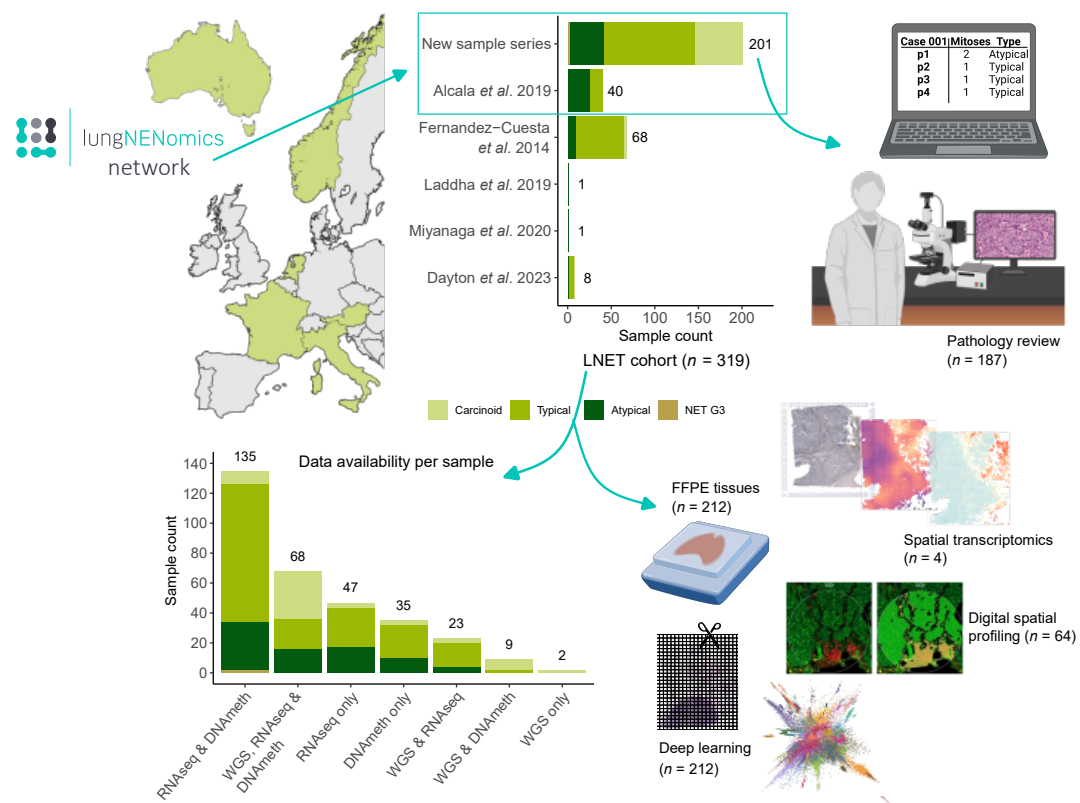


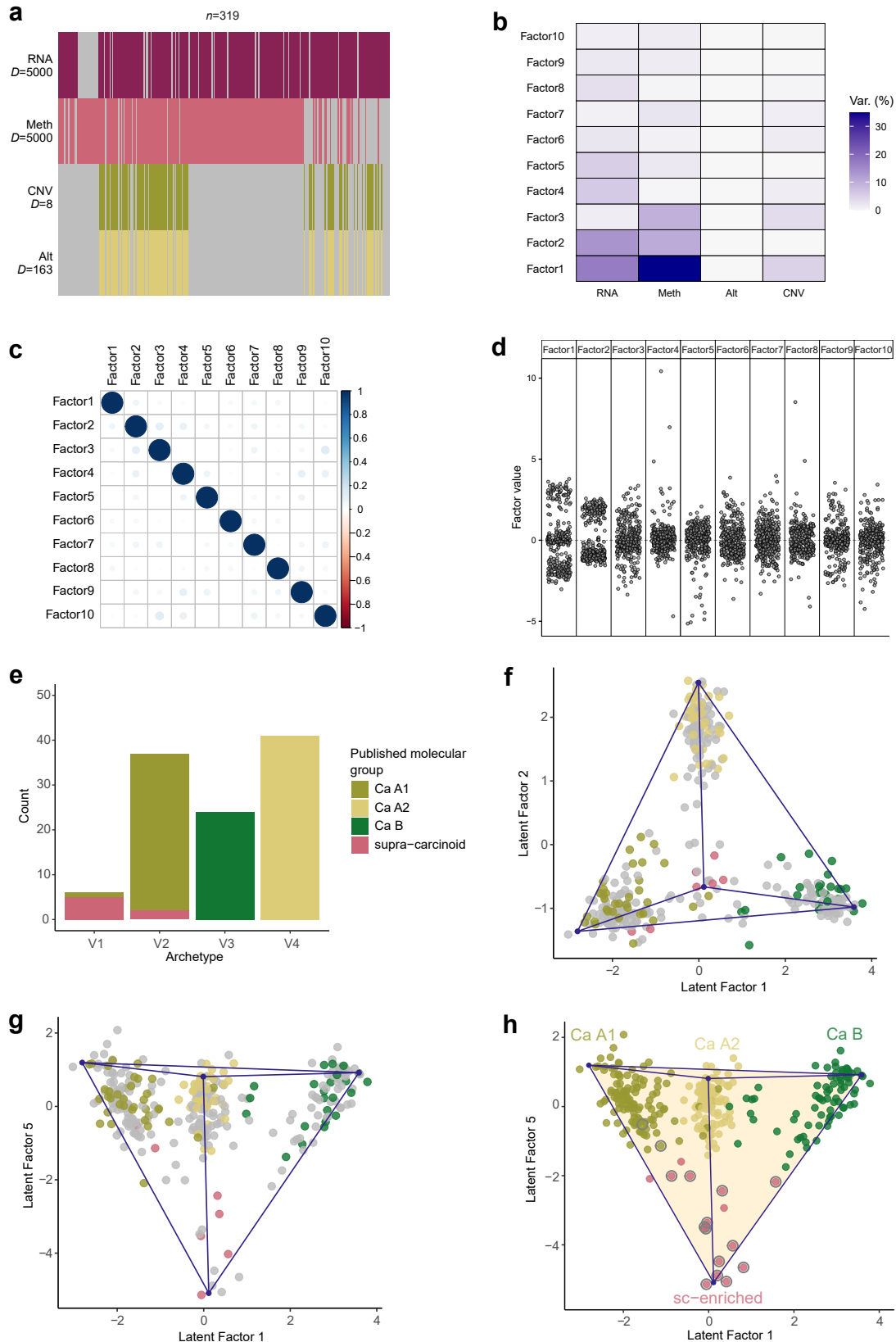
Supplementary Figures for article "Deep molecular profiling of lung neuroendocrine tumours and supra-carcinoids"

CONTENTS

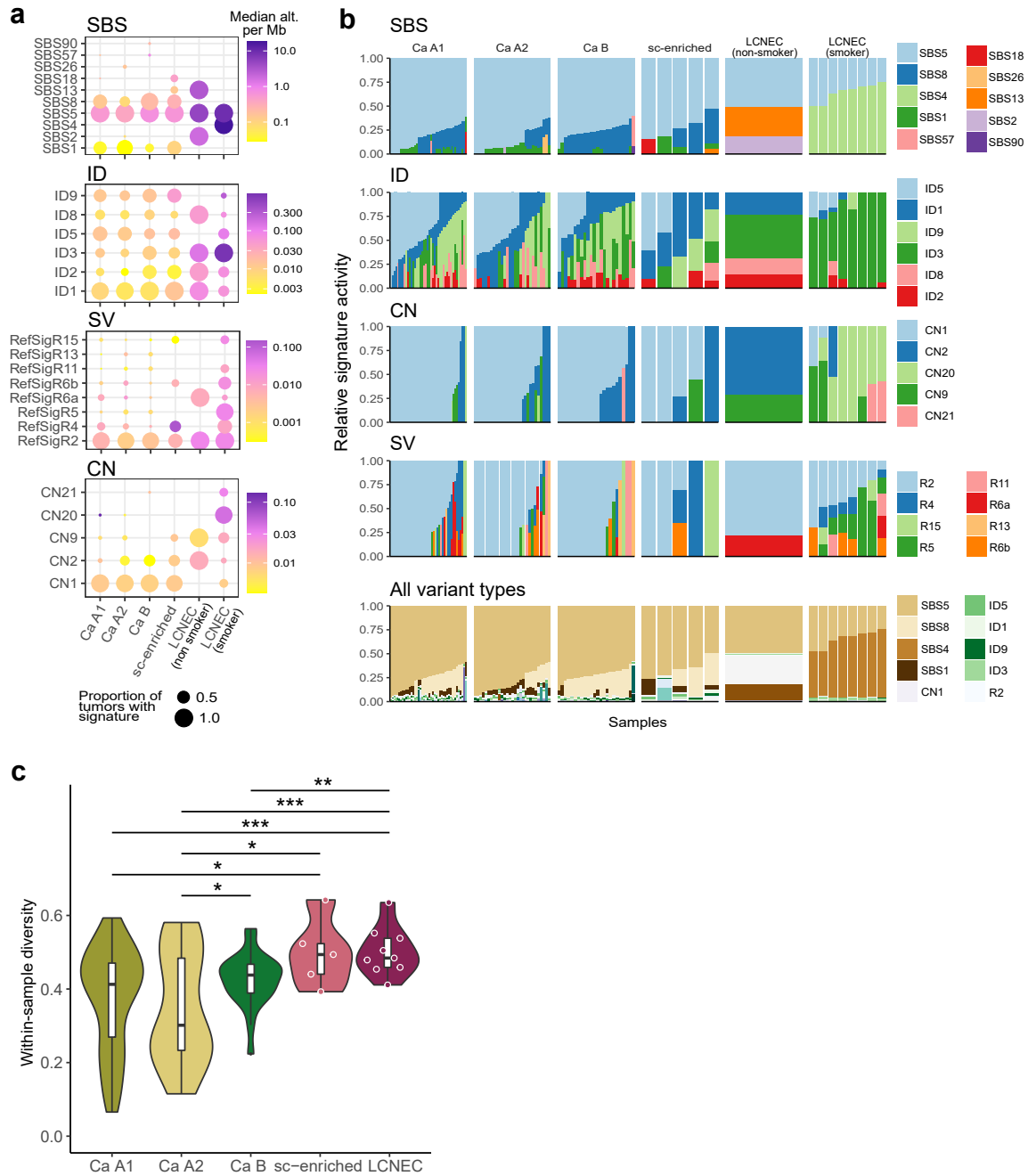
Supplementary Figure S1	2
Supplementary Figure S2	3
Supplementary Figure S3	4
Supplementary Figure S4	5
Supplementary Figure S5	6
Supplementary Figure S6	7
Supplementary Figure S7	8
Supplementary Figure S8	9
Supplementary Figure S9	10
Supplementary Figure S10	11
Supplementary Figure S11	12
Supplementary Figure S12	13
Supplementary Figure S13	14
Supplementary Figure S14	15
Supplementary Figure S15	16
Supplementary Figure S16	17
Supplementary Figure S17	18
Supplementary Figure S18	19
Supplementary Figure S19	20



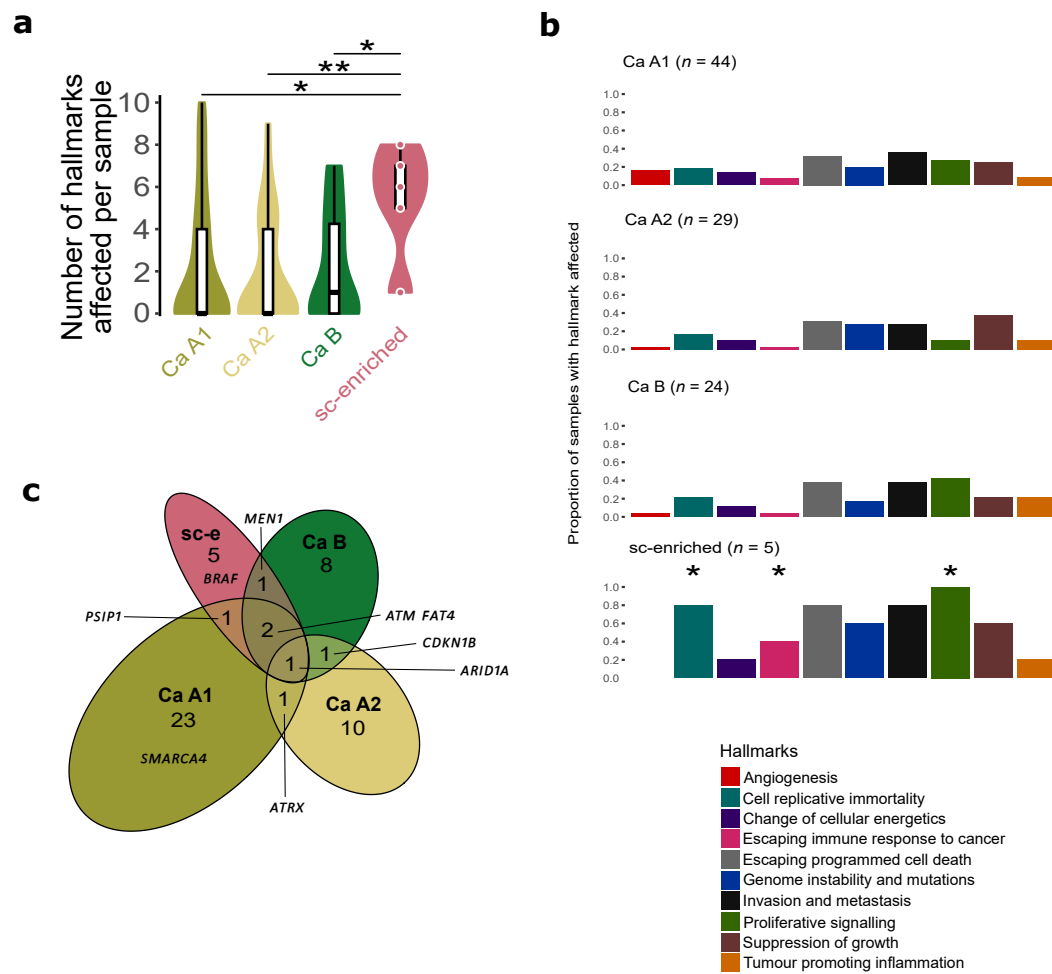
Supplementary Figure S1. Study design overview. Origin of samples within the lung NET cohort ($n = 319$, lungNENomics network or previous publications), and number of samples per histological type by origin (where “carcinoid” indicates tumours that could not be definitively classified as typical or atypical). Depiction of central pathology review. Distribution of omics data generated, or publicly available, by sample type for the lung NET cohort. Analysis techniques performed with formalin-fixed paraffin-embedded (FFPE) tissues from the lungNENomics network series. Figure created, in part, with BioRender.com.



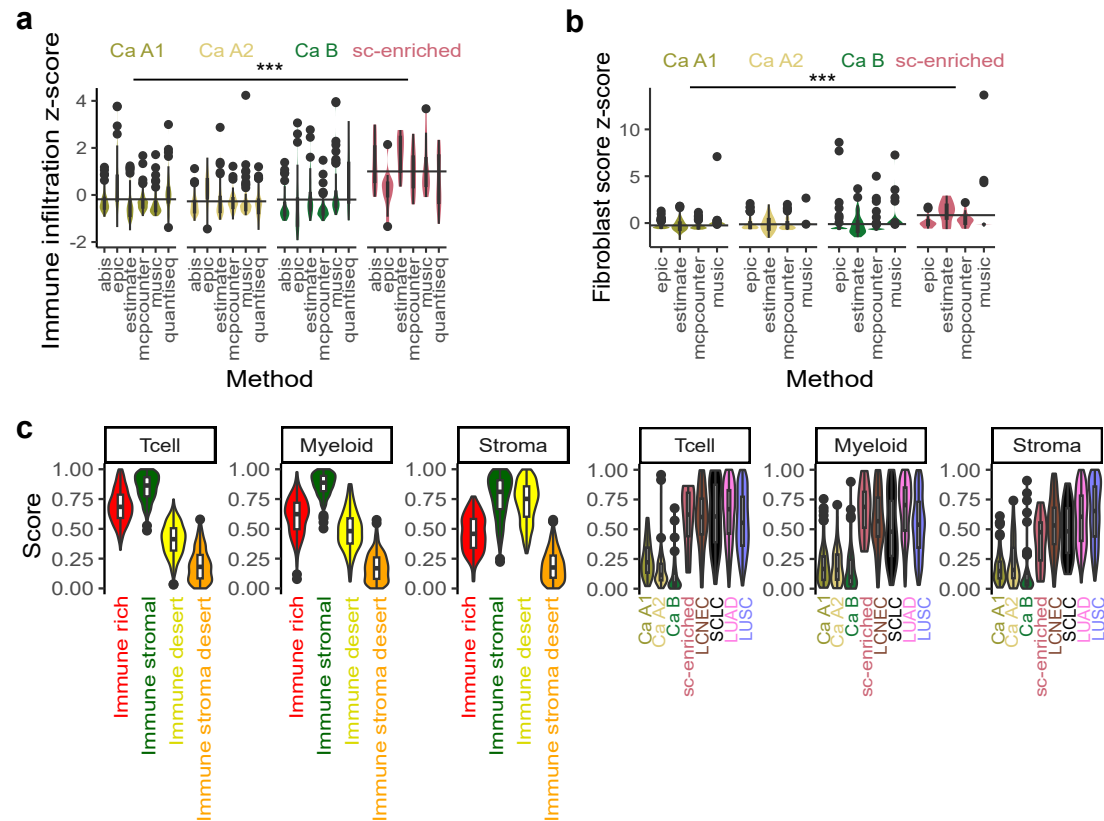
Supplementary Figure S2. Overview of lung NET cohort MOFA and ParetoTI analyses. **a**, MOFA input data matrices. D is the number of omic features incorporated per matrix [expression levels for 5000 genes (RNA), DNA methylation levels at 5000 CpG sites (Meth), copy number values at eight broad and focal regions (CNV), small and/or structural variant status at 163 genes (Alt)]. Grey indicates missing data. **b**, Variance explained by each MOFA latent factor for each data type. **c**, Correlation between sample coordinates along MOFA latent factors. **d**, Distribution of samples along each MOFA latent factor. **e** Count of samples from Alcala *et al.* 2019 ($n = 107$) and Dayton *et al.* 2023 ($n = 1$) coloured by published molecular group within archetypes found in the current study when considering $k = 4$ archetypes. **f–g**, Scatterplots showing the coordinates of samples from Alcala *et al.* 2019 and Dayton *et al.* 2023 (coloured by previously published molecular group), remaining lung NET cohort samples ($n = 211$, grey points), and archetype positions (blue points) along MOFA latent factors for $k = 4$ archetypes derived from LFs 1, 2 and 5. Colours are as per (e). **h**, Data as in Fig. 1a plotted over MOFA Latent Factors 1 and 5.



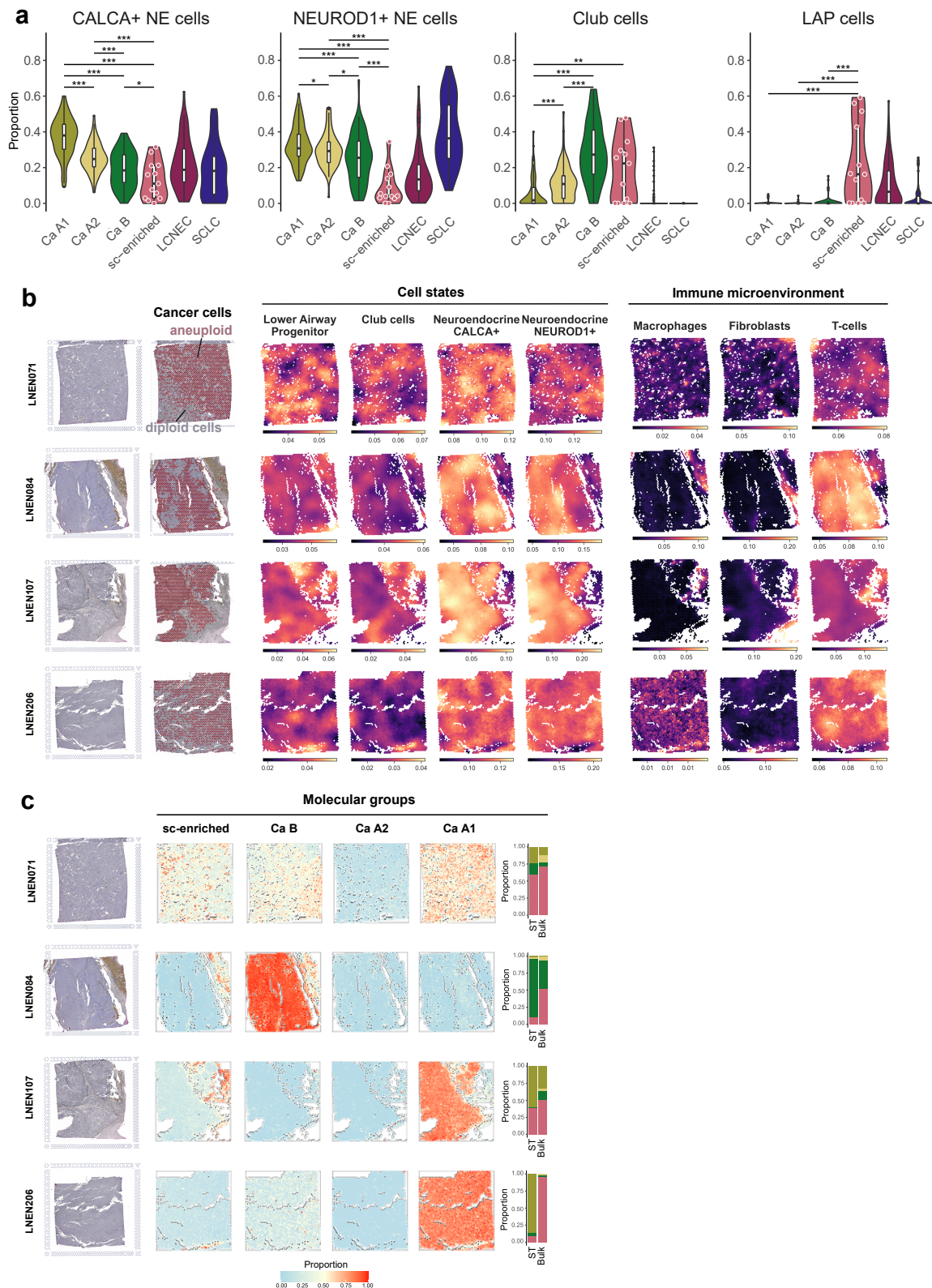
Supplementary Figure S3. The repertoire of mutational signatures in lung NETs. **a**, Dot plot of mutational signature burdens for different types of variants. Dot size represents the proportion of tumours with a given signature, and dot colours represent the median number of alterations per megabase among tumours with the signature present. SBS, single base substitution; ID, Indel; CN, copy number; SV, structural variant. **b**, Barplots of the relative proportion of signatures (y-axis) across samples (x-axis) for different types of variants. **c**, Signature variability analysis showing the within-sample diversity measure (y-axis) which corresponds to the mean Gini-Simpson index of each molecular group (x-axis). Analysis takes into account mutational signature proportions from all variant types. Asterisks represent significance codes for two-sided t-tests. In all plots and analyses, signatures with a relative activity below 5% were removed, and plots represent the lung NEN cohort with WGS data available ($n = 111$, $n = 102$ samples of the lung NET cohort plus $n = 9$ LCNec). * $0.01 < P \text{ value} < 0.05$; ** $0.001 < P \text{ value} \leq 0.01$; *** $P \text{ value} \leq 0.001$.



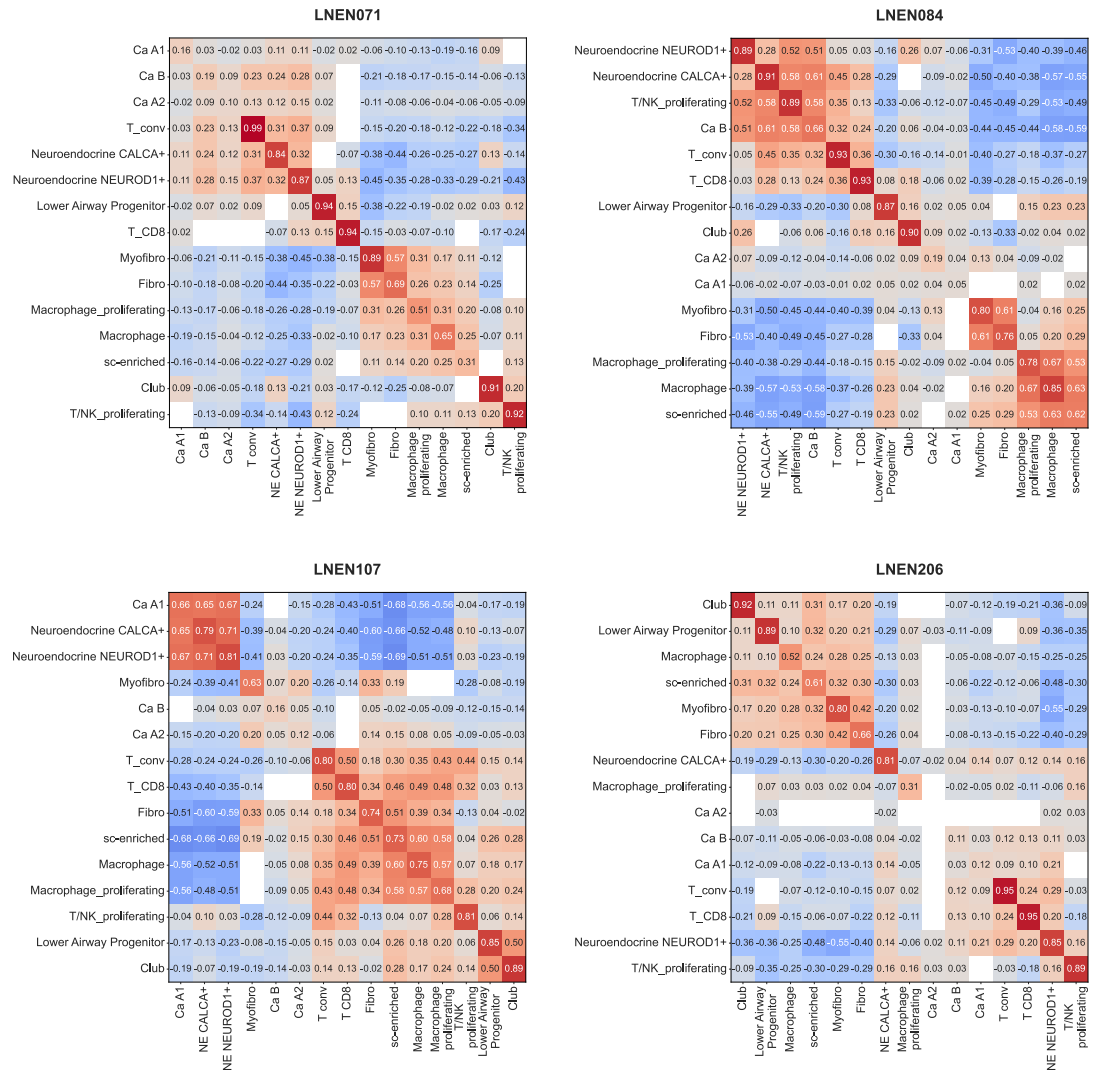
Supplementary Figure S4. Genomic hallmarks of cancer affected by damaging small variants within lung NET molecular groups. **a**, Distribution of the number of hallmarks affected by small and structural variants per patient for each group. Stars represent the significance level of Mann-Whitney U Tests. **b**, Average hallmark profile for each lung NET molecular group. Each bar represents the proportion of samples where that hallmark is affected by damaging small variants. Asterisks indicate significance level (P value) obtained from logistic regression models, or in the case of hallmarks Angiogenesis and Proliferative signalling, Fisher's exact tests, performed between archetype Ca A1 and each other archetype individually. **c**, Euler diagram of genes involved in hallmark acquisition. Each gene can only belong to one molecular group on the diagram. The genes shown on the diagram as examples are those attributed to at least two different participants. * $0.01 < P \text{ value} < 0.05$; ** $0.001 < P \text{ value} \leq 0.01$; *** $P \text{ value} \leq 0.001$.



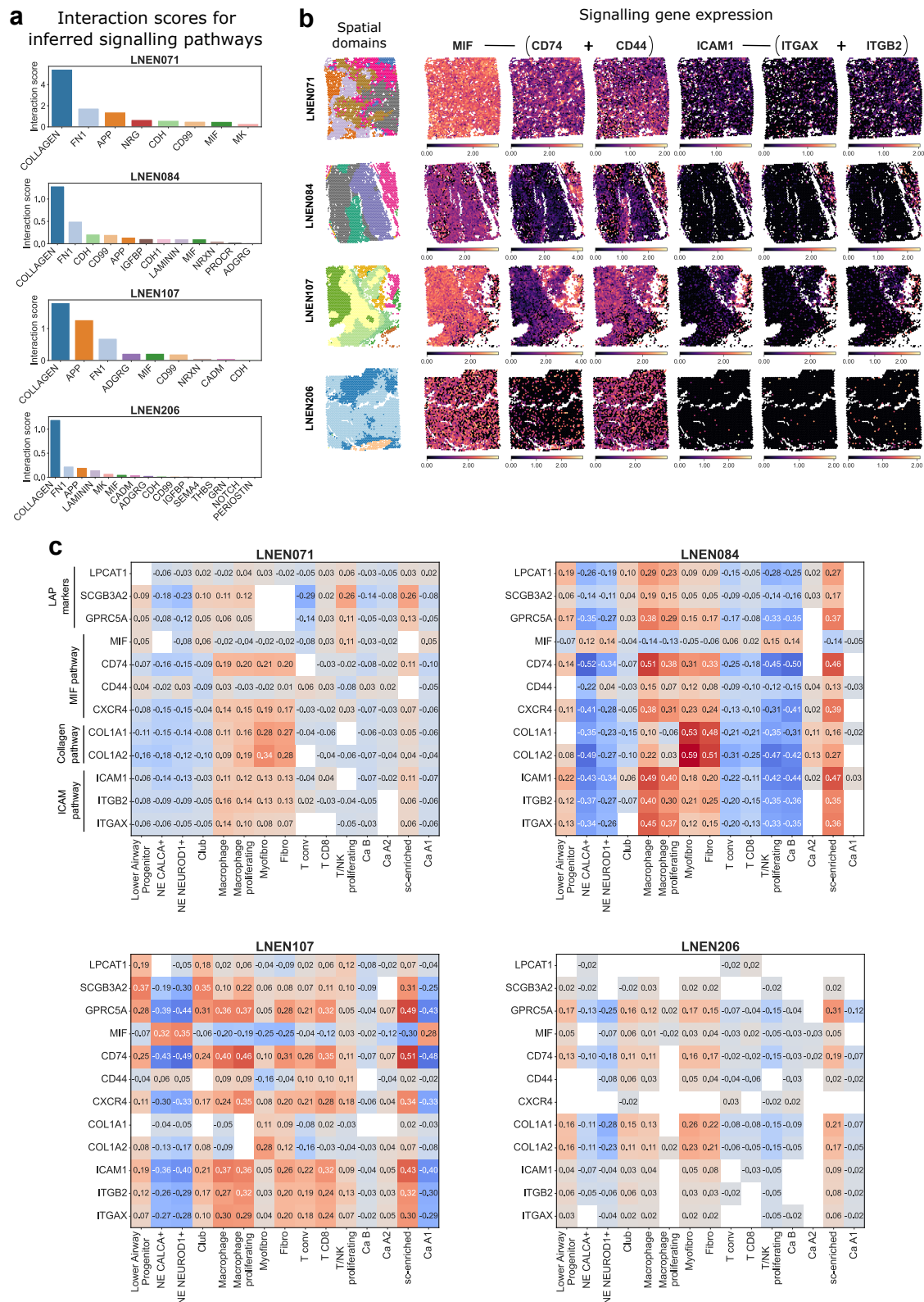
Supplementary Figure S5. Immune landscape of lung NET molecular groups. **a**, Infiltrating immune cell scores from six estimation methods per molecular group ($n = 273$). Mean z-scores per molecular group were compared by t-tests (linear mixed model with methods as random effects). **b**, Distribution of fibroblast score z-scores from six estimation methods for each molecular group ($n = 273$). Mean z-scores per molecular group were compared by t-tests. *** P value ≤ 0.001 . **c-d**, Identification of coarse immune archetypes in lung neoplasms using the three features from Combes *et al.* 2022. **c**, Violin plots of each of the three features used for the classification (T cell, Myeloid, and Stroma), in units of percentile rank of expression among samples, across the four immune clusters we identified. **d**, Same as (c) but comparing lung NET molecular groups and lung carcinoma types. In (c) and (d) samples correspond to cohort lung NEN plus SCLC, and LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma) TCGA datasets ($n = 1420$).

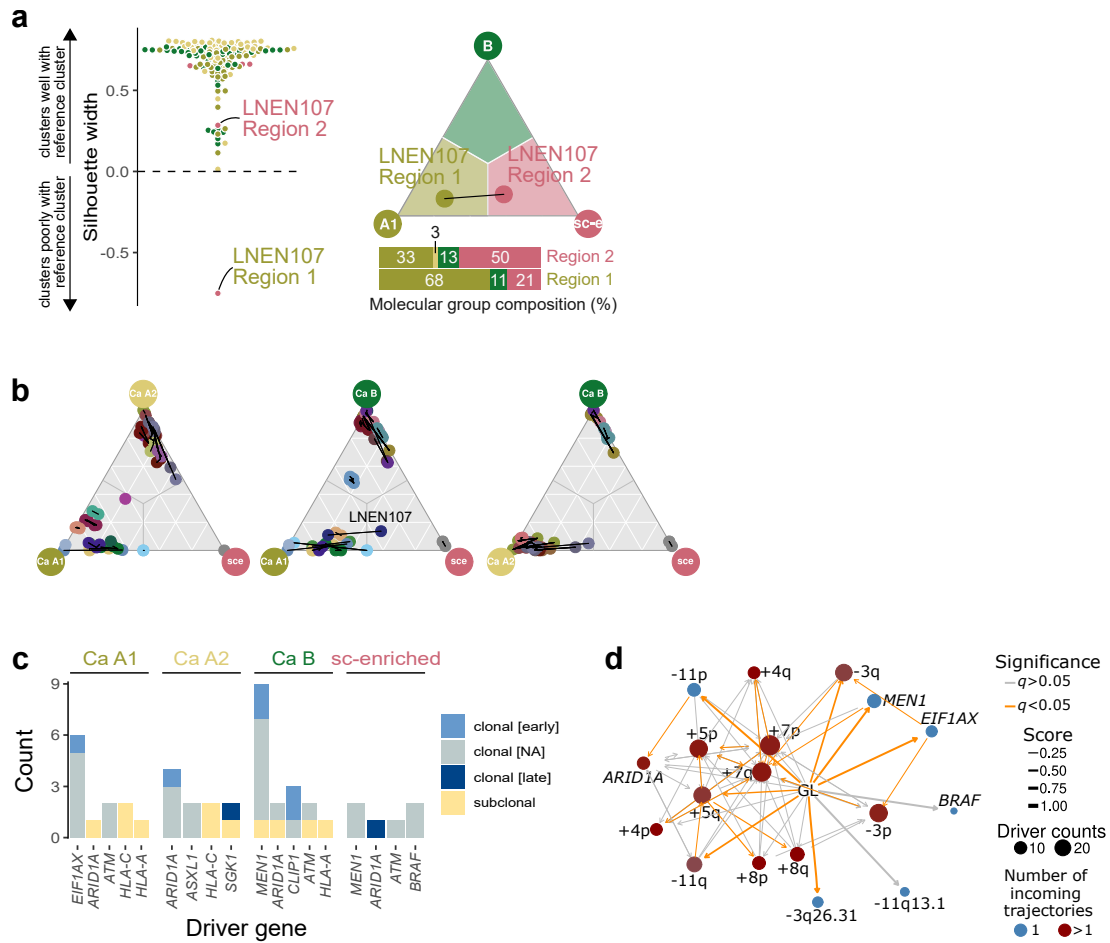


Supplementary Figure S6. Cell type analysis of bulk RNA-seq and spatial transcriptomics. **a**, Estimated proportion (y-axis) of cell types per sample (x-axis), using deconvolution with a custom single-cell reference database including a diverse repertoire of neuroendocrine cells. NE: neuroendocrine; LAP, lower airway progenitor. **(a)** displays samples from the lung NET cohort (n = 273) and LCNEC (n = 69) and SCLC (n = 51) for which RNA-seq data is available. Asterisks correspond to significance assessed by t-tests. * 0.01 < P value < 0.05; ** 0.001 < P value ≤ 0.01; *** P value ≤ 0.001. **b-c**, Spatial transcriptomics cell type analysis of four supra-carcinoid samples. **b**, H&E slide images of four supra-carcinoid tumour samples used in spatial transcriptomic analyses (column 1). Visualisation of cancer cells, defined as aneuploid cells (red), versus non-cancer cells, defined as diploid cells (grey), within each sample slide (column 2). Location of cells with one of four cell profiles, corresponding to the potential cells of origin for the tumours, estimated using spot expression deconvolution (columns 3-6). Location of the main cell types from the tumour immune microenvironment, estimated using spot expression deconvolution (columns 7-9). **c**, H&E slide images of four supra-carcinoid tumour samples used in spatial transcriptomic analyses (column 1). Location of cells with a profile matching that of one of the four lung NET molecular groups, computed using deconvolution of the spots with the average bulk gene expression profile of each group as reference (columns 2-5). Bar charts comparing the proportion of molecular group within the spatial transcriptomic slide (ST; averaged from columns 2-5), and multi-omic sequenced bulk tumour tissue from the same patient, for each of the four samples.

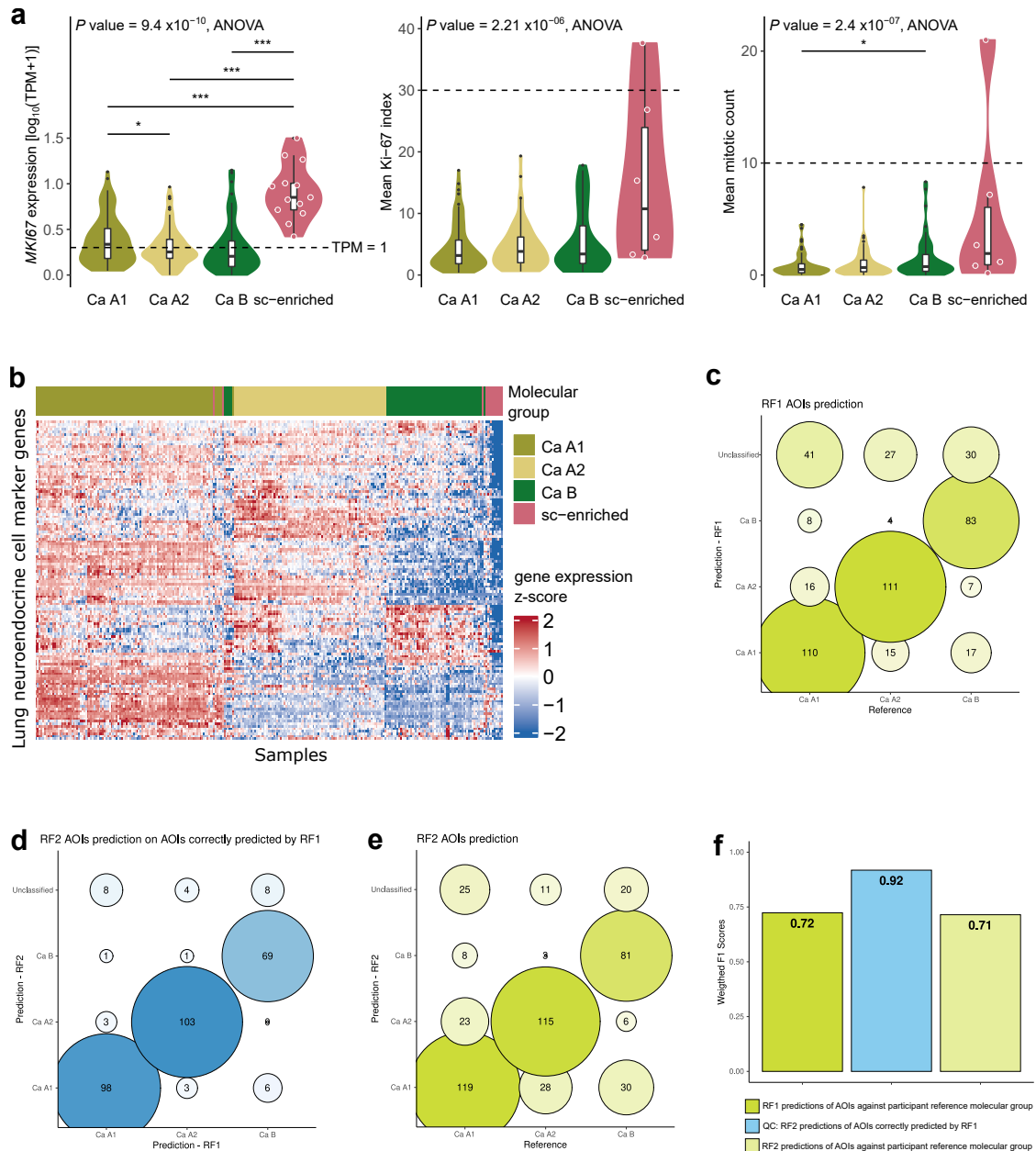


Supplementary Figure S7. Spatial transcriptomics cell type correlations of four supra-carcinoid samples. Correlation matrices displaying bivariate spatial cross-correlation (bivariate Moran's R) coefficients for cell type proportions for each supra-carcinoid sample analysed. White areas indicate non-significant results.

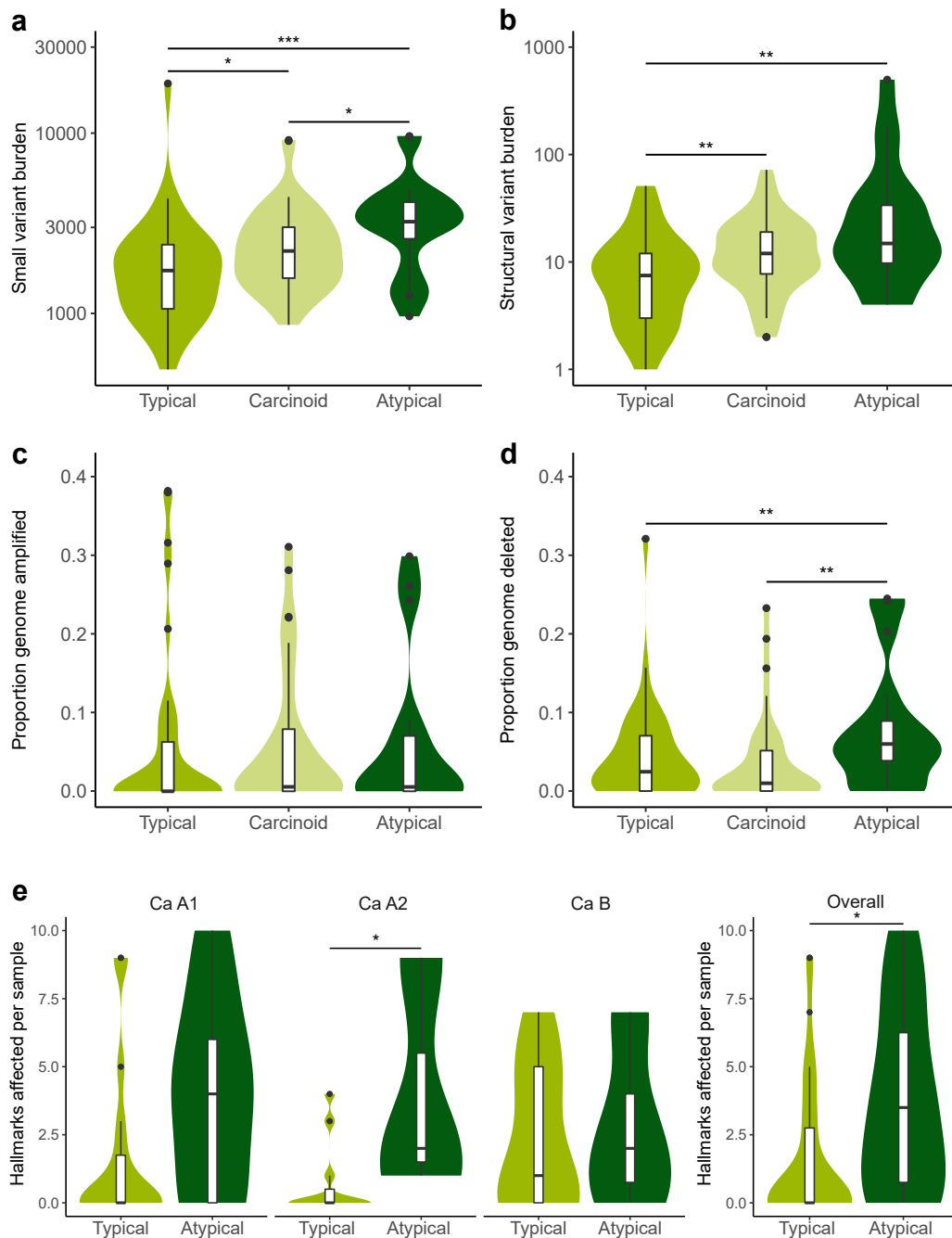




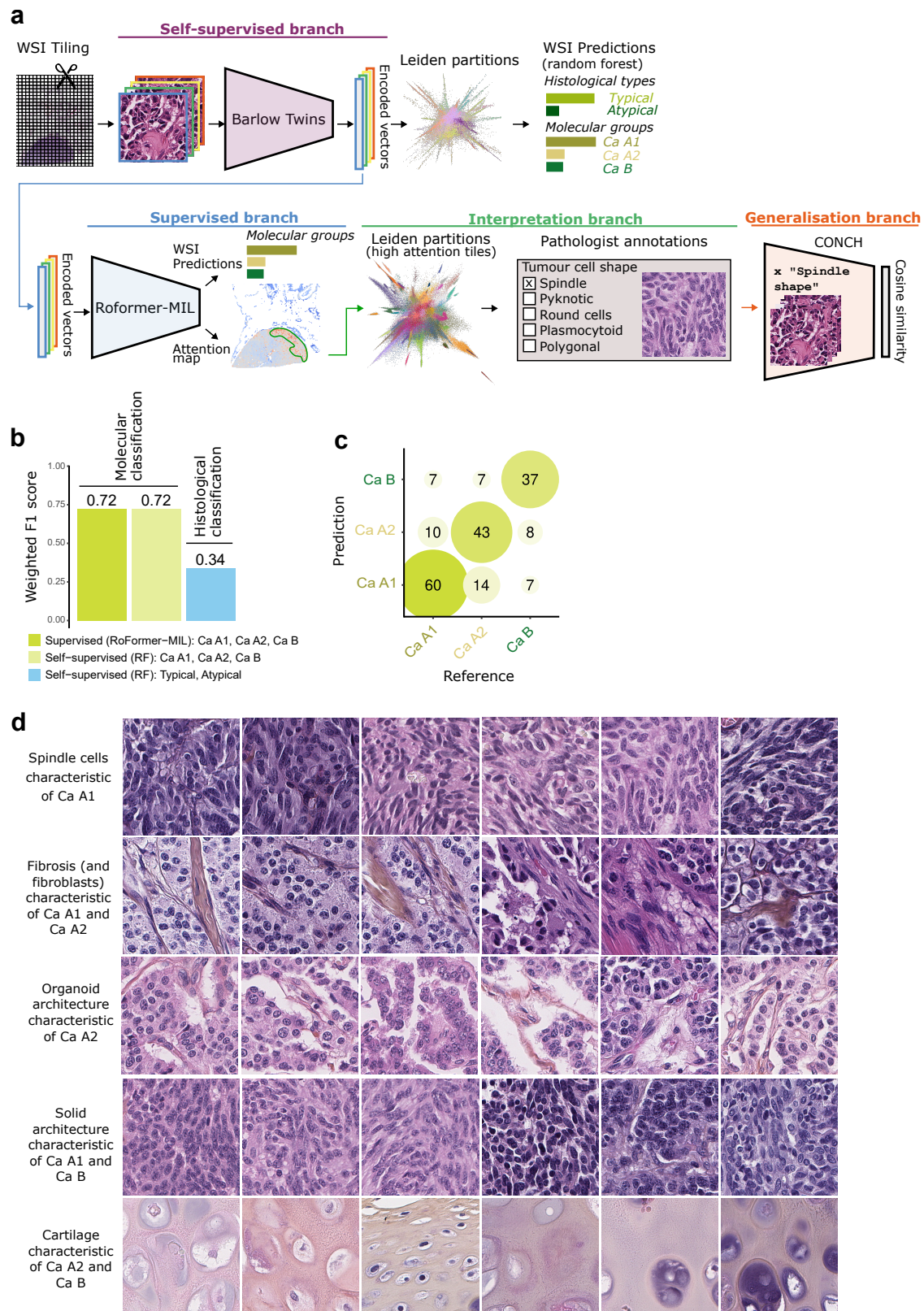
Supplementary Figure S9. Intra-tumoural heterogeneity and genetic evolutionary analyses. **a**, Silhouette width of intra-tumoural heterogeneity (ITH) samples, measuring the distance between the sample and its patient reference molecular group (molecular group of the ITH piece included in lung NET cohort MOFA) (left). Archetype analysis of a heterogeneous multi-regional tumour (LNEN107, right). Barplots correspond to the molecular group proportion composition of each region. **b**, Ternary plots depicting the relative position of ITH samples in space between different molecular groups. Black lines join ITH samples from the same patients. ITH samples are coloured by patient. Patient LNEN107 highlighted as the patient whose ITH samples were assigned to two different molecular groups. **c**, Clonality (colour) of driver alterations identified in each molecular group. **d**, Driver to driver trajectories among tumours with identified driver alterations. Arrow colours correspond to the significance of Fisher's exact test of the association between pairs of drivers (FDR corrected using the Benjamini-Hochberg method), arrow size corresponds to the strength of the inferred association, point size to the number of driver events, and point colour indicates the number of paths leading to a given driver (blue: a single trajectory leads to this driver, brown: multiple trajectories lead to it). GL, germline.



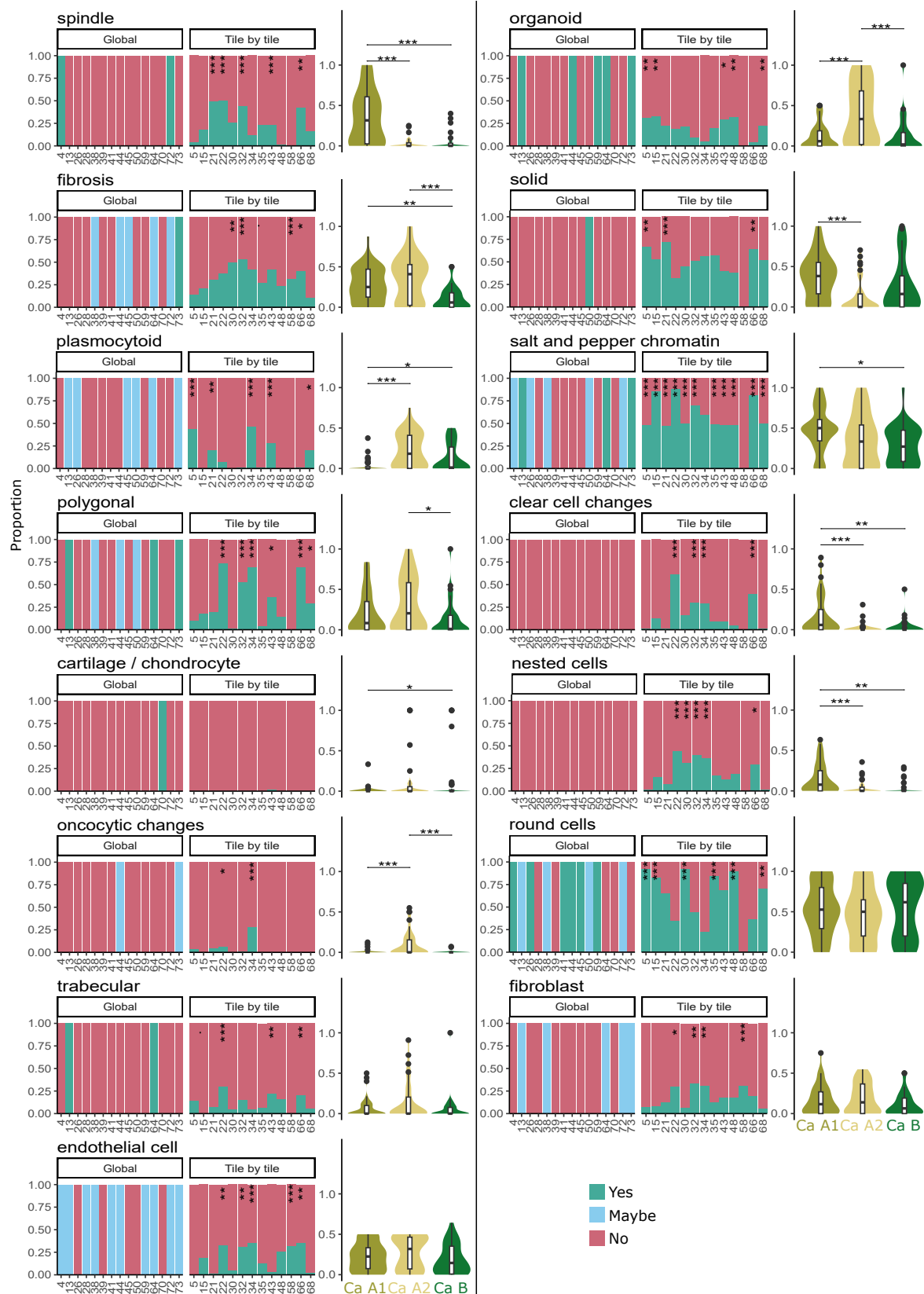
Supplementary Figure S10. Additional features of lung NET molecular groups and results of digital spatial profiling. **a**, Distribution of *MKI67* gene expression per molecular group ($n = 273$, left), distribution of mean Ki-67 index (mean of six pathologist's readings) per molecular group ($n = 196$, middle), distribution of mean mitotic count (mean of six pathologist's readings) per molecular group ($n = 197$, right). All pairwise comparisons performed with t-tests. * $0.01 < P$ value < 0.05 ; ** $0.001 < P$ value ≤ 0.01 ; *** P value ≤ 0.001 . TPM, transcripts per kilobase million. **b**, Gene expression values of lung NET samples ($n = 273$), plotted as z-score of variance stabilised transformed expression values, for 128 genes highly expressed in lung neuroendocrine cells. **c**, Confusion matrix of molecular group predictions from digital spatial profiling (DSP, y-axis) against ground truth from tumour bulk multi-omic analyses (x-axis) of the areas of interest (AOIs) selected for DSP, by a first random forest model, RF1, trained on all AOIs using five-fold cross-validation. **d**, Confusion matrix of molecular group predictions of AOIs for which the predictions of RF1 matched the molecular group defined at the participant level ($n = 304$ AOIs on the diagonal in panel (c)). These predictions result from a second random forest model, called RF2, trained only on AOIs correctly predicted by RF1 in order to reduce training data noise due to intra-tumour heterogeneity. **e**, Confusion matrix of molecular group predictions (y-axis) of all AOIs according to the RF2 models as a function of the molecular group of the participants (x-axis). **f**, Weighted F1 scores of the three sets of predictions presented in the confusion matrices (c-e).



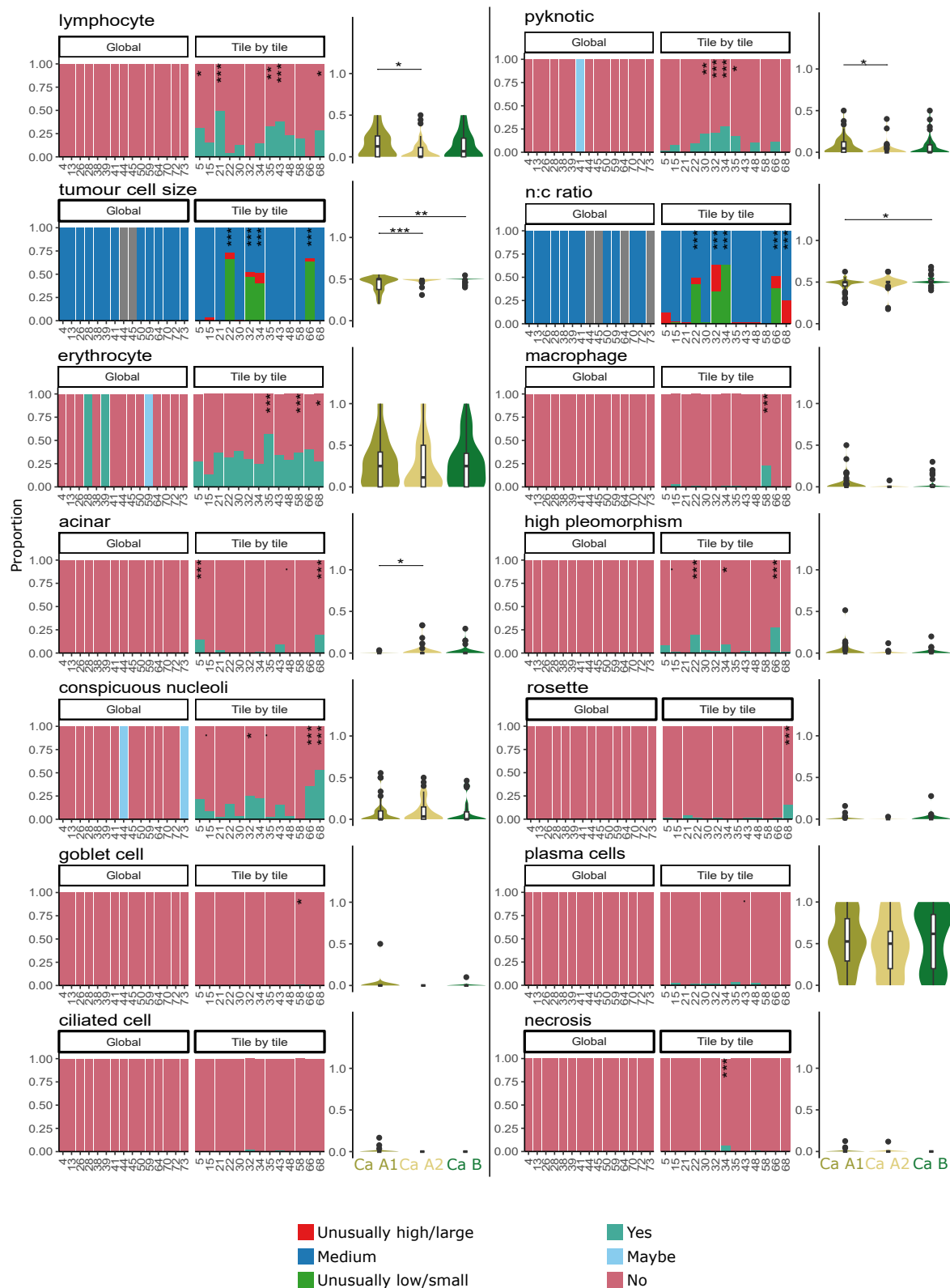
Supplementary Figure S11. Genomic features of lung NETs by histopathological type. a, Violin plot of the distribution of the tumour mutational burden (TMB, number of small variants) in the lung NET cohort (y-axis) as a function of the histopathological classification (x-axis). b, As per (a) for the structural variant burden (SVB, number of structural variants). c, As per (a) for the amplified copy number burden (proportion of the genome amplified). d, As per (a) for the deleted copy number burden (proportion of the genome deleted). In (a-d), significance levels correspond to t-tests. e, Distribution of the number of hallmarks acquired through damaging small variants per sample in typical versus atypical tumours by molecular group, and overall. The overall distribution does not include sc-enriched samples. Comparisons made with Mann-Whitney U tests. For all plots, * $0.01 < P \text{ value} < 0.05$; ** $0.001 < P \text{ value} \leq 0.01$; *** $P \text{ value} \leq 0.001$.



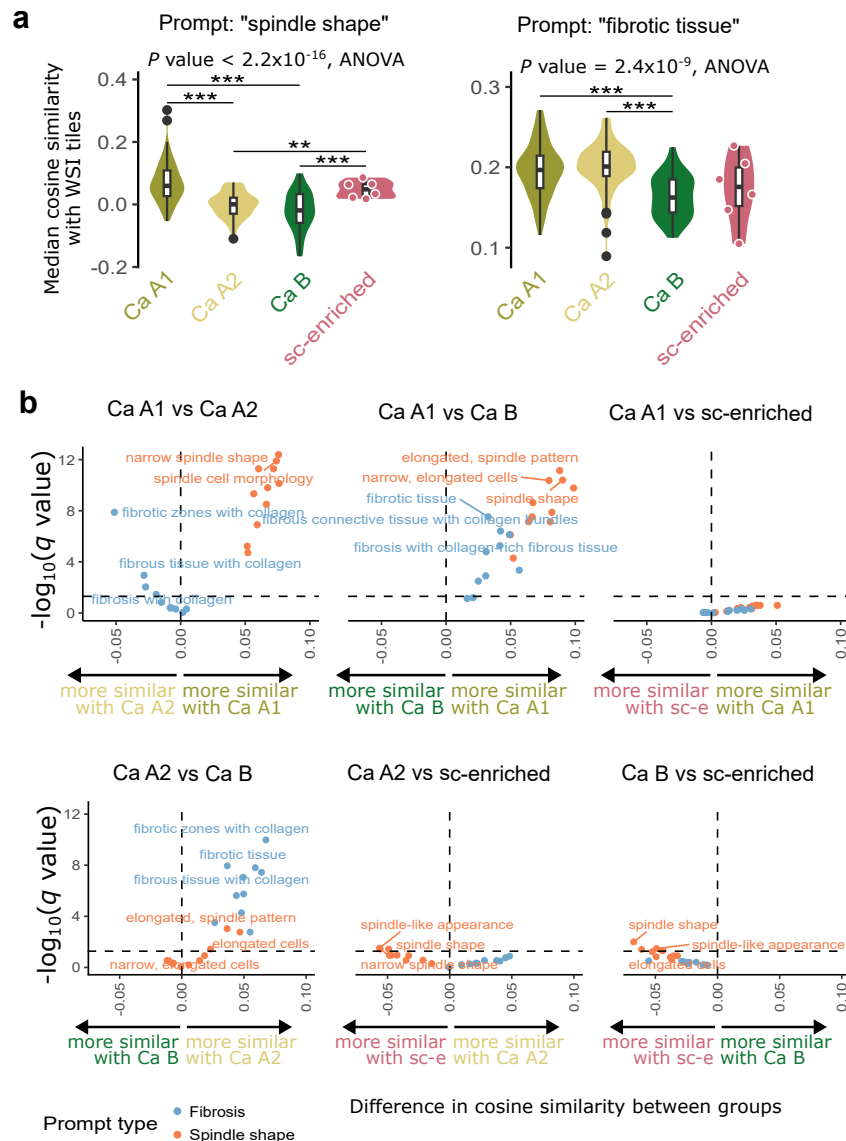
Supplementary Figure S12. Deep learning analyses of lung NETs. **a**, Schematic representation of the deep learning algorithm applied to H&E and H&E/S stained whole-slide images (WSIs). WSIs are divided into smaller regions, referred to as tiles. These tiles are used to train the Barlow Twins algorithm in a self-supervised manner, aiming to generate similar encoded vectors for tiles with common morphological features. The encoded vectors are grouped into morphological partitions using the Leiden clustering algorithm. Tile proportions within each Leiden partition are then used to train random forest models to predict either the tumour histology or the molecular group. A supervised branch, based on RoFormer-MIL, was subsequently employed to identify the most informative tiles for molecular group classification. RoFormer-MIL processes a matrix of encoded vectors from Barlow Twins for each WSI and predicts the molecular group by considering all interactions between the vectors. Attention scores from RoFormer-MIL were used to extract the most relevant tiles, which were clustered using the Leiden algorithm and interpreted by a panel of expert pathologists (Interpretation branch). To generalise the representation of key features across all tiles, cosine similarity scores between tile embedding vectors and text-based prompts were computed using the CONCH visual-language foundation model. **b**, Comparison of deep learning performance for each classification system, measured by weighted F1 score. **c**, Confusion matrix of molecular group predictions by RoFormer-MIL. **d**, Representative tiles of five key features, illustrating the morphological characteristics of the molecular groups. Each tile was independently reviewed by two pathologists, and only tiles where the feature was reported by both pathologists are shown.



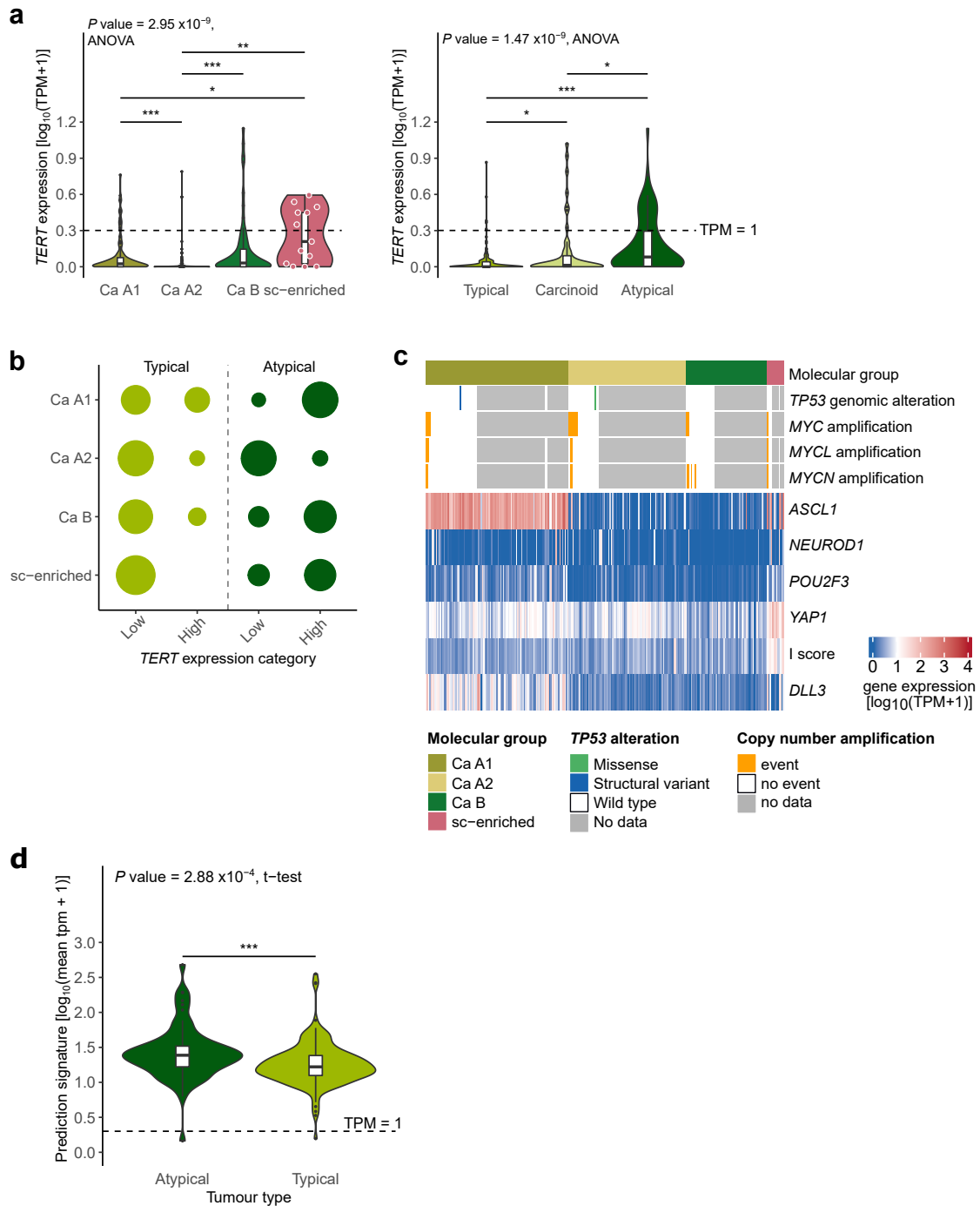
Supplementary Figure S13. Results part A of pathology review to interpret deep learning partition/tile selection. Bar charts display the proportion of partition/tiles classified as 'yes' (green), 'maybe' (blue), or 'no' (red) with regard to the presence of each proposed feature, e.g. spindle cells, (y-axis) per partition (x-axis). Partitions evaluated globally are described in column one, those annotated on a tile-by-tile basis are described in column two. Asterisks in column two indicate significance level (q value) obtained from Fisher's exact tests. * $0.01 < q \text{ value} < 0.05$; ** $0.001 < q \text{ value} \leq 0.01$; *** $q \text{ value} \leq 0.001$. Violin and boxplots in column three indicate the proportion of tiles per molecular group classified as 'yes' and 'maybe' with regard to the presence of the feature. For partitions annotated tile-by-tile, the proportions of 'yes' and 'maybe' were summed, with the 'maybe' level assigned half the weight of the 'yes' level. For globally annotated partitions, 'yes' corresponded to 100% of the tiles displaying the feature, 'maybe' to 50% of the tiles, and 'no' to 0% of the tiles displaying the feature. Asterisks in column three indicate the significance level (q value) obtained from permutation tests performed between groups. * $0.01 < q \text{ value} < 0.05$; ** $0.001 < q \text{ value} \leq 0.01$; *** $q \text{ value} \leq 0.001$.



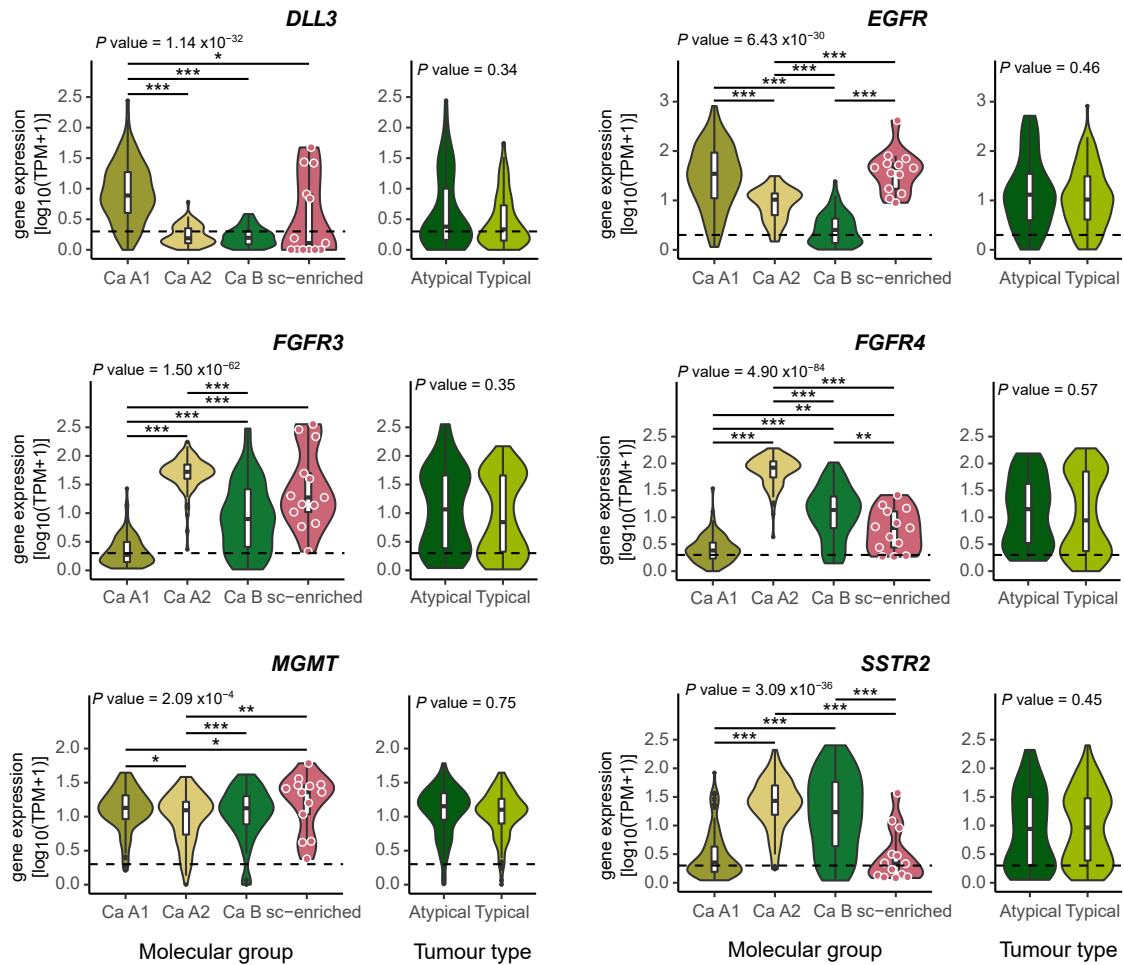
Supplementary Figure S14. Results part B of pathology review to interpret deep learning partition/tile selection. As per Supplementary Figure S13, for different features. N:C ratio, nucleus to cytoplasm ratio. Classification labels for N:C ratio and tumour cell size differ to all other characteristics evaluated. Labels were only assigned to partitions/tiles if both pathologists agreed, otherwise the partition/partition was marked as NA (grey).



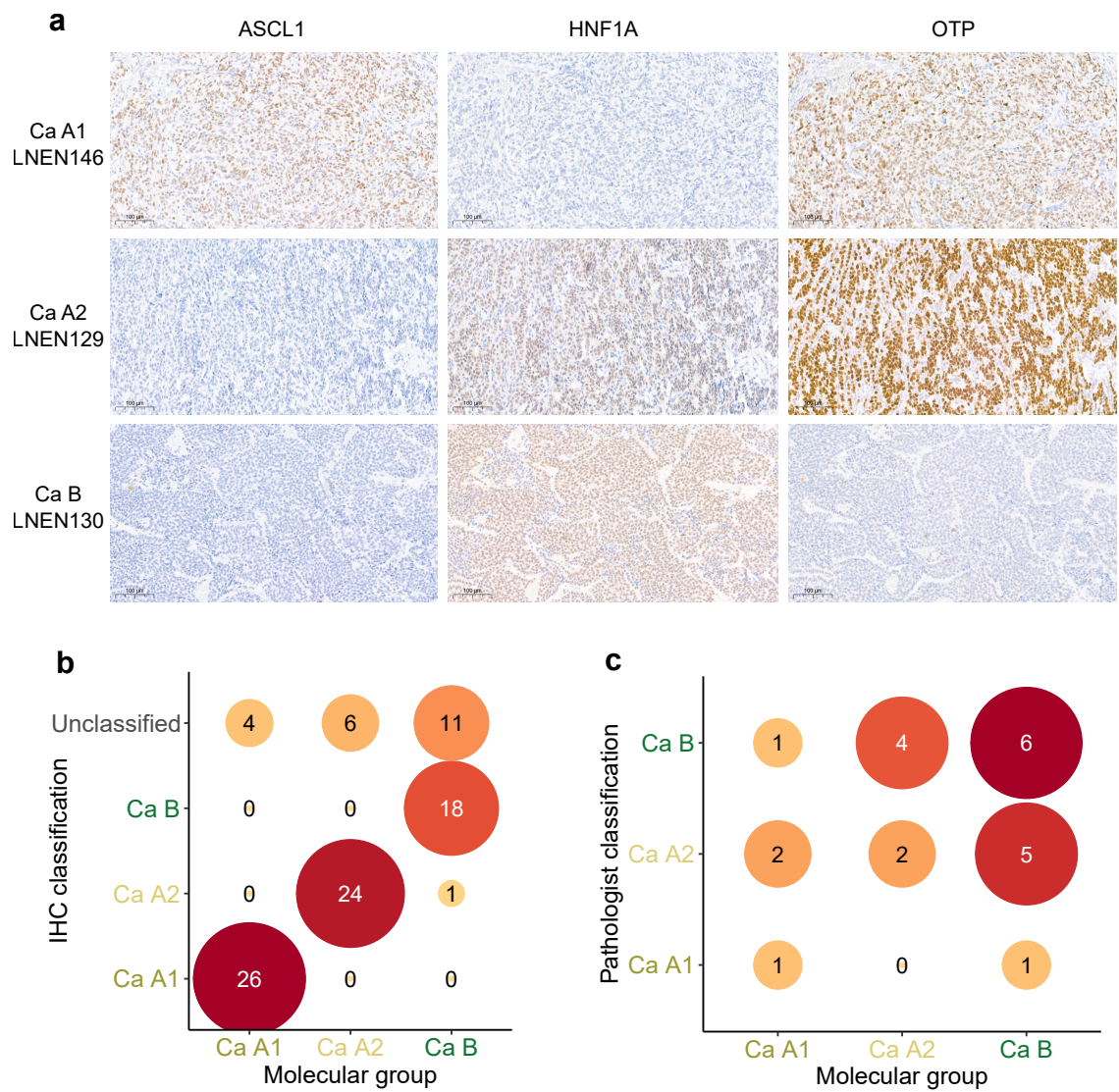
Supplementary Figure S15. Results of CONCH visual-language foundation model analyses. **a**, Distribution of median cosine similarity scores per WSI for the text prompts indicated above the figure, by molecular groups (x-axis). **b**, Robustness analysis of the results presented in (a) using alternative prompts. Volcano plots represent the statistical significance (q value, y-axis) versus the effect size (x-axis) of the differences between molecular groups in terms of cosine similarity between CONCH WSI tile embeddings and text prompt embeddings, for different prompts (points).



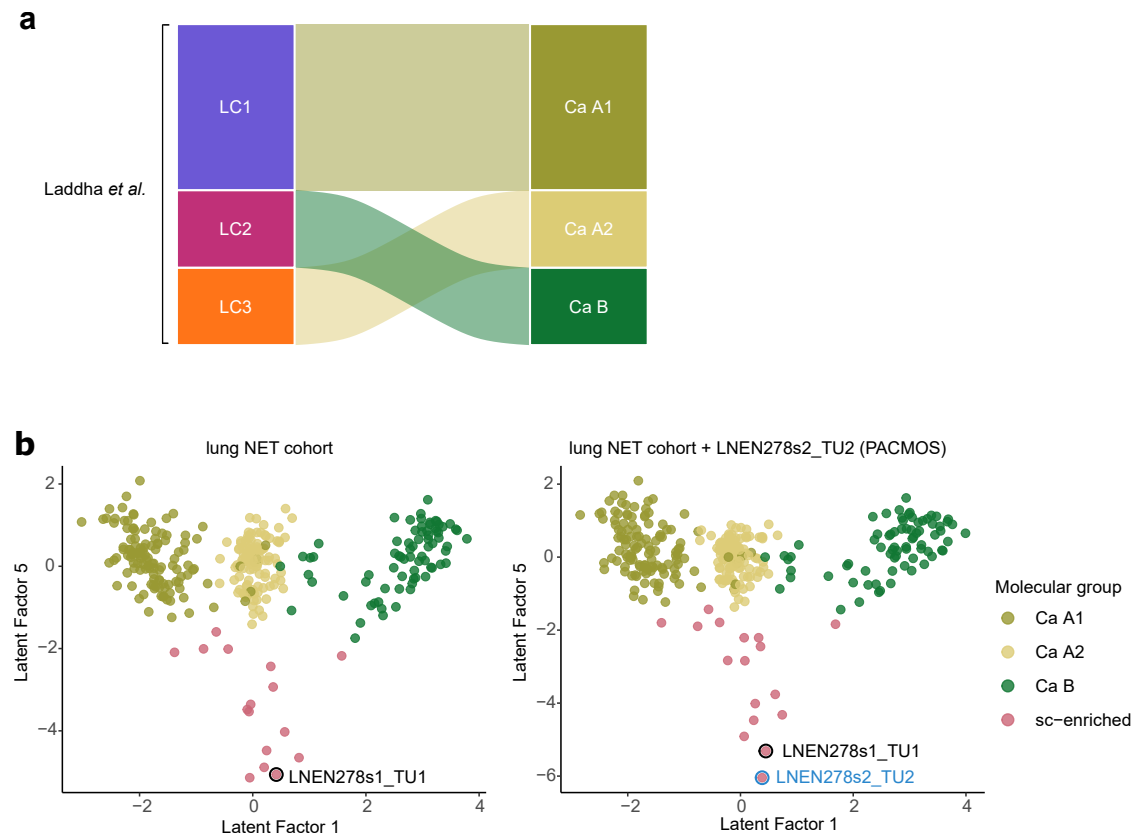
Supplementary Figure S16. Comparison of lung NET molecular groups with lung NEN molecular markers. **a**, Distribution of *TERT* expression by molecular group (left), and by tumour type (right). **b**, Bubble chart displaying the proportion of typical and atypical tumours, respectively, within each molecular group to be classified as *TERT* low or *TERT* high. **c**, Diagram of characteristic SCLC genomic variants and heatmap of gene expression of SCLC molecular group markers in $\log_{10}(\text{TPM}+1)$ units, for lung NETs with gene expression data ($n = 273$). Coloured bars along the top represent sample molecular group and genomic alteration status. Immune score (I score) is the $\log_{10}(\text{mean}(\text{TPM}+1))$ of selected immune genes. **d** Distribution of mean expression (in $\log_{10}(\text{TPM}+1)$) of a panel of 18 genes that capture a T-cell inflamed phenotype and is associated with response to PD-1 inhibitor pembrolizumab by histological type. $n = 223$, TPM, transcripts per kilobase million. Pair-wise comparisons performed with t-tests. * $0.01 < P$ value < 0.05 ; ** $0.001 < P$ value ≤ 0.01 ; *** P value ≤ 0.001 .



Supplementary Figure S17. Expression of clinically relevant gene markers by lung NET molecular group and histological type. Violin plots displaying the distribution of expression by molecular group ($n = 273$), and by histological type ($n = 223$), in log₁₀(TPM+1) units, dashed line represents TPM (transcripts per kilobase million) = 1. P values displayed on molecular group panels correspond to results of ANOVA, all pair-wise comparisons performed with t-tests. * $0.01 < P \text{ value} < 0.05$; ** $0.001 < P \text{ value} \leq 0.01$; *** $P \text{ value} \leq 0.001$.



Supplementary Figure S18. IHC and morphological classifications. **a**, Representative images of one patient slide per molecular group (rows) stained for ASCL1, HNF1A and OTP (columns), to identify molecular groups Ca A1, Ca A2, and Ca B. **b**, Confusion matrix of the performance of the immunohistochemistry (IHC) protein classification for molecular groups, $n = 90$. **c**, Confusion matrix of the performance of the morphological classification for molecular groups, performed on the $n = 22$ tumour images that were unclassified or incorrectly classified in (b).



Supplementary Figure S19. Projection and classification of additional lung NETs with PACMOS. a, Alluvial plot comparing gene expression-based classification of lung NETs derived from Laddha *et al.* (LC1, LC2, LC3) with the proposed molecular classification in the current manuscript. Molecular group was determined by PACMOS for $n = 29$ publicly available lung NETs from Laddha *et al.*, box sizes correspond to sample sizes. b, 319 tumours from the lung NET cohort plotted over MOFA Latent Factors 1 and 2, coloured by molecular group as identified by ParetoTI, with LNEN278s1_TU1 circled (left, figure is identical to **Figure 1a**). 320 tumours from the lung NET cohort plus LNEN278s2_TU2, plotted over MOFA Latent Factors 1 and 2, coloured by molecular group, as identified through PACMOS analysis, with LNEN278s1_TU1 and LNEN278s2_TU2 (resected two years after LNEN278s1_TU1) circled (right).