Genomic–transcriptomic evolution in lung cancer and metastasis

https://doi.org/10.1038/s41586-023-05706-4

Received: 21 October 2021

Accepted: 4 January 2023

Published online: 12 April 2023

Open access

Check for updates

Carlos Martínez-Ruiz[1,2,73], James R. M. Black[1,2,73], Clare Puttick[1,2,3,73], Mark S. Hill[3,73], Jonas Demeulemeester[4,5,6], Elizabeth Larose Cadieux[4,7], Kerstin Thol[1,2], Thomas P. Jones[1,2], Selvaraju Veeriah[1], Cristina Naceur-Lombardelli[1], Antonia Toncheva[1], Paulina Prymas[1], Andrew Rowan[3], Sophia Ward[1,3,8], Laura Cubitt[8], Foteini Athanasopoulou[1,3,8], Oriol Pich[3], Takahiro Karasaki[1,3,9], David A. Moore[1,3,10], Roberto Salgado[11,12], Emma Colliver[3], Carla Castignani[4,7], Michelle Dietzen[1,2,3], Ariana Huebner[1,2,3], Maise Al Bakir[1,3], Miljana Tanić[7,13], Thomas B. K. Watkins[3], Emilia L. Lim[1,3], Ali M. Al-Rashed[14], Danny Lang[15], James Clements[15], Daniel E. Cook[3], Rachel Rosenthal[3], Gareth A. Wilson[3], Alexander M. Frankell[1,3], Sophie de Carné Trécesson[16], Philip East[17], Nnennaya Kanu[1], Kevin Litchfield[1,18], Nicolai J. Birkbak[1,3,19,20,21], Allan Hackshaw[22], Stephan Beck[7], Peter Van Loo[4,23,24], Mariam Jamal-Hanjani[1,9,25], TRACERx Consortium*, Charles Swanton[1,3,25] ✉ & Nicholas McGranahan[1,2] ✉

Intratumour heterogeneity (ITH) fuels lung cancer evolution, which leads to immune evasion and resistance to therapy[1]. Here, using paired whole-exome and RNA sequencing data, we investigate intratumour transcriptomic diversity in 354 non-small cell lung cancer tumours from 347 out of the first 421 patients prospectively recruited into the TRACERx study[2,3]. Analyses of 947 tumour regions, representing both primary and metastatic disease, alongside 96 tumour-adjacent normal tissue samples implicate the transcriptome as a major source of phenotypic variation. Gene expression levels and ITH relate to patterns of positive and negative selection during tumour evolution. We observe frequent copy number-independent allele-specific expression that is linked to epigenomic dysfunction. Allele-specific expression can also result in genomic–transcriptomic parallel evolution, which converges on cancer gene disruption. We extract signatures of RNA single-base substitutions and link their aetiology to the activity of the RNA-editing enzymes ADAR and APOBEC3A, thereby revealing otherwise undetected ongoing APOBEC activity in tumours. Characterizing the transcriptomes of primary–metastatic tumour pairs, we combine multiple machine-learning approaches that leverage genomic and transcriptomic variables to link metastasis-seeding potential to the evolutionary context of mutations and increased proliferation within primary tumour regions. These results highlight the interplay between the genome and transcriptome in influencing ITH, lung cancer evolution and metastasis.

An understanding of the causes of cancer cell-to-cell variation is essential to understand tumour evolution. Recent work has emphasized that much of this variation is transcriptomic, arising from diverse mechanisms that relate to, or are independent of, genomic variation[4]. In mouse models of non-small cell lung cancer (NSCLC), transcriptomic plasticity has been shown to underpin ITH[5]. While genomic variation reflects the relics of past somatic events acquired during the evolutionary history of a tumour, transcriptomic variation may provide an accurate approximation of the phenotypic state of a tumour at the time of sampling[1]. To date, most studies of tumour evolution in humans have focused on the impact of genomic alterations on cancer. Transcriptomic studies that leverage bulk tumour RNA sequencing (RNA-seq) data tend to focus on the amplitude of gene expression in a single biopsy taken at a single time point. This approach might fail to capture poorly understood transcriptomic processes, including allele-specific expression (ASE) and RNA editing that can exert important effects on cancer evolution[1,4].

Here we leverage multiregion sequencing data from patients recruited into the TRACERx study[2] to better understand the impact of multiple transcriptomic features and their interplay with genomic and phenotypic diversity in NSCLC evolution at different spatial and temporal scales.

Cohort overview

We analysed matched RNA-seq and whole-exome sequencing data from 347 patients recruited into the prospective study TRACERx (TRACERx 421 cohort). Samples from the cohort comprised 947 tumour regions from 354 NSCLC tumours (6 patients harboured multiple

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

primaries at diagnosis), as well as 96 tumour-adjacent normal lung tissue regions (see the consolidated standards of reporting trials (CONSORT) diagram in Supplementary Information)[6,7]. Of these patients, 344 had 886 primary tumour regions, 21 also had 29 metastatic lymph node (LN) regions sampled at surgical resection of the primary tumour and 24 patients had 30 metastatic tumour regions sampled at relapse or progression. In total, 168 primary tumour regions and 4 LN regions from 64 patients in this cohort were previously described in the TRACERx 100 cohort[8]. The cohort of paired primary–metastatic regions analysed here (and reported in a companion paper[6]) comprises 61 metastatic regions including LN regions and intrapulmonary metastases resected at surgery (henceforth termed primary LN/satellite lesions) and LN and metastatic regions at recurrence or progression.

Expression diversity in NSCLC evolution

We first examined patterns of gene expression across tumour samples. A uniform manifold approximation and projection (UMAP) analysis (Extended Data Fig. 1a) based on gene expression across the cohort revealed that samples clustered in three main groups dominated by lung adenocarcinomas (LUADs), lung squamous cell carcinomas (LUSCs) and tumour-adjacent normal lung tissue. Notably, 27 out of 184 non-LUAD tumours, defined by central pathological review, clustered with LUADs. These tumours, which included four LUSCs, were 23 times more likely to harbour a LUAD-specific driver mutation (Methods) than other non-LUADs ($P = 2.7 \times 10^{-11}$, Fisher's exact test; Extended Data Fig. 1b). Although not classified as LUADs, 67% of these tumours (18 out of 27) were positive for common LUAD immunohistochemical staining markers such as TTF-1 or exhibited LUAD morphology (Extended Data Table 1). This enrichment for LUAD driver mutations among non-LUAD NSCLC tumours that cluster with LUADs suggests that phenotypically, this subset of tumours may be similar to LUADs. This result is also consistent with some such tumours harbouring an adenocarcinomatous component[9] and with other reports of LUAD drivers in non-LUAD tumours[10].

Next, to establish determinants of intertumour and intratumour transcriptomic diversity, we performed independent principal component analyses (PCAs) within the two major NSCLC histologies (LUAD and LUSC) and related these to 39 underlying genomic and clinico-pathological variables (Fig. 1a; see Methods for the rationale of feature selection). Principal components (PCs) were more frequently significantly correlated with genomic variables in LUAD than in LUSC. This trend persisted when LUADs were downsampled to account for differences in the sample size (Extended Data Fig. 1c). PCs exhibited lower relative ITH in LUADs compared to LUSCs; that is, the ratio of intratumour to intertumour heterogeneity of the PC amplitude was lower within LUADs (Fig. 1a). Taken together, these results are suggestive of more deterministic genomic–transcriptomic relationships within LUADs than LUSCs. Furthermore, LUAD PC activity correlated with orthogonal signatures that quantify RAS pathway activation[11], which highlights that PCs might represent transcriptional programmes that are preserved across datasets (Extended Data Fig. 1d).

In LUADs, this analysis further revealed two relationships consistent with mutual exclusivity, with separate features showing significant and opposing correlations with a given PC. First, PC5 was positively associated with predicted driver mutations in *KRAS* and invasive mucinous adenocarcinomas (IMAs). IMAs were enriched in tumours harbouring non-G12C *KRAS* predicted driver mutations ($P = 0.003$, χ^2 test; Extended Data Fig. 1e), which were less likely to be associated with a history of smoking[12]. This result provides transcriptomic context to previous work suggesting that IMAs are more common in never-smokers[13]. Second, PC1 was strongly negatively correlated with MSigDB Hallmark gene sets related to proliferation[14] (Extended Data Fig. 1f), yet positively associated with activating mutations in *EGFR* (linear mixed-effect, model false discovery rate (FDR) = 0.0008). In keeping with this, *EGFR* driver mutations were associated with low Ki-67 levels ($P = 0.028$, Wilcoxon test; Extended Data Fig. 1g). This finding suggests that the phenotype of *EGFR* mutant LUADs is one of reduced proliferation compared with *EGFR* wild-type LUADs.

To further assess transcriptomic ITH independently from the number of tumour regions sampled, we developed the intratumour expression distance (I-TED) metric, which is calculated as the mean normalized gene expression correlation distance for a given region paired with every other region from the same tumour (Methods and Fig. 1b). A high I-TED value reflects high expression ITH. Hierarchical clustering of all samples based on the gene expression correlation distance revealed that tumour regions from a given patient tended to cluster together (in 231 out of 280 multiregion primary tumours, all regions within a given tumour clustered together). Within the 49 tumours for which constituent regions did not all cluster, those regions clustering apart harboured increased weighted genome instability index scores ($P = 0.002$, linear regression, 104 regions). Consistently, the fraction of the genome affected by subclonal somatic copy number alterations (SCNAs) and intratumour variation in purity were independently associated with increased I-TED values (Fig. 1c; 13.3% and 2.8% of variance explained, respectively). Conversely, I-TED was not associated with the heterogeneity of subclonal mutations nor the number of regions sampled per tumour. This result underlines the link between SCNAs and changes in gene expression.

To further evaluate the relationship between tumour purity and transcriptomic heterogeneity, we estimated the tumour transcript fraction (a ploidy-adjusted estimate of the proportion of all transcripts that were derived from the tumour) from RNA-seq reads (Methods). We observed that the tumour transcript fraction was consistently greater than the tumour purity (Fig. 1d). This result suggests that per chromosome copy, gene expression from tumour cells tends to exceed that of non-tumour cells within a bulk sample, which is in keeping with results from another study[15]. Of note, the tumour transcript fraction was a better predictor of I-TED than purity, which highlights that DNA-derived estimates of tumour diversity may not always be representative of phenotypic diversity ($P = 5.03 \times 10^{-8}$, linear regression; Extended Data Fig. 1h).

Next, we sought to understand whether patterns of gene expression and their heterogeneity are related to selection during tumour evolution. We measured selection within established lung cancer and non-cancer genes using the ratio between the observed number of nonsynonymous mutations per nonsynonymous site and the number synonymous mutations per synonymous site (dN/dS), calculated through the dNdScv method[16]. Genes were grouped into tertiles according to the average amplitude of their expression across the cohort (Fig. 1e). Within cancer genes, significant positive selection (implied when dN/dS with ±95% confidence intervals is >1) was most readily observed within truncating mutations in genes in the highest expression tertile. Notably, within non-cancer genes, signals of negative selection (dN/dS ± 95% confidence intervals of <1) were identified within truncating mutations in genes within the highest expression tertile only (242 truncating mutations, relative to 3,932 observed truncating mutations, were estimated to have been lost through negative selection in these genes). Similar patterns were observed when dividing the data by different expression quantiles (Extended Data Fig. 1i).

Expanding on this analysis, we next explored the relationship between the ITH of gene expression (measured as the standard deviation of normalized gene expression among all regions within a tumour) and selection in tumour evolution[17]. Cancer genes within the lowest tertile of expression ITH exhibited the strongest signals of positive selection. By contrast, within non-cancer genes of the same tertile, negative selection was identified (188 truncating mutations, relative to 3,083 observed truncating mutations, were estimated to have been lost through negative selection; Fig. 1e). Furthermore, the lowest ITH quantile and highest expression quantile were significantly enriched for NSCLC essential genes as identified in the Project Achilles study[18]

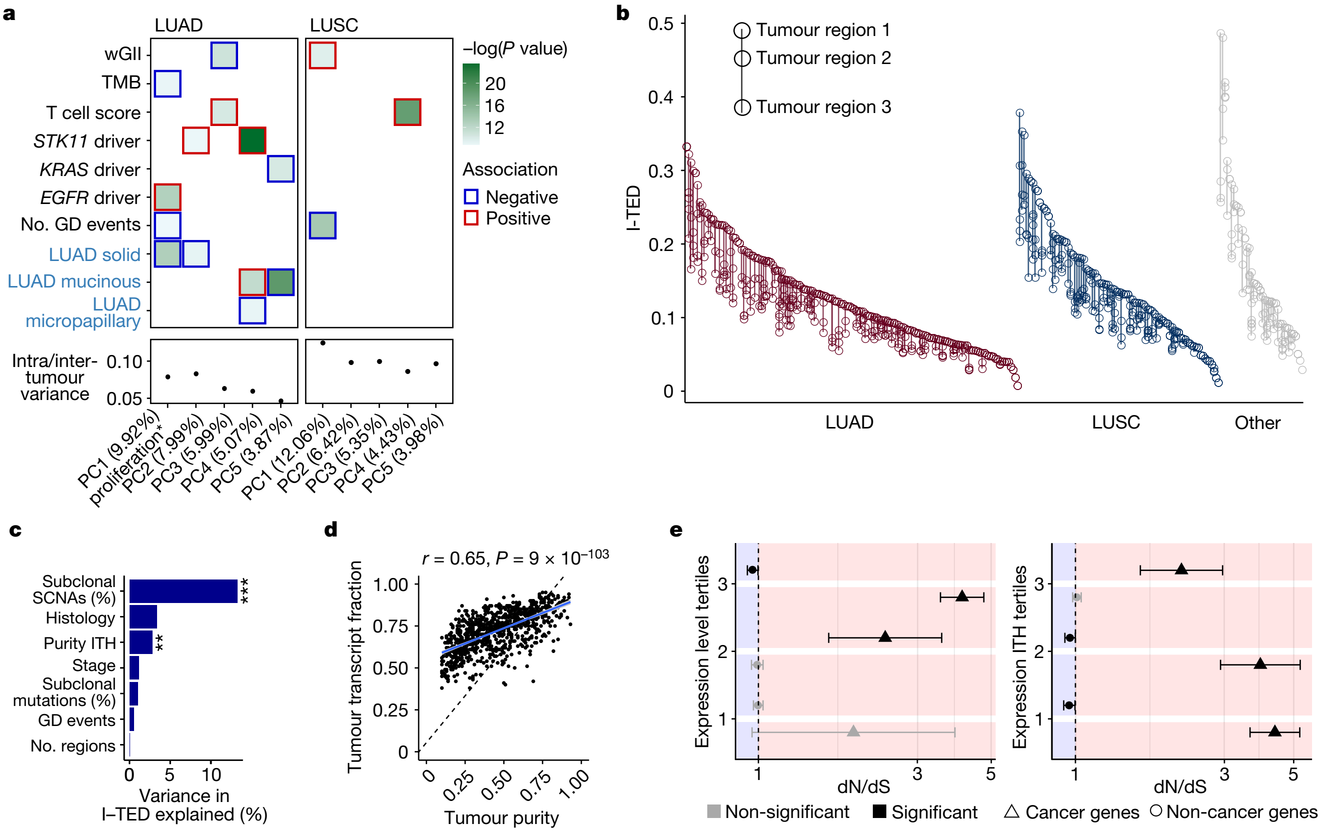


Fig. 1 | Expression diversity in the TRACERx 421 cohort. a, Relationship between PCs of transcriptomic diversity and genomic (black labels) and clinical (blue labels) variables. Displayed are the top PCs within LUADs (n = 480 regions from 190 tumours) and LUSCs (n = 303 regions from 119 tumours) that together explain at least 30% of the total variance, alongside their median ratio of heterogeneity (intratumour heterogeneity of PC activity divided by intertumour heterogeneity of PC activity). The colour of the border around each square indicates the direction of the association between each covariate and PC. In total, 39 variables were tested (Methods). Significance was determined using a mixed-effects linear model with purity as a fixed covariate and tumour as a random variable. Only features significant (P < 0.05) after FDR correction with at least one PC are displayed. *PC1 in LUAD was strongly negatively associated with the expression of hallmark gene sets related to proliferation (Extended Data Fig. 1f, Methods). GD, genome doubling; TMB, tumour mutational burden; wGII, weighted genome instability index. **b**, I-TED, calculated as the mean normalized gene expression correlation distance for a given region paired with every other region from the same tumour, displayed by histology. **c**, Proportion of variance in I-TED explained by selected genomic and clinical features from a linear model using 260 tumours with at least 2 primary tumour regions, and purity and genome instability estimates. Histological types represented by only a single tumour were excluded to ensure a sufficiently large sample size to estimate the effect of histology. **P = 0.003, ***P = 5.15 × 10^{-10}. **d**, ASCAT-derived tumour purity and RNA estimate of the tumour transcripts fraction. Each dot represents one tumour region. A modified version of ASCAT[50] was used to estimate the proportion of tumour and non-tumour cells within an admixed sequencing sample. **e**, dN/dS, inferring positive and negative selection of truncating somatic mutations, for cancer genes and non-cancer genes, by tertiles of median gene expression across the cohort (left) and by tertiles of gene expression ITH across the cohort (right). Dots represent the estimated dN/dS and the error bars represent the 95% confidence intervals calculated using the genesetdnds function in R from the package dNdScv. A dN/dS estimate is considered significant if the 95% confidence intervals do not overlap 1. Expression level tertiles contained 76, 24 and 9 cancer genes, and 4,856, 5,100 and 5,166 non-cancer genes, for tertiles 3, 2 and 1, respectively. Expression ITH tertiles contained 54, 24 and 31 cancer genes and 4,994, 5,082 and 5,046 non-cancer genes, for tertiles 3, 2 and 1, respectively. Median expression levels and expression ITH were based on the total number of tumour samples collected at surgical resection from tumours with more than one sample at that time point (n = 845 regions from 283 tumours).

(χ^2 test, P = 2.8 × 10^{-81}; Extended Data Fig. 1j). These results are consistent with the idea that a subset of highly and homogeneously expressed non-cancer genes are conserved during somatic evolution and with the presence of weak negative selection among mutations in cancer.

ASE in NSCLC

Next, we focused on transcriptomic diversity arising from ASE, which may result from genomic allelic imbalance (termed copy number (CN)-dependent ASE) or from unequal allelic expression per chromosome copy (CN-independent ASE).

We analysed genes that contained at least one heterozygous germline single-nucleotide polymorphism (SNP) with an RNA coverage of >8 reads (Methods). It was possible to evaluate ASE in a total of 16,378 different genes across all samples within the cohort at an average of 3,809 (s.d. ± 885) and 4,064 (s.d. ± 485) genes per tumour and normal tissue sample, respectively.

We evaluated CN-dependent and CN-independent ASE using an approach that controls for the difference in tumour purity and tumour transcript fraction of each sample (Figs. 1d and 2a and Methods)[19]. The mean percentage of evaluable genes with CN-dependent ASE in each tumour region was 17.4% (s.d. ± 12.7%), compared with 1.01% with CN-independent ASE (s.d. ± 0.47%), which partially reflects our stringent approach to calling CN-independent ASE (Fig. 2b).

ASE can result from genomic imprinting or truncating mutations that lead to nonsense-mediated decay of the mutant allele. In keeping with this, imprinted genes[20] were significantly enriched among genes most frequently affected by CN-independent ASE (odds ratio (OR) = 71.3, P < 2.2 × 10^{-16}) and explained 5.4% of the observed CN-independent ASE. CN-independent ASE was also enriched in genes that contained

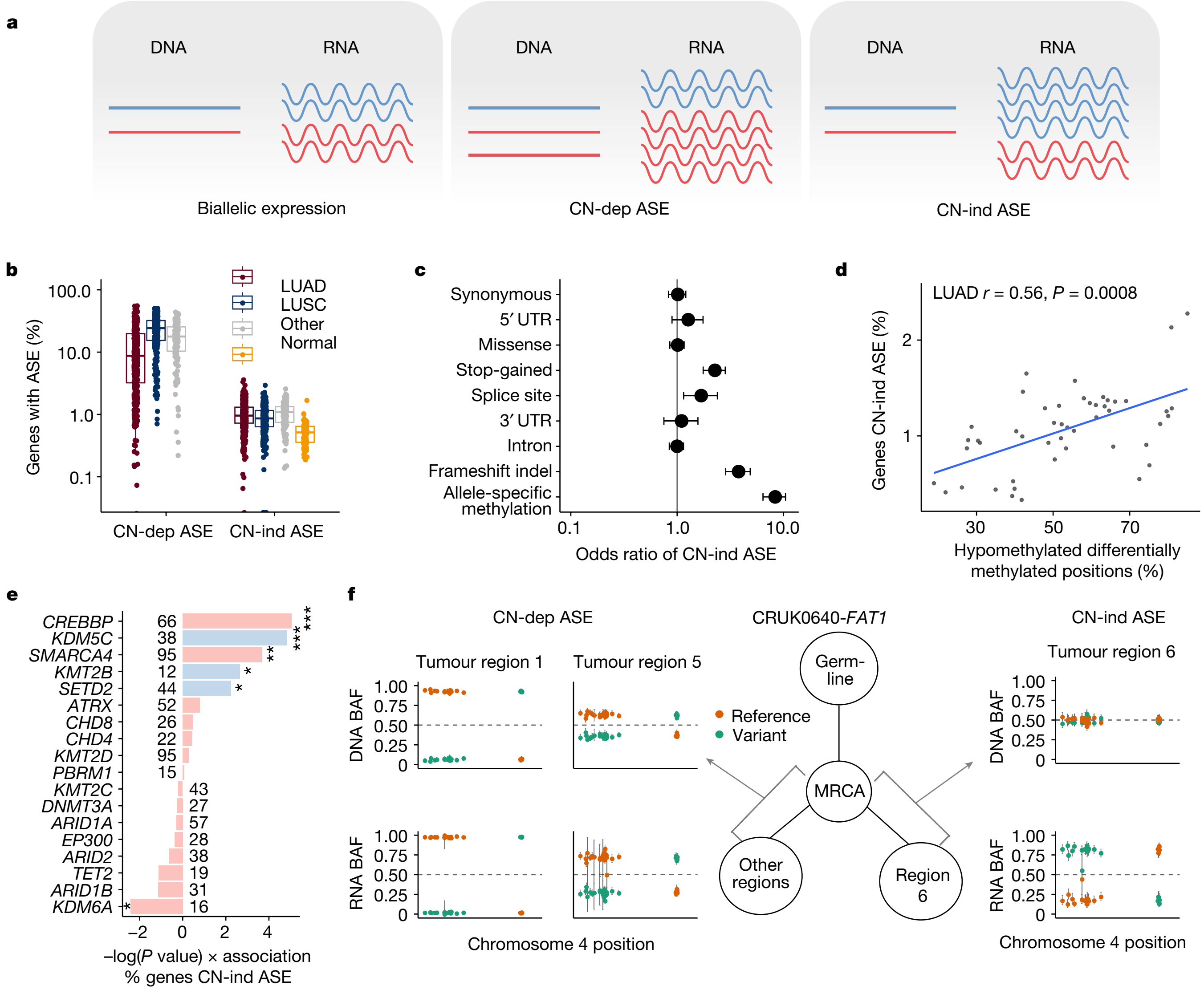


Fig. 2 | ASE in NSCLC. a, Schematic displaying the concepts of biallelic expression, CN-dependent ASE (CN-dep ASE) and CN-independent ASE (CN-ind ASE). **b**, Proportion of evaluable (containing an expressed SNP) genes affected by CN-dependent ASE and CN-independent ASE in tumours and normal tissue samples. LUAD, $n = 454$ regions from 144 tumours; LUSC, $n = 293$ regions from 88 tumours; Other, other subtypes, $n = 130$ regions from 38 tumours; Normal, tumour-adjacent normal lung tissue, $n = 95$. **c**, Points indicate odds ratio estimates for CN-independent ASE when somatic point mutations, or ASM (in samples for which both RRBS and RNA-seq were available) was concomitantly detected in the same gene, by type of alteration. Bars indicate 95% confidence intervals. Odds ratios for the links between CN-independent ASE and mutations and between CN-independent ASE and ASM are based on 876 primary tumour regions from 332 tumours and on 96 tumour regions from 31 tumours, respectively. **d**, Relationship in LUAD between the proportion of evaluable genes with CN-independent ASE and the ratio of differentially hypomethylated clusters of neighbouring CpGs compared to all differentially methylated genomic positions. The *P* value was calculated using a linear mixed-effects model with tumour as the random variable. **e**, Linear mixed-effects model showing the impact of driver mutations in candidate epigenetic modifier genes[22] (mutated in more than five tumours) and tumour mutational burden on the proportion of evaluable genes with CN-independent ASE. Factors independently associated with increased CN-independent ASE in a multivariable model are coloured blue. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. **f**, An example of genomic–transcriptomic mirrored subclonal allelic imbalance occurring in *FAT1* within CRUK0640. DNA and RNA B allele frequencies (BAFs) for each SNP in *FAT1* are plotted and coloured according to the reference and variant status of each allele for each region sampled within the tumour. In this instance, there is evidence of CN-dependent ASE in two regions and CN-independent ASE in one region. These events favour overexpression of different parental alleles and occur on different branches of the phylogenetic tree; a simplified version is displayed. MRCA, most recent common ancestor.

a stop-gain mutation (OR = 2.23, $P = 1.4 \times 10^{-11}$), an insertion or deletion leading to a frameshift (OR = 3.73, $P < 2.2 \times 10^{-16}$) or a splice-site mutation (OR = 1.66, $P = 0.006$) (Fig. 2c). Such mutations explained 0.7% of the total observed CN-independent ASE and had a reduced impact on CN-dependent ASE (Extended Data Fig. 2a).

ITH of CN-independent ASE, defined as the proportion of events that were detected in only a subset of the tumour regions in which it was possible to evaluate ASE, was correlated with I-TED (Pearson's $r = 0.25$, $P = 4 \times 10^{-5}$; Extended Data Fig. 2b). A linear model of the determinants of I-TED revealed that the heterogeneity of SCNAs and CN-independent ASE were independent predictors of I-TED, accounting for 13.9% and 2.7% of variance, respectively (Extended Data Fig. 2c, $P = 2.4 \times 10^{-10}$ and $P = 0.004$, respectively). This result highlights the link between ASE and transcriptional diversity.

Next, we assessed whether patterns of CN-independent ASE varied between tumour and tumour-adjacent normal tissue samples. The lack of expressed SNPs within many genes necessitated the imputation of missing data; therefore we considered genes in which ASE was evaluable in ≥100 tumour regions across the cohort (Methods). PCA revealed that normal tissue samples were distinguishable from tumour samples, which suggests that patterns of CN-independent ASE are fundamentally different between normal tissue and tumour samples (Extended Data Fig. 2d). Gene-level analysis showed that 11 genes were subject to differential CN-independent ASE between normal and tumour tissue when controlling for repeated measures: *NTM* (more frequent CN-independent ASE in normal tissue); and *NLRP2*, *PRIM2*, *CSNK2A3*, *GALNT18*, *ZNF597*, *RAB5B*, *RRM1*, *CAST*, *PDE4DIP* and *LOC653513* (more frequent CN-independent ASE in tumours) (Extended Data Fig. 2e).

To investigate the mechanisms that underpin CN-independent ASE, we examined tumour regions (96 regions from 31 tumours) with DNA methylation data from reduced representation bisulphite sequencing (RRBS). Copy-number-aware methylation deconvolution analysis of cancers (CAMDAC)[21] was used to estimate allele-specific methylation (ASM) rates, excluding the signal from non-cancer cells (Methods). ASM was 8.4 times more likely to occur at the promoters of genes showing CN-independent ASE than those without CN-independent ASE ($P < 2.2 \times 10^{-16}$, Fisher's exact test; Fig. 2c). When global levels of methylation in a tumour region (measured as the percentage of all differentially methylated positions that comprise hypomethylated CpG loci) were compared with the proportion of evaluable genes with CN-independent ASE, a correlation was observed in LUADs but not LUSCs (Pearson's $r = 0.56$, $P = 0.0008$, linear mixed-effects model; Fig. 2d and Extended Data Fig. 2f). In LUADs, CN-independent ASE might therefore represent a surrogate for methylation patterns.

Given the relationship between CN-independent ASE and epigenetic variation, we proposed that tumours that harbour driver mutations in epigenetic modifier genes[22] might contain more CN-independent ASE. Consistent with this hypothesis, univariate linear regression analysis revealed that mutations within epigenetic modifier genes, in particular *CREBBP*, *KDM5C*, *SMARCA4*, *SETD2* and *KMT2B*, were associated with higher levels of CN-independent ASE. By contrast, *KDM6A* predicted driver mutations were associated with decreased CN-independent ASE ($P < 0.05$; Fig. 2e). A multivariable linear mixed-effects model, controlling for tumour mutational burden and repeated measures, confirmed that mutations in *SETD2*, *KDM5C* and *KMT2B* were independently predictive of higher levels of CN-independent ASE. To validate this observation, we explored publicly available RNA-seq data from *SETD2*-deficient isogenic human cell lines. Across H1650 (lung)[23], 786-0 (renal)[24] and HepG2 (liver)[25] cell lines, we observed an increase in CN-independent ASE in *SETD2*-deficient cells compared with wild type ($P = 0.009$, linear mixed-effects model; Extended Data Fig. 2g).

Cataloguing CN-dependent and CN-independent ASE within multiregion tumours also enabled the identification of examples of parallel evolution in which genomic and transcriptomic events affecting the same gene evolve independently in different subclones within a tumour. Such events would not be detected with a genomic-only approach. We utilized haplotype phasing to explore evidence of mirrored subclonal allelic imbalance (MSAI), in which the maternal allele is gained or lost in one subclone of a tumour but the paternal allele is gained or lost in another subclone independently. We provide an example of this phenomenon in the context of allelic expression data in tumour CRUK0640 (Fig. 2f). Here the tumour suppressor gene *FAT1* contained a loss of heterozygosity with associated CN-dependent ASE in two tumour regions. However, in one other tumour region, *FAT1* did not contain a SCNA but instead showed evidence of CN-independent ASE, which might represent transcript repression favouring the expression of the parental allele subject to copy number loss in the other two regions. Phylogenetic reconstruction demonstrated that the tumour regions showing CN-dependent and CN-independent ASE were found on different branches, suggestive of parallel evolution, with convergence upon the loss of different alleles of *FAT1* through different mechanisms. This example of genomic–transcriptomic MSAI highlights that CN-independent ASE can provide an alternative source of diversity to genomic variation in an evolving cancer.

RNA-editing diversity in NSCLC

Another potential source of transcriptomic diversity is RNA editing, a post-transcriptional process characterized by changes in the nucleotide sequence of RNA molecules. We applied a stringent approach to define exonic RNA substitutions (single nucleotide changes exclusive to RNA molecules and absent in DNA) and identified 40,057 RNA substitutions across 6,019 specific sites across the cohort (mean of 1.26 RNA substitutions per Mb per tumour; Fig. 3a,b and Extended Data Fig. 3a). The majority (mean 59.7% per tumour region) were A>G substitutions, in keeping with ADAR-linked RNA editing, which deaminates adenosine to inosine, a nucleotide that is then read as guanosine by the translation machinery[26] and sequencing platforms. Of these substitutions, 65% were present in the REDIportal database[27] of known A>G editing events in human tissues. C>T substitutions[28] represented 11.8% of the total substitutions detected. Of all the RNA substitutions detected, 67% were tumour specific (not present within a TRACERx panel of samples of normal tissue), and of these, 29.4% were shared between two or more tumours.

To investigate the molecular processes that underlie RNA editing in an unbiased manner, we generated RNA single-base substitution (RNA-SBS) signatures, which considered not only the mutated base but also the two adjacent bases and the strand on which the mutation occurred[29]. We detected five RNA-SBS signatures: RNA-SBS1 to RNA-SBS5 (Fig. 3c). RNA-SBS1 consisted predominantly of A>G transitions, whereas RNA-SBS2 consisted mainly of C>T transitions. RNA-SBS3 consisted mainly of A>G and T>C transitions, RNA-SBS4 of G>A transitions and RNA-SBS5 of G>T transversions. RNA-SBS1 and RNA-SBS3 were identified in most tumours (RNA-SBS1 in 98% and RNA-SBS3 in 85%). RNA-SBS1 exhibited the lowest ITH and was detected within all regions of 87.4% of multiregion tumours.

An unbiased correlation of the activity of each RNA-SBS signature with gene expression (Extended Data Fig. 3b) revealed a relationship between RNA-SBS1 and *ADAR* expression (Pearson's $r = 0.42$, FDR = 2.4×10^{-14}, linear mixed-effects model; Fig. 3d). The RNA 192-channel substitution spectrum (encompassing all possible trinucleotide contexts of RNA substitutions across both transcribed strands) derived when considering only those events that overlapped curated A>G sites from REDIportal was highly similar to RNA-SBS1 (cosine similarity = 0.97), consistent with the A>G activity of ADAR underpinning RNA-SBS1.

RNA-SBS2 was dominated by C>T transitions at TpC sites (67%), a motif consistent with the RNA editing activity of APOBEC3A (ref. [30]). In keeping with this, an unbiased analysis showed that RNA-SBS2 correlated more strongly with *APOBEC3A* expression than with any other gene in the transcriptome (Pearson's $r = 0.73$, FDR = 4.7×10^{-108}; Fig. 3d). A multiple linear regression considering all APOBEC enzymes revealed that the expression of *APOBEC3A* was the strongest independent predictor of RNA-SBS2 activity, although *APOBEC3F* was also significant ($P = 2.6 \times 10^{-57}$ and $P = 0.01$ for *APOBEC3A* and *APOBEC3F*, respectively, linear mixed-effects model). Investigating the link between RNA-SBS2 and C>T enrichment at APOBEC3A-specific motifs[30,31] further confirmed that RNA-SBS2 was strongly influenced by *APOBEC3A* expression (Extended Data Fig. 3c,d). Associations between gene expression or genomic features and the activity of the three remaining RNA-SBS signatures did not produce any obvious explanations for their aetiology.

Next we tested whether the processes that underlie RNA substitutions were also identified within paired normal tissue samples and whether they were preserved over time during cancer evolution (Fig. 3e and Extended Data Fig. 3e). For all RNA-SBS signatures, activity was correlated between metastatic regions and their paired primary tumours. However, signature activity was also preserved between tumour regions and paired normal

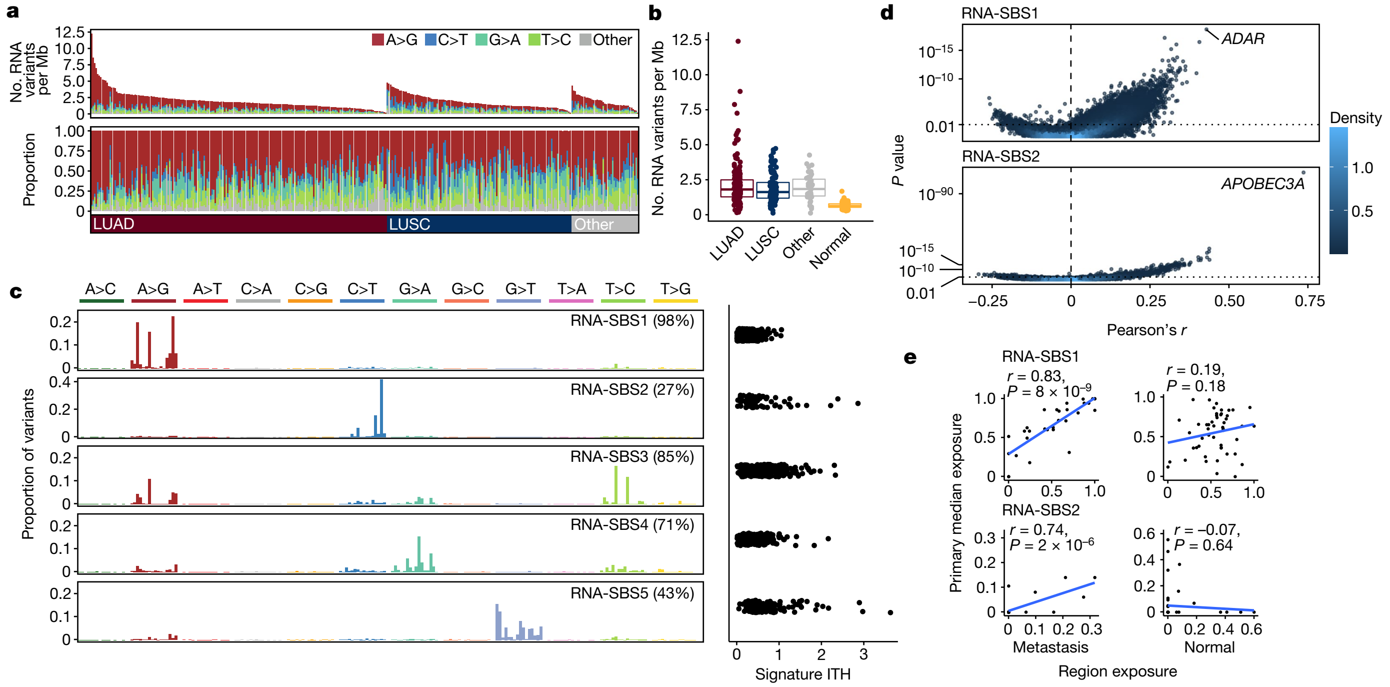


Fig. 3 | RNA-SBS signatures in NSCLC. a, RNA-editing overview (from top to bottom): number and type of RNA substitutions per Mb per primary tumour, tumours are sorted from left to right by histological subtype and by number of substitutions; proportion of each editing type per tumour; NSCLC histological subtype per tumour. **b**, Number of RNA substitutions detected per tumour by histological subtype of NSCLC and in normal adjacent lung tissue. LUAD, $n = 190$; LUSC, $n = 119$; Other, other subtypes, $n = 43$; Normal, tumour-adjacent normal lung tissue, $n = 96$. Boxes represent the lower quartile, median and upper quartile. **c**, Left, trinucleotide profile of each RNA-SBS signature (left). Only samples from patients with more than 20 RNA variants were considered, $n = 333$. Right, signature ITH measured as standard deviation of each signature exposure across tumour regions divided by the mean exposure of each signature across the cohort, based on 280 tumours with more than 20 RNA variants and more than one region. The percentage of tumours with signature activity in at least one primary region is indicated in parentheses. **d**, Volcano plot showing the Pearson's *r* correlations between the number of RNA-SBS1 (top) or RNA-SBS2 (bottom) substitutions with the expression of all genes in the transcriptome. *P* values were calculated using a linear mixed-effects model, using the tumour of origin of each region as random effect. *P* values were adjusted for repeated measures. Correlations were based on 765 primary tumour regions with at least 20 RNA variants from 329 tumours. Colour indicates dot density, with light coloured points belonging to areas of high density in the plot. **e**, Correlation between the exposure of RNA-SBS signatures within tumour-adjacent normal lung tissue and their respective primary tumour regions, and metastatic tumour regions and their respective seeding regions in the primary tumour. Primary tumour exposure was calculated as the median exposure across all primary regions for the comparison with tumour-adjacent normal tissue, and across all seeding regions for the comparison with metastases. Only primary–metastasis pairs where more than 20 RNA substitutions were detected in the metastasis and primary region were used ($n = 50$ pairs for normal samples, $n = 31$ for metastases). *P* values were computed with a two-sided *t* test testing the null hypothesis that the Pearson correlation coefficient $r = 0$.

tissue samples in the case of RNA-SBS3, RNA-SBS4 and RNA-SBS5, but not in RNA-SBS1 or RNA-SBS2. These findings suggest that the processes that underlie changes in RNA-SBS1 and RNA-SBS2 might be tumour-specific. Moreover, they might occur de novo within the regions of primary tumours that seed metastasis and persist within their metastases. By contrast, those that fuel RNA-SBS3, RNA-SBS4 and RNA-SBS5 might be influenced by germline, environmental or technical factors.

ADAR has previously been linked to epigenomic dysregulation[32]. Accordingly, we tested whether RNA-SBS1 activity might be influenced by epigenomic dysregulation within tumours. We observed a significant correlation between global levels of hypomethylation and RNA-SBS1 activity in tumour regions but not in paired normal tissue samples. This result highlights that these processes might be linked in NSCLC evolution (Pearson's $r = -0.35$, $P = 0.008$, linear mixed-effects for tumour regions; $P = 1$ for normal tissue samples; Extended Data Fig. 3f).

Multi-omic features of metastasis

Finally, we evaluated the dynamics of transcriptomic diversity during metastatic progression. We observed significantly higher transcriptomic diversity between paired primary–metastatic tumour regions than between primary regions derived from the same tumour (Fig. 4a). This relationship remained consistent when considering only intrathoracic non-LN metastases (Extended Data Fig. 4a), which suggests that there are consistent differences between primary and metastatic transcriptomes that cannot be fully explained by microenvironmental differences between metastatic organ sites. To further explore this finding, we compared transcriptomic diversity between metastasis-seeding or non-seeding primary tumour regions and their paired metastatic tumours (Fig. 4b). Across the cohort, expression patterns in metastases were more similar to the metastasis-seeding primary regions than non-seeding primary regions ($n = 22$ primary–metastasis pairs from 18 tumours[6]; $P = 0.0019$, two-tailed paired Wilcoxon test; Fig. 4c). Gene set enrichment analyses between these regions showed an enrichment within seeding regions for gene sets linked to proliferation and a depletion in immune-linked groups[14] (Extended Data Fig. 4b). Taken together, these results suggest that a proportion of the transcriptomic patterns observed in metastatic tumours are underpinned by somatic changes that originated in the primary tumour and are capable of influencing ongoing evolution, including metastasis.

To further explore this result, we evaluated the impact of relevant molecular features on the metastatic potential of a tumour region. In particular, we tested whether transcriptomic features are informative for inferring metastatic potential. To achieve this, we built an ensemble

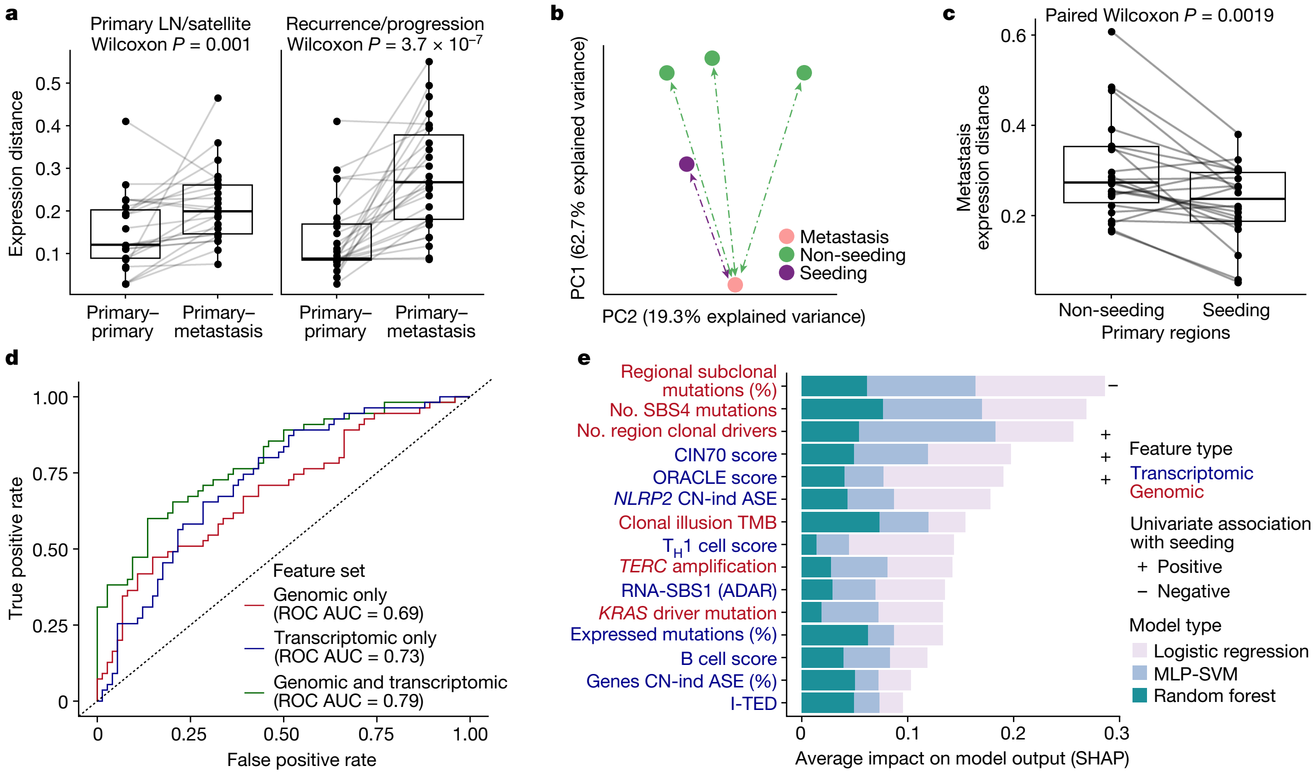


Fig. 4 | Transcriptional landscape of seeding tumour regions. a, Expression distance between primary regions compared to either metastatic LN regions or pulmonary nodules resected at the time of surgery (left) or metastatic regions resected at relapse within the same patient (right). Only tumours containing two or more regions with at least one metastatic region sampled are shown ($n = 50$ primary–metastasis pairs from 35 tumours). **b**, First two PCs for all available primary and metastatic tumour regions in an example tumour, CRUK0361, based on gene expression levels. The region containing the seeding clone was more proximal to the metastatic sample than other primary regions. **c**, Expression distance between metastatic samples and their paired primary samples across the cohort depending on whether the region contained a seeding clone(s). The analysis was run on 22 metastatic samples that had gene expression data for both seeding and non-seeding primary regions. **d**, ROC curves for ensemble models trained on each feature set: genomic only (red), transcriptomic only (blue), combined genomic and transcriptomic (green) and assessed against the held-out test dataset. The predictions are based on 516 primary tumour regions from 206 tumours for which seeding status could be established and for which all metrics tested could be measured (307 non-seeding regions, 209 seeding), with a 75/25% training/test dataset split. **e**, Mean Shapley additive explanations (SHAP) values (calculated across the held-out test dataset) for each feature in the combined ensemble model, capturing the importance of each feature for model prediction. Label colours indicate the feature type, genomic (red) or transcriptomic (blue), and box colours indicate the model type from which the SHAP values were extracted. The symbols at the end of the bars indicate either a significantly positive (+) or negative (−) association, with increased weight for seeding potential based on a two-sided Wilcoxon test comparing seeding to non-seeding regions. MLP-SVM, multilayer perceptron with support vector machine. All box plots in this figure represent the lower quartile, median and upper quartile, whiskers represent lower and higher bound ±1.5× interquartile range. All Wilcoxon tests shown here (paired or unpaired) were two sided.

machine-learning classifier to predict whether a tumour region contained a seeding clone (or seeding clones) (Methods and Extended Data Fig. 4c). Leveraging a recently published approach[33], we defined three feature sets: genomic only; transcriptomic only; and combined genomic and transcriptomic (see Methods for details on initial feature selection). For each feature set, we trained three model types (logistic regression, random forest, multilayer perceptron with support vector machine terminal layer) and selected the best model in each case (hyperparameter tuning using a randomized search grid across relevant parameters with *K*-fold stratified cross validation; the best model was selected as that with the highest balanced accuracy – see Methods for full details). The combined genomic and transcriptomic feature set generated a marginally better classifier relative to the classifiers generated from the independent genomic and transcriptomic feature sets (Fig. 4d, combined receiver operator characteristic (ROC) area under the curve (AUC) = 0.79; genomic only ROC AUC = 0.69, transcriptomic only ROC AUC = 0.73). Overall, the combined feature classifier showed promising performance with an accuracy of 71% (significantly greater than the no-information rate, $P = 0.0007$), although with much greater specificity than sensitivity (sensitivity = 0.51, specificity = 0.86). Of the variables tested, the two most important to infer metastatic potential related to the evolutionary context of the mutations within the tumour region: a decreased proportion of subclonal mutations that were present in only a subset of tumour cells within the tumour region (that is, were not regionally dominant)[7] and a decreased number of mutations linked to the smoking signature SBS4 (likely a proxy for the trunk length of the phylogenetic tree)[34,35] (Fig. 4e and Extended Data Fig. 4d). Similarly, increased CIN70 (ref. [36]) and ORACLE[17] expression signature scores, both associated with proliferation, were also associated with metastatic potential and demonstrated strong weighting across different models. Other variables described in this work also helped the classifier to discriminate regions with metastatic potential. These included CN-independent ASE within *NLRP2*, RNA-SBS1 activity and the proportion of genes with CN-independent ASE per region.

Discussion

Multiregion sequencing studies in the past decade have highlighted important genomic alterations in cancer, including point mutations and CN alterations, that drive ITH and fuel cancer evolution[2,3,37–39].

Through paired genomic–transcriptomic analysis and multiregion sampling of NSCLC, we highlight sources of variation that would be missed by an exclusively genomic approach. Highly expressed cancer genes with low intratumour variance were more likely to be under positive selection. This finding could inform studies seeking to discover new cancer genes. Our results also imply the presence of limited yet significant negative selection in cancer evolution, consistent with constraints to tumour development[16,40]. The additional resolution gained by restricting to uniformly expressed genes mirrors results reported in a previous publication, in which an expression signature composed of such genes represented a robust biomarker in NSCLC[17].

We find pervasive ASE in NSCLC and find that, as expected in a disease characterized by significant chromosomal instability, in the majority of tumours, ASE is predominantly explained by SCNAs. However, we highlight an important fraction of ASE that is linked to epigenomic dysfunction, in particular to changes in DNA methylation and inactivating mutations in the histone methyltransferase *SETD2*, the lysine demethylase *KDM5C* and the lysine methyltransferase *KMT2B* genes. Our observations build on in vivo single-cell studies of mouse models of LUAD, which have highlighted the importance of a dynamic epigenome in governing tumour progression[41]. Furthermore, our results provide orthogonal insight into the role of these genes in transcription: loss of *SETD2* has been linked to increased oncogenic transcriptional output[42], and *KDM5C* regulates transcription through H3K4 demethylation[43]. In addition, previous work has linked ASE to epigenomic changes through enhancer activity[44]. Future work should focus on the interplay between mutated epigenetic modifiers and enhancer activity in cancer.

We also utilize approaches that have previously been leveraged to define DNA mutational signatures to extract unbiased trinucleotide signatures of RNA single-base substitutions from paired DNA and RNA-seq. We show that these signatures underlie RNA editing. Importantly, two signatures, RNA-SBS1 (linked to ADAR editing) and RNA-SBS2 (linked to APOBEC3A editing) seemed to be underpinned by heritable somatic mechanisms. That is, their activities were preserved between paired primary and metastatic samples, which is consistent with recent work linking genomic variants with RNA editing[45]. Their potential importance to tumour evolution is underlined by the roles that APOBEC enzymes play in driving genome instability[46] and by the observed relationship between RNA-SBS1 activity and patterns of aberrant hypomethylation (Extended Data Fig. 3f). *ADAR*, which we suggest underpins RNA-SBS1, has been proposed as a tumour suppressor gene that is essential to cancer cells in the context of epigenomic dysfunction owing to its ability to target otherwise immunogenic double-stranded RNAs within short interspersed nuclear elements[32]. Furthermore, our data strongly suggests that *APOBEC3A* drives the observed RNA-SBS2 signature, consistent with a stronger role for *APOBEC3A* compared with other *APOBEC3* gene family members, including *APOBEC3B* in RNA[47,48].

Multiregion paired genomic–transcriptomics enables characterization of the phenotype of the metastasis-seeding region of the primary tumour. To elucidate the influence of transcriptomic features on the biology of metastasis, we used a combined machine-learning approach, which revealed key features of the metastatic transition and demonstrated that both genomic and transcriptomic features are able to predict metastatic potential. In particular, we found that two genomic metrics relating to the evolutionary context of mutations (proxies for the dominance of a subclone within a tumour region and decreased phylogenetic trunk length) and two gene expression signatures related to proliferation were markers of metastatic potential. In a companion paper[7], we report that the presence of recent subclonal expansions (that is, large subclones at the terminus of a phylogenetic branch) are associated with shorter disease-free survival. We also observed that within both circulating tumour DNA and primary tumour tissue, the size of subclonal mutational clusters is linked to their metastatic potential[6,49]. Conceivably, recent subclonal expansions might be driven by increased proliferation, which is captured by transcriptomic signatures. Also, for a given tumour region, metrics related to the evolutionary background of its constituent mutations were associated with metastatic potential. The finding that such metrics tended to be more useful at differentiating primary tumour regions with and without metastasis-seeding potential than the specific genes in which mutations occurred may have important implications for biomarker discovery.

Of note, this machine-learning approach combined multiple variables, which together might render some features redundant within the classifier. This could mean that variables not presented within Fig. 4e might be biologically important.

A limitation of this work is that it is unlikely to have captured the true extent of transcriptomic variation in these tumours. We did not consider all forms of transcriptional variation, including alternative splicing. Furthermore, we could only study ASE in the minority of genes that contain an expressed SNP, and we applied a strict, but specific, approach to defining CN-independent ASE. Similarly, the filters applied to identify RNA variants, rather than transcribed DNA mutations, were stringent, and only exonic events (capturing only a fraction of RNA editing)[26] were considered. Therefore, it is likely that we have underestimated the variation and biological impact attributable to these processes in tumours. Nevertheless, in this way, we were able to identify previously unknown RNA-SBS signatures, including three of unknown aetiology. This result highlights the need for further studies that examine patterns of RNA editing in larger datasets. Finally, our machine-learning approach utilized tumour region transcriptomic data that was not deconvolved at the subclone level. Therefore we could not test subclone-specific metrics, which are also likely to affect metastatic potential[6].

Despite these limitations, this work has shown that transcriptional variation is likely to play an important part in NSCLC evolution. It has revealed sources of diversity that would not have been identified by a focused analysis of the cancer genome and underlined the importance of multi-omic sequencing and systems biology approaches to the study of tumour evolution.

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/s41586-023-05706-4.

1. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).
2. Bailey, C. et al. Tracking cancer evolution through the disease course. *Cancer Discov.* **11**, 916–932 (2021).
3. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
4. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
5. Marjanovic, N. D. et al. Emergence of a high-plasticity cell state during lung cancer evolution. *Cancer Cell* **38**, 229–246.e13 (2020).
6. Al Bakir, M. et al. The evolution of non-small cell lung cancer metastases in TRACERx. *Nature* https://doi.org/10.1038/s41586-023-05729-x (2023).
7. Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* https://doi.org/10.1038/s41586-023-05783-5 (2023).
8. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
9. Travis, W. D. et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* **10**, 1243–1260 (2015).
10. Lam, V. K. et al. Targeted tissue and cell-free tumor DNA sequencing of advanced lung squamous-cell carcinoma reveals clinically significant prevalence of actionable alterations. *Clin. Lung Cancer* **20**, 30–36.e3 (2019).
11. East, P. et al. Oncogenic RAS activity predicts response to chemotherapy and outcome in lung adenocarcinoma. *Nat. Commun.* **13**, 5632 (2022).
12. Dogan, S. et al. Molecular epidemiology of *EGFR* and *KRAS* mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin. Cancer Res.* **18**, 6169–6177 (2012).
13. Buettner, R. Invasive mucinous adenocarcinoma: genetic insights into a lung cancer entity with distinct clinical behavior and genomic features. *Mod. Pathol.* **35**, 138–139 (2022).

14. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
15. Cao, S. et al. Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. *Nat. Biotechnol.* **40**, 1624–1633 (2022).
16. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
17. Biswas, D. et al. A clonal expression biomarker associates with lung cancer mortality. *Nat. Med.* **25**, 1540–1548 (2019).
18. Cowley, G.S. et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* **1**, 140035 (2014).
19. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
20. Baran, Y. et al. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).
21. Larose Cadieux, E. et al. Copy number-aware deconvolution of tumor-normal DNA methylation profiles. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.03.366252 (2020).
22. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17**, 284–299 (2016).
23. Zhou, Y. et al. Histone methyltransferase SETD2 inhibits tumor growth via suppressing CXCL1-mediated activation of cell cycle in lung adenocarcinoma. *Aging* **12**, 25189–25206 (2020).
24. Ho, T. H. et al. High-resolution profiling of histone H3 lysine 36 trimethylation in metastatic renal cell carcinoma. *Oncogene* **35**, 1565–1574 (2016).
25. Chen, K. et al. Methyltransferase SETD2-mediated methylation of STAT1 is critical for interferon antiviral activity. *Cell* **170**, 492–506.e14 (2017).
26. Bazak, L. et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24**, 365–376 (2014).
27. Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).
28. Sharma, S., Patnaik, S. K., Taggart, R. T. & Baysal, B. E. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci. Rep.* **6**, 39100 (2016).
29. Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **101**, 1566–1581 (2006).
30. Sharma, S. & Baysal, B. E. Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. *PeerJ* **5**, e4136 (2017).
31. Sharma, S. et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat. Commun.* **6**, 6881 (2015).
32. Mehdipour, P. et al. Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. *Nature* **588**, 169–173 (2020).
33. Sammut, S.-J. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
34. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
35. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
36. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
37. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
38. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
39. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
40. López, S. et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
41. LaFave, L. M. et al. Epigenomic state transitions characterize tumor progression in mouse lung adenocarcinoma. *Cancer Cell* **38**, 212–228.e13 (2020).
42. Xie, Y. et al. SETD2 loss perturbs the kidney cancer epigenetic landscape to promote metastasis and engenders actionable dependencies on histone chaperone complexes. *Nat. Cancer* **3**, 188–202 (2022).
43. Outchkourov, N. S. et al. Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. *Cell Rep.* **3**, 1071–1079 (2013).
44. Sungalee, S. et al. Histone acetylation dynamics modulates chromatin conformation and allele-specific interactions at oncogenic loci. *Nat. Genet.* **53**, 650–662 (2021).
45. Li, Q. et al. RNA editing underlies genetic risk of common inflammatory diseases. *Nature* **608**, 569–577 (2022).
46. Venkatesan, S. et al. Induction of APOBEC3 exacerbates DNA replication stress and chromosomal instability in early breast and lung cancer evolution. *Cancer Discov.* **11**, 2456–2473 (2021).
47. Jalili, P. et al. Quantification of ongoing APOBEC3A activity in tumor cells by monitoring RNA editing at hotspots. *Nat. Commun.* **11**, 2971 (2020).
48. Petljak, M. et al. Mechanisms of *APOBEC3* mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
49. Abbosh, C. et al. Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. *Nature* https://doi.org/10.1038/s41586-023-05776-4 (2023).
50. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).

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

[1]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [2]Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [3]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute and University College London Cancer Institute, London, UK. [4]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [5]Integrative Cancer Genomics Laboratory, Department of Oncology, KU Leuven, Leuven, Belgium. [6]VIB–KU Leuven Center for Cancer Biology, Leuven, Belgium. [7]Medical Genomics, University College London Cancer Institute, London, UK. [8]Advanced Sequencing Facility, The Francis Crick Institute, London, UK. [9]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK. [10]Department of Cellular Pathology, University College London Hospitals, London, UK. [11]Department of Pathology, ZAS Hospitals, Antwerp, Belgium. [12]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [13]Experimental Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia. [14]Centre for Nephrology, Division of Medicine, University College London, London, UK. [15]Scientific Computing STP, Francis Crick Institute, London, UK. [16]Oncogene Biology Laboratory, The Francis Crick Institute, London, UK. [17]Bioinformatics and Biostatistics, The Francis Crick Institute, London, UK. [18]Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK. [19]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [20]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [21]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [22]Cancer Research UK & UCL Cancer Trials Centre, London, UK. [23]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [24]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [25]Department of Medical Oncology, University College London Hospitals, London, UK. [73]These authors contributed equally: Carlos Martínez-Ruiz, James R. M. Black, Clare Puttick, Mark S. Hill. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: Charles.Swanton@crick.ac.uk; nicholas.mcgranahan.10@ucl.ac.uk

TRACERx Consortium

Nicholas McGranahan[1,2], Charles Swanton[1,3,25], Carlos Martínez-Ruiz[1,2], James R. M. Black[1,2], Clare Puttick[1,2,3], Mark S. Hill[3], Jonas Demeulemeester[4,5,6], Elizabeth Larose Cadieux[4,7], Kerstin Thol[1,2], Selvaraju Veeriah[1], Cristina Naceur-Lombardelli[1], Antonia Toncheva[1], Paulina Prymas[1], Andrew Rowan[3], Sophia Ward[1,3,8], Foteini Athanasopoulou[1,3,8], Oriol Pich[3], Takahiro Karasaki[1,3,9], David A. Moore[1,3,10], Roberto Salgado[11,12], Emma Colliver[3], Carla Castignani[4,7], Michelle Dietzen[1,2,3], Ariana Huebner[1,2,3], Maise Al Bakir[1,3], Miljana Tanić[7,13], Thomas B. K. Watkins[3], Emilia L. Lim[1,3], Rachel Rosenthal[3], Gareth A. Wilson[3], Alexander M. Frankell[1,3], Nnennaya Kanu[1], Kevin Litchfield[1,18], Nicolai J. Birkbak[1,3,19,20,21], Allan Hackshaw[22], Stephan Beck[7], Peter Van Loo[4,23,24], Mariam Jamal-Hanjani[1,9,25], Jason F. Lester[26], Amrita Bajaj[27], Apostolos Nakas[27], Azmina Sodha-Ramdeen[27], Keng Ang[27], Mohamad Tufail[27], Mohammed Fiyaz Chowdhry[27], Molly Scotland[27], Rebecca Boyles[27], Sridhar Rathinam[27], Claire Wilson[28], Domenic Marrone[28], Sean Dulloo[28], Dean A. Fennell[27,28], Gurdeep Matharu[29], Jacqui A. Shaw[29], Joan Riley[29], Lindsay Primrose[29], Ekaterini Boleti[30], Heather Cheyne[31], Mohammed Khalil[31], Shirley Richardson[31], Tracey Cruickshank[31], Gillian Price[32,33], Keith M. Kerr[33,34], Sarah Benafif[25], Kayleigh Gilbert[35], Babu Naidu[36], Akshay J. Patel[37], Aya Osman[37], Christer Lacson[37], Gerald Langman[37], Helen Shackleford[37], Madava Djearaman[37], Salma Kadiri[37], Gary Middleton[37,38], Angela Leek[39], Jack Davies Hodgkinson[39], Nicola Totten[39], Angeles Montero[40], Elaine Smith[40], Eustace Fontaine[40], Felice Granato[40], Helen Doran[40], Juliette Novasio[40], Kendadai Rammohan[40], Leena Joseph[40], Paul Bishop[40], Rajesh Shah[40], Stuart Moss[40], Vijay Joshi[40], Philip Crosbie[40,41,42], Fabio Gomes[43], Kate Brown[43], Mathew Carter[43], Anshuman Chaturvedi[42,43], Lynsey Priest[42,43], Pedro Oliveira[42,43], Colin R. Lindsay[44], Fiona H. Blackhall[44], Matthew G. Krebs[44], Yvonne Summers[44], Alexandra Clipson[42,45], Jonathan Tugwood[42,45], Alastair Kerr[42,45], Dominic G. Rothwell[42,45], Elaine Kilgour[42,45], Caroline Dive[42,45], Hugo J. W. L. Aerts[46,47,48], Roland F. Schwarz[49,50], Tom L. Kaufmann[50,51], Zoltan Szallasi[52,53,54], Judit Kisistok[19,20,21], Mateo Sokac[19,20,21], Miklos Diossy[52,53,55], Abigail Bunkum[1,9,56], Aengus Stewart[57], Alastair Magness[57], Angeliki Karamani[58], Benny Chain[58], Brittany B. Campbell[3], Chris Bailey[3], Christopher Abbosh[1], Clare E. Weeden[57], Claudia Lee[3], Corentin Richard[1], Crispin T. Hiley[1,3], David R. Pearce[58], Despoina Karagianni[58], Dhruva Biswas[1,3,59], Dina Levi[57], Elena Hoxha[58], Emma Nye[60], Eva Grönroos[57], Felip Gálvez-Cancino[58], Francisco Gimeno-Valiente[1], George Kassiotis[61,62], Georgia Stavrou[58], Gerasimos Mastrokalos[58], Haoran Zhai[1,3], Helen L. Lowe[58], Ignacio Garcia Matos[58], Jacki Goldman[57], James L. Reading[58], Javier Herrero[59], Jayant K. Rane[3,58], Jerome Nicod[8], Jie Min Lam[1,9,25], John A. Hartley[58], Karl S. Peggs[63,64], Katey S. S. Enfield[3], Kayalvizhi Selvaraju[58], Kevin W. Ng[61], Kezhong Chen[58], Krijn Dijkstra[65,66], Kristiana Grigoriadis[1,2,3], Krupa Thakkar[1], Leah Ensell[58], Mansi Shah[58], Marcos Vasquez Duran[58], Maria Litovchenko[58], Mariana Werner Sunderland[1],

Michelle Leung[1,2,3], Mickael Escudero[57], Mihaela Angelova[3], Monica Sivakumar[1], Olga Chervova[58], Olivia Lucas[1,3,25,56], Othman Al-Sawaf[1,3,9], Philip Hobson[57], Piotr Pawlik[58], Richard Kevin Stone[60], Robert Bentham[1,2], Robert E. Hynds[58], Roberto Vendramin[57], Sadegh Saghafinia[1], Saioa López[58], Samuel Gamble[58], Seng Kuong Anakin Ung[58], Sergio A. Quezada[1,67], Sharon Vanloo[1], Simone Zaccaria[1,56], Sonya Hessey[1,9,56], Stefan Boeing[57], Supreet Kaur Bola[58], Tamara Denner[57], Teresa Marafioti[10], Thanos P. Mourikis[58], Victoria Spanswick[58], Vittorio Barbè[57], Wei-Ting Lu[57], William Hill[57], Wing Kin Liu[1,9], Yin Wu[58], Yutaka Naito[57], Zoe Ramsden[57], Catarina Veiga[68], Gary Royle[69], Charles-Antoine Collins-Fekete[70], Francesco Fraioli[71], Paul Ashford[72], Tristan Clark[73], Martin D. Forster[1,25], Siow Ming Lee[1,25], Elaine Borg[10], Mary Falzon[10], Dionysis Papadatos-Pastos[25], James Wilson[25], Tanya Ahmad[25], Alexander James Procter[74], Asia Ahmed[74], Magali N. Taylor[74], Arjun Nair[74,75], David Lawrence[76], Davide Patrini[76], Neal Navani[77,78], Ricky M. Thakrar[77,78], Sam M. Janes[77], Emilie Martinoni Hoogenboom[79], Fleur Monk[79], James W. Holding[79], Junaid Choudhary[79], Kunal Bhakhri[79], Marco Scarci[79], Martin Hayward[79], Nikolaos Panagiotopoulos[79], Pat Gorman[79], Reena Khiroya[10], Robert C. M. Stephens[79], Yien Ning Sophia Wong[79], Steve Bandula[79], Abigail Sharp[22], Sean Smith[22], Nicole Gower[22], Harjot Kaur Dhanda[22], Kitty Chan[22], Camilla Pilotti[22], Rachel Leslie[22], Anca Grapa[80], Hanyun Zhang[80], Khalid AbdulJabbar[80], Xiaoxi Pan[80], Yinyin Yuan[81], David Chuter[82], Mairead MacKenzie[82], Serena Chee[83], Aiman Alzetani[83], Judith Cave[84], Lydia Scarlett[83], Jennifer Richards[83], Papawadee Ingram[83], Silvia Austin[83], Eric Lim[85,86], Paulo De Sousa[86], Simon Jordan[86], Alexandra Rice[86], Hilgardt Raubenheimer[86], Harshil Bhayani[86], Lyn Ambrose[86], Anand Devaraj[86], Hema Chavan[86], Sofina Begum[86], Silviu I. Buderi[86], Daniel Kaniu[86], Mpho Malima[86], Sarah Booth[86], Andrew G. Nicholson[87,88], Nadia Fernandes[86], Pratibha Shah[86], Chiara Proli[86], Madeleine Hewish[89,90], Sarah Danson[91], Michael J. Shackcloth[92], Lily Robinson[93], Peter Russell[93], Kevin G. Blyth[94,95,96], Craig Dick[97], John Le Quesne[94,95,98], Alan Kirk[99], Mo Asif[99], Rocco Bilancia[99], Nikos Kostoulas[99] & Mathew Thomas[99]

[26]Singleton Hospital, Swansea Bay University Health Board, Swansea, UK. [27]University Hospitals of Leicester NHS Trust, Leicester, UK. [28]University of Leicester, Leicester, UK. [29]Cancer Research Centre, University of Leicester, Leicester, UK. [30]Royal Free Hospital, Royal Free London NHS Foundation Trust, London, UK. [31]Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [32]Department of Medical Oncology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [33]University of Aberdeen, Aberdeen, UK. [34]Department of Pathology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [35]The Whittington Hospital NHS Trust, London, UK. [36]Birmingham Acute Care Research Group, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. [37]University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. [38]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. [39]Manchester Cancer Research Centre Biobank, Manchester, UK. [40]Wythenshawe Hospital, Manchester University NHS Foundation Trust, Wythenshawe, UK. [41]Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK. [42]Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. [43]The Christie NHS Foundation Trust, Manchester, UK. [44]Division of Cancer Sciences, The University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. [45]Cancer Research UK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, UK. [46]Artificial Intelligence in Medicine (AIM) Program, Massachusetts General Brigham, Harvard Medical School, Boston, MA, USA. [47]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [48]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands. [49]Institute for Computational Cancer Biology, Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [50]Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany. [51]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [52]Danish Cancer Society Research Center, Copenhagen, Denmark. [53]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [54]Department of Bioinformatics, Semmelweis University, Budapest, Hungary. [55]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. [56]Computational Cancer Genomics Research Group, University College London Cancer Institute, London, UK. [57]The Francis Crick Institute, London, UK. [58]University College London Cancer Institute, London, UK. [59]Bill Lyons Informatics Centre, University College London Cancer Institute, London, UK. [60]Experimental Histopathology, The Francis Crick Institute, London, UK. [61]Retroviral Immunology Group, The Francis Crick Institute, London, UK. [62]Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK. [63]Department of Haematology, University College London Hospitals, London, UK. [64]Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [65]Department of Molecular Oncology and Immunology, the Netherlands Cancer Institute, Amsterdam, The Netherlands. [66]Oncode Institute, Utrecht, The Netherlands. [67]Immune Regulation and Tumour Immunotherapy Group, Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [68]Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [69]Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. [70]Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [71]Institute of Nuclear Medicine, Division of Medicine, University College London, London, UK. [72]Institute of Structural and Molecular Biology, University College London, London, UK. [73]University College London, London, UK. [74]Department of Radiology, University College London Hospitals, London, UK. [75]UCL Respiratory, Department of Medicine, University College London, London, UK. [76]Department of Thoracic Surgery, University College London Hospital NHS Trust, London, UK. [77]Lungs for Living Research Centre, UCL Respiratory, University College London, London, UK. [78]Department of Thoracic Medicine, University College London Hospitals, London, UK. [79]University College London Hospitals, London, UK. [80]The Institute of Cancer Research, London, UK. [81]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [82]Independent Cancer Patients' Voice, London, UK. [83]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [84]Department of Oncology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [85]Academic Division of Thoracic Surgery, Imperial College London, London, UK. [86]Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [87]Department of Histopathology, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [88]National Heart and Lung Institute, Imperial College London, London, UK. [89]Royal Surrey Hospital, Royal Surrey Hospitals NHS Foundation Trust, Guilford, UK. [90]University of Surrey, Guilford, UK. [91]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [92]Liverpool Heart and Chest Hospital, Liverpool, UK. [93]Princess Alexandra Hospital, The Princess Alexandra Hospital NHS Trust, Harlow, UK. [94]School of Cancer Sciences, University of Glasgow, Glasgow, UK. [95]Cancer Research UK Beatson Institute, Glasgow, UK. [96]Queen Elizabeth University Hospital, Glasgow, UK. [97]NHS Greater Glasgow and Clyde, Glasgow, UK. [98]NHS Greater Glasgow and Clyde Pathology Department, Queen Elizabeth University Hospital, Glasgow, UK. [99]Golden Jubilee National Hospital, Clydebank, UK.

Methods

Data generation and processing

TRACERx cohort. The TRACERx study (ClinicalTrials.gov identifier NCT01888601) is a prospective observational cohort study that aims to transform our understanding of NSCLC, the design of which has been approved by an independent research ethics committee (13/LO/1546). Informed consent for entry into the TRACERx study was mandatory and obtained from every participant. All participants were assigned a study identity number that was known to the individual. These were subsequently converted to linked study identities such that the participants could not identify themselves in study publications. All human samples (tissue and blood) were linked to the study identity number and barcoded such that they were anonymized and tracked on a centralized database, which was overseen by the study sponsor only.

The cohort in this manuscript includes the fraction of samples from the first 421 participants (described in detail in two companion manuscripts[6,7]) with RNA-seq data available after quality checking before and after sequencing (CONSORT diagram in the Supplementary Information).

Seven samples from individuals with disease relapse (CRUK0046_BR_T1-R1, CRUK0046_BR_T1-R2, CRUK0069_MR_T1-R1, CRUK0069_MR_T1-R2, CRUK0280_BR_T1-R1, CRUK0280_BR_T1-R2 and CRUK0679_BP_T1-R1) were not associated with any primary tumour, and one normal sample (CRUK0643_SU_N01) was not paired with any tumour sample with RNA-seq data. These eight samples are present in the raw data but were not included in downstream analyses. Additionally, for four LN samples (CRUK0099_SU_LN01, CRUK0227_SU_LN01, CRUK0240_SU_LN01 and CRUK0240_SU_LN02) the seeding tumour region could not be established[6], and these samples were therefore not used in paired primary–metastatic analyses.

RNA extraction and sequencing. For each sample, total RNA was extracted using a dual extraction method for RNA and DNA using AllPrep DNA/RNA Mini kits (Qiagen). Frozen samples were transferred onto cold Petri dishes kept on dry ice and dissected into 20–30 mg samples. Before extraction, the freshly dissected tissue was transferred directly to homogenization tubes with RLT plus lysis buffer. Homogenization of tissues was carried out using a TissueRuptor II probe or bead methods and by passing the lysate through a QIAshredder column (Qiagen). Libraries were prepared using a minimum of 100 ng input of total RNA, where possible, using an Illumina TruSeq Stranded Total RNA Human/Mouse/Rat ribo-depletion library preparation kit (20020597), and PCR was amplified for 15 cycles according to the manufacturer's guidelines. Final libraries were quality checked using Agilent Tapestation and Promega QuantiFluor dsDNA and were then pooled in equimolar solutions. Sequencing was performed using an Illumina HiSeq 4000 platform at a depth of 50 million paired reads per sample, with a length of 75 bp or 100 bp per read.

DNA methylation sequencing. A subset of previously published primary NSCLC data of the first 100 participants of the TRACERx study with multiple tumour regions were selected for RRBS[21].

The NuGEN Ovation RRBS Methyl-Seq system adapted for automation was applied by using 100 ng of gDNA per sample, digested with MspI, ligated to sequencing adapters and processed using Qiagen's EpiTect Fast DNA Bisulfite kit. Bisulfite-converted DNA libraries were then amplified by PCR (12 cycles) and purified using Agencourt RNAClean XP magnetic beads. The quantity and quality of the resulting libraries were evaluated using a Qubit dsDNA HS assay (Invitrogen) and an Agilent Bioanalyzer High Sensitivity DNA assay (Agilent Technologies), respectively. Samples were multiplexed in pools of 8 and sequenced on a HiSeq2500 system in single-end 100 bp runs. Sequencing outputs were converted into fastq files. FastQC (v.0.11.2)[51] was used for quality control, and Trim Galore! (Babraham Institute), a wrapper around Cutadapt[52], was applied with default settings to perform quality and adapter trimming for each set of paired-end fastq files. The bisulfite-converted DNA sequence aligner Bismark (v.0.14.4)[53] was used to align reads to the UCSC reference genome hg19 build, and PCR deduplication was carried out using NuDup, leveraging NuGEN's molecular tagging technology (v.2.3; https://github.com/nugentechnologies/nudup).

In total, 96 tumour and 31 tumour-adjacent normal regions from 31 patients (average 3 tumour regions per patient) had matched RRBS and RNA-seq data available.

RNA-seq alignment and gene expression. Illumina adapters were trimmed from raw sequencing reads using Cutadapt (v.2.10)[52] with standard parameters, and the quality of the trimmed reads were estimated per flow cell lane using FastQC (v.0.11.9)[51]. Only fastq files with less than 80% of total reads being duplicates were kept for alignment. Fastq read files passing these quality checks were aligned to the UCSC hg19 human reference genome build using STAR (v.2.5.2a)[54] in two-pass mode with ENCODE 3 parameters, generating one BAM file per tumour region. The same reads were also mapped to the human transcriptome (RefSeq GCA_000001405.1 build) using the same STAR parameters to generate gene expression data. Next, we marked duplicates using the MarkDuplicates function from GATK (v.4.1.7.0)[55]. Aligned reads were quality checked using QoRTs (v.1.3.6)[56] to assess RNA integrity. Somalier (v.0.2.7)[57] was used to detect potential instances of sample mislabelling and FastQ Screen (0.14.0)[58] was used to detect potential instances of contamination. FastQC, QoRTs and Somalier outputs were visualized using MultiQC (v.1.9)[59]. RSEM (v.1.3.3)[60] was used with default parameters to quantify gene expression from the BAM files aligned to the transcriptome. Gene expression patterns were used for further quality checking of each sample. Tumour regions with >40% of all genes with zero counts (estimated using the QoRTS output Genes_WithZeroCounts) were excluded. Additionally, samples with <20% of reads mapping to a genomic area covered by exactly one gene in a coding sequence genomic region (estimated using the QoRTS output ReadPairs_UniqueGene_CDS) were excluded. Next, RNA coverage was calculated for single nucleotide variants (SNVs) detected in matched whole-exome sequencing data per tumour region using SAMtools (v.1.9)[61] mpileup. Mutation expression was used to further quality check the mapping of RNA reads. The expression of SNVs exclusive to a given tumour region was used to detect potential instances of within-patient mislabelling of RNA–DNA matched tumour regions as well as to exclude normal adjacent lung tissue regions that expressed mutations present in paired tumour regions. A similar approach was applied to germline SNPs to further assess potential sample swaps based on patterns of CN variation from matched DNA per tumour region. Tumour regions in which fewer than 10 mutations, or fewer than 25% of the total mutation count, had evidence of expression, and/or less than 10% of SNPs had evidence of biallelic expression, were excluded. Finally, tumour regions clustering with tumour-adjacent normal tissue regions (see the section 'UMAP clustering') and tumour regions with a low purity were also excluded from further analyses. To ensure the reproducibility and portability of the above pipeline, all steps described were implemented through the Nextflow (v.20.07.1)[62] pipeline manager.

Analyses

Unless explicitly specified otherwise, all Wilcoxon tests performed in this work are two-sided, using the function wilcox.test() in base R. To account for the effect of each individual tumour when comparing tumour regions in the cohort, we use linear mixed-effects models throughout the manuscript. These were fitted using the package lmerTest (v.3.1-3)[63] in R, using the parent tumour from which the tumour region was derived as a random effect. Significance was obtained comparing a null model with the model containing the variable of interest using the base R function anova(), and setting refit = TRUE to test the impact of fixed effects.

Unless stated otherwise, plots were generated in the R environment (v.3.6.3), using ggplot2 (v.3.2.1)[64], ggpubr (v.0.4.0), cowplot (v.1.0.0), scales(v.1.0.0) and ggrepel (v.0.8.1).

Gene expression distance. RSEM raw read counts were first normalized using the median of ratios method implemented in DESeq2 (v.1.24.0)[65]. Genes with more than 5 read counts in at least 20% of the cohort (a total of 20,136 genes) were kept after filtering. Variant stabilizing transformation (VST) was performed on the normalized reads. Distance correlation was calculated for each sample to all other samples in the cohort for the top 500 most variable genes using the dcor() function from the R package energy (v.11.7-6)[66]. In total, 500 was chosen as the number of variable genes to keep, as previous tests showed that this number represented the variance in the cohort while reducing the computational resources needed to calculate a cohort wide correlation distance. The correlation distance provides a measure from 0 (no similarity) to 1 (maximum similarity); to transform this metric into distance, we subtracted the resulting distance correlation from 1. Primary tumour regions were then clustered on the basis of the minimum distance to other samples across the cohort.

UMAP clustering. VST counts from all samples in the cohort were used to generate a UMAP[67] of expression patterns across the cohort. UMAP was performed using the umap (v.2.7.7.0) package in R with default parameters.

LUAD drivers within non-LUAD NSCLCs. Samples that were considered to fall within the LUAD cluster (UMAP 1<2.5 and UMAP 2>0) were evaluated for an enrichment in driver mutations more commonly associated with LUADs.

Differential mutation analysis was conducted to establish driver alterations that were enriched in LUADs relative to non-LUAD NSCLCs. For each driver mutation (a total of 266 genes were considered; see the section ‘Epigenetic drivers’ and a companion manuscript[7] for how these mutations were annotated), a Fisher’s exact test was performed comparing the numbers of non-LUAD and LUAD tumour regions that harboured the mutation with those that did not. After adjusting for repeated measures using the Benjamini–Hochberg method[68], the four genes in which driver mutations were significantly enriched among LUADs compared with non-LUADs were *KRAS*, *EGFR*, *STK11* and *RBM10*. A Fisher’s exact test was then performed to test the relative enrichment of these ‘LUAD-favoured’ events within non-LUADs clustering with LUADs in the UMAP compared with non-LUADs not clustering with LUADs in the UMAP.

Although these non-LUAD tumours had been subjected to independent histological review, we further reviewed histological features of these tumours in terms of morphological heterogeneity and immunohistochemistry staining profiles, including TTF-1, p63 and p40, CD56, synaptophysin and chromogranin, and pan-cytokeratin to confirm the histological diagnosis and to investigate the presence and extent of the adenocarcinomatous components in these non-LUAD tumours.

PCA. A PCA was performed using VST counts for tumour regions extracted at surgery, in LUAD and LUSC tumours separately. PCA was performed using the prcomp() function in base R, centring the data but not scaling, as expression data had already been scaled through VST. The PCs adding up to 30% of variance explained in LUAD and LUSC separately were subjected to further analysis.

Each PC was then linked at the tumour region level with multiple genomic and clinical features using a linear-mixed effects model accounting for the tumour from which the regions were derived and tumour purity as a covariate. PC ITH was calculated as the standard deviation for each PC divided by the median PC value. The selection of features is described in the section ‘PCA feature selection’.

RNA ASCAT. Estimates of tumour fraction were calculated using the tumour purity values from ASCAT[50]. The RNA-derived estimates of tumour fraction (referred to in the main text as tumour transcript fraction) were calculated using a modified version of ASCAT. In brief, using this approach, at SNP sites, the B allele frequency (BAF) was calculated using the non-duplicated RNA-seq reads aligning to each allele using SAMtools (v.1.9)[61] mpileup. This RNA BAF was then used as input to ASCAT instead of DNA BAF, whereas DNA-derived logR was maintained. The RNA-derived estimate of tumour fraction was the estimated tumour purity when ploidy was identical to that of the DNA.

I-TED. To estimate ITH, we focussed only on tumours with more than one region sampled (813 samples from 280 tumours). Because standard measures of heterogeneity might be affected by the number of regions available per tumour, a pairwise approach was used to estimate ITH. For each region of a tumour, 1 minus the correlation distance between VST gene expression to all other regions in the same tumour was calculated for the top 500 most variable genes. The correlation distance was calculated using the function dcor() in the R package energy (v.1.7-6)[66]. Only genes with read counts above 5 in at least 20% of the cohort (a total of 20,136 genes) were considered for this analysis. Tumour-level I-TED was calculated as the median I-TED of all regions. This metric was independent of the number of regions sampled (Fig. 1c).

The relationship between I-TED and purity, CN and mutation heterogeneity as well as histological subtype and number of regions per tumour was tested using a multivariable linear regression. The percentage of variance explained by each type of alteration was calculated using the Anova function from the R package car (v.3.0-6)[69].

Mutational and CN heterogeneity were calculated based on metrics from a companion manuscript[7]. In brief, for CN heterogeneity, the total proportion of region-specific CN events compared with the total number of CN events was determined. For mutation heterogeneity, the proportion of subclonal mutations at the tumour level was obtained compared with the total number of mutations per tumour. Both metrics were bootstrapped by resampling to account for differences in the number of regions samples per tumour.

Differential gene expression and gene set enrichment analyses. All differential gene expression and subsequent gene set enrichment analyses (GSEAs) were performed using the following approach. First, trimmed mean of *M*-values normalization from the edgeR (v.3.26.5)[70] R package was performed on RSEM raw counts. Genes with expression below 30 counts per million in <70% of the smallest group size were removed using the function filterByExpr() with min.count set to 30. Expression differences were performed at the region level through the limma-voom analytical pipeline, taking tumour as a blocking factor, by performing within-tumour expression correlations and including them within the voom model estimate using the duplicateCorrelation() function. This method is analogous to using tumour as a random effect in a linear mixed-effects model. The raw *P* values provided by limma for differential expression were then corrected for multiple testing using the Benjamini–Hochberg (FDR) method[68].

The *t*-statistic generated by limma was used as input for GSEA for MSigDB hallmark gene sets[14] using the R package fgsea (v.1.10.1)[71] with default parameters.

ASE analysis. To understand patterns of ASE in the TRACERx cohort, we focused on genes containing an expressed heterozygous SNP and quantified the number of unique reads aligning to each parental allele with a minimum mapping quality of 0 and a minimum base quality score of 13 using SAMtools (v.1.9)[61] mpileup. Only SNPs with a total coverage of greater than eight such reads were considered to be expressed. Further filtering removed SNPs in blacklisted regions of the genome with poor mappability. The blacklisted genomic regions were obtained from

UCSC Genome Table Browser and include regions excluded from the ENCODE project (both DAC and Duke list), simple repeats, segmental duplications and microsatellite regions[72].

In accordance with the expected distribution of allelic expression, a beta-binomial test was used to test for ASE[73] using the pbetabinom function from the R package VGAM (v.1.1-2)[74] and an overdispersion parameter σ of 0.05. We attribute allele-specific RNA reads to the major and minor alleles, as inferred from multiregion DNA-derived allele-specific CN data. Specifically, for each SNP, the allele with the greatest number of reads in the corresponding whole-exome DNA sequencing (DNA-seq) data was considered the major allele. RNA-seq reads reporting this allele were designated major allele RNA reads and vice versa for the minor allele RNA reads.

To assess the probability of obtaining allele-specific read counts at least as disparate as the observed distribution, given an expected allelic expression ratio of 0.5, the following beta-binomial (Betabin) test was performed with the following parameters:

$$\mathrm{Betabin}(X \geq m, t, 0.5) \tag{1}$$

where m represents the major allele RNA reads and t the total RNA reads at that heterozygous SNP. To alleviate the multiple testing burden and preserve statistical power, we performed independent filtering by using the following binomial test criterion and retaining only those heterozygous SNPs for which $P < 0.001$:

$$\mathrm{Bin}(X \geq t, t, \mathrm{CPNratio}); P < 0.001 \tag{2}$$

where t is the total RNA reads at that heterozygous SNP and CPNratio is the raw major allele copy number divided by the total copy number at that site. In effect, this removes sites with low read counts and/or extreme CN ratios such as regions with loss of heterozygosity or high-level allele-specific amplifications.

To test whether ASE was CN-dependent or CN-independent, two further beta-binomial tests were performed using the following parameters:

$$\mathrm{Betabin}(X < m, t, 0.5) \tag{3}$$

$$\mathrm{Betabin}(X \geq m, t, \mathrm{CPNratio}) \tag{4}$$

where m again represents the major allele RNA-seq read count, t the total RNA reads at that heterozygous SNP, and CPNratio the raw local major allele CN divided by the total CN. Following this, two combined P values were generated from all SNPs within each gene using the Fisher method: one (A) using the P value from test equation (1); and the second (B) using the smallest P value from either test equation (3) or equation (4). The Benjamini–Hochberg approach was used to adjust for multiple hypothesis testing across all genes considered. Genes with an adjusted P value (FDR) < 0.05 from test equation (1) but not either equation (3) or (4) were considered to show CN-dependent ASE, whereas those with an adjusted P value < 0.05 from either test equation (3) or (4) were considered to show CN-independent ASE. The adjusted P value threshold of <0.05 for either one of two one-tailed tests was chosen given the stringency of this approach to investigate CN-independent ASE.

The RNA allelic ratio can vary between 0.5 and the CPNratio, given the tumour allele-specific CN, owing to expression levels in the tumour and admixed non-tumour cells. This approach therefore in effect tests for ASE beyond that which would be seen given expression only from the tumour, or non-tumour component, of the bulk sample, accounting for the estimated CN status of the gene of interest within the tumour.

Combining point mutation and insertion–deletion calls within each gene and their corresponding gene ASE classifications, we computed the Fisher exact test statistics of the odds of observing each mutation type listed at ASE versus non-ASE genes, within genes in which we were powered to detect ASE (that is, containing an expressed heterozygous SNP).

We also considered the potential impact of reference bias, whereby heterozygous sites are incorrectly assigned as being homozygous as a result of the use of a generic reference genome, potentially resulting in false-positive ASE calls, on our results. To account for this, we quantified the extent to which reference bias was present in each sample. For each sample, the total instances of CN-independent ASE in which the reference allele was overexpressed relative to the alternative allele was divided by the total instances in which the reverse was true. To ensure this phenomenon was not affecting our results, we tested the impact of adding this per-sample quantification of reference bias as an additional covariate to the linear mixed-effects models within Fig. 2d,e. The statistical associations presented in those figures remained consistent after this additional test.

ASM analysis. To evaluate the relationship between ASM and expression, we leveraged previously published CAMDAC pure tumour methylation rates, m_t, derived from multiregion bulk tumour and adjacent normal RRBS performed on a subset of TRACERx samples[21]. We subset these methylomes to promoter-associated CpGs in CpG islands with read depths of ≥10 in the adjacent-matched normal sample and at the promoters of genes with CN-dependent or CN-independent ASE information in at least one sample. In total, 11,254 genes met these criteria. On average, 4,345 and 2,771 genes had promoter methylation information and could be tested for CN-dependent and CN-independent ASE per sample, respectively.

For each sample, genes were classified with respect to their CN-dependent and CN-independent ASE test statistics. Genes with $P < 0.05$ were deemed as ASE and those with $P > 0.5$ as not significantly ASE. In the case of CN-dependent ASE, genes were required to show no significant ASE, irrespective of CN, to be categorized as not significantly ASE. Genes with no phasing information were not tested for ASE.

Previously reported findings[21] indicated that intermediate m_t signals are in large part due to subclonal ASM. We therefore used intermediate CAMDAC m_t values as a proxy for ASM. CpGs with methylation rates 99% highest density intervals (HDI^{99}) $\subseteq$ [0.15,0.75] and point estimates ϵ [0.2, 0.7] were deemed ASM. We required three consecutive ASM loci to classify a gene promoter as ASM.

Combining promoter ASM and corresponding gene ASE classifications, we computed Fisher's exact test statistics of the odds of observing ASM at ASE versus non-ASE genes, within genes in which we were powered to detect ASE (that is, containing an expressed heterozygous SNP).

Tumour–normal differential methylation analysis. For a subset of TRACERx samples with previously published RRBS data, we obtained a list of tumour–normal differentially methylated positions based on CAMDAC m_t values and using the tumour-adjacent normal methylation rate as a proxy for the cell of origin, m_n ($P < 0.01$ and $|m_t - m_n| > 0.2$). For each of these, we computed the number of CpGs that were significantly hypomethylated and hypermethylated in tumour samples compared to the normal samples, taking only loci that had coverage in all samples ($\min_{\mathrm{normal}} = 10$, $\min_{\mathrm{tumour}} = 3$). We then calculated the fraction of differentially methylated positions that were hypomethylated. Using a linear mixed effects model, with tumour identity as random effect, we then compared this metric to the percentage of genes showing evidence of CN-independent ASE per sample (separately for LUAD and LUSC).

ASE PCA and imputation. PCA was performed to test for differences in patterns of CN-independent ASE between tumour and normal tissue samples. Only genes in which it was possible to test for CN-independent ASE (that is, having an expressed SNP that was not in a region of extreme CN) in at least 100 samples across the cohort were considered. The negative natural logarithm of the FDR from the test for CN-independent

ASE in the section 'ASE analysis' for each gene was computed. In samples for which it was not possible to test for CN-independent ASE for a given gene, the median negative natural logarithm for that gene in all tumour and normal tissue samples across the cohort was imputed. Data were scaled and centred, and PCA performed using the function prcomp() in base R.

Epigenetic drivers. A list of epigenetic modifier genes was obtained from previous work[22]. We collated a cancer gene list out of all genes identified in the COSMIC cancer gene census (v.75)[75], supplemented with those identified in large-scale pan-cancer analyses (using FDR < 0.05 as cut-off)[76] and previous large-scale NSCLC sequencing studies[77–79] (this list is also utilized in a companion paper[7]). Genes overlapping the epigenetic modifier and cancer gene lists (see the companion paper[7] for the definition of cancer genes in our cohort) were considered. Any non-silent variant located within one of these genes underwent further categorization on the basis of the following criteria: if the mutation was found to be deleterious (either a stop-gain or predicted deleterious in two of the three computational approaches applied (Sift[80], Polyphen[81] and MutationTaster[82])) and the gene was annotated as being recessive by COSMIC (tumour suppressor), the variant was classified as a driver mutation. Also, if the gene was annotated as being dominant (oncogene) by COSMIC and we could identify ≥3 exact matches of the specific variant in COSMIC, it was classified as a putative driver mutation. Frequently mutated genes, containing more than five putative driver mutations across the cohort, were incorporated into the below model.

A univariable linear-mixed effects model using region-level mutation and CN-independent ASE data was run to establish the effect of driver mutations in individual genes on the proportion of genes tested showing CN-independent ASE. *P* values were adjusted for repeated measures using the Benjamini–Hochberg method[68]. Independent predictors were defined using a multivariable linear-mixed effects model, using all epigenetic modifiers taking the tumour containing the region as a random factor.

Validation of link between *SETD2*-inactivating mutation and CN-independent ASE with cell line data. A literature search was conducted to find publicly available cell line data with which it would be possible to test the impact of *SETD2*, *KDM5C* or *KMT2B* knockdown or knockout after ASE in an isogenic human setting.

Three separate relevant publications[23–25] were identified for *SETD2*, only one for *KDM5C*[83] and none for *KMT2B*. Therefore, we proceeded to focus solely on the impact of *SETD2* on CN-independent ASE. This was done in lung cells (H1650; three biological replicates with shRNA knockdown[23]), kidney cells (786-0; single replicate with ZFN knockout[24]) and liver cells (HepG2; single replicate with CRISPR-mediated knockout[25]).

In each study, DNA-seq was not performed alongside RNA-seq, and it was therefore not possible to obtain a highly accurate and contemporaneous record of heterozygous sites across the genome (used to measure ASE) and CN events. Conceivably, both of these sources of variation might fluctuate with passaging. A proxy was therefore sought: SNPs and SCNAs listed within the Cancer Cell Line Encyclopaedia[84] (in the case of H1650 and 786-0); or within another publication (in the case of HepG2, SNPs were derived from variant calling of whole-genome sequencing data[85]). To mitigate the possibility of false-positive ASE calls arising from inaccurate genotyping and subsequent misclassifying of homozygous sites as heterozygous, a site was only considered as heterozygous if it was both annotated as such in the relevant DNA-seq study while also harbouring at least one expressed read from both alleles within the RNA-seq data. At sites of DNA allelic imbalance, the allele with the majority of available RNA reads assigned to it was considered likely to be the major allele. With this record of heterozygous sites, analysis of CN-independent ASE was performed in the same way as described in section 'ASE analysis'. Finally, the impact of *SETD2* on CN-independent ASE was evaluated using a linear mixed-effects model, with the study added as a random factor to control for the additional biological replicates within ref. [23].

Detection of RNA variants. RNA-specific variants were called using the somatic variant caller Mutect2 from GATK (v.4.1.7.0)[55,86]. Each BAM file was first pre-processed following GATK's best practices for RNA-variant calling. In brief, marked duplicated reads were removed and splice junctions split followed by a base quality recalibration to ensure the compatibility of the mapped reads with GATK's variant callers. The somatic variant caller Mutect2 was then run to generate raw putative RNA variant calls in exonic regions only, integrating information from multiple regions per tumour, using the multiple-sample mode. To filter germline variants, a blood DNA sample was added as a 'normal' region per tumour, along with GATK's panel of normal samples based on DNA-seq from 4,136 normal samples from The Cancer Genome Atlas and Genome Aggregation Database (gnomAD) sites. The option-tumour-lod-to-emit was set to 2.0 to ensure a maximum number of raw calls. Variant calls were run per chromosome in parallel using GNU parallel (v.20210422)[87]. FilterMutectCalls was then used to filter the raw calls with the additional option -read-filter NotSupplementaryAlignmentReadFilter to exclude variants supported exclusively by supplementary reads. After this first filtering step, BCFtools (v1.10.2)[88] was run to select only PASS biallelic SNVs.

Next, bam-readcount (v.0.8)[89] was used to obtain RNA reads with a base and mapping quality above 20 supporting the variants called by Mutect2 as an orthogonal measure of variant calls at these sites. On the basis of the bam-readcount output, the following criteria were applied to remove variants: variants with fewer than 30 reads in the germline DNA; with fewer than 30 reads in total for all DNA tumour regions; with an RNA coverage below 10 reads; for which the alternative base was supported by fewer than 3 reads; or present at less than 1% variant allele frequency. Additionally, further filtering was applied to variants in regions of the genome with poor mappability such as centromeres, repetitive regions, genomic regions with high nucleotide variability in the sample. These included blacklisted genomic regions obtained from UCSC Genome Table Browser, excluded from the Encode project (both DAC and Duke list)[72] as well as regions coding for immunoglobulin antibodies in hg19: chromosome 14, positions beyond 106000000, chromosome 22 between positions 22385572 and 23265082, and chromosome 2 between positions 132032200 and 133174000. In positions at which the RNA variant was supported by one or more reads from the DNA tumour samples, support in the DNA might arise from sequencing errors in very high coverage regions. To distinguish between this scenario and expressed mutations, a one-tailed Fisher's test was performed comparing the number of DNA reads supporting the RNA variant to the number of reads supporting other variants compared to the total coverage at the same position. If the number of DNA reads supporting the RNA variant was distinguishable from sequencing noise (with a non-stringent *P* value threshold of 0.1), the RNA variant was excluded. Furthermore, variants flanked by the same four nucleotides as either the reference or alternative allele were also excluded.

Because the libraries used for RNA-seq were stranded, variant reads from each strand were compared to obtain the difference in strandedness relative to the total depth at the variant position.

Additionally, we selected ten putative editing events for Sanger sequencing, all of which were validated with this orthogonal method (Extended Data Table 2).

RNA-SBS signatures. The R package hdp (v.0.1.5)[90], available on GitHub (https://github.com/nicolaroberts/hdp), was used to call de novo RNA-SBS signatures using default parameters and 15 iterations using the trinucleotide context of strand-independent variants (192 possibilities in total). To prevent sample size biases for which RNA variants present in all regions in highly sampled tumours could be artificially over-represented during de novo signature calling, we ran this step using unique RNA variants across all samples per patient. Only de novo signatures with significant exposure (as determined by hdp) in at least 1% of the cohort were considered for further analyses. A signature present in only one patient was therefore discarded for downstream analyses. deconstructSigs (v1.9.0)[91] was then applied to each individual sample to estimate RNA-SBS signatures per sample. Only tumour regions with more than 20 RNA variants were considered for further analyses.

To test the potential relationship between signature activity and the expression of specific genes, we performed a linear mixed-effects model using the number of RNA variants attributed to each signature as the dependent variable, and gene expression of all genes in our dataset ($n = 20{,}136$). Expression was measured as $\log_{10}$(transcripts per million + 1) for genes with at least 5 read counts in 20% of the RNA-seq cohort.

To further test the relationship between *APOBEC* expression and RNA-SBS2, a linear mixed-effects model was performed, using the number of RNA variants attributed to RNA-SBS2 as the dependent variable, the expression of all *APOBEC* genes in the transcriptome as explanatory variables and the tumour identifier as a random effect.

Detection of RNA loops. RNA loops were detected in the flanking regions of RNA variants. Flanking regions were derived using the flanks() function from the R package GenomicRanges (v.1.36.0)[92]. Loops were defined by the flanking regions 3′ and 5′ of a 3–5-nucleotide-long sequence containing an RNA variant being complementary at a length of at least 3 nucleotides.

RNA-editing motif enrichment. To confirm the role of specific *APOBEC* enzymes in RNA-SBS2, we tested the relative proportions of C>T events at known RNA-editing *APOBEC* motifs. *APOBEC* enzymes typically edit C>T variants at the fourth position of 4-nucleotide-long RNA hairpin loops. In particular, *APOBEC3A* favours the CAT[C>T] motif[30,31].

APOBEC motif enrichment analyses were performed based on a previously reported local enrichment method[93]. In brief, for each C>T variant site, a Fisher's test was performed to test whether C>T changes within 20 upstream or downstream nucleotides occurred more than expected by chance at specific motifs (CAT[C>T]) in either strand.

ITH of CN-independent ASE. The ITH of CN-independent ASE was calculated for each tumour as follows. The total number of genes in a tumour showing CN-independent ASE in all of two or more tumour regions was divided by the total number of genes in that tumour showing CN-independent ASE in at least two regions. A gene showed homogeneous CN-independent ASE if it was detected in all regions of a tumour for which it was possible to test as outlined in the section 'ASE analysis'.

The relationship between I-TED and the ITH of other forms of alterations was tested using a multivariable linear regression in a similar fashion as that detailed in the section 'I-TED'.

dN/dS analysis. The dndscv function in R from the dNdScv package (v.0.1.0.0)[16] was run on all mutations available in the cohort. The function genesetdnds() was then run on the resulting object on various subsets of gene lists divided by expression quantiles for ITH, intertumour heterogeneity or amplitude. This ensured that the global dN/dS metrics obtained for each group were based on the same mutational background, making them more comparable.

Expression amplitude was measured as VST counts, whereas ITH was measured as the standard deviation in expression amplitude across all regions in multiregion tumours. Intertumour heterogeneity was measured as the bootstrapped (ten iterations) standard deviation in expression per gene sampling one tumour region per tumour per iteration, as in the previously described[17].

Cancer genes were defined as specified in[7] the section 'Epigenetic drivers' as well as in a companion manuscript[7]. Non-cancer genes were those not present in the pan-cancer COSMIC database (v.75)[75] or the list of cancer genes from ref. [94]. Essential genes were identified from the Project Achilles list of essential genes for NSCLC[18].

PCA feature selection. The following clinical and genomic features per primary tumour region were tested for association with the foremost PC of gene expression:

(1) Clinical features, including age of the patient, sex, years spent smoking cigarettes and TNM stage of the primary tumour at resection. See methods in a companion manuscript[7] for details on how these features were obtained.
(2) LUAD-specific subtype as defined by central pathological review (acinar, lepidic, cribriform, micropapillary, mucinous, papillary or solid). This feature was available only for LUAD tumours and is described in more detail in a companion paper[95].
(3) Tumour mutation burden: the number of mutations per region. Only mutations that are likely to have a phenotypic effect are included, in line with calculations of a harmonized tumour mutation burden[96]. These include all exonic single-nucleotide mutations, except synonymous changes, as well as insertions and deletions. All metrics below that depend on mutation numbers are based on this set of mutations.
(4) Presence or absence of driver mutations in cancer genes with a driver mutation in at least 5% of the cohort. This included driver mutations in *ARID1A*, *ATM*, *ATRX*, *CDKN2A*, *COL5A2*, *CREBBP*, *EGFR*, *FAT1*, *KEAP1*, *KMT2D*, *KRAS*, *MGA*, *NF1*, *PIK3CA*, *RBM10*, *SMARCA4*, *STK11* and *TP53*. See a companion manuscript[7] on the definition of cancer genes in our cohort.
(5) Proportion of subclonal mutations: the number of exonic mutations in the focal tumour region belonging to subclonal mutational clusters in the tumour, divided by the total number of exonic mutations in that region. Subclonal mutations were defined as those belonging to any mutation cluster with a cancer cell fraction below 1 across the tumour (that is, not present in all cells in the focal tumour). Details on how clonal clusters are determined are available from a companion manuscript[7]. This metric gives a measure of the proportion of smaller clones present in the tumour region.
(6) Genome instability at the tumour region level, a common feature in tumour evolution[38], was measured through the weighted genome instability index (wGII), which measures the extent of genome instability per tumour region. See methods in a companion manuscript[7] for details on the calculation of this index.
(7) Similarly, the number of whole genome duplication events per tumour region was also considered. See methods in a companion manuscript[7] for details on the calculation of genome doubling events.
(8) COSMIC mutational signatures SBS1, SBS2, SBS4, SBS5, SBS13 and SBS92 (ref. [34]). Signature activity was measured as the fraction of mutations per tumour region corresponding to each signature's weight. SBS2 and SBS13 were combined into a single SBS DNA signature for APOBEC activity. See methods in a companion manuscript[7] for details on the mutational signature analysis.
(9) The immune microenvironment was assessed by estimating the T cell fraction from DNA using the R package T-Cell ExTRECT(v.1.0.1)[97].

Additionally, we performed a single sample GSEA (ssGSEA) for the 50 MSigDB hallmark gene sets using the R package fgsea (v.1.10.1)[71] on VST counts using a Gaussian distribution and default parameters. The resulting enrichment scores per sample were correlated to each PC using a linear mixed-effects model that controlled for the tumour of origin. The resulting *P* values were merged by MSigDB functional group[14] (Extended Data Table 3) using the harmonic mean and corrected for multiple testing using FDR.

Classifier feature selection. The seeding region classifier was based on a cohort of regions from primary tumours that had metastasized or that had not metastasized after 3 years of follow up. Only tumour regions with a seeding clone at >0.2 CCF were considered as seeding for this analysis. In total, 516 primary tumour regions from 206 tumours for which seeding status could be established and for which all metrics tested could be measured (307 non-seeding regions, 209 seeding) were analysed. The following features were also considered for the classifier:

(1) Tumour mutation burden: the number of mutations per region. Only mutations that are likely to have a phenotypic effect are included, in line with calculations of a harmonized tumour mutation burden[96]. These include all exonic single-nucleotide mutations, except synonymous changes, as well as insertions and deletions. All metrics below that depend on mutation numbers are based on this set of mutations.

(2) Regionally truncal and clonal mutation burden: the number of clonal mutations per tumour region. Clonal mutations were defined as those belonging to a mutation cluster with a cancer cell fraction of 1 (that is, present in all cells) in the tumour region. This included mutations that are clonal in the entire tumour (trunk mutations) as well as mutations that were clonal in the focal region, but subclonal or absent in other regions of the same tumour. See methods in a companion manuscript[7] for details on the determination of the truncal cluster.

(3) Clonal illusion tumour mutation burden: the number of mutations that were clonal in the focal region, but not clonal within all other regions of the same tumour. Only mutations belonging to clusters with a cancer cell fraction of 1 (that is, present in all tumour cells) in the focal region, but not in the rest of the tumour, were counted. See methods in a companion manuscript[7] for more details on the definition of the truncal cluster.

(4) Proportion of regionally subclonal mutations: the number of mutations belonging to subclonal mutational clusters in the focal tumour region divided by the total number of mutations in that region. Subclonal mutations were defined as those belonging to any mutation cluster with a cancer cell fraction below 1 (that is, not present in all cells in the focal tumour region). Details on how clonal clusters were determined are available in a companion paper[7]. This metric gives a measure of the proportion of smaller clones present in the tumour region.

(5) Proportion of expressed mutations: the number of expressed mutations divided by the total mutation burden in the tumour region. A mutation is considered expressed if it had at least three reads with the mutated allele in the RNA-seq data. This metric serves as a proxy for the proportion of tumour mutations that were present in the bulk RNA-seq transcripts.

(6) Number of region clonal driver mutations: the number of driver mutations that belong to clonal mutation clusters in the focal region. This included both truncal mutations (that is, clonal across the tumour) and clonal illusion mutations (that is, clonal in the focal region but not in the rest of tumour regions). Details on how driver mutations and clonal clusters were determined are available in a companion manuscript[7].

(7) Number of region subclonal driver mutations: the number of driver mutations that belong to subclonal mutation clusters (cancer cell fraction below 1) in the focal region. Details on how driver mutations and clonal clusters were determined are available in a companion manuscript[7].

(8) Presence or absence of driver mutations in cancer genes that contained a driver mutation in at least 10% of the cohort. This included driver mutations in *CDKN2A*, *KEAP1*, *KMT2D*, *KRAS*, *SMARCA4*, *STK11* and *TP53*.

(9) Presence or absence of CN drivers, that is, amplification of a subset of oncogenes or the homozygous loss of a subset of tumour suppressor genes. Genes with at least copy number driver alterations in 10% of the cohort were included. These include *SOX2, TERT, TERC, CDKN2A, MYC, CCND1, FGFR1, NKX2-1, AKT2, EGFR* and *CCNE1*. Details on how driver mutations and clonal clusters are determined are available in a companion paper[7].

(10) COSMIC mutational signatures SBS1, SBS2, SBS4, SBS5, SBS13 and SBS92 (ref. [34]). Signature activity was measured as the fraction of mutations per tumour region corresponding to each signature's weight. SBS2 and SBS13 were combined into a single SBS DNA signature for APOBEC activity. Details on how mutational signatures were extracted are available in a companion manuscript[7].

(11) Genome instability at the tumour region level, a common feature in tumour evolution[38], was measured through the wGII, which measures the extent of genome instability per tumour region. Details on how this metric was calculated per tumour region are available in a companion paper[7].

(12) Similarly, the number of genome-doubling events per tumour region was also considered. Details on how genome-doubling events per tumour region were calculated are available in a companion paper[7].

(13) In a companion paper[7], the presence of expanded subclones in a tumour were taken as evidence of recent subclonal sweeps and linked to poor prognosis. To test the potential impact of this metric on seeding potential, per tumour region, we calculated the maximum cancer cell fraction of all mutation clusters on terminal nodes of the phylogenetic tree as a measure of clone dominance. A higher cancer cell fraction indicates a larger terminal mutation cluster in the focal region. Details on how driver mutations and clonal clusters are determined are available in the companion paper[7].

(14) To measure the impact of expression diversity within a tumour, we included the per tumour region I-TED score. I-TED was imputed as the median score across the cohort for samples for which only one region per tumour was available.

(15) To characterize the phenotype of the tumour region, we measured the tumour-region enrichment score through ssGSEA (using the R package fgsea (v.1.10.1)[71] using a Gaussian distribution and default parameters on VST counts) for three cancer-specific gene sets: (1) CIN70 (ref. [36]): an expression signature linked to genome instability and cell proliferation, phenotypes that have been linked to poor prognosis and metastasis[98]; (2) Oracle[17]: a lung cancer-specific prognostic maker, in which increased expression of this gene set is linked to poor prognosis; (3) a high-plasticity cell state: an expression signature for phenotypic plasticity extracted from the recent publication from ref. [5]. biomaRt (v.2.40.1)[99] 1:1 orthologues between human and mice genes from cluster 5 were used to calculate this signature, as described in the publication.

(16) The tumour microenvironment was characterized using expression markers consistent with previously described immune cell types[100].

(17) Additionally, tumour purity as calculated using ASCAT (v.2.3)[50] was included.

(18) To test the potential effect of overall tumour-specific expression in the metastatic potential, we added the differential between the transcript tumour fraction (described in section 'RNA ASCAT') and tumour purity as calculated using ASCAT from DNA.

(19) We tested the potential impact of CN-independent ASE on the metastatic potential of tumour regions by including the proportion of genes with CN-independent ASE compared with the total number of genes for which ASE could be measured per tumour region.

(20) We also included the ASE status of genes with significant enrichment in CN-independent ASE in tumours compared to tumour-adjacent normal lung tissue. These included *CSN2KA3*, *DNAH11*, *DOCK1*, *GALNT18*, *NLRP2*, *PRIM2* and *ZNF597*. In cases when ASE could not be measured in a tumour region, the ASE status was encoded as unknown to prevent missing values.

(21) The potential impact of RNA editing on seeding potential was included through the two RNA-editing signatures characterized in this paper: RNA-SBS1 (ADAR) and RNA-SBS2 (APOBEC3A). Their activity was measured as the fraction of RNA variants per tumour region corresponding to each signature's weight.

(22) We additionally included the RNA-editing levels (fraction of RNA molecules with edited sites) of three genes reported to play a role in cancer development from the literature[101–103]: *AZIN1*, *COPA* and *COG3*. This feature was added only for tumour regions with at least 30 unique RNA reads covering the editing sites of interest.

Classifier to predict seeding and non-seeding tumour regions. We built the machine-learning framework in Python using Tensorflow (v.2.6.0)[104] and sklearn (v.0.0)[105]. Specifically, we built an ensemble classifier that used three different model types: (1) logistic regression, (2) random forest and (3) multilayer perceptron with support vector machine embedded in the final layer. We describe the structure of the machine-learning pipeline in more detail below.

Pre-processing. To pre-process the input data, we first explored the correlation structure among potential explanatory features ($n = 61$) and removed those features with high correlation coefficients ($r > 0.75$, $n = 11$). We one-hot-encoded categorical features using get_dummies from Pandas (v1.3.3)[106] and then split the data into training and test datasets (75/25 split). After encoding, we had a total of 60 features. We scaled the continuous features using MinMaxScaler from sklearn.preprocessing (v.0.0)[107] and used SMOTENC from imblearn.over_sampling (v.0.8.0)[105] to improve the balance of the dataset in terms of numbers of seeding and non-seeding regions. Finally, we used the sklearn (v.0.0)[105] framework to perform additional variable selection before training using a LinearSVC model (penalty = "l1"), keeping those features with importance ≥0.015. This threshold removed 15 out of 60 features. Following this initial pre-processing, we generated different subsets of the dataset depending on the source of the input features, thus downstream processes within the pipeline operated on three datasets: (1) genomic only features, (2) transcriptomic only features, and (3) all features.

Model training. For each model type, to tune model hyperparameters, we performed a randomized grid search with RandomizedSearchCV (sklearn.model_selection, v.0.0)[104] and StratifiedKFold cross-validation (n_splits = 10, n_iter = 500).

We implemented a sequential model from tensorflow.keras (v.2.6.0)[104] with dropout layers (dropout = 0.2) to reduce overfitting and used a categorical hinge loss function with an l2 kernel regularizer and sigmoid activation function in the final layer. This approach effectively constitutes a support vector machine in the final layer of the sequential model. We used the Adam optimizer from tensorflow.keras.optimizers (v.2.6.0)[104]. Specifically, we defined a search grid to tune the following parameters: learning rate, batch size, epochs, number of hidden layers and sizes of hidden layers. Following the cross-validated training across the randomized search grid, we selected the best performing model according to the greatest balanced accuracy. We then extracted feature weights from this selected model using PermutationImportance from eli5.sklearn (v.0.11.0). Finally, to assess the performance of the selected model on the held-out test dataset, we used the model to predict whether a test region was seeding or not and compared this to the true labels. The machine-learning pipeline was developed using Python (v.3.5.5), and plots of results were generated in R (v.4.0.3) using ggplot2 (v.3.3.5).

Reporting summary

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

Data availability

The RNA-seq, whole-exome sequencing and RRBS data (in each case from the TRACERx study) used during this study have been deposited at the European Genome–phenome Archive, which is hosted by the European Bioinformatics Institute and the Centre for Genomic Regulation, under the accession codes EGAS00001006517 (RNA-seq), EGAS00001006494 (whole-exome sequencing) and EGAS00001006523 (RRBS). Access is controlled by the TRACERx data access committee. Details on how to apply for access are available at the linked page.

Code availability

Code used to process data and generate figures is available at the following link: https://doi.org/10.5281/zenodo.7603386.

51. Andrews, S. et al. FastQC v0.11.2 (Babraham Bioinformatics, 2010).
52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J* **17**, 10–12 (2011).
53. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
54. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. McKenna, A. et al. The Genome Analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
56. Hartley, S. W. & Mullikin, J. C. QoRTs: a comprehensive toolset for quality control and data processing of RNA-seq experiments. *BMC Bioinformatics* **16**, 224 (2015).
57. Pedersen, B. S. et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
58. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* **7**, 1338 (2018).
59. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
60. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
61. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
63. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* https://doi.org/10.18637/jss.v082.i13 (2017).
64. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016).
65. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
66. Rizzo, M. L. & Szekely, G. J. energy: E-Statistics: multivariate inference via the energy of data. R package version 1.7-0 (2017).
67. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1802.03426 (2018).
68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
69. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (SAGE Publications, 2018).
70. 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).
71. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. Preprint at *bioRxiv* https://doi.org/10.1101/060012 (2016).
72. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
73. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21**, 1728–1737 (2011).
74. Yee, T. W., Stoklosa, J. & Huggins, R. M. The VGAM package for capture-recapture data using the conditional likelihood. *J. Stat. Softw.* **65**, 1–33 (2015).
75. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids. Res.* **43**, D805–D811 (2015).
76. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
77. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
78. Collisson, E. et al. Comprehensive molecular profiling of lung adenocarcinoma: the Cancer Genome Atlas Research Network. *Nature* **511**, 543–550 (2014).
79. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
80. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
81. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **7**, Unit 7.20 (2013).
82. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
83. Shen, H. F. et al. The dual function of KDM5C in both gene transcriptional activation and repression promotes breast cancer cell growth and tumorigenesis. *Adv. Sci.* **8**, 2004635 (2021).
84. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

85. Zhou, B. et al. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res.* **47**, 3846–3861 (2019).

86. Benjamin, D. et al. Calling somatic SNVs and indels with Mutect2. Preprint at *bioRxiv* https://doi.org/10.1101/861054 (2019).

87. Tange, O. GNU Parallel 2018. *Zenodo* https://doi.org/10.5281/zenodo.1146014 (2018).

88. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

89. Khanna, A. et al. Bam-readcount–rapid generation of basepair-resolution sequence metrics. Preprint at https://arxiv.org/abs/2107.12817 (2021).

90. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

91. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

92. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

93. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).

94. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).

95. Karasaki, T. et al. Evolutionary characterization of lung adenocarcinoma morphology in TRACERx. *Nat. Med.* https://doi.org/10.1038/s41591-023-02230-w (2023).

96. Merino, D. M. et al. Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project. *J. Immunother. Cancer* **8**, e000147 (2020).

97. Bentham, R. et al. Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature* **597**, 555–560 (2021).

98. Caswell, D. R. & Swanton, C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Med.* **15**, 133 (2017).

99. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

100. Danaher, P. et al. Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* **5**, 18 (2017).

101. Han, L. et al. The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell* **28**, 515–528 (2015).

102. Zhang, M. et al. RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat. Commun.* **9**, 3919 (2018).

103. Chen, L. et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat. Med.* **19**, 209–216 (2013).

104. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1603.04467 (2016).

105. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 559–563 (2017).

106. McKinney, W. Data structures for statistical computing in python. In *Proc. 9th Python in Science Conf.* (eds van der Walt, S. & Millman, J.) 51–56 (SciPy, 2010).

107. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements The TRACERx study (Clinicaltrials.gov number NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent research ethics committee (13/LO/1546). TRACERx is funded by Cancer Research UK (CRUK) (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre, which has a core grant from CRUK (C444/A15953). We gratefully acknowledge the patients and relatives who participated in the TRACERx study; all site personnel, investigators, funders and industry partners who supported the generation of the data within this study; and staff at the Scientific Computing, the Advanced Sequencing Facility and Experimental Histopathology departments at the Francis Crick Institute. This work was supported by the Cancer Research UK Lung Cancer Centre of Excellence and the CRUK City of London Centre Award (C7893/A26233), as well as the UCL Experimental Cancer Medicine Centre. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (grant number MR/L016311/1). For the generation of the RRBS data, we acknowledge technical support from the CRUK–UCL Centre-funded Genomics and Genome Engineering Core Facility of the UCL Cancer Institute and grant support from the NIHR BRC (BRC275/CN/SB/101330) and the Wellcome Trust (218274/Z/19/Z). N.M. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (grant number 211179/Z/18/Z) and receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. C.S. is a Royal Society Napier Research Professor (RSRP\R\210001). C.S. is supported by the Francis Crick Institute that receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (CC2041) and the Wellcome Trust (CC2041). C.S. is funded by Cancer Research UK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; NovoNordisk Foundation (ID16584); Royal Society Professorship Enhancement Award (RP/EA/180007); National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US); and The Mark Foundation for Cancer Research Aspire Award (Grant 21-029-ASP). This work was supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (grant number SU2C-AACR-DT23-17 to S.M. Dubinett and A.E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. is in receipt of an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 835297). C.M.-R. is supported by the Rosetrees (M630) and Wellcome trusts. J.R.M.B. is supported by Cancer Research UK. C.P. is supported by the Francis Crick Institute. M.S.H. is supported by Cancer Research UK (TRACERx (C11496/A17786)). M.J.-H. is a Cancer Research UK Career Establishment Awardee and has received funding from CRUK, IASLC International Lung Cancer Foundation, Lung Cancer Research Foundation, Rosetrees Trust, UKI NETs and NIHR University College London Hospitals Biomedical Research Centre. J.D. is a postdoctoral fellow of the European Union's Horizon 2020 research programme (Marie Skłodowska-Curie Grant Agreement No. 703594-DECODE) and the Research Foundation–Flanders (FWO 12J6916N). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. P.V.L. is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). T.K. is supported by a JSPS Overseas Research Fellowships Program (202060447). J.D., E.L.C., C.C. and P.V.L. were supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008) and the Wellcome Trust (CC2008). T.B.K.W. is supported by the Francis Crick Institute, as well as the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), the Breast Cancer Research Foundation (BCRF), a Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. M.A.B. is supported by Cancer Research UK, the Rosetrees Trust and the Francis Crick Institute. M.D. is supported by Cancer Research UK and the Lung Cancer Centre of Excellence. A. Huebner is supported by Cancer Research UK. E.L.L. receives funding from the NovoNordisk Foundation (ID 16584). N.J.B. is a fellow of the Lundbeck Foundation (R272-2017-4040) and acknowledges funding from Aarhus University Research Foundation (AUFF-E-2018-7-14) and the Novo Nordisk Foundation (NNF21OC0071483).R.R. is supported by a Royal Society Research Professorships Enhancement Award (RP/EA/180007). K.L. is funded by the UK Medical Research Council (MR/P014712/1 and MR/V033077/1), Rosetrees Trust and Cotswold Trust (A2437) and Cancer Research UK (C69256/A30194). E.C. is funded by Cancer Research UK (TRACERx (C11496/A17786)) and the Francis Crick Institute. S.d.C.T. was funded in part by a Marie Skłodowska-Curie Individual Fellowship from the European Union (MSCA-IF-2015-EF-ST 703228-iGEMMdev). A.M.F. is supported by Stand Up To Cancer (SU2C-AACR-DT23-17). A.R. is supported by the Francis Crick Institute.D.A.M. is supported the cancer Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025).

Author contributions A. Hackshaw, N.M., M.J.-H. and C.S. performed study design and conduct and clinical and laboratory oversight. C.M.-R., J.R.M.B. and C.P. performed all expression analyses. D.E.C., R.R., G.A.W., T.P.J. and M.S.H. developed the alignment and quality control (QC) pipeline (jointly led by C.P. and C.M.R.). C.M.-R. and J.R.M.B. performed the QC and curation of the RNA-seq data of the cohort (led by C.P.). M.S.H. developed and ran the ensemble machine-learning classifier. C.M.-R. and J.R.M.B. curated and collated the variables for the classifier. J.D., E.L.C. and P.V.L. developed the ASE pipeline (led by J.R.M.B.). J.R.M.B. analysed the ASE data. C.M.-R. developed the RNA variant pipeline call. K.T., A.M.A.-R., J.R.M.B. and T.P.J. analysed the RNA variant data (led by C.M-R.). A.R. performed Sanger sequencing to validate RNA variants. Fresh-frozen samples were extracted by C.N.-L. and A.R. (led by S.V.). Clinical annotation was performed and curated by T.K., D.A.M., A. Hackshaw, C.N.-L., R.S., A.T. and P.P. (led by M.A.B.). RNA-seq was performed by L.C. and F.A. (led by S.W.). RRBS was performed and curated by G.A.W. (led by M.T.). Methylation data analyses were performed by J.R.M.B., J.D., C.C., M.T., P.V.L., S.B. and N.K. (led by E.L.C.). Analyses of whole exonic data, including calling mutations, classifying driver mutations and CN analyses were performed by A.M.F., T.B.K.W., E.L.L., E.C., M.D., A. Huebner and O.P. Metastatic seeding patterns were determined by M.A.B. and A. Huebner. RAS activation groups in LUAD samples were obtained by P.E., S.d.C.T., C.M.-R. and J.R.M.B. The optimal use of the high-performance cluster from the Francis Crick Institute used for most analyses in this manuscript was enabled by M.S.H. and J.C. (led by D.L.). N.J.B., K.L. and A. Hackshaw supervised the bioinformatic and statistical analyses (led by N.M.). M.S.H., K.T., C.P., N.M. and T.K. wrote the manuscript (jointly led by J.R.M.B. and C.M.-R.). E.C., N.J.B., K.L. and C.S. gave feedback on the manuscript. M.J.-H., N.M. and C.S. jointly designed and supervised the study.

Competing interests S.V. is a co-inventor to a patent to detect molecules in a sample (US patent no. 10578620). M.A.B. has consulted for Achilles Therapeutics. A.M.F. is co-inventor to a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987). M.J-H. has consulted for, and is a member of, the Achilles Therapeutics Scientific Advisory Board and Steering Committee; has received speaker honoraria from Pfizer, Astex Pharmaceuticals and Oslo Cancer Cluster; and holds patent PCT/US2017/028013 relating to methods for lung cancer detection. This patent has been licensed to commercial entities and under terms of employment. M.J.-H. is due a share of any revenue generated from such license(s). A. Hackshaw has received fees for being a member of independent data monitoring committees for Roche-sponsored clinical trials and academic projects co-ordinated by Roche. C.S. acknowledges grant support from AstraZeneca, Boehringer-Ingelheim, Bristol Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx Inc — collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical. C.S. is an AstraZeneca Advisory Board member and chief investigator for the AZ MeRmaiD 1 and 2 clinical trials, and is also co-chief investigator of the NHS Galleri trial, funded by GRAIL, and a paid member of GRAIL's Scientific Advisory Board. He receives consultant fees from Achilles Therapeutics (where he is also a Scientific Advisory Board member), Bicycle Therapeutics (where he is also a Scientific Advisory Board member), Genentech, Medicxi, Roche Innovation Centre – Shanghai, Metabomed (until July 2022) and the Sarah Cannon Research Institute, had stock options in Apogen Biotechnologies and GRAIL until June 2021, currently has stock options in Epic Bioscience and Bicycle Therapeutics, and has stock options in and is co-founder of Achilles Therapeutics. C.S. is an inventor on a European patent application relating to assay technology to detect tumour recurrence (PCT/GB2017/053289); the patent has been licensed to commercial entities and under his terms of employment C.S is due a revenue share of any revenue generated from such license(s). C.S. holds patents relating to targeting neoantigens (PCT/EP2016/059401), identifying clinical response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA loss of heterozygosity (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), identifying patients whose cancer responds to treatment (PCT/GB2018/051912), detecting tumour mutations (PCT/US2017/28013), methods for lung cancer detection (US20190106751A1) and identifying insertion/deletion mutation targets (European and US, PCT/GB2018/051892), and is co-inventor to a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987). C.S. is a named inventor on a provisional patent protection related to a ctDNA detection algorithm. G.A.W. is employed by and has stock options in Achilles Therapeutics. N.M. has received consultancy fees and has stock options in Achilles Therapeutics. N.M. holds European patents relating to targeting neoantigens (PCT/EP2016/059401), identifying clinical response to immune checkpoint blockade (PCT/ EP2016/071471), determining HLA loss of heterozygosity (PCT/GB2018/052004) and predicting survival rates of patients with cancer (PCT/GB2020/050221). D.A.M. reports speaker fees from AstraZeneca, Eli Lilly and Takeda, consultancy fees from AstraZeneca, Thermo Fisher, Takeda, Amgen, Janssen and Eli Lilly and has received educational support from Takeda and Amgen. S.C.T. has acted as a consultant for Revolution Medicines. N.J.B. is a co-inventor to a patent to identify patients whose cancer responds to treatment (PCT/GB2018/051912), a co-inventor on a patent for methods for predicting anti-cancer response (US14/466,208) and has a patent application (PCT/GB2020/050221) on methods for cancer prognostication. R.S. reports non-financial support from Merck and Bristol Myers Squibb (BMS); research support from Merck, Puma Biotechnology and Roche; and personal fees from Roche, BMS and Exact Sciences for advisory boards. E.L.C. is employed by and has stocks in Achilles Therapeutics.

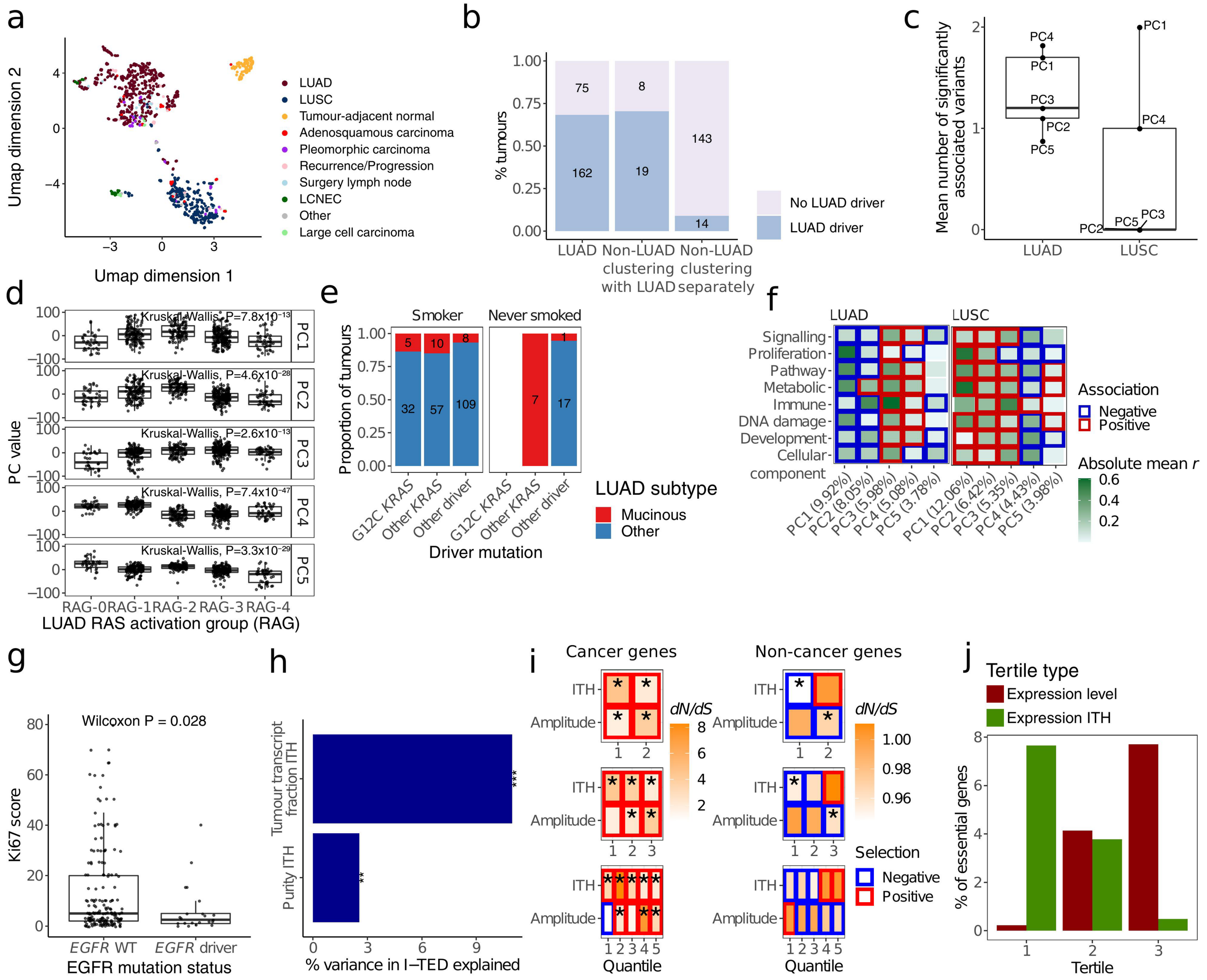
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05706-4.

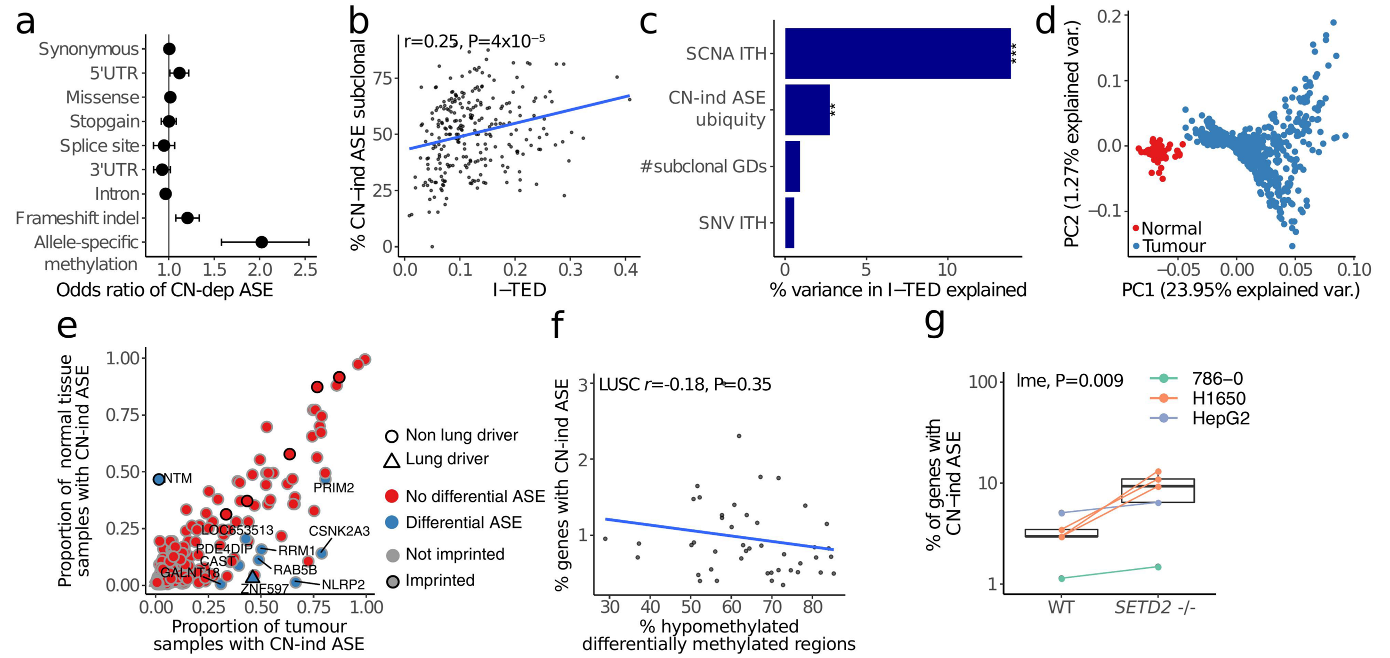
Correspondence and requests for materials should be addressed to Charles Swanton or Nicholas McGranahan.

Peer review information *Nature* thanks David Adams, Matthew Meyerson and Jüri Reimand for their contribution to the peer review of this work.

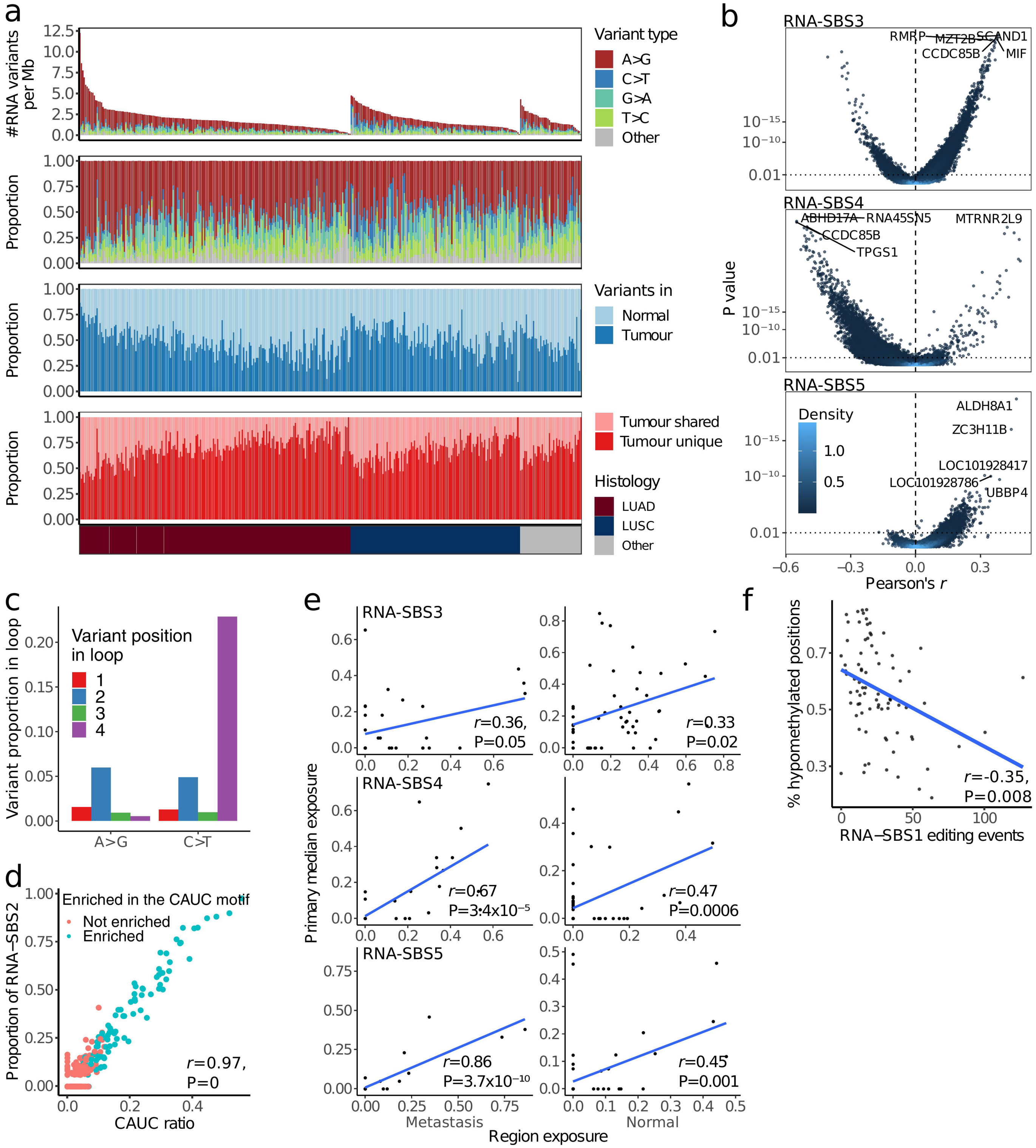
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Patterns of expression diversity in the TRACERx cohort. a. Uniform manifold approximation and projection (UMAP) showing the distribution of each primary tumour region in the cohort based on gene expression. n = 914 tumour regions collected at surgical resection from 352 primary tumours, n = 33 recurrence/relapse samples from 24 tumours and n = 96 paired normal samples from 96 tumours. LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; LCNEC: Large cell neuroendocrine carcinoma. **b**. Percentage of tumours with and without 'LUAD drivers' (driver mutations enriched in LUADs) in LUAD, non-LUADs clustering with LUADs in the UMAP and non-LUADs clustering apart from LUADs. Number of tumours within each category is annotated. **c**. Mean number of variables significantly associated with each principal component (PC) of gene expression after randomly sub-sampling the number of LUAD regions to match that of LUSC regions (n = 303) for 50 iterations. LUAD subtypes were not included in this comparison to ensure an equal number of variables between LUAD and LUSC. **d**. PC associations with each of the different RAS activation groups (RAG) developed by East and colleagues[11]. PC activity different significantly between RAGs. Analysis based on 480 tumour regions collected at surgical resection from 190 LUAD tumours where RAG could be estimated. **e**. Proportion of LUAD tumours in smokers (comprising current and ex-smokers) and never smokers, split by LUAD subtype, with either G12C KRAS driver mutations, non-G12C KRAS driver mutations or driver mutations in other genes. Numbers annotated indicate the number of tumours per category. **f**. Pearson's *r* between each PC and functional groups comprising the fifty MSigDb Hallmark gene sets[14]. Pearson's *r* values were averaged within the functional group to which each hallmark was assigned[14] across LUAD, n = 480 tumour regions from 190 tumours; and LUSC, n = 303 tumour regions from 119 tumours. The colour of the border around each square indicates the direction of the association between each covariate and PC for significant (FDR<0.05) associations. Significance was determined through a mixed effects linear model using purity as a fixed covariate and tumour as a random variable; P values were calculated by hallmark and combined within MSigDB functional group using the harmonic mean. **g**. Immuno-histochemical staining for Ki67 proliferation marker in LUAD tumours with and without *EGFR* driver mutations. Only the 196 LUAD tumours within which Ki67 was measured are displayed. Significance was calculated through a two-sided unpaired Wilcoxon test. WT: Wild type. **h**. Percentage of variance in Intra-Tumour Expression Distance (I-TED) that was explained by intra-tumour variance in tumour transcript fraction and intra-tumour variance in tumour purity, in a linear regression. Analysis based on 258 tumours with at least two primary tumour regions, and purity and tumour transcript fraction estimates. ***:P value = 5.03 × 10^{-8}; **:P value = 0.007. **i**. dN/dS in non-cancer and cancer genes for different quantiles of ITH or expression amplitude. Asterisks indicate significance whereby the 95% confidence interval of the dN/dS estimate did not overlap 1 signalling either negative (blue square) or positive (red square) selection. Broadly, lower quantiles of ITH tended towards negative selection in non-cancer genes, whereas the opposite was true for cancer genes. Results based on bootstrapping from the total number of tumour samples resected at surgery of the primary tumour from tumours with more than one sample at that time point, 845 regions from 285 tumours. **j**. Percentage of all essential genes from the Project Achilles list[18] (n = 604) in lung cancer for tertiles of expression ITH or amplitude. All box plots in this figure represent lower quartile, median and upper quartile, whiskers represent lower/higher bound +/− 1.5 x interquartile range.

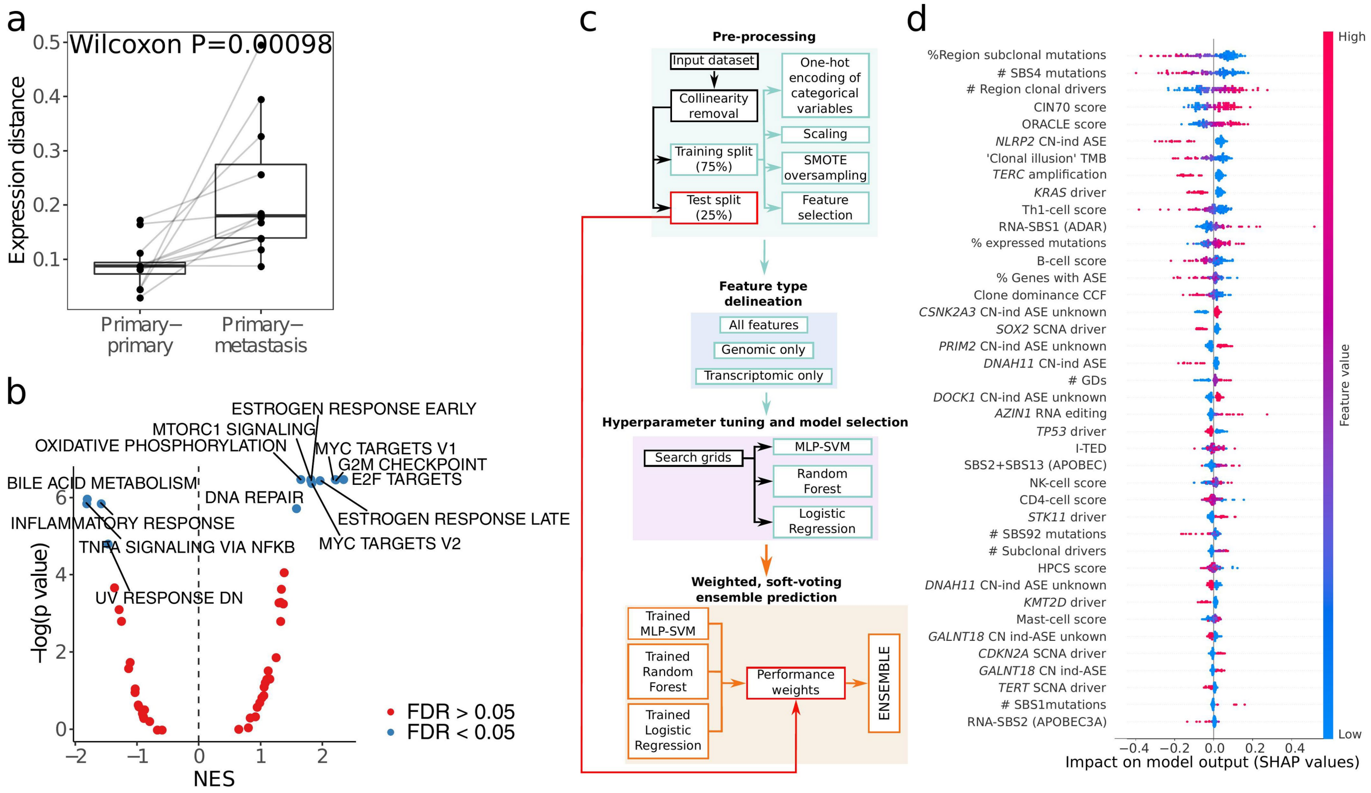


Extended Data Fig. 2 | Genomic and transcriptomic links with allele-specific expression. a. Points indicate odds ratio estimates for copy-number dependent allele-specific expression (CN-dependent ASE) when somatic point mutations, or allele-specific methylation (where both RRBS and RNA-Seq were available) were concomitantly detected in the same gene, by type of alteration. Bars indicate 95% confidence intervals. Odds ratio for the links between CN-dependent ASE and mutations; and CN-dependent ASE and ASM are based on 876 primary tumour regions from 332 tumours, and 96 tumour regions from 31 tumours, respectively. **b**. Relationship between the proportion of CN-independent ASE in a tumour that is subclonal, being found in a subset of regions within a given tumour, and intra-tumour expression diversity. The Pearson correlation coefficient is shown ($r = 0.25$, $P = 4 \times 10^{-5}$). **c**. Percentage of variation in I-TED that was explained by single nucleotide variant (SNV), SCNA and CN-independent ASE ITH, as well as the number of subclonal whole genome duplication events (GDs) per tumour. The linear regression was based on 269 tumours where all variables could be calculated. ***:$P = 2.4 \times 10^{-10}$; **:P = 0.004. **d**. PCA of CN-independent ASE patterns in TRACERx421 tumours (n = 877 tumour regions) and normal tissue (n = 95) samples where CN-independent ASE could be estimated. Samples are coloured by tissue type. Values within parentheses on the axes indicate the proportion of variance explained by each principal component. **e**. Genes with CN-independent ASE in either tumour or normal tissue samples. Genes with an enrichment of CN-independent ASE in tumours are marked in blue, lung cancer genes are represented by triangles and imprinted genes have a black outline. Enrichment was defined as FDR < 0.05 from a Fisher's exact test per gene. The number of regions used to calculate enrichment varied per gene between 5 and 850 (median = 164) for tumours and between 5 and 95 (median = 35) for normal tissue. **f**. Relationship in LUSC between the proportion of evaluable genes with CN-independent ASE and the ratio of differentially hypo-methylated clusters of neighbouring CpGs compared to all differentially methylated genomic positions. The Pearson correlation coefficient is shown; P value was calculated using a linear mixed-effects model with tumour as random variable ($r = -0.18$, $P = 0.35$). **g**. Percentage of evaluable genes affected by CN-independent ASE in wild type (WT) and *SETD2* deficient isogenic cell lines. Expression data was obtained from publicly available datasets from three separate studies in three different cell lines[23–25]: in total, data from 10 cell lines across 3 experiments (n = 6, 2 and 2). Boxes represent lower quartile, median and upper quartile. P values were calculated using a linear mixed effects model, using the study of origin of each sample as a random effect. *SETD2*-/-: inactivation of the *SETD2* gene.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Patterns of RNA variant diversity in TRACERx. **a**. Overview of RNA substitutions in the primary tumour lung TRACERx cohort, from top to bottom: Number and type of RNA variants per megabase per tumour, tumours are sorted from left to right by histological subtype and by number of variants; Proportion of each variant type per tumour; Proportion of variants present in any of the normal samples; Proportion of tumour-specific RNA variant sites shared across at least two tumours. NSCLC histological subtype per patient. LUAD, lung adenocarcinomas, n = 190; LUSC, lung squamous cell carcinomas, n = 119; Other, other subtypes, n = 43; tumour-adjacent normal lung tissue, n = 96. **b**. Volcano plots showing Pearson correlations between the number of RNA variant signature substitutions and gene expression for all genes in the transcriptome, split by RNA single-base substitution (SBS) signature. P values were calculated using a linear mixed effects model, using tumour of origin of each region as random effect. The genes with the 5 most significant correlations with each signature are labelled. P values were adjusted for repeated measures. Correlations were based on 765 primary tumour regions with at least 20 RNA variants from 329 tumours. Colour indicates dot density, with light coloured points belonging to areas of high density in the plot. **c**. Proportion of RNA variants relative to variant type (A>G or C>T) in 4nt RNA loops. C>T substitutions were more prevalent in the 4th nucleotide of 4nt RNA hairpin loops, consistent with APOBEC RNA editing activity. **d**. Proportion of substitutions assigned to RNA-SBS2 activity compared to the proportion of RNA variants at CAT[C>T] motif sites per tumour region (CAUC ratio). Blue dots represent regions where RNA editing at these motifs was enriched (Fisher's test $P<0.05$ for C>T substitutions at each site compared to C sites in a 40nt genomic region). P values were computed based on a two-sided t test testing the null hypothesis that the Pearson correlation coefficient (r) = 0, within 892 tumour regions and 77 tumour-adjacent normal tissue samples with at least 10 C>T variants. **e**. Pearson correlation between the exposure of RNA-SBS signatures within metastatic tumour regions and their respective seeding regions in the primary tumour (left); and tumour-adjacent normal lung tissue and their respective primary tumour regions (right). Primary tumour exposure was calculated as the median exposure across all primary regions for the comparison with normal tumour-adjacent tissue, and of all seeding regions for the comparison with metastases. Only primary-metastasis pairs where more than 20 RNA substitutions were detected in the metastasis and primary region were used (n = 50 pairs for normals, n = 31 for metastases). P values were computed based on a two-sided t test testing the null hypothesis that the Pearson correlation coefficient = 0. **f**. Pearson correlation between the activity of RNA-SBS1 and the global levels of methylation in a tumour region (measured as the percentage of all differentially methylated positions that are differentially hypomethylated clusters of neighbouring CpGs). Methylation data and sufficient RNA substitutions for signature deconvolution were available for 80 regions from 31 tumours. P values were calculated using a linear mixed effects model, using tumour of origin of each region as a random effect.



Extended Data Fig. 4 | Transcriptional features of metastasis. a. Expression distance between paired primary tumour regions; compared to distance between paired primary and non-LN intrathoracic metastatic tumour regions. Only patients with two or more primary regions and at least one metastatic region sampled are shown (12 primary-metastasis pairs from 8 tumours). Boxes represent lower quartile, median and upper quartile, whiskers represent lower/higher bound +/− 1.5 x interquartile range. Significance was tested using a paired Wilcoxon test (P = 0.00098). **b**. Gene set enrichment analysis (GSEA) of functional groups from hallmark gene sets[14] between metastasis seeding and non-seeding regions. Only tumours where both seeding and non-seeding regions had RNA-seq were included (n = 37, 122 regions). Dots coloured by a significant enrichment after FDR correction. Mean normalised enrichment score (NES) is displayed on the x-axis and indicates the enrichment for a given gene set, and the negative log of the adjusted P value is displayed on the y-axis. **c**. Overview schematic of the machine learning framework used to predict whether a region contains a metastasis-seeding clone(s). MLP-SVM: multilayer-perceptron with support vector machine terminal layer. **d**. Individual Shapley Additive Explanations (SHAP) values for the most important features across the combined ensemble. Positive SHAP values indicate weighting towards a prediction of metastasis seeding whereas negative SHAP values indicate a weighting towards prediction of metastasis non-seeding. Colour scale represents the value of the feature across the test dataset (red=high values, blue=low values). For instance, high values of the ORACLE expression marker (red dots) were associated with a higher likelihood of a region being seeding (positive SHAP values) in the combined ensemble. The predictions were based on 516 primary tumour regions from 206 tumours where seeding status could be established and where all metrics tested could be measured (307 non-seeding regions, 209 seeding), with a 75%-25% training-test dataset split. TMB: tumour mutational burden; CN-ind ASE: Copy number-independent allele specific expression; HPCS: High Plasticity Cell State[5]; GD: genome doubling; CCF: cancer cell fraction; Clone dominance CCF: maximum CCF at terminal nodes of a phylogenetic tree; SCNA: somatic copy number alteration.

Extended Data Table 1 | Central pathological review of non-LUAD tumours clustering with LUADs based on expression patterns

Tumour ID	Histology	Any LUAD morphology and/or TTF-1+	Presence of adenocarcinomatous morphology (H&E)	TTF-1	p40/p63/CK5	CD56/synaptophysin/ chromogranin	pan-cytokeratin/CK AE1/3
CRUK0013	LUAD combined LCNEC	yes	yes	pos	-	pos	-
CRUK0096	Adenosquamous carcinoma	yes	yes	neg	-	-	-
CRUK0098	Pleomorphic carcinoma	yes	yes	pos	neg	neg	-
CRUK0099	Adenosquamous carcinoma	yes	yes	pos	pos	-	-
CRUK0100	LCNEC	yes	no	pos	neg	pos	-
CRUK0271	Adenosquamous carcinoma	yes	yes	neg	neg	-	-
CRUK0273	LCC	no	no	neg	neg	neg	pos
CRUK0334	Squamous cell carcinoma	-	-	-	-	-	-
CRUK0372 Tumour1	Pleomorphic carcinoma	yes	yes	-	-	-	-
CRUK0420	Adenosquamous carcinoma	yes	yes	pos	pos	-	-
CRUK0422	Adenosquamous carcinoma	yes	yes	pos	pos	-	-
CRUK0434	Pleomorphic carcinoma	yes	yes	pos	-	-	-
CRUK0467	Squamous cell carcinoma	no	no	-	-	-	-
CRUK0476	Pleomorphic carcinoma	yes	yes	-	-	-	-
CRUK0514	Adenosquamous carcinoma	yes	yes	pos	-	-	-
CRUK0524	Pleomorphic carcinoma	no	no	neg	neg	-	pos
CRUK0527	Pleomorphic carcinoma	no	no	neg	pos	-	pos
CRUK0549	Squamous cell carcinoma	no	no	neg	pos	-	pos
CRUK0557	Pleomorphic carcinoma	yes	yes	pos	pos	-	-
CRUK0566	LCC	no	no	neg	neg	neg	pos
CRUK0584	Squamous cell carcinoma	no	no	neg	pos	-	-
CRUK0598	Adenosquamous carcinoma	yes	yes	-	-	-	-
CRUK0692	Pleomorphic carcinoma	yes	yes	pos	pos	-	-
CRUK0698	LCNEC	yes	no	pos	neg	pos	-
CRUK0719	LCC	no	no	neg	neg	pos	pos
CRUK0735	LUAD combined LCNEC	yes	yes	-	-	-	-
CRUK0769	LCNEC	yes	no	pos	pos	pos	-

Histological subtype was determined by centralized pathological slide review, as was the presence of adenocarcinomatous morphology within these tumours. Additionally, immunohistochemical staining profiles were summarized according to the pathology reports. LCC, large cell carcinoma; LCNEC, large cell neuroendocrine carcinoma.

Extended Data Table 2 | Sanger sequencing-validated RNA variant sites

Tumour region	Position	Alternative allele count	Total coverage	Editing type	Gene Symbol
CRUK0034_SU_T1-R2	chr8:103841636	34	107	A>G	AZIN1
CRUK0206_BR_T1-R3	chr13:46090371	52	96	A>G	COG3
CRUK0129_SU_T1-R7	chr13:46090371	46	95	A>G	COG3
CRUK0129_SU_T1-R7	chr1:160302244	254	466	A>G	COPA
CRUK0748_SU_T1-R4	chr3:58141791	112	166	A>G	FLNB
CRUK0418_SU_T1-R2	chr4:57976234	102	117	A>G	IGFBP7
CRUK0036_SU_T1-R1	chr8:103841637	75	342	A>G	AZIN1
CRUK0293_SU_T1-R2	chr8:143845821	55	185	C>T	LYNX1
CRUK0451_SU_T1-R2	chr1:20981977	40	90	C>T	DDOST
CRUK0094_SU_T1-R4	chr14:94844819	82	226	C>T	SERPINA1

RNA variant sites detected through our bioinformatics pipeline, and validated using Sanger sequencing. The variant was detected in the RNA but not the DNA of the same tumour region.

Extended Data Table 3 | Hallmark gene set functional groups

Hallmark	Functional group
Adipogenesis	Development
Allograft rejection	Immune
Androgen response	Signalling
Angiogenesis	Development
Apical junction	Cellular component
Apical surface	Cellular component
Apoptosis	Pathway
Bile acid metabolism	Metabolic
Cholesterol homeostasis	Metabolic
Coagulation	Immune
Complement	Immune
DNA repair	DNA damage
E2F targets	Proliferation
Epithelial mesenchymal transition	Development
Estrogen response early	Signalling
Estrogen response late	Signalling
Fatty acid metabolism	Metabolic
G2M checkpoint	Proliferation
Glycolisis	Metabolic
Hedgehog signalling	Signalling
Heme metabolism	Metabolic
Hypoxia	Pathway
IL2 STAT5 signalling	Signalling
IL6 JAK STAT3 signalling	Immune
Inflammatory response	Immune
Interferon alpha response	Immune
Interferon gamma response	Immune
Kras signalling down	Signalling
Kras signalling up	Signalling
Mitotic spindle	Proliferation
MTORC1 signalling	Signalling
MYC targets v1	Proliferation
MYC targets v2	Proliferation
Myogenesis	Development
NOTCH signalling	Signalling
Oxidative phosphorylation	Metabolic
P53 pathway	Proliferation
Pancreas beta cells	Development
Peroxisome	Cellular component
PI3K AKT MTOR signalling	Signalling
Protein secretion	Pathway
Reactive oxygen species pathway	Pathway
Spermatogenesis	Development
TGF beta signalling	Signalling
TNFA signalling via NfKB	Signalling
Unfolded protein response	Pathway
UV response down	DNA damage
UV response up	DNA damage
WNT beta catenin signalling	Signalling
Xenobiotic signalling	Metabolic

MSigDB hallmark gene set and their functional group as assigned by the authors[14].

nature portfolio

Corresponding author(s): Nicholas McGranahan
Charles Swanton

Last updated by author(s): Feb 3, 2023

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	RNA-seq alignment and QC Illumina adapters were trimmed from raw sequencing reads using Cutadapt (v2.10) The quality of the trimmed reads estimated per flow cell lane using FASTQC (v.0.11.9) Fastq read files were aligned to the Hg19 human reference genome using STAR (v2.5.2a) Duplicated reads in each BAM file were marked with the MarkDuplicates function from GATK (v4.1.7.0) Aligned reads were quality checked using QoRTs (v1.3.6) to assess RNA integrity Somalier (v0.2.7) was used to detect potential instances of sample mislabelling. FASTQC, QoRTs and Somalier outputs were visualised using MultiQC (v1.9) RSEM (v1.3.3) was used with default parameters to quantify gene expression based on the BAM files aligned to the transcriptome RNA coverage was calculated for single nucleotide variants (SNVs) detected in matched whole exome sequencing data per tumour region using SAMtools (v1.9) mpileup All steps described were implemented through the Nextflow (v20.07.1) pipeline manager Reduced-representation Bisulfite Sequencing (RRBS) FastQC v0.11.2 was used for quality control Trim Galore! (Babraham Institute, https://www.babraham.ac.uk/) a wrapper around Cutadapt (v2.10), was used to trim reads The bisulfite converted DNA sequence aligner Bismark (v0.14.4) was used to align reads to the UCSC reference genome Hg19 PCR deduplication was carried out using NuDup (v2.3), leveraging NuGEN's molecular tagging technology (https://github.com/nugentechnologies/nudup) Most analyses were run using the R coding environment (v3.6.3)

RNA clustering
RSEM raw read counts were normalised using the median of ratios method implemented in DESeq2 (v1.24.0)
uniform manifold approximation and projection was performed using the R package umap (v2.7.7.0)
ASCAT (v2.3) and SAMTools mpileup (v1.9) were used to obtain RNA-derived estimates of tumour fraction

Gene expression differences
The R package edgeR (v3.26.5) was used to obtain gene expression differences
The R package fgsea (v1.10.1) was used to perform a gene set enrichment analysis on the gene expression differences results

Allele-specific expression
RNA read counts were compared to DNA copy number estimates through beta-binomial tests using the R package VGAM (v1.1-2)
CAMDAC (https://doi.org/10.1101/2020.11.03.366252) was used for allele-specific methylation calls

RNA variant calling
RNA-specific variants were called using the somatic variant caller Mutect2 and FilterMutectCalls from GATK (v4.1.7.0)
Mutect2 processes were run in parallel using GNU parallel (v20210422)
BCFtools (v1.10.2) was run to keep only biallelic PASS variants
bam-readcount (v0.8) was used to extract RNA reads at variant locations called by Mutect2 for further filtering, based on read depth and on the location of variants in the genome to prevent false positives arising from sequencing and mapping errors
RNA editing signatures were extracted using the R package hdp (v0.1.5)
Signatures were assigned to each tumour region using the R package deconstructSigs (v1.9.0)

All linear mixed effects models were performed using the R package lmerTest (v3.1-3)

The packages GenomicRanges (v1.36.0), stringr (v1.4.0) and TxDb.Hsapiens.UCSC.hg.knownGene (v3.2.2) were used to handle sequence data in R

dNdS analyses for detecting selection were performed usning the R package dndscv (v0.1.0.0)

The metastatic potential classifier was performed in Python (v3.3.5) using the packages pandas(v1.3.3), sklearn(v0.0) and tensorflow(v2.6.0)

The packages dplyr (v1.0.3), tidyr(v1.1.0) and reshape2 (v1.4.2) were used for data handling in R

Visualisation
Data was visualised using the R packages ggplot2 (v3.2.1), ggpubr (v0.4.0), cowplot (v1.0.0), gridExtra(v2.3), scales (v1.0.0) and ggrepel (v0.8.1)

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The RNA sequencing (RNA-seq), Whole exome sequencing (WES) and Reduced representation bisulfite sequencing (RRBS) data data (in each case from the TRACERx study) used during this study have been deposited at the European Genome–phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession codes EGAS00001006517 (RNAseq), EGAS00001006494 (WES) and EGAS00001006523 (RRBS); access is controlled by the TRACERx data access committee. Details on how to apply for access are available at the linked page.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size

The sample size (421 patients) represents the half-way point of the TRACERx longitudinal study. In total, we analysed paired whole exome sequencing and RNA-seq paired data from 347 patients that passed quality check filters for RNA.

TRACERx is a programme of work of multiple projects built around a single observational cohort study. It is not possible to perform a sample size calculation for each project, especially post hoc. The study size of the cohort was done in relation to tumour heterogeneity and disease free survival:

The sample size is based on demonstrating a relationship between tumours with divergent intratumour heterogeneity index values and clinical outcome. Patients will be split evenly into those with a low and high intratumour heterogeneity index value (and other splits will be considered). Assuming a median Disease Free Survival (DFS) of 30 months and a hazard ratio (HR) of 0.77, with a 2-sided 5% significance level, 90% power, accrual period of 3 years and 5 years follow-up after the end of accrual, the sample size required is almost 400 per group (total of 800 patients). Assuming a 5% dropout rate, a total of 842 patients (421 per group) are required. At 85% power, 705 patients would be required in total, which could be the minimum target. However, we will instead aim for 750 patients and recruitment will continue for the length of time which is funded for accrual in order to get as close as possible to the ideal target of 842 patients. A study size of 842 is also large enough to detect a 10% improvement in a 5 year OS rate from 46% in the high Intratumour Heterogeneity Index (ITB) to 56% in the low Intratumour Heterogeneity Index group (HR=0.75), with 80% power and a 2 sided type I error set at 5% (logrank test). A high/low ITB value will be defined as values above/below the 50th percentile (median ITB). We have a target DFS effect of a 23% reduction in risk (hazard ratio 0.77), which means that our study is powered for an effect at least this large, including a 30% difference (which has been the target for progression-free survival in trials of advanced NSCLC, in relation to expected effects on OS).

Data exclusions

Data was excluded only on the basis of:
- Non-elegibility for the TRACERx clinical trial due to failure of the patient's data to comply with the study protocol (see below)
- The sequenced data did not pass our quality check filters

Replication

TRACERx is a prospective longitudinal study. As such, the results shown here are not the result of an experimental setup. This is the half-way point of the TRACERx 421 and reflects hypothesis generating analysis.

Randomization

Given the observational nature of the TRACERx longitudinal study, no experimental groups were allocated beforehand. Factors that could affect the interpretation of our results such as the background genetic makeup of each patient or the histological subtype of tumours were taken into account in all our statistical analyses. These were accounted for by including them as covariates in hypothesis testing. For instance, we used tumour ID as a random effect factor in linear mixed effects models for many of our analyses.

Blinding

Blinding is not relevant as this is an observational study. Patients were not allocated to any intervention and they were followed up and assessed as per routine practice. No biomarker results (tissue and bloods) are reported back to patients, so there is no likelihood of people changing their behaviours based on these findings. The laboratory analyses were all performed without knowing the outcome (DFS or survival) status of the patients, which represents a form of blinding.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics

421 patients are included in this TRACERx cohort. 44.6% are females, 55.4% males; 93% are smokers or have a smoking history, 7% are never smokers; 25% of patients were diagnosed at stage IA, 25% at IB, 17.8% at IIA, 13.5% at IIB, 18.5% at IIIA and 0.2% and IIIB; 52% of diagnosed tumours were adenocarcinomas, 28.8% squamous cell carcinomas and 19.2, other histological subtypes; 93% of the cohort is from a white ethnic background and the mean age of the patients is 69, ranging between 34 and 92.

Please note that the study started recruiting patients in 2016, when TNM version 7 was standard of care. The up-to-date inclusion/exclusion criteria now utilizes TNM version 8.

TRACERx inclusion and exclusion criteria

Inclusion Criteria:
_Written Informed consent
_Patients ≥18 years of age, with early stage I-IIIB disease (according to TNM 8th edition) who are eligible for primary surgery.
_Histopathologically confirmed NSCLC, or a strong suspicion of cancer on lung imaging necessitating surgery (e.g. diagnosis determined from frozen section in theatre)
_Primary surgery in keeping with NICE guidelines planned

_Agreement to be followed up at a TRACERx site
_Performance status 0 or 1
_Minimum tumor diameter at least 15mm to allow for sampling of at least two tumour regions (if 15mm, a high likelihood of nodal involvement on pre-operative imaging required to meet eligibility according to stage, i.e. T1N1-3)
Exclusion Criteria:
_Any other* malignancy diagnosed or relapsed at any time, which is currently being treated (including by hormonal therapy).
_Any other* current malignancy or malignancy diagnosed or relapsed within the past 3 years**.
*Exceptions are: non-melanomatous skin cancer, stage 0 melanoma in situ, and in situ cervical cancer
**An exception will be made for malignancies diagnosed or relapsed more than 2, but less than 3, years ago only if a pre-operative biopsy of the lung lesion has confirmed a diagnosis of NSCLC.
_Psychological condition that would preclude informed consent
_Treatment with neo-adjuvant therapy for current lung malignancy deemed necessary
_Post-surgery stage IV
_Known Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV) or syphilis infection.
_Sufficient tissue, i.e. a minimum of two tumor regions, is unlikely to be obtained for the study based on pre-operative imaging

Patient ineligibility following registration
_There is insufficient tissue
_The patient is unable to comply with protocol requirements
_There is a change in histology from NSCLC following surgery, or NSCLC is not confirmed during or after surgery.
_Change in staging to IIIC or IV following surgery
_The operative criteria are not met (e.g. incomplete resection with macroscopic residual tumors (R2)). Patients with microscopic residual tumors (R1) are eligible and should remain in the study
_Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy is administered.

Recruitment

When patients are initially diagnosed with stage I-III lung cancer and then referred for surgical resection, a research nurse identifies them on a clinic/operating list. The patient has an initial eligibility assessment and then provided with written information about the TRACERx study and he/she can ask the research nurse any questions.

Patients have to agree to provide serial blood samples whenever they attend clinic for routine blood sampling, so this represents the only main potential self-selecting bias (i.e. only patients willing to do this would participate). However, it is unclear how this would affect the biomarker analyses. Also, the gender and ethnicity characteristics are in line with patients seen in routine practice.

Inclusion and exclusion criteria are summarised above.
Informed consent for entry into the TRACERx study was mandatory and obtained from every patient.

Ethics oversight

The study was approved by the NRES Committee London with the following details:
Study title: TRAcking non small cell lung Cancer Evolution through therapy (Rx)
REC reference: 13/LO/1546
Protocol number: UCL/12/0279
IRAS project ID: 138871

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

Clinical data

Policy information about clinical studies
All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

TRACERx Lung https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546

Study protocol

https://clinicaltrials.gov/ct2/show/NCT01888601

Data collection

Clinical and pathological data is collected from patients during study follow up - this period is a minimum of five years. Data collection is overseen by the sponsor of the study (Cancer Research UK & UCL Cancer Trials Centre) and takes place in hospitals across the United Kingdom. A centralised database called MACRO is used for this purpose. Recruitment started in 2014 across 6 sites (London, Leicester, Manchester, Aberdeen, Birmingham, and Cardiff) in the United Kingdom.

Outcomes

The main clinical outcomes are:
Disease-free survival (DFS) – measured from the time of study registration to date of first lung recurrence or death from any cause. Patients who do not have these events are censored at the date last known to be alive (including patients who developed a new primary tumour that has been shown biologically to not be linked to the initial primary lung tumour).
Overall survival - measured from the time of study registration to date of death from any cause.