Dearth of smoking-induced mutations in oncogene-driven non-small-cell lung cancer despite smoking exposure

Chen-Yang Huang[1,2,3], Nanhai Jiang[1,2], Meixin Shen[4], Gillianne Lai[4], Aaron C. Tan[4], Amit Jain[4], Stephanie P. Saw[4], Mei-Kim Ang[4], Quan Sing Ng[4], Darren Wan-Teck Lim[4,5], Ravindran Kanesvaran[4], Eng-Huat Tan[4], Wan Ling Tan[4], Boon-Hean Ong[6], Kevin L. Chua[7], Devanand Anantham[8], Angela Takano[9], Tony K.H. Lim[9], Wai Leong Tam[10,11,12], Ngak Leng Sim[10], Anders J. Skanderup[10*], Daniel S.W. Tan[4,10,13,14*], Steven G. Rozen[1,2,15*]

Affiliations

[1]Centre for Computational Biology, Duke-NUS Medical School, Singapore, 169857, Singapore

[2]Programme in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore, 169857, Singapore

[3]Division of Hematology-Oncology, Department of Internal Medicine, Linkou Chang Gung Memorial Hospital and Chang Gung University, Taoyuan 333, Taiwan

[4]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, 169610, Singapore

[5]Institute of Molecular and Cell Biology, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore

[6]Department of Cardiothoracic Surgery, National Heart Centre Singapore, Singapore 169609, Singapore

[7]Division of Radiation Oncology, National Cancer Centre Singapore, Singapore 168583, Singapore

[8]Department of Respiratory and Critical Care Medicine, Singapore General Hospital, Singapore

[9]Department of Pathology, Singapore General Hospital, Singapore, 169608, Singapore

[10]Genome Institute of Singapore, Singapore, 138672, Singapore

[11]Cancer Science Institute of Singapore, National University of Singapore, Singapore, 117599, Singapore

[12]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 117597, Singapore

[13]Duke-NUS Medical School Singapore, Singapore, 169857, Singapore

[14]Cancer Therapeutics Research Laboratory, Division of Medical Sciences, National Cancer Center Singapore, Singapore, 169610, Singapore

[15]NUS Graduate School for Integrative Sciences and Engineering, Singapore 117456, Singapore

*To whom correspondence may be addressed:
Anders J. Skanderup, PhD
Genome Institute of Singapore, Singapore
60 Biopolis St, Singapore 138672
skanderupamj@gis.a-star.edu.sg

Dr. Daniel S.W. Tan, MD, PhD
Division of Medical Oncology, National Cancer Centre Singapore, Singapore
11 Hospital Cres, Singapore 169610
daniel.tan.s.w@singhealth.com.sg

Steven G. Rozen, PhD
Centre for Computational Biology, Duke-NUS Medical School, Singapore
8 College Rd, Singapore 169857
steve.rozen@duke-nus.edu.sg

Abstract

Background: Unlike smoking-related non-small cell lung cancers (NSCLCs), oncogene-driven NSCLCs (including those driven by epidermal growth factor receptor – *EGFR*) are characterized by low mutational burdens and complex genomic landscapes. However, the clonal architecture and genomic landscape of the oncogene-driven NSCLCs in smokers remain unknown. Here, we investigate the impact of tobacco smoking on genomic and transcriptomic alterations in the context of oncogene-driven NSCLC.

Methods: Patients undergoing resection for NSCLC at the National Cancer Centre Singapore were enrolled in this study. Resected tumors were divided into multiple regions, which then underwent whole-exome sequencing and bulk RNA sequencing. We investigated tumor mutational burden, intra-tumor heterogeneity, tumor phylogeny, mutational signatures, and transcriptomes across the regions of each tumor.

Results: We studied a total of 173 tumor sectors from 48 patients. Tumors were classified into three groups: “oncogene-driven non-smoking” (n=25, 52%), “oncogene-driven smoking” (n=12, 25%) and “typical smoking” (n=11, 23%). Oncogene-driven smoking versus non-smoking tumors did not differ significantly in terms of tumor mutational burden, intra-tumor heterogeneity, and driver mutation composition. Surprisingly, the mutational signature caused by tobacco smoking was essentially absent in oncogene-driven smoking tumors, despite prominent smoking histories. Compared to oncogene-driven non-smoking tumors, oncogene-driven smoking tumors had higher activity in pathways related to regulation of cell cycle, especially mitotic exit.

Conclusions: Oncogene-driven tumors in smokers shared similar clonal architecture and genomic features with archetypical oncogene-driven tumors in non-smokers. Oncogene-driven tumors in smokers had low tumor mutational burden and high intra-tumor heterogeneity and the mutational signature of smoking was largely absent. However, among oncogene-driven tumors, the differences in transcriptomic pathway activities between smokers and non-smokers suggest that smoking may foster a tumor phenotype distinct from that in non-smokers.

Highlights

- Like oncogene-driven NSCLC tumors in smokers, oncogene-driven NSCLC tumors in non-smokers have low mutational burden and high intra-tumor heterogeneity.
- The mutational signature of smoking was prevalent in typical smoking-related NSCLC but not in oncogene-driven NSCLC in smokers.
- Oncogene-driven NSCLC in smokers had high activity of pathways related to cell cycle, especially mitotic exit.
- This study highlights the genomic and transcriptomic features of oncogene-driven NSCLC in smokers, which suggest further investigation into optimizing treatment strategies.

Keywords

Oncogene-driven non-small cell lung cancer, *EGFR*-mutated lung cancer, oncogene-driven lung cancers with smoking history, tumor mutational burden, intra-tumor heterogeneity, mutational signature, SBS4

Introduction

Lung cancer remains the most lethal cancer worldwide and causes more than 1.8 million deaths annually, even though the worldwide prevalence of tobacco smoking is decreasing[1,2]. However, in East Asia, lung cancer in non-smokers is increasing and has become an emerging health problem[3,4]. Many of these are non-small-cell lung cancers (NSCLCs) that are driven by specific oncogenic mutations; usually activating mutations in oncogenes such as *EGFR*, *ERBB2*, or *MET* or activating fusions involving genes such as *ALK*, *ROS1*, or *RET*[5-9]. These oncogene-driven tumors constitute approximately half of NSCLC in East Asia[10-13] and tend to have lower mutational burdens and favorable responses to targeted therapies such as tyrosine-kinase inhibitors[14-21]. In contrast, non-oncogene-driven NSCLCs typically are smoking-related and have high mutational burdens and favorable responses to immune checkpoint inhibitors[22]. While oncogene-driven NSCLCs have been most studied in never-smokers, in East Asia, approximately 30% to 40% of patients with oncogene-driven NSCLC have histories of tobacco smoking[23].

To investigate similarities and differences between oncogene-driven NSCLC in non-smokers, oncogene-driven NSCLC in smokers, and typical smoking NSCLC, we carried out an integrated genomic and transcriptomic study of clonal architecture and intra-tumor heterogeneity across 173 tumor sectors in 48 patients representing all 3 groups.

Material and methods

Patients and clinical outcomes

Patients diagnosed with NSCLC at the National Cancer Centre Singapore (between 2013 and 2017) who underwent surgical resection of their tumors prior to receiving any form of anti-cancer therapy were enrolled in this study. Clinical information and histopathological features were curated by the Lung Cancer Consortium Singapore (Supplementary Table S1). Written informed consent was obtained from all participants. The study was approved by the SingHealth Centralized Institutional Review Board (CIRB reference 2018/2963).

Definition of index oncogenes in non-smoking tumor

To define oncogene-driven NSCLC, based on previous reports[24,25], we assembled a list of driver mutations and gene rearrangements (index oncogenes) characteristic of NSCLCs in non-smokers in East Asia (Supplementary Table S2). These included *EGFR* exon 18-21 activating mutations, *ALK* fusions, *ERBB2* exon 20 insertions, *RET* fusions, and *MET* exon 14 skipping mutations. In contrast, activating mutations in the *KRAS* and *BRAF* genes were characteristic of NSCLCs arising from smokers.

Tumor/normal sample processing and whole-exome sequencing

Resected tumors and paired normal samples were sectioned and processed as previously described[26]. Peripheral blood, or if peripheral blood was not available, normal lung tissue adjacent to the tumor was taken as a normal sample. The median number of sectors for an individual tumor was 3 (range 2-7, Supplementary Table S1). For whole exome sequencing, genomic DNA was extracted with the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen), and 500 ng to 1 μg of genomic DNA was sheared using Covaris to a size of 300 to 400 bp. Libraries were prepared with NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs). Regions to sequence were selected with the SeqCap EZ Human Exome Library v3.0 (Roche Applied Science) according to the manufacturer's instructions and underwent 2 × 151 base-pair sequencing on Hiseq 4000 (Illumina) sequencers. The median coverage of the capture target was 55.1X and 54.4X for normal and tumor samples, respectively (Supplementary Table S3).

Somatic single nucleotide variation and insertion-deletion calling

Exome reads were trimmed with trimmomatic (version 0.39) to remove adaptor-containing or poor-quality sequences[27]. Trimmed reads were mapped to the human reference sequence GRCh38.p7 (accession number GCA_000001405.22) using the BWA-mem software (version 0.7.15) with default parameters[28]. Duplicate reads were marked and removed from variant calling using sambamba (version 0.7.0)[29]. Global mapping quality was evaluated by Qualimap 2 (version 2.2.1, Supplementary Table S3)[30]. Somatic single nucleotide variations (SNVs) and insertion-deletions (indels) were called by MuTect2 (version 4.1.6.0) and Strelka2 (version 2.9.2) with default parameters[31,32]. We considered only variants called by both variant callers and with (i) ≥ 3 reads supporting the variant allele in the tumor sample, (ii) sequencing depth ≥ 20 in both the normal and tumor samples, and (iii) variant allele fraction ≥ 0.05. The somatic SNVs and indels were provided at https://github.com/Rozen-Lab/oncogene-NSCLC/supp-table-gene-mutation.csv. Variants were annotated by wANNOVAR (https://wannovar.wglab.org/)[33]. Driver status of genes was based on the Catalog of Somatic Mutations in Cancer (COSMIC) database, downloaded 24 February 2021 (https://cancer.sanger.ac.uk/census)[34].

We excluded 12 out of 185 sectors (6.5%) that had tumor purity < 0.1 from downstream analysis. We estimated tumor purity as follows: for tumors with a known oncogenic *EGFR*, *ERBB2*, *MET,* or *KRAS* mutation, we used the variant allele fraction (VAF) for each sector as calculated by Integrative Genomics Viewer (version 2.6.3, https://igv.org/)[35]. We reasoned that these oncogenic mutations were likely clonal and therefore would appear in all sectors[36]. We also reasoned that the index oncogenic mutations would be present in at least one-half of the chromosomes. From this, we estimated the tumor purity to be 2 times the VAF of the oncogenic mutation.

Definitions of truncal mutation, branch mutation, tumor mutational burden, and intra-tumor heterogeneity

We refer to mutations present in every sector of a tumor as "truncal", and we refer to other mutations as "branch". We defined tumor mutational burden (TMB) as the mean number of unique non-silent (nonsynonymous or splice-site) mutations across all sectors of a tumor. We defined intra-tumor heterogeneity (ITH) as the mean proportion of the number of unique branch mutations across all sectors.

Phylogenetic analysis

We used the Python PTI package (https://github.com/bioliyezhang/PTI, version 1.0) using the input of a "binary matrix" to infer phylogenetic relationships based on non-silent mutations[37].

Mutational signature assignment and spectrum reconstruction

Mutational signature assignment was carried out with mSigAct R package (version 2.3.2, https://github.com/steverozen/mSigAct) and COSMIC mutational signature database version 3.2 (https://cancer.sanger.ac.uk/signatures/)[38]. To better estimate the impact of smoking on cancer evolution, we first used the SignaturePresenceTest function with default parameters on all individual sectors within each group to decide whether the SBS4 mutational signature (the signature of tobacco smoking) was present in the sector's mutational spectrum. In brief, SignaturePresenceTest estimates optimal coefficients for the reconstruction of the observed spectrum using the mutational signatures previously detected in NSCLC[39]. The test does this without the SBS4 signature (null hypothesis) and with the SBS4 signature (alternative hypothesis). The test then carries out a standard likelihood ratio test on these two hypotheses to calculate a p-value. We then calculated Benjamini-Hochberg false discovery rates across all sectors of all tumors within the group. To estimate the contribution of signatures to each spectrum we used the SparseAssignActivity function and the signatures found in lung adenocarcinomas in reference[39], except that SBS4 was included only if the false discovery rate based on the SignaturePresenceTest was < 0.5. We also excluded SBS3 (caused by defective homologous recombination DNA damage repair mechanism) from sparse assignment after ensuring no pathogenic mutation of the germline and somatic *BRCA1* or *BRCA2* genes in all samples. Supplementary Table S4 and S5 show the mutational spectra and signature activity of each sector.

Detection of fusion transcripts

We used STAR-Fusion (version 1.10.0)[40] to detect transcript fusions in the RNA-sequencing data with default parameters. We required candidate fusions to satisfy the following criteria:

- spanning fragment count ≥ 1
- junction read count + spanning fragment count ≥ 5
- presence of a large anchor-support read, and
- for intrachromosomal fusion partners, a genomic distance ≥ 1MB between fusion breakpoints.

Of the putative transcript fusions detected, three are considered oncogenic variants according to the literature: *EML4-ALK*, *KLC1-ALK*, and *PARG-BMS1*[41-43]. Supplementary Table S7 provides the full list of putative fusions.

RNA sequencing and gene expression subtype

Total RNA was extracted and processed from 103 tumor samples as previously described[44]. We used the STAR software (version 2.7.3a) to align raw RNA sequence reads to the human genome (GRCh38p7 build) and to estimate transcript abundance based on the reference transcriptome (GRCh38.85 build)[45]. Only the counts of protein-coding genes were included for downstream analysis. The raw gene expression matrix is provided at https://github.com/Rozen-Lab/oncogene-NSCLC/supp-table-gene-expression-count-matrix.csv.

Transcriptomic pathway analysis

Raw gene expression levels were transformed to transcript levels in transcripts per million (TPM) values[46]. We computed pathway enrichment scores with the GSVA R package (version 1.40.1) and the Reactome subset of the Molecular Signatures Database (MSigDB version 7.5.1, https://www.gsea-msigdb.org/gsea/msigdb/)[47-49]. The pathway activity is provided at https://github.com/Rozen-Lab/oncogene-NSCLC/supp-table-pathway-activities-matrix.csv. Differential pathway expression was conducted using limma R package (version 3.48.0)[50]. Pathways with a Benjamini-Hochberg false discovery rate < 0.05

were taken as significant. Assignment of gene expression subtypes (terminal respiratory unit, TRU, versus non-TRU) was carried out as described[51]. Gene expression values and pathway enrichment scores were transformed to Z-scores (mean of 0 and standard deviation of 1) before downstream analysis. Heatmaps were constructed with the ComplexHeatmap R package (version 2.8.0)[52]. Heatmap columns were first clustered based on all rows using ComplexHeatmap::Heatmap function using default arguments for clustering distance and method, and then ordered by main group, patient, and gene expression status accordingly.

Data and code availability

All WES and RNA sequencing data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), under the accession number EGAS00001006942. R code used in this study is provided at https://github.com/Rozen-Lab/oncogene-NSCLC/.

Results

Clinical and histopathological characteristics

We studied a total of 173 tumor sectors from 48 patients with resected NSCLC. Table 1 summarizes clinical and histopathological characteristics. We classified tumors into three groups: “oncogene-driven non-smoking”: tumors with index oncogene mutations in never-smokers (n=25, 52%), “oncogene-driven smoking”: tumors with index oncogene mutations in smokers (n=12, 25%), and “typical smoking”: tumors without index oncogene mutation in smokers (n=11, 23%). The smoking history was similar between oncogene-driven smoking (median 34.5 pack-years, range 0.5-99) and typical smoking groups (median 38, range 2-168, Wilcoxon rank-sum test, p value = 0.5792). Besides the differences in smoking status and oncogene mutation, gender distribution also differed significantly across the three groups.

Oncogene-driven NSCLC with and without smoking histories have similar genomic architectures

Overall, we identified 6,251 single nucleotide variants and 314 small indels affecting the exons of 4,738 genes and the splicing junctions of 177 genes. Oncogene mutations were detected in 23 of 25 (92%) oncogene-driven non-smoking tumors. We classified the remaining 2 tumors (8%) as oncogene-driven because they arise in never-smokers and phylogenetically resemble oncogene-driven non-smoking tumors (Supplementary Figure S1). There were 18 tumors with *EGFR* mutations, 3 tumors with *MET* exon 14 skipping mutations, 1 tumor with an *ERBB2* mutation, and 1 tumor with an *ALK* fusion. Table 1 details these in the oncogene-driven non-smoking tumors. Oncogene mutations were detected in all 12 oncogene-driven smoking tumors. Eight of these tumors had *EGFR* mutations. We identified *KRAS* mutations in 7 out of 11 (64%) typical smoking tumors (2 with G12D, 2 with G12V, 1 with G12A, 1 with G12C, and 1 with Q61H). Mutations in *EGFR*, *MET*, *ERBB2*, and *KRAS* and *ALK* fusions did not co-occur in this study. Across all three groups of tumors, after *EGFR*, *TP53* was the second most mutated gene (22 of 48, 46%), consistent with previously published East-Asian cohorts[9,10].

In this study, all *EGFR*, *ERBB2*, *MET,* and *KRAS* mutations were known oncogenic mutations and were truncal, underscoring their central roles in early oncogenesis (Supplementary Table S8). By contrast, only 1,813 out of 6,215 mutations in non-drivers were truncal. Two of the three presumed oncogenic fusions were also truncal: 1 *EML4-ALK* fusion in the oncogene-driven non-smoking group and 1 *KLC1-ALK* fusion in the oncogene-driven smoking group (Figure 1A).

Oncogene-driven smoking tumors had slightly higher TMB than non-smoking tumors (median 55.5 vs. 40 mutations, $p = 0.039$, two-sided Wilcoxon rank-sum test). By contrast, compared to oncogene-driven smoking tumors, typical smoking tumors had much higher TMB (median 144 vs. 55.5, p = 0.017, two-sided Wilcoxon rank-sum test), more truncal mutations (median 56 vs. 22.5, p = 0.031), and more mutations in COSMIC driver genes (median 14 vs. 7.5, p = 0.002, Figure 1B). Although ITH (see Methods) was similar across the three groups (medians 0.593, 0.543, and 0.580, for oncogene-driven non-smoking tumors, oncogene-driven smoking tumors, and typical smoking tumors, respectively, Figure 1B), “coconut-tree” phylogenies, characterized by a combination of high TMB (> 100) and low ITH (< 0.5), occurred exclusively among the typical smoking tumors (5 out of 11, Figure 2, Supplementary Figure S1).

In addition to those mutations used to categorize typical-smoking versus oncogene-driven tumors, *CSMD3* mutations were statistically more common in typical smoking tumors (Figure 1C, left and middle). In comparing oncogene-driven smoking versus non-smoking tumors, there was no significant difference in the prevalence of mutations in COSMIC driver genes (Figure 1C, right).

Previous studies reported that whole-genome doubling (WGD) had occurred in 30% to 80% of NSCLC, and the WGD is associated with poor clinical outcome[26,53,54]. Moreover, cancers in which large fractions of the genome had copy number gain or loss (high genome instability index) and subclonal copy number change (also known as subclonal allelic imbalance) were prevalent among NSCLC[26,55]. In the present study, we found no significant difference across the three groups in terms of WGD rate, tumor ploidy, genome instability index, and subclonal allelic imbalance index (Supplementary Figure S2). Supplementary Figure S3 provides details of copy number variation profiles for all groups collectively and individually.

We also note that gender distribution differed strongly across the three groups. Among patients with oncogene-driven non-smoking tumors, only 32% were male, whereas among the oncogene-driven smoking and typical smoking groups 92% and 100% were male, respectively ($p = 0.0011$ and 0.0001 by two-sided Fisher's exact tests compared to the oncogene-driven non-smoking group). We analyzed genomic landscapes in oncogene-driven tumors by gender and found no significant differences (Supplementary Figure S4).

Mutational signatures of oncogene-driven versus typical smoking NSCLC

We next investigated the impact of smoking on the mutational landscape across three groups. We used a signature presence test followed by signature attribution with mSigAct software to detect the mutational signature SBS4, which is caused by tobacco smoking in lung cancers (Figure 3A)[39,56]. We were able to detect SBS4 in 30 of 34 (88%) typical smoking tumor sectors. Surprisingly, however, SBS4 was found in only 7 of 48 (15%) oncogene-driven smoking tumor sectors, significantly less than typical smoking tumor sectors despite similar smoking histories (two-sided Fisher's exact test, $p < 2.1\text{x}10^{-10}$, Figure 3B). For tumor sectors with SBS4 activity, the median number of mutations attributed to SBS4 was 216 for typical smoking tumors versus 53 for oncogene-driven smoking tumors (two-sided Wilcoxon rank-sum test, $p < 9\text{x}10^{-5}$, Figure 3C). T-distributed stochastic neighbor embedding (tSNE) based on single-base-substitution spectra identified different mutational patterns in typical smoking sectors compared with oncogene-driven sectors (Figure 3D).

To confirm the surprising paucity of SBS4 activity in oncogene-driven smoking tumors, we applied the same signature assignment algorithm to a subset of the TCGA-LUAD (lung adenocarcinoma) cohort[57]. This subset consisted of 406 tumors with mutational spectra reported in reference [39] and smoking-history data from TCGA. For each tumor, Supplementary Table S6 provides the clinical information, including smoking history, index oncogenes and their mutations, and signature activity. SBS4 was found in 2 of 24 (8%) oncogene-driven non-smoking tumors, 8 of 21 (38%) oncogene-driven smoking tumors, and 260 of 290 (90%) typical smoking tumors (Supplementary Figure S5). Thus, all the genomic data indicates that oncogene-driven tumors, whether in smokers or non-smokers, have origins and oncogenic histories distinct from those of typical smoking tumors.

Previous studies found that, in NSCLC, APOBEC mutations were enriched in branch versus truncal mutations[26,58]. Thus, we investigated differences in the activities of APOBEC and other signatures in branch versus truncal mutations in the entire data set and in each of the three groups in our study (Supplementary Figure S6A, S6B). Levels of APOBEC mutations were elevated in branches than trunks across the entire data set (Supplementary Figure S6C). Unexpectedly, we found that mutations due to reactive oxygen species (ROS, SBS18) were significantly higher in the branches compared to trunks for every group of tumors (all q values < 0.0068 by two-sided Wilcoxon rank-sum tests with Benjamini-Hochberg correction, Supplementary Figure S6C). This may suggest a potential role of ROS in promoting tumor evolution and clonal mutagenesis in NSCLC. The smoking signature (SBS4) was a major contributor only in typical smoking tumors, in which it contributed similar activities in trunks and branches (Supplementary Figure S6C).

Transcriptomic features of oncogene-driven smoking tumors

The similarity of genomic landscapes between oncogene-driven tumors with and without smoking histories was surprising because many clinical studies have shown that smoking is an indicator of poor prognosis in patients with advanced EGFR-mutated NSCLC treated with tyrosine kinase inhibitors[59-61]. Therefore, we investigated whether transcriptomic factors might help reveal the reason for this clinical observation. To this end, we profiled the transcriptomes of 103 of the 173 sectors from 32 out of the 48 patients. UMAP dimension reduction did not reveal a strong separation between oncogene-driven tumors in smoking versus non-smoking patients (Supplementary Figure S7A). Indeed, the primary separation seems to be between tumors with the terminal-respiratory-unit (TRU) subtype and those without it (Supplementary Figure S7B).

To further explore the transcriptomic activities associated with tobacco smoking, we conducted differential expression pathway analysis between oncogene-driven non-smoking tumors and both groups of smoking tumors (typical smoking and oncogene-driven smoking tumors grouped together) across 1,259 pathways from the Reactome Database[49] (Figure 4, Supplementary Figure S8, Supplementary Tables S9, S10). Pathways with high activities in non-smoking tumors included those related to NOTCH signaling and to glycosaminoglycans and arylsulfatases, which catabolize glycosaminoglycans. Pathways with low activities in non-smoking tumors included many due to cell cycle regulation, especially mitotic exit. These observations underscore differences between smoking and non-smoking tumors in the activities of pathways related to two of the major hallmarks of cancer.

It has been proposed that smoking-associated lung cancers are more likely to trigger an anti-tumor immune response that would confer a better response to immunotherapy[62,63]. Antitumor responses mediated by immune checkpoint inhibitors (ICIs) correlate with the immune repertoires of the tumor microenvironment (TME)[64-66]. “Immune-hot” tumors are characterized by the infiltration of cytotoxic T cells capable of recognizing and killing tumor cells, the expression of proinflammatory and effector cytokine genes, and higher tendencies to respond to ICIs[64,67]. Therefore, we investigated whether smoking, independent of genomic alterations, can foster an immune-hot TME in NSCLC. To detect immune-hot TME, we performed hierarchical clustering of sectors based on the transcript levels of T-cell inflammation and immune checkpoint genes (Supplementary Figure S9A)[68-70]. We did not see strong evidence for enrichment of the immune-hot TME in oncogene-driven smoking tumors (1 of 9, 11%) and typical smoking tumor sectors (1 of 5, 20%) compared to oncogene-driven non-smoking tumor sectors (2 of 18, 11%, Supplementary Figure S9B). However, we found that immune-hot TME was more prevalent in tumor sectors harboring a TP53 mutation (16 of 43; 37.2%) as compared with those without TP53 mutations (5 of 60, 8.3%, q value of 0.0023 using a two-sided Fisher’s exact test with Benjamini-Hochberg correction, Supplementary Figure S9C), which is consistent with a previous finding of pro-inflammatory tendencies in these tumors[71].

Discussion

To our knowledge, this is the first integrated genomic study to conduct direct comparisons across oncogene-driven non-smoking, oncogene-driven smoking, and typical smoking NSCLCs through multi-region exome and RNA sequencing. Surprisingly, we found that tobacco smoking has almost no influence on the genomic features and clonal architectures of *EGFR*-mutated and other oncogene-driven NSCLCs. Despite prominent smoking histories, oncogene-driven smoking tumors were similar to oncogene-driven non-smoking tumors in terms of mutational burden, mutational signature activity, and intra-tumor heterogeneity. In contrast, compared to both groups of oncogene-driven NSCLC, typical smoking tumors showed higher TMBs. Furthermore, “coconut-tree” phylogenies, which are defined by a combination of high TMB (> 100) and low ITH (< 0.5), occurred in nearly half of the typical smoking tumors but were absent from oncogene-driven NSCLC.

As noted in the Results section, gender distribution differed significantly across the three groups. Across all groups, tobacco smoking was more prevalent among males compared to females (Table 1). This male preponderance reflects the extreme gender imbalance of smoking in East Asia. For example, in the population we studied, 6.8% of women are smokers compared to 20.6% of men[2,72,73]. Previously, oncogene-driven NSCLC was sometimes viewed as a disease of non-smokers, often women. This view may have been partly driven by this gender imbalance, due to which oncogene-driven tumors were particularly salient among women, since they were usually non-smokers. The current study confirms oncogene-driven NSCLC occurs in both smokers and non-smokers and in both sexes, and it shows that genomic features are similar in both smokers and non-smokers and in both sexes. Because of the strong differences in smoking rates between women and men in the study population, it is not possible to disentangle the effects of gender from the effects of smoking.

Nevertheless, we note that available evidence suggests that oncogene-driven tumors are more common among women. In both non-smokers and smokers, *EGFR*-mutated NSCLC is more common among women: for non-smokers, odds-ratio 1.38 ($p < 8 \times 10^{-9}$); and for smokers, odds ratio 1.40 ($p < 0.006$, analyses by two-sided Fisher's exact tests on data from Tseng et al[74]).

An unexpected discovery from this study was the paucity of mutations due to tobacco smoking in oncogene-driven smoking tumors. We confirmed this discovery in the 21 patients with oncogene-driven smoking tumors in the TCGA-LUAD cohort[57]. It is unclear why oncogene-driven smoking tumors rarely acquire mutations caused by smoking, while typical smoking tumors with similar exposures have abundant smoking mutations. Indeed, studies suggest that the cell of origin of oncogene-driven NSCLC may be different from that of typical smoking NSCLC[75-79]. Thus, it may be that oncogene-driven smoking tumors are less prone to mutation because their cells of origin are less exposed to tobacco smoke or have more effective DNA damage repair.

Although oncogene-driven smoking and non-smoking tumors have similar clonal architectures and genomic features, they differ in transcriptomic pathway activities, especially those related cell cycle and mitotic exit. Indeed, for these pathway activities, oncogene-driven smoking tumors are more similar to typical smoking tumors than to oncogene-driven non-smoking tumors (Figure 4, Supplementary Figure S8). Of note, advanced *EGFR*-mutated NSCLCs treated with tyrosine kinase inhibitors (TKIs) had worse outcomes in smokers than in non-smokers[60,61]. The transcriptomic activities of oncogene-driven NSCLC in smokers might account for these cancers' higher resistance to standard TKIs but could potentially lead to higher susceptibility to therapies such as chemotherapy or CDK4/6 inhibitors that target the cell cycle. This warrants further investigation regarding the selection of treatments for patients with advanced-stage, oncogene-driven smoking NSCLC.

In summary, based on the multi-region whole-exome and RNA sequencing, we have elucidated the clonal architectures and genomic features of three groups of East-Asian NSCLC: oncogene-driven non-smoking, oncogene-driven smoking, and typical smoking tumors. In the context of oncogene-driven disease, we found no evidence that tobacco smoking affects the clonal evolution or genomic alteration of NSCLC. However, the transcriptomic pathway activities were more similar between oncogene-driven smoking tumors and typical smoking tumors than between smoking and non-smoking oncogene-driven cancers. The in-depth analysis of oncogene-driven NSCLC in smokers and non-smokers presented here may provide a guide to optimizing treatment approaches.

Funding

This work was funded by the National Medical Research Council (NMRC; Singapore) through the Translational and Clinical Research Program (NMRC/TCR/007-NCC/2013), Open Fund Large Collaboration Grant (NMRC/OFLCG/002/2018). This work was also funded in part by grants from the Chang Gung Memorial Hospital (CGMH) Foundation (Grant No. CMRPG3F1911) to Chen-Yang Huang.

Disclosure

GL reports receiving personal fees from Astra Zeneca and grants from Merck, Astra Zeneca, Pfizer, BMS, Amgen and Roche outside the submitted work and sponsorship from DKSH. **ACT** reports receiving personal fees from ASLAN Pharmaceuticals, and Illumina, consultation fees from Pfizer, Amgen, Bayer, and honoraria from Amgen, Thermo Fisher Scientific, Janssen, Pfizer, Juniper Biologics, and Guardant Health. **SPS** reports receiving personal fees from MSD, consultation fees from Pfizer, and Bayer, and grants from Astra Zeneca, and Guardant Health. **RV** reports receiving honoraria and consultation fees from MSD, BMS, Astellas, Novartis, Pfizer, Merck, J&J, and AstraZeneca. **WLT** reports receiving grant support from AstraZeneca, honoraria from Novartis, Merck, Amgen, personal fee from AstraZeneca, Ipsen, Boehringer Ingelheim, Bristol-Myers-Squibb, and DKSH. **BHO** reports receiving personal fees from AstraZeneca, Medtronic, Stryker, and MSD. All remaining authors have declared no conflicts of interest.

Acknowledgements

We thank Jacob J.S. Alvarez, Jia Chi Yeo from the Genome Institute of Singapore, for data transfer. We thank Willie Yu from Duke-NUS Medical School, for helping with some Unix/Python scripts and setting up some bioinformatic pipelines. We thank Xinyi Yang from Duke-NUS Medical School for testing the R codes. We thank the Lung Cancer Consortium Singapore (LCCS) for assisting with specimen collection and clinical data compilation. Finally, we are grateful to the patients, physicians, and pathologists at the National Cancer Centre Singapore and Singapore General Hospital who contributed patient material.

References

1 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249, doi:10.3322/caac.21660 (2021).

2 G. B. D. Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. *Lancet* **397**, 2337-2360, doi:10.1016/S0140-6736(21)01169-7 (2021).

3 Tseng, C. H. *et al.* The Relationship Between Air Pollution and Lung Cancer in Nonsmokers in Taiwan. *J Thorac Oncol* **14**, 784-792, doi:10.1016/j.jtho.2018.12.033 (2019).

4 Toh, C. K. *et al.* A Decade of Never-smokers Among Lung Cancer Patients-Increasing Trend and Improved Survival. *Clin Lung Cancer* **19**, e539-e550, doi:10.1016/j.cllc.2018.03.013 (2018).

5 Cho, J. *et al.* Proportion and clinical features of never-smokers with non-small cell lung cancer. *Chin J Cancer* **36**, 20, doi:10.1186/s40880-017-0187-6 (2017).

6 Zhang, Y. *et al.* Frequency of driver mutations in lung adenocarcinoma from female never-smokers varies with histologic subtypes and age at diagnosis. *Clin Cancer Res* **18**, 1947-1953, doi:10.1158/1078-0432.CCR-11-2511 (2012).

7 Li, C. *et al.* Spectrum of oncogenic driver mutations in lung adenocarcinomas from East Asian never smokers. *PLoS One* **6**, e28204, doi:10.1371/journal.pone.0028204 (2011).

8 Sun, Y. *et al.* Lung adenocarcinoma from East Asian never-smokers is a disease

largely defined by targetable oncogenic mutant kinases. *J Clin Oncol* **28**, 4616-4620, doi:10.1200/JCO.2010.29.6038 (2010).

9 Chen, Y. J. *et al.* Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell* **182**, 226-244 e217, doi:10.1016/j.cell.2020.06.012 (2020).

10 Chen, J. *et al.* Genomic landscape of lung adenocarcinoma in East Asians. *Nat Genet* **52**, 177-186, doi:10.1038/s41588-019-0569-6 (2020).

11 Hsu, K.-H. *et al.* Identification of five driver gene mutations in patients with treatment-naive lung adenocarcinoma in Taiwan. *Plos one* **10**, e0120852 (2015).

12 Melosky, B. *et al.* Worldwide prevalence of epidermal growth factor receptor mutations in non-small cell lung cancer: A meta-analysis. *Molecular Diagnosis & Therapy* **26**, 7-18, doi:10.1007/s40291-021-00563-1 (2022).

13 Tan, A. C. & Tan, D. S. W. Targeted therapies for lung cancer patients With oncogenic driver molecular alterations. *Journal of Clinical Oncology* **40**, 611-625, doi:10.1200/JCO.21.01626 (2022).

14 Li, B. T. *et al.* Trastuzumab Deruxtecan in HER2-Mutant Non-Small-Cell Lung Cancer. *N Engl J Med* **386**, 241-251, doi:10.1056/NEJMoa2112431 (2022).

15 Wolf, J. *et al.* Capmatinib in MET Exon 14-Mutated or MET-Amplified Non-Small-Cell Lung Cancer. *N Engl J Med* **383**, 944-957, doi:10.1056/NEJMoa2002787 (2020).

16 Paik, P. K. *et al.* Tepotinib in Non-Small-Cell Lung Cancer with MET Exon 14 Skipping Mutations. *N Engl J Med* **383**, 931-943, doi:10.1056/NEJMoa2004407 (2020).

17 Drilon, A. *et al.* Efficacy of Selpercatinib in RET Fusion-Positive Non-Small-Cell Lung Cancer. *N Engl J Med* **383**, 813-824, doi:10.1056/NEJMoa2005653 (2020).

18 Wu, Y. L. *et al.* Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. *Lancet Oncol* **15**, 213-222, doi:10.1016/S1470-2045(13)70604-1 (2014).

19 Solomon, B. J. *et al.* First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N Engl J Med* **371**, 2167-2177, doi:10.1056/NEJMoa1408440 (2014).

20 Zhou, C. *et al.* Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol* **12**, 735-742, doi:10.1016/S1470-2045(11)70184-X (2011).

21 Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* **361**, 947-957, doi:10.1056/NEJMoa0810699 (2009).

22 Dai, L. *et al.* The effect of smoking status on efficacy of immune checkpoint inhibitors in metastatic non-small cell lung cancer: A systematic review and meta-analysis. *EClinicalMedicine* **38**, 100990, doi:10.1016/j.eclinm.2021.100990 (2021).

23 Tseng, C.-H. *et al.* EGFR mutation, smoking, and gender in advanced lung adenocarcinoma. *Oncotarget* **8**, 98384-98393, doi:10.18632/oncotarget.21842 (2017).

24 Zheng, D. *et al.* MET exon 14 skipping defines a unique molecular class of non-small cell lung cancer. *Oncotarget* **7**, 41691-41702, doi:10.18632/oncotarget.9541 (2016).

25 Gou, L. Y., Niu, F. Y., Wu, Y. L. & Zhong, W. Z. Differences in driver genes between smoking-related and non-smoking-related lung cancer in the Chinese population. *Cancer* **121 Suppl 17**, 3069-3079, doi:10.1002/cncr.29531 (2015).

26 Nahar, R. *et al.* Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing. *Nat Commun* **9**, 216, doi:10.1038/s41467-017-02584-z (2018).

27 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

28 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

29 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034, doi:10.1093/bioinformatics/btv098 (2015).

30 Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294, doi:10.1093/bioinformatics/btv566 (2016).

31 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594, doi:10.1038/s41592-018-0051-x (2018).

32 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

33 Chang, X. & Wang, K. wANNOVAR: annotating genetic variants for personal

genomes via the web. *J Med Genet* **49**, 433-436, doi:10.1136/jmedgenet-2012-100918 (2012).

34 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

35 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

36 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128, doi:10.1038/s41586-019-1907-7 (2020).

37 Wu, P., Hou, L., Zhang, Y. & Zhang, L. Phylogenetic Tree Inference: A Top-Down Approach to Track Tumor Evolution. *Front Genet* **10**, 1371, doi:10.3389/fgene.2019.01371 (2019).

38 Ng, A. W. T. *et al.* Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aan6446 (2017).

39 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).

40 Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* **20**, 213, doi:10.1186/s13059-019-1842-9 (2019).

41 Ou, S. I., Zhu, V. W. & Nagasaka, M. Catalog of 5' Fusion Partners in ALK-positive NSCLC Circa 2020. *JTO Clin Res Rep* **1**, 100015, doi:10.1016/j.jtocrr.2020.100015 (2020).

42 Piscuoglio, S. *et al.* Genomic and transcriptomic heterogeneity in metaplastic carcinomas of the breast. *NPJ Breast Cancer* **3**, 48, doi:10.1038/s41523-017-0048-0 (2017).

43 Gotoh, M. *et al.* Comprehensive exploration of novel chimeric transcripts in clear cell renal cell carcinomas using whole transcriptome analysis. *Genes Chromosomes Cancer* **53**, 1018-1032, doi:10.1002/gcc.22211 (2014).

44 Chua, K. P. *et al.* Integrative Profiling of T790M-Negative EGFR-Mutated NSCLC Reveals Pervasive Lineage Transition and Therapeutic Opportunities. *Clin Cancer Res* **27**, 5939-5950, doi:10.1158/1078-0432.CCR-20-4607 (2021).

45 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

46 Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281-285, doi:10.1007/s12064-012-0162-3 (2012).

47 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene

set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

48 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7, doi:10.1186/1471-2105-14-7 (2013).

49 Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142, doi:10.1186/s12859-017-1559-2 (2017).

50 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).

51 Wilkerson, M. D. *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One* **7**, e36530, doi:10.1371/journal.pone.0036530 (2012).

52 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849, doi:10.1093/bioinformatics/btw313 (2016).

53 Lopez, S. *et al.* Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat Genet* **52**, 283-293, doi:10.1038/s41588-020-0584-7 (2020).

54 Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* **50**, 1189-1195, doi:10.1038/s41588-018-0165-1 (2018).

55 Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).

56 Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622, doi:10.1126/science.aag0299 (2016).

57 The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

58 de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256, doi:10.1126/science.1253462 (2014).

59 Chang, J. W. *et al.* Risk Stratification Using a Novel Nomogram for 2190 EGFR-Mutant NSCLC Patients Receiving the First or Second Generation EGFR-TKI. *Cancers (Basel)* **14**, doi:10.3390/cancers14040977 (2022).

60 Kim, I. A., Lee, J. S., Kim, H. J., Kim, W. S. & Lee, K. Y. Cumulative smoking dose affects the clinical outcomes of EGFR-mutated lung adenocarcinoma patients

treated with EGFR-TKIs: a retrospective study. *BMC Cancer* **18**, 768, doi:10.1186/s12885-018-4691-0 (2018).

61 Zhang, Y. *et al.* Impact of smoking status on EGFR-TKI efficacy for advanced non-small-cell lung cancer in EGFR mutants: a meta-analysis. *Clin Lung Cancer* **16**, 144-151 e141, doi:10.1016/j.cllc.2014.09.008 (2015).

62 Desrichard, A. *et al.* Tobacco Smoking-Associated Alterations in the Immune Microenvironment of Squamous Cell Carcinomas. *J Natl Cancer Inst* **110**, 1386-1392, doi:10.1093/jnci/djy060 (2018).

63 Sun, Y. *et al.* The Effect of Smoking on the Immune Microenvironment and Immunogenicity and Its Relationship With the Prognosis of Immune Checkpoint Inhibitors in Non-small Cell Lung Cancer. *Front Cell Dev Biol* **9**, 745859, doi:10.3389/fcell.2021.745859 (2021).

64 Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321-330, doi:10.1038/nature21349 (2017).

65 Bai, R., Lv, Z., Xu, D. & Cui, J. Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. *Biomark Res* **8**, 34, doi:10.1186/s40364-020-00209-0 (2020).

66 Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer* **19**, 133-150, doi:10.1038/s41568-019-0116-x (2019).

67 Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563-567, doi:10.1038/nature14011 (2014).

68 Liu, S. *et al.* Efficient identification of neoantigen-specific T-cell responses in advanced human ovarian cancer. *J Immunother Cancer* **7**, 156, doi:10.1186/s40425-019-0629-6 (2019).

69 Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, doi:10.1126/science.aar3593 (2018).

70 Ayers, M. *et al.* IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest* **127**, 2930-2940, doi:10.1172/JCI91190 (2017).

71 Fu, J. *et al.* A special prognostic indicator: tumor mutation burden combined with immune infiltrates in lung adenocarcinoma with TP53 mutation. *Transl Cancer Res* **10**, 3963-3978, doi:10.21037/tcr-21-565 (2021).

72 Yang, T. *et al.* Gender balance and its impact on male and female smoking rates in Chinese cities. *Soc Sci Med* **154**, 9-17, doi:10.1016/j.socscimed.2016.02.035 (2016).

73 Tsai, Y. W., Tsai, T. I., Yang, C. L. & Kuo, K. N. Gender differences in smoking behaviors in an Asian population. *J Womens Health (Larchmt)* **17**, 971-978, doi:10.1089/jwh.2007.0621 (2008).

74 Tseng, C. H. *et al.* EGFR mutation, smoking, and gender in advanced lung adenocarcinoma. *Oncotarget* **8**, 98384-98393, doi:10.18632/oncotarget.21842 (2017).

75 Chen, F. *et al.* Cellular Origins of EGFR-Driven Lung Cancer Cells Determine Sensitivity to Therapy. *Adv Sci (Weinh)* **8**, e2101999, doi:10.1002/advs.202101999 (2021).

76 Spella, M. *et al.* Club cells form lung adenocarcinomas and maintain the alveoli of adult mice. *Elife* **8**, doi:10.7554/eLife.45571 (2019).

77 Hynds, R. E. & Janes, S. M. Airway Basal Cell Heterogeneity and Lung Squamous Cell Carcinoma. *Cancer Prev Res (Phila)* **10**, 491-493, doi:10.1158/1940-6207.CAPR-17-0202 (2017).

78 Kadur Lakshminarasimha Murthy, P. *et al.* Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature* **604**, 111-119, doi:10.1038/s41586-022-04541-3 (2022).

79 Basil, M. C. *et al.* Human distal airways contain a multipotent secretory cell that can regenerate alveoli. *Nature* **604**, 120-126, doi:10.1038/s41586-022-04552-0 (2022).

Table 1. Baseline clinical and genomic characteristics of patients with oncogene-driven non-smoking, oncogene-driven smoking, and typical smoking non-small-cell lung cancers

Clinical or genomic characteristic, n (%)	**Oncogene-driven non-smoking (n = 25)**		**Oncogene-driven smoking (n = 12)**		**Typical smoking (n = 11)**		**All patients (n = 48)**		***P* value[b]**
Number of patients	25		12		11		48		
Number of tumor sectors with WES[a]	91	(100)	48	(100)	34	(100)	173	(100)	
Number of tumor sectors with RNA seq	57	(62.6)	32	(67)	14	(41)	103	(60)	0.34
Age, median (range)	66	(44-82)	70	(39-79)	66	(49-74)	67	(39-82)	n.s.[c]
Gender									
Male	8	(32)	11	(91.7)	11	(100)	30	(62.5)	<0.0001
Female	17	(68)	1	(8.3)	0	(0)	18	(37.5)	
Cigarette smoking status									
Never	25	(100)	0	(0)	0	(0)	25	(52.1)	<0.0001
Current/Former	0	(0)	12	(100)	11	(100)	23	(47.9)	
Pack years, median (range)	0	(0-0)	34.5	(0.5-99)	38	(2-168)	0	(0-168)	0.58[d]
Ethnicity									
Chinese	22	(88)	10	(83.3)	10	(90.9)	42	(87.5)	1
Non-Chinese	3	(12)	2	(16.7)	1	(9.1)	6	(12.5)	
Stage at diagnosis									
Early (I & II)	21	(84)	9	(75)	11	(100)	41	(85.4)	0.23
Late (III & IV)	4	(16)	3	(25)	0	(0)	7	(14.6)	
Histology									
Adenocarcinoma	24	(96)	12	(100)	10	(90.9)	46	(95.8)	0.47
Squamous cell carcinoma	1	(4)	0	(0)	1	(9.1)	2	(4.2)	
Oncogene mutation status									
EGFR exon 18-21	18	(72)	9	(75)	0	(0)	27	(56.3)	<0.0001
MET exon 14 skipping	3	(12)	1	(8.3)	0	(0)	4	(8.3)	
ALK fusion	1	(4)	1	(8.3)	0	(0)	2	(4.2)	
ERBB2 exon 20	1	(4)	1	(8.3)	0	(0)	2	(4.2)	
KRAS exon 2-3	0	(0)	0	(0)	7	(63.6)	7	(14.6)	
No oncogene	2	(8)	0	(0)	4	(36.4)	6	(12.5)	

[a] WES, whole exome sequencing.
[b] *P* value by two-sided Fisher's exact tests across all 3 group within category (e.g. Gender, Cigarette smoking, status, etc.).
[c] *P* values by two-sided Wilcoxon rank-sum test were insignificant among any 2 of the 3 groups.
[d] *P* values by two-sided Wilcoxon rank-sum test were insignificant between oncogene-driven smoking and typical smoking groups.

Figure Legends

Figure 1. (A) Genomic landscape of tumors and tumor sectors of (i) all branch and truncal mutations (ii) selected patient clinical information, (iii) whole-genome doubling, (iv) gene expression subtype, and (v) the presence or absence of driver mutations of interest. For tumors without RNA-sequencing data, there is no information on gene expression subtype or the 3 fusions at the bottom of the grid. (B) Counts of total mutations, truncal mutations, driver mutations, and levels of intra-tumor heterogeneity in the 3 groups. (C) Enrichment of driver mutations in comparisons among the 3 groups.

Figure 2. Intra-tumor heterogeneity (ITH) versus tumor mutation burden (TMB) for each tumor. Five tumors with “coconut-tree” phylogenies are labeled. These phylogenies occurred only in the typical-smoking group, and the corresponding phylogenies are shown below.

Figure 3. Single-base substitution (SBS) mutational signatures. (A) The first two sections show mutational-signature activity in the three groups by absolute mutation counts. and by proportion. The remaining sections show smoking status, the presence of mutations in selected oncogenes, and whether the phylogenetic pattern is a “coconut tree” pattern. Colors indicate various mutational signatures (e.g., SBS40, SBS5, etc.), as indicated by the legend above. (B) The proportions of tumor sectors with SBS4 (caused by tobacco smoking) by tumor group. (C) Counts of mutations due to SBS4 in tumor sectors that have SBS4 mutations. (D) tSNE (t-distributed stochastic neighbor embedding) dimension reduction based on the mutational spectra. For information on the mutational signatures, see COSMIC (https://cancer.sanger.ac.uk/signatures/).

Figure 4. Heatmap of activities of the top 10 pathways up- and down-regulated in oncogene-driven non-smoking tumors. Each column is a tumor sector, and sectors are grouped by patient as shown in the row labelled “Patient”. SBS4 is the mutational signature of tobacco smoking in lung cancers. Z-scores are of pathway activity. The Benjamini-Hochberg false discovery rates (q values) of differential pathway activity were based on p-values calculated using limma[50]. Supplementary Figure S8 and Supplementary Tables S9 and S10 provide details.

Figure 1

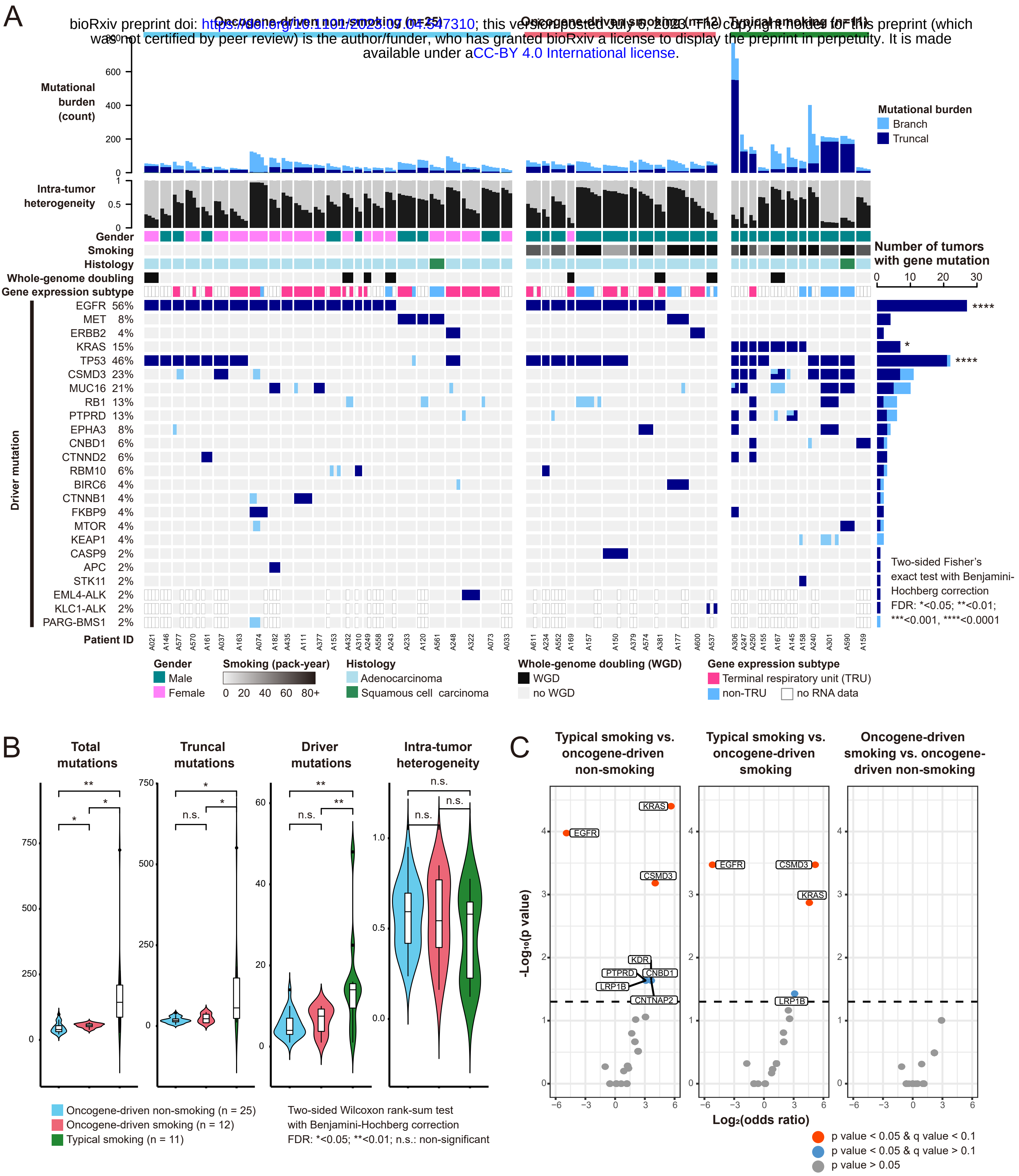

A
Oncogene-driven non-smoking (n=25)
Oncogene-driven smoking (n=12)
Typical smoking (n=11)
Mutational burden (count)
800
600
400
200
0
Mutational burden
Branch
Truncal
Intra-tumor heterogeneity
1
0.5
0
Gender
Smoking
Histology
Whole-genome doubling
Gene expression subtype
Number of tumors with gene mutation
0
10
20
30
Driver mutation
EGFR 56%
MET 8%
ERBB2 4%
KRAS 15%
TP53 46%
CSMD3 23%
MUC16 21%
RB1 13%
PTPRD 13%
EPHA3 8%
CNBD1 6%
CTNND2 6%
RBM10 6%
BIRC6 4%
CTNNB1 4%
FKBP9 4%
MTOR 4%
KEAP1 4%
CASP9 2%
APC 2%
STK11 2%
EML4-ALK 2%
KLC1-ALK 2%
PARG-BMS1 2%
****
*
****
Two-sided Fisher's exact test with Benjamini-Hochberg correction
FDR: *<0.05; **<0.01; ***<0.001, ****<0.0001
Patient ID
A021 A146 A577 A570 A161 A037 A163 A074 A182 A435 A111 A377 A153 A432 A310 A249 A558 A243 A233 A120 A561 A248 A322 A073 A033
A611 A234 A552 A169 A157 A150 A379 A574 A381 A177 A600 A537
A306 A247 A250 A155 A167 A145 A158 A240 A301 A590 A159
Gender
Male
Female
Smoking (pack-year)
0 20 40 60 80+
Histology
Adenocarcinoma
Squamous cell carcinoma
Whole-genome doubling (WGD)
WGD
no WGD
Gene expression subtype
Terminal respiratory unit (TRU)
non-TRU
no RNA data
B
Total mutations
Truncal mutations
Driver mutations
Intra-tumor heterogeneity
**
*
*
n.s.
n.s.
n.s.
750
500
250
0
60
40
20
1.0
0.5
0.0
Oncogene-driven non-smoking (n = 25)
Oncogene-driven smoking (n = 12)
Typical smoking (n = 11)
Two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction
FDR: *<0.05; **<0.01; n.s.: non-significant
C
Typical smoking vs. oncogene-driven non-smoking
Typical smoking vs. oncogene-driven smoking
Oncogene-driven smoking vs. oncogene-driven non-smoking
-Log₁₀(p value)
Log₂(odds ratio)
KRAS
EGFR
CSMD3
KDR
PTPRD
CNBD1
LRP1B
CNTNAP2
EGFR
CSMD3
KRAS
LRP1B
4
3
2
1
0
−6
−3
0
3
6
p value < 0.05 & q value < 0.1
p value < 0.05 & q value > 0.1
p value > 0.05

Figure 2

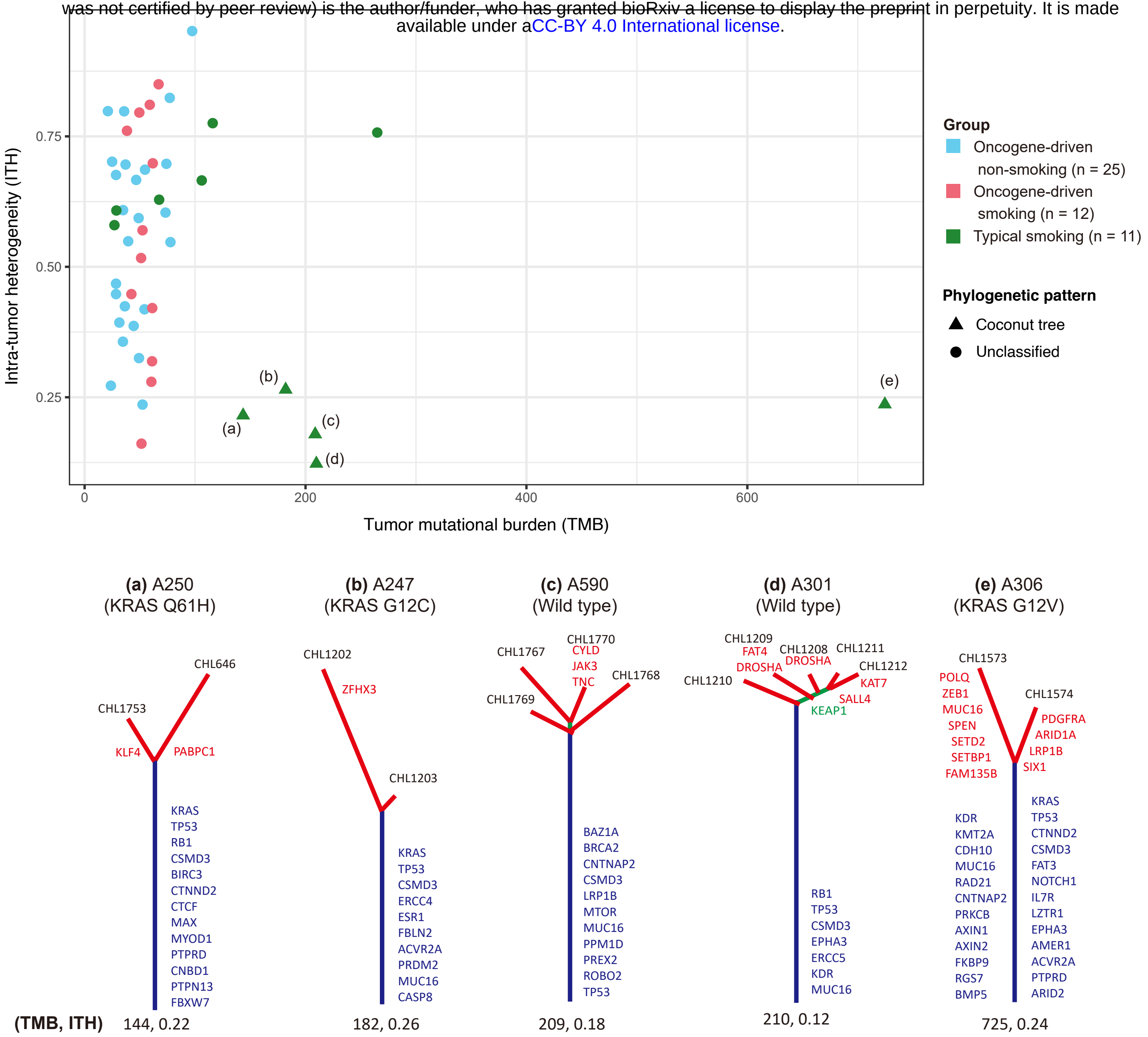

Intra-tumor heterogeneity (ITH)
0.75
0.50
0.25
0
200
400
600
Tumor mutational burden (TMB)
(a)
(b)
(c)
(d)
(e)
Group
🟦 Oncogene-driven non-smoking (n = 25)
🟥 Oncogene-driven smoking (n = 12)
🟩 Typical smoking (n = 11)
Phylogenetic pattern
▲ Coconut tree
● Unclassified
(a) A250
(KRAS Q61H)
CHL1753
CHL646
KLF4
PABPC1
KRAS
TP53
RB1
CSMD3
BIRC3
CTNND2
CTCF
MAX
MYOD1
PTPRD
CNBD1
PTPN13
FBXW7
(TMB, ITH)
144, 0.22
(b) A247
(KRAS G12C)
CHL1202
ZFHX3
CHL1203
KRAS
TP53
CSMD3
ERCC4
ESR1
FBLN2
ACVR2A
PRDM2
MUC16
CASP8
182, 0.26
(c) A590
(Wild type)
CHL1767
CHL1770
CYLD
JAK3
TNC
CHL1768
CHL1769
BAZ1A
BRCA2
CNTNAP2
CSMD3
LRP1B
MTOR
MUC16
PPM1D
PREX2
ROBO2
TP53
209, 0.18
(d) A301
(Wild type)
CHL1209
FAT4
DROSHA
CHL1208
DROSHA
CHL1211
CHL1212
KAT7
CHL1210
SALL4
KEAP1
RB1
TP53
CSMD3
EPHA3
ERCC5
KDR
MUC16
210, 0.12
(e) A306
(KRAS G12V)
CHL1573
POLQ
ZEB1
MUC16
SPEN
SETD2
SETBP1
FAM135B
CHL1574
PDGFRA
ARID1A
LRP1B
SIX1
KDR
KMT2A
CDH10
MUC16
RAD21
CNTNAP2
PRKCB
AXIN1
AXIN2
FKBP9
RGS7
BMP5
KRAS
TP53
CTNND2
CSMD3
FAT3
NOTCH1
IL7R
LZTR1
EPHA3
AMER1
ACVR2A
PTPRD
ARID2
725, 0.24

Figure 3

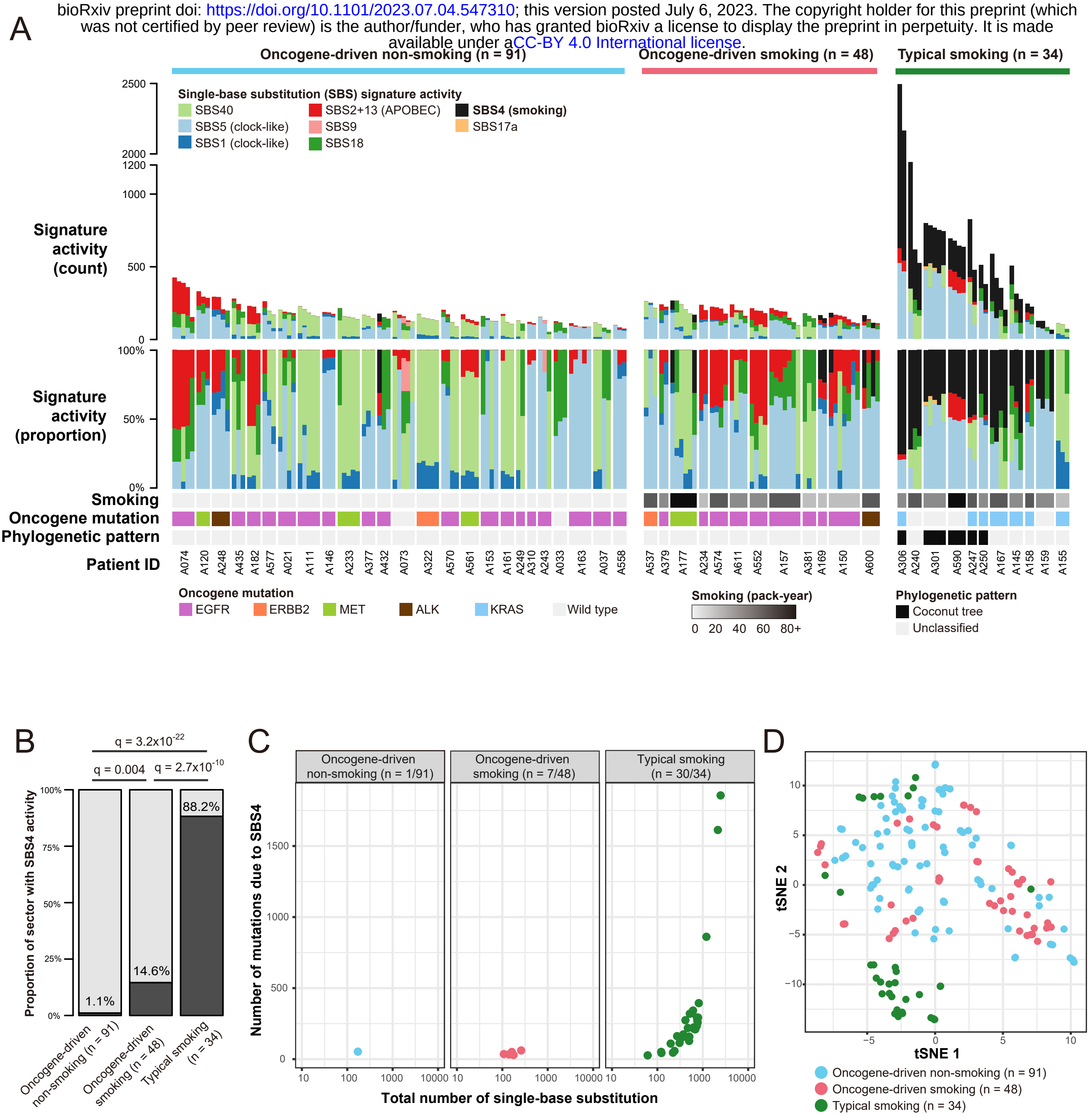
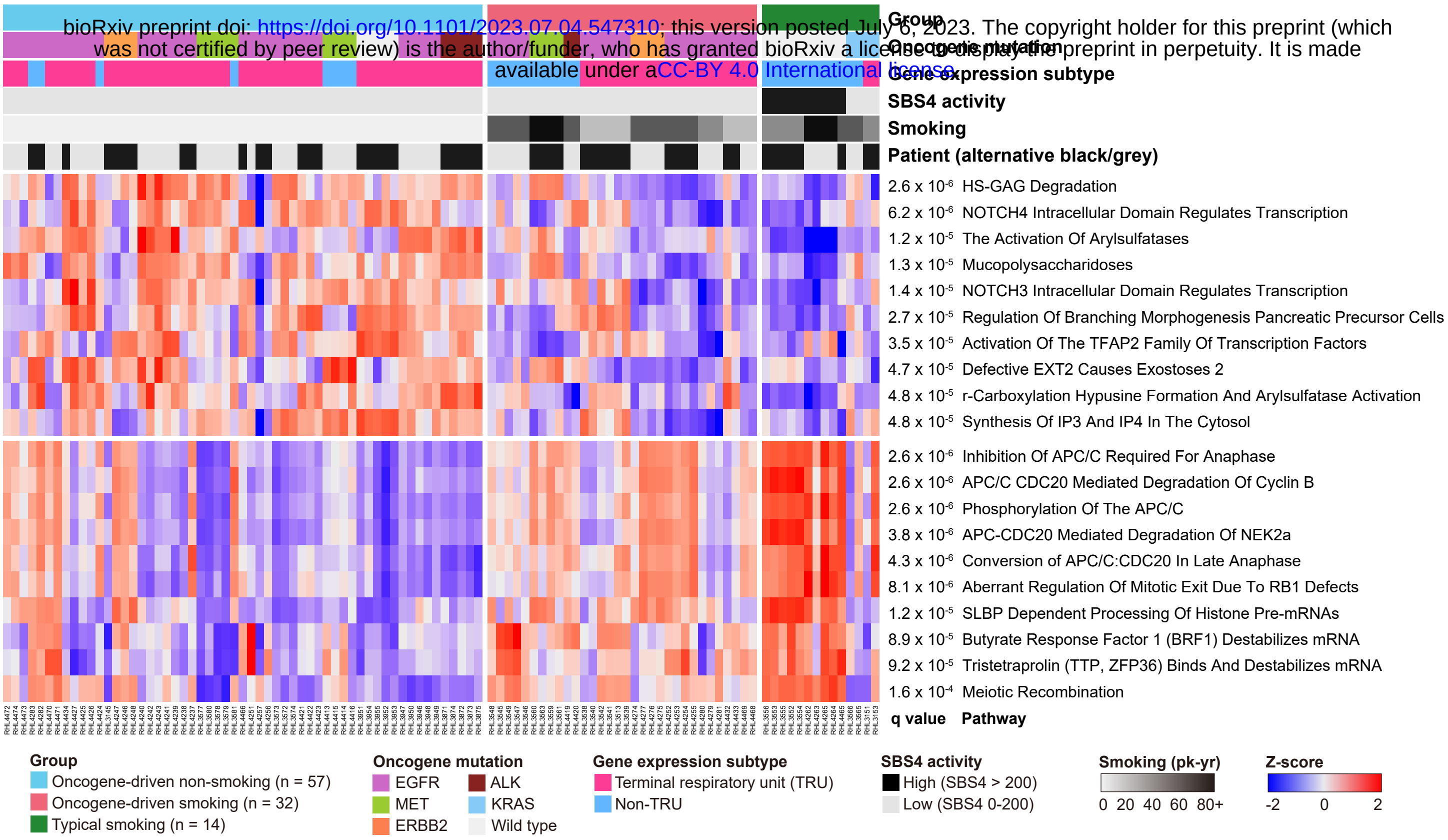

A
Oncogene-driven non-smoking (n = 91)
Oncogene-driven smoking (n = 48)
Typical smoking (n = 34)
Single-base substitution (SBS) signature activity
SBS40
SBS5 (clock-like)
SBS1 (clock-like)
SBS2+13 (APOBEC)
SBS9
SBS18
SBS4 (smoking)
SBS17a
Signature activity (count)
2500
2000
1200
1000
500
0
Signature activity (proportion)
100%
50%
0%
Smoking
Oncogene mutation
Phylogenetic pattern
Patient ID
A074
A120
A248
A435
A182
A577
A021
A111
A146
A233
A377
A432
A073
A322
A570
A561
A153
A161
A249
A310
A243
A033
A163
A037
A558
A537
A379
A177
A234
A574
A611
A552
A157
A381
A169
A150
A600
A306
A240
A301
A590
A247
A250
A167
A145
A158
A159
A155
Oncogene mutation
EGFR
ERBB2
MET
ALK
KRAS
Wild type
Smoking (pack-year)
0
20
40
60
80+
Phylogenetic pattern
Coconut tree
Unclassified
B
q = 3.2x10^-22
q = 0.004
q = 2.7x10^-10
Proportion of sector with SBS4 activity
100%
75%
50%
25%
0%
1.1%
14.6%
88.2%
Oncogene-driven non-smoking (n = 91)
Oncogene-driven smoking (n = 48)
Typical smoking (n = 34)
C
Oncogene-driven non-smoking (n = 1/91)
Oncogene-driven smoking (n = 7/48)
Typical smoking (n = 30/34)
Number of mutations due to SBS4
1500
1000
500
0
10
100
1000
10000
Total number of single-base substitution
D
10
5
0
−5
−10
tSNE 2
−5
0
5
10
tSNE 1
Oncogene-driven non-smoking (n = 91)
Oncogene-driven smoking (n = 48)
Typical smoking (n = 34)

Figure 4


Group
Oncogenic mutation
Gene expression subtype
SBS4 activity
Smoking
Patient (alternative black/grey)
2.6×10^{-6} HS-GAG Degradation
6.2×10^{-6} NOTCH4 Intracellular Domain Regulates Transcription
1.2×10^{-5} The Activation Of Arylsulfatases
1.3×10^{-5} Mucopolysaccharidoses
1.4×10^{-5} NOTCH3 Intracellular Domain Regulates Transcription
2.7×10^{-5} Regulation Of Branching Morphogenesis Pancreatic Precursor Cells
3.5×10^{-5} Activation Of The TFAP2 Family Of Transcription Factors
4.7×10^{-5} Defective EXT2 Causes Exostoses 2
4.8×10^{-5} r-Carboxylation Hypusine Formation And Arylsulfatase Activation
4.8×10^{-5} Synthesis Of IP3 And IP4 In The Cytosol
2.6×10^{-6} Inhibition Of APC/C Required For Anaphase
2.6×10^{-6} APC/C CDC20 Mediated Degradation Of Cyclin B
2.6×10^{-6} Phosphorylation Of The APC/C
3.8×10^{-6} APC-CDC20 Mediated Degradation Of NEK2a
4.3×10^{-6} Conversion of APC/C:CDC20 In Late Anaphase
8.1×10^{-6} Aberrant Regulation Of Mitotic Exit Due To RB1 Defects
1.2×10^{-5} SLBP Dependent Processing Of Histone Pre-mRNAs
8.9×10^{-5} Butyrate Response Factor 1 (BRF1) Destabilizes mRNA
9.2×10^{-5} Tristetraprolin (TTP, ZFP36) Binds And Destabilizes mRNA
1.6×10^{-4} Meiotic Recombination
q value
Pathway
Group
Oncogene-driven non-smoking (n = 57)
Oncogene-driven smoking (n = 32)
Typical smoking (n = 14)
Oncogene mutation
EGFR
ALK
MET
KRAS
ERBB2
Wild type
Gene expression subtype
Terminal respiratory unit (TRU)
Non-TRU
SBS4 activity
High (SBS4 > 200)
Low (SBS4 0-200)
Smoking (pk-yr)
0 20 40 60 80+
Z-score
-2 0 2