Mutational signatures associated with tobacco smoking in human cancer

Ludmil B. Alexandrov,1,2,3* Young Seok Ju,4 Kerstin Haase,5 Peter Van Loo,5,6 Ilkka Martinkocera,7 Serena Nik-Zainal,7,8 Yassyshi Totoki,9 Akihiro Fujimoto,10,11 Hidewaki Nakagawa,12,13 Tatsuhiko Shibata,9,12 Peter J. Campbell,7,13 Paolo Vineis,14,15 David H. Phillips,16 Michael R. Stratton16

Tobacco smoking increases the risk of at least 17 classes of human cancer. We analyzed somatic mutations and DNA methylations in 5243 cancers of types for which tobacco smoking confers an elevated risk. Smoking is associated with increased mutation burdens of multiple distinct mutational signatures, which contribute to different extents in different cancers. One of these signatures, mainly found in cancers derived from tissues directly exposed to tobacco smoke, is attributable to misreplication of DNA damage caused by tobacco carcinogens. Others likely reflect indirect activation of DNA editing by APOBEC cytidine deaminases and of an endogenous clocklike mutational process. Smoking is associated with limited differences in methylation. The results are consistent with the proposition that smoking increases cancer risk by increasing the somatic mutation load, although direct evidence for this mechanism is lacking in some smoking-related cancer types.

CANCER ETIOLOGY

The similarity extends to the presence of a transcriptional strand bias indicative of transcription-coupled nucleotide excision repair (NER) of bulky DNA adducts on guanine (fig. S1), the proposed mechanism of DNA damage by benzo[a]pyrene. Thus, signature 4 is likely the direct mutational consequence of misreplication of DNA damage induced by tobacco carcinogens.

Most lung and larynx cancers from smokers had signature 4 mutations. Signature 4 mutations occurred more often in cancers from

signatures 2, 4, 5, 13, and 16 [the mutational signature nomenclature is that used in the Catalogue of Somatic Mutations in Cancer (COSMIC) and in (16–18)]. There was sufficient statistical power to show that these increases were of clonal mutations (mutations present in all cells of each cancer) for signatures 4 and 5 (q < 0.05), as expected if these mutations are due to cigarette smoke exposure before neoplastic change (supplementary text).

Signature 4 is characterized mainly by C>A mutations with smaller contributions from other base substitution classes (Fig. 2B and fig. S1). This signature was found only in cancer types in which tobacco smoking increases risk and mainly in those derived from epithelia directly exposed to tobacco smoke (figs. S2 and S3). Signature 4 is very similar to the mutational signature induced in vitro by exposing cells to benzo[a]pyrene (cosine similarity = 0.94) (Fig. 2B and fig. S3), a tobacco smoke carcinogen (19). The similarity extends to

...
smokers compared with nonsmokers in all cancer types together (table S2) and in lung squamous, lung adenocarcinoma, and larynx cancers (table S2). This finding largely accounts for differences in total numbers of base substitutions (Table 1). In nonsmokers, 13.8% of lung cancers showed many signature 4 mutations (Fig. 2A; >1 mutation per megabase), which may be due to passive smoking, misreporting of smoking habits, or annotation errors. Signature 4 mutations were also detected in cancers of the oral cavity, pharynx, and esophagus, albeit in much smaller numbers than in lung and larynx cancers, perhaps because of reduced exposure to tobacco smoke or more efficient clearance. Differences in mutation burden attributed to signature 4 between smokers and nonsmokers were not observed in these cancer types (Fig. 1). Signature 4 mutations were found at low levels in cancers of the liver, an organ not directly exposed to tobacco smoke, and were elevated in smokers versus nonsmokers (Fig. 1).

Signature 4 was not extracted from bladder, cervical, kidney, or pancreatic cancers, despite the known risks conferred by smoking and the presence of many smokers in these series. Additionally, this mutational signature was not extracted from cancers of the stomach, colorectum, and ovary, nor from acute myeloid leukemia (in the analyzed series, the smoking status of patients with these cancers was unknown, but it is likely that many have been smokers). The tissues from which all of these cancer types are derived are not directly exposed to tobacco smoke. Simulations indicate that the lack of signature 4 is not due to statistical limitations (supplementary text and fig. S4). The absence of signature 4 suggests that misreplication of direct DNA damage due to tobacco smoke constituents does not contribute substantially to mutation burden in these cancers, even though DNA adducts indicative of tobacco-induced DNA damage are present in the tissues from which they arise (7).

Signatures 2 and 13 are characterized by C>T and C>G mutations, respectively, at TpC dinucleotides and have been attributed to overactive DNA editing by APOBEC deaminases (20, 21). The cause of the overactivity in most cancers has not been established, although APOBECs are implicated in the cellular response to the entrance of foreign DNA, retrotransposon movement, and local inflammation (22). Signatures 2 and 13 showed more mutations in smokers versus nonsmokers with lung adenocarcinoma (table S2). Because these signatures are found in many other cancer types, where they are apparently unrelated to tobacco smoking, it seems unlikely that the signature 2 and 13 mutations associated with smoking in lung adenocarcinoma are direct consequences of misreplication of DNA damage induced by tobacco smoke. More plausibly, the cellular machinery underlying signatures 2 and 13 is activated by tobacco smoke, perhaps as a result of inflammation arising from the deposition of particulate matter or by indirect consequences of DNA damage. Signature 5 is characterized by mutations distributed across all 96 subtypes of base substitution, with a predominance of T>C and C>T mutations (Fig. 2B) and evidence of transcriptional strand bias for T>C mutations (28). Signature 5 is found in all cancer types, including those unrelated to tobacco smoking, and in most cancer samples. It is “clocklike” in that the number of mutations attributable to this signature correlates with age at the time of diagnosis in many cancer types (27). Signature 5, together with signature 1, is thought to contribute to mutation accumulation in most normal somatic cells and in the germline (27, 29). The mechanisms underlying signature 5 are not well understood, although an enrichment of signature 5 mutations was found in bladder cancers harboring inactivating

### Table 1. Mutational signatures and cancer types associated with tobacco smoking

Information about the age-adjusted odds ratios for current male smokers to develop cancer is taken from (2–4). Odds ratios for small cell lung cancer, squamous cell lung cancer, and lung adenocarcinoma are for an average daily dose of more than 30 cigarettes. Odds ratios for cervical and ovarian cancers are for current female smokers. Detailed information about all mutation types, all mutational signatures, and DNA methylation is provided in table S2. Nomenclature for signature identification numbers is consistent with the COSMIC database (http://cancer.sanger.ac.uk/cosmic/signatures). The numbers of smokers and nonsmokers are unknown (i.e., not reported in the original studies) for acute myeloid leukemia, stomach, ovarian, and colorectal cancers. The patterns of all mutational signatures with elevated mutation burden in smokers are displayed in Fig. 2B. N/A denotes lack of smoking annotation for a given cancer type. Asterisks indicate that a signature correlates with pack years smoked in a cancer type. N.S. reflects cancer types without statistically significant elevation of mutational signatures. The odds ratio for all cancer types is not provided.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Odds ratio</th>
<th>Nonsmokers</th>
<th>Smokers</th>
<th>Total number of mutational signatures found in the cancer type</th>
<th>Signature 4 found in cancer type</th>
<th>Mutational signatures with elevated mutation burden in smokers versus nonsmokers (q &lt; 0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cancer types</td>
<td>ND</td>
<td>1062</td>
<td>2490</td>
<td>26</td>
<td>Y</td>
<td>4, 5*</td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>111.3</td>
<td>3</td>
<td>145</td>
<td>6</td>
<td>Y</td>
<td>N.S.</td>
</tr>
<tr>
<td>Lung squamous</td>
<td>103.5</td>
<td>7</td>
<td>168</td>
<td>8</td>
<td>Y</td>
<td>4*, 5</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>21.9</td>
<td>120</td>
<td>558</td>
<td>7</td>
<td>Y</td>
<td>2*, 4*, 5*, 13*</td>
</tr>
<tr>
<td>Larynx</td>
<td>13.2</td>
<td>6</td>
<td>117</td>
<td>5</td>
<td>Y</td>
<td>4*, 5</td>
</tr>
<tr>
<td>Pharynx</td>
<td>6.6</td>
<td>27</td>
<td>49</td>
<td>5</td>
<td>Y</td>
<td>5*</td>
</tr>
<tr>
<td>Oral cavity</td>
<td>4.2</td>
<td>98</td>
<td>265</td>
<td>5</td>
<td>Y</td>
<td>5*</td>
</tr>
<tr>
<td>Esophagus squamous</td>
<td>3.9</td>
<td>99</td>
<td>193</td>
<td>9</td>
<td>Y</td>
<td>5*</td>
</tr>
<tr>
<td>Esophagus adenocarcinoma</td>
<td>3.9</td>
<td>67</td>
<td>175</td>
<td>9</td>
<td>Y</td>
<td>N.S.</td>
</tr>
<tr>
<td>Bladder</td>
<td>3.8</td>
<td>111</td>
<td>288</td>
<td>5</td>
<td>N</td>
<td>5*</td>
</tr>
<tr>
<td>Liver</td>
<td>2.9</td>
<td>157</td>
<td>235</td>
<td>19</td>
<td>Y</td>
<td>4*, 5, 16</td>
</tr>
<tr>
<td>Stomach</td>
<td>2.1</td>
<td>472</td>
<td>13</td>
<td>N/A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Acute myeloid leukemia</td>
<td>2.0</td>
<td>202</td>
<td>2</td>
<td>N/A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>1.9</td>
<td>458</td>
<td>3</td>
<td>N/A</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Cervix</td>
<td>1.8</td>
<td>94</td>
<td>74</td>
<td>8</td>
<td>N</td>
<td>N.S.</td>
</tr>
<tr>
<td>Kidney</td>
<td>1.7</td>
<td>154</td>
<td>103</td>
<td>6</td>
<td>N</td>
<td>5*</td>
</tr>
<tr>
<td>Pancreas</td>
<td>1.6</td>
<td>119</td>
<td>120</td>
<td>11</td>
<td>N</td>
<td>N.S.</td>
</tr>
<tr>
<td>Colorectal</td>
<td>1.3</td>
<td>559</td>
<td>4</td>
<td>N/A</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>
mutations in ERCC2, which encodes a component of NER (24).

Signature 5 (or a similar signature that is difficult to differentiate from signature 5 because of the relatively flat profiles of these signatures) was increased by a factor of 1.3 to 5.1 ($q < 0.05$; table S2) in smokers versus nonsmokers in all cancer types together and in lung squamous, lung adenocarcinoma, larynx, pharynx, oral cavity, esophageal squamous, bladder, liver, and kidney cancers. The association of smoking with signature 5 mutations across these nine cancer types therefore includes some for which the risks conferred by smoking are modest and for which normal progenitor cells are not directly exposed to cigarette smoke (Table 1). Given the clocklike nature of signature 5 (17), its presence in the human germline (23), its ubiquity in cancer types unrelated to tobacco smoking (18), and its widespread occurrence in nonsmokers, it seems unlikely that signature 5 mutations associated with tobacco smoking are direct consequences of misreplication of DNA damaged by tobacco carcinogens. It is more plausible that smoking affects the machinery generating signatures 5 mutations (24). Presumably as a consequence of the effects of smoking, signature 5 mutations correlated with age at the time of diagnosis in nonsmokers ($P = 0.001$) but not in smokers ($P = 0.59$).

Signature 16 is predominantly characterized by T>C mutations at ApT dinucleotides (Fig. 2B); exhibits a strong transcriptional strand bias consistent with almost all damage occurring on adenine (fig. S5); and, thus far, has been detected only in liver cancer. The underlying mutational process is currently unknown. Signature 16 exhibited a higher mutation burden in smokers versus nonsmokers (table S2). For smokers with lung, larynx, pharynx, oral cavity, esophageal, bladder, liver, cervical, kidney, and pancreatic cancers, quantitative data on cumulative exposure to tobacco smoke were available (table S1). Total numbers of base substitution mutations were positively correlated with pack years smoked (1 pack year is defined as smoking one pack per day for 1 year) for all cancer types together ($q < 0.05$) and for lung adenocarcinoma (table S3). For individual mutational signatures, correlations with pack years smoked were found in multiple cancer types for signatures 4 and 5 (table S3). Signature 4 correlated with pack years in lung squamous, lung adenocarcinoma, larynx, and liver cancers. Signature 5 correlated with pack years in all cancers together, as well as in lung adenocarcinoma, pharynx, oral cavity, and...
bladder cancers (table S3). In lung adenocarcinoma, correlations with pack years smoked were also observed for signatures 2 and 13. The rates of these correlations allow estimation of the approximate numbers of mutations accumulated in a normal cell of each tissue due to smoking a pack of cigarettes a day for a year: lung, 150 mutations; larynx, 97; pharynx, 39; oral cavity 23; bladder, 18; liver, 6 (table S3).

Consistent with our results, previous studies have reported higher numbers of total base substitutions in lung adenocarcinoma in smokers versus nonsmokers (mainly due to C->A substitutions) (25, 26). The same is true of signatures 4 and 5 in lung adenocarcinoma (28), signature 4 in liver cancer (27), and signature 5 in bladder cancer (24).

Differential methylation of the DNA of normal cells of smokers compared to nonsmokers has been reported (28). Using data from methylation arrays, each containing ~470,000 of the ~28 million CpG sites in the human genome, we evaluated whether differences in methyla-

Fig. 2. Mutational signatures associated with tobacco smoking. (A) Each panel contains 25 randomly selected cancer genomes (represented by individual bars) from either smokers or nonsmokers in a given cancer type. The y axes reflect numbers of somatic mutations per megabase. Each bar is colored proportionately to the number of mutations per megabase attributed to the mutational signatures found in that sample. The naming of mutational signatures is consistent with previous reports (16–18). (B) Each panel contains the pattern of a mutational signature associated with tobacco smoking. Signatures are depicted using a 96-substitution classification defined by the substitution type and sequence context immediately 5' and 3' to the mutated base. Different colors are used to display the various types of substitutions. The percentages of mutations attributed to specific substitution types are on the y axes, whereas the x axes display different types of substitutions. Mutational signatures are depicted based on the trinucleotide frequency of the whole human genome. Signatures 2, 4, 5, 13, and 16 are extracted from cancer samples. The percentages of mutations attributed to specific substitution types are on the y axes, whereas the x axes display different types of substitutions. Mutational signatures are depicted based on the trinucleotide frequency of the whole human genome. Signatures 2, 4, 5, 13, and 16 are extracted from cancer samples.
Fig. 3. Differentially methylated individual CpGs in tobacco smokers across cancers associated with tobacco smoking. Each dot represents an individual CpG. The axes reflect differences in methylation between lifelong nonsmokers and smokers, where positive values correspond to hypermethylation and negative values to hypomethylation. The y axes depict levels of statistical significance. Results satisfying a Bonferroni threshold of $10^{-7}$ (above the red line) are considered statistically significant.

### Average Difference in Methylation (Smokers - Lifelong non-smokers)

- **Lung Adenocarcinoma**
  - 369 CpGs
- **Pharyngeal Cancer**
  - 65 CpGs
- **Oral Cancer**
  - 3 CpGs
- **Esophageal Adenocarcinoma**
  - none
- **Esophageal Squamous**
  - none
- **Papillary Renal Carcinoma**
  - none
- **Pancreatic Adenocarcinoma**
  - none
- **Bladder Cancer**
  - none
- **Cervical Cancer**
  - 3 CpGs

### REFERENCES AND NOTES


### ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust (grant 098850). S.N.-Z. is a Wellcome-Bet Prize Fellow and is supported through a Wellcome Trust Intermediate Fellowship (grant WT003383MA).
P.J.C. is personally funded through a Wellcome Trust Senior Clinical Research Fellowship (grant WT088340MA). M.R.S. is a paid advisor for GRAIL, a company developing technologies for sequencing of circulating tumor DNA for the purpose of early cancer detection. L.B.A. is personally supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy (DOE) National Nuclear Security Administration under contract no. DE-AC52-06NA25396. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the DOE. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (grant FC010201), the UK MRC (grant FC010202), and the Wellcome Trust (grant FC010203). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation’s support toward the establishment of The Francis Crick Institute. D.H.P. is funded by Cancer Research UK (grant C131/A4326), the Wellcome Trust (grants 101126/Z/13/Z and 101126/B/12/Z), the National Institute for Health Research (NIHR) Health Protection Research Unit in Health Impact of Environmental Hazards at King’s College London in partnership with PHE [the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health, or PHE], and by the project EXPOSIMICS (grant agreement 308610-FP7) (European Commission). P.V. was partially supported by the project EXPOSIMICS (grant agreement 308610-FP7) (European Commission). Y.T. and T.S. are supported by the Practical Research for Innovative Cancer Control from Japan Agency for Medical Research and Development (grant 15ck010604h0002) and National Cancer Center Research and Development Funds (15-A-5). We thank The Cancer Genome Atlas, the International Cancer Genome Consortium, and the authors of all studies cited in table S1 for providing free access to their somatic mutational data.

### SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/354/6312/618/suppl/DC1

Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S6
References (27–54)

2 May 2016; accepted 23 September 2016
10.1126/science.aag0299
Assessing smoke damage in cancer genomes

We have known for over 60 years that smoking tobacco is one of the most avoidable risk factors for cancer. Yet the detailed mechanisms by which tobacco smoke damages the genome and creates the mutations that ultimately cause cancer are still not fully understood. Alexandrov et al. examined mutational signatures and DNA methylation changes in over 5000 genome sequences from 17 different cancer types linked to smoking (see the Perspective by Pfeifer). They found a complex pattern of mutational signatures. Only cancers originating in tissues directly exposed to smoke showed a signature characteristic of the known tobacco carcinogen benzo[a]pyrene. One mysterious signature was shared by all smoking-associated cancers but is of unknown origin. Smoking had only a modest effect on DNA methylation.