bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Title: Dearth of smoking-induced mutations in NSRO-driven non-small-cell lung cancer despite smoking exposure

Running title: Dearth of smoking mutations in NSRO-driven NSCLC

Authors:
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 168583, 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 169608, 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[14]Cancer Therapeutics Research Laboratory, Division of Medical Sciences, National Cancer Centre Singapore, Singapore 168583, 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

Daniel S.W. Tan, MD, PhD
Division of Medical Oncology, National Cancer Centre Singapore, Singapore
30 Hospital Boulevard, Singapore 168583
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

Disclosure of Potential Conflicts of Interest

G. Lai reports receiving personal fees from AstraZeneca and grants from Merck, AstraZeneca, Pfizer, Bristol Myers Squibb, Amgen, and Roche outside the submitted work and sponsorship from DKSH. A.C. Tan 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. S.P. Saw reports receiving personal fees from MSD, consultation fees from Pfizer, and Bayer, and grants from AstraZeneca, and Guardant Health. R. Kanesvaran reports receiving honoraria and consultation fees from MSD, Bristol Myers Squibb, Astellas, Novartis, Pfizer, Merck, Johnson & Johnson, and AstraZeneca. D.W.T. Lim reports receiving grant support from AstraZeneca, honoraria from Novartis, Merck, Amgen, personal fees from AstraZeneca, Ipsen, Boehringer Ingelheim, Bristol Myers Squibb, and DKSH. B.H. Ong reports receiving personal fees from AstraZeneca, Medtronic, Stryker, and MSD. All remaining authors have declared no conflicts of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Non-small cell lung cancers (NSCLCs) in non-smokers are mostly driven by mutations in the oncogenes *EGFR, ERBB2,* and *MET,* and fusions involving *ALK* and *RET*. We term these "non-smoking-related oncogenes" (NSROs). In addition to occurring in non-smokers, NSRO-driven tumors also occur in smokers, and the clonal architecture and genomic landscape of these tumors remain unknown. We investigated genomic and transcriptomic alterations in 173 tumor sectors from 48 patients with NSRO-driven or typical-smoking NSCLCs. NSRO-driven NSCLCs in smokers and non-smokers have similar genomic landscapes. Surprisingly, even in patients with prominent smoking histories, the mutational signature caused by tobacco smoking was essentially absent in NSRO-driven NSCLCs. However, NSRO-driven NSCLCs in smokers had higher transcriptomic activities related to regulation of the cell cycle, suggesting that smoking still affects tumor phenotype independently of genomic alterations.

Statement of significance

This study highlights the lack of genomic scars caused by smoking in NSCLCs driven by non-smoking-related oncogenes regardless of smoking history. The impact of smoking on these tumors is mainly non-genomic. The transcriptomic features of NSCLCs associated with smoking may help in the development of therapeutic approaches.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 among non-smokers is increasing, has become an emerging health problem, and accounts for more than half of cases in Taiwan and Singapore (3, 4). Most of these are non-small-cell lung cancers (NSCLCs) driven by a specific set of oncogenic mutations, usually activating mutations in the oncogenes *EGFR*, *ERBB2*, or *MET*, or rearrangements involving *ALK* or *RET* (5-9). Here, we refer to these genes as "non-smoking-related oncogenes" (NSROs). Including both smokers and non-smokers, NSRO-driven NSCLCs constitute approximately half of all East-Asian lung cancer (10-13). These tumors tend to have similar clinical trajectories and favorable responses to tyrosine kinase inhibitors (TKIs) (14-21). In contrast, tumors that occur in smokers and that have activating mutations in the *KRAS* or *BRAF* genes or that lack mutations in known oncogenes altogether constitute a group that we term "typical-smoking NSCLCs". Typical-smoking NSCLCs are more common in populations of European descent than in East-Asian populations. They often have high mutational burdens, sometimes respond well to immune checkpoint inhibitors, and have many somatic mutations caused by tobacco smoking (22, 23). In contrast, NSRO-driven NSCLCs tend to have lower mutational burdens and complex genomic architectures (24, 25).

While NSRO-driven NSCLCs have been most studied in non-smokers, in East Asia, approximately 20% to 40% of patients with these cancers have histories of tobacco smoking (4, 5, 26-28). Furthermore, among patients with *EGFR*-mutated NSCLC treated with TKIs, smokers have worse prognoses than non-smokers (29-31). However, the impact of tobacco smoking on clonal architecture, somatic genomic alterations, and transcriptomic phenotypes in NSRO-driven NSCLCs remains largely unknown. By means of an integrated study of genomic and transcriptomic landscapes, clonal architecture, and intra-tumor heterogeneity, we investigated similarities and differences between (i) NSRO-driven NSCLCs in non-smokers (ii) NSRO-driven NSCLCs in smokers and (iii) typical-smoking NSCLCs.

Results

Characteristics of the study cohort

We performed multi-region exome sequencing in a total of 173 sectors from 48 NSCLCs with clinical and histopathological characteristics shown in Supplementary Tables S1 and S2. Overall, we identified 6,251 single-nucleotide variants (SNVs) and

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

314 small indels (insertions or deletions) affecting the exons of 4,738 genes and the splicing junctions of 177 genes. We also performed RNA sequencing on 103 of the 173 sectors from 32 out of the 48 tumors (Supplementary Tables S1, S2).

Consistent with previous studies in East-Asian NSCLC, *EGFR* and *TP53* were the most frequently mutated genes (56%, 27 of 48, and 46%, 22 of 48, respectively) (9, 10). In addition to the SNVs and indels, we also studied gene fusions using RNA sequencing data. Of the putative transcript fusions detected, three were known oncogenic variants: *EML4-ALK*, *KLC1-ALK*, and *PARG-BMS1* (Supplementary Table S3), and the two *ALK* fusions were confirmed by fluorescence in situ hybridization (32-34). As expected, mutations in *EGFR*, *MET*, *ERBB2*, and *KRAS* and *ALK* fusions did not co-occur, suggesting that these were key initiating events (35).

All *EGFR*, *ERBB2*, and *MET* mutations were truncal (occurred in every sector) and were clonal in every sector (Supplementary Table S4). Compared to mutations in non-driver genes, mutations in *EGFR* were statistically more likely to be truncal (Supplementary Table S5). These findings underscore the impact of these oncogene mutations on the clonal architecture of NSCLCs. In addition, mutations in *KRAS* were truncal in every tumor, but were subclonal in single sectors in 2 out of 7 of the tumors. A similar pattern was reported in another multi-regional study (36).

Three categories of NSCLC and their genomic characteristics

For the purposes of this study, we first defined NSRO-driven tumor as those with any of the mutations or other genetic alterations detailed in Supplementary Table S6, based on the incidences of these alterations reported in the literature (27, 28). The NSRO mutations observed in this study were activating mutations in *EGFR* exons 18 through 21, *ALK* fusions, insertions in *ERBB2* exon 20, and skipping of *MET* exon 14.

We then focused on the following three groups: (i) NSRO-driven tumors in non-smokers (n = 23, 48%), (ii) NSRO-driven tumors in smokers (n = 12, 25%), and (iii) typical-smoking tumors (n = 11, 23%). In addition, there were 2 tumors without NSRO-mutations in non-smokers. This is typical of East-Asian populations, in which tumors without NSROs are rare in non-smokers (7, 8, 28). The proportion of non-smokers without NSROs in this study (8%, 2 out of 25) was lower than those reported in population of European descent (e.g., 51%, 96 out of 189, $p = 0.05$ by Fisher's two-sided exact test) (37). Fig. 1A shows the genomic landscape of the three groups.

Table 1 summarizes clinical characteristics and NSRO mutations in the three groups. Notably, smoking exposure was similar between NSRO-driven tumors in

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

smokers (median 34.5 pack-years, range 0.5-99) and typical-smoking tumors (median 38 pack-years, range 2-168, Wilcoxon rank-sum test, p value = 0.5792).

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

Overall, NSRO-driven tumors in smokers and non-smokers had similar genomic architectures, including tumor mutational burdens (TMB), number of truncal mutations, and number of mutations in driver genes. In contrast, compared to NSRO-driven tumors in smokers, typical-smoking tumors had much higher TMBs (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 all driver genes (including NSROs, median 14 vs. 7.5, p = 0.002, Fig. 1B, Supplementary Data and Code). Intra-tumoral heterogeneity (ITH, defined as the mean ratio of the numbers of branch mutations to the total number of mutations, see Methods) was similar across the three groups (medians 0.549, 0.543, and 0.580, for NSRO-driven non-smokers, NSRO-driven NSCLCs in smokers, and typical-smoking NSCLCs, respectively, Fig. 1B). However, “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, Fig. 2). Supplementary Fig. S1 details phylogenetic trees of all tumors across the three groups.

In addition to the oncogenes used to categorize the three groups of tumors, *CSMD3* mutations were statistically more common in typical-smoking tumors (Fig. 1C, left and middle). In comparing NSRO-driven NSCLCs in smokers versus NSRO-driven NSCLCs in non-smokers, there was no significant difference in the prevalence of mutations in COSMIC driver genes (including NSROs, Fig. 1C, right, https://cancer.sanger.ac.uk/census).

Previous studies reported that whole-genome doubling and chromosomal instability are common features of NSCLC (24, 25, 38, 39). In the present study, we found no significant difference across the three groups in tumor ploidy, proportions of tumors with whole-genome doubling, and chromosomal instability (Supplementary Fig. S2, Supplementary Table S2). Supplementary Fig. S3 provides details of somatic copy number alterations for all groups.

We also noted that gender distribution differed strongly across the three groups. Among non-smokers with NSRO-driven tumors, only 30% (7 of 23) were male, whereas among the NSRO-driven smoking and typical-smoking groups, 92% and 100% were male, respectively (p = 0.0024 and 0.0001 by two-sided Fisher’s exact tests compared to the NSRO-driven non-smoking group). We analyzed genomic

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

landscapes in NSRO-driven tumors by gender and found no significant differences (Supplementary Fig. S4).

Mutational signatures of NSRO-driven NSCLCs in smokers

Next, we investigated the impact of smoking on the mutational landscape across three groups. Overall, the median number of single-base substitution (SBS) per sector was 173 (range 47 to 2472). We used a signature presence test followed by signature attribution for each tumor sector to detect the mutational signature SBS4, which is caused by tobacco smoking in lung cancers (22, 40). Fig. 3A shows the activities of mutational signatures of the three groups.

We detected SBS4 in 30 of 34 (88%) sectors in typical-smoking tumors. Surprisingly, however, SBS4 was found in only 7 of 48 (15%) sectors in NSRO-driven tumors in smokers, significantly less than sectors in typical-smoking tumors despite similar exposures to tobacco smoking (two-sided Fisher's exact test, $p < 2.1\times10^{-11}$, Fig. 3D). For tumor sectors with SBS4 activity, the median number of mutations attributed to SBS4 was also significantly higher in typical-smoking tumors than in NSRO-driven tumors in smokers (216 vs. 53, two-sided Wilcoxon rank-sum test, $p < 9\times10^{-5}$, Fig. 3E). T-distributed stochastic neighbor embedding (tSNE) based on SBS spectra identified different mutational patterns in typical-smoking tumors compared to the other two groups (Fig. 3F, Supplementary Table S7 and S8).

To confirm the paucity of SBS4 activity in NSRO-driven NSCLCs, we applied the same tumor classification and signature assignment algorithm to two large, previously reported cohorts of NSCLC (10, 41). Supplementary Table S9 provides clinical information, including smoking history, oncogenes and their mutations, and signature activities for patients in these cohorts. The SBS4 signature was found in only 38% (8 of 21) and 29% (9 of 31) of NSRO-driven NSCLCs in smokers, significantly fewer than in typical-smoking tumors (90% and 78%, odds ratio of 0.07 and 0.12, two-sided Fisher's exact tests, p values of 1.1×10^{-7} and 4.5×10^{-6}, respectively, Supplementary Figs. S5, S6). Thus, all the genomic data indicates that NSRO-driven NSCLCs, whether in smokers or non-smokers, have origins and oncogenic histories distinct from those of typical-smoking NSCLCs.

Previous studies of NSCLC suggested that mutations caused by smoking and *APOBEC* activities dominate at different stages of cancer evolution (24, 42). For smoking mutagenesis, in the current study, SBS4 contributed similar activities in trunks and branches in typical-smoking tumors, suggesting ongoing exposure to tobacco smoke during cancer development (Supplementary Fig. S7). Analysis of SBS4 was not meaningful for the other two groups of NSCLCs, because they had

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

almost no SBS4 mutations. For *APOBEC*, consistent with previous studies, there were more mutations in branches than in trunks across the entire data set (Supplementary Fig. S7). Unexpectedly, we found that mutations due to reactive oxygen species (ROS, SBS18) were significantly higher in the branches than in the trunks for every group of tumors (all q values < 0.018 by two-sided Wilcoxon paired rank-sum tests with Benjamini-Hochberg correction, Supplementary Fig. S7). Thus, ROS might contribute to NSCLC evolution.

Transcriptomic features of NSRO-driven smoking tumors

The similarity of genomic landscapes between NSRO-driven tumors with and without smoking histories was surprising because clinical studies have shown that smoking indicates poor prognosis in NSRO-driven NSCLC (29-31). Therefore, we investigated whether transcriptomic features may account for this clinical observation. To this end, we profiled the transcriptomes of 103 of the 173 sectors from 32 out of the 48 tumors (Table 1 and Supplementary Table S2). UMAP dimension reduction did not reveal a separation between NSRO-driven NSCLCs in smokers versus non-smokers in the first 2 dimensions (Supplementary Fig. S8A). Indeed, the primary separation seems to reflect membership in the terminal-respiratory-unit (TRU) expression subtype. There is a trend for association of the TRU subtype with NSRO-driven tumors in both non-smokers and smokers compared to typical-smoking tumors (Supplementary Fig. S8B).

To further explore the transcriptomic activities associated with tobacco smoking, we investigated differential expression pathway activity between non-smokers (57 sectors from 18 tumors) and smokers (46 sectors from 14 tumors, including both typical-smoking tumors and NSRO-driven tumors in smokers). We examined activities of 1,259 pathways from the Reactome Database (43) (Fig. 4, Supplementary Fig. S9, Supplementary Table S10). In this comparison, tumors in non-smokers had higher activities in pathways related to NOTCH signaling and to metabolism and lower activities in pathways related to cell cycle regulation and mitotic exit (Fig. 4). These differences in pathway activity were also evident in a comparison of NSRO-driven tumors in smokers versus in non-smokers (i.e., after excluding typical-smoking tumors and NSRO-negative tumors, Supplementary Table S10). These observations underscore the phenotypic differences associated with tobacco smoking in some transcriptomic pathways independent of genomic alterations.

It has been proposed that smoking-associated lung cancers are associated with the immune repertoires of tumor microenvironments (TME) that would confer better responses to immunotherapy (44-47). TMEs with infiltration of cytotoxic T cells and

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression of proinflammatory cytokine genes are often called “immune-hot”, and have higher tendencies to respond to immune checkpoint inhibitors (48, 49). In this study, we investigated the TME of NSRO-driven NSCLCs in smokers versus the other two groups. To define immune-hot TME, we performed hierarchical clustering of sectors based on the transcript levels of T-cell inflammation and immune checkpoint genes (Supplementary Fig. S10A) (50-52). However, there was no strong evidence for enrichment for immune-hot TMEs in NSRO-driven NSCLCs in smokers (3 out of 9 tumors, 33%) and typical-smoking-related NSCLCs (1 out of 5 tumors, 20%) compared to NSRO-driven tumors in non-smokers (5 out of 17 tumors, 29%, $p = 0.7332$ by Fisher’s exact test across the three groups, Supplementary Fig. S10B).

Discussion

To our knowledge, this is the first integrated genomic study to directly compare NSRO-driven NSCLCs in non-smokers, NSRO-driven NSCLCs in smokers, and typical-smoking NSCLCs using multi-region exome and RNA sequencing. We found that tobacco smoking had almost no influence on the genomic features and clonal architectures of *EGFR*-mutated and other NSRO-driven NSCLCs. Despite prominent smoking histories, NSRO-driven tumors in smokers were similar to those in non-smokers in terms of mutational burden, intra-tumor heterogeneity, and mutational signature activity. In contrast, compared to both NSRO-driven groups, typical-smoking tumors showed higher TMBs and more mutations in driver genes. 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 NSCLCs but were absent from NSRO-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 than 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, 53, 54). Previously, NSRO-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 NSRO-driven tumors were particularly noticeable among women, since they were usually non-smokers. The current study confirms that NSRO-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 (Supplementary Fig. S4). 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

gender from the effects of smoking. Nevertheless, we note that available evidence suggests that NSRO-driven tumors are more common among women. Notably, in data from a previous study (55), in both non-smokers and smokers, *EGFR*-mutated NSCLC was 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.

While it has been recently reported that NSCLCs that occur in smokers but that lack SBS4 are enriched for mutations in *EGFR* and other NSROs (36, 56), here we have shown that the paucity of mutations due to tobacco smoking is a nearly-universal characteristic of NSRO-driven tumors in smokers. We further confirmed this finding in two large cohorts of patients with lung adenocarcinomas (10, 41). It is unclear why NSRO-driven tumors rarely acquire mutations caused by smoking, while typical-smoking tumors with similar exposures have abundant smoking mutations. Possibly, as suggested by some studies, the cell of origin of NSRO-driven NSCLC may be different from that of typical-smoking NSCLC (57-61). Thus, one possibility is that NSRO-driven 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 NSRO-driven tumors in non-smokers and in smokers have similar clonal architectures and genomic features, they differ in their transcriptomic pathway activities, especially those related to the cell cycle and mitotic exit. Consistent with previous studies (62, 63), for these pathway activities, NSRO-driven tumors in smokers are more similar to typical-smoking tumors than to NSRO-driven tumors in non-smokers (Fig. 4, Supplementary Fig. S9). Despite the lack of somatic mutations caused by smoking, it is still possible that chronic tobacco exposure causes epigenomic changes to bronchial epithelial cells, leading to a carcinogenic phenotype that is independent of genomic alterations (64). Of note, advanced *EGFR*-mutated NSCLCs treated with TKIs had worse outcomes in smokers than in non-smokers (30, 31). The transcriptomic activities of NSCLC driven by NSROs, including *EGFR*, in smokers, might account for these cancers' higher resistance to standard TKIs and suggests the possibility of better responses to therapies such as chemotherapy or CDK4/6 inhibitors that target the cell cycle (65). This warrants further investigation regarding treatment selection for patients with NSRO-driven NSCLC.

In summary, based on multi-region whole-exome and RNA sequencing data, we have elucidated the clonal architectures and genomic features of three groups of East-Asian NSCLC: NSRO-driven in non-smokers, NSRO-driven in smokers, and typical-smoking tumors. We found no evidence that tobacco smoking affects clonal architecture or patterns of genomic alteration in NSRO-driven NSCLC. However, some transcriptomic pathway activities were more similar between NSRO-driven

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tumors in smokers and typical-smoking tumors than between NSRO-driven cancers in smokers and non-smokers. The in-depth analysis of NSRO-driven NSCLC in smokers and non-smokers presented here may provide guidance for optimizing treatment approaches.

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 obtained from the Lung Cancer Consortium Singapore. Written informed consent was obtained from all participants. This study was approved by the SingHealth Centralized Institutional Review Board (CIRB reference 2018/2963).

Tumor/normal sample processing and whole-exome sequencing

Resected tumors and paired normal samples were sectioned and processed as previously described (24). 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 S2).

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 (66). 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 (67). Duplicate reads were marked and removed from variant calling using Sambamba (version 0.7.0) (68). Global mapping quality was evaluated by Qualimap 2 (version

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.2.1, Supplementary Table S2) (69). 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 (70, 71). 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. Variants were annotated by wANNOVAR (https://wannovar.wglab.org/) (72). Driver status of genes was based on the Catalog of Somatic Mutations in Cancer (COSMIC) database, downloaded February 24, 2021 (https://cancer.sanger.ac.uk/census) (73).

We excluded 12 out of 185 sectors (6.5%) that had tumor purity < 0.1 as assessed by HATCHet (74) (Supplementary Table S2). In tumor sectors with low tumor purity, fewer variants were called than in other sectors of the same tumor, supporting the estimation of low tumor purity (Supplementary Fig. S11).

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. The clonality of somatic variants was evaluated by using MutationTimeR R package (75).

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 (76).

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) (77). For mutational signature analysis, we used all single-base substitution (SBS), including exonic and non-exonic variants. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (40). 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 (40), except that SBS4 was included only if the false discovery rate based on the SignaturePresenceTest was < 0.05. We excluded SBS3 (caused by defective homologous recombination DNA damage repair mechanism) from sparse assignment after ensuring no pathogenic germline or somatic alterations in the *BRCA1* or *BRCA2* genes. Supplementary Table S7 and S8 show the SBS mutational spectra and signature activities of each sector. We did not assign activities for indels or double-base substitutions because there were too few of these mutations (median exome-wide counts of 10 and 1, respectively).

Detection of fusion transcripts

We used STAR-Fusion (version 1.10.0) (78) 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 $\geq$ 1MB between fusion breakpoints

Supplementary Table S3 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 (79). We used STAR (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) (80). Only the counts of protein-coding genes were included for downstream analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transcriptomic pathway analysis

Raw gene expression levels were transformed to transcript levels in transcripts per million (TPM) values (81). 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) (43, 82, 83). Differential pathway expression was assessed using the limma R package (version 3.48.0) (84). 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 (85). 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) (86). 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. Immune cell deconvolution was performed by using the CIBERSORT web application (https://cibersortx.stanford.edu/) (87).

Data Availability

All WES and RNA sequencing data were deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), under the accession number EGAS00001006942. Supplementary Data and Code, including lists of somatic mutations, mutation-timing and clonality output from the MutationTimeR package, gene fusions, gene expression matrices, transcriptomic pathway activities, and immune cell deconvolution, are publicly available at https://github.com/Rozen-Lab/oncogene-NSCLC/.

Authors' Contributions

C.-Y. Huang: Data curation, formal analysis, visualization, methodology, writing original drafts, reviewing, and editing. **N. Jiang**: Formal analysis, methodology, writing, reviewing, and editing. **M. Shen**: Project coordination. **G. Lai**, **A.C. Tan**, **A. Jain**, **S.P. Saw**, **M.K. Ang**, **Q.S. Ng**, **D.W.T. Lim**, **R. Kanesvaran**, **E.H. Tan**, **W.L. Tan**, **B.H. Ong**, **K.L. Chua**, and **D. Anantham**: Clinical data curation and patient sample collection. **A. Takano** and **T.K.H. Lim**: Data curation, investigation, and methodology. **W.L. Tam**: Data curation and resources. **N.L. Sim**: Data curation, investigation, and resources. **A.J. Skanderup**: Conceptualization, data curation,

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

resources, methodology, funding acquisition, writing, reviewing, and editing. **D.S.W. Tan**: Conceptualization, data curation, patient sample collection, resources, methodology, funding acquisition, supervision, writing, reviewing, and editing. **S.G. Rozen**: Conceptualization, supervision, resources, methodology, writing, reviewing, and editing.

Acknowledgements

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 funded in part by grants from the Chang Gung Memorial Hospital (CGMH) Foundation (Grant No. CMRPG3F1911) to Chen-Yang Huang.

We thank Jacob J.S. Alvarez, and 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 the 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, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.
2. Reitsma MB, Flor LS, Mullany EC, Gupta V, Hay SI, Gakidou E. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and initiation among young people in 204 countries and territories, 1990-2019. Lancet Public Health. 2021;6(7):e472-e81.
3. Tseng CH, Tsuang BJ, Chiang CJ, Ku KC, Tseng JS, Yang TY, et al. The Relationship Between Air Pollution and Lung Cancer in Nonsmokers in Taiwan. J Thorac Oncol. 2019;14(5):784-92.
4. Toh CK, Ong WS, Lim WT, Tan DS, Ng QS, Kanesvaran R, et al. A Decade of Never-smokers Among Lung Cancer Patients-Increasing Trend and Improved Survival. Clin Lung Cancer. 2018;19(5):e539-e50.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5. Cho J, Choi SM, Lee J, Lee CH, Lee SM, Kim DW, et al. Proportion and clinical features of never-smokers with non-small cell lung cancer. Chin J Cancer. 2017;36(1):20.
6. Zhang Y, Sun Y, Pan Y, Li C, Shen L, Li 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. 2012;18(7):1947-53.
7. Li C, Fang R, Sun Y, Han X, Li F, Gao B, et al. Spectrum of oncogenic driver mutations in lung adenocarcinomas from East Asian never smokers. PLoS One. 2011;6(11):e28204.
8. Sun Y, Ren Y, Fang Z, Li C, Fang R, Gao B, et al. Lung adenocarcinoma from East Asian never-smokers is a disease largely defined by targetable oncogenic mutant kinases. J Clin Oncol. 2010;28(30):4616-20.
9. Chen YJ, Roumeliotis TI, Chang YH, Chen CT, Han CL, Lin MH, et al. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell. 2020;182(1):226-44 e17.
10. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, et al. Genomic landscape of lung adenocarcinoma in East Asians. Nat Genet. 2020;52(2):177-86.
11. Hsu K-H, Ho C-C, Hsia T-C, Tseng J-S, Su K-Y, Wu M-F, et al. Identification of five driver gene mutations in patients with treatment-naive lung adenocarcinoma in Taiwan. Plos one. 2015;10(3):e0120852.
12. Melosky B, Kambartel K, Häntschel M, Bennetts M, Nickens DJ, Brinkmann J, et al. Worldwide prevalence of epidermal growth factor receptor mutations in non-small cell lung cancer: A meta-analysis. Molecular Diagnosis & Therapy. 2022;26(1):7-18.
13. Tan AC, Tan DSW. Targeted therapies for lung cancer patients With oncogenic driver molecular alterations. Journal of Clinical Oncology. 2022;40(6):611-25.
14. Li BT, Smit EF, Goto Y, Nakagawa K, Udagawa H, Mazieres J, et al. Trastuzumab Deruxtecan in HER2-Mutant Non-Small-Cell Lung Cancer. N Engl J Med. 2022;386(3):241-51.
15. Wolf J, Seto T, Han JY, Reguart N, Garon EB, Groen HJM, et al. Capmatinib in MET Exon 14-Mutated or MET-Amplified Non-Small-Cell Lung Cancer. N Engl J Med. 2020;383(10):944-57.
16. Paik PK, Felip E, Veillon R, Sakai H, Cortot AB, Garassino MC, et al. Tepotinib in Non-Small-Cell Lung Cancer with MET Exon 14 Skipping Mutations. N Engl J Med. 2020;383(10):931-43.
17. Drilon A, Oxnard GR, Tan DSW, Loong HHF, Johnson M, Gainor J, et al. Efficacy of Selpercatinib in RET Fusion-Positive Non-Small-Cell Lung Cancer. N Engl J Med. 2020;383(9):813-24.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. Wu YL, Zhou C, Hu CP, Feng J, Lu S, Huang Y, 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. 2014;15(2):213-22.
19. Solomon BJ, Mok T, Kim DW, Wu YL, Nakagawa K, Mekhail T, et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. N Engl J Med. 2014;371(23):2167-77.
20. Zhou C, Wu YL, Chen G, Feng J, Liu XQ, Wang 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. 2011;12(8):735-42.
21. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361(10):947-57.
22. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354(6312):618-22.
23. Dai L, Jin B, Liu T, Chen J, Li G, Dang J. 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. 2021;38:100990.
24. Nahar R, Zhai W, Zhang T, Takano A, Khng AJ, Lee YY, et al. Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing. Nat Commun. 2018;9(1):216.
25. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med. 2017;376(22):2109-21.
26. Tseng C-H, Chiang C-J, Tseng J-S, Yang T-Y, Hsu K-H, Chen K-C, et al. EGFR mutation, smoking, and gender in advanced lung adenocarcinoma. Oncotarget. 2017;8(58):98384-93.
27. Zheng D, Wang R, Ye T, Yu S, Hu H, Shen X, et al. MET exon 14 skipping defines a unique molecular class of non-small cell lung cancer. Oncotarget. 2016;7(27):41691-702.
28. Gou LY, Niu FY, Wu YL, Zhong WZ. Differences in driver genes between smoking-related and non-smoking-related lung cancer in the Chinese population. Cancer. 2015;121 Suppl 17:3069-79.
29. Chang JW, Huang CY, Fang YF, Chang CF, Yang CT, Kuo CS, et al. Risk Stratification Using a Novel Nomogram for 2190 EGFR-Mutant NSCLC Patients Receiving the First or Second Generation EGFR-TKI. Cancers (Basel). 2022;14(4).

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30. Kim IA, Lee JS, Kim HJ, Kim WS, Lee KY. Cumulative smoking dose affects the clinical outcomes of EGFR-mutated lung adenocarcinoma patients treated with EGFR-TKIs: a retrospective study. BMC Cancer. 2018;18(1):768.
31. Zhang Y, Kang S, Fang W, Hong S, Liang W, Yan 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. 2015;16(2):144-51 e1.
32. Ou SI, Zhu VW, Nagasaka M. Catalog of 5' Fusion Partners in ALK-positive NSCLC Circa 2020. JTO Clin Res Rep. 2020;1(1):100015.
33. Piscuoglio S, Ng CKY, Geyer FC, Burke KA, Cowell CF, Martelotto LG, et al. Genomic and transcriptomic heterogeneity in metaplastic carcinomas of the breast. NPJ Breast Cancer. 2017;3:48.
34. Gotoh M, Ichikawa H, Arai E, Chiku S, Sakamoto H, Fujimoto H, et al. Comprehensive exploration of novel chimeric transcripts in clear cell renal cell carcinomas using whole transcriptome analysis. Genes Chromosomes Cancer. 2014;53(12):1018-32.
35. Gainor JF, Varghese AM, Ou SH, Kabraji S, Awad MM, Katayama R, et al. ALK rearrangements are mutually exclusive with mutations in EGFR or KRAS: an analysis of 1,683 patients with non-small cell lung cancer. Clin Cancer Res. 2013;19(15):4273-81.
36. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023;616(7957):525-33.
37. Zhang T, Joubert P, Ansari-Pour N, Zhao W, Hoang PH, Lokanga R, et al. Genomic and evolutionary classification of lung cancer in never smokers. Nat Genet. 2021;53(9):1348-59.
38. Lopez S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M, et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat Genet. 2020;52(3):283-93.
39. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018;50(8):1189-95.
40. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101.
41. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543-50.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

42. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346(6206):251-6.
43. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics. 2017;18(1):142.
44. Desrichard A, Kuo F, Chowell D, Lee KW, Riaz N, Wong RJ, et al. Tobacco Smoking-Associated Alterations in the Immune Microenvironment of Squamous Cell Carcinomas. J Natl Cancer Inst. 2018;110(12):1386-92.
45. Sun Y, Yang Q, Shen J, Wei T, Shen W, Zhang N, 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. 2021;9:745859.
46. Bai R, Lv Z, Xu D, Cui J. Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. Biomark Res. 2020;8:34.
47. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer. 2019;19(3):133-50.
48. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541(7637):321-30.
49. Herbst RS, Soria JC, Kowanetz M, Fine GD, Hamid O, Gordon MS, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature. 2014;515(7528):563-7.
50. Liu S, Matsuzaki J, Wei L, Tsuji T, Battaglia S, Hu Q, et al. Efficient identification of neoantigen-specific T-cell responses in advanced human ovarian cancer. J Immunother Cancer. 2019;7(1):156.
51. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science. 2018;362(6411).
52. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest. 2017;127(8):2930-40.
53. Yang T, Barnett R, Jiang S, Yu L, Xian H, Ying J, et al. Gender balance and its impact on male and female smoking rates in Chinese cities. Soc Sci Med. 2016;154:9-17.
54. Tsai YW, Tsai TI, Yang CL, Kuo KN. Gender differences in smoking behaviors in an Asian population. J Womens Health (Larchmt). 2008;17(6):971-8.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

55. Tseng CH, Chiang CJ, Tseng JS, Yang TY, Hsu KH, Chen KC, et al. EGFR mutation, smoking, and gender in advanced lung adenocarcinoma. Oncotarget. 2017;8(58):98384-93.
56. Ernst SM, Mankor JM, van Riet J, von der Thüsen JH, Dubbink HJ, Aerts JGJV, et al. Tobacco Smoking-Related Mutational Signatures in Classifying Smoking-Associated and Nonsmoking-Associated NSCLC. Journal of Thoracic Oncology. 2023;18(4):487-98.
57. Chen F, Liu J, Flight RM, Naughton KJ, Lukyanchuk A, Edgin AR, et al. Cellular Origins of EGFR-Driven Lung Cancer Cells Determine Sensitivity to Therapy. Adv Sci (Weinh). 2021;8(22):e2101999.
58. Spella M, Lilis I, Pepe MA, Chen Y, Armaka M, Lamort AS, et al. Club cells form lung adenocarcinomas and maintain the alveoli of adult mice. Elife. 2019;8.
59. Hynds RE, Janes SM. Airway Basal Cell Heterogeneity and Lung Squamous Cell Carcinoma. Cancer Prev Res (Phila). 2017;10(9):491-3.
60. Kadur Lakshminarasimha Murthy P, Sontake V, Tata A, Kobayashi Y, Macadlo L, Okuda K, et al. Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. Nature. 2022;604(7904):111-9.
61. Basil MC, Cardenas-Diaz FL, Kathiriya JJ, Morley MP, Carl J, Brumwell AN, et al. Human distal airways contain a multipotent secretory cell that can regenerate alveoli. Nature. 2022;604(7904):120-6.
62. Wang J, Chen T, Yu X, N OU, Tan L, Jia B, et al. Identification and validation of smoking-related genes in lung adenocarcinoma using an in vitro carcinogenesis model and bioinformatics analysis. J Transl Med. 2020;18(1):313.
63. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One. 2008;3(2):e1651.
64. Vaz M, Hwang SY, Kagiampakis I, Phallen J, Patil A, O'Hagan HM, et al. Chronic Cigarette Smoke-Induced Epigenomic Changes Precede Sensitization of Bronchial Epithelial Cells to Single-Step Transformation by KRAS Mutations. Cancer Cell. 2017;32(3):360-76 e6.
65. Hosomi Y, Morita S, Sugawara S, Kato T, Fukuhara T, Gemma A, et al. Gefitinib Alone Versus Gefitinib Plus Chemotherapy for Non-Small-Cell Lung Cancer With Mutated Epidermal Growth Factor Receptor: NEJ009 Study. J Clin Oncol. 2020;38(2):115-23.
66. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
67. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

68. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032-4.
69. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.
70. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591-4.
71. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.
72. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet. 2012;49(7):433-6.
73. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696-705.
74. Zaccaria S, Raphael BJ. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. Nat Commun. 2020;11(1):4301.
75. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578(7793):122-8.
76. Wu P, Hou L, Zhang Y, Zhang L. Phylogenetic Tree Inference: A Top-Down Approach to Track Tumor Evolution. Front Genet. 2019;10:1371.
77. Ng AWT, Poon SL, Huang MN, Lim JQ, Boot A, Yu W, et al. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. Sci Transl Med. 2017;9(412).
78. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20(1):213.
79. Chua KP, Teng YHF, Tan AC, Takano A, Alvarez JJS, Nahar R, et al. Integrative Profiling of T790M-Negative EGFR-Mutated NSCLC Reveals Pervasive Lineage Transition and Therapeutic Opportunities. Clin Cancer Res. 2021;27(21):5939-50.
80. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
81. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281-5.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

82. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.
83. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
84. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
85. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One. 2012;7(5):e36530.
86. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.
87. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol. 2018;1711:243-59.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Baseline clinical and genomic characteristics of NSRO-driven NSCLCs in non-smokers, NSRO-driven NSCLCs in smokers, and typical-smoking NSCLCs.

Clinical or genomic characteristic, n (%)	NSRO-driven in non-smokers (n = 23)		NSRO-driven in smokers (n = 12)		Typical-smoking (n = 11)		P value
Number of patients	23	(48)	12	(25)	11	(23)	
Number of tumor sectors with WES[a]	83	(100)	48	(100)	34	(100)	0.8608
Number of tumors with RNA seq	17	(74)	9	(75)	5	(45)	0.2139
Number of tumor sectors with RNA seq	52	(63)	32	(67)	14	(41)	0.3392
Age, median (range)	67	(53-82)	70	(39-79)	66	(49-74)	n.s.[c]
Gender							
Male	7	(30)	11	(92)	11	(100)	<0.0001
Female	16	(70)	1	(8)	0	(0)	
Cigarette smoking status							
Never	23	(100)	0	(0)	0	(0)	<0.0001
Current/Former	0	(0)	12	(100)	11	(100)	
Pack years, median (range)	0	(0-0)	34.5	(0.5-99)	38	(2-168)	0.58[d]
Ethnicity							
Chinese	20	(87)	10	(83)	10	(91)	1
Non-Chinese	3	(13)	2	(17)	1	(9)	
Stage at diagnosis							
Early (I & II)	19	(83)	9	(75)	11	(100)	0.2719
Late (III & IV)	4	(17)	3	(25)	0	(0)	
Histology							
Adenocarcinoma	22	(96)	12	(100)	10	(91)	0.7333
Squamous cell carcinoma	1	(4)	0	(0)	1	(9)	
Key mutations Non-smoking related oncogenes							
EGFR exon 18-21	18	(78)	9	(75)	0	(0)	<0.0001
MET exon 14 skipping	3	(13)	1	(8)	0	(0)	
ALK fusion	1	(4)	1	(8)	0	(0)	
ERBB2 exon 20	1	(4)	1	(8)	0	(0)	
KRAS exon 2-3	0	(0)	0	(0)	7	(64)	
No *NSRO*, *KRAS*, or *BRAF* mutation	0	(0)	0	(0)	4	(36)	

[a]WES, whole exome sequencing.

[b]*P* value by two-sided Fisher's exact tests across all three group within category (e.g., Gender, Cigarette smoking, status, etc.).

[c]*P* values by two-sided Wilcoxon rank-sum test > 0.05 in all pairwise comparisons across the three groups.

[d]*P* values by two-sided Wilcoxon rank-sum test > 0.05 between NSRO-driven smoking and typical-smoking groups.

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

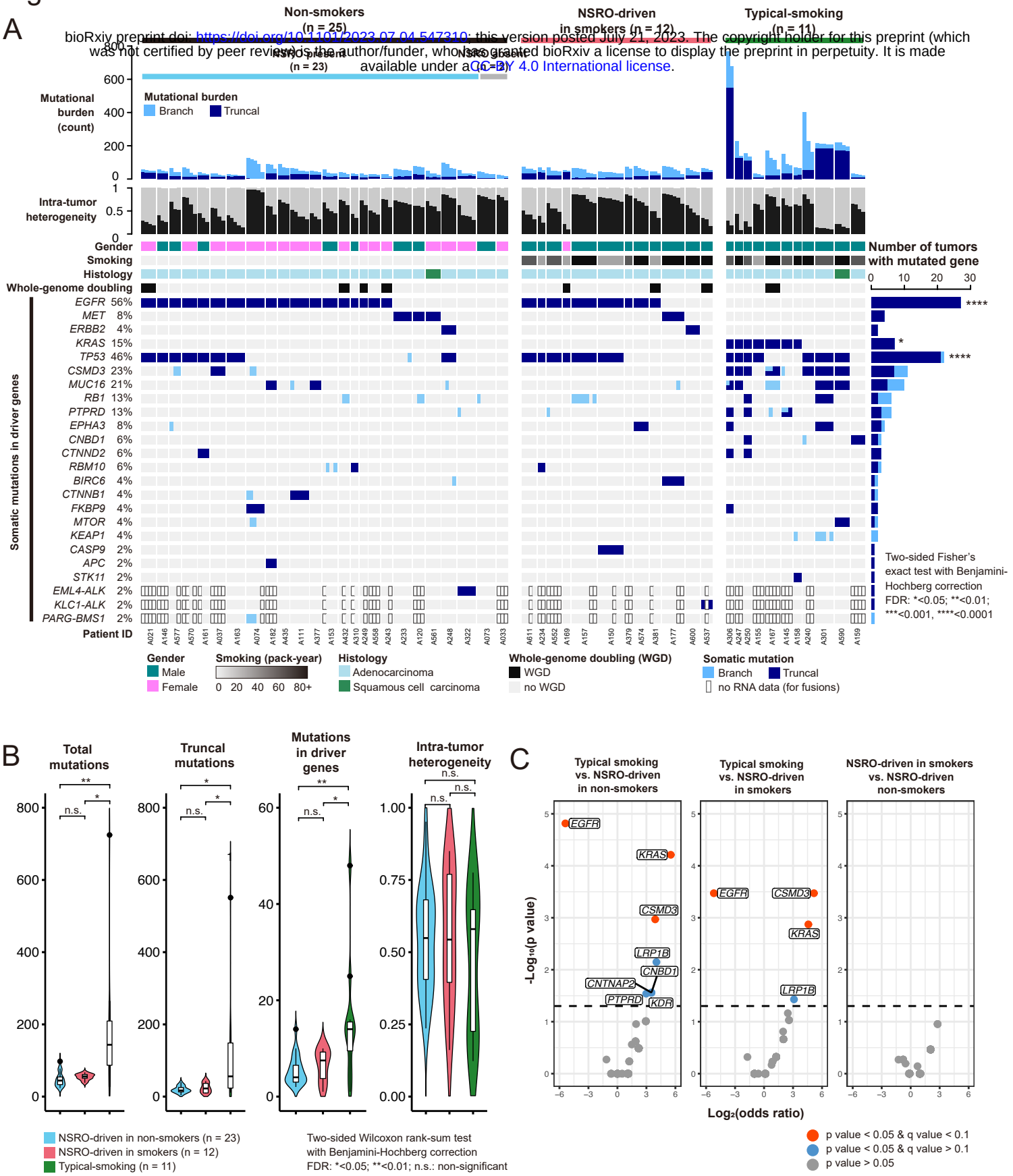
Fig. 1. (A) Overview of genomic alterations in tumors and tumor sectors. (B) Counts of total mutations, truncal mutations, numbers of mutations in driver genes, and levels of intra-tumor heterogeneity in the three groups. We used the list of driver genes from COSMIC (https://cancer.sanger.ac.uk/census). (C) Enrichment for mutations in driver genes in pair-wise comparisons among the three groups.

Fig. 2. (A) Intra-tumor heterogeneity (ITH) versus tumor mutation burden (TMB) for each tumor. Five tumors with "coconut-tree" phylogenies are labeled (a) through (e) and the corresponding phylogenies are in panel (B). These phylogenies occurred only in the typical-smoking-related group.

Fig. 3. Single-base substitution (SBS) mutational signatures. (A, B) Mutational-signature activities in the three groups by absolute mutation counts (A) and by proportions (B). Colors indicate various mutational signatures (e.g., SBS1, SBS5, etc.), as indicated by the legend above. (C) Smoking history, key mutated genes, and whether the tumor has a coconut tree pattern. (D) Proportions of tumor sectors with SBS4 (caused by tobacco smoking). (E) Counts of mutations due to SBS4 in tumor sectors that have SBS4 mutations. (F) tSNE (t-distributed stochastic neighbor embedding) dimension reduction based on the mutational spectra. For information on 10the mutational signatures, see COSMIC (https://cancer.sanger.ac.uk/signatures/).

Fig. 4. Heatmap of activities of the top 10 pathways up- and down-regulated in all non-smokers compared to all smokers. Each column is a tumor sector, and sectors are grouped by patient as shown in the row labeled "Patient". 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 (84). Supplementary Fig. S9 and Supplementary Tables S10 and S11 provide details.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1. (**A**) Overview of genomic alterations in tumors and tumor sectors. (**B**) Counts of total mutations, truncal mutations, numbers of mutations in driver genes, and levels of intra-tumor heterogeneity in the three groups. We used the list of driver genes from COSMIC (https://cancer.sanger.ac.uk/census). (**C**) Enrichment for mutations in driver genes in pair-wise comparisons among the three groups.

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

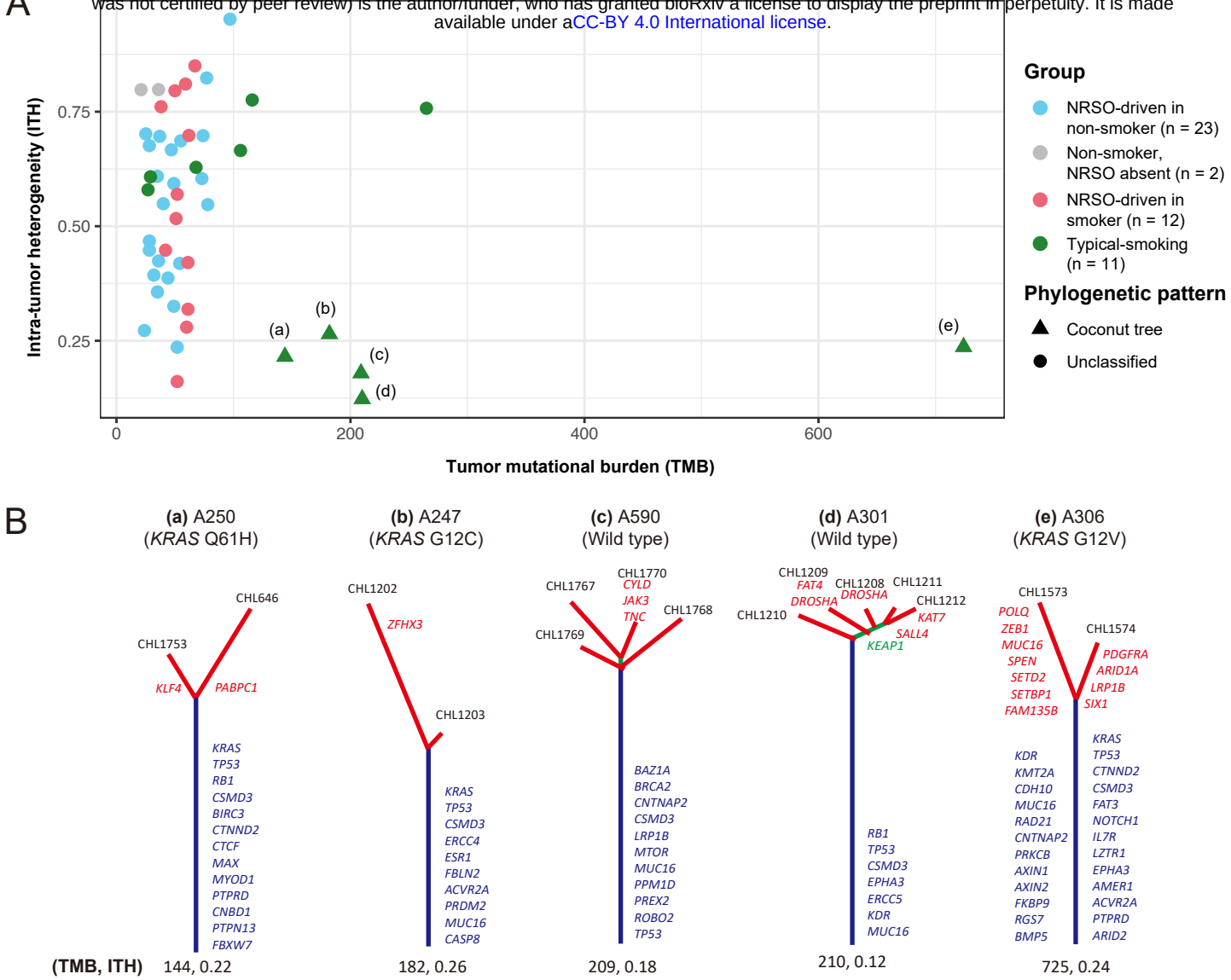


Figure 2. (**A**) Intra-tumor heterogeneity (ITH) versus tumor mutation burden (TMB) for each tumor. Five tumors with “coconut-tree” phylogenies are labeled (a) through (e) and the corresponding phylogenies are in panel (**B**). These phylogenies occurred only in the typical-smoking-related group.

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

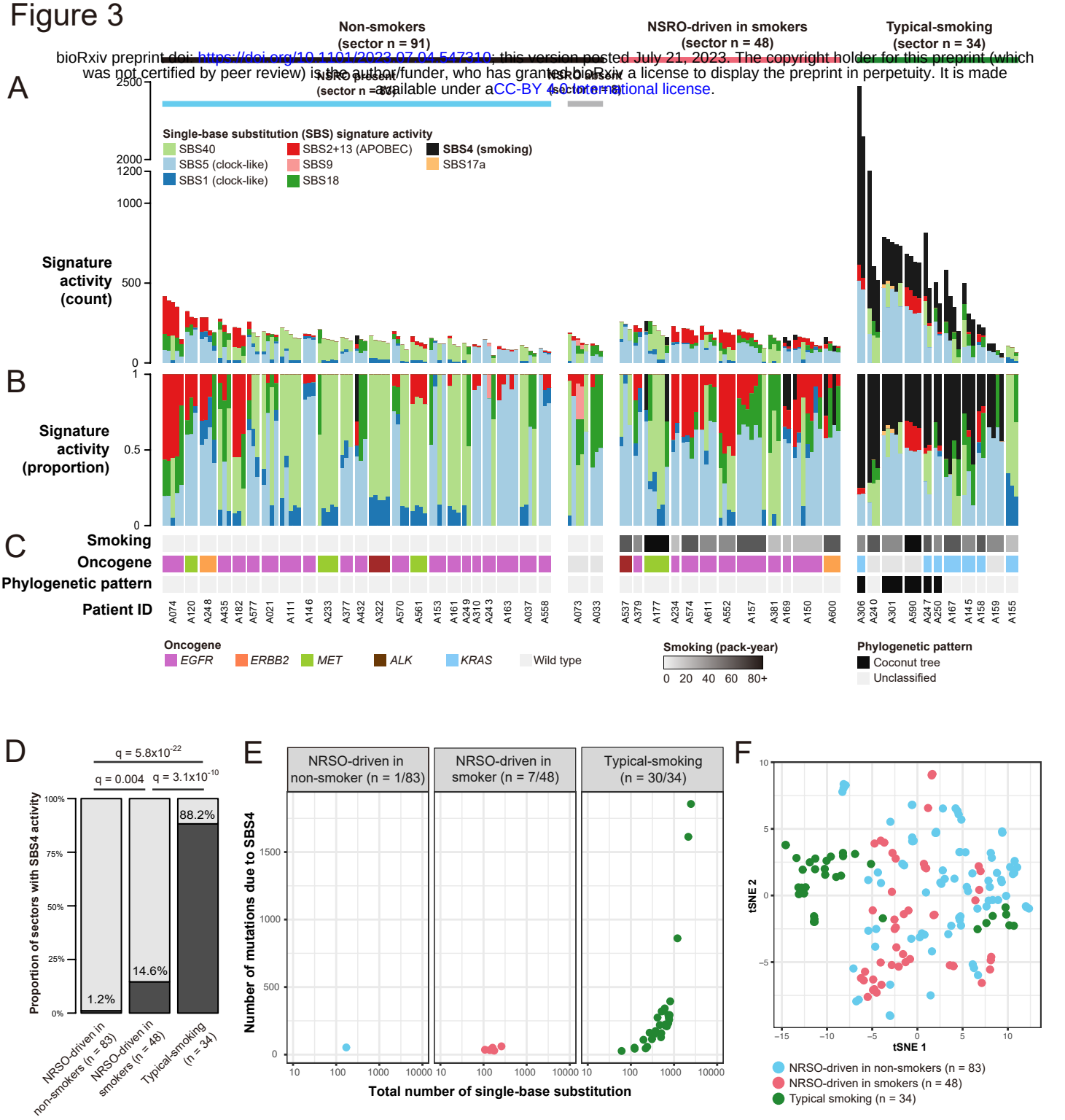


Figure 3. Single-base substitution (SBS) mutational signatures. (**A**, **B**) Mutational-signature activities in the three groups by absolute mutation counts (**A**) and by proportions (**B**). Colors indicate various mutational signatures (e.g., SBS1, SBS5, etc.), as indicated by the legend above. (**C**) Smoking history, key mutated genes, and whether the tumor has a coconut tree pattern. (**D**) Proportions of tumor sectors with SBS4 (caused by tobacco smoking). (**E**) Counts of mutations due to SBS4 in tumor sectors that have SBS4 mutations. (**F**) 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.07.04.547310; this version posted July 21, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

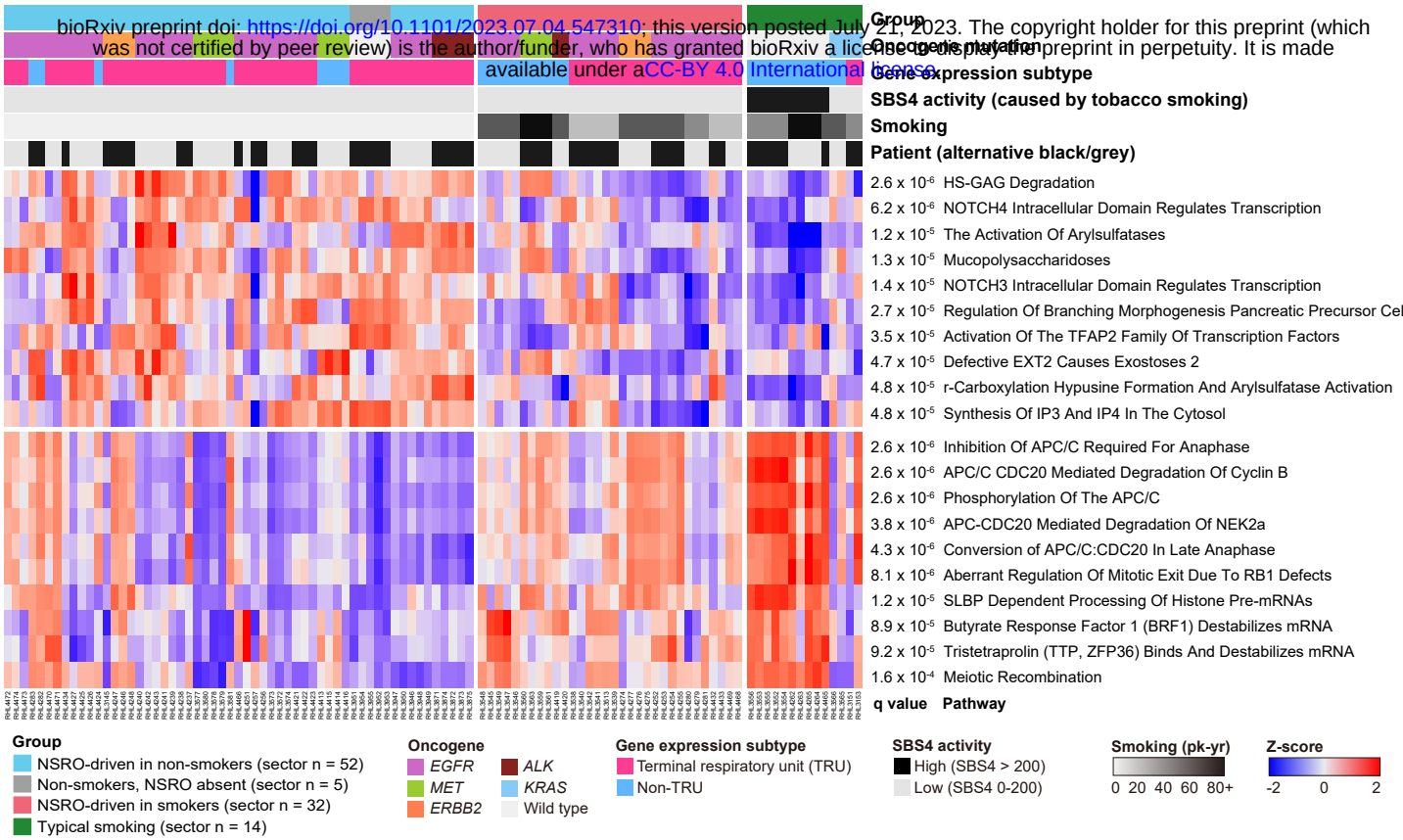


Figure 4. Heatmap of activities of the top 10 pathways up- and down-regulated in all non-smokers compared to all smokers. Each column is a tumor sector, and sectors are grouped by patient as shown in the row labelled “Patient”. Z-scores are of pathway activity. The Benjamini-Hochberg false discovery rates (q values) of differential pathway activity were based on p-values calculated using limma. Supplementary Fig. S9 and Supplementary Tables S10 and S11 provide details.