Tobacco smoking related mutational signatures in classifying smoking associated and non-smoking associated non-small cell lung cancer.

Sophie M. Ernst, Joanne Mankor, Job van Riet, Jan H. von der Thüsen, Hendrikus J. Dubbink, Joachim G.J.V. Aerts, Adrianus J. de Langen, Egbert F. Smit, Anne-Marie C. Dingemans, Kim Monkhorst

PII: S1556-0864(22)01961-X
DOI: https://doi.org/10.1016/j.jtho.2022.11.030
Reference: JTHO 2599

To appear in: Journal of Thoracic Oncology

Received Date: 19 July 2022
Revised Date: 9 November 2022
Accepted Date: 29 November 2022


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 International Association for the Study of Lung Cancer. Published by Elsevier Inc.
Title page

Title: Tobacco smoking related mutational signatures in classifying smoking associated and non-smoking associated non-small cell lung cancer.

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

* These authors contributed equally;
# Corresponding author.

Affiliations:
a Department of Respiratory Medicine, Erasmus MC Cancer Institute, Doctor Molewaterplein 40, 3015 GD Rotterdam, the Netherlands.
b Department of Medical Oncology, Erasmus MC Cancer Institute, Doctor Molewaterplein 40, 3015 GD Rotterdam, The Netherlands.
c Department of Pathology, Erasmus University Medical Center, Doctor Molewaterplein 40, 3015 GD Rotterdam, the Netherlands.
d Department of Thoracic Oncology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands.
e Department of Pathology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands.

E-mail addresses authors: s.ernst@erasusmc.nl; j.mankor@erasusmc.nl; j.vanriet@erasusmc.nl; j.vonderthusen@erasusmc.nl; h.dubbink@erasusmc.nl; j.aerts@erasusmc.nl; j.d.langen@nki.nl; e.smit@nki.nl; a.dingemans@erasusmc.nl; k.monkhorst@nki.nl.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: JHT reports research grants and/or institutional fees from AstraZeneca, BMS, Roche Diagnostics; personal fees from Eli Lilly, MSD, Pfizer, Roche outside the submitted work. HJD reports research funding and support from AstraZeneca, MSD, Illumina; personal fees from Abbvie, AstraZeneca, Bayer, Janssen, Eli Lilly, MSD, Pfizer, Bayer outside the submitted work. JGIVA reports personal fees from Eli Lilly, Amphera, MSD, Takeda, BIOCAD, Pfizer, BMS; stock/ownership interest in Amphera, Pamgene outside the submitted work. AJL reports institutional fees from AstraZeneca, BMS, Boehringer Ingelheim, Eli Lilly, MSD, Pfizer, Merus outside the submitted work. ACD reports institutional fees from Roche, Takeda, Sanofi, Amsgen, Bayer, Eli Lilly, Janssen, Pfizer, AstraZeneca outside the submitted work. KM reports institutional fees from AstraZeneca, MSD, Roche, Benecke, Pfizer, BMS, AbbVie, Diaceutics, Eli Lilly, Bayer, Boehringer Ingelheim outside the submitted work. All other authors have declared no conflicts of interest.

Corresponding author
Prof Anne-Marie C. Dingemans
Department of Respiratory Medicine, Erasmus MC Cancer Institute
Dr. Molewaterplein 40
Rotterdam, 3015 GD, the Netherlands
(+31) 10-07043739
a.dingemans@erasusmc.nl
Abstract

Introduction: Patient-reported smoking history is frequently used as stratification factor in NSCLC-directed clinical research. However, this classification does not fully reflect the mutational processes in a tumour. Next-generation sequencing can identify mutational signatures associated with tobacco smoking, such as single-base signature 4 (SBS4) and indel-based signature 3 (ID3). This provides an opportunity to redefine the classification of smoking and non-smoking associated NSCLC based on individual genomic tumour 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 SBS4 and ID3 were combined with relative contributions of age-related signatures to divide the cohort into smoking (‘smoking high’) and non-smoking associated (‘smoking low’) clusters.

Results: The smoking high (N=169) and smoking low (N=147) clusters differed significantly in tumour 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% harboured 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.

Conclusions: A substantial subset of metastatic NSCLC is differently classified into smoking and non-smoking associated tumours based on smoking-related mutational signatures than based on smoking history. This signature-based classification more accurately classifies patients based on genome-wide context and should therefore be considered as 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 non-small cell lung cancer (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, tumours with squamous cell carcinoma histology or KRAS transversion mutations such as KRAS G12C or G12V mutations, are more prevalent in patients who smoke or have previously smoked; while tumours harbouring 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 non-smoking associated tumours based on patient-reported smoking history. However, this division falls short since tumours with non-smoking associated carcinogenesis may also occur in patients who smoke. Additionally, 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 (NGS) and programmed death-ligand 1 (PD-L1) tumour 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. However, the current guidelines recommend testing all patients with a denocarcinoma 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’ based on clinical smoking history is still frequently used as a stratification factor and as 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. NGS, including targeted panels, whole-exome sequencing (WES) 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 mutational signature 4 (SBS4) is
characterised 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 mutational signature 3 (ID3) is also associated with tobacco smoking.\textsuperscript{10} Therefore, these signatures seem to provide an accurate way of classifying smoking and non-smoking associated tumours. However, 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 non-smoking associated NSCLC based on 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 no. NCT01855477), the Whole genome sequencing Implementation in standard Diagnostics for Every cancer patient study (Samsom et al.\textsuperscript{11}) and the Drug Rediscovery Protocol study (ClinicalTrial.gov no. 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 2-4 biopsies with 18 G needles. A minimum tumour percentage of 20% and a minimum DNA yield of 50ng was needed. Matched whole blood samples were collected to discriminate somatic mutations from germline DNA background variations. The handling, processing and sequencing of the samples has previously been described in detail for these cohorts.\textsuperscript{11-13} The WGS data was 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 5 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 pathological information from the local pathology reports including histology 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 is described in detail in Supplementary Data \textsuperscript{1,7,12,14-22} Mutational signature contribution was determined by the number of somatic mutations falling into the 96 single nucleotide substitution (SBS), 78 double base substitutions (DBS) and 83 indel (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 characterised by the predominant substitutional class(es), and also the predominant sequences in those classes.\textsuperscript{23} The relative
mutational signature contribution was determined relative to the total tumour mutational burden (TMB).

To classify the samples as smoking or non-smoking associated, we calculated the proportion of the relative contribution of single base mutational signature SBS4 (tobacco smoking) compared to 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 to the summed relative contributions of ID1 (mismatch repair deficiency (MMRd)/age), ID2 (MMRd/age) and ID3. As inspired by Lee et al.\textsuperscript{7}, we utilized 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 \textit{smoking high} and \textit{smoking low}; respective to the presence of these proportions and absolute counts. Prior to clustering, these values were centred and scaled appropriately.

\textbf{Statistics}

Statistical analysis of the clinical characteristics was performed using IBM SPSS Statistics software, version 25. Continuous data was compared with Student’s T test or Mann-Whitney U test as appropriate. Means are presented with standard deviations (SD), medians with interquartile ranges (IQR). Categorical data was compared with a Chi Square test. Correlations were analysed with the Spearman’s rho. Genomic differences were tested in the statistical platform R (v4.1.1) using the Wilcoxon Rank-Sum Test with multiple testing correction (Benjamini-Hochberg). Mutational enrichment (or depletion) of genes were 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 analysed (Supplementary Figure S1). The cohort consisted of a majority of females (57%), mean age at diagnosis was 62 (±10) years, and adenocarcinoma was the most prevalent histology (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 6 (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.

Based on the supervised signature-based clustering, we categorized 169 samples as smoking high, and 147 samples as smoking low (Figure 1a, Supplementary Figure S2). The signature-based clusters differed significantly with regard to sex, smoking history, prior treatment and histology (Table 1). The distribution of recorded smoking history within these signature-based clusters is depicted in Figure 1b. 4% of 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). 30% of squamous cell carcinomas (N=10) were included in the smoking low cluster, of which 5 patients had an active smoking history, 4 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).
<table>
<thead>
<tr>
<th>Table 1. Demographical and clinical characteristics of the patients stratified by signature-based clustering.*</th>
<th>NSCLC smoking high ( (N = 169) )</th>
<th>NSCLC smoking low ( (N = 147) )</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex (N, %)</strong></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Female</td>
<td>82 (49%)</td>
<td>99 (67%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>87 (51%)</td>
<td>48 (33%)</td>
<td></td>
</tr>
<tr>
<td><strong>Age at diagnosis</strong></td>
<td>Mean (SD)</td>
<td>62 (10)</td>
<td>62 (11)</td>
</tr>
<tr>
<td><strong>Tumour stage at diagnosis</strong> ( (N, %) )</td>
<td></td>
<td></td>
<td>0.147</td>
</tr>
<tr>
<td>I</td>
<td>6 (4%)</td>
<td>7 (5%)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>8 (5%)</td>
<td>4 (3%)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>40 (24%)</td>
<td>20 (14%)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>112 (66%)</td>
<td>114 (78%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>3 (2%)</td>
<td>2 (1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Smoking history (N, %)</strong></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Never</td>
<td>4 (2%)</td>
<td>85 (58%)</td>
<td></td>
</tr>
<tr>
<td>Former</td>
<td>133 (79%)</td>
<td>49 (33%)</td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>29 (17%)</td>
<td>7 (5%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>3 (2%)</td>
<td>6 (4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Pack years</strong></td>
<td>Median (IQR)</td>
<td>28 (17–40)</td>
<td>14 (6–25)</td>
</tr>
<tr>
<td>Unknown</td>
<td>30</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td><strong>Histology (N, %)</strong></td>
<td></td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>114 (67%)</td>
<td>124 (84%)</td>
<td></td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>23 (14%)</td>
<td>10 (7%)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>32 (19%)</td>
<td>13 (9%)</td>
<td></td>
</tr>
<tr>
<td><strong>PD-L1 score (N, %)</strong></td>
<td></td>
<td></td>
<td>0.70</td>
</tr>
<tr>
<td>&lt;1%</td>
<td>41 (24%)</td>
<td>39 (27%)</td>
<td></td>
</tr>
<tr>
<td>1-50%</td>
<td>36 (21%)</td>
<td>41 (28%)</td>
<td></td>
</tr>
<tr>
<td>≥50%</td>
<td>26 (15%)</td>
<td>22 (15%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>66 (39%)</td>
<td>45 (31%)</td>
<td></td>
</tr>
<tr>
<td><strong>Pre-biopsy regimen (N, %)</strong></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chemotherapy/immuno-therapy/other</td>
<td>103 (61%)</td>
<td>36 (24%)</td>
<td></td>
</tr>
<tr>
<td>TKI</td>
<td>20 (12%)</td>
<td>91 (62%)</td>
<td></td>
</tr>
</tbody>
</table>
Overview and differences of the genomic landscape between signature-based clusters

An overview of genomic characteristics has been summarized in Figure 2, grouped by signature-based clustering and ordered by descending tumour mutational burden (TMB). Recurrent copy-number alterations and chromosomal-arm aberrations (from GISTIC2 analysis) for the entire cohort, and per signature-based cluster, can be found within Supplementary Figure 3. Comparing the samples with and without putative chromothripsis events, chromothripsis did not significantly impact driver genes, mutational signatures or TMB within the entire cohort, as well as 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 (Figure 3a,c). No major differences between genome-wide ploidy and chromosomal arms copy-number alterations could be observed (Figure 3b, Supplementary Figure 3b-c). Compared to the smoking low cluster, median TMB was almost a fivefold higher within the smoking high cluster (4 versus 19 mutations/Mb, P < 0.001) (Figure 3a; Supplementary Table S1). In addition, TMB was strongly correlated to absolute SBS4 contribution (Spearman’s rho: 0.87, P < 0.001). However, 26 (15%) of samples in the smoking high cluster had a low TMB (<5). In addition, 12 (8%) samples within the smoking low cluster had a high TMB (≥10), yet generally harboured mutational signatures related to MMRd and APOBEC activity rather than those related to smoking; 3 of these samples indeed were classified as microsatellite instable (MSI) tumours.

Beyond the expected differences in SBS4 and ID3 abundance due to their usage as clustering features (Supplementary Table S1), we also observed differences in additional mutational signatures between the two clusters (Figure 3c). Within the smoking high cluster, we observed significantly higher contributions of SBS signatures SBS18, SBS29, and SBS5 compared to the smoking low cluster. SBS4, SBS18 and SBS29 are all characterised 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 (ROS), and SBS29 is associated with tobacco chewing. SBS5 also has an unknown aetiology, although its mutational spectrum seems to be increased in cancers related to tobacco exposure and seems to be related to age. With regard to the DBS signatures, DBS2, which

<table>
<thead>
<tr>
<th></th>
<th>Untreated</th>
<th>Median (IQR)</th>
<th>20 (14%)</th>
<th>&lt;0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines of systemic treatment before biopsy</td>
<td>46 (27%)</td>
<td>1 (0–2)</td>
<td>2 (1–3)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
| * Percentages may not add up to 100% due to rounding.
is also associated with exposure to tobacco smoke, had the highest contribution in the smoking high cluster.\textsuperscript{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 characterised 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.\textsuperscript{10} It has been suggested that the activation of APOBEC in cancer could be caused by previous viral infection or tissue inflammation, which suggest a role of inflammation in the development of non-smoking associated lung cancer.\textsuperscript{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.\textsuperscript{10} DBS6 had the highest DBS signature contribution and ID1 had the highest ID signature contribution in the smoking low cluster. DBS6 still has an unknown aetiology, but seems to be related to age.\textsuperscript{10} ID1 is associated with age and with slippage of the template DNA strand during DNA replication. It is often found in cancers with DNA MMRd and MSI.\textsuperscript{10}

The differences in TMB and relative mutational signature contribution between the two clusters (figure 3a,c), were also investigated in the different histological subtypes (Supplementary Table S2). Most observed differences were consistent over the histological subtypes, however, both APOBEC signatures (SBS2 and SBS13) did not differ significantly between the smoking high and smoking low cluster in the squamous cell subgroup (3\% versus 5\%, \(P=0.062\) and 5\% versus 13\%, \(P=0.237\), respectively), but did in the adenocarcinoma subgroup (\(P<0.001\)). The differences in smoking related signatures were consistent over all histological subgroups.

Altered landscape of putative driver genes

Next, we investigated differences and/or enrichment within the somatic inventory of perturbed genes between the two signature-based clusters (Figure 3d, Supplementary Figure S4-5). Of the known driver oncogenes, \textit{EGFR} mutations and \textit{ALK} fusions were significantly more prevalent in the smoking low cluster than in the smoking high cluster (50\% vs 9\%, \(P<0.001\); 13\% vs 0\%, \(P<0.001\); respectively). \textit{KRAS} mutations were significantly more prevalent in the smoking high cluster (28\% vs 5\%, \(P<0.001\)). Table 2 illustrates the differences in frequency of oncogene driver alterations in the non-squamous cell carcinoma samples between the signature-based clusters and the groups based on clinical smoking history.
Table 2. Frequency of oncogene driver alterations in non-squamous cell NSCLC.

<table>
<thead>
<tr>
<th>Alteration</th>
<th>Total N=</th>
<th>Smoking low cluster N (%)</th>
<th>Never smoked N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EGFR alteration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 19 deletion</td>
<td>38</td>
<td>35 (92%)</td>
<td>24 (63%)</td>
</tr>
<tr>
<td>L858R</td>
<td>28</td>
<td>22 (79%)</td>
<td>17 (61%)</td>
</tr>
<tr>
<td>Exon 20 insertion</td>
<td>4</td>
<td>2 (50%)</td>
<td>1 (25%)</td>
</tr>
<tr>
<td>Other</td>
<td>15</td>
<td>12 (80%)</td>
<td>7 (47%)</td>
</tr>
<tr>
<td><strong>ALK fusion</strong></td>
<td>19</td>
<td>19 (100%)</td>
<td>14 (74%)</td>
</tr>
<tr>
<td><strong>RET fusion</strong></td>
<td>3</td>
<td>3 (100%)</td>
<td>1 (33%)</td>
</tr>
<tr>
<td><strong>ROS1 fusion</strong></td>
<td>6</td>
<td>5 (83%)</td>
<td>4 (67%)</td>
</tr>
<tr>
<td><strong>MET amplification</strong></td>
<td>21</td>
<td>10 (48%)</td>
<td>9 (43%)</td>
</tr>
<tr>
<td><strong>BRAF mutation</strong></td>
<td>21</td>
<td>11 (52%)</td>
<td>7 (33%)</td>
</tr>
<tr>
<td>V600E</td>
<td>12</td>
<td>10 (83%)</td>
<td>6 (50%)</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>1 (13%)</td>
<td>1 (13%)</td>
</tr>
<tr>
<td><strong>ERBB2/Her2 amplification</strong></td>
<td>13</td>
<td>7 (54%)</td>
<td>3 (23%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alteration</th>
<th>Total N=</th>
<th>Smoking high cluster N (%)</th>
<th>Active smoking history N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KRAS mutation</strong></td>
<td>50</td>
<td>45 (90%)</td>
<td>43 (86%)</td>
</tr>
<tr>
<td>G12C</td>
<td>19</td>
<td>19 (100%)</td>
<td>18 (95%)</td>
</tr>
<tr>
<td>G12A</td>
<td>6</td>
<td>5 (83%)</td>
<td>5 (83%)</td>
</tr>
<tr>
<td>G12V</td>
<td>13</td>
<td>11 (85%)</td>
<td>10 (77%)</td>
</tr>
<tr>
<td>G12D</td>
<td>5</td>
<td>4 (80%)</td>
<td>4 (80%)</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>6 (86%)</td>
<td>6 (86%)</td>
</tr>
<tr>
<td>Codon 12</td>
<td>3</td>
<td>3 (100%)</td>
<td>3 (100%)</td>
</tr>
<tr>
<td>Codon 61</td>
<td>3</td>
<td>2 (67%)</td>
<td>2 (67%)</td>
</tr>
<tr>
<td>K117N</td>
<td>1</td>
<td>1 (100%)</td>
<td>1 (100%)</td>
</tr>
<tr>
<td><strong>MET exon 14 skipping alteration</strong></td>
<td>6</td>
<td>3 (50%)</td>
<td>4 (67%)</td>
</tr>
</tbody>
</table>

Discordances

We further analysed the samples of patients who had never smoked but were included in the smoking high cluster (N=4) and found that all harboured high (>10) or medium-high (>5) TMB. In addition, these samples also reflected the mutational signatures found within the rest of the smoking
high cluster and showed 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 harbour a known driver oncogene.

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% harboured a known oncogene driver alteration: 22 EGFR alterations, 5 BRAF mutations, 5 ALK fusions, 4 ERBB2/Her2 amplifications, 2 KRAS mutations (one G12V and one G12D), 1 MET amplification, 1 MET exon 14 skipping mutation, 2 RET fusions and 1 ROS1 fusion.

### Table 3. TMB and relative dominant mutational signature contribution in discordant samples.

<table>
<thead>
<tr>
<th></th>
<th>Smoking high without smoking history (N=4)</th>
<th>Smoking low with active smoking history (N=56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMB (IQR)</td>
<td>11.7 (7.8–17.9)</td>
<td>5.1 (3.6–6.7)</td>
</tr>
<tr>
<td>SBS4, % (IQR)</td>
<td>23.2 (19.0–25.6)</td>
<td>5.1 (2.2–9.4)</td>
</tr>
<tr>
<td>SBS18, % (IQR)</td>
<td>23.9 (21.1–27.0)</td>
<td>3.7 (0.4–10.5)</td>
</tr>
<tr>
<td>SBS29, % (IQR)</td>
<td>12.1 (11.2–15.3)</td>
<td>4.7 (0.6–7.4)</td>
</tr>
<tr>
<td>SBS2, % (IQR)</td>
<td>1.7 (0.5–3.0)</td>
<td>5.8 (3.5–8.2)</td>
</tr>
<tr>
<td>SBS13, % (IQR)</td>
<td>2.1 (1.1–2.9)</td>
<td>8.5 (3.3–17.1)</td>
</tr>
<tr>
<td>ID3, % (IQR)</td>
<td>50.4 (42.6–63.9)</td>
<td>17.9 (12.4–23.5)</td>
</tr>
</tbody>
</table>

**In silico analysis targeted panel**

Lastly, 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, California, US) 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 roughly ~1.94Mb spread throughout the genome and has also been used for signature analysis previously.25 90% of the samples (N=285) harboured enough SBS mutations to derive the SBS signatures for subsequent analysis. Of those samples, 54% (N=155) also harboured enough InDel mutations to derive the ID signatures. Next, we investigated the concordance between the TSO500 and WGS clustering, and if this concordance was effected by extension of the TSO500 regions. Utilizing 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.
The distribution of smoking history within the clusters again revealed a considerable degree of discordance between smoking history and mutational signatures (Figure S6c,d, Figure 1b 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 to WGS, as several (non-coding) 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 (Figure S6a (upper track) and S6b). Extensions of the TSO500 regions also generally improved the concordance (F1) of both approaches between WGS and genomic subsets (Figure S6a).
Discussion

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

Our classification shows that there is large overlap in clinical smoking history and classification based on SBS4 and ID3 contributions. This also revealed a degree of discordance between these two grouping methods. A small minority of patients who had never smoked, had tumours 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 healthcare professionals or passive smoke exposure. In the tumours 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, since patients with an active smoking history in the smoking low cluster had fewer pack years than those in the smoking high cluster. 26% of 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. showed that in their cohort of lung adenocarcinoma about one-third of tumours from patients with an active smoking history had no or minor SBS4 contribution. Devarakonda et al. used a statistical model, including TMB and SBS4, to infer smoking status and excluded 4 out of 88 tumour samples of patients who reportedly had never smoked. This confirms that the mutational processes that have occurred in the tumour 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 tumours with a high TMB, high SBS4, SBS18 and SBS29 contribution and with KRAS mutations were predominantly classified as smoking high. These signatures and most KRAS mutations in NSCLC are characterised by transversion mutations, which would explain why they group together. Tumours 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 tumours. The tumours from patients who had never smoked but were clustered as smoking high, had a similar genotype to the other tumours in the smoking high cluster, which suggests smoke exposure despite a negative smoking history. Tumours 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 NSCLC’s based on similar genomic context rather than reported smoking history. As TMB was strongly correlated to SBS4 contribution, this might suggest that classification based on TMB would yield similar results. However, 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 upregulation 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 based on clinical history could have several clinical implications. First, as we have shown 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 rarer 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 (KDD). Although rare, these oncogenic drivers can provide an important target for treatment, e.g. several reports have demonstrated sensitivity of tumours with an *EGFR*-KDD to *EGFR* TKI’s.\(^{31, 32}\) 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’ wellbeing should not be underestimated. Next, in randomized trials investigating immunotherapy smoking history can be of special interest as a predictive biomarker due 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. However, as a considerable percentage of patients with an active smoking history did not actually harbour high (or any) smoking signature contribution, the reliability of smoking history as stratification factor and predictive biomarker in these trials can be questioned. As 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.\(^{34, 35}\) Also, it is possible that the subgroup of patients with low SBS4 contribution
would derive less benefit from immunotherapy than would be expected based on smoking history or PD-L1 expression, since PD-L1 expression did not differ between the smoking high and smoking low cluster. Additionally, 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 that do derive benefit from immunotherapy. However, 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 this data is already available. Also, the WIDE study investigators have recently shown that WGS for patients with metastatic cancer is feasible in routine clinical practice. 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 tumour tissue to be sent in for genome testing during the screening period. Our TSO500 in silico analysis showed that the TSO500 panel allows for SBS mutational signature calling in the vast majority of cases, whereas the ID signatures are more challenging to retrieve. However, as 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 common practice, mutational signature analysis with targeted panels could still help identify those with a higher likelihood of harbouring a (rare) oncogenic driver. However, the optimal cut off between a high or low smoking signature contribution does still warrant further investigation.

This study has certain limitations that should be considered. Most importantly, a golden standard in distinguishing smoking associated from non-smoking associated carcinogenesis is lacking. In the absence of a golden standard, accuracy analyses are unreliable and were therefore not performed. Additionally, mutational signatures infer the dominant processes of mutagenesis within a tumour 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. Our samples of EGFR-driven NSCLC with high SBS4 contribution further illustrates 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. Lastly, 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. Additionally, previously published genomic cohorts are often small, only focus on patients without smoking history, or primarily included early-stage lung cancer.\textsuperscript{8, 9, 38} Our comprehensive cohort allows for an in-depth analysis of metastatic tumours from patients with and without smoking history.

To conclude, our mutational signature-based classification of smoking associated and non-smoking associated NSCLC is more accurate in grouping tumours with similar genomic contexts together compared to classification based on 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. Additionally, it could aid in optimizing diagnostic strategies in daily practice that are currently still influenced by clinical smoking history, such as the pursuit of identification of rarer 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.
Acknowledgements

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 (CPCT) have made available to the study.

The authors thank Arne van Hoeck for sharing the lung cancer mutational signatures.

References

Figure legends

Figure 1. Signature-based clustering of NSCLC into smoking high and low groups.

a) Overview of two of the employed clustering features 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 vs. the genomic clusters. N is indicated on the outer dial.

Figure 2. Overview of the genomic landscape.

For each sample (column), we display common genomic characteristics. a) Signature-based cluster designation. b) Genome-wide Tumour Mutational Burden (TMB). c) Total no. 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 f, but for DBS signatures. h) Same as f, but for ID signatures. i) Histology. j) Reported clinical smoking history. k) Microsatellite instability (MSI) status. l) Homologous recombination deficient (HRD) status as determined by CHORD. m) Presence of chromothripsis. n) Presence of known driver alteration(s). o) Previous systemic therapy before biopsy.

Figure 3. Genomic differences between signature-based smoking high and smoking low samples.

a) Box plot of tumour mutational burden (TMB) between the signature-based classification of samples; the median, Q1 and Q3 are represented. Statistical significance is shown on top. Cut-off of high TMB (≥10 mutations/Mb) is indicated with a red horizontal line. b) Same as a) but depicting genome-wide ploidy. c) Same as a) but depicting mutational signatures (SBS, DBS and ID) with statistically significant differences (q < 0.05) and harbouring a minimal median relative contribution of ≥5% in either group. Proposed aetiologies are shown on top. d) Statistically significant enriched genes between the signature-based classification of samples. Size of points represent the adjusted P (q-value), with genes having greater enrichment (or depletion) depicted as larger.
List of supplementary material

- Supplementary Data 1: Methods (Microsoft Word file)
- Supplementary Data 2: Supplementary Figure Captions (Microsoft Word file)
- Supplementary Figure S1 (PDF file)
- Supplementary Figure S2 (PDF file)
- Supplementary Figure S3 (PDF file)
- Supplementary Figure S4 (PDF file)
- Supplementary Figure S5 (PDF file)
- Supplementary Figure S6 (PDF file)
- Supplementary Table S1 (Microsoft Word file)
- Supplementary Table S2 (Microsoft Word file)
Tobacco smoking related mutational signatures in classifying smoking associated and non-smoking associated non-small cell lung cancer.

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

Author contributions

Sophie M. Ernst: Conceptualization, methodology, formal analysis, investigations, data curation, writing – original draft preparation, review & editing.

Joanne Mankor: Conceptualization, methodology, investigations, data curation, writing – review & editing.

Job van Riet: Conceptualization, software, formal analysis, writing – review & editing.

Jan H. von der Thüsen: Conceptualization, writing – review & editing.

Hendrikus J. Dubbink: Conceptualization, writing – review & editing.

Joachim G.J.V. Aerts: Conceptualization, writing – review & editing.

Adrianus J. de Langen: Conceptualization, writing – review & editing.

Egbert F. Smit: Conceptualization, writing – review & editing.

Anne-Marie C. Dingemans: Conceptualization, writing – review & editing, supervision.

Kim Monkhorst: Conceptualization, writing – review & editing, supervision.