IJC
International Journal of Cancer

The genomic landscape of nonsmall cell lung carcinoma in never smokers

Bram Boeckx[1,2]*, Rajendra B. Shahi[3]*, Dominiek Smeets[1,2], Sylvia De Brakeleer[3], Lore Decoster[3], Thomas Van Brussel[1,2], Daniella Galdermans[4], Piet Vercauter[5], Lynn Decoster[6], Patrick Alexander[7], Guy Berchem[8], Sebahat Ocak[9,10], Peter Vuylsteke[11], Koen Deschepper[12], Marc Lambrechts[13], Nadia Cappoen[3], Erik Teugels[3], Diether Lambrechts[1,2] and Jacques De Greve[3]

[1]Laboratory of Translational Genetics, VIB Center for Cancer Biology, VIB, Leuven, Belgium
[2]Laboratory of Translational Genetics, Department of Human Genetics, University of Leuven (KULeuven), Leuven, Belgium
[3]Laboratory of Medical and Molecular Oncology (LMMO), Department of Medical Oncology, Oncologisch Centrum, UZ Brussel, Vrije Universiteit Brussel (VUB), Brussels, Belgium
[4]Department of Pneumology, ZNA Middelheim, Antwerp, Belgium
[5]Department of Pneumology, OLVziekenhuis, Aalst, Belgium
[6]Department of Pulmonology, AZ Turnhout, Turnhout, Belgium
[7]Thoracale Oncologie, AZ Glorieux, Ronse, Belgium
[8]Centre Hospitalier de Luxembourg, Luxembourg City, Luxemburg
[9]CHU UCL Namur (Godinne Site), Yvoir, Belgium
[10]Institut de Recherche Expérimentale et Clinique (IREC), Pneumology Pole, Université Catholique de Louvain (UCL), Ottignies-Louvain-la-Neuve, Belgium
[11]Université Catholique de Louvain, CHU UCL Namur, Site Sainte Elisabeth, Namur, Belgium
[12]Algemeen Ziekenhuis Nikolaas, Sint-Niklaas, Belgium
[13]Algemeen Ziekenhuis St Maarten, Mechelen, Belgium

*B.B. and R.B.S. shared equally first authorship

Additional Supporting Information may be found in the online version of this article.

Key words: lung cancer, whole-exome sequencing, never smokers

Abbreviations: ALK: ALK receptor tyrosine kinase; APOBEC: apolipoprotein B mRNA editing enzyme, catalytic polypeptide; ARID1A: AT-rich interaction domain 1A; ASCAT: allele-specific copy number analysis of tumors; ATM: ATM serine/threonine kinase; BRAF: B-Raf proto-oncogene, serine/threonine kinase; BRCA1: BRCA1 DNA repair associated; BRCA2: BRCA2 DNA repair associated; BWA: Burrows–Wheeler aligner; COSMIC: Catalog of Somatic Mutations in Cancer; CCNE1: cyclin E1; CDKN2A: cyclin-dependent kinase inhibitor 2A; CDKN2B: cyclin-dependent kinase inhibitor 2B; CNA: copy number alteration; CSMD1: CUB and Sushi multiple domains 1; DNA: deoxyribonucleic acid; EGFR: epidermal growth factor receptor; ERBB2: erb-b2 receptor tyrosine kinase 2; ERBB3: Erb-B2 Receptor Tyrosine Kinase 3; EWI: Flemish Government, Department of Economy, Science and Innovation; FANCA: FA complementation group A; FDR: false discovery rate; FFPE: formalin-fixed and paraffin-embedded tumor sample; FGFR1: fibroblast growth factor receptor 1; GATK: Genome Analysis Toolkit; GISTIC: Genomic Identification of Significant Targets In Cancer; HR: homologous recombination; IL7R: interleukin 7 receptor; IGV: integrative genomics viewer; IPA: ingenuity pathway analysis; KRAS: KRAS proto-oncogene, GTPase; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MECOM: MDS1 and EVI1 complex locus; MET: MET proto-oncogene, receptor tyrosine kinase; MRE11A: MRE11 homolog, double-strand break repair nuclease; NCBI: National Center for Biotechnology Information; NSCLC: nonsmall cell lung cancer; PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; POLE: DNA polymerase epsilon, catalytic subunit; RBM10: RNA binding motif protein 10; ROS: proto-oncogene 1, receptor tyrosine kinase; sCNA: somatic copy number alteration; SNP: single nucleotide polymorphism; TCGA: The Cancer Genome Atlas; TERT: telomerase reverse transcriptase; TK: tyrosine kinase; TP53: tumor protein p53; TSC22D1: TSC22 domain family member 1; U2AF1: U2 small nuclear RNA auxiliary factor 1; VSC: Flemish Supercomputer Center; WES: whole-exome-sequencing; WGS: whole-genome sequencing; WHSC1L1: Wolf-Hirschhorn syndrome candidate-1 gene

Conflict of interest: Peter Vuylsteke received travel grants and advisory board fees from Roche and Piet Vercauter received lecture fees from Astra Zeneca, Boehringer-Middelheim, BMS, Daiichi Sankyo and Novartis, research grants from MSD and Roche and has an advisory function for Astra Zeneca, Bayer and Chiesi.

Grant sponsor: National Cancer Plan; **Grant number:** Project 29-03; **Grant sponsor:** Department of Economy, Science and Innovation from the Flemish government; **Grant sponsor:** Hercules Foundation

DOI: 10.1002/ijc.32797

History: Received 28 Mar 2019; Accepted 15 Oct 2019; Online 20 Nov 2019

Correspondence to: Diether Lambrechts, E-mail: diether.lambrechts@vib-kuleuven.be; or Jacques De Greve, E-mail: jacques.degreve@uzbrussel.be

Lung cancer is the number one cause of cancer-related death worldwide with cigarette smoking as its major risk factor. Although the incidence of lung cancer in never smokers is rising, this subgroup of patients is underrepresented in genomic studies of lung cancer. Here, we assembled a prospective cohort of 46 never-smoking, nonsmall cell lung cancer (NSCLC) patients and performed whole-exome and low-coverage whole-genome sequencing on tumors and matched germline DNA. We observed fewer somatic mutations, genomic breakpoints and a smaller fraction of the genome with chromosomal instability in lung tumors from never smokers compared to smokers. The lower number of mutations, enabled us to identify *TSC22D1* as a potential driver gene in NSCLC. On the other hand, the frequency of mutations in actionable genes such as *EGFR* and *ERBB2* and of amplifications in *MET* were higher, while the mutation rate of *TP53*, which is a negative prognostic factor, was lower in never smokers compared to smokers. Together, these observations suggest a more favorable prognosis for never smokers with NSCLC. Classification of somatic mutations into six-substitution type patterns or into 96-substitution type signatures revealed distinct clusters between smokers and never smokers. Particularly, we identified in never smokers signatures related to aging, homologous recombination damage and APOBEC/AID activity as the most important underlying processes of NSCLC. This further indicates that second-hand smoking is not driving NSCLC pathogenesis in never smokers.

What's new?

About 15 percent of lung cancers occur in people who never smoked but never smokers are usually underrepresented in genomic studies of lung cancer. Here the authors sequenced a unique cohort of 46 never smokers with non-small cell lung carcinoma (NSCLC). They identified the transcription factor TSC22D1 as a new potential driver gene of NSCLC and observed higher mutation rates in receptor tyrosine-protein kinases EGFR HER2 and MET in tumors of never smokers compared to smokers. Targeted agents against each of these 3 genes are available, underscoring the relevance of studying never smokers as a separate group.

Introduction

Lung cancer is the number one cancer-related mortality worldwide, leading to over a million deaths each year. The main risk factor of lung cancer is cigarette smoking.[1] In the Western world, approximately 10–15% of lung cancers, however, affect never-smokers,[2,3] which are usually defined by smoking <100 cigarettes during a lifetime. Lung cancer classification is traditionally based on histological subtyping, dividing lung cancer into small cell lung cancer and nonsmall cell lung cancer (NSCLC). The latter is by far the most common type (85%), consisting of three major subtypes: squamous cell carcinoma, large cell carcinoma and adenocarcinoma.[4,5] Treatment with platinum-based doublet chemotherapy which is guided by these histological subtypes prolongs survival only modestly.[6] However, recently the approval of immune checkpoint inhibitors provides unexpected survival benefits compared to chemotherapy alone, both as a first or second-line treatment option.[7–9]

In oncogene-driven NSCLC tumors, the development of targeted therapies against tyrosine kinase (TK) receptors has made major treatment advances. Particularly, these TK inhibitors generate high response rates and significant survival benefits in tumors carrying a somatic mutation in the TK receptor targeted by the TK inhibitor. The identification of actionable driver mutations in these receptors, including *EGFR*, *ALK*, *ROS*, *BRAF* and *ERBB2* mutations, combined with the decreasing cost of genomic testing was key for the clinical implementation of these treatments.[10] However, all NSCLC tumors at some point acquire resistance to these treatments.

Despite the large amount of genetic data generated by the large-scale efforts of the TCGA[11,12] with over 1,000 patients being sequenced, it remains challenging to distinguish driver mutations from passenger mutations, particularly in NSCLC from (ex-) smokers. This is due to the high mutation burden which is a direct consequence of exposure to tobacco smoke. Genomic profiling of lung carcinoma revealed that these mutations mainly consist of C:G>A:T mutations, which arise due to the mutagenic effect of tobacco. The vast majority of patients with clinical information about smoking status in the TCGA cohort are self-declared current or former smokers. As a result, the pathogenesis of lung cancer in never-smokers is much less characterized, especially with respect to whole-exome sequencing (WES) or whole-genome sequencing (WGS) studies.

The French BioCast study has assessed the mutation status of *EGFR*, *KRAS*, *ALK*, *BRAF*, *ERBB2* and *PIK3CA* in 359 never smokers of which 66% had a passive smoking history.[13] This study represents the largest study of lung cancer in never smokers in the European population. Interestingly, an actionable molecular alteration was found in 73% of the cases, with *EGFR* (51%) being the most frequently altered gene, followed by *ALK* (8%), *KRAS* (6%), *ERBB2* (3%), *BRAF* (3%) and *PIK3CA* (1%). There were no significant differences in the frequency of these specific gene mutations between never smokers with or without a passive smoking history. In contrast to the BioCast study, which was limited to just a few genes, Govindan *et al.*[14] performed WGS on a very small ($n = 5$) cohort of never

smokers and revealed that the genomic landscape of lung cancer in never smokers is remarkably distinct from that in smokers.

Interestingly, the prevalence of lung cancer in never smokers is higher in Asian countries. WGS on 16 Asian never smokers revealed a mutation signature that was distinct from the Asian smokers but similar to European never smokers.[15] In another study, the landscape of lung adenocarcinoma was assessed in a cohort of 188 patients, which included a subset of 27 never smokers. Our study confirmed a distinct mutational pattern in never smokers and identified new potential driver mutations in *U2AF1*, *RBM10* and *ARID1A* in lung adenocarcinoma.[16]

To better characterize the mutation landscape of lung cancer in never smokers, we here prospectively selected and analyzed NSCLC from never smokers and assessed their somatic mutation landscapes by performing WES and low-coverage WGS.

Materials and Methods

Patient selection

In accordance with the study protocol approved by institutional review boards of Universitair Ziekenhuis Brussel and Antwerp University Hospital, NSCLC patients (n = 46) with history of nonsmoking (strictly never smoked or smoked <100 cigarettes in their lifetime) or former limited smoking (smoked less than five packed years with a smoking free interval of more than 20 years) were enrolled from 2013 to 2015 for the present study, which is called the Belgian FIELT-2 study. From each patient, and after written informed consent, formalin-fixed and paraffin-embedded tumor sample and matched normal tissue (blood) were obtained for an explorative genomic analysis. Patient characteristics are reported in Supporting Information Table S1A.

DNA extraction, whole-exome and whole-genome sequencing and downstream analysis

Genomic DNA was extracted from FFPE and blood with Qiagen DNA extraction kits and sequencing libraries were prepared with Illumina's Truseq V2 library preparation kit and sequenced with an Illumina HiSeq 2500 or HiSeq 4000. The exome sequencing data were analyzed with the well-established BWA-GATK pipeline and variants were further filtered and annotated with Annovar using publicly available databases. A subset of the mutations identified was validated with an orthogonal method: Sanger sequencing (n = 33), Sequenom MassArray (n = 137) or target sequencing (n = 3). The analysis and plotting of the somatic signatures were performed using deconstructSigs[17] and SomaticSignatures,[18] both available in Bioconductor. The low-coverage whole-genome sequencing data were also mapped with BWA and analyzed with QDNAseq[19] and ASCAT.[20] Extended details on DNA extraction, whole-exome sequencing, validation of the somatic mutations, low-coverage whole-genome sequencing and the analysis of the mutational signatures and pathways are described in Supporting Information.

NSCLC in smokers and Asian never-smokers

To compare our findings with NSCLC in smokers, we downloaded TCGA datasets of lung squamous cell carcinoma (LUSC; n = 492) and lung adenocarcinoma (LUAD; n = 508). In these datasets we subselected (*i*) a cohort of current smokers with either LUSC or LUAD (n = 134 and n = 122, respectively), (*ii*) a cohort with strict never smokers (n = 18 and n = 75 with LUSC or LUAD, respectively), which includes self-declared patients that have never smoked and (*iii*) a cohort combining lung cancer patients that refrained from smoking >15 years ago and those that never smoked (n = 212 and n = 101 with LUSC and LUAD, respectively). Notably, in the last two cohorts (n = 93 and n = 313), the smoking signature (Signature 4) explained ~23% of the observed mutation spectrum of exome data, suggesting that both cohorts were very similar, but did not purely consist of strict never smokers. WGS mutation calls in 16 Asian never smokers[15] were accessed from the European Genome–Phenome Archive under accession number EGAD00001001240. Clinical characteristics of these cohorts are summarized in Supporting Information Table S1*b*.

Data availability

Raw sequencing data have been deposited in the ArrayExpress database under accession numbers E-MTAB-7811 (WGS data) and E-MTAB-7833 (WES data).

Results

FIELT-2 is a prospective clinical study investigating the somatic mutation landscape of NSCLC in patients that are never smokers. Particularly, it aims to identify novel driver mutations especially in those tumors in which no actionable primary driver mutations were identified.

Overall, 36 strict nonsmoking and 10 former light smokers defined as having smoked <5 pack-years and with a smoking-free interval of >20 years, were enrolled at the time of diagnosis of lung cancer. In FIELT-2, we refer to all these 46 patients as never smokers. Lung adenocarcinoma was the major histological subtype (n = 43, 93%) followed by squamous cell carcinoma (n = 2, 4%) and large cell carcinoma (n = 1, 2%). Our mainly Caucasian (n = 43, 91%) cohort consists of both female (n = 27, 59%) and male (n = 19, 41%) patients (Supporting Information Table S1). Interestingly, we also observed a higher prevalence of female never smokers with lung cancer in other cohorts. For instance, Krishnan *et al.*[15] reported that 62% of never smokers with lung cancer were women (10/16; 62%). Remarkably, both these cohorts consisted almost completely of lung adenocarcinomas. Indeed, also in the TCGA cohort of never smokers with LUAD, 73% of patients (55/75) were women, while in the patients with lung squamous cell carcinoma, (8/18) only 44% were women (Supporting Information Table S1*b*).

Whole-exome sequencing of NSCLC in never smokers

Whole-exome sequencing (WES) was performed on 46 tumors and their matched germline. One tumor-normal pair was

excluded due to insufficient quality of the germline DNA resulting in 45 successfully sequenced patient pairs. Overall, samples were sequenced at 60 ± 30× coverage. Tumor samples were sequenced slightly deeper than the corresponding germline samples (64 ± 33 *vs.* 57× ± 26× coverage). On average 91.3% ± 5.5 of the exome was covered at a sequencing depth >10× coverage and 80.1% ± 14.3 was covered at a depth >20× coverage (Supporting Information Fig. S1 and Table S2). In the tumor samples, we called on average 58,684 ± 3,171 [50,730–62,662] variants and 68,023 ± 62,561 [22,297–333,274] indels (Supporting Information Table S3). Somatic mutations were identified by removing mutations detected in the corresponding germline samples and by filtering out known common mutations, resulting in on average 38 substitutions and 10 indels per tumor, which corresponds, as expected, to low tumor mutational burden of 0.8 mutations/Mbase. A subset of these mutations (n = 173) was validated by an orthogonal technology, yielding a true-positive rate of 89% (Supporting Information Table S4). Thus, the large majority of somatic mutations detected were true-positive mutations. The tumor mutation burden detected in this cohort was sixfold lower (Fig. 1*a*) compared to that of strictly smokers in the TCGA LUSC and LUAD datasets (p-value <2.2 $\times 10^{-16}$, Student's t-test). This strong reduction in mutation burden in never smokers has also previously been observed in WGS data.[14] The number of mutations detected in our cohort of NSCLC never smokers is in concordance with the number of mutations detected in the NSCLC Asian never smokers (48 *vs.* 43 mutations, respectively).[15] The ratio of silent to nonsilent mutations was comparable for light smokers (0.32) and never smokers (0.35), suggesting that selective pressure on somatic mutations was comparable between both. Notably, we also did not observe a significant difference in tumor mutation burden between never smokers and former light smokers in FIELT-2 (p-value = 0.7, Student's t-test, Supporting Information Fig. 2*a*), supporting the fact that we can consider both our cohorts as one entity.

Somatic mutation patterns in NSCLC in never smokers

Next, we assessed whether the tumor mutation landscape differed between never smokers and smoke-induced lung cancer. Hereto, we classified somatic mutations based on six subtypes of substitutions (i.e., A:T>G:C, G:C>A:T, A:T>T:A, G:C>C:G, G:C>T:A and A:T>C:G) for each sample (Fig. 1*b*). We found

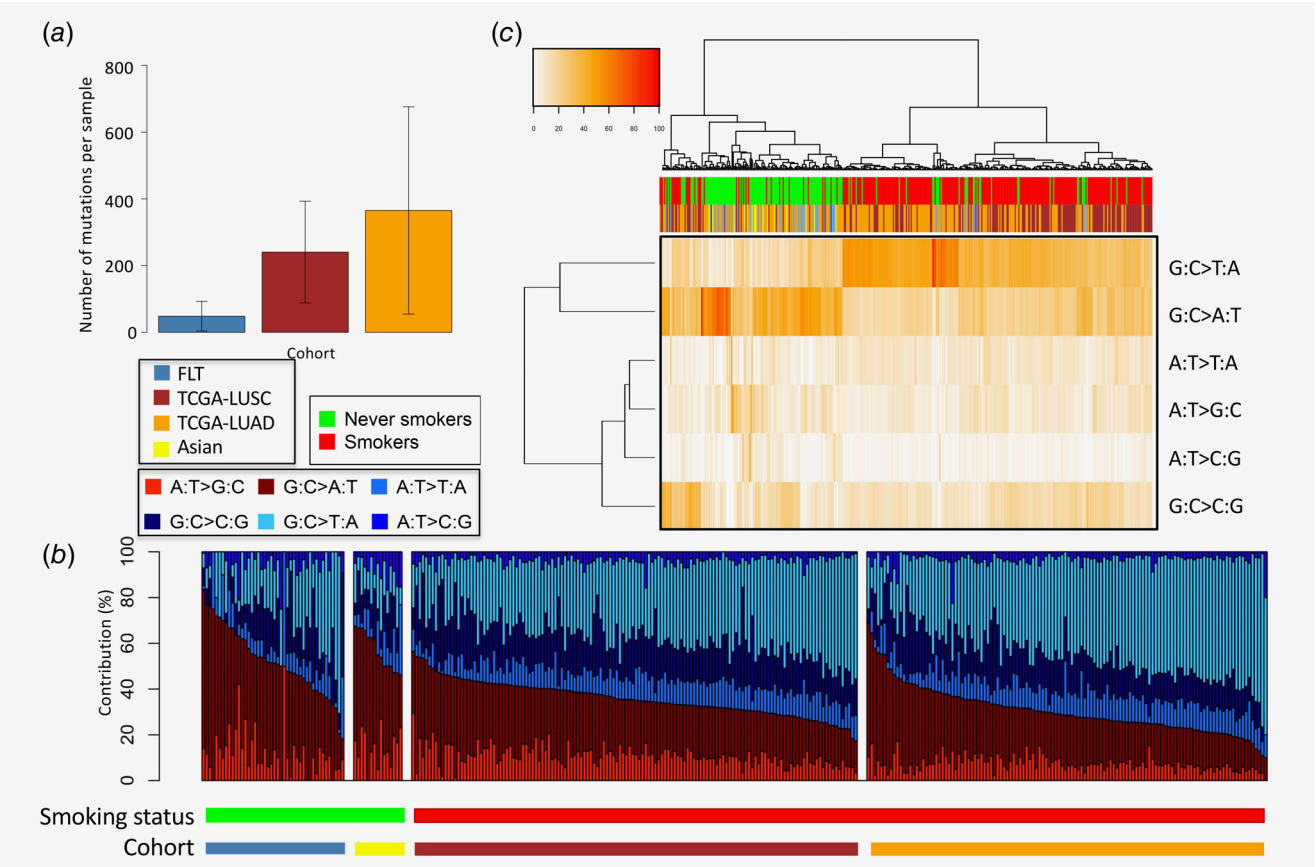

Figure 1. Mutation characteristics in never smokers and strictly smokers: FIELT-2 never smokers, Asian never smokers, TCGA data of strict smokers with lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). (*a*) Number of mutations; (*b*) Six substitution types; (*c*) Unsupervised clustering of the six substitution types for strictly smokers and never smokers reveals two distinct clusters. [Color figure can be viewed at wileyonlinelibrary.com]

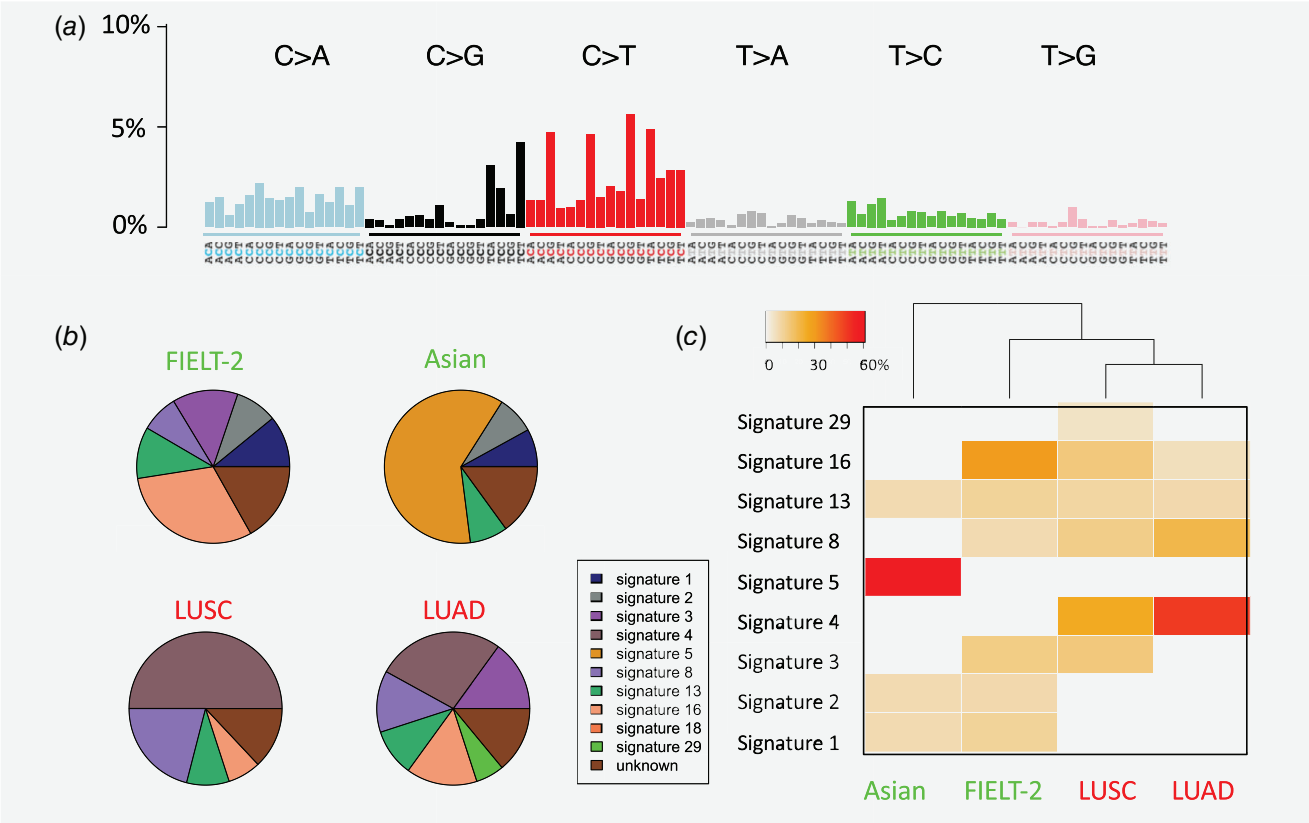

Figure 2. (*a*) Ninety-six subtype classification of the mutational signature of NSCLC in never smokers. (*b*) Contributions of the established COSMIC somatic signatures to the 96-subtype classification of the FIELT-2 and Asian never smokers and smokers of the lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) TCGA datasets. (*c*) Unsupervised clustering of the significant contributions. [Color figure can be viewed at wileyonlinelibrary.com]

that 40.1% of all mutations occurring in never smokers represented G:C>A:T transitions. On the contrary, G:C>T:A transversions, which we observed in 21.8% of the never smokers, are relatively more abundant in lung tumors of smokers. Indeed, in current smokers selected from the TCGA G:C>T:A transversions were more frequent, that is, 41.5% in LUAD and 33.0% in LUSC. G:C>T:A mutations are known to be induced by tobacco smoke. Unsupervised clustering (Fig. 1*c*) of NSCLC tumors included in the FIELT-2 cohort, the Asian never smokers study, as well as both TCGA-based LUSC and LUAD cohorts confirmed the existence of distinct subgroups that are determined by their smoking status. Obviously, unsupervised clustering differentiated smokers from never smokers (Fisher's exact test, p-value $<2.2 \times 10^{-16}$). The classification based on six mutation subtypes also revealed that former light smokers clustered together with the never smokers as both are dominated by G:C>A:T transitions (Supporting Information Fig. 2*b*; Fisher's exact test, p-value $<2.2 \times 10^{-16}$).

We then analyzed the trinucleotide context around each substitution to reveal the etiology of the somatic mutations accumulating in tumors from never smokers.[21] This 96-substitution classification was developed for WGS data but can also be used for exonic variants after correction for the nucleotide content of the exome. However, since the tumor mutation burden in our cohort of never smokers is rather low, we combined all mutations detected in all 45 tumors from never smokers (Fig. 2*a*) and did not analyze them on a per-patient basis. The resulting never smoker signature based on the trinucleotide context of the substitutions was decomposed into the 30 signatures of mutational processes in human processes, allowing us to estimate the relative abundance of the mutational processes contributing to the etiology of NSCLC in never smokers. (Fig. 2*b*). We observed that Signature 4, which is dominantly present in the signatures of LUSC (27%) and LUAD (50%) tumors from smokers is absent in NSCLC tumors of never smokers. Indeed, this mutational signature is related to exposure to tobacco mutagens and is associated with smoking. Notably, Signature 4 was absent in tumors from never smokers from FIELT-2. Interestingly, this absence suggests that tumorigenesis of NSCLC in never smokers is not due to second-hand smoking.[22] Furthermore, it suggests that the etiology of lung cancer in former light smokers is unrelated to their past tobacco exposure.

Instead, the mutational signature of lung tumors from FIELT-2 never smokers was composed of Signature 16 (31%), 3 (14%), 1 (11%), 13 (11%), 2 (9%) and 8 (8%; Fig. 2*b*). These

signatures are also observed in NSCLC tumors from smokers that were sequenced by TCGA, with the exception of Signatures 1 and 2. Notably, Signature 16 (31%) is the most abundant contributor to tumors of never smokers and is much less abundant in lung cancer of smokers, that is, 7% (LUAD) and 15% (LUSC). This signature is mainly characterized by T>C mutations in an ApTpN context, occurring almost exclusively on the transcribed strand. The etiology underlying this mutation is however still unknown. Signature 3, contributing to 14% of the observed never smoker signature in our study is associated with the failure of DNA double-strand break repair by homologous recombination. DNA damage due to HR repair deficiency in our tumor samples is also supported by the presence of somatic mutations in *ATM* (n = 2, 4% of the tumors), *MRE11A* (n = 2, 4%), *BRCA1* (n = 1, 2%) and *BRCA2* (n = 1, 2%) and significant focal deletions of 13q12.12 (GISTIC q-value = 0.01) containing *BRCA2* (n = 26, 57%) and 16q24.1 (GISTIC q-value = 0.003) containing *FANCA* (n = 24, 52%). Overall, a somatic mutation or a focal deletion in HR genes is observed in 34 patients (74%). Signature 1 (11%) correlates with aging and is dominated by C>T transitions, which are initiated by spontaneous deamination of 5-methylcytosine. Because the smoking signature in tumors from smokers is abundant, Signature 1 might not be detectable as its frequency will be quite low. In addition, the mean age of our cohort of never smokers is 67 (46–86), which is significantly (p-value = 0.026, Student's t-test) higher compared to the smokers in the TCGA cohort with a mean age of 63 (33–90) years, which might possibly contribute to the absence of Signature 1. Signature 13 (11%), consisting mainly of mainly C>G substitutions, is assigned to the activity of *AID*/*APOBEC* cytidine deaminases, which convert cytidine into uracil. Some hypothesize that the activation of these cytidine deaminases is caused by viral infection, retrotransposon activity or inflammation. Signature 13 often co-occurs with Signature 2, which contributes 9% to the never smokers signature. Signature 2 is not observed in smokers with LUSC and LUAD and is also linked to cytidine deaminases by *AID*/*APOBEC* activity. Signature 8 (8%) contributes <10% to the signature of lung cancer in never smokers and has an unknown etiology. The other mutational processes that contribute to the spectrum of never smokers also seem to be present in strictly smokers, despite the large difference in G:C>T:A substitutions.

The mutational signature of Asian never smokers also lacks smoking Signature 4, and it displays the same features as the never smoker signature observed in FIELT-2. The strong enrichment for Signature 5 (61%) in the Asian never smokers, can be explained by its high similarity to Signature 16 (observed FIELT-2), since both are dominated by T>C substitutions in an ApTpN context meaning that the etiology of NSCLC in Asian and Caucasian never smokers is similar. Finally, unsupervised clustering based on the 96-substitution signature deconstruction also revealed distinct cluster patterns between never smokers and smokers with lung cancer (Fig. 2*c*).

EGFR and TP53 mutations are most common in never smokers with NSCLC

In FIELT-2, *EGFR* was by far the most frequently mutated gene (Supporting Information Table S1). Particularly, we observed 21 *EGFR* mutations in 18 patients (18/45 = 40%). The frequency of *EGFR* mutations in never smokers is considerably higher compared to the lung cancer TCGA[23,24] cohort (10.2%) or compared to smokers in either the LUSC (8%) or LUAD (8%) TCGA subcohorts. The higher incidence of *EGFR* mutations in never smokers was also observed in BioCast (61%),[25] Asian (67%)[15] and TCGA (39%) never smokers. This increase has clinical implications since these patients can be treated with *EGFR* tyrosine-kinase inhibitors (TKI). The most frequent *EGFR* mutation (Supporting Information Fig. S3*a*) is a deletion in exon 19 that was detected in 10 out of 45 patients (22%), which again represents a considerable increase compared to the pan-lung cancer analysis of TCGA detecting this deletion only in 3.5% (41 out of 1,144) of patients. We observe three different nonframeshift deletions: p.745_750del (n = 7, 16%), p.746_748del (n = 2, 4.5%) and p.746_750del (n = 1, 2.3%). The most abundant point mutation L858R was observed in 7 (16%) never smokers, which interestingly also represents the most abundant point mutation in the pan-lung cancer dataset (n = 28 out of 1,144; 2.4%).

Interestingly, *TP53* ranks as the second most significantly mutated gene. Particularly, we observed *TP53* mutations in 12 patients (27%), which is in accordance with the Asian never-smoker cohort (20%),[15] as well as the TCGA never smoker cohort, in which we observed a 29 and 56% mutation rate for LUSC and LUAD, respectively. Notably, these rates are considerably lower than the 69% (LUSC) and 82% (LUAD) mutation rates observed in smokers from TCGA.[23]

Identifying new driver genes active in NSCLC

The search for novel driver mutations active in NSCLC is complicated by the fact that smoke-induce lung cancer is characterized by a high tumor mutation burden, which obviously includes a large number of passenger mutations. Since the tumor mutation burden is lower in NSCLC of never smokers, which should enable the identification of novel driver mutations in NCCLC, we applied several pipelines dedicated to identify novel driver mutations. Particularly, we used GenomeMuSiC[26] and MutSigCV[27] (Supporting Information Table S6). Next to the recurrently mutated genes *EGFR* and *TP53*, we also identified *TSC22D1* as a potential driver gene in NSCLC (Supporting Information Fig. S3*c*). Particularly, we observed three patients with a deletion in *TSC22D1* and one with a nonsynonymous substitution. Interestingly, *TSC22D1* was the third most significantly mutated gene in GenomeMuSiC (q = 0.0057, Supporting Information Table S6). and represents a putative tumor suppressor that inhibits Ras/Raf signalling[28] and that is coexpressed with *ACSL1* which has a potential tumor-suppressing role in lung cancer[29] and

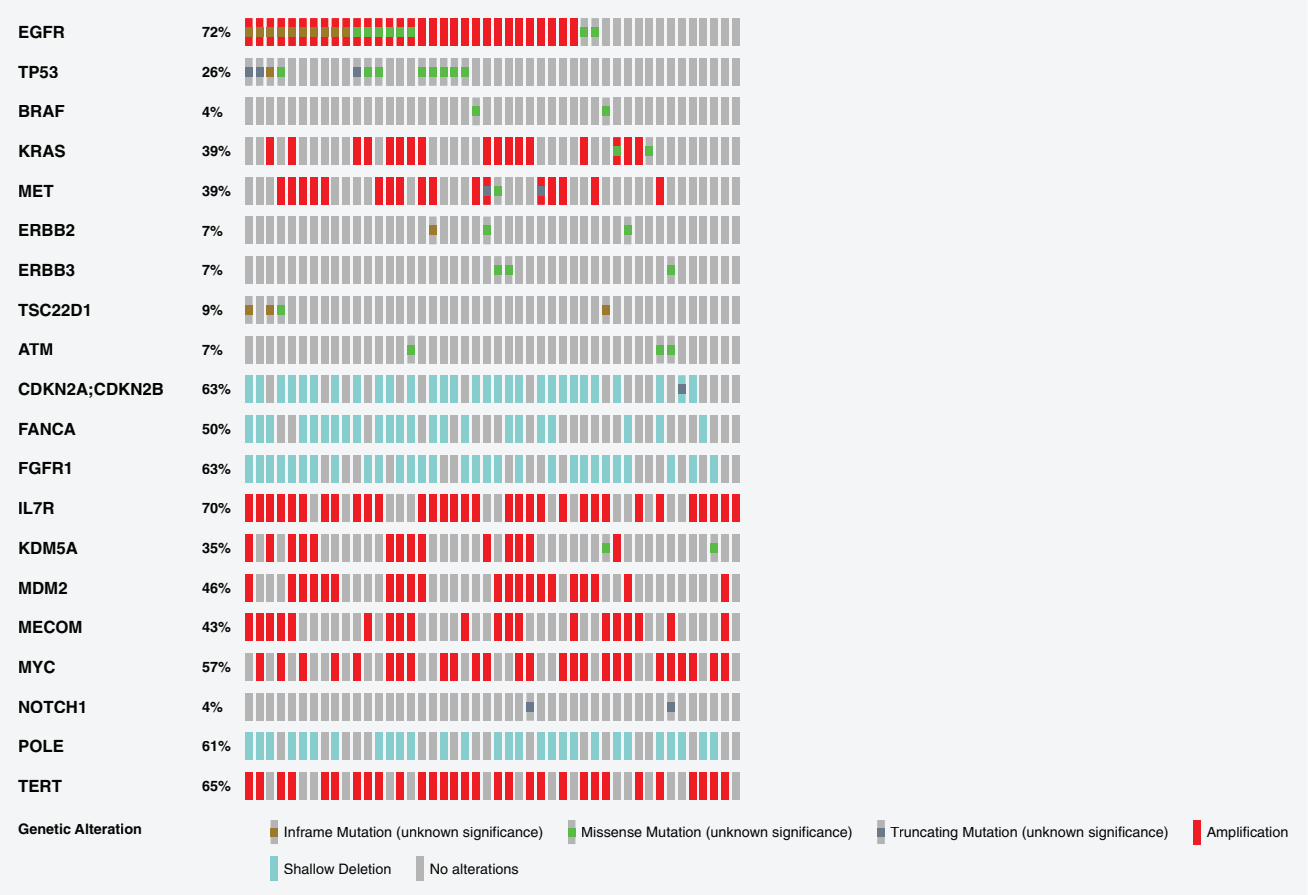

Figure 3. Oncoprint of NSCLC in never smokers for 45 patients with mutation data and 45 patients with copy-number data. The percentages represent the affected patients. The samples without mutation data (°) and copy-number data (*) are indicated. The percentage of affected samples in the FIELT-2 cohort (n = 45) is shown at the left, while the percentage of affected samples in never smokers of the TCGA cohort (including n = 93 strict never smokers and n = 220 former light-smokers that stopped smoking >15 years ago) is presented at the right. [Color figure can be viewed at wileyonlinelibrary.com]

previously has been linked to the development of pulmonary adenoma in mice.[30]

We then assessed whether *TSC22D1* was also mutated in other cohorts. The Asian study was too small (n = 16) to really explore this, while *TSC22D1* was not included in the gene panel of the BioCast study. However, in the TCGA cohort of never smokers, 6 (2%) patients harbored a *TSC22D1* mutation, which is slightly higher than in TCGA smokers. Indeed, although TCGA tumors from smokers are hypermutated, we observed *TSC22D1* mutations in only 1.32% of patients. In none of the cohorts, mutations and copy number alterations affecting *TSC22D1* were prognostic (Logrank test p-value = 0.9, Supporting Information Fig. S4). When assessing other cancer types (n = 10,953 *via* the cBioPortal), *TSC22D1* was mutated in 147 (1.34%) of cases (Supporting Information Fig. S5). At the level of copy number alterations, *TSC22D1* was altered in 307 patients (2.80%).

We observed *ERBB2* mutations in 3 out of 45 (7%) patients. Similar frequencies were observed in the Asian[15] (7%), BioCast[25] (3%) and TCGA (1%) lung cancer cohorts of never smokers, and were higher compared to smoke-induced lung cancer cohorts, where it was observed in 2–4% of patients.[31–34] *ERBB2* mutations may be more relevant in lung carcinogenesis compared to *ERBB2* amplification, as several kinase inhibitors such as afatinib have already shown considerable activity in *ERBB2* mutant lung cancer.[35,36] Interestingly, also *RBM10*, which represents a potential driver of lung adenocarcinoma, was found to be mutated in 2 out of 45 samples (7%).[16] Interestingly, strict mutual exclusivity was found for driver mutations in *EGFR*, *KRAS*, *BRAF* and *MET*, with 26 out of 45 tumors harboring a mutation in one of these genes (Fig. 3).

Copy number alterations in NSCLC of never smokers

To also define the role of chromosomal alterations in NSCLC in never smokers, we successfully performed low-coverage WGS on 45 tumor samples (Supporting Information Table S7). We observed somatic copy number alterations (sCNAs) in most tumor samples and assessed whether the number of sCNA in never smokers is lower than in smokers,

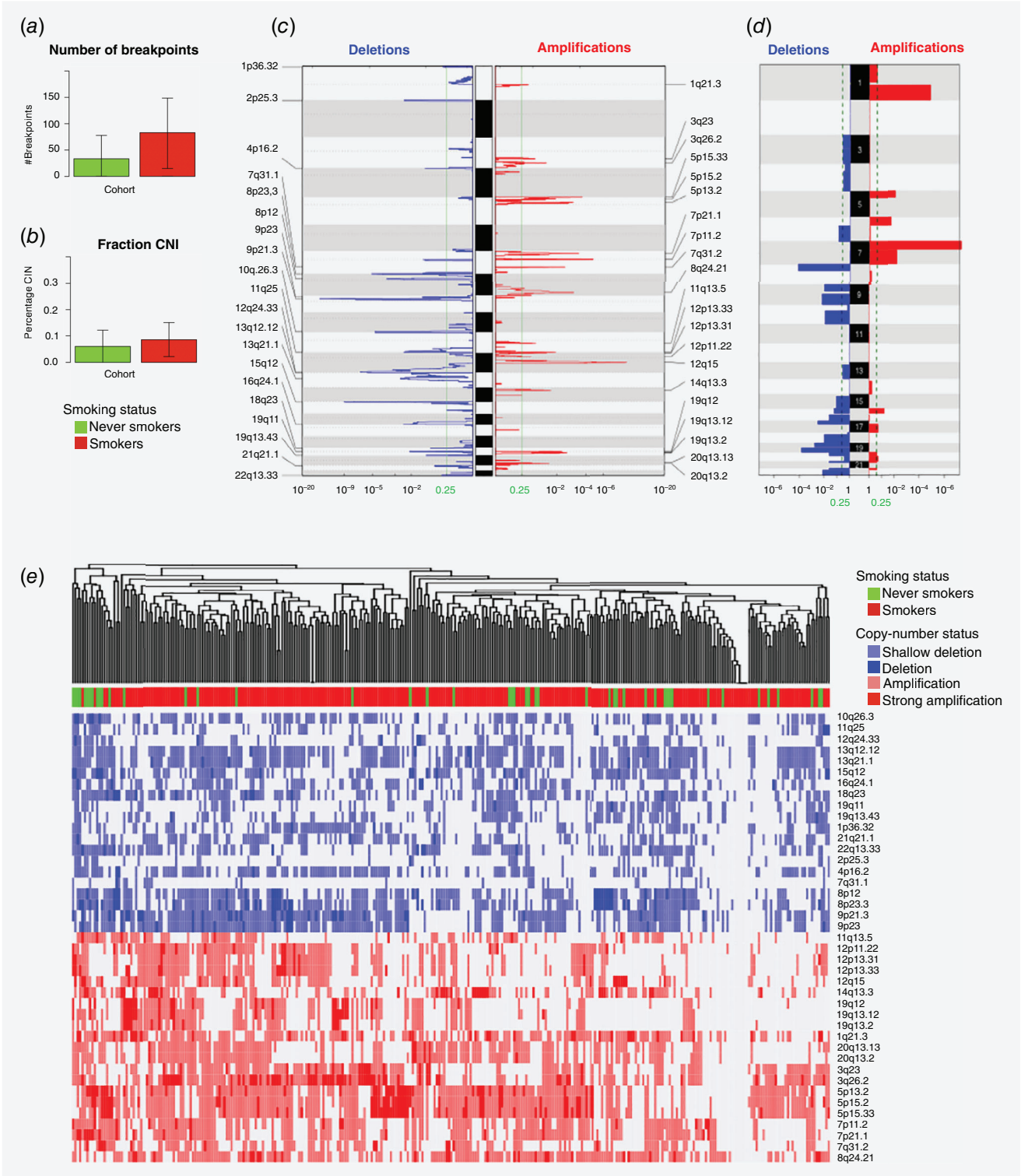

Figure 4. Barplots comparing never smokers and smokers: (*a*) the number of breakpoints and (*b*) the fraction of sCNA. Significant (*c*) focal and (*d*) broad amplification and deletions identified by GISTIC in NSCLC of never smokers (*q*-value ‹0.25). (*e*) Unsupervised clustering of significantly altered focal peaks by GISTIC on never smokers and smokers. [Color figure can be viewed at wileyonlinelibrary.com]

similar to what is observed for mutations. Hereto, we calculated the number of breakpoints and the aberrant fraction of the genome, the latter is defined by the fraction of the genome that is amplified or deleted (Materials and Methods, Supporting Information Table S8). Interestingly, the number of breakpoints observed in never smokers was on average

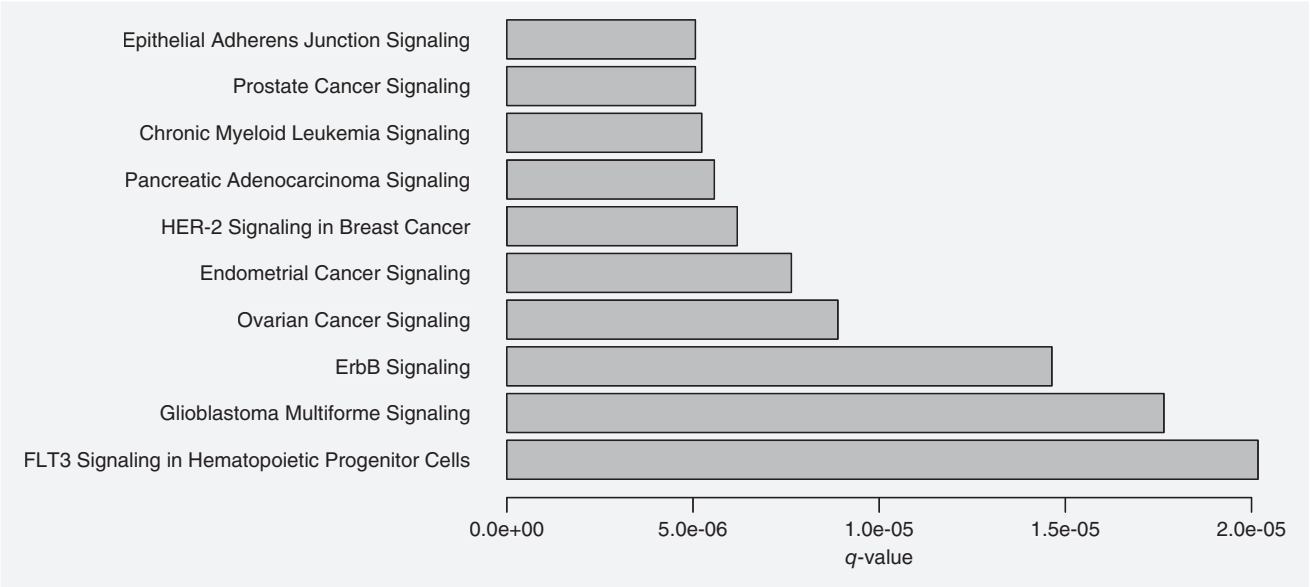

Figure 5. Integrated ingenuity pathway analysis on focal copy-number alterations and recurrently mutated genes, *q* values of the top 10 pathways are shown.

33 ± 45 (0–277), while in current smokers this was 83 ± 66 (0–444). This difference was significant (p-value = 4.12×10^{-8}, Student's t-test; Fig. 4*a*) and was observed both for LUSC and LUAD (82 ± 56 (2–294) and 84 ± 77 (0–444); Student's t-test p-value = 7.36×10^{-8} and 5.04×10^{-7}, respectively).

The aberrant fraction of the genome was $6.0 \pm 6.1\%$ (0–31%) for never smokers and $8.6 \pm 6.5\%$ (0–34%) for smokers, confirming a significant increase in sCNAs in the latter (Student's t-test, p-value = 0.011; Fig. 4*b*). More in detail, the average aberrant fraction of the genome was $7.9 \pm 6.8\%$ (0–30%) in LUAD (Student's t-test, p-value = 0.11) and $9.3 \pm 6.1\%$ (0–34%) in LUSC (Student's t-test, p-value = 0.0038). These results indicate that the genomic profile of never smokers is not only characterized by fewer somatic mutations but also by a reduced number of sCNAs.

To identify which focal and whole-arm or whole-chromosome sCNAs were enriched for in NSCLC, a GISTIC analysis was performed (Figs. 4*c* and 4*d*, Supporting Information Tables S8–S11). One of the most significant focal amplifications that we observed affected *EGFR* (69% of the cases, Supporting Information Table S10). In addition, other significantly amplified regions that contained oncogenes involved chromosome 5p15.33 (*TERT*, 69%), 5p13.2 (*IL7R*, 71%), 8q24.21 (*MYC*, 58%) and 7q31.2 (*MET*, 38%). Amplifications of 19q12 (22%) and 3q26.2 (44%) including, respectively, *CCNE1* and *MECOM* were also observed. Remarkably, focal amplifications of 7q31.2 (containing *MET*) were not frequently found in the pan-lung cancer TCGA cohort, suggesting this is characteristic of NSCLC in never smokers. The large abundance of *MET* amplifications in our cohort (38%) can have therapeutic implications, since it has been reported that NSCLC patients with a *MET* amplified tumor respond to crizotinib, an inhibitor of *MET* and *ALK*.[37] In a Phase I trial, a 33% response rate to crizotinib was reported in a cohort of 12 *MET*-amplified NSCLC tumors.[38] Interestingly, focal *MET* amplification have also been observed in NSCLC (5–22%) with acquired resistance to *EGFR* inhibitors in patients with *EGFR*-mutated tumors, due to the activation of *ERBB3* signaling.[39,40]

Besides these focal amplifications, several focal deletions were also observed. The most significant focal deletions were in 9q21.3, which contains the tumor suppressor genes *CDKN2A* and *CDKN2B* (67%), 8p23.3 (*CSMD1*, 71%) and 12q.24.33 (*POLE*, 62%). The former two regions are also observed in the pan-lung cancer tumors of TCGA. Expression of the tumor suppressor gene *CDKN2A* is also often silenced by promotor methylation which is known to occur more often in smokers compared to never smokers.[41,42] Since we observe a focal deletion of *CDKN2A* in 67% of our cases, the high frequency of this deletion may compensate for the reduced promotor methylation reported in never smokers.

Other focal deletions that we also detected involved the following chromosomal regions: 15q12, 10q26.3 and 8p12 (*FGFR1*, *WHSC1L1*), which are not observed in the pan-lung cancer dataset of TCGA. Focal deletions of 13q12.12 and 16q24.1 containing respectively *BRCA2* and *FANCA*, both involved in the HR pathway, were also often observed.

Interestingly, all except three tumors (FLT050, FLT052 and FLT032) had an amplification or mutation in *EGFR* or *MYC*, suggesting that these genes play an important role in driving NSCLC in never smokers. This might have future therapeutic

implications as *MYC* is also an emerging therapeutic target in cancer.[43]

In addition to the quantitative differences in sCNAs (i.e., number of breakpoints and fraction sCNA), we assessed whether recurrent amplified or deleted regions observed in our nonsmoker cohort are characteristic for never smokers. Therefore, we performed unsupervised clustering on the recurrent focal alterations defined in never smokers on our cohort of never smokers and the current smokers of the TCGA cohort. Statistical analysis revealed two distinct clusters (Fig. 4*e*), one containing mainly never smokers and another one containing mainly smokers (Fisher's exact test, p value = 1.3×10^{-6}).

Integration of mutation and copy number data

Finally, we integrated sCNA and mutation data to reveal which pathways were affected. Hereto, ingenuity pathway analysis (IPA) was performed on the significantly affected focal regions containing <50 genes as well as on all recurrently mutated genes. Results and p values of both analyses were integrated by Fisher's method. Interestingly, the 10 most significantly affected pathways (Fig. 5, Supporting Information Table 12) were all related to cancer, epithelial cell adhesion and *HER* signaling.

Discussion

During the past decade, knowledge about the genomic landscape of different tumor types including lung carcinoma has emerged. Since tobacco smoking is the major cause of lung carcinoma, one of the limitations of these studies is that most of the reported cohorts contain a vast majority of smokers. To specifically elucidate the genomic landscape and features of NSCLC in never smokers, we prospectively selected a cohort of 46 never-smoking lung cancer patients and performed an integrative WES and low-coverage WGS study. This led to several interesting findings.

First of all, we observed quantitative differences between the NSCLC genome in never smokers compared to smokers. We could not only confirm that the number of somatic mutations is strongly reduced in never smokers compared to current smokers, but we also identified that the number of chromosomal breakpoints and the fraction of the genome with an aberrant copy-number was significantly reduced in never smokers. Classification of somatic mutations into the six substitution type patterns indicated that the mutational pattern in NSCLC of never smokers is dominated by G:C>A:T mutations, which is in contrast to the pattern observed in smokers and which is dominated by G:C>T:A mutations or more in general, the dominant substitutions are transitions in never smokers and transversions in smokers. Unsupervised clustering of a cohort of NSCLC on these somatic mutational patterns results in a cluster with smokers and a cluster with never smokers.

Next, we identified the collective somatic mutational signature which is not only based on the type of substitutions but also on the sequence context for each substitution. This collective signature of NSCLC from never smokers is composed out of six mutational signatures, each underlying a mutagenic process. Among these contributing signatures, Signature 3 is contributing 14% to the collective signature and is related to perturbation in the mechanism of DNA repair by HR. This is further supported by the detection of somatic mutations in HR genes such as *ATM*, *BRCA1/2* and *MRE11A*, and also focal deletions of *BRCA2* and *FANCA*. In our entire cohort, we observe 34 patients (74%) with an alteration in the HR genes. We also observe a frequent focal deletion (n = 28, 62%) of *POLE*, which is involved in DNA synthesis and base excision repair. Both observations stress the importance of damaged DNA repair mechanisms in the carcinogenesis of lung cancer in nonsmokers. Interestingly, unsupervised clustering of a cohort of NSCLC based on somatic mutational signatures, distinguishes between smokers and never smokers demonstrating a different etiology in these groups. While the etiology of NSCLC in smokers is due to tobacco smoke, our results indicate that NSCLC in never smokers is not due to passive smoking as the smoking signature (Signature 4) is not observed in these patients.

The higher prevalence of women in our (59%) and other never smoker lung cancer cohorts (TCGA LUAD 73%, Asian 62%) has not yet been explained. A strong candidate causal factor for such imbalance could be estrogen exposure, either through its tumor-promoting effects or the genotoxic effects of certain estrogen adduct forming genotoxic catabolites.[44] However, neither the Estrogen-Dependent Breast Cancer Signaling pathway nor the Estrogen receptor pathway was enriched for somatic mutations. We also did not observe mutational hallmarks reflecting the effect of estrogen catabolites in never smokers. We therefore failed to confirm a potential etiologic role for estrogens in lung cancer development of never smokers.

Previously observed driver mutations in lung cancer including but not limited to *EGFR*, *TP53* and *ERBB2* are also recurrently mutated in our and other cohorts of never smokers but interestingly at a different frequency. The frequency of *EGFR* and *ERBB2*, two targets in personalized medicine is higher in never smokers while the frequency of *TP53*, a negative prognostic factor is reduced. Like *EGFR*, *MET* is also significantly focally amplified in never smokers, which is not observed in NSCLC of smokers and uncommon in untreated patients. Patients with a *MET* amplification respond to crizotinib. These observations might give rise to a more favorable disease progression in never smokers.

One of the advantages of our cohort composed of never smokers, is that the amount of passenger mutations is strongly reduced, which makes the identification of less abundant driver mutations less cumbersome. We could identify *TSC22D1* as a potential driver gene in NSCLC. This gene is significantly mutated in NSCLC in never smokers and is

known to be involved in the development of pulmonary adenoma in mice.

In conclusion, our unique prospectively selected cohort of NSCLC in never smokers revealed (*i*) significantly less mutations, breakpoints and copy-number instability in never smokers compared to smokers, (*ii*) a different six-substitution type pattern and 96-substitution type signature both discriminating between smokers and never smokers and indicating a different etiology of NSCLC, (*iii*) well-known clinically relevant mutations in genes such as *EGFR*, *TP53* and *ERBB2* are observed at different more beneficial, mutation rates in NSCLC in never smokers compared to smokers and (*iv*) *TSC22D1* can be considered as a new driver mutation in NSCLC from nonsmokers.

Acknowledgements

Our study was supported by the National Cancer Plan Project 29-03 (JDG and DL). The computational resources and services used in this work were provided by the Flemish Supercomputer Center (VSC), funded by the Hercules Foundation, Department of Economy, Science and Innovation from the Flemish government (EWI).

Scientific Fund W Gepts (JDG and LDC) of the UZ Brussel.

References

1. Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat Rev Cancer* 2003;3:733–44.
2. Samet JM, Avila-Tang E, Boffetta P, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clin Cancer Res* 2009;15:5626–45.
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers — a different disease. *Nat Rev Cancer* 2007;7:778–90.
4. Gridelli C, Rossi A, Carbone DP, et al. Non-small-cell lung cancer. *Nat Rev Dis Primers* 2015; 1:15009.
5. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol* 2015;10:1243–60.
6. Schiller JH, Harrington D, Belani CP, et al. Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N Engl J Med* 2002;346:92–8.
7. Melosky B. Treatment algorithms for patients with metastatic non-small cell, non-squamous lung cancer. *Front Oncol* 2014;4:256.
8. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med* 2016;375:1823–33.
9. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med* 2015;373: 1627–39.
10. Zhu Q, Zhang S, Ding X, et al. Driver genes in non-small cell lung cancer: Characteristics, detection methods, and targeted therapies. *Oncotarget* 2017;8:57680–92.
11. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
12. Collisson E, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543–50.
13. Couraud S, Debieuvre D, Moreau L, et al. No impact of passive smoke on the somatic profile of lung cancers in never-smokers. *Eur Respir J* 2015; 45:1415–25.
14. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012;150: 1121–34.
15. Krishnan VG, Ebert PJ, Ting JC, et al. Whole-genome sequencing of Asian lung cancers: second-hand smoke unlikely to be responsible for higher incidence of lung cancer among Asian never-smokers. *Cancer Res* 2014;74:6071–81.
16. Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 2012;150: 1107–20.
17. Rosenthal R, McGranahan N, Herrero J, et al. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* 2016;17:31.
18. Gehring JS, Fischer B, Lawrence M, et al. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* 2015;31:3673–5.
19. Scheinin I, Sie D, Bengtsson H, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Biol* 2014;24:2022–32.
20. Van Loo P, Nordgard SH, Lingjærde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010;107:16910–5.
21. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
22. Martin P, Heavner DL, Nelson PR, et al. Environmental tobacco smoke (ETS): a market cigarette study. *Environ Int* 1997;23:75–90.
23. Campbell JD, Alexandrov A, Kim J, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* 2016;48:607–16.
24. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2:401–4.
25. Couraud S, Souquet P-J, Paris C, et al. Bio-CAST/IFCT-1002: epidemiological and molecular features of lung cancer in never-smokers. *Eur Respir J* 2015;45:1403–14.
26. Dees ND, Zhang Q, Kandoth C, et al. MuSiC : Identifying mutational significance in cancer genomes. *Genome Res* 2012;22:1589–98.
27. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499: 214–8.
28. Nakamura M, Kitaura J, Enomoto Y, et al. Transforming growth factor-b-stimulated clone-22 is a negative-feedback regulator of Ras/Raf signaling: implications for tumorigenesis. *Cancer Sci* 2012; 103:26–33.
29. Chen WC, Wang CY, Hung YH, et al. Systematic analysis of gene expression alterations and clinical outcomes for long-chain acyl-coenzyme a synthetase family in cancer. *PLoS One* 2016;11:1–23.
30. Berndt A, Cario CL, Silva K a, et al. Identification of Fat4 and Tsc22d1 as novel candidate genes for spontaneous pulmonary adenomas. *Cancer Res* 2011;71:5779–91.
31. Arcila ME, Chaft JE, Nafa K, et al. Prevalence, clinicopathologic associations, and molecular spectrum of ERBB2 (HER2) tyrosine kinase mutations in lung adenocarcinomas. *Clin Cancer Res* 2012;18:4910–8.
32. Buttitta F, Barassi F, Fresu G, et al. Mutational analysis of the HER2 gene in lung tumors from Caucasian patients: mutations are mainly present in adenocarcinomas with bronchioloalveolar features. *Int J Cancer* 2006;119:2586–91.
33. Shigematsu H, Takahashi T, Nomura M, et al. Somatic mutations of the HER2 kinase domain in lung adenocarcinomas. *Cancer Res* 2005;65: 1642–6.
34. Stephens P, Hunter C, Bignell G, et al. Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* 2004;431:525–6.
35. De Grève J, Teugels E, Geers C, et al. Lung cancer clinical activity of afatinib (BIBW 2992) in patients with lung adenocarcinoma with mutations in the kinase domain of HER2/neu. *Lung Cancer* 2012;76:123–7.
36. De Grève J, Moran T, Graas M, et al. Lung cancer phase II study of afatinib, an irreversible ErbB family blocker, in demographically and genotypically defined lung adenocarcinoma. *Lung Cancer* 2015;88:63–9.
37. Ou S-HI, Kwak EL, Siwak-Tapp C, et al. Activity of Crizotinib (PF02341066), a dual Mesenchymal-epithelial transition (MET) and anaplastic lymphoma kinase (ALK) inhibitor, in a non-small cell lung cancer patient with De novo MET amplification. *J Thorac Oncol* 2011;6:942–6.
38. Camidge DR, Ou S-HI, Shapiro G, et al. Efficacy and safety of crizotinib in patients with advanced c-MET-amplified non-small cell lung cancer (NSCLC). *J Clin Oncol* 2014;32:abstr 8001.
39. Engelman JA, Zejnullahu K, Mitsudomi T, et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* 2007;438:1039–43.
40. Turke AB, Zejnullahu K, Wu Y, et al. Pre-existence and clonal selection of MET

amplification in EGFR mutant NSCLC. *Cancer Cell* 2011;17:77–88.
41. Bowman RV, Yang IA, Semmler ABT, et al. Epigenetics of lung cancer. *Respirology* 2006;2006 (February):355–65.
42. Toyooka S, Maruyama R, Toyooka KO, et al. Smoke exposure, histologic type and geography-related differences in the methylation profiles of non-small cell lung cancer. *Int J Cancer* 2003;103: 153–60.
43. Dang CV, Reddy EP, Shokat KM, et al. Drugging the 'undruggable' cancer targets. *Nature* 2017;17:502–8.
44. Cavalieri E, Rogan E. The molecular etiology and prevention of estrogen-initiated cancers. *Mol Aspects Med* 2014;36:1–55.