Tobacco smoking and somatic mutations in human bronchial epithelium

Tobacco smoking causes lung cancer1–3, a process that is driven by more than 60 carcinogens in cigarette smoke that directly damage and mutate DNA4,5. The profound effects of tobacco on the genome of lung cancer cells are well-documented6–10, but equivalent data for normal bronchial cells are lacking. Here we sequenced whole genomes of 632 colonies derived from single bronchial epithelial cells across 16 subjects. Tobacco smoking was the major influence on mutational burden, typically adding from 1,000 to 10,000 mutations per cell; massively increasing the variance both within and between subjects; and generating several distinct mutational signatures of substitutions and of insertions and deletions. A population of cells in individuals with a history of smoking had mutational burdens that were equivalent to those expected for people who had never smoked: these cells had less damage from tobacco-specific mutational processes, were fourfold more frequent in ex-smokers than current smokers and had considerably longer telomeres than their more-mutated counterparts. Driver mutations increased in frequency with age, affecting 4–14% of cells in middle-aged subjects who had never smoked. In current smokers, at least 25% of cells carried driver mutations and 0–6% of cells had two or even three mutated counterparts. Driver mutations increased in frequency with age, affecting 4–14% of cells in middle-aged subjects who had never smoked. In current smokers, at least 25% of cells carried driver mutations and 0–6% of cells had two or even three drivers. Thus, tobacco smoking increases mutational burden, cell-to-cell heterogeneity and driver mutations, but quitting promotes replenishment of the bronchial epithelium from mitotically quiescent cells that have avoided tobacco mutagenesis.

Lung cancer kills more people globally than any other cancer, and 80–90% of those deaths are attributable to tobacco exposure1,2. Our model for how tobacco causes lung cancer emphasizes direct mutagenesis from the numerous (more than 60) carcinogens in cigarette smoke4,5, combined with indirect effects such as inflammation, immune suppression and infection. As recognized first in the sequencing of the TP53 gene6 and more recently in genome-wide sequencing of lung cancers6–10, tobacco exposure leads to both an increase in somatic mutational burden and an altered spectrum of mutations. Clones of lung cancer cells from a smoker typically have tens of thousands of somatic mutations6–7,9; of these, a small handful—probably fewer than 20—drive the biology of the tumour10–11.

Epidemiological studies have quantified the relationships between lung cancer and duration of smoking, intensity of smoking, type of smoking and timing of smoking cessation11–13. Interpreting these observations from population cohorts in terms of the molecular basis for tobacco carcinogenesis is challenging. Under a model in which lung cancer requires $n$ driver mutations, an exposure that, say, increases $k$-fold should increase incidence by around $k^n$ across a range of growth patterns12. However, in a paradox first noted in 197118, the dose–response relationship between the number of cigarettes smoked per day and the risk of lung cancer is linear13,14—that is, $k^1$—or, at most, weakly quadratic15. The benefits of smoking cessation likewise do not fit into multistage models of cancer in a straightforward manner13. By stopping smoking in middle age or earlier, smokers avoid most of the risk of tobacco-associated lung cancer. This benefit begins to emerge almost immediately and accrues steadily with time16. Of two people who smoked the same total number of cigarettes across their lifetime, why the person with longer duration of cessation should have a lower risk of lung cancer is difficult to explain if tobacco induces carcinogenesis exclusively by increasing the mutational burden.

Sequencing single-cell-derived colonies

We recruited 16 individuals to assess the landscape of somatic mutations in normal bronchial epithelium: 3 children, 4 individuals who had never smoked (‘never-smokers’), 6 ex-smokers and 3 current smokers (Supplementary Table 1). For ethical reasons, samples could only be

1Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. 2Lungs For Living Research Centre, UCL Respiratory, University College London, London, UK. 3Department of Thoracic Medicine, University College London Hospital, London, UK. 4Department of Pathology and Tumor Biology, Kyoto University Hospital, Kyoto, Japan. 5CRUK Lung Cancer Centre of Excellence, UCL Cancer Institute, University College London, London, UK. 6Stem Cell Institute, University of Cambridge, Cambridge, UK. 7These authors contributed equally: Kenichi Yoshida, Kate H. C. Gowers. 8These authors jointly supervised this work: Sam M. Janes, Peter J. Campbell. *e-mail: s.janes@ucl.ac.uk; pc8@sanger.ac.uk

https://doi.org/10.1038/s41586-020-1961-1

Received: 7 June 2019
Accepted: 29 December 2019
Published online: 29 January 2020

Tobacco smoking causes lung cancer1–3, a process that is driven by more than 60 carcinogens in cigarette smoke that directly damage and mutate DNA4,5. The profound effects of tobacco on the genome of lung cancer cells are well-documented6–10, but equivalent data for normal bronchial cells are lacking. Here we sequenced whole genomes of 632 colonies derived from single bronchial epithelial cells across 16 subjects. Tobacco smoking was the major influence on mutational burden, typically adding from 1,000 to 10,000 mutations per cell; massively increasing the variance both within and between subjects; and generating several distinct mutational signatures of substitutions and of insertions and deletions. A population of cells in individuals with a history of smoking had mutational burdens that were equivalent to those expected for people who had never smoked: these cells had less damage from tobacco-specific mutational processes, were fourfold more frequent in ex-smokers than current smokers and had considerably longer telomeres than their more-mutated counterparts. Driver mutations increased in frequency with age, affecting 4–14% of cells in middle-aged subjects who had never smoked. In current smokers, at least 25% of cells carried driver mutations and 0–6% of cells had two or even three drivers. Thus, tobacco smoking increases mutational burden, cell-to-cell heterogeneity and driver mutations, but quitting promotes replenishment of the bronchial epithelium from mitotically quiescent cells that have avoided tobacco mutagenesis.
obtained from subjects who underwent a bronchoscopy for clinical indications. The never-smokers and current smokers had a bronchoscopy to investigate changes that were eventually diagnosed as benign. Of the ex-smokers, two had a previous cancer treated with curative intent, and five had a carcinoma in situ or invasive squamous cell carcinoma that was the indication for the bronchoscopy. The children in the cohort underwent a bronchoscopy for investigation or follow-up of congenital anomalies: all had normal bronchial epithelium.

Samples of airway epithelium were obtained from biopsies or brushings of main or secondary bronchi. These were dissociated into single cells, and epithelial cells positive for epithelial cell adhesion molecule (EpCAM) were flow-sorted (one to a well) onto mouse feeder cells allowing basal cell attachment and growth (Extended Data Fig. 1a). Each cell was independently cultured to obtain single-cell-derived colonies that expressed the transcripts expected for basal cells of pseudostratified bronchial epithelium (Extended Data Fig. 1b). Around 15–40% of flow-sorted cells typically produced colonies (Extended Data Fig. 1c), confirming that the sequenced cells were drawn from a prevalent and representative population of epithelial cells. Colonies underwent whole-genome sequencing to an average coverage of 16× (Supplementary Table 2, Extended Data Fig. 2a, b). Using a xenograft pipeline to flag non-human sequencing reads, somatically acquired mutations were identified from reads specific to the human genome. In nearly all colonies, the variant allele fraction (VAF) of mutations was around 50% on average, which is consistent with contamination-free colonies derived from a single bronchial cell (Extended Data Fig. 2c). To remove variants that had possibly been acquired in vitro, we excluded mutations with a VAF of less than 30% that were present in only a single colony (Extended Data Fig. 2c). Occasionally colonies had a low mean VAF (Extended Data Fig. 2d), consistent with seeding by two bronchial cells; these colonies were excluded from downstream analyses. We estimated that a sequencing depth of 8× gave a sensitivity for variants of 70–75%, and this increased to more than 95% at a depth of 15× (Extended Data Fig. 2e). The majority of colonies had a sequencing depth greater than 15×, and we set a minimum cut-off of 8× for inclusion.

The final dataset comprises catalogues of somatic mutations from the whole genomes of 632 single bronchial cells. Five patients had a squamous cell carcinoma or carcinoma in situ, three of which we also sequenced. Normal basal cells from these patients shared no clonal relationships with the carcinomas, and we found no systematic differences in mutational burden between normal cells in the vicinity of carcinoma in situ lesions and cells in regions that were histologically normal (Extended Data Fig. 2f).

**Mutational burden**

The burden of somatic substitutions per cell showed considerable heterogeneity both across the cohort and within individual patients (Fig. 1a). Using linear mixed-effects (LME) models, we assessed factors that influenced the mutational burden (Supplementary Code). Single-base substitutions increased significantly with age, at an estimated rate of 22 per cell per year (95% confidence interval (CI), 20–25; \( P = 10^{-8} \); Fig. 1b). Previous or current smoking significantly increased the mean burden of substitutions (\( P = 0.0002 \)) by an estimated 2,330 per cell (95% CI, 1,180–3,480) in ex-smokers and 5,300 per cell (95% CI, 3,660–6,930) in current smokers.

The effects of age and smoking were expected but, more surprisingly, smoking also markedly increased the variability in mutational burden from cell to cell, even within the same individual. Among closely collocated cells from a small biopsy of normal airway from a given subject, the estimated standard deviation was 2,350 per cell for ex-smokers and 2,100 per cell for current smokers, compared with 140 per cell for children and 290 per cell for adult never-smokers (\( P = 10^{-16} \) by LME for within-subject heterogeneity of variance across smoking categories). There was also heterogeneity between subjects: the estimated standard deviation in mean substitution burden across individuals was 1,200 per cell for ex-smokers and 1,260 per cell for current smokers, compared to 90 per cell for non-smokers (\( P = 10^{-8} \) by LME for heterogeneity of variance).

Although most cells in ex-smokers or current smokers had a considerably higher substitution burden than cells in never-smokers, a fraction of cells in these patients had burdens within the range expected for never-smokers of an equivalent age (Fig. 1c). For many of these patients, the distribution of mutational burden was distinctly bimodal, with one mode in the near-normal range and the other mode exhibiting a substantially increased mutational burden (Extended Data Fig. 3a).

Notably, although cells with a near-normal mutational burden were rarely present in current smokers, their relative frequency was on average fourfold higher in ex-smokers (95% CI, 2.0–7.5-fold; \( P = 3 \times 10^{-4} \)) by log-linear model, typically accounting for 20–40% of all cells studied. Colonies with a near-normal mutational burden expressed the same set of airflow basal cell genes as did colonies with an increased mutational burden, and had the same tightly associated, cobbled architecture in culture (Extended Data Fig. 3b, c), confirming that they derived from bronchial epithelial cells.

Among current and ex-smokers, we found that mutational burden was not significantly correlated with the duration of cigarette smoking or the number of cigarettes smoked per day, even if near-normal cells were excluded. However, the small numbers of subjects and large within-subject heterogeneity limits our statistical power for this analysis, and definitive analysis will require much larger sample sizes.

Insertions and deletions (indels) showed similar associations as substitutions, increasing steadily with age (0.7 indels per cell per year; 95% CI, 0.6–0.8; \( P = 10^{-12} \)) and tobacco smoking (101 extra indels per cell in smokers; 51 in ex-smokers; \( P = 0.001 \); Extended Data Fig. 4a). Generally, the normal bronchial epithelial cells had few copy-number changes or structural variants (Extended Data Fig. 4b) —this represents a qualitative difference from lung cancers, which tend to have large numbers of structural abnormalities13,14. There were occasional examples of more-complex structural events in the bronchial epithelial cells, including chromoplexy (Extended Data Fig. 4c) and even chromothripsis in a cell from a child (Extended Data Fig. 4d). The latter is particularly interesting, given that recent data suggest that driver-gene fusions in lung adenocarcinoma can arise through complex structural events that occur early in life15.

**Mutational signatures**

A range of mutational processes operate in lung cancers, driven both by the exogenous carcinogens present in tobacco smoke and by endogenous DNA damage. These processes leave characteristic signatures in the genome16. We built phylogenetic trees for each patient, and applied a Bayesian de novo mutational-signature discovery algorithm to mutational histories assigned to each branch. We also included samples from squamous cell lung cancers17 and control samples cultured in vitro18 in the signature analysis to maintain comparability with previous analyses (Fig. 2). Few mutations in our samples (typically 10–30 per cell) were attributed to SBS-18, the signature that accounted for all variants in the control samples18, which confirmed that mutations acquired in vitro were minimal in our dataset. Similar results emerged using a different mutational-signature algorithm19 (Extended Data Fig. 5a–c).

A large proportion of mutations in all subjects was attributed to the endogenous mutational signature SBS-5, which accumulated linearly with age (Fig. 2c, d). As reported previously13, the absolute number of mutations attributed to this signature was higher in those with a smoking history (ex-smokers 1,140 per cell, 95% CI, 590–1,700; current smokers 2,200 per cell, 95% CI, 1,590–2,810; \( P < 10^{-18} \)). Signature SBS-1—which comprises C>T mutations at CpG dinucleotides—contributed larger proportions of mutations in children than adults, but the absolute numbers of SBS-1-attributed mutations continued to increase linearly with age through adulthood (Fig. 2c, d). Presumably, then,
SBS-1 is enriched during early lung development and continues steadily throughout life, but other signatures become proportionally more active in adulthood. A novel signature (Sig-A; Fig. 2b) was universally present across samples. It has some resemblance to SBS-5, and likewise increased linearly with age.

Signatures SBS-2 and SBS-13, which are caused by mutagenesis mediated by APOBEC3A or APOBEC3B, showed striking heterogeneity: they were mostly absent from bronchial cells, but occasionally contributed hundreds of mutations in an individual cell, even in children. This activity appeared to be temporally restricted: individual branches of a phylogenetic tree had high proportions of SBS-2 or SBS-13 despite their absence from antecedent and descendent branches (Fig. 3a, Extended Data Fig. 6). This implies that the episodic activity of APOBEC-mediated mutagenesis observed in cell lines21 extends to somatic cells in vivo, as the proportion of mutations attributed to APOBEC enzymes on a given branch of the phylogenetic tree does not predict past or future rates of mutagenesis in that lineage.

Three substitution signatures were largely restricted to current or ex-smokers. Signature SBS-4 was expected—this is the predominant signature in lung cancers from smokers7,8 and is recapitulated by in vitro exposure to polycyclic aromatic hydrocarbons19. SBS-16 comprised 5–15% of mutations in several current or ex-smokers, but was absent from never-smokers. This signature, with its distinctive pattern of transcription-coupled damage and repair22 (Extended Data Fig. 5d), correlates with alcohol and tobacco exposure in hepatocellular carcinomas8,23, but has not been linked with tobacco exposure in lung cancers previously.

A new mutational signature (Sig-B) was extracted, which comprised predominantly T>A and T>C mutations and was evident only in patients with a history of smoking (Fig. 2b). The signature was mostly present at low rates, but in one ex-smoker it contributed up to 15% of mutations per cell. We found a strong transcriptional strand bias, whereby the transcribed strand showed decreased rates of mutation at the adenine in the
A-T pairing. This is consistent with in vitro data that show that purines are more reactive than pyrimidines with mutagens in tobacco smoke. As described above, an unexpectedly high fraction of cells in ex-smokers had a near-normal mutational burden. These cells had considerably lower proportions of SBS-4 mutations than cells with an increased mutational burden in the same patients. Instead, the distribution of signatures in these near-normal cells resembled that seen in never-smokers, with prominent endogenous signatures such as SBS-5, SBS-1 and with each group of eight by whether A, C, G or T is 3′ to the mutated base. The activity of the mutational signature on the untranscribed strand is shown in a pale colour; on the transcribed strand it is shown in a darker colour. The blue shaded area represents the 95% CI for the fitted line. The black line represents the fitted effect of age, which was estimated from LME models after correction for smoking status and within-patient correlation structure. The blue shaded area represents the 95% CI for the fitted line. The quoted P values for the fixed effects of age and smoking are derived from the full LME models. Effect size of smoking status

**Fig. 2 | Mutational signatures in normal bronchial epithelium.** a. Stacked bar plot showing the proportional contribution of mutational signatures to single-base substitutions across the n = 632 colonies from normal bronchial cells, extracted using a hierarchical Dirichlet process (HDP). Within each patient, colonies are sorted from left to right by increasing mutational burden (bar chart in dark grey above coloured signature-attribution stacks). The dashed black vertical lines in current and ex-smokers denote the cut-off between cells with a near-normal and an increased mutational burden. b. Trinucleotide context spectrum on transcribed and untranscribed strands of two new SBS signatures (Sig-A and Sig-B). The six substitution types are shown across the top. Within each substitution type, the trinucleotide context is shown as four sets of eight bars, grouped by whether an A, C, G or T, respectively, is 5′ to the mutated base.

Nature | Vol 578 | 13 February 2020 | 269
Fig. 3 | Driver mutations in normal bronchial epithelial cells. a, Phylogenetic trees showing clonal relationships among normal bronchial cells in three representative subjects. Branch lengths are proportional to the number of mutations (x axis) specific to that clone or subclone. Each branch is coloured by the proportion of mutations on that branch that are attributed to the various SBS signatures. The driver mutations that were identified in each branch are also shown (black, SBS; red, indel). b, Total number of colonies with mutations (left) and number of unique mutations (right) in key cancer genes across the sample set (n = 632). ** represents genes that are significant (q < 0.05 by dSNdScv) when correction for multiple-hypothesis testing is applied across known driver genes in lung cancers and normal squamous tissues (exact q values are provided in Supplementary Table 4). c, Fraction of colonies with 0, 1, 2 or 3 driver mutations across the 16 subjects. d, Distribution of driver mutations across colonies in the cohort, coloured by type of mutation. Loss of heterozygosity (LOH) that affects driver mutations is also shown. e, Frequency of driver mutations that are shared by more than one colony in a patient (dark blue) versus those found in a single colony (light blue) across different cancer genes.

and Sig-A. Phylogenetically, cells with a near-normal mutational burden showed polyclonal origins (Fig. 3a, Extended Data Fig. 6), suggesting that they do not arise from the expansion of a single ancestral cell.

Signatures of indels and double-base substitutions that were observed in normal bronchial epithelium matched those extracted from lung cancers24 and those generated in vitro by exposure of cells to polycyclic aromatic hydrocarbons25 (Extended Data Figs. 7, 8). A history of tobacco smoking was particularly associated with a signature of double-base substitutions at CpC (equivalently GpG) dinucleotides—a finding that is in accordance with the high rates of C>A (G>T) single-base substitutions in SBS-4. Similarly, tobacco exposure was associated with an indel signature of single-base deletions of cytosines (guanines) in our dataset. Together, these data suggest that the propensity of polycyclic aromatic hydrocarbons in tobacco smoke to bind guanine nucleotides can result in a range of mutation types even in normal bronchial epithelial cells, including single-base substitutions, dinucleotide substitutions and small indels.

**Driver mutations**

To assess whether any mutations are under positive selection in normal bronchial epithelium, we applied an algorithm, dSNdScv, which identifies and quantifies the number of excess non-synonymous mutations compared with the number expected from the rate of synonymous mutations.
variants, correcting for local variation in mutation rates\(^1\). With hypothesis testing applied across all coding genes, three were significant: NOTCH1 (20 unique non-synonymous variants; \(q = 1 \times 10^{-5}\)); TP53 (7 unique non-synonymous variants; \(q = 2 \times 10^{-5}\)); and ARID2 (7 unique non-synonymous variants; \(q = 4 \times 10^{-5}\)) (Fig. 3b). When hypothesis testing was restricted to genes that are mutated in lung cancers\(^{12,13,18,25,26}\) and normal squamous tissues\(^{27-30}\), FAT1, PTEN, CHEK2 and ARID1A were also significant, showing the expected patterns of protein-truncating mutations (Supplementary Table 3–5, Extended Data Fig. 9a). This set of significant genes closely resembles those under positive selection in squamous cell lung cancers\(^{13,18}\) and other normal squamous tissues\(^{27-30}\).

Driver mutations were more frequent in patients with a history of tobacco smoking (Fig. 3c, Extended Data Fig. 9b). No candidate driver mutations were identified in cells from children, and 4–14% of cells in adult never-smokers had drivers; by contrast, in current smokers, at least 25% of cells carried at least one driver. Furthermore, a small fraction of cells in smokers had two or even three coding driver point mutations (Fig. 3d) – as many as is seen in some lung cancers\(^{12}\). We used generalized LME models to quantify these effects (Supplementary Code). Driver mutations were significantly more frequent in individuals with a smoking history and showed an increase of 1.5-fold (95% CI, 1.2–2.1; \(P = 0.004\)) – a pattern reminiscent of the increasing number of driver mutations with age in the oesophagus\(^{18,29}\). Finally, the number of driver mutations doubled on average for every 5,000 extra somatic mutations per cell, independent of the other variables (95% CI, 1.4–2.7; \(P = 0.0003\)).

Layering driver mutations onto phylogenetic trees revealed that driver mutations occurred throughout molecular time (Fig. 3a, Extended Data Fig. 6). Mutations in TP53 were much more likely to be shared by two or more sequenced cells (Fig. 3e), however, suggesting that they either occur earlier in molecular time or drive larger clonal expansions.

**Fig. 4** | Relationship of telomere length with mutational burden. Split by smoking status, the graphs show the relationship between telomere length (x-axis) and mutational burden (y-axis) for colonies with less than 10% contamination from the mouse feeder cells (n = 398 colonies). Individual cells are shown as points and fitted lines for each patient are shown as coloured lines (slopes were estimated using LME models). The difference in slopes according to smoking status is highly significant (\(P = 0.0009\) for interaction term; LME models). One outlying cell from an ex-smoker, which had more than 10,000 mutations, was excluded from the plot to improve visualization.

**Telomere lengths**

To assess historic mitotic activity, we estimated telomere lengths from the sequencing data (Fig. 4). Bronchial cells from children had longer telomeres than did cells from adults (Extended Data Fig. 10), as expected, and telomere length showed no correlation with mutational burden in children. Among never-smokers, there was also minimal correlation between mutational burden and telomere length. In current smokers, however – and especially in ex-smokers – there was a strong inverse relationship between telomere length and mutational burden, independent of the number of driver mutations (\(P = 0.0009\) for interaction between smoking status and telomere length by LME models; Supplementary Code). In particular, the cells with a near-normal mutational burden in ex-smokers had considerably longer telomeres than did their more-mutated counterparts, suggesting that they have historically undergone fewer cell divisions.

**Discussion**

The simplicity of the notion that cigarette smoking causes lung cancer through its mutagenic effects belies the underlying complexity of how tobacco shapes clonal dynamics, mutation acquisition and the selective environment in the bronchus. As expected, exposure to tobacco smoke increases the number of somatic mutations (by an average of a few thousand mutations per normal bronchial cell); the excess mutations are attributable to signatures of carcinogens in cigarette smoke; and the increased mutational burden generates more driver mutations. What is unexpected, however, is the pronounced within-patient variation in mutational burden among smokers: cells from one small biopsy from the bronchus can vary tenfold in their mutational burden, from 1,000 to over 10,000 mutations per cell.

Our cohort may be affected by recruitment bias, as samples could only ethically be obtained from individuals who underwent a clinically indicated bronchoscopy. Nonetheless, such a recruitment bias cannot explain the considerable within-patient variance in mutational.
burden, and this finding probably therefore applies to smokers more generally. Understanding how heterogeneity in mutational burden among competing cells contributes to clonal evolution will be important for refining our models of lung cancer development, which usually assume that the effects of carcinogens are homogeneous across a population of cells. We recently described similar heterogeneity in tobacco-induced mutagenesis among neighbouring clones within non-malignant liver, suggesting that this phenomenon is not restricted to bronchial epithelium.

We found that a qualitatively distinct population of bronchial epithelial cells with a near-normal mutational burden exists in subjects with a history of smoking. These cells have the same mutational burden as age-matched never-smokers; they have low proportions of signatures from tobacco carcinogens and longer telomeres than more-mutated cells; and occur at a fourfold higher frequency in ex-smokers compared to current smokers. These cells are clearly protective against cancer—lung cancers that emerge in ex-smokers do not have a near-normal mutational burden, instead typically showing the high mutational burden that is associated with tobacco-induced signatures.

Two points remain unclear: how these cells have avoided the high rates of mutations that are exhibited by neighbouring cells, and why this particular population of cells expands after smoking cessation. The longer telomeres of these cells imply that cells with a near-normal burden have undergone fewer cell divisions, and therefore potentially represent recent descendants of quiescent stem cells. Although they remain elusive in human lung, quiescent stem cells have been identified through lineage tracing in mouse models, and have been shown to occupy a protected niche in submucosal glands and expand after lung injury. A physically protected niche could explain how such stem cells avoid exposure to tobacco carcinogens, but so too could mitotic quiescence itself, as replication is required to convert added DNA bases to mutations.

It is tempting to assume that the expansion of cells with a near-normal burden after smoking cessation arises through better fitness in the altered selection landscape—perhaps because these cells have longer telomeres or fewer mutations, or because aberrant NOTCH or TP53 signalling confers less advantage in the absence of tobacco smoke. These explanations notwithstanding, the apparent expansion of the near-normal cells could represent the expected physiology of a two-compartment model in which relatively short-lived proliferative progenitors are slowly replenished from a pool of quiescent stem cells, but the progenitors are more exposed to tobacco carcinogens.

Only in ex-smokers would the difference in mutagenic environment be sufficient to distinguish newly produced progenitors from long-term occupants of the bronchial epithelial surface.

Epidemiological studies show that the health benefits of stopping smoking begin immediately, accrue with time since cessation and are evident even after quitting late in life. That these benefits could be facilitated by replenishment of the bronchial epithelium with cells that are essentially impervious to decades of sustained cigarette smoking attests to the resilience and regenerative capacity of the lungs. The message for public health is that stopping smoking—at any age—does not slow the accumulation of further damage, but can also reawaken cells that have not been damaged by past lifestyle choices.
Methods

Data reporting
No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Subjects
Subjects were recruited at University College London Hospitals (UCLH) or Great Ormond Street Hospital (GOSH) and gave written informed consent with approval of the Research Ethics Committee (REC reference 06/Q0505/12 and 11/LO/152, respectively). Details of the patients studied are listed in Supplementary Table 1. All patients underwent bronchoscopy as part of their clinical care. In adults, the bronchoscopy procedure was performed for diagnostic or surveillance indications; in children, it was undertaken for investigational procedures on congenital tracheal abnormalities. For five patients with squamous cell carcinomas or carcinoma in situ, biopsy of normal bronchial tissue was taken from a site distant from the tumour.

Single-cell-derived colonies
Endobronchial biopsies were dissociated using 16 U/ml dispase in RPMI for 20 min at room temperature. The epithelium was dissociated away from the underlying stroma and fetal bovine serum (FBS) was added to a final concentration of 10%. Both epithelium and stroma were combined and digested in 0.1% trypsin/EDTA at 37 °C for 30 min. The solution was neutralized with FBS to a final concentration of 10% and added to the neutralized dispase solution. Cells were passed through a 100-μm cell strainer before staining. EPCAM-APC (Biolegend 324208, 1:50) antibodies and DAPI (1 μg/ml). For 555483, 1:200), anti-CD31-PE (BD Pharminogen 555446, 1:200), anti-CD45-PE-Cy7 (BD Pharminogen, 1:200), and anti-EpCAM-APC (Biolegend 324208, 1:50) antibodies and DAPI (1 μg/ml). For endobronchial brushings, no dissociation was carried out and the cell suspension was passed through a 100-μm cell strainer before staining.

Clonality of samples
To ensure that each sample was single-cell-derived, we visually inspected the distribution of VAFs of mutations: 632 clones had VAFs distributed around 50%, confirming that they were derived from a single cell, but 10 clones had lower allele fractions, suggesting that these colonies were oligoclonal (Extended Data Fig. 2d). These samples were removed from further analyses owing to lower estimated sensitivity, as described later (Extended Data Fig. 2e).

Single-base-substitution calling
Single-base substitutions were called using the Cancer Variants through Expectation Maximization (CaVEMan) algorithm with copy-number options of major copy number 5, minor copy number 2 and normal contamination 0.1. To allow the discovery of early embryonic mutations, we ran CaVEMan using an unmatched normal control. In addition to the default ‘PASS’ filter, we removed variants with a median alignment score (ASMD) < 120 and those with a clipping index (CLPM) > 0, to remove mapping artefacts. Variants identified in the mouse feeder fibroblast DNA sample were also removed, if they persisted in the call-set. Subsequently, for every mutation identified in any colonies from each patient, we counted the number of mutant and wild-type reads in all bronchial DNA samples from the same patient using the bam2R function of the R package deepSNV44, for which bases with ≥30 base quality and sequencing reads with ≥30 mapping quality were used. Further filters described below were applied to identify true somatic mutations and separate them from either germline variants or recurrent sequencing errors.

Removing germline variants (binomial filter)
We fitted a binomial distribution to the total variant counts and total depth at each single-base substitution site across all samples from one patient. To differentiate somatic variants from germline variants, we used a one-sided exact binomial test, with the null hypothesis that these variants were drawn from a binomial distribution with a success
probability of 0.5 (0.95 for sex chromosomes in males). The alternative hypothesis was that these variants were drawn from distributions with lower success probabilities. Variants with P-value > 10^{-50} were considered as germline variants.

Removing errors (beta-binomial filter)
We fitted a beta-binomial distribution to the variant counts and depths of all single-base substitutions across samples from the same patient for the remaining somatic variants. The beta-binomial was used as it captures the difference between artefactual variant sites and true somatic variants. Many artefacts appear to be randomly distributed across samples and can be modelled as drawn from a binomial distribution. True somatic variants will be present at a high VAF in some samples, but absent in others, and are hence best captured by a highly dispersed beta-binomial. For each variant site, the maximum likelihood of the overdispersion factor (\(\rho\)) was calculated using a grid-based method (ranging from a value of 10^{-4} to 10^{-0.00}). Variants with \(\rho > 0.1\) were filtered out and considered to be artefactual. The code for this filter is based on the Shearwater variant caller\(^4\).

Removing mutations that were induced in vitro
We observed peaks of lower VAFs in a subset of samples (Extended Data Fig. 2c), which suggested that mutations were present that had arisen during the in vitro expansion of the single cell. These peaks were more prominent in samples from children, suggesting that the number of this kind of mutation is relatively small. They would, however, be more prominent in samples with a low true mutational burden, such as in children. We discarded mutations with a median VAF ≤ 0.3 for autosomal regions and median VAF ≤ 0.6 for sex chromosomes across all samples from the same patient; these cut-offs were determined on the basis of the observed distribution of VAFs here and in a previous report\(^3\).

We quantified sensitivity by measuring how well our algorithms called heterozygous germline polymorphisms in the colonies depending on sequencing depth; as our colonies are single-cell-derived, we would expect heterozygous germline single-nucleotide polymorphisms to have the same VAF distribution as true somatic mutations in that original single cell. We find that a sequencing depth of 8× leads to an estimated sensitivity of 70–75%, increasing to more than 95% at a sequencing depth of 15×. The majority of the colonies that we sequenced had depths of greater than 15×, and we set a minimum cut-off depth of 8× for inclusion of a colony within the study (Extended Data Fig. 2e). Finally, we visually inspected allelic counts of removed germline variants with two or more samples without any mutant reads, and rescued embryonic mutations. Somatic variants were annotated using ANNOVAR\(^4\).

Indel calling
Indels were called using cgpPindel\(^4\), and an unmatched normal sample was used as the germline control. Indels that were detected in mouse fibroblast feeder cells were removed as mouse-derived artefacts. For all indels, indel-positive or -negative sequencing reads were counted using cgpVAF across all samples from each patient.

To remove germline variants and recurrent sequencing errors, the same binomial and beta-binomial filters were used as described above for single-base substitutions. We discarded mutations with a median VAF ≤ 0.25 for autosomal regions and median VAF ≤ 0.3 for sex chromosomes across all samples from the same patient to remove mutations that were induced in vitro.

Double-base-substitution calling
We first identified candidate double-base substitutions based on side-by-side single-base substitutions that were called using CaVEMan for each patient, and ran cgpVAF across all samples from each patient to remove those called in independent reads. Double-base substitutions with three or more mutant reads in at least one sample were considered as true positives. Germline variants, errors and mutations induced in vitro were filtered as for single-base substitutions and indels.

Structural-variant calling
Structural variants were called using the BRASS algorithm\(^4\), and matched normal samples (including blood samples and normal bronchial samples that were assigned on distantly located branches in phylogenetic trees) were used as controls. To remove germline structural variants, we filtered structural variants that were detected in the descendant colonies of both of the earliest two branches at the top of the phylogenetic tree for each patient. If the earliest branch had three or more branches (polytomy), those detected in both descendant and non-descendant samples of the earliest branch with the highest number were removed. We further filtered structural variants that were not identified using an unmatched normal control, to remove structural variants that were not filtered owing to the lower sequencing coverage of the matched normal control sample. Structural variants that were detected in other patients were also removed as germline variants or errors. Finally, all remaining structural-variant calls were manually inspected using the Integrated Genomics Viewer (IGV) to confirm somatic variants.

Copy-number calling
Copy-number changes were called using the ASCAT algorithm\(^4\), and the same matched normal control samples as those used in the structural-variant analysis were used as germline controls. Copy-number gains, losses and copy-neutral LOHs were visually confirmed using LogR and BAF plots in ascatNgs. For amplification, those copy-number changes that were greater than 100 kb were visually confirmed using ascatNgs and JBrowse\(^4\).

Mutational burden and estimation of the effects of age and smoking
For single-base substitutions, indels and double-base substitutions, samples with three or more mutant reads and a VAF of 0.2 or higher were considered to be mutated, and the number of each class of genetic lesions was counted for all bronchial cells. For structural variants, chromoplexy\(^4\) (Extended Data Fig. 4c), chromothripsis\(^4\) (Extended Data Fig. 4d) and translocation pairs with similar breakpoints were considered single structural variants. Genetic lesions that were identified both as structural variants and copy-number changes were also considered as single events.

An LME model was then fitted to estimate the effects of age and smoking status on the number of single-base substitutions or indels using the nlm package in R (Supplementary Code). In addition to the fixed effects of age and smoking, patient was used as a grouping variable in the random effect, in which smoking status was used as a modifier of between-patient difference. Difference of within-group heterogeneity (heteroscedasticity) according to smoking status was also fitted in this model. The intercept of this model was probably derived from embryonic mutations and mutations that were introduced in vitro. Models were fitted using maximum likelihood estimation, and nested models were compared using likelihood ratio tests.

Identification of near-normal lung cells
We define cells as having a near-normal mutational burden if they have a mutational burden that is less than two non-smoker within-patient standard deviations plus two non-smoker between-patient standard deviations above the estimated number of mutations accumulated at the age of that patient using an LME model (Supplementary Code). The fraction of cells with a near-normal mutational burden was compared between current smokers and ex-smokers with log-linear regression using the logarithm of the total number of cells sequenced per patient as an offset.
Construction of phylogenetic trees

Phylogenetic trees were built with maximum parsimony using substitutions for each patient. First, the input matrix of mutations was used, in which samples with a VAF of 0.2 or higher and three or more mutations were considered as mutated samples and labelled as ‘1’, and remaining samples were labelled as ‘0’. Among samples labelled as 0, samples with (i) a sequencing depth of 6× or less for each mutated base and (ii) one or more mutant reads were considered as undetermined and labelled as ‘?’. For every individual, phylogenetic trees were constructed using the Camin–Sokal method of the Mix program of the RPhylip package, and consensus trees of all the trees were then constructed using the Consensus program in RPhylip.

Subsequently, all mutations were reassigned to branches in the phylogenetic trees. If mutations were called in all the descendants of a given branch and in no samples that were not descendants of the branch, mutations were perfectly assigned to those branches. Given the existence of samples with relatively lower sequencing depths for each mutated position, we also assigned mutations to branches if mutations were called in all but one undetermined descendant labelled as ‘?’ of a given branch, and all samples that were not descendants of the branch were wild-type (0). Given the smaller number of indels and double-base substitutions, these were assigned to each branch using the tree defined from single-base substitutions, rather than generating new trees for the other mutation types.

Extraction of SBS signatures

To analyse mutational signatures for single-base substitutions, those assigned to each branch of the phylogenetic trees were categorized into 288 subtypes, consisting of 6 mutation classes by 16 5′- and 3′-base contexts on the transcribed strand, non-transcribed strand or intergenic region. Mutational signatures were extracted using the HDP package50, relying on the hierarchical Bayesian Dirichlet process (https://github.com/nicolaroberts/hdp). Owing to the lack of reference signatures categorized into 288 subtypes, we conducted a de novo signature extraction. We included somatic mutations from squamous cell lung carcinomas sequenced by The Cancer Genome Atlas (TCGA) and from in vivo single-cell culture controls as separate samples to maintain comparability with signatures that have already been established in previous studies. For identified SBS signatures, signatures with ≥0.90 cosine similarity with reported signatures, in terms of distribution to 96 or 192 subtypes24, were considered as the same signatures, including SBS-1, SBS-4, SBS-5, SBS-16 and SBS-18. For the remaining new signatures, the expectation-maximization algorithm was used to deconvolute these signatures into the five signatures above and other known signatures in lung cancers (SBS-2, SBS-8 and SBS-13), because it is difficult to separate signatures that are strongly correlated across samples. If a signature reconstituted from the components that expectation maximization extracted (only including signatures that accounted for at least 10% of mutations in each sample to avoid overfitting) had a ≥0.90 cosine similarity to the original HDP signature, the signature was presented as its expectation-maximization deconvolution. Two HDP signatures met these criteria: one new signature was deconvoluted into a mixture of SBS-4 and SBS-5, and another new signature was deconvoluted in SBS-2 and SBS-13. After these analyses, seven known and two new SBS signatures were identified.

To validate these signatures that were identified using the HDP, we also analysed SBS signatures using the MutationalPatterns package20, which relies on non-negative matrix factorization. The optimal factorization rank (7) was determined on the basis of the slope of the cophenetic correlation coefficient. MutationalPatterns identified similar signatures to SBS-5 (Signature A), SBS-4 (Signature B), Sig-B (Signature D), SBS-18 (Signature E), SBS-1 (Signature F), SBS-2 and SBS-13 (Signature G) (Extended Data Fig. 5a, b).

Extraction of indel and DBS signatures

For indels and double-base substitutions, each type of genetic alteration that was assigned to each branch of the phylogenetic trees was categorized into 83 and 78 subtypes, as previously reported24. First, the algorithm was conditioned on the set of mutational signatures that have been detected in lung cancers (ID-1, ID-2, ID-3, ID-5, ID-6, ID-8, ID-9, DBS-2, DBS-4, DBS-5, DBS-6, DBS-11). This allows simultaneous discovery of known and new signatures. For known signatures, signatures identified by HDP with a cosine similarity ≥0.90 with corresponding reported signatures were accepted as known signatures. Deconvolution of new signatures to the above known signatures was also performed, and one new indel signature was deconvoluted in ID-5 and ID-8. Finally, ten known signatures and one new signature were identified.

Analysis of A→G transcriptional strand bias

First, we measured the distance from mutations to the nearest transcription start sites (TSSs) of all the expressed genes in the lung; expressed genes were defined as those with a median of one or more transcripts per million in lung samples in the GTEx database (https://gtexportal.org/home/). Mutations in regions of bidirectional transcription were excluded from further analysis. We tiled 10 kb up and downstream of the TSSs into 1-kb bins, and counted the number of A→G mutations on transcribed and untranscribed regions in each tile. This number was further divided by the average number of bins in intergenic regions.

Analysis of driver variants

To systematically identify genes under positive selection in normal bronchial epithelium, we used the dN/dS method12. We performed exome-wide dN/dS analysis and also analysed global dN/dS ratios for driver genes (n = 86) reported in lung cancer12,13,18,26 or normal skin or oesophagus tissues17–20,22 using dNdScv (Supplementary Table 3). Genes with q value ≤ 0.05 were reported as driver genes (Supplementary Tables 4, 5). Finally, hot-spot mutations reported in COSMIC for four or more patients were also considered as driver mutations, in addition to those in the seven driver genes identified by dNdScv (Fig. 3b). The proportion of shared mutations (found in more than one colony) and private mutations (found in a single colony) was calculated for patients other than PD30160 (who had a low number of sequenced samples n = 13). For known lung cancer driver genes, the distributions of mutations were compared between bronchial cells and lung squamous cell carcinoma14 (Extended Data Fig. 9b).

To estimate the effect of smoking status on the number of driver mutations, a generalized LME model was fitted using the lme4 package in R (Supplementary Code). Patient was modelled as a random effect, and the fixed effects of age, smoking status and total mutational burden were fitted into the model.

Estimation of telomere length

The average telomere lengths of bronchial epithelium cells were estimated from the whole-genome sequencing data using Telomerecat23. Considering the similarity of telomere sequences between human and mouse, we aligned all sequencing reads to the human reference genome using BWA-MEM without using Xenome, and then ran Telomerecat on the bam files. Samples with reported mouse contamination of more than 10% were excluded from further analysis to prevent a possible effect of mouse cells on telomere length. The average telomere length for the mouse fibroblast feeder samples was estimated at 1,745 bp, which is within the range of estimates of human telomere length, so a low level of mouse contamination will not substantially affect the estimates.

An LME model was then fitted to estimate the effect of telomere length on the number of single-base substitutions using the lme4 package in R (Supplementary Code). Patient was modelled as a random effect, and the fixed effects of telomere length and its interaction with...
smoking status, as well as the fixed effects of age and smoking status, were fitted into the model.

### Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

### Data availability
Sequencing data have been deposited at the European Genome-phenome Archive (http://www.ebi.ac.uk/ega/) under the accession number EGAD00001005193. Somatic-mutation calling, including single-base substitutions, indels and structural variants, from all 632 samples have been deposited on Mendeley Data with the identifier: https://doi.org/10.17632/b53h2kwppy.2.

### Code availability
Detailed method and custom R scripts for the analysis of mutational burden in bronchial epithelium are available in Supplementary Code. Other packages used in the analysis are as follows: R v.3.5.1; BWA-MEM v.0.7.17-r1188 (https://sourceforge.net/projects/bio-bwa/); CaVEMan v.1.11.2 (https://github.com/cancerit/CaVEMan); Pindel v.2.2.5 (https://github.com/cancerit/cgpPindel); Brass v.6.1.2 (https://github.com/cancerit/BRASS); ASCAT NGS v. 4.1.2 (https://github.com/cancerit/ascatNgs); Xenome (https://github.com/data61/gossamer/blob/master/docs/xenome.md); deepSNV v.1.28.0 (https://biocductor.org/packages/release/bioc/html/deepSNV.html); ANNOVAR (http://wannovar.wglab.org/); IGV (https://software.broadinstitute.org/software/igv/); JBrowse (https://jbrowse.org/); cgVAF (https://github.com/cancerit/vafCorrect); RPhylip v.0.1.23 (http://www.phytools.org/Rphylip/); hdph v.0.1.5 (https://github.com/nicolaroberts/hdph); MutationalPatterns v.1.8.0 (https://biocductor.org/packages/release/bioc/html/MutationalPatterns.html); dNdScv v.0.0.1 (https://github.com/im3sanger/dndscv); and Telomerecat v.3.1.2 (https://github.com/jhrf/telomerecat).

### Acknowledgements
This work was supported by a Cancer Research UK Grand Challenge Award (C98/A24032) and the Wellcome Trust. P.J.C. and S.M.J. are Wellcome Trust Senior Clinical Fellows (WTD88340MA). S.M.J. receives funding as a member of the UK Regenerative Medicine Platform (UKRMP2) Engineered Cell Environment Hub (MRC, MR/R015635/1) and the Longfonds BREATH Lung regeneration consortium, and is further supported by The Rosetrees Trust, the Stoneygate Trust, the British Lung Foundation and the UCLH Charitable Foundation; K.Y. is supported by a Japan Society for the Promotion of Science (JSPS) Overseas Research Fellowship and The Mochida Memorial Foundation for Medical and Pharmaceutical Research; S.M.J. and R.E.H. are supported by the Roy Castle Lung Cancer Foundation; R.E.H. is a Wellcome Trust Sir Henry Wellcome Fellow (WT209199/2/17/Z), and I.M. is funded by Cancer Research UK (CS7387/A21777). The authors thank S. Broad and F. Wutt for providing 3T3-J2 fibroblasts, and B. Carroll for help with sample collection.

### Author contributions
D.P.C. conceived and designed the experiments. K.Y. and S.M.J. performed most of the data curation and statistical analysis, with help from D.P.C., E.F.M. and F.R.M. E.F.M. and C.R.B. collected the paediatric samples, and performed all of the sample collection, cell isolation, clonal expansion and DNA extraction. L.A.R. performed the qPCR characterization of the clones. M.R.S. oversaw the analysis of mutational signatures. P.J.C. and I.M. oversaw statistical analyses. R.E.H., A.P., K.H.C.G., K.Y., S.M.J. and D.P.C. conducted the experiments. P.J.C. performed data interpretation and, together with D.P.C., helped to draft and revise the manuscript.

### Competing interests
The authors declare no competing interests.

### Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1961-1.

### Correspondence and requests for materials
should be addressed to S.M.J. or P.J.C.

### Peer review information
Nature thanks Gerd P. Pfeifer, Roman Thomas and the other anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Flow-sorting strategy of single basal bronchial epithelial cells. a, Sorting of EpCAM⁺ epithelial cells from human airway biopsies. Human haematopoietic and endothelial cells were stained with antibodies against CD45 and CD31, respectively. Within the population of cells negative for those markers, EpCAM-expressing cells were gated. Single, live (DAPI-negative) cells were flow-sorted from this population into individual wells of 96-well plates. b, Quantitative PCR (qPCR) analysis of cultures of clonally derived airway epithelial cells. Airway basal cells express integrin subunit α 6 (ITGA6), keratin 5 (KRT5), cadherin 1 (CDH1) and TP63. Expression is shown in clonally derived cell cultures (n = 13 from 3 donors, coloured blue, green and orange) compared to control bulk human bronchial epithelial cell cultures (HBECs) that were expanded in the same culture conditions and lung fibroblast cell cultures (lung fibs) that served as a negative control. The centre values and error bars indicate mean and s.e.m., respectively. Conditions in which no expression was detected are shown as 0. c, Colony-forming efficiency of CD45⁻CD31⁻EPCAM⁺ cells after single-cell sorting from endobronchial biopsy samples (n = 16). For one ex-smoker, EpCAM was not used to select cells and only CD45⁻CD31⁻ cells were sorted; as expected, this was the patient with the lowest colony-forming efficiency.
Extended Data Fig. 2 | Quality assurance of mutation calls. a, Stacked bar chart showing the proportion of reads attributed to the human genome, mouse genome, both, neither, or with ambiguous mapping for the pure mouse fibroblast feeder line (left) or a pure human sample (right), assessed with the Xenome pipeline. b, Clean-up of mutation calls using the Xenome pipeline for one of the samples that was more heavily contaminated by the mouse feeder layer. The Venn diagram on the left shows the overlap in mutation calls before and after removing non-human reads by Xenome. c, Histograms of VAFs for two representative colonies in the sample set. The plot on the left shows a tight distribution around 50%, as expected for a colony derived from a single cell without contamination. The plot on the right shows a bimodal distribution with one peak at 50% (mutations present in the original basal cell) and a second peak at around 25% (probably representing mutations that were acquired in vitro during colony expansion). These second peaks at less than 50% are more evident in colonies from children, owing to the low number of mutations in the original basal cell. d, Histogram of VAFs for a colony seeded by more than one basal cell, leading to a peak at much less than 50%. e, Estimated sensitivity of mutation calling according to sequencing depth. Heterozygous germline polymorphisms were identified in each subject; for each colony sequenced, we calculated the fraction of these polymorphisms that was recalled by our algorithms. f, Comparison of mutational burden in normal bronchial epithelial cells that neighbour a carcinoma in situ (CIS) versus cells distant from the CIS in five patients. The box-and-whisker plots show the distribution of mutational burden per colony within each subject, with the boxes indicating median and interquartile range and the whiskers denoting the range. The overlaid points are the observed mutational burden of individual colonies.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Colonies with a near-normal mutational burden.

a, Density distribution of mutational burden in cells from ex-smokers (green) and current smokers (purple). The black vertical line shows the threshold for near-normal mutational burden derived for each patient. The x-axis is on a logarithmic scale. Note the frequently bimodal distribution of mutational burden, especially in the ex-smokers, with the modes separated at the threshold for near-normal mutational burden.

b, Flow cytometric analysis of clones for expression of KRT5, EpCAM, ITGA6, podoplanin (PDPN), NGFR and CD45 or CD31. Lung fibroblasts are included as a comparison. Fluorescence minus one (FMO) is shown. Plots for one clone with a near-normal mutational burden (low-mutant clone) and one with an increased burden (high-mutant clone) are shown, and are representative of five clones from one patient.

c, Bright-field images of expanded clones at passage 3, showing cobblestone epithelial morphology. Images are representative of five clones from one patient. A clone with an increased mutational burden is shown at the top, and a clone from an ex-smoker with a near-normal mutational burden is shown at the bottom. For the left images, the magnification is ×10 and the scale bar is 200 μm; for the right images, the magnification is ×20 and the scale bar is 100 μm.
Extended Data Fig. 4 | Indels, copy-number changes and structural variants in normal bronchial epithelial cells. **a**, Relationship of burden of indels per cell with age. The points represent individual colonies (\(n = 632\)) and are coloured by smoking status. The black line represents the fitted effect of age on indel burden, which was estimated from LME models after correction for smoking status and within-patient correlation structure. The blue shaded area represents the 95% CI for the fitted line. **b**, Stacked bar plot showing the distribution of colonies with 0–7 copy-number changes and structural variants across the 16 subjects. **c**, Three examples of chromoplexy in normal bronchial cells. Structural variants are shown as coloured arcs that join two positions in the genome around the circumference. The instances of chromoplexy all consist of three translocations (purple). **d**, An example of chromothripsis in a cell from an 11-month old child. The plot on the right shows the copy number of genomic windows in the relevant region of chromosome 1 (black points); the lines and arcs denote the positions of observed structural variants.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | Comparison of mutational signatures that were extracted using two algorithms. a, Trinucleotide contexts for the signatures extracted by the hierarchical Dirichlet process (HDP) (left) and MutationalPatterns non-negative matrix factorization (right). The six substitution types are shown across the top of each signature. Within each signature, the trinucleotide context is shown as four sets of four bars, grouped by whether an A, C, G or T respectively is 5′ to the mutated base, and within each group of four by whether A, C, G or T is 3′ to the mutated base (the order of bars is the same as that shown in Fig. 2b). Where signatures show high cosine similarity scores between algorithms, they are lined up horizontally. We note that Signature C in MutationalPatterns does not have a match in the signatures extracted by the HDP algorithm, but appears very similar to Signature A in MutationalPatterns (or SBS-5 from the HDP). This means that it probably represents over-splitting of the signatures. b, Heat map showing the cosine similarities of signatures extracted by MutationalPatterns with those extracted by the HDP. Only cosine-similarity scores that are greater than 0.75 are coloured. c, Scatter plots showing the fraction of mutations in each colony (n = 632) assigned to each signature by the HDP algorithm (x axis) versus the MutationalPatterns algorithm (y axis). The correlation values quoted are Pearson's correlation coefficients (R²). d, Transcriptional strand bias of A>G mutations in an N[A]T context before and after TSSs. Note the absence of transcriptional strand bias in intergenic regions but evidence for both transcription-coupled damage and repair after the TSS, applying similarly in both never-smokers and ex- or current smokers.
Extended Data Fig. 6 | Phylogenetic trees of 13 subjects. Phylogenetic trees showing clonal relationships among normal bronchial cells in the 13 subjects not shown in Fig. 3a. Branch lengths are proportional to the number of mutations (x axis) specific to that clone or subclone. Each branch is coloured by the proportion of mutations on that branch that are attributed to the various SBS signatures.
Extended Data Fig. 7 | Indel signatures in the sample set. a, Five indel signatures (ID-1, ID-2, ID-3, ID-5 and ID-8) were extracted by the HDP. The contributions of different types of indels to each signature are shown, grouped by whether variants are deletions or insertions; the size of the event; whether they occur at repeat units; and the sequence content of the indel. b, Stacked bar plot showing the proportional contribution of mutational signatures to indels across the 632 colonies derived from normal bronchial cells, extracted using the HDP. Within each patient, colonies are sorted from left to right by increasing indel burden (bar chart in dark grey above coloured signature-attribution stacks).
**Extended Data Fig. 8 | DBS signatures in the sample set.**

**a.** Six DBS signatures were extracted by the HDP. The contributions of different types of double-base substitution to each signature are shown, grouped by the sequence that is mutated and by what it is mutated to. Five of the signatures have been observed in cancer genomes, and one (DBS Sig-C) is a novel signature that was extracted here.

**b.** Stacked bar plot showing the proportional contribution of mutational signatures to double-base substitutions across the 632 normal bronchial cells, extracted using the HDP. Note that some of the colonies in children have no double-base substitutions. Within each patient, colonies are sorted from left to right by increasing burden of double-base substitutions (bar chart in dark grey above coloured signature-attribution stacks).
Extended Data Fig. 9 | Driver mutations in normal bronchial epithelium.

a, Stick plots showing distribution of mutations in TP53, NOTCH1 and other genes that were significantly mutated in our sample set. Mutations are coloured by type. The gene structure is shown horizontally in the centre of each plot, with domains as coloured bars. Above the gene are mutations in this sample set, and below the gene are mutations found in squamous cell carcinomas from the TCGA sample set. b, Fraction of cells with driver mutations in TP53 (left), NOTCH1 (middle) or all other significant cancer genes (right), split by smoking status.
Extended Data Fig. 10 | Relationship of telomere length with age. Scatter plot of estimated telomere lengths (y-axis) against the age of the subject (x-axis). Individual points represent colonies (n = 398 colonies in which less than 10% of the DNA was derived from the mouse feeder layer). Cells with a near-normal mutational burden are coloured gold.
**Statistical parameters**

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| Item                                                                 | Confirmed |
|---------------------------------------------------------------------|-----------|
| The exact sample size \( n \) for each experimental group/condition, given as a discrete number and unit of measurement | Yes       |
| An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly | Yes       |
| The statistical test(s) used AND whether they are one- or two-sided  | Yes       |
| *Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |           |
| A description of all covariates tested                              | Yes       |
| A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons | Yes       |
| A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) | Yes       |
| For null hypothesis testing, the test statistic (e.g. \( F \), \( t \), \( r \)) with confidence intervals, effect sizes, degrees of freedom and \( P \) value noted | Yes       |
| *Give \( P \) values as exact values whenever suitable.* |           |
| For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings | Yes       |
| For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes | Yes       |
| Estimates of effect sizes (e.g. Cohen's \( d \), Pearson's \( r \)), indicating how they were calculated | Yes       |
| Clearly defined error bars                                          | Yes       |
| *State explicitly what error bars represent (e.g. SD, SE, CI)*       |           |

Our web collection on [statistics for biologists](https://www.nature.com/articles/nbt3483) may be useful.

**Software and code**

Policy information about [availability of computer code](https://www.nature.com/nature/researchasctime-condition).

| Data collection | Image processing from sequencing data using standard Illumina X10 pipeline |
|-----------------|--------------------------------------------------------------------------|
| Data analysis   | Alignment and variant calling performed using Sanger institute's custom pipeline. Single-nucleotide substitutions were called using the CaVEMan (cancer variants through expectation maximization) algorithm (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm (https://github.com/cancerit/BRASS). |
| List of programs and softwares: | |
| R – version 3.5.1 | |
| BWA-MEM – version 0.7.17-r1188 (https://sourceforge.net/projects/bio-bwa/) | |
| CaVEMan – version 1.11.2 | |
| Pindel – version 2.2.5 | |
| Brass – version 6.1.2 | |
| ASCAT NGS – version 4.1.2 | |
| Xenome (https://github.com/data61/gossamer/blob/master/docs/xenome.md) | |
| deepSNV – version 1.28.0 (https://bioconductor.org/packages/release/bioc/html/deepSNV.html) | |
| ANNOVAR (http://wannovar.wglab.org/) | |
| IGV (http://software.broadinstitute.org/software/igv/) | |
| JBrowse (https://jbrowse.org/) | |
cgpVAF (https://github.com/cancerit/vafCorrect)
RPhylip - version 0.1.23 (http://www.phytools.org/Rphylip/)
hdp - version 0.1.5 (https://github.com/nicolaroberts/hdp)
MutationalPatterns - version 1.8.0 (https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html)
dNdScv - version 0.0.1 (https://github.com/im3sanger/dndscv)
Telomerecat - version 3.1.2 (https://github.com/jhrf/telomerecat)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data that support the findings of this study have been deposited in the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) under accession number EGAD00001005193. Somatic mutation calls, including single base substitutions, indels and structural variants, from all 632 samples have been deposited on Mendeley Data with the identifier: http://dx.doi.org/10.17632/b53h2kwppy.2.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Sample size was chosen to give good representation of inter-patient and intra-patient variability in mutation burden.

Data exclusions
Oligoclonal colonies or with low mean coverage (<8x) (Extended Figure 2e) were excluded due to the inaccuracy of mutation catalogues. One outlying cell in an ex-smoker with >10,000 mutations is excluded from the plot of Figure 4 to improve visualisation.

Replication
No experimental replication has yet been attempted.

Randomization
Not applicable - this is a descriptive study, not an intervention study.

Blinding
Not applicable - all dependent variables were computationally generated (mutation counts, signatures etc) and statistical analyses were prespecified.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a
- Involved in the study
- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

Methods

n/a
- Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Antibodies

Antibodies used
Antibodies used in this study were described in OnlineMethods.
Validation

Antibodies were validated by the manufacturer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
3T3-J2 feeder cells were kindly provided by Prof. Fiona Watt (King’s College London)

Authentication
The feeder cells had whole genome sequencing performed, confirming their murine origin and clonal derivation.

Mycoplasma contamination
They were tested negative for Mycoplasma by PCR test (PMCID: PMC202165)

Commonly misidentified lines
No commonly misidentified cell lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics
We analysed single cell-derived colonies from bronchial epithelium of 16 subjects, including 3 children, 4 never-smokers, 6 ex-smokers and 3 current smokers. Clinical characteristics of the cohort are described in Supplementary Table 1. Of the ex-smokers, 2 had had a previous cancer treated with curative intent, and 5 had a carcinoma in situ or invasive squamous cell carcinoma. The children in the cohort had bronchoscopy for investigation or follow-up of congenital anomalies.

Recruitment
Recruited through University College Hospitals, London, UK. Our cohort does potentially suffer from recruitment bias, since samples could only ethically be obtained from individuals undergoing a clinically indicated bronchoscopy.

Flow Cytometry

Plots

Confirm that:
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation
The epithelium was dissected away from the underlying stroma and fetal bovine serum (FBS) was added to a final concentration of 10%. Both the epithelium and stroma were combined and digested in 0.1% trypsin/EDTA at 37 °C for 30 minutes. The solution was neutralised with FBS to a final concentration of 10% and added to the neutralised dispase solution. Cells were passed through a 100 μm cell strainer and stained in sorting buffer (1x PBS, 1% FBS, 25 mM HEPES and 1 mM EDTA) with anti-CD45-PE (BD Pharmingen 555483, 1:200), anti-CD31-PE (BD Pharmingen 555446, 1:200), anti-EPCAM-APC (Biolegend 324208, 1:50) antibodies and DAPI (1 μg/ml).

Instrument
BD FACSAria Fusion

Software
Supplementary Methods

Cell population abundance
Supplementary Methods

Gating strategy
Supplementary Methods

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.