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

nature genetics

Genomic and evolutionary classification of lung cancer in never smokers

Tongwu Zhang[1], Philippe Joubert[2], Naser Ansari-Pour[3], Wei Zhao[1], Phuc H. Hoang[1], Rachel Lokanga[4], Aaron L. Moye[5], Jennifer Rosenbaum[6], Abel Gonzalez-Perez[7], Francisco Martínez-Jiménez[7], Andrea Castro[8], Lucia Anna Muscarella[9], Paul Hofman[10], Dario Consonni[11], Angela C. Pesatori[11,12], Michael Kebede[1], Mengying Li[1], Bonnie E. Gould Rothberg[13,14], Iliana Peneva[15,16], Matthew B. Schabath[17], Maria Luana Poeta[18], Manuela Costantini[19], Daniela Hirsch[4], Kerstin Heselmeyer-Haddad[4], Amy Hutchinson[1,20], Mary Olanich[1,20], Scott M. Lawrence[1,20], Petra Lenz[1,20], Maire Duggan[21], Praphulla M. S. Bhawsar[1], Jian Sang[1], Jung Kim[1], Laura Mendoza[1], Natalie Saini[22], Leszek J. Klimczak[23], S. M. Ashiqul Islam[24], Burcak Otlu[24], Azhar Khandekar[24], Nathan Cole[1,20], Douglas R. Stewart[1], Jiyeon Choi[1], Kevin M. Brown[1], Neil E. Caporaso[1], Samuel H. Wilson[22], Yves Pommier[25], Qing Lan[1], Nathaniel Rothman[1], Jonas S. Almeida[1], Hannah Carter[8], Thomas Ried[4], Carla F. Kim[5,26], Nuria Lopez-Bigas[7,27], Montserrat Garcia-Closas[1], Jianxin Shi[1], Yohan Bossé[2,28], Bin Zhu[1], Dmitry A. Gordenin[22], Ludmil B. Alexandrov[24], Stephen J. Chanock[1], David C. Wedge[3,29] and Maria Teresa Landi[1]✉

Lung cancer in never smokers (LCINS) is a common cause of cancer mortality but its genomic landscape is poorly characterized. Here high-coverage whole-genome sequencing of 232 LCINS showed 3 subtypes defined by copy number aberrations. The dominant subtype (piano), which is rare in lung cancer in smokers, features somatic *UBA1* mutations, germline *AR* variants and stem cell-like properties, including low mutational burden, high intratumor heterogeneity, long telomeres, frequent *KRAS* mutations and slow growth, as suggested by the occurrence of cancer drivers' progenitor cells many years before tumor diagnosis. The other subtypes are characterized by specific amplifications and *EGFR* mutations (mezzo-forte) and whole-genome doubling (forte). No strong tobacco smoking signatures were detected, even in cases with exposure to secondhand tobacco smoke. Genes within the receptor tyrosine kinase–Ras pathway had distinct impacts on survival; five genomic alterations independently doubled mortality. These findings create avenues for personalized treatment in LCINS.

Lung cancer is the leading cause of cancer-related deaths with approximately two million people diagnosed each year[1]. LCINS account for 10–25% of all lung cancers, with most LCINS being lung adenocarcinomas (LUAD)[2]. Several studies have profiled the genomic landscape of LUAD[3–10] and the rarer carcinoid subtype[11]. Previous LUAD samples were mostly from smokers, primarily undergoing whole-exome sequencing (WES). The largest moderate to high-coverage whole-genome sequencing (WGS)-based LUAD studies cumulatively total less than 100 LCINS individuals, mostly of Asian ancestry[4,8,10,12–14]. As part of the Sherlock-*Lung* study[15], we evaluated the genomic landscape and mutational processes in 232 treatment-naïve LCINS using high-coverage WGS (tumor: 70.6–141.5×, mean: 85×; normal: 26.2–57.2×, mean: 31.6×) (Supplementary Table 1). Three subtypes based on somatic copy number alterations (SCNAs) were observed, with major genomic differences from LUAD in smokers and distinct clonal evolutionary patterns affecting diagnosis and possibly survival. Our findings suggest developmental processes and possible new therapeutic approaches for LCINS.

Results

Characteristics of Sherlock-*Lung* patients with cancer. Fresh-frozen tumor tissue and matched germline DNA were obtained from 232 treatment-naïve never smoker patients with lung cancer with unknown exposures to lung cancer risk factors, with the exception of secondhand (passive) tobacco smoking in 27.6% of patients (Methods and Supplementary Table 1). Patients were diagnosed with non-small-cell lung carcinoma (NSCLC), including 189 patients with adenocarcinomas, 36 patients with carcinoids and 7 patients with other tumors of various subtypes (Methods). Patients were predominantly of European ancestry ($n=226$; 97.4%), with the rest of Asian ($n=4$; 1.7%) or African ($n=2$; 0.9%) ancestry (Supplementary Fig. 1).

Genomic characteristics of LCINS. The median tumor mutational burden (TMB) was 1.1 Mut Mb^{-1} (single-nucleotide variation (SNV) = 1.0; insertions and deletions (indels) = 0.06), more than sevenfold lower than in smokers[4] ($P=7\times10^{-73}$) (Fig. 1). TMB was significantly associated with tumor stage, histology and age at diagnosis but not tumor purity (Supplementary Fig. 2).

A full list of affiliations appears at the end of the paper.

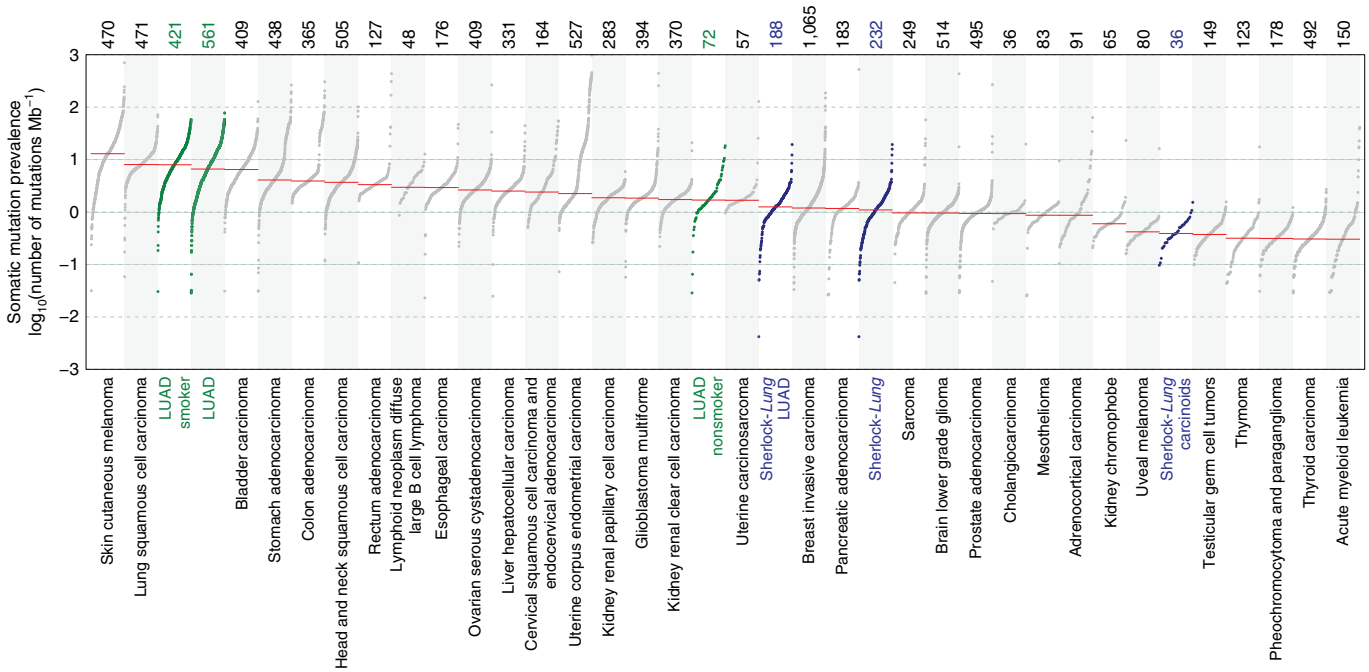


Fig. 1 | TMB across LCINS from the Sherlock-*Lung* study and 33 cancer types from the TCGA study. The Sherlock-*Lung* samples (blue) are shown overall and by histological type. The TCGA LUAD samples (green) are shown overall and by smoking status. Each dot represents a sample; the total sample numbers for each type are shown at the top. The red horizontal lines are the median numbers of mutations per megabase ($\log_{10}$).

The major genomic characteristics of LCINS are summarized in Fig. 2 and in the Supplementary Note. Among genes in the receptor tyrosine kinase–Ras (RTK–Ras) pathway, *EGFR* was the most frequently altered (30.6%), followed by *KRAS* (7.3%), *ALK* (6.0%), *MET* (4.3%), *ERBB2* (3.9%, all indels), *ROS1* (2.6%) and *RET* (1.3%). A strong mutually exclusive distribution was observed across these 7 genes, which were altered in total in 54.3% of tumors (Extended Data Fig. 1a). The pattern of genomic alterations was strikingly different between RTK–Ras$^+$ and RTK–Ras$^-$ groups (Extended Data Fig. 1b). The former had a significantly higher burden of SNVs/indels, SCNAs, structural variants (SVs), kataegis, whole-genome doubling (WGD) and *BRCA2* loss of heterozygosity (LOH) but lower tumor/normal telomere length (T/N:TL) ratios. The 49 (21.1%) tumors bearing both *TP53* deficiency and activating RTK–Ras mutations had higher TMB, as observed previously[16] and also higher kataegis, WGD and LOH in genes associated with DNA homologous repair than tumors with either *TP53* deficiency or RTK–Ras alterations alone (Supplementary Fig. 3).

As expected[17], *TP53* mutations were mutually exclusive with *MDM2* amplifications ($P=0.03$; Extended Data Fig. 2a) and tumors with mutations in either gene (25.4% in total) were enriched with genomic alterations including SNVs, SCNAs, SVs, kataegis, WGD, human leukocyte antigen (HLA) LOH and LOH in *BRCA1* (Extended Data Fig. 2b).

SVs were enriched in hotspot regions, including *MDM2*, *TERT*, *6p21*, *MYC*, *CDKN2A*, *NKX2-1* and *GNAS*, which together contributed 16.7% of SVs (>200 breakpoints within a 5-Mb window; Extended Data Fig. 3) as observed in multiple tumor types[18]. Known driver fusion oncogene-generating rearrangements were observed in 24 (10.3%) tumors (Supplementary Table 2) and were mutually exclusive with *EGFR* mutations ($P=1.1\times10^{-4}$). Non-clustered SVs were enriched in *TP53*-deficient tumors and RTK–Ras$^+$ tumors (Supplementary Fig. 4a–c).

Copy number alteration subtypes. Unsupervised clustering of arm-level SCNAs identified three distinct subtypes, with increasing levels of SCNAs (Fig. 3a). Subtype 1 (49.6% of all tumors) largely lacked SCNAs despite relatively high purity and included 33 of 36 carcinoids and 78 adenocarcinomas. Subtype 2 (30.2%) was enriched with chromosome arm-level amplifications, primarily of 1q, 5p, 7p, 7q (each with $P<0.001$ subtype 2 versus other subtypes, Fisher's exact test) and 8q (exclusively in subtype 2). Subtype 3 (20.2%) was dominated by WGD. Hereafter we refer to these three subtypes respectively as 'piano', 'mezzo-forte' and 'forte', borrowing the terms from musical dynamics. Combining our copy number profiles with those of LUAD from smokers ($n=38$)[12], the majority of LUAD from smokers (20 out of 38, 52.6%) fell into the subtype forte ($P=6.6\times10^{-5}$) (Supplementary Fig. 5a). Focal amplifications of *MDM2* and *EGFR* were significantly less frequent in the subtype piano than in forte and mezzo-forte ($P=0.001$ and $P=0.02$ respectively; Supplementary Table 3 and Supplementary Fig. 5b). Mitochondrial DNA copy numbers were higher than previously reported in LUAD from smokers ($P=0.01$)[19] (Supplementary Note and Supplementary Fig. 6a–c). HLA LOH was previously identified in nearly 40% of lung cancer cases, particularly squamous cell carcinomas[20]. In our cohort, only 5.2% of tumors (all LUAD, mostly in forte and mezzo-forte) harbored HLA LOH (Methods and Supplementary Table 1).

Genomic features across SCNA subtypes. Notably, several other genomic features of LCINS differed between SCNA-defined subtypes. TMB was much lower in the piano subtype (0.7 Mut Mb^{-1}), particularly in carcinoids (0.4 Mut Mb^{-1}), compared to forte (1.4 Mut Mb^{-1}) and mezzo-forte (1.6 Mut Mb^{-1}) ($P=2.0\times10^{-7}$ and $P=3.2\times10^{-11}$, respectively) (Fig. 3b).

While 24 out of 25 recurrently mutated genes (Supplementary Table 4 and Supplementary Fig. 7) were previously identified as drivers in the TCGA Pan-Cancer cohort[21], many of them had substantial frequency differences from LUAD in smokers and across SCNA subtypes (Extended Data Fig. 4 and Fig. 2). For example, *TP53* mutations were most common in forte (31.9%, 18.6% and 7.0% for forte, mezzo-forte and piano; $P=1.2\times10^{-3}$), while

Fig. 2 | Genomic characteristics of LCINS. Numbered sections from top to bottom: (1) distribution of genomic alteration numbers; (2) most frequently mutated or potential driver genes; (3) oncogenic fusions; (4) Somatic mutations in surfactant associated genes; (5) significant focal SCNAs; (6) significant arm-level SCNAs; (7) genes with rare germline mutations; (8) and (9) different genomic features. The numbers on the right show the overall frequency (1–8) or median values (9).

remaining lower than in LUAD from smokers (53.4%)[21]. Furthermore, we identified one likely new driver gene, *UBA1* (6 out of 9 in piano), which encodes an E1 ubiquitin-conjugating enzyme that acts as one of the main orchestrators of the cellular DNA damage response[22]. All 25 recurrently mutated genes exhibited signals of positive selection[23] (Extended Data Fig. 5).

Over half of the tumors in mezzo-forte had *EGFR* mutations (51.4% versus 9.0% in LUAD from smokers), while only 1.4% had *KRAS* mutations (versus 34.0% in LUAD from smokers). Tumors in piano were less likely to have *TP53* deficiency or aberrations in *EGFR* and other RTK–Ras genes ($P = 4.4 \times 10^{-6}$ and $P = 4.3 \times 10^{-7}$, respectively), with the exception of *KRAS* (76.5% of *KRAS*$^+$ tumors

were in piano, $P=0.02$). While carcinoids in piano were enriched ($P=7.5\times10^{-4}$) with mutations in chromatin-remodeling genes (for example, *ARID1A*) as observed previously[11], LUAD in this subtype rarely harbored a known recurrent driver, with the exception of mutations in *NKX2-1* ($n=4$ only in piano), *SETD2* ($n=8$ out of 10 in piano) and *UBA1*. These piano tumors exhibited low burden of SNVs/indels, SCNAs, SVs, kataegis and WGD and a high T/N:TL ratio and subclonal mutation ratio, with carcinoids being exceptionally quiet (Fig. 3b).

The median number of SVs per tumor varied widely between SCNA subtypes, with 73, 63 and 10 in forte, mezzo-forte and piano, respectively, distributed as translocations (52.4%), deletions (32.7%) and tandem duplications (14.5%) (Supplementary Fig. 8). Of note (Extended Data Fig. 4), *RET* fusions were present only in the piano subtype (2.6%).

Rare, predicted deleterious germline variants were identified[24] (Methods and Supplementary Table 5) recurrently in *CYP21A2*, which encodes the 21-hydroxylase enzyme involved in the synthesis of cortisol and aldosterone ($n=8$, 6 with an identical stop-gain variant, more common in forte and mezzo-forte) and *GLUD2*, which encodes glutamate dehydrogenase 2 in the mitochondria ($n=6$, identical variant). Among known cancer susceptibility genes, *AR* was the most frequently mutated ($n=5$, 4 of which in piano). Variants in both *CYP21A2* and *AR* suggest a role for hormones in driving LCINS, warranting further investigation. A handful of tumors had germline variants in homologous recombination genes[25] including *BRCA1* ($n=3$, 2 of which in piano), *ATM* ($n=2$) and *RAD51D* ($n=2$). Single variants, one per tumor, were identified in *CDK4* in forte, *SOS1* in mezzo-forte and *RET* and *MSH6* in piano.

Mutational signatures in LCINS. Mutational signature deconvolution of single base substitutions (SBS) using SigProfiler[26,27] identified 14 previously reported signatures from the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Supplementary Table 6, Fig. 4 and Supplementary Fig. 9). Notably, SBS18, related to damage by reactive oxygen species (ROS)[28], was observed in 46% of samples, particularly in the SCNA subtypes forte and mezzo-forte (59.6% and 67.1%, respectively, $P=2.2\times10^{-9}$ in comparison to piano). SBS8, linked to nucleotide excision repair deficiency[29] and late replication errors[30], was present in 13% of samples, particularly in carcinoids (30.3%, $P=2.7\times10^{-4}$). In the 38 LUAD from the Pan-Cancer Analysis of Whole Genomes (PCAWG)[12], SBS8 was not identified, indicating possible differences in the etiology of tumors from smokers and never smokers. Signature extraction using indel (ID-83) and double-base substitution (DBS-78) profiles[26] identified six indel (ID1, 2, 3, 5, 8 and 9) and four DBS (DBS2, 4, 9 and 11) signatures (Supplementary Fig. 10).

About 58% of samples ($n=135$) had $\geq$100 SNVs, mostly subclonal (median fold change = 1.2, interquartile range (IQR) = 1.0–1.5), assigned to APOBEC mutational signatures SBS2 and SBS13 (Supplementary Note and Supplementary Fig. 11a,b), with substantial intertumor heterogeneity (Fig. 4). APOBEC mutational loads were verified using pattern of mutagenesis by APOBEC cytidine deaminases (P-MACD)[31] (Supplementary Note and Supplementary Fig. 11c). APOBEC signatures were dominant (>80%) in 4 hypermutated tumors (TMB > 8 Mut Mb^{-1}), 1 in forte and 3 in piano. Significant enrichment of the APOBEC signature was observed in *TP53*-deficient ($P=1.9\times10^{-8}$) and RTK–Ras^{+} ($P=3.5\times10^{-8}$) tumors.

In all tumors, endogenous processes (Supplementary Table 7) predominated over exogenous processes[32] (median cosine similarities of 0.96 and 0.82, respectively, $P=4.8\times10^{-53}$; Extended Data Fig. 6). Only a few samples showed higher cosine similarities combining exogenous and endogenous signatures compared with endogenous signatures alone (Supplementary Fig. 12), particularly 6 tumors (4 in piano) with a signature of nitrated polycyclic aromatic hydrocarbons, 1,8 dinitropyrene)[32], contributing 18.7% of SNVs on average in these samples. No additional dominant or recurrent genomic events were apparent in these tumors (Supplementary Fig. 13). Nitrated polycyclic aromatic hydrocarbons are derived mostly from diesel exhaust and are associated with cancer risk[33].

The mutational signature associated with direct exposure to tobacco smoking (SBS4) was not observed, even in 62 cases with reported exposure to secondhand tobacco smoking (passive smoking) (Fig. 4). Our simulations demonstrated that signature SBS4, if present, is below the detection threshold of 15% of somatic mutations (Supplementary Note and Supplementary Fig. 14). The lack of passive smoking signatures was not explained by tumor purity differences between passive and nonpassive smokers ($P=0.39$; Fig. 5a) and was confirmed by measuring alkylation-induced mutagenesis[34] (Supplementary Note and Fig. 5b). Directly comparing the mutational patterns in passive versus nonpassive smokers, we found strong similarities between the two groups (Fig. 5c–e) ($Q>0.05$ for all SBS, DBS and indel signatures), even comparing highly and lowly exposed individuals (Methods, Supplementary Fig. 15 and Supplementary Table 8) or comparing strand asymmetries for mutation types or SBS signatures (Supplementary Figs. 16 and 17). Of note, tumors from passive smokers had shorter telomere lengths ($P=0.005$; Supplementary Fig. 18).

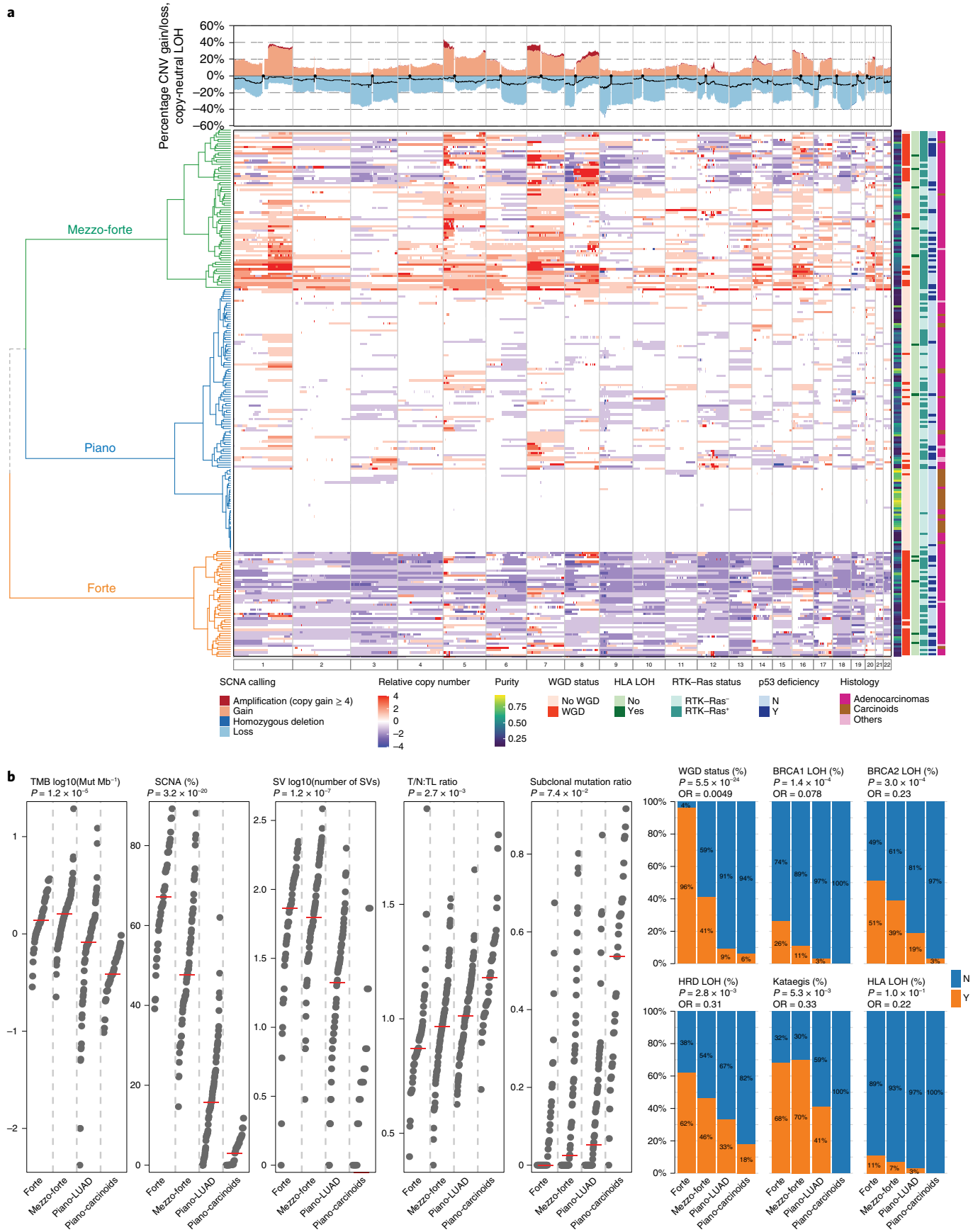
Genomic instability. Overall, 36.2% of tumors had WGD, with much higher prevalence in the forte subtype (95.7% versus 41.4% and 8.7% in mezzo-forte and piano, respectively) (Extended Data Fig. 4). Similarly, the proportion of the genome affected by SCNAs was much lower in piano than forte or mezzo-forte tumors ($P=6.8\times10^{-35}$; Supplementary Fig. 19).

Kataegis was identified in 49.6% of tumors, with an average of 4 events per sample (range: 1–55), rarely in piano tumors (29.6% versus 68.1% and 70% in forte and mezzo-forte, respectively; $P=1.3\times10^{-9}$). As expected, mutations within kataegis events had APOBEC-related signatures[26] in both APOBEC3A- and APOBEC3B-like tumors[35] (Supplementary Note and Supplementary Fig. 20). Kataegis frequently occurred within the *MDM2* locus ($P=1.3\times10^{-15}$) and often colocalized with SVs (Supplementary Fig. 21).

We estimated TL using two previously published methods[36,37], with comparable results, confirming an inverse correlation with age ($r=-0.14$, $P=0.04$) and no association with tumor purity (Supplementary Fig. 22). Notably, tumor TL in the LUAD of LCINS was significantly longer than that observed in the LUAD of smokers[36] (6.4 kilobases (kb), 95% confidence intervals (CIs): 5.3–7.6 kb, $P=7.1\times10^{-11}$; Extended Data Fig. 7a and Supplementary Fig. 23). Losses of 9q, 9p and 22q and HLA LOH were significantly associated with TL shortening (two-sided *t*-test, $Q<0.05$) and were

Fig. 3 | Genomic classification of LCINS based on SCNAs. a, Left: unsupervised clustering of arm-level SCNA events, piano, mezzo-forte and forte. The relative copy number was calculated as: total copy number − ploidy (non-WGD = 2 and WGD = 4). Samples in the rows are annotated by tumor purity, WGD status, HLA LOH, RTK–Ras status, *TP53* deficiency and tumor histological type. Top: SCNA frequency including amplification, deletion and copy-neutral LOH (black line). **b**, Comparison of genomic aberrations or features (Y = with, N = without) among forte, mezzo-forte, piano LUAD and piano carcinoids tumors. Left: TMB, percentage of genome with SCNAs, SV burden, T/N:TL ratio and subclonal mutation ratio. *P* values were calculated using a two-sided Mann–Whitney *U* test. Right: enrichments for WGD, kataegis, *BRCA2* LOH, *BRCA1* LOH, HRD LOH and HLA LOH. *P* values and OR were calculated using a two-sided Fisher's exact test. All statistical analyses were performed between forte and piano LUAD.

most frequent in forte and mezzo-forte (Extended Data Fig. 7b,c). While tumors in forte had significantly shorter telomeres (mean T/N:TL ratio 0.9, $P=0.01$, *t*-test), mezzo-forte tumors displayed no significant differences and piano had significantly longer telomeres than their matched normal tissues (mean T/N:TL ratio 1.1, $P=4.7\times10^{-3}$).



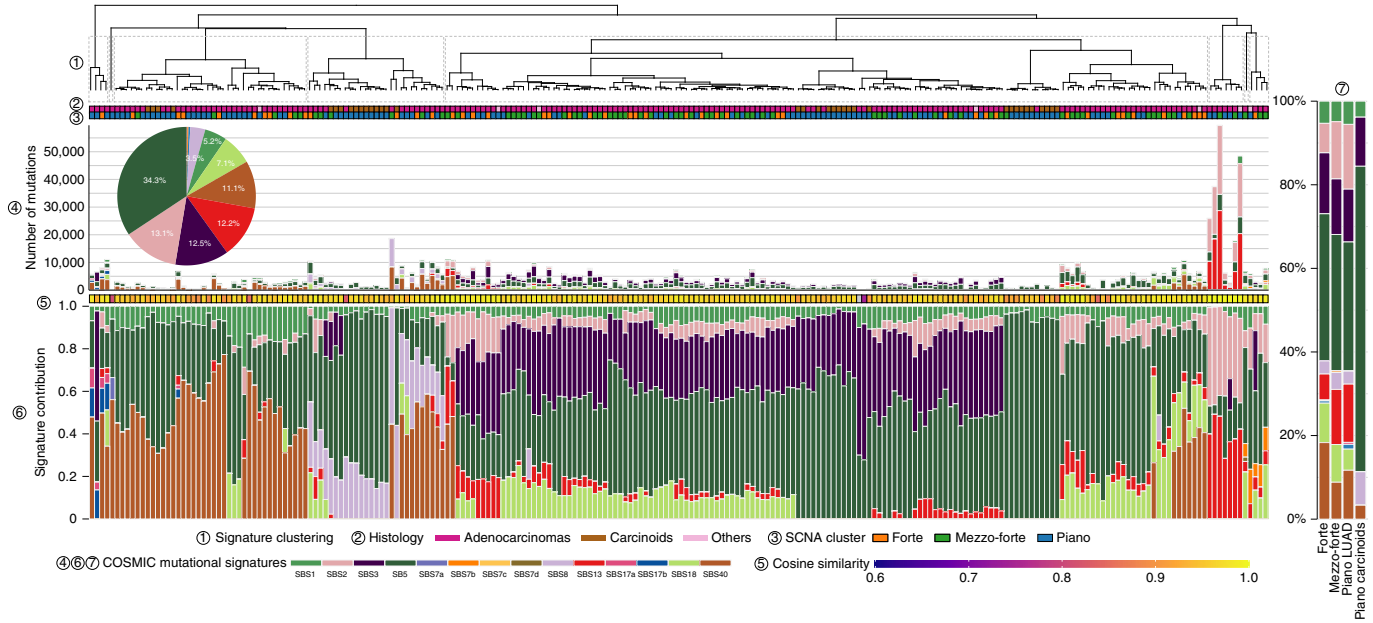


Fig. 4 | Landscape of mutational processes in Sherlock-*Lung*. Mutational signature profile of SBS across 232 Sherlock-*Lung* samples. Top to bottom: (1) unsupervised clustering based on the proportion of SBS signatures; (2) tumor histological type; (3) SCNA cluster; (4) pie chart showing the percentage of mutations contributed to each SBS signature and bar plot presenting the total number of SNVs assigned to each SBS signature; (5) cosine similarity between the original mutational profile and signature decomposition result; (6) proportions of SBS mutational signatures in each sample; and (7) proportions of SBS mutational signatures in each SCNA subtype.

Approximately 16.0% ($n=37$) of tumors had a high homologous recombination deficiency detect (HRDetect) score[25,38,39] (>0.7), with genomic aberrations predictive of homologous recombination deficiency (HRD) (Extended Data Fig. 8), particularly in forte and mezzo-forte ($P=1.4\times10^{-3}$ versus piano) (Fig. 3b). Biallelic loss of *ATM* in one tumor and monoallelic loss of HRD-associated genes in 42% of tumors had higher HRDetect scores (Extended Data Fig. 9).

The evolutionary history of LCINS. We reconstructed the likely order of acquisition of recurrent genomic aberrations, including SCNAs, WGD and common cancer driver genes within each of the SCNA subtypes in LUAD (Methods and Fig. 6). In all three subtypes, mutations in the driver genes *TP53*, *RBM10*, *KRAS* or *EGFR* were generally early events, occurring before both WGD and most other SCNAs. Two exceptions in mezzo-forte were the earlier occurrences of LOH on 17p targeting *TP53* and LOH on 3p12.2, likely targeting transcription factor *ZNF717*, suggesting that these events are also early drivers of mezzo-forte tumors. Whereas putative copy number drivers in mezzo-forte were balanced between gains and LOH, forte was dominated by LOH events. Compared to mezzo-forte, WGD generally occurred after other key SCNAs in forte. Early events in piano included mutations in *SETD2*, LOH of 8p and 17p, as well as focal gain at 3p12.2 and at 2p11.2 involving the immunoglobulin gene *IGKV1-5*.

Using the proportion of mutations on two or more chromosome copies allows for the relative timing of clonal copy number gains and copy-neutral LOH (CN LOH)[40,41]. Gains of 5q, 16p, 1p and 14q occurred early during tumor development, whereas gain of 7q and CN LOH events occurred relatively late (Supplementary Fig. 24a). Reversing this method to time driver mutations relative to clonal gains or CN LOH identified that mutations in *EGFR*, *MET*, *KRAS*, *ERBB2*, *TP53* and *UBA1* generally occurred before the corresponding copy number gain (Supplementary Fig. 24b). In contrast, mutations in *PIK3CA* and *SFTPB* occurred after gain events.

We adopted a previously validated model[42] using clock-like mutations (CpG>TpG in an NpCpG context) to time the appearance of the most recent common ancestor (MRCA) of all tumor cells. We used an estimated acceleration rate of 1×, given the low mutational burden and the paucity of exogenous mutational signatures in LCINS. The MRCA, by definition, possesses all driver mutations for tumorigenesis. Grouping tumors according to common driver events (>3% frequency) (Fig. 7a) enables the estimation of the occurrence of these events in an individual's lifetime. For example, in tumors with *EGFR* mutations, the MRCA was estimated to appear at 61 years of age but the tumors were clinically evident a median of 8 years later. There were substantial latency differences across tumors with different drivers. For example, in tumors harboring *ERBB2*, *CDKN2A*, or *TP53* mutations, or *NKX2-1*, *STK11* or chr22q SCNAs, the MRCA appeared more than a decade before clinical diagnosis. In contrast, tumors with *MDM2* amplifications, or *MET*, *RBM10*, *HUWE1* or *KRAS* mutations, had much shorter latency. Notably, tumors in piano had significantly longer latency (median: 9.10 years) than forte (median: 0.08 years) and mezzo-forte (median: 0.28 years) ($P=8.3\times10^{-4}$, Fig. 7b), suggesting that a large amount of time passed between the last clonal sweep and diagnosis, during which mutations continued to accumulate. This observation was robust to assumed acceleration parameter values between 1× and 20× (Supplementary Fig. 25). We also observed a lower age of appearance of the MRCA in piano (median: 60.4 years), particularly the piano with carcinoid histology (median: 55.0 years) compared to forte (median: 63 years) (all piano: $P=0.038$; piano carcinoids: $P=0.062$; Fig. 7c), which requires further confirmation in larger future studies.

Impact of molecular pathways on survival. Cases with *TP53* mutations or *MDM2* amplifications had poor survival (hazard ratio (HR) = 2.9, 95% CI = 1.6–5.2, $P=4.5\times10^{-4}$; Supplementary Fig. 26a,b), as reported previously in LCINS[43] and NSCLC[44], with a suggestive stronger impact of *TP53* mutations compared to *MDM2* amplifications (Fig. 8a). Similarly, *EGFR* mutations, *CHEK2* LOH, 22q loss and 15q loss were associated with poor survival (Fig. 8b–e). A risk score calculated as the mutational burden of these five independent genomic alterations (Fig. 8f) showed an increment of

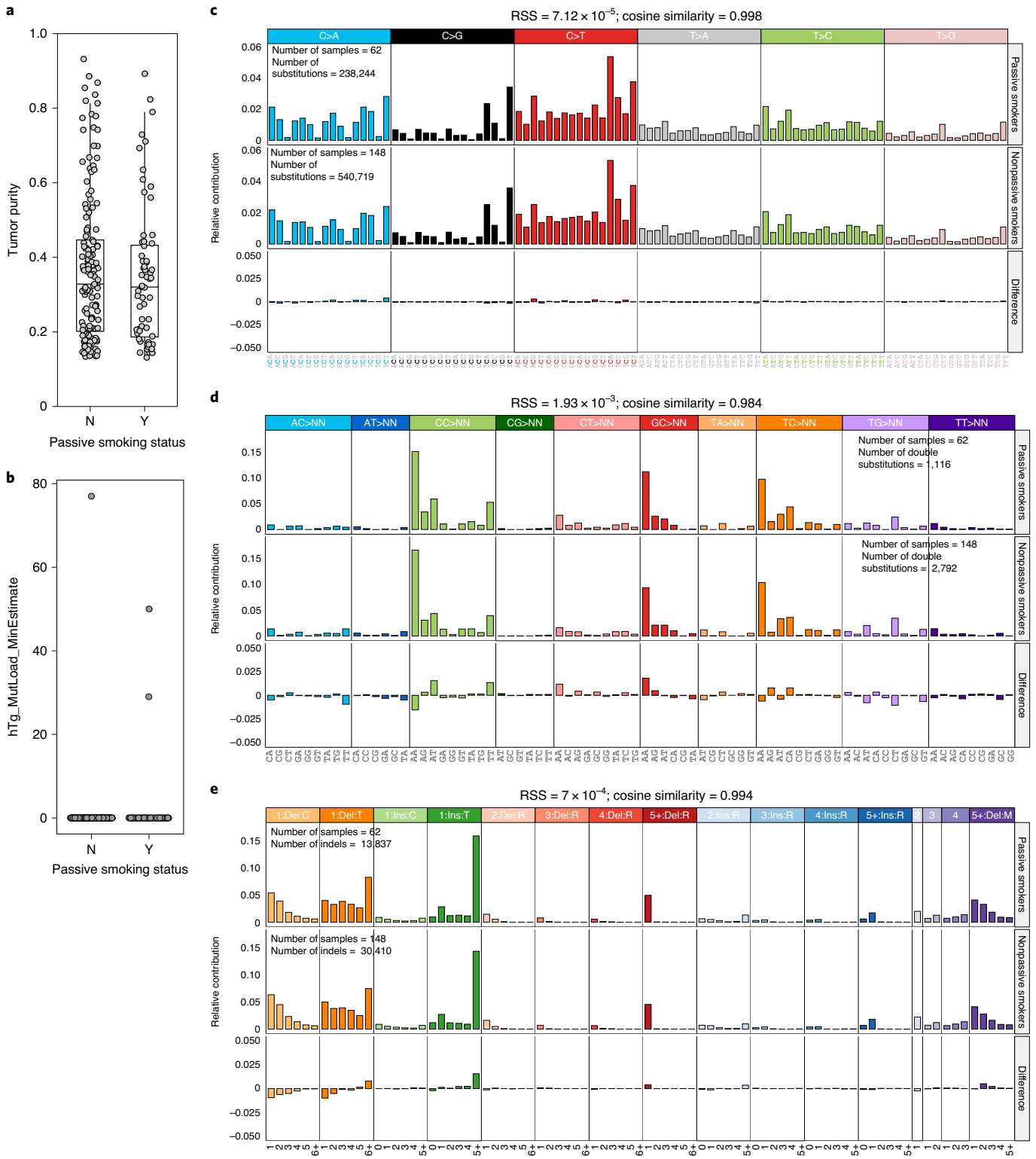


Fig. 5 | Comparison of mutational spectra between passive and nonpassive smokers in Sherlock-*Lung*. **a**,**b**, Identification of tumor purity (**a**) and alkylation-induced mutagenesis (hTg→hGg signature) (**b**) between passive (Y, $n=62$) and nonpassive smokers (N, $n=148$). **c**–**e**, Mutational spectra comparison of SBS (**c**), DBS (**d**) and indels (**e**) between passive and nonpassive smokers.

mortality risk for each genomic alteration of approximately 1.9 (95% CI = 1.5–2.4, $P = 3.7 \times 10^{-7}$).

Interestingly, no significant association was found between RTK–Ras status and overall survival (Supplementary Fig. 26c). However, there were strong differences in clinical association patterns across different genes in the pathway (Fig. 8b). Patients with *ERBB2* mutations had poor overall survival (HR = 5.7, 95% CI = 1.6–20.4, $P = 7.2 \times 10^{-3}$), although >50% of *ERBB2*$^+$ tumors (4 out of 7) also harbored *TP53* alterations, requiring further confirmation in *ERBB2*$^+$/*TP53*$^-$ tumors. *KRAS* mutations and *ALK*

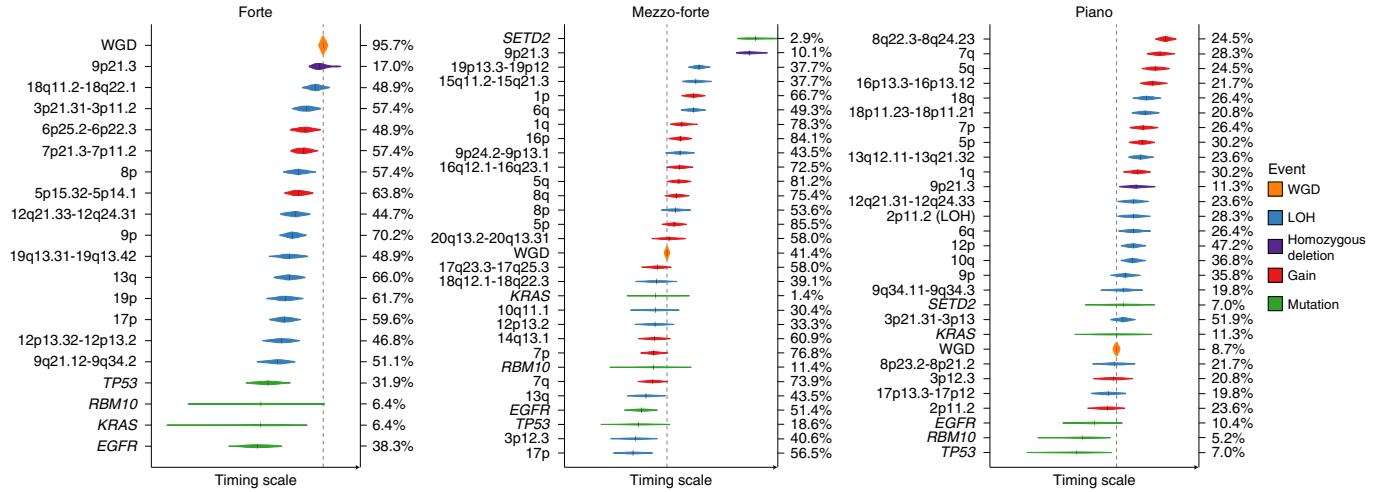


Fig. 6 | Diagrams of estimated ordering of significant SCNAs (including chromosome gains/losses and mutations) relative to WGD in three lung cancer subtypes based on the SCNA clusters forte, mezzo-forte and piano. The size of the violin plots denotes the uncertainty of timing for specific events across all samples; the short black solid lines represent the median time. The vertical dashed line indicates the median time for WGD events. Ordering of genomic events was based on the PlackettLuce package (v.0.2.2) model with 95% CIs. The frequency of each event is labeled on the *y* axis.

fusions were also associated with poor survival but not significantly. In contrast, patients with *MET*-altered tumors had better overall survival than the RTK–Ras$^-$ group. The small number of patients with both *TP53*-deficient and RTK–Ras$^-$ tumors ($n=8$) had poorer survival (HR = 5.3, 95% CI = 1.8–15.2, $P=2.0\times10^{-3}$; Supplementary Fig. 26d).

Patients with piano tumors had overall better survival (HR = 0.52, 95% CI = 0.3–0.9, $P=0.03$), particularly patients with carcinoids (HR = 0.24, 95% CI = 0.06–1.0, $P=0.05$), as did patients with *SETD2*$^+$ tumors (HR = 0.13, 95% CI = 0.02–1, $P=0.05$) (Supplementary Fig. 26e–g).

Discussion

WGS of 232 LCINS samples revealed 3 subtypes based on SCNAs and profound differences from adenocarcinomas in smokers. Whereas WGD is observed in over 60% of LUADs in smokers[42,45] and is considered to be a major driver of aggressive lung adenocarcinomas[46,47], it occurs in 36% of LCINS overall but in 95.7% of the forte subtype. While mezzo-forte is enriched for specific chromosomal arm-level amplifications and has frequent *EGFR* mutations, tumors in the quiet piano have a low mutational burden, infrequent WGD, small numbers of known drivers and a larger proportion of subclonal mutations indicative of extensive intratumor heterogeneity.

Forte tumors and tumors from passive smokers had shorter telomeres than their matched normal samples, while piano had longer telomeres, suggesting fewer cell divisions. *TERT* was amplified in only 11.6% tumors and had promoter mutations in only 0.9%; they were rarely in piano, excluding a major role for TERT reactivation in TL elongation.

Notably, we found no major difference between passive and nonpassive smokers for mutational signatures or mutation types, while we observed a few tumors with diesel exhaust signatures. Simulation studies showed that smoking-related mutations in the 62 tumors from passive smokers had to be below the detection threshold of 15%. It is possible that SBS4 is present in some passive smokers below this mutation threshold. Secondhand tobacco smoke has been causally linked to lung cancer[48] but it is a weak carcinogen compared to active smoking[48,49] and may also act through alternative tumorigenic processes and selective constraints[50]. Larger studies including highly exposed cases and in vitro or animal models are needed to definitively characterize the tumors arising from these exposures.

The long telomeres, low growth rate suggested by the occurrence of MRCA approximately a decade before tumor diagnosis, scarcity and heterogeneity of driver mutations, low mutation rate, high intratumoral heterogeneity and paucity of SBS18 indicating low ROS activity[51], are all consistent with piano tumors being derived from adult stem cells that have exited their quiescent state[52,53]. Driver genes specifically mutated in piano also suggest stem-like features. Oncogenic mutations in *KRAS*, the most frequently mutated driver gene in piano, have been shown to induce proliferation of bronchioalveolar stem cells, giving rise to lung adenocarcinoma[54]. Similarly, *KRAS*[55,56] and *UBA1* (ref. [57]) have important regulatory roles in hematopoietic and pluripotent stem cells. The presence of fusions and germline variants in *RET* (as well as mutations in *NKX2*, a regulator of *RET*[58]) uniquely in piano suggests a role for *RET* in these tumors. *RET* expression and activity are enriched in human hematopoietic stem cells[59] and are involved in murine hematopoietic stem cell regulation[60]. Notably, *ARID1A* is essential for telomere cohesion[61]; deleting *Arid1a* in mice greatly enhances the ability to regenerate organ tissues[62–64]. *ARID1A* depletion in humans promotes cells to enter the cell cycle[65]. Mutations in *ARID1A*, as well as *NOTCH1*, another gene whose signaling has a role in stem cell expansion and progenitor cell survival[66], have been found in normal and near-normal bronchial epithelial cells from former smokers[67], which are also characterized by long telomeres and polyclonal origins. Notably, alterations in *KRAS, UBA1, RET* and *ARID1A* were mutually exclusive in piano. Hypothetically, mutations in *NOTCH1*, *ARID1A* or other genes with similar function could promote exit from a quiescent cell state, resulting in high intratumoral heterogeneity, and could drive some of the tumors with no detected known cancer driver gene mutations or fusions. Carcinoids and LUAD in piano would then represent tumors diagnosed before acquisition of a dominant clone. Using RNA sequencing (RNA-seq) for an orthogonal assessment of stemness and cell of origin (Methods and Supplementary Note), we found that both a 'development score', incorporating expression of the *SOX2*, *SOX9* and *HMGA2* genes[68–70], and a marker of basal cells suggesting lineage infidelity[71] were higher in piano (Supplementary Fig. 27), which is consistent with piano representing a stem cell-like state. Larger studies are needed to verify the WGS-based stemness hypothesis, possibly using single-cell RNA-seq and methylome analyses, particularly in tumors with no apparent drivers.

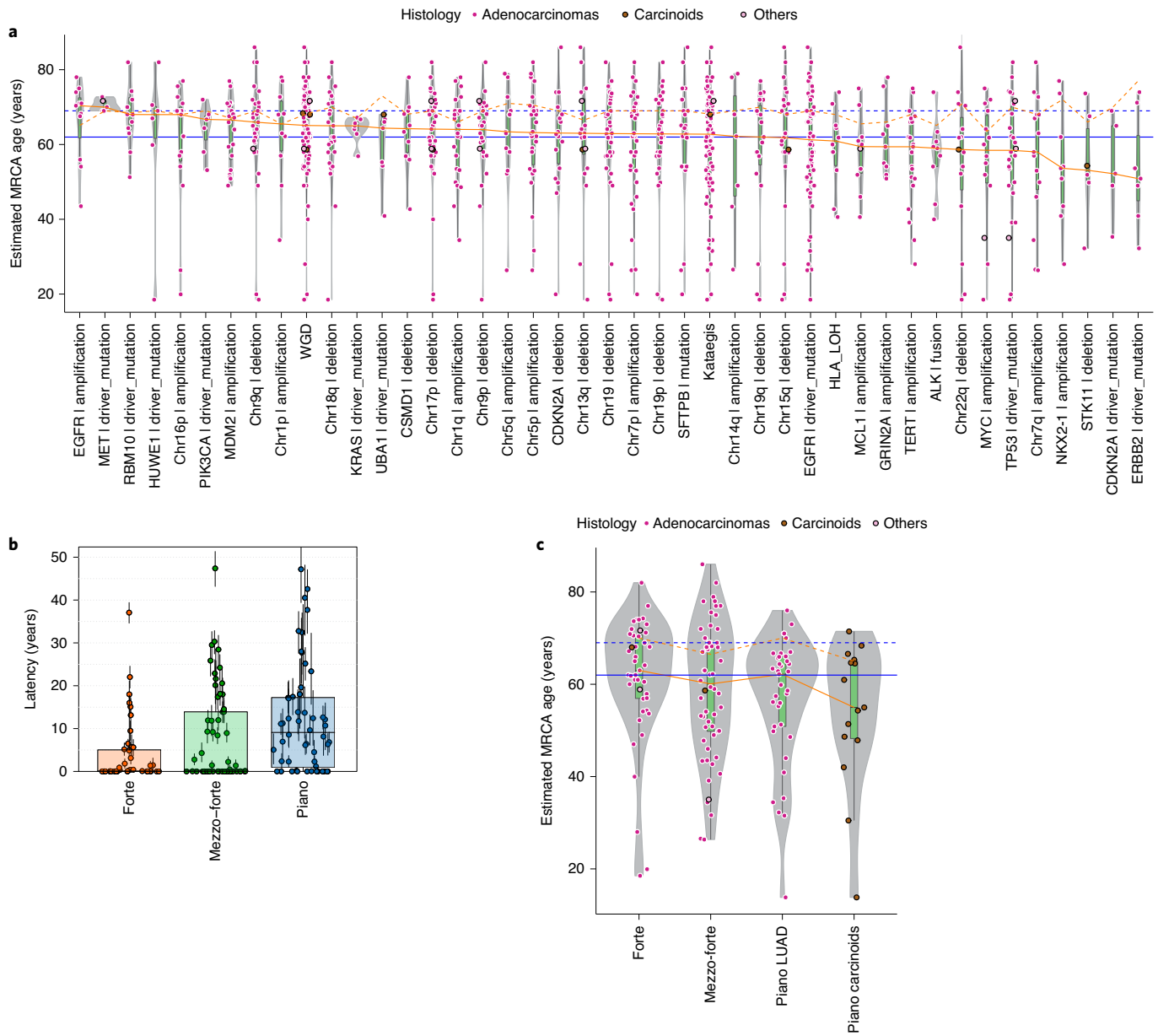


Fig. 7 | Reconstruction of the evolutionary history of LCINS. a, Estimated age at which the MRCA emerged in tumors (*y* axis), grouped by genomic alterations or features (*x* axis, frequency >3%) as shown in Fig. 2. The color of each dot represents the tumor histological subtype. The orange solid and dashed lines indicate the median estimated MRCA age and the median age at diagnosis in the same group, respectively. The blue solid and dashed lines indicate the median estimated MRCA age and median age at diagnosis in all samples, respectively. **b**, Box plots showing the latency between MRCA and age at diagnosis based on a 1× acceleration rate across forte, mezzo-forte and piano subtypes with 95% CIs for each tumor. **c**, Like **a**, estimated MRCA age among SCNA subtypes: forte; mezzo-forte; piano LUAD; and piano carcinoids. **a**–**c**, The center lines show the median; the box limits indicate the 25th and 75th percentiles; the whiskers extend 1.5 times the IQR from the 25th and 75th percentiles.

The founder cells of piano appear around a decade before diagnosis and provide an optimal time window for early detection. In contrast, driver gene mutations and WGD or gross SCNAs in the forte and mezzo-forte subtypes are generally clonal, with later onset followed by rapid expansion of a single ancestral cell. Their clonal nature could facilitate identification with a single biopsy and successful treatment.

Currently, treatments targeting the most recurrent genomic alterations in forte and mezzo-forte are available or are under investigation in clinical trials, namely for *TP53*[72] or *MDM2-TP53* interaction[73] and for mutations in *EGFR* or *ERBB2* (refs. [74–77]), genes that conveyed the poorest survival among the RTK–Ras pathway, and even for tumors with both *TP53* deficiency and RTK–Ras mutations (21% of our tumors)[78]. Together with *TP53* and *EGFR* alterations, tumors with loss of chromosomes 22q and 15q or *CHEK2* LOH were frequently identified, particularly in the forte subtype. A twofold higher mortality risk was estimated for each of these five independent genomic alterations, suggesting that compounds targeting bystander genes that are deleted together with tumor suppressor genes in chromosome arm losses (collateral lethality)[79,80] should be explored in these subtypes. Moreover, >15% of tumors had LOH of an HRD-associated gene. Targeting these genes could be a promising therapeutic option to explore. Moreover, mutations in HRD-associated genes could act as predictors of immune

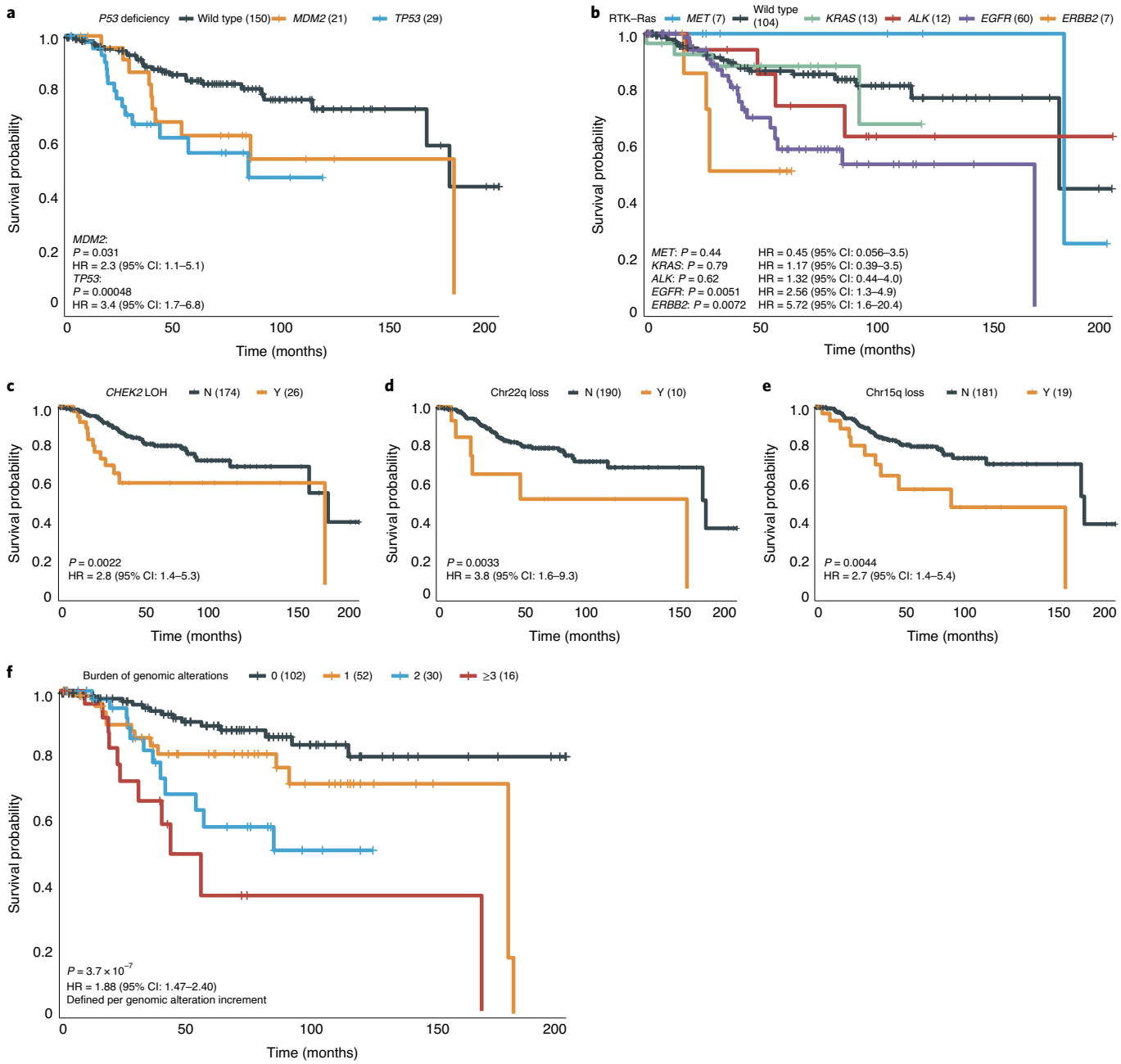


Fig. 8 | Association between genomic aberrations and clinical outcomes in never smoker patients with lung cancer. a–f, Kaplan–Meier survival curves for overall survival stratified by *TP53* mutations and *MDM2* amplifications (**a**), activation of individual driver genes in the RTK–Ras pathway (**b**), *CHEK2* LOH (**c**), Chr22q and Chr15q loss (**d**,**e**) and risk score based on the burden of five genomic alterations (TP53 deficiency, *CHEK2* LOH, Chr22q loss, Chr15q loss, and *EGFR* mutations) (**f**). *P* values for significance and HRs of the difference were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex and tumor stage. No multiple-testing correction was applied. Y, with aberration; N, without aberration. The numbers in brackets indicate the number of patients.

checkpoint inhibitor response[81], broadening the options for treatment for this subgroup. In contrast, piano has a scarcity of driver mutations, offering limited targets for therapeutic intervention. Furthermore, due to low TMB[82–84] and HLA LOH[20], these patients may not benefit from immunotherapy. However, targeting the KRAS[85] and stem cell-associated signaling pathways[51,86], or regulating the stem cell microenvironment[87], are promising for this subtype.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00920-0.

Received: 25 November 2020; Accepted: 15 July 2021;
Published online: 6 September 2021

References

1. *The Cancer Atlas: Lung Cancer* (American Cancer Society, 2021); https://canceratlas.cancer.org/the-burden/lung-cancer/

2. Cho, J. et al. Proportion and clinical features of never-smokers with non-small cell lung cancer. *Chin. J. Cancer* **36**, 20 (2017).
3. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
4. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
5. Chen, J. et al. Genomic landscape of lung adenocarcinoma in East Asians. *Nat. Genet.* **52**, 177–186 (2020).
6. Govindan, R. et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).
7. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
8. Lee, J. J.-K. et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* **177**, 1842–1857.e21 (2019).
9. Shi, J. et al. Somatic genomics and clinical features of lung adenocarcinoma: a retrospective study. *PLoS Med.* **13**, e1002162 (2016).
10. Wang, C. et al. Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nat. Commun.* **9**, 2054 (2018).
11. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
12. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
13. Wu, K. et al. Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. *Nat. Commun.* **6**, 10131 (2015).
14. Carrot-Zhang, J. et al. Whole-genome characterization of lung adenocarcinomas lacking the RTK/RAS/RAF pathway. *Cell Rep.* **34**, 108707 (2021).
15. Landi, M. T. et al. Tracing lung cancer risk factors through mutational signatures in never smokers: the Sherlock-Lung study. *Am. J. Epidemiol.* **190**, 962–976 (2021).
16. Skoulidis, F. et al. Co-occurring genomic alterations define major subsets of *KRAS*-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. *Cancer Discov.* **5**, 860–877 (2015).
17. Moll, U. M. & Petrenko, O. The MDM2-p53 interaction. *Mol. Cancer Res.* **1**, 1001–1008 (2003).
18. Wala, J. A. et al. Selective and mechanistic sources of recurrent rearrangements across the cancer genome. Preprint at *bioRxiv* https://doi.org/10.1101/187609 (2017).
19. Reznik, E. et al. Mitochondrial DNA copy number variation across human cancers. *eLife* **5**, e10769 (2016).
20. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271.e11 (2017).
21. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
22. Moudry, P. et al. Ubiquitin-activating enzyme UBA1 is required for cellular response to DNA damage. *Cell Cycle* **11**, 1573–1582 (2012).
23. Martínez-Jiménez, F. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
24. Huang, K.-L. et al. Pathogenic germline variants in 10,389 adult cancers. *Cell* **173**, 355–370.e14 (2018).
25. Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
26. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
27. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
28. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
29. Jager, M. et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019).
30. Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.* **3**, 421 (2020).
31. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
32. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
33. Tokiwa, H. & Sera, N. Contribution of nitrated polycyclic aromatic hydrocarbons in diesel particles to human lung cancer induction. *Polycycl. Aromat. Compd.* **21**, 231–245 (2000).
34. Saini, N. et al. Mutation signatures specific to DNA alkylating agents in yeast and cancers. *Nucleic Acids Res.* **48**, 3692–3707 (2020).
35. 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).
36. Barthel, F. P. et al. Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
37. Feuerbach, L. et al. TelomereHunter—in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, 272 (2019).
38. Davies, H. et al. HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
39. Zhao, E. Y. et al. Homologous recombination deficiency and platinum-based therapy outcomes in advanced breast cancer. *Clin. Cancer Res.* **23**, 7521–7530 (2017).
40. Letouzé, E. et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
41. 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).
42. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
43. Halvorsen, A. R. et al. *TP53* mutation spectrum in smokers and never smoking lung cancer patients. *Front. Genet.* **7**, 85 (2016).
44. Gu, J. et al. *TP53* mutation is associated with a poor clinical outcome for non-small cell lung cancer: evidence from a meta-analysis. *Mol. Clin. Oncol.* **5**, 705–713 (2016).
45. López, S. et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
46. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
47. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
48. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, International Agency for Research on Cancer. *A Review of Human Carcinogens: Personal Habits and Indoor Combustions* (International Agency for Research on Cancer, 2012).
49. United States Public Health Service. Office of the Surgeon General. *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General* (US Department of Health and Human Services, Public Health Service, Office of the Surgeon General, 2006).
50. Lopez-Bigas, N. & Gonzalez-Perez, A. Are carcinogens direct mutagens? *Nat. Genet.* **52**, 1137–1138 (2020).
51. Cho, I. J. et al. Mechanisms, hallmarks, and implications of stem cell quiescence. *Stem Cell Reports* **12**, 1190–1200 (2019).
52. Fukada, S.-I., Ma, Y. & Uezumi, A. Adult stem cell and mesenchymal progenitor theories of aging. *Front. Cell Dev. Biol.* **2**, 10 (2014).
53. Li, L. & Clevers, H. Coexistence of quiescent and active adult stem cells in mammals. *Science* **327**, 542–545 (2010).
54. Kim, C. F. B. et al. Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121**, 823–835 (2005).
55. Van Meter, M. E. M. et al. K-Ras^{G12D} expression induces hyperproliferation and aberrant signaling in primary hematopoietic stem/progenitor cells. *Blood* **109**, 3945–3952 (2007).
56. Kubara, K. et al. Status of KRAS in iPSCs impacts upon self-renewal and differentiation propensity. *Stem Cell Reports* **11**, 380–394 (2018).
57. Bax, M. et al. The ubiquitin proteasome system is a key regulator of pluripotent stem cell survival and motor neuron differentiation. *Cells* **8**, 581 (2019).
58. Leon, T. Y. Y. et al. Transcriptional regulation of *RET* by Nkx2-1, Phox2b, Sox10, and Pax3. *J. Pediatr. Surg.* **44**, 1904–1912 (2009).
59. Grey, W. et al. Activation of the receptor tyrosine kinase, RET, improves long-term hematopoietic stem cell outgrowth and potency. *Blood* **136**, 2535–2547 (2020).
60. Fonseca-Pereira, D. et al. The neurotrophic factor receptor RET drives haematopoietic stem cell survival and function. *Nature* **514**, 98–101 (2014).
61. Zhao, B. et al. ARID1A promotes genomic stability through protecting telomere cohesion. *Nat. Commun.* **10**, 4067 (2019).
62. Sun, X. et al. Suppression of the SWI/SNF component *Arid1a* promotes mammalian regeneration. *Cell Stem Cell* **18**, 456–466 (2016).
63. van der Vaart, A. & van den Heuvel, S. Switching on regeneration. *Stem Cell Investig.* **3**, 41 (2016).
64. Wu, S., Zhang, R. & Bitler, B. G. Arid1a controls tissue regeneration. *Stem Cell Investig.* **3**, 35 (2016).
65. Nagl, N. G. Jr, Wang, X., Patsialou, A., Van Scoy, M. & Moran, E. Distinct mammalian SWI/SNF chromatin remodeling complexes with opposing roles in cell-cycle control. *EMBO J.* **26**, 752–763 (2007).
66. Chiba, S. Notch signaling in stem cell systems. *Stem Cells* **24**, 2437–2447 (2006).
67. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

68. Maeda, Y., Davé, V. & Whitsett, J. A. Transcriptional control of lung morphogenesis. *Physiol. Rev.* **87**, 219–244 (2007).
69. 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).
70. Singh, I. et al. *Hmga2* is required for canonical WNT signaling during lung development. *BMC Biol.* **12**, 21 (2014).
71. Laughney, A. M. et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat. Med.* **26**, 259–269 (2020).
72. Duffy, M. J. et al. p53 as a target for the treatment of cancer. *Cancer Treat. Rev.* **40**, 1153–1160 (2014).
73. Shaikh, M. F. et al. Emerging role of MDM2 as target for anti-cancer therapy: a review. *Ann. Clin. Lab. Sci.* **46**, 627–634 (2016).
74. 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).
75. Harvey, R. D., Adams, V. R., Beardslee, T. & Medina, P. Afatinib for the treatment of *EGFR* mutation-positive NSCLC: a review of clinical findings. *J. Oncol. Pharm. Pract.* **26**, 1461–1474 (2020).
76. Park, K. et al. Afatinib versus gefitinib as first-line treatment of patients with *EGFR* mutation-positive non-small-cell lung cancer (LUX-Lung 7): a phase 2B, open-label, randomised controlled trial. *Lancet Oncol.* **17**, 577–589 (2016).
77. Shen, X. et al. A systematic analysis of the resistance and sensitivity of $HER2^{YVMA}$ receptor tyrosine kinase mutant to tyrosine kinase inhibitors in HER2-positive lung cancer. *J. Recept. Signal Transduct. Res.* **36**, 89–97 (2016).
78. Miyazaki, M. et al. The p53 activator overcomes resistance to ALK inhibitors by regulating p53-target selectivity in ALK-driven neuroblastomas. *Cell Death Discov.* **4**, 56 (2018).
79. Dey, P. et al. Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature* **542**, 119–123 (2017).
80. Muller, F. L., Aquilanti, E. A. & DePinho, R. A. Collateral lethality: a new therapeutic strategy in oncology. *Trends Cancer* **1**, 161–173 (2015).
81. Hsiehchen, D. et al. DNA repair gene mutations as predictors of immune checkpoint inhibitor response beyond tumor mutation burden. *Cell Rep. Med.* **1**, 100034 (2020).
82. Rizvi, N. A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
83. Hellmann, M. D. et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
84. Ready, N. et al. First-line nivolumab plus ipilimumab in advanced non-small-cell lung cancer (CheckMate 568): outcomes by programmed death ligand 1 and tumor mutational burden as biomarkers. *J. Clin. Oncol.* **37**, 992–1000 (2019).
85. Canon, J. et al. The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).
86. Yang, L. et al. Targeting cancer stem cell pathways for cancer therapy. *Signal Transduct. Target. Ther.* **5**, 8 (2020).
87. Medema, J. P. & Vermeulen, L. Microenvironmental regulation of stem cells in intestinal homeostasis and cancer. *Nature* **474**, 318–326 (2011).

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [2]Institut Universitaire de Cardiologie et de Pneumologie de Québec, Laval University, Quebec City, Quebec, Canada. [3]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [4]Cancer Genomics Section, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. [5]Stem Cell Program and Divisions of Hematology/Oncology and Pulmonary Medicine, Boston Children's Hospital, Boston, MA, USA. [6]Westat, Rockville, MD, USA. [7]Institute for Research in Biomedicine Barcelona, The Barcelona Institute of Science and Technology, Barcelona, Spain. [8]Department of Medicine, Division of Medical Genetics, University of California San Diego, San Diego, CA, USA. [9]Laboratory of Oncology, Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy. [10]Laboratory of Clinical and Experimental Pathology, University Hospital Federation OncoAge, Nice Hospital, University Côte d'Azur, Nice, France. [11]Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy. [12]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy. [13]Smilow Cancer Hospital, Yale-New Haven Health, New Haven, CT, USA. [14]Yale Comprehensive Cancer Center, New Haven, CT, USA. [15]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [16]National Institute for Health Research Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [17]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [18]Department of Bioscience, Biotechnology and Biopharmaceutics, University of Bari, Bari, Italy. [19]Department of Urology, Istituto di Ricovero e Cura a Carattere Scientifico Regina Elena National Cancer Institute, Rome, Italy. [20]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [21]Department of Pathology and Laboratory Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. [22]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle, NC, USA. [23]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, Research Triangle, NC, USA. [24]Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California San Diego, San Diego, CA, USA. [25]Developmental Therapeutics Branch and Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. [26]Department of Genetics, Harvard Medical School, Boston, MA, USA. [27]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [28]Department of Molecular Medicine, Laval University, Quebec City, Quebec, Canada. [29]Manchester Cancer Research Centre, The University of Manchester, Manchester, UK. ✉e-mail: landim@mail.nih.gov

Methods

A detailed description of the methods used in this paper and many additional results are described in the Supplementary Note. In this article, we summarize the key aspects of the analysis.

Ethics declarations. Since the National Cancer Institute only received de-identified samples and data from collaborating centers, had no direct contact or interaction with the study participants and did not use or generate identifiable private information, Sherlock-*Lung* has been determined to constitute 'non-human subject research' based on the federal Common Rule (45 CFR 46; https://www.ecfr.gov/cgi-bin/ECFR?page=browse).

Collection of lung cancer samples. Fresh-frozen tumor tissue and matched germline DNA from whole-blood samples or fresh-frozen normal lung tissue sampled approximately 3 cm from the tumor were obtained from 256 treatment-naïve lung cancer patients from five institutions/centers (Supplementary Note). Among the 256 samples, 20 were excluded after quality check and 4 were excluded based on mutational signatures analysis (Supplementary Note). The resulting 232 samples and associated demographic and clinical data were included in the final analysis. For these 232 individuals, the mean age at lung cancer diagnosis was 64.8 years (range: 21–86); 75.4% of patients were female. To confirm the ancestry of these patients, we estimated the admixture proportions based on WGS data using the fastNGSadmix v.1.0[88] tool.

Of the 232 tumors, 189 were adenocarcinomas, 36 carcinoids, 5 sarcomatoid carcinomas or undifferentiated non-small-cell carcinomas with sarcomatoid features and 2 squamous cell carcinomas. Three pathologists reviewed the histological diagnoses. Histological images can be found at https://episphere.github.io/svs. All 232 matched tumor and germline samples underwent DNA WGS. Of these, 35 (all adenocarcinomas) also underwent RNA-seq.

Genome-wide somatic variant calling. The analysis-ready BAM files were processed using four different algorithms, including MuTect[89], MuTect2, Strelka v.2.9.0 (ref. [90]) and TNscope[91]. To improve the performance of variant calling, we used Sentieon's genomics package v.201808.03 to run MuTect, MuTect2 and TNscope. Only those SNVs that passed calling by a minimum of three algorithms were kept. To reduce false positive calling, we applied an in-house filtering script (https://github.com/xtmgah/Sherlock-Lung) similar to our previous publication[92]. To summarize, variant calling was considered only at the genome positions with (1) read depth >12 in tumor and >6 in normal samples and (2) variant read count >5 in tumor and variant allele frequency <0.02 in normal samples. To remove possible germline variants from the called somatic variants, somatic variants were filtered against the dbSNP build 138, 1000 Genomes (phase 3 v.5), ExAC v.0.3.1, Genome Aggregation Database (gnomAD) v.2.1.1 (ref. [93]) databases and an in-house Italian germline variant database from the EAGLE WES study (database of Genotypes and Phenotypes (dbGAP) accession no. phs002496.v1.p1) for commonly occurring SNPs (somatic variant frequency <0.001). The filtered variants were annotated with Oncotator v.1.9.1.0 (ref. [94]) and ANNOVAR v.2019-10-24[95]. For the indel calling, only variants called by three algorithms were kept (MuTect2, TNscope and Strelka). The UPS-indel[96] algorithm was used to compare and combine different indel call sets. Similar filtering steps as those used for SNV calling were also applied to indel calling. The final set of indels were left-normalized (left-aligned and trimmed) for the downstream analysis. Clustering of subclonal somatic mutations was analyzed using a Bayesian Dirichlet Process (DPClust v.2.28) as described previously[92,97,98]. Further details are available in the Supplementary Note.

Germline variant calling. Final BAM files from paired normal samples were used to call germline variants using the GATK Haplotyper algorithm in Sentieon's genomics package v.201808.03. Default parameters or suggested input files, such as the most recent dbSNP VCF file were applied. The final joint callings from all normal samples were generated and annotated with ANNOVAR[95]. Strict filtering criteria were used to identify the potential pathogenic variants: (1) minor allele frequencies <0.05% in the gnomAD noncancer and non-Finnish European ancestry dataset v.2.1.1; (2) estimated CharGer score >4 to include the pathogenic or likely pathogenic variants based on the CharGer algorithm v.0.5.2 (ref. [99]). The default parameters for CharGer were used and the most damaging interpretation from ClinVar[100], excluding Online Mendelian Inheritance in Man and GeneReviews as submitters, was used for annotation; (3) variants predicted to have 'silent' functional activity were removed, including variants in the UTR, upstream or intron regions. All the final germline variants were manually inspected through the Integrative Genomics Viewer and suspicious variants were removed.

Driver gene discovery. The IntOGen pipeline v.2020.02.01[23], which combines seven state-of-the-art computational methods, was employed to detect signals of positive selection in the mutational pattern of genes across the cohort. Default parameters were used; in the postprocessing phase, the *CSMD3* gene was filtered out based on warnings provided by the pipeline. The 25 genes identified as drivers in the cohort were classified according to their mode of action in tumorigenesis (that is, tumor suppressor genes or oncogenes) based on the relationship between the excess of observed nonsynonymous and truncating mutations computed by dNdScv v.0.0.1[101] and their annotations in the Cancer Gene Census. In the case of *UBA1*, only the excess values were used. Genes with conflicting computed and annotated modes of action were labeled ambiguous. To identify potential driver mutations across the 24 cancer driver genes annotated in the Cancer Gene Census, we used boostDM from IntOGen pipeline with gene-tumor type-specific (LUAD) or more general models depending on their availability and accuracy[102].

SCNA analysis. We used the updated Battenberg v.2.2.8 algorithm[98] to estimate the clonality of each segmentation, tumor purity and ploidy (Supplementary Note). Unsupervised clustering of copy number profiles including both major clone and subclone segmentations were generated based on relative copy number $\log_2$ (copy number/Tumor_Ploidy) using the Euclidean distance and Ward's method. Recurrent copy number alterations from WGS at the gene level were identified using GISTIC v.2.0 (ref. [103]) based on the major clonal copy number for each segmentation (Supplementary Note).

WGD identification. Multiple methods were used to determine the genome doubling status for each tumor. First, tumors were considered to have undergone WGD if >50% of their autosomal genome had a major copy number (MCN) (that is, the more frequent allele in a given segment) ≥2 (ref. [46]). Also, the number of chromosomes with 50% of the segment with an MCN ≥ 2 had to be greater than 11. Next, we applied a modified version of the published WGD algorithm EstimateClonality v.1.0[104], where a *P* value was obtained using 10,000 simulations with observed probabilities of copy number events. For samples with ploidy ≤ 3 and ploidy = 4, *P* value thresholds of 0.001 and 0.05 were used, respectively. All samples were classified as genome-doubled if the ploidy exceeded 4. Tumors were determined to have undergone WGD if the tumors met the criteria for WGD for both methods. Finally, to improve WGD calling, we manually checked the Battenberg CNA profile to resolve tumors with ambiguous WGD calling (for example, MCN close to 0.5), evaluating features such as presence of multiple copy losses after WGD (total copy of 3) and/or LOH events having 2:0 copy number state. For the chronological reconstruction of genomic aberrations, we limited the WGD samples with average ploidy > 3.

SV calling and clustering. The Meerkat algorithm v.0.185[105] was used to call somatic SVs and estimate the corresponding genomic positions of breakpoints (Supplementary Note). 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[92]. Driver oncogenic fusions were selected from SVs based on the driver gene list[21] and an oncogenic fusion list previously reported in LUAD[8]. We selected the fusions with the following SV features in the 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 fusions were found.

We used the algorithm developed by Li et al.[106] to cluster the SVs in each sample. The algorithm groups SVs into clusters based on the proximity of breakpoints, the number of events in that cluster region and the size distribution of those events. A cluster contains SVs that have arisen from the same event and are significantly closer than expected by chance, given the orientation and the number of SVs in that patient. In addition, to visualize the hotspots of the breakpoints, we counted the number of breakpoints across the whole genome using a 5-Mb window. A similar approach was also applied to visualize the kataegis hotspots.

TL estimation. We estimated TL in kb using TelSeq v.0.0.2[107]. We used seven as the threshold for the number of TTAGGG/CCCTAA repeats in a read for the read to be considered telomeric. The TelSeq calculation was done individually for each read group within a sample and the total number of reads in each read group was used as weight to calculate the average TL for each sample. To validate the estimation of TL by TelSeq, telomere content was quantified using TelomereHunter v.1.1.0[37] using ten telomere variant repeats including TCAGGG, TGAGGG, TTGGGG, TTCGGG, TTTGGG, ATAGGG, CATGGG, CTAGGG, GTAGGG and TAAGGG.

To compare TL in Sherlock-*Lung* with previous studies, we collected the TL estimation from the same algorithms across the TCGA cohort and applied the same linear mixed model to predict the mean TL as described by Barthel et al.[36].

Mutational signature analysis. Mutational signatures were analyzed by the updated computational framework SigProfiler[26,108]. SigProfilerExtractor v.0.0.5.77 with default parameters was used to perform both de novo extraction and decomposition to the known global COSMIC mutation signatures (v.3). 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.

To investigate the endogenous and exogenous mutational processes in our Sherlock-*Lung* study, we collected 4 mutational signature sets according to the likely etiologies, including 65 COSMIC SBS mutational signatures, 22 COSMIC SBS endogenous signatures, 53 environmental mutagen signatures[32] and 75 combined endogenous and exogenous mutational signatures (see Supplementary Table 7 for the signatures included). We then performed the SBS mutational

signature analyses as described above. To maximally deconvolute all mutations to these global signatures in SigProfilerExtractor, we decreased the cosine similarity threshold for de novo mutational signatures until no new mutational signatures were found. Among these four mutational signature sets, we compared the cosine similarity between the reconstructed mutational profiles with the original mutational profiles for each sample.

Analysis of passive smoking. To investigate tumor mutational patterns between passive and nonpassive smokers, we first excluded four hypermutated tumors (two from passive smokers, one from a nonpassive smoker and one with unknown passive smoking exposure) driven by APOBEC mutagenesis (TMB > 8 Mut Mb^{-1}). We then compared each mutation type among SBS, DBS and indels between passive and nonpassive smokers using the Mann–Whitney *U* test followed by multiple-testing correction using the Benjamini–Hochberg method. To quantify and visualize differences in the overall mutational patterns between passive and nonpassive smokers, we combined all mutations in each tumor group into a single profile and estimated their cosine similarity and residual sum of squares (RSS). As a sensitivity analysis, we replicated these analyses within the samples from two studies (EAGLE and Yale) with high-quality passive smoking exposure data. The EAGLE study had details on exposure during childhood, during adulthood at home and during adulthood at work. Thus, we created a score from the highest exposure (1, during all three periods) to the lowest (4, only one exposure setting during adulthood). We then extracted the mutational patterns across the groups and estimated the cosine similarity of the two extremes (1 and 4).

HRD by HRDetect. We applied HRDetect to assess the HRD as described in previous studies[25,38,39]. Mutations including SNVs and indels, Battenberg segmentation profile, SVs and tumor purity and ploidy were included for HRDetect. HRDetect scores were computed by aggregating 6 features associated with HRD including SNV signature 3, SNV signature 8, SV signature 3, SV signature 5, HRD index from copy number profile and the fraction of deletions with microhomology. All features were normalized and log-transformed. A logistic model was used to predict the HRDetect scores using previously trained data[38].

Assessment of LOH. HLA LOH was identified by the LOHHLA algorithm v.1.0[20]. Patient-specific HLA genotypes were inferred by POLYSOLVER v.1.0[109] based on the normal samples. Then, tumor and normal BAM files, HLA calls, HLA FASTA file and tumor purity and ploidy were used as input to LOHHLA. A copy number < 0.5 was classified as subject to loss and thereby indicative of LOH. Allelic imbalance was determined if $P < 0.01$ using a paired *t*-test between the two distributions.

LOH analysis for HRD genes was based on the overlapping gene location with copy number profile by Battenberg. LOH segmentation was called if the clonal minor copy number was 0. The HRD gene list was based on a previous publication[25].

Prediction of chronological timing. We adopted the approach from the PCAWG[42] to estimate the elapsed time between the appearance of the MRCA and the last observable subclone in our Sherlock-*Lung* study. Briefly, the number of clock-like CpG>TpG mutations in an NpCpG context was counted for all tumors, accounting for tumor ploidy as well as clonal and subclonal mutations. Tumors with no age information, insufficient number of clonal and subclonal clock-like mutations (<50 mutations per sample) to estimate mutation rate, abnormal mutation rate or high fraction of APOBEC-associated mutations (SBS2 and SBS13 fraction > 70%) were excluded from the analysis as advocated previously[8,42], leaving 153 samples for this analysis. The latency of the MRCA was estimated for each tumor, adopting an estimated tumor acceleration rate of 1×. We subtracted the estimated latency from age at diagnosis to obtain the real-time age at which the MRCA likely emerged, grouping tumors by the presence of specific genomic alterations or features with >3% frequency (such as SCNA subtypes, groups with RTK–Ras alterations, *TP53* deficiency, *ALK* fusions and *ARID1A* mutations). Significant differences between subgroups were assessed using a Wilcoxon rank-sum test.

Timing model of ordering events. Mutational drivers and CNAs were simultaneously incorporated into the timing model based on the clonality of the events. For CNAs, Battenberg copy number calls were used to assign the clonality of CNAs (whether the cancer cell fraction (CCF) = 1 or <1), describe the type of CNA (that is, gain, LOH and homozygous deletion) and whether WGD occurred in the overall copy number profile. To include only recurrent regions, first, CNA events of each type were piled up across all samples along the chromosomes to get the frequency landscape of each CNA type based on all observed breakpoints. Next, a permutation test ($n = 1{,}000$) followed by false discovery rate (FDR)-based multiple-testing correction was undertaken to identify regions that were significantly enriched above the random background copy change rate. The enriched regions that encompassed the HLA region (6p21) or specific to telomeric ends or present as a singleton were excluded. For each mutational driver (with ≥5% recurrence), the CCF of each variant was estimated by adjusting the variant allele frequency according to the CNA status of the locus and purity of the tumor sample as described previously[110]. Variants were then classified as clonal (CCF = 1) and subclonal (CCF < 1) using DPClust. All events were combined per sample and ordered based on the CCF. Where more than one tree could be inferred based on subclonal events, all possible trees were generated and randomly chosen in each iteration of the ordering events. To time the events based on the entire dataset, events were ordered based on clonality (randomized clonal events followed by a sampled tree of subclonal events) in each sample. To classify events regarding WGD, we used the estimated number of chromosomes bearing the mutation and major/minor copy number status to call pre-WGD and post-WGD mutations and CNA, respectively. The PlackettLuce model[111,112] for ordering partial rankings was implemented (https://github.com/hturner/PlackettLuce) to infer the order of events based on the ordering matrix of the entire dataset while allowing for unobserved events. This analysis was undertaken for 1,000 iterations to obtain the 95% CI of the timing estimate of each event. We repeated this analysis across the three subtypes of tumors based on SCNA clusters (forte, mezzo-forte and piano).

Statistical and survival analysis. All statistical analyses were performed using the R software v3.5 (https://www.r-project.org/). To investigate the functional relevance of potential driver mutations of each pair of genes, we performed mutual exclusivity analysis and co-occurrence analysis using a two-sided Fisher's exact test. $P < 0.05$ was considered statistically significant. If multiple testing was required, we applied the FDR correction based on the Benjamini–Hochberg method. For the survival analyses, a proportional-hazards model was used to investigate the associations between genomic features and overall survival, adjusting for age at diagnosis, sex and stage. The multiple-testing correction for survival analysis was performed based on 33 different genomic alteration events with at least 5% frequency including mutations, focal SCNA, arm-level SCNA and gene fusions. Genomic alterations were identified as significant if $Q < 0.1$. A risk score was calculated as the mutational burden of these significant independent genomic alterations; we then performed association between risk score and overall survival using the same method.

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

Data availability

The 232 normal and tumor-paired raw data (BAM files) of the WGS datasets have been deposited in the dbGaP under accession no. phs001697.v1.p1. Researchers will need to obtain authorization from the dbGaP to download these data. The RNA-seq raw data (FASTQ files) have been submitted to the Gene Expression Omnibus under accession no. GSE171415. The germline variant dataset from the EAGLE whole-exome sequencing study can be access at the dbGaP with accession no. phs002496.v1.p1. In addition, histological images of these tumors can be found at https://episphere.github.io/svs. Public datasets were used in this study including gnomAD v.2.1.1/ExAC v.0.3.1 (https://gnomad.broadinstitute.org/), 1000 Genomes (phase 3 v.5, https://www.internationalgenome.org/) and dbSNP (v.138, https://www.ncbi.nlm.nih.gov/snp/).

Code availability

The code for the WGS subclonal copy number caller can be found at https://github.com/Wedge-lab/battenberg (v.2.2.8). The code for somatic mutation filtering can be found at https://github.com/xtmgah/Sherlock-Lung. The code for the Dirichlet process-based methods for subclonal reconstruction of tumors can be found at https://github.com/Wedge-lab/dpclust (v.2.2.8). The code for the mutational signature analysis can be found at https://pypi.org/project/sigproextractor/ (SigProfilerExtractor v.0.0.5.77). The code for inferring the order of genomic events can be found at https://github.com/hturner/PlackettLuce (v.0.2-2). The code for the chronological timing analysis can be found at https://gerstung-lab.github.io/PCAWG-11/. The code for P-MACD can be found at https://github.com/NIEHS/P-MACD.

References

88. Jørsboe, E., Hanghøj, K. & Albrechtsen, A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33**, 3148–3150 (2017).
89. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
90. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
91. Freed, D., Pan, R. & Aldana, R. TNscope: accurate detection of somatic mutations with haplotype-based variant candidate detection and machine learning filtering. Preprint at *bioRxiv* https://doi.org/10.1101/250647 (2018).
92. Zhu, B. et al. The genomic and epigenomic evolutionary history of papillary renal cell carcinomas. *Nat. Commun.* **11**, 3096 (2020).
93. Karczewski, K. J. et al. The mutational constraints spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
94. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).

95. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
96. Hasan, M. S., Wu, X., Watson, L. T. & Zhang, L. UPS-indel: a universal positioning system for indels. *Sci. Rep.* **7**, 14106 (2017).
97. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
98. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
99. Scott, A. D. et al. CharGer: clinical Characterization of Germline variants. *Bioinformatics* **35**, 865–867 (2019).
100. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
101. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
102. Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *Nature* **596**, 428–432 (2021).
103. 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).
104. Dewhurst, S. M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
105. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
106. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
107. Ding, Z. et al. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42**, e75 (2014).
108. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
109. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
110. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
111. Luce, R. D. *Individual Choice Behavior: a Theoretical Analysis* (Wiley, 1959).
112. Plackett, R. L. The analysis of permutations. *Appl. Stat.* **24**, 193 (1975).

Acknowledgements

This work has been supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, and the Intramural Research Program of the National Institute of Environmental Health Sciences (project nos. Z01 ES050159 to S.H.W. and Z1AES103266 to D.A.G.), National Institutes of Health (NIH). This project was funded in whole or in part with federal funds from the National Cancer Institute, NIH, under contract nos. 75N91019D00024 and HHSN261201800001I. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government. The research was supported by the Wellcome Trust Core Award, grant no. 203141/Z/16/Z with funding from the National Institute for Health Research Oxford Biomedical Research Centre. L.B.A. is an Abeloff V scholar and he is personally supported by an Alfred P. Sloan Research Fellowship and a Packard Fellowship for Science and Engineering. Research at the L.B.A. laboratory was supported by a National Institute of Environmental Health Sciences grant no. R01ES032547. The views expressed are those of the authors and not necessarily those of the National Health Service, National Institute for Health Research or Department of Health. The collection of samples from the Institut Universitaire de Cardiologie et de Pneumologie de Québec (IUCPQ), Université Laval was supported by the IUCPQ Foundation. The GR Program 2010-2316264 supported L.A.M. for sample collection by the Istituto di Ricovero e Cura a Carattere Scientifico Fondazione Casa Sollievo della Sofferenza. A.L.M. is supported by a Damon Runyon Cancer Research Foundation postdoctoral fellowship (no. DRG:2368-19) and a Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund (no. 1019903). C.F.K. is supported in part by grant no. R35HL150876-01, the Thoracic Foundation, Ellison Foundation, American Lung Association (no. LCD-619492) and the Harvard Stem Cell Institute. N.L-B. acknowledges funding from the European Research Council (consolidator grant no. 682398). P.H. is supported in part by the Association pour la Recherche contre le Cancer (CANC'AIR GENExposomics project). This work has been supported in part by the Tissue Core at the H. Lee Moffitt Cancer Center & Research Institute, a comprehensive cancer center designated by the National Cancer Institute and funded in part by a Moffitt Cancer Center Support Grant (no. P30-CA076292). B.E.G.R. is supported by NIH grant nos. 1P50 CA196530-01 and NIH 1K08 CA151645-01. We thank the Sherlock-*Lung* study scientific advisory board (M. Meyerson, J. Samet, M. Spitz, R. Summers, M. Thun and W. Travis) for their support. We also thank Y. Rubanova from Toronto University for her help with the TrackSig analysis. We thank the staff at the IUCPQ Université Laval Biobank, Nice Biobank Centre de Ressources Biologiques, Yale University and Moffitt Cancer Center & Research Institute for their valuable assistance in collecting samples and corresponding clinical data. This work utilized the computational resources of the NIH high-performance computational capabilities Biowulf cluster (http://hpc.nih.gov).

Author contributions

M.T.L. and T.Z. conceptualized the study. T.Z., D.C.W., J. Shi., B.Z., N.A-P., N.L-B., B.Z., S.H.W., Y.P., H.C., T.R., D.R.S., D.A.G., L.B.A. and M.T.L. devised the methodology. T.Z., N.A-P., W.Z., P.H.H., R.L., K.H.-H., A.G.-P., F.M.-J., A.C., I.P., J. Sang, J. Shi, J.K., N.S., L.J.K., S.M.A.I., B.O., A.K., A.L.M. and C.F.K. carried out the formal analysis. A.H., N.C., J.C., D.H. and K.M.B. carried out the laboratory work. M.O., S.M.L., M.D., P.L., P.M.S.B. and J.S.A. carried out the pathology work. P.J., Y.B., P.H., D.C., A.C.P., L.A.M., B.E.G.R., M.L.P., M.C., M.B.S., N.E.C., M.L. and S.J.C. managed the resources. M.K., L.M. and J.R. curated the data. T.Z. and M.T.L. wrote the original draft. D.C.W., S.J.C., Y.B., Q.L., N.R., M.G-C., D.A.G., L.B.A., N.L-B., B.Z., J. Sang, J. Shi, T.Z., P.H.H. and M.T.L. reviewed and edited the draft. All authors carried out the data visualization. M.T.L. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

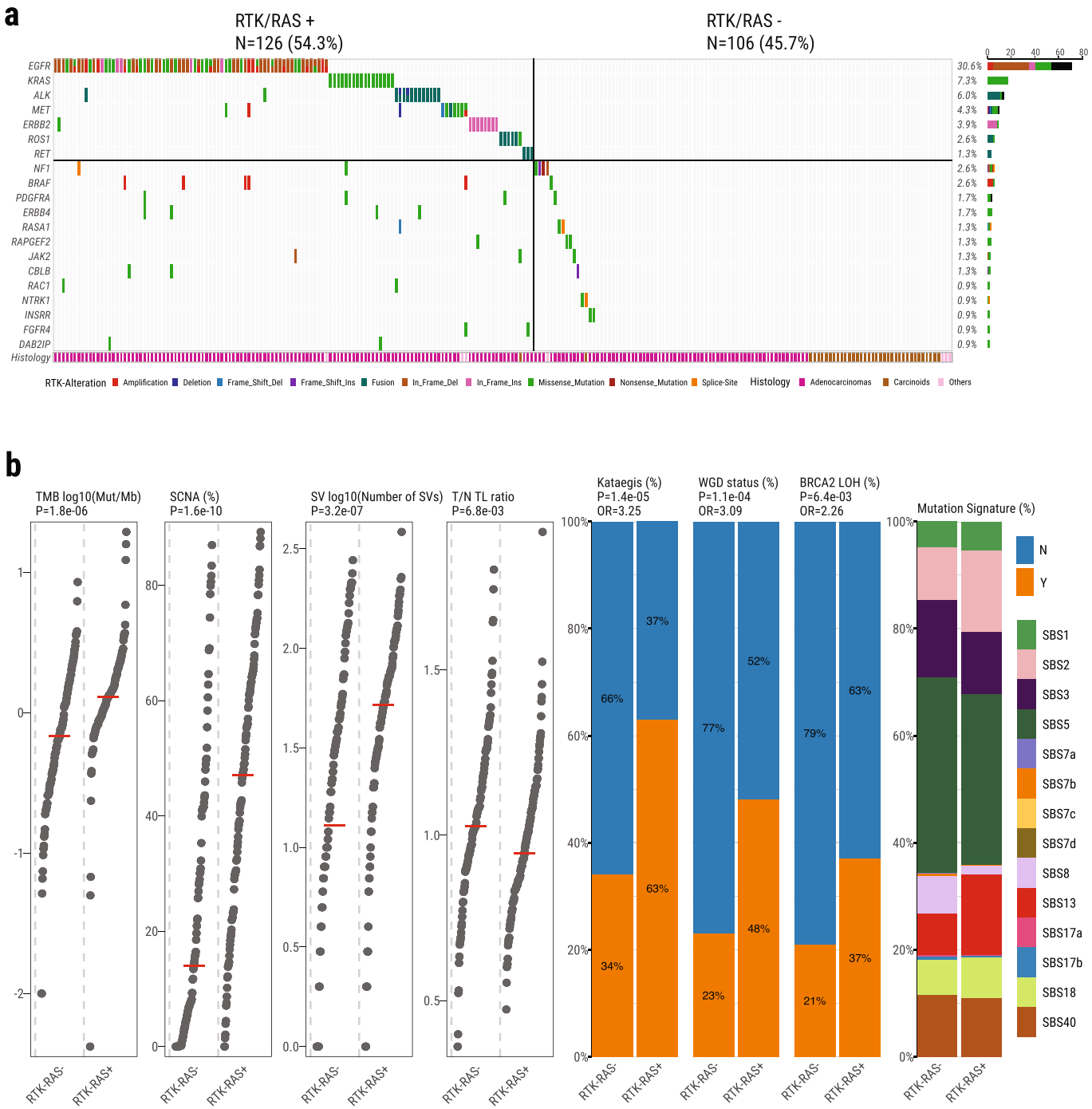
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00920-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00920-0.

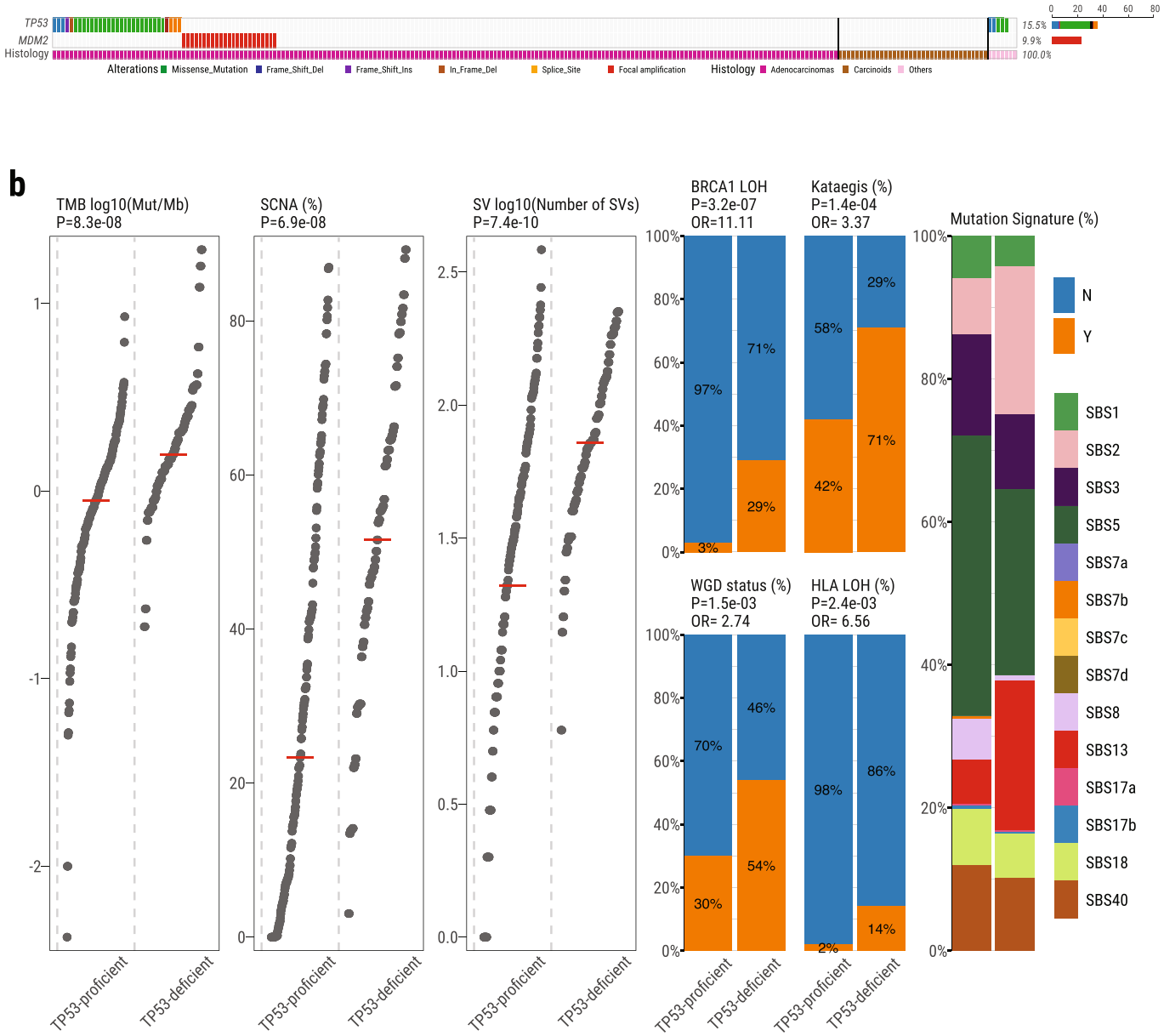
Correspondence and requests for materials should be addressed to M.T.L.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

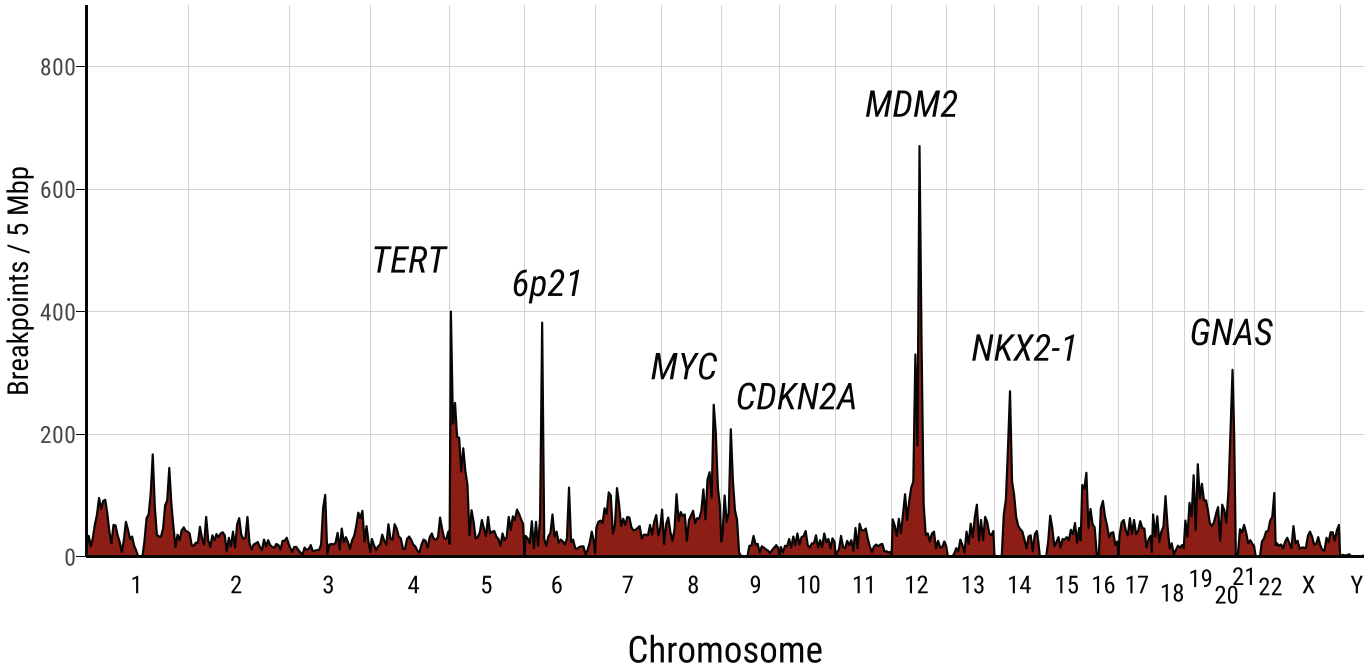
Reprints and permissions information is available at www.nature.com/reprints.



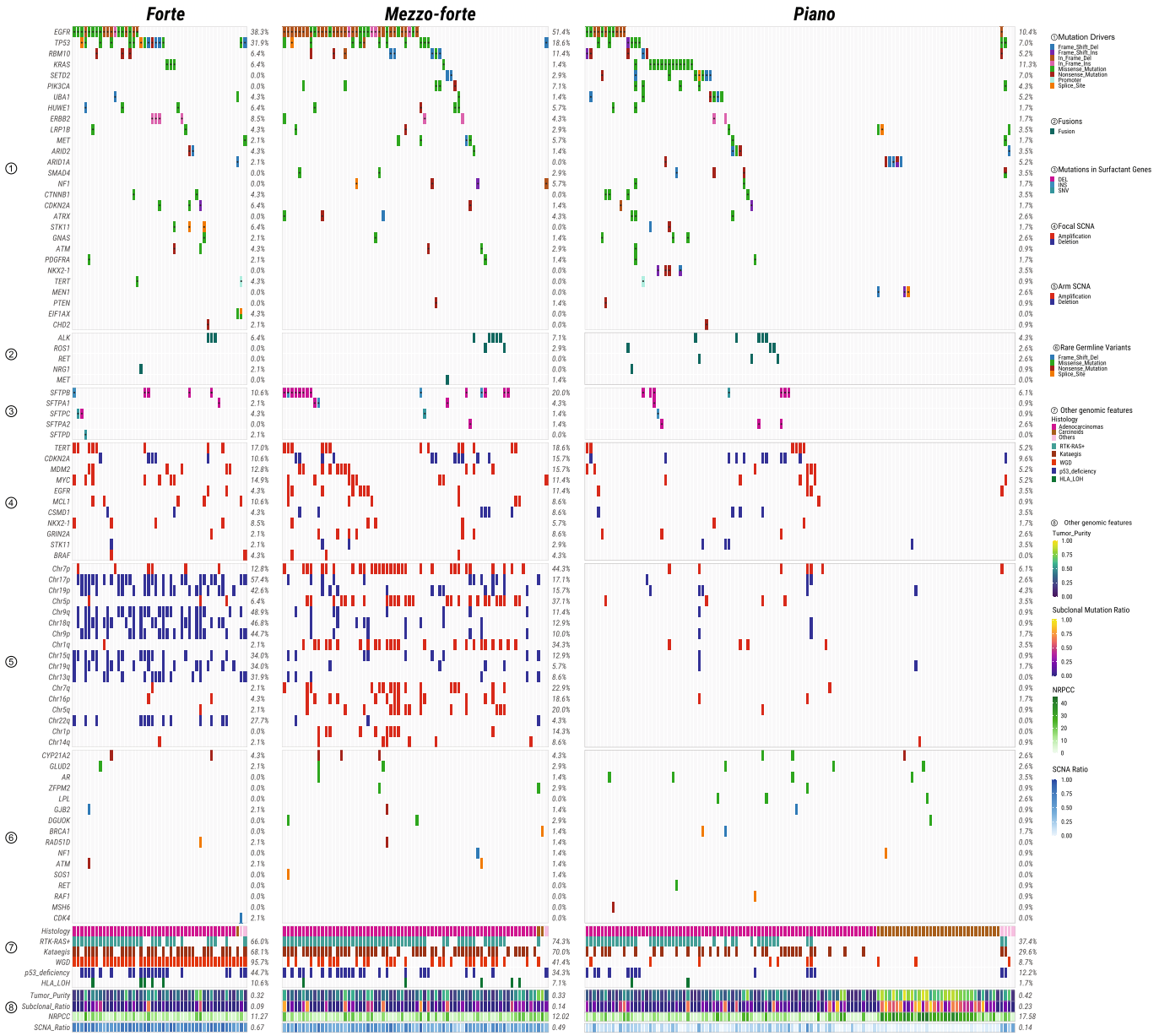
Extended Data Fig. 1 | Genomic alterations of RTK-RAS pathway in Sherlock-*Lung*. a, Oncoplot showing mutual exclusivity of genes within the RTK-RAS pathway, which were used to define the RTK-RAS status. The bottom bar shows tumor histological types. **b**, Comparison of genomic features between RTK-RAS negative and positive tumors. Left four 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; Middle three panels: enrichments for Kataegis events, WGD events, and *BRCA2* LOH. *P*-values and *OR* are calculated using Fisher's exact test (two-sided); Right panel: Contributions of each SBS signature.



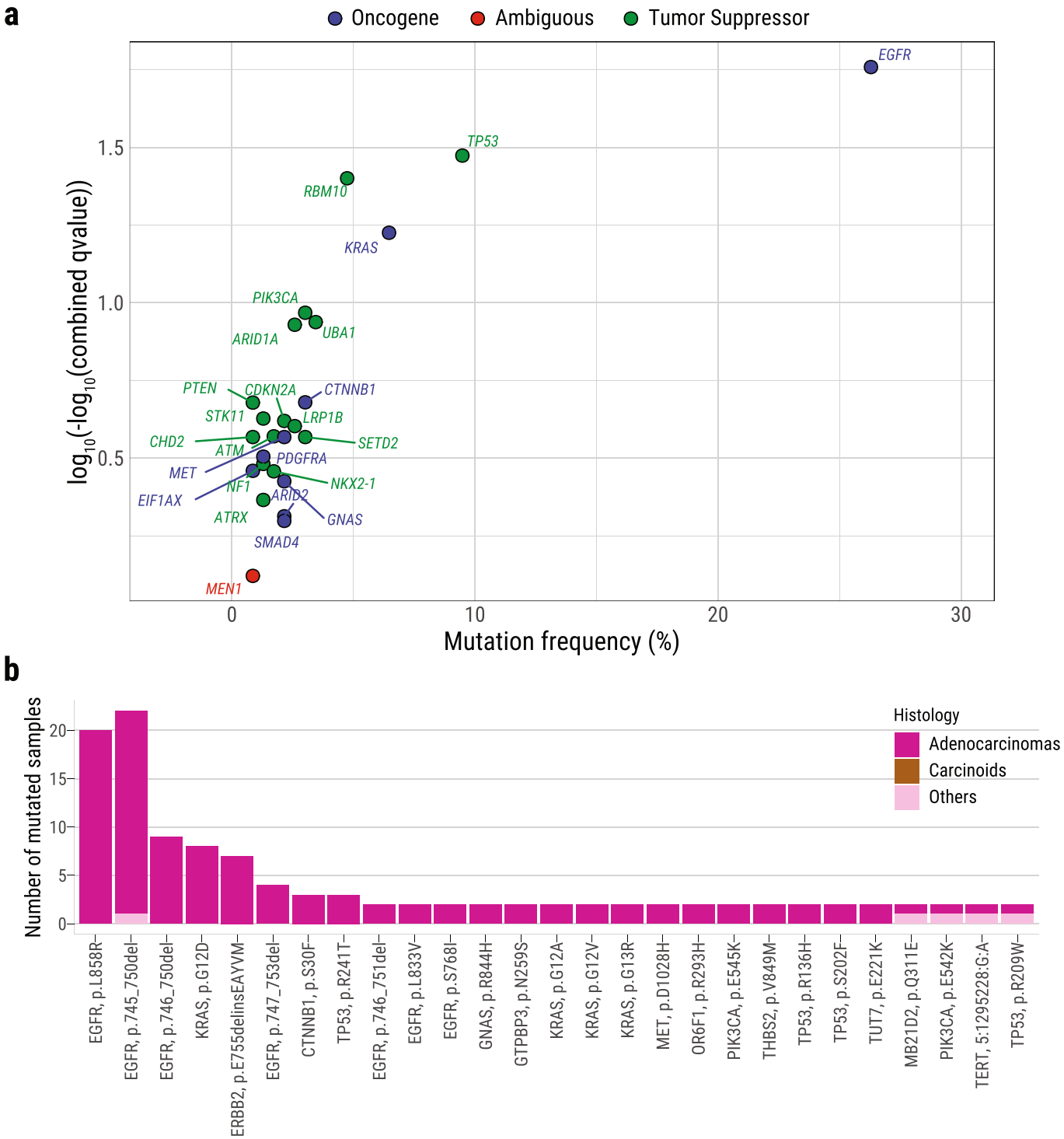
Extended Data Fig. 2 | Genomic alterations of *TP53* pathway in Sherlock-*Lung*. a, Oncoplot showing the mutual exclusivity between *TP53* mutations and *MDM2* amplification, which was used to define the TP53 proficient and deficient groups. The bottom bar shows tumor histological types. **b**, Comparison of genomic features between TP53-proficient and TP53-deficient tumors. Left three panels: tumor mutation burden, percentage of genome with SCNA and SV burden. *P*-values are calculated using the two-sided Mann-Whitney *U* test. Middle four panels: enrichments for *BRCA1* LOH, Kataegis events, WGD events, and HLA LOH. *P*-values and OR are calculated using Fisher's exact test (two-sided). Right panel: Contributions of each SBS signature.



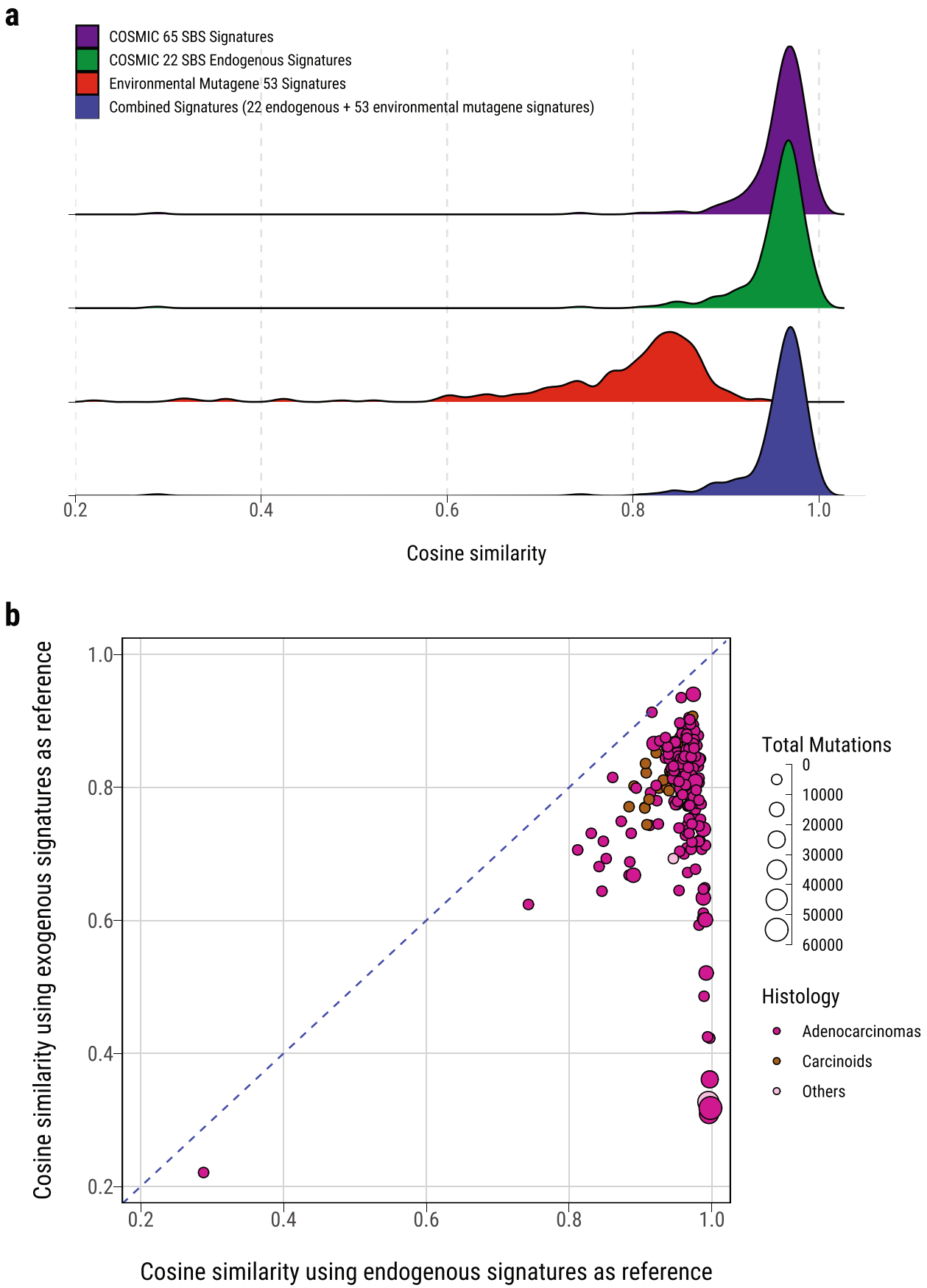
Extended Data Fig. 3 | Recurrence of SV breakpoints in Sherlock-*Lung*. The frequencies of chromosomal breakpoints are calculated using 5 Mb as a window across the whole genome.



Extended Data Fig. 4 | Summary of genomic features in LCINS based on different SCNA clusters. Panels from top to bottom describe: 1) most frequently mutated or potential driver genes; 2) oncogenic fusions; 3) somatic mutations in surfactant associated genes; 4) significant focal SCNAs; 5) significant arm-level SCNAs; 6) genes with rare germline mutations; 7) and 8) other genomic features. The numbers on the right panel show the overall frequency (1-7) or median values (8). NRPCC: the number of reads per clonal copy.

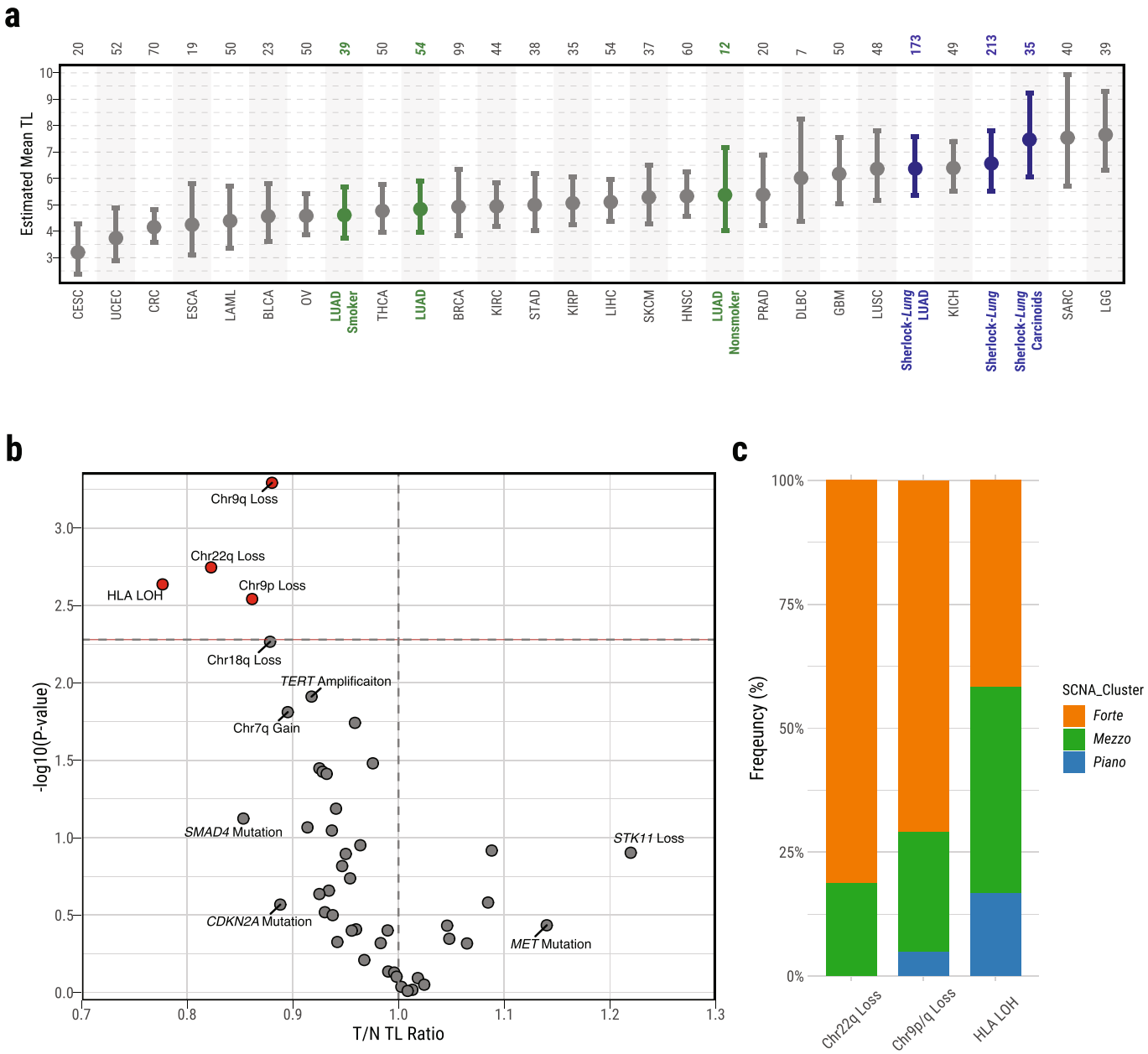


Extended Data Fig. 5 | Genes with signals of positive selection in Sherlock-*Lung*. a, The scatter plot showing significantly mutated genes according to IntOGen *q*-value <0.05 (y-axis) and mutational frequency in the cohort (x-axis). Genes are colored according to their inferred mode of action in tumorigenesis. **b**, Recurrent non-synonymous driver mutations (in ≥2 patients).

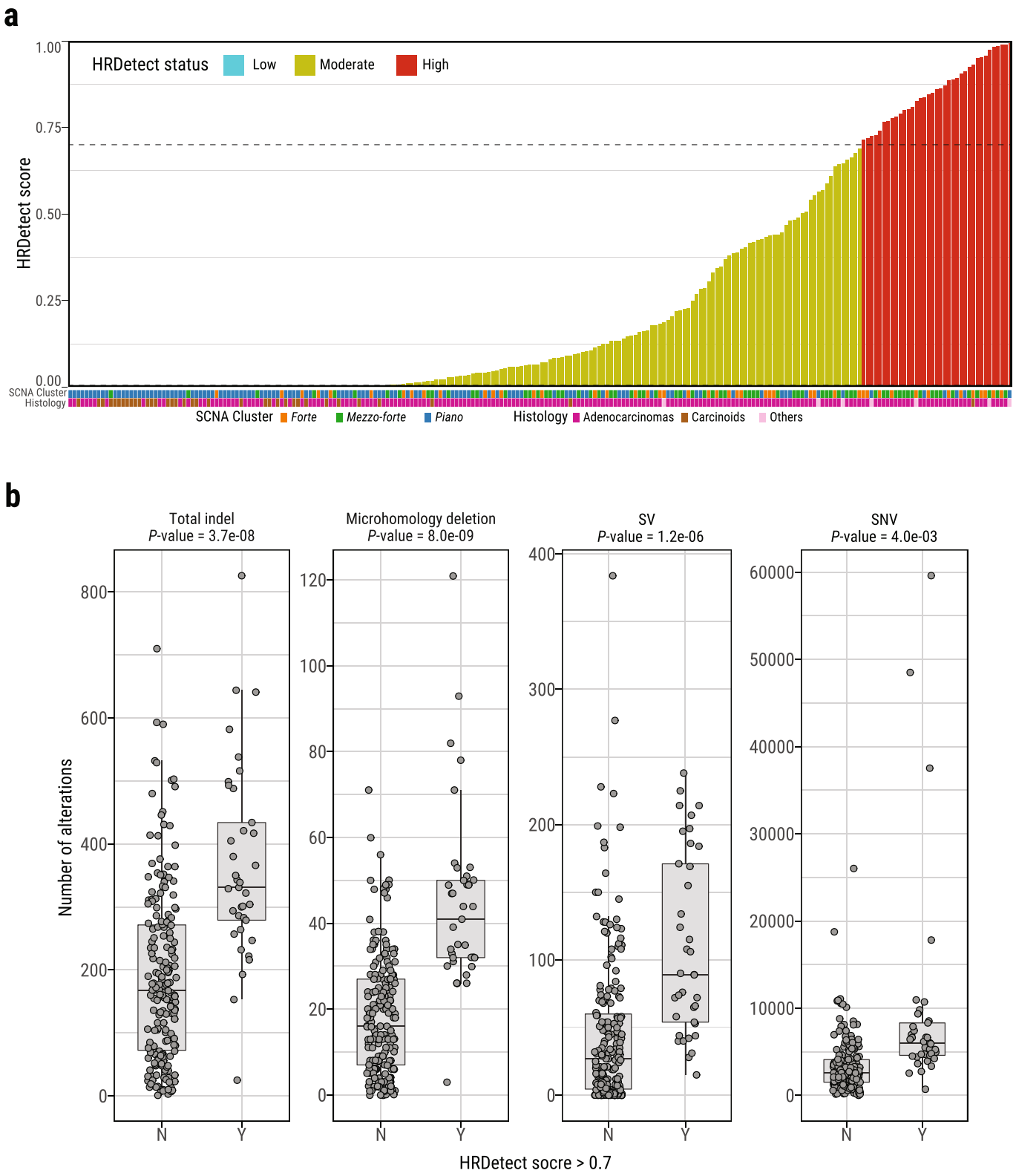


Extended Data Fig. 6 | See next page for caption.

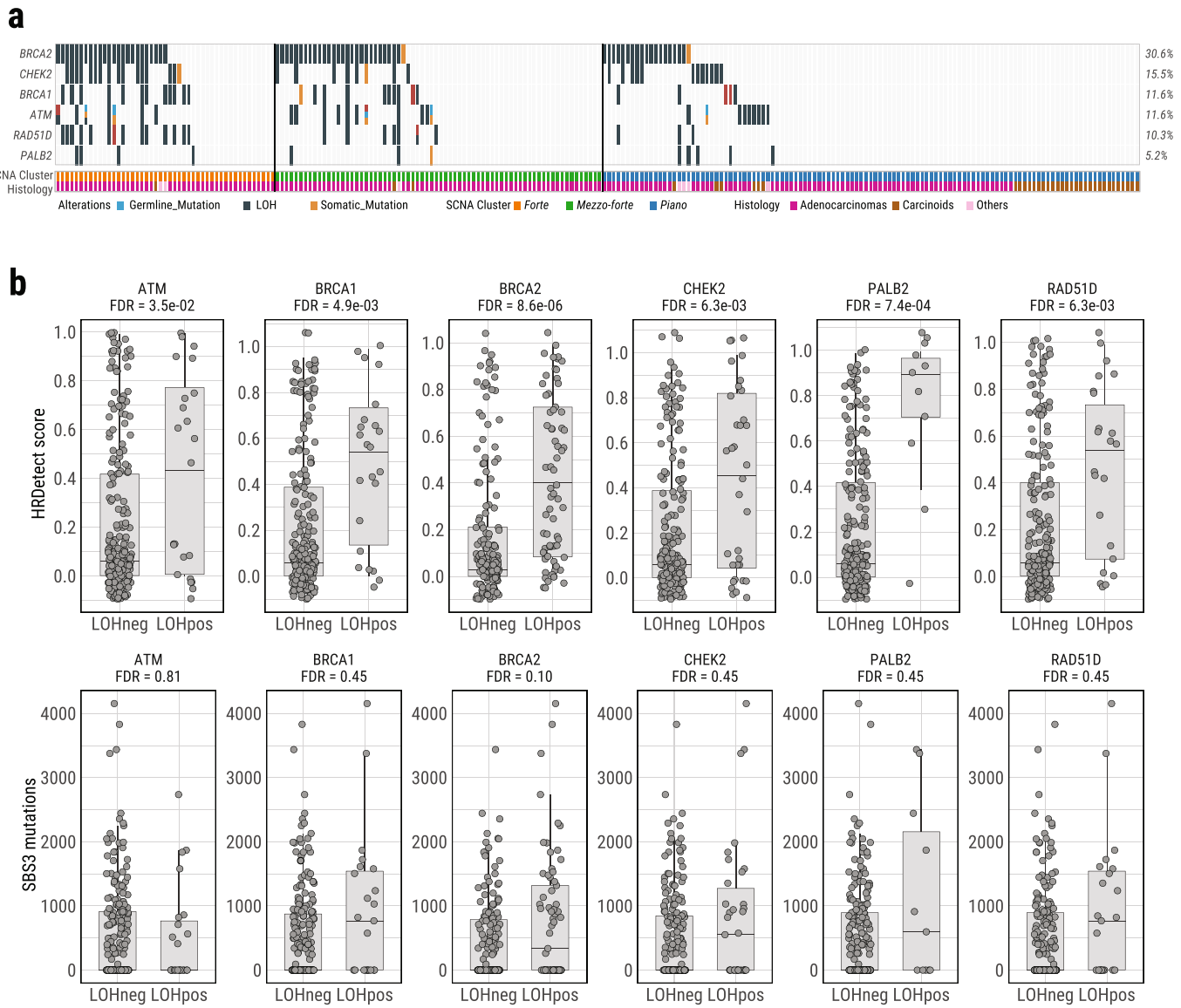
Extended Data Fig. 6 | Dominant endogenous processes in Sherlock-*Lung*. a, Density plot of cosine similarity between original mutational profile and reconstructed mutational profile using reference signatures from (top to bottom): 65 COSMIC SBS signatures, 22 COSMIC SBS signatures for endogenous processes, 53 MutaGene SBS signatures of environmental exposures, and a combined set of signatures including the 22 endogenous and 53 environmental exposure signatures. **b**, Comparison of the cosine similarity between the original mutational profiles and reconstructed mutational profiles using endogenous and exogenous signatures (similar to **a**). Each dot represents one sample. The size and color represent the total number of mutations and tumor histological type, respectively.



Extended Data Fig. 7 | Association between T/N TL ratio and somatic alterations in Sherlock-*Lung*. a, Distribution of mean telomere lengths (TL) in Sherlock-*Lung* (dark blue, overall and by histological type), TCGA LUAD (green, overall and by smoking status) and TCGA other cancer types (Grey). Total sample numbers for each type are shown at the top. Error bars, 95% CIs from linear mixed model. **b**, Scatterplot showing association between T/N TL ratio and somatic alterations. Association *P*-values (two-sided t-test; *FDR* adjusted using Benjamini-Hochberg method) are shown on the y-axis. Genomic alterations with *FDR* <=0.1 or T/N TL ratio >1.1 or <0.9 are labeled and further highlighted in red when significant (*FDR*=0.05; horizontal dashed line). **c**, The proportion of each SCNA cluster among the group of tumors with somatic alterations significantly associated with shorten T/N TL including Chr22q Loss, Chr9p/q Loss or HLA LOH.



Extended Data Fig. 8 | Homologous recombination deficiency (HRD) in Sherlock-*Lung*. a, HRDetect scores of Sherlock-Lung samples. HRD-high: >0.7, HRD-low: < 0.005. **b**, Comparison of the number of total indels, microhomology deletions, SVs, and SNVs between samples with HRDetect score below 0.7 (group N) and above 0.7 (group Y). *P*-values are calculated using the two-sided Mann-Whitney *U* test. 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.



Extended Data Fig. 9 | Genomic alterations in HRD associated genes in Sherlock-*Lung*. a, Oncoplot of genomic alterations in HRD associated genes, including germline mutations, somatic mutations and LOH. Samples with biallelic alterations are represented by bars with two different colors. The bottom bar shows tumor histological types. **b**, Boxplots of HRDetect scores (top) and SBS mutation loads (bottom) in tumors with and without LOH of six HR associated genes. *FDR* are calculated using the two-sided Mann-Whitney *U* test with multiple testing correction based on the Benjamini & Hochberg method. 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.

nature research

Corresponding author(s): Maria Teresa Landi

Last updated by author(s): Jun 24, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Whole Genome Sequence data were generated from the Illumina CASAVA (v1.8) software package.

Data analysis	mosdepth (v0.2.5), https://github.com/brentp/mosdepth; VerifyBamID (v1.1.3), https://genome.sph.umich.edu/wiki/VerifyBamID; somalier (v0.2.6), https://github.com/brentp/somalier; Picard Toolkit (v2.20.8), https://broadinstitute.github.io/picard/; GATK (v3.8.0), https://software.broadinstitute.org/gatk/; Strelka (v2.9.0), https://github.com/Illumina/strelka; MuTect (v1.1.7), https://software.broadinstitute.org/cancer/cga/mutect; Sentieon-genomics (v201808.03, including TNsnv, TNhaplotyper, TNscope and Haplotyper algorithms), https://www.sentieon.com; Oncotator (v1.9.1.0), https://software.broadinstitute.org/cancer/cga/oncotator; Annovar (v2019-10-24), http://annovar.openbioinformatics.org/en/latest/; Maftools (v1.6.05), https://github.com/PoisonAlien/maftools; Battenberg (v2.2.8), https://github.com/Wedge-Oxford/battenberg; DPClust (v2.28), https://github.com/Wedge-Oxford/dpclust; GISTIC2 (v2.0.23), ftp://ftp.broadinstitute.org/pub/GISTIC2.0; dNdScv (v0.0.1.0), https://github.com/im3sanger/dndscv; MutSigCV (v1.41), https://software.broadinstitute.org/cancer/cga/mutsig; OncodriveFM (v1.0), http://bg.upf.edu/group/projects/oncodrive-fm.php; CharGer (v0.5.2), https://github.com/ding-lab/CharGer; Meerkat (v0.185), http://compbio.med.harvard.edu/Meerkat/; TraFiC-mem (v.1.1.0), https://gitlab.com/mobilegenomesgroup/TraFiC; TelSeq (v0.0.2), https://github.com/zd1/telseq; TelomereHunter (v1.1.0), https://www.dkfz.de/en/applied-bioinformatics/telomerehunter/telomerehunter.html; LOHHLA (v1), https://bitbucket.org/mcgranahanlab/lohhla/src/master/; POLYSOLVER (v1.0), https://software.broadinstitute.org/cancer/cga/polysolver; MSISensor (v0.5), https://github.com/ding-lab/msisensor; MANTIS (v1.0.4), https://github.com/OSU-SRLab/MANTIS; KataegisPortal (v1.0.3), https://github.com/MeichunCai/KataegisPortal; SigProfilerMatrixGenerator (v1.0.20), https://github.com/AlexandrovLab/SigProfilerMatrixGenerator; SigProfilerExtractor (v0.0.5.77), https://pypi.org/project/sigproextractor/; HRDetect (v1), https://github.com/eyzhao/hrdetect-pipeline; TrackSig (v0.1.0), https://github.com/morrislab/TrackSig; Palimpsest (v1.0.0), https://github.com/FunGeST/Palimpsest; PlackettLuce (v0.2.2), https://github.com/hturner/PlackettLuce; IntOGen pipeline, https://www.intogen.org/; Integrative genomics viewer (IGV, v2.0), http://software.broadinstitute.org/software/igv/; P-MACD, https://github.com/NIEHS/P-MACD; PCAWG Chronological timing analysis, https://gerstung-lab.github.io/PCAWG-11/; Somatic mutation filtering, https://github.com/xtmgah/Sherlock-Lung;

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

Data

Policy information about availability of data

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

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

232 normal and tumor paired raw data (BAM files) of the whole genome sequencing datasets have been deposited in dbGaP with accession number: phs001697.v1.p1. Researchers will need to obtain dbGaP authorization to download these data. RNA-seq raw data (FASTQ files) have been submitted to NCBI GEO database with access number GSE171415. Germline variant dataset from EAGLE whole exome sequencing study can be access in dbGaP with access number phs002496.v1.p1. In addition, histological images of these tumors can be found at https://episphere.github.io/svs. Public datasets were used in this study including: gnomAD (v2.1.1)/ExAC (v0.3.1) (https://gnomad.broadinstitute.org/), 1000 genomes (phase 3 v5, https://www.internationalgenome.org/) and dbSNP (v138, https://www.ncbi.nlm.nih.gov/snp/).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	In this study, we evaluated the genomic landscape and mutational processes in 232 treatment-naïve lung cancer in never smokers using 85x WGS after excluding low quality data. Currently, our study is the largest high-coverage whole genome sequencing-based LUAD analysis. In addition, we also generated RNA-Seq data from 35 sherlock-lung tumor samples and 32 paired adjacent normal tissue samples. Sample sizes for both WGS and RNA-Seq were determined by sample availability.
Data exclusions	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; and no driver mutations in any known lung adenocarcinoma drive genes as reported by the TCGA study; and had no copy number alteration detected; 2) Tumor or normal samples had contamination >5% identified by VerifyBamID; 3) Tumor and normal sample were swapped based on pairwise sample relatedness metrics, as detected by Somalier; or 4) Normal samples were identified with large proportion copy number alterations using the Battenberg algorithm. 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.
Replication	This is the largest WGS study of lung cancer in never smokers. No additional studies are available from the same cancer type and WGS sequencing coverage to date.
Randomization	We did not investigate the effect of different treatments, so it is not applicable to this study.

Blinding	We did not investigate the effect of different treatments, so it is not applicable to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	We collected the following population characteristics in this study, including age at diagnosis (median 68 yrs, range 21-86 yrs), gender (75% female and 25% male), tumor stage (41% IA, 21% IB, 19% II and 16% III), tumor grade (35% grade 1, 36% grade 2 and 13% grade 3), passive smoking status (28% passive smokers and 64% non-passive smokers), survival status and overall survival time. All tumor samples were from treatment-naïve lung cancers in never smokers.
Recruitment	We collected bio-specimens from all lung cancer never smoker cases that we could identify across different centers without applying any selection criteria. Patients were predominantly of European descent (97.4%) with the remainder of Asian or African ancestry.
Ethics oversight	Since NCI only received de-identified samples and data from collaborating centers, had no direct contact or interaction with study subjects, and did not use or generate identifiable private information, Sherlock-Lung has been determined to constitute "Not Human Subject Research (NHSR)" based on the Federal Common Rule (45 CFR 46; eCFR.gov).

Note that full information on the approval of the study protocol must also be provided in the manuscript.