Science
AAAS

Supplementary Materials for

Somatic mutant clones colonize the human esophagus with age

Iñigo Martincorena[*,1,&], Joanna C. Fowler[*,1], Agnieszka Wabik[1], Andrew R. J. Lawson[1], Federico Abascal[1], Michael W. J. Hall[1,2], Alex Cagan[1], Kasumi Murai[1], Krishnaa Mahbubani[5], Michael R. Stratton[1], Rebecca C. Fitzgerald[2], Penny A. Handford[3], Peter J. Campbell[1,4], Kourosh Saeb-Parsy[5], Philip H. Jones[1,&]
Correspondence to: im3@sanger.ac.uk; pj3@sanger.ac.uk

This PDF file includes:

Materials and Methods

1. Sample collection and preparation

Esophageal tissue was obtained from deceased organ donors from whom organs were being retrieved for transplantation. Informed consent for the use of tissue was obtained from the donor's family (REC reference: 15/EE/0152 NRES Committee East of England - Cambridge South). A full thickness segment of mid-esophagus was excised within 60 minutes of circulatory arrest and preserved in University of Wisconsin (UW) organ preservation solution (Belzer UW® Cold Storage Solution, Bridge to Life, USA) until processing.

Esophageal samples were then opened longitudinally and the muscle and submucosa removed. Samples were cut into approximately 0.5x0.5cm pieces and incubated in 20mM EDTA at 37˚C for 2 hours. After this the epithelium was peeled away from the remaining submucosa using fine forceps. The epithelium was fixed in 4% paraformaldehyde (FD Neurotechnologies) for 30 minutes before being washed three times in 1xPBS.

For sequencing the esophageal epithelium was cut into 2 mm^2 samples and DNA extracted using QIAMP DNA microkit (Qiagen) by digesting overnight and following manufacturer's instructions. DNA was eluted using pre-warmed AE buffer where the first eluent was passed through the column two further times. Flash frozen esophageal muscle DNA was used as the germline control and DNA extracted as for the epithelial samples.

1.1. Histology images

In order to obtain histology images from each donor, esophageal tissue with the muscle removed was placed in 10% neutral buffered formalin (Sigma) to fix. Tissue was paraffin embedded using a Sakura Tissue_Tek VIP tissue processor. 5µm sections were cut and hematoxylin and eosin stained and coverslipped using a Leica ST5020-CV5030 autostainer. Images were taken using a NanoZoomer 2.0HT (Hammatsu) slide scanner. Example images from each donor are shown in Fig. S1.

1.2. Immunofluorescence staining and confocal imaging

Excess tissue was imaged by confocal microscopy for all donors except PD30987 and PD30273 (Fig. S2). PFA-fixed wholemounts were blocked for 2 hours in blocking buffer (0.5% bovine serum albumin, 0.25% fish skin gelatine, 0.5% Triton X-100 and 10% donkey serum) dissolved in PHEM buffer (60 mM PIPES, 25 mM HEPES, 10 mM EGTA, and 4 mM $MgSO_4 \cdot 7H_20$). Wholemounts were co-stained with KRT4 (GTX11215, Genetex) and KI67 (ab15580, Abcam) primary antibodies diluted 1:500 in blocking buffer and incubated for 24 hours at room temperature with continuous rocking. Samples were washed for a minimum of 24 hours with 0.2% Tween-20 in PHEM buffer changing daytime washes every 2-3 hours. Appropriate Alexa Fluor-conjugated secondary antibodies (1:500 dilution) and 1 µg/ml DAPI were diluted in blocking buffer without donkey serum and samples were incubated for 24 hours at room temperature with continuous rocking. Samples were washed for a minimum of 24 hours with

0.2% Tween-20 in PHEM buffer, changing daytime washes every 2-3 hours before imaging on a Leica SP8 confocal microscope. Z-stacks were rendered using Imaris software.

2. DNA sequencing and coverage metrics

In this study, we used an Agilent SureSelect custom bait capture design covering 74 cancer genes. In addition, we targeted 610 SNPs regularly scattered across the genome for copy number analysis and 1,124 SNPs within or around the 74 target genes for targeted copy number analysis. This was the same design used in our previous study on sun-exposed skin (*7*). The list was designed to include frequently mutated genes in squamous carcinomas and frequent driver genes from other cancer types, and includes most of the main driver genes of esophageal cancers. The bait set was design using the Agilent SureDesign software and custom filters, removing repetitive and low-complexity regions to maximize on-target coverage. The total size of the targeted regions was 0.67 Mb, of which 0.33 Mb correspond to coding sequences.

The list of genes selected for ultra-deep targeted sequencing is shown below:
ADAM29, ADAMTS18, AJUBA, AKT1, AKT2, APOB, ARID1A, ARID2, AURKA, BAI3, BRAF, CASP8, CCND1, CDH1, CDKN2A, CR2, CREBBP, CUL3, DICER1, EGFR, EPHA2, ERBB2, ERBB3, ERBB4, EZH2, FAT1, FAT4, FBXW7, FGFR1, FGFR2, FGFR3, FLG2, GRIN2A, GRM3, HRAS, IRF6, KCNH5, KEAP1, KMT2A, KMT2C, KMT2D, KRAS, MET, MUC17, NF1, NFE2L2, NOTCH1, NOTCH2, NOTCH3, NOTCH4, NRAS, NSD1, PCED1B, PIK3CA, PLCB1, PPP1R3A, PREX2, PTCH1, PTEN, PTPRT, RB1, RBM10, SALL1, SCN11A, SCN1A, SETD2, SMAD4, SMO, SOX2, SPHKAP, SUFU, TP53, TP63 and *TRIOBP*.

We note that two genes reported as significantly mutated in esophageal squamous carcinomas (ESCCs) after the design of this bait set were not included in this study (*ZNF750* and *TGFBR2*) (*21, 40*). As a result, the mutation frequency of these genes in normal esophageal epithelium cannot be evaluated in this study.

Samples were multiplexed and sequenced on Illumina HiSeq 2000 machines using paired-end 75bp reads. Paired-end reads were aligned with BWA (*41*) and PCR duplicates were marked using Pircard (http://broadinstitute.github.io/picard/). We then performed indel realignment on the resulting bam files using *IndelRealigner* from GATK.

When detecting mutations at very low allele fractions, low frequency contamination of DNA or libraries with material from a different individual can complicate the analysis. To minimize the impact of inter-individual contamination, only samples from the same donor were multiplexed and sequenced together. Further, HiSeq 2000 machines were used to avoid index hopping. Lack of inter-individual contamination was confirmed by deep genotyping of all samples using the high-coverage data.

After removing off-target reads, PCR duplicates and bases with base quality below 30 and mapping quality below 25, the mean effective coverage across all samples and genes was 870.7x. Across donors, median coverage varied from 722x (PD36806) to 968x (PD30986). Fig. S3A-B shows the variation in coverage across genes and samples. We note that the density of mutations per gene was not strongly influenced by differences in coverage, as a result of the dominant effect of selection. The correlation between the number of mutations per gene per kb and median coverage of the gene was: Pearson's r=0.029, *P*-value=0.81; Spearman's ρ=0.19, *P*-value=0.10.

3. Mutation calling

3.1. *ShearwaterML*: mutation calling from deep targeted sequencing data

The nature of our data requires the identification of somatic mutations present in a small fraction of the cells of a sample. As in our previous work on sun-exposed skin, here we used the *ShearwaterML* algorithm (*7, 8*) for variant calling on deep targeted data. This algorithm is publicly available as part of the *deepSNV* R package. The strength of *ShearwaterML* relies on using a collection of deeply-sequenced normal samples to learn a base-specific error model for each site of interest in the genome. This is achieved by fitting a beta-binomial distribution to each site combining the error rates across all normal samples, learning both the mean error rate at the site and the variation across libraries, and comparing the observed mutation rate in the sample of interest against this background model using a likelihood-ratio test.

ShearwaterML was used in the way described in our previous study (*7*), with three modifications. First, we extended the original model to detect insertions as well as deletions. Second, we allowed the overdispersion parameter to vary per site. Third, as described below, instead of using all other samples from a donor as a matched normal panel, we used samples from different donors as controls, filtering germline mutations after variant calling. The algorithm is described briefly below, but for additional details we refer the reader to the relevant original publications (*7, 8*).

For each position j in the genome and for each potential change at that position, $k \in$ (A,C,G,T,-,INS), where "-" denotes all deletions and "INS" denotes all insertions, let X_{ijk}, X'_{ijk} denote the number of sequencing reads reporting that nucleotide in each genomic strand orientation in sample i. The coverage at site j from each strand is denoted as n_{ij} and n'_{ij}. *ShearwaterML* then models the counts as drawn from a beta-binomial (*BB*) distribution:

$X_{ijk} \sim BB(n_{ij}, \upsilon_{ijk}, \rho_{jk})$
$X'_{ijk} \sim BB(n'_{ij}, \upsilon'_{ijk}, \rho_{jk})$

The parameters υ_{ijk} and υ'_{ijk} define the fraction of reads across all normal samples supporting a given base (*i.e.* the average error rate at site j for change k). The overdispersion parameter (ρ_{jk}) reflects how much the error rate varies across the collection of normal samples. As mentioned above, in this study we estimated a separate ρ per site, to control for differences in overdispersion across sites.

To identify mutations in a given sample, *ShearwaterML* uses a likelihood-ratio test for every site (j) and change (k). A real mutation will be present in both strands with approximately equal rates, $\mu_{jk}=\mu'_{jk}$ that must be higher than the background error rates (ν_{jk}, ν'_{jk}) to be detectable. The null hypothesis is that the counts from the sample of interest were drawn from the background beta-binomial distributions. The alternative hypothesis states that a somatic mutation is present at the site and uses an extra parameter ($\mu_{jk}=\mu'_{jk}$) for the mutation rate at this site in the sample of interest. A *P*-value is obtained from each strand using a likelihood ratio test with one degree of freedom (df) for the extra parameter μ, and *P*-values from both strands are combined using Fisher's method.

H_0: $\mu_{jk} = \nu_{jk}$ ($\mu'_{jk} = \nu'_{jk}$)
H_1: $\mu_{jk} > \nu_{jk}$ ($\mu'_{jk} > \nu'_{jk}$)

In our previous study on sun-exposed skin (*7*), for any given sample we used all other samples from the same donor as normal samples. Although this has the advantage of removing germline SNPs during variant calling, it has the risk of losing variants shared by multiple samples from the same donor. In normal esophagus, we used a continuous array of rectangular samples instead of the separate punch samples that we used previously in skin (*7*). The contiguous sampling increases the chance of mutations spanning multiple samples. Hence, instead of using all samples from the same donor to learn the background error rates, in this study we used samples from other donors from the study. In particular, we used 311 low-burden normal samples, by combining one biopsy of muscle from each donor and all esophageal samples from the three youngest donors (Table S1). For any given donor, between 189 and 311 of these samples were used as background in *Shearwater* (excluding any esophageal sample from the same donor in the background), ensuring an average background coverage higher than 150,000x.

To reduce the impact of sequencing and alignment errors we used a minimum base Phred quality of 30 and a minimum mapping quality of 25. Overdispersion estimates were estimated within the interval [10^{-6}, 0.32]. *P*-values were subject to multiple testing correction using Benjamini & Hochberg's False Discovery Rate (*42*) and a q-value cutoff of 0.01 was used to call somatic mutations. Variants were also required to have at least one supporting read from each strand. Mutations within 10bp of an indel were filtered out as they typically reflect mapping errors near the end of the read caused by the indel, although we noted that this filter had limited relevance after using indel realignment. Pairs or groups of mutations closer than 10bp were flagged for visual inspection, allowing the manual annotation of dinucleotides and complex substitution events and the removal of a small number of clustered artefacts.

3.2. Reads supporting the mutation calls and quality assessment

ShearwaterML is able to identify somatic mutations at very low allele frequencies thanks to using a site-specific error model. Some sites of the genome have high error rates owing to sequencing errors or recurrent misalignment. *ShearwaterML* adjust the sensitivity at each site according to the typical error rate of the site in the panel of background normal samples. This allows *ShearwaterML* to detect mutations at low allele frequencies at most sites in the genome where error rates are very low, while avoiding false positives at sites with high error rates.

Using the base and mapping quality scores described above, the median background mismatch rate at mutant sites (v_{jk} estimated from both strands) in this study was 8e-5 errors/bp (10% percentile: 4e-6; 90% percentile: 4e-4). This low background error rate enabled the detection of mutations at very low allele fraction, with a median variant allele fraction (VAF) from all detected mutations of 0.016 (10% percentile: 0.0050, 90% percentile: 0.10). As expected, the local coverage at mutant sites tended to be slightly higher than the average sample coverage, with a median coverage at mutant sites of 894x (10% percentile: 559, 90% percentile: 1266). Thus, all mutations detected by *ShearwaterML* were present in multiple independent mutant reads, with a median across mutations of 13 mutant reads per mutation and over 82% of all mutations supported by 5 or more independent reads. Information on the number of supporting mutant reads from each strand and the local coverage at the mutant site for each mutation is shown in Fig. S3.

To evaluate the reproducibility of the variant calls we performed a validation experiment on 16 samples. New sequencing libraries were generated from surplus DNA and sequenced at a high coverage to confirm the presence of the mutations originally called in these samples. Since the new libraries were generated from the original genomic DNA, we can confidently eliminate PCR, multiplexing, inter-library contamination and sequencing errors as potential sources of false calls in the original mutation set. A total of 135 mutations in the 16 samples had coverage higher than 800x in this validation experiment. Comparison of the original and new VAFs for these mutations showed an excellent agreement between pairs of libraries (Fig. S4A). 129/135 (95.5%) of the original calls are supported by at least 1 read in the validation experiment. 93.3% and 90.4% are supported by at least 2 or 3 mutant reads, respectively. Randomizing the sample names reveals that only ~3% of the mutations would be expected to be supported by two or more mutant reads in two unrelated libraries. Only 6/135 mutations were unsupported by the validation dataset. However, all of these originally had VAF<1% and so the lack of mutant reads supporting these calls is not unexpected given the coverage achieved in the validation experiment (>800x). Overall, despite the technical difficulties in experimentally validating variants with very low allele fractions, this experiment strongly supports the validity of >90% of our calls.

To further study the quality of the mutation calls, we studied the context-specific mutation spectra for mutations identified at different levels of statistical significance. A set of 3,618 very high-confidence mutations (*ShearwaterML* q-value<1e-10) was chosen as reference. Comparison of mutation calls generated with increasingly relaxed q-value thresholds confirms that the quality of the calls remains very high until around the q-value threshold of 0.01 used in this study and chosen a-priori (cosine similarity >0.99). However, the quality of the calls, as determined by their mutation spectra quickly drops for more relaxed q-value thresholds (Fig. S4B-C). This offers additional evidence of the overall quality of the mutation calls used in this study and supports the choice of the q-value threshold used in this study.

3.3. Collapsing mutations by distance

In this study, samples were collected using a continuous rectangular grid of 2 mm^2 samples for every biopsy of esophageal tissue processed. This means that we sequenced many adjacent samples. Large clones or clones on the edge between two samples are expected to be detected in two or more samples. Such mutations need to be collapsed into individual events, to avoid counting the same mutation multiple times in analyses of mutational signatures and selection and to obtain more accurate estimates of clone sizes (particularly important for larger clones).

To do so, we used the information about the spatial location of all samples within each piece of tissue. As expected given the small size of the mutant clones observed, the mean number of shared mutations is much higher for immediately adjacent samples and decreases rapidly with increasing Euclidean distance between pairs of samples. For example, pooling the targeted data from all nine donors, the mean number of shared mutations between two samples as a function of the distance between the samples was 0.20 shared mutations per pair (binomial CI99%: 0.195, 0.214) for pairs of samples separated by 0-1mm, 0.0048 (CI99%: 0.0037, 0.0061) for pairs separated by 4-5mm and 0.0008 for pairs separated by more than 1.5cm (CI99%: 0.00013, 0.0026). For distances beyond 8mm we noticed that the level of sharing between samples plateaus and is not significantly different from the level of sharing between samples 1.5cm away from each other. Based on these estimates, we decided to collapse mutations shared between samples closer than 10 mm. This approach greatly reduces the risk of double counting mutations from clones spanning multiple samples, without excessively collapsing genuinely independent mutations from a given donor. For mutations detected in more than two samples, we used the *igraph* R package to identify connected components of mutations shared by samples closer than 10 mm.

When more than one biopsy of tissue was available from the same donor, mutations between these biopsies were not merged as distances between biopsies were not available. In theory, this could lead to occasional clones spanning two biopsies being double-counted. However, analogous results to those presented in this paper were obtained by the overly conservative approach of collapsing all mutations within the same donor, independently of their physical distance. Table S2 shows the results of running *dNdScv* on both sets of mutation calls.

3.4. Copy number analysis of targeted sequencing samples

Copy number analysis of the ultra-deep targeted sequencing data was performed as described earlier (*7*). For a detailed description of the method please refer to the original publication. Briefly, this method uses statistical phasing of heterozygous SNPs within a gene to detect subclonal copy number changes from targeted sequencing data. Only copy number alterations leading to allelic imbalances are detectable by this method, including loss of heterozygosity and copy number gains. Thus, homozygous loss of both copies of a gene by large deletions are not detectable by this method, since these changes do not lead to allelic imbalances and merely manifest as a reduction in local coverage, which tends to be less reliable from targeted data. The use of statistical phasing of SNPs and the accuracy of B-allele fractions (BAF) obtained from the ultra-deep targeted sequencing data enabled the detection of copy number changes leading to allelic imbalances in a small percentage of cells of a sample. This method also quantifies and corrects the small allelic imbalance introduced during the hybridization capture in favor of the reference allele (*7*).

3.5. Mutation calling in whole-genome sequencing (WGS) data

To better understand the landscape of somatic mutation in normal esophagus, 21 samples that were found to be dominated by a mutant clone from the targeted data were whole-genome sequenced to a median coverage of ~37x, using 75 base-pair clipped reads sequenced in *Illumina XTen* machines. Muscle biopsies were used as matched normal samples to remove germline variation.

3.5.1. Substitutions and small insertions and deletions from WGS data

Substitutions were called using the CaVEMan (Cancer Variants through Expectation Maximization) variant caller (http://cancerit.github.io/CaVEMan). To increase the sensitivity to subclonal variants we used the following parameters: copy number = 10/2, normal contamination = 0.5, as in (*7*). To detect insertions and deletions (indels) we used cgpPindel, an adaptation of the Pindel algorithm, which uses split-read mapping (http://cancerit.github.io/cgpPindel) (*43*). Calling of genomic rearrangement was performed using BRASSI and BRASSII (BReakpoint AnalySiS), which use clusters of discordant read pairs and de novo assembly to reconstruct breakpoints (https://github.com/cancerit/BRASS). Subclonal copy number calling was performed using the Battenberg algorithm (see below).

To reduce the risk of SNP contamination, we removed calls at SNP sites present in 1,000 genomes with equal or higher than 5% population frequency, calls at sites where the matched normal genome had a coverage <10x and mutations with supporting reads in the matched normal. We restricted the signature and burden analyses to calls supported by at least 5 mutant reads with at least one supporting read from each direction and excluded a small fraction of clustered calls. Overall, conservatively 31,937 substitutions were identified in the 21 genomes.

3.5.2. Patterns of copy number alterations across the 21 whole-genomes sequenced

Subclonal copy number calling was performed on unclipped 150 base-pair reads using the Battenberg algorithm (https://github.com/cancerit/cgpBattenberg) (*44-46*). Esophageal muscle samples from each patient were used as germline controls. Fig. S9 shows the genome-wide B-allele fractions (BAF) and LogR ratios for heterozygous SNPs that were calculated and segmented by Battenberg. Table S3 lists the segments that were deemed to be both somatic (variants shared by biopsies further than 10 mm apart were excluded) and of high confidence (variants with estimated cell fraction < 5% and those spanning regions of low SNP density were excluded). The only recurrent events observed across the 21 WGS samples were those spanning the *NOTCH1* locus on chromosome 9q. These were assigned a copy number state of 2:0 (copy-neutral loss of heterozygosity) due to deviation in BAF with no concomitant change in LogR observed in the WGS samples with a high degree of clonality. Cell fractions were calculated using 2*(BAF – 0.5) for the 2:0 events and (2*BAF-1)/(1-BAF) for the sole 2:1 event detected (a whole chromosome 3 gain in PD30273ap).

A single genomic rearrangement was identified with confidence in the 21 whole genomes using BRASSII, a ~1kb intergenic tandem duplication in chromosome 2 (chr2:19476879-19477797) in sample PD30274x.

The presence of frequent copy-neutral LOH events affecting *NOTCH1* without detectable genomic rearrangements, suggests a possible mechanism of mitotic homologous recombination behind these events. To gain further insights into the places where these recombination events take place beyond the information available from the 21 whole-genomes, we exploited the targeted sequencing data. Using targeted samples with copy number variation calls in *NOTCH1* and/or *PTCH1*, the most likely breakpoints for copy-neutral LOH events affecting chromosome 9q were inferred as follows. High quality heterozygous SNPs were identified for each patient by piling up reads at common SNP sites within the bait footprint on chromosome 9 and excluding those with VAF > 0.75 and/or a mean coverage < 200 across samples. Upstream of a breakpoint, heterozygous SNPs would be expected to have BAF values around 0.5, significantly deviating from this value downstream of the breakpoint. To identify the most likely breakpoint segments, we used the minimum *P*-value obtained using one-sided Fisher's exact tests comparing the cumulative reads supporting each allele before and after each pair of consecutive SNPs. Fig. S10 shows that the breakpoints do not occur at a single recurrent position but instead appear to occur throughout the q arm of chromosome 9.

4. Mutational signatures and transcription-coupled damage

Fig. 4C,D show the mutational spectra of all 21 whole-genomes together. Fig. S7 shows the trinucleotide spectra of each of the whole-genomes separately. Overall, there was limited variation in the spectra across mutant clones and donors. This limited variation precludes *de novo* discovery of mutational signatures in these genomes, but the contribution of currently known mutational signatures can be approximately estimated using linear decomposition. To do so we used the *deconstructSigs* R package with the 30 mutational signatures described in the COSMIC website (option "signatures.ref = signatures.cosmic") (*30*). Since they use strict representations of known signatures, linear decomposition approaches can underestimate the contribution of the dominant signatures in a sample and wrongly assign unexplained residual variation to other mutational signatures. To reduce the extent of overfitting, we restricted the analysis to mutational signatures estimated to contribute at least 5% of all of the observed mutations across the 21 genomes. Fig. S8 shows the relative contribution of different mutational signatures to the entire collection of mutations and to each of the whole-genomes separately.

Globally, signature 1 (characterized by C>T mutations at CpG dinucleotides) appears to be the dominant mutational signature in the 21 whole-genomes, contributing approximately a third of all mutations observed according to *deconstructSigs*. Signatures 5 and 16 follow in frequency, with signatures 1, 5 and 16 contributing approximately 60% of all of the observed mutations.

Signature 16 has been mainly described in liver cancers and is characterized by a very strong transcription strand asymmetry. It is characterized by T>C mutations at ApT sites and currently seems to be incompletely resolved from signature 5, which also displays these peaks and strand asymmetry at those sites. Transcription strand asymmetry can result from transcription-coupled repair or transcription-associated mutagenesis. To better understand this mutational process, we stratified the mutational spectra according to gene expression level. To this end, we used the median expression level (transcripts per million) of each gene across ESCC samples from TCGA. This revealed that highly expressed genes display a strikingly higher rate of T>C mutations at ApT sites (Fig. 4D). Analysis of the mutation rates in the gene body as well as upstream and downstream of genes suggests a process of transcription-coupled mutagenesis and transcription-coupled repair variably active in these genomes, with an increase in the rate of T>C mutations at ApT sites in the transcribed strand of highly expressed genes (Fig. S8). This resembles a process of transcription-coupled mutagenesis and repair described in liver cancers (*31*).

deconstructSigs also identified two additional mutational signatures in the esophageal whole-genomes, signatures 8 and 19. Signature 8 is a C>A-rich signature observed in some breast cancers and medulloblastomas. The esophageal whole-genomes display an apparent excess of C>A mutations, not explained by signatures 1, 5 and 16, but it remains unclear whether these mutations truly correspond to signature 8 or whether this represents an instance of overfitting. The identification of signature 19, only observed in some pilocytic astrocytomas may also be the result of overfitting by *deconstructSigs*.

5. Analyses on mutant clones and allele frequencies

5.1. Mutation burden per cell

As described before (*7*), the average number of mutations per cell in a given sample (*s*) can be estimated as follows:

$$\beta_s = \sum_j \rho_j / L_{Mb} \approx 2 \sum_j v_j / L_{Mb} \qquad \text{(Eq. 1)}$$

Where L_{Mb} is the number of megabases sequenced, ρ_j is the fraction of cells of a sample carrying a mutation (j) and v_j is the variant allele fraction (VAF) of the mutations. As explained in the section below, in the absence of copy number changes, ρ_j can be approximated as $2v_j$. Aggregating all of the samples from a donor we can obtain an estimate of the mutation burden per cell in normal esophageal epithelium for each donor.

$$\beta \approx \frac{2}{L_{Mb} S} \sum_s \sum_j v_j \qquad \text{(Eq. 2)}$$

Where S is the total number of samples from a patient. As discussed in the main text, this is a lower-bound estimate as it is restricted to detectable mutations.

As shown in Fig. 2H, there is a very strong enrichment of non-synonymous mutations in the genes sequenced in this study as a result of positive clonal selection, with approximately a 2-fold increase in missense substitutions and an 8-fold increase in truncating substitutions over neutral expectation. If not accounted for, such strong positive selection could inflate estimates of mutation burden from targeted data. As described before (*7*), this can be avoided by estimating mutation burden exclusively from synonymous sites, which is how the estimates used in this manuscript were obtained.

These estimates of mutation burden are consistent with the observed number of mutations per sample in the 21 whole-genomes (Fig. S7B). We note, however, that estimation of mutation burden from the whole-genomes in this study is complicated by subclonality and limited coverage.

The estimated mutation burden and mutational signatures detected in this study are similar to those in colon, small intestine and liver from whole-genome sequencing of clonal organoids (*9*). Considering the absence of clear APOBEC mutations, the burden and signatures observed here are also compatible with those seen in ESCC tumors. This contrasts with a report of a mutation burden around 15-24 substitutions/Mb in normal esophagus and unusual mutation spectra, using a new PCR-based method for mutation detection down to single DNA molecules (*47*). We note that this mutation burden is ~20-40 times higher than that seen in the present study or in organoids from colon, small intestine and liver. In fact, this burden is higher than that seen in the vast majority of ESCC tumors, despite a lack of APOBEC mutations in (*47*), suggesting technical problems in the identification of somatic mutations in that study.

5.2. <u>Clone sizes</u>

With our experimental design, the fraction of cells in a sample that carry a mutation can be estimated using the variant allele frequency of the mutation and the local copy number at the mutant site. This enables us to estimate clone sizes as the product of the fraction of mutant cells in a sample and the area of the sample.

As explained in detail in (*7*), the relationship between the variant allele frequency of a mutation (v) and the fraction of mutant cells (ρ) can be expressed by the following equation:

$$\rho = \frac{n_{wt}v}{n_m + n_{wt}v\text{-}(n_m + n_{nm})v}$$

(Eq. 3)

Where n_{wt} is the average ploidy of cells not carrying the mutation in the sample, n_m is the number of DNA copies carrying the mutation in mutant cells, and n_{nm} is the number of copies not carrying the mutation in mutant cells. However, this expression requires precise knowledge of the local copy number, which is often unavailable in this study. Fortunately, our analyses show that cells in normal esophagus are largely diploid without frequent copy number changes except around the *NOTCH1* locus. In the simple case of heterozygous mutations in diploid cells, we have n_m=1, n_{nm}=1 and n_{wt}=2, and the equation simplifies to:

$$\rho = 2v$$

Thus, in diploid regions the fraction of cells carrying a heterozygous mutation can be estimated as twice the variant allele fraction of the mutation.

In the common case of homozygous loss of *NOTCH1* by copy-neutral LOH, we would have n_m=2, n_{nm}=0 and n_{wt}=2, and the equation simplifies to:

$$\rho = v$$

In this manuscript, owing to the unavoidable uncertainties, we largely avoid relying on estimates of clone sizes, with the following exceptions. For simplicity, the estimates used in Fig. S5B were calculated using $\rho = 2v$, although we note that these estimates are likely to be overestimates for some *NOTCH1* mutations. As described in detail in Methods S5.4, for the generation of the patchwork plots (Fig. 3) we used a conservative approach to *NOTCH1*, representing lower bound estimates of the sizes of *NOTCH1* clones by assuming bi-allelic loss. The estimate of the range of mutant clone sizes in our study from 0.01 mm^2 to 8 mm^2 is also conservative, with the largest clone being one with compound heterozygous loss of *NOTCH1*.

Thus, the clone sizes reported in this manuscript should be considered approximate estimates whose accuracy can be affected by occasional undetected copy number changes as well as differences in cell density within a sample.

5.3. Lower and upper bound estimates of the fraction of mutant epithelium

Our data allow us to derive estimates of the fraction of cells in a sample that carry non-synonymous mutations in a given gene. When only one mutation is observed in a gene in a given sample, estimation of the fraction of mutant cells requires knowing how many of the copies of the gene are affected in mutant cells. The equation below corresponds to Eq. 3, assuming that n_{wt}=2.

$$\rho = \frac{2v}{n_m + 2v\text{-}(n_m + n_{nm})v}$$

(Eq. 4)

As shown above, in the absence of a copy number change at this locus (n_m=1, n_{nm}=1), the fraction of mutant cells is: $\rho = 2v_j$. In the presence of copy-neutral LOH (n_m=2, n_{nm}=0) we obtain: $\rho = v_j$. And in the presence of hemizygous loss of the wild-type allele (n_m=1, n_{nm}=0) we obtain $\rho = 2v_j/(1+v_j) < 2v_j$. Thus, accounting for the possible existence of undetected CN-LOH or hemizygous loss at the site, we can say that ρ falls within the range (v_j, $2v_j$).

When more than one mutation is observed in a gene in a given sample, estimation of the fraction of mutant cells (ρ) requires knowing whether these mutations affect different cells of a sample or whether some co-occur within the same cells (*e.g.* as in the case of bi-allelic loss of a gene by compound heterozygous mutations). For example, if two mutations are heterozygous at a diploid locus with VAFs v_1 and v_2, the fraction of mutant cells would be ~2(v_1+v_2) if they affect different cells or ~2(max(v_1,v_2)) if one mutation is subclonal to the other or both are co-clonal (*i.e.* present in the same fraction of cells).

Assuming the biologically reasonable scenario of a maximum of two non-synonymous mutations per cell per gene, and considering the different possibilities of compound heterozygosity, copy neutral LOH and hemizygous losses, the total fraction of mutant cells for a given gene in a sample ranges from the lower bound $\Sigma_j v_j$ (which corresponds to bi-allelic mutation of the gene in all mutant cells of the sample by either CN-LOH or compound heterozygous mutations) to the higher bound $2\Sigma_j v_j$ (which corresponds to one mutant allele per gene at a diploid locus).

5.4. Generation of the patchwork summary plots

The panels in Fig. 3 aim to provide a graphical summary representation of the mutant clones detected in a typical 1cm^2 of normal esophagus. The figure summarizes the number, density and estimated sizes of mutant clones (each shown as a circle) as well as the genes affected (shown as different colors). An analogous representation was used in our previous study on aged sun-exposed skin (*7*). Only non-synonymous mutations in the 14 genes under positive selection were represented in the figure. The number, density and estimated sizes of the clones are entirely based on the sequencing data, using conservative assumptions about bi-allelic hits when dealing with *NOTCH1* mutant clones (see below). The nesting of clones and subclones is informed by the sequencing data whenever possible or randomly simulated within a sample when the data is uninformative.

This point below explains how the figures were generated.

1. Selection of samples: For each donor, a number of samples were randomly selected to amount to at least 1cm^2 in total area (*i.e.* at least 50 biopsies from every patient were randomly selected). This was done making sure that all of the samples sharing a clone were included in the random selection of samples to avoid underestimating clone sizes in the figure by selecting only a fraction of the samples in which a clone is present.

2. Clustering of mutations into clones: Typically, multiple mutations were detected per sample in the targeted sequencing data (Fig. 1). These mutations needed to be clustered into clones and subclones. Whenever possible, and exclusively for representation purposes, the clustering of mutations into likely clones and subclones was informed by the sequencing data using two approaches. First, when multiple mutations were present in a single sample, if possible we used the pigeonhole principle to nest mutations (see section S5.5). Second, when two or more mutations were shared between two or more samples, they are more likely to fall within the same clone. In such case, we grouped mutations within a clone as co-clonal if their VAFs were consistent with co-clonality across samples (VAFs not significantly different across the samples sharing these mutations, tested using a Poisson test per sample), or we nested them if the VAFs did not violate a nesting structure (*i.e.* if the VAFs of a mutation across samples were always smaller or equal to the VAFs of another mutation). When neither mutation sharing across samples nor the pigeonhole principle could be applied to cluster or nest mutations within a sample, mutations were nested randomly (see below).

3. Random nesting of clones: When the sequencing data could not inform about the way clones are nested, mutations were nested randomly, assuming no interaction between

mutations. For example, if two mutations are observed in a sample, one in 20% of cells and one in 5% of cells, we randomly nested the smaller clone into the larger clone with probability equal to the fraction of the area of the sample covered by the larger clone (*i.e.* 20%). This simple approach effectively assumes no tendency of mutations to preferentially co-occur or be mutually exclusive. Two constraints were imposed to the random nesting: (1) a maximum of two hits per gene in a clone, (2) there must be space left within a clone to draw a circular subclone without overlapping with previously drawn subclones. If the latter condition is violated, the subclone was placed within a different clone in the patchwork plot. In summary, nesting of clones was informed by the sequencing data whenever possible or assumed to be random when this information was not available.

4. Clone sizes: Given that copy number changes were found to be rare in genes other than *NOTCH1*, for the purpose of representing clone sizes in the patchwork plots we assumed that mutations in genes other than *NOTCH1* were heterozygous. Mutations in *NOTCH1* were conservatively assumed to be homozygous, effectively representing a scenario close to the lower bounds shown in Fig. 2C,G. When a single mutation was observed in *NOTCH1* in a cluster, the mutation was assumed to be in copy neutral LOH, which is consistent with our analyses showing that virtually all of the samples with an isolated *NOTCH1* mutation also show LOH in *NOTCH1* and that LOH in *NOTCH1* is most often copy neutral. This was a conservative choice to avoid inflating clone sizes for *NOTCH1* due to undetected copy number changes.

5. Representation in space: Clones were drawn randomly in the available space of the patchwork plot.

In summary, the patchwork plots are a summary representation of the sequencing data across donors, with each circle representing an observed mutation and clone sizes and clone densities estimated from the sequencing data. Clone sizes for *NOTCH1* are conservative lower-bound estimates. Nesting of clones was informed by the sequencing data whenever possible but randomly simulated in most cases, particularly when involving small subclones. The placement of clones in space was randomly simulated. Thus, these plots can be a useful summary representation of the variation in the frequency, size and density of mutant clones across donors and genes, but the nesting and placement of clones in space is typically simulated and should not be relied on.

5.5. Statistical pigeonhole principle

As described above, allele frequencies and copy number data can be used to estimate the fraction of cells carrying a mutation in a sample. The pigeonhole principle is widely used to infer whether two mutations observed in a sample must be co-occurring within the same cells (*45*). Let ρ_A and ρ_B be the fraction of cells carrying mutations A and B in a given sample. The pigeonhole principle states that if $\rho_A+\rho_B>1$, then both mutations cannot be in unrelated clones but must be nested or co-clonal. For example, if a mutation is present in 80% of all cells in a sample and another mutation is present in 30%, the latter must be a subclone of the former (*45*).

In practice, however, allele frequencies have associated sampling uncertainty. In this study, for pairs of mutation that appear to fulfill the pigeonhole principle, we conservatively used a statistical test to determine whether $\rho_A+\rho_B$ is significantly higher than 1. To do so, allele frequencies for mutations A and B are assumed to be drawn from binomial distributions and a one-sided Likelihood-Ratio Test is used to compared two hypotheses:

H_0: $\rho_A+\rho_B \leq 1$
H_1: unconstrained estimates for ρ_A and ρ_B

Since copy number changes were seen frequently in *NOTCH1* and occasionally in *TP53*, and large clones were commonly observed for both genes, in samples where a copy number change was detected in *NOTCH1* or *TP53*, mutant cell fraction estimates were halved to conservatively account for the likely scenario of copy neutral LOH. This conservative assumption can reduce the sensitivity to detect pairs of co-occurring mutations, but should lead to a higher specificity and results robust to LOH in *NOTCH1* and *TP53*. Similarly, genes in the X-chromosome of male donors were treated accordingly.

Significant pairs or groups of mutations identified using the test above (*P*-value<0.05) are shown in Figs. 4A (a subset) and Fig. S6 (full list).

5.6. Subclonal reconstruction across samples

Phylogenetic reconstruction from bulk polyclonal sequencing data requires the clustering of mutations into subclones (*45*). Different approaches exist to cluster mutations into clones, in which mutations are grouped into clusters that show consistent mutant cell fractions across related samples.

Given the simplicity of the copy number landscape in the whole-genomes that we sequenced (Fig. S9), to reconstruct the tree shown in Fig. 4B, we clustered allele frequencies directly using a mixture model across the six samples shown in the figure. Allele frequencies for each mutation in each sample were assumed to be drawn from a mixture of beta-binomial distributions, with each component corresponding to a cluster of mutations. Beta-binomial distributions were used instead of binomial distributions to account for the likely possibility of a certain level of overdispersion in the allele frequencies due to mapping biases or overlapping reads. An overdispersion parameter ρ=0.01 was used in Fig. 4B. An Expectation-Maximization (EM) algorithm was used to optimize the parameters of the mixture model (the mutant cell fraction and the weight of each cluster in each sample). The EM algorithm was run 2000 times with different random starting values of the parameters and the best overall solution was selected. The Bayesian Information Criterion (BIC) was used to select the optimal number of clusters, with a total of $n(2k\text{-}1)$ parameters (where n corresponds to the number of samples and k the number of clusters).

6. Selection analyses (dN/dS)

To study the extent of selection, we used the *dNdScv* algorithm, a maximum-likelihood implementation of dN/dS particularly adapted to somatic mutation data.

For a detailed description of this method please refer to (*4*). Briefly, dN/dS measures the ratio of the rate of non-synonymous mutations per non-synonymous site and the rate of synonymous mutations per synonymous site. dN/dS ratios can be calculated at the level of individual genes or groups of genes. When mutational biases are properly considered, dN/dS values significantly higher than 1 reveal an excess of non-synonymous mutations in a gene, reflecting the action of positive selection. Conversely, dN/dS values significantly lower than 1 reflect the action of negative selection (*4*).

In the *dNdScv* implementation, dN/dS ratios are calculated by controlling gene sequence composition and variable substitution rates using a 192-rate trinucleotide substitution model, which largely controls for context-dependent mutational processes and transcription strand biases (*4*). Mutation rates have been shown to vary considerably across genes in the genome, often depending on the expression level of a gene or genomic features such as the local chromatin state or the replication timing of the genomic region where the gene is located. In *dNdScv*, the background mutation rate of a gene is estimated by combining local information (the observed number of synonymous mutations observed in the gene) and information on the mutation density across all genes (using a negative binomial regression, which effectively fits a Gamma distribution for the variation of the mutation rate across genes) (*4*).

6.1. *dNdScv*: screen for positively selected genes

In order to identify genes under positive selection in normal esophageal epithelium we combined all of the observed mutations from the nine donors, after collapsing mutations shared across nearby biopsies (Methods S3.3).

In this study, we used the default settings of *dNdScv* (version 0.0.0.9), with the exception of restricting the fitting of the background model for indels to genes not found as significant using substitutions alone (*dNdSloc* model). *dNdSloc* is a simpler model that only uses substitution data. This was the model originally used in our earlier study on sun-exposed skin (*7*). This model is typically less sensitive than the *dNdScv* model as it does not use information about indels or the Gamma constraint. Nevertheless, *dNdSloc* detected 12 genes under positive selection in this dataset (q-value<0.01): *TP53*, *NOTCH1*, *NOTCH2*, *NOTCH3*, *FAT1*, *ARID1A*, *CUL3*, *AJUBA*, *CCND1*, *PIK3CA*, *TP63*, *KMT2D.*

One particular synonymous mutation occurring at an exon boundary in *TP53* has been shown to be a recurrent driver mutation by affecting *TP53* splicing (T125T) (*4, 48*). Consistent with this observation, 5 of the 9 synonymous mutations observed in *TP53* in this dataset were T125T mutations. These changes can be considered genuine driver events and were excluded from the dN/dS analyses to avoid artifactually inflating the mutation rate estimation in *TP53*.

Of interest, we noted that *CCND1*, which is amplified in ~57% of ESCCs (*21*), has an unusual pattern of recurrent nonsense mutations. These nonsense mutations mostly cluster on a small track of glutamic acids at the end of the Cyclin_C domain (E272-E280) and are predicted to lead to the truncation of the C-terminal of *CCND1*. The product of this truncated gene is predicted to resemble a shorter and more oncogenic protein isoform of *CCND1* often expressed in cancers by alternative splicing (cyclin D1b) (*49*).

In addition to the 14 genes found as significant by *dNdScv*, we detected canonical hotspot mutations in *HRAS* (G13C), *NRAS* (G12D) and *EGFR* (G598A) (Supplementary Dataset S1).

6.2. Interpretation of dN/dS ratios in terms of clone sizes

In our experimental design, only mutations that reach certain clone sizes are detectable by *ShearwaterML*. This gives dN/dS ratios a simple interpretation in terms of clonal expansions. For example, a dN/dS ratio of 10 for a given gene in our experiment suggests that a cell that acquires a non-synonymous mutation in that gene has a 10-fold higher probability of reaching detectable clone sizes than a cell that acquires a synonymous mutation in the same gene.

6.3. Estimation of the number of driver mutations

dN/dS ratios can be used to estimate the excess of non-synonymous mutations favored by positive selection in a given collection of mutations (*4*). In this study, the slightly conservative figure of 3,915 (CI95%: 3,829-3,988) positively-selected mutations was estimated by calculating the excess of non-synonymous substitutions and indels in the 14 genes found under significant positive selection by *dNdScv*. Confidence intervals for the excess of non-synonymous substitutions were calculated from the global Poisson regression used by *dNdScv*. For consistency with the background model used in *dNdScv*, the predicted excess for indels was calculated by estimating the number of indels per coding bp in the 62 genes not detected as significant by *dNdSloc* (see above), and confidence intervals were calculated using the confidence interval for the ratio of two Poisson observations.

6.4. Variation in selection between donors across genes

To evaluate whether certain genes are more strongly positively selected in one individual than in other individuals, we compared dN/dS ratios between donors (Fig. S5C-E) using Likelihood-Ratio Tests. A simpler and more approximate approach would be to test for differences in the fraction of non-synonymous mutations seen in a given gene in a donor compared to all other donors, which could be tested using a Fisher's Exact Test. An advantage of using dN/dS ratios instead of this simpler approach is that differences in mutational signatures across donors are corrected.

Let $\omega_{g,1}$ and $\omega_{g,2}$ be the maximum-likelihood estimates (MLEs) for the dN/dS ratios for gene g in datasets 1 and 2, respectively. Dataset 1 corresponds to all mutations observed in a given individual, and dataset 2 to the mutations present in all other individuals. Here we used MLEs under the *dNdScv* model, which, as we explained above, combines local and global information to estimate the background mutation rate in a gene (*4*). We can test for higher dN/dS ratios in a gene in a given individual using a one-sided test with the following null and alternative hypotheses:

H_0: $\omega_{g,1} \leq \omega_{g,2}$
H_1: unconstrained $\omega_{g,1}$ and $\omega_{g,2}$

In normal esophagus, we noticed that the signal of positive selection across genes tends to vary across donors with a general increase with age. When the overall signal of positive selection is higher across all genes in a given donor, the test above can detect significantly higher dN/dS ratios for a given gene that are reflective of this global difference rather than gene-specific differences. Thus, we used a more conservative test by removing the effect of global differences in dN/dS ratios across all genes. Let $\boldsymbol{\omega}_1$ be the MLE of the global dN/dS ratio from all genes other than the gene being tested. We can then estimate a gene-specific relative enrichment of non-synonymous mutations over $\boldsymbol{\omega}_1$ as a multiplicative factor ($\omega'_{g,1}$): $\omega_{g,1}=\boldsymbol{\omega}_1\omega'_{g,1}$. Based on this, a more conservative Likelihood-Ratio Test can be used to test for a gene-specific enrichment of non-synonymous mutations in a donor compared to all other donors, while removing the effect of differences in mutational signatures and global differences in the signal of selection:

H_0: $\omega'_{g,1} \leq \omega'_{g,2}$
H_1: unconstrained $\omega'_{g,1}$ and $\omega'_{g,2}$

This test was used for the analyses in Fig. S5C-E. This detected extensive differences in selection intensity across genes between donors. The strongest results in terms of significance and effect size were an enrichment of *TP53* mutations in the oldest donor, PD31182 (shown in Fig. S5E), and enrichments of *NOTCH3* mutations in PD30988 and PD36712 (shown in Fig. S5D).

As mentioned above, a one-sided Fisher's Exact Test could also be used to determine whether a given gene contains a higher fraction of all of the non-synonymous mutations in a donor than expected from all other donors. Although this test does not correct for differences in mutational signatures across donors, these differences are modest in this dataset and the Fisher's Exact Test approach yielded largely comparable results to the dN/dS-based Likelihood-Ratio Test above, with differences only around the limit of significance.

6.5. Gene mutation frequency in ESCCs and EACs

Fig. 2D,G and S5A show the frequency of non-synonymous mutations (substitutions and small indels) in different cancer genes in ESCC and EAC tumors. These numbers were obtained from TCGA, using somatic mutation calls for esophageal carcinomas from a previous study (*4*), and classifying them into ESCCs (n=88) and EACs (n=58) according to (*21*).

Although the power to detect recurrently mutated genes in EAC and ESCC from TCGA tumors is modest given the limited numbers of samples, we also ran *dNdScv* on both datasets restricted to the 74 genes sequenced in our study. This found the following genes as significantly mutated with q-value (FDR, restricted hypothesis testing) < 0.05:

- ESCC: *TP53* (q=0), *NFE2L2* (q=0), *KMT2D* (q=4e-08), *PIK3CA* (q=2.3e-07), *NOTCH1* (q=3.6e-07), *PTCH1* (q=1.1e-06), *PTEN* (q=2.7e-05), *SMAD4* (q=0.0016), *NOTCH3* (q=0.01).
- EAC: *TP53* (q=0), *ARID1A* (q=0.00056), *FBXW7* (q=0.00089), *SMAD4* (q=0.0059).

Six of the nine genes detected as significantly mutated in ESCCs were found to be under positive clonal selection in normal esophagus (Fig. 2). In addition to the six genes found under positive selection in TCGA ESCCs using *dNdScv*, other genomic studies of ESCC tumors have reported evidence of frequent mutations in several of the other genes that we find under positive selection in normal esophagus: *CUL3* (*21*) (often lost by homozygous deletions), *CCND1* (*21*) (very frequently affected by amplifications in ESCCs), *TP63* (*21*) (typically amplifications), *FAT1* (*22*) and *AJUBA* (*20, 22*). Overall, at least 11 of the 14 genes that we find under positive selection in normal esophagus appear to be recurrently mutated in ESCCs.

Interestingly, *ARID1A* has not been reported as frequently mutated in ESCC, but it is frequently mutated in esophageal adenocarcinomas (*21*).

7. Mixed effect regression models

Using the mutations from all nine donors, we studied whether donor age, the main risk factor in ESCC, correlate with clone sizes and numbers of mutations per sample, when accounting for the potential confounding effects of smoking history and gender. To do this, we used the *lme4* package in R to fit mixed effects regression models with age, gender and smoking as fixed effects and a random effect for the intercept across donors. Unlike simpler regression models, the use of this random intercept model controls for the non-independence of multiple observations per donor.

In R code, the regression model for clone sizes was:

lmer(logsize ~ age + gender + packyears + coverage + (1|donor/sample), REML=F)

And the regression model for testing differences in the total number of mutations per sample (*n*) was:

glmer(n ~ age + gender + packyears + coverage + (1|donor/sample), family="poisson")

logsize is the logarithm of the estimated clone size (assuming heterozygous mutations for simplicity given that this is a comparison across donors), and *packyears* is a continuous variable for the cumulative smoking history in units of pack years (Table S1). Since multiple clones are detected in the same samples and donors, we used a random intercept model for donors and samples. Using only a random intercept per donor yielded analogous results. Since the number of mutations and clone sizes in a sample could be affected by the sequencing coverage of the sample, we also included coverage as a fixed effect in both models. Analogous results were obtained with or without using coverage as a fixed effect.

Given the small number of donors studied (n=9), statistical testing of the effect of age was performed using non-parametric bootstrapping, generating 1,000 bootstrapped datasets by randomly sampling with replacement the donors as well as the samples within each selected donor. The resulting p-values are provided in the main text. We note that these p-values were considerably more conservative than those using Likelihood-Ratio Tests or the asymptotic Wald test, but should be more robust in a cohort of this size.

Alcohol intake was also tested as a potential confounding factor in these models. However, there were doubts about the reliability of the alcohol intake clinical records as these were collected from family members under stressful conditions near the time of death of the donor. Only two donors were recorded as consuming alcohol on a daily or weekly basis and they were among the three youngest donors of the cohort (Table S1), making this variable largely uninformative and potentially confounding. Larger cohorts will be needed to study the association between behavioral risk factors, such as smoking and alcohol consumption, and the mutational landscape in normal tissues.

1. Sample collection and preparation

Esophageal tissue was obtained from deceased organ donors from whom organs were being retrieved for transplantation. Informed consent for the use of tissue was obtained from the donor's family (REC reference: 15/EE/0152 NRES Committee East of England - Cambridge South). A full thickness segment of mid-esophagus was excised within 60 minutes of circulatory arrest and preserved in University of Wisconsin (UW) organ preservation solution (Belzer UW® Cold Storage Solution, Bridge to Life, USA) until processing.

Esophageal samples were then opened longitudinally and the muscle and submucosa removed. Samples were cut into approximately 0.5x0.5cm pieces and incubated in 20mM EDTA at 37˚C for 2 hours. After this the epithelium was peeled away from the remaining submucosa using fine forceps. The epithelium was fixed in 4% paraformaldehyde (FD Neurotechnologies) for 30 minutes before being washed three times in 1xPBS.

For sequencing the esophageal epithelium was cut into 2 mm^2 samples and DNA extracted using QIAMP DNA microkit (Qiagen) by digesting overnight and following manufacturer's instructions. DNA was eluted using pre-warmed AE buffer where the first eluent was passed through the column two further times. Flash frozen esophageal muscle DNA was used as the germline control and DNA extracted as for the epithelial samples.

1.1. Histology images

In order to obtain histology images from each donor, esophageal tissue with the muscle removed was placed in 10% neutral buffered formalin (Sigma) to fix. Tissue was paraffin embedded using a Sakura Tissue_Tek VIP tissue processor. 5µm sections were cut and hematoxylin and eosin stained and coverslipped using a Leica ST5020-CV5030 autostainer. Images were taken using a NanoZoomer 2.0HT (Hammatsu) slide scanner. Example images from each donor are shown in Fig. S1.

1.2. Immunofluorescence staining and confocal imaging

Excess tissue was imaged by confocal microscopy for all donors except PD30987 and PD30273 (Fig. S2). PFA-fixed wholemounts were blocked for 2 hours in blocking buffer (0.5% bovine serum albumin, 0.25% fish skin gelatine, 0.5% Triton X-100 and 10% donkey serum) dissolved in PHEM buffer (60 mM PIPES, 25 mM HEPES, 10 mM EGTA, and 4 mM $MgSO_4 \cdot 7H_20$). Wholemounts were co-stained with KRT4 (GTX11215, Genetex) and KI67 (ab15580, Abcam) primary antibodies diluted 1:500 in blocking buffer and incubated for 24 hours at room temperature with continuous rocking. Samples were washed for a minimum of 24 hours with 0.2% Tween-20 in PHEM buffer changing daytime washes every 2-3 hours. Appropriate Alexa Fluor-conjugated secondary antibodies (1:500 dilution) and 1 µg/ml DAPI were diluted in blocking buffer without donkey serum and samples were incubated for 24 hours at room temperature with continuous rocking. Samples were washed for a minimum of 24 hours with 0.2% Tween-20 in PHEM buffer, changing daytime washes every 2-3 hours before imaging on a Leica SP8 confocal microscope. Z-stacks were rendered using Imaris software.

2. <u>DNA sequencing and coverage metrics</u>

In this study, we used an Agilent SureSelect custom bait capture design covering 74 cancer genes. In addition, we targeted 610 SNPs regularly scattered across the genome for copy number analysis and 1,124 SNPs within or around the 74 target genes for targeted copy number analysis. This was the same design used in our previous study on sun-exposed skin (*7*). The list was designed to include frequently mutated genes in squamous carcinomas and frequent driver genes from other cancer types, and includes most of the main driver genes of esophageal cancers. The bait set was design using the Agilent SureDesign software and custom filters, removing repetitive and low-complexity regions to maximize on-target coverage. The total size of the targeted regions was 0.67 Mb, of which 0.33 Mb correspond to coding sequences.

The list of genes selected for ultra-deep targeted sequencing is shown below:
ADAM29, ADAMTS18, AJUBA, AKT1, AKT2, APOB, ARID1A, ARID2, AURKA, BAI3, BRAF, CASP8, CCND1, CDH1, CDKN2A, CR2, CREBBP, CUL3, DICER1, EGFR, EPHA2, ERBB2, ERBB3, ERBB4, EZH2, FAT1, FAT4, FBXW7, FGFR1, FGFR2, FGFR3, FLG2, GRIN2A, GRM3, HRAS, IRF6, KCNH5, KEAP1, KMT2A, KMT2C, KMT2D, KRAS, MET, MUC17, NF1, NFE2L2, NOTCH1, NOTCH2, NOTCH3, NOTCH4, NRAS, NSD1, PCED1B, PIK3CA, PLCB1, PPP1R3A, PREX2, PTCH1, PTEN, PTPRT, RB1, RBM10, SALL1, SCN11A, SCN1A, SETD2, SMAD4, SMO, SOX2, SPHKAP, SUFU, TP53, TP63 and *TRIOBP*.

We note that two genes reported as significantly mutated in esophageal squamous carcinomas (ESCCs) after the design of this bait set were not included in this study (*ZNF750* and *TGFBR2*) (*21, 40*). As a result, the mutation frequency of these genes in normal esophageal epithelium cannot be evaluated in this study.

Samples were multiplexed and sequenced on Illumina HiSeq 2000 machines using paired-end 75bp reads. Paired-end reads were aligned with BWA (*41*) and PCR duplicates were marked using Pircard (http://broadinstitute.github.io/picard/). We then performed indel realignment on the resulting bam files using *IndelRealigner* from GATK.

When detecting mutations at very low allele fractions, low frequency contamination of DNA or libraries with material from a different individual can complicate the analysis. To minimize the impact of inter-individual contamination, only samples from the same donor were multiplexed and sequenced together. Further, HiSeq 2000 machines were used to avoid index hopping. Lack of inter-individual contamination was confirmed by deep genotyping of all samples using the high-coverage data.

After removing off-target reads, PCR duplicates and bases with base quality below 30 and mapping quality below 25, the mean effective coverage across all samples and genes was 870.7x. Across donors, median coverage varied from 722x (PD36806) to 968x (PD30986). Fig. S3A-B shows the variation in coverage across genes and samples. We note that the density of mutations per gene was not strongly influenced by differences in coverage, as a result of the dominant effect of selection. The correlation between the number of mutations per gene per kb and median coverage of the gene was: Pearson's r=0.029, *P*-value=0.81; Spearman's ρ=0.19, *P*-value=0.10.

3. Mutation calling

3.1. *ShearwaterML*: mutation calling from deep targeted sequencing data

The nature of our data requires the identification of somatic mutations present in a small fraction of the cells of a sample. As in our previous work on sun-exposed skin, here we used the *ShearwaterML* algorithm (*7, 8*) for variant calling on deep targeted data. This algorithm is publicly available as part of the *deepSNV* R package. The strength of *ShearwaterML* relies on using a collection of deeply-sequenced normal samples to learn a base-specific error model for each site of interest in the genome. This is achieved by fitting a beta-binomial distribution to each site combining the error rates across all normal samples, learning both the mean error rate at the site and the variation across libraries, and comparing the observed mutation rate in the sample of interest against this background model using a likelihood-ratio test.

ShearwaterML was used in the way described in our previous study (*7*), with three modifications. First, we extended the original model to detect insertions as well as deletions. Second, we allowed the overdispersion parameter to vary per site. Third, as described below, instead of using all other samples from a donor as a matched normal panel, we used samples from different donors as controls, filtering germline mutations after variant calling. The algorithm is described briefly below, but for additional details we refer the reader to the relevant original publications (*7, 8*).

For each position j in the genome and for each potential change at that position, $k \in$ (A,C,G,T,-,INS), where "-" denotes all deletions and "INS" denotes all insertions, let X_{ijk}, X'_{ijk} denote the number of sequencing reads reporting that nucleotide in each genomic strand orientation in sample i. The coverage at site j from each strand is denoted as n_{ij} and n'_{ij}. *ShearwaterML* then models the counts as drawn from a beta-binomial (*BB*) distribution:

$X_{ijk} \sim BB(n_{ij},\upsilon_{ijk},\rho_{jk})$
$X'_{ijk} \sim BB(n'_{ij},\upsilon'_{ijk},\rho_{jk})$

The parameters υ_{ijk} and υ'_{ijk} define the fraction of reads across all normal samples supporting a given base (*i.e.* the average error rate at site j for change k). The overdispersion parameter (ρ_{jk}) reflects how much the error rate varies across the collection of normal samples. As mentioned above, in this study we estimated a separate ρ per site, to control for differences in overdispersion across sites.

To identify mutations in a given sample, *ShearwaterML* uses a likelihood-ratio test for every site (j) and change (k). A real mutation will be present in both strands with approximately equal rates, μ_{jk}=μ'_{jk} that must be higher than the background error rates (v_{jk}, v'_{jk}) to be detectable. The null hypothesis is that the counts from the sample of interest were drawn from the background beta-binomial distributions. The alternative hypothesis states that a somatic mutation is present at the site and uses an extra parameter (μ_{jk}=μ'_{jk}) for the mutation rate at this site in the sample of interest. A *P*-value is obtained from each strand using a likelihood ratio test with one degree of freedom (df) for the extra parameter μ, and *P*-values from both strands are combined using Fisher's method.

H_0: $\mu_{jk} = v_{jk}$ ($\mu'_{jk} = v'_{jk}$)
H_1: $\mu_{jk} > v_{jk}$ ($\mu'_{jk} > v'_{jk}$)

In our previous study on sun-exposed skin (*7*), for any given sample we used all other samples from the same donor as normal samples. Although this has the advantage of removing germline SNPs during variant calling, it has the risk of losing variants shared by multiple samples from the same donor. In normal esophagus, we used a continuous array of rectangular samples instead of the separate punch samples that we used previously in skin (*7*). The contiguous sampling increases the chance of mutations spanning multiple samples. Hence, instead of using all samples from the same donor to learn the background error rates, in this study we used samples from other donors from the study. In particular, we used 311 low-burden normal samples, by combining one biopsy of muscle from each donor and all esophageal samples from the three youngest donors (Table S1). For any given donor, between 189 and 311 of these samples were used as background in *Shearwater* (excluding any esophageal sample from the same donor in the background), ensuring an average background coverage higher than 150,000x.

To reduce the impact of sequencing and alignment errors we used a minimum base Phred quality of 30 and a minimum mapping quality of 25. Overdispersion estimates were estimated within the interval [10^{-6}, 0.32]. *P*-values were subject to multiple testing correction using Benjamini & Hochberg's False Discovery Rate (*42*) and a q-value cutoff of 0.01 was used to call somatic mutations. Variants were also required to have at least one supporting read from each strand. Mutations within 10bp of an indel were filtered out as they typically reflect mapping errors near the end of the read caused by the indel, although we noted that this filter had limited relevance after using indel realignment. Pairs or groups of mutations closer than 10bp were flagged for visual inspection, allowing the manual annotation of dinucleotides and complex substitution events and the removal of a small number of clustered artefacts.

3.2. Reads supporting the mutation calls and quality assessment

ShearwaterML is able to identify somatic mutations at very low allele frequencies thanks to using a site-specific error model. Some sites of the genome have high error rates owing to sequencing errors or recurrent misalignment. *ShearwaterML* adjust the sensitivity at each site according to the typical error rate of the site in the panel of background normal samples. This allows *ShearwaterML* to detect mutations at low allele frequencies at most sites in the genome where error rates are very low, while avoiding false positives at sites with high error rates.

Using the base and mapping quality scores described above, the median background mismatch rate at mutant sites (v_{jk} estimated from both strands) in this study was 8e-5 errors/bp (10% percentile: 4e-6; 90% percentile: 4e-4). This low background error rate enabled the detection of mutations at very low allele fraction, with a median variant allele fraction (VAF) from all detected mutations of 0.016 (10% percentile: 0.0050, 90% percentile: 0.10). As expected, the local coverage at mutant sites tended to be slightly higher than the average sample coverage, with a median coverage at mutant sites of 894x (10% percentile: 559, 90% percentile: 1266). Thus, all mutations detected by *ShearwaterML* were present in multiple independent mutant reads, with a median across mutations of 13 mutant reads per mutation and over 82% of all mutations supported by 5 or more independent reads. Information on the number of supporting mutant reads from each strand and the local coverage at the mutant site for each mutation is shown in Fig. S3.

To evaluate the reproducibility of the variant calls we performed a validation experiment on 16 samples. New sequencing libraries were generated from surplus DNA and sequenced at a high coverage to confirm the presence of the mutations originally called in these samples. Since the new libraries were generated from the original genomic DNA, we can confidently eliminate PCR, multiplexing, inter-library contamination and sequencing errors as potential sources of false calls in the original mutation set. A total of 135 mutations in the 16 samples had coverage higher than 800x in this validation experiment. Comparison of the original and new VAFs for these mutations showed an excellent agreement between pairs of libraries (Fig. S4A). 129/135 (95.5%) of the original calls are supported by at least 1 read in the validation experiment. 93.3% and 90.4% are supported by at least 2 or 3 mutant reads, respectively. Randomizing the sample names reveals that only ~3% of the mutations would be expected to be supported by two or more mutant reads in two unrelated libraries. Only 6/135 mutations were unsupported by the validation dataset. However, all of these originally had VAF<1% and so the lack of mutant reads supporting these calls is not unexpected given the coverage achieved in the validation experiment (>800x). Overall, despite the technical difficulties in experimentally validating variants with very low allele fractions, this experiment strongly supports the validity of >90% of our calls.

To further study the quality of the mutation calls, we studied the context-specific mutation spectra for mutations identified at different levels of statistical significance. A set of 3,618 very high-confidence mutations (*ShearwaterML* q-value<1e-10) was chosen as reference. Comparison of mutation calls generated with increasingly relaxed q-value thresholds confirms that the quality of the calls remains very high until around the q-value threshold of 0.01 used in this study and chosen a-priori (cosine similarity >0.99). However, the quality of the calls, as determined by their mutation spectra quickly drops for more relaxed q-value thresholds (Fig. S4B-C). This offers additional evidence of the overall quality of the mutation calls used in this study and supports the choice of the q-value threshold used in this study.

3.3. Collapsing mutations by distance

In this study, samples were collected using a continuous rectangular grid of 2 mm^2 samples for every biopsy of esophageal tissue processed. This means that we sequenced many adjacent samples. Large clones or clones on the edge between two samples are expected to be detected in two or more samples. Such mutations need to be collapsed into individual events, to avoid counting the same mutation multiple times in analyses of mutational signatures and selection and to obtain more accurate estimates of clone sizes (particularly important for larger clones).

To do so, we used the information about the spatial location of all samples within each piece of tissue. As expected given the small size of the mutant clones observed, the mean number of shared mutations is much higher for immediately adjacent samples and decreases rapidly with increasing Euclidean distance between pairs of samples. For example, pooling the targeted data from all nine donors, the mean number of shared mutations between two samples as a function of the distance between the samples was 0.20 shared mutations per pair (binomial CI99%: 0.195, 0.214) for pairs of samples separated by 0-1mm, 0.0048 (CI99%: 0.0037, 0.0061) for pairs separated by 4-5mm and 0.0008 for pairs separated by more than 1.5cm (CI99%: 0.00013, 0.0026). For distances beyond 8mm we noticed that the level of sharing between samples plateaus and is not significantly different from the level of sharing between samples 1.5cm away from each other. Based on these estimates, we decided to collapse mutations shared between samples closer than 10 mm. This approach greatly reduces the risk of double counting mutations from clones spanning multiple samples, without excessively collapsing genuinely independent mutations from a given donor. For mutations detected in more than two samples, we used the *igraph* R package to identify connected components of mutations shared by samples closer than 10 mm.

When more than one biopsy of tissue was available from the same donor, mutations between these biopsies were not merged as distances between biopsies were not available. In theory, this could lead to occasional clones spanning two biopsies being double-counted. However, analogous results to those presented in this paper were obtained by the overly conservative approach of collapsing all mutations within the same donor, independently of their physical distance. Table S2 shows the results of running *dNdScv* on both sets of mutation calls.

3.4. Copy number analysis of targeted sequencing samples

Copy number analysis of the ultra-deep targeted sequencing data was performed as described earlier (*7*). For a detailed description of the method please refer to the original publication. Briefly, this method uses statistical phasing of heterozygous SNPs within a gene to detect subclonal copy number changes from targeted sequencing data. Only copy number alterations leading to allelic imbalances are detectable by this method, including loss of heterozygosity and copy number gains. Thus, homozygous loss of both copies of a gene by large deletions are not detectable by this method, since these changes do not lead to allelic imbalances and merely manifest as a reduction in local coverage, which tends to be less reliable from targeted data. The use of statistical phasing of SNPs and the accuracy of B-allele fractions (BAF) obtained from the ultra-deep targeted sequencing data enabled the detection of copy number changes leading to allelic imbalances in a small percentage of cells of a sample. This method also quantifies and corrects the small allelic imbalance introduced during the hybridization capture in favor of the reference allele (*7*).

3.5. Mutation calling in whole-genome sequencing (WGS) data

To better understand the landscape of somatic mutation in normal esophagus, 21 samples that were found to be dominated by a mutant clone from the targeted data were whole-genome sequenced to a median coverage of ~37x, using 75 base-pair clipped reads sequenced in *Illumina XTen* machines. Muscle biopsies were used as matched normal samples to remove germline variation.

3.5.1. Substitutions and small insertions and deletions from WGS data

Substitutions were called using the CaVEMan (Cancer Variants through Expectation Maximization) variant caller (http://cancerit.github.io/CaVEMan). To increase the sensitivity to subclonal variants we used the following parameters: copy number = 10/2, normal contamination = 0.5, as in (*7*). To detect insertions and deletions (indels) we used cgpPindel, an adaptation of the Pindel algorithm, which uses split-read mapping (http://cancerit.github.io/cgpPindel) (*43*). Calling of genomic rearrangement was performed using BRASSI and BRASSII (BReakpoint AnalySiS), which use clusters of discordant read pairs and de novo assembly to reconstruct breakpoints (https://github.com/cancerit/BRASS). Subclonal copy number calling was performed using the Battenberg algorithm (see below).

To reduce the risk of SNP contamination, we removed calls at SNP sites present in 1,000 genomes with equal or higher than 5% population frequency, calls at sites where the matched normal genome had a coverage <10x and mutations with supporting reads in the matched normal. We restricted the signature and burden analyses to calls supported by at least 5 mutant reads with at least one supporting read from each direction and excluded a small fraction of clustered calls. Overall, conservatively 31,937 substitutions were identified in the 21 genomes.

3.5.2. Patterns of copy number alterations across the 21 whole-genomes sequenced

Subclonal copy number calling was performed on unclipped 150 base-pair reads using the Battenberg algorithm (https://github.com/cancerit/cgpBattenberg) (*44-46*). Esophageal muscle samples from each patient were used as germline controls. Fig. S9 shows the genome-wide B-allele fractions (BAF) and LogR ratios for heterozygous SNPs that were calculated and segmented by Battenberg. Table S3 lists the segments that were deemed to be both somatic (variants shared by biopsies further than 10 mm apart were excluded) and of high confidence (variants with estimated cell fraction < 5% and those spanning regions of low SNP density were excluded). The only recurrent events observed across the 21 WGS samples were those spanning the *NOTCH1* locus on chromosome 9q. These were assigned a copy number state of 2:0 (copy-neutral loss of heterozygosity) due to deviation in BAF with no concomitant change in LogR observed in the WGS samples with a high degree of clonality. Cell fractions were calculated using 2*(BAF – 0.5) for the 2:0 events and (2*BAF-1)/(1-BAF) for the sole 2:1 event detected (a whole chromosome 3 gain in PD30273ap).

A single genomic rearrangement was identified with confidence in the 21 whole genomes using BRASSII, a ~1kb intergenic tandem duplication in chromosome 2 (chr2:19476879-19477797) in sample PD30274x.

The presence of frequent copy-neutral LOH events affecting *NOTCH1* without detectable genomic rearrangements, suggests a possible mechanism of mitotic homologous recombination behind these events. To gain further insights into the places where these recombination events take place beyond the information available from the 21 whole-genomes, we exploited the targeted sequencing data. Using targeted samples with copy number variation calls in *NOTCH1* and/or *PTCH1*, the most likely breakpoints for copy-neutral LOH events affecting chromosome 9q were inferred as follows. High quality heterozygous SNPs were identified for each patient by piling up reads at common SNP sites within the bait footprint on chromosome 9 and excluding those with VAF > 0.75 and/or a mean coverage < 200 across samples. Upstream of a breakpoint, heterozygous SNPs would be expected to have BAF values around 0.5, significantly deviating from this value downstream of the breakpoint. To identify the most likely breakpoint segments, we used the minimum *P*-value obtained using one-sided Fisher's exact tests comparing the cumulative reads supporting each allele before and after each pair of consecutive SNPs. Fig. S10 shows that the breakpoints do not occur at a single recurrent position but instead appear to occur throughout the q arm of chromosome 9.

4. Mutational signatures and transcription-coupled damage

Fig. 4C,D show the mutational spectra of all 21 whole-genomes together. Fig. S7 shows the trinucleotide spectra of each of the whole-genomes separately. Overall, there was limited variation in the spectra across mutant clones and donors. This limited variation precludes *de novo* discovery of mutational signatures in these genomes, but the contribution of currently known mutational signatures can be approximately estimated using linear decomposition. To do so we used the *deconstructSigs* R package with the 30 mutational signatures described in the COSMIC website (option "signatures.ref = signatures.cosmic") (*30*). Since they use strict representations of known signatures, linear decomposition approaches can underestimate the contribution of the dominant signatures in a sample and wrongly assign unexplained residual variation to other mutational signatures. To reduce the extent of overfitting, we restricted the analysis to mutational signatures estimated to contribute at least 5% of all of the observed mutations across the 21 genomes. Fig. S8 shows the relative contribution of different mutational signatures to the entire collection of mutations and to each of the whole-genomes separately.

Globally, signature 1 (characterized by C>T mutations at CpG dinucleotides) appears to be the dominant mutational signature in the 21 whole-genomes, contributing approximately a third of all mutations observed according to *deconstructSigs*. Signatures 5 and 16 follow in frequency, with signatures 1, 5 and 16 contributing approximately 60% of all of the observed mutations.

Signature 16 has been mainly described in liver cancers and is characterized by a very strong transcription strand asymmetry. It is characterized by T>C mutations at ApT sites and currently seems to be incompletely resolved from signature 5, which also displays these peaks and strand asymmetry at those sites. Transcription strand asymmetry can result from transcription-coupled repair or transcription-associated mutagenesis. To better understand this mutational process, we stratified the mutational spectra according to gene expression level. To this end, we used the median expression level (transcripts per million) of each gene across ESCC samples from TCGA. This revealed that highly expressed genes display a strikingly higher rate of T>C mutations at ApT sites (Fig. 4D). Analysis of the mutation rates in the gene body as well as upstream and downstream of genes suggests a process of transcription-coupled mutagenesis and transcription-coupled repair variably active in these genomes, with an increase in the rate of T>C mutations at ApT sites in the transcribed strand of highly expressed genes (Fig. S8). This resembles a process of transcription-coupled mutagenesis and repair described in liver cancers (*31*).

deconstructSigs also identified two additional mutational signatures in the esophageal whole-genomes, signatures 8 and 19. Signature 8 is a C>A-rich signature observed in some breast cancers and medulloblastomas. The esophageal whole-genomes display an apparent excess of C>A mutations, not explained by signatures 1, 5 and 16, but it remains unclear whether these mutations truly correspond to signature 8 or whether this represents an instance of overfitting. The identification of signature 19, only observed in some pilocytic astrocytomas may also be the result of overfitting by *deconstructSigs*.

5. Analyses on mutant clones and allele frequencies

5.1. Mutation burden per cell

As described before (*7*), the average number of mutations per cell in a given sample (s) can be estimated as follows:

$$\beta_s = \sum_j \rho_j / L_{Mb} \approx 2 \sum_j v_j / L_{Mb} \qquad \text{(Eq. 1)}$$

Where L_{Mb} is the number of megabases sequenced, ρ_j is the fraction of cells of a sample carrying a mutation (j) and v_j is the variant allele fraction (VAF) of the mutations. As explained in the section below, in the absence of copy number changes, ρ_j can be approximated as $2v_j$. Aggregating all of the samples from a donor we can obtain an estimate of the mutation burden per cell in normal esophageal epithelium for each donor.

$$\beta \approx \frac{2}{L_{Mb}S} \sum_s \sum_j v_j \qquad \text{(Eq. 2)}$$

Where S is the total number of samples from a patient. As discussed in the main text, this is a lower-bound estimate as it is restricted to detectable mutations.

As shown in Fig. 2H, there is a very strong enrichment of non-synonymous mutations in the genes sequenced in this study as a result of positive clonal selection, with approximately a 2-fold increase in missense substitutions and an 8-fold increase in truncating substitutions over neutral expectation. If not accounted for, such strong positive selection could inflate estimates of mutation burden from targeted data. As described before (*7*), this can be avoided by estimating mutation burden exclusively from synonymous sites, which is how the estimates used in this manuscript were obtained.

These estimates of mutation burden are consistent with the observed number of mutations per sample in the 21 whole-genomes (Fig. S7B). We note, however, that estimation of mutation burden from the whole-genomes in this study is complicated by subclonality and limited coverage.

The estimated mutation burden and mutational signatures detected in this study are similar to those in colon, small intestine and liver from whole-genome sequencing of clonal organoids (*9*). Considering the absence of clear APOBEC mutations, the burden and signatures observed here are also compatible with those seen in ESCC tumors. This contrasts with a report of a mutation burden around 15-24 substitutions/Mb in normal esophagus and unusual mutation spectra, using a new PCR-based method for mutation detection down to single DNA molecules (*47*). We note that this mutation burden is ~20-40 times higher than that seen in the present study or in organoids from colon, small intestine and liver. In fact, this burden is higher than that seen in the vast majority of ESCC tumors, despite a lack of APOBEC mutations in (*47*), suggesting technical problems in the identification of somatic mutations in that study.

5.2. Clone sizes

With our experimental design, the fraction of cells in a sample that carry a mutation can be estimated using the variant allele frequency of the mutation and the local copy number at the mutant site. This enables us to estimate clone sizes as the product of the fraction of mutant cells in a sample and the area of the sample.

As explained in detail in (*7*), the relationship between the variant allele frequency of a mutation (v) and the fraction of mutant cells (ρ) can be expressed by the following equation:

$$\rho = \frac{n_{wt}v}{n_m + n_{wt}v\text{-}(n_m + n_{nm})v}$$

(Eq. 3)

Where n_{wt} is the average ploidy of cells not carrying the mutation in the sample, n_m is the number of DNA copies carrying the mutation in mutant cells, and n_{nm} is the number of copies not carrying the mutation in mutant cells. However, this expression requires precise knowledge of the local copy number, which is often unavailable in this study. Fortunately, our analyses show that cells in normal esophagus are largely diploid without frequent copy number changes except around the *NOTCH1* locus. In the simple case of heterozygous mutations in diploid cells, we have n_m=1, n_{nm}=1 and n_{wt}=2, and the equation simplifies to:

$$\rho = 2v$$

Thus, in diploid regions the fraction of cells carrying a heterozygous mutation can be estimated as twice the variant allele fraction of the mutation.

In the common case of homozygous loss of *NOTCH1* by copy-neutral LOH, we would have n_m=2, n_{nm}=0 and n_{wt}=2, and the equation simplifies to:

$$\rho = v$$

In this manuscript, owing to the unavoidable uncertainties, we largely avoid relying on estimates of clone sizes, with the following exceptions. For simplicity, the estimates used in Fig. S5B were calculated using $\rho = 2v$, although we note that these estimates are likely to be overestimates for some *NOTCH1* mutations. As described in detail in Methods S5.4, for the generation of the patchwork plots (Fig. 3) we used a conservative approach to *NOTCH1*, representing lower bound estimates of the sizes of *NOTCH1* clones by assuming bi-allelic loss. The estimate of the range of mutant clone sizes in our study from 0.01 mm^2 to 8 mm^2 is also conservative, with the largest clone being one with compound heterozygous loss of *NOTCH1*.

Thus, the clone sizes reported in this manuscript should be considered approximate estimates whose accuracy can be affected by occasional undetected copy number changes as well as differences in cell density within a sample.

5.3. Lower and upper bound estimates of the fraction of mutant epithelium

Our data allow us to derive estimates of the fraction of cells in a sample that carry non-synonymous mutations in a given gene. When only one mutation is observed in a gene in a given sample, estimation of the fraction of mutant cells requires knowing how many of the copies of the gene are affected in mutant cells. The equation below corresponds to Eq. 3, assuming that n_{wt}=2.

$$\rho = \frac{2v}{n_m + 2v\text{-}(n_m + n_{nm})v}$$

(Eq. 4)

As shown above, in the absence of a copy number change at this locus (n_m=1, n_{nm}=1), the fraction of mutant cells is: $\rho = 2v_j$. In the presence of copy-neutral LOH (n_m=2, n_{nm}=0) we obtain: $\rho = v_j$. And in the presence of hemizygous loss of the wild-type allele (n_m=1, n_{nm}=0) we obtain $\rho = 2v_j/(1+v_j) < 2v_j$. Thus, accounting for the possible existence of undetected CN-LOH or hemizygous loss at the site, we can say that ρ falls within the range (v_j, $2v_j$).

When more than one mutation is observed in a gene in a given sample, estimation of the fraction of mutant cells (ρ) requires knowing whether these mutations affect different cells of a sample or whether some co-occur within the same cells (*e.g.* as in the case of bi-allelic loss of a gene by compound heterozygous mutations). For example, if two mutations are heterozygous at a diploid locus with VAFs v_1 and v_2, the fraction of mutant cells would be ~2(v_1+v_2) if they affect different cells or ~2(max(v_1,v_2)) if one mutation is subclonal to the other or both are co-clonal (*i.e.* present in the same fraction of cells).

Assuming the biologically reasonable scenario of a maximum of two non-synonymous mutations per cell per gene, and considering the different possibilities of compound heterozygosity, copy neutral LOH and hemizygous losses, the total fraction of mutant cells for a given gene in a sample ranges from the lower bound $\Sigma_j v_j$ (which corresponds to bi-allelic mutation of the gene in all mutant cells of the sample by either CN-LOH or compound heterozygous mutations) to the higher bound $2\Sigma_j v_j$ (which corresponds to one mutant allele per gene at a diploid locus).

5.4. Generation of the patchwork summary plots

The panels in Fig. 3 aim to provide a graphical summary representation of the mutant clones detected in a typical 1cm^2 of normal esophagus. The figure summarizes the number, density and estimated sizes of mutant clones (each shown as a circle) as well as the genes affected (shown as different colors). An analogous representation was used in our previous study on aged sun-exposed skin (*7*). Only non-synonymous mutations in the 14 genes under positive selection were represented in the figure. The number, density and estimated sizes of the clones are entirely based on the sequencing data, using conservative assumptions about bi-allelic hits when dealing with *NOTCH1* mutant clones (see below). The nesting of clones and subclones is informed by the sequencing data whenever possible or randomly simulated within a sample when the data is uninformative.

This point below explains how the figures were generated.

6. Selection of samples: For each donor, a number of samples were randomly selected to amount to at least 1cm^2 in total area (*i.e.* at least 50 biopsies from every patient were randomly selected). This was done making sure that all of the samples sharing a clone were included in the random selection of samples to avoid underestimating clone sizes in the figure by selecting only a fraction of the samples in which a clone is present.

7. Clustering of mutations into clones: Typically, multiple mutations were detected per sample in the targeted sequencing data (Fig. 1). These mutations needed to be clustered into clones and subclones. Whenever possible, and exclusively for representation purposes, the clustering of mutations into likely clones and subclones was informed by the sequencing data using two approaches. First, when multiple mutations were present in a single sample, if possible we used the pigeonhole principle to nest mutations (see section S5.5). Second, when two or more mutations were shared between two or more samples, they are more likely to fall within the same clone. In such case, we grouped mutations within a clone as co-clonal if their VAFs were consistent with co-clonality across samples (VAFs not significantly different across the samples sharing these mutations, tested using a Poisson test per sample), or we nested them if the VAFs did not violate a nesting structure (*i.e.* if the VAFs of a mutation across samples were always smaller or equal to the VAFs of another mutation). When neither mutation sharing across samples nor the pigeonhole principle could be applied to cluster or nest mutations within a sample, mutations were nested randomly (see below).

8. Random nesting of clones: When the sequencing data could not inform about the way clones are nested, mutations were nested randomly, assuming no interaction between mutations. For example, if two mutations are observed in a sample, one in 20% of cells and one in 5% of cells, we randomly nested the smaller clone into the larger clone with probability equal to the fraction of the area of the sample covered by the larger clone (*i.e.* 20%). This simple approach effectively assumes no tendency of mutations to

preferentially co-occur or be mutually exclusive. Two constraints were imposed to the random nesting: (1) a maximum of two hits per gene in a clone, (2) there must be space left within a clone to draw a circular subclone without overlapping with previously drawn subclones. If the latter condition is violated, the subclone was placed within a different clone in the patchwork plot. In summary, nesting of clones was informed by the sequencing data whenever possible or assumed to be random when this information was not available.

9. Clone sizes: Given that copy number changes were found to be rare in genes other than *NOTCH1*, for the purpose of representing clone sizes in the patchwork plots we assumed that mutations in genes other than *NOTCH1* were heterozygous. Mutations in *NOTCH1* were conservatively assumed to be homozygous, effectively representing a scenario close to the lower bounds shown in Fig. 2C,G. When a single mutation was observed in *NOTCH1* in a cluster, the mutation was assumed to be in copy neutral LOH, which is consistent with our analyses showing that virtually all of the samples with an isolated *NOTCH1* mutation also show LOH in *NOTCH1* and that LOH in *NOTCH1* is most often copy neutral. This was a conservative choice to avoid inflating clone sizes for *NOTCH1* due to undetected copy number changes.

10. Representation in space: Clones were drawn randomly in the available space of the patchwork plot.

In summary, the patchwork plots are a summary representation of the sequencing data across donors, with each circle representing an observed mutation and clone sizes and clone densities estimated from the sequencing data. Clone sizes for *NOTCH1* are conservative lower-bound estimates. Nesting of clones was informed by the sequencing data whenever possible but randomly simulated in most cases, particularly when involving small subclones. The placement of clones in space was randomly simulated. Thus, these plots can be a useful summary representation of the variation in the frequency, size and density of mutant clones across donors and genes, but the nesting and placement of clones in space is typically simulated and should not be relied on.

5.5. Statistical pigeonhole principle

As described above, allele frequencies and copy number data can be used to estimate the fraction of cells carrying a mutation in a sample. The pigeonhole principle is widely used to infer whether two mutations observed in a sample must be co-occurring within the same cells (*45*). Let ρ_A and ρ_B be the fraction of cells carrying mutations A and B in a given sample. The pigeonhole principle states that if $\rho_A+\rho_B>1$, then both mutations cannot be in unrelated clones but must be nested or co-clonal. For example, if a mutation is present in 80% of all cells in a sample and another mutation is present in 30%, the latter must be a subclone of the former (*45*).

In practice, however, allele frequencies have associated sampling uncertainty. In this study, for pairs of mutation that appear to fulfill the pigeonhole principle, we conservatively used a statistical test to determine whether $\rho_A+\rho_B$ is significantly higher than 1. To do so, allele frequencies for mutations A and B are assumed to be drawn from binomial distributions and a one-sided Likelihood-Ratio Test is used to compared two hypotheses:

H_0: $\rho_A+\rho_B \leq 1$
H_1: unconstrained estimates for ρ_A and ρ_B

Since copy number changes were seen frequently in *NOTCH1* and occasionally in *TP53*, and large clones were commonly observed for both genes, in samples where a copy number change was detected in *NOTCH1* or *TP53*, mutant cell fraction estimates were halved to conservatively account for the likely scenario of copy neutral LOH. This conservative assumption can reduce the sensitivity to detect pairs of co-occurring mutations, but should lead to a higher specificity and results robust to LOH in *NOTCH1* and *TP53*. Similarly, genes in the X-chromosome of male donors were treated accordingly.

Significant pairs or groups of mutations identified using the test above (*P*-value<0.05) are shown in Figs. 4A (a subset) and Fig. S6 (full list).

5.6. Subclonal reconstruction across samples

Phylogenetic reconstruction from bulk polyclonal sequencing data requires the clustering of mutations into subclones (*45*). Different approaches exist to cluster mutations into clones, in which mutations are grouped into clusters that show consistent mutant cell fractions across related samples.

Given the simplicity of the copy number landscape in the whole-genomes that we sequenced (Fig. S9), to reconstruct the tree shown in Fig. 4B, we clustered allele frequencies directly using a mixture model across the six samples shown in the figure. Allele frequencies for each mutation in each sample were assumed to be drawn from a mixture of beta-binomial distributions, with each component corresponding to a cluster of mutations. Beta-binomial distributions were used instead of binomial distributions to account for the likely possibility of a certain level of overdispersion in the allele frequencies due to mapping biases or overlapping reads. An overdispersion parameter ρ=0.01 was used in Fig. 4B. An Expectation-Maximization (EM) algorithm was used to optimize the parameters of the mixture model (the mutant cell fraction and the weight of each cluster in each sample). The EM algorithm was run 2000 times with different random starting values of the parameters and the best overall solution was selected. The Bayesian Information Criterion (BIC) was used to select the optimal number of clusters, with a total of $n(2k\text{-}1)$ parameters (where n corresponds to the number of samples and k the number of clusters).

6. Selection analyses (dN/dS)

To study the extent of selection, we used the *dNdScv* algorithm, a maximum-likelihood implementation of dN/dS particularly adapted to somatic mutation data.

For a detailed description of this method please refer to (*4*). Briefly, dN/dS measures the ratio of the rate of non-synonymous mutations per non-synonymous site and the rate of synonymous mutations per synonymous site. dN/dS ratios can be calculated at the level of individual genes or groups of genes. When mutational biases are properly considered, dN/dS values significantly higher than 1 reveal an excess of non-synonymous mutations in a gene, reflecting the action of positive selection. Conversely, dN/dS values significantly lower than 1 reflect the action of negative selection (*4*).

In the *dNdScv* implementation, dN/dS ratios are calculated by controlling gene sequence composition and variable substitution rates using a 192-rate trinucleotide substitution model, which largely controls for context-dependent mutational processes and transcription strand biases (*4*). Mutation rates have been shown to vary considerably across genes in the genome, often depending on the expression level of a gene or genomic features such as the local chromatin state or the replication timing of the genomic region where the gene is located. In *dNdScv*, the background mutation rate of a gene is estimated by combining local information (the observed number of synonymous mutations observed in the gene) and information on the mutation density across all genes (using a negative binomial regression, which effectively fits a Gamma distribution for the variation of the mutation rate across genes) (*4*).

6.1. *dNdScv*: screen for positively selected genes

In order to identify genes under positive selection in normal esophageal epithelium we combined all of the observed mutations from the nine donors, after collapsing mutations shared across nearby biopsies (Methods S3.3).

In this study, we used the default settings of *dNdScv* (version 0.0.0.9), with the exception of restricting the fitting of the background model for indels to genes not found as significant using substitutions alone (*dNdSloc* model). *dNdSloc* is a simpler model that only uses substitution data. This was the model originally used in our earlier study on sun-exposed skin (*7*). This model is typically less sensitive than the *dNdScv* model as it does not use information about indels or the Gamma constraint. Nevertheless, *dNdSloc* detected 12 genes under positive selection in this dataset (q-value<0.01): *TP53*, *NOTCH1*, *NOTCH2*, *NOTCH3*, *FAT1*, *ARID1A*, *CUL3*, *AJUBA*, *CCND1*, *PIK3CA*, *TP63*, *KMT2D.*

One particular synonymous mutation occurring at an exon boundary in *TP53* has been shown to be a recurrent driver mutation by affecting *TP53* splicing (T125T) (*4, 48*). Consistent with this observation, 5 of the 9 synonymous mutations observed in *TP53* in this dataset were T125T mutations. These changes can be considered genuine driver events and were excluded from the dN/dS analyses to avoid artifactually inflating the mutation rate estimation in *TP53*.

Of interest, we noted that *CCND1*, which is amplified in ~57% of ESCCs (*21*), has an unusual pattern of recurrent nonsense mutations. These nonsense mutations mostly cluster on a small track of glutamic acids at the end of the Cyclin_C domain (E272-E280) and are predicted to lead to the truncation of the C-terminal of *CCND1*. The product of this truncated gene is predicted to resemble a shorter and more oncogenic protein isoform of *CCND1* often expressed in cancers by alternative splicing (cyclin D1b) (*49*).

In addition to the 14 genes found as significant by *dNdScv*, we detected canonical hotspot mutations in *HRAS* (G13C), *NRAS* (G12D) and *EGFR* (G598A) (Supplementary Dataset S1).

6.2. Interpretation of dN/dS ratios in terms of clone sizes

In our experimental design, only mutations that reach certain clone sizes are detectable by *ShearwaterML*. This gives dN/dS ratios a simple interpretation in terms of clonal expansions. For example, a dN/dS ratio of 10 for a given gene in our experiment suggests that a cell that acquires a non-synonymous mutation in that gene has a 10-fold higher probability of reaching detectable clone sizes than a cell that acquires a synonymous mutation in the same gene.

6.3. Estimation of the number of driver mutations

dN/dS ratios can be used to estimate the excess of non-synonymous mutations favored by positive selection in a given collection of mutations (*4*). In this study, the slightly conservative figure of 3,915 (CI95%: 3,829-3,988) positively-selected mutations was estimated by calculating the excess of non-synonymous substitutions and indels in the 14 genes found under significant positive selection by *dNdScv*. Confidence intervals for the excess of non-synonymous substitutions were calculated from the global Poisson regression used by *dNdScv*. For consistency with the background model used in *dNdScv*, the predicted excess for indels was calculated by estimating the number of indels per coding bp in the 62 genes not detected as significant by *dNdSloc* (see above), and confidence intervals were calculated using the confidence interval for the ratio of two Poisson observations.

6.4. Variation in selection between donors across genes

To evaluate whether certain genes are more strongly positively selected in one individual than in other individuals, we compared dN/dS ratios between donors (Fig. S5C-E) using Likelihood-Ratio Tests. A simpler and more approximate approach would be to test for differences in the fraction of non-synonymous mutations seen in a given gene in a donor compared to all other donors, which could be tested using a Fisher’s Exact Test. An advantage of using dN/dS ratios instead of this simpler approach is that differences in mutational signatures across donors are corrected.

Let $\omega_{g,1}$ and $\omega_{g,2}$ be the maximum-likelihood estimates (MLEs) for the dN/dS ratios for gene *g* in datasets 1 and 2, respectively. Dataset 1 corresponds to all mutations observed in a given individual, and dataset 2 to the mutations present in all other individuals. Here we used MLEs under the *dNdScv* model, which, as we explained above, combines local and global information to estimate the background mutation rate in a gene (*4*). We can test for higher dN/dS ratios in a gene in a given individual using a one-sided test with the following null and alternative hypotheses:

H_0: $\omega_{g,1} \leq \omega_{g,2}$

H_1: unconstrained $\omega_{g,1}$ and $\omega_{g,2}$

In normal esophagus, we noticed that the signal of positive selection across genes tends to vary across donors with a general increase with age. When the overall signal of positive selection is higher across all genes in a given donor, the test above can detect significantly higher dN/dS ratios for a given gene that are reflective of this global difference rather than gene-specific differences. Thus, we used a more conservative test by removing the effect of global differences in dN/dS ratios across all genes. Let $\boldsymbol{\omega}_1$ be the MLE of the global dN/dS ratio from all genes other than the gene being tested. We can then estimate a gene-specific relative enrichment of non-synonymous mutations over $\boldsymbol{\omega}_1$ as a multiplicative factor ($\omega'_{g,1}$): $\omega_{g,1}=\boldsymbol{\omega}_1\omega'_{g,1}$. Based on this, a more conservative Likelihood-Ratio Test can be used to test for a gene-specific enrichment of non-synonymous mutations in a donor compared to all other donors, while removing the effect of differences in mutational signatures and global differences in the signal of selection:

H_0: $\omega'_{g,1} \leq \omega'_{g,2}$
H_1: unconstrained $\omega'_{g,1}$ and $\omega'_{g,2}$

This test was used for the analyses in Fig. S5C-E. This detected extensive differences in selection intensity across genes between donors. The strongest results in terms of significance and effect size were an enrichment of *TP53* mutations in the oldest donor, PD31182 (shown in Fig. S5E), and enrichments of *NOTCH3* mutations in PD30988 and PD36712 (shown in Fig. S5D).

As mentioned above, a one-sided Fisher's Exact Test could also be used to determine whether a given gene contains a higher fraction of all of the non-synonymous mutations in a donor than expected from all other donors. Although this test does not correct for differences in mutational signatures across donors, these differences are modest in this dataset and the Fisher's Exact Test approach yielded largely comparable results to the dN/dS-based Likelihood-Ratio Test above, with differences only around the limit of significance.

6.5. Gene mutation frequency in ESCCs and EACs

Fig. 2D,G and S5A show the frequency of non-synonymous mutations (substitutions and small indels) in different cancer genes in ESCC and EAC tumors. These numbers were obtained from TCGA, using somatic mutation calls for esophageal carcinomas from a previous study (*4*), and classifying them into ESCCs (n=88) and EACs (n=58) according to (*21*).

Although the power to detect recurrently mutated genes in EAC and ESCC from TCGA tumors is modest given the limited numbers of samples, we also ran *dNdScv* on both datasets restricted to the 74 genes sequenced in our study. This found the following genes as significantly mutated with q-value (FDR, restricted hypothesis testing) < 0.05:

- ESCC: *TP53* (q=0), *NFE2L2* (q=0), *KMT2D* (q=4e-08), *PIK3CA* (q=2.3e-07), *NOTCH1* (q=3.6e-07), *PTCH1* (q=1.1e-06), *PTEN* (q=2.7e-05), *SMAD4* (q=0.0016), *NOTCH3* (q=0.01).
- EAC: *TP53* (q=0), *ARID1A* (q=0.00056), *FBXW7* (q=0.00089), *SMAD4* (q=0.0059).

Six of the nine genes detected as significantly mutated in ESCCs were found to be under positive clonal selection in normal esophagus (Fig. 2). In addition to the six genes found under positive selection in TCGA ESCCs using *dNdScv*, other genomic studies of ESCC tumors have reported evidence of frequent mutations in several of the other genes that we find under positive selection in normal esophagus: *CUL3* (*21*) (often lost by homozygous deletions), *CCND1* (*21*) (very frequently affected by amplifications in ESCCs), *TP63* (*21*) (typically amplifications), *FAT1* (*22*) and *AJUBA* (*20, 22*). Overall, at least 11 of the 14 genes that we find under positive selection in normal esophagus appear to be recurrently mutated in ESCCs.

Interestingly, *ARID1A* has not been reported as frequently mutated in ESCC, but it is frequently mutated in esophageal adenocarcinomas (*21*).

7. Mixed effect regression models

Using the mutations from all nine donors, we studied whether donor age, the main risk factor in ESCC, correlate with clone sizes and numbers of mutations per sample, when accounting for the potential confounding effects of smoking history and gender. To do this, we used the *lme4* package in R to fit mixed effects regression models with age, gender and smoking as fixed effects and a random effect for the intercept across donors. Unlike simpler regression models, the use of this random intercept model controls for the non-independence of multiple observations per donor.

In R code, the regression model for clone sizes was:

```
lmer(logsize ~ age + gender + packyears + coverage + (1|donor/sample), REML=F)
```

And the regression model for testing differences in the total number of mutations per sample (*n*) was:

```
glmer(n ~ age + gender + packyears + coverage + (1|donor/sample), family=”poisson”)
```

logsize is the logarithm of the estimated clone size (assuming heterozygous mutations for simplicity given that this is a comparison across donors), and *packyears* is a continuous variable for the cumulative smoking history in units of pack years (Table S1). Since multiple clones are detected in the same samples and donors, we used a random intercept model for donors and samples. Using only a random intercept per donor yielded analogous results. Since the number of mutations and clone sizes in a sample could be affected by the sequencing coverage of the sample, we also included coverage as a fixed effect in both models. Analogous results were obtained with or without using coverage as a fixed effect.

Given the small number of donors studied (n=9), statistical testing of the effect of age was performed using non-parametric bootstrapping, generating 1,000 bootstrapped datasets by randomly sampling with replacement the donors as well as the samples within each selected donor. The resulting p-values are provided in the main text. We note that these p-values were considerably more conservative than those using Likelihood-Ratio Tests or the asymptotic Wald test, but should be more robust in a cohort of this size.

Alcohol intake was also tested as a potential confounding factor in these models. However, there were doubts about the reliability of the alcohol intake clinical records as these were collected from family members under stressful conditions near the time of death of the donor. Only two donors were recorded as consuming alcohol on a daily or weekly basis and they were among the three youngest donors of the cohort (Table S1), making this variable largely uninformative and potentially confounding. Larger cohorts will be needed to study the association between behavioral risk factors, such as smoking and alcohol consumption, and the mutational landscape in normal tissues.

Supplementary Text

40. D. C. Lin et al., Genomic and molecular characterization of esophageal squamous cell carcinoma. Nature genetics 46, 467-473 (2014).

41. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25, 1754-1760 (2009).

42. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289-300 (1995).

43. K. M. Raine et al., cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. Current protocols in bioinformatics 52, 15.17.11-12 (2015).

44. S. Nik-Zainal et al., Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47-54 (2016).

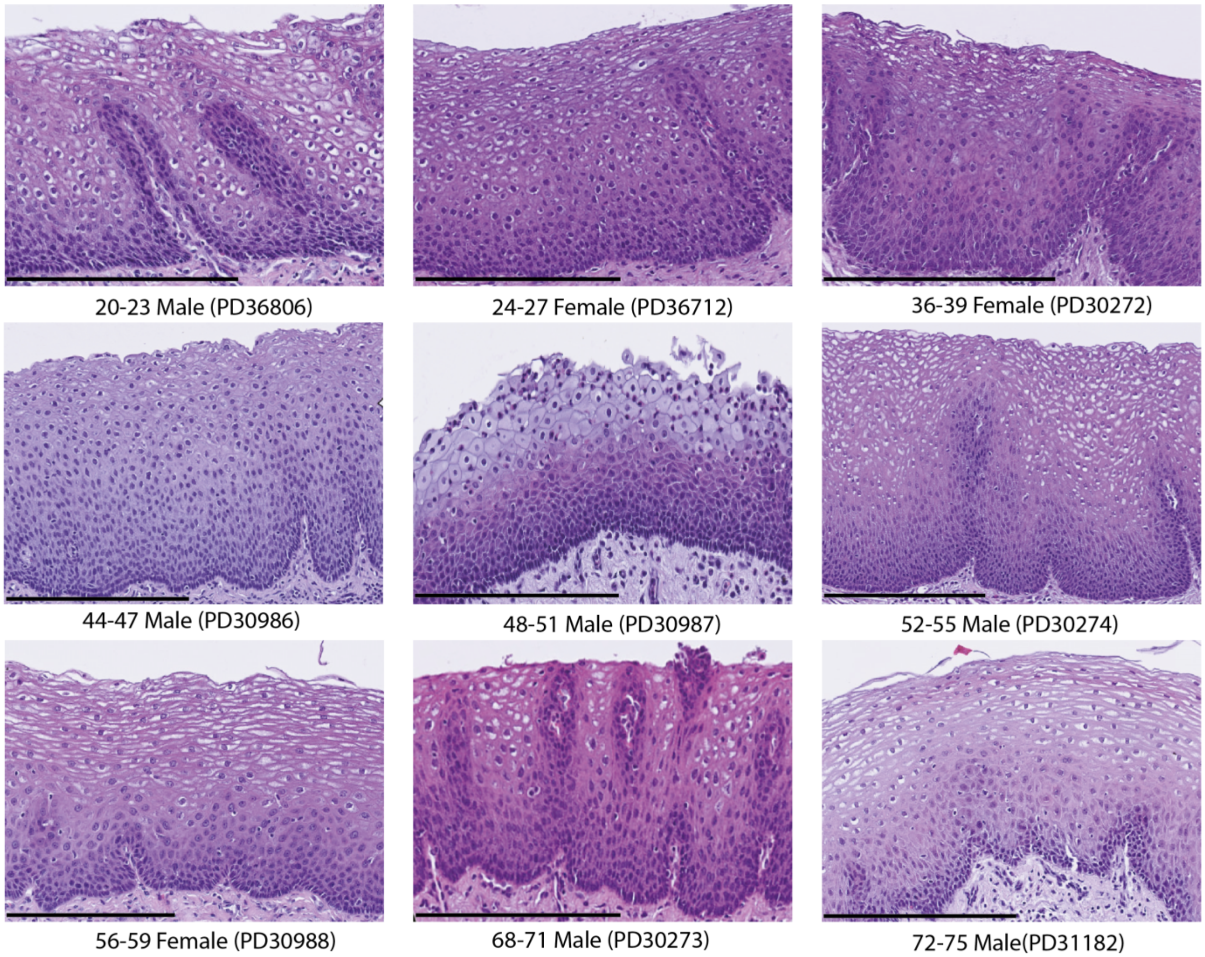
45. S. Nik-Zainal et al., The life history of 21 breast cancers. Cell 149, 994-1007 (2012).

46. K. M. Raine et al., ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. Current protocols in bioinformatics 56, 15.19.11-15.19.17 (2016).

47. S. Yamashita et al., Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. Proceedings of the National Academy of Sciences of the United States of America 115, 1328-1333 (2018).

48. F. Supek, B. Minana, J. Valcarcel, T. Gabaldon, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers. Cell 156, 1324-1335 (2014).

49. K. E. Knudsen, J. A. Diehl, C. A. Haiman, E. S. Knudsen, Cyclin D1: polymorphism, aberrant splicing and cancer risk. Oncogene 25, 1620-1628 (2006).



20-23 Male (PD36806)

24-27 Female (PD36712)

36-39 Female (PD30272)

44-47 Male (PD30986)

48-51 Male (PD30987)

52-55 Male (PD30274)

56-59 Female (PD30988)

68-71 Male (PD30273)

72-75 Male(PD31182)

Fig. S1.

Representative histology images from the 9 donors. Histology performed as described in Methods S1. Scale bar: 250µm.

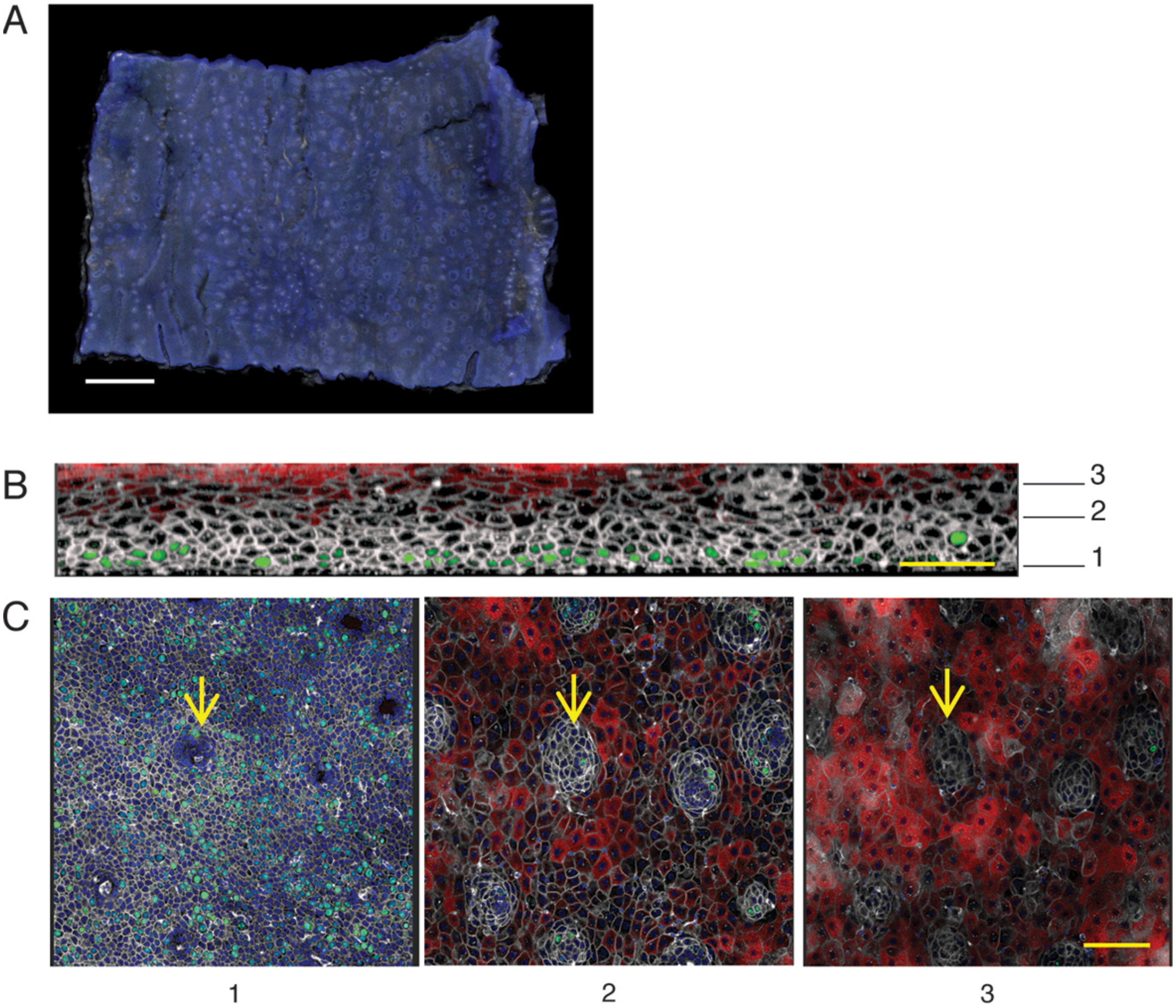


Fig. S2.

Supplementary Fig. S2. Confocal imaging of oesophageal epithelia. (A) Typical low magnification projected 3D confocal z-stack image of peeled oesophageal epithelia. There are no visible areas of dysplasia or other abnormalities. Scale bar 1 mm, blue is DAPI staining DNA, grey is wheat germ agglutinin staining cell membranes. (B) Lateral view of rendered confocal z-stack from a 44-47 year-old male smoker (patient PD30986), KRT4 is red, DAPI blue and WGA grey. KI67 (green) is confined to the basal layer unless associated with the papillae. Scale bar 50 µm. Numbers indicate the level of optical sections shown in C. (C) Top down views of optical sections of wholemount indicated at levels 1-3 in panel B. KRT4 is red, KI67 green, DAPI blue and WGA grey. Scale bar 100 µm. KI67 and KRT4 staining was normal and no lesions were seen in the imaged wholemounts.

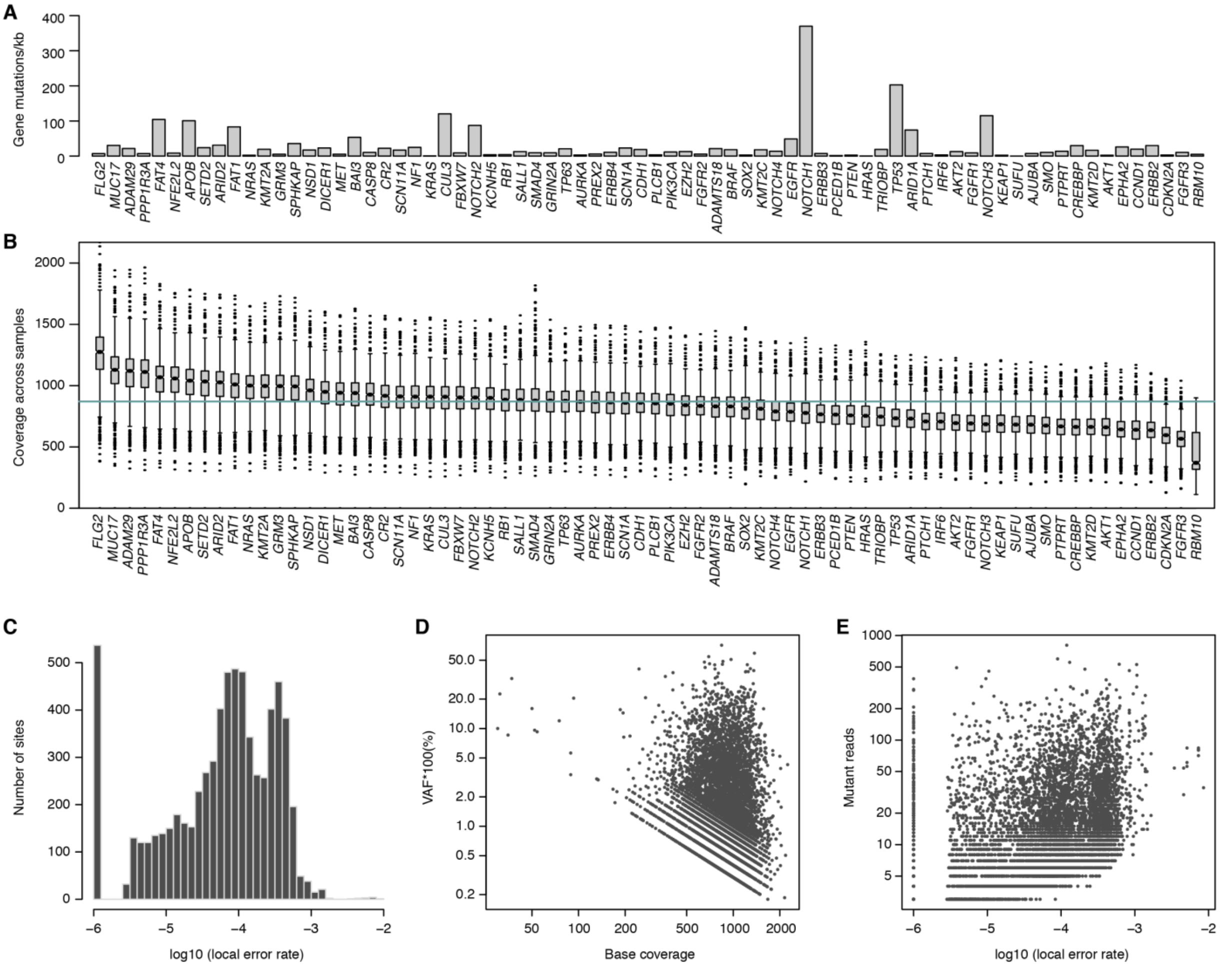


Fig. S3.

Number of reads supporting each mutation. (**A**) Barplot depicting the number of mutations per coding megabase of sequence for every gene across all samples. (**B**) Boxplot showing the mean coverage per gene and sample, with genes in the x-axis sorted by median coverage across samples. (**C**) Histogram of the MLEs of mutation rates per site from *ShearwaterML*. (**D**) Scatter plot showing the mutant site coverage and VAF of each of the mutations detected. As expected, mutations with lower VAFs can be detected with increasing coverage. (**E**) Scatter plot showing the site-specific error rate (MLE) estimated by Shearwater at a mutant site and the number of mutant reads supporting a mutation, for each of the mutations detected. The plot shows how sites with higher error rates require a higher minimum number of supporting mutant reads to detect a mutation.

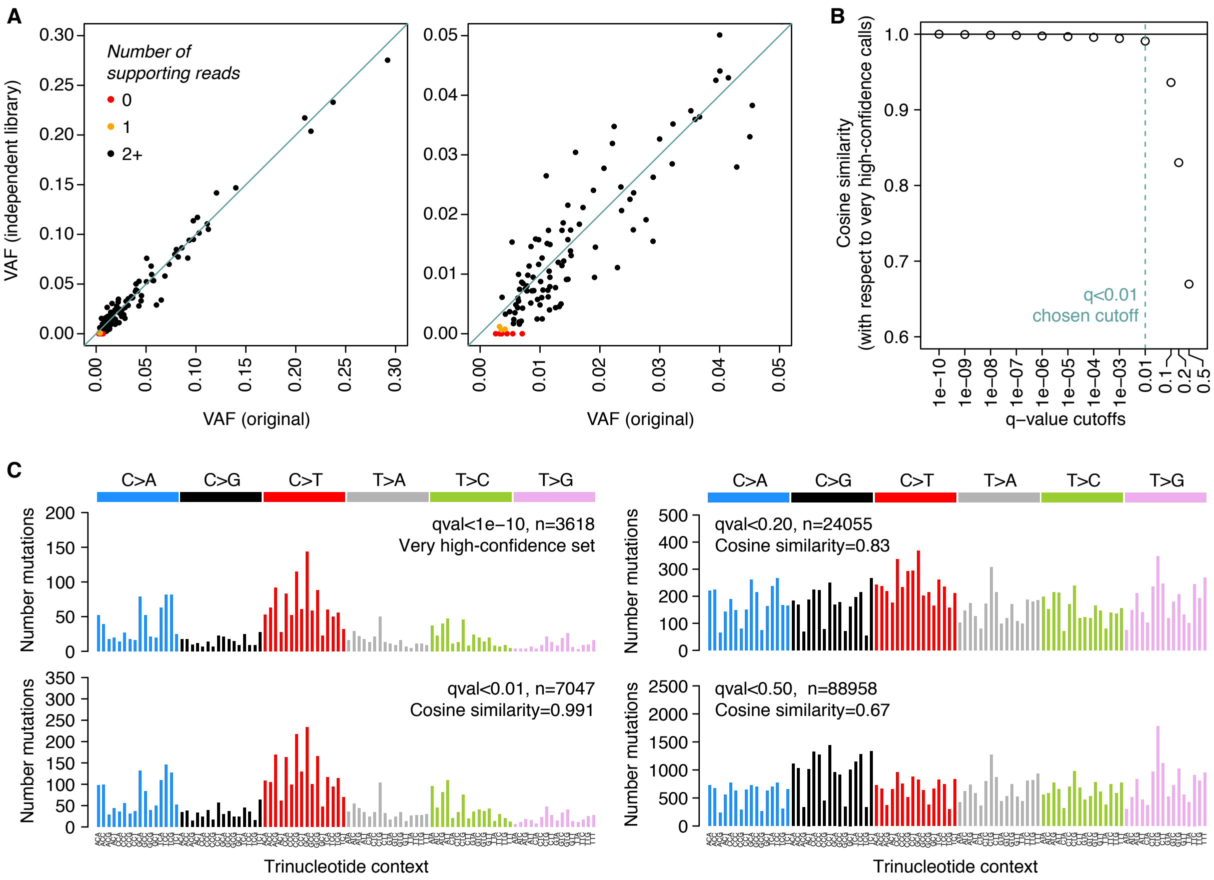


Fig. S4.

Quality and reproducibility of the mutation calls. (A) Validation of mutation calls. Scatter plots comparing the VAFs of 135 substitutions from 16 samples in the original samples (x-axis) and in independent validation libraries (y-axis). The panel on the left represents the full range of VAFs in these 135 mutations, while the panel on the right focuses on low VAF mutations. Overall, 93.3% (126/135) of the mutations were supported by two or more independent mutant reads in the validation experiment (see Methods S3.2). (B-C) Quality of the mutation calls based on the mutation spectra. Panel B shows the cosine similarity between the 96-trinucleotide spectrum of a reference set of very high-confidence mutations (ShearwaterML q-value<1e-10, see panel C) and the spectra of mutation sets generated with increasingly higher q-value thresholds for variant calling. Examples of the spectra at different q-value thresholds are shown in C. This confirms that the originally-chosen q-value threshold of 0.01 yields seemingly clean mutation calls with a spectrum closely matching that of the reference high-confidence set, and that relaxing this threshold would introduce large numbers of likely false positive calls with a very different spectrum.

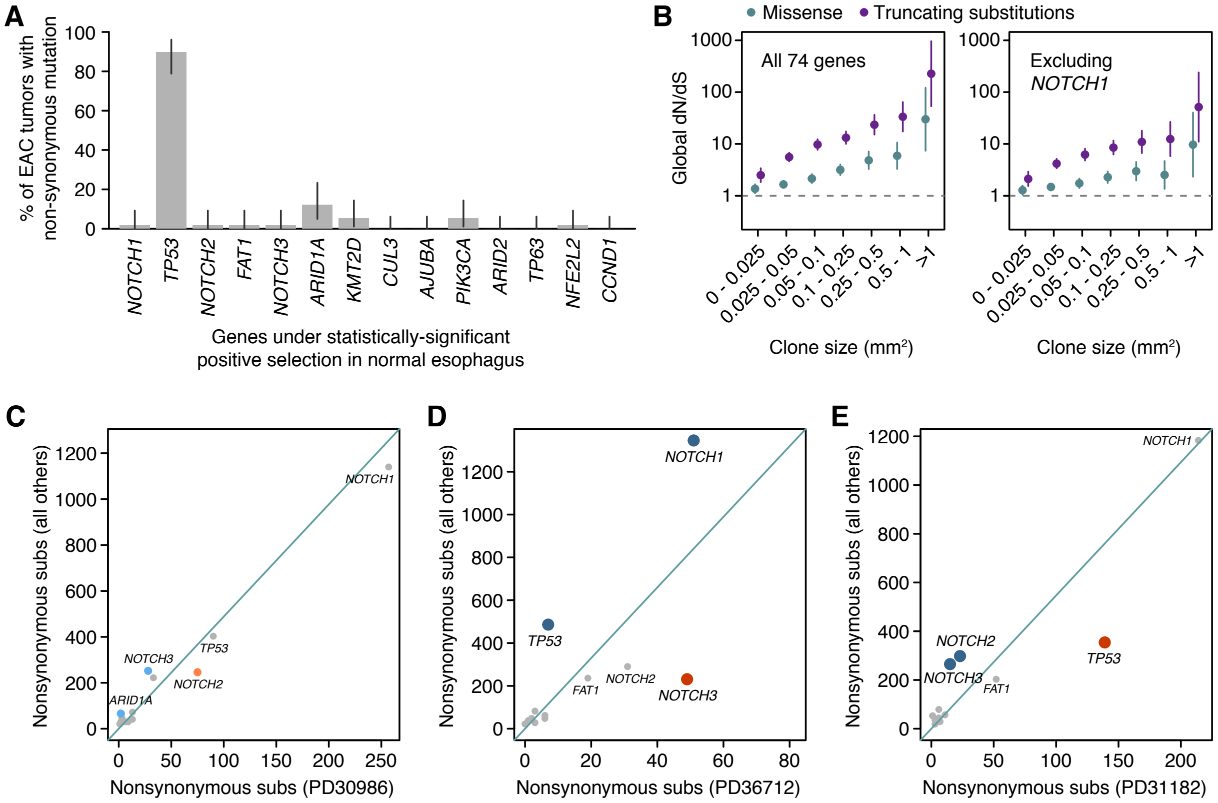


Fig. S5.

Supporting analyses on selection across genes, clone sizes and patients. (**A**) Percentage of EAC tumors with a non-synonymous substitution or an indel in each gene (supplementary panel to Fig. 2D). Error bars depict 95% Poisson confidence intervals. (**B**) dN/dS ratios from all 74 genes together as a function of estimated clone size, including and excluding *NOTCH1*. This figure shows that the enrichment of non-synonymous mutations increases drastically with larger clone sizes. (**C-E**) Differential selection across individuals. Colored genes reflect genes more strongly (orange) or weakly (blue) selected in a given donor compared to the average across all other donors, using a likelihood ratio test comparing two dN/dS ratios (Methods S6.4). The larger and smaller colored dots depict genes significant with q-value<0.001 or q-value<0.01, respectively.

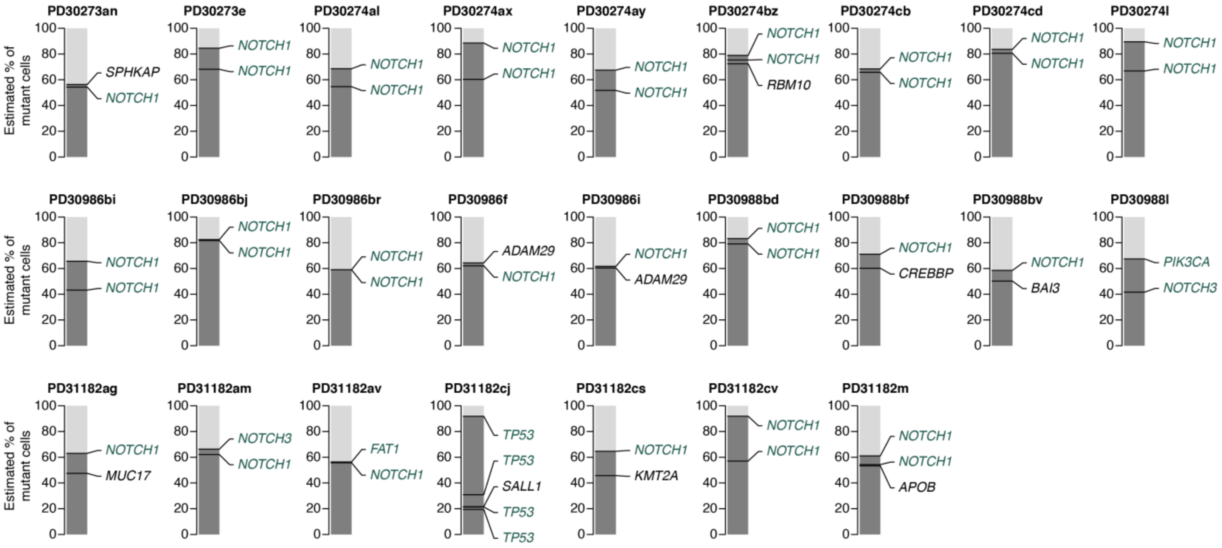


Fig. S6.

Significant results from the statistical pigeonhole principle. Representation of mutations co-occurring in the same clones using the pigeonhole principle. For a description of this approach see section S5.5.

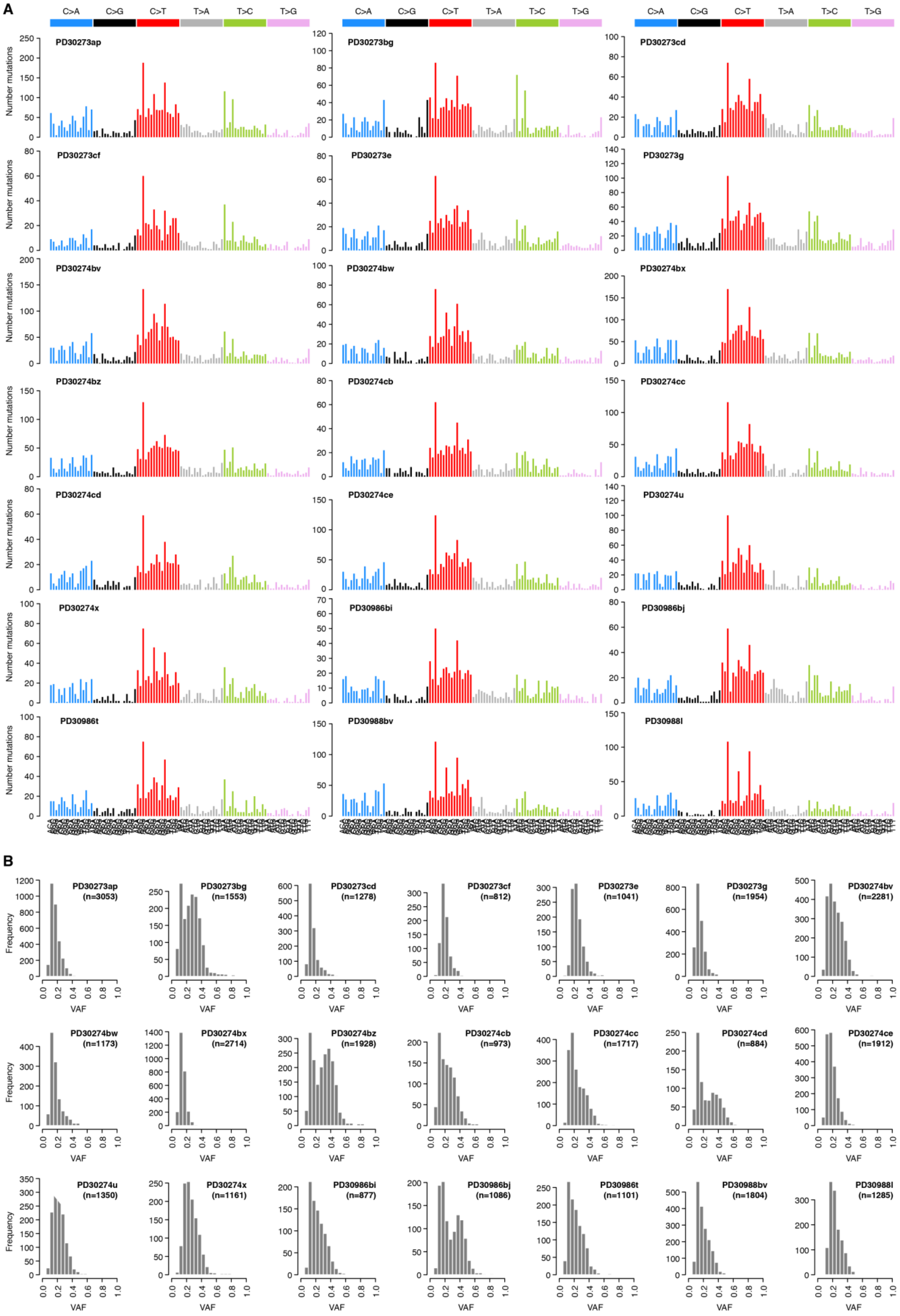

A
C>A
C>G
C>T
T>A
T>C
T>G
Number mutations
PD30273ap
PD30273bg
PD30273cd
PD30273cf
PD30273e
PD30273g
PD30274bv
PD30274bw
PD30274bx
PD30274bz
PD30274cb
PD30274cc
PD30274cd
PD30274ce
PD30274u
PD30274x
PD30986bi
PD30986bj
PD30986t
PD30988bv
PD30988l
B
Frequency
VAF
PD30273ap (n=3053)
PD30273bg (n=1553)
PD30273cd (n=1278)
PD30273cf (n=812)
PD30273e (n=1041)
PD30273g (n=1954)
PD30274bv (n=2281)
PD30274bw (n=1173)
PD30274bx (n=2714)
PD30274bz (n=1928)
PD30274cb (n=973)
PD30274cc (n=1717)
PD30274cd (n=884)
PD30274ce (n=1912)
PD30274u (n=1350)
PD30274x (n=1161)
PD30986bi (n=877)
PD30986bj (n=1086)
PD30986t (n=1101)
PD30988bv (n=1804)
PD30988l (n=1285)
0.0
0.2
0.4
0.6
0.8
1.0

Fig. S7.

Trinucleotide spectra and VAF from the 21 whole genomes. (**A**) 96-barplot depicting the number of mutations observed at each trinucleotide in each whole genome. Mutations are shown using the pyrimidine base of each base pair as the reference. (**B**) Histograms of the VAFs for each whole genome sample with the total number of mutations detected in each genome shown in brackets.

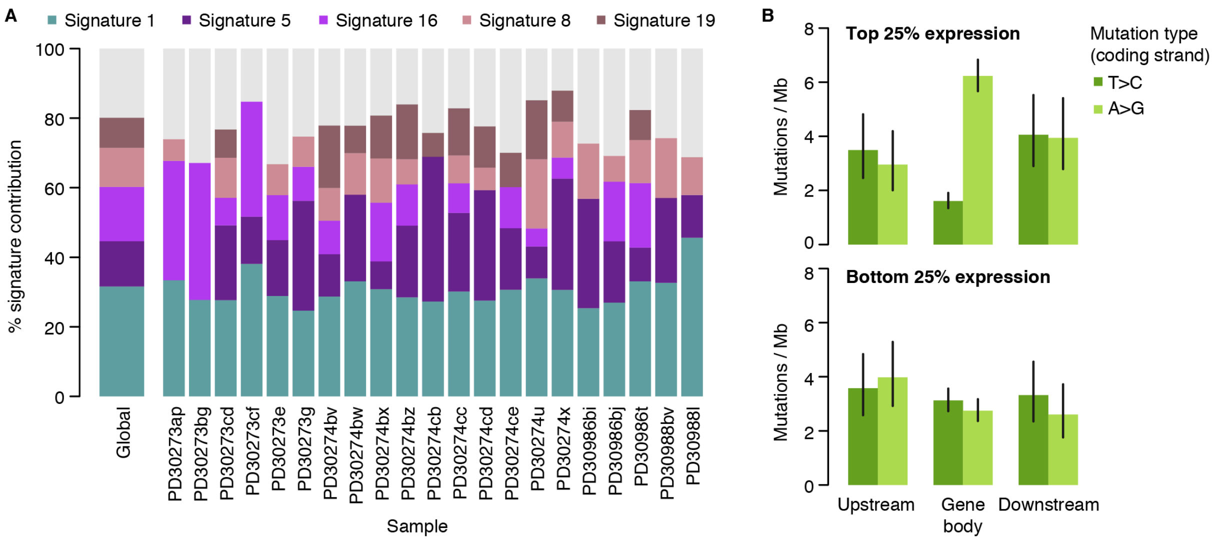


Fig. S8.

Mutational signatures and transcription-coupled damage. (A) Percentage of substitutions attributed to each mutational signature (COSMIC-30 set, Methods S4) per whole genome and for all mutations from the 21 whole genomes together ("Global"). (B) Barplot depicting the number of mutations per megabase observed upstream, in the gene body and downstream of highly (top panel) and lowly (bottom) expressed genes (Methods S4). The figure shows a significant increase in the rate of A>G mutations (mapped to the coding strand, T>C mutations in the transcribed strand) in the gene body of highly expressed genes with respect to flanking sequences and a decrease of T>C changes.

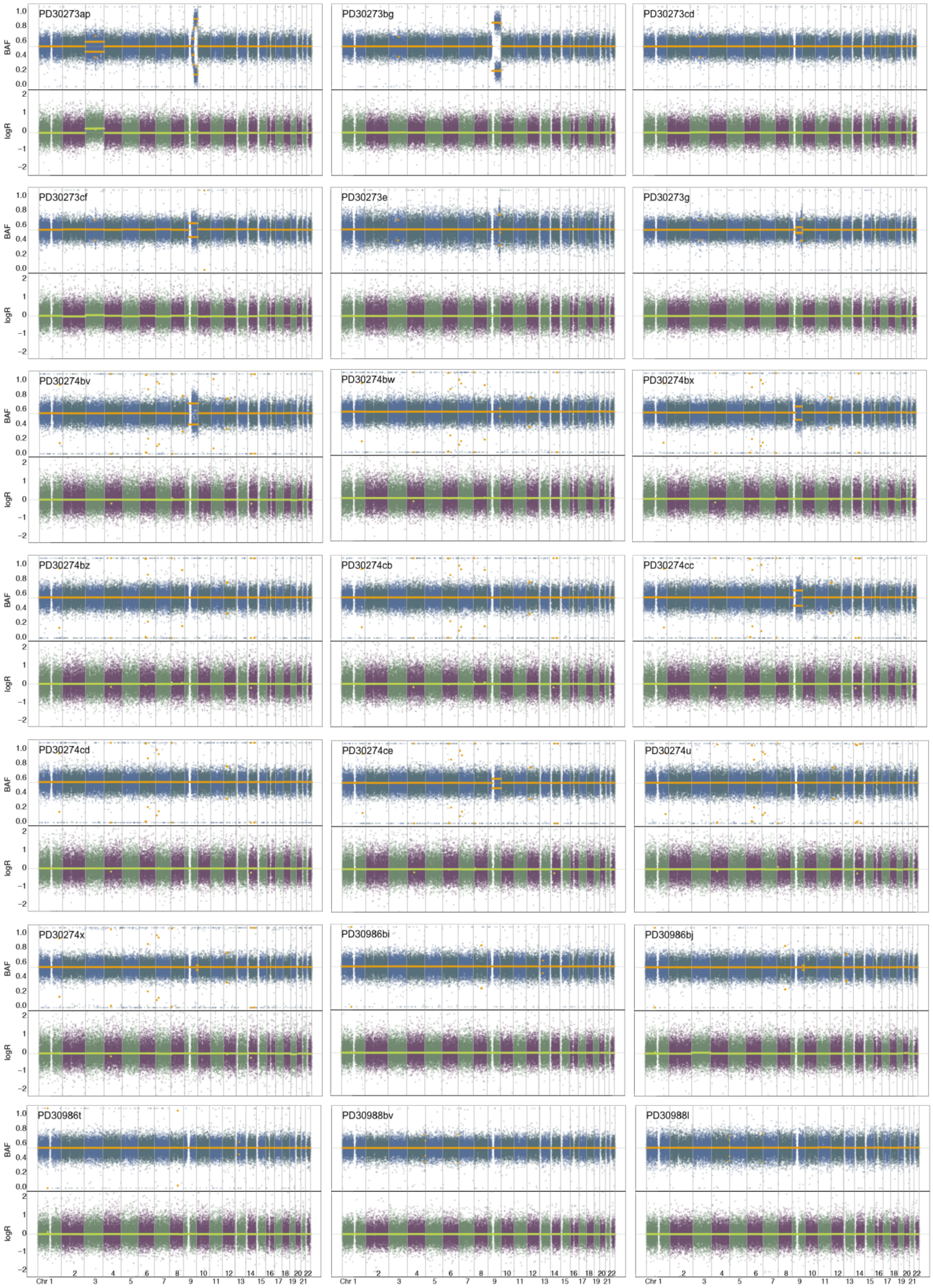

PD30273ap
PD30273bg
PD30273cd
PD30273cf
PD30273e
PD30273g
PD30274bv
PD30274bw
PD30274bx
PD30274bz
PD30274cb
PD30274cc
PD30274cd
PD30274ce
PD30274u
PD30274x
PD30986bi
PD30986bj
PD30986t
PD30988bv
PD30988l
BAF
logR
1.0
0.8
0.6
0.4
0.2
0.0
2
1
0
-1
-2
Chr 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Fig. S9.

BAF and logR plots for the 21 whole genomes. (**A**) Each of the 21 panels shows a scatter plot of the BAF (top) and logR (bottom) values of the heterozygous SNPs used by the Battenberg algorithm (section S3.5.2). Orange and green lines represent the mean values for the BAF and logR, respectively, of the segments identified by Battenberg. 12 of the 21 whole-genomes were found to have copy neutral LOH at the *NOTCH1* locus by Battenberg (Table S3).

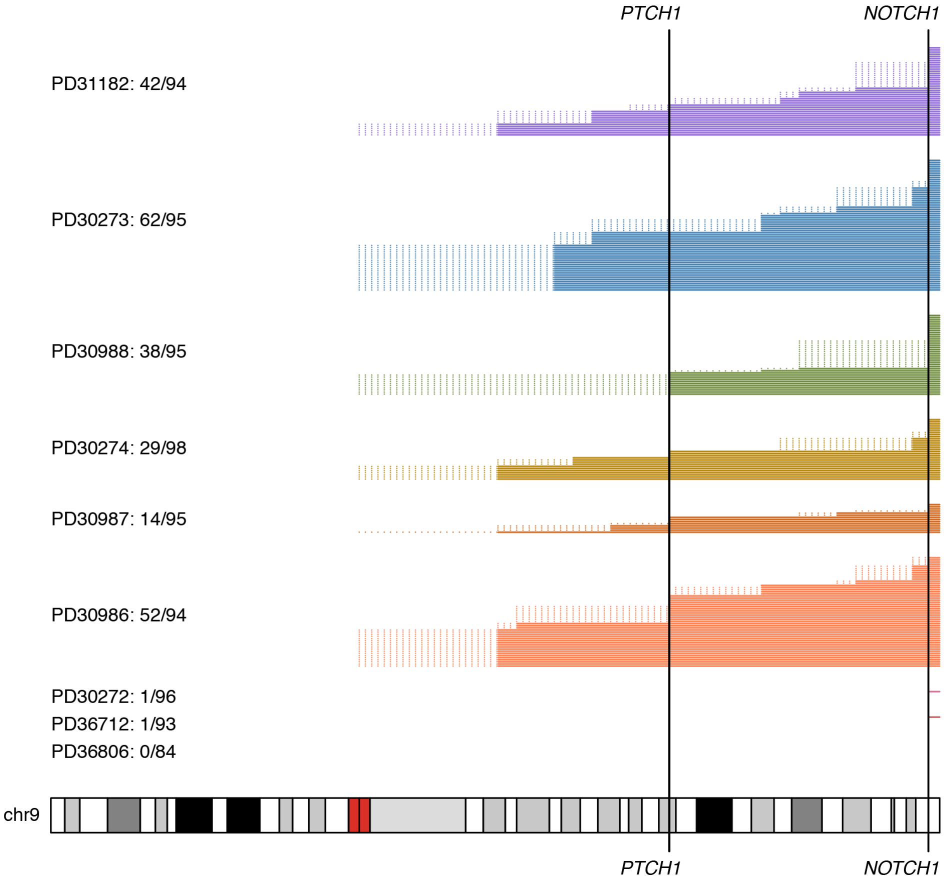


Fig. S10.

Loss of heterozygosity breakpoints from targeted data. (**A**) Representation of the inferred starting sites of LOH events in chromosome 9q, from the analysis of heterozygous SNPs in the targeted sequencing samples from the nine patients. This analysis is described in section S3.5.2 and shows that LOH events (likely caused by mitotic homologous recombination) start at multiple places throughout the q arm of chromosome 9. The spatial resolution of this analysis is restricted to the segment in between two consecutive SNPs in the chromosome and so it is limited by the sparsity of intergenic heterozygous SNPs targeted in our bait capture design. Dotted lines reveal the segments where the putative mitotic recombination event likely took place. The ideogram below shows a representation of chromosome 9 indicating the position of the centromere (red) and the location of the *PTCH1* and *NOTCH1* genes (vertical lines).

Table S1. List of donors and metadata

Table_S1_donor_information.xlsx

Table S2. Table of coding mutations found in the 844 samples

Table_S2_coding_mutations_across_844_samples.xlsx

Table S3. ***dNdScv*** **selection results**

Table_S3_dNdScv_results.xlsx

Table S4. Copy number calls from the whole-genome sequencing data

Table_S4_WGS_copy_number_results.xlsx