2$), which we termed smoking high and smoking low, respective to the presence of these proportions and absolute counts. Before clustering, these values were centered and scaled appropriately.

Statistics

Statistical analysis of the clinical characteristics was performed using IBM SPSS Statistics software, version 25. Continuous data were compared with Student's *t* test or Mann-Whitney *U* test as appropriate. Means are presented with SDs, and medians with interquartile ranges (IQRs). Categorical data were compared with a chi-square test. Correlations were analyzed with the Spearman's rho. Genomic differences were tested in the statistical platform R (version 4.1.1) using the Wilcoxon ranked sum test with multiple testing correction (Benjamini-Hochberg). Mutational enrichment (or depletion) of genes was tested using a Fisher's exact test with multiple testing correction (Benjamini-Hochberg). For visualization, *p* values (or *q* values) are visualized as * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

Results

Patient Cohort

The WGS data of 316 biopsies of metastatic NSCLC were analyzed (Supplementary Fig. 1). The cohort consisted mostly of females (57%), mean age at diagnosis was 62 (±10) years, and adenocarcinoma was the most prevalent histologic subtype (75%). With respect to recorded clinical smoking history, 11% were currently smoking, 58% had previously smoked, 28% had never smoked, and 3% had an unknown smoking history. Patients with an active smoking history had a median of 25 pack years (IQR: 13–39).

Supervised Clustering Based on Tobacco Smoking-Related Mutational Signatures

Median relative SBS4 contribution was 15% (IQR: 3.8–23.8) and median relative ID3 contribution was 38% (IQR: 13.1–62.3) in the entire cohort. Of the patients with an active smoking history, 10% of samples had low (<3.8%) SBS4 contribution, of which six (3%) samples had no SBS4 contribution at all and 7% had low (<13.1%) ID3 contribution. Of the patients who had never smoked, one sample had high (>23.8%) SBS4 contribution and one sample had high (>62.3%) ID3 contribution.

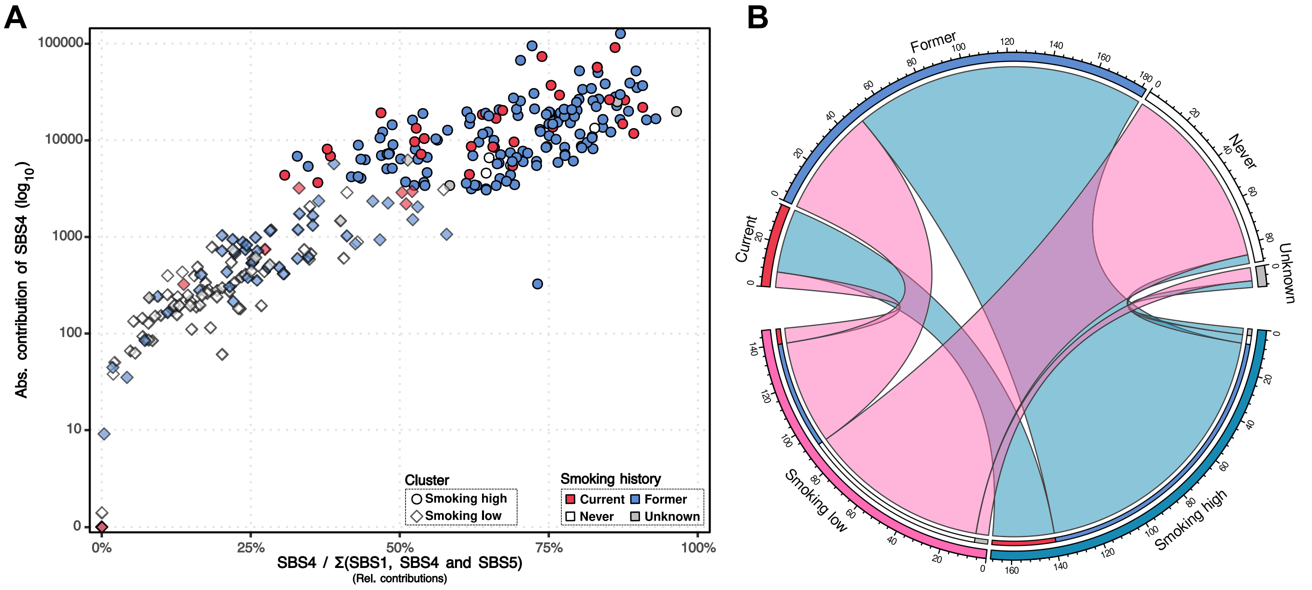

Figure 1. Signature-based clustering of NSCLC into smoking high and low groups. (*A*) Overview of two clustering features which are used, with relationship between the absolute contribution of SBS4 ($\log_{10}$; y axis) and the proportion of relative SBS4 contribution out of the sum of the relative contributions of SBS1, SBS4, and SBS5 for each cluster and smoking history category. (*B*) Chord diagram depicting the distribution of patients based on smoking history versus the genomic clusters. *N* is indicated on the outer dial. SBS, single-base signature.

On the basis of the supervised signature-based clustering, we categorized 169 samples as smoking high and 147 samples as smoking low (Fig. 1*A* and Supplementary Fig. 2). The signature-based clusters differed significantly with regard to sex, smoking history, prior treatment, and histopathological diagnosis (Table 1). The distribution of recorded smoking history within these signature-based clusters is depicted in Figure 1*B*. Furthermore, 4% of the patients who had never smoked were included in the smoking high cluster and 26% of the patients with an active smoking history were included in the smoking low cluster. Patients with an active smoking history in the smoking high cluster had significantly more pack years than those with an active smoking history in the smoking low cluster (28 versus 14, $p < 0.001$). In addition, 30% of squamous cell carcinomas ($n = 10$) were included in the smoking low cluster, of which five patients had an active smoking history, four patients had never smoked, and one patient had an unknown smoking history. PD-L1 expression did not differ between clusters ($p = 0.70$) or between those with an active smoking history and those who had never smoked ($p = 0.16$).

Overview and Differences of the Genomic Landscape Between Signature-Based Clusters

An overview of genomic characteristics has been summarized in Figure 2*A*-*O*, grouped by signature-based clustering and ordered by descending TMB. Recurrent copy number alterations and chromosomal arm aberrations (from GISTIC2 analysis) for the entire cohort, and per the signature-based cluster, can be found within Supplementary Figure 3. Comparing the samples with and without putative chromothripsis events, chromothripsis did not significantly affect driver genes, mutational signatures, or TMB within the entire cohort and within the separate clusters (adjusted $p > 0.1$).

When comparing the signature-based clusters, we observed significant differences in genomic characteristics regarding TMB and mutational signatures (Fig. 3*A* and *C*). No major differences between genome-wide ploidy and chromosomal arms copy number alterations could be observed (Fig. 3*B* and Supplementary Fig. 3*B* and *C*). Compared with the smoking low cluster, median TMB was almost a fivefold higher within the smoking high cluster (4 versus 19 mutations per megabase, $p < 0.001$) (Fig. 3*A* and Supplementary Table 1). In addition, TMB was strongly correlated to absolute SBS4 contribution (Spearman's rho: 0.87, $p < 0.001$). Nevertheless, 26 (15%) of the samples in the smoking high cluster had a low TMB (<5). In addition, 12 samples (8%) within the smoking low cluster had a high TMB (≥10), yet they generally harbored mutational signatures related to MMRd and APOBEC activity rather than those related to smoking; three of these samples indeed were classified as microsatellite instable (MSI) tumors.

Beyond the expected differences in SBS4 and ID3 abundance because of their usage as clustering features (Supplementary Table 1), we also observed differences

Table 1. Demographic and Clinical Characteristics of the Patients Stratified by Signature-Based Clustering

Characteristic	NSCLC Smoking High (n = 169)	NSCLC Smoking Low (n = 147)	*p* Value
Sex, n (%)			0.001
Female	82 (49)	99 (67)	
Male	87 (51)	48 (33)	
Age at diagnosis, n (%)			0.685
Mean (SD)	62 (10)	62 (11)	
Tumor stage at diagnosis, n (%)			0.147
I	6 (4)	7 (5)	
II	8 (5)	4 (3)	
III	40 (24)	20 (14)	
IV	112 (66)	114 (78)	
Unknown	3 (2)	2 (1)	
Smoking history, n (%)			<0.001
Never	4 (2)	85 (58)	
Former	133 (79)	49 (33)	
Current	29 (17)	7 (5)	
Unknown	3 (2)	6 (4)	
Pack years			<0.001
Median (IQR)	28 (17-40)	14 (6-25)	
Unknown	30	10	
Histopathological diagnosis, n (%)			0.002
Adenocarcinoma	114 (67)	124 (84)	
Squamous cell carcinoma	23 (14)	10 (7)	
Other	32 (19)	13 (9)	
PD-L1 score, n (%)			0.70
<1%	41 (24)	39 (27)	
1%-50%	36 (21)	41 (28)	
≥50%	26 (15)	22 (15)	
Unknown	66 (39)	45 (31)	
Prebiopsy regimen, n (%)			<0.001
Chemotherapy/immunotherapy/other	103 (61)	36 (24)	
TKI	20 (12)	91 (62)	
Untreated	46 (27)	20 (14)	
Lines of systemic treatment before biopsy			<0.001
Median (IQR)	1 (0-2)	2 (1-3)	

Note: Percentages may not add up to 100% owing to rounding.
IQR, interquartile range; PD-L1, programmed death-ligand 1; TKI, tyrosine kinase inhibitor.

in additional mutational signatures between the two clusters (Fig. 3*C*). Within the smoking high cluster, we observed significantly higher contributions of SBS signatures SBS18, SBS29, and SBS5 compared with the smoking low cluster. SBS4, SBS18, and SBS29 are all characterized by transcriptional strand bias for C>A substitutions, which would explain why these signatures cluster together. SBS18 is suggested to be related to damage by reactive oxygen species, and SBS29 is associated with tobacco chewing. SBS5 is still of unknown cause, although its mutational spectrum seems to be increased in cancers related to tobacco exposure and seems to be related to age.[10] With regard to the DBS signatures, DBS2, which is also associated with exposure to tobacco smoke, had the highest contribution in the smoking high cluster.[10]

In the smoking low cluster, SBS40, SBS1, SBS13, and SBS2 had the highest SBS contributions. SBS1 has a clock-like mechanism related to age, and it is characterized by C>T mutations caused by spontaneous deamination of 5-methylcytosine. SBS2, mainly consisting of C>T mutations, and SBS13, mainly consisting of C>G mutations, are both linked to activity of the AID/APOBEC enzymes.[10] It has been suggested that the activation of APOBEC in cancer could be caused by previous viral infection or tissue inflammation, which suggests a role of inflammation in the development of nonsmoking-associated lung cancer.[24] The relative contribution of the two APOBEC signatures was significantly higher in the smoking low cluster than in the smoking high cluster (16% versus 6%, $p < 0.001$). SBS40 is also correlated with age and has a large similarity to SBS5.[10] DBS6 had the highest DBS signature contribution and ID1 had the highest ID signature contribution in the smoking low cluster. DBS6 is still of unknown cause, but it seems to be related to age.[10] ID1

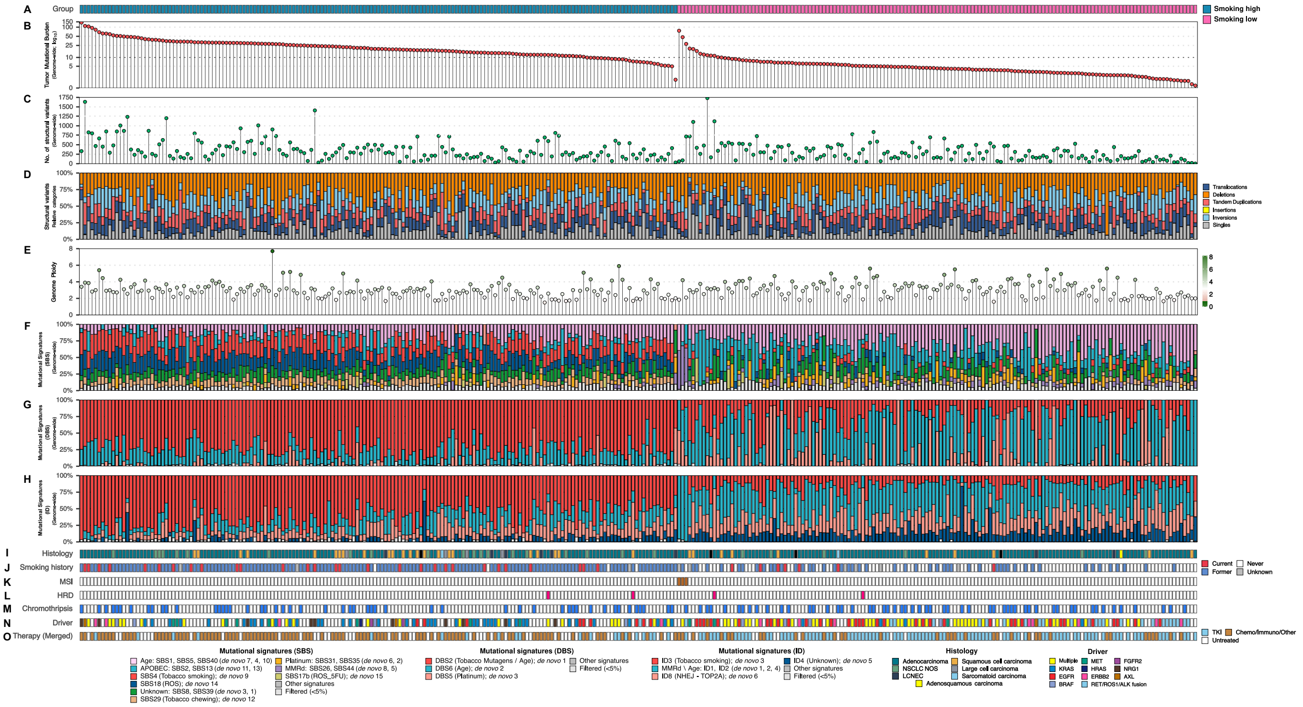

Figure 2. Overview of the genomic landscape. For each sample (column), we display common genomic characteristics. (*A*) Signature-based cluster designation. (*B*) Genome-wide TMB. (*C*) Total number of structural variants. (*D*) Relative frequency per structural variant category; translocations, deletions, tandem duplications, insertions, inversions, and single breakpoints. (*E*) Mean genome-wide ploidy. (*F*) Relative contribution of de novo SBS mutational signatures. (*G*) Same as in panel *F*, but for DBS signatures. (*H*) Same as in panel *F*, but for ID signatures. (*I*) Histopathological diagnosis. (*J*) Reported clinical smoking history. (*K*) MSI status. (*L*) HRD status as determined by CHORD. (*M*) Presence of chromothripsis. (*N*) Presence of known driver alteration(s). (*O*) Previous systemic therapy before biopsy. DBS, doublet base substitution; HRD, homologous recombination deficient; ID, indel-based; MSI, microsatellite instability; SBS, single-base signature; TMB, tumor mutational burden.

is associated with age and slippage of the template DNA strand during DNA replication. It is often found in cancers with DNA MMRd and MSI.[10]

The differences in TMB and relative mutational signature contribution between the two clusters (Fig. 3*A* and *C*) were also investigated in the different histologic subtypes (Supplementary Table 2). Most observed differences were consistent over the histologic subtypes; however, both APOBEC signatures (SBS2 and SBS13) did not differ significantly between the smoking high and smoking low clusters in the squamous cell subgroup (3% versus 5%, $p = 0.062$, and 5% versus 13%, $p = 0.237$, respectively), but it did in the adenocarcinoma subgroup ($p < 0.001$). The differences in smoking-related signatures were consistent over all histologic subgroups.

Altered Landscape of Putative Driver Genes

Next, we investigated differences and enrichment within the somatic inventory of perturbed genes between the two signature-based clusters (Fig. 3*D* and Supplementary Figs. 4 and 5). Of the known driver oncogenes, *EGFR* mutations and *ALK* fusions were significantly more prevalent in the smoking low cluster than in the smoking high cluster (50% versus 9%, $p < 0.001$; 13% versus 0%, $p < 0.001$; respectively). *KRAS* mutations were significantly more prevalent in the smoking high cluster (28% versus 5%, $p < 0.001$). Table 2 illustrates the differences in frequency of oncogene driver alterations in the nonsquamous cell carcinoma samples between the signature-based clusters and the groups on the basis of clinical smoking history.

Discordances

We further analyzed the samples of patients who had never smoked but were included in the smoking high cluster (n = 4) and found that all harbored high (>10) or medium-high (>5) TMB. In addition, these samples reflected the mutational signatures found within the rest of the smoking high cluster and revealed an absence of, or minor, APOBEC signature contribution (Table 3). With regard to oncogene driver alterations, we identified a *KRAS* G12C mutation, an *EGFR* L858R mutation with concomitant *EGFR* amplification, and a *BRAF* G469A-activating mutation. The fourth sample did not harbor a known driver oncogene.

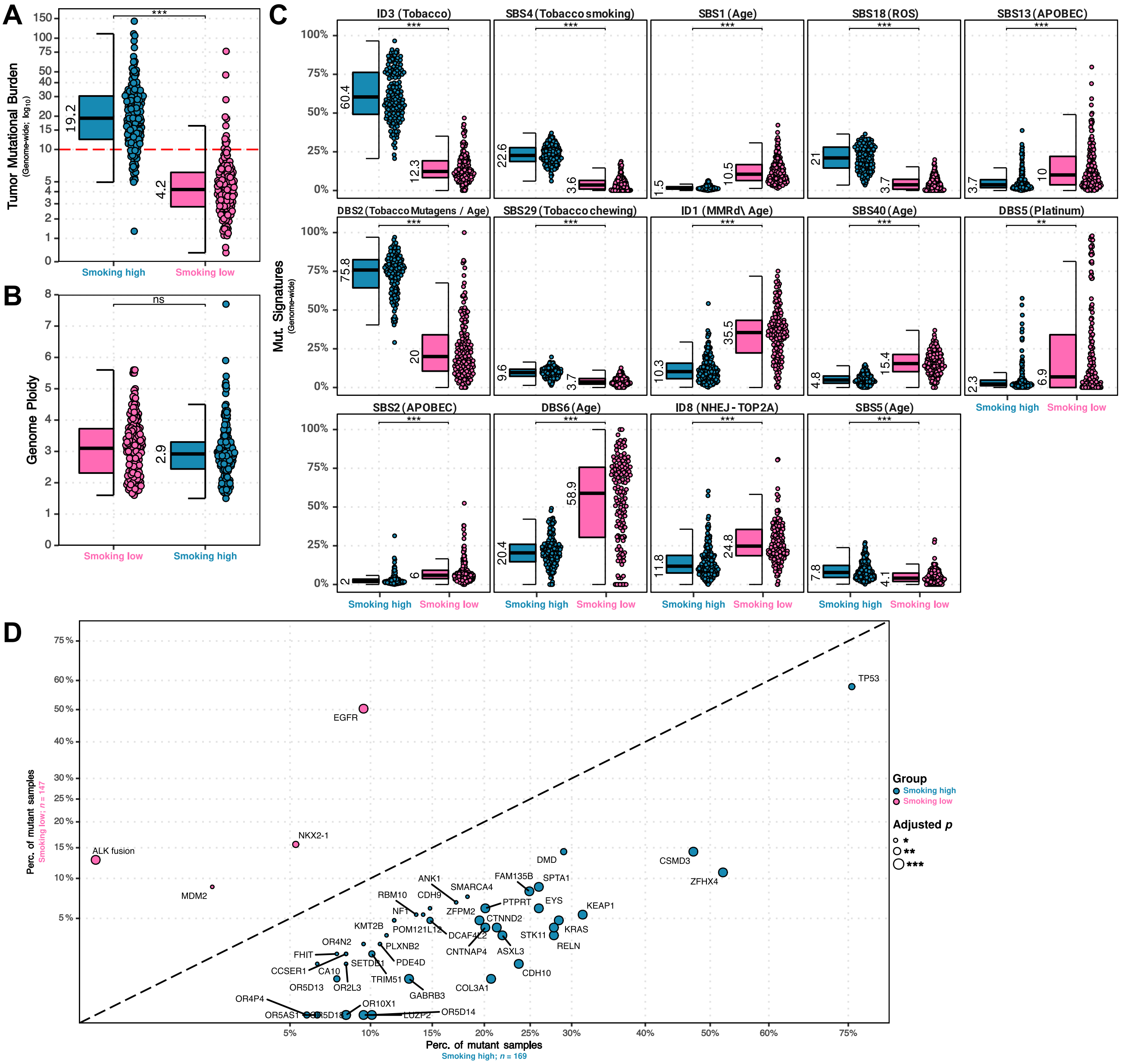

Figure 3. Genomic differences between signature-based smoking high and smoking low samples. (*A*) Box plot of TMB between the signature-based classification of samples; the median, Q1, and Q3 are represented. Statistical significance is found on the top. Cutoff of high TMB (≥10 mutations/Mb) is indicated with a red horizontal line. (*B*) Same as in panel *A* but depicting genome-wide ploidy. (*C*) Same as in panel *A* but depicting mutational signatures (SBS, DBS, and ID) with statistically significant differences ($q < 0.05$) and harboring a minimal median relative contribution of greater than or equal to 5% in either group. Proposed causes are found on the top. (*D*) Statistically significant enriched genes between the signature-based classification of samples. Size of points represents the adjusted *p* (*q* value), with genes having greater enrichment (or depletion) depicted as larger. DBS, doublet base substitution; ID, indel-based; Mb, megabase; Q1, quartile 1; Q3, quartile 3; SBS, single-base signature; TMB, tumor mutational burden.

The samples of patients with an active smoking history captured within the smoking low cluster (n = 56) contained only minor contributions of the dominant signatures of the smoking high cluster (Table 3). The APOBEC signatures SBS13 and SBS2 had median contributions that were in line with the median contributions of these signatures in the rest of the smoking low cluster. Of these samples, 77% harbored a known oncogene driver alteration: 22 *EGFR* alterations, five *BRAF* mutations, five *ALK* fusions, four *ERBB2/Her2* amplifications, two *KRAS* mutations (one G12V and one G12D), one *MET* amplification, one *MET* exon 14 skipping mutation, two *RET* fusions, and one *ROS1* fusion.

Table 2. Frequency of Oncogene Driver Alterations in Nonsquamous Cell NSCLC

Oncogene Driver Alterations	Total N	Smoking Low Cluster, n (%)	Never Smoked, n (%)
*EGFR*alteration	85	72 (85)	49 (58)
Exon 19 deletion	38	35 (92)	24 (63)
L858R	28	22 (79)	17 (61)
Exon 20 insertion	4	2 (50)	1 (25)
Other	15	12 (80)	7 (47)
ALK fusion	19	19 (100)	14 (74)
RET fusion	3	3 (100)	1 (33)
ROS1 fusion	6	5 (83)	4 (67)
MET amplification	21	10 (48)	9 (43)
*BRAF*mutation	21	11 (52)	7 (33)
V600E	12	10 (83)	6 (50)
Other	8	1 (13)	1 (13)
ERBB2/Her2 amplification	13	7 (54)	3 (23)
	Total N	Smoking High Cluster,n (%)	Active Smoking History,n (%)
*KRAS*mutation	50	45 (90)	43 (86)
G12C	19	19 (100)	18 (95)
G12A	6	5 (83)	5 (83)
G12V	13	11 (85)	10 (77)
G12D	5	4 (80)	4 (80%)
Other	7	6 (86)	6 (86)
Codon 12	3	3 (100)	3 (100)
Codon 61	3	2 (67)	2 (67)
K117N	1	1 (100)	1 (100)
*MET*exon 14 skipping alteration	6	3 (50)	4 (67)

In Silico Analysis Targeted Panel

Last, to investigate the translatability of WGS mutational signature analysis to targeted panels used in daily clinical practice, we performed an in silico analysis for the TruSight Oncology 500 (TSO500; Illumina, San Diego, CA) panel to retain only those somatic variants which were captured within the target regions of the TSO500. The TSO500 is a widely used large pan-cancer panel that captures 523 (onco)genes within approximately 1.94 megabase spread throughout the genome and has also been used for signature analysis previously.[25] Of the samples, 90% (n = 285) harbored enough SBS mutations to derive the SBS signatures for subsequent analysis. Of those samples, 54% (n = 155) also harbored enough indel mutations to derive the ID signatures. Next, we investigated the concordance between the TSO500 and WGS clustering, and if this concordance was affected by extension of the TSO500 regions. Using somatic variants within only TSO500 target regions for signature-based clustering yielded similar results as performing this on the whole genome with an F1 score (SBS only on 285 samples) of 0.813 and an F1 score (SBS + ID on 155 samples) of 0.753 (Supplementary Fig. 6*A*). The distribution of smoking history within the clusters again revealed a considerable degree of discordance between smoking history and mutational signatures (Fig. 1*B* and Supplementary Fig. 6*C* and *D* for the WGS clusters). A few more patients with an active smoking history were classified as smoking low with the TSO500 than with WGS. This can be explained by the lower discriminatory power of a targeted panel compared with WGS, as several (noncoding) genomic regions which are affected by (smoking-related) mutational processes are not included in the target regions of the TSO500. This is further illustrated by the fact that the number of informative samples improves when extending the TSO500 regions to allow for the capture of additional mutations (Supplementary Fig. 6*A* [upper track] and *B*). Extensions of the TSO500 regions also generally improved the concordance (F1) of both approaches between WGS and genomic subsets (Supplementary Fig. 6*A*).

Table 3. TMB and Relative Dominant Mutational Signature Contribution in Discordant Samples

Characteristic	Smoking High Without Smoking History (n = 4)	Smoking Low With Active Smoking History (n = 56)
TMB (IQR)	11.7 (7.8-17.9)	5.1 (3.6-6.7)
SBS4, % (IQR)	23.2 (19.0-25.6)	5.1 (2.2-9.4)
SBS18, % (IQR)	23.9 (21.1-27.0)	3.7 (0.4-10.5)
SBS29, % (IQR)	12.1 (11.2-15.3)	4.7 (0.6-7.4)
SBS2, % (IQR)	1.7 (0.5-3.0)	5.8 (3.5-8.2)
SBS13, % (IQR)	2.1 (1.1-2.9)	8.5 (3.3-17.1)
ID3, % (IQR)	50.4 (42.6-63.9)	17.9 (12.4-23.5)

ID3, indel-based signature 3; IQR, interquartile range; SBS, single-base signature; TMB, tumor mutational burden.

Discussion

In this study, we aimed to investigate a more accurate classification than clinical smoking history in NSCLC. We revealed that clustering metastatic NSCLC tumors into smoking-associated and nonsmoking-associated mutagenesis on the basis of the SBS4 and ID3 mutational signatures derived from WGS data is a feasible classification method.

Our classification reveals that there is a large overlap in clinical smoking history and classification on the basis of SBS4 and ID3 contributions. This also revealed a degree of discordance between these two grouping methods. A few patients who had never smoked had tumors in which smoking-associated mutational signatures were considerably present within their somatic genome despite a negative smoking history. Possible explanations for this are recall bias of previous tobacco smoke exposure, inaccurate history taking by health care professionals, or passive smoke exposure. In the tumors of patients with an active smoking history, the amount of SBS4 contribution varied greatly. This could possibly be explained by the extent of tobacco smoke exposure, because patients with an active smoking history in the smoking low cluster had fewer pack years than those in the smoking high cluster. Furthermore, 26% of the patients with an active smoking history had relatively little SBS4 and ID3 contribution and were therefore included in our smoking low cluster. The discordance between smoking history and SBS4 contribution has previously been reported. Lee et al.[7] revealed that in their cohort of lung adenocarcinoma approximately one-third of tumors from patients with an active smoking history had no or minor SBS4 contribution. Devarakonda et al.[26] used a statistical model, including TMB and SBS4, to infer smoking status and excluded four of 88 tumor samples of patients who reportedly had never smoked. This confirms that the mutational processes that have occurred in the tumor are not fully reflected by patient-reported smoking history.

Our signature-based clustering resulted in two distinct clusters with different TMB, mutational signature contributions, and distinct mutational landscapes. We found that tumors with a high TMB, with high SBS4, SBS18, and SBS29 contributions, and with *KRAS* mutations were predominantly classified as smoking high. These signatures and most *KRAS* mutations in NSCLC are characterized by transversion mutations, which would explain why they group together. Tumors with a low TMB, high APOBEC signature contribution, and *EGFR* mutations or *ALK* fusions were predominantly classified as smoking low. Genome-based studies have found similar findings when using clinical smoking history to classify tumors.[8,9,26] The tumors from patients who had never smoked but were clustered as smoking high had a similar genotype to the other tumors in the smoking high cluster, which suggests smoke exposure despite a negative smoking history. Tumors from patients with an active smoking history in the smoking low cluster had similar genotypes to the rest of this cluster, including a high percentage of oncogenic driver alterations such as *EGFR* mutations and *ALK* fusions. This suggests that our classification based on SBS4 and ID3 is more accurate in grouping NSCLCs on the basis of similar genomic context rather than reported smoking history. Because TMB was strongly correlated to SBS4 contribution, this might suggest that classification on the basis of TMB would yield similar results. Nevertheless, other causes of high TMB, such as MSI, might lead to more misclassifications. Interestingly, PD-L1 expression did not differ between the two clusters. Several studies have suggested the up-regulation of PD-L1 expression in patients with an active smoking history,[27,28] which leads to the expectation of higher PD-L1 in the smoking high cluster. The fact that we found no difference between the smoking high and smoking low cluster supports studies that have found no association between PD-L1 expression and smoking status.[29,30]

A signature-based classification based on genomic SBS4 and ID3 contribution instead of classifying on the basis of clinical history could have several clinical implications. First, as we have found that the frequency of targetable driver oncogenes is higher in those with a low smoking signature contribution than in those who have never smoked, a low smoking signature contribution suggests an increased likelihood of oncogene-driven NSCLC regardless of smoking status. Therefore, if a driver alteration has not been detected during routine diagnostics, a low smoking signature contribution could warrant further investigation to identify more rare oncogenic driver alterations. Further investigation should include comprehensive RNA analysis for the detection of gene fusions, including those with unknown fusion partners and kinase domain duplications (KDDs). Although rare, these oncogenic drivers can provide an important target for treatment, for example, several reports have revealed sensitivity of tumors with an EGFR-KDD to *EGFR* tyrosine kinase inhibitors.[31,32] Second, replacing the terms “smoker” and “never smoker” that are coined by clinical smoking history could contribute to reducing the stigma and self-blame around lung cancer. It has been suggested that the stigma of lung cancer is still a significant barrier in reducing the lung cancer burden in global society.[33] Therefore, the potential impact of the label “smoker” on patients’ well-being should not be underestimated. Next, in randomized trials investigating immunotherapy, smoking history can be of special interest as a predictive biomarker owing to the current assumption that smoking leads to an accumulation of mutations that in turn could generate a higher number of neoantigens. These neoantigens could potentially predict response to immunotherapy. Nevertheless, as a considerable percentage of patients with an active smoking history did not actually harbor high (or any) smoking signature contribution, the reliability of smoking history as stratification factor and predictive biomarker in these trials can be questioned. Because SBS4 has been found to have a potential predictive value for response to immunotherapy, it could therefore potentially provide a more accurate stratification factor than smoking history.[34,35] Furthermore, it is possible that the subgroup of patients with low SBS4 contribution would derive less benefit from immunotherapy than would be expected on the basis of smoking history or PD-L1 expression, because PD-L1 expression did not differ between the smoking high and smoking low clusters. In addition, a small group of *EGFR*-mutated samples were classified as smoking high, which might suggest that these patients are part of the limited subpopulation of *EGFR*-

mutated NSCLC who do derive benefit from immunotherapy.[36] Nevertheless, the predictive value of SBS4, and ID3, in oncogene-driven NSCLC is yet to be determined.

Implementing such a classification method in genome-based research should constitute little additional effort as these data are already available. In addition, the WIDE study investigators have recently found that WGS for patients with metastatic cancer is feasible in routine clinical practice.[13] We do, however, appreciate that the implementation in clinical trials would still be a hurdle to overcome. Currently, many clinical trials already require archival or fresh tumor tissue to be sent in for genome testing during the screening period. Our TSO500 in silico analysis revealed that the TSO500 panel allows for SBS mutational signature calling in most cases, whereas the ID signatures are more challenging to retrieve. Nevertheless, because the ID3 signature is associated with the SBS4 signature, the lack of the ID signatures should not vastly differ conclusions. This provides an opportunity for clinical trials to incorporate mutational signature analysis in the genome testing procedure during screening without the need to perform WGS. Similarly, in daily practices where WGS is currently not a common practice, mutational signature analysis with targeted panels could still help identify those with a higher likelihood of harboring a (rare) oncogenic driver. However, the optimal cutoff between a high or low smoking signature contribution does still warrant further investigation.

This study has certain limitations that should be considered. Most importantly, an accepted standard in distinguishing smoking-associated from nonsmoking-associated carcinogenesis is lacking. In the absence such a standard, accuracy analyses are unreliable and were therefore not performed. In addition, mutational signatures infer the dominant processes of mutagenesis within a tumor genome; however, they do not necessarily reflect the driving cause of carcinogenesis. For instance, cells of normal lung epithelium can also have SBS4 contribution without this leading to carcinogenesis.[37] Our samples of *EGFR*-driven NSCLC with high SBS4 contribution further illustrate this. Next, many of the patients in our cohort were included because the treating physician deemed WGS to have added clinical value in the patient’s treatment course, which could have led to a selection bias. Most patients in our cohort had also received previous systemic therapy; however, these therapies do not induce the same mutations as tobacco exposure and thus have no influence on the smoking-related signatures. Last, we did not collect outcome data. Consequently, the prognostic or predictive value of our classification method has not been determined. Despite these limitations, our study also has several strengths. To the best of our knowledge, it is the first to focus on the discordances between clinical smoking history and smoking-associated mutational processes in the NSCLC genome. In addition, previously published genomic cohorts are often small, only focus on patients without smoking history, or primarily include early stage lung cancer.[8,9,38] Our comprehensive cohort allows for an in-depth analysis of metastatic tumors from patients with and without smoking history.

To conclude, our mutational signature-based classification of smoking-associated and nonsmoking-associated NSCLC is more accurate in grouping tumors with similar genomic contexts together compared with classification on the basis of clinical smoking history. Implementing such a signature-based classification aids in defining more accurate subgroups for future genome-based research and should be considered as a stratification factor in clinical trials. In addition, it could aid in optimizing diagnostic strategies in daily practice which are currently still influenced by clinical smoking history, such as the pursuit of identification of more rare oncogenic drivers. Importantly, with a signature-based classification, there is less focus on the act of smoking in lung cancer development, and it can thus be an important achievement in overcoming the self-blame and stigma around lung cancer.

CRediT Authorship Contribution Statement

Sophie M. Ernst: Conceptualization, Methodology, Formal analysis, Investigations, Data curation, Writing—original draft preparation, Review and editing.

Joanne Mankor: Conceptualization, Methodology, Investigations, Data curation, Writing—review and editing.

Job van Riet: Conceptualization, Software, Formal analysis, Writing—review and editing.

Jan H. von der Thüsen: Conceptualization, Writing—review and editing.

Hendrikus J. Dubbink: Conceptualization, Writing—review and editing.

Joachim G.J.V. Aerts: Conceptualization, Writing—review and editing.

Adrianus J. de Langen: Conceptualization, Writing—review and editing.

Egbert F. Smit: Conceptualization, Writing—review and editing.

Anne-Marie C. Dingemans: Conceptualization, Writing—review and editing, supervision.

Kim Monkhorst: Conceptualization, Writing—review and editing, supervision.

Acknowledgments

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment have made available to the study. The authors thank Arne van Hoeck for sharing the lung cancer mutational signatures.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at http://www.jto.org and at 10.1016/j.jtho.2022.11.030.

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. *CA Cancer J Clin*. 2021;71:209-249.
2. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc*. 2008;83:584-594.
3. Khuder SA. Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. *Lung Cancer*. 2001;31:139-148.
4. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers-a different disease. *Nat Rev Cancer*. 2007;7:778-790.
5. Chapman AM, Sun KY, Ruestow P, Cowan DM, Madl AK. Lung cancer mutation profile of EGFR, ALK, and KRAS: meta-analysis and comparison of never and ever smokers. *Lung Cancer*. 2016;102:122-134.
6. Planchard D, Popat S, Kerr K, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2018;29(suppl 4):iv192-iv237.
7. Lee JJ, Park S, Park H, et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell*. 2019;177:1842-1857.e21.
8. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*. 2012;150:1121-1134.
9. Boeckx B, Shahi RB, Smeets D, et al. The genomic landscape of nonsmall cell lung carcinoma in never smokers. *Int J Cancer*. 2020;146:3207-3218.
10. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578:94-101.
11. Samsom KG, Bosch LJW, Schipper LJ, et al. Study protocol: whole genome sequencing implementation in standard Diagnostics for Every cancer patient (WIDE). *BMC Med Genomics*. 2020;13:169.
12. Priestley P, Baber J, Lolkema MP, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*. 2019;575:210-216.
13. Samsom KG, Schipper LJ, Roepman P, et al. Feasibility of whole-genome sequencing-based tumor diagnostics in routine pathology practice. *J Pathol*. 2022;258:179-188.
14. Cameron DL, Baber J, Shale C, et al. GRIDSS, PURPLE, LINX: unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. https://www.biorxiv.org/content/10.1101/781013v1. Accessed February 1, 2022.
15. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17:122.
16. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443.
17. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:D1062-D1067.
18. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. *Nat Commun*. 2020;11:5584.
19. Cortés-Ciriano I, Lee JJ, Xi R, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet*. 2020;52:331-341.
20. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. 2018;10:33.
21. Martincorena I, Raine KM, Gerstung M, et al. Universal patterns of selection in cancer and somatic tissues. *Cell*. 2017;171:1029-1041.e21.
22. 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:R41.
23. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415-421.
24. Koito A, Ikeda T. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Front Microbiol*. 2013;4:28.
25. Kroeze LI, de Voer RM, Kamping EJ, et al. Evaluation of a hybrid capture-based pan-cancer panel for analysis of treatment stratifying oncogenic aberrations and processes. *J Mol Diagn*. 2020;22:757-769.
26. Devarakonda S, Li Y, Martins Rodrigues F, et al. Genomic profiling of lung adenocarcinoma in never-smokers. *J Clin Oncol*. 2021;39:3747-3758.
27. Calles A, Liao X, Sholl LM, et al. Expression of PD-1 and its ligands, PD-L1 and PD-L2, in smokers and never smokers with KRAS-mutant lung cancer. *J Thorac Oncol*. 2015;10:1726-1735.
28. Norum J, Nieder C. Tobacco smoking and cessation and PD-L1 inhibitors in non-small cell lung cancer (NSCLC): a review of the literature. *ESMO Open*. 2018;3:e000406.
29. Brody R, Zhang Y, Ballas M, et al. PD-L1 expression in advanced NSCLC: insights into risk stratification and treatment selection from a systematic literature review. *Lung Cancer*. 2017;112:200-215.
30. Lafuente-Sanchis A, Zúñiga Á, Estors M, et al. Association of PD-1, PD-L1, and CTLA-4 gene expression and clinicopathologic characteristics in patients with

non-small-cell lung cancer. *Clin Lung Cancer*. 2017;18:e109–e116.

31. Wang J, Li X, Xue X, et al. Clinical outcomes of EGFR kinase domain duplication to targeted therapies in NSCLC. *Int J Cancer*. 2019;144:2677–2682.
32. Taek Kim J, Zhang W, Lopategui J, Vail E, Balmanoukian A. Patient with Stage IV NSCLC and CNS metastasis with EGFR Exon 18–25 kinase domain duplication with response to osimertinib as a first-line therapy. *JCO Precis Oncol*. 2021;5:88–92.
33. Hamann HA, Ver Hoeve ES, Carter-Harris L, Studts JL, Ostroff JS. Multilevel opportunities to address lung cancer stigma across the cancer control continuum. *J Thorac Oncol*. 2018;13:1062-1075.
34. Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348:124–128.
35. Anagnostou V, Niknafs N, Marrone K, et al. Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. *Nat Cancer*. 2020;1:99-111.
36. Qiao M, Jiang T, Liu X, et al. Immune checkpoint inhibitors in EGFR-mutated NSCLC: dusk or dawn? *J Thorac Oncol*. 2021;16:1267–1288.
37. Yoshida K, Gowers KHC, Lee-Six H, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*. 2020;578:266–272.
38. Luo W, Tian P, Wang Y, et al. Characteristics of genomic alterations of lung adenocarcinoma in young never-smokers. *Int J Cancer*. 2018;143:1696-1705.