www.sciencemag.org/content/354/6312/618/suppl/DC1

Supplementary Materials for

Mutational signatures associated with tobacco smoking in human cancer

Ludmil B. Alexandrov,* Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J. Campbell, Paolo Vineis, David H. Phillips, Michael R. Stratton*

*Corresponding author. Email: lba@lanl.gov (L.B.A.); mrs@sanger.ac.uk (M.R.S.)

Published 4 November 2016, *Science* **354**, 618 (2016)
DOI: 10.1126/science.aag0299

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S10
Captions for Tables S1 to S6
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/354/6312/618/suppl/DC1)

Tables S1 to S6

METHODS

Data Curation

Freely available data for cancers from cancer types associated with tobacco smoking were curated from data portals and previously published articles. In total, we were able to curate 5,243 samples, consisting of at least sequencing data for the cancer genome and the genome of a matched-normal tissue. For all 5,243 samples, we were able to examine somatic substitutions, small insertions and deletions (indels), dinucleotide substitutions and mutational signatures. Smoking annotation was available for 3,552 of the 5,243 samples, allowing direct comparison between smokers and non-smokers. The comparison between smokers and non-smokers was extended to include copy number data, genomic rearrangements, and methylation data. Copy number data were available for 2,270 of the 3,552 samples, genomic rearrangements were available for 349 samples, and 1,770 samples had data about their genomic methylation. Overall, data were retrieved from three sources: (i) The Cancer Genome Atlas (TCGA) data portal, (ii) the International Cancer Genome Consortium (ICGC) data portal, and (iii) data previously generated for 17 articles published in peer-reviewed journals. Table S1 provides detailed information about each sample as well as the available data for individual samples used in the comparisons between smokers and non-smokers. Additionally, information about data sources and accession numbers for sequencing, copy number and methylation data are also provided in Table S1 allowing reproducibility of the analyses performed in this study.

Filtering somatic mutations and generating mutational catalogues for cancer samples

This study relies on previously generated DNA sequencing data and somatic mutations identified in these data. Additionally, the study uses previously generated copy number data and previously generated and processed methylation data. The examined sequencing data were originally generated by a variety of different laboratories, leveraging different experimental platforms and using a diverse set of computational algorithms. To remove any residual germline mutations as well as technology, institute, and/or laboratory specific sequencing artifacts, extensive filtering was performed prior to analyzing the somatic mutation data. Germline mutations were filtered out from the lists of reported somatic mutations using the complete list of germline mutations from dbSNP (*32*), 1000 genomes project (*33*), NHLBI GO Exome Sequencing Project (*34*) and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/). Technology specific sequencing artifacts were filtered out by using panels of BAM files of unmatched normal tissues containing more than 500 normal whole-genomes and 1,000 normal whole-exomes. Any somatic mutation present in at least two well-mapping reads in at least two normal BAM files was discarded. The remaining somatic mutations constituted the mutational catalogue for every matched-normal pair. The immediate 5′ and 3′ sequence context for each somatic mutation was extracted using the ENSEMBL Core APIs for the human genome build that was originally used to identify these somatic mutations.

Identifying mutational signatures and their exposures in cancer genomes

Mutational catalogues were generated for all 5,243 samples. These catalogues were examined following two independent and distinct steps as previously done in refs. (*17*, *35*). The first step encompasses *de novo* extraction of mutational signatures based on somatic substitutions and their immediate sequence context, while the second step focuses on accurately estimating the number of somatic mutations associated with each extracted mutational signature in each sample. Briefly, mutational signatures were deciphered independently for each of the 17 cancer types associated with tobacco smoking as well as for all cancer types together using our previously developed computational MATLAB framework (*18*). The computational framework for deciphering mutational signatures is freely available and it can be downloaded from: http://www.mathworks.com/matlabcentral/fileexchange/38724. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type found in each catalogue and then estimates the contribution of each signature to each mutational catalogue. Mutational signatures were also extracted separately for genomes and exomes. Mutational signatures extracted from exomes were normalized (*i.e.*, the trinucleotide values of the exome extracted mutational signatures were divided by the observed trinucleotide frequency in an exome and multiplied by the observed trinucleotide frequency in a genome) from the observed trinucleotide frequency in the human exome to the trinucleotide frequency of the human genome. Overall, we identified multiple distinct mutational signatures, all of which were previously reported (and named) by others or us (*17, 18, 36*). The names of mutational signatures used in this manuscript are consistent with the ones on the COSMIC website, http://cancer.sanger.ac.uk/cosmic/signatures, which, in turn, are consistent with all previous publications. Signatures 1 through 28 (with the exception of signatures 9, 11, 14, 19, and 25, which were not found in these samples) were previously validated (*17, 18, 36*) and, thus, these processes most likely reflect genuine and biologically relevant mutational processes. Signatures R1 through R2 were previously found in ref. (*18*), where these three signatures failed validation by orthogonal sequencing and were attributed to sequencing artifacts. We were not able to perform validation for signatures U2, U5, U6 and U7 as we did not have access to the samples in order to perform validation by orthogonal sequencing or visual validation of BAM files by experienced curators. The patterns of almost all extracted mutational signatures were the same as the ones already reported on the COSMIC website: http://cancer.sanger.ac.uk/cosmic/signatures. The only exceptions were signatures 4 and 16, which patterns of mutations were slightly updated (Fig. 2 and Table S6).

The *de novo* extraction was used to identify the complete set of mutational signatures across the 5,243 curated tobacco smoking associated cancers. Next, we quantified the rates of somatic mutations attributed to each signature in each sample following our previously developed methodology (*17*, *35*). Briefly, the contributions of the mutational signatures were estimated independently for each of the 5,243 samples with a subset of mutational signatures. For each sample, the estimation algorithm consists of finding the minimum of the Frobenius norm of a constrained linear function for a set of previously extracted mutational signatures. This set of signatures is determined based on the known operative mutational processes in the cancer type of the examined sample from the hitherto described mutational signature extraction process. For example, for any acute myeloid leukemia sample, only signatures 1 and 5 will be used since these are the only known signatures of mutational processes identified in acute myeloid leukemia. The prevalence of somatic mutations and mutational signatures in each sample was estimated based on a haploid human genome after all filtering was performed. Briefly, the prevalence of somatic mutations in a whole-exome sample was calculated based on the identified mutations in protein coding genes and assuming that an average whole-exome has sufficient coverage of 30.0 megabase-pairs in protein coding genes. The prevalence of somatic mutations in a whole-genome sample was calculated based on all identified mutations and assuming that an average whole-genome has sufficient coverage of 2.78 gigabase-pairs. No mutations (*i.e.*, value of zero mutations per megabase) were attributed to all signatures that were not found in a given sample.

Factors that influence extraction of mutational signatures and mutational signatures robustness and reproducibility

In a previous analysis, we used simulated data to describe multiple factors that can influence the extraction mutational signatures (*15*). Examples of such factors are the number of somatic mutations found in an individual sample, the *bona fide* numbers of mutations contributed by different mutational signatures, the resemblance of patterns amongst mutational signatures, the total number of available samples, and computational limitations of our framework (*15*). Regardless of these limitations, throughout the past four years, our framework has demonstrated its ability to identify robust and novel mutational signatures. Multitude of studies performed by us and others using our computational framework (or other computational frameworks) on different datasets have revealed the stability and reproducibility of extracting mutational signatures from human cancers (*18, 20, 24, 37-45*). Additionally, this study demonstrates once more the reproducibility of our computational framework. Here, we analyzed 10 distinct datasets containing different numbers of samples from different tobacco smoking associated cancers and consistently deciphered signature 4 in these cancer types (Fig. S2). Furthermore, the newly identified pattern of signature 4 (Table S6) has a cosine similarity of 0.97, where 1.00 is a perfect match, with the previously reported pattern of signature 4 on the COSMIC website, http://cancer.sanger.ac.uk/cosmic/signatures. Additionally, this *in silico* derived mutational signature also matches a mutational signature induced *in vitro* by exposing cells to the carcinogen benzo[*a*]pyrene (cosine similarity=0.94), a constituent of tobacco smoke (*19*). It should be noted that the benzo[*a*]pyrene mutational signature is based on our previous experimental work (*19*).

Analysis of copy number changes and structural rearrangements

Analysis for copy number changes was performed for all samples with annotated tobacco smoking history for which copy number data were available. Where raw Affymetrix SNP 6.0 data could be obtained, we inferred sample purity and ploidy, as well as copy number changes, using the ASCAT computational framework (*46*). Where raw SNP array data wasn't available (a subset of pancreatic cancers), processed data was obtained from the ICGC data portal and harmonized before analysis. More specifically, since only aberrant regions were reported, we filled all unreported positions of chromosomes 1 through 22 with a major + minor copy number state of 1 + 1. Additionally, we had to transform the three reported mutation type classes (copy neutral loss of heterozygosity (LOH), gain, loss) and overall copy number into copy number states for major and minor allele using the following classification: (i) copy neutral LOH: 2+0, (ii) loss: overall copy number+0, and (iii) gain: when the overall copy number is 4, 2+2, otherwise (overall copy number-1)+1. Gains are not easy to derive from overall copy number only, but as our further analysis only relies on the length of aberrant regions, this approach is sufficient.

From the adjusted copy number information, we could directly extract the number of breakpoints in each sample and compare the distributions across samples between smokers and non-smokers. To determine the genomic instability per sample, we calculated the fraction of the genome that is copy number aberrant. As we had observed multiple polyploid samples when examining the ploidy, we decided to define the aberrant regions based on a majority-rule of copy number states per sample. If the segmented data shows more 2+2 than 1+1 regions, it is classified as whole genome duplicated and all segments with a copy number not at 2+2 are considered aberrant. For all other samples we assumed a normal diploid state and only regions not at 1+1 are counted towards the aberrant genomic fraction. As such, the overall genomic instability of a sample is then calculated as the length of all aberrant segments divided by the length of all regions for which copy number information was available.

Analysis for genomic structural rearrangements was performed for all samples with annotated tobacco smoking history for which genomic rearrangements were originally reported. The number of genomic rearrangements for each sample was derived based on the provided data in the curated data source from which the sample was curated.

Analysis of methylation data

Comparison of overall methylation between smokers and non-smokers was performed for all tobacco-associated cancer types for which there were available data from Illumina Infinium HumanMethylation450 BeadChip array, where each array contains 473,864 autosomal CpG probes. The examined data were downloaded from the original data source (Table S1) after standard pipelines have already pre-processed the data (TCGA level 3 data). This processing includes: quality control, normalization, and beta value calculations. All probes targeting X- and Y-chromosomes were removed from our downstream analyses.

For each cancer type, we categorized samples into two distinct non-overlapping groups according to their smoking history (*i.e.*, smokers and non-smokers). Principal component analysis was applied using prcomp function in “stats” package of the language R. We compared interquartile ranges (IQR) of beta values of each probe between smoker and non-smoker subgroups. Smokers showed statistically significant increase of IQR values in all cancer types, albeit at low levels in most cancer types. The exceptions were larynx and lung adenocarcinoma where the inter-sample variability of methylation levels of CpGs, measured using the interquartile range, had more than 20% average increase in smokers compared to non-smokers (Fig. S9). Additionally, for each cancer type, we derived two average methylation profiles: one for smokers and another one for non-smokers. In each cancer type, we compared each individual smoker to the average methylation profile of smokers in that cancer type as well as each individual non-smoker to the average methylation profile of non-smokers in that cancer type. The mean methylation deviation for a given sample was determined by:

$$MeanDeviation = \sum_{i=1}^{N} \frac{|\beta_i - \bar{\beta}|}{N}$$

where β_i is the methylation level (*i.e.*, beta value) of the i-th CpG probe, $\bar{\beta}$ is the average methylation level of the probe, and N is the total number of the methylation probes (N is 473,864 in a Illumina Infinium HumanMethylation450 BeadChip array). The distributions of the derived mean deviations were used as methylation features, which distributions were subsequently compared between smokers and non-smokers using a two-sample Student’s t-test. Results were considered significant for Bonferroni threshold of 10^{-7}.

Statistical comparison between smokers and non-smokers

Comparisons between smokers and non-smokers were performed both across all examined samples as well as in samples from individual types of cancers associated with tobacco smoking. In individual cancer types, the distributions of multiple distinct features were compared between smokers and non-smokers. These features include: total somatic substitutions, total long indels (>=3bp) with overlapping microhomology, total short indels (<3bp) found at mono/polynucleotide repeats, total dinucleotide substitutions, somatic substitutions of a given type (C>A, C>G, C>T, T>A, T>C, or T>G) and numbers of somatic mutations attributed to the mutational signatures found in that cancer type. Additionally, when data were available for a cancer type, we compared numbers of breakpoints, fraction of the genome that shows copy number changes, ploidy of the genome, purity of the tumor, and overall methylation. For each feature, in a given cancer type, a p-value was derived using a two-sample Kolmogorov-Smirnov test to compare the distributions of this feature between smokers and non-smokers. The calculated p-values were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure for all features in a cancer type.

In addition to examining individual cancer types, we compared smokers to non-smokers using all samples across all cancer types. The examination was performed analogously to the one described for individual cancer types. However, to make sure that mixing samples across cancer types does not bias our results, the mixing process was iterated multiple times and overall average values and p-values were derived for each feature. In each of the iterations, 30 smokers and 30 non-smokers were randomly sampled with replacement from each cancer type and mixed together. The distributions of each feature in the mixture were compared between smokers and non-smokers using two-sample Kolmogorov-Smirnov tests. In total, we performed 10,000,000 mixing iterations. For each feature, an iteration for which the two-sample Kolmogorov-Smirnov test returned a p-value<0.05 was considered statistically significant, whereas an iteration with p-value>=0.05 was considered not statistically significant. The overall p-value per feature was calculated as follows: $\frac{NumberOfNotSignificantIterations+1}{10,000,000+1}$. It should be noted that the number of iterations limits the minimum possible p-value, in this case 9.99E-08, and p-values reported as 9.99E-08 are most likely lower. The derived p-values were further corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure for all examined features.

The p-values for the correlations between the ages of diagnosis and the mutations attributed to signature 5 were derived separately across all smokers and all non-smokers using robust linear regression for all samples in which signature 5 was found.

Analysis of clonal and subclonal mutations in smokers and non-smokers

We managed to obtain access and to download aligned sequencing data (BAM files) for 1,506 of the 3,553 samples with annotated smoking information (Table S1). In order to analyze what fraction of mutations belong to a subclonal population, we determined the variant allele fraction (VAF) of variants in the coding regions of samples for which we had access to sequencing data. The initial VAF is transformed into mutation copy number (MCN), *i.e.*, the number of genomic segments carrying a given variant, as described in equation 1.

$$\text{MCN} = \frac{VAF}{\rho} * (\rho\psi_t + (1-\rho)\psi_n) \tag{1}$$

Each sample's purity ρ and the copy number of the surrounding region in the tumor ψ_t was determined by ASCAT (*46*), the copy number in the normal ψ_n is assumed to be two, except for variants on sex chromosomes in male patients where it is set to one. Additionally, the expected VAF of a variant with a mutation copy number of one that is located in the same genomic region than the observed mutation was calculated as described in equation 2.

$$\text{Expected VAF}_{\text{MCN=1}} = \frac{\rho}{(\rho\psi_t + (1-\rho)\psi_n)} \tag{2}$$

To estimate if a given mutation is subclonal, a one-sided proportion test was applied by using the number of reads supporting the variant and the overall coverage of the locus as counts of successes and trials, respectively. The calculated expected VAF is used in the test as the probability of success. All tested mutations that have a significantly lower VAF than expected for a variant with a mutation copy number of one (p-value<0.05) are defined as subclonal.

Statistical comparisons for mutational signatures in clonal and subclonal mutations were performed across all tobacco smokers and all non-smokers in a manner analogous to the one described in the previous section.

Statistical analysis of relationships between pack years smoked and mutational signatures

We performed global analysis for linear relationship between pack years smoked and mutational signatures. In addition to mutational signatures, total somatic mutations were examined in the same way as a mutational signature. The data heteroscedasticity as well as the presence of many outliers requires leveraging an appropriate statistical method (*47*). We made use of robust linear regression to evaluate linear dependencies between the number of mutations associated with each mutational signature across all examined samples and the pack years smoked.

Examination was first performed across all samples in all cancer types by combining data for the identified mutational signatures. The calculated p-values from the applied robust regression analysis were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure (Table S3). In addition to examining all samples across all cancer types, individual cancer types were examined separately for a linear dependence between pack years smoked and the number of somatic mutations attributed to each of the signatures of the operative mutational processes found in that cancer type. Taking into consideration that the majority of traditional or generalized linear regression approaches are very sensitive to outliers (*47*) and since many of the examined cancer samples are hypermutators (*i.e.*, outliers), we again leveraged a robust regression model to evaluate for linear dependencies. Briefly, robust regression iteratively reweights least squares with a bi-square weighting function and overcomes some, if not the majority, of limitations attributed to traditional approaches (*48-50*). Similarly, we have chosen to report results using

Spearman's rank correlation coefficient since it is more robust to data outliers when compared to the traditional Pearson's product-moment correlation coefficient (*51*). It should be noted that only samples with known pack years smoked were included in this analysis. Each mutational signature was examined separately in each of the cancer types in which that signature had been identified. The examination was based on: a robust linear regression model that estimates the slope of the line and whether this slope is significantly different from a horizontal line with a slope of zero (F-test; p-value<0.05) as well as by calculating the Spearman's rank correlation coefficient. The calculated p-values from the applied robust regression analysis were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure for each cancer type (Table S3).

SUPPLEMENTARY TEXT

Theoretical considerations on the increase in mutation rate and the increase in cancer risk

Using age-incidence epidemiological data, in the early 1950s, Nordling (*52*) as well as Armitage and Doll (*53*) first modeled cancer as a result of *n* rate-limiting steps. For example, under their assumption, if there are 5 rate-limiting steps that a cell has to undergo to become a cancer, the risk of cancer can be modeled as:

$$\text{risk} = k\, p_1\, p_2\, p_3\, p_4\, p_5\, t^4 \sim k' * \mathrm{t}^4$$

Where *k* and *k'* are constant terms and *p* reflect the relative rates of different driver events. This equation offers a way to derive a simple relationship between the mutation rate and cancer risk. Let us assume that *n* of the rate limiting steps are somatic mutations whose rate is increased by a mutagen by *r*-fold. For example, if n=3, we have:

$$\text{risk}_{\text{with mutagen}} = k\, r\, p_1\, r\, p_2\, r\, p_3\, p_4\, p_5\, t^4 \sim k' * \mathrm{r}^3 * \mathrm{t}^4$$

Dividing the expressions for the risk with and without a mutagen one obtains the following equation for the expected increase in cancer risk as a function of an increase in mutation rate:

$$\text{risk-fold} = r^{\mathrm{n}} \qquad (3)$$

This is an intuitive solution: since an increase in mutation rate (r) linearly increases the probability of each of the independent driver mutations, a moderate increase in mutation rate can lead to a much larger increase in risk.

An analogous approach was recently used to estimate the number of driver mutations required in lung adenocarcinoma by comparing the incidence of cancer in smokers and non-smokers and the change in mutation burden observed in cancers from smokers and non-smokers (*29*). This study concluded that n=3 was the integer number of driver mutations that best explained how an increase in mutation burden of 3.23-fold in smokers can cause an increase in cancer incidence of 16.2-fold (based on estimates available at the time).

Equation 3 is conceptually important as it offers an explanation for how a moderate change in mutation burden can cause a large change in cancer risk. However, there are several reasons why this model is simplistic and should be applied with caution, including:

1. The model assumes that a mutagen acts constantly throughout life.
2. The model assumes that a mutagen increases equally the rate of all driver mutations, which may not be a generally valid assumption.
3. The model does not consider clonal expansions inducible by the first driver events in precancerous tissue, which is known to affect these equations.
4. The model assumes that all cells from all individuals of each group (smokers and never-smokers) have the same mutation rate, which is unrealistic.

Sensitivity for detecting mutational signature 4 in cancer types

To evaluate whether the lack of signature 4 in some cancer types is not due to limitations in our computational approach, we performed simulations based on actual data from bladder, cervix and kidney cancers since signature 4 was not detected in any of these cancer types. A discrete number of somatic mutations (*i.e.*, 1, 5, 10, 20, 50, 75, 100, 200, 500 or 1,000) from signature 4 was *in silico* added to each sample in each of the cancer types. We evaluated the ability of our computational approach to identify the presence of signature 4 based on the respective additions of somatic mutations (Fig. S4). Adding single mutations to mutational catalogues did not allow us to find signature 4 in any sample (Fig. S4); however, as low as 5 somatic mutations were sufficient to identify signature 4 in approximately 20% of samples. These 20% of samples were ones with low mutational burden usually with two or less C>A mutations (the main feature of signature 4). In contrast, even 200 somatic mutations were not sufficient to identify signature 4 in some samples (Fig. S4) as these samples usually contained many C>A substitutions from other mutational signatures. The addition of 500 or more somatic mutations proved sufficient for the identification of signature 4 across all samples in the examined three cancer types. Overall, our simulations indicate that as low as 10 somatic mutations, reflecting 0.20 mutations per megabase in an exome, are sufficient to identify the presence of signature 4 in more than 25% of the samples in the three cancer types indicating that signature 4 would have been detected if present at this levels (Fig. S4).

Aromatic amines, an important class of carcinogens found in tobacco smoke, have also been linked to bladder cancer and experimentally shown to generate C>A mutations (*54*). However, we were unable to identify a C>A predominant mutational signature in bladder cancer and our simulations indicate that as few as 10 somatic mutations in an exome will be sufficient for detecting this signature. One interpretation of this result is that the carcinogenicity of aromatic amines in bladder cancer is due not to direct DNA damage but potentially to an activation of other mutational processes generating different mutational signatures. Another interpretation is that aromatic amines generate too few somatic mutations (<10 mutations) to be detected by our signatures approach. A further interpretation is that aromatic amines are not responsible for the increased bladder cancer risk associated with cigarette smoking.

Mutation rate variation across normal cells can lead to complex relationships between cancer mutation burden and risk

In reality, there is likely to be significant variation in the mutation rate across lung epithelial cells within an individual and across individuals in both the smoking and non-smoking population. As shows below, this can lead to very unexpected and even counterintuitive relationships between mutation rate and cancer risk.

Under the multistage model of cancer, when the mutation rate varies across cells and across individuals, cancers are much more likely to evolve from normal cells with higher mutation rates than from those with lower rates. This means that the mutation rates observed in cancers are expected to be a biased representation of the mutation rates in normal cells. Being $f(\mu)$ the probability density function of the mutation rate across normal cells, assuming a simple multi-stage model of cancer, the density function of the mutation rate across cancers can be approached by: $f_{cancer}(\mu) \propto f(\mu) * \text{risk}(\mu)$.

An example of the impact of this phenomenon is exemplified by simulations in Fig S6. For these simulations, risk(μ) is modeled as the probability of a single-cell having n=5 driver genes with at least one mutation at a given time (the product of five cumulative Poisson probabilities), as a function of an overall mutation rate per gene (μ). To explore the effect of mutation rate variation among normal cells, most cells have low mutation rates (~0.001 mutations/gene) and a small subpopulation (0.1%) of cells have a ten-fold higher basal rate (Fig. S10*A*). When evolving cancers from this population of normal cells, one can see that most cancers evolve from the hypermutator subpopulation of cells (Fig S6*A*, bottom panel).

To evaluate the effect of a mutagen, in a second simulation we increase the mutation rate of all cells by 0.001 mutations per gene (Fig. S10*B*). This nearly doubles the average mutation burden in the population of normal cells and also increases the cancer risk in the mutagen-exposed population by ~3.5-fold. When evolving cancers from this population of cells, we can now see that many cancers evolve from non-hypermutator cells, thanks to the increase in mutation rate due to the mutagen. Paradoxically, however, this causes the average mutation burden of the cancers to be lower in smokers than in non-smokers.

In this scenario, a signature analysis of the mutations observed in cancers from both mutagen-exposed and non-exposed individuals would reveal the contribution of the mutagen to cancers from exposed individuals, and a higher contribution of other signatures to cancers from non-exposed individuals. This shows that more complex relationships than suggested by equation (3) can emerge when studying the changes in cancer incidence and cancer mutation rates induced by a mutagen.

Comparison of mutational signatures in clonal and subclonal mutations in tobacco smokers and non-smokers

Somatic mutations were classified in each sample as either clonal or subclonal as described in Supplementary Methods. In total, approximately 14% of the examined mutations were classified as subclonal with the remaining 86% being clonal. Each sample was split into two virtual samples: one containing only the clonal mutations and one containing only the subclonal mutations. We then performed mutational signatures analysis using all of these virtual samples by extracting mutational signatures and estimating the contributions of each signature to each sample.

Next, we compared the clonal mutations assigned to each signature in smokers to the ones in non-smokers. The results revealed that signatures 4 and 5 are enriched in clonal mutations of smokers compared to clonal mutations of non-smokers (q-value<0.05; two sample Kolmogorov-Smirnov tests corrected for multiple hypothesis testing for all examined mutational signatures). No differences were observed for any other mutational signatures in clonal mutations. Furthermore, our analysis did not reveal significant differences for any mutational signature when comparing subclonal mutations assigned to signatures of tobacco smokers to the ones of non-smokers. In summary, signatures 4 and 5 were the only signatures enriched in clonal mutations of tobacco smokers compared to non-smokers but no difference was observed for any signature in regards to subclonal mutations.

Additionally, for all non-smokers, we compared the percentages of mutations assigned to each signature in clonal samples with the ones in subclonal samples. The results revealed statistically significant differences (q-value<0.05) for two mutational signatures: signatures 1 and 2 were enriched in clonal mutations of non-smokers. Analogous comparison in smokers revealed statistically significant differences (q-value<0.05) for four mutational signatures: signatures 1, 2, 4 and 5 were enriched in clonal mutations of smokers.

Overall, the results from these analyses are consistent with the hypothesis that signatures 4 and 5 are tobacco associated signatures since the increases in these signatures are due to cigarette smoke exposure prior to neoplastic change.

SUPPLEMENTARY FIGURES

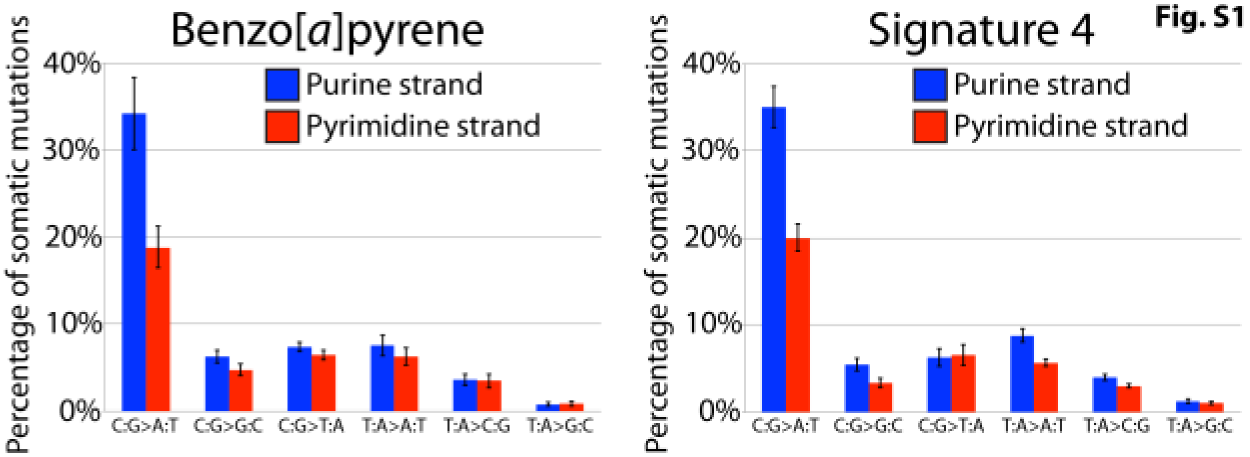


Fig. S1. Comparison of transcriptional strand bias between *in vitro* exposure to benzo[*a*]pyrene and signature 4 extracted from all tobacco-associated cancer types. The six possible types of substitutions are show on the x-axes, while the y-axes reflect the percentage of mutations attributed to each strand oriented substitution type. Red reflects pyrimidine-oriented substitutions (*e.g.*, C>A), while blue depicts purine-oriented substitutions (*e.g.*, G>T).

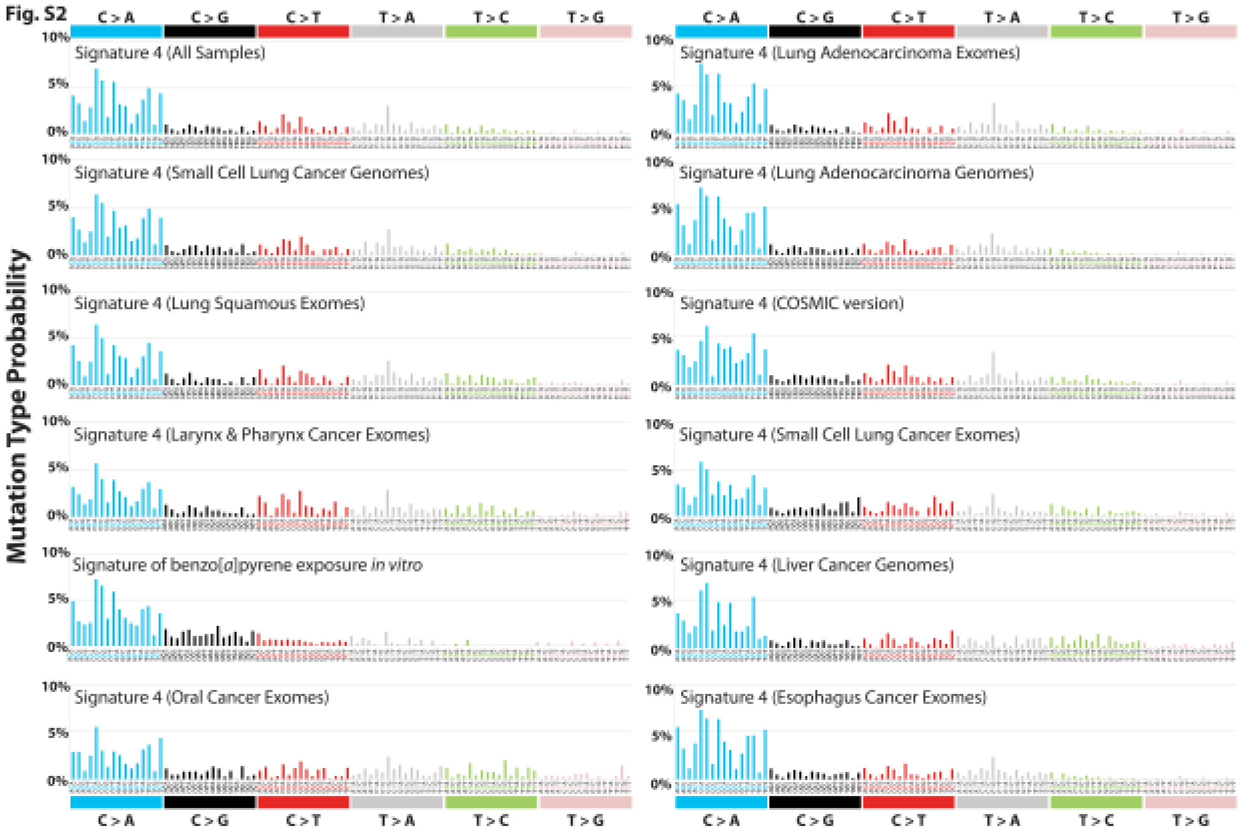


Fig. S2. Extracting the pattern of signature 4 from different cancer types. Signatures are depicted using a 96 substitution classification defined by the substitution type and sequence context immediately 5’ and 3’ to the mutated base. Different colors are used to display different types of substitutions. The percentages of mutations attributed to specific substitution types are on the vertical axes, while the horizontal axes display different types of substitutions. Mutational signatures are depicted based on the trinucleotide frequency of the whole human genome. Each panel is clearly label to indicate the caner type(s) used for extracting signature 4. The quantified similarities between signature 4 extracted from all cancer types and signature 4 extracted from individual cancer types is depicted in Fig. S3.

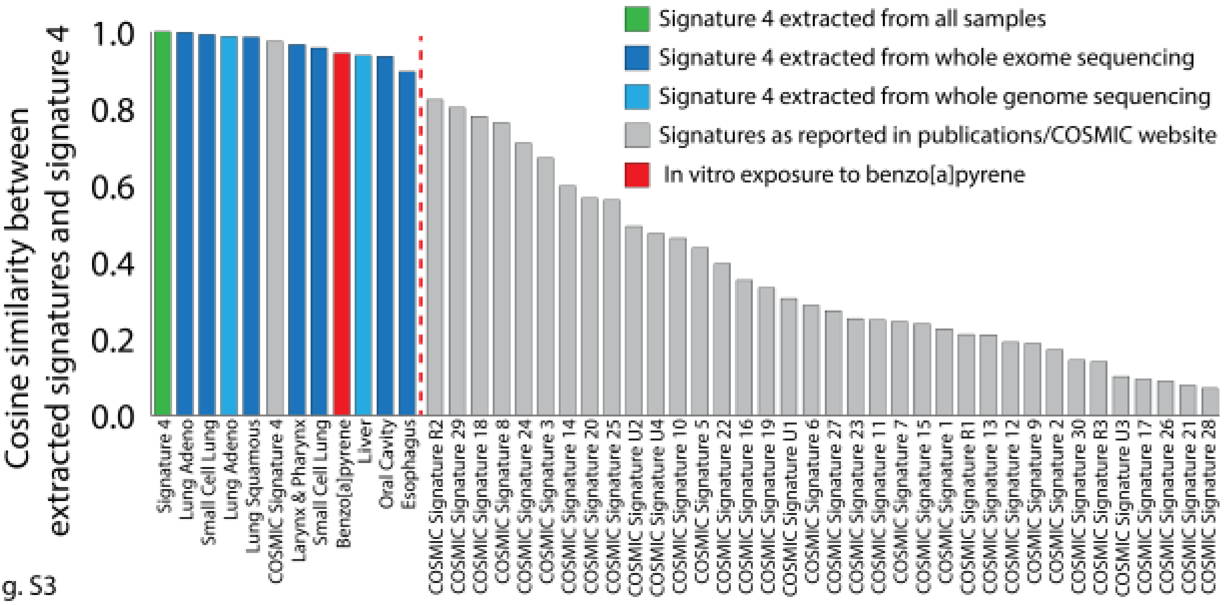


Fig. S3

Fig. S3. Similarity between extraction of signature 4 across different cancer types and other known mutational signatures. Each bar represents a mutational signature, where the height of the bar reflects the similarity between the signature and consensus signature 4 extracted from all 5,243 samples (*i.e.*, green bar). Similarity of 1.00 reflects a perfect match between mutational signatures, whereas similarity of 0.00 indicates orthogonal (*i.e.*, completely different) mutational signatures. Dark blue bars are used for mutational signatures extracted from whole exome sequencing data, light blue bars for mutational signatures extracted from whole genome sequencing data, gray bars corresponds to signatures as reported on the COSMIC website and in refs. (*17, 18*), and the red bar reflects a signature due to *in vitro* exposure by benzo[*a*]pyrene.

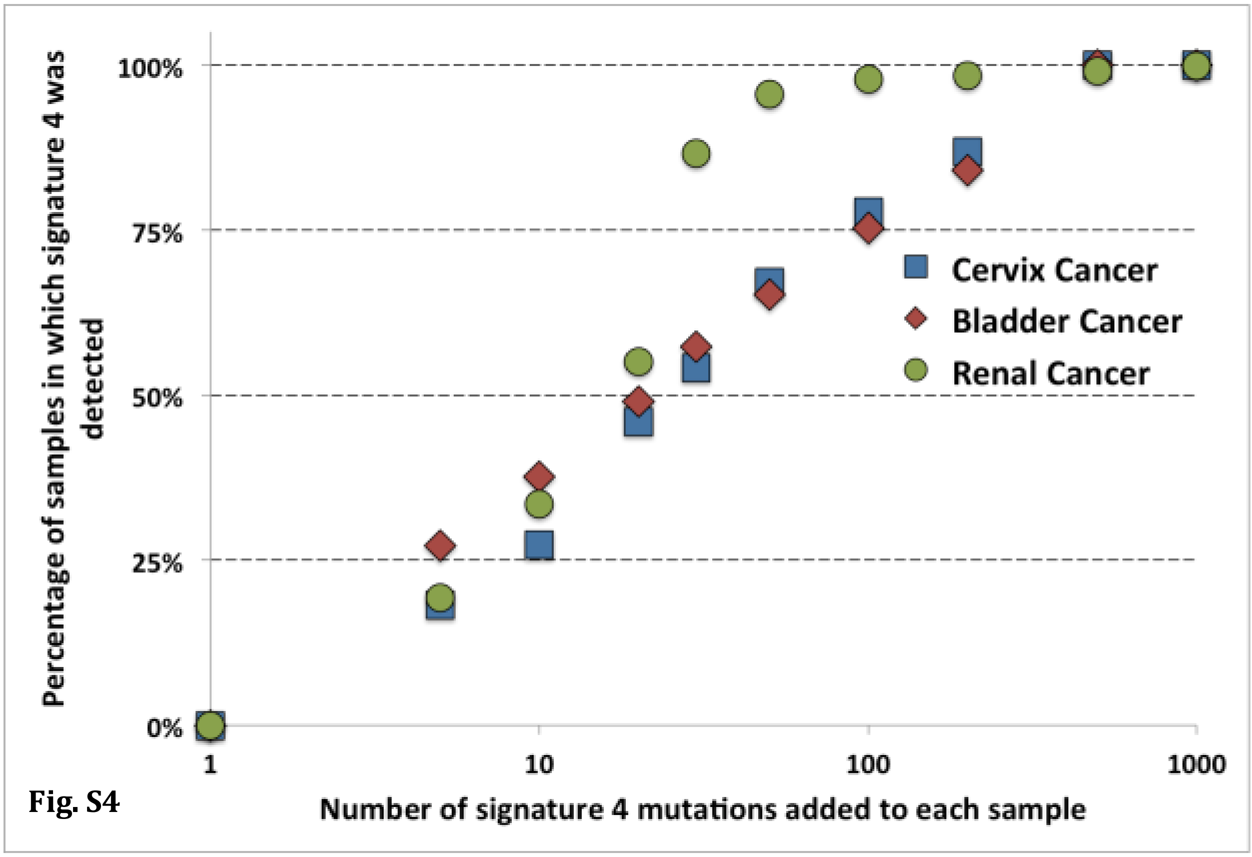


Fig. S4. Sensitivity for detecting signature 4 in cervix, bladder and renal cancer. The x-axis reflects the number of somatic mutations added to each sample (discrete values: 1, 5, 10, 20, 50, 75, 100, 200, 500 and 1,000) in logarithmic scale. The y-axis reflects the percentage of samples in which our approach can identify signature 4 after adding the respective number of somatic mutations.

Fig. S5

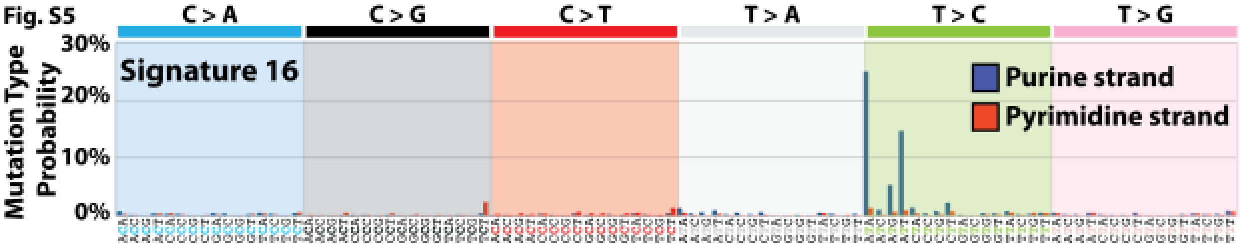


Fig. S5. Strong transcriptional strand bias of signature 16. Mutations are shown according to the 192 mutation classification incorporating the type of substitution, the sequence context immediately 5’ and 3’ to the mutated base and whether the mutated base is on the purine or pyrimidine strand of the human reference genome. Different colors are used to shade different types of substitutions. Mutations occurring on the purine strand are shown in blue, while mutations occurring on the pyrimidine strand are shown in red. The mutational signature is depicted based on the trinucleotide frequency of the whole human genome. The percentages of mutations attributed to specific substitution types are on the vertical axes, while the horizontal axes display different types of substitutions.

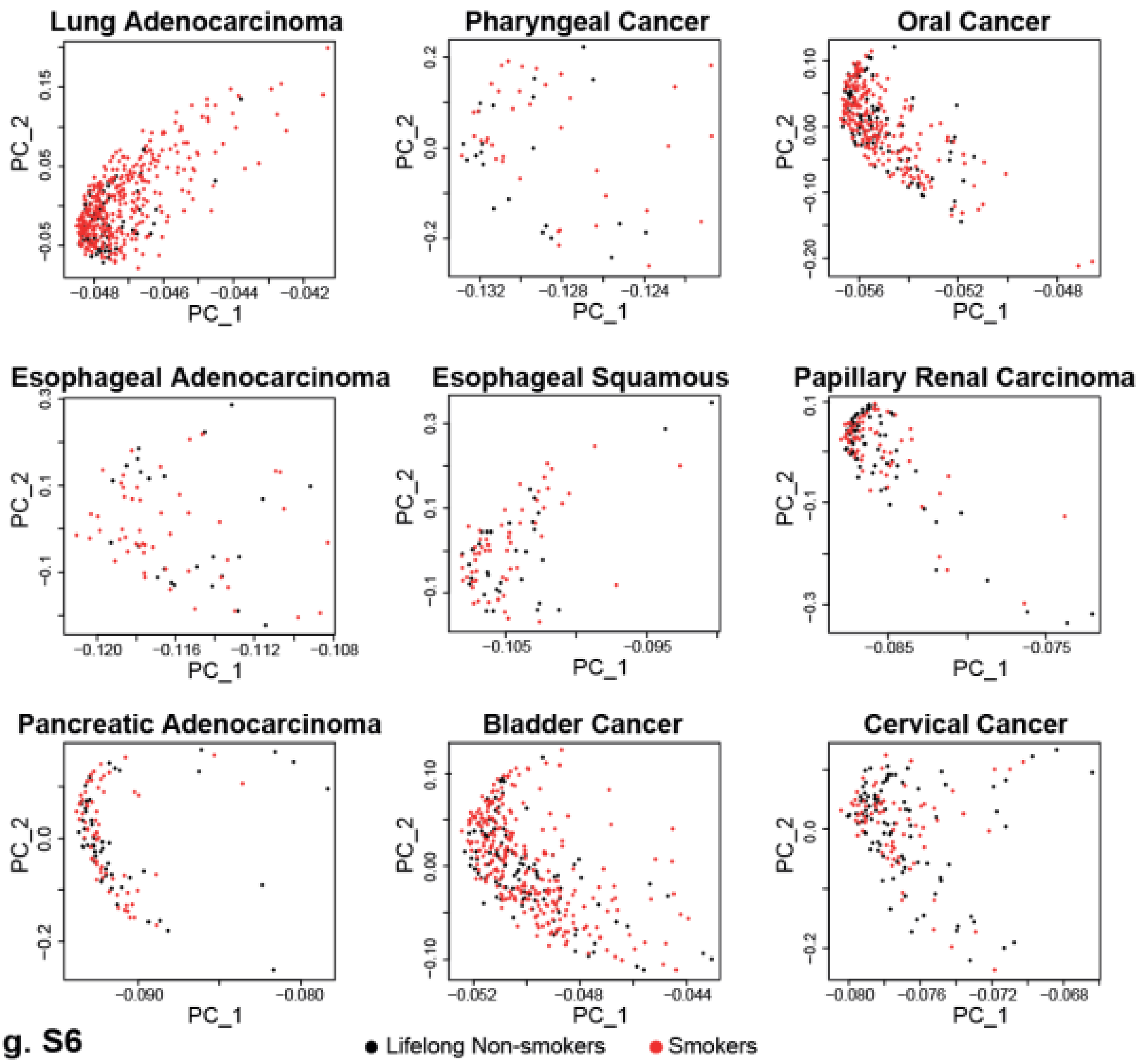
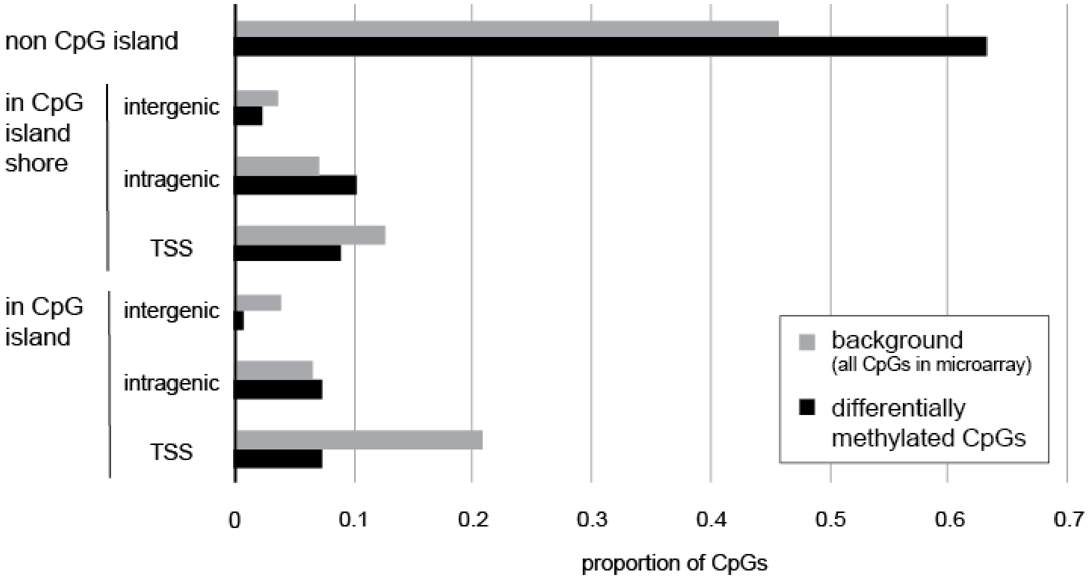


Fig. S6

Fig. S6. Comparison between lifelong non-smokers and smokers based on overall CpG methylation profiles. Results from principle component analyses reveal no remarkable difference in overall CpG methylation profiles between smokers and non-smokers. The analyses were performed using CpG methylation profiles obtained by Infinium HumanMethylation450 BeadChip. Please note that some of the red circles are overlapping with some of the black circles.



TSS: transcription start site

Fig. S7

Fig. S7. CpG islands and differentially methylated CpGs in smokers compared to non-smokers. Gray bars reflect the proportion of CpG probes in the Illumina Infinium HumanMethylation450 BeadChip. Black bars reflect the number of differentially methylated CpGs in lung adenocarcinoma and oral cancer (Tables S4 and S5). TSS stands for transcription start site.

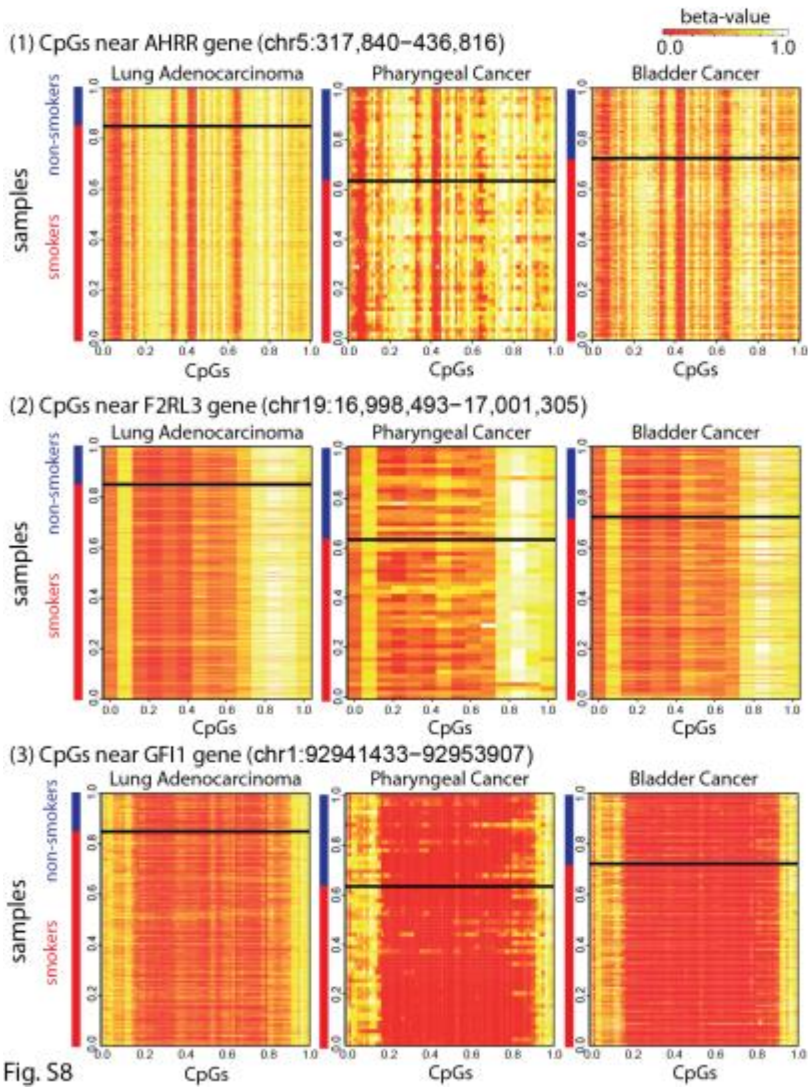


Fig. S8

Fig. S8. Cancer tissue methylation of individual CpGs near genes differentially methylated in blood or buccal cells of smokers. Heatmaps of beta values of CpG methylation are shown for three genes (AHRR, F2RL3, and GFI1) known to be differentially methylated in normal tissues of smokers (*28*). The CpGs are on the vertical axes, while the horizontal axes reflect samples in the respective cancer types. Lifelong non-smokers (top blue part of each panel) and smokers (bottom red part of each panel) have been separated manually. No difference was observed

between smokers and non-smokers. All differentially methylated CpGs in tobacco-associated cancers are provided in Tables S4 and S5.

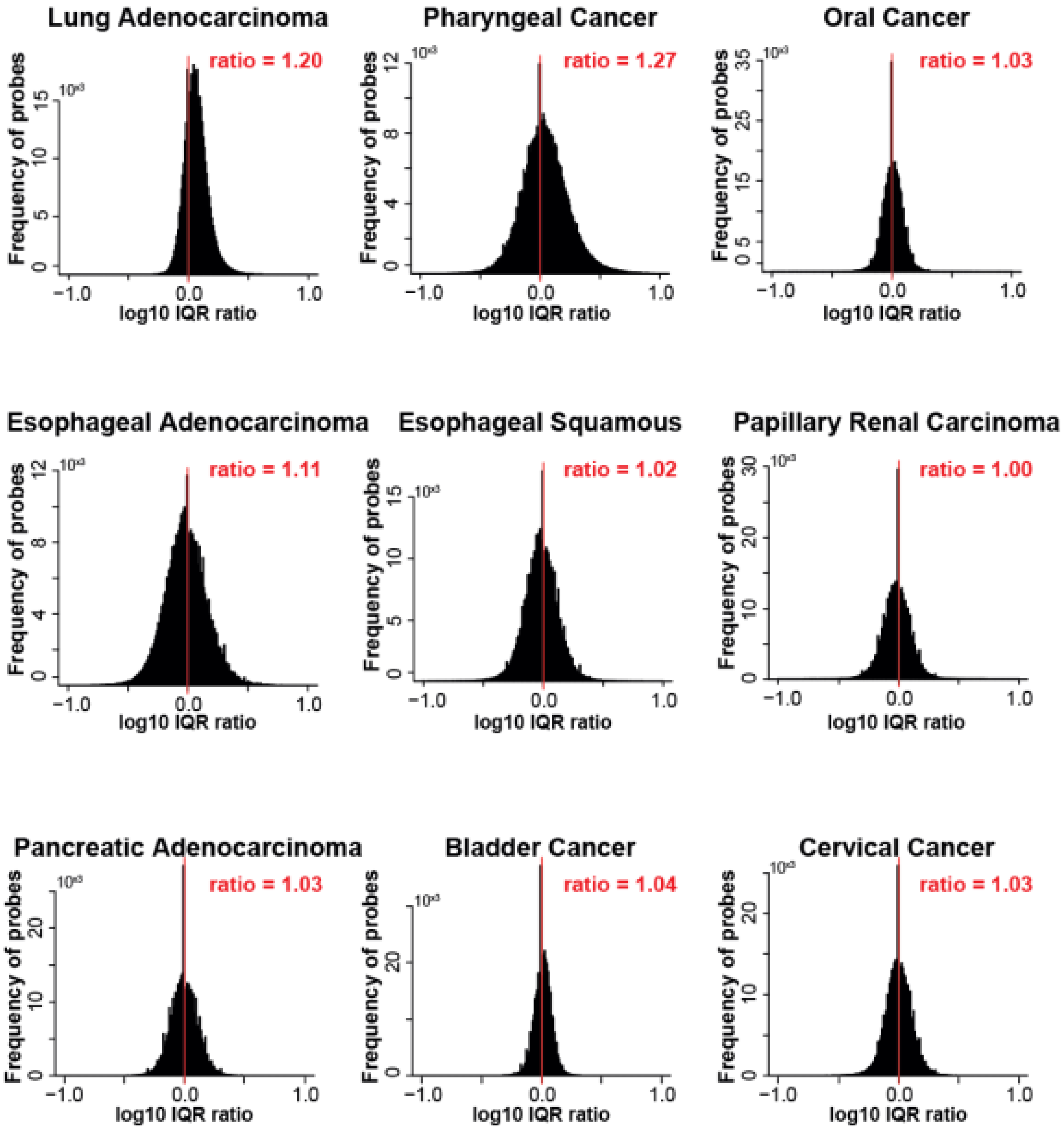


Fig. S9

Fig. S9. Interquartile range of methylation levels in tobacco-associated cancer types. The horizontal axes reflect logarithmic interquartile range, while the vertical axes reflect frequency of probes. Ratio is calculated by dividing the interquartile range of methylation levels of smokers to the one of non-smokers.

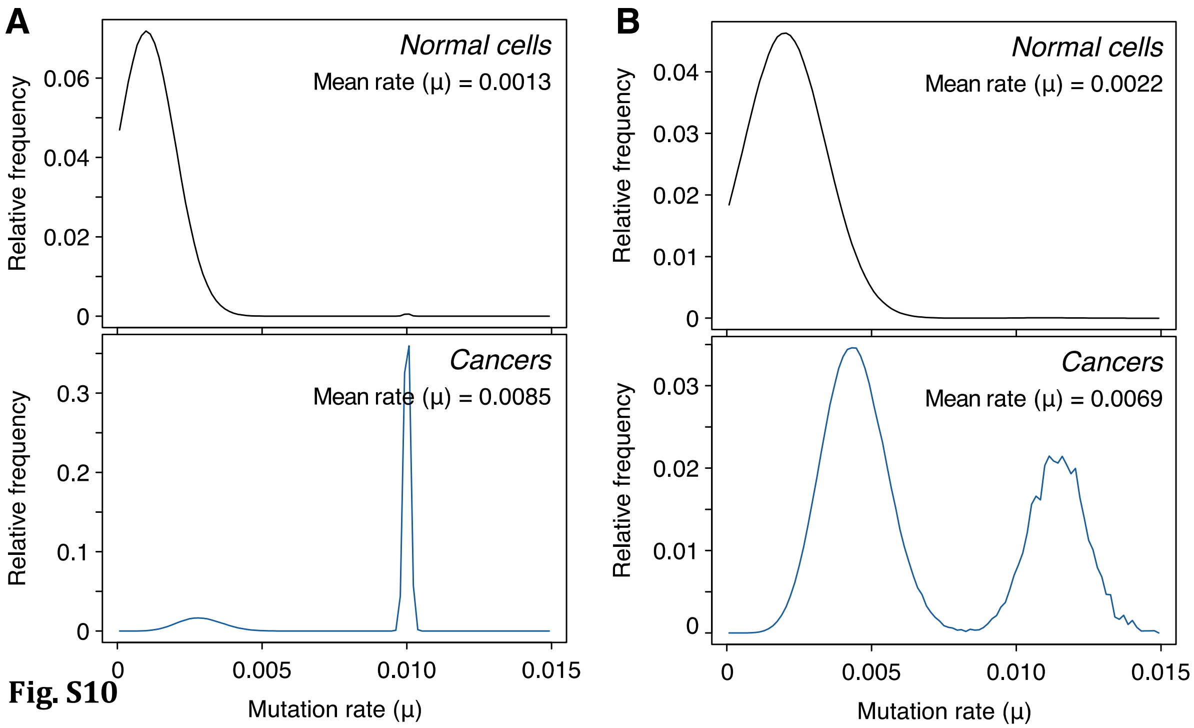


Fig. S10

Fig. S10. Mutation rate variation across normal cells can lead to complex relationships between cancer mutation burden and cancer risk. To exemplify the impact of mutation rate variation across cells, we simulate two different scenarios (see supplementary text for a detailed description). Top panels show the underlying mutation rate in normal cells in (***A***) the absence and (***B***) presence of a mutagen used in the simulation (notice the shift towards higher mutation rate in ***B*** due to the mutagen). As described in the supplementary text, a small subpopulation of normal cells are hypermutators. In this simulation, cancer emerged when a cell acquired a mutation in five driver genes. Under these conditions, it can be seen that in the absence of a mutagen, the vast majority of tumors evolve from the subpopulation of hypermutator cells (***A***). In contrast, in the presence of a mutagen, the much more frequent non-hypermutator cells have sufficiently increased mutation rates to generate tumors more frequently than hypermutator cells (***B***). As a result, in this simulation the mean cancer mutation burden is higher in patients without a mutagen exposure. This simple model exemplifies how, in the presence of variable mutation rates across cells, the relationship between mutation burden and cancer risk need not be a simple

one and that, under certain situations, some signatures could be more prevalent in tumors from patients not exposed to a mutagen.

Captions for supplementary tables

Table S1: Detailed information about each examined tobacco-associated cancer sample.
Table S2: Comparison of features of tobacco smokers to the ones of lifelong non-smokers.
Table S3: Relationships between mutational signatures and pack years smoked.
Table S4: Individual CpGs with differential methylation in lung adenocarcinoma.
Table S5: Individual CpGs with differential methylation in oral cancer.
Table S6: Numerical patterns of mutational signatures associated with tobacco smoking.

AUTHOR CONTRIBUTIONS

L.B.A. and M.R.S. conceived the overall approach and wrote the manuscript. L.B.A. and M.R.S. carried out signatures and/or statistical analyses with assistance from S.N.-Z. and P.J.C. Theoretical considerations for mutation burden in normal cells and cancer risk were developed by I.M. Methylation analysis was performed by Y.S.J., while K.H. and P.V.L. performed copy-number analysis. I.M., K.H., and P.V.L. performed the clonal and sub-clonal analyses. Y.T, A.F., H.N. and T.S. contributed data, provided expertise in regards to liver cancer, and assisted in the interpretation of the overall results. P.V. and D.H.P. assisted with the writing of the manuscript and provided expertise and advice in regards to tobacco epidemiology, mechanisms of DNA damage, and interpreting the overall results. All authors have read and edited the manuscript. Correspondence should be addressed to L.B.A. (lba@lanl.gov) and M.R.S. (mrs@sanger.ac.uk).

References and Notes

1. B. Secretan, K. Straif, R. Baan, Y. Grosse, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, V. Cogliano; WHO International Agency for Research on Cancer Monograph Working Group, A review of human carcinogens—Part E: Tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol.* **10**, 1033–1034 (2009).doi:10.1016/S1470-2045(09)70326-2 Medline

2. S. S. Lim, T. Vos, A. D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, M. Amann, H. R. Anderson, K. G. Andrews, M. Aryee, C. Atkinson, L. J. Bacchus, A. N. Bahalim, K. Balakrishnan, J. Balmes, S. Barker-Collo, A. Baxter, M. L. Bell, J. D. Blore, F. Blyth, C. Bonner, G. Borges, R. Bourne, M. Boussinesq, M. Brauer, P. Brooks, N. G. Bruce, B. Brunekreef, C. Bryan-Hancock, C. Bucello, R. Buchbinder, F. Bull, R. T. Burnett, T. E. Byers, B. Calabria, J. Carapetis, E. Carnahan, Z. Chafe, F. Charlson, H. Chen, J. S. Chen, A. T. Cheng, J. C. Child, A. Cohen, K. E. Colson, B. C. Cowie, S. Darby, S. Darling, A. Davis, L. Degenhardt, F. Dentener, D. C. Des Jarlais, K. Devries, M. Dherani, E. L. Ding, E. R. Dorsey, T. Driscoll, K. Edmond, S. E. Ali, R. E. Engell, P. J. Erwin, S. Fahimi, G. Falder, F. Farzadfar, A. Ferrari, M. M. Finucane, S. Flaxman, F. G. Fowkes, G. Freedman, M. K. Freeman, E. Gakidou, S. Ghosh, E. Giovannucci, G. Gmel, K. Graham, R. Grainger, B. Grant, D. Gunnell, H. R. Gutierrez, W. Hall, H. W. Hoek, A. Hogan, H. D. Hosgood 3rd, D. Hoy, H. Hu, B. J. Hubbell, S. J. Hutchings, S. E. Ibeanusi, G. L. Jacklyn, R. Jasrasaria, J. B. Jonas, H. Kan, J. A. Kanis, N. Kassebaum, N. Kawakami, Y. H. Khang, S. Khatibzadeh, J. P. Khoo, C. Kok, F. Laden, R. Lalloo, Q. Lan, T. Lathlean, J. L. Leasher, J. Leigh, Y. Li, J. K. Lin, S. E. Lipshultz, S. London, R. Lozano, Y. Lu, J. Mak, R. Malekzadeh, L. Mallinger, W. Marcenes, L. March, R. Marks, R. Martin, P. McGale, J. McGrath, S. Mehta, G. A. Mensah, T. R. Merriman, R. Micha, C. Michaud, V. Mishra, K. Mohd Hanafiah, A. A. Mokdad, L. Morawska, D. Mozaffarian, T. Murphy, M. Naghavi, B. Neal, P. K. Nelson, J. M. Nolla, R. Norman, C. Olives, S. B. Omer, J. Orchard, R. Osborne, B. Ostro, A. Page, K. D. Pandey, C. D. Parry, E. Passmore, J. Patra, N. Pearce, P. M. Pelizzari, M. Petzold, M. R. Phillips, D. Pope, C. A. Pope 3rd, J. Powles, M. Rao, H. Razavi, E. A. Rehfuess, J. T. Rehm, B. Ritz, F. P. Rivara, T. Roberts, C. Robinson, J. A. Rodriguez-Portales, I. Romieu, R. Room, L. C. Rosenfeld, A. Roy, L. Rushton, J. A. Salomon, U. Sampson, L. Sanchez-Riera, E. Sanman, A. Sapkota, S. Seedat, P. Shi, K. Shield, R. Shivakoti, G. M. Singh, D. A. Sleet, E. Smith, K. R. Smith, N. J. Stapelberg, K. Steenland, H. Stöckl, L. J. Stovner, K. Straif, L. Straney, G. D. Thurston, J. H. Tran, R. Van Dingenen, A. van Donkelaar, J. L. Veerman, L. Vijayakumar, R. Weintraub, M. M. Weissman, R. A. White, H. Whiteford, S. T. Wiersma, J. D. Wilkinson, H. C. Williams, W. Williams, N. Wilson, A. D. Woolf, P. Yip, J. M. Zielinski, A. D. Lopez, C. J. Murray, M. Ezzati, M. A. AlMazroa, Z. A. Memish, A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2224–2260 (2012).doi:10.1016/S0140-6736(12)61766-8 Medline

3. B. Pesch, B. Kendzia, P. Gustavsson, K.-H. Jöckel, G. Johnen, H. Pohlabeln, A. Olsson, W. Ahrens, I. M. Gross, I. Brüske, H.-E. Wichmann, F. Merletti, L. Richiardi, L. Simonato, C. Fortes, J. Siemiatycki, M.-E. Parent, D. Consonni, M. T. Landi, N. Caporaso, D. Zaridze, A. Cassidy, N. Szeszenia-Dabrowska, P. Rudnai, J. Lissowska, I. Stücker, E.

Fabianova, R. S. Dumitru, V. Bencko, L. Foretova, V. Janout, C. M. Rudin, P. Brennan, P. Boffetta, K. Straif, T. Brüning, Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int. J. Cancer* **131**, 1210–1219 (2012).doi:10.1002/ijc.27339 Medline

4. A. Agudo, C. Bonet, N. Travier, C. A. González, P. Vineis, H. B. Bueno-de-Mesquita, D. Trichopoulos, P. Boffetta, F. Clavel-Chapelon, M.-C. Boutron-Ruault, R. Kaaks, A. Lukanova, M. Schütze, H. Boeing, A. Tjonneland, J. Halkjaer, K. Overvad, C. C. Dahm, J. R. Quirós, M.-J. Sánchez, N. Larrañaga, C. Navarro, E. Ardanaz, K.-T. Khaw, N. J. Wareham, T. J. Key, N. E. Allen, A. Trichopoulou, P. Lagiou, D. Palli, S. Sieri, R. Tumino, S. Panico, H. Boshuizen, F. L. Büchner, P. H. M. Peeters, S. Borgquist, M. Almquist, G. Hallmans, I. Johansson, I. T. Gram, E. Lund, E. Weiderpass, I. Romieu, E. Riboli, Impact of cigarette smoking on cancer risk in the European prospective investigation into cancer and nutrition study. *J. Clin. Oncol.* **30**, 4550–4557 (2012).doi:10.1200/JCO.2011.41.0183 Medline

5. S. S. Hecht, Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat. Rev. Cancer* **3**, 733–744 (2003).doi:10.1038/nrc1190 Medline

6. D. H. Phillips, in *The Cancer Handbook*, M. R. Allison, Ed. (Macmillan, 2002), pp. 293–306.

7. D. H. Phillips, Smoking-related DNA and protein adducts in human tissues. *Carcinogenesis* **23**, 1979–2004 (2002).doi:10.1093/carcin/23.12.1979 Medline

8. D. H. Phillips, S. Venitt, DNA and protein adducts in human tissues resulting from exposure to tobacco smoke. *Int. J. Cancer* **131**, 2733–2753 (2012).doi:10.1002/ijc.27827 Medline

9. L. B. Alexandrov, M. R. Stratton, Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).doi:10.1016/j.gde.2013.11.014 Medline

10. P. Hainaut, M. Hollstein, p53 and human cancer: The first ten thousand mutations. *Adv. Cancer Res.* **77**, 81–137 (1999).doi:10.1016/S0065-230X(08)60785-X Medline

11. M. F. Denissenko, A. Pao, M. Tang, G. P. Pfeifer, Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in *P53*. *Science* **274**, 430–432 (1996).doi:10.1126/science.274.5286.430 Medline

12. G. P. Pfeifer, M. F. Denissenko, Formation and repair of DNA lesions in the p53 gene: Relation to cancer mutations? *Environ. Mol. Mutagen.* **31**, 197–205 (1998).doi:10.1002/(SICI)1098-2280(1998)31:3<197::AID-EM1>3.0.CO;2-I Medline

13. L. E. Smith, M. F. Denissenko, W. P. Bennett, H. Li, S. Amin, M. Tang, G. P. Pfeifer, Targeting of lung cancer mutational hotspots by polycyclic aromatic hydrocarbons. *J. Natl. Cancer Inst.* **92**, 803–811 (2000).doi:10.1093/jnci/92.10.803 Medline

14. F. Le Calvez, A. Mukeria, J. D. Hunt, O. Kelm, R. J. Hung, P. Tanière, P. Brennan, P. Boffetta, D. G. Zaridze, P. Hainaut, TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: Distinct patterns in never, former, and current smokers. *Cancer Res.* **65**, 5076–5083 (2005).doi:10.1158/0008-5472.CAN-05-0551 Medline

15. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).doi:10.1016/j.celrep.2012.12.008 Medline

16. L. B. Alexandrov, Understanding the origins of human cancer. *Science* **350**, 1175–1177 (2015).doi:10.1126/science.aad7363 Medline

17. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, M. R. Stratton, Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).doi:10.1038/ng.3441 Medline

18. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van ’t Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).doi:10.1038/nature12477 Medline

19. S. Nik-Zainal, J. E. Kucab, S. Morganella, D. Glodzik, L. B. Alexandrov, V. M. Arlt, A. Weninger, M. Hollstein, M. R. Stratton, D. H. Phillips, The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015). Medline

20. S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jönsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerød, A. Tutt, J. W. M. Martens, S. A. J. R. Aparicio, Å. Borg, A. V. Salomon, G. Thomas, A.-L. Børresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton; Breast Cancer Working Group of the International Cancer Genome Consortium, Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).doi:10.1016/j.cell.2012.04.024 Medline

21. S. A. Roberts, J. Sterling, C. Thompson, S. Harris, D. Mav, R. Shah, L. J. Klimczak, G. V. Kryukov, E. Malc, P. A. Mieczkowski, M. A. Resnick, D. A. Gordenin, Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).doi:10.1016/j.molcel.2012.03.030 Medline

22. C. Swanton, N. McGranahan, G. J. Starrett, R. S. Harris, APOBEC enzymes: Mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov.* **5**, 704–712 (2015).doi:10.1158/2159-8290.CD-15-0344 Medline

23. R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton, M. E. Hurles; UK10K Consortium, Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).doi:10.1038/ng.3469 Medline

24. J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, G. Tiao, D. J. Kwiatkowski, J. E. Rosenberg, E. M. Van Allen, A. D. D'Andrea, G. Getz, Somatic *ERCC2* mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).doi:10.1038/ng.3557 Medline

25. R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, K. Chen, J. Walker, S. McDonald, R. Bose, D. Ornitz, D. Xiong, M. You, D. J. Dooling, M. Watson, E. R. Mardis, R. K. Wilson, Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).doi:10.1016/j.cell.2012.08.024 Medline

26. M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansén, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Jänne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, M. Meyerson, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).doi:10.1016/j.cell.2012.08.029 Medline

27. A. Fujimoto, M. Furuta, Y. Totoki, T. Tsunoda, M. Kato, Y. Shiraishi, H. Tanaka, H. Taniguchi, Y. Kawakami, M. Ueno, K. Gotoh, S. Ariizumi, C. P. Wardell, S. Hayami, T. Nakamura, H. Aikata, K. Arihiro, K. A. Boroevich, T. Abe, K. Nakano, K. Maejima, A. Sasaki-Oku, A. Ohsawa, T. Shibuya, H. Nakamura, N. Hama, F. Hosoda, Y. Arai, S. Ohashi, T. Urushidate, G. Nagae, S. Yamamoto, H. Ueda, K. Tatsuno, H. Ojima, N. Hiraoka, T. Okusaka, M. Kubo, S. Marubashi, T. Yamada, S. Hirano, M. Yamamoto, H. Ohdan, K. Shimada, O. Ishikawa, H. Yamaue, K. Chayama, S. Miyano, H. Aburatani, T. Shibata, H. Nakagawa, Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).doi:10.1038/ng.3547 Medline

28. A. E. Teschendorff, Z. Yang, A. Wong, C. P. Pipinikas, Y. Jiao, A. Jones, S. Anjum, R. Hardy, H. B. Salvesen, C. Thirlwell, S. M. Janes, D. Kuh, M. Widschwendter, Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncol.* **1**, 476–485 (2015). Medline

29. C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, B. Vogelstein, Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 118–123 (2015).doi:10.1073/pnas.1421839112 Medline

30. M. Sopori, Effects of cigarette smoke on the immune system. *Nat. Rev. Immunol.* **2**, 372–377 (2002).doi:10.1038/nri803 Medline

31. H. Rubin, Selective clonal expansion and microenvironmental permissiveness in tobacco carcinogenesis. *Oncogene* **21**, 7392–7411 (2002).doi:10.1038/sj.onc.1205800 Medline

32. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).doi:10.1093/nar/29.1.308 Medline

33. G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean; 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).doi:10.1038/nature11632 Medline

34. W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, J. M. Akey; NHLBI Exome Sequencing Project, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).doi:10.1038/nature11690 Medline

35. L. B. Alexandrov, S. Nik-Zainal, H. C. Siu, S. Y. Leung, M. R. Stratton, A mutational signature in gastric cancer suggests therapeutic strategies. *Nat. Commun.* **6**, 8683 (2015).doi:10.1038/ncomms9683 Medline

36. K. Schulze, S. Imbeaud, E. Letouzé, L. B. Alexandrov, J. Calderaro, S. Rebouissou, G. Couchy, C. Meiller, J. Shinde, F. Soysouvanh, A.-L. Calatayud, R. Pinyol, L. Pelletier, C. Balabaud, A. Laurent, J.-F. Blanc, V. Mazzaferro, F. Calvo, A. Villanueva, J.-C. Nault, P. Bioulac-Sage, M. R. Stratton, J. M. Llovet, J. Zucman-Rossi, Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).doi:10.1038/ng.3252 Medline

37. S. Behjati, M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem, P. S. Tarpey, S. Roerink, J. Blokker, M. Maddison, L. Mudie, B. Robinson, S. Nik-Zainal, P. Campbell, N. Goldman, M. van de Wetering, E. Cuppen, H. Clevers, M. R. Stratton, Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).doi:10.1038/nature13448 Medline

38. N. Bolli, H. Avet-Loiseau, D. C. Wedge, P. Van Loo, L. B. Alexandrov, I. Martincorena, K. J. Dawson, F. Iorio, S. Nik-Zainal, G. R. Bignell, J. W. Hinton, Y. Li, J. M. C. Tubio, S. McLaren, S. O' Meara, A. P. Butler, J. W. Teague, L. Mudie, E. Anderson, N. Rashid, Y.-T. Tai, M. A. Shammas, A. S. Sperling, M. Fulciniti, P. G. Richardson, G. Parmigiani, F. Magrangeas, S. Minvielle, P. Moreau, M. Attal, T. Facon, P. A. Futreal, K. C. Anderson, P. J. Campbell, N. C. Munshi, Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).doi:10.1038/ncomms3997 Medline

39. Y. S. Ju, L. B. Alexandrov, M. Gerstung, I. Martincorena, S. Nik-Zainal, M. Ramakrishna, H. R. Davies, E. Papaemmanuil, G. Gundem, A. Shlien, N. Bolli, S. Behjati, P. S. Tarpey, J. Nangalia, C. E. Massie, A. P. Butler, J. W. Teague, G. S. Vassiliou, A. R. Green, M.-Q. Du, A. Unnikrishnan, J. E. Pimanda, B. T. Teh, N. Munshi, M. Greaves, P.

Vyas, A. K. El-Naggar, T. Santarius, V. P. Collins, R. Grundy, J. A. Taylor, D. N. Hayes, D. Malkin, C. S. Foster, A. Y. Warren, H. C. Whitaker, D. Brewer, R. Eeles, C. Cooper, D. Neal, T. Visakorpi, W. B. Isaacs, G. S. Bova, A. M. Flanagan, P. A. Futreal, A. G. Lynch, P. F. Chinnery, U. McDermott, M. R. Stratton, P. J. Campbell; ICGC Breast Cancer Group; ICGC Chronic Myeloid Disorders Group; ICGC Prostate Cancer Group, Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, e02935 (2014).doi:10.7554/eLife.02935 Medline

40. E. P. Murchison, D. C. Wedge, L. B. Alexandrov, B. Fu, I. Martincorena, Z. Ning, J. M. C. Tubio, E. I. Werner, J. Allen, A. B. De Nardi, E. M. Donelan, G. Marino, A. Fassati, P. J. Campbell, F. Yang, A. Burt, R. A. Weiss, M. R. Stratton, Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. *Science* **343**, 437–440 (2014).doi:10.1126/science.1247167 Medline

41. T. Helleday, S. Eshtad, S. Nik-Zainal, Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).doi:10.1038/nrg3729 Medline

42. S. Nik-Zainal, D. C. Wedge, L. B. Alexandrov, M. Petljak, A. P. Butler, N. Bolli, H. R. Davies, S. Knappskog, S. Martin, E. Papaemmanuil, M. Ramakrishna, A. Shlien, I. Simonic, Y. Xue, C. Tyler-Smith, P. J. Campbell, M. R. Stratton, Association of a germline copy number polymorphism of *APOBEC3A* and *APOBEC3B* with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487–491 (2014).doi:10.1038/ng.2955 Medline

43. M. Gerlinger, S. Horswell, J. Larkin, A. J. Rowan, M. P. Salm, I. Varela, R. Fisher, N. McGranahan, N. Matthews, C. R. Santos, P. Martinez, B. Phillimore, S. Begum, A. Rabinowitz, B. Spencer-Dene, S. Gulati, P. A. Bates, G. Stamp, L. Pickering, M. Gore, D. L. Nicol, S. Hazell, P. A. Futreal, A. Stewart, C. Swanton, Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).doi:10.1038/ng.2891 Medline

44. L. R. Yates, M. Gerstung, S. Knappskog, C. Desmedt, G. Gundem, P. Van Loo, T. Aas, L. B. Alexandrov, D. Larsimont, H. Davies, Y. Li, Y. S. Ju, M. Ramakrishna, H. K. Haugland, P. K. Lilleng, S. Nik-Zainal, S. McLaren, A. Butler, S. Martin, D. Glodzik, A. Menzies, K. Raine, J. Hinton, D. Jones, L. J. Mudie, B. Jiang, D. Vincent, A. Greene-Colozzi, P.-Y. Adnet, A. Fatima, M. Maetens, M. Ignatiadis, M. R. Stratton, C. Sotiriou, A. L. Richardson, P. E. Lønning, D. C. Wedge, P. J. Campbell, Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).doi:10.1038/nm.3886 Medline

45. R. Wagener, L. B. Alexandrov, M. Montesinos-Rongen, M. Schlesner, A. Haake, H. G. Drexler, J. Richter, G. R. Bignell, U. McDermott, R. Siebert, Analysis of mutational signatures in exomes from B-cell lymphoma cell lines suggest APOBEC3 family members to be involved in the pathogenesis of primary effusion lymphoma. *Leukemia* **29**, 1612–1615 (2015).doi:10.1038/leu.2015.22 Medline

46. P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Børresen-Dale, V. N. Kristensen, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16910–16915 (2010).doi:10.1073/pnas.1009843107 Medline

47. V. Barnett, T. Lewis, *Outliers in Statistical Data* (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley, ed. 3, 1994).

48. P. W. Holland, R. E. Welsch, Robust regression using iteratively reweighted least-squares. *Commun. Stat. Theory Methods* **6**, 813–827 (1977). doi:10.1080/03610927708827533

49. P. J. Huber, E. Ronchetti, *Robust Statistics* (Wiley Series in Probability and Statistics, Wiley, ed. 2, 2009).

50. J. Street, R. Carroll, D. Ruppert, A note on computing robust regression estimates via iteratively reweighted least squares. *Am. Stat.* **42**, 152–154 (1988).

51. M. B. Abdullah, On a robust correlation coefficient. *Statistician* **39**, 455–460 (1990). doi:10.2307/2349088

52. C. O. Nordling, A new theory on cancer-inducing mechanism. *Br. J. Cancer* **7**, 68–72 (1953).doi:10.1038/bjc.1953.8 Medline

53. P. Armitage, R. Doll, The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8**, 1–12 (1954).doi:10.1038/bjc.1954.1 Medline

54. D. T. Silverman, P. Hartge, A. S. Morrison, S. S. Devesa, Epidemiology of bladder cancer. *Hematol. Oncol. Clin. North Am.* **6**, 1–30 (1992). Medline