Whole-Genome Sequencing of Asian Lung Cancers: Second-Hand Smoke Unlikely to Be Responsible for Higher Incidence of Lung Cancer among Asian Never-Smokers

Vidhya G. Krishnan, Philip J. Ebert, Jason C. Ting, Elaine Lim, Swee-Seong Wong, Audrey S.M. Teo, Yong G. Yue, Hui-Hoon Chua, Xiwen Ma, Gary S.L. Loh, Yuhao Lin, Joanna H.J. Tan, Kun Yu, Shenli Zhang, Christoph Reinhard, Daniel S.W. Tan, Brock A. Peters, Stephen E. Lincoln, Dennis G. Ballinger, Jason M. Laramie, Geoffrey B. Nilsen, Thomas D. Barber, Patrick Tan, Axel M. Hillmer, and Pauline C. Ng

Abstract
Asian nonsmoking populations have a higher incidence of lung cancer compared with their European counterparts. There is a long-standing hypothesis that the increase of lung cancer in Asian never-smokers is due to environmental factors such as second-hand smoke. We analyzed whole-genome sequencing of 30 Asian lung cancers. Unsupervised clustering of mutational signatures separated the patients into two categories of either all the never-smokers or all the smokers or ex-smokers. In addition, nearly one third of the ex-smokers and smokers classified with the never-smoker–like cluster. The somatic variant profiles of Asian lung cancers were similar to that of European origin with G>C>T=A being predominant in smokers. We found EGFR and TP53 to be the most frequently mutated genes with mutations in 50% and 27% of individuals, respectively. Among the 16 never-smokers, 69% had an EGFR mutation compared with 29% of 14 smokers/ex-smokers. Asian never-smokers had lung cancer signatures distinct from the smoker signature and their mutation profiles were similar to European never-smokers. The profiles of Asian and European smokers are also similar. Taken together, these results suggested that the same mutational mechanisms underlie the etiology for both ethnic groups. Thus, the high incidence of lung cancer in Asian never-smokers seems unlikely to be due to second-hand smoke or other carcinogens that cause oxidative DNA damage, implying that routine EGFR testing is warranted in the Asian population regardless of smoking status. Cancer Res; 74(21); 6071–81. ©2014 AACR.

Introduction
Lung cancer is one of the leading causes of cancer-associated deaths worldwide and non–small-cell lung carcinoma (NSCLC) accounts for nearly 85% of lung cancers (1). NSCLC accumulates somatically acquired variants (single base changes, insertions, deletions, substitutions, structural variations, and copy number variations). A subset of these somatic variants are called “drivers,” which are causally implicated in tumorigenesis (2). In the past, several studies on cancer genomics have revealed the relationship between mutational signatures, carcinogenic exposures, and DNA repair processes. For example, G>C>T=A transversions are predominant in lung cancers with tobacco exposure due to adduct formation by carcinogens in tobacco (3–6).

Recent applications of exome and whole-genome analysis of tumors (5–12) revealed multiple mutational patterns in lung cancer, melanoma, and others. To date, most of the next-generation sequencing reports on NSCLC mainly focused on tumors of European origin (7, 13–16), and there are little data on genomes from Asian patients with NSCLC (1, 17). The frequency of mutations in driver genes, such as EGFR, is different among ethnic groups, demonstrating that population differences exist (18). Therefore, it is important to study NSCLC in various human populations, including Asians.

The rate of never-smoker NSCLC in Asia is substantially higher than never-smoker NSCLC in the West. It is posited that the higher incidence of lung cancer in Asian never-smokers may be due to second-hand smoke, cooking fumes, or other Asia-specific environmental factors (11, 12, 19). If the higher
incidence of lung cancer in Asian never-smokers is due to second-hand smoke, then it is expected that the molecular signature of never-smokers would resemble that of smokers because many mainstream smoking carcinogens have been shown to also be present in second-hand smoke (20–24). If Asian never-smokers have increased lung cancer due to Asian-specific environmental factors, then their molecular signature may differ from that of European never-smokers. Therefore, we performed a comprehensive analysis of somatic variants in 30 Asian lung cancer tumors and explored how Asian lung cancer differs between smokers and never-smokers.

Patients and Methods

Patient samples and clinical information

Patients with NSCLC were staged according to the American Joint Committee on Cancer version 6 and selected on the basis of consecutive recruitment and known smoking status. Age, gender, ethnicity, histology, and tumor stage were collected for these 30 patients (Supplementary Table S1). Majority of the selected population had stage I and II NSCLC.

Tumors were collected at two centers (National University Hospital and Tan Tock Seng Hospital, Singapore, Singapore) between 2002 and 2006. All samples were surgical specimens and underwent pathologic review. Tissue samples were snap-frozen and stored at −80°C or stored in RNAlater (Ambion; Supplementary Table S1). Only samples with tumor cells more than 50% were selected for DNA extraction. DNA was extracted by the Blood and Cell Culture Kit (Qiagen) and by the RecoverEase DNA Isolation Kit (Stratagene) as indicated in Supplementary Table S1.

For each patient, two samples were taken: a tumor sample and a normal sample. The normal sample was taken either from tumor-free lung tissue from the same lobe (n = 26) or from blood (n = 4) as provided in Supplementary Table S1.

Informed consent was obtained from all individuals and the study was approved by the Institutional Review Board. Patients were interviewed about their smoking history 1 day before surgery. We asked patients how many cigarettes they smoked, and for how long. On the basis of this information, we calculated pack-years for each patient. In addition, we asked if the patient had stopped smoking, and if yes, how long had they stopped smoking. We defined never smokers as subjects who never smoked or smoked less than 100 cigarettes in their lifetime. Current smokers were defined as smokers. Ex-smokers were those who had smoked previously, but stopped smoking at least 2 months before the interview. Smoking history is included in Supplementary Table S1 to enable re-analysis using alternative smoking pattern definitions. Summary of clinical characteristics is provided in Table 1 and detailed clinical and smoking information, such as pack-years, is provided in Supplementary Table S1. Two patients (CTS21 and CT219) have incomplete smoking information that highlights the importance of collecting all clinical characteristics of patients.

Library construction, whole-genome sequencing, and variant calling

Complete Genomics, Inc. performed whole-genome sequencing (WGS) of DNA samples of 30 tumor/normal pairs using DNA nanoball and combinatorial probe anchor ligation technology (25). All samples were sequenced to 50× average coverage. Please see Supplementary Materials and Methods for more details on sequencing statistics, variant calls, and their annotation.

Validation of somatic single-nucleotide variants and indels

We tested a total of 331 "SQHIGH-filtered" somatic variants for which we could design primers and obtain PCR products. Of these, we were able to validate 279 (84%). Validation was

| Table 1. Clinical characteristics of 30 Asian lung cancer patients used for the analysis |
|---------------------------------------------------------------|----------------|----------------|
| Gender | All patients (n = 30) | Smokers (n = 14) | Never-smokers (n = 16) |
| Female | 10 | 0 | 10 |
| Male | 20 | 14 | 6 |
| Age at diagnosis | 65 (41–81) | 67 (41–81) | 62 (47–73) |
| Tumor stage | | | |
| I | 22 | 8 | 14 |
| II | 6 | 6 | 0 |
| III | 1 | 1 | 0 |
| IV | 1 | 1 | 0 |
| Histology | | | |
| Adenocarcinoma | 23 | 10 | 13 |
| Squamous cell carcinoma | 5 | 5 | 0 |
| Adeno squamous | 2 | 1 | 1 |
| Ethnicity | | | |
| Chinese | 27 | 14 | 13 |
| Vietnamese | 1 | 1 | 0 |
| Malay | 2 | 1 | 1 |
performed in two tiers on two different platforms. For tier 1, we selected four individuals, two with a high number of somatic single-nucleotide variants (SNV) and two with a low number of somatic SNVs. Of note, 194 somatic SNVs/indels were randomly selected across four individuals and sequenced by Ion Torrent technology (Life Technologies). PCR amplicons up to 250 bp in length were generated around the somatic variants and were sequenced by two Ion Torrent 316 chips (tumor- and normal-derived amplicons on separate chips) by 200-bp read length according to the manufacturer’s recommendations.

Sequence was analyzed by the Torrent Suite software (Life Technologies). Somatic variants were considered “tested” if they had >100 reads in the tumor and normal sample and were considered “validated” if the nonreference allele had a frequency of >1% in the tumor but not in the normal sample of the same patient. One hundred and fifty-eight somatic variants have been validated by this procedure (81%; Supplementary Table S6). In a second tier, a total of 137 somatic variants were tested by Sanger sequencing based on three categories: (i) recurrent mutations, (ii) mutations in WNT-related genes, and (iii) mutations in never-smokers (because this category was underrepresented in the first tier). PCR products were sequenced in both directions (forward and reverse) and sequence chromatograms were independently analyzed by two individuals. Somatic variants (121 of 137) have been validated in the second tier as somatic variants, resulting in a validation rate of 88% (Supplementary Table S6). When stratifying the validation rates based on smoking status, we had a validation rate of 93% for smokers (n = 200) and 78% for never-smokers (n = 131).

As EGFR is known to be highly mutated at a few key residues in Asian NSCLC, we used for EGFR less stringent somatic variant calling criteria from the WGS data and included low-quality somatic variants. These were either (i) non-SQHIGH or (ii) did not pass the Fisher exact test for allele-ratio differences or were within 5 bp of indels, or both. For EGFR, these were 18 coding mutations in 15 individuals. We attempted to validate all EGFR mutations by Sanger sequencing in all samples except for one individual where no DNA for validation work was available. We were able to validate all 17 tested EGFR mutations with one mutation (p.L858R in subject CTS177), which we also detected in the normal tissue (Supplementary Table S7). Because p.L858R is a frequently observed somatic mutation in EGFR, it is likely this is also a somatic mutation in this patient and that the presence of p.L858R in the normal tissue is due to a contamination by malignant cells.

Gene analysis

From the “Final functional variant list,” we identified variants that were previously observed in COSMIC (release 60). We also identified “recurrent” variants (two or more mutations at same chromosome and position in different samples). Recurrent mutations were further restricted by (i) eliminating variants that were likely germline based on searching master variant files from all normal samples (using five or more reads in any normal sample as a cutoff); (ii) eliminating variants that had equal to or less than twice the number of reads, supporting the variant in the tumor than in the normal (i.e., four reads in tumor/two reads in normal would be excluded). The list of variants previously observed in COSMIC was pooled with the "recurrent" variant list (as we felt these two classifiers were likely of the most importance) to form the "recurrent and COSMIC" variant list. In addition, we identified genes that contained at least three high-quality somatic variants in three separate samples, with no requirement for recurrence at the nucleotide level. These variants were subjected to the same criteria as above: (i) eliminating variants that were likely germline based on searching master variant files from all normal samples (using five or more reads in any normal sample as a cutoff); (ii) eliminating variants that had equal to or less than twice the number of reads, supporting the variant in the tumor than in the normal. These variants formed the "multi-hit" variant list. These groupings (COSMIC, recurrent, and multi-hit) were not mutually exclusive (e.g., EGFR populated both groups).

Details on structural variant analysis and strand bias of somatic variations can be found in Supplementary Materials and Methods.

Results

All the analyses presented here are based on defining never-smokers as subjects who never smoked or smoked less than 100 cigarettes in their lifetime. Ex-smokers were those who had smoked previously, but stopped smoking at least two months before the interview. Current smokers were defined as smokers. The average number of pack-years for the current and ex-smoker groups is 49 and 41 pack-years, respectively. Smoking details such as number of pack-years, cigarettes smoked per day, and duration can be found in Supplementary Table S1.

We analyzed the genomes of 30 tumor-normal pairs from patients with lung cancer sequenced by Complete Genomics, Inc. Of the 30 patients with lung cancer in our cohort, 7 (23%), 7 (23%), and 16 (53%) were former, current, and never-smokers, respectively. Sixty-six percent of the patients were men. All the never-smoker patients were stage I and the majority (88%) of the patients with smoking history were either stage I or II. Clinical characteristics of patients with lung cancer in our study are detailed in Table 1 and Supplementary Table S1.

Mutational signatures of Asian never-smokers suggest no evidence of second-hand smoke

It has been hypothesized that second-hand smoke, cooking fumes, or Asia-specific environmental factors are responsible for the higher incidence of lung cancer in Asian population (11, 12). If second-hand smoking causes mutational processes similar to mainstream smoking, the molecular signature of never-smokers would be predicted to resemble that of smokers. We therefore characterized the somatic SNVs in patients with lung cancer (Fig. 1). After unsupervised clustering on mutational signatures, our patients were categorized into two major groups based on smoking status (Fig. 1A). One cluster consisted of only smokers and ex-smokers while the other cluster consisted of all never-smokers and several ex-smokers and smokers. We defined the first group the smoker-only group because it consisted of six smokers, four ex-smokers, and zero never-smokers and the second group the never-smoker–like group.

\[\text{Published OnlineFirst September 4, 2014; DOI: 10.1158/0008-5472.CAN-13-3195}\]
The never-smoker–like group contained all of the never-smokers, as well as one smoker and three ex-smokers. Interestingly, the smoking dosage of smoker/ex-smokers in the never-smoker–like group (median, 36 pack-years or 262,800 cigarettes/lifetime; \(n = 4\)) was similar to that of smoker/ex-smokers in the smoker-only group (median, 32.5 pack-years or 228,125 cigarettes/lifetime; \(n = 10\); \(P = 0.5\) for pack-years and 0.7 for cigarettes/lifetime). This suggests the importance of considering the molecular signature irrespective of heavy smokers or light smokers or ex-smokers.

G.C > T.A transversions are known to be predominant in smokers, and result from the formation of polycyclic aromatic hydrocarbon adducts with deoxyguanosine (3–7, 13, 15). This substitution type dominated our smoker-only group (smoker-only group, 40%; never-smoker–like group, 16%; \(P = 3.1 \times 10^{-0.3}\), t test). We found the mutation signature pattern to be

Figure 1. A, unsupervised hierarchical clustering of Asian lung cancer genomes based on somatic substitutions. Patient ID along with clinical phenotype is shown on the right side of the heatmap. S25, CTS107, and CTS25 are patients with more than one EGFR mutation. B, percentage of somatic variants in Asian lung cancer genomes.
similar between coding and noncoding regions (Supplementary Fig. S2). In contrast, G>C>A.T transition dominated the never-smoker–like group. The proportion of G>C>A.Ts among the never-smoker–like group was higher relative to G>C>A. Ts in smokers (smoker-only group, 19%; never-smoker–like group, 34%; \( P = 1.2 \times 10^{-9} \), t test; Fig. 1B). Thus the never-smoker–like group’s mutation signature differed from the smoker-only group’s mutation signature. The one smoker and three ex-smokers who clustered with the never-smokers had the same molecular signature as a never-smoker, despite self-reporting exposure to tobacco (Fig. 1A).

The higher frequency of G>C>A.T transition in never-smokers versus smokers and the higher frequency of G>C>T.A transversions in smokers versus never-smokers was observed in another study based on RNA sequencing of Korean patients with lung adenocarcinoma (Supplementary Fig. S1; ref. 26). Thus, the mutation signatures were confirmed in another study. The range for Koreans is much larger than that seen in this study. One possible explanation is the number of somatic mutations detected by RNA sequencing in the Korean study is small (~25 per sample). Because of the small values, the percentages for the Korean population are more susceptible to noise, and hence a large range is observed. The compact range detected by WGS in this study is probably due to more accurate numbers as we observed >1,000 somatic mutations per sample. This could be an advantage of sequencing the whole genome when mutation numbers are low.

The two major clusters of smoker-only and never-smoker–like groups hold even after taking into account the bases surrounding the mutation (Fig. 2). The most frequent mutation type for the never-smoker–like group is a C>T transition where the C is flanked 3’ by a G, that is, XpCpG (Fig. 2). This sequence context has been reported in other cancers, such as melanoma, breast cancer, and is thought to be due to deamination of methylated cytosines (5, 6). Mutation at CpG sites was not as frequent in the smoker-only group.

We observed a significantly higher mutation load (somatic point mutations) in the smoker-only group (median, 18,794; range, 2,452–79,859) compared with the never-smoker–like group (median, 4,139; range, 764–27,037; \( P = 0.001 \), Wilcoxon rank-sum test; Fig. 3A). The higher mutational load in smokers is expected because of the mutagenic properties of tobacco smoke (7, 13). Interestingly, the ex-smokers/smoker with a never-smoker–like molecular signature had a mutational load not significantly different from the never-smokers (\( P = 0.3 \), Wilcoxon rank-sum test).

We validated the smoking-related differences for mutation load in an independent cohort using the Imielinski and colleagues data (14). We found that smokers in their never-smoker–like group had a lower mutation load compared with smokers in the smoker-only group (average mutation load: smokers with smoker-only signature, 3,305; smokers with never-smoker–like signature, 1,248; never-smokers, 465). This is similar to what we observe in this study. Also, the number of somatic SNVs observed in our population is corroborated by a published report using Complete Genomics, Inc. sequencing of two never-smoker genomes with 1,802 and 1,169 mutations.

Figure 2. Mutation spectrum of trinucleotides of Asian lung cancers. Heatmap representation of flanking bases (±1 base) of single base substitutions for all patients with lung cancer. Patient ID along with clinical phenotype is shown on the right side of the heatmap. S25, CTS107, and CTS25 are patients with more than one EGFR mutations.
each and one smoker genome with approximately 23,000 mutations (15). A second study reported roughly similar counts for never-smokers (median, 888; range, 842–1,268; ref. 13). However, we cannot directly compare with this study due to difference in sequencing technologies and bioinformatics filters.

Our study is consistent with previous studies but there are a few individuals in our study that appear to be outliers. An ex-smoker with a never-smoker–like signature had an abnormally high number of somatic SNVs, compared with the other three in his group (CTS153 with 27,037 SNVs). This patient’s age, tumor stage, and smoking dosage were similar to that of other

![Figure 3. A, somatic mutation load in Asian lung cancer samples based on molecular signature. S25, CTS107, and CTS25 are patients with more than one EGFR mutations. B, number of somatic point mutations observed in different lung cancer stages of patients.](https://cancerres.aacrjournals.org/doi/10.1158/0008-5472.CAN-13-3195)
patients in the group (Supplementary Table S1). This patient had a missense and nonsense mutation in the DNA repair genes RAD51 and RIF1, respectively, which might have contributed to the high mutation rate. Conversely, while the majority of the individuals in the smoker-only group had a lower number of mutations (>10,000 somatic SNVs), CTS181 and S27 had low mutation loads with 2,452 and 3,903 mutations, respectively. Age and smoking dosage for these patients were similar to that of other patients in the group (Supplementary Table S1). Such outliers might be explained by different mechanistic mutation subcategories that are active to variable degrees in the different patients, the sampling of atypical tumor sections for genome analysis or patient-specific predispositions for better or worse DNA repair.

Because increasing age and tumor stage could correlate with larger numbers of mutations, we tested for such relationships in our data. We could not find a correlation between age and number of mutations ($r = -0.1$) and a moderately negative correlation between tumor stage and number of mutations ($r = -0.4$; Fig. 3B). Tumors in stage IIB showed a wide range of somatic point mutations. Tumors classified as stage II are known to be heterogeneous (27). The clinical heterogeneity might underlie the wide range of mutations in stage IIB.

Carcinogens in tobacco smoke induce oxidative DNA damage that gives rise to G.C>T.A transversions. This DNA damage can be removed by the transcription-coupled repair process (28, 29). There are two lines of evidence supporting the hypothesis that transcription-coupled repair occurs, to a greater extent, in the smoker-only group than in the never-smoker–like group. First, smokers in the smoker-only group have fewer somatic SNVs on the transcribed strand compared with the nontranscribed strand (Fig. 4). The median number of somatic SNVs on the transcribed strand was 2,432, compared with 3,394 on the nontranscribed strand, which represents an average 39% reduction ($P < 0.01$, paired $t$ test). This is consistent with transcription-coupled repair actively occurring in the smoker-only group and reducing the number of somatic substitutions in the genic regions. In contrast, for never-smokers, the difference between the transcribed and nontranscribed strand was much smaller. For never-smokers, the median number of somatic SNVs on the transcribed strand was 640, compared with 686 on the nontranscribed strand (7% reduction, $P < 0.01$, paired $t$ test; Fig. 4). In this case, the ex-smokers/smokers in the never-smoker–like group were similar to the never-smokers.

The second piece of evidence is that smokers in the smoker-only group had a lower proportion of somatic substitutions in genic regions compared with never-smokers (smokers, 34%; never-smokers, 40%; $P = 3.7 \times 10^{-8}$, $t$ test; Fig. 5). Thus, transcription-coupled repair occurred to a much larger extent in smokers in the smoker-only group that have a smoking molecular signature, than it did in never-smokers and ex-smokers/smokers that classify with never-smokers.

In conclusion, Asian never-smokers have a distinct mutation signature from smokers. However, a subgroup of patients exposed to smoke had a mutation signature that appeared similar to never-smokers.

**Candidate driver genes in Asian lung cancers**

We compared the frequencies of mutations in the known NSCLC driver genes EGFR, KRAS, and TP53 with previous reports. Fifty percent ($n = 15$) of our patients harbored EGFR mutations (including variants of low initial quality,
see below) of which 73% (n = 11) were never-smokers, 63% of which (n = 7) were women (Fig. 3). This is in agreement with a previous study (30). Mutations in EGFR primarily occurred in two locations, encoding the L858R variant (11 samples) and AA746-750 deletion (four samples), which are common activating mutations (Table 2). Seventy-three percent of these activating mutations occurred in never-smokers, which is consistent with previously published reports of enrichment of EGFR mutations in never-smokers in Asian NSCLC (31). Apart from these previously observed mutations, we also noticed a novel mutation, R889G, in EGFR. TP53 mutations in lung cancer have been reported to be more prevalent in squamous cell carcinoma relative to adenocarcinoma (Cosmic v62; refs. 7, 32). However, while the TP53 mutation frequency in patients in the adenocarcinoma cohort of our dataset is largely in agreement with these data (30%; 7 of 23), the mutation frequency in TP53 in our squamous cell carcinoma cohort is significantly less (20%; 1 of 5), albeit with a limited representation of this tumor type (two samples with mixed adenosquamous histology were excluded in the comparison of adenocarcinoma vs. squamous, neither of which have TP53 mutations). We observed a KRAS mutation in one Asian smoker lung cancer patient (1 of 30; 3%, G12D). This KRAS mutation rate is lower than what has been described for European populations (>20%; ref. 33) but comparable with what has been reported for Asian populations (3.8%; refs. 34, 35), underscoring the importance of investigating population-specific molecular features in lung cancer.

To identify genes with potential driver mutations, we classified 2,246 somatic variants ("Final functional variant list"); Materials and Methods) which alter coding sequence using three metrics: (i) presence of the same variant in the COSMIC database (COSMIC); (ii) recurrence of the same variant within the dataset (two or more separate tumors but not in any normal sample; "recurrent"); or (iii) multiple different variants (three or more) within a gene ("multi-hit"). For the first category, we identified 41 variants in 16 genes that exist in the COSMIC database (Table 2 and Supplementary Table S2). For the second category, we observed 22 recurrent variants across six genes (Table 2 and Supplementary Table S2). (Note that the Cosmic and recurrent categories are not mutually exclusive.) For the third multiple-hit

### Table 2. Genes with recurrent variants or variants previously observed in COSMIC

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Mutation count</th>
<th>Mutation (COSMIC in bold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
<td>18 (5/13)</td>
<td>E709A,L833V,(11)L858R,(4)E746_A750(del), R889G</td>
</tr>
<tr>
<td>SATB2</td>
<td>SATB homeobox 2</td>
<td>4 (1/3)</td>
<td>E3422 ,T4061M,D4374Y</td>
</tr>
<tr>
<td>RYR2</td>
<td>Ryanodine receptor 2</td>
<td>3 (2/1)</td>
<td>(2)E714D,R374K,D477Y</td>
</tr>
<tr>
<td>C1orf88</td>
<td>Chromosome 1 open reading frame 88</td>
<td>2 (0/2)</td>
<td>(2)F3L</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin-dependent kinase inhibitor 2A</td>
<td>2 (2/0)</td>
<td>E69fs,77,D84N</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>Catenin β1, 88 kDa</td>
<td>2 (1/1)</td>
<td>D32N,S37F</td>
</tr>
<tr>
<td>FERMT1</td>
<td>Fermitin family member 1</td>
<td>2 (1/1)</td>
<td>(2)E330V</td>
</tr>
<tr>
<td>HOXB1</td>
<td>Homeobox B1</td>
<td>2 (2/0)</td>
<td>P77L,G281V</td>
</tr>
<tr>
<td>KLC3</td>
<td>Kinesin light chain 3</td>
<td>2 (1/1)</td>
<td>(2)R170G</td>
</tr>
<tr>
<td>PLOD1</td>
<td>Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCN5A</td>
<td>Sodium channel, voltage-gated, type V, α subunit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMA3D</td>
<td>Semaphorin 3D</td>
<td>2 (2/0)</td>
<td>H532L,R180K</td>
</tr>
<tr>
<td>ASB15</td>
<td>Ankyrin repeat and SOCS box containing 15</td>
<td>1 (1/0)</td>
<td>E9K</td>
</tr>
<tr>
<td>DSCR6</td>
<td>Down syndrome critical region gene 6</td>
<td>1 (1/0)</td>
<td>E50</td>
</tr>
<tr>
<td>ERBB3</td>
<td>v-erb-b2 erythroleukemia viral oncogene homolog 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRAF</td>
<td>v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPRE3</td>
<td>Microtubule-associated protein, RP/Eβ family, member 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLXNA4</td>
<td>Plexin A4</td>
<td>1 (1/0)</td>
<td>W1652</td>
</tr>
<tr>
<td>TIGD2</td>
<td>Tigger transposable element derived 2</td>
<td>1 (1/0)</td>
<td>L186V</td>
</tr>
</tbody>
</table>

NOTE: Genes with a nucleotide-level recurrent SQHIGH variant (two or more) or that had an SQHIGH variant in the same chromosome and position as a variant observed in COSMIC were rescanned for additional variants. Mutation count includes all nonsynonymous variants observed, segregated by smokers and ex-smokers/never-smokers; some patients have more than one mutation per gene for EGFR (3) and TP53 (1). Mutation lists the amino acid changes observed in the gene, with those in bold previously observed in COSMIC and recurrent variants preceded by the variant count in parenthesis. Also, "ins" represents an amino acid insertion with the corresponding residues, whereas "del" represents a deletion with the corresponding residues. Additional supporting data can be found in Supplementary Table S2.
list category, we find 179 variants in 47 genes (Supplementary Table S3).

We searched the "final functional variant list" for additional variants in genes included in the combined recurrent/COSMIC list. This approach identified an additional seven variants in six existing genes in the recurrent/COSMIC list. *EGFR* plays a prominent role in Asian NSCLC, we further examined all predicted *EGFR* variants regardless of variant quality status. This approach identified three additional *EGFR* mutations (two L858R; one A746-750 deletion), which were experimentally validated. One of the additional activating *EGFR* mutations was found in an ex-smoker with the never-smoker–like phenotype. The other two *EGFR* mutations were observed in never smokers. Thus all of the additional *EGFR* mutations were found in the never-smoker–like group.

The majority of our gene list (recurrent, COSMIC, and multi-hit variant lists) overlapped with existing NSCLC publications (7, 14, 15, 26, 35). Fifty-three genes are seen in at least one other study, and 33 genes are seen in two or more studies. *TP53, KRAS, EGFR, CTNNB1,* and *RYRI* had identical amino acid changes with the other studies. However, we found *DSCR6, HOXB1, KL3C, MAPR3, TIGD2,* and *C10ORF88* (in our COSMIC/recurrent list) and *DILGAP2, DUSP27, PLEC1,* and *SLC27A3* (in our multi-hit variants list) to be unique to our cohort. These genes are possibly specific to Asian NSCLC. It is unlikely that these genes explain the phenotype of smokers with never-smoker–like signature because none of the mutations in the genes noted above was found in the smokers classified in the never-smoker–like group.

The axonal pathfinding/WNT signaling pathway is implicated in various cancers (36, 37). The multi-hit variant list revealed mutations in *NTRK3* and *SLITRK1* as well as a putative translocation involving *NTRK2* (observed in structural variant analysis; Supplementary Table S4). Also, mutations in *CTNNB1, PLXNA4,* and *SEMA3D* were observed in the recurrent/COSMIC list (see Table 2 and Supplementary Table S2). As the products of these genes are implicated in axonal pathfinding/WNT signaling pathways and some have previously been implicated in cancer, we further analyzed the "final functional variant list" for genes whose products have a similar function. We found somatic mutations in *CCND1, CTNNN2, DVL3, EFN4A, EPUB1/A6, EPUB1, HGF, NTRK1, PLXNA1, RELN, ROBO2/4, SEMA3D/3G/5A/6C, SLITRK3, SLIT2/3, UNC5C,* and *WNT8A,* many of which have been linked to WNT signaling. Analysis of the mutations at the pathway level, however, did not confirm significant enrichment of the WNT pathway mutations in our dataset (*P > 0.05*).

Analysis of structural variants predicted numerous events per tumor, including interchromosomal translocations, deletions, insertions, and duplications. Among the 30 subjects, we predict 99 events that involve a gene at both the 5′ and 3′ side of the junction (Supplementary Table S4). This includes genes previously observed in translocation events in cancer (*ETV5, ETV6, MIL11, CAMTA1, LIRF,* and *TCF12*) as well as structural variants in genes of interest: (i) the p38/MAPK pathway member *RIT1,* (ii) the netrin receptor *UNC5D* (related to oncogene *DCC,* (iii) receptor tyrosine kinase *MSR1,* (iv) *GDNF* (a ligand of the oncogene *RET,* and (v) *NRG3* (a ligand of the oncogene *ERBB4*). PCR-based analysis of a subset of the structural variants confirmed six of nine events surveyed (Supplementary Table S4). The vast majority of these translocation events occurred in intronic regions and would thus have been missed by an exome-only approach, highlighting the added value of the whole-genome approach relative to exome sequencing. Further analysis of noncoding regions was beyond the scope of this study. We have made the genomic data available to the community so researchers have the opportunity to explore noncoding regions further.

**Discussion**

Asian never-smokers are currently underrepresented among WGS studies of lung cancer. Therefore, a major component of our study was to compare mutation profiles between Asian smokers and never-smokers. We analyzed 30 Asian NSCLC lung tumors with both smoking and never-smoking history.

Asians have higher lung cancer–related death rates and Asian never-smoking women have a higher incidence of lung cancer compared with European counterparts (19). It has been speculated that the high rate of lung cancer in Asian never-smokers is due to environmental factors such as second-hand smoke or cooking style (28, 29, 38). Exposure to these carcinogens would lead to increased oxidative damage and an increase in the G>T transversion mutation rate (28). We did not survey the degree of second-hand smoke exposure in our never-smoking patients. This raises the possibility secondary tobacco smoke could be a confounding factor in these patients. However, we did not observe a smoker-like mutation signature in any of our never-smoker patients, suggesting that this confounder might not be significant for our conclusions.

Our study demonstrates that the mutation signature of Asian never-smokers is distinct from that seen in smokers and therefore, it is unlikely that tobacco-related environmental signals are responsible for the increased incidence rate in NSCLC in Asian never-smokers compared with European populations. First, the molecular signature of Asian never-smokers resembles that of the never-smoking signature observed in the West, and the signature of Asian smokers resembles that of European smokers. Second, the mutational load in Asian never-smokers is lower than in Asian smokers, which is what has been observed in Europeans. Third, a stronger signature of transcription-coupled repair, which is initiated in response to oxidative damage, such as tobacco or cooking (28, 29, 38), is observed in Asian smokers compared with Asian never-smokers. Specifically, somatic variations in Asian smokers are considerably lower in genic regions and on transcribed strand compared with intergenic regions and nontranscribed strand, respectively, and this phenomenon is not true for Asian never-smokers. Because carcinogens for mainstream tobacco smoking have also been detected under second-hand smoking conditions (20–24), the elevated rate of NSCLC in Asian never-smokers might not be due to second-hand smoke. Instead, intrinsic factors, gene–environment interactions, or epigenetic aspects might be responsible for the epidemiologic differences in NSCLC frequencies across populations.
The $G>C\rightarrow A\rightarrow T$ transition that is most frequent among never-smokers (13) is also observed in other cancers such as melanoma, lung, and breast cancers (5, 6, 9), suggesting that the mutational mechanism is not cancer type specific and/or that different mechanisms can result in the same transition. The mutation signature of smokers is expected to be diluted when smokers are not stratified into those with or without the smoker signature and the analysis of only expressed coding regions provides less statistical power to define mutational signatures for each sample. Despite these limitations for a direct comparison, the RNA-sequencing data of Soe and colleagues are in agreement with our study. Hence, the described signature might be considered a general smoker/never-smoker phenomenon.

Although patients self-reported as smokers or ex-smokers, the molecular signature of a subgroup of patients with smoking history resembles that of never-smokers, despite having cigarette consumption similar to the smokers with the smoker-only signature. What could be responsible for these patients being classified with never-smokers? Driver mutations in the $EGFR$ oncogene have been reported in never-smokers and are more frequent in the Asian population (18, 30). The majority of $EGFR$ mutations ($n=15$) was found among the tumors with the never-smoker–like signature ($n=11$). However, 50% (2 of 4) of ex-smokers/smokers with never-smoker–like signature had $EGFR$ mutations and, in contrast, only 20% (2 of 10) of the smokers in the smoker-only group had $EGFR$ mutations. Thus, a higher fraction of the ex-smokers/smokers in the never-smoker–like group carry a driver mutation in $EGFR$. This raises the possibility that people who smoke or have a smoking history may have a driver mutation, and regardless of smoking status, have the never-smoker–like signature. It is possible that the never-smoker–like signature reflects an oncogene-driven mutation mechanism in which mutated $EGFR$ or other oncogenes drive the cancer. In contrast, the smoker signature is due to mutations caused by tobacco exposure. On the basis of this model, the presence of oncoplastic drivers such as $EGFR$ mutations in smokers could result in a dominant never-smoker–like mutation profile, which can be stronger than the smoking signature.

One would expect long-term quitters to have a mutational pattern similar to that of never-smokers, and short-term quitters to resemble that of current smokers. It is known that 5 to 9 years of smoking cessation can lower the risk for lung cancer (39). Interestingly, the one smoker and three ex-smokers with the never-smoker–like signature had similar smoking dosage as that of the smoker-only group, and two of the ex-smokers quitted smoking only 3 years ago (Supplementary Table S1). Therefore, the never-smoker–like signature in these patients cannot be accounted for by quitting smoking a long time ago. This highlights the importance of checking the molecular signature of patients with lung cancer irrespective of smoking status.

Physicians in both Western and Asian hospitals tend to order $EGFR$ first-line testing for never-smoking patients with NSCLC (40, 41). In Singapore, all patients are tested for $EGFR$ mutations regardless of smoking status. Our results suggest $EGFR$ testing could be useful in the case of patients with lung cancer regardless of smoking status, especially in Asian populations, as 50% of smokers in our cohort whose mutation signature resembled that of never-smokers harbored $EGFR$ mutations. This is supported by the observation that patients harboring $EGFR$ mutations have similar clinical outcomes to $EGFR$ tyrosine kinase inhibitors, regardless of smoking status (42).

Limitations of this study are small sample size and the lack of an independent validation cohort. As discussed earlier, our findings are supported by other recent sequencing studies (14, 26), but additional sequencing will be needed before our findings can be generalized. In conclusion, we show that NSCLC in Asian never-smokers is unlikely due to tobacco exposure and other oxidative damaging agents, and molecular signature may provide additional information beyond clinical phenotype for a better understanding of the underlying etiology. Future research may show that the genomic signature could be a better classifier for lung cancer than actual smoking status.

Disclosure of Potential Conflicts of Interest

Y.G. Yue is a Principal Research Scientist at Eli Lilly and Company, R.A. Peters has ownership interest (including patents) in Complete Genomics, Inc. S.E. Lincoln is a current employee of Invitae, however, at the time of the study he was an employee of Complete Genomics, Inc. and has ownership interest (including patents) in Invitae. G.B. Nilsen has ownership interest (including patents) in Complete Genomics, Inc. P. Tan received a commercial research grant from Eli Lilly and Company. P.C. Ng has ownership interest (including patents) in Illumina. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: V.G. Krishnan, C. Reinhardt, T.D. Barber, P. Tan, A.M. Hillmer

Development of methodology: V.G. Krishnan, J.C. Ting, Y.G. Yue, T.D. Barber

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): E. Lim, A.S.M. Teo, Y. Lin, S. Zhang, D.S.W. Tan, R.A. Peters, S.E. Lincoln, P. Tan


Writing, review, and/or revision of the manuscript: V.G. Krishnan, P.J. Ebert, J.C. Ting, E. Lim, S.-S. Wong, C. Reinhardt, D.S.W. Tan, D.G. Ballinger, J.M. Laramie, T.D. Barber, P. Tan, A.M. Hillmer, P.C. Ng

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): P.J. Ebert, A.S.M. Teo, H.-H. Chua, G.S.L. Loh, Y. Lin, J.M. Laramie, T.D. Barber

Study supervision: D.G. Ballinger, T.D. Barber, A.M. Hillmer, P.C. Ng

Acknowledgments

The authors thank Dawn Choi Poh Sum and Leong See Ting for performing Ion Torrent sequencing and Yao Fei for help on structural variation validation work, and John Calley, Marcio Chedid, and Adam West for helpful conversations. The authors also thank Sebastian Ribi for his help with image formatting.

Grant Support

This work was supported by Agency for Science Technology and Research (A*STAR) and the National Medical Research Council [Translational & Clinical Research Flagship Program “Non-Small Cell Lung Cancer: Targeting Cancer Stem Cell and Drug Resistance” (TC1012016)], Singapore. 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 November 11, 2013; revised July 30, 2014; accepted August 6, 2014; published OnlineFirst September 4, 2014.
References


Cancer Research

Whole-Genome Sequencing of Asian Lung Cancers: Second-Hand Smoke Unlikely to Be Responsible for Higher Incidence of Lung Cancer among Asian Never-Smokers

Vidhya G. Krishnan, Philip J. Ebert, Jason C. Ting, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-13-3195

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2014/09/05/0008-5472.CAN-13-3195.DC1

Cited articles
This article cites 41 articles, 11 of which you can access for free at:
http://cancerres.aacrjournals.org/content/74/21/6071.full.html#ref-list-1

Citing articles
This article has been cited by 7 HighWire-hosted articles. Access the articles at:
/content/74/21/6071.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.