nature portfolio

https://doi.org/10.1038/s41588-021-00920-0

Supplementary information

Genomic and evolutionary classification of lung cancer in never smokers

In the format provided by the authors and unedited

Genomic and evolutionary classification of lung cancer in never smokers

Supplementary Notes

Estimation of Tumor Purity and Sample Exclusion

We systematically compared the tumor purity estimation by our method (based on the Battenberg algorithm for the SCNA profiles and DPClust for the VAF/CCF distribution of all mutations) with other algorithms, including ascatNgs (SCNA-based method) and CCUBE (bayesian mixture models for estimation based on both SCNAs and mutations). As shown in **Supplementary Fig. 28**, the purity estimation was very similar across the different algorithms: pearson correlation coefficient 1 and 0.95, for ascatNgs and CCUBE, respectively. We note that the CCUBE analysis was conducted on 208 of the 232 samples because of insufficient SNVs in 24 samples.

For our analyses, we selected tumor samples with high tumor cell percentage (>60%) based on histologic assessment. As reference, we mostly used blood samples. In the cases (n=48) for which we relied on normal lung tissue, these samples had no histological evidence of tumor nuclei. We did not exclude samples based on tumor purity because the percentage of genome alterations (PGA) in low purity samples was not lower than in high purity samples (**Supplementary Fig. 29)**. We excluded some samples based on low quality DNA sequencing, sample contamination or swapping, and dominant mutational signatures of known exposures (e.g. smoking, platinum treatment). Moreover, we excluded samples based on the following criteria: tumor samples with less than 100 total number of genomic alterations including SNV, INDEL, SV, and TE; no driver mutations in any known lung adenocarcinoma driver genes reported in the TCGA LUAD study; and no copy number alterations detected. We believe these criteria are reasonable, given the following results: 1) As shown in the manuscript, the genome of lung cancer in never smokers is very quiet with low mutational burden. However, we were able to detect driver events even in the low purity tumors due to the high WGS sequencing depth. For example, even in a sample with very low numbers of mutations, we found driver mutations with CCF close to 1 in *KRAS* (p.G12D) in NSLC-0214-T01 (**Supplementary Fig. 30)**. Similarly, we found few additional samples with very low mutation numbers but driver mutations in *EGFR*/*SETD2*/*ERBB2*/*CDKN2A*. 2) Although the VAF for the driver or other mutations was very low (<0.1), we could still capture these mutations with high sequencing quality. For example, we detected double *EGFR* mutations in cis (p.V834L and p.L858R) in tumor NSLC-0105-T01, but not in the blood control (**Supplementary Fig. 31**). A similar finding of such complex mutations in *EGFR* has been reported in lung cancer previously[1]. 3) We verified that the frequency of driver gene mutations was similar across tumors with different purity thresholds. When we stratified samples between low (n=88) and high (n=144) purity (purity cut-off 0.3), we observed no major difference in the frequencies of the top mutated genes: *EGFR*=28.4%, *TP53*=16.0%, *KRAS*=7.8%, *RBM10*=7.8% (high purity) and *EGFR*=25.9%, *TP53*=16.8%, *KRAS*=7.0%, *RBM10*=8.4% (low purity). Similarly, we observed no major

differences in number of SVs, frequency of kataegis, or prevalence of mutational signatures between the two purity groups (**Supplementary Fig. 32**). 4) We extracted the mutations with low VAF in the low purity tumors and checked the mutational spectra. We confirmed that we had sufficient power to detect different mutational signatures in the low purity tumors as shown in **Supplementary Fig. 33**.

Validation of the SCNA Subtypes

To validate the three SCNA subtypes (*forte, mezzo-forte* and *piano*), we extracted the TCGA lung adenocarcinoma (LUAD) SNP array data (N=544) from the GDC database, which includes 429 smokers and 75 non-smokers. We used the same SCNA clustering method to validate SCNA subtypes in TCGA LUAD alone or in combination with our Sherlock-*Lung* data. The SCNA clustering in TCGA LUAD alone identified the same three SCNA subtypes with similar SCNA features observed in our study (**Supplementary Figure 34a**). Similarly, when we combined TCGA LUAD SNP array data from non-smokers with our Sherlock-*Lung* data, we observed three SCNA subtypes with similar frequencies for Sherlock (*forte:* 20.0%, *mezzo-forte*: 32.3%, *piano:* 47.8%) and TCGA LUAD samples (*forte:* 22.7%, *mezzo-forte*: 29.3%, *piano*: 48.0%). We also found that 94% of Sherlock-*Lung* samples remained in the same SCNA cluster as in our original analysis after combination with TCGA LUAD samples (**Supplementary Figure 34b**).

Additional Genomic Characteristics of Lung Cancer in Never Smokers

Originally we had identified a group of 4 Sherlock-*Lung* tumors (NSLC-0038-T01, NSLC-0143-T01, NSLC-0068-T01 and NSLC-0059-T01) appearing to carry a homozygous deletion of the entire chr.19. However, a subsequent fluorescence in situ hybridization (FISH) analysis did not confirm this finding. **Supplementary Fig. 35** shows an example of miFISH results in tumor NSLC-0038-T01. We then refitted Battenberg SCNA profiles for chr.19 as normal copy number stage for these samples.

As reported previously[2], we observed a higher frequency of *EGFR* mutations (P=0.092) in female (31.4%) than in male patients (19.3%). *EGFR* and *KRAS* mutations were mutually exclusive (P=0.004). In contrast, mutations in *RBM10* (P=0.001) and *TP53* (P=0.004) were found to co-occur with *EGFR* more frequently than expected by chance. Co-occurring patterns were also found between *RBM10* and *PIK3CA* (P=0.003), as well as *TP53* and *ERBB2* (P=0.006). We also observed marginally significant enrichment of *SETD2* mutations in samples with oncogene fusions, particularly in *TP53*-proficient tumors (P=0.06), as previously reported[3].

A small number of tumors carried two nonsynonymous mutations in the same cancer driver genes with similar cancer cell fraction (CCF), including *EGFR* (n=8), *ARID1A* (n=3), and *TP53* (n=2). Among them, two tumors with double *TP53* mutations (NSLC-0020-T01 and NSLC-0183-T01) and one tumor with a double *EGFR* mutation (NSLC-0057-T01) on the same allele had also experienced loss of the other gene copy through loss-of-heterozygosity (LOH).

Of nine (4%) *ERBB2* mutations observed, eight were potentially targetable in-frame indels located in exon 20[4] with seven insertions E755delinsEAYVM and one deletion G761delinsVC (**Extended Data Fig. 5**).

Overall, 74 (31.9%) tumors, particularly in the *piano* subtype (50.4%), had no detectable known cancer driver gene mutations or fusions, with a TMB of 0.57 Mut/Mb in median, despite relatively higher purity (0.42 in average) in comparison to the other samples (0.34). Cancer stem cells can potentially drive some of these tumors. The addition of transcriptome and methylome data, as well as the study of medium-impact putative SNV passengers[5] will be necessary to further investigate these tumors' drivers in the future.

20% of LUAD had recurrent somatic indels affecting lineage-specific surfactant protein genes, comparable to previous reports[6] (**Figure 1**, **Supplementary Fig. 36**). As expected, no other histologic types harbored surfactant gene alterations. Three tumors carried *TERT* somatic mutations, including two recurrent promoter mutations (*C228T*) and one missense mutation (*p.R972H*). Low prevalence of somatic promoter mutations in *TERT* (2-3%) has been noted in previous studies of non-small cell lung cancer (NSCLC)[7,8].

The rate of somatic L1 insertions was low (mean 1.5, range: 0-34), of which 70.7% were Solo-L1 with no apparent transduction activity (**Supplementary Fig. 37a**), in sharp contrast to LUAD in smokers[9]. We identified known germline L1 source insertion hotspots (22q12.1, Xp22.2)[9,10] and a new recurrently inserted source element at 8q21.3 (**Supplementary Fig. 37b**).

In addition, tumors with kataegis events generally had significantly more genomic alterations, including SVs (P=1.3e-21), SNVs (P=1.9e-17), and SCNAs (P=1.4e-08), and were enriched for mutations in EGFR (P=3.9e-05), focal amplifications in *MDM2*, *TERT*, *GRIN2A*, and deletions in *CDKN2A* (P<0.01). Interestingly, mitochondrial DNA content and telomere length were correlated[11] (r=0.23, P=0.0004), particularly in tumors with STK11 focal deletion (r=0.96, P=0.0011, FDR=0.05; **Supplementary Fig. 38**).

Mutational Signatures in LCINS

Notably, our study included 62 patients who had been exposed to second-hand tobacco smoking, yet we did not identify appreciable contributions of SBS4 signatures or signatures of alkylation-induced mutagenesis previously associated with tobacco smoking. However, we cannot exclude that these signatures were not detected due to overall low TMB, algorithm sensitivity for identifying low-contribution signatures, or signature bleeding[12]. In contrast, mutational signature SBS8 was identified in 13% of LCINS samples (30% in carcinoids), but not observed in TCGA LUAD[13].

Comparing the replication and transcription strand asymmetries between never smokers and passive smokers did not reveal any differences for the six types of single base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G; **Supplementary Fig. 16**). Further comparisons between strand asymmetries of somatic mutations preferentially attributed to individual mutational signatures common in passive and never smokers (SBS1, SBS2, SBS3, SBS5, SBS13, SBS18, and SBS40)

revealed only one statistically significant result **(Supplementary Fig. 17**), Specifically, signature SBS18 exhibited a weak transcriptional strand asymmetry in never smokers which was not observed in the passive smokers (*Q*=0.034).

Simulations were performed to evaluate our resolution for detecting SBS4 in the 62 passive smokers. Specifically, SBS4 was injected at different average levels in all 62 passive smokers (1%, 5%, 10%, 15%, and 20% of mutations in each sample) and mutational signatures were re-extracted as done for the original data. Our results demonstrate that 62 passive smokers allow for detection of SBS4 as long as it contributes at least 15% of mutations in these 62 samples (**Supplementary Fig. 14)**. Indeed, in approximately 40% of simulations, our methodology is also able to detect SBS4 if it contributes 10% of mutations in these 62 samples. The method is unable to detect SBS4 in any of the simulations where SBS4 is injected at 5% or lower average levels. The lack of SBS4 in the original analysis indicates that signature SBS4, if present, has to be below the detection threshold of 15% of somatic mutations, which is possible and likely given that second-hand tobacco smoking is a weak carcinogen in comparison to active smoking. Moreover, there is the possibility that exposures to carcinogens in second-hand tobacco smoke affect cancer development not only through mutagenesis. Some carcinogens may contribute mutations as well as selective constraints to the evolutionary dynamics of somatic tissues. Further work in a larger sample size, with highly exposed individuals and with laboratory and animal models, including the analysis of the tumor microenvironment possibly at a single cell level is needed to characterize the carcinogenic process of second-hand tobacco smoking.

Approximately 16.0% (n=37) of tumors in our study had a high HRDetect score >0.7[14–16], which correlated with multiple genomic aberrations predictive of Homologous Repair Deficiency (HRD) (**Extended Data Fig. 8**). Biallelic loss (germline mutation and LOH) of *ATM* was found in a single patient, with a HRDetect score of 0.95. However, 42% of tumors had monoallelic loss due to LOH of HRD associated genes, resulting in higher HRDetect scores (**Extended Data Fig. 9**), therefore suggesting that in LCINS, loss of a single copy of an HRD associated gene may be sufficient to cause an HRD signature. Alternatively, different genomic features may need to be considered to identify HRD in lung tumors in comparison to breast tumors[13]. Notably, HRD tumors were enriched in subtypes *forte* and *mezzo-forte* (*P*=1.4e-3 versus *piano*) (**Fig. 3b**).

Stem cell-like features in the *piano* subtype

The low growth rate suggested by the occurrence of the MRCA approximately a decade prior to tumor diagnosis, scarcity and heterogeneity of driver mutations, low mutation rate, long telomeres, high level of intra-tumour heterogeneity (ITH), paucity of SBS18 indicating low ROS activity[17], and infrequent deficiency of TP53, which has been shown to regulate stemness of airway epithelial progenitors[18], are all consistent with *piano* tumors being derived from adult stem cells that have exited their quiescent state[19,20]. For an orthogonal assessment of stemness and cell of origin, we conducted RNA sequencing analyses on 35 Sherlock-*Lung* adenocarcinomas and 70 TCGA LUAD samples from nonsmokers. Since gene expression can be affected by copy number state[21], we focused on the subgroup of samples that were copy neutral in the regions studied, and combined the *forte* and *mezzo-forte* groups for comparison with the *piano* subtype. We tested the

expression of several genes and created a “development score”, which included the *SOX2*, *SOX9*, and *HMGA2* genes[22–24]. The development score suggested an increase of expression in the *piano* group in both the Sherlock and TCGA datasets (**Supplementary Fig. 27a**), which is consistent with a stem cell-like state of the *piano* subtype. In addition, we tested previously demonstrated hallmarks of LUAD that are also suggestive of loss of differentiation, such as decreased expression of genes associated with AT2 cells (proposed LUAD cell of origin), and aberrant expression of multiple lineage-specific signatures, or “lineage infidelity”[25]. We found a higher expression of *TP63*, a marker of basal cells[26] (**Supplementary Figure 27b**), in the *piano* subtype in both datasets. Since basal cells are not typically identified in LUAD, if confirmed in a larger dataset, this finding suggests that there is “lineage infidelity” in the *piano* subtype. We note, however, that these analyses were based on very small numbers of tumors. Much larger studies are needed to verify the WGS-based stemness hypothesis, possibly using single-cell RNA sequencing.

Supplementary Methods

Specimen Collection and Processing

Of the 232 samples, 125 were from the Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval (IUCPQ-UL), Quebec, Canada; 47 from Université Côte d’Azur, Nice, France; 27 from the EAGLE study from Italy[27]; 22 from Yale University, New Haven, Connecticut, USA; and 11 from H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, USA.

Frozen tumor tissue with matched blood or normal tissue samples were immediately put into 1ml of 0.2 mg/ml Proteinase K (Qiagen) in DNA lysis buffer (10 mMTris-Cl (pH 8.0), 0.1 M EDTA (pH 8.0), and 0.5% (w/v) SDS) for 24 hrs at 56°C with shaking at 850 rpm in Thermomixer R (Eppendorf) until the tissue was completely lysed. Genomic DNA was extracted from fresh frozen tissue using the QIAmp DNA Mini Kit (Qiagen) according to the manufacturer’s instructions. Each sample was eluted in 200 µl AE buffer and DNA concentration was determined by Nanodrop spectrophotometer. All DNA samples were aliquoted and stored at −80°C until use.

DNA was quantified using the QuantiFluor® dsDNA System (Promega Corporation, USA). DNA was normalized to 25ng/ul and underwent fragment analysis via AmpFLSTR™ Identifiler™ PCR Amplification Kit (ThermoFisher Scientific, USA). DNA samples are required to meet minimum mass and concentration thresholds for each assay, as well as show no evidence of contamination or profile discordance in the Identifiler assay. Samples meeting these requirements were aliquoted at the appropriate mass needed for downstream assay processing.

Illumina Library Construction and Sequencing

The resulting post-capture enriched multiplexed sequencing libraries were used in cluster formation on an Illumina cBOT (Illumina, San Diego, CA, USA). Paired-end sequencing was performed by the Broad Institute (https://www.broadinstitute.org) using the Illumina HiSeq X system following Illumina-provided protocols for 2x150bp paired-end sequencing.

FASTQ files were generated after Illumina base-calling. Next, paired FASTQ were converted to umapped BAM using the GATK pipeline (https://github.com/gatk-workflows/seq-format-conversion). The unmapped BAM files were then processed using GATK on the cloud-based platform TERRA workspaces (https://app.terra.bio). The sequence data were aligned to the human reference genome GRCh37 (Homo_sapiens_assembly19.fasta), and the aligned BAM files were transferred to the NIH HPC system (https://hpc.nih.gov) for downstream analyses.

Sample Quality Checking

Average sequencing depths for normal and tumor samples were 32x and 85x, respectively. We analyzed WGS data in a unified pipeline and applied strict criteria for sample and data QC, including mutational signature analysis to remove samples misclassified as primary LCINS or from never smokers.

Sequencing depth was calculated using mosdepth[28], and sequencing library insert size was estimated using Picard Toolkit. Tumors (N=20) were excluded based on the following criteria: 1) Tumor samples had less than 100 total number of genomic alterations including SNV, INDEL, SV, and TE, no driver mutations in any known lung adenocarcinoma driver genes as reported by the TCGA study[29], and had no copy number alteration detected; 2) Tumor or normal samples had contamination >5% identified by VerifyBamID[30]; 3) Tumor and normal sample were swapped based on pairwise sample relatedness metrics, as detected by Somalier[31]; or 4) Normal samples were identified with large proportion copy number alterations using the Battenberg algorithm[32]. In addition, four tumors were excluded based on mutational signatures: two samples with signature SBS7a/b/c/d that turned out to be from squamous cell carcinomas of the skin metastasized to the lung; one sample with signature SBS4 that was confirmed to belong to a current smoker; and one sample with signature 31 originated from previous platinum-based treatment. In summary, a total of 232 subjects were included in this study after QC.

We compared tumor mutational burden (normalized as number of mutations per megabase) against TCGA cohorts using maftools[33]. The distribution of mutations was compiled from over 10,000 WES samples across 33 TCGA landmark cohorts[34]. The smoking status for the TCGA LUAD samples was extracted from the TCGA LUAD clinical file under the column: “Patient Smoking History Category”.

Acronyms of cancer types in TCGA include SKCM: skin cutaneous melanoma; LUSC: lung squamous cell carcinoma; LUAD: lung adenocarcinoma; BLCA: bladder urothelial carcinoma; STAD: stomach adenocarcinoma; COAD: colon adenocarcinoma; HNSC: head & neck squamous cell carcinoma; READ: rectum adenocarcinoma; DLBC: lymphoid neoplasm diffuse large B-cell adenocarcinoma; ESCA: esophageal carcinoma; OV: ovarian serous cystadenocarcinoma; LIHC: liver hepatocellular carcinoma; CESC: cervical squamous cell carcinoma and endocervical adenocarcinoma; UCEC: uterine corpus endometrial carcinoma; KIRP: kidney renal papillary cell carcinoma; GBM: glioblastoma multiforme; KIRC: kidney renal clear cell carcinoma; UCS: uterine carcinosarcoma; BRCA: breast invasive carcinoma; PAAD: pancreatic adenocarcinoma; SARC: sarcoma; LGG: brain lower grade glioma; PRAD: prostate adenocarcinoma; CHOL:

cholangiocarcinoma; MESO: mesothelioma; ACC: adrenocortical carcinoma; KICH: kidney chromophobe; UVM: uveal melanoma; TGCT: testicular germ cell tumors; THYM: thymoma; PCPG: pheochromocytoma and paraganglioma; THCA: thyroid carcinoma; LAML: acute myeloid leukemia.

Estimation of Tumor Purity, Ploidy, Clonality, and Allele-Specific Copy Numbers

We used the updated Battenberg (v2.2.8) algorithm[32] to estimate the clonality of each segmentation, tumor purity, and tumor ploidy. In general, we ran the algorithm without providing "rho" (purity) and "psi" (ploidy) as a first step, so that it could estimate the best values for these parameters. However, a few samples showed that tumor ploidy and purity were not consistent with estimated copy number status of segmentations for the same tumor (e.g., tumor genome identified as whole-genome doubling (WGD) but ploidy <3). We then refitted the Battenberg algorithm by picking a reference segment with a confident copy number and calculated the "rho" and "psi" values using the "calc_rho_psi_refit" function. The newly estimated "rho" and "psi" values were then used as input in a rerun of theBattenberg algorithm. Finally, we manually checked the Battenberg average and subclone profiles to ensure the accuracy of each estimate. The final output generated two files (i.e., clone and subclone segmentation file "subclones.txt.gz" and tumor purity and ploidy file "rho_and_psi.txt") that were used in downstream analyses. Unsupervised clustering of copy number profiles including both major clone and subclone segmentations were generated based on relative copy number log2(copy number/Tumor_Ploidy) using the euclidean distance and Ward's method.

Clustering of subclonal somatic mutations was analyzed using the Bayesian Dirichlet Process as previously described[32,35,36]. Allele-specific copy number alterations called by the Battenberg algorithm and read count information of each mutation were used to calculate the cancer cell fraction (CCF). Clone clusters were identified as local peaks in the posterior mutation density estimated from the Dirichlet process using DPClust. For each cluster, a region representing a 'basin of attraction' was defined as the point of minimum density between each pair of clusters. Each mutation was then assigned to the cluster in whose basin of attraction they were most likely to fall, using the posterior probabilities obtained from the Dirichlet process. We separated clonal and subclonal clusters using the average CCF for mutations in the same cluster. To refine the tumor purity estimation from the Battenberg algorithm and estimate the purity in the absence of any copy number alterations, we multiplied the purity estimate by the position of the clonal peak in DPClust. The average tumor purity estimated from copy number alterations and mutations was 37.3%.

Structural Variants Calling

The Meerkat algorithm[37] was used to call somatic SVs and estimate the corresponding genomic positions of breakpoints. The parameters were selected based on the sequencing depth for both tumor and normal tissue samples and the library insert size as in a previous publication[36]. Candidate breakpoints were first found based on soft-clipped and split reads, which required identifying at least two discordant read pairs, with one read covering the actual breakpoint

junction, and then confirmed to be the precise breakpoints by local alignments ('meerkat.pl'). Mutational mechanisms were predicted based on homology and sequencing features ('mechanism.pl). SVs from tumor genomes were filtered by those in normal genomes. SVs found in simple or satellite repeats were also excluded from the output ('somatic_sv.pl'). The final somatic SVs were annotated as a uniformed format for all breakpoints ('fusions.pl').
Driver oncogenic fusions were selected from SVs based on the driver gene list[29] and an oncogenic fusion list previously reported in LUAD[3]. We selected the fusions with the following SV features in meerkat output: "gene-gene", "head-tail" and "in_frame" or "out_of_frame". All driver oncogenic fusions in our study were reported with the same partners and no other new recurrent oncogenic fusion was found.

Recurrent Copy Number Alterations

To identify recurrent copy number alterations, we used the major clonal copy number for each segmentation from the Battenberg algorithm. Recurrent copy number alterations from WGS at a gene level (-genegistic) were identified using GISTIC 2.0[38] as previously described[39] with a broad analysis (-broad) and arm-level peel off (-armpeel), a threshold for deletions and amplifications of 0.25 (-ta,-td), and a threshold for broad events of 0.98 (-brlen). The confidence level for calculating driver regions was 0.90 (-conf). Marker-level copy number data was collapsed to gene-level data using the extreme method (-gcm). Significant copy number alteration callings at gene-level and arm-level were obtained from the "all_thresholded.by_genes.txt" and "broad_values_by_arm.txt" files, where those with *q*-values <0.01 were considered statistically significant. Only GISTIC2 calls of -2 (deletion) and +2 (amplification) were considered. To compare the significance of the identified driver genes as well as the copy number profile, we downloaded available data (MutSigCV and Gistic2 results) from the TCGA cohort [40].

Kataegis Identification

We followed previous methods to identify kataegis events[41,42]. KataegisPortal (https://github.com/MeichunCai/KataegisPortal) was used to detect the kataegis loci and visualize the kataegis events from SNVs. A minimum of 5 mutations and maximal distance of 1,000 bp were required for kataegis events. Few features were observed for the kataegis loci when manually reviewing the 'rainfall' plots: 1) a preponderance to C>T and C>G mutations; 2) The enrichment for TpC sequence context; 3) processivity consecutive mutations within a kataegis locus were on the same strand. In addition, SeqKat[43] was used to score the hypermutation and APOBEC mediated kataegis, and all kataegis locus found in our study showed extremely high APOBEC mediated kataegis score.

Strand asymmetry analysis between never and passive smokers

Analysis of strand asymmetry was performed using SigProfilerTopography (v1.0.19; https://github.com/AlexandrovLab/SigProfilerTopography). Briefly, SigProfilerTopography was run separately for passive smokers and for never smokers. SigProfilerTopography distinguishes whether mutations in a given sample occur on transcribed or un-transcribed strand with respect to transcription, and on leading or lagging strand with respect to replication.

SigProfilerTopography identifies transcription and replication strand bias for aggregated mutations and for mutations preferentially attributed to a mutational signature. For each sample, SigProfilerTopography was used, first, to simulate the observed somatic mutations with their patterns using SigProfilerSimulator[44], and then to compare the transcription and replication strand asymmetries between the simulated and real somatic mutations across all samples using Fisher's exact tests. Multiple testing correction was performed to control the false discovery rate using the Benjamini–Hochberg procedure.

Additionally, for each sample, SigProfilerTopography calculates the ratio of the transcribed and un-transcribed strands, as well as the ratio of leading and lagging strands. As such, for each of the six types of single base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) and mutational signatures common in passive and never smokers (SBS1, SBS2, SBS3, SBS5, SBS13, SBS18, and SBS40), we have two ratios in each examined sample: transcribed/un-transcribed strand ratio and leading/lagging strand ratio. We performed Mann Whitney rank tests comparing these ratios between never smokers and passive smokers to evaluate the strand asymmetry for each type of single base substitution and for each common mutational signature. Multiple testing correction was performed to control the false discovery rate using the Benjamini–Hochberg procedure.

Statistical power for detecting mutational signatures

Prior research[45] has attributed SBS4 to tobacco smoking mutagenesis in lung cancer. To evaluate the statistical power for detecting SBS4 in passive smokers, we performed simulations where SBS4 was injected in the mutational vectors of each of the 64 passive smokers. For the mutational vector corresponding to each sample, a percentage of mutations attributed to SBS4 was drawn from a Gaussian distribution with an average of 1%, 5%, 10%, 15%, and 20% with a standard deviation reflecting ten percent of the particular average value. For example, if a sample has 1,000 mutations and SBS4 was injected at 1% on average, then a sample is expected to get an average of 10 mutations from SBS4 with a standard deviation of 1 mutation. The simulations were repeated 10 times for each average level (1%, 5%, 10%, 15%, and 20%) and *de novo* extraction of mutational signatures with SigProfilerExtractor was performed for the data in each simulated dataset as done for the original data. SBS4 was considered successfully identified if SigProfilerExtractor reported it as one of the operative COSMIC signatures.

RNA-sequencing and stemness analysis

RNA-seq libraries from 35 Sherlock-Lung tumor samples and 32 paired adjacent normal tissue samples were prepared using the Illumina TruSeq Stranded Total RNA Sequencing protocol, and then sequenced on the Illumina HiSeq platform, generating 100 or 150 bp paired-end reads. The FASTQ files were aligned to the human reference genome GRCh38/hg38 using STAR[46], and the BAM files were processed by RSEM[47] for gene quantification. The resulting gene expression data were normalized to transcripts per million (TPM) estimates, followed by log2 transformation.

The RNA-seq data and copy number alteration data from the TCGA-LUAD study were downloaded from the GDC Data Portal[48]. We extracted data of 70 tumor samples from non-

smokers and 59 normal lung tissue samples (including 7 paired normal samples for non-smokers) and used the $\log_2$ transformed normalized gene expression data for the analyses.

For each tumor and normal sample in both data sets, the lung development score was determined using the average levels of three marker genes: *SOX2*, *SOX9* and *HMGA2*. Tumor samples with neutral copy numbers in all three genes were analyzed. Gene expression in the tumor samples was compared with that from paired normal samples for the Sherlock-*Lung* set or from pooled normal samples for the TCGA-LUAD set. We used the Student's t-test to compare the development scores across tumor subtypes (forte+mezzo-forte versus piano) and between tumor and normal samples. Similar analyses were conducted to estito the basal cell marker, *TP63*, with copy neutral tumor samples determined by the copy number of *TP63*.

Multiplex interphase fluorescence in situ hybridization (miFISH) for validation of chr.19 homozygous deletion

Preparation of cytospins from archived tumor blocks

All FISH experiments were performed on formalin-fixed, paraffin-embedded (FFPE) tumor blocks, from which two sections (thickness: 5 µm, and 50 µm) were consecutively prepared. The 5 µm sections were stained with hematoxylin and eosin (H&E) and representative tumor areas were marked on all slides by a pathologist. Using this H&E stained section for guidance, the pathologist outlined the corresponding tumor area on the unstained 50 µm sections. This 50 µm section was used to prepare cytospins for the miFISH analysis using the modified Hedley-method as described previously[49]. In brief, the representative area on the unstained 50 µm section was macrodissected, deparaffinized in xylene, rehydrated, and digested with protease. Tissue disintegration was monitored under microscopic control. Once it was judged to be optimal, the reaction was stopped and the density of the single cell suspension was adjusted to achieve medium dense, monolayered cytospins.

Multiplex interphase fluorescence in situ hybridization (miFISH)

A multicolor FISH probe panel consisting of three gene-specific probes: COX2 (1q33.2) labeled in red, TERC (3q26.2) labeled in gold and CCNE1 (19q12) labeled in aqua and two centromere probes: CEP7 labeled in green and CEP10 labeled in far red were hybridized onto the cytospin slide and imaged by the DUET scanning imaging workstation (BioView Ltd., Rehovot, Israel, software version 3.6.0.15) as previously described[49,50]. All signals were manually counted and reviewed for accuracy using the SOLO workstation (BioView Ltd.) applying the following exclusion criteria: (i) overlap with other nuclei, (ii) visibly damaged or incomplete nuclei or (iii) insufficient hybridization quality, e.g., one or more not clearly discernible probe signals. 100 aberrant nuclei were evaluated per case (64 nuclei for case NSLC-0059-T01).

Retrotransposon analysis

For the identification of putative solo-L1 and L1-transduction integration sites, we used a pipeline called TraFiC-mem (Transposome Finder in Cancer)[9]. Four standard steps have been applied to identify somatic TEs including solo-L1, Alu, SINE and ERV: 1) selection of candidate reads; 2) transposable element masking; 3) clustering and prediction of TE integration sites; 3) filtering of

germline events; and 4) identification of L1-transductions. Due to the short read lengths and library inserts generated, we were unable to attribute somatic solo-L1 transpositions to their germline source.

Microsatellite Instability Detection

Microsatellite instability (MSI) is an important indicator of genome instability and has been linked to many genetic diseases. MANTIS[51] was used to detect MSI status for each patient. Paired tumor, normal BAM files, and default parameters were used as input for this algorithm. We found no MSI in our samples, in line with the very low MSI frequency detected in TCGA LUADs[52].

Mitochondrial DNA Copy Number Estimation

We estimated mtDNA copy number using a similar method as previously described by Reznik *et al*[53]. Briefly, sequencing reads were analyzed to determine the number aligning to each chromosome. Relative abundance of mitochondrial DNA was calculated as the ratio of mtDNA reads to nuclear DNA reads with correcting tumor purity and ploidy. Pairs of matched tumor and adjacent-normal samples were compared to uncover patterns of mtDNA accumulation and depletion. To ensure a meaningful comparison, we only used adjacent-normal tissue for the analysis. In addition, we also collected previously estimated mtDNA content from TCGA WES and WGS analyses[53] for comparison with our Sherlock-*Lung* study. For the mutations in mtDNA, we used the same mutation calling method as for the nuclear DNA.

Mutational Signature Analysis

Mutational signature analysis was analyzed by the updated computational framework SigProfiler[41,54]. To summarize, SigProfilerMatrixGenerator[55] was first used to create mutational matrices for all types of somatic mutations including SBS, DBS and ID. In particular, mutational profiles of SBS-1536, DBS-78 and ID-83 were used as the inputs for SigProfilerExtractor (v0.0.5.77) (https://pypi.org/project/sigproextractor/). SigProfilerExtractor with default parameters was used to perform both *de-novo* extraction and decomposition to known global cosmic mutation signatures (v3). For *de-novo* extraction, we tested the number of signatures to be extracted from 1 to 18 and selected the number of signatures based on minimum stability and mean L2 % as suggested by the SigProfilerExtractor. The extracted *de-novo* signatures were then compared to known global signature compositions, and new signatures were assigned if cosine similarities were less than 0.8. In our study, no new signature was identified for all types of mutational signatures, including SBS, DBS and ID. Suggested global signatures were then selected for the decomposition solution. SigProfilerPlotting was used to visualize all types of mutational signatures as well as all types of mutational patterns. The relative contributions of mutational signatures were calculated based on the number of mutations assigned to each signature. Of note, SBS18 contribution was highest in subject NSLC-0147, who carried an known rs34612342 germline mutation in *MUTYH*[56]. Mutation probabilities for each mutation type in each sample were generated for grouping samples based on different genomic features. Hierarchical clustering of contribution of mutational signatures was performed using “euclidean” distance and Ward’s minimum variance clustering method.

Enrichment and Mutation Load for hTg → hGg

We used this approach to confirm the absence of tobacco smoking related Alkylating signatures in passive smokers. To determine enrichment, the fraction of base substitutions of a given type in a signature-defining trinucleotide motif among all base substitutions of this type was calculated. This ratio was further compared to the prevalence of the trinucleotide motif versus the single base in the 40 bases flanking the mutated residue[57]. Enrichment calculation for hTg → hGg is provided below:

$$Enrich_{(hTg \rightarrow hGg)} = \frac{mutations_{(hTg \rightarrow hGg)} \times context\ (T)}{mutations_{(T \rightarrow G)} \times context_{(hTg)}}$$

Fisher's exact test was utilized to determine if enrichment was statistically significant and Benjamini-Hochberg corrected *p*-values were applied for multiple testing. The minimum estimate of mutation loads was only calculated for those samples whose enrichment > 1 and the corrected *p*-value < 0.05. The calculation for minimum estimate of mutations loads is shown below:

$$MutLoad_{(hTg \rightarrow hGg)} = \frac{mutations_{(hTg \rightarrow hGg)} \times (Enrich_{(hTg \rightarrow hGg)} - 1)}{Enrich_{(hTg \rightarrow hGg)}}$$

Analysis of Mutagenesis by APOBEC Cytidine Deaminases (P-MACD)

The analysis of APOBEC-related mutations was performed using the P-MACD (Pattern of Mutagenesis by APOBEC Cytidine Deaminases) tools[58]. Mutation annotation file (MAF) was used as input and sample specific summaries were generated, including fold-enrichment with APOBEC mutagenesis signature "APOBEC_enrich", benjamini-Hochberg-corrected p-values "BH-Fisher_p-value_tCw", and minimum estimate of the number of APOBEC induced mutations in a sample "APOBEC_MutLoad_MinEstimate". "APOBEC_MutLoad_MinEstimate" was then compared to the number of APOBEC mutations (SBS2+SBS13) estimated by SigProfilerExtractor.

Y/RTCA Enrichment Analysis and Identification of A3A-like or A3B-like tumor

An APOBEC3A hypermutation signature is distinguishable from the signatures of background mutagenesis by APOBEC3B in human cancers[59]. We followed their procedure to calculate the Y/RTCA enrichment and identify A3A-like or A3B-like tumors. To summarize, the $\chi 2$ test for goodness of fit was used to identify samples that had a ratio of YTCA to RTCA mutations which differed statistically from random, by comparing observed *vs.* expected mutation counts. The expected number of YTCA mutations, given the null hypothesis of random mutagenesis, simply scales with fraction of motifs at YTCA:

$$Exp_{YTCA} = mutations_{TCA} \times \frac{context_{YTCA}}{Ccontext_{TCA}}$$

P-values were corrected by the BH method with *q*-values < 0.05 considered significant. The expected number of RTCA and NTCA mutations was computed analogously.

Reconstructing Mutational Signature Trajectories

TrackSig was applied to reconstruct evolutionary trajectories of mutation signature activities based on the variant allele frequency (VAF) data from a single heterogeneous tumor sample[60]. In brief, SNVs were sorted according to their estimated CCF using the VAFs and copy number profile of the same sample. Then, a trajectory of the mutational signature activities was inferred over the estimated ordering of the SNVs. We estimated the activity trajectory for each signature as a piecewise constant function of the SNV ordering with a small number of changepoints. We specified a separate set of active signatures for each sample as we identified using SigProfilerExtractor.

To summarize the average signature trajectories over time across all the tumors, we used the TrackSig output for all tumors ("mixtrues.csv"), and removed signature trajectories from all tumors if activity or activity change less than 5%. We also removed signature trajectories from consideration if a particular signature was observed in less than 10 tumors. We rescaled the trajectories so that all tumors had the same number of time points and the same subclonal boundary. First, we used the median of positions of clonal-subclonal boundaries in each tumor and set the number of time points (N=97). Then, we used the splines to rescale the clonal and subclonal parts of the timeline to have the same clonal-subclonal boundary. In addition, we computed per-time-point mean across trajectories for the same signature and subtracted the mean of signature trajectories to make the activity as zero at clonal-subclonal boundary. Finally, we plotted the change in signature activity across the time points.

Estimating the Timing of Duplication

The relative timing of each copy number gain or UPD (uniparental disomy) were estimated using the recently published R package Palimpsest[61]. In general, we estimated the molecular timing of duplicates using the proportion of duplicated/non-duplicated somatic mutations as previously described[62].

For a simple case of a chromosome with absolute copy number N_t = 3, the molecular time at which the extra copy of the chromosome was gained can be estimated as:

$$T = \frac{N_{dup}}{N_{dup} + \frac{N_{ndup} - N_{dup}}{3}} \times 100$$

where, N_{dup} and N_{ndup} are the number of duplicated and non-duplicated mutations, respectively.

This formula can be adjusted for chromosomes with $N_t >= 4$ to estimate the first duplication event using:

$$T = \frac{N_{dup}}{N_{dup} + \frac{N_{ndup} - N_{dup}}{(3 + N_t)/2}} \times 100$$

where, N_{dup} is the number of mutations at the maximal level of multiplicity and N_{ndup} the number of mutations at intermediate levels of multiplicity or non-duplicated.

In addition, if two parental chromosome copies were duplicated, we adapted the formula as follows:

$$T = \frac{N_{dup}/2}{N_{dup}/2 + \frac{N_{ndup}}{(3 + N_t)/2}} \times 100$$

We applied these formulas to time every clonal duplication identified by Battenberg in our study with at least 30 somatic mutations located within the duplicated chromosome region.

Subtype Analyses Based on Molecular Alterations

Two pathways were most frequently altered, including TP53 deficiency and RTK-RAS pathway. We observed a total exclusive pattern between *TP53* mutations and *MDM2* amplification. Thus, we defined tumors with *TP53* mutations or *MDM2* amplifications as the “TP53 deficiency group”, while the rest of tumor samples characterized the TP53 proficient group. For the RTK-RAS pathway status, we used the following criteria to select the driver genes: 1) genes with alteration frequency > 3%; or 2) driver genes included in known oncogenic fusions with the same known partners; or 3) genes included in the significant GISTIC peak with focal amplifications or deletions (*q*-value < 0.01). As a result, seven genes within the RTK-RAS pathway were selected, including *EGFR*, *KRAS*, *ALK*, *MET*, *ERBB2*, *ROS1*, and *RET*. Most of these genes showed exclusive patterns with each other. *BRAF* amplifications co-occurred with alterations in these seven genes, and no recurrent mutations were found in *NF1*. Therefore, *BRAF* and *NF1* were not considered. In addition, we defined a third group characterized by the copy number cluster ***piano***, ‘quiet tumors’, and compared genomic features in this group with those in the rest of the tumors. In this group, there were almost no copy numbers and very few driver mutations. We examined all groups for any potential driver events as shown in **Fig. 2**.

LUAD WGS Analytical Data from PCAWG

We used the Battenberg profile and the WGD status of 38 LUAD WGS samples from the PCAWG study (PCAWG, 2020). The mutational signature SBS4 related to tobacco smoking was extracted from these data by SigProfilerExtractor as previously described[54].

Additional statistical analysis

For enrichment analyses, categorical variables were compared using the two-sided Fisher's exact test. For two group comparisons, group variables were compared using the Mann-Whitney U test, or the Wilcoxon signed-rank test or the Student's *t*-test. Survival analyses were performed using the R survival package and visualized using the R survminer package.

References

1. Hata, A. *et al.* Complex mutations in the epidermal growth factor receptor gene in non-small cell lung cancer. *J. Thorac. Oncol.* **5**, 1524–1528 (2010).
2. Yuan, Y. *et al.* Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer Cell* **29**, 711–722 (2016).
3. Lee, J. J.-K. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177**, 1842–1857.e21 (2019).
4. Chuang, J. C. *et al.* ERBB2-Mutated Metastatic Non-Small Cell Lung Cancer: Response and Resistance to Targeted Therapies. *J. Thorac. Oncol.* **12**, 833–842 (2017).
5. Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* **180**, 915–927.e16 (2020).
6. Imielinski, M., Guo, G. & Meyerson, M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**, 460–472.e14 (2017).
7. Jung, J.-S. *et al.* Rare but poor prognosis of TERT promoter mutation in non-small cell lung cancer patients. *J. Thorac. Oncol.* **11**, S29–S30 (2016).
8. Ma, X. *et al.* Recurrent TERT promoter mutations in non-small cell lung cancers. *Lung Cancer* **86**, 369–373 (2014).
9. Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
10. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
11. Zheng, Q., Huang, J. & Wang, G. Mitochondria, Telomeres and Telomerase Subunits. *Front Cell Dev Biol* **7**, 274 (2019).
12. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
13. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249–263 (2020).
14. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
15. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
16. Zhao, E. Y. *et al.* Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clin. Cancer Res.* **23**, 7521–7530 (2017).
17. Cho, I. J. *et al.* Mechanisms, Hallmarks, and Implications of Stem Cell Quiescence. *Stem Cell Reports* **12**, 1190–1200 (2019).
18. McConnell, A. M. *et al.* p53 Regulates Progenitor Cell Quiescence and Differentiation in the

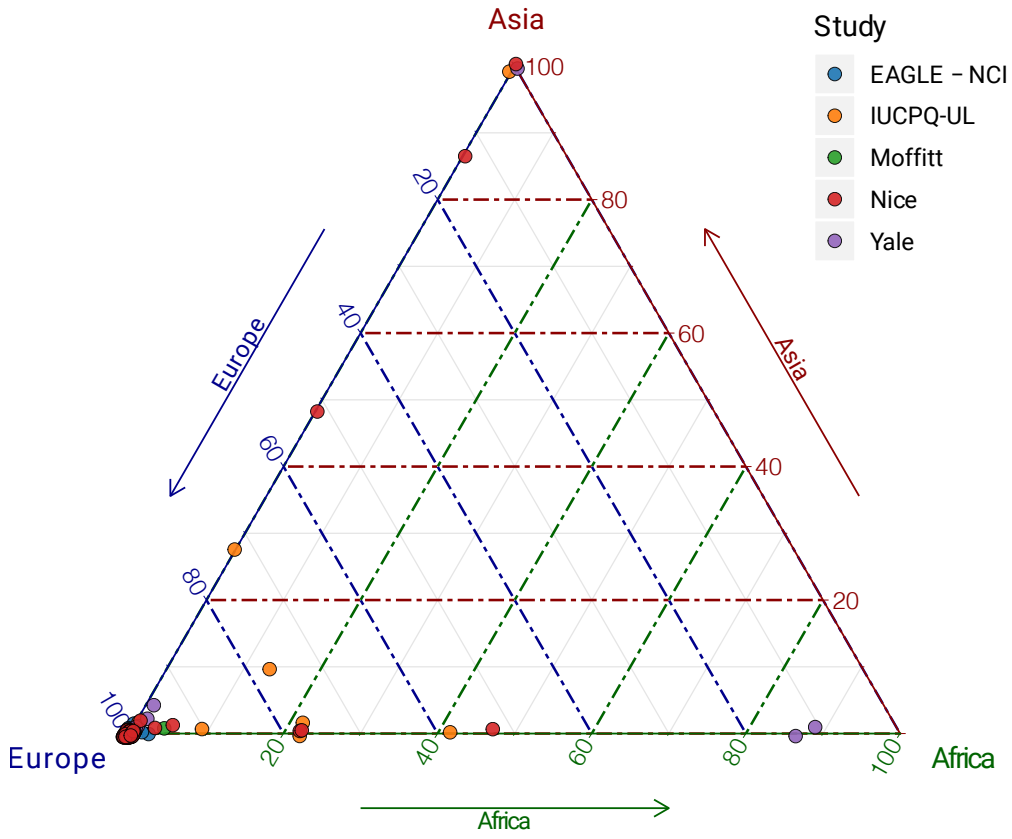
Airway. *Cell Rep.* **17**, 2173–2182 (2016).
19. Fukada, S.-I., Ma, Y. & Uezumi, A. Adult stem cell and mesenchymal progenitor theories of aging. *Front Cell Dev Biol* **2**, 10 (2014).
20. Li, L. & Clevers, H. Coexistence of quiescent and active adult stem cells in mammals. *Science* **327**, 542–545 (2010).
21. Shao, X. *et al.* Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* **20**, 175 (2019).
22. Maeda, Y., Davé, V. & Whitsett, J. A. Transcriptional control of lung morphogenesis. *Physiol. Rev.* **87**, 219–244 (2007).
23. Alanis, D. M., Chang, D. R., Akiyama, H., Krasnow, M. A. & Chen, J. Two nested developmental waves demarcate a compartment boundary in the mouse lung. *Nat. Commun.* **5**, 3923 (2014).
24. Singh, I. *et al.* Hmga2 is required for canonical WNT signaling during lung development. *BMC Biol.* **12**, 21 (2014).
25. Laughney, A. M. *et al.* Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat. Med.* **26**, 259–269 (2020).
26. Rock, J. R. *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12771–12775 (2009).
27. Landi, M. T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008).
28. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
29. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
30. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
31. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *bioRxiv* 839944 (2019) doi:10.1101/839944.
32. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
33. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
34. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).
35. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb. Perspect. Med.* **7**, (2017).
36. Zhu, B. *et al.* The genomic and epigenomic evolutionary history of papillary renal cell carcinomas. *Nat. Commun.* **11**, 3096 (2020).
37. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
38. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
39. Steele, C. D. *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary

Pathways. *Cancer Cell* **35**, 441–456.e8 (2019).
40. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
41. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
42. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
43. Yousif, F. *et al.* The Origins and Consequences of Localized and Global Somatic Hypermutation. *bioRxiv* 287839 (2018) doi:10.1101/287839.
44. Bergstrom, E. N., Barnes, M., Martincorena, I. & Alexandrov, L. B. Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics* **21**, 438 (2020).
45. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
48. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
49. Heselmeyer-Haddad, K. *et al.* Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *Am. J. Pathol.* **181**, 1807–1822 (2012).
50. Oltmann, J. *et al.* Aneuploidy, TP53 mutation, and amplification of MYC correlate with increased intratumor heterogeneity and poor prognosis of breast cancer patients. *Genes Chromosomes Cancer* **57**, 165–175 (2018).
51. Kautto, E. A. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* **8**, 7452–7463 (2017).
52. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis Oncol* **2017**, (2017).
53. Reznik, E. *et al.* Mitochondrial DNA copy number variation across human cancers. *Elife* **5**, (2016).
54. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
55. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
56. Thibodeau, M. L. *et al.* Base excision repair deficiency signatures implicate germline and somatic MUTYH aberrations in pancreatic ductal adenocarcinoma and breast cancer oncogenesis. *Cold Spring Harb Mol Case Stud* **5**, (2019).
57. Saini, N. *et al.* The Mutation Signatures Specific To DNA Alkylating Agents In Yeast And Cancers. in *ENVIRONMENTAL AND MOLECULAR MUTAGENESIS* vol. 60 51–51 (WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2019).
58. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
59. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the

signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).
60. Rubanova, Y. *et al.* Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. *Nat. Commun.* **11**, 731 (2020).
61. Shinde, J. *et al.* Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics* **34**, 3380–3381 (2018).
62. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).

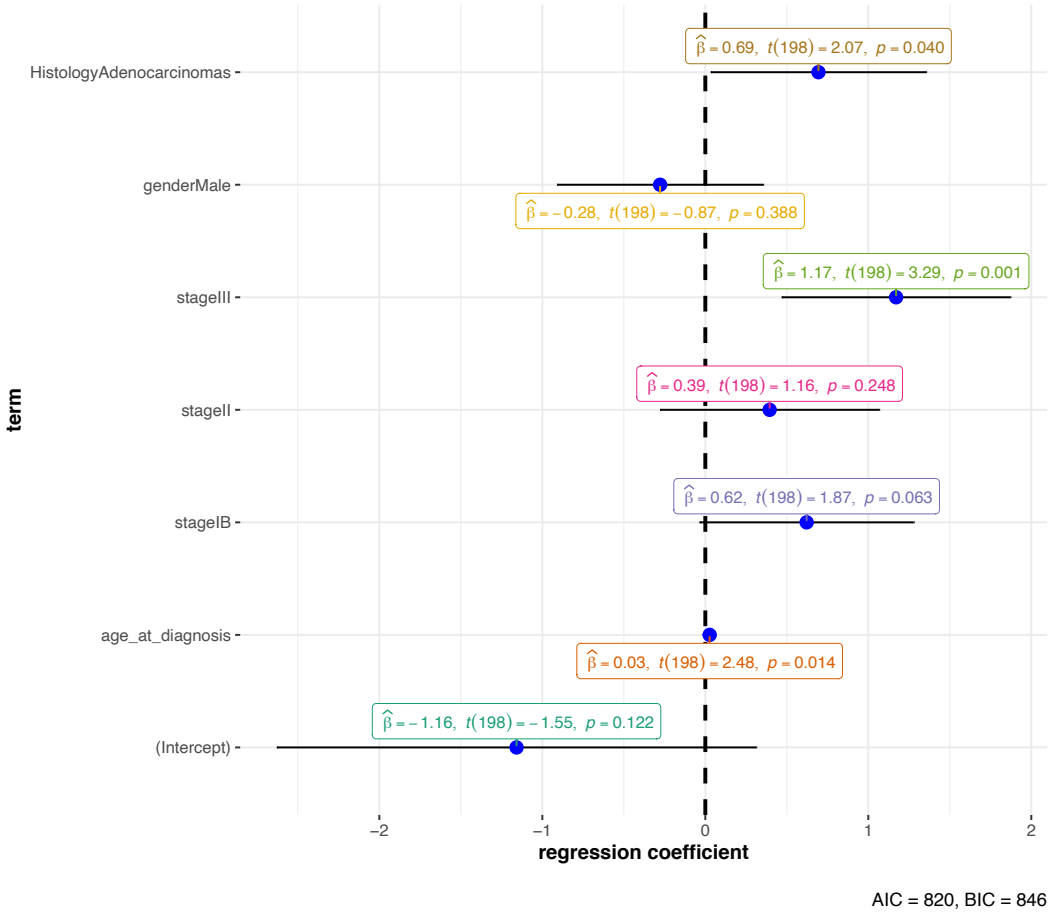
Supplementary Figures

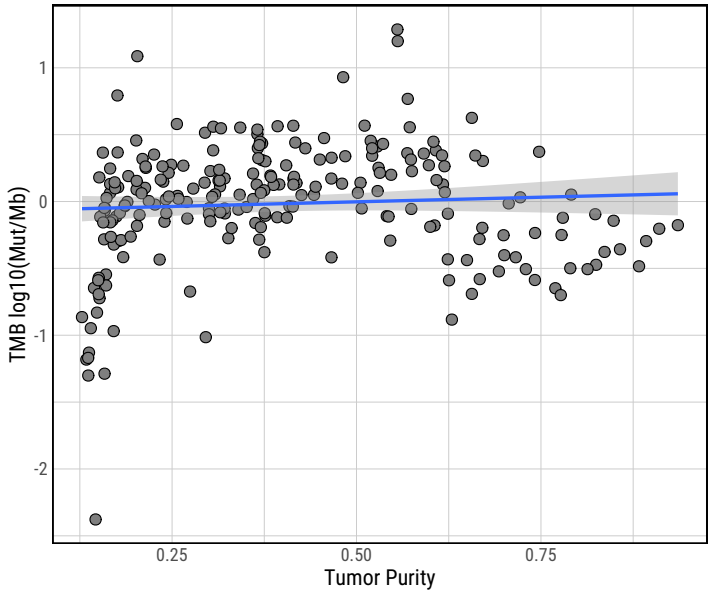
Supplementary Fig. 1



Supplementary Fig. 1 Patients' ancestry proportions based on ADMIXTURE analysis.
Each dot represents an individual color-coded by study centers.

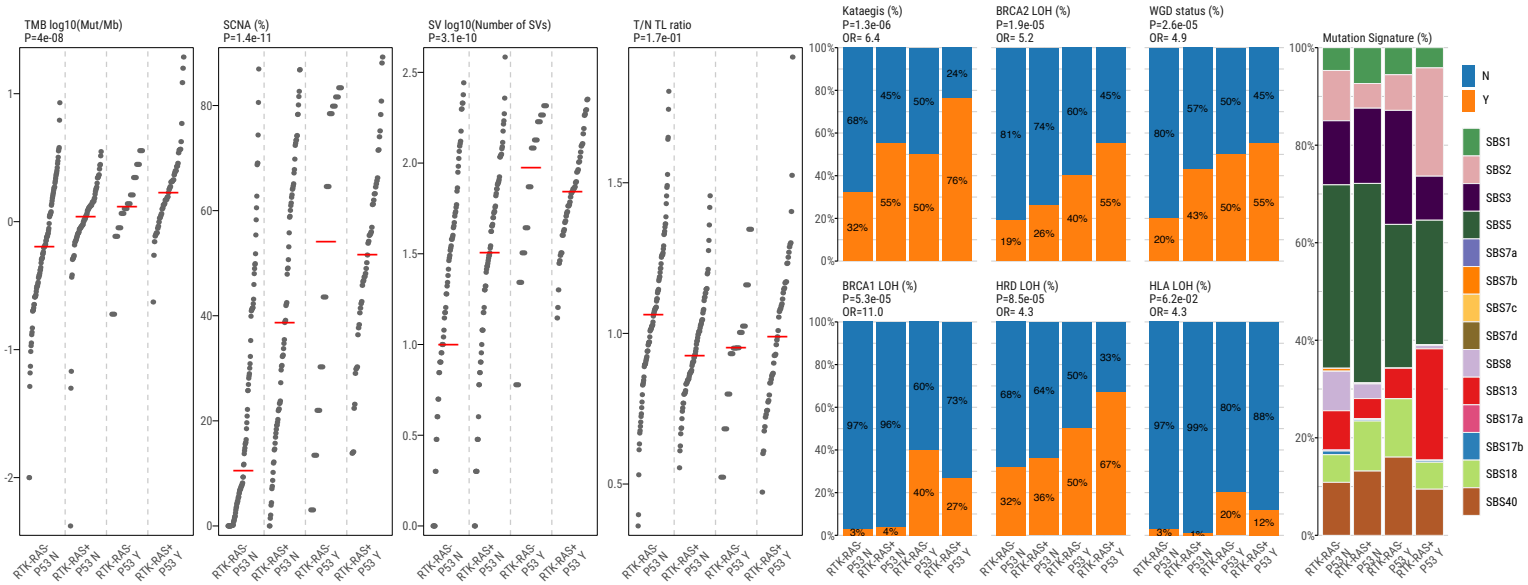
Supplementary Fig. 2





Supplementary Fig. 2 Association between tumor mutational burden (TMB) and clinical features. **a**, A multivariate linear regression (two-sided) for association testing between tumor mutational burden (TMB) and clinical features including histology, gender, tumor stage, and age at diagnosis (n=232). No multiple-testing correction applied. The blue dot shows the coefficient and the black solid line indicates the standard error. **b**, no association found between tumor purity and TMB. Gray error band indicates the 95% CI for coefficient of linear regression.

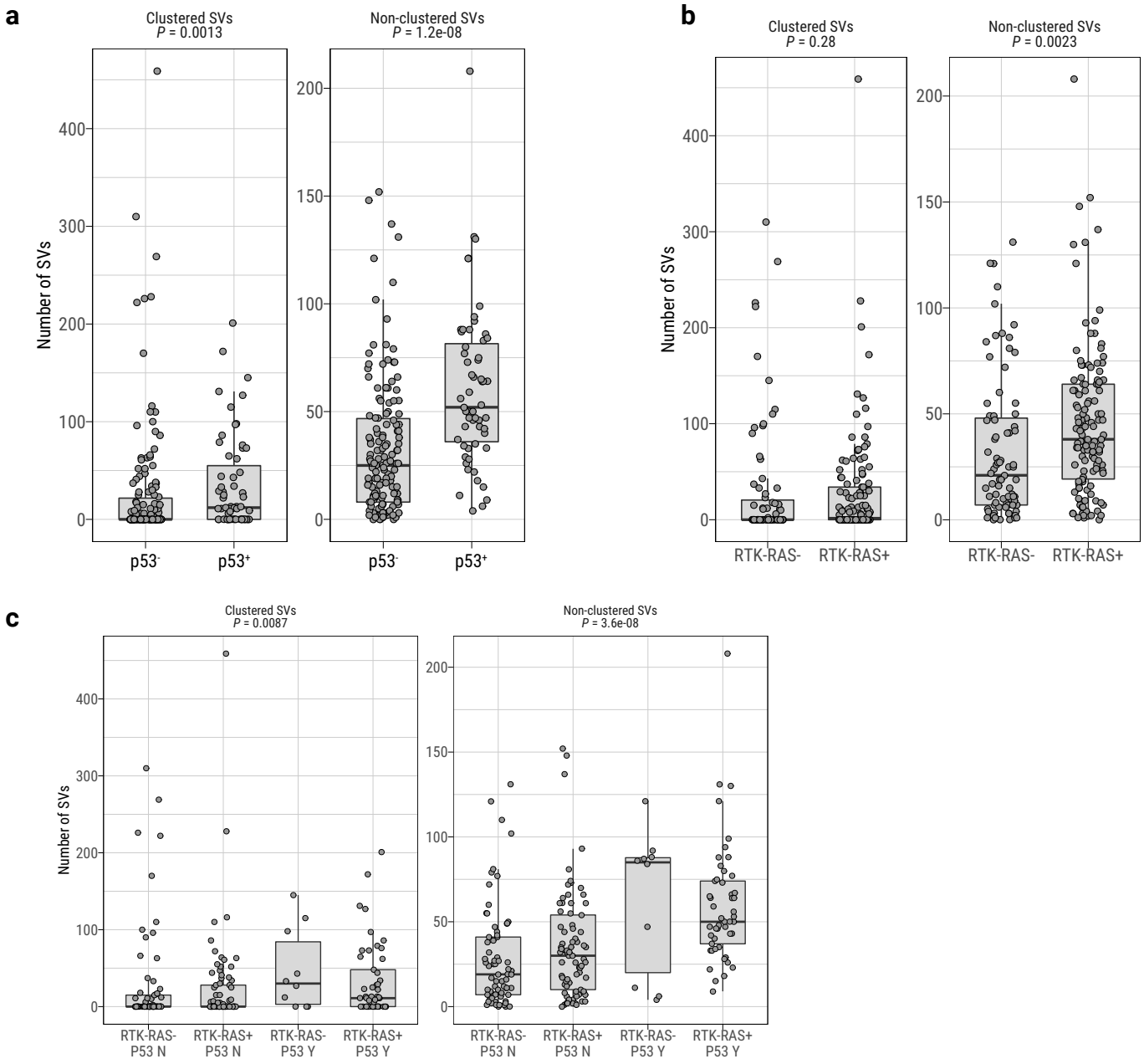
Supplementary Fig. 3



Supplementary Fig. 3 Comparison of genomic features combining TP53 and RTK-RAS status.

Left four panels: tumor mutation burden, percentage of genome with SCNAs, SV burden and T/N TL ratio. *P*-values are calculated using the two-sided Mann-Whitney U test. Middle six panels (Y="with", N="without"): enrichments for Kataegis events, *BRCA2* LOH, WGD events, *BRCA1* LOH, HRD LOH, and HLA LOH. *P*-values and *OR* are calculated using Fisher's exact test (two-sided). Right panel: Contributions of each SBS signature. All statistical analyses were performed between "RTK-RAS$^-$ & TP53 N" and "RTK-RAS$^+$ & TP53 Y".

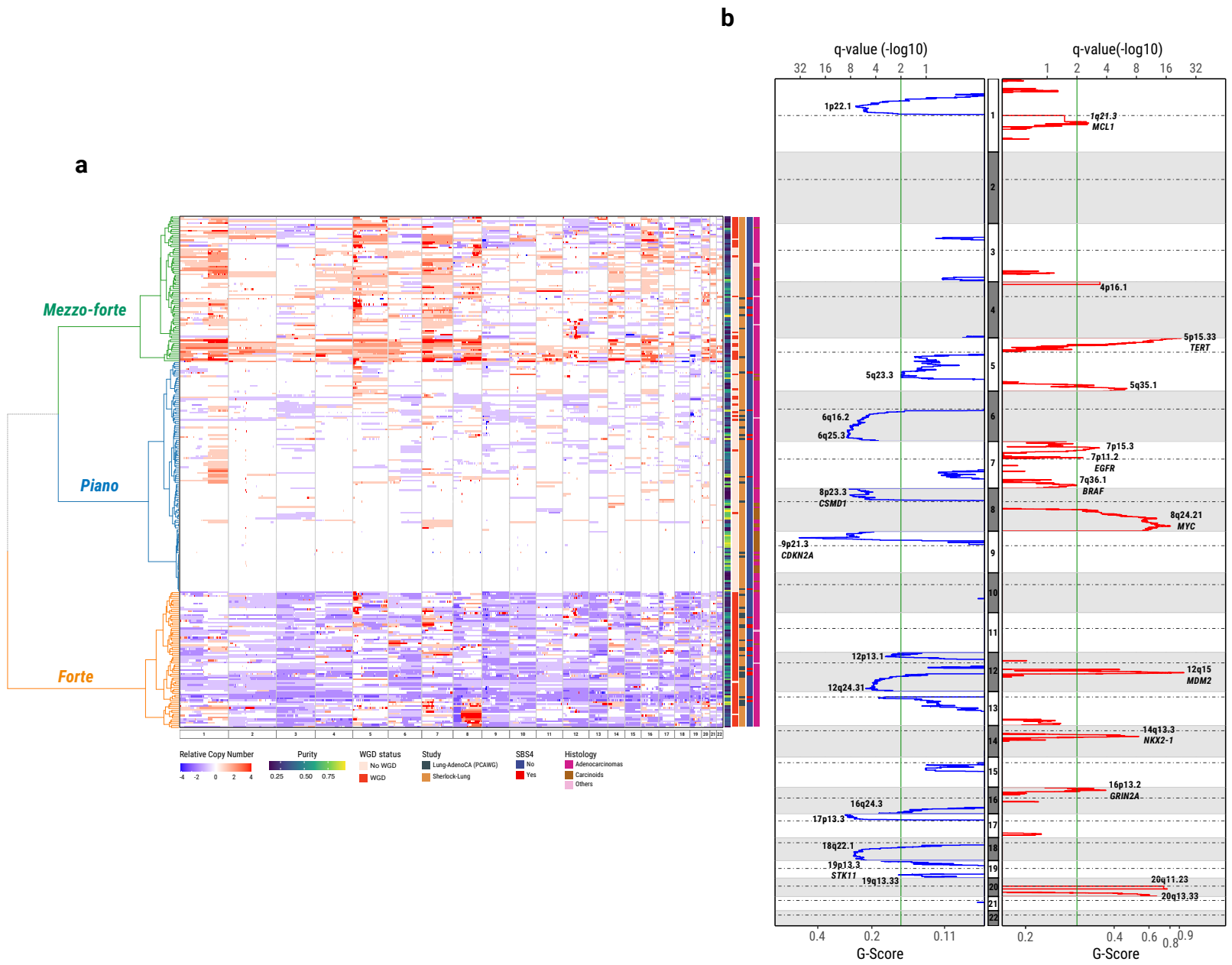
Supplementary Fig. 4



Supplementary Fig. 4 Enrichment of non-clustered SVs in tumors with somatic alteration in TP53 or RTK-RAS pathway.

a, Comparison of the total number of clustered SVs (left) and non-clustered SVs (right) between TP53-proficient (group N) and TP53-deficient (group Y) tumors. **b**, Comparison of the total number of clustered SVs (left) and non-clustered SVs (right) between RTK-RAS negative and positive tumors. **c**, Comparison of the total number of clustered SVs (left) and non-clustered SVs (right) among subtypes combining TP53 and RTK-RAS status. *P*-values are calculated using the two-sided Mann-Whitney U test between "RTK-RAS$^-$ & TP53 N" and "RTK-RAS$^+$ & TP53 Y". For box plots from **a** to **c**, center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

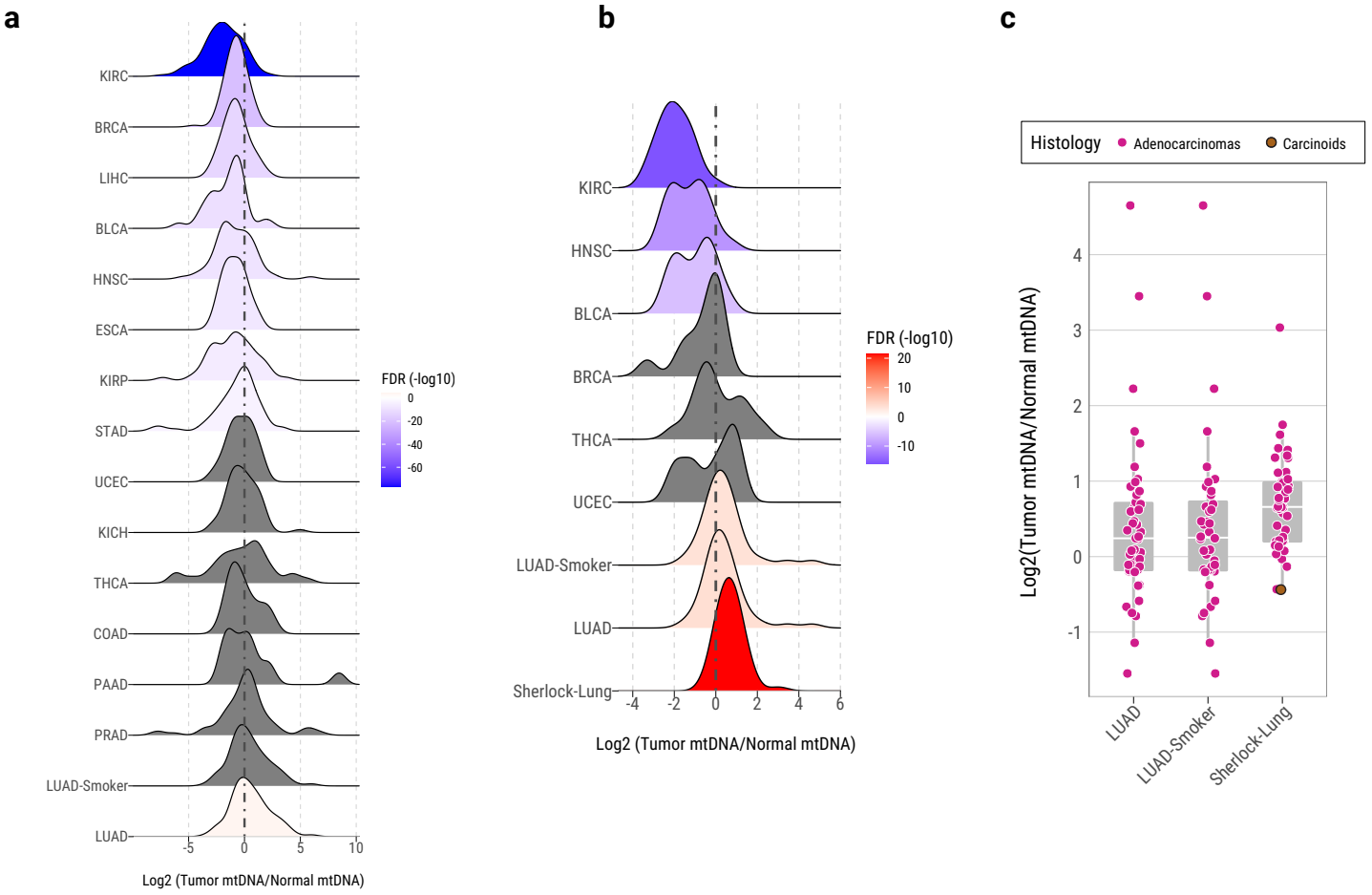
Supplementary Fig. 5



Supplementary Fig. 5 Somatic copy number alterations analysis in both arm- and focal-level.

a, The landscape of somatic copy number alterations combining 38 TCGA LUAD samples from PCAWG with the Sherlock-*Lung* samples. Left panel presents unsupervised clustering of arm-level SCNA events: *forte*, *mezzo-forte* and *piano*. Relative copy number is calculated as: total copy number - ploidy (non-WGD=2 and WGD=4). Samples in rows are annotated by tumor purity, WGD status, study, presence of SBS4 signatures and tumor histological type. **b**, GISTIC analysis of focal amplifications (red) and deletions (blue) in Sherlock-*Lung*. Significantly amplified and deleted (*q*-value < 0.01, green line) regions are annotated with cytoband or gene names.

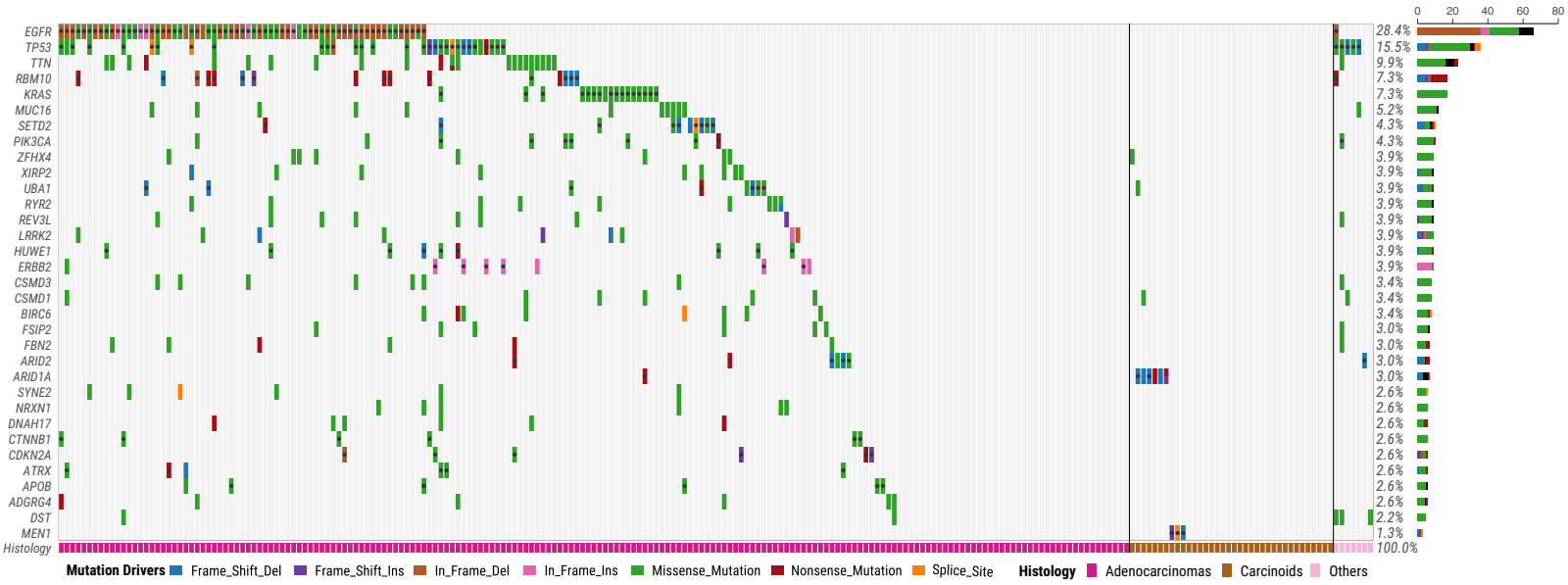
Supplementary Fig. 6



Supplementary Fig. 6 Mitochondrial DNA copy number accumulation in tumor compared to normal tissue.

a, Accumulation (light red) or depletion (blue) of mtDNA in tumor samples relative to the adjacent normal tissues based on WES data from the TCGA study. Normalized density plots illustrate $\log_2$(Tumor mtDNA/ Normal mtDNA). Each row represents a tumor type. Statistical significance of trends is assessed using Wilcoxon sign rank test (two-sided), and *P*-values are corrected using the Benjamini-Hochberg procedure.
b, The accumulation (red) or depletion (blue) of mtDNA in tumor samples relative to adjacent normal tissue based on the WGS data from the TCGA and Sherlock-*Lung* studies. **c**, Comparison of $\log_2$(Tumor mtDNA/ Normal mtDNA) among TCGA LUAD, TCGA LUAD smokers only and Sherlock-*Lung*. Center lines in box plots show the medians; gray box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

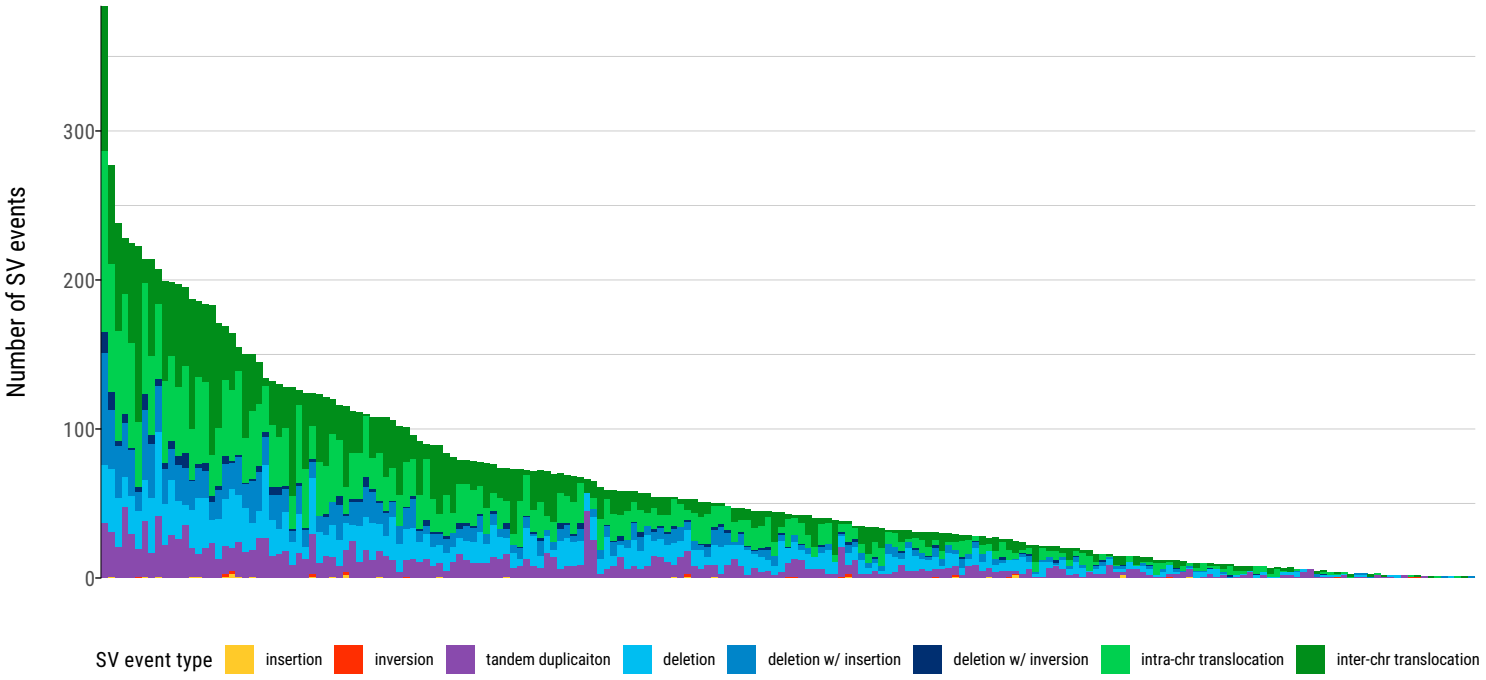
Supplementary Fig. 7



Supplementary Fig. 7 Landscape of most frequent somatic non-synonymous mutations.

Numbers and barplot on the right show the overall frequencies and number of mutated samples for each gene. The bottom bar shows tumor histological subtypes.

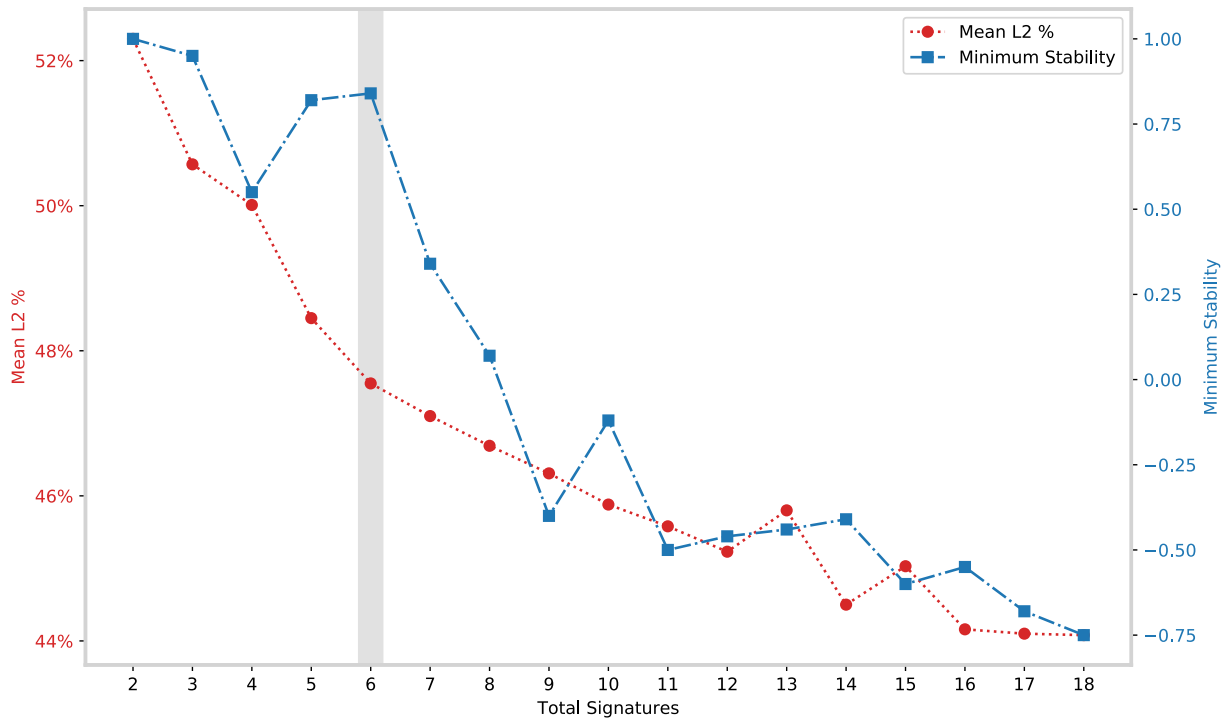
Supplementary Fig. 8



Supplementary Fig. 8 Frequencies of types of somatic SVs in each patient.

The y-axis shows the number of events. The colors indicate different SV types.

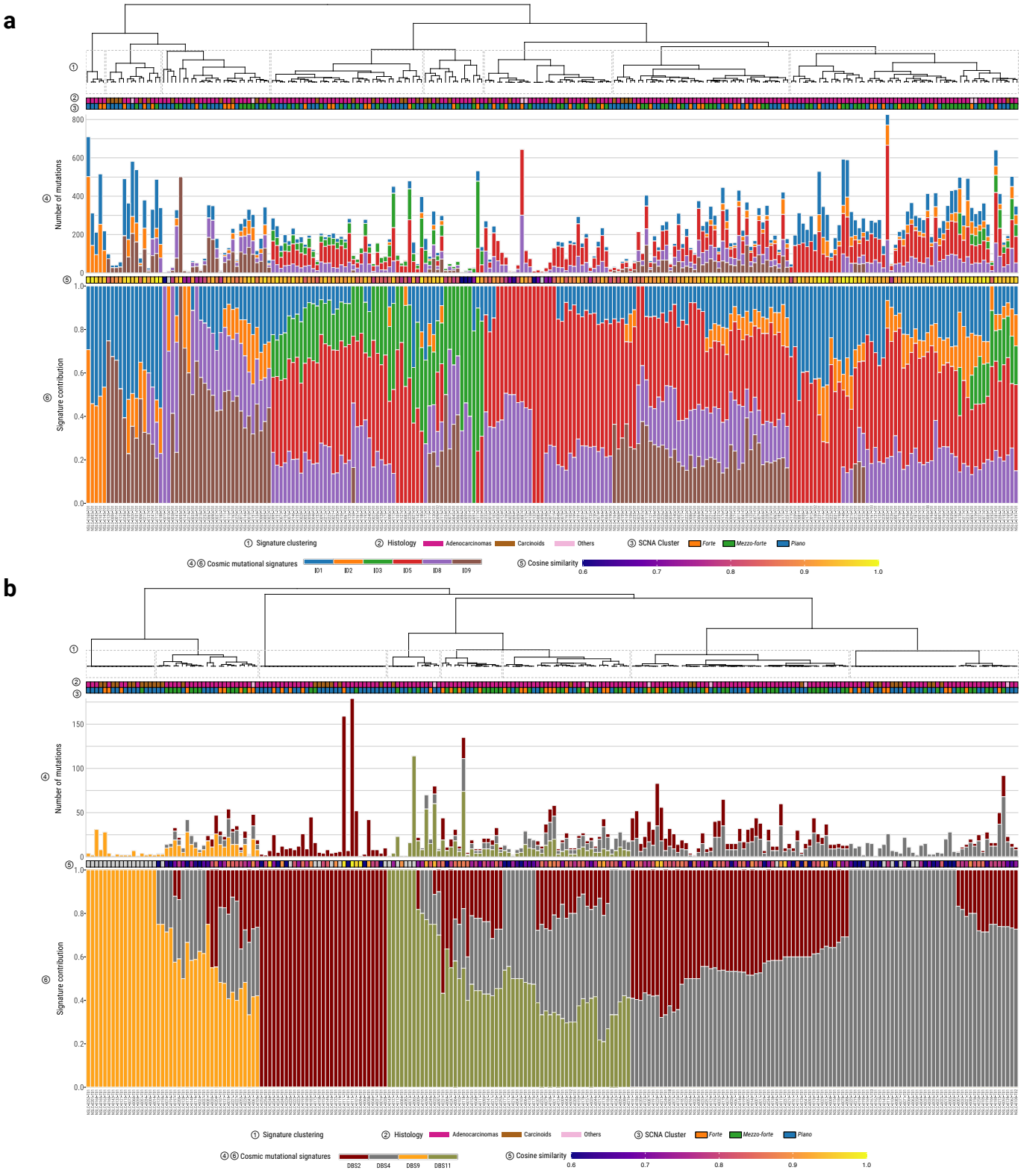
Supplementary Fig. 9



Supplementary Fig. 9 *De-novo* extraction of mutational signatures.

Assessment of signature decomposition stability during *de-novo* SBS signature extraction using SBS-1536 profile. Mean L2 (red) is calculated as the square root of the sum of the squared vector values, and minimum stabilities (blue) are calculated as average silhouette coefficient. The gray bar indicates the final selection for the total number of *de-novo* signatures.

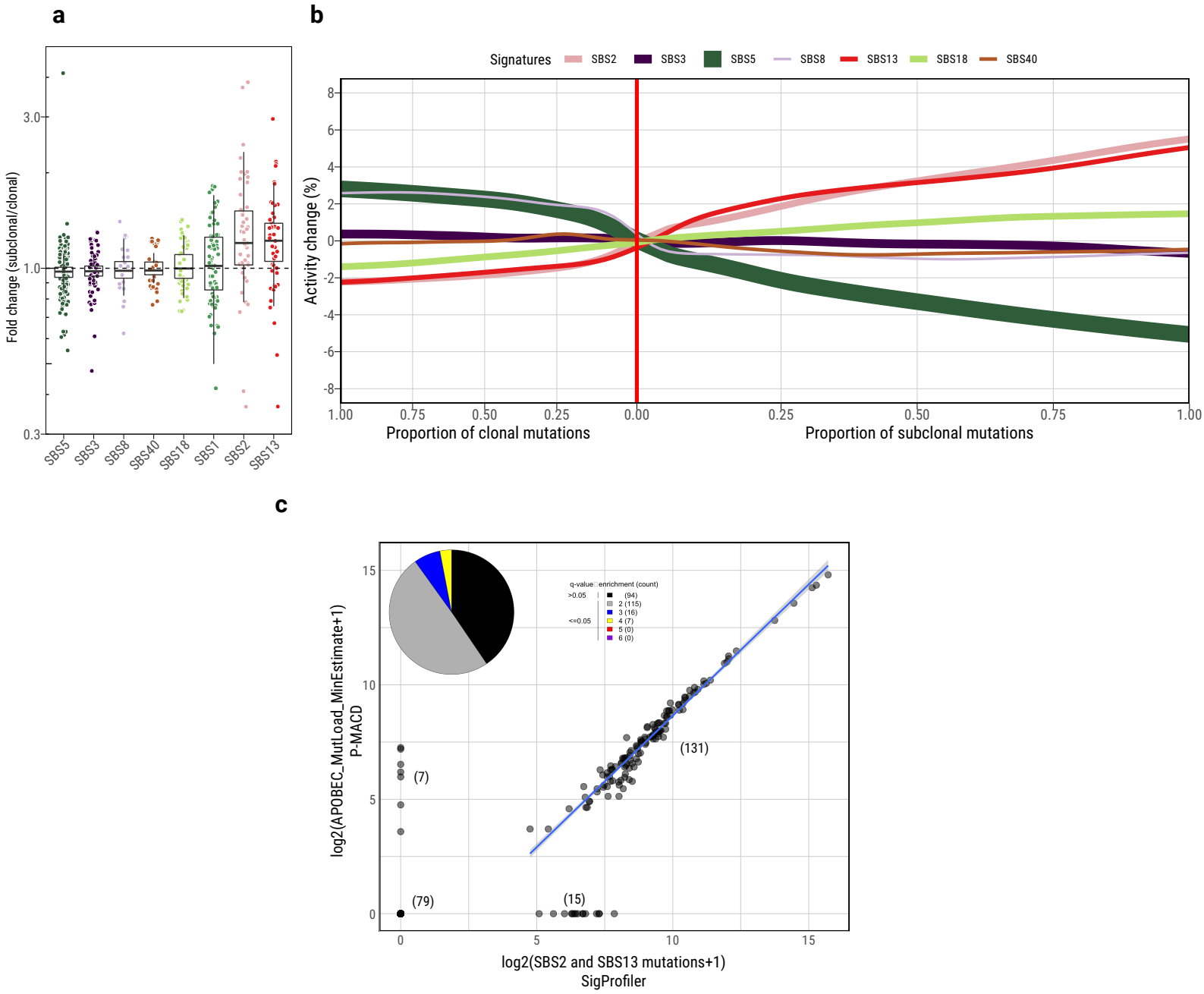
Supplementary Fig. 10



Supplementary Fig. 10 Mutational signature profile of indels (ID) and doublet substitutions (DBS).

a, Mutational signature profile of indels across 232 Sherlock-*Lung* samples. Panels from top to bottom: 1) Unsupervised clustering based on the proportion of ID signatures decomposed by SigProfilerExtractor; 2) Histology; 3) SCNA Cluster; 4) Total number of ID assigned to each ID signature; 5) Cosine similarity between the original mutational profile and signature decomposition results; 6) Proportions of ID mutational signatures. **b**, Mutational signature profile of doublet substitutions across 232 Sherlock-*Lung* samples. Panels from top to bottom: 1) Unsupervised clustering based on the proportion of DBS signatures decomposed by SigProfilerExtractor; 2) Histology; 3) SCNA cluster; 4) Total number of DBS assigned to each DBS signature; 5) Cosine similarity between the original mutational profile and signature decomposition results; 6) Proportions of DBS mutational signatures.

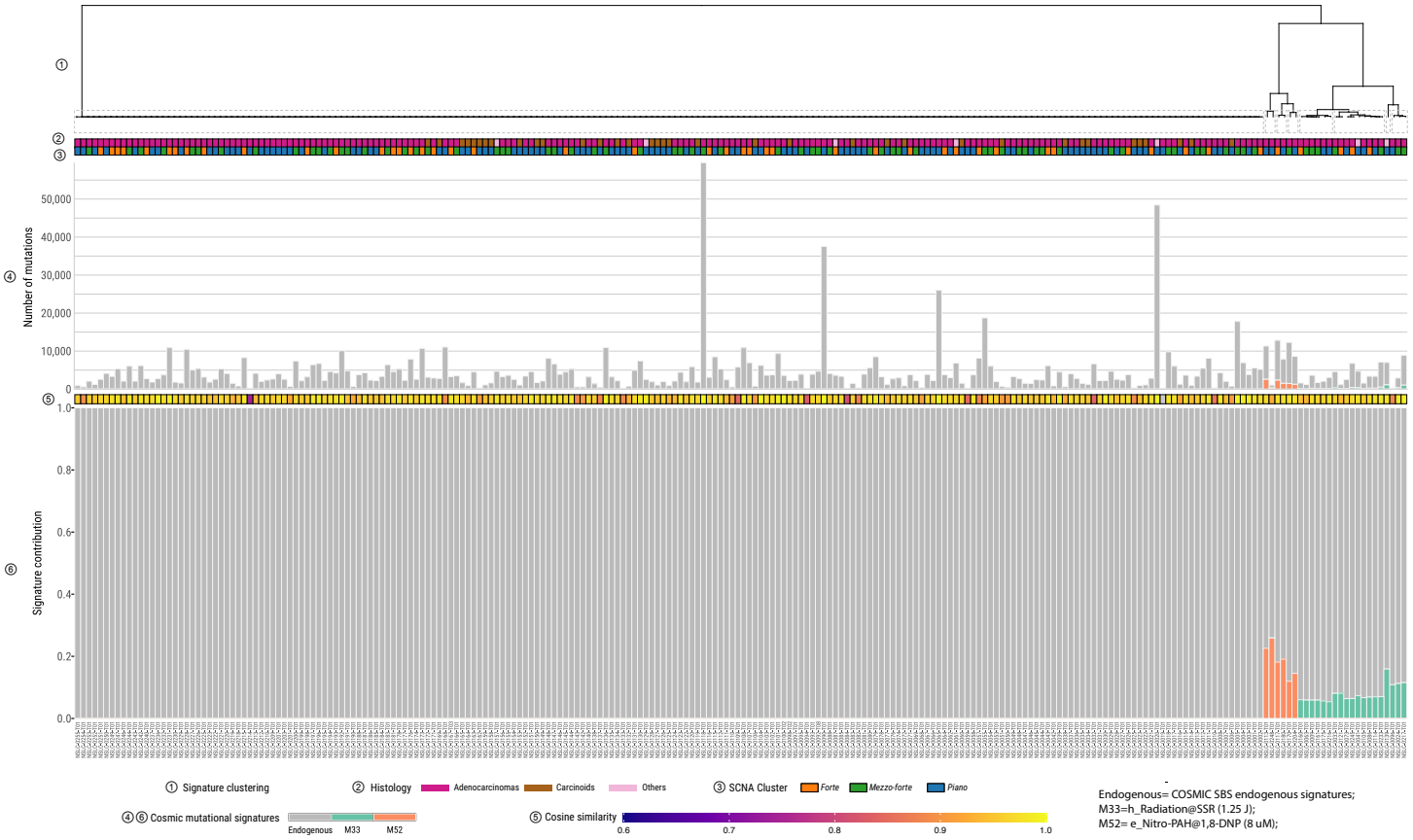
Supplementary Fig. 11



Supplementary Fig. 11 Dynamic mutational process during clonal and subclonal tumor evolution.

a, Fold changes between subclonal and clonal periods are shown in samples with measurable changes in mutation spectra based on mutational signatures. For box plots, center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. **b**, Reconstruction of the trajectory of signatures activity. Each horizontal line is a trajectory of changes of one SBS signature activity relative to its activity in the most recent common ancestor (MRCA) (y-axis) as a function of decreasing cancer cell fraction (x-axis). The width of the lines indicates the number of samples of estimated activity changes. The vertical red line depicts the time point of the MRCA. **c**, Comparison of the APOBEC mutation loads identified by P-MACD (APOBEC_mutLoad_MinEstimate) and SigProfilerExtractor (SBS2 and SBS13). The pie chart with legend indicates the fold-enrichment and significance of APOBEC mutagenesis signatures over the expected occurrence for random mutagenesis estimated by P-MACD. Numbers in brackets show the number of samples in each subset on the scatter plot.

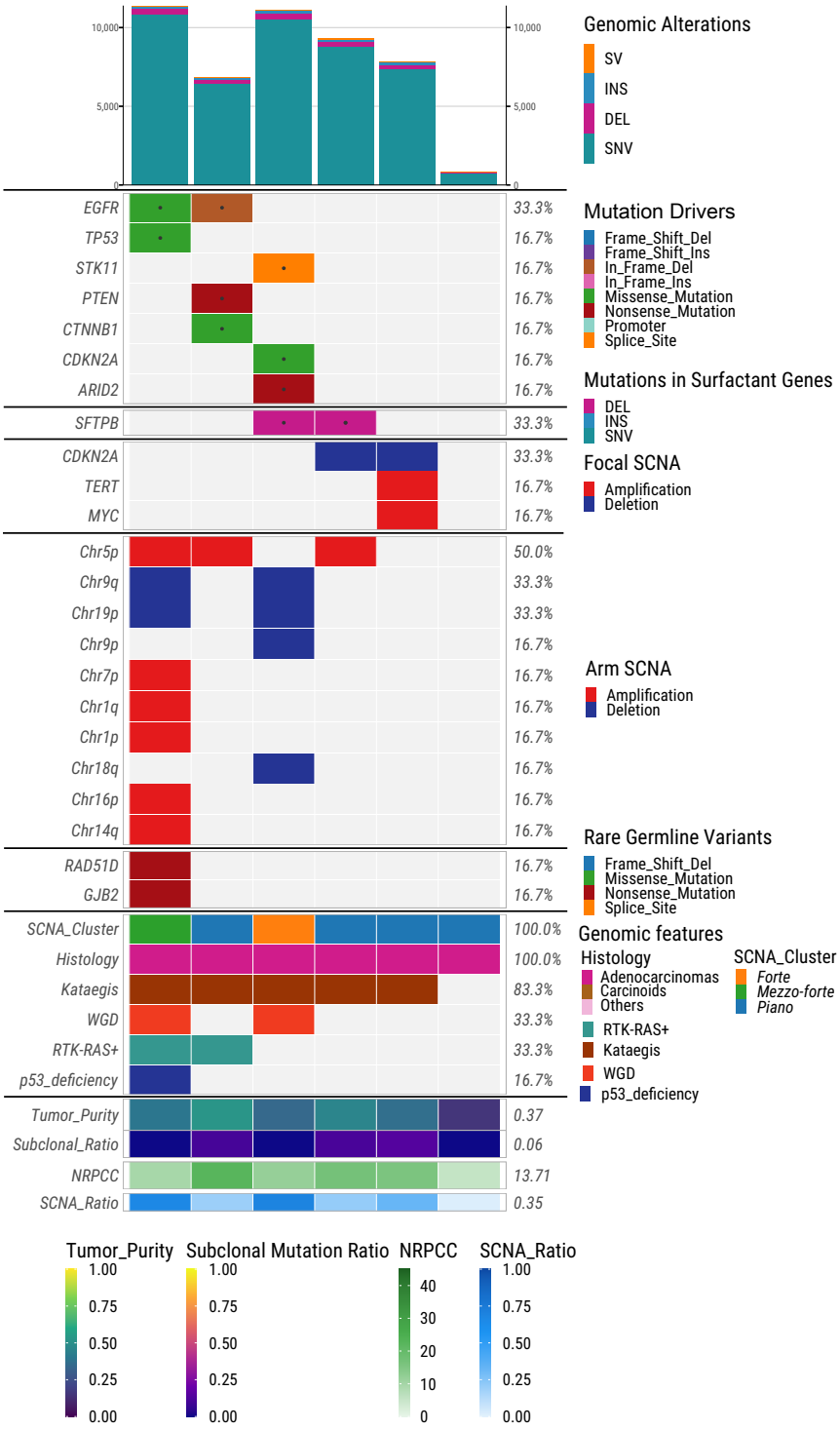
Supplementary Fig. 12



Supplementary Fig. 12 SBS mutational signature profile using combined endogenous and environmental signatures.

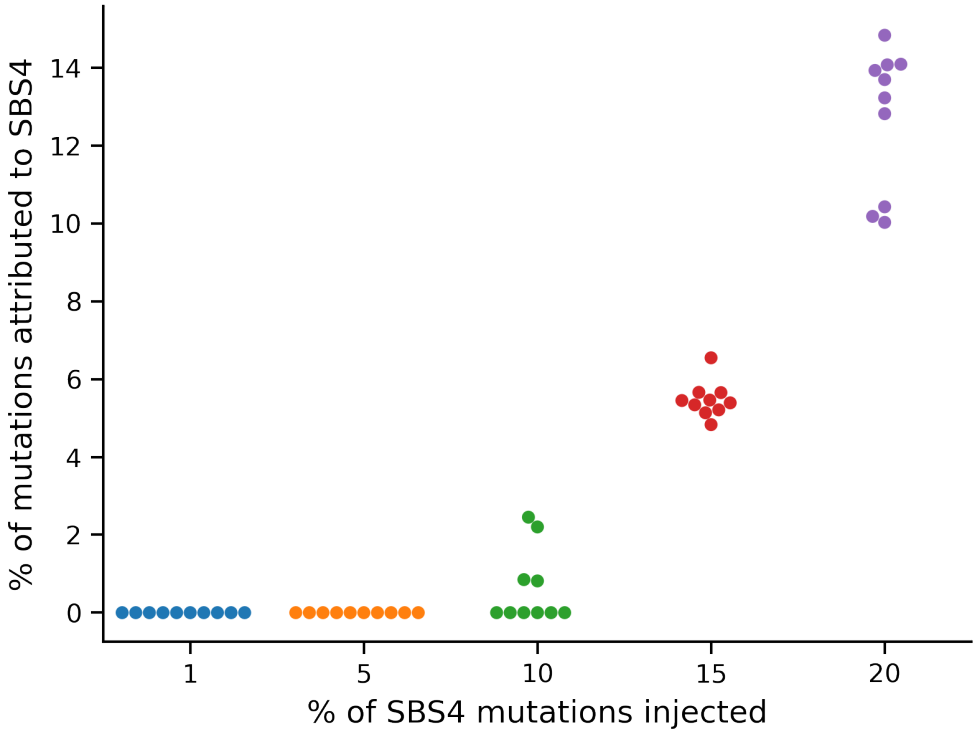
Panels from top to bottom: 1) Unsupervised clustering based on the proportion of SBS signatures decomposed by SigProfiler; 2) Histology; 3) SCNA cluster; 4) Total number of SBS assigned to each signature; 5) Cosine similarity between the original mutational profile and signature decomposition results; 6) Proportions of SBS assigned to each mutational signatures.

Supplementary Fig. 13



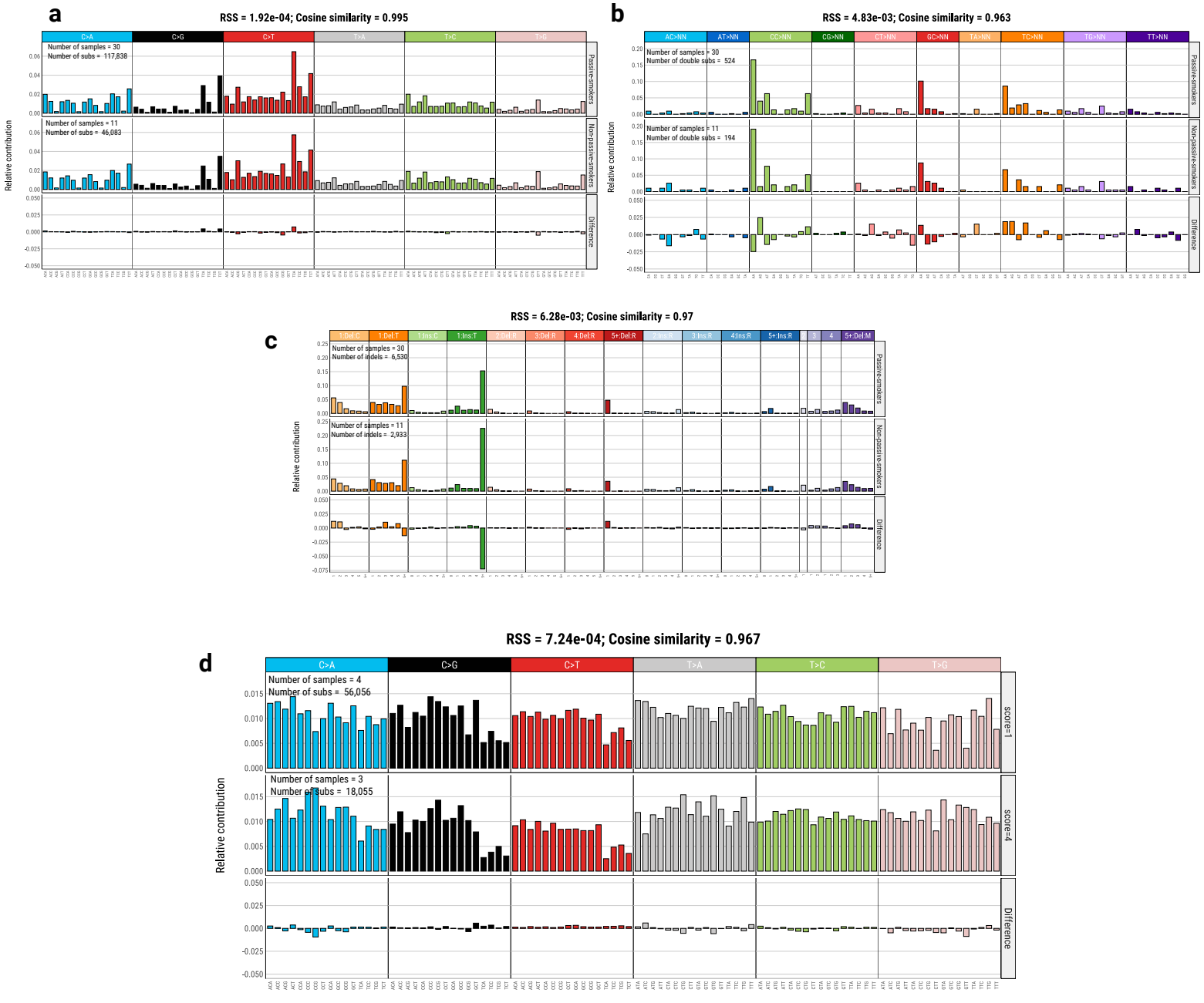
Supplementary Fig. 13 Genomic characteristics of 6 tumors with potential signature of diesel fumes.

Supplementary Fig. 14



Supplementary Fig. 14 Evaluating the resolution for detecting signature SBS4 from passive smokers. Each dot represents the results from a simulation in which signature SBS4 was injected at a specific average percentage of mutations in all passive smokers and a complete extraction of mutational signatures was performed. The x-axis reflects the average percentage of mutations injected in the passive smokers, while the y-axis corresponds to the average detected percentage of mutations in the passive smokers. When the y-axis is 0, SBS4 is not detected in the extraction of mutational signatures.The assignment algorithm uses clock-like signatures SBS1 and SBS5 as a background model in each sample and adds additional signatures only if they significantly improve the assignment of the mutational pattern within each sample. Due to similarity of their T>C profiles, mutations from SBS4 may be misassigned to SBS5 resulting in higher levels of SBS5 and lower levels of SBS4. This may explain the lower number of mutations attributed to SBS4 (y-axis) than the number of synthetically injected SBS4 mutations (x-axis).

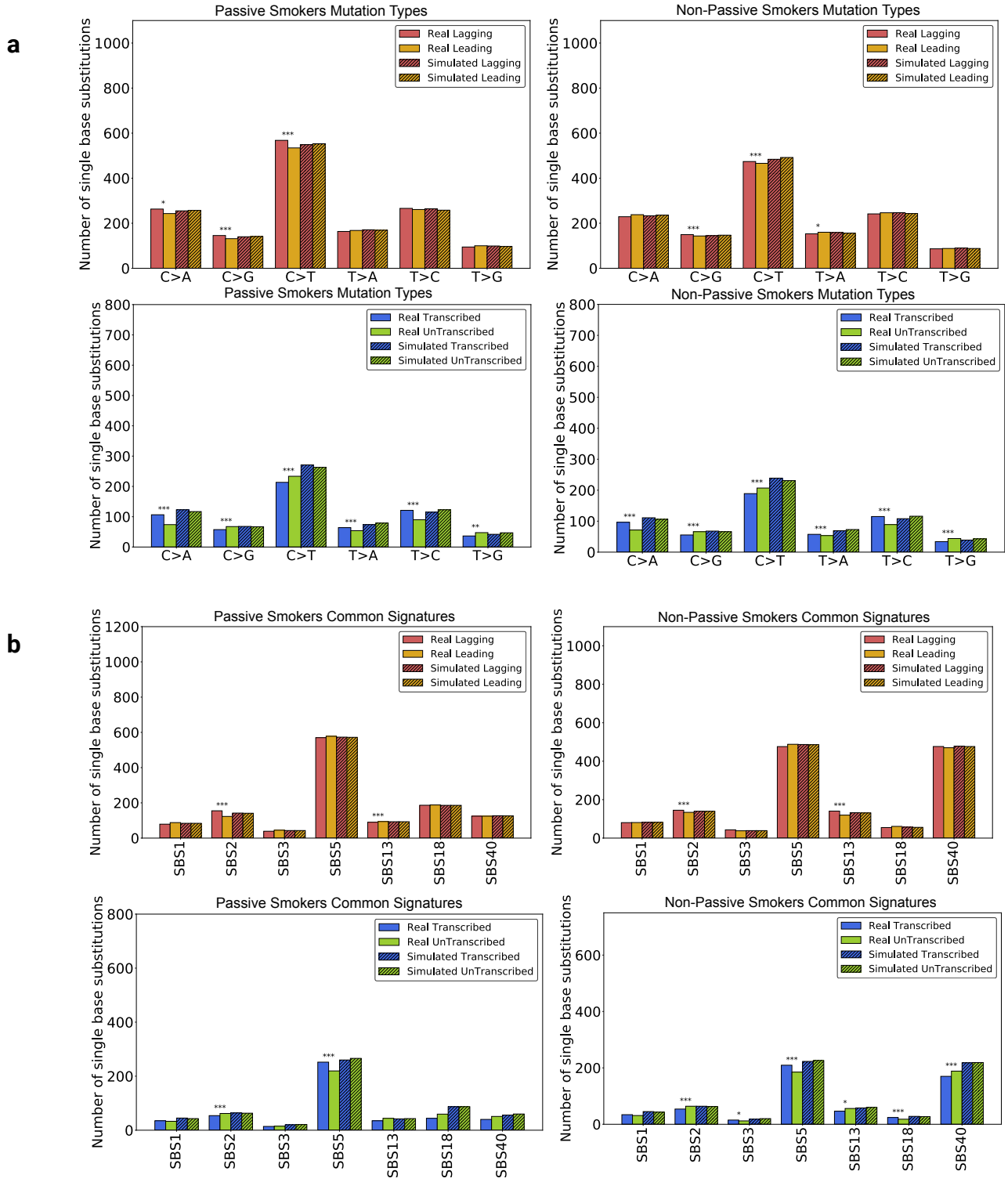
Supplementary Fig. 15



Supplementary Fig. 15 Mutational spectra comparison between passive-smokers and non-passive smokers.

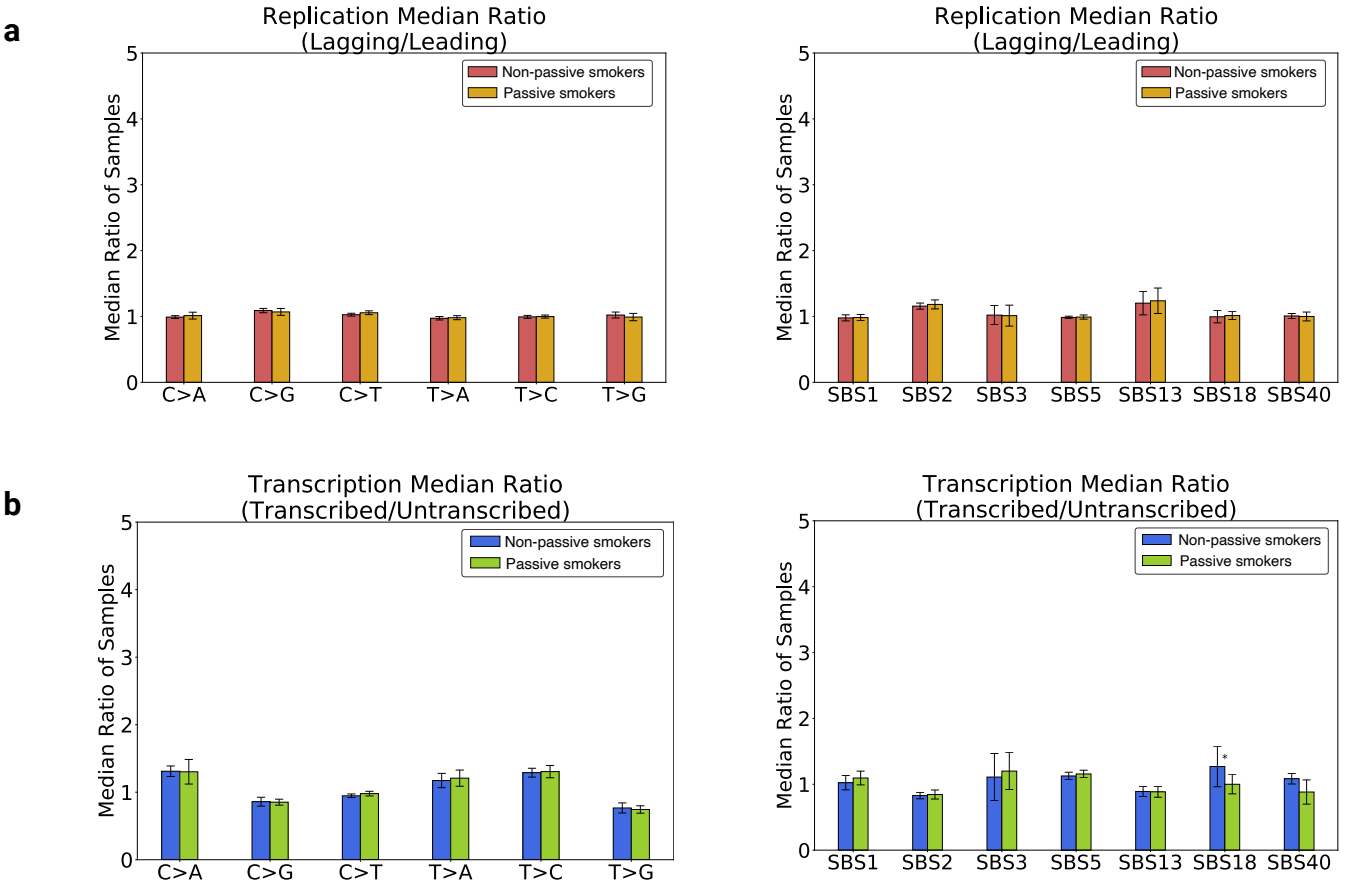
a-c) Mutational spectra comparison of SBS (**a**), DBS (**b**) and ID (**c**) between LCINS patients with passive smoking (n=30), and non-passive smoking (n=11) from EAGLE and Yale studies. **d**, Mutational spectra comparison of SBS between LCINS patients with passive smoking exposure score 1 (n=4), and score 4 (n=3) from EAGLE study.

Supplementary Fig. 16



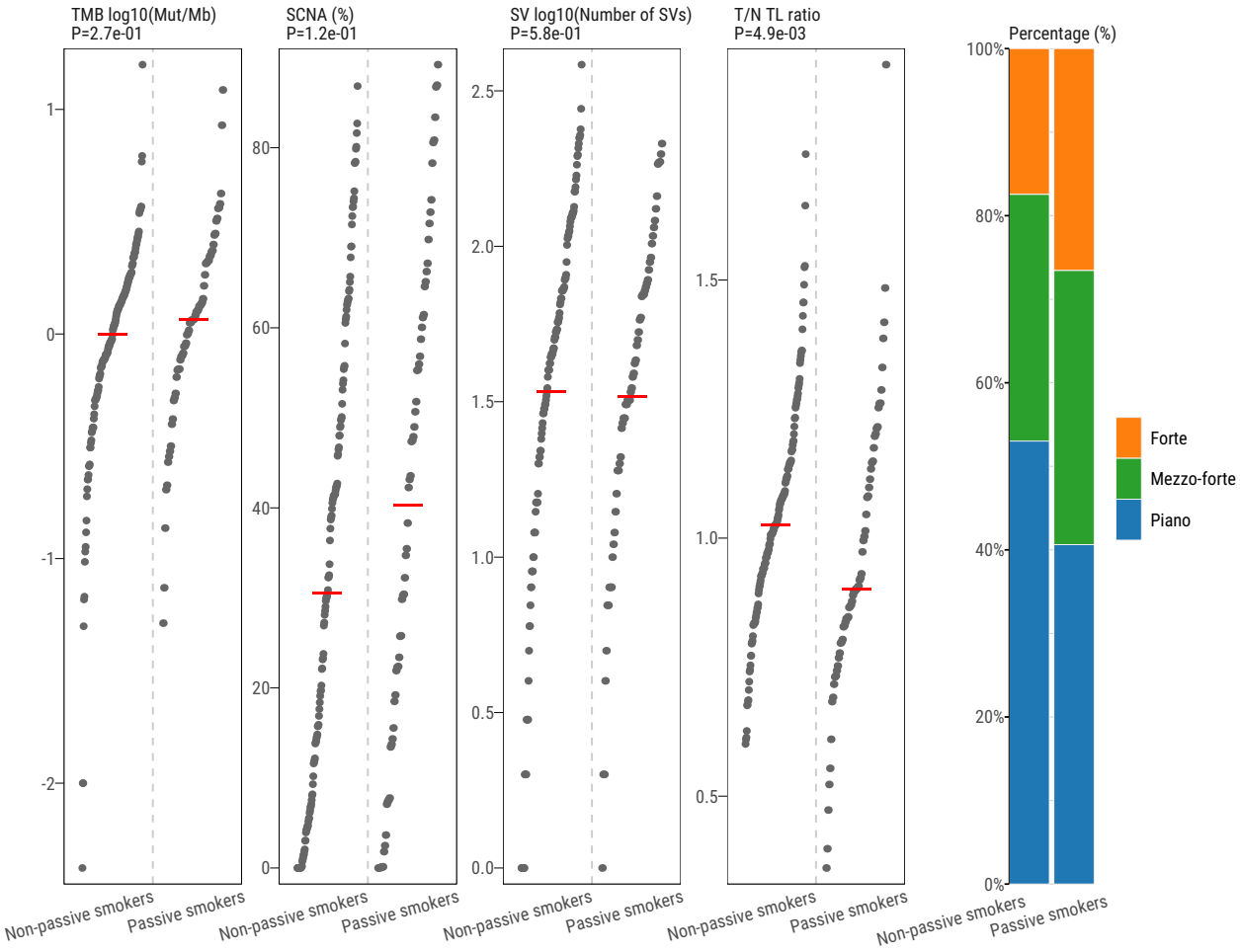
Supplementary Fig. 16 Strand bias comparisons between observed and simulated data in passive smokers (n=62) and non-passive smokers (n=148). *(**a**)* Median number of *observed* somatic mutations attributed to different mutation types are shown in solid red (when occurring on the lagging strand), solid yellow (leading strand), solid blue (transcribed strand), and solid green (untranscribed strand). Median number of *simulated* somatic mutations attributed to different mutation types are shown in stripped red (lagging strand), stripped yellow (leading strand), stripped blue (transcribed strand), and stripped green (untranscribed strand).(***b***) Median number of *observed* somatic mutations attributed to different mutational signatures are shown in solid red (when occurring on the lagging strand), solid yellow (leading strand), solid blue (transcribed strand), and solid green (untranscribed strand). Median number of *simulated* somatic mutations attributed to different mutational signatures are shown in stripped red (lagging strand), stripped yellow (leading strand), stripped blue (transcribed strand), and stripped green (untranscribed strand). In all panels, statistically significant differences are denoted with: *q-value<0.05; **q-value<0.01; and **q-value<0.001.

Supplementary Fig. 17



Supplementary Fig. 17 Comparison of transcription and replication strand asymmetry in passive smokers (n=62) and non-passive smokers (n=148). *(**a**)* For replication asymmetry, median ratios between mutations on the lagging strands and leading strands are shown. Mutation types are shown on the left while Mutational signatures are shown on the right. Red reflects strand asymmetry in non-passive smokers, while yellow reflects strand asymmetry in passive smokers. *(**b**)* For transcription asymmetry, median ratios between mutations on the transcribed strands and untranscribed strands are shown. Mutation types are shown on the left while mutational signatures are shown on the right. Blue reflects strand asymmetry in non-passive smokers, while green reflects strand asymmetry in passive smokers. In all cases, 95% confidence intervals are shown for each asymmetry ratio is shown. In all cases, 95% confidence intervals are shown for each asymmetry ratio is shown. In all cases, 95% confidence intervals are shown for each asymmetry ratio shown. In all panels, statistically significant differences are denoted with: *q-value<0.05; **q-value<0.01; and **q-value<0.001.

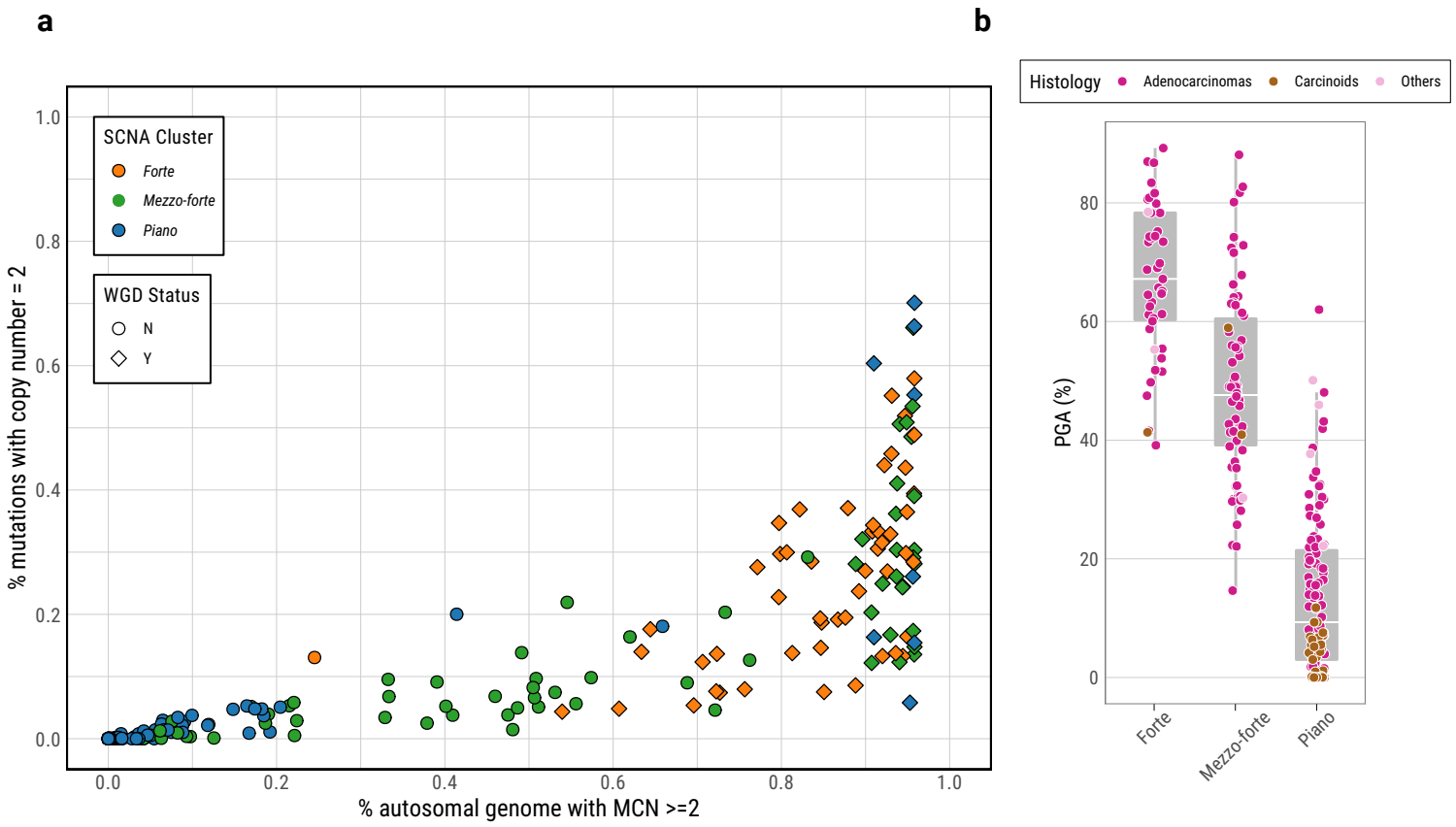
Supplementary Fig. 18



Supplementary Fig. 18 Genomic alteration differences in tumors from passive smokers vs non-passive smokers.

Left three panels: tumor mutational burden, percentage of genome with SCNAs, SV burden and T/N TL ratio. *P*-values are calculated using the two-sided Mann-Whitney U test; Right panel: percentage of each SCNA subtype. *P*-values and *OR* are calculated using Fisher's exact test (two-sided) after combining *forte* and *mezzo-forte* as one group.

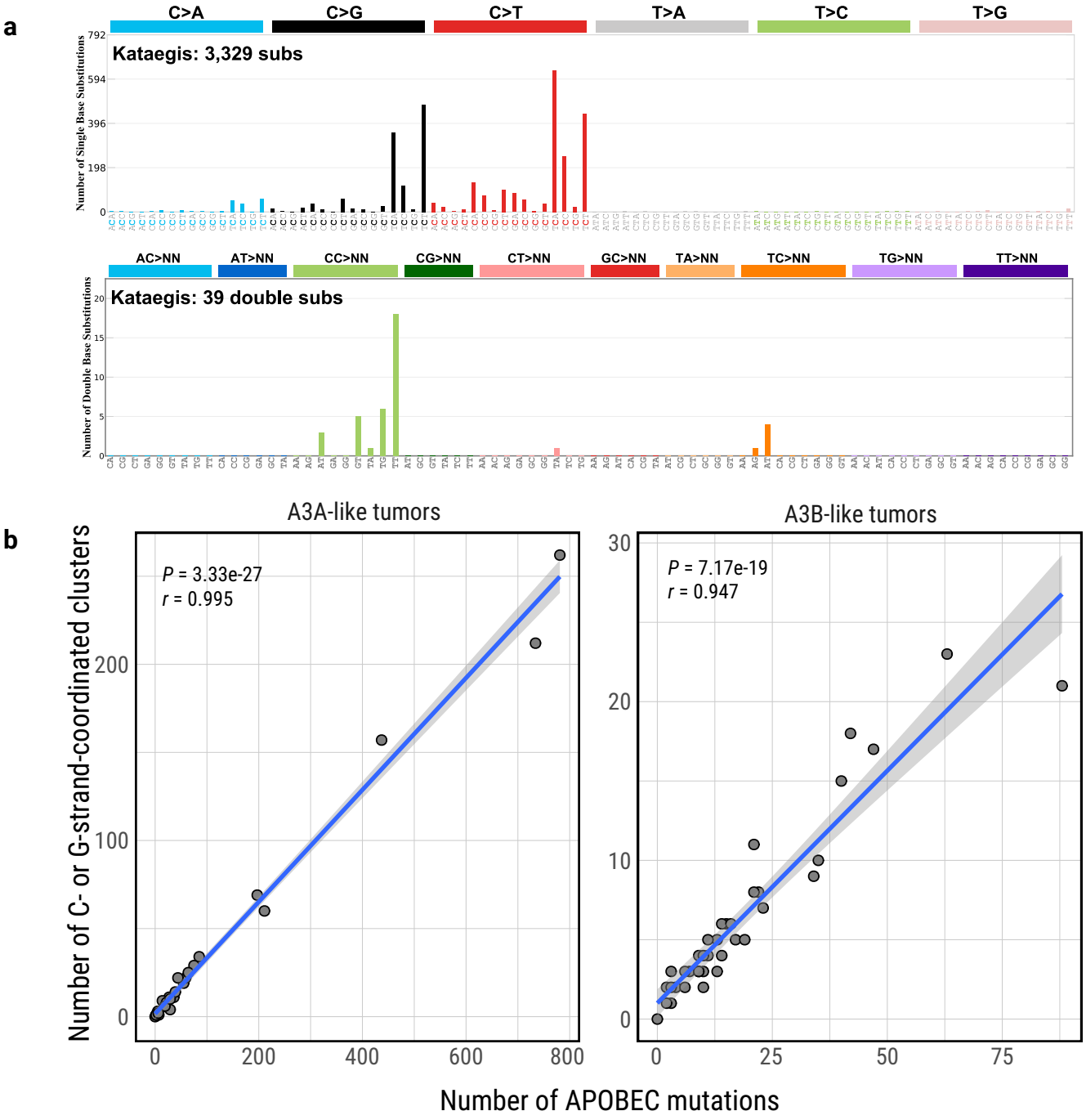
Supplementary Fig. 19



Supplementary Fig. 19 Genomic instability based on SCNAs.

a, Prevalence of whole genome doubling (WGD) in Sherlock-*Lung*. The x-axis shows the fraction of the autosomal genome with a major copy number (MCN) of two or greater. The y-axis shows the percentage of mutations with copy number=2. Each dot represents an individual (color-coded for the SCNA clusters; WGD status shown in shapes). **b**, Boxplot showing the percentage of genomic alterations (PGA) across different SCNA clusters. The white center lines show the medians; gray box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

Supplementary Fig. 20

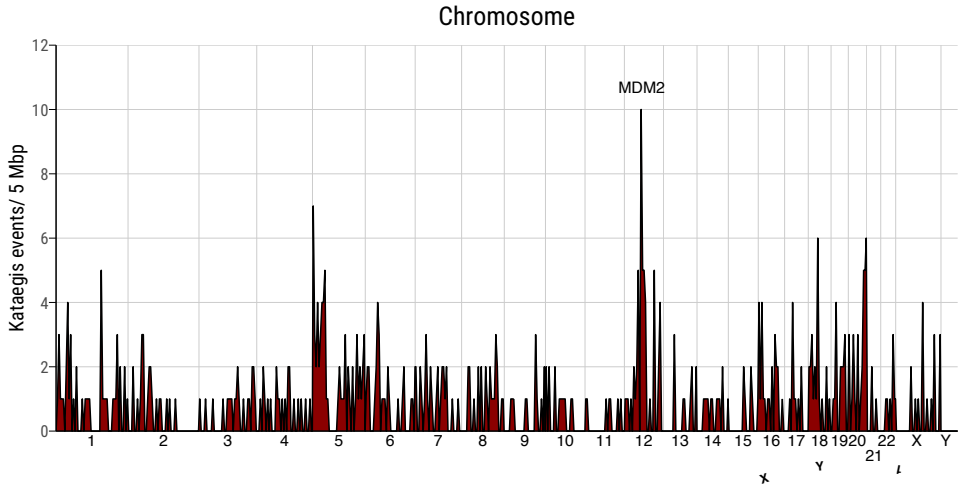


Supplementary Fig. 20 APOBEC emerged as the dominant mutational process in kataegis loci.

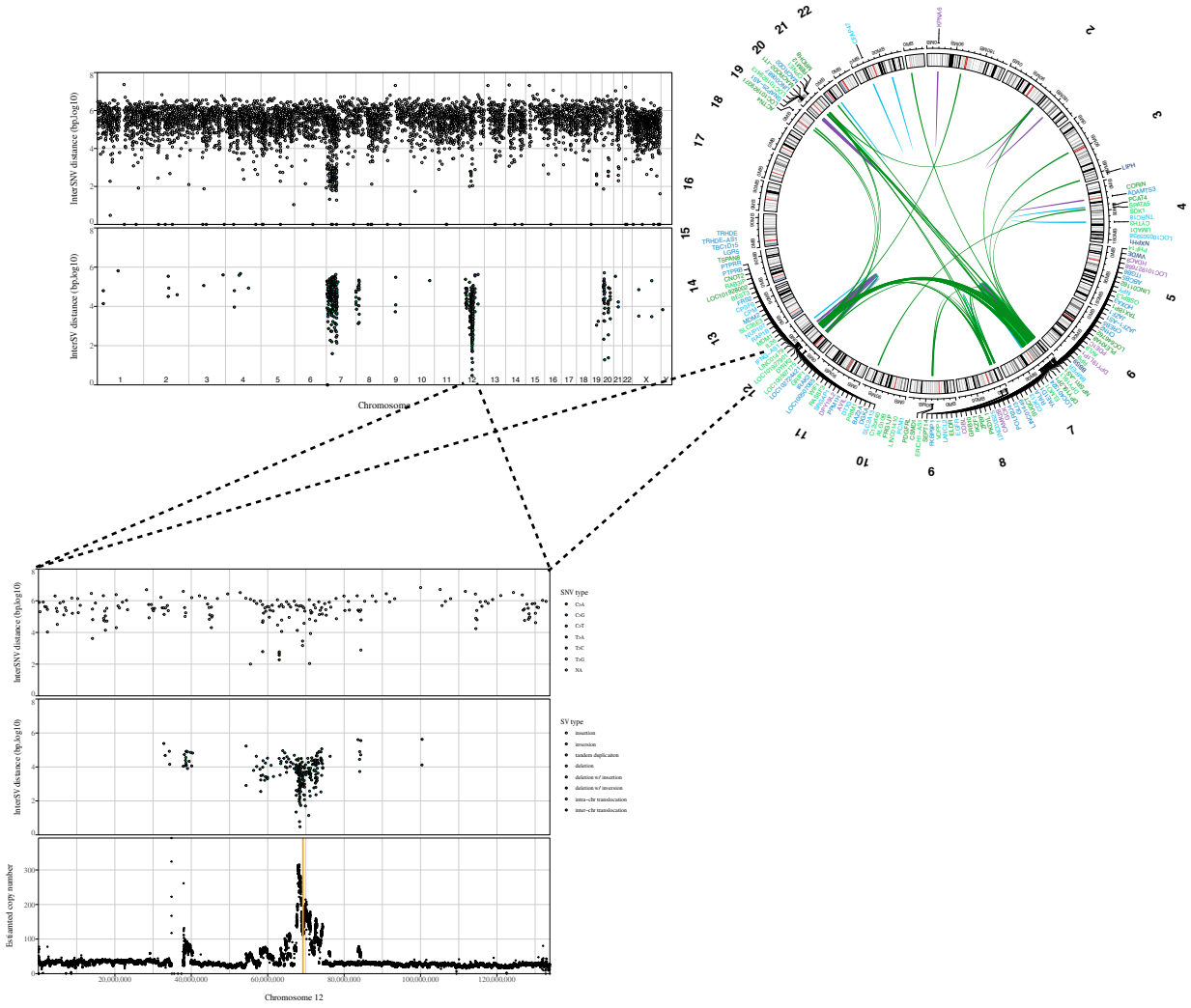
a, Spectra of single base substitutions and double base substitutions in samples within kataegis. **b**, Association between APOBEC mutation loads and the number of C- or G-strand-coordinated clusters related to kataegis in both A3A-like tumors and A3B-like tumors. Gray error band indicates the 95% CI for coefficient of linear regression.

Supplementary Fig. 21

a



b

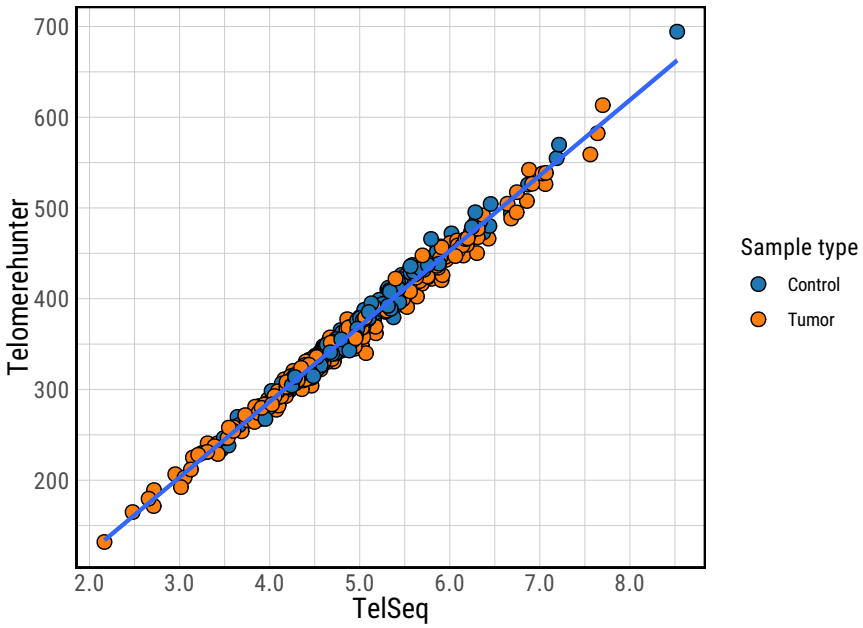


Supplementary Fig. 21 Co-localization between kataegis loci and *MDM2* amplification due to chromothripsis processes.

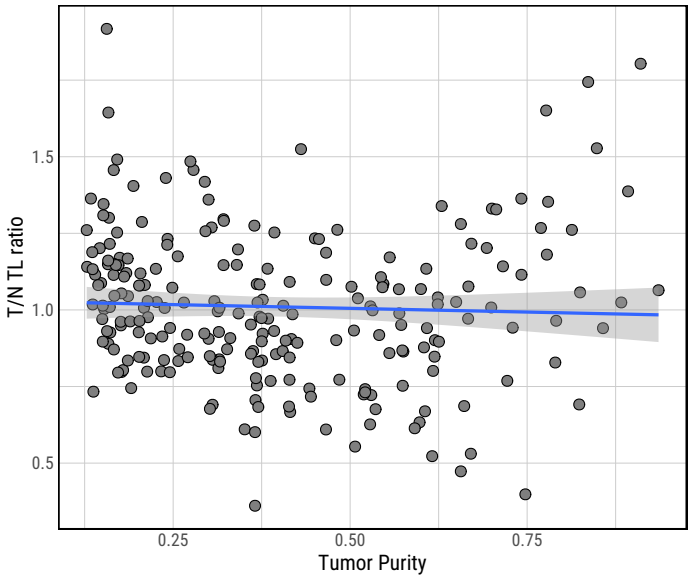
a, Frequency of kataegis events using a 5 Mbp window across the whole genome. **b**, Representative kataegis events in sample NSLC-0012-T01. Left top panels show rainfall plots of SNVs and SVs. Each dot represents an SNV or SV. SNVs and SVs are ordered by the genomic locations on the x-axis. The distances between two consecutive SNVs and SVs are plotted on the y-axis. The top right panel shows a circos plot of SV rearrangements. The bottom panels show an enlargement of rainfall plots on chromosome 12 with the corresponding estimated copy numbers. The genomic location of *MDM2* is indicated by the orange line.

Supplementary Fig. 22

a



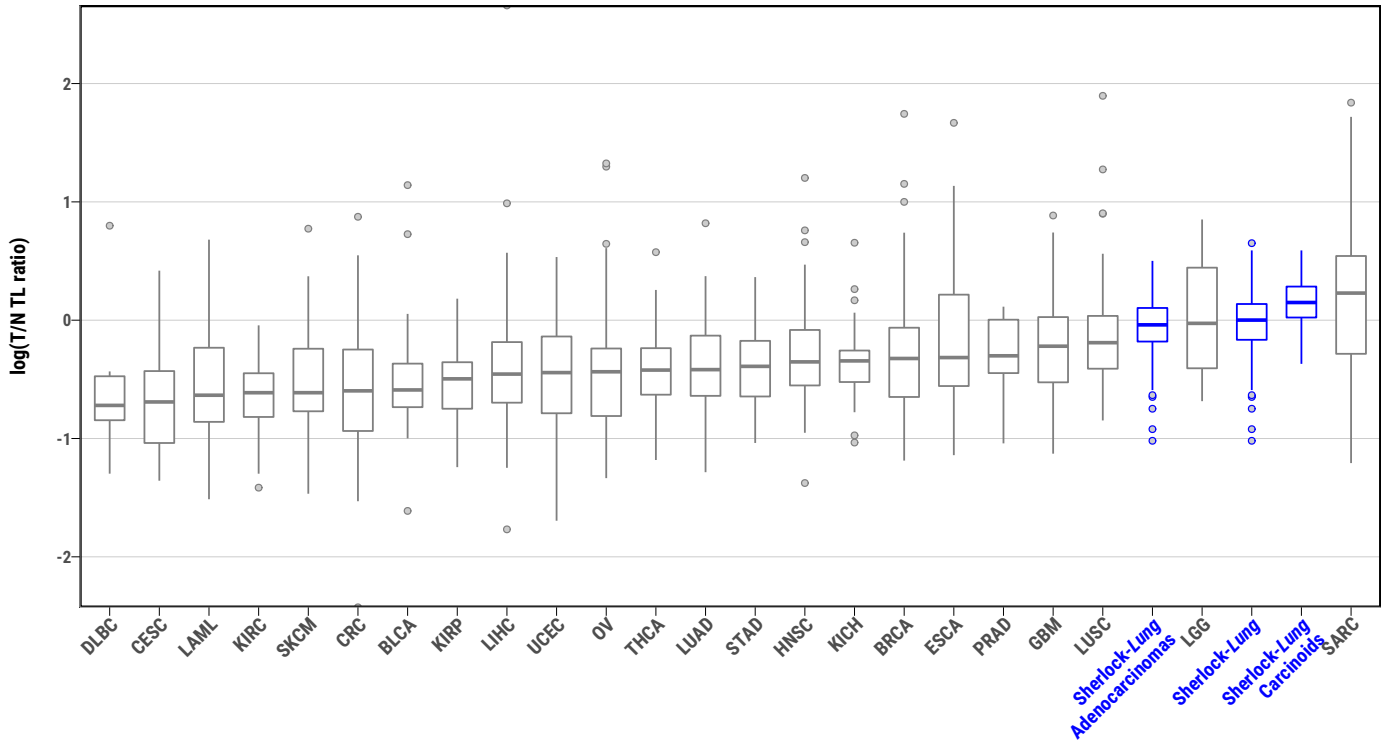
b



Supplementary Fig. 22 Telomere length estimation and association with tumor purity.

a, Correlation of telomere length estimations between TelSeq and TelomereHunter by sample type. **b**, Correlation between tumor purity and Tumor/Normal telomere length (T/N TL) ratio. Gray error band indicates the 95% CI for coefficient of linear regression.

Supplementary Fig. 23

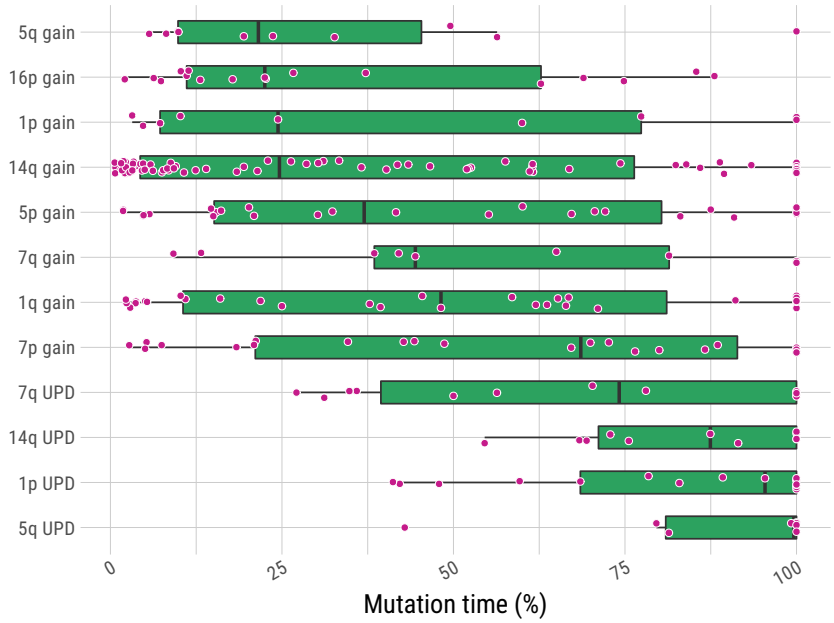


Supplementary Fig. 23 The comparison of T/N TL ratio between Sherlock-*Lung* and other cancer types in the TCGA study.

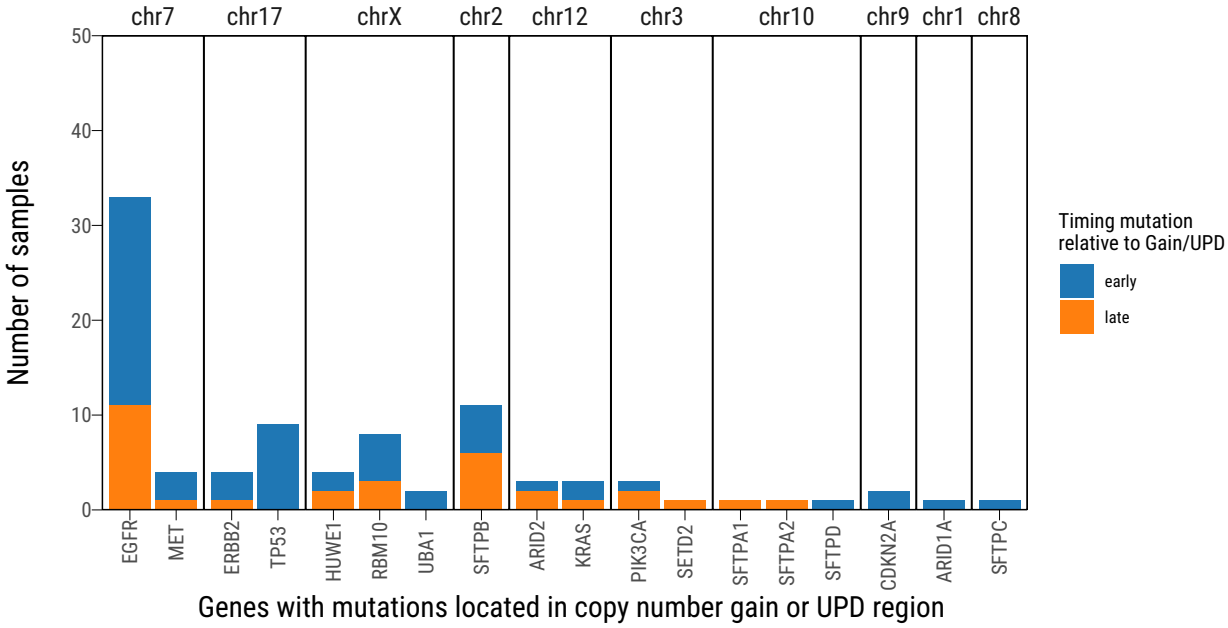
Boxplot of the T/N TL ratio (natural logarithms) in Sherlock-*lung* (blue) and TCGA Cancer Types (gray). The name abbreviation of the TCGA cancer types can be found in Supplementary Methods. Center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

Supplementary Fig. 24

a



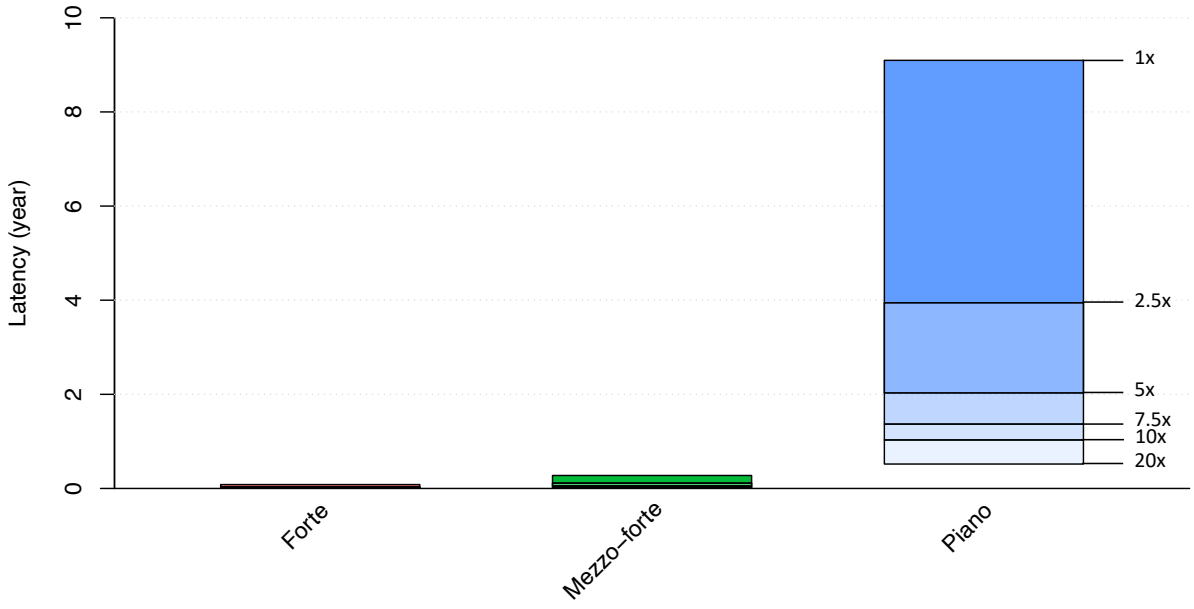
b



Supplementary Fig. 24 Estimation of the timing of duplications and driver mutations.

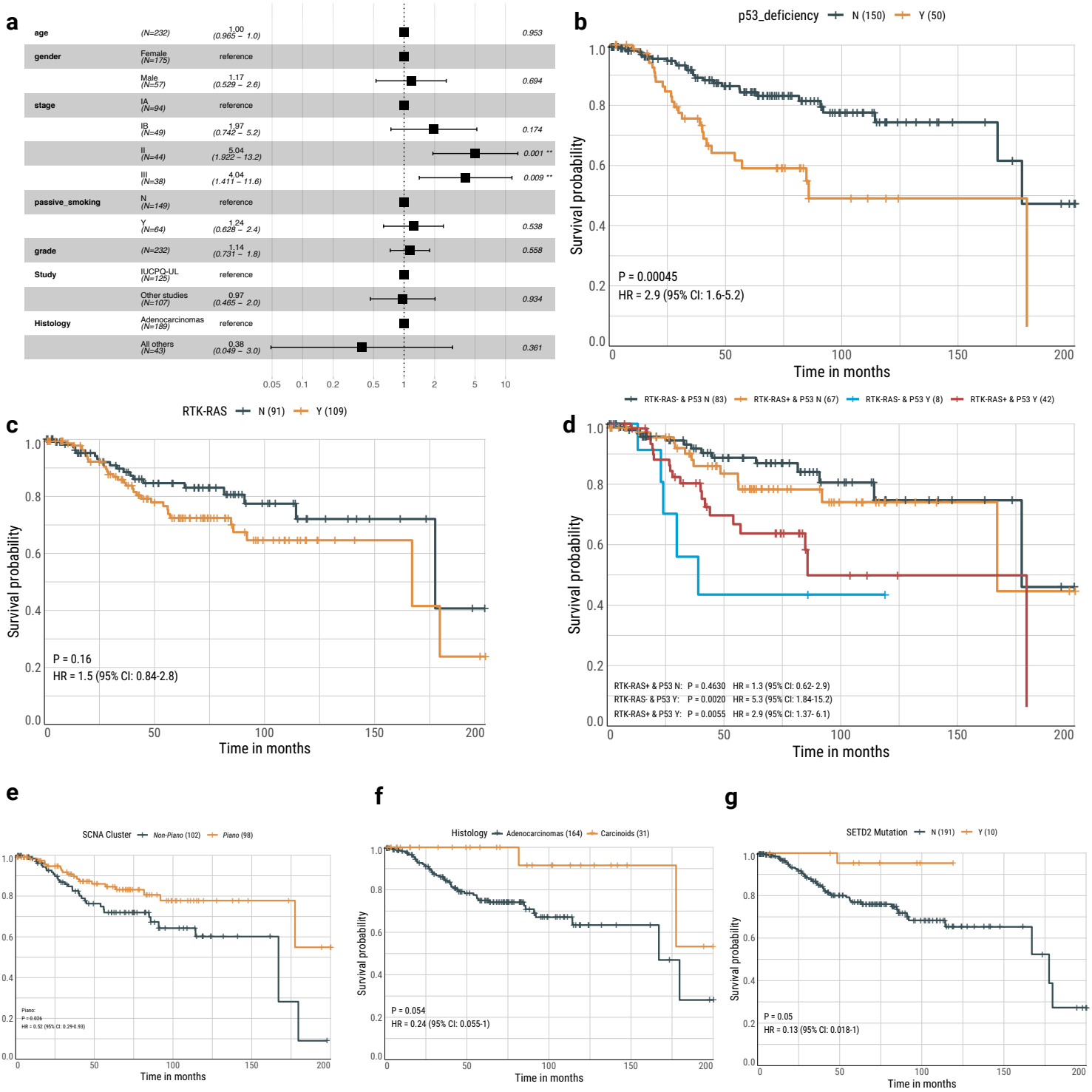
a, Molecular timing estimation of the arm-level chromosome duplications. Each dot represents a tumor sample harboring arm-level chromosome duplication. The molecular time is estimated based on co-amplified point mutations in individual samples. UPD: uniparental disomy (UPD=CN-LOH). Center lines in box plots show the medians; green box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. **b**, Timing mutations in driver genes relative to copy number gain or uniparental disomy (UPD) events of the corresponding chromosomal regions. The bar plots indicate the number of samples with mutations in driver genes occurring earlier (blue) or later (orange) than the SCNA events at the corresponding chromosomal regions.

Supplementary Fig. 25



Supplementary Fig. 25 Median latency between the MRCA and the age at diagnosis for different CpG>TpG mutation rate change estimates across SCNA subtypes.

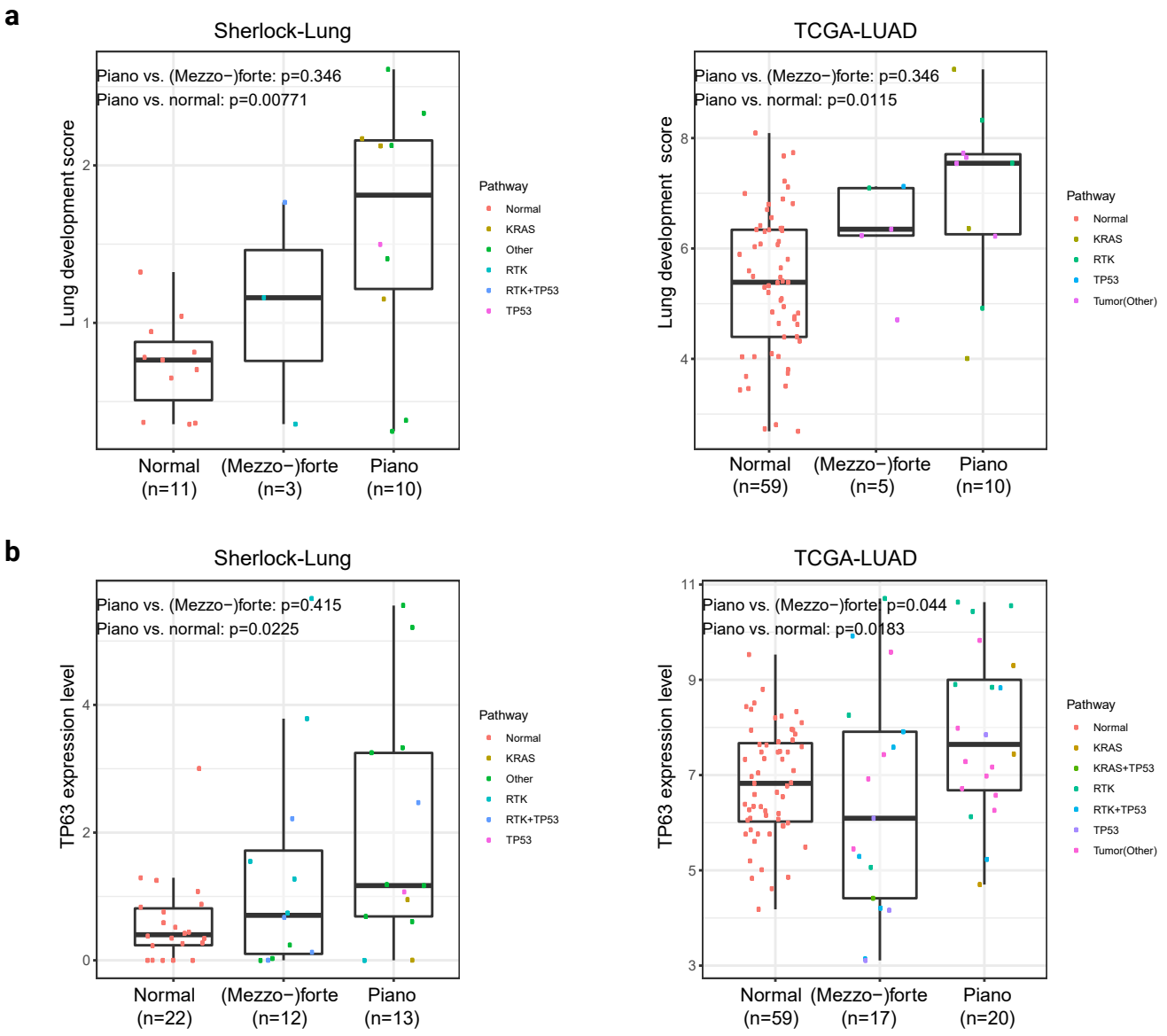
Supplementary Fig. 26



Supplementary Fig. 26 Additional associations between genomic aberrations and clinical outcomes.

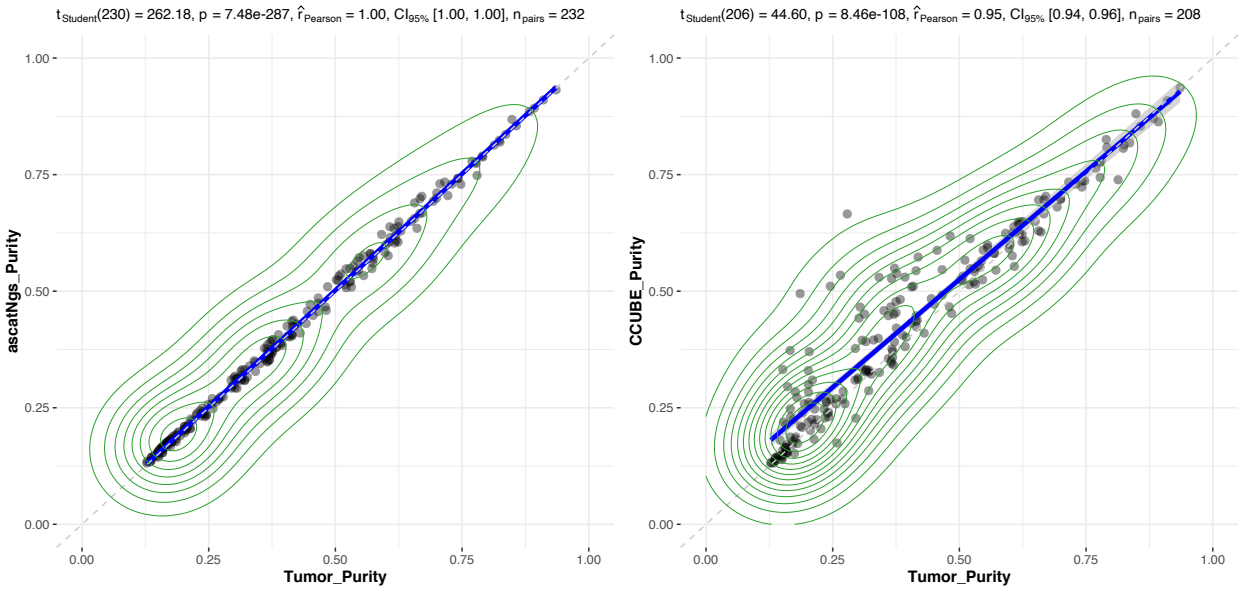
a, Forest plot showing hazard ratios with 95% confidence intervals of multivariate analysis of clinical and pathological features for overall survival probability, including age at diagnosis, gender, tumor stage, tumor grade, passive smoking status, study sites, and tumor histological types (n=232). Kaplan-Meier survival analyses for overall survival probability stratified by overall TP53 deficiency status (**b**), overall RTK-RAS pathway activation status (**c**), combining TP53 deficiency and RTK-RAS pathway activation status (**d**), *piano* tumor (**e**), tumor histology type (**f**), and *SETD2* mutation status (**g**). *P*-values for significance and hazard ratios (HR) of difference are calculated using the cox proportional hazards regression (two-sided) with adjustment for age, gender and tumor stage. No multiple-testing correct applied. For groups in each plot, Y= “with” aberration; N=“without” aberration. The number in brackets indicates the number of patients.

Supplementary Fig. 27



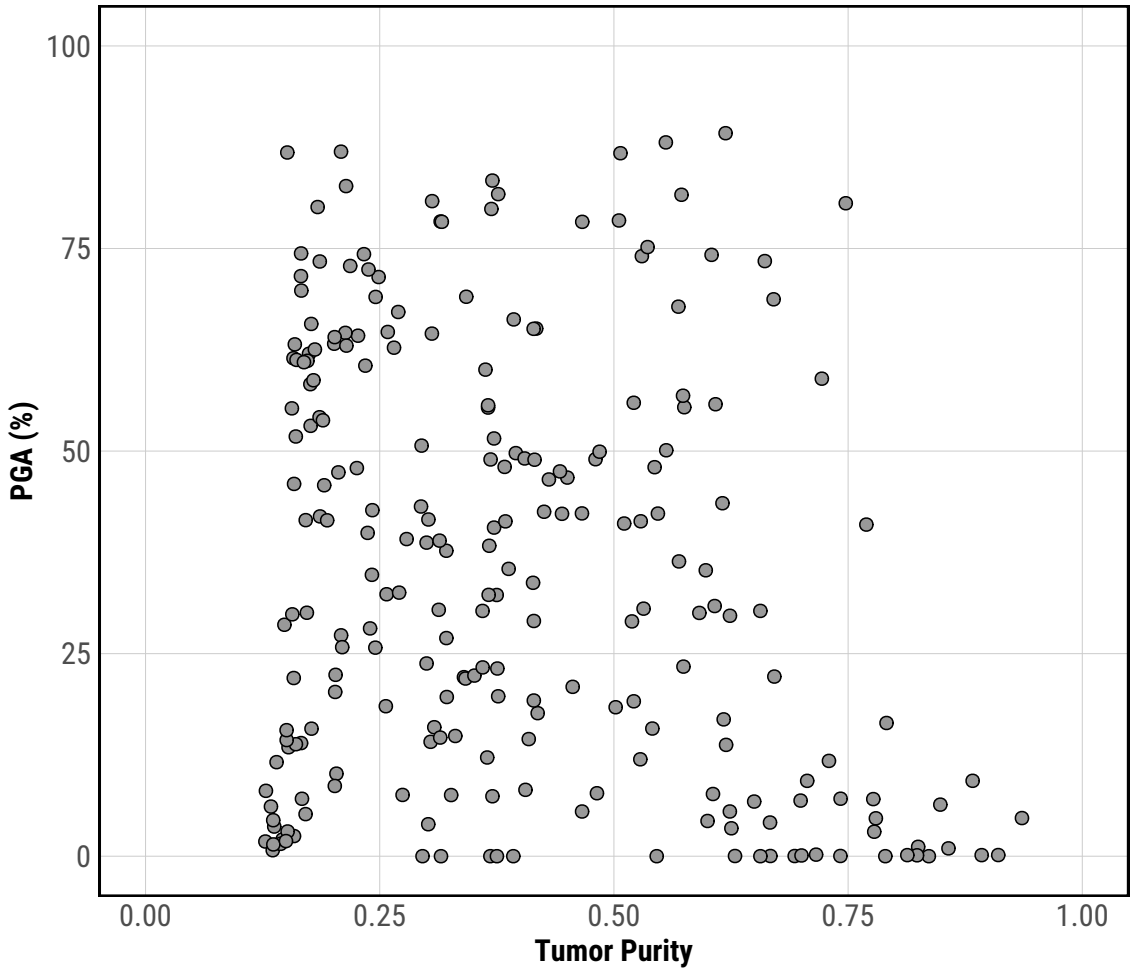
Supplementary Fig. 27 Analysis of gene expression signatures for stemness and cell of origin. (**a**) Comparison of lung development score in copy neutral tumor samples and normal lung tissue samples in (left) Sherlock-*Lung* and in (right) non-smokers from TCGA-LUAD study. (**b**) Comparison of the expression levels of TP63 in copy neutral tumor samples and normal lung tissue samples in (left) Sherlock-*Lung* and in (right) non-smokers from TCGA-LUAD study. For box plots, center lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. *P*-values were derived from two-sided Student's t-test with no multiple-testing correction applied.

Supplementary Fig. 28



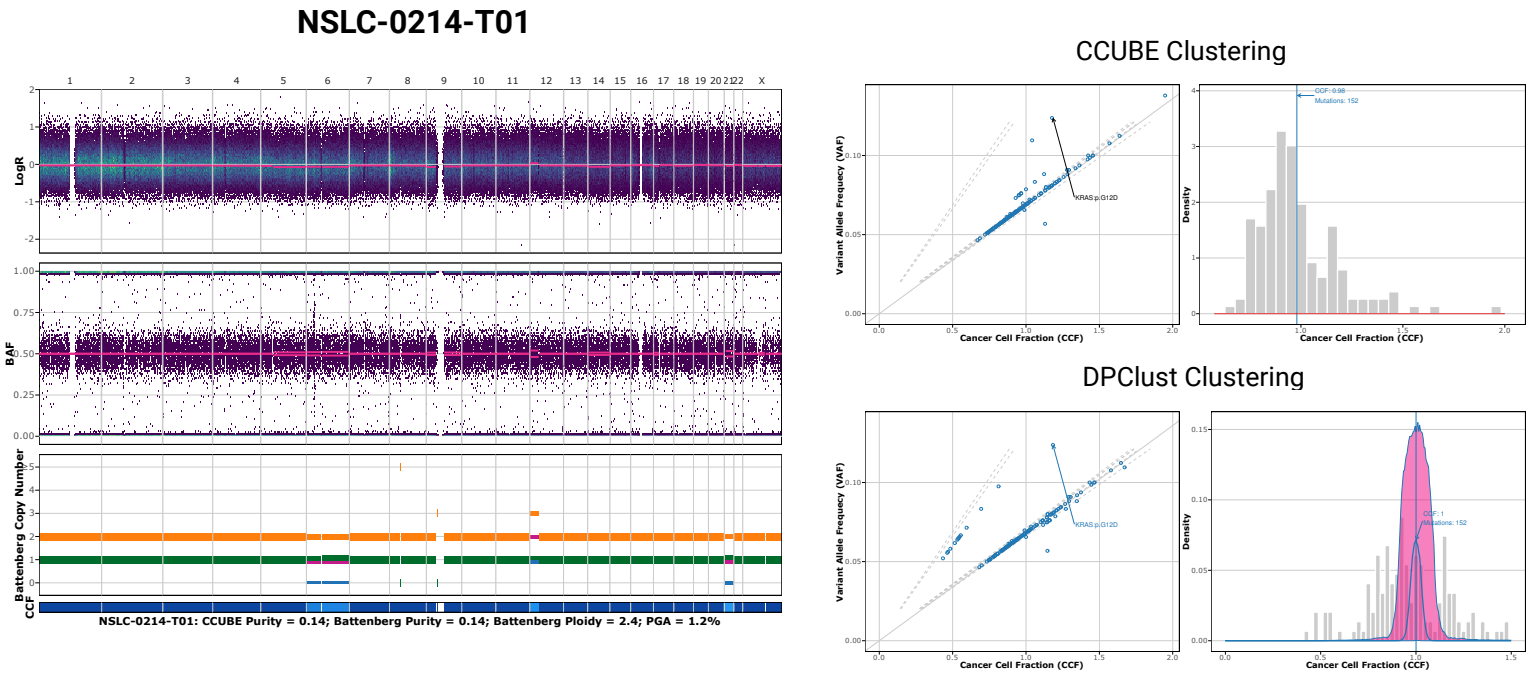
Supplementary Fig. 28 Association between final tumor purity and purity estimated by amscatngs (left) and CCUBE (right). The statistical test reported on the top of each figure. Note, due to insufficient SNVs in 24 samples, the CCUBE purity estimations were conducted only on 208 of the 232 tumor samples.

Supplementary Fig. 29



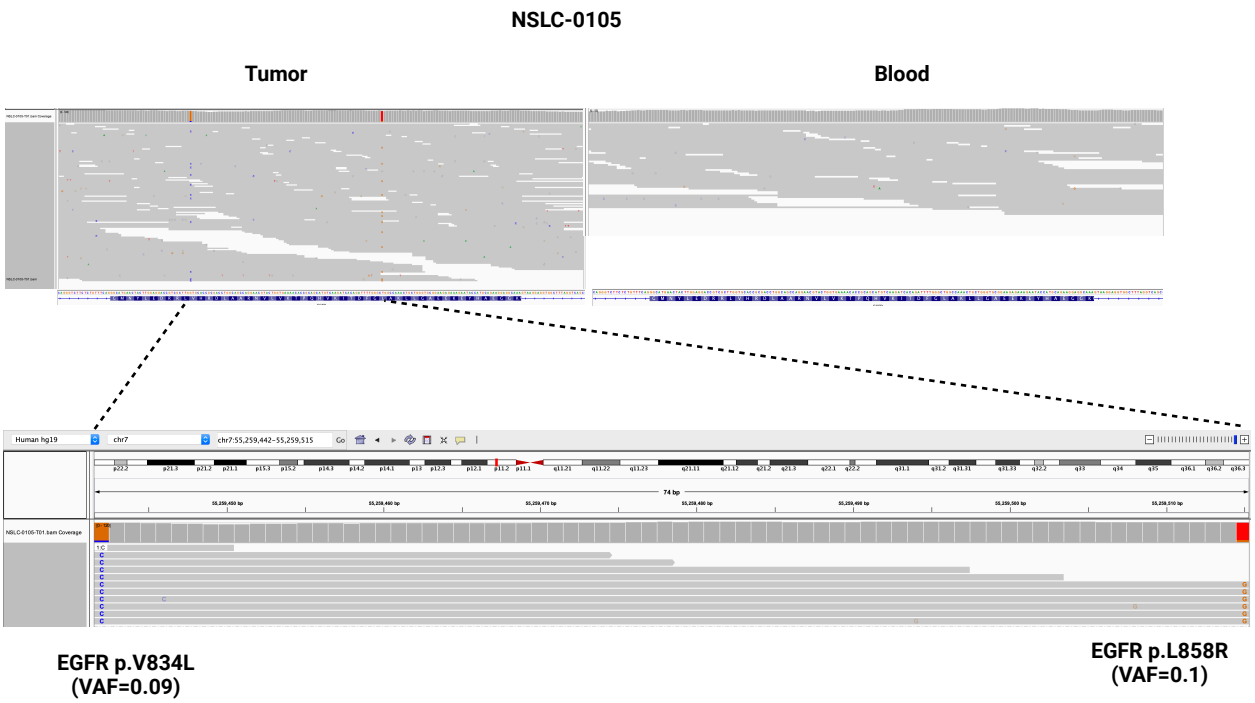
Supplementary Fig. 29 Association between tumor purity and percentage of Genomic Alteration (PGA) by SCNA.

Supplementary Fig. 30



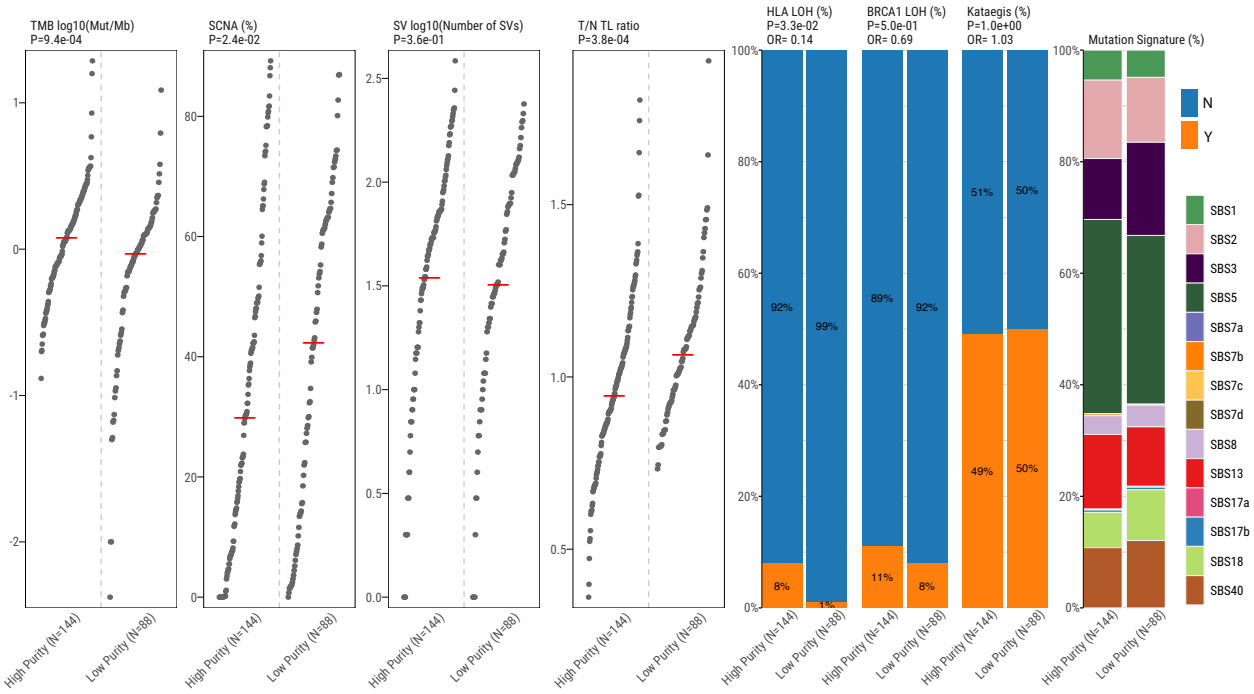
Supplementary Fig. 30 SCNA profiling using Battenberg (left) and mutation clustering using both CCUBE (top right) and DPClust (bottom right) for tumor NSLC-0214-T01. Each Battenberg SCNA profile includes LogR, BAF, absolute copy number (orange: total copy number for major clone; green: minor copy number for major clone; purple: total copy number for additional subclone; blue: minor copy number for additional subclone) and clonality of each segment (dark blue to light blue: ccf from 1 to 0). The tumor sample name, purity estimated by CCUBE, the final purity (Battenberg Purity), final ploidy (Battenberg Ploidy) and percentage of genome altered is shown at the bottom of each Battenberg profile. Mutation clustering figures show both CCF vs VAF scatter plot and CCF distribution. Different colors represent mutations included in different clusters. Potential drive genes are labeled on the clustering plots.

Supplementary Fig. 31



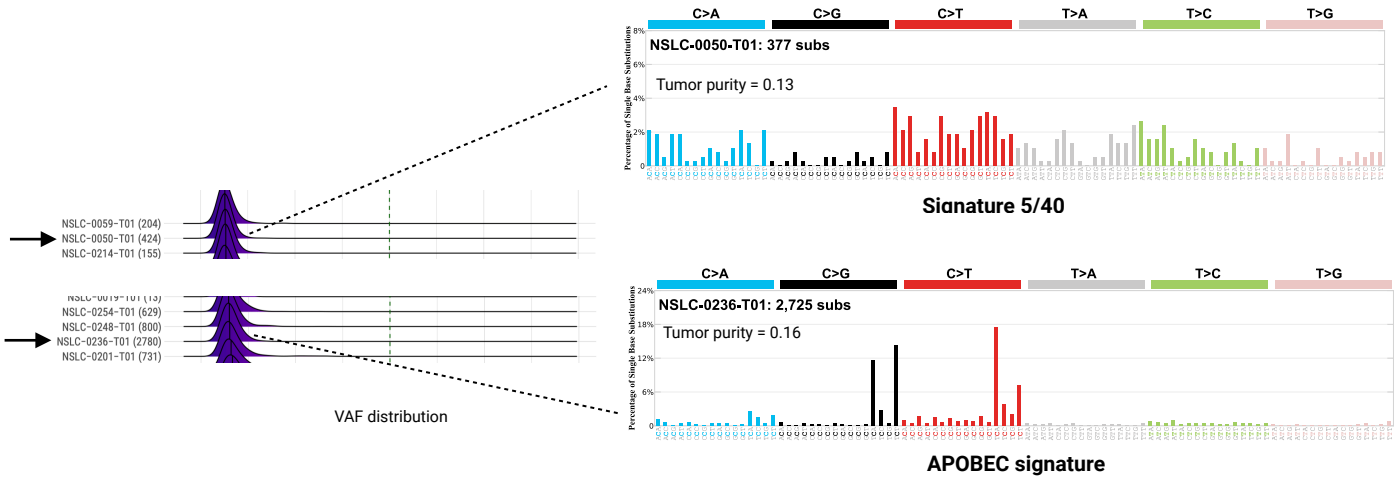
Supplementary Fig. 31 IGV visualization of *EGFR* double mutations *in cis* from patient NSLC-0105.

Supplementary Fig. 32



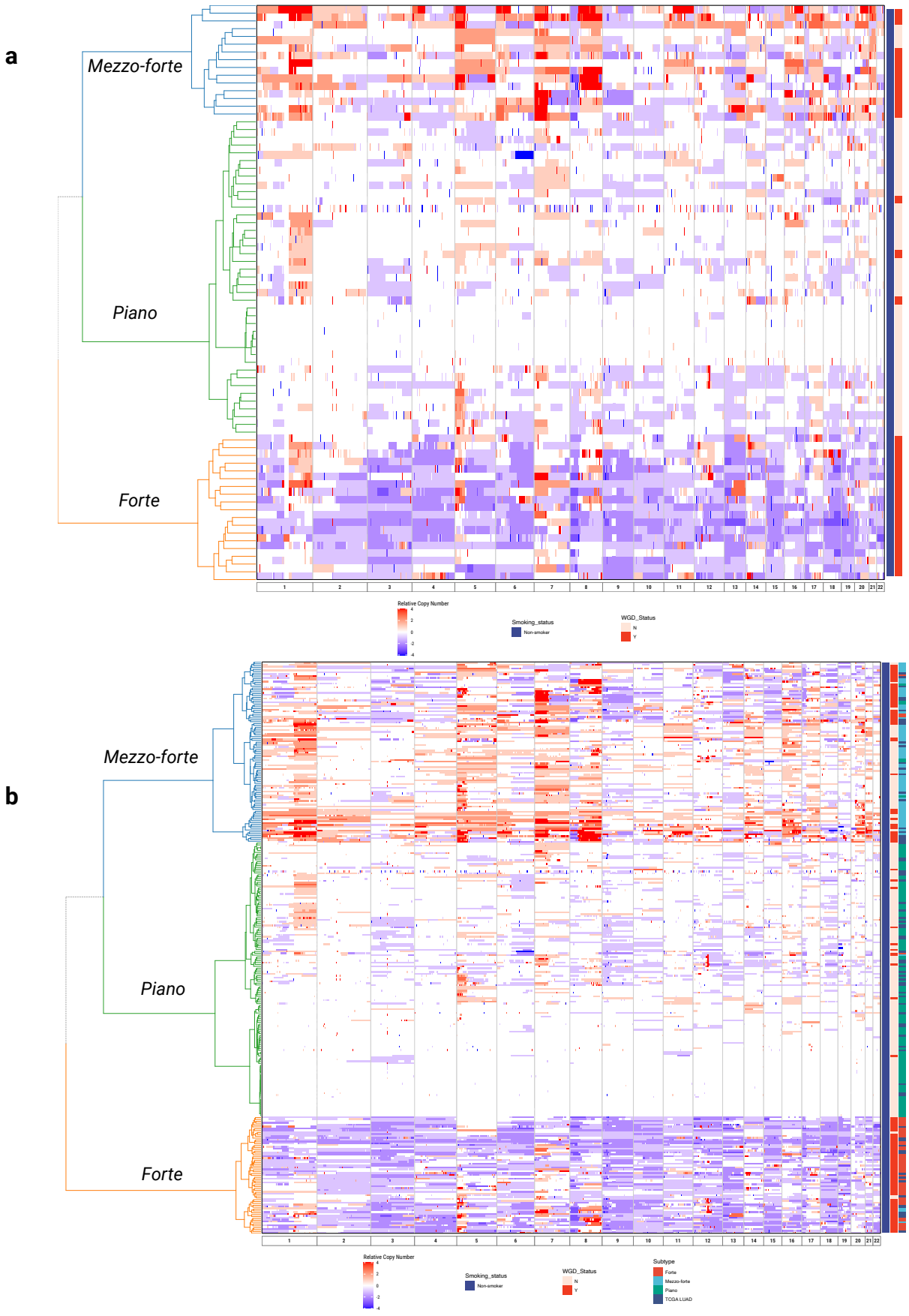
Supplementary Fig. 32 Comparison of genomic features between high purity tumors and low purity tumors. Left four panels: tumor mutational burden, percentage of genome with SCNA, SV burden and T/N TL ratio. *P*-values are calculated using the two-sided Mann-Whitney U test. Middle three panels (Y="with", N="without"): enrichments for *HLA* LOH, BRCA LOH, and Kataegis events. *P*-values and *OR* are calculated using Fisher's exact test (two sided). Right panel: Contributions of each SBS signature.

Supplementary Fig. 33



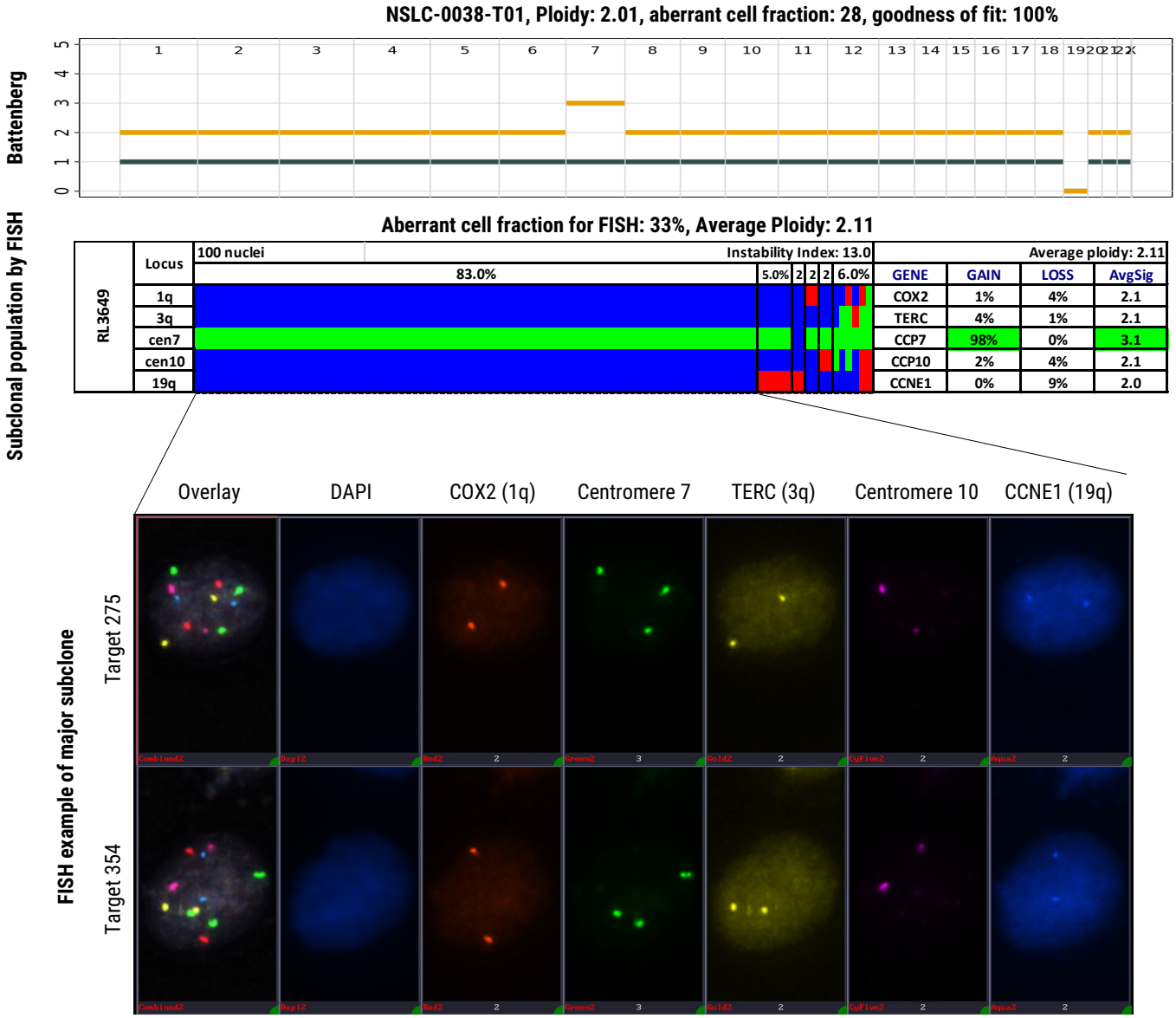
Supplementary Fig. 33 Examples of different mutational spectrum in two low purity tumors with only low VAF peak.

Supplementary Fig. 34



Supplementary Fig. 34 Unsupervised clustering of arm-level SCNA events in TCGA LUAD non-smokers (N=75; a) and in combination of our Sherlock-*Lung* data (b). The relative copy number is calculated as: total copy number - ploidy (non-WGD=2 and WGD=4).

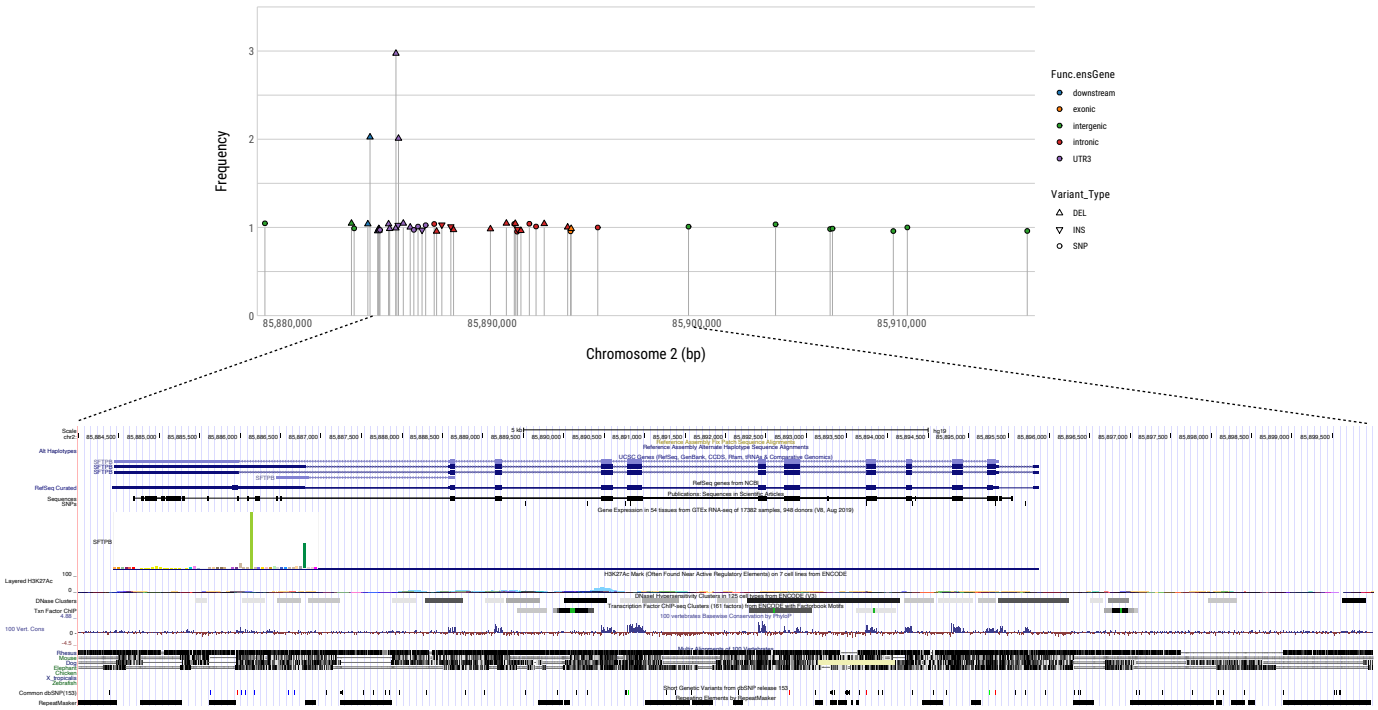
Supplementary Fig. 35



Supplementary Fig. 35 Example of Chr.19 homozygous deletion validation originally identified by Battenberg algorithm.

Multicolor Fluorescence *in situ* hybridization (mFISH) of samples with apparent Chr.19 homozygous deletion (HD) from Battenberg analysis. One representative tumor (NSLC-0038-T01) where Battenberg showed a gain of Chr.7 and HD loss of Chr.19 (top panel) but the mFISH only confirmed the gain of Chr.7 (lower panels). 100 aberrant nuclei were counted. In the mFISH color chart, blue=copy number neutral, green=copy number gain, red=copy number loss. The mFISH panel consisting of three gene-specific probes: *COX2* (1q33.2; red), *TERC* (3q26.2; gold), and *CCNE1* (19q12;aqua), and two centromere probes: *CEP7* (green) and *CEP10* (far red), revealed a gain in Chr.7 in ~98% of aberrant nuclei and no HD of Chr.19.

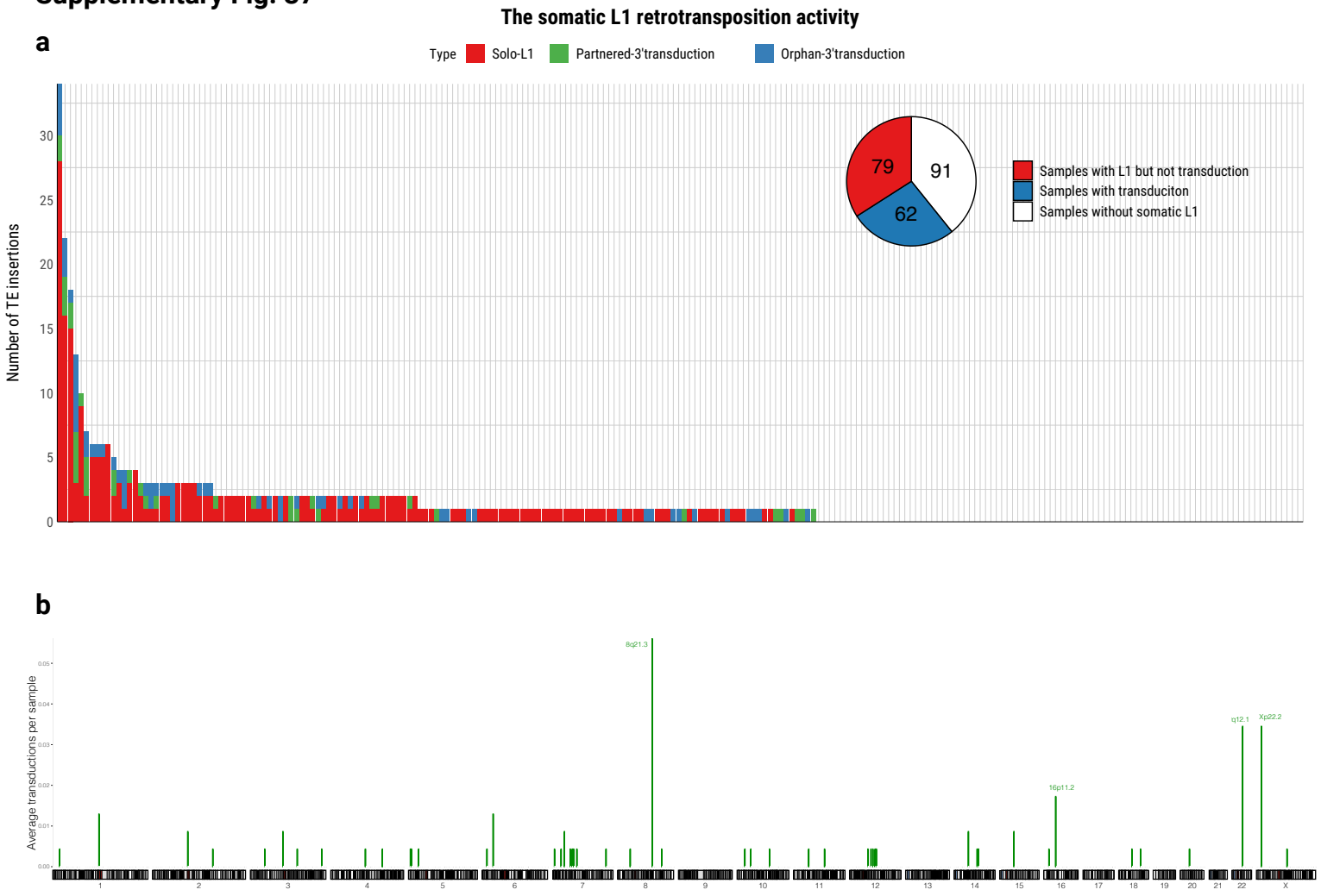
Supplementary Fig. 36



Supplementary Fig. 36 Identification of somatic mutations in the surfactant associated gene *SPFTB*.

The top plot shows the mutation functional classifications, with types are highlighted in different colors and shapes, respectively. The bottom plot indicates the *SPFTB* gene annotated tracks from UCSC Genome Browser.

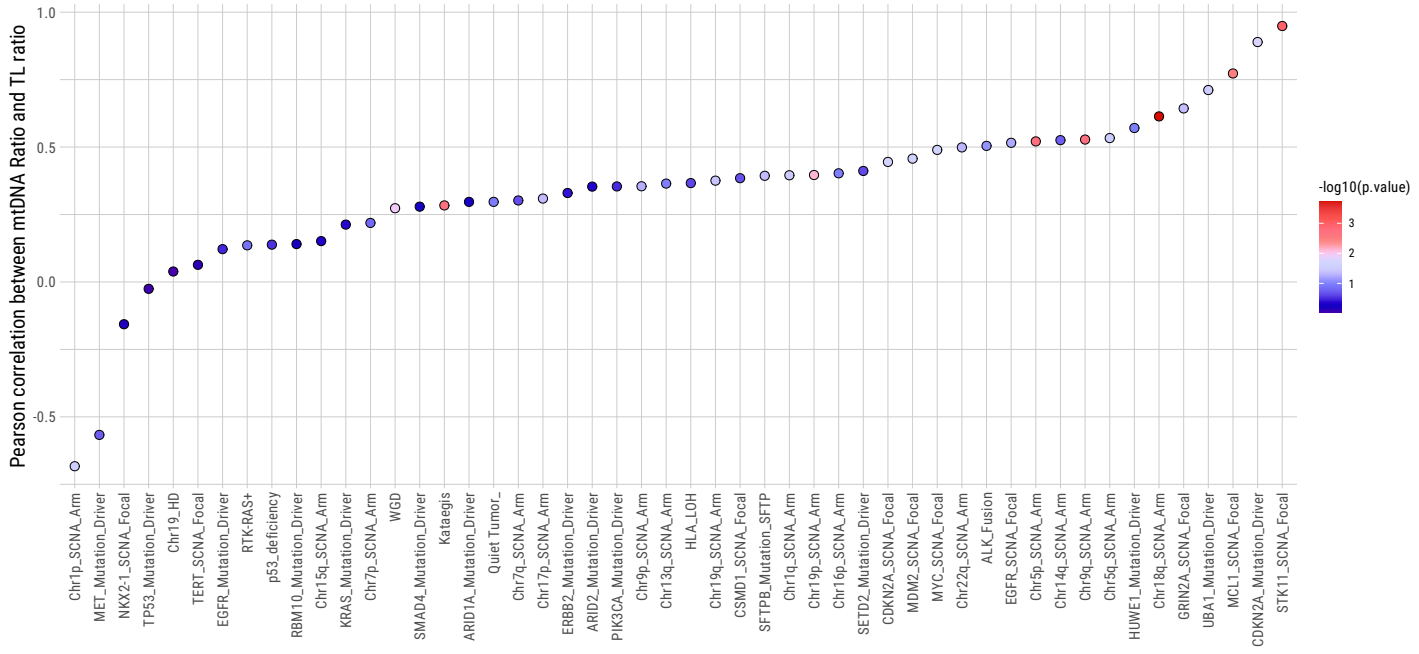
Supplementary Fig. 37



Supplementary Fig. 37 Activity of L1 retrotransposition insertion.

a, Frequencies of L1 retrotransposition activity. Pie chart represents the number of samples with transduction (blue), at least one solo L1 but no transduction (red) and no L1 retrotransposition (white). Bar plot represents the numbers of transposon insertions (y-axis) in each patient (x-axis). **b**, Summary of the rate of source element activity. Average number of transductions involving the given source element per sample (y-axis) are shown across the chromosomes (x-axis).

Supplementary Fig. 38



Supplementary Fig. 38 Pearson correlation between T/N mtDNA ratio and T/N telomere length (TL)TL ratio in a subset of tumors with the same genomic alterations (N>5).