Genomic and Transcriptional Alterations in Lung Adenocarcinoma in Relation to Smoking History

Anna Karlsson1,2, Markus Ringnér1,2, Martin Lauss1, Johan Botling3, Patrick Micke3, Maria Planck1,2, and Johan Staaf1,2

Abstract

Purpose: Cigarette smoking is the major pathogenic factor for lung cancer. The precise mechanisms of tobacco-related carcinogenesis and its effect on the genomic and transcriptional landscape in lung cancer are not fully understood.

Experimental Design: A total of 1,398 (277 never-smokers and 1,121 smokers) genomic and 1,449 (370 never-smokers and 1,079 smokers) transcriptional profiles were assembled from public lung adenocarcinoma cohorts, including matched next-generation DNA-sequencing data (n = 423). Unsupervised and supervised methods were used to identify smoking-related copy-number alterations (CNAs), predictors of smoking status, and molecular subgroups.

Results: Genomic meta-analyses showed that never-smokers and smokers harbored a similar frequency of total CNAs, although specific regions (5q, 8q, 16p, 19p, and 22q) displayed a 20% to 30% frequency difference between the two groups. Importantly, supervised classification analyses based on CNAs or gene expression could not accurately predict smoking status (balanced accuracies ~60% to 80%). However, unsupervised multicohort transcriptional profiling stratified adenocarcinomas into distinct molecular subgroups with specific patterns of CNAs, oncogenic mutations, and mutation transversion frequencies that were independent of the smoking status. One subgroup included approximately 55% to 90% of never-smokers and approximately 20% to 40% of smokers (both current and former) with molecular and clinical features of a less aggressive and smoking-unrelated disease. Given the considerable intragroup heterogeneity in smoking-defined subgroups, especially among former smokers, our results emphasize the clinical importance of accurate molecular characterization of lung adenocarcinoma.

Conclusions: The landscape of smoking-related CNAs and transcriptional alterations in adenocarcinomas is complex, heterogeneous, and with moderate differences. Our results support a molecularly distinct less aggressive adenocarcinoma entity, arising in never-smokers and a subset of smokers. Clin Cancer Res; 20(18); 4912–24. ©2014 AACR.

Introduction

Lung cancer is the leading cause of cancer-related death worldwide, with cigarette smoking as the principal cause (1). Cigarette smoke consists of a complex mixture of chemicals causing direct or indirect damage to the respiratory epithelium and its genome (2). Consistently, accumulation of genomic alterations, including increased mutation frequencies and differences in mutation spectra, is observed in lung cancers arising in smokers compared with never-smokers (3, 4). However, up to 25% of lung cancer cases have been estimated to arise in never-smokers, which would rank it as the seventh cause of cancer death worldwide if considered a separate disease (5). Several etiological factors for lung cancer in never-smokers have been suggested, including environmental tobacco exposure, indoor and outdoor pollution, various occupational carcinogens, and genetic susceptibility (5). Lung cancer in never-smokers has been suggested to represent a distinct disease entity compared with tumors arising in smokers (1, 5). Specifically, lung cancer in never-smokers has been associated with female sex, East Asian ethnicity, adenocarcinoma histology, differences in mutational spectra and overall number of mutations, higher frequency of ALK rearrangements and EGFR mutations, and lower frequency of KRAS mutations compared with tumors arising in smokers (4–7). Recent sequencing studies have indicated that a large fraction of lung cancers in never-smokers harbor mutually exclusive
Smoking and Genomic Alterations in Lung Adenocarcinoma

Translational Relevance
Smoking is the major pathogenic factor for lung cancer. Lung cancer in never-smokers has been proposed to represent a separate disease entity, primarily presenting as adenocarcinomas. However, whether specific genomic and/or transcriptional alterations are robustly associated with smoking status in lung adenocarcinoma is conflicting. Here, we demonstrate a few consistent smoking-related genomic and transcriptional differences, but also a considerable heterogeneity within the studied smoking groups, which can be explained by clinicopathologic and/or tumor-associated factors. However, we did not observe a distinct molecular entity based on genomic and/or transcriptional alterations for adenocarcinomas arising in never-smokers. Instead, most tumors arising in never-smokers, together with a specific subset of smoker’s tumors, seem to represent a more distinct molecular entity of less aggressive adenocarcinomas unrelated to smoking. The extent of shared carcinogenesis pathways and differences in response to treatment and patient outcome for these cancers remains to be elucidated.

oncogenic driver mutations that may be vital for the viability of the tumor cells (6, 8).

In the literature, a varying spectrum of genomic alterations have been reported in adenocarcinomas arising in never-smokers and smokers, including regions on chromosome 5q, 7p, 7q, 8q, 10q, and 16p (9–13). Moreover, conflicting reports exist on whether smokers overall display more or less copy-number alterations (CNAs) than never-smokers (12, 14, 15). Together, this indicates a significant heterogeneity within smoking-defined subgroups of adenocarcinoma. Numerous studies have reported transcriptional differences between never-smokers and smokers in both normal airway epithelium and adenocarcinoma tumor tissue (16–23). In addition, gene expression–based molecular subgroups in lung adenocarcinoma, e.g., the bronchioid (24) and terminal respiratory unit (TRU; ref. 25) molecular subtypes, have also been associated with patient smoking history. For instance, in the Wilkerson and colleagues (26) meta-analysis, 60% of all never-smokers were classified as bronchioid, representing 30% of this subgroup. However, unsupervised analyses of genome-wide expression patterns in adenocarcinomas have not yet identified never-smokers as a separate and distinct transcriptional entity without notable inclusion of smokers, challenging the hypothesis of a separate disease entity (13, 17, 22, 25, 27, 28). Thus, further investigations are warranted for improved understanding of the molecular pathogenesis, especially into whether specific CNAs or transcriptional differences are actually acquired depending on smoking status in otherwise clinically and pathologically similar tumors.

In this study, we aim to provide a comprehensive survey of genomic and transcriptional alterations in lung adenocarcinomas associated with patient smoking history. On the basis of a multicohort study design, including independent discovery and validation cohorts, we analyzed 1,398 genomic and 1,449 transcriptional profiles for smoking-related alterations (Fig. 1). We demonstrate a considerable heterogeneity at the genomic and transcriptional level within the smoking-defined subgroups that precludes stringent classification of smoking status based on CNAs or transcriptional patterns by supervised methods. Overall, our results indicate that the genomic and transcriptional landscape of lung adenocarcinomas of smokers and never-smokers is not as distinct, and that there are common mechanisms in the tumorigenesis in never-smokers and smokers.

Materials and Methods
Genomic tumor cohort
Published genomic profiles from 1,398 adenocarcinomas with available patient smoking history were collected into a genomic discovery cohort as previously described (29, 30; Table 1 and Supplementary Tables S1 and S2). Heavy smokers were defined as smokers with >60 pack-years consistent with Huang and colleagues (14).

Gene expression cohorts
Gene expression profiles from 841 adenocarcinomas with available patient smoking history were collected from five microarray-based studies (13, 18, 27, 31, 32). Patients from Chitale and colleagues (32) were divided into two cohorts according to their different Affymetrix platforms (U133A and U133 2plus) creating six microarray discovery cohorts (Table 1; Supplementary Table S2; and Fig. 1B). A total of 221 patients overlapped between the 1,398-sample genomic cohort and the 841-sample expression cohort.

Supervised gene expression classification results were validated in 115 patients from Der and colleagues (33) and 70 patients from Tarca and colleagues (34; Fig. 1B). Validation and extension of unsupervised gene expression results were performed in whole transcriptome RNA-sequencing profiles from 435 independent adenocarcinomas from The Cancer Genome Atlas (TCGA; 423 with smoking status) including matched whole-exome sequencing data (Fig. 1B; Table 1; and Supplementary Methods). Included studies were performed in both western and Asian countries.

Genomic analyses
Normalized copy-number estimates were generated and/or assembled, and the fraction of the genome altered by CNA (CN-FGA) was calculated as described (30). Smoking-related CNAs were identified through a genome-wide screen of 12.698 sequential ~200-kbp segments as described (29). Multinomial logistic regression analysis, similar to Broet and colleagues (11), was used to investigate the relationship with possible confounding factors (gender, EGFR mutation, cohort, and stage) for smoking-related CNAs. Identified CNA regions were cross-compared with a set of 90 recurrent focal CNAs (59 gains, 31 losses) identified in our previous study (mGISTIC regions; ref. 29).
Supervised classification was performed using the caret R package between never-smokers and smokers (current or former), never-smokers and current smokers, and never-smokers and former smokers based on smoking-related CNAs from the current study (above) or smoking-related genomic signatures from Massion and colleagues (15), Thu and colleagues (12), Broet and colleagues (11), Weir and colleagues (10), or recurrent genome-wide regions from Planck and colleagues (29). Classifier performance was evaluated using balanced accuracy, which avoids inflated performance estimates on imbalanced cohorts. The 1,398-sample cohort was divided into 50%/50% training and test sets, which were balanced for individual cohort, smoking status, and EGFR mutation status. Three different types of classifier models [partitioning around medoids (PAM), linear support vector machine, and linear discriminant analysis] with parameter tuning based on 4-fold cross validation were used to derive classifiers in the training set. Each classifier model was iterated up to 10 times with different training and test samples to assess average performance. Data processing steps are further described in Supplementary Methods and refs. 29 and 30.

Gene expression analyses

Affymetrix gene expression cohorts were individually normalized as described by ref. 29, whereas non-Affymetrix cohorts were processed as described in Supplementary Methods. Supervised classification was performed using the caret R package between never-smokers and smokers (current or former), and between never-smokers and current smokers, in six and four microarray cohorts, respectively. Each cohort was used to train a classifier that was next applied to remaining cohorts for evaluation. Different feature selection criteria, four different types of classifier models (PAM, linear support vector machine, random forest, and generalized boosted regression) with parameter tuning based on 4-fold cross validation were used for each training cohort. In total, >29 combinations were constructed for each training cohort and evaluated in separate test sets (Supplementary Methods).

Unsupervised group discovery was performed by consensus clustering using the ConsensusClusterPlus R package (35). For microarray cohorts, consensus clustering was performed individually in each cohort after three different probe set variance filters (expression SD >0.3,
### Table 1. Clinical characteristics of patients with adenocarcinoma in genomic and gene expression cohorts

<table>
<thead>
<tr>
<th>Genomic cohort</th>
<th>Okayama et al. (27)</th>
<th>Chitale U133A (32)</th>
<th>Chitale U133 2plus (32)</th>
<th>Fouret et al. (13)</th>
<th>Landi et al. (18)</th>
<th>Shedden et al. (31)</th>
<th>TCGA</th>
<th>Der et al. (33)</th>
<th>Tarca et al. (34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>Discovery</td>
<td>Discovery</td>
<td>Discovery</td>
<td>Discovery</td>
<td>Discovery</td>
<td>Discovery</td>
<td></td>
<td>Discovery</td>
<td>Discovery</td>
</tr>
<tr>
<td>Data type</td>
<td>Copy number</td>
<td>Copy number</td>
<td>Copy number</td>
<td>Copy number</td>
<td>Copy number</td>
<td>Copy number</td>
<td></td>
<td>Copy number</td>
<td>Copy number</td>
</tr>
<tr>
<td>Total number of patients</td>
<td>1,398</td>
<td>226</td>
<td>91</td>
<td>102</td>
<td>103</td>
<td>58</td>
<td>356</td>
<td>435</td>
<td>115</td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never-smokers</td>
<td>277</td>
<td>115</td>
<td>17</td>
<td>19</td>
<td>63</td>
<td>16</td>
<td>33</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>Smokers</td>
<td>1,121</td>
<td>111</td>
<td>73</td>
<td>83</td>
<td>40</td>
<td>42</td>
<td>229</td>
<td>358</td>
<td>92</td>
</tr>
<tr>
<td>Current smokers</td>
<td>391</td>
<td>—</td>
<td>13</td>
<td>12</td>
<td>—</td>
<td>24</td>
<td>20</td>
<td>102</td>
<td>36</td>
</tr>
<tr>
<td>Former smokers</td>
<td>567</td>
<td>—</td>
<td>60</td>
<td>71</td>
<td>—</td>
<td>18</td>
<td>209</td>
<td>256</td>
<td>56</td>
</tr>
<tr>
<td>Pack-years (median)</td>
<td>40</td>
<td>—</td>
<td>37</td>
<td>34</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>40</td>
<td>—</td>
</tr>
<tr>
<td>Heavy smokers (%)</td>
<td>22%</td>
<td>—</td>
<td>21%</td>
<td>28%</td>
<td>—</td>
<td>—</td>
<td>17%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gender</td>
<td>Male/female</td>
<td>586/695</td>
<td>105/121</td>
<td>41/50</td>
<td>42/60</td>
<td>15/84</td>
<td>35/23</td>
<td>189/166</td>
<td>201/234</td>
</tr>
<tr>
<td>Mutation status</td>
<td>EGFR-mutated</td>
<td>205</td>
<td>127</td>
<td>15</td>
<td>24</td>
<td>49</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>EGFRwt and KRASwt</td>
<td>564</td>
<td>79</td>
<td>65</td>
<td>42</td>
<td>33</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>739</td>
<td>168</td>
<td>53</td>
<td>70</td>
<td>57</td>
<td>22</td>
<td>224</td>
<td>238</td>
<td>83</td>
</tr>
<tr>
<td>II</td>
<td>235</td>
<td>58</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>21</td>
<td>77</td>
<td>100</td>
<td>32</td>
</tr>
<tr>
<td>III</td>
<td>272</td>
<td>0</td>
<td>18</td>
<td>17</td>
<td>32</td>
<td>12</td>
<td>51</td>
<td>74</td>
<td>—</td>
</tr>
<tr>
<td>IV</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>22</td>
<td>—</td>
</tr>
<tr>
<td>Platform</td>
<td>—</td>
<td>Affymetrix</td>
<td>Affymetrix</td>
<td>Affymetrix</td>
<td>Affymetrix</td>
<td>Affymetrix</td>
<td>RNAseq</td>
<td>Affymetrix U133 2plus</td>
<td>Affymetrix U133 2plus</td>
</tr>
</tbody>
</table>

* EGFR and KRAS mutations taken from nonsilent mutations in Mutation Annotation Format (MAF) file and not listed here.

b Heavy smoker defined as smoker with >60 pack-years. Value corresponds to percentage of all smokers with pack-year annotation in a given cohort.
>0.5, and >1) representing increased stringency in selection of the most variant genes across a cohort (Fig. 1B and Supplementary Methods). An expression SD >1 was used as filter before consensus clustering of the 435 TCGA cases using a three-group consensus cluster solution.

Tumors were classified as bronchioid, squamoid, and magnoi (26), and scored according to different expression metagenes, including a proliferation/chromosome instability signature (CIN70; ref. 36), a TRU-like signature (25), an alveolar/bronchial signature (28), and a distal airway stem cell (DASC)–like subtype (37) as described (Supplementary Methods).

Results

Cohort demographics with respect to smoking history

In total, we analyzed 1,398 genomic and 1,449 transcriptional profiles for smoking-related alterations (Table 1, Supplementary Tables S1 and S2). Consistent with the literature (5, 7), we observed higher rates of EGFR mutations and female gender in never-smokers (29%–61% of never-smokers carried EGFR mutations and 72%–84% were females), whereas more KRAS mutations were found in smokers (22%–35% of smokers carried KRAS mutations). This analysis was performed in (i) the 1,398-sample genomic cohort, (ii) the combined six discovery gene expression cohorts, and (iii) the TCGA gene expression cohort (P < 0.001, Fisher exact test). Tumor stage was not associated with smoking history in any of the cohorts. Overall survival was associated with smoking history only in one of the cohorts studied [Okayama and colleagues (27), log-rank P = 0.02, 5-year censored data].

Overall pattern of CNAs in adenocarcinoma stratified by smoking status

Stratification of adenocarcinomas based on patient smoking history revealed both common regions of CNAs across groups, e.g., gain of chromosome 1q, 5p, and loss of 3p, 6q, 9, 13q, and regions with apparently different prevalence between groups, e.g., gains on 5q, 7p and 16p in never-smokers and 8q in current smokers (Supplementary Fig. S1A–S1D). No significant difference in the overall amount of CNAs per sample, CN-FGA, was observed between never-smokers and smokers, whereas current smokers showed a minor increase in CN-FGA compared with never-smokers and former smokers (Fig. 2A). However, the higher CN-FGA in current smokers was not significant in a bootstrap resampling (P = 0.11, 10,000 bootstraps), implying that the observed differences are to some extent related to individual studies. Supporting this hypothesis, we found that individual cohorts varied in whether never-smokers or smokers showed more or less CNAs (Fig. 2B).

Moreover, we did not find significant differences in total CN-FGA, or FGA for copy-number gain or loss specifically, between heavy (>60 pack-years) and light smokers in (i) all smokers, (ii) current smokers only, or (iii) former smokers only, in contrast to Huang and colleagues (14; P > 0.05, Wilcoxon test). Similar results were found with different pack-year cutoffs between 10 to 80 pack-years, and individually in the four largest cohorts (TCGA, Chitale and colleagues, Weir and colleagues, and Huang and colleagues; refs. 10, 14, 32) for all smokers (P > 0.05, Wilcoxon test).

Genome-wide screen of smoking-related CNAs

CNAs associated with patient smoking status were identified through a genome-wide screen of CNA frequency in...
approximately 200-kbp sequential segments. For the three-group comparison of never-smokers, current smokers, and former smokers, this analysis identified regions of copy-number gain at 5q, 8q, and 16p, and copy-number loss at 5q, 19p, and 22q to differ between the groups, involving in total 8% of all genome-wide segments [false discovery rate (FDR) adjusted Fisher exact test P < 0.05 and >20% difference in CNA frequency; Table 2 and Supplementary Table S3]. All identified regions appeared robust based on bootstrap resampling (P < 0.05, 1,000 permutations per region) and were significantly associated with smoking status in multivariate analysis (Holm adjusted P < 0.05). To further evaluate identified regions from Table 2, we analyzed them in each of the four largest cohorts individually [Weir and colleagues (10), Chitale and colleagues (32), TCGA, and The Clinical Lung Cancer Genome Project (CLCGP); Table 1]. For all regions, there was a strong agreement between cohorts regarding which smoking subgroup showed highest CNA frequency (Supplementary Fig. S1E compared with Table 2). However, only a subset of regions, such as gains at 5q31.3-q32, 5q33.1-q35.3, and 16p13.3-p12.1, and losses at 5q13.3-q35.3 and 22q13.1-q13.33, seemed robustly altered in ≥2 individual cohorts (Fisher exact P value <0.05 and >20% frequency difference; Supplementary Fig. S1F). These cohort-specific analyses highlight the existence of both general and more study-dependent smoking-related CNAs, stressing the need of a multicohort approach.

Similar analysis of smoking-related CNAs for never-smokers versus smokers identified only higher frequencies of copy-number gain at 5qter and 16p in never-smokers. Together, these results indicate that the group of former smokers is more heterogeneous than never-smokers and current smokers.

**Supervised classification of smoking-related CNAs**

To investigate the predictive power of smoking-related CNAs, we performed supervised classification in the 1,398-sample cohort, using three different classifier methods combined with smoking-related CNAs from Table 2, four reported smoking-related genomic signatures (10–12, 15), and a set of recurrent genome-wide focal CNAs (29). Here, we acknowledge that the regions from Table 2 are used for prediction of the same cohort they were derived from, which could inflate results for these regions. Throughout comparisons, the PAM method performed the best, although reaching only moderate accuracies (generally <70% balanced accuracy) in classification of never-smokers/smokers, never-smokers/current smokers, and never-smokers/former smokers (Fig. 3A). Classification accuracies were consistently higher for never-smokers/current smokers compared with never-smokers/former smokers irrespectively of classifier model, supporting that former smokers is a more heterogeneous group.

To further investigate the moderate prediction accuracies, we performed principal component analysis (PCA; ref. 38) in the 1,398-sample cohort, including different clinicopathologic factors (gender, stage, clinical smoking status, EGFR and KRAS mutations), unsupervised gene expression clusters (see below), and molecular adenocarcinoma subtypes (26). Supporting the moderate supervised classification results, and the identification of a smaller set of smoking-related CNAs, we found that smoking status was not a dominant contributor to the total copy-number variation in the cohort (Supplementary Fig. S1G).

**Supervised classification of smoking status based on gene expression patterns**

Given the moderate accuracies in predicting smoking history based on CNAs, we questioned whether classification models built on transcriptional patterns performed better. To identify gene expression signatures predictive of smoking status (never-smokers vs. smokers, or never-

### Table 2. Smoking-related CNAs with >20% frequency difference between never-smokers, current smokers, and former smokers

<table>
<thead>
<tr>
<th>Type</th>
<th>Cytoband</th>
<th>Region*</th>
<th>Size (Mbp)</th>
<th>Number of genes</th>
<th>Most altered group</th>
<th>Focal CNAs (29)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td>5q31.3-q32</td>
<td>chr5:142071001-145089001</td>
<td>3.02</td>
<td>3</td>
<td>Never-smoker</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>5q33.1-q35.3</td>
<td>chr5:147723001-180711001</td>
<td>32.99</td>
<td>201</td>
<td>Never-smoker</td>
<td>Amp_5q35.1</td>
</tr>
<tr>
<td>Gain</td>
<td>8q13.1-q13.2</td>
<td>chr8:67043001-69050001</td>
<td>2.01</td>
<td>15</td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>8q13.3</td>
<td>chr8:72872001-73673001</td>
<td>0.8</td>
<td>3</td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>8q21.11-q21.12</td>
<td>chr8:76325001-80141001</td>
<td>3.82</td>
<td>6</td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td>Gain</td>
<td>8q21.3-q22.1</td>
<td>chr8:93355001-94154001</td>
<td>0.8</td>
<td>0</td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>5q12.1-q13.2</td>
<td>chr5:62799001-73107001</td>
<td>10.31</td>
<td>38</td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>5q13.3-q35.3</td>
<td>chr5:76125001-180711001</td>
<td>104.59</td>
<td>559</td>
<td>Current</td>
<td>Del_5q13.3</td>
</tr>
</tbody>
</table>

*hg18 coordinates.
smokers vs. current smokers) in lung adenocarcinoma, we therefore performed supervised classification in six microarray cohorts (Fig. 1B, discovery cohorts). For each cohort, we trained classifiers based on various feature selection criteria and classifier models, including previously reported smoking-related gene signatures from both normal and tumor tissues (16, 19–21, 39; Supplementary Fig. S2A and S2B and Supplementary Methods). Next, we evaluated each trained classifier from a specific training cohort in remaining cohorts for never-smokers versus smokers (5 test cohorts, 34 different classifiers totaling 170 different tests), or never-smokers versus current smokers (3 test cohorts, 29 different classifiers). However, despite the large number of tested models for each training cohort, no classifier showed consistently high accuracy (>80%) for prediction of never-smoker and smoker status in the test cohorts (Fig. 3B and Supplementary Fig. S2C).

For prediction of never-smoker versus current smoker status, results were slightly better, with classification accuracies of approximately 80% for several models (Fig. 3C and Supplementary Fig. S2D). Notably, for all tested models, a strong correlation existed between expression of proliferation-related genes (estimated by the CIN70 metagene; ref. 36) and correct classification of never-smokers/smokers and never-smokers/current smokers. Specifically, correctly identified never-smokers showed lower expression of proliferation-related genes, whereas correctly identified smokers/current smokers showed higher expression (Supplementary Fig. S2E and S2F). The improved performance in predicting never-smokers and current smokers specifically is consistent with an observed higher fraction of misclassified former smokers than current smokers across classifier models trained to predict never-smoker/smoker status (Supplementary Fig. S2G).
This suggests a higher degree of overlapping transcriptional patterns between never-smokers and former smokers and/or higher heterogeneity in former smokers compared with current smokers.

Excessive comparisons of different models in the same training and test sets may lead to bias and/or overinterpretation of the results. We therefore selected a single well-performing model (~80% balanced accuracy) for prediction of never-smoker/current smoker status and applied it to two novel cohorts, Tarca and colleagues (34) and Der and colleagues (33; Fig. 1B). In these cohorts, the selected model, a PAM classifier based on the Bosse and colleagues (20) gene signature trained in the Sheden and colleagues cohort, showed balanced accuracies of 73% in Tarca and colleagues and 80% in Der and colleagues for prediction of never-smoker/current smoker status, on par with the original results (Supplementary Fig. S2D). Again, the correctly identified never-smokers showed lower CIN70 expression, whereas correctly identified smokers/current smokers showed higher CIN70 expression in both cohorts (data not shown).

Unsupervised gene expression class discovery in adenocarcinoma identifies a subgroup of smokers aggregating with never-smokers

To further investigate why smoking-related CNAs and transcriptional patterns did not fully predict patient smoking status in supervised analysis, we performed unsupervised investigations of the genome-wide transcriptional pattern in adenocarcinoma, by individual consensus clustering (35) of the six discovery microarray cohorts (Fig. 1B). The aims of these analyses were to determine the impact of patient smoking status on the global transcriptional landscape, and the relationship of smoking status with transcriptional subgroups in lung adenocarcinoma. We found that never-smoker–enriched clusters (referred to as NS-enriched clusters, comprising ~70%–95% of all never-smokers) could be identified, however always including a notable fraction of smokers (~20%–60% of all smokers; Supplementary Fig. S3). Importantly, results were independent of the number of evaluated consensus clusters (n = 2, 3, 4), and there was high agreement in the grouping of cases for the different expression variation thresholds (Supplementary Fig. S3). Smokers in never-smoker–enriched clusters included both current smokers (~20%–60% of all current smokers) and former smokers (~40%–70% of all former smokers), with a reported lower number of smoking pack-years compared with non–never-smoker-enriched smokers (Supplementary Fig. S4A). In support of a shared expression pattern between never-smokers and smokers in the never-smoker–enriched clusters, we found only the noncoding X inactive specific transcript gene (XIST, chromosome Xq13.2) to be differentially expressed between these two groups in >2 microarray cohorts [Student t test FDR P < 0.05]. This result is presumably due to the higher frequency of females among the never-smokers.

Smokers and never-smokers aggregating together in consensus clusters share molecular and clinical characteristics

The relevance of the expression-based consensus clusters was supported by both molecular and clinical characteristics independent of patient smoking status. For instance, never-smokers and smokers in never-smoker–enriched clusters showed (i) significantly better overall survival, (ii) more differentiated tumors, (iii) less CNAs, (iv) strong enrichment of bronchioid-classified tumors (26), (v) lower expression of proliferation-related genes (the CIN70 metagene), (vi) TRU-like/alveolar-like/non–DASC-like tumor expression patterns (25, 28, 37), and (vii) higher expression of lineage-specific genes for alveolar/peripheral airway cells such as surfactant genes (SFTPB, SFTPC, CC10, GATA6, HOPX, and NKX2-1/TTF-1 (master transcription factor for the peripheral airways; refs. 25, 28, 37, 40) compared with respective non–never-smoker-enriched cases (Fig. 4A and B and Supplementary Fig. S4).

To validate the microarray-based results, we performed consensus clustering of 435 independent adenocarcinoma RNAseq profiles (423 with smoking history) from the TCGA project. Here, approximately 55% of all never-smokers aggregated with one third of all smokers in a single cluster (never-smoker–enriched; Supplementary Fig. S4C and S4D). Convincingly, we found similar molecular and clinical patterns within, as well as between, never-smoker–enriched and non–never-smoker-enriched cases, independent of smoking status (Fig. 4C and D and Supplementary Fig. S4D and S4E).

PCA analysis (38) performed in the TCGA RNAseq and Chitale and colleagues (32) gene expression microarray cohorts confirmed that clinical smoking status together with other clinicopathologic factors such as stage, gender, EGFR, and KRAS mutation status were not strong contributors to the total variation in gene expression compared with for instance reported adenocarcinoma subtypes (26; Supplementary Fig. S5). Together, the unsupervised gene expression analyses and the PCA analyses provide an explanation to why smoking-related classifiers do not reach 100% performance, i.e., there is a notable subgroup of smokers with similar transcriptional pattern to the majority of never-smokers.

Smokers aggregating with never-smokers based on transcriptional patterns show signs of less tobacco-related carcinogenesis on the DNA level

To further characterize the never-smoker–enriched and non–never-smoker-enriched consensus clusters, we took advantage of matched whole-exome DNA sequencing data available for the TCGA cohort. First, never-smokers and smokers in the never-smoker–enriched cluster showed less mutations overall compared with respective non–never-smoker-enriched cases (Fig. 4C and Supplementary Fig. S4F). Second, we observed significant or borderline non-significant differences in specific mutation transversions, especially a higher C>T and lower C>A transversion frequency (C>A transversion is a recognized smoking
Figure 4. Transcriptional patterns in adenocarcinomas stratified by smoking status. A, consensus clustering was performed in six adenocarcinoma cohorts (k = 3 clusters, expression SD > 0.5 as prefilter). For each cohort, the cluster with the highest number of never-smokers was identified (never-smoker-enriched cluster). Next, all cohorts were pooled to a meta-cohort (n = 841 samples). The heatmap shows mean z-score transformed values for different features (rows) for respective group (columns). The z-score transformation allows a common heatmap scale to be applied to all features. (Continued on the following page.)
Smoking and Genomic Alterations in Lung Adenocarcinoma

signature; refs. 4, 41) in never-smoker–enriched compared with non–never-smoker-enriched cases (Supplementary Fig. S4G). Combined mutation frequency for seven oncogenic driver mutations in lung cancer (EGFR, KRAS, ERBB2, BRAF, and gene fusions involving ALK, RET, and ROS1; similar to ref. 8) showed that the never-smoker–enriched cases (CCL2) together with one non–never-smoker-enriched cluster (CCL1) showed more alterations in these genes compared with the remaining non–never-smoker-enriched cluster (CCL3), irrespective of smoking status (Fig. 4C). However, for the two most frequently mutated oncogenic drivers in lung adenocarcinoma, KRAS and EGFR, the patterns were less distinct between consensus clusters. This seems consistent with results of a recent study investigating the impact of these mutations on the genomic (CNAs) and transcriptional landscape in adenocarcinoma (29).

Next, we identified 174 significantly mutated genes in lung adenocarcinoma by MutSigCV (42) analysis (402 analyzed cases, q value < 0.05), and screened these for association with the consensus clusters using a permutation-based approach (see Supplementary Methods). This analysis demonstrated association of the mutation pattern in four well-known tumor suppressor genes (TP53, STK11, KEAP1, and SMARCA4), as well as in ELTD1 and SNRPN, with the three consensus clusters (FDR < 10%; Fig. 4C, ELTD1 and SNRPN excluded due to lower mutation frequencies). Interestingly, mutation frequencies within consensus clusters were comparable between smokers and never-smokers for the four tumor suppressor genes, of which TP53, STK11, and KEAP1 mutations have been associated with smoking in lung cancer (4, 6, 43). Similar permutation-based mutation analysis between smokers and never-smokers in the TCGA cohort identified TP53 to be associated with smoking, but not the other five genes.

In summary, this unsupervised characterization of the global expression pattern in lung adenocarcinoma identifies a fraction of smokers with molecular and clinical features suggestive of less smoking-related carcinogenesis.

Discussion

In the current study, we have systematically explored genomic and transcriptional alterations in lung adenocarcinomas arising in never-smokers and smokers. We demonstrate that prediction of smoking history, based on CNAs and gene expression, is intrinsically difficult due to a heterogeneous pattern of alterations within and overlap between smoking subgroups. However, molecular stratification (based on transcriptional and clinicopathologic characteristics) of lung adenocarcinoma suggests that most tumors arising in never-smokers together with a specific subset of tumors from smokers form a more distinct and relevant molecular entity of less aggressive and potentially more smoking-unrelated disease.

Herein, we show that conflicting results from previous studies about smoking-related genomic and/or transcriptional alterations may be due to selection of different patient populations, tumor characteristics, and cohort sizes. Previous studies reporting overall more CNAs in smokers (14, 15), and more CNAs in heavy compared with light smokers (based on pack-years, a composite index of smoking intensity and duration; ref. 14), have included a notable fraction of squamous cell carcinomas (predominantly smokers). Importantly, squamous cell lung carcinoma has been shown to harbor overall more, as well as specific, CNAs compared with adenocarcinomas (30), which could influence these results. Besides tumor histology, other patient characteristics may also influence the pattern of smoking-related CNAs, such as ethnicity and EGFR mutation status (11, 12, 29, 44, 45). Different cohort characteristics could therefore be an important explanation for the observed differences in amount of CNAs in smoking-defined subgroups between individual cohorts in the current study (Fig. 2), but also between previous studies reporting contradicting results (12, 14, 15). In addition, the smoking group definitions themselves are a source of variation due to their self-reported nature and potentially different definitions between studies. Moreover, current definitions do not capture the intensity and duration of cigarette exposure, the exposure to environmental tobacco smoke and other pollutants for never-smokers, or the time of smoking cessation for former smokers. Here, the group of former smokers seems especially heterogeneous with (i) intermediate expression of genes separating never-smokers and current smokers in both tumor and normal airway tissues, (ii) less characteristic CNAs, (iii) higher cooccurring frequency with never-smokers in never-smoker–enriched clusters than current smokers, and accordingly lower prediction accuracies observed in both our and other studies (15, 16, 18, 20, 46; Supplementary Figs. S1, S2, S3, and S6). Importantly, our results stress the importance of a multicohort approach in determining consistent and robust smoking-related genomic and transcriptional alterations.

(Continued.) Red, higher expression/frequency values of a feature; blue, lower values. The heatmap shows the consistency between never-smokers (NS) and smokers (S) within non–never-smoker–enriched or never-smoker–enriched clusters, and the strong differences between never-smoker–enriched and non–never-smoker-enriched cases for the different features. P values computed using Wilcoxon or Fisher test, referring to group comparisons of all samples, never-smokers only, smokers only, rs, nonsignificant. B, overall survival (OS) (censored at 5 years) for smokers (top) and never-smokers (bottom) in never-smoker–enriched and non–never-smoker-enriched clusters from A. C, characterization of consensus clusters from analysis of 435 TCGA cases. Cluster 2 represents the never-smoker–enriched cluster (see also Supplementary Fig. S3). Heatmap shows mean z-score values for different features (rows) for respective group (columns) as in A. For mutations, the percentage of mutated cases in each group is shown. Oncogene drivers include mutations in EGFR, KRAS, ERBB2, BRAF, and gene fusions involving ALK, RET, and ROS1. Percentage of amplified cases refers to number of cases in each group with >1 high-level amplification in the focal CNA regions reported by Planck et al. (29). P values calculated using the Fisher or Kruskal–Wallis test similar to A, D, overall survival for all TCGA patients in consensus clusters (top) and smokers specifically (bottom) in clusters from C. For never-smokers, CCL2 patients showed borderline nonsignificant association with better overall survival (log-rank P = 0.07).
Our investigations of smoking-related CNAs identified only a few, variably sized, regions with moderate frequency differences (20%-30%; Supplementary Fig. S1, Table 2, and Supplementary Table S3). Corroborating previous smaller studies, regions with higher frequency of copy-number gain in never-smokers were found at 5q and 16p (9–12). Other reported smoking-related CNAs, e.g., gain of 7p (including EGFR) in never-smokers (10, 12), showed just below 20% frequency difference in the current study. However, the most reproducible smoking-related CNAs in the current study were higher frequencies of copy-number gain at 5q33.1-q35.3 and 16p13.3-p12.1 in never-smokers, and losses at 5q and 22q in current smokers when investigated in individual cohorts (Supplementary Fig. S1). Copy-number gain of 16p13.13 and 16p13.11 was recently reported as ethnic-specific events in east-Asian patients with adenocarcinoma (11). In our study, both regions had significantly higher frequency of copy-number gain in never-smokers in the total 1,398-sample cohort, as well as individually in the Chitale and colleagues (32), TCGA, and CLCGP cohorts. As the never-smoker/current smoker/former smoker groups in the TCGA cohort all consisted of 85% to 98% Caucasians (based on 358 annotated patients), our findings argue against 16p13.13 and 16p13.11 being only ethnic-specific events.

Our comprehensive supervised and unsupervised analyses together highlight an intrinsic heterogeneity within smoking-defined subgroups about CNAs and transcriptional alterations, but also considerable overlap between the clinically defined smoking groups. For instance, although we report smoking-related CNAs (with only moderate frequency differences), we acknowledge that the majority of the investigated genome was not significantly altered between never-smokers and smokers. Supported by the moderate performance of supervised genomic classification and the genomic PCA analysis (Fig. 3 and Supplementary Fig. S1G), this implies that the landscape of CNAs in lung adenocarcinoma is likely driven more by other patient and/or tumor-specific characteristics.

On the transcriptional level, a majority of never-smokers together with a specific fraction of smokers (both current smokers and former smokers) seem to display similar gene expression patterns, including expression of proliferation-associated genes (Fig. 4A and C; Supplementary Figs. S2 and S3). Cell proliferation generally has a strong impact on genome-wide expression patterns in tumors. Hence, the intrinsic heterogeneity in expression of proliferation-related genes within the smoking-defined subgroups could be a major reason for the consistent lack of success in identifying adenocarcinomas arising in never-smokers as a separate transcriptional entity with little or no inclusion of smokers by both supervised and unsupervised methods. Proliferation differences may also explain the better results in separating never-smokers (generally lower proliferation) from current smokers (generally highest proliferation) compared with former smokers (generally intermediate proliferation) in supervised classification (Supplementary Figs. S2 and S6). Moreover, cell proliferation provides a possible explanation to the association of never-smoking status with the bronchioid molecular subtype (26; lower proliferation; Supplementary Fig. S6). Our findings of prediction accuracies of approximately 80% for never-smokers and current smokers using gene expression classifiers are in agreement with reports from the literature based on analysis of both tumor and histologically normal airway tissues (4, 15, 46). For instance, Beane and colleagues (46) derived a 28-gene expression classifier with 80% accuracy in predicting smoking history in histologically normal airway epithelial cells. Imliminski and colleagues (4) reported a mutation-signature-based classifier with 79% balanced accuracy for prediction of never-smokers and smokers in adenocarcinoma tissue, whereas Massion and colleagues (15) reported 73% balanced accuracy for a CNA-based classifier. Although not an aim of the current study, we acknowledge that combining different measurements, e.g., whole-exome sequencing and gene expression data, may create a smoking status predictor with higher performance.

Although smoking increases the overall incidence of lung cancer, tumors unrelated to smoking can still occur in heavy smokers as smoking does not prevent the incidence of such cancers. Our gene expression analyses suggest that other factors than the actual smoking status, such as cell of origin, tumor microenvironment, mutation status of key oncogenic drivers, and overall genomic instability, may be more prominent in forming the genomic and transcriptional landscape in adenocarcinoma. Such factors may explain the intrinsic heterogeneity within smoking-defined subgroups, and the shared molecular features and carcinogenesis pathways between never-smokers and a fraction of lung cancers occurring in smokers (47).

Here, the two broad transcriptional subgroups of patients identified by unsupervised analysis in both the multicohort discovery set and the TCGA validation cohort in the current study are of interest: the subset of smokers that aggregates with the majority of never-smokers (smokers in never-smoker–enriched clusters), and the never-smokers aggregating in the more smoking-dominated clusters (non–never-smoker-enriched clusters). Smokers in the former group display clinical and molecular characteristics of a more smoking-unrelated tumorigenesis (e.g., less pack-years, less mutations, different mutation pattern; refs. 4, 6, 41), seem more genomically stable (less CNAs, mutations, and amplifications), and show transcriptional associations with the peripheral airways (6, 25, 28, 37, 40, 48). Whether these smokers have been long reformed and/or exposed to the same environmental or genetic factors that underlie lung cancer in never-smokers is unclear given the available patient annotations. Together, this group may represent a more differentiated and less aggressive road of tumor progression, less related to smoking and more dependent on the accumulation of further key oncogene mutations and/or rearrangements. In contrast, we show that tumors from never-smokers aggregating in the non–never-smoker-enriched/non-TRU/DASC-like/bronchial/non-bronchioid smoking-dominated clusters represent a more aggressive subset of smoking-unrelated disease.
with higher expression of proliferation-associated genes, higher genomic instability, less differentiated tumors, and poorer patient outcome. Whether these tumors arise more centrally in the lung or are a product of genomic instability caused by other factors, share carcinogenesis pathways with tobacco-related lung cancers, and respond differently to targeted treatment or adjuvant chemotherapy remain to be investigated. In addition, our findings of similar mutation frequencies of reported smoking-related tumor suppressors (TP53, STKI1, and KEAP1) between smokers and never-smokers within unsupervised clusters, while different between clusters, suggest a role for these genes also in smoking-unrelated disease. Here, forthcoming integrative analyses of mutational spectrum, CNAs, DNA methylation, and gene expression may further unravel the effect of smoking on the genomic and transcriptional landscape in the disease.

Together, our multicohort analyses illustrate the complex and heterogeneous landscape of genomic and transcriptional alterations between and within smoking-defined adenocarcinoma subgroups. On the basis of CNAs or gene expression patterns, adenocarcinomas arising in never-smokers do not seem to be readily resolved into a distinct molecular cluster, without notable inclusion of smokers. Instead, most tumors arising in never-smokers together with a specific subset of tumors from smokers seem to represent a more distinct and relevant molecular entity of less aggressive and potentially more smoking-unrelated disease. The possible predisposing factors or the extent of shared carcinogenesis pathways, and their relevance for, e.g., treatment response, in these lung cancer subgroups remain to be elucidated. We and others have recently shown that prognostic high-risk groups in non-small cell lung cancer (characterized by high expression of proliferation-associated genes) benefit more from adjuvant chemotherapy than less-proliferative low-risk cases (49, 50).

Irrespective, improved molecular characterization of lung adenocarcinoma may not only delineate the effect and impact of smoking on tumorigenesis, but is also clinically relevant. Molecular characterization could lead to identification of new targets for synergistic treatment, provide new insights into resistance mechanisms, and derive new predictors of treatment response and prognosis for the benefit of the patients.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: A. Karlsson, M. Plank, J. Staaf
Development of methodology: A. Karlsson, J. Staaf
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Boiling
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Karlsson, M. Ringné, M. Lauss, J. Staaf
Writing, review, and/or revision of the manuscript: A. Karlsson, M. Ringné, M. Lauss, J. Boiling, P. Micke, M. Plank, J. Staaf
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A. Karlsson, P. Micke

Study supervision: J. Staaf

Acknowledgments
The authors thank Dr. David Lindgren, Lund University, Sweden, and the editors at Elevate Scientific for fruitful discussions and comments on the article.

Grant Support
This study was financially supported by the Swedish Cancer Society, the Knut and Alice Wallenberg Foundation, the Foundation for Strategic Research through the Lund Centre for Translational Cancer Research (CREATE Health), the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Swedish Research Council, the Lund University Hospital Research Funds, the Gustav V:s Jubilee Foundation, and the Ingabritt and Arne Lundberg Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 29, 2014; revised May 15, 2014; accepted July 9, 2014; published OnlineFirst July 18, 2014.

References
Genomic and Transcriptional Alterations in Lung Adenocarcinoma in Relation to Smoking History

Anna Karlsson, Markus Ringnér, Martin Lauss, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-14-0246

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2014/07/19/1078-0432.CCR-14-0246.DC1

Cited articles
This article cites 50 articles, 17 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/20/18/4912.full.html#ref-list-1

Citing articles
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
/content/20/18/4912.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.