
Supplementary information

**Author Correction: Uncovering the role of
LINE-1 in the evolution of lung
adenocarcinoma**

In the format provided by the
authors and unedited

Uncovering the role of LINE-1 in the evolution of lung adenocarcinoma

<https://doi.org/10.1038/s41586-025-09825-y>

Received: 16 May 2024

Accepted: 28 October 2025

Published online: 10 December 2025

 Check for updates

Tongwu Zhang¹, Wei Zhao¹, Christopher Wirth², Marcos Díaz-Gay^{3,4,5,6}, Jinhu Yin¹, Monia Cecati⁷, Francesca Marchegiani⁸, Phuc H. Hoang¹, Charles Leduc⁹, Marina K. Baine¹⁰, William D. Travis¹⁰, Lynette M. Sholl¹¹, Philippe Joubert¹², Jian Sang¹, John P. McElderry¹, Michelle Antony¹, Alyssa Klein¹, Azhar Khandekar^{1,3,4,5}, Caleb Hartman¹, Jennifer Rosenbaum¹³, Frank J. Colón-Matos¹, Mona Mirafteb¹, Monjoy Saha¹, Olivia W. Lee¹, Kristine M. Jones^{1,14}, Neil E. Caporaso^{1,14}, Maria Pik Wong¹⁵, Kin Chung Leung¹⁶, Chao Agnes Hsiung¹⁷, Chih-Yi Chen^{18,19}, Eric S. Edell²⁰, Jacobo Martínez Santamaria^{21,22}, Matthew B. Schabath²³, Sai S. Yendamuri²⁴, Marta Manczuk²⁵, Jolanta Lissowska²⁵, Beata Świątkowska²⁶, Anush Mukeria²⁷, Oxana Shangina²⁷, David Zaridze²⁷, Ivana Holcatova^{28,29}, Dana Mates³⁰, Sasa Milosavljevic³¹, Milan Savic³², Yohan Bossé¹², Bonnie E. Gould Rothberg³³, David C. Christiani^{34,35}, Valerie Gaborieau³⁶, Paul Brennan³⁶, Geoffrey Liu³⁷, Paul Hofman³⁸, Robert Homer³⁹, Soo-Ryum Yang¹⁰, Angela C. Pesatori^{40,41}, Dario Consonni⁴¹, Lixing Yang^{42,43,44}, Bin Zhu¹, Jianxin Shi¹, Kevin Brown¹, Nathaniel Rothman¹, Stephen J. Chanock¹, Ludmil B. Alexandrov^{3,4,5,45}, Jiyeon Choi¹, Maurizio Cardelli⁷, Qing Lan¹, Martin A. Nowak^{46,47}, David C. Wedge^{2,48} & Maria Teresa Landi^{1,50}

Understanding lung cancer evolution can identify tools for intercepting its growth^{1,2}. Here, in a landscape analysis of 1,024 lung adenocarcinomas (LUADs) with deep whole-genome sequencing integrated with multiomic data, we identified 542 LUADs with a diverse clonal architecture. In this group, we observed divergent evolutionary trajectories based on tobacco smoking exposure, ancestry and sex. LUAD from smokers showed an abundance of tobacco-related C:G>A:T driver mutations³ in *KRAS* and short subclonal diversification. LUAD in people who have never smoked (hereafter, never-smokers) showed early occurrence of copy-number alterations and *EGFR* mutations associated with SBS5 and SBS40a mutational signatures. Tumours containing *EGFR* mutations exhibited long latency, particularly in female individuals of European-ancestry. Tumours from Asian never-smokers showed a short clonal evolution. Importantly, we found that the mutational signature ID2⁴ is a marker of a previously unrecognized mechanism for LUAD evolution. Tumours with ID2 showed short latency and high long interspersed nuclear element-1 (LINE-1, hereafter L1) retrotransposon activity linked to L1 promoter demethylation. These tumours exhibited an aggressive phenotype with genomic instability, elevated hypoxia scores, low neoantigen burden, metastasis propensity and poor overall survival. Reactivated L1-retrotransposition-induced mutagenesis probably contributes to the mutational signature ID2, including through the regulation of the transcriptional factor *ZNF695*, a member of the KZFP family⁵. The complex nature of LUAD evolution creates both challenges and opportunities for screening and treatment plans.

Lung cancer is the second most common cancer and the leading cause of cancer-related death worldwide. LUAD is the predominant histological subtype⁶. Understanding tumour evolution can reveal strategies to halt progression and inform treatment. Previous work has examined evolution across several cancers⁷, including LUAD^{2,8}, largely using targeted or exome sequencing, which limits the detection of complex structural events and transposable elements.

L1 elements are mobile elements that can retrotranspose and create new genomic insertions^{9,10}. Normally silenced in somatic cells, L1

becomes aberrantly active in many cancers, inducing double-stranded breaks (DSBs) and genome instability, including copy-number changes and subclonal heterogeneity^{10,11}. Although L1 activity is implicated in cancer heterogeneity^{12,13}, its role in LUAD evolution remains unclear owing to limited whole-genome sequencing (WGS) data. Beyond our previous work in 232 lung cancers in never-smokers (LCINS)¹⁴, the other LUAD WGS evolution study analysed 36 LUADs from the Pan-Cancer Analysis of Whole Genomes⁷ (PCAWG), which lacked sufficient power to explore evolutionary trajectories by demographic or exposure context.

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

We examined 1,024 LUAD tumours from the Sherlock-Lung study^{14,15}, which integrates genomic, questionnaire, imaging and exposure data to identify factors contributing to lung tumorigenesis. Here we focused on 542 tumours with sufficient subclonal resolution by deep WGS and multi-omics to reconstruct evolutionary timing, identify trajectories by ancestry, smoking and sex, and characterize L1 contributions. We identify an aggressive LUAD subgroup marked by specific mutational processes, link L1 to mutational signature ID2 and implicate ZNF695—a KZFP transcription factor—as a potential regulator of L1 in LUAD. These findings illuminate how L1 contributes to LUAD progression and heterogeneity.

Temporal order of genomic events in LUAD

We analysed deep WGS data and matched multi-omic profiles of 1,024 LUAD and selected 542 tumours with sufficient subclonal resolution for evolutionary analysis (Methods, Supplementary Table 1 and Extended Data Fig. 1). Tumours were grouped by ancestry and smoking: East Asian never-smokers (AS_N, $n = 180$), European never-smokers (EU_N, $n = 184$), European smokers (EU_S, $n = 120$) and others (Fig. 1a and Supplementary Fig. 1). The timing analysis of genomic events (Methods) revealed that whole-genome duplication (WGD) occurs significantly earlier in never-smokers than in smokers (two-sided Wilcoxon rank-sum test, $P = 5.11 \times 10^{-11}$; Fig. 1b,c and Supplementary Fig. 2). While most chromosomal gains occurred later in smokers, early gains of chromosomes 21 and 22 (in all groups) and chromosome X (in EU_S) in tumours suggest possible roles in LUAD initiation (Extended Data Fig. 2).

LUAD from smokers had a higher fraction of early clonal mutations (median, 24.7%) compared with in tumours from never-smokers (9.0% in AS_N and 6.0% in EU_N), including frequent driver mutations in *KRAS* (Supplementary Fig. 3) enriched for tobacco-associated C:G>A:T changes (SBS4; Extended Data Fig. 3). By contrast, LCINS carried more subclonal mutations (15.1% and 18.4% in AS_N and EU_N, versus 3.9% in EU_S; linear regression, $P = 2.61 \times 10^{-13}$ adjusted for ancestry and tumour purity; Fig. 1d and Supplementary Table 2). However, we observed a dominant presence of early clonal *EGFR* driver mutations in never-smokers (66.7% AS_N, 74.7% EU_N, among *EGFR*⁺ tumours), while no *EGFR* driver mutations were observed as early clonal in EU_S (never-smokers versus smokers, false-discovery rate (FDR)-adjusted $P = 0.025$; Fig. 1e and Supplementary Fig. 3).

Overall, 81.4% of *EGFR* mutations were attributed to SBS5 and SBS40a (AS_N, 83.6%; EU_N, 85.8%; EU_S, 35.3%; never-smokers versus smokers, $P = 0.011$; Extended Data Fig. 3), consistent with endogenous processes^{4,16,17}, and highlighting the early, endogenous-driven origin of *EGFR* mutations in never-smokers. After multiple-testing correction, no other drivers (such as *TP53* or *KRAS*) were enriched as early clonal in LCINS versus EU_S. Similar temporal patterns in AS_N and EU_N tumours suggest no major ancestry-related evolutionary differences in LCINS.

Dynamic mutational processes in LUAD

To understand how mutational processes contribute to LUAD evolution, we stratified mutations by clonality and evaluated the timing of mutational signatures identified in the Sherlock-Lung study¹⁸. Notably, exogenous mutational processes tended to act early during tumour evolution. Specifically, the aristolochic-acid-associated SBS22a signature was significantly enriched in early clonal mutation in AS_N tumours (subclonal/clonal mutation fold change (FC) = 0.68, two-sided Wilcoxon rank-sum test with Benjamini–Hochberg correction, FDR-adjusted $P = 0.026$), and the tobacco-smoking-associated SBS4 and SBS100 signatures showed similar early clonal enrichment in EU_S tumours (SBS4: FC = 0.79, FDR-adjusted $P = 6.97 \times 10^{-7}$; SBS100: FC = 0.81, FDR-adjusted $P = 2.21 \times 10^{-3}$; Fig. 1f). Similar patterns of SBS4 have also been reported in previous studies^{7,19}. By contrast, several endogenous or unknown-aetiology signatures were

enriched in subclonal mutations, including APOBEC signatures (SBS2: FC = 1.24, FDR-adjusted $P = 2.33 \times 10^{-3}$; SBS13: FC = 1.20, FDR-adjusted $P = 2.95 \times 10^{-3}$) and SBS17a/b (unknown aetiology; SBS17a: FC = 1.55, FDR-adjusted $P = 0.65$; SBS17b: FC = 1.91, FDR-adjusted $P = 5.37 \times 10^{-3}$). However, most APOBEC-related mutations occurred during the clonal stage overall (Supplementary Fig. 4). Notably, SBS17a was absent in AS_N tumours, and SBS17b was observed in only one AS_N tumour (Supplementary Fig. 5), highlighting potential ancestry- or exposure-specific differences.

Trends were consistent when comparing early versus late clonal mutations (Extended Data Fig. 4 and Supplementary Fig. 6). The insertion–deletion mutation (indel) signatures ID8 and ID6⁴ (DNA repair deficiency) were enriched in subclonal indels (Supplementary Fig. 7), although low event counts limited group comparisons. In ancestry comparisons, group-specific signatures (such as SBS4, SBS22a, SBS17a/b) could not be statistically evaluated, whereas common signatures (such as, SBS2/13) showed no timing differences.

Factors influencing lung tumour latency

We used stably accumulating CpG>TpG mutations as a molecular clock to map evolutionary time^{21,24} and define tumour latency as the interval between the age at diagnosis and the estimated age of most recent common ancestor (MRCA) (Methods and Supplementary Fig. 8). CpG>TpG burden correlated with the age at diagnosis ($P = 2.09 \times 10^{-5}$; Supplementary Fig. 9). Mapping molecular timing estimates to chronological time under different scenarios of increased CpG>TpG mutation rates revealed similar tumour latency across groups (Supplementary Fig. 10a,b and Supplementary Table 3). Notably, AS_N tumours showed earlier age at onset (AS_N, 63.0 years; EU_N, 67.9 years; EU_S, 66.6 years; $P \leq 0.003$; Supplementary Fig. 10c) and earlier MRCA age than other groups (Supplementary Fig. 10d).

Short latency suggests aggressive but clonal and therapeutically targetable tumours, whereas long latency implies greater subclonal diversity and a longer detection window but therapeutic challenges. We examined the impact of driver mutations on tumour latency across 50 driver genes (Fig. 2a and Supplementary Table 4). Overall, *EGFR*-mutant tumours had a significantly longer latency than *EGFR*-wild-type mutations (median latency difference, 6.3 years; FDR-adjusted $P = 8.25 \times 10^{-5}$). Within LCINS, *EGFR*⁺ tumours showed increased latency in AS_N (6.2 years; $P = 0.086$) and EU_N (8.9 years; $P = 0.00031$) (Extended Data Fig. 5a). By contrast, *KRAS*-mutant tumours exhibited shorter latency (−5.2 years; FDR-adjusted $P = 0.028$), particularly in never-smokers (AS_N, −9.2 years; EU_N, −10 years; based on 3 AS_N and 21 EU_N tumours, respectively; Extended Data Fig. 5b). Notably, tumours from female patients showed longer latency than male (3.3 years; $P = 0.0034$), mostly in EU_N (6.4 years; Extended Data Fig. 5c). Multivariable analyses suggest *EGFR* mutations and female sex were independently associated with longer latency, reaching statistical significance in EU_N (Fig. 2b,c). No latency difference by sex was observed between tumours with and without *KRAS* mutations.

We next investigated mutational processes in relation to tumour latency. SBS17a was nominally associated with shorter latency across exposure groups (Fig. 2d and Supplementary Fig. 11). Importantly, the indel signature ID2 was significantly associated with shorter tumour latency after correcting for testing all signature types (ID2⁺ versus ID2[−] tumours: −6.2 years; FDR-adjusted $P = 1.94 \times 10^{-4}$; Fig. 2d and Supplementary Table 5), and ID2 deletion burden was negatively correlated with latency ($R = -0.16$; $P = 0.0027$; Supplementary Fig. 12). ID2 frequencies were 11.1%, 16.8% and 35.8% in AS_N, EU_N and EU_S, respectively (Supplementary Fig. 13), and 83.7% of ID2 deletions were clonal. Consistent with a previous report²⁰, we observed that ID2 deletions were associated with late replication timing in lung cancer cell lines (Methods and Supplementary Fig. 14), consistent with increased mutation rates in late replication²¹. In a multivariable analysis including

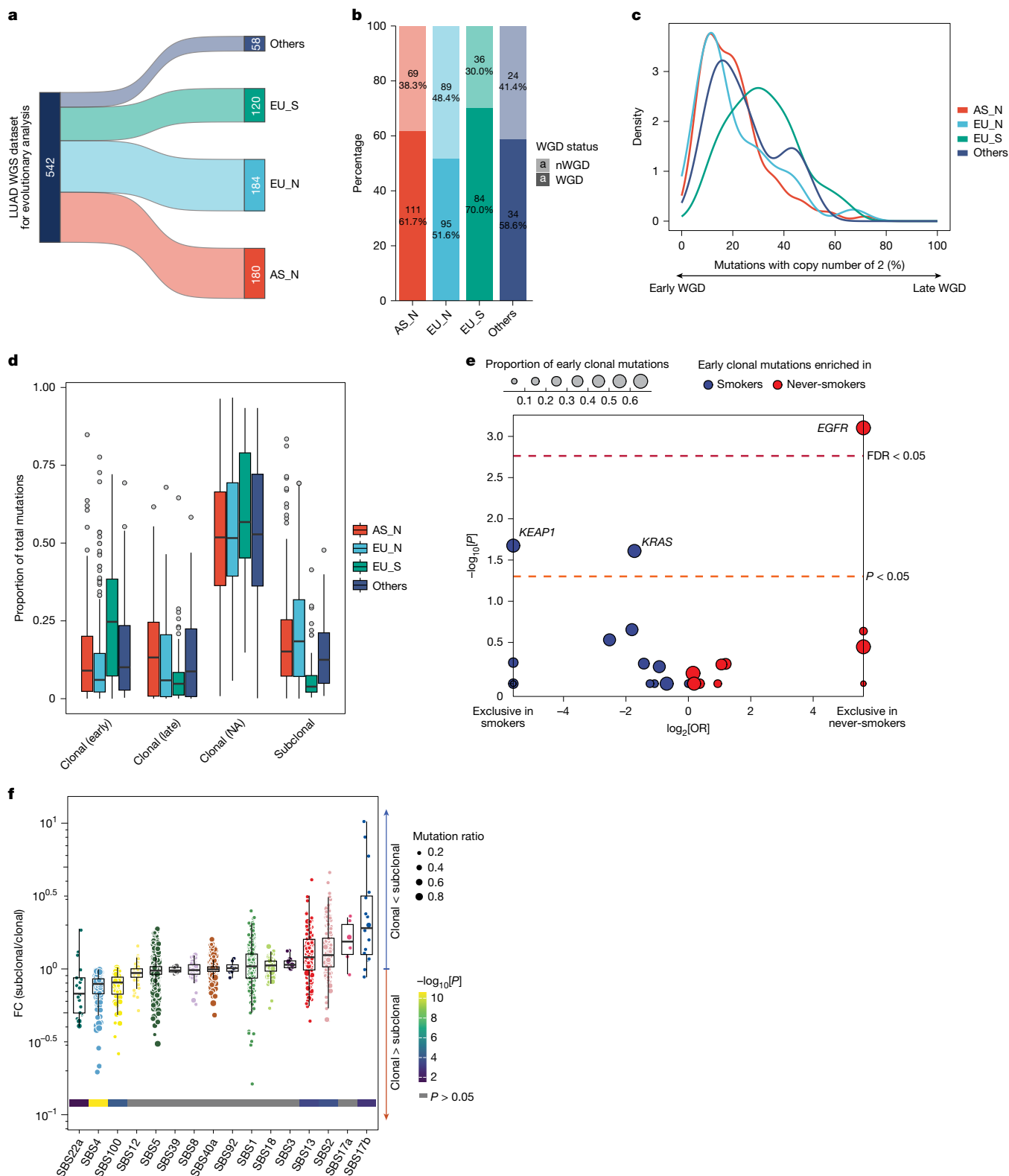


Fig. 1 | Evolutionary dynamics of lung cancer. **a**, Summary of high-clonality WGS data from the Sherlock-Lung study. **b**, The proportion of tumour samples ($n = 542$) with WGD across the AS_N, EU_N, EU_S and other groups. **c**, The distribution of the percentage of mutations with copy number 2, considering only those attributed to clock-like signatures (SBS1, SBS5). **d**, The proportion of mutations attributed to clonal versus subclonal states. **e**, Enrichment of early clonal mutations within cancer-driver genes, stratified by smoking status. Each point represents a driver gene, coloured by direction of enrichment. Fisher's exact test was applied to 2×2 contingency tables (smoking status versus early clonal mutation status). Significance thresholds are shown as dashed lines:

$P < 0.05$ (orange) and FDR-adjusted $P < 0.05$ (red); FDR correction was performed using the Benjamini–Hochberg method. **f**, Dynamic mutational processes during clonal versus subclonal tumour evolution. Fold changes represent the difference in the relative proportions of clonal and subclonal mutations attributed to each mutational signature. Each point corresponds to a tumour sample in which the signature was detected (total sample size $n = 542$), coloured by signature. P values were calculated using two-sided Wilcoxon rank-sum tests. The box plots show the median (centre line) and interquartile range (IQR; box limits), and the whiskers extend to $1.5 \times$ IQR.

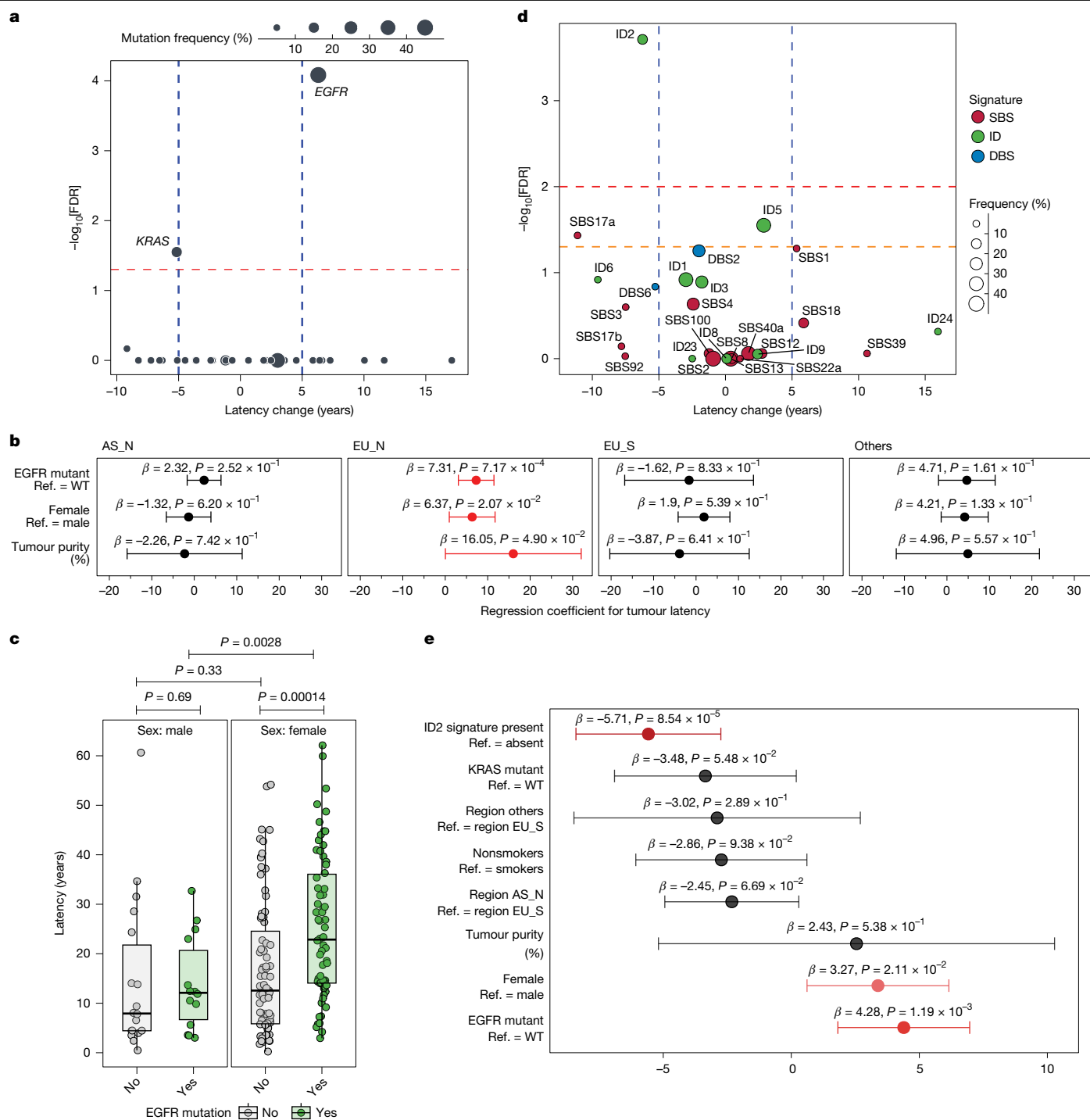


Fig. 2 | Features associated with lung tumour latency. **a**, Associations between tumour latency and driver gene mutation status were assessed using two-sided Wilcoxon rank-sum tests with Benjamini–Hochberg correction. The vertical blue dashed line indicates a 5-year latency difference between wild-type and mutant groups; the horizontal line marks the significance threshold ($\text{FDR} < 0.05$). **b**, Associations between *EGFR* mutation status and tumour latency, as determined using linear regression adjusted for sex and tumour purity (the percentage of cancer cells within a sample). Sample sizes were as follows: AS_N ($n = 180$), EU_N ($n = 184$), EU_S ($n = 120$) and other ($n = 58$). Points represent regression coefficients (β), and the error bars indicate the 95% confidence intervals (CIs) with significant associations ($P < 0.05$) shown in red. Ref., reference. **c**, The estimated latency by *EGFR* mutation status and sex ($n = 542$).

P values from two-sided Wilcoxon rank-sum tests are shown above the plots. The box plots show the median (centre line), IQR (box limits), $1.5 \times \text{IQR}$ (whiskers) and individual tumour samples. **d**, Associations between latency and specific mutational signatures were evaluated using two-sided Wilcoxon rank-sum tests with Benjamini–Hochberg correction. The vertical dashed line represents a 5-year difference between tumours with and without each signature, and the horizontal line denotes the $\text{FDR} < 0.05$ threshold. **e**, Multivariate linear regression ($n = 542$) analysis of the relationship between latency and sex, ancestry, smoking, *EGFR* and *KRAS* mutations, and the ID2 mutational signature. Two-sided tests were applied without adjustment for multiple comparisons. Points represent regression coefficients (β) and error bars indicate 95% CIs, with significant associations ($P < 0.05$) shown in red.

sex, ancestry, smoking, *EGFR*, *KRAS* and ID2 (Fig. 2e), ID2 showed the strongest effect size on latency, followed by *EGFR* and sex, suggesting an independent impact of these factors on LUAD evolution.

Characterization of ID2⁺ LUAD

ID2, although of unknown origin, has been found elevated in cancer types with defective DNA mismatch repair and microsatellite instability⁴, which were absent in our samples. Short tumour latency suggests aggressive tumour behaviour. We assessed the ID2 signature correlation with tumour proliferation markers²² using RNA-sequencing (RNA-seq). As hypothesized, ID2⁺ tumours had elevated proliferation marker expression (for example, *MKI67*: FC = 1.2, FDR-adjusted $P = 9.35 \times 10^{-13}$; Fig. 3a and Supplementary Fig. 15), strongly correlating with the burden of ID2 deletions (Fig. 3b; *MKI67*: $R = 0.48$, $P = 8.88 \times 10^{-16}$), a pattern not observed in matched normal samples (Supplementary Fig. 16). Pathway analyses indicated activation of cell cycle/proliferation programs (for example, cell cycle checkpoints, $Z = 9.43$, $P = 5.23 \times 10^{-31}$; Extended Data Fig. 6a–d). Clinically, ID2⁺ tumours were associated with worse survival (hazard ratio (HR) = 1.8, $P = 0.0021$) and the occurrence of metastases (Fisher's exact test: odds ratio (OR) = 2.5, $P = 5.58 \times 10^{-3}$; logistic regression: $P = 0.039$ adjusted tumour purity, age, sex and *TP53* mutation status) (Fig. 3c,d and Extended Data Fig. 6e–h).

A multivariable logistic model including age, sex, ancestry, smoking and tumour purity as covariates showed that *TP53* driver mutations, WGD and L1 insertions were the most significantly enriched genomic events in ID2⁺ tumours (Fig. 3e). *TP53* mutations were associated with the burden of ID2 deletions (Supplementary Fig. 17), and upstream regulator analysis indicated *TP53* downregulation in ID2⁺ tumours ($Z = -6.81$, $P = 1.81 \times 10^{-31}$; Supplementary Fig. 18). Both *TP53* mutations and WGD are consistent with genomic instability. Hypoxia, driven by tumour growth exceeding oxygen supply²³, profoundly impacts cancer genomics and evolution²⁴. Consistently, we found a considerably elevated hypoxia score in ID2⁺ tumours across groups (Methods and Fig. 3f).

A shorter tumour latency may, in part, result from impaired immune editing. To investigate this, we predicted neoantigens from mutation data and quantified the propensity of each SBS and ID signature to generate neoantigens. Notably, ID2 exhibited the lowest propensity to generate neoantigens (Extended Data Fig. 7), suggesting that diminished immunoediting that could promote rapid tumour growth. Consistently, RNA-seq-based immune deconvolution showed reduced T cell and myeloid dendritic cell infiltration in ID2⁺ tumours (Supplementary Fig. 19), indicating that weakened immune surveillance may drive their aggressive progression.

L1 is enriched in ID2-positive LUAD

Given the aggressive phenotype and molecular correlates of ID2⁺ tumours, we next investigated whether L1 retrotransposition, a known driver of genomic instability, could mechanistically contribute to the ID2 emergence. L1 insertions were strongly associated with ID2⁺ tumours (Fig. 3e), independent of *TP53* mutations (*TP53*⁺ tumours: $P = 0.000129$, OR = 5.57; *TP53*[−] tumours: $P = 0.00211$, OR = 3.73). Pathway analysis of RNA-seq data revealed that ID2⁺ tumours significantly activated L1-regulated gene sets, identified from genome-wide CRISPR–Cas9 screens of L1 control genes in two distinct human cell lines^{25,26}, ranking immediately after the top cell proliferation pathways (Methods, Extended Data Fig. 6b–d and Supplementary Fig. 20).

Previous studies have highlighted the role of L1 retrotransposition in both tumour initiation and progression²⁷, predominantly in epithelial cancers such as lung cancer. We therefore tracked the origins of L1 insertions from both germline and somatic sources, using our deep-WGS data (Fig. 4a,b and Methods). L1 insertions were significantly elevated in tumours from smokers than never-smokers ($P = 4.87 \times 10^{-7}$) and 70.4%

of those with known sources originated from pre-existing germline elements, suggesting a role in early clonal expansion. Consistent with previous reports^{10,28}, we identified several master germline sources of L1 elements across all groups, including chromosome 22q12.1 (28.9%), chromosome Xp22.2 (14.2%), chromosome 2q24.1 (4.6%), chromosome 14q23.1 (4.4%) and others (Fig. 4c).

L1 activity correlates with ID2 burden

L1 retrotransposon activity promotes genomic instability and tumorigenesis by inducing DNA damage²⁹ and/or disrupting driver genes³⁰. Using logistic regression (Methods), we observed significant associations between germline-derived L1 insertions and mutational signature ID2 (FDR-adjusted $P = 2.8 \times 10^{-5}$) and ID1 (FDR-adjusted $P = 5.71 \times 10^{-4}$; Fig. 4d). Notably, there were strong correlations among the numbers of indels assigned to these two signatures (Fig. 4e), suggesting a shared underlying aetiology. ID1 and ID2 are characterized by 1 bp insertions and deletions, respectively, occurring within homopolymers of ≥ 5 A or T bases. Motif analysis confirmed minimum homopolymer lengths of 5 and 6 T bases for ID1 and ID2, respectively (Fig. 4f,g). Despite fewer somatic-derived L1 insertions, we still observed a significant association with ID2 (Supplementary Fig. 21). Moreover, L1 insertions strongly correlated with ID2-associated deletions across all three major groups (overall: $R = 0.55$; AS_N: $R = 0.37$; EU_N: $R = 0.46$; EU_S: $R = 0.44$; all $P < 0.005$; Fig. 4h). We validated this association in an independent TCGA LUAD dataset ($R = 0.55$, $P = 4.1 \times 10^{-8}$; Supplementary Fig. 22). Furthermore, ID2⁺ tumours were more enriched for germline source L1 insertions (OR = 3.57, $P = 1.23 \times 10^{-8}$) than somatic transposed L1 insertions (OR = 2.33, $P = 1.94 \times 10^{-4}$; Supplementary Fig. 23). These findings suggest a likely contribution of L1 retrotransposition to the development of ID2 and its associated early clonal indels and short tumour latency. Consistently, tumours with L1 insertions, particularly from a germline source, showed a shorter tumour latency (Supplementary Fig. 24).

To further validate the association between L1 activity and ID2, we reanalysed publicly available single-cell clone WGS datasets from two independent studies (Methods). In longitudinal WGS data from 100 cancer cell line daughter clones³¹, in vitro-acquired L1 insertions significantly correlated with ID1/ID2 mutation burden (Supplementary Fig. 25). Similarly, in 406 colorectal single-cell clones³², strong correlations were observed across epithelial, cancer and MUTYH-associated polyp clones (Extended Data Fig. 8). These results confirm that the L1–ID2 link is robust, generalizable across tissues and detectable even in non-tumour epithelial clones.

L1 hypomethylation in ID2⁺ LUAD

Retrotransposition proceeds through targeted-site-primed reverse transcription (TPRT)³³. During TPRT, the endonuclease activity of ORF2p, encoded by the L1 ORF2 gene, cleaves target DNA or other TPRT intermediates, resulting in DNA DSBs³⁴ or DNA lesions that subsequently activate DNA repair pathways (for example, non-homologous end-joining)^{35–37}. Notably, ORF2p specifically recognizes and cleaves the genomic consensus sequence 5'-TTTTT-3' (canonical endonuclease motifs)^{9,38}, which closely resembles the motifs characteristic of mutational signature ID1 and ID2. Normally, L1 activity is suppressed epigenetically, primarily through methylation of L1 promoters³⁹. We hypothesized that L1 is reactivated through promoter demethylation. Reactivated L1 ORF2-encoded endonuclease, coupled with DNA repair mechanisms, could contribute to the emergence of single-base-pair indels⁴⁰ like ID2 (Fig. 5a).

Consistent with global hypomethylation, differentially methylated probe (DMP) analysis showed that 95.4% of DMPs were hypomethylated in ID2⁺ versus ID2[−] tumours (FDR-adjusted $P < 0.01$; Supplementary Fig. 26), aligning with tobacco-smoking-associated global CpG

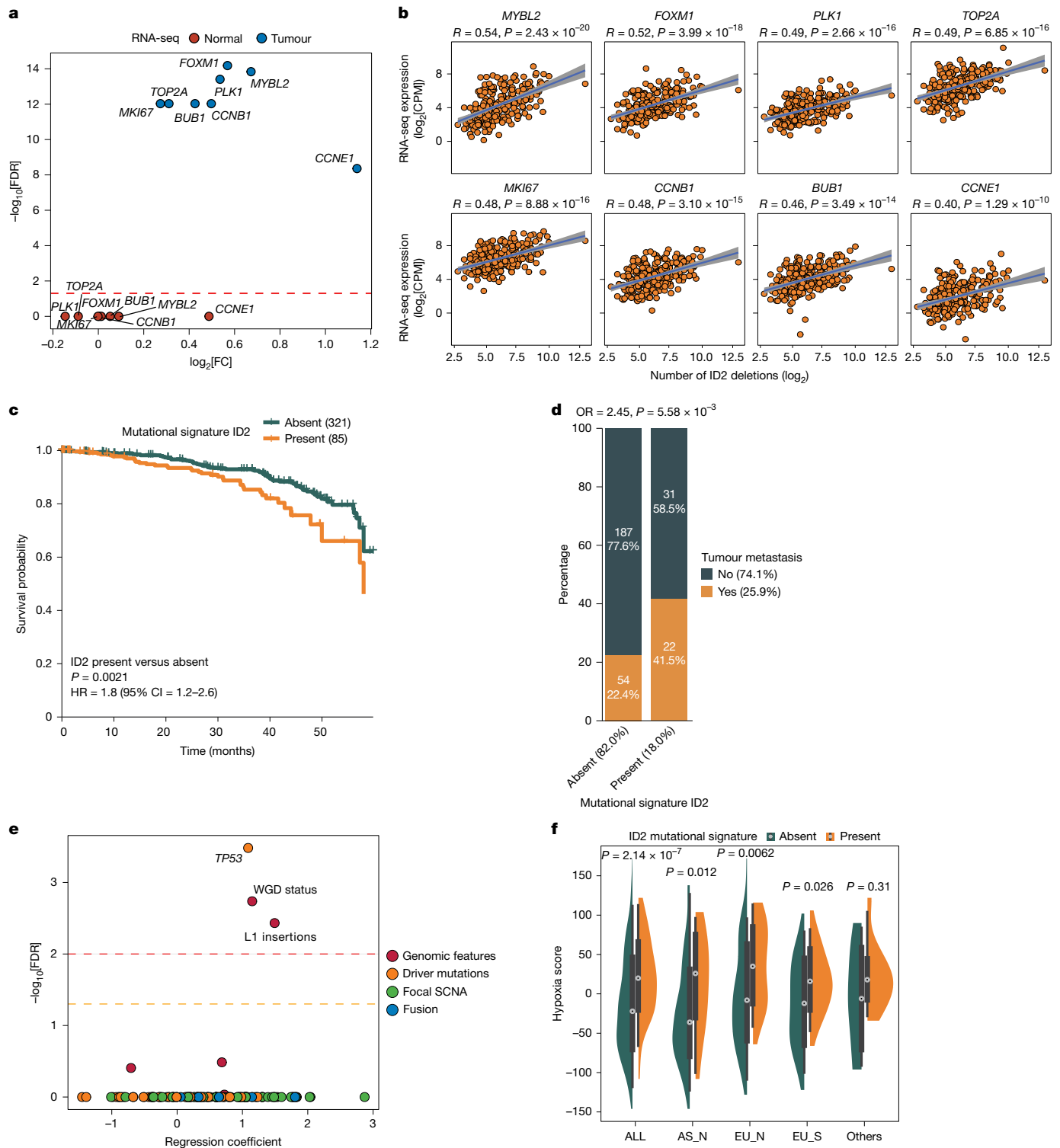


Fig. 3 | Characterization of tumours with the mutational signature ID2.

a, Associations between mutational signature ID2 and expression of tumour proliferation markers were assessed from tumour and normal tissue RNA-seq using two-sided Wilcoxon rank-sum tests with Benjamini–Hochberg correction. The horizontal line indicates the significance threshold ($\text{FDR} < 0.05$).

b, Correlations between ID2 deletions and proliferation marker expression were measured using Pearson correlation. The fitted regression line (blue) with 95% CIs (grey) is shown. Two-sided correlation coefficients and P values are reported above. **c**, Kaplan–Meier curves illustrate overall survival stratified by ID2 presence. Significance was tested with two-sided Cox proportional-hazards regression adjusted for age, sex, smoking and tumour stage, without additional multiple-comparison adjustment. P values, HRs and patient numbers (brackets)

are reported. **d**, Enrichment of metastasis in ID2-positive tumours was evaluated using two-sided Fisher's exact test. ORs and P values are shown above the plot. **e**, Genomic alterations enriched in ID2-positive tumours were assessed using logistic regression adjusted for ancestry, sex, smoking, age and tumour purity. Multiple-testing correction was applied using the Benjamini–Hochberg method. The horizontal lines represent significance thresholds ($\text{FDR} < 0.05$ in orange, $\text{FDR} < 0.01$ in red). SCNA, somatic copy-number alterations. **f**, The distribution of hypoxia scores between ID2-positive and ID2-negative tumours. Sample sizes: AS_N ($n = 159$), EU_N ($n = 101$), EU_S ($n = 72$) and other ($n = 36$). P values were calculated using two-sided Wilcoxon rank-sum tests. The box plots show the median (centre line), IQR (box limits), $1.5 \times \text{IQR}$ (whiskers) and individual tumours.

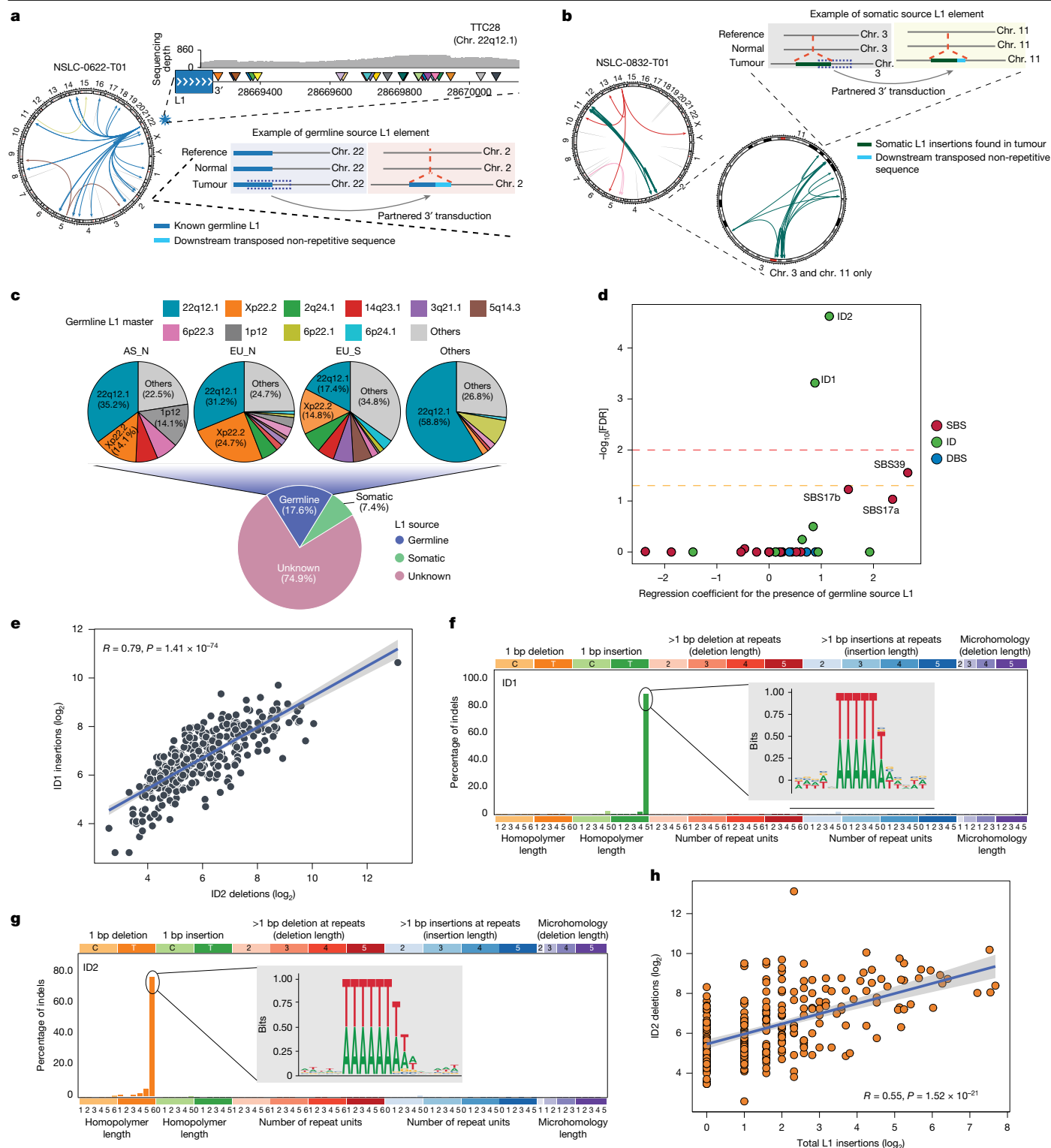


Fig. 4 | See next page for caption.

hypomethylation in lung and other tissues^{41,42}, and L1 hypomethylation in the blood of both cigarettes⁴³ or e-cigarettes⁴⁴ smokers. In fact, ID2 was more frequent in tumours from smokers. Targeted bisulfite sequencing of the germline L1 promoter on chromosome 22q12.1 in 80 tumours (40 ID2⁺, 40 ID2⁻) revealed notable demethylation in ID2⁺L1^{high} tumours, hypermethylation in matched normals and intermediate methylation in ID2⁻ tumours (Fig. 5b,c and Supplementary Fig. 27a,b). CpG methylation was inversely correlated with ID2 deletions (Supplementary Fig. 27c). Moreover, L1 RNA was higher in tumours than

normal tissue across all groups (Fig. 5d and Supplementary Fig. 28) and higher in ID2⁺ versus ID2⁻ tumours ($P = 0.021$; Fig. 5e). Individuals with a smoking history (including former smokers) exhibited higher L1 RNA in both tumour and normal tissues compared with never-smokers⁴³, suggesting a long-term impact of tobacco-related L1 retrotransposition (Supplementary Fig. 29). This effect may result from persistent epigenomic or transcriptional alterations that sustain L1 activity even after smoking cessation⁴⁵. Collectively, these findings support a role for reactivated L1 elements in driving the ID2 mutational signature in LUAD.

Fig. 4 | The association between L1 retrotransposition and the mutational signatures ID2 and ID1. **a**, Example tumour NSLC-0622-T01 containing L1 insertions from a germline source. In the Circos plot, the arrows indicate directions from germline L1 elements to somatic insertions in the tumour genome. The grey lines without arrowheads indicate insertions with unknown sources. Retrotranspositions from the master L1 on chromosome (chr.) 22q12.1 (highlighted in blue) are magnified. Partnered 3' transductions are shown as coloured triangles on chromosome 22 with a sequence depth plot (grey bars). One partnered 3' transduction between chromosome 22 and chromosome 2 illustrates germline-source retrotransposition. **b**, Example tumour NSLC-0832-T01 with predominantly somatic-source L1 insertions. The green arrows highlight multiple tumour-specific insertions. A magnified view of a somatic retrotransposition between chromosome 3 and chromosome 11 is shown.

ZNF695 contributes to L1 reactivation

To further investigate factors influencing L1 promoter methylation in both smokers and never-smokers, we analysed the Kruppel-associated box zinc-finger proteins (KRAB-ZFPs), the largest transcription factor family that silences transposable elements, including L1⁵ with cell-type-specific effects⁴⁶. To identify KRAB-ZFPs potentially involved in germline L1 retrotransposition activation in LUAD, particularly in tumours enriched for ID2 signature, we analysed RNA-seq data for 471 previously identified KZFP genes⁴⁶. Most expressed KZFP genes (390 out of 471) were positively correlated with each other (Supplementary Fig. 30), and 302 (77.4%) were differentially expressed in tumours versus normal lung (62.9% upregulated, 37.1% downregulated; Supplementary Fig. 31a). Among these, 43 also differed between smokers and never-smokers. Notably, *ZNF695* showed the greatest upregulation in tumours (FC = 4.8; FDR-adjusted $P = 1.73 \times 10^{-45}$) and the highest differential expression in smokers (FC = 2.7, FDR-adjusted $P = 0.01$; Supplementary Fig. 31b–d). *ZNF695* was also the most significantly upregulated KZFP gene in ID2⁺ tumours (FC = 3.5, FDR-adjusted $P = 1.14 \times 10^{-8}$; Fig. 6a,b), independent of other covariates (Supplementary Fig. 32). *ZNF695* expression was most significantly correlated with the burden of ID2 deletions ($R = 0.47$, $P = 1.11 \times 10^{-14}$; Fig. 6c,d) and L1 insertions (Supplementary Fig. 33). Moreover, the *ZNF695*-binding motif was enriched near demethylated CpG probes (± 1 kb) in germline L1 promoters ($P = 0.0028$; FDR-adjusted $P = 0.016$; Methods), and *ZNF695* expression was inversely correlated with L1 promoter methylation at chromosome 22q12.1 in tumours but not in normal samples (Fig. 6e). As expected, *ZNF695* expression correlated with the expression of proliferation markers (Supplementary Fig. 34). Single-cell analysis revealed elevated *ZNF695* expression in lung epithelial cells but low in immune cell populations, suggesting a potential inverse relationship with immunogenicity (Extended Data Fig. 9a). Notably, type I pneumocytes—the well-differentiated alveolar epithelial cells involved in oxygen exchange—showed much lower *ZNF695* expression than stem-like type II pneumocytes, potential cells of origin for LUAD⁴⁷, supporting the hypothesis of a role of *ZNF695* in adenocarcinoma genesis. This hypothesis was further supported by our lung single-nucleus multi-omics study integrating gene expression and chromatin accessibility⁴⁸. In this dataset, *ZNF695* expression was highest in a proliferating subpopulation of alveolar type 2 (AT2pro) cells among other normal lung cell types identified (Extended Data Fig. 9b). Cells with detectable *ZNF695* expression also exhibited significantly elevated L1 expression ($P = 0.04$; Extended Data Fig. 9c), further supporting an association between *ZNF695* and L1 activity.

To further validate the link between *ZNF695* expression and ID2 signature, we examined the expression of the *ZNF695*-target genes. Among 28 potential *ZNF695* targets identified from curated databases (Methods), 79% were differentially expressed between ID2⁺ and ID2⁻ tumours (Fig. 6f). Similarly, analysis of 70 downstream target genes regulated by KRAB-ZFPs revealed that 67% were differentially expressed with respect to ID2 status (Supplementary Fig. 35). These results strongly

A specific partnered 3' transduction illustrates somatic-source activity. **c**, The distribution of L1 insertion sources: germline, somatic and unknown (bottom), and proportions from specific germline masters (top). **d**, Enrichment of mutational signatures in tumours with germline-source L1 insertions, assessed by logistic regression adjusted for ancestry, sex, smoking status, age and tumour purity. Multiple-testing correction was applied using the Benjamini–Hochberg method. The horizontal lines indicate FDR thresholds <0.05 (orange) and <0.01 (red). **e**, Pearson correlation between deletions attributed to ID2 and insertions attributed to ID1, with two-sided coefficients and P values. **f,g**, Mutational signature profiles and motifs for ID1 (**f**) and ID2 (**g**). **h**, Pearson correlation analysis between deletions attributed to ID2 and total somatic L1 insertions; two-sided coefficients and P values are shown. The blue lines in **e** and **h** show the fitted regression with 95% CIs (grey).

implicate KRAB-ZFPs, including *ZNF695*, in the transcriptional regulation linked to ID2/L1 activity.

We extracted all RNA-seq reads corresponding to *ZNF695* and examined junction reads alongside sequencing coverage. Notably, non-canonical *ZNF695* transcripts were elevated in LUAD compared with in matched normal tissues (Supplementary Fig. 36). Furthermore, ID2⁺ tumours showed higher expression of non-canonical transcripts than ID2⁻ tumours. Canonical *ZNF695* transcripts encode zinc-finger proteins that recruit DNA methyltransferases, maintaining the DNA methylation required to repress germline L1 retrotransposition⁴⁹. By contrast, non-canonical transcripts, which lack the DNA-binding zinc-finger domain, may reduce methylation at the L1 loci through a dominant-negative effect (Extended Data Fig. 10). This observation suggests a dosage competition mechanism, whereby non-canonical *ZNF695* transcripts contribute to L1 reactivation in ID2⁺ tumours.

Discussion

In this comprehensive study of LUAD evolution, we used deep WGS and multi-omics analyses of high-clonality samples to determine the relative timing of various genomic changes and examine their associations with tobacco exposure, demographic factors and clinical outcomes.

Major cancer-driver genes, such as *EGFR* (dominant in never-smokers) and *KRAS* (in smokers), have been found to be mutated in histologically normal lung tissues⁵⁰ and in adenocarcinoma in situ⁵¹. We report here that they are mutated very early in LUAD evolution. The timing of the mutational signatures show that the signatures associated with exogenous exposures, such as SBS4 and SBS100 (tobacco smoking), mostly include clonal mutations, reflecting an early impact on tumour initiation, while endogenous-related signatures (APOBEC, SBS17a/b) appear later in the evolution. Taken together, one could expect to see early clonal driver gene mutations enriched with early occurring mutational patterns. Indeed, in smokers' tumours we found *KRAS* oncogene hotspots enriched with C:G>A:T transversions, the most frequent mutational type associated with tobacco smoking exposure. However, *EGFR* mutations in never-smokers were mostly attributed to age or unknown endogenous processes (SBS5/SBS40a signatures). Moreover, we observed that tumours with these driver genes have a profoundly different latency, with slow and fast subclonal diversification in *EGFR*-mutant and *KRAS*-mutant tumours, respectively. These results suggest that there could be two divergent LUAD evolutionary trajectories between smokers and never-smokers.

Specifically, in never-smokers, the epithelial tissue and related cancer-driver genes accumulate mutations with increasing age⁵². Mutant clones compete with one another⁵³ until one or a few clones (for example carrying *EGFR* mutations) acquires a fitness advantage and expands. The selection and expansion do not appear to be due to exogenous mutational processes, but they may be affected by a still unknown, probably endogenous, mutagenic process, although generating a low number of mutations (for example, SBS40a) or a promotion effect, for example, by stimulating inflammation^{14,50}. The

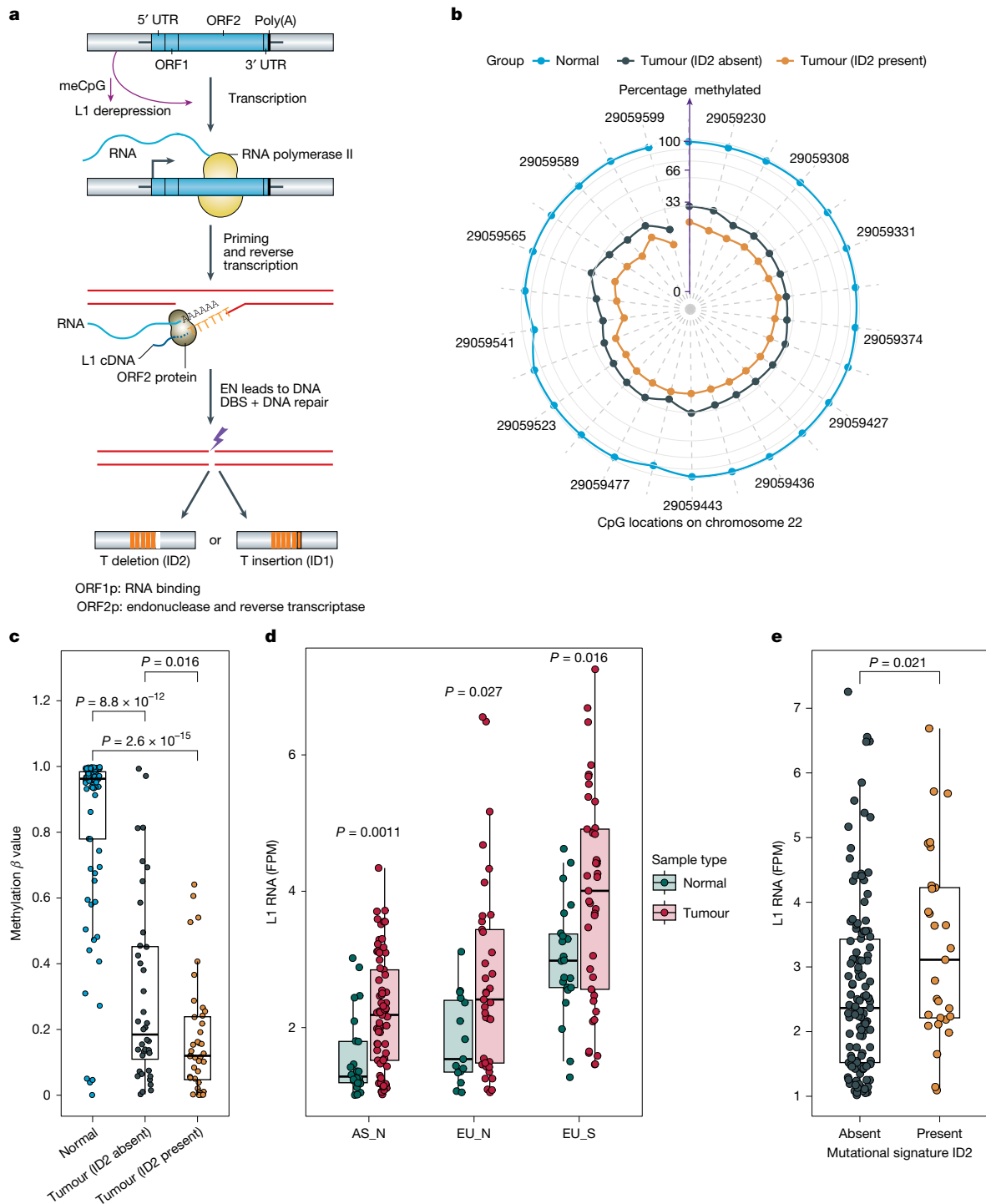


Fig. 5 | Activation of germline L1 retrotransposition due to DNA demethylation of L1 promoter. **a**, Schematic of the mobilization mechanism for L1. The diagram was adapted from ref. 33 (Springer Nature Limited), which illustrates non-LTR retrotransposons mobilizing through TPRT. ORF2-encoded endonuclease (EN) generates a single-strand nick, freeing a 3'-OH to prime reverse transcription of L1 RNA. Demethylation of CpG sites in the L1 promoter (purple arrow) is hypothesized to activate retrotransposition. Endonuclease activity, coupled with DNA repair, may produce single base-pair insertions or deletions at poly(A/T) tracts. **b**, DNA methylation levels at the L1 promoter region of a germline insertion on chromosome 22q12.1 were validated by targeted bisulfite sequencing. The polar plot shows the median methylation across CpG sites on chromosome 22q12.1, stratified by normal lung ($n = 80$),

tumours without ID2 signature ($n = 40$) and tumours with the ID2 signature ($n = 40$). Controls reflect 0%, 33.3%, 66.6% and 100% methylation at each CpG site (y-axis). **c**, The median methylation across CpG sites on chromosome 22q12.1, stratified by sample type and ID2 status. Sample sizes: normal lung ($n = 80$), tumours without ID2 ($n = 40$) and tumours with ID2 ($n = 40$). **d**, Total L1 RNA expression estimated from RNA-seq data, stratified by sample type. Sample sizes: AS_N ($n = 84$), EU_N ($n = 40$) and EU_S ($n = 41$). **e**, Total L1 RNA expression compared between tumours with ($n = 33$) and without ($n = 133$) the ID2 signature. All of the box plots show the median (centre line), IQR (box limits), $1.5 \times$ IQR (whiskers) and individual datapoints. P values from two-sided Wilcoxon rank-sum tests are displayed above the box plots.

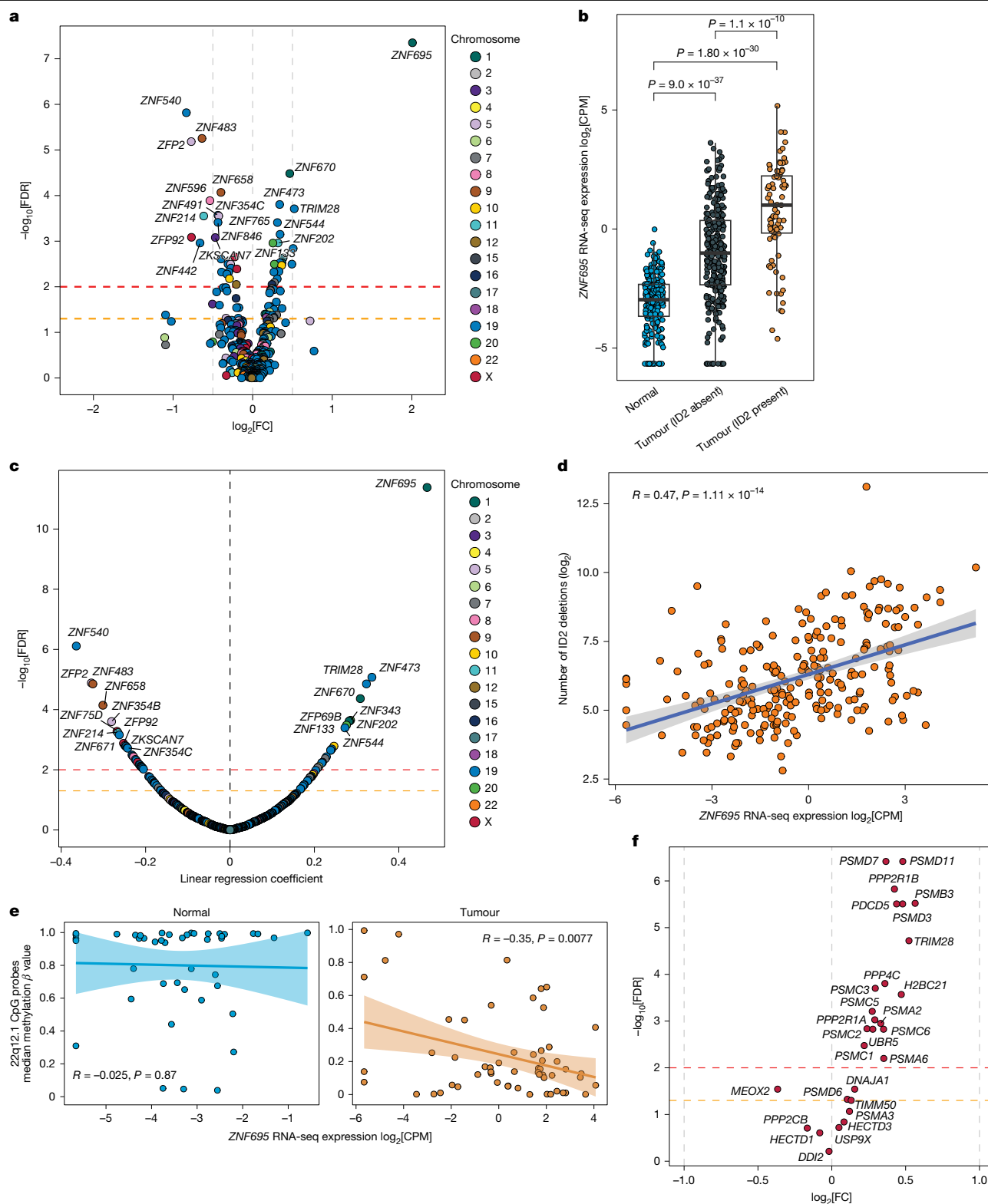


Fig. 6 | ZNF695 upregulation in tumours and its association with mutational signature ID2. **a**, Differential expression of KZFP protein-coding genes between tumours with and without ID2 signatures was assessed using two-sided Wilcoxon rank-sum tests with Benjamini–Hochberg correction. The horizontal dashed lines indicate significance thresholds (FDR-adjusted $P < 0.05$, orange; FDR-adjusted $P < 0.01$, red). The top 20 significant genes are annotated. **b**, ZNF695 expression in normal tissue or blood, tumours without ID2 and tumours with ID2. The box plots show the median (centre line), IQR (box limits), $1.5 \times$ IQR (whiskers) and individual datapoints. P values determined using two-sided Wilcoxon rank-sum tests are shown above the plots. **c**, Pearson correlations between KZFP gene expression and indels attributed to ID2 were computed

using Benjamini–Hochberg correction. The horizontal dashed lines mark the FDR thresholds (<0.05 (orange) and <0.01 (red)). The top 20 significant genes are annotated. **d**, Correlation between ZNF695 expression and ID2 deletions; two-sided Pearson correlation coefficients and P values are reported. The fitted regression line (blue) and 95% CIs (grey) are shown. **e**, Correlation between ZNF695 RNA-seq expression and median DNA methylation across CpG sites on chromosome 22q12.1. Two-sided Pearson correlation coefficients and P values are reported. Regression line (blue) and 95% CIs (grey) are shown. **f**, Differential expression of ZNF695-target genes between tumours with and without ID2 signatures, assessed using two-sided Wilcoxon rank-sum tests with Benjamini–Hochberg correction.

competition among subclones⁵³ continues during tumour progression, maintained by the lack of strong mutagenic effects facilitating specific subclones' positive selection and by poor immunoediting effects due to the low mutational burden⁵⁴. Together, these factors contribute to the long latency, which was particularly evident in tumours from female patients, possibly reflecting sex-differences in immune responses⁵⁵. This suggests that future or ongoing (such as NCT05164757) lung cancer screening programs for female non-smokers could consider long intervals between scans.

On the contrary, in smokers, the age-related accumulation of mutations is increased by the strong mutagenic effect of tobacco smoking over time. Once *KRAS*-associated clones are positively selected (*KRAS* has enrichment of C:G>A:T mutations in the hotspot), or sensitized through smoke-induced epigenomic changes⁵⁶, they rapidly expand, facilitated by the genomic instability, chromosomal rearrangements, frequent DSBs associated with L1 retrotransposition and other tobacco smoking-related factors. In fact, tumours with *KRAS* mutations have much shorter subclonal diversification than *KRAS*-negative tumours.

Overall, we did not observe substantial differences in the evolutionary patterns between AS_N and EU_N tumours, suggesting no major ancestry-related divergence in the evolutionary trajectories of never-smoker LUAD. However, future studies using analytical frameworks that offer formal statistical support and can model complex mutation timing relationships will be important to confirm or refine these observations. Notably, we observed the SBS22a signature almost exclusively in Taiwanese individuals. Although SBS22a appears to be mostly clonal, *EGFR* and other major cancer-driver genes in the Taiwanese samples are not enriched with the typical T:A>A:T mutation patterns of SBS22a. Thus, it is unclear whether this signature reflects a causative link between aristolochic acid exposure and LCINS carcinogenesis as has been observed in mice models of lung cancer⁵⁷, or just the exposure to aristolochic acid in this population.

Importantly, we found that the mutational signature ID2 is a marker of a mechanism for LUAD tumour evolution, which, if validated, could be added to the arsenal of precision medicine tools for prognosis and treatment. The indels contributing to the mutational signature ID2 were mostly clonal, and therefore present at tumour initiation. Similarly, ID2 was strongly enriched in tumours with short tumour latency. Both the associations of ID2 with tumour initiation and short latency were independent from major driver genes or demographic features affecting tumour evolution. This signature, which was two-to-three times more frequent in smokers than non-smokers, was enriched in fast-growing tumours with poor prognosis. Crucially, we showed that this signature, previously of unknown aetiology, could be a by-product of the activation of L1 retrotransposition, often from germline sources. The aggressive phenotype of ID2-positive tumours could be due to L1-related genomic instability, oncogene activation, tumour suppressor inactivation, metabolic reprogramming and immune modulation. L1 is the most abundant transposon element family in humans, occupying around 17% of genomic DNA. L1 could be reactivated in the genome through demethylation consequent to tobacco smoking exposure⁵⁸ and reduced epigenetic regulation. We found that *ZNF695*, a transcription factor that is upregulated in ID2-positive tumours and is located on chromosome 1q—a region that is typically amplified in lung cancers—was enriched with non-canonical transcripts lacking the DNA-binding zinc-finger domain and unable to recruit DNA methyltransferases, probably resulting in decreased methylation levels at the L1 locations due to a dominant-negative effect. While these findings provide strong correlative evidence linking *ZNF695* to L1 regulation, further experiments, such as CRISPR-based functional studies and single-cell approaches, will be required to validate this relationship and elucidate the mechanistic role of *ZNF695* in regulating L1 retrotransposition as well as its potential as a therapeutic target. Notably, previous studies may have failed to recognize the important role of L1 retrotransposition on LUAD evolution because they were largely based on whole-exome sequencing data

(for example, TRACERx), which cannot capture non-coding genomic regions, including transposable elements.

Different evolutionary trajectories may impact clinical course and treatment responses. The longer latency of *EGFR*-mutant tumours may allow the accumulation of mutations increasing resistance to tyrosine kinase inhibitors (for example, the *EGFR* T790M mutation)⁵⁹, highlighting the potential value of upfront combination therapies to suppress treatment resistance mediated by the outgrowth of competing clones under selective therapeutic pressure. At the same time, the long latency of *EGFR*-mutant cancers provides a wide time window for tumour early detection. By contrast, the short subclonal diversification seen in *KRAS*-mutation-positive tumours suggests a more focused and potent *KRAS*-directed therapeutic strategy. Given the distinctly aggressive phenotype of ID2-positive LUADs, the presence of ID2 signature could be explored as a potential biomarker to escalate therapies. While the low-mutation-related neoantigen in ID2-positive LUADs could lead to poor immune recognition and may limit the clinical efficacy of currently available immune checkpoint inhibitors, the interplay of other immune cells in these tumours (Supplementary Fig. 19) suggests that novel immunotherapies harnessing these immune cell subsets could represent a more effective strategy. L1, the ORF2 protein and transposable elements are the subject of increasing research into their role in stabilizing the genome, their dynamic relationship with DNA methylation and their potential to be therapeutically targeted³⁸, possibly exploiting tumour-specific antigens derived from transposable elements⁶⁰. These findings highlight the need for future research to explore the relationship between tumour latency, genomic changes, transposable elements and clinical outcomes in LUAD.

In summary, we delineate distinct LUAD evolutionary paths, highlight ID2 as a signature of rapid tumour growth linked to L1 activity and nominate *ZNF695* as a putative regulator of L1 reactivation, providing mechanistic insights with potential prognostic and therapeutic implications.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-025-09825-y>.

- Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).
- Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
- Alexandrov, L. B. et al. Mutational signatures R. *Science* **354**, 618–622 (2016).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Yang, P., Wang, Y. & Macfarlan, T. S. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet.* **33**, 871–881 (2017).
- Zhang, Y. et al. Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. *Lancet Oncol.* **24**, 1206–1218 (2023).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
- Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **616**, 525–533 (2023).
- Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
- Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
- Mendez-Dorantes, C. et al. Chromosomal rearrangements and instability caused by the LINE-1 retrotransposon. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.12.14.628481> (2024).
- Rodić, N. et al. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat. Med.* **21**, 1060–1064 (2015).
- Nguyen, T. H. M. et al. L1 retrotransposon heterogeneity in ovarian tumor cell evolution. *Cell Rep.* **23**, 3730–3740 (2018).
- Zhang, T. et al. Genomic and evolutionary classification of lung cancer in never smokers. *Nat. Genet.* **53**, 1348–1359 (2021).
- Landi, M. T. et al. Tracing lung cancer risk factors through mutational signatures in never-smokers. *Am. J. Epidemiol.* **190**, 962–976 (2021).
- Zhu, B. et al. The genomic and epigenomic evolutionary history of papillary renal cell carcinomas. *Nat. Commun.* **11**, 3096 (2020).

17. Senkin, S. et al. Geographic variation of mutagenic exposures in kidney cancer genomes. *Nature* **629**, 910–918 (2024).
18. Diaz-Gay, M. et al. The mutagenic forces shaping the genomes of lung cancer in never smokers. *Nature* **644**, 133–144 (2025).
19. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
20. Otlu, B. et al. Topography of mutational signatures in human cancer. *Cell Rep.* **42**, 112930 (2023).
21. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
22. Whitfield, M. L., George, L. K., Grant, G. D. & Perou, C. M. Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).
23. Emami Nejad, A. et al. The role of hypoxia in the tumor microenvironment and development of cancer stem cell: a novel approach to developing treatment. *Cancer Cell Int.* **21**, 62 (2021).
24. Bhandari, V., Li, C. H., Bristow, R. G., Boutros, P. C. & PCAWG Consortium. Divergent mutational processes distinguish hypoxic and normoxic tumours. *Nat. Commun.* **11**, 737 (2020).
25. Liu, N. et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* **553**, 228–232 (2018).
26. Li, X. et al. LINE-1 transcription activates long-range gene expression. *Nat. Genet.* **56**, 1494–1502 (2024).
27. Scott, E. C. & Devine, S. E. The role of somatic L1 retrotransposition in human cancers. *Viruses* **9**, 131 (2017).
28. Tubio, J. M. C. et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
29. McKerrow, W. et al. LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint. *Proc. Natl Acad. Sci. USA* **119**, e2115999119 (2022).
30. Kazazian, H. H. Jr & Moran, J. V. Mobile DNA in health and disease. *N. Engl. J. Med.* **377**, 361–370 (2017).
31. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294 (2019).
32. Nam, C. H. et al. Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* **617**, 540–547 (2023).
33. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
34. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357**, 1383–1393 (2006).
35. Morrish, T. A. et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31**, 159–165 (2002).
36. Farkash, E. A. & Luning Prak, E. T. DNA damage and L1 retrotransposition. *J. Biomed. Biotechnol.* **2006**, 37285 (2006).
37. Suzuki, J. et al. Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. *PLoS Genet.* **5**, e1000461 (2009).
38. Baldwin, E. T. et al. Structures, functions and adaptations of the human LINE-1 ORF2 protein. *Nature* **626**, 194–206 (2024).
39. Freeman, B. et al. Analysis of epigenetic features characteristic of L1 loci expressed in human cells. *Nucleic Acids Res.* **50**, 1888–1907 (2022).
40. Rodgers, K. & McVey, M. Error-prone repair of DNA double-strand breaks. *J. Cell. Physiol.* **231**, 15–24 (2016).
41. Wangsri, S., Subbalekha, K., Kitkumthorn, N. & Mutirangura, A. Patterns and possible roles of LINE-1 methylation changes in smoke-exposed epithelia. *PLoS ONE* **7**, e45292 (2012).
42. Stueve, T. R. et al. Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Hum. Mol. Genet.* **26**, 3014–3027 (2017).
43. Caliri, A. W., Caceres, A., Tommasi, S. & Besaratinia, A. Hypomethylation of LINE-1 repeat elements and global loss of DNA hydroxymethylation in vapers and smokers. *Epigenetics* **15**, 816–829 (2020).
44. Camila, B. et al. Genotoxicity and hypomethylation of LINE-1 induced by electronic cigarettes. *Ecotoxicol. Environ. Saf.* **256**, 114900 (2023).
45. Joehanes, R. et al. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
46. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
47. Han, G. et al. An atlas of epithelial cell states and plasticity in lung adenocarcinoma. *Nature* **627**, 656–663 (2024).
48. Long, E. et al. Context-aware single-cell multiomics approach identifies cell-type-specific lung cancer susceptibility genes. *Nat. Commun.* **15**, 7995 (2024).
49. Rossopopoff, O. & Trono, D. Take a walk on the KRAB side. *Trends Genet.* **39**, 844–857 (2023).
50. Hill, W. et al. Lung adenocarcinoma promotion by air pollutants. *Nature* **616**, 159–167 (2023).
51. Haga, Y. et al. Whole-genome sequencing reveals the molecular implications of the stepwise progression of lung adenocarcinoma. *Nat. Commun.* **14**, 8375 (2023).
52. Huang, Z. et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat. Genet.* **54**, 492–498 (2022).
53. Colom, B. et al. Mutant clones in normal epithelium outcompete and eliminate emerging tumours. *Nature* **598**, 510–514 (2021).
54. Jardim, D. L., Goodman, A., de Melo Gagliato, D. & Kurzrock, R. The challenges of tumor mutational burden as an immunotherapy biomarker. *Cancer Cell* **39**, 154–173 (2021).
55. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
56. Vaz, M. et al. Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. *Cancer Cell* **32**, 360–376 (2017).
57. Mengs, U. Tumour induction in mice following exposure to aristolochic acid. *Arch. Toxicol.* **61**, 504–505 (1988).
58. Ambatipudi, S. et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* **8**, 599–618 (2016).
59. Kobayashi, S. et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
60. Shah, N. M. et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet.* **55**, 631–639 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. ²Manchester Cancer Research Centre, The University of Manchester, Manchester, UK. ³Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. ⁴Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. ⁵Moore's Cancer Center, University of California San Diego, La Jolla, CA, USA. ⁶Digital Genomics Group, Structural Biology Program, Spanish National Cancer Research Center (CNIO), Madrid, Spain. ⁷Advanced Technology Center for Aging Research, IRCCS INRCA, Ancona, Italy. ⁸Clinic of Laboratory and Precision Medicine, IRCCS INRCA, Ancona, Italy. ⁹Department of Pathology, Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada. ¹⁰Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ¹¹Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. ¹²Institut Universitaire de Cardiologie et de Pneumologie de Québec—Université Laval, Quebec City, Quebec, Canada. ¹³Westat, Rockville, MD, USA. ¹⁴Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. ¹⁵Queen Mary Hospital, The University of Hong Kong, Hong Kong, China. ¹⁶Department of Pathology, The University of Hong Kong, Hong Kong, China. ¹⁷Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan. ¹⁸Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan. ¹⁹Department of Surgery, Division of Thoracic Surgery, Chung Shan Medical University Hospital, Taichung, Taiwan. ²⁰Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, USA. ²¹Biobanco IBSP-CV FISABIO, Valencia, Spain. ²²Red Valenciana de Biobancos FISABIO, Valencia, Spain. ²³Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. ²⁴Thoracic Surgery, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. ²⁵Department of Cancer Epidemiology and Primary Prevention, Maria Skłodowska-Curie National Research Institute of Oncology, Warsaw, Poland. ²⁶Department of Environmental Epidemiology, Nofer Institute of Occupational Medicine, Łódź, Poland. ²⁷Department of Clinical Epidemiology, N.N. Blokhin National Medical Research Centre of Oncology, Moscow, Russia. ²⁸Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic. ²⁹Department of Oncology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic. ³⁰Department of Occupational Health and Toxicology, National Center for Environmental Risk Monitoring, National Institute of Public Health, Bucharest, Romania. ³¹International Organisation for Cancer Prevention and Research (IOCPR), Belgrade, Serbia. ³²Department of Thoracic Surgery, Clinical Center of Serbia, Belgrade, Serbia. ³³Sylvester Comprehensive Cancer Center, Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, USA. ³⁴Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ³⁵Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. ³⁶Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France. ³⁷Princess Margaret Cancer Center, University of Toronto, Toronto, Ontario, Canada. ³⁸IHU RespirERA, Biobank-BB-0033-0025, Côte d'Azur University, Nice, France. ³⁹Department of Pathology, Yale School of Medicine, New Haven, CT, USA. ⁴⁰Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy. ⁴¹Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy. ⁴²Ben May Department for Cancer Research, The University of Chicago, Chicago, IL, USA. ⁴³Department of Human Genetics, The University of Chicago, Chicago, IL, USA. ⁴⁴The University of Chicago Medicine Comprehensive Cancer Center, The University of Chicago, Chicago, IL, USA. ⁴⁵Sanford Stem Cell Institute, University of California San Diego, La Jolla, CA, USA. ⁴⁶Department of Mathematics, Harvard University, Cambridge, MA, USA. ⁴⁷Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. ⁴⁸Manchester NIHR Biomedical Research Centre, Manchester, UK. [✉]e-mail: landim@mail.nih.gov

Methods

Collection of lung cancer samples

In the Sherlock-Lung study, we identified 1,024 patients with histologically confirmed LUAD (Supplementary Table 1). Samples were collected from multiple centres across four continents¹⁵, including 286 smokers, 737 never-smokers and 1 individual of unknown smoking status. The fresh-frozen tumour tissue with matched whole-blood samples or fresh frozen normal lung tissue sampled, when possible, at least 3 cm from the tumour were collected from these treatment-naïve patients. We followed rigorous criteria for sample inclusion, similar to previous publications¹⁴: (1) Sequencing coverage. We maintained a minimum average sequencing coverage of $>40\times$ for tumour samples and $>25\times$ for normal samples. (2) Contamination and relatedness. Cross-sample contamination was limited to $<1\%$ by Conpair⁶¹, and detected relatedness was maintained <0.2 by Somailier⁶². (3) Copy-number analysis. Participants with abnormal copy-number profiles in normal samples were excluded, as determined by Battenberg⁶³. (4) Mutational signatures. Tumour samples exhibiting mutational signatures SBS7 (two samples, associated with ultraviolet light exposure) and SBS31 (one sample, associated with platinum chemotherapy) were removed from the analysis. (5) WGS quality control. Tumour samples with a total genomic alteration count of <100 or $<1,000$ along with NRPPC (number of reads per chromosomal copy)⁶⁴ of <10 were excluded. (6) Multiple-region sequencing. In cases in which multiple regions of a tumour were sequenced, only one high-purity tumour sample was included to avoid redundancy. These stringent criteria were applied consistently to ensure the robustness and reliability of the data collected for the Sherlock-Lung study.

As NCI exclusively received deidentified samples and data from collaborating centres, had no direct contact or interaction with the study participants, and did not use or generate any identifiable private information, Sherlock-Lung has been classified as ‘not human subject research’ in accordance with the Federal Common Rule (45 CFR 46; eCFR.gov). Collaborating centres obtained informed consent for publication of human data from participants under protocols approved by their respective Institutional Review Boards.

Sample selection for tumour evolution analysis

To ensure adequate statistical power for identifying subclone architectures and constructing lung cancer evolutionary histories, we used a metric known as NRPPC⁶⁴ to select the appropriate samples for clonal evolution analyses. As expected, our ability to detect subclones was found to depend on the NRPPC, rather than the number of identified mutations. By setting a threshold of NRPPC >10 , we could reliably identify subclones with a cancer cell fraction exceeding 40%. Overall, approximately 53.2% of the LUAD samples exhibited evidence of at least one subclone expansion, which is higher than the LUAD tumours (mostly from smokers) in the PCAWG study (47.4%), possibly due to the higher sequencing depth in the Sherlock-Lung study. We selected 542 LUAD tumour samples with NRPPC >10 , purity >0.3 and high-quality copy-number profiles as we need clonality data for the downstream tumour evolution analyses. We grouped these tumours into four categories based on the patients’ tobacco smoking status and genetic ancestry, including East Asian-ancestry never-smokers, mostly from Taiwan and Hong Kong (AS_N, $n = 180$), European-ancestry never-smokers, mostly from Europe and North America (EU_N, $n = 184$), European-ancestry smokers mostly from Europe and North America (EU_S, $n = 120$) and an indeterminate group (others). The ‘others’ group comprised 6 African-ancestry individuals, 16 individuals who did not cluster with 1KGP EUR, EAS or AFR reference panels, and 36 from either East Asian or European-ancestry with uncertain tobacco exposure, based on the presence of SBS4 mutational signature and tobacco smoking history.

Specimen processing

The methods for specimen collection and processing have been described in our previous publication¹⁴. Genomic DNA was extracted from fresh-frozen tissue according to the protocols outlined in the QIAmp DNA Mini Kit (Qiagen) user manual. The concentration of DNA was determined using a Nanodrop spectrophotometer. Subsequently, DNA quantification, normalization and fragmentation were performed using the same procedure as our standard WGS process¹⁴. Each DNA sample was required to meet minimum mass and concentration thresholds and had to exhibit no signs of contamination or profile discordance as indicated by the identifier assay. Samples meeting these requirements were then subdivided into aliquots, with mass adjusted to meet the specific requirements for downstream assay processing.

WGS processing

WGS library construction followed our previously reported protocol¹⁴. The Broad Institute (<https://www.broadinstitute.org>) conducted WGS sequencing on the Illumina HiSeq X and NovaSeq6000 platforms following Illumina protocols for 2×150 bp paired-end sequencing. FASTQ files were generated after Illumina base-calling. Paired FASTQ were converted into unmapped BAM using the GATK pipeline, available at <https://github.com/gatk-workflows/seq-format-conversion>. Subsequently, the unmapped BAM files were further processed using GATK on the cloud-based platform TERRA workspaces (<https://app.terra.bio>). The sequencing data were aligned to the human reference genome GRCh38, and the resulting aligned BAM files were then transferred to the NIH HPC system (<https://hpc.nih.gov>) for subsequent downstream analyses.

For the public WGS data, the preprocessed aligned BAM/CRAM files (including unmapped reads) were first converted back to FASTQ files using Bazam (v.1.0.1)⁶⁵ to retain the sequencing lane and read group information and then processed using the same pipeline as for the Sherlock-Lung WGS dataset. As a validation dataset to examine the relationship between ZNF695, mutational signature ID2 and L1 insertions, we downloaded newly released raw BAM files of TCGA LUAD samples from the NCI GDC data portal (<https://portal.gdc.cancer.gov/>). We excluded tumour samples previously analysed as part of the Sherlock-Lung dataset, resulting in a total of 258 TCGA LUAD samples for validation. We followed the same analytical pipelines as in the Sherlock-Lung dataset.

Defining high-quality WGS dataset

As indicated in the PCAWG study⁶⁴, our ability to detect subclones was determined by the NRPPC value, rather than the quantity of identified mutations⁶⁴. Thus, for tumour evolution analysis, we selected a set of 542 LUAD tumour samples with high-quality WGS data as our discovery set. The selection was based on the following criteria. First, we required a tumour purity greater than 0.3, indicating a substantial proportion of tumour cells within the samples. Second, we set a minimum NRPPC value of 10 to ensure the robustness of subclone detection throughout tumour evolution analysis. Last, we included only samples with high-quality copy-number profiles generated by Battenberg⁶³, without any involvement of refitting steps, thereby eliminating the potential somatic copy-number artifacts. We used VerifyBamID2 to project germline genetic principal components for these samples into the principal component space defined by the 1000 Genome Project (1KGR) reference panel (Supplementary Fig. 1). We labelled samples that clustered with the 1KGP EUR sample as European ancestry ($n = 335$); those that clustered with the 1KGP EAS sample as East Asian ancestry ($n = 185$); and those that clustered with the 1KGP AFR sample as African ancestry ($n = 6$). These analyses included 400 non-smokers and 141 smokers, and 362 female and 180 male individuals.

Identifying genome-wide alterations

The somatic mutation calling was performed using our established bioinformatics pipeline as previously described¹⁴. Our approach

involved applying four distinct mutation calling algorithms for tumour-normal paired analysis, including Strelka (v.2.9.10), MuTect, MuTect2 and TNscope implemented in the Sentieon's genomics software (v.202010.01). We used an ensemble method to merge the results from these different callers followed by additional filtering to reduce false-positive calling. The final mutation calls for both single-nucleotide variants (SNVs) and indels were required to meet the following criteria: (1) read depth >12 in tumour samples and >6 in normal samples; (2) variant allele frequency <0.02 in normal samples; and (3) overall allele frequency < 0.001 in multiple genetics databases, including 1000 Genomes (phase 3 v.5), gnomAD exomes (v.2.1.1) and gnomAD genomes (v.3.0). For indel calling, we retained only variants identified by at least three algorithms. The IntOGen pipeline (v.2020.02.0123)⁶⁶, which combines seven state-of-the-art computational methods, was used with the default parameters to detect signals of positive selection in the mutational patterns of driver genes across the cohort. However, additional functional studies are needed to validate the driver potential of certain genes, particularly large genes such as *LRP1B*.

We used the Battenberg algorithm (v.2.2.9)⁶³ to conduct analyses of SCNAs. Initial SCNA profiles were generated, followed by an assessment of the clonality of each segment, purity and ploidy. Any SCNA profile determined to be low-quality after manual inspection underwent a refitting process. This process required new tumour purity and ploidy inputs, either estimated by ccube (v.1.0)⁶⁷ or recalculated from local copy-number status. The Battenberg refitting procedures were iteratively executed until the final SCNA profile was established and met the criteria of manual validation check. GISTIC (v.2.0) was used to identify the recurrent copy-number alterations at the gene level based on the major clonal copy number for each segmentation. Meerkat (v.0.189)⁶⁸ and Manta (v.1.6.0)⁶⁹ were applied with the recommended filtering for identifying structural variants (SVs), and the union set of these two callers was merged as the final SV dataset.

To identify putative transposable elements (TE), we used a pipeline called TraFiC-mem (v.1.1.0) (Transposome Finder in Cancer)²⁸ retaining only TE insertions that passed the default filtering criteria. The source element of L1 retrotransposition can be classified into two categories^{10,28}: germline and somatic source L1 elements. Germline source L1 elements are identified by unique DNA regions retrotransposed somatically elsewhere in the cancer genome from a single locus of a reference full-length L1 element or a putative non-reference polymorphic L1 element. Somatic source L1 elements are those derived from the downstream region of a putative L1 event present in the tumour genome but absent in the matched normal genome.

Collectively, we define the overall somatic alterations as the summary of SNVs, indels, SVs and TE insertions.

Defining driver mutations

To identify driver mutations within the set of identified driver genes, we implemented a rigorous and multifaceted strategy, considering multiple criteria: (1) the presence of truncating mutations specifically in genes annotated as tumour suppressors; (2) the recurrence of missense mutations in a minimum of 3 samples; (3) mutations designated as likely drivers with a boostDM score⁷⁰ exceeding 0.5; (4) mutations categorized as either oncogenic or likely oncogenic based on the criteria established by OncoKB⁷¹, an expert-guided precision oncology knowledge base; (5) mutations previously recognized as drivers in the TCGA MC3 drivers paper⁷²; and (6) missense mutations characterized as likely pathogenic in genes that are annotated as tumour suppressors, as described in AlphaMissense previously⁷³. Any mutation meeting one or more of these criteria will be recognized as a potential driver mutation. This comprehensive approach ensures a thorough and exhaustive identification of driver mutations within the context of the identified driver genes.

Timing mutations and copy-number gains

We used the R package MutationTimeR (v.1.00.2) to estimate the timing of somatic mutations relative to clonal and subclonal copy-number states and to calculate the relative timing of copy-number gains⁷. This analysis used all somatic mutations, including SNVs and indels, alongside the final copy-number profiles as input data for the MutationTimeR algorithm.

MutationTimeR integrates mutation allele frequencies, local copy-number states and tumour purity to determine whether mutations occurred before or after copy-number gains. The algorithm identifies whether mutations are present on alleles in chromosome regions with gains or losses. Using allele-specific copy-number profiles, it assigns each mutation to a specific copy-number state.

For diploid regions without copy-number gains, MutationTimeR classifies mutations as clonal if they are present on one copy per cell or as subclonal if present on fewer than one copy per cell. In regions with copy-number gains (for example, biallelic gains), clonal mutations are further classified into two temporal categories based on their relationship to the copy-number event:

- (1) Early clonal mutations: mutations occurring before the copy-number gain and present on all tumour cell copies (≥ 2 copies per cell). These mutations are clonal and often represent early driver events.
- (2) Late clonal mutations: mutations occurring after the copy-number gain, present on only one copy per cell.

Using Bayesian inference, MutationTimeR estimates the likelihood of mutations occurring in the context of specific copy-number changes, enabling precise temporal classification.

For cancer-driver genes, we calculated the proportion of each mutational clonality category across all samples. Moreover, we aggregated clonality information for each mutational signature by considering the probability of each mutation type being assigned to its respective signature.

Chronological time analysis

To determine the temporal expansion of tumours, we used a methodology of chronological time analysis derived from previous studies^{7,14}, including the estimation of the emergence of the MRCA and the tumour latency between the MRCA and the age at diagnosis (Supplementary Fig. 8). The approach assumes a constant mutation rate per year, particularly for CpG>TpG mutations caused by the spontaneous deamination of 5-methylcytosine to thymine at CpG dinucleotides, which exhibit clock-like behaviour. In summary, a hierarchical Bayesian linear regression model was used to establish the relationship between the age at diagnosis and the scaled number of clock-like CpG>TpG mutations occurring in an NpCpG context (SBS1). This model accounted for key variables, including tumour ploidy, tumour purity, and subclonal architecture, as described in the PCAWG evolutionary study⁷ (<https://github.com/gerstung-lab/PCAWG-11/>).

For each tumour, the age at the occurrence of the MRCA was estimated. Given the low mutational burden observed in LCINS, the most frequent group in our cohort, an estimated tumour acceleration rate of $1\times$ was applied. Tumour latency, representing the time from subclonal diversification to diagnosis, was calculated as the difference between the estimated age at the MRCA and the age at tumour diagnosis. An implicit assumption is that the latency between the emergence of the last detectable subclone and the age at diagnosis is relatively short compared to the overall timeline of tumour development. This approach provides critical insights into the evolutionary dynamics of tumours, highlighting differences in clonal expansion and subclonal diversification across tumour subtypes.

Mutational signature analysis

The detailed methods for the mutational signature analysis have been included in a separate mutational signature article¹⁸. In brief,

SigProfilerMatrixGenerator⁷⁴ was used to create the mutational matrices for all types of somatic mutations, including single base substitutions (SBSs), doublet base substitutions (DBSs) and indels (IDs). De novo mutational signatures for SBSs, DBSs and IDs were extracted using SigProfilerExtractor⁷⁵ (v.1.1.21) using the default parameters, normalization set to 10,000 mutations and splitting the cohort between smokers, and non-smokers and unknown smoking status, to limit the effect of hypermutators in the signature extraction process. De novo signatures were extracted using the SBS-288, DBS-78 and ID-83 mutational contexts, respectively, and later decomposed into COSMIC v.3.4 reference signatures⁷⁶, based on the GRCh38 reference genome, and assigned to individual samples using SigProfilerAssignment⁷⁷ (v.0.1.0). Two novel signatures were identified, one smoking-associated SBS signature, termed SBS100, extracted de novo in the smoker cohort and presenting similarities to previous signatures reported in Signal⁷⁸; and one ID signature, termed ID24, extracted de novo in the non-smoker cohort and subsequently used to decompose other de novo extracted signatures in the smoker cohort.

RNA-seq data analysis

The RNA-seq data from the Sherlock-Lung study were generated at the Cancer Genomics Research Laboratory (CGRL), Division of Cancer Epidemiology and Genetics, National Cancer Institute, using the Illumina HiSeq platform based on the Illumina TruSeq Stranded Total RNA-seq protocol, which resulted in the production of 2 × 100 bp paired-end reads. The RNA-seq data from the TCGA-LUAD study⁷⁹ were downloaded from the GDC Data Portal⁸⁰. The FASTQ files were aligned to the human reference genome GRCh38/hg38 using STAR (v.2.7.3a)⁸¹, annotated using the GENCODE v35 and processed by HTSeq (v.0.11.4)⁸² for gene quantification. The resulting expression data were corrected for batch effects using the R package ComBat-seq (v.3.48.0)⁸³ and normalized to counts per million (CPM), followed by log₂ transformation using the R package edgeR (v.3.42.4)⁸⁴. Only ‘expressed’ genes defined by CPM > 0.1 in at least 10% quality-control-passed samples were included in the final quantification data. Pathway analyses were performed using Ingenuity Pathway Analysis software and also the traditional gene set enrichment analysis (GSEA) using the hallmark gene sets from mSigDB database (<https://www.gsea-msigdb.org/>)⁸⁵. The TIMER2.0 web tool⁸⁶ was used for immune cell decomposition based on the normalized gene expression matrix.

Differential expression of KRAB-ZFP targets

For the transcription factor ZNF695, we curated target genes from two major databases: the TE enrichment KRAB-ZFP database⁸⁷ (KRABopedia: <https://tronoapps.epfl.ch/web/krabopedia/>) and the IntAct molecular interaction database⁸⁸ (<https://www.ebi.ac.uk/intact/home>). Notably, the KRABopedia database indicated that L1 elements are among the most significantly enriched peaks in ZNF695 chromatin immunoprecipitation–sequencing experiments. Integration of these two sources resulted in a collection of 28 potential ZNF695 target genes. Moreover, we obtained 70 downstream target genes regulated by general KRAB-ZFPs from a previous study⁸⁹. To confirm the link between ZNF695 expression and ID2 signatures, we examined the association between ID2 signatures and the transcriptional activity of KRAB-ZFPs, as indicated by the expression levels of these target genes. Differential expression analysis was performed using the Wilcoxon rank-sum test on the RNA-seq data.

L1 RNA quantification from bulk RNA-seq

L1 locus-specific RNA expression was quantified using the LIEM software⁹⁰ based on RNA-seq data. LIEM uses the expectation-maximization algorithm to estimate gene expression specific to L1 loci and to differentiate proper L1 expression from passive co-transcription, which includes L1 RNA but does not support retrotransposition. To quantify intact L1 RNA, we considered full-length L1 loci with no stop codon

present in either ORF1 or ORF2 as expressed if they meet the following criteria: at least two read pairs per million (FPM) were assigned to that locus, and less than 10% of the RNA assigned to that locus was estimated to be result of passive co-transcription. The total expression of intact L1 RNA was estimated by aggregating the FPM values for each locus.

ZNF695 and L1 expression in scRNA-seq

To investigate ZNF695 expression at the single-cell level, we queried the cellxgene database (<https://cellxgene.cziscience.com/gene-expression>), a comprehensive repository of curated single-cell RNA-seq (scRNA-seq) datasets, to evaluate ZNF695 expression across diverse cell types from lung tissue studies. Expression data, including raw and scaled values, were extracted from lung-specific datasets, with cell types annotated based on the original study metadata.

Moreover, we analysed ZNF695 and L1 expression using our previously published lung single-cell multiomics dataset⁴⁸. Raw sequencing data in FASTQ format were downloaded from this study to quantify L1 expression. To improve the accuracy of TE expression profiling, particularly for L1 elements, we applied the Stellarscope pipeline⁹¹—a bioinformatics tool that reassigns multimapped reads to specific TE loci using an expectation-maximization algorithm. Stellarscope was run with the default parameters to generate a per-cell L1 count matrix. This matrix, along with cell annotations, was processed using standard scRNA-seq workflows in Seurat (v.5.0). Normalized L1 expression values from specific loci were aggregated to calculate overall L1 expression per cell. Cells were then classified based on detectable ZNF695 expression, determined from gene expression profiles generated in Seurat. Differential L1 expression between ZNF695-positive and ZNF695-negative cells was evaluated using a Wilcoxon rank-sum test, with statistical significance defined as $P < 0.05$.

Genome-wide DNA methylation analysis

Genome-wide DNA methylation was profiled on the Illumina Human-MethylationEPIC BeadChip (Illumina). Genomic DNA was extracted, and DNA methylation was measured according to Illumina’s standard procedure at the Cancer Genomics Research Laboratory (CGRL), Division of Cancer Epidemiology and Genetics, National Cancer Institute. Raw DNA methylation data (.idat files) were generated for our study, combined with TCGA LUAD⁷⁹ methylation data collected from the GDC Data Portal⁸⁰ (Illumina Human Methylation 450k BeadChip) and processed using RnBeads (v.2.0)⁹². Background correction (enmix.oob) and beta-mixture quantile normalization (BMIQ) were applied. Unreliable probes (GreedyCut algorithm with detection $P < 0.05$), cross-reactive probes and probes mapping to sex chromosomes were removed. Samples with outlier intensities in 450k/EPIC array control probes were removed from the dataset described in the RnBeads vignette. Principal component analysis of these samples was performed based on the genotyping probes and we removed individuals with a Euclidean distance between matched tumour/normal pair > upper quartile + 3 × IQR. Moreover, we performed Combat batch correction separately for tumours and normal samples using the sva R package⁹³ on M -values for known technical factors, including studies, collection sites, sample plates, sample wells, array types (450K and EPIC), chip ID and positions. We used MBatch (v.2.0; <https://github.com/MD-Anderson-Bioinformatics/BatchEffectsPackage>) to assess the presence of batch effects. The Dispersion Separability Criteria (DSC) values were <0.5 for all technical factors, suggesting that batch effects were negligible after batch correction. The corrected methylation-level beta values were used for downstream association analyses.

Neoantigen analysis

To predict neoantigens, patient-specific HLA haplotypes were identified using HLA-HD (v.1.2.1)⁹⁴. The software NetMCHpan4.1⁹⁵ was then run on 9–11 neo peptides derived from all nonsynonymous mutations, taking into account the patient’s specific HLA genotypes. NetMCHpan4.1 was

included in the NeoPredPipe pipeline (v.1.1)⁹⁶. The decomposed mutation probabilities generated by the SigProfilerExtractor were used to predict the neoantigens derived from specific mutational signatures. For TCGA LUAD data, the predicted SNV neoantigen counts per sample were collected from a previous publication⁹⁷. We used the number of peptides predicted to bind to MHC proteins as the neoantigen load for our analysis.

Enrichment of *ZNF695* motifs in L1

The DMP analysis between tumours exhibiting ID2 signatures and those without was performed by the Wilcoxon rank-sum testing. We identified significantly demethylated probes in ID2 positive tumours with a FDR < 0.05 after multiple-testing correction. Subsequently, the sequences surrounding the demethylated CpG probes of germline L1 insertions (± 1 kb) were extracted and used as input for the transcription factor enrichment analysis. This analysis was performed using the Simple Enrichment Analysis algorithm from the MEME suite (v.5.5.4). The input motifs were selected from the *Homo sapiens* motif database (CIS-BP 2.00 Single Species DNA), which includes the transcription factor *ZNF695* motif.

Hypoxia score calculation

Hypoxia scores were calculated using a previously described method²⁴. Gene markers of hypoxia identified previously^{98–102} were synthesized into one gene list of 286 genes, and 28 genes that were filtered during the RNA-seq processing pipeline were removed. The median tumour expression for each gene was calculated, and each tumour was given gene scores of +1 or –1 based on whether gene expression was above or below the median for each gene. The total hypoxia score for each tumour was then calculated from the sum of its gene scores, where +258 was the most hypoxic score and –258 was the most normoxic.

Replication timing for signature ID2

The replication timing data for two lung cancer cell lines, A549 and NCI-H460, were obtained from Repli-Seq assays available in the ENCODE database (accession IDs: ENCFF454EFB for A549 and ENCFF098QZG for NCI-H460). For each chromosome, we calculated the median replication timing using the $\log_2$ -transformed ratio of early (E) to late (L) replicating regions. Moreover, we determined the number of ID2 deletions on each chromosome. To investigate the relationship between replication timing and ID2 mutational signature, we performed Pearson correlation analyses between $\log_2[E/L]$ and number of ID2 deletions across all chromosomes.

Targeted bisulfite-seq of L1HS promoter

Sodium bisulfite conversion of DNA samples. The DNA samples from tumour or normal tissues (29 tumour and normal pairs) were treated with sodium bisulfite to convert unmethylated cytosines to uracils. Moreover, 6 standard human DNA samples containing 100% methylated, 66% methylated (in triplicate), 33% methylated or 0% methylated human standard DNA from a commercial source (Human Methylated & Non-methylated DNA Set, Zymo Research) were treated along with the experimental samples to check the accuracy of the method in detecting the percentage of CpG DNA methylation.

Sodium bisulfite treatment was performed on 120 ng of each experimental or standard DNA sample using the EpiTect Fast DNA Bisulfite kit (Qiagen) according to the manufacturer's instructions. The reaction was conducted by adding 85 μ l of bisulfite solution and 35 μ l of DNA protection buffer to the DNA sample, in a total reaction volume of 140 μ l. The reaction was conducted on an Applied Biosystems 9700 thermal cycler (cycles: 95 °C for 5 min, 60 °C for 20 min, 95 °C for 5 min, 60 °C for 20 min and hold at 20 °C). After bisulfite treatment, the bisulfite-converted DNA was cleaned-up by MiniElute Spin columns (Qiagen) according to the manufacturer instructions and eluted in 15 μ l of buffer EB.

Targeted bisulfite PCR amplification. PCR was performed by preparing a 60 μ l reaction mix containing 6 μ l of purified bisulfite-converted DNA, 0.2 mM (final concentration) of each dNTP, 1.2 U (0.6 μ l) of Phusion U Hot Start DNA Polymerase (Thermo Fisher Scientific), 12 μ l of 5 $\times$ HF Phusion Buffer, 0.5 mM (final concentration) for each of the two primers (respectively, Chr22_Forw and L1_rev_upper_360). The total reaction volume of 60 μ l was divided into three vials of approximately 20 μ l and the PCR reaction was performed on an Applied Biosystems 9700 thermal cycler (cycles: 98 °C for 1 min 30 s; 42–44 cycles of 98 °C for 30 s, 59 °C for 20 s, 68 °C for 1 min 30 s; and a final extension at 68 °C for 7 min). The sequence and target specificity of the forward and reverse primers are listed in Supplementary Table 7.

The PCR-amplified samples were then purified using Magtivio MagSi-NGSPREP Plus magnetic beads (with bead ratio 1.8 $\times$ the input sample volume). After binding the DNA to the beads, the beads were washed twice with 70% ethanol (500 μ l). The purified DNA samples were eluted in 50 μ l of low-EDTA TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8). Purification was performed in 1.5 ml Eppendorf LoBind tubes. The size and concentration of the amplicons were checked by automated electrophoretic analysis on an Agilent TapeStation instrument, D1000 ScreenTape Assay.

Targeted bisulfite sequencing (NGS). Approximately 70 ng of PCR amplified and purified DNA sample was used for next-generation sequencing (NGS) library preparation using the Ion Plus Fragment Library Kit (Thermo Fisher Scientific). The manufacturer's instructions for the preparation of amplicon libraries were followed. In summary, the DNA sample was subjected to an end-prep reaction in a total volume of 50 μ l for 25 min at room temperature, then to a cleanup with 1.8 $\times$ MagSi beads (as reported above) with final elution in 12.5 μ l of IDTE. Finally, a ligation reaction was performed by adding 1 μ l of Ion Torrent P1 Adapter, 1 μ l of Ion Xpress Barcode (different for each sample) and 35.5 μ l of a ligation mix (containing 5 μ l of 10 $\times$ ligase buffer, 1 μ l dNTPs, 1 μ l ligase enzyme, 4 μ l nick-repair polymerase and 24.5 μ l H₂O) to the purified and pre-purified DNA sample (12.5 μ l). The reaction was performed in an Applied Biosystems 9700 thermal cycler (hold: 25 °C for 15 min, 72 °C for 5 min, 4 °C for up to 1 h). After adapter/index ligation, the library was purified from the unligated adapters by a double round of purification: the first round in MagSi beads 1 $\times$ (50 μ l of sample and 50 μ l of beads), two washes in 300 μ l 70% ethanol, elution in 50 μ l IDTE; the second purification in SPRI select beads 0.9 $\times$ (50 μ l of DNA sample + 45 μ l of beads), with three washes in 300 μ l 70% ethanol, elution in 20 μ l low-EDTA TE buffer.

After checking the size and concentration of each library on the D1000 HS Screen Tape Assay (Agilent), a pool of 455 pM of each library was prepared. The library pool was then treated with a semi-nested PCR assay of few cycles, to produce shorter amplicons and to increase that target specificity. The PCR reaction was performed by using a forward primer specific for the target sequence and fused with a Ion Torrent P1 adapter (trP1_Alu_Met_Forw_3), and a reverse primer specific for the external part of the indexed adapter (trA). The sequences of the forward and reverse primers are reported in Supplementary Table 7.

The reaction was prepared by mixing 76.9 μ l of Platinum SuperMix High Fidelity (Thermo Fisher Scientific), 0.5 mM forward primer and 0.5 mM reverse primer. The final volume of 100 μ l was divided into two tubes (50 μ l each) and the reaction was run in the Applied Biosystems 9700 thermal cycler (hold 95 °C for 5 min, 5 cycles of 95 °C for 15 s, 55 °C for 30 s, 70 °C for 1 min and hold at 4 °C).

The reamplified library pool was then purified with 1.5 $\times$ MagSi beads (2 washes with 70% ethanol, elution in 13 μ l low-EDTA TE). After checking the size and concentration of the reamplified library pool using the D1000 HS Screen Tape Assay (Agilent), it was diluted to 65 pM for loading onto the Ion Chef instrument. Template preparation and chip loading were performed using the reagents included in the Ion 510 & Ion 520 & Ion 530—Chef Kit and the Ion 510 Chip (Thermo Fisher

Scientific). The chip was loaded onto an Ion S5 instrument for NGS (Thermo Fisher Scientific).

After sequencing (850 run flows) and raw data processing, the sequences of each sample were mapped to the target region (bisulfite converted sequence) on the hg19 human genome sequence.

Validation of L1–ID1/ID2 associations. To further evaluate the association between L1 retrotransposition and the ID1/ID2 mutational signatures, we reanalysed publicly available WGS data from cancer cell lines reported previously³¹. This dataset includes WGS profiles of multiple daughter and granddaughter clones derived from various cancer cell lines, including LUAD, cultured under conditions that allow the accumulation of mutations in vitro. Aligned BAM files were downloaded from the EGA repository (EGAD00001004203), converted and lifted over to GRCh38 and processed into analysis-ready CRAM files. Mutation calling and filtering were performed using the same pipeline described above for the Sherlock-Lung WGS dataset. Indel mutational matrices were generated using SigProfilerMatrixGenerator, and ID signatures were extracted and quantified using SigProfilerExtractor.

Somatic L1 retrotransposition events were obtained from the supplementary materials of the original publication, which identified insertions using the same algorithm TraFiC-mem as our study, restricted to insertions acquired during in vitro expansion. To model the relationship between L1 insertions and mutational signature burden, we applied a generalized linear mixed model with Poisson regression (R package lme4), as described in the original study³¹. Two models were compared: a null model excluding L1 activity, and an alternative model including a log-transformed L1 insertion rate as a predictor (Supplementary Fig. 25). Both models adjusted for time in culture and included random effects for tumour type and experimental replicates. The Wald test was used to evaluate model fit. Separate association analyses were performed for ID1, ID2 and combined ID1/ID2 mutation counts.

We also analysed single-cell WGS data from ref. 32, which included 899 clones from normal colorectal epithelium and adenomatous polyps across 28 individuals. For our analysis, we focused on 406 clones derived from 19 donors with colorectal cancer or MUTYH-associated polyps. Somatic mutation and L1 insertion counts per clone were obtained from the original study's processed callsets (Supplementary Tables). The relationship between the number of L1 insertions and ID1/ID2 mutations per clone was assessed using Spearman correlation, including stratified analysis of clones from MUTYH-associated polyps. These analyses were used to test the reproducibility of the L1–ID1/ID2 association in an independent dataset from a different tissue context.

Statistical analyses

All statistical analyses were performed using R software (v.4.1.2) (<https://www.r-project.org/>). In general, the Mann–Whitney *U*-test (Wilcoxon rank-sum test) was used for two-group continuous variables, and the two-sided Fisher's exact tests were applied for the enrichment analysis of two categorical variables. $P < 0.05$ was considered to be statistically significant. If multiple testing was required, we applied FDR correction based on the Benjamini–Hochberg method. For the survival analyses, a proportional-hazards model was used to investigate the associations between the presence of ID2 mutational signature and overall survival, adjusting for age, sex and tumour stage.

To assess and account for cohort heterogeneity, we performed uniform manifold approximation and projection (UMAP) analysis using 3,491 genomic features, including mutation burden, mutational signatures, driver mutations, copy-number alterations, gene fusions and SVs. UMAP results showed that samples from different sources were intermixed, with no distinct clustering by site. By contrast, separation by smoking status was clearly observed, suggesting that biological—rather than technical—differences drive the main variation. Accordingly, we included study sites as a covariate in all relevant statistical models and

conducted stratified analyses by ancestry and smoking status to ensure robustness of key findings across subgroups.

Reporting summary

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

Data availability

Normal and tumour-paired CRAM files, as well as raw methylation intensity (IDAT) files for the WGS participants from the Sherlock-Lung study, have been deposited in dbGaP under accession numbers phs001697.v2.p1 and phs002992.v1.p1. RNA-seq FASTQ files for the same individuals are available through dbGaP under the accession number phs002346.v1.p1. The human reference genome (GRCh38) was obtained from the GATK resource repository (https://github.com/broadinstitute/gatk/blob/master/src/test/resources/large/Homo_sapiens_assembly38.fasta.gz). Publicly available LUAD multi-omics datasets can be accessed through dbGaP (accessions phs000178.v9.p8 and phs000488.v1.p1) and the European Genome-phenome Archive (EGA) (EGAS00001001757, EGAS00001002801 and EGAS00001003830). Detailed dataset descriptions and accession information are also provided in Supplementary Table 1.

Code availability

The WGS bioinformatics pipelines can be accessed at GitHub (<https://github.com/xtmgah/Sherlock-Lung>). Battenberg SCNA calling algorithm can be found at GitHub (<https://github.com/Wedge-lab/battenberg>). The Dirichlet process-based method for the subclonal reconstruction of tumours can be found at GitHub (<https://github.com/Wedge-lab/dpclus>). The bioinformatic pipeline for identifying TE insertion is available at GitLab (<https://gitlab.com/mobilegenom-esgroup/TraFiC>).

- Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
- Pedersen, B. S. et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
- Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
- Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239–2254 (2021).
- Sadedin, S. P. & Oshlack, A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. *Genome Biol.* **20**, 78 (2019).
- Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
- Yuan, K., Macintyre, G., Liu, W., PCAWG-11 Working Group & Markowitz, F. Ccube: a fast and robust method for estimating cancer cell fractions. Preprint at *bioRxiv* <https://doi.org/10.1101/484402> (2018).
- Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
- Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
- Muñoz, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *Nature* **596**, 428–432 (2021).
- Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **1**, 1–16 (2017).
- Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
- Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
- Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).
- Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.* **2**, 100179 (2022).
- Sondka, Z. et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
- Díaz-Gay, M. et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, btad756 (2023).
- Degasperi, A. et al. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* **376**, abt9283 (2022).
- Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

80. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
81. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
82. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945 (2022).
83. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.* **2**, lqaa078 (2020).
84. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
85. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
86. Li, T. et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* **48**, W509–W514 (2020).
87. de Tribolet-Hardy, J. et al. Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Res.* **33**, 1409–1423 (2023).
88. Del Toro, N. et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* **50**, D648–D653 (2022).
89. Oleksiewicz, U. et al. TRIM28 and interacting KRAB-ZNFs control self-renewal of human pluripotent stem cells through epigenetic repression of pro-differentiation genes. *Stem Cell Rep.* **9**, 2065–2080 (2017).
90. McKerrrow, W. & Fenyő, D. LIEM: a tool for accurate locus specific LINE-1 RNA quantification. *Bioinformatics* **36**, 1167–1173 (2020).
91. Reyes-Gopar, H. et al. A single-cell transposable element atlas of human cell identity. *Cell Rep. Methods* **5**, 101086 (2025).
92. Müller, F. et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* **20**, 55 (2019).
93. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
94. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
95. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
96. Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A. & Anderson, A. R. A. NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinform.* **20**, 264 (2019).
97. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 (2018).
98. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
99. Winter, S. C. et al. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67**, 3441–3449 (2007).
100. Ragnum, H. B. et al. The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. *Br. J. Cancer* **112**, 382–390 (2015).
101. Elvidge, G. P. et al. Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1α, HIF-2α, and other pathways. *J. Biol. Chem.* **281**, 15215–15226 (2006).
102. Sørensen, B. S., Toustrup, K., Horsman, M. R., Overgaard, J. & Alsner, J. Identifying pH independent hypoxia induced genes in human squamous cell carcinomas in vitro. *Acta Oncol.* **49**, 895–905 (2010).

Acknowledgements This work was supported by the Intramural Research Program of the National Cancer Institute, US National Institute of Health (NIH) (project ZIACP101231 to M.T.L.); by the NIH grants R01ES032547-01, R01CA269919-01 and 1U01CA290479-01 to L.B.A. as well as by L.B.A.'s Packard Fellowship for Science and Engineering. The research performed in L.B.A.'s laboratory was also supported by UC San Diego Sanford Stem Cell Institute. The funders had no roles in study design, data collection and analysis, decision to publish or preparation of the manuscript. The M.D.-G. fellowship, within the "Generación D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (CO05/24-ED CV1), is funded by the European Union NextGenerationEU funds, through PRTR. The computational

analyses reported in this manuscript have used the NIH high-performance Biowulf Cluster and the Triton Shared Computing Cluster at the San Diego Supercomputer Center of UC San Diego. We thank the study participants; P. Kraft for reading manuscript and comments; and the staff at Westat for their assistance in collecting samples and corresponding clinical data. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization. This research was supported by the Intramural Research Program of the NIH. The contributions of the NIH authors were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered works of the United States Government. However, the findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services. We want to acknowledge the patients and the INCLIVA Biobank (PT17/0015/0049) integrated in the Spanish National Biobanks Network and in the Valencian Biobanking Network for their collaboration. This study was supported by the Health and Medical Research Fund of Hong Kong SAR, HMRF 03142856. The related studies of Taiwan site were supported by grants from the Ministry of Health and Welfare, Taiwan DOH97-TD-G-111-026 (C.A.H.), DOH98-TD-G-111-015 (C.A.H.), DOH99-TD-G-111-028 (C.A.H.); DOH97-TD-G-111-029 (C.Y.C.), DOH98-TD-G-111-018 (C.Y.C.), DOH99-TD-G-111-015 (C.Y.C.), DOH97-TD-G-111-028 (I.S.C.), DOH98-TD-G-111-017 (I.S.C.), DOH99-TD-G-111-014 (I.S.C.), and the Ministry of Science and Technology, Taiwan MOST109-2740-B-400-002 (C.A.H.), MOST110-2740-B-400-002 (C.A.H.), MOST111-2740-B-400-002 (C.A.H.). 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). And, in part, by NIH (NCI) grant # U01CA209414 to the Boston Lung Cancer Survival Study of the Dana-Farber/ Harvard Cancer Center (D.C.C.). The authors would like to thank the team at the IUCPO site of the Quebec Respiratory Health Network Biobank of the FRQS for their valuable assistance, and would like to thank the staff at Harvard University, Yale University, Roswell Park Cancer Institute and Roswell PI, Instituto Nacional de Cancerología, Nice University Hospital Centre (Nice UHC) - University Côte d'Azur and the Nice Biobank CRB, Toronto University Health Network, and Mayo Clinic for their assistance providing samples and corresponding clinical data.

Author contributions T.Z. and M.T.L. conceived and designed the study. T.Z., L.Y., B.Z., J. Shi, L.B.A., D.C.W. and M.T.L. developed the methodology. T.Z., W.Z., C.W., M.D.-G., P.H.H., J.M.S., J.P.M., A. Klein, C.H., L.Y., B.Z., A. Khandekar and J. Sang carried out the formal analyses. J.Y., M.A., M. Cardelli, M. Cecati, F.M., K.B., J.C. and K.M.J. performed validation of findings. C.L., M.K.B., W.D.T., L.M.S., P.J., R.H. and S.-R.Y. conducted pathology review and analysis. M.P.W., K.C.L., C.A.H., C.-Y.C., N.E.C., A.C.P., D.C., E.S.E., J.M.S., M.B.S., S.S.Y., M. Manczuk, J.L., B.S., A.M., O.S., D.Z., I.H., D.M., S.M., M. Savic, Y.B., B.E.G.R., D.C.C., V.G., P.B., G.L., P.H., N.R., Q.L., M.T.L. and S.J.C. provided resources and essential materials. T.Z., J.R., M. Mirafteb, F.J.C.-M., M. Saha, P.H.H. and O.W.L. undertook data curation. The original manuscript draft was written by T.Z. and M.T.L. All of the authors, including L.Y., B.Z., J. Shi, J.C., T.Z., M.A.N., D.C.W., S.J.C., L.B.A. and M.T.L., contributed to the review and editing of the manuscript. T.Z. performed visualization. M.T.L. supervised the study. All of the authors read and approved the final version of the manuscript.

Competing interests L.B.A. is a co-founder, CSO, scientific advisory member and consultant for io9, has equity and receives income. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. L.B.A. is also a compensated member of the scientific advisory board of Inocras. L.B.A.'s spouse is an employee of Biotheranostics. L.B.A. declares US provisional applications filed with UCSD with serial numbers: 63/269,033, 63/366,392, 63/289,601, 63/483,237, 63/412,835 and 63/492,348. L.B.A. is also listed as an inventor of a US patent 10,776,718 for source identification by non-negative matrix factorization. S.-R.Y. has received consulting fees from AstraZeneca, Sanofi, Amgen, AbbVie and Sanofi; and received speaking fees from AstraZeneca, Medscape, PRIME Education and Medical Learning Institute. The other authors declare no competing interests.

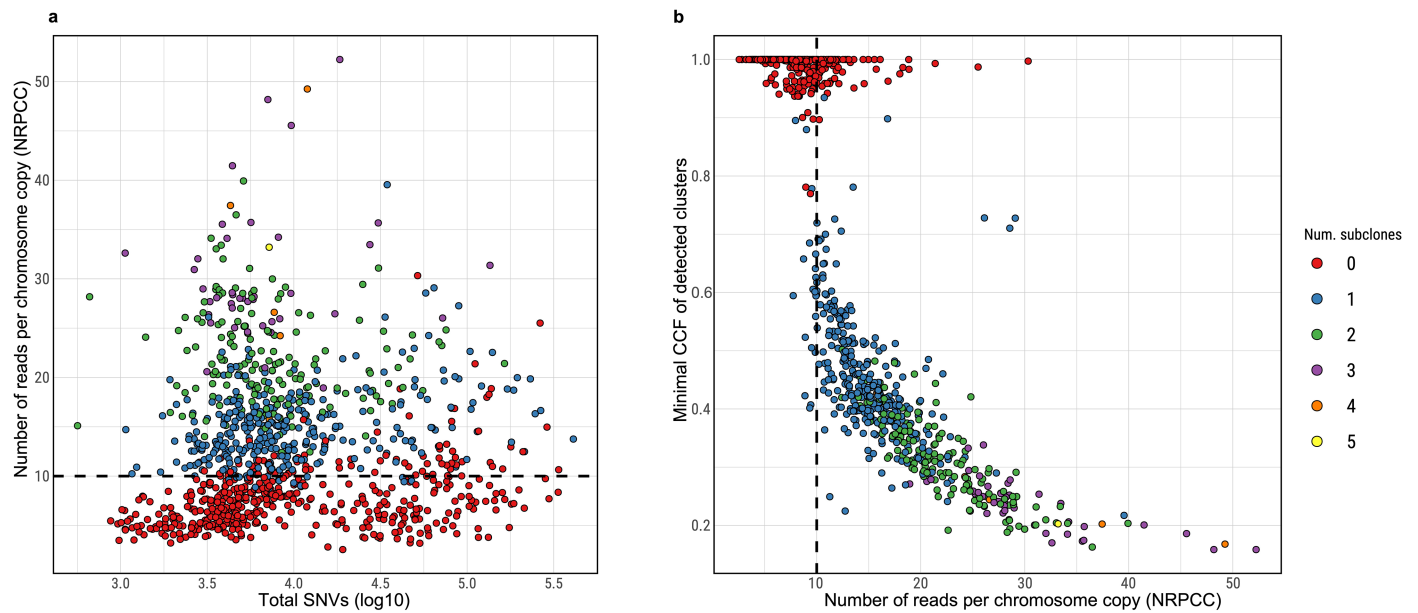
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09825-y>.

Correspondence and requests for materials should be addressed to Maria Teresa Landi.

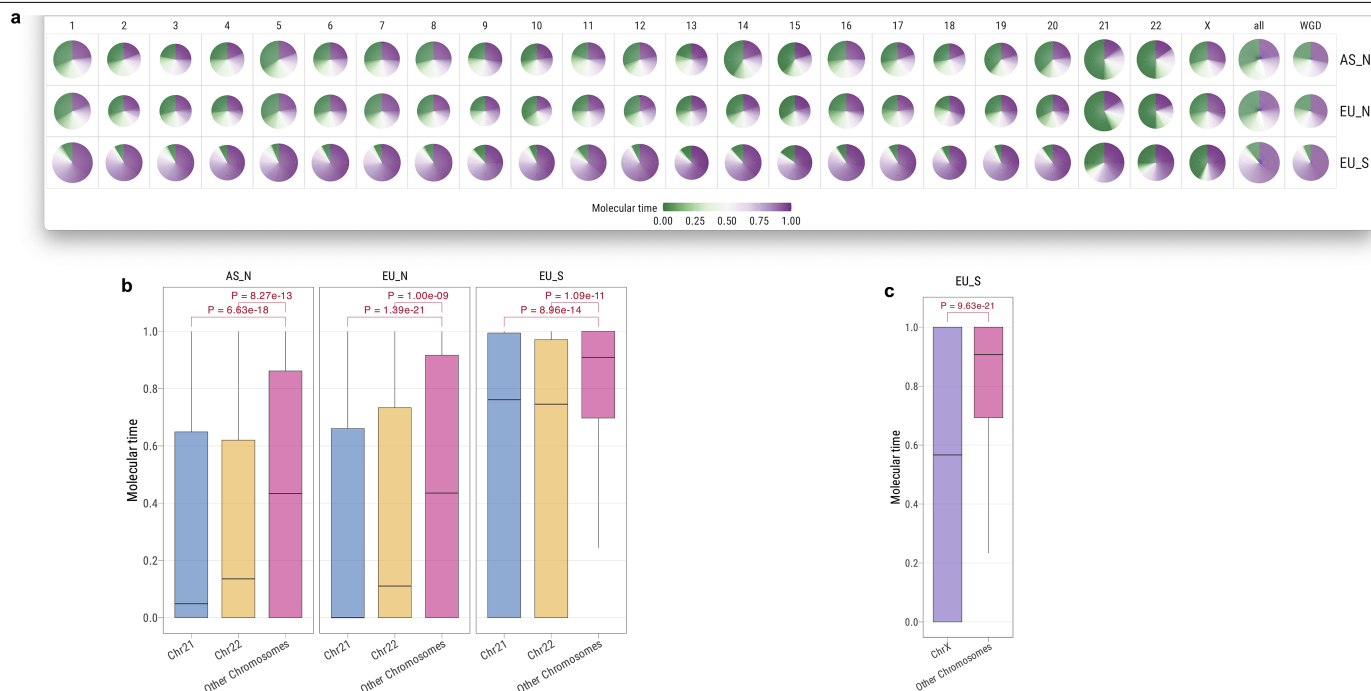
Peer review information Nature thanks David Adams and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



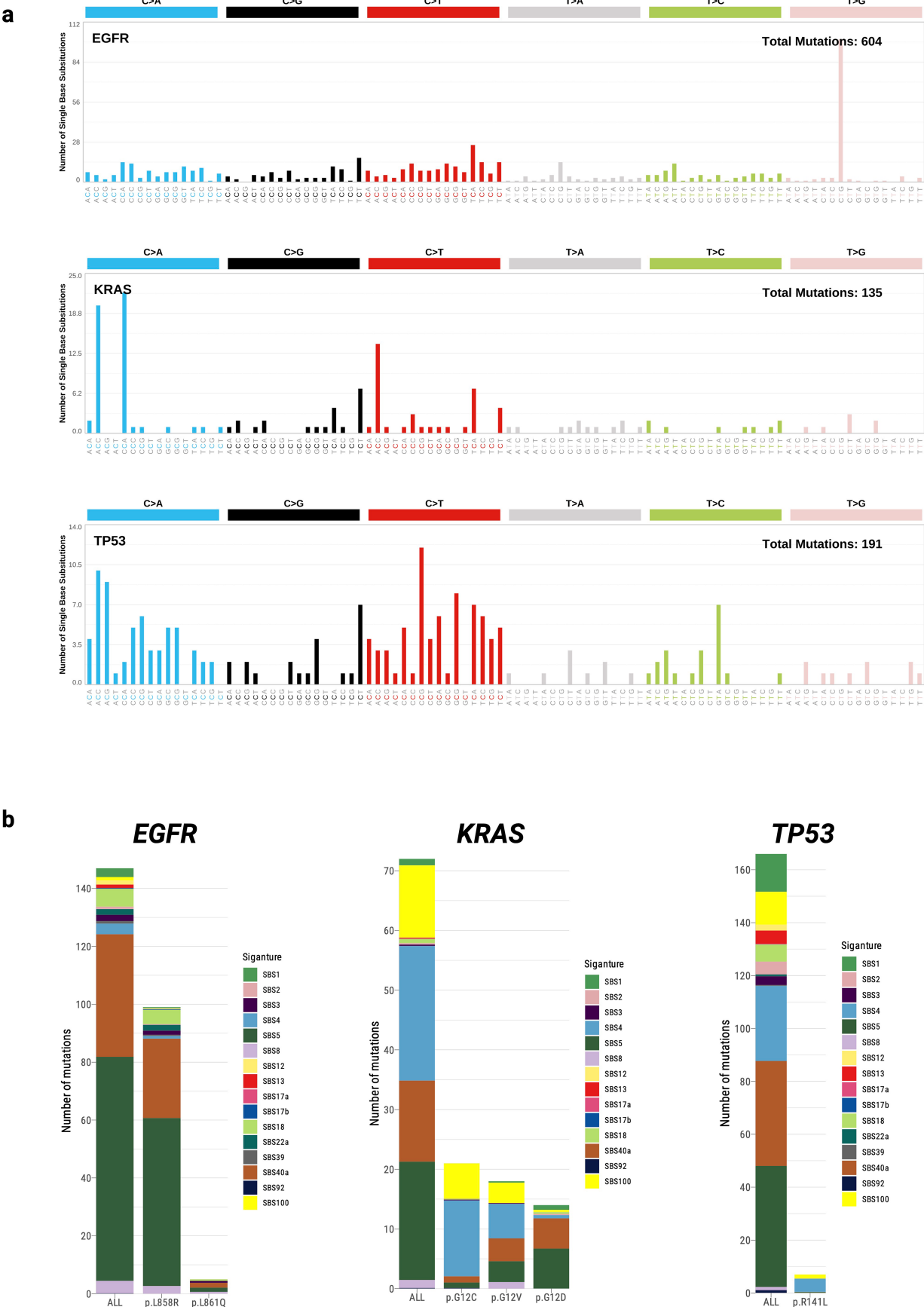
Extended Data Fig. 1 | Power analysis for detecting a diverse clonal architecture in 1024 LUAD. a, The scatter plot illustrates the relationship between the number of reads per chromosome copy (NRPCC) and the total detected single nucleotide variants (SNVs). Our ability to detect subclones relies not on the number of identified SNVs, but on the number of reads per tumour chromosomal copy. NRPCC accounts for tumour purity, ploidy, and

sequencing coverage. **b,** The minimum cancer cell fraction (CCF) of the detected clusters in each tumour is plotted against NRPCC. To mitigate biases, we exclusively considered tumours with $\text{NRPCC} \geq 10$. In these tumours, our analysis is sufficiently powered to identify a subclone with a CCF $\geq 30\%$. The suggested NRPCC threshold is denoted by the dashed line.



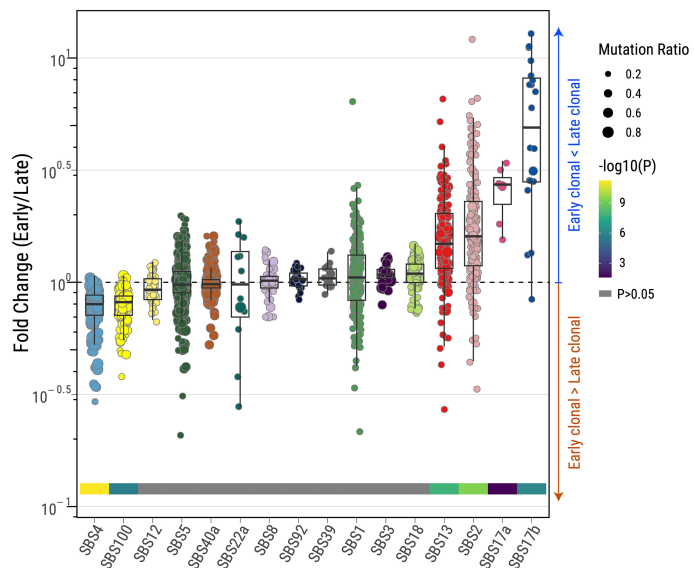
Extended Data Fig. 2 | Overview of the molecular timing distribution of copy number gains per chromosome. a, Pie charts depict the distribution of the inferred mutation time for a given copy number gain in a group. Green denotes early clonal gains, with a gradient to purple for late gains. The size of each pie chart is proportional to the frequency of recurrence of each event. **b,** Comparison of the molecular timing of copy number gains on chromosomes 21 and 22 versus other chromosomes across all three groups. **c,** Comparison of

the molecular timing of copy number gains on chromosome X versus other chromosomes in tumours from the EU_S group. Two-sided Wilcoxon rank-sum test p-values are shown above the box plots. All box plots show the median (centre line), IQR (box), and whiskers extending to 1.5× the IQR by default in ggplot2. The P-values derived from the two-sided Wilcoxon rank-sum test are shown above the box plots.

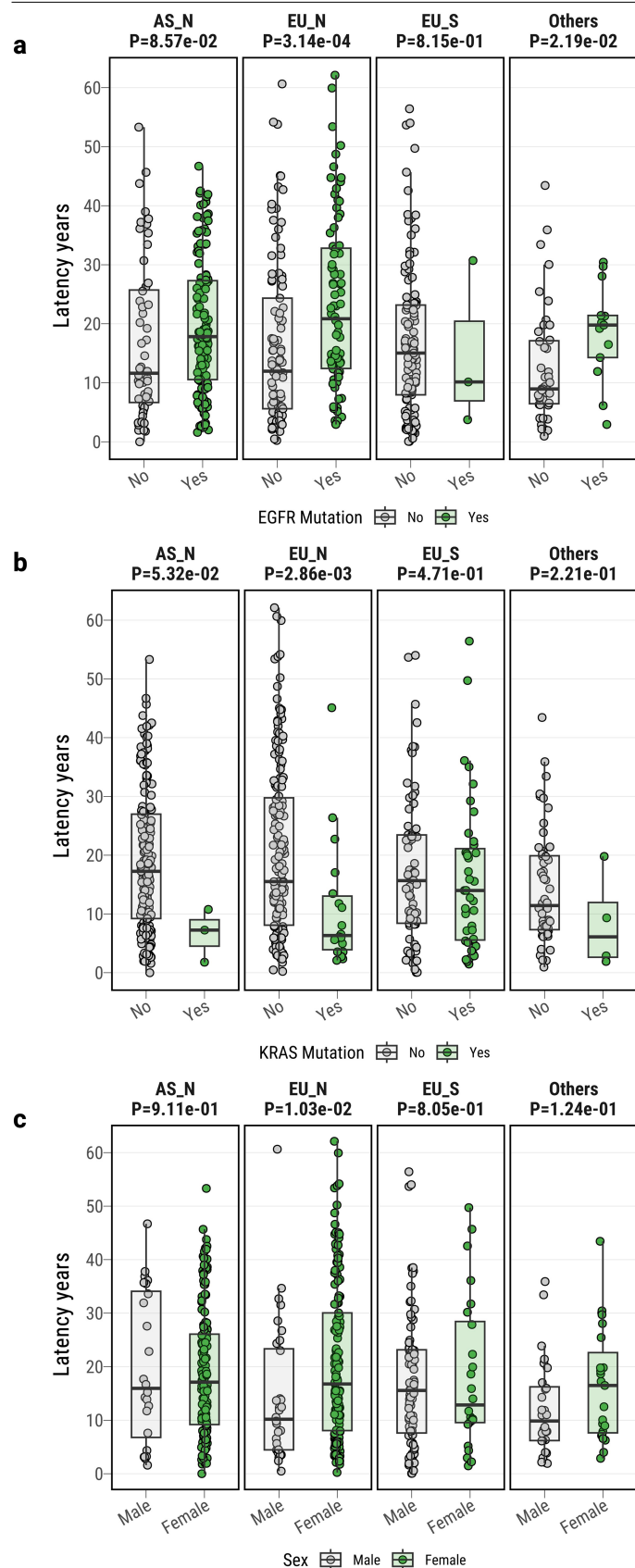


Extended Data Fig. 3 | Contribution of mutational signatures to major driver genes. a, Single base substitution (SBS) mutational patterns for the driver mutations of EGFR, KRAS, and TP53. **b,** Distribution of all driver mutations

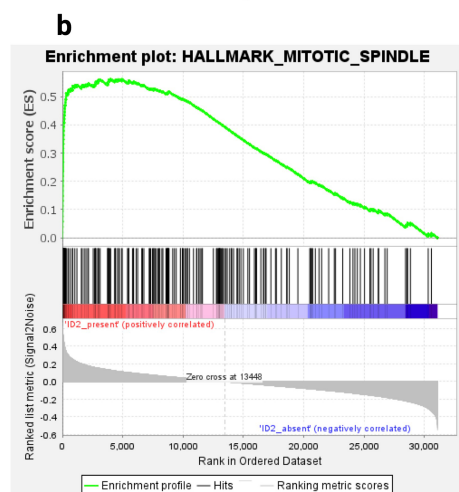
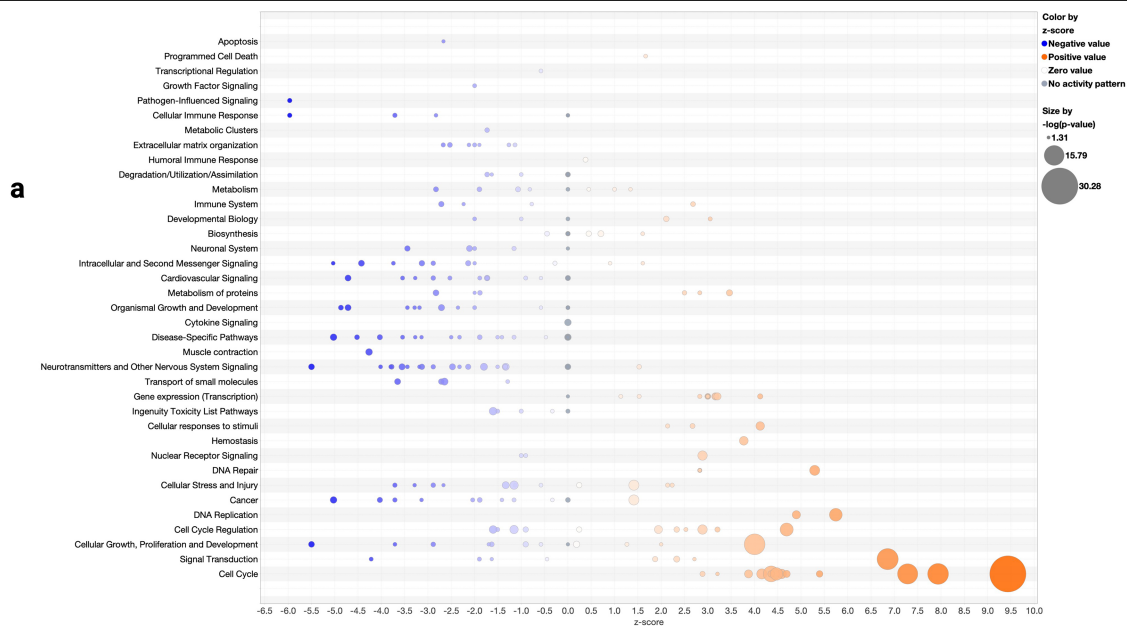
and hotspot mutations (recurrence > 5) in EGFR, KRAS, and TP53 attributed to each mutational signature.



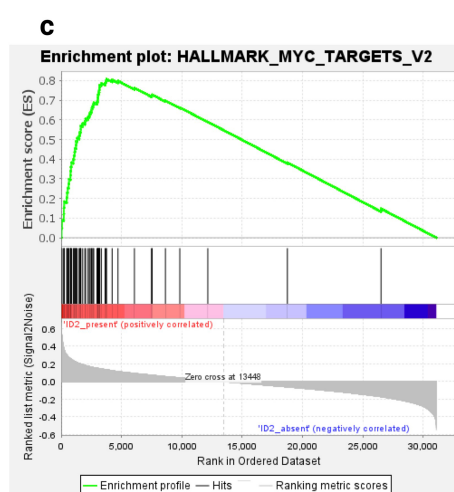
Extended Data Fig. 4 | Dynamic SBS mutational processes during early and late tumour evolution. Fold changes between the relative proportions of early and late clonal mutations attributed to individual SBS mutational signatures. Box plots display the median (centre line), interquartile range (box), and whiskers extending to $1.5 \times$ the IQR. Each point represents a tumour sample ($n = 542$) and is coloured by mutational signature. P values from two-sided Wilcoxon rank-sum tests are shown below the box plots.



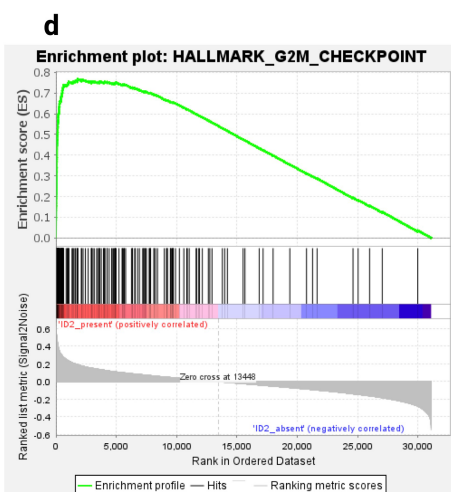
Extended Data Fig. 5 | Associations between tumour latency and *EGFR* mutation status (a), *KRAS* mutation status (b) and sex (c), stratified by ancestry and tobacco smoking group. All box plots display the median (centre line), interquartile range (box), and whiskers extending to 1.5× the IQR. Sample size: n = 542. P values were calculated using two-sided Wilcoxon rank-sum tests.



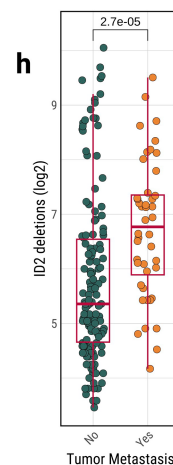
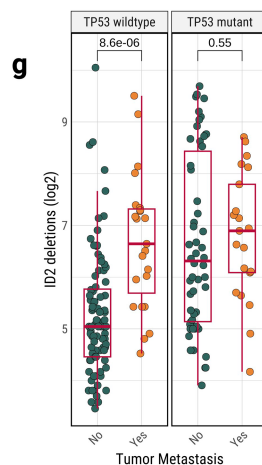
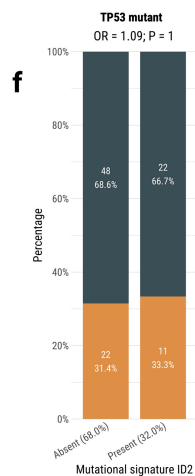
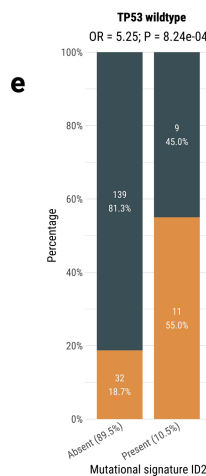
Normalized Enrichment Score (NES) = 2.11
 FDR q-value = 0.02



Normalized Enrichment Score (NES) = 2.08
 FDR q-value = 0.023



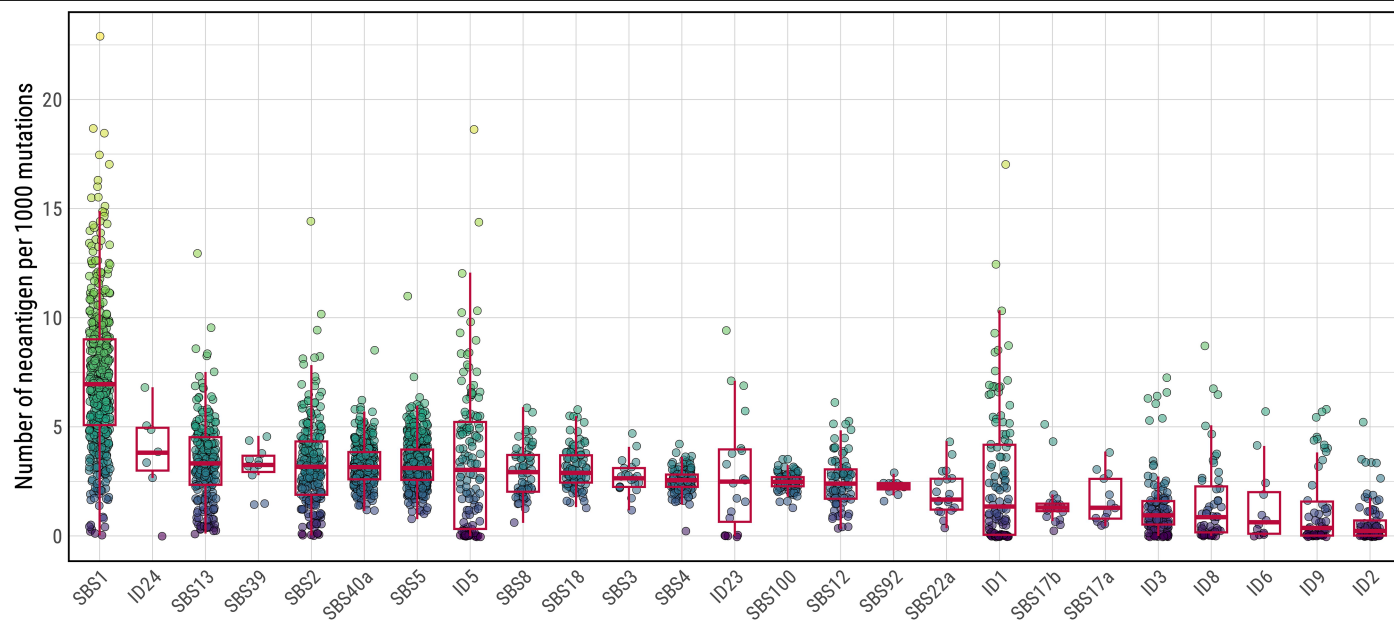
Normalized Enrichment Score (NES) = 1.99
 FDR q-value = 0.028



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

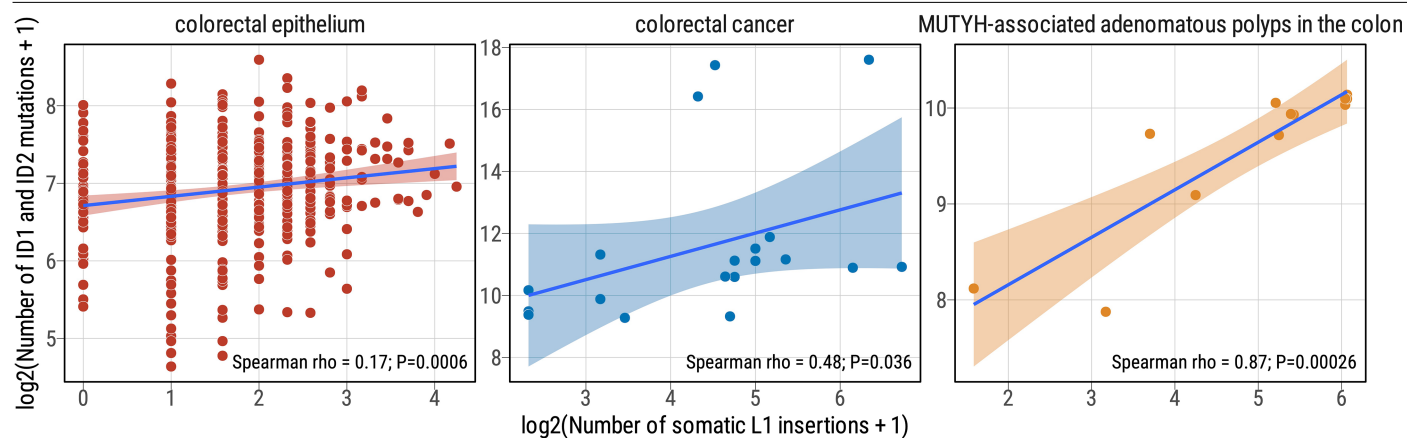
Extended Data Fig. 6 | Association between tumours with mutational signature ID2 and increased tumour proliferation. **a**, Ingenuity pathway analysis (IPA) of differentially expressed genes in tumours with and without mutational signature ID2. The x-axis represents the Z-score, and the y-axis indicates the corresponding canonical pathways. Suppressed pathways are shown in blue, while activated pathways are shown in orange. Circle size reflects the level of significance. **b–d**, Gene set enrichment analysis (GSEA) highlights key proliferation-related pathways significantly associated with tumours harbouring mutational signature ID2, including hallmark mitotic spindle (**b**), hallmark MYC targets V2 (**c**), and hallmark G2M checkpoint (**d**). The normalized

enrichment score (NES) and FDR-adjusted q-value are displayed at the bottom of each GSEA plot. **e,f**, Enrichment of tumour metastasis in tumours with mutational signature ID2, stratified by *TP53* mutation status. Odds ratios and p-values from the Fisher’s exact test are shown above the plot. **g,h**, Comparison of ID2 deletion burden between tumours with and without metastasis, stratified by *TP53* mutation status (**g**) or across all tumours (n = 294) (**h**). All box plots show the median (centre line), IQR (box), and whiskers extending to 1.5× the IQR by default in ggplot2. The *P*-values derived from the two-sided Wilcoxon rank-sum test are shown above the box plots.



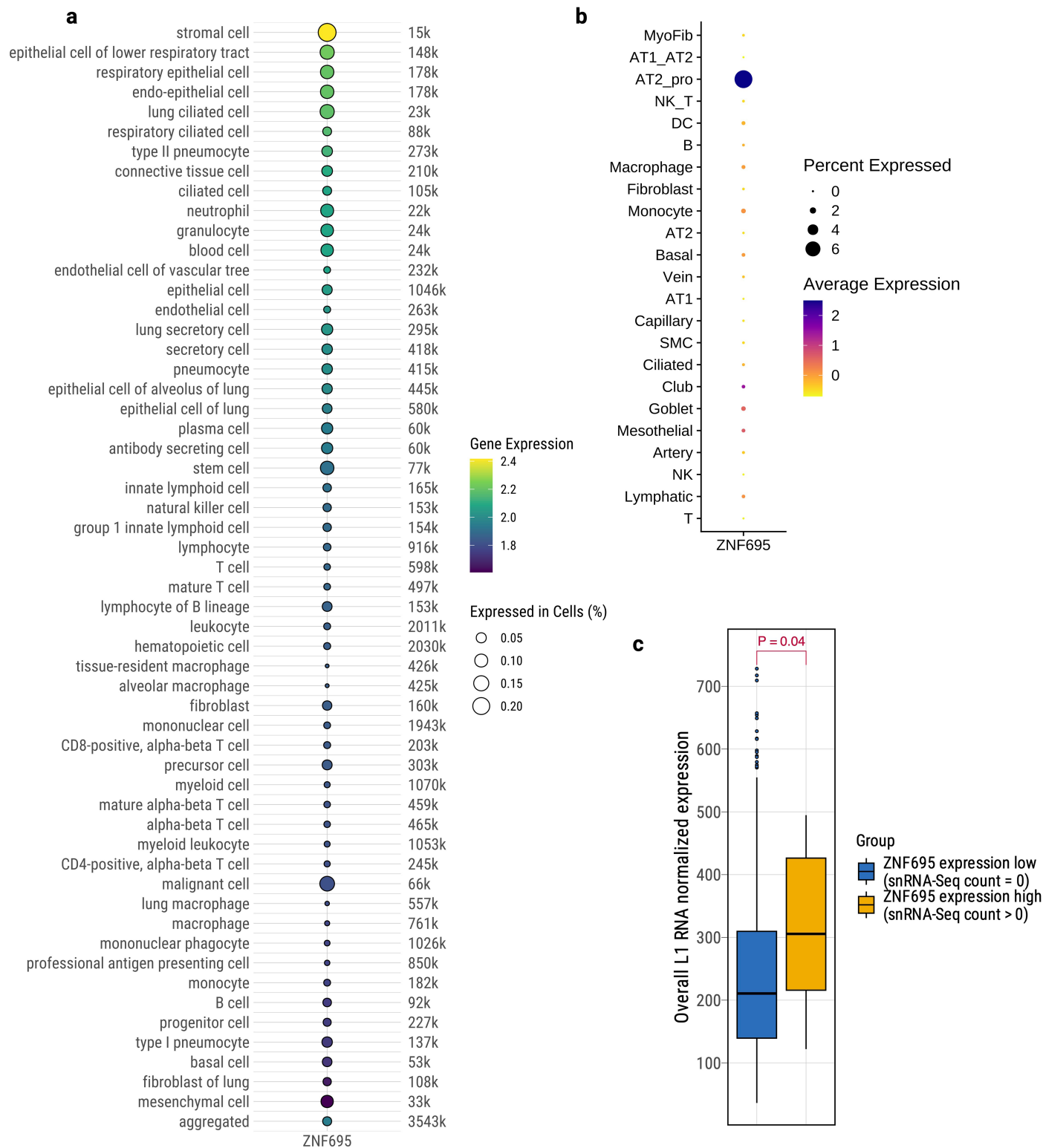
Extended Data Fig. 7 | Box plot illustrating neoantigen predictions across all mutational signatures. Neoantigen burden is normalized as the number of neoantigens per 1000 mutations. All box plots show the median (centre line),

IQR (box), and whiskers extending to 1.5× the IQR by default in ggplot2. Each data point is shown as an individual sample.



Extended Data Fig. 8 | Validation of the association between L1 retrotransposition and mutational signatures ID2 using WGS data from normal and matched colorectal cancer clones³². Scatter plots show Pearson correlations between the number of somatic L1 insertions and the number of

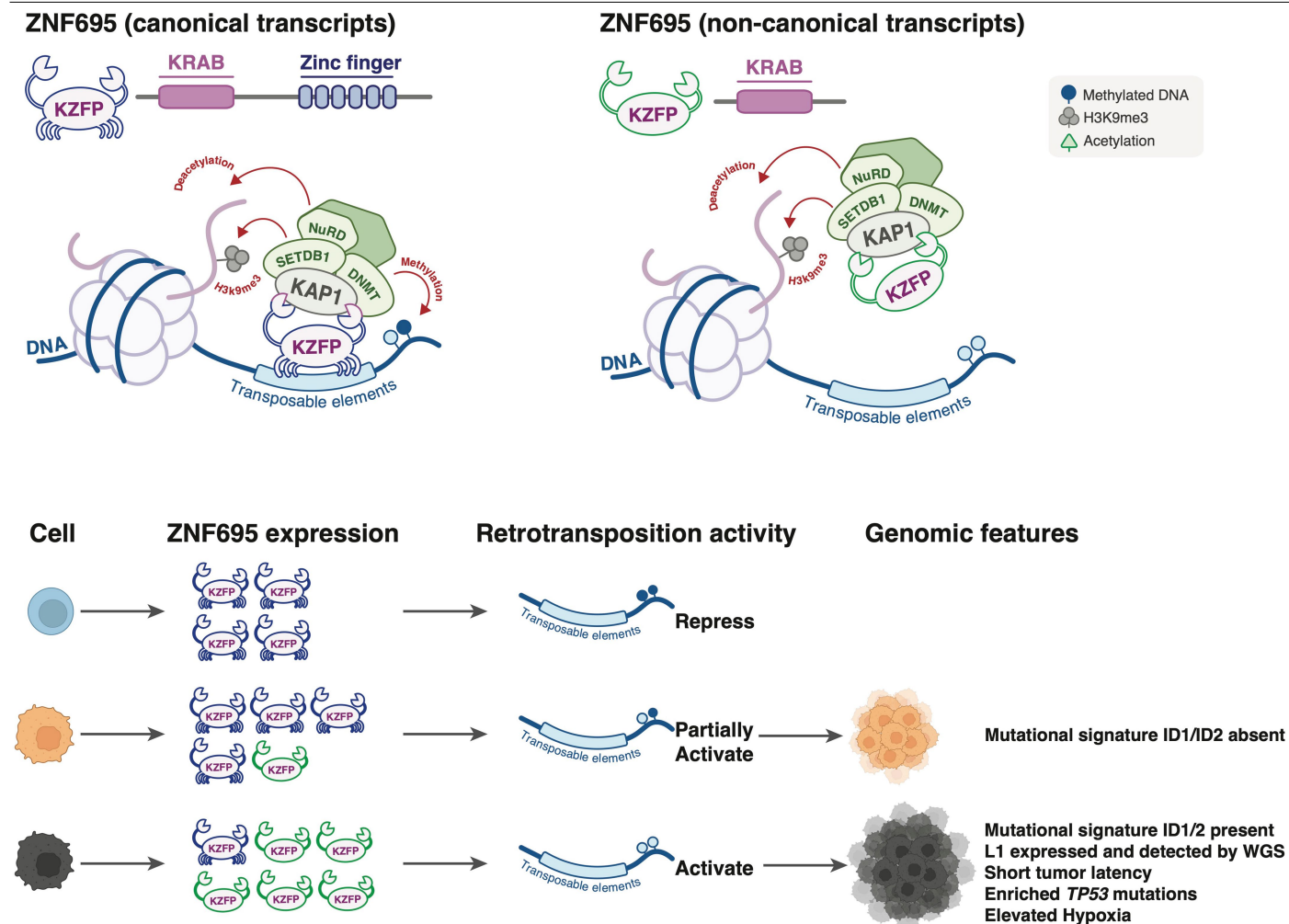
ID1/ID2 mutations in normal colorectal epithelium (red), matched colorectal cancer (blue), and MUTYH-associated adenomatous polyps (orange). Spearman correlation coefficients (R) and p-values (P) are indicated in the plots. The fitted regression line (blue) and 95% confidence intervals (grey) are shown.



Extended Data Fig. 9 | ZNF695 expression in single cell RNA-seq studies.

a, Identification of lung cell types with detectable *ZNF695* expression across multiple single cell types from RNA-seq data in the CZ CELLxGENE database. The numbers on the right y-axis represent the number of cells identified from the single-cell studies. **b**, Enrichment of *ZNF695* expression in alveolar type 2 proliferating (AT2pro) cells from a normal lung single-nucleus multiome dataset.

Dot plot visualizes the normalized RNA expression of *ZNF695* by cell type. The colour and size of each dot correspond to the scaled average expression level and fraction of expressing cells, respectively. **c**, Enrichment of L1 RNA expression in cells with detectable *ZNF695* expression. P-values from the Wilcoxon rank-sum test are shown on the top of the boxplot.



Extended Data Fig. 10 | Schematic model of canonical and non-canonical ZNF695 function in L1 regulation. This schematic, adapted from ref. 49, illustrates ZNF695-mediated regulation of L1 retrotransposons. In normal cells, canonical KZFP transcripts encode proteins with an N-terminal KRAB domain and a C-terminal zinc finger array. The zinc finger domain binds transposable elements (TEs), while the KRAB domain recruits a heterochromatin-inducing complex—including TRIM28 (KAP1), SETDB1, the NuRD complex, HP1, and DNMTs. This complex mediates H3K9 trimethylation and directs DNMTs to methylate CpG sites in L1 promoter regions, repressing L1 retrotransposition.

In tumour cells, however, non-canonical *ZNF695* transcripts, lacking the zinc finger domain, predominate. This impairs TE binding and DNMT recruitment, leading to L1 promoter hypomethylation and subsequent L1 activation. Between the first and second strand synthesis by activated ORF2p (ref. 38), intermediates may be cleaved and integrated into the genome. Coupled with DNA repair mechanisms, this process may generate single-base-pair indels (e.g., ID1/ID2). This suggests a dominant-negative effect, where the balance between canonical and non-canonical transcripts governs L1 methylation status.