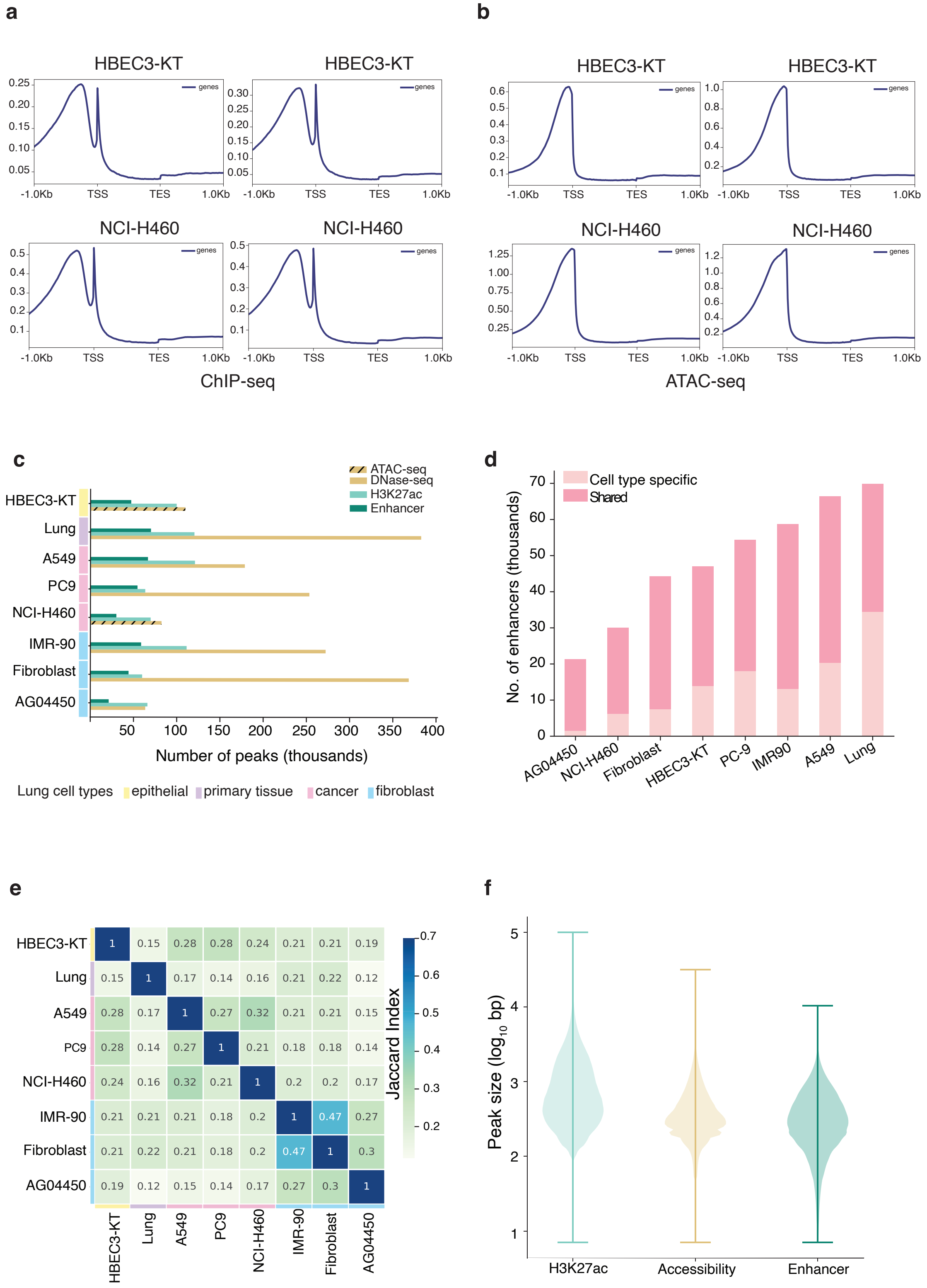
Enhancer somatic mutations with functional impact in lung cancers identified by leveraging the tissue-specific enhancer-target genes regulatory network.

Judith Mary Hariprakash, Elisa Salviato, Federica La Mastra, Endre Sebestyén, Ilario Tagliaferri, Raquel Sofia Silva, Federica Lucini, Lorenzo Farina, Mario Cinquanta, Ilaria Rancati, Mirko Riboni, Simone Paolo Minardi, Luca Roz, Francesca Gorini, Chiara Lanzuolo, Stefano Casola, Francesco Ferrari

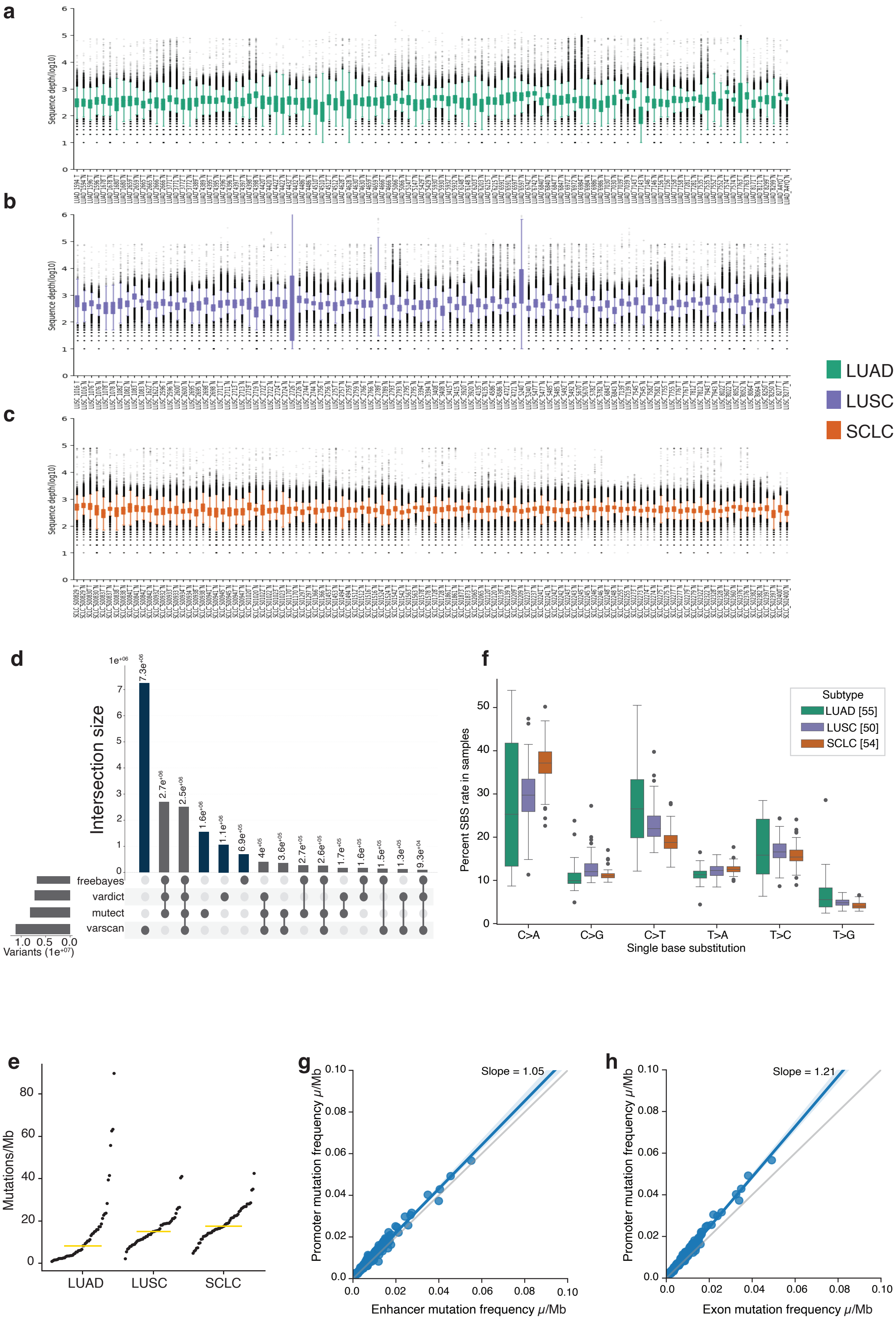
Supplementary data



	HBEC3-KT	Lung	A549	PC9	NCI-H460	IMR-90	Fibroblast	AG04450
HBEC3-KT	1	0.15	0.28	0.28	0.24	0.21	0.21	0.19
Lung	0.15	1	0.17	0.14	0.16	0.21	0.22	0.12
A549	0.28	0.17	1	0.27	0.32	0.21	0.21	0.15
PC9	0.28	0.14	0.27	1	0.21	0.18	0.18	0.14
NCI-H460	0.24	0.16	0.32	0.21	1	0.2	0.2	0.17
IMR-90	0.21	0.21	0.21	0.18	0.2	1	0.47	0.27
Fibroblast	0.21	0.22	0.21	0.18	0.2	0.47	1	0.3
AG04450	0.19	0.12	0.15	0.14	0.17	0.27	0.3	1

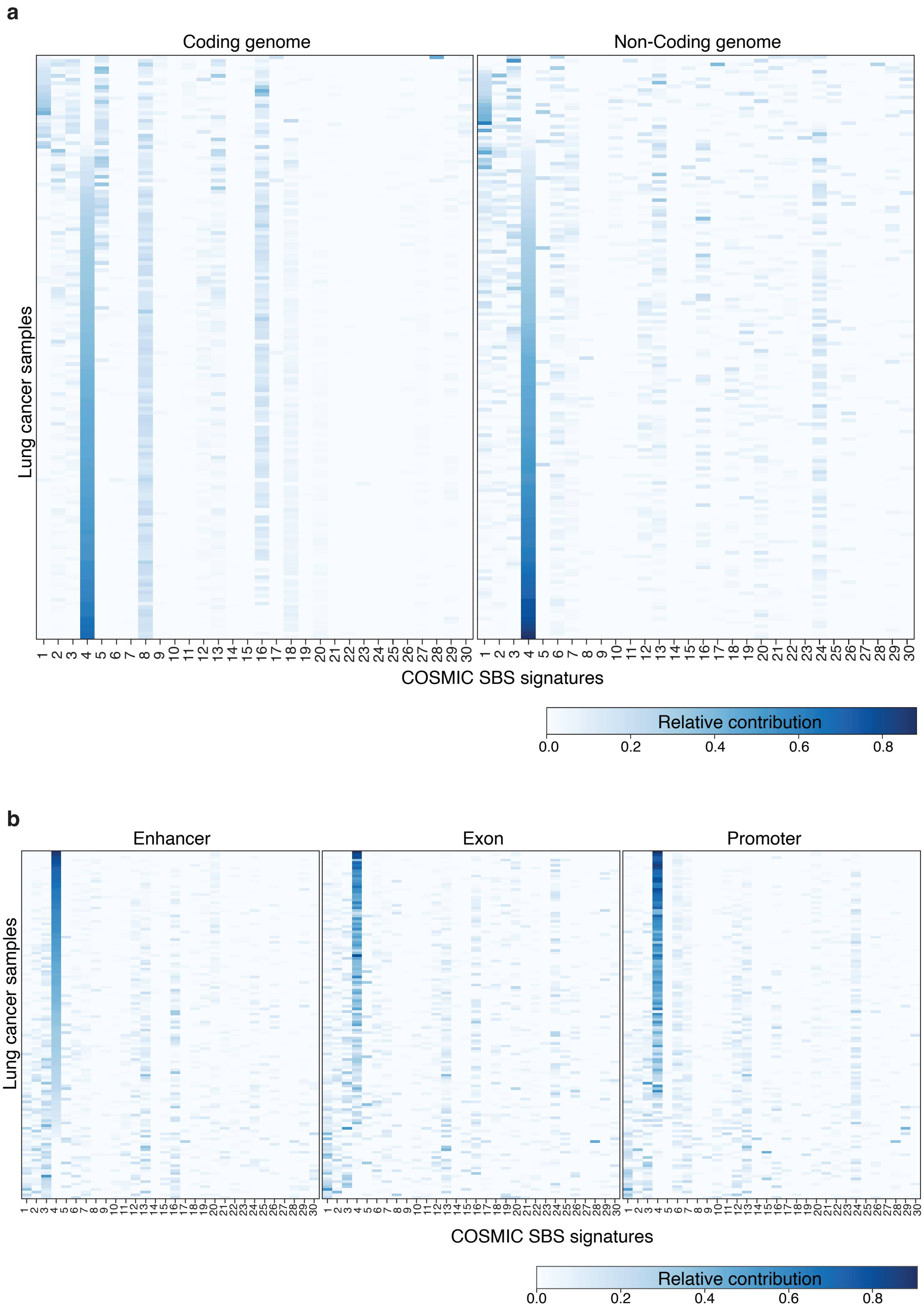
Supplementary Figure S1. Definition of lung-specific enhancer catalogue.

a) Distribution of ChIP-seq reads around the transcription start site (TSS) generated in-house for HBEC3-KT (top panel) and NCI-H460 (bottom panel) cell lines. **b**) Distribution of ATAC-seq reads around the transcription start site (TSS) generated in-house for HBEC3-KT (top panel) and NCI-H460 (bottom panel) cell lines. The TSS plot shows the accumulated view of the distribution of sequence reads related to the closest annotated gene. All annotated genes have been normalized to the same size. The colour blocks denote the annotation of the genomic region, *i.e.*, green: -3.0Kb upstream of gene to TSS yellow: TSS to TES (transcription end site) and pink: TES to 3.0Kb downstream of gene. **c)** Number of cell type-specific enhancer regions (dark cyan) resulting from the intersection of H3K27ac ChIP-seq peaks (light cyan) and chromatin accessibility (light yellow) obtained from DNase-seq or ATAC-seq (striated) in the cell and tissue types used for lung-specific enhancer cataloguing. The nature of the cell or tissue type is represented as coloured boxes along the Y-axis, such as primary tissue (lilac), cancer cell line (pink), epithelial cell line (yellow) and fibroblast (blue). **d)** The stacked barplot shows the cell-specific enhancers (lighter colour shade) and those shared by at least another sample (darker colour shade) for each cell or tissue sample used to define the reference catalogue enhancers. **e**) Similarity (Jaccard index on overlap) among cell-specific set of enhancers. Each square in the heatmap represents the ratio between the intersection of two cell-specific sets of enhancers over their union. The cell or tissue type are mentioned in coloured boxes along the axis. **f)** Violin plots representing size distributions (Y-axis in log scale) of H3K27ac, chromatin accessibility and enhancer peaks.

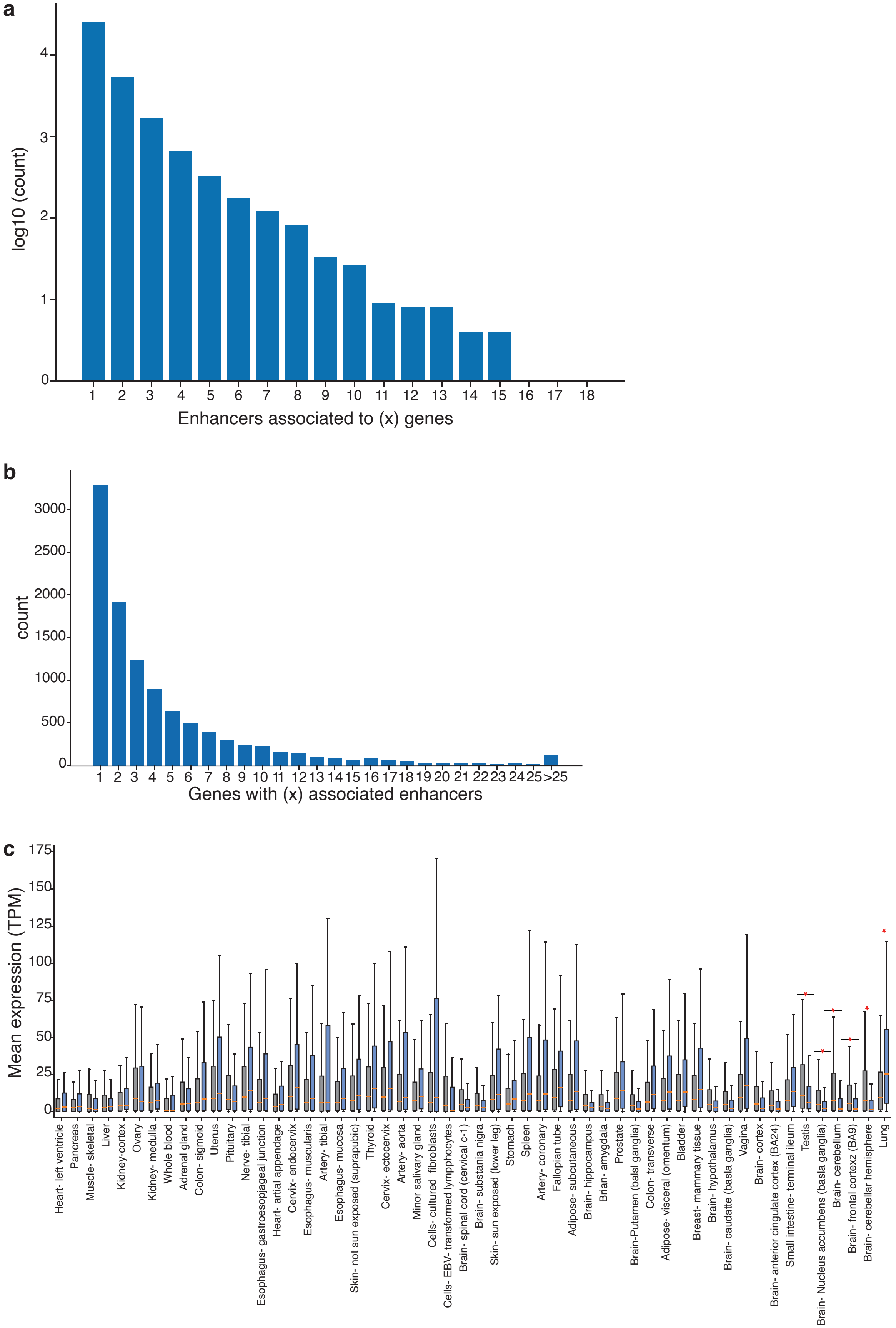


Supplementary Figure S2. Mutational landscape of lung cancer cohort.
Whole genome sequence coverage plot of **a)** LUAD **b)** LUSC and **c)** SCLC samples. Box and whiskers plots show the distribution of the number of reads for each mutation called in the sample. **d**) Ensemble mutation calling. UpsetR plot of the 4 mutation callers *viz*., Varscan, Mutect (including Scalpel), Vardict and Freebayes. The left horizontal bars show the number of somatic mutations (substitutions and small indels) called by each tool. The vertical bars

show the number of mutations in each intersection of sets specified by dark circles. Mutations called by only one tool (Dark blue bars) were removed from further analysis. **e)** Mutational burden of LUAD, LUSC and SCLC. The median mutation burden (mutations/Mb) is shown as a dot plot (SNVs and small indels); yellow bars denote the median burden of all samples. **f)** Single base pair substitution (SBS) rates are reported for each individual patient in the three lung cancer subtypes as a boxplot. Each data point in the boxplots correspond to one patient (n indicated in the colour legend). A separate boxplot is shown for each substitution type as indicated in the x-axis labels. In each box, the horizontal line marks the median, the box margins mark the interquartile range (IQR), the whiskers extend up to 1.5 IQR and individual outliers beyond this range are marked. **g)** Mutation burden comparisons. Scatter plots showing the mutation burden per Mb in enhancers (X-axis) and promoters (Y-axis). **h)** Mutation burden per Mb in exons (X-axis) and promoters (Y-axis). Each blue dot in scatter plots represents a sample, the grey line represents the bisectors, the blue line represents the line of regression, and the slope of the regression is mentioned in the plot.

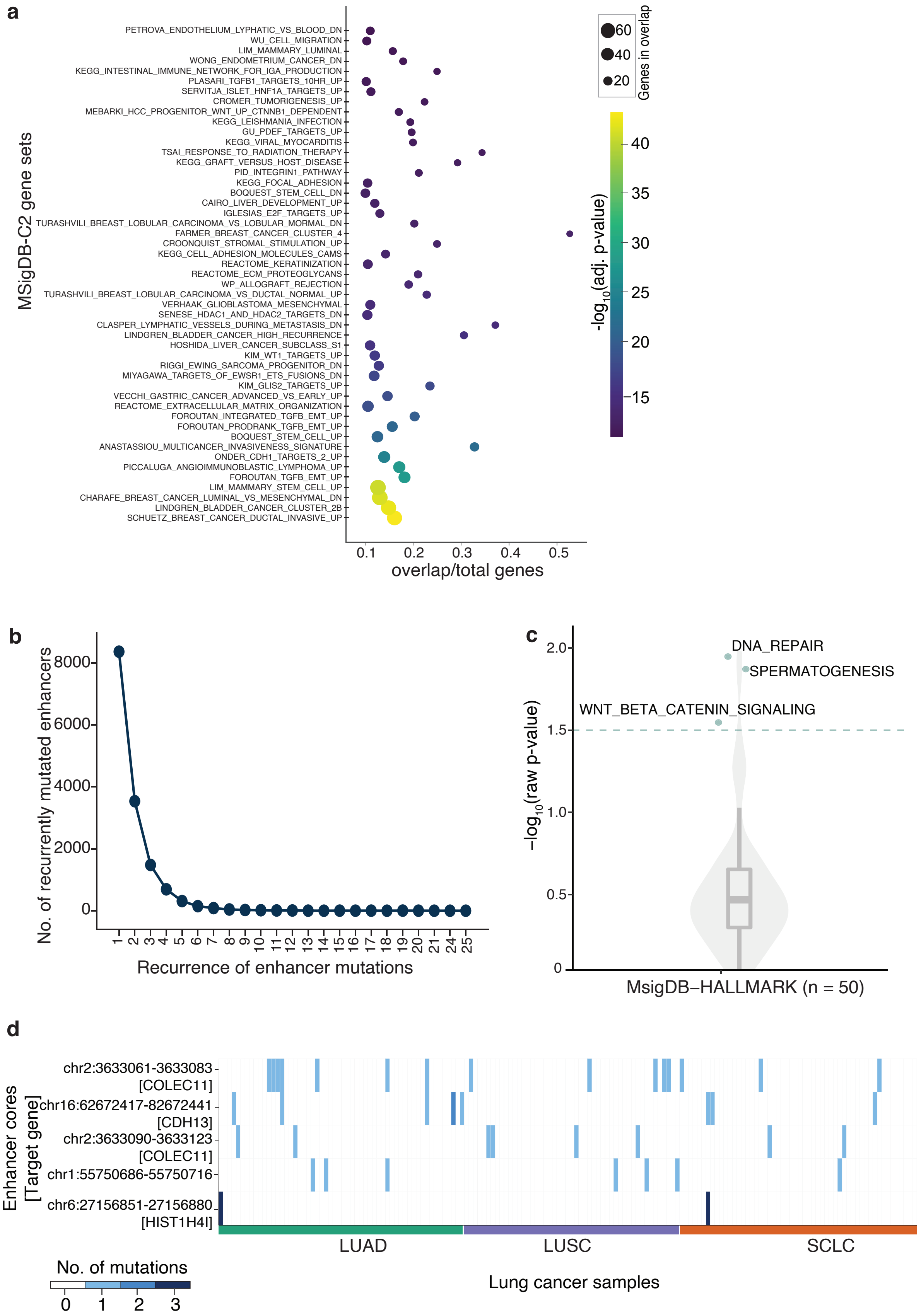


Supplementary Figure S3. Mutational signature differences between the coding and non-coding genome. **a)** Heatmap of the relative contribution of each COSMIC single base substitutions (SBS) signature for each sample in the coding and non-coding genome. The samples are ordered based on the relative contribution of signature 4. **b)** Heatmap of the relative contribution of each COSMIC single base substitutions (SBS) signature for each sample in the enhancer, exon and promoter. The samples are ordered based on the relative contribution of signature 4.



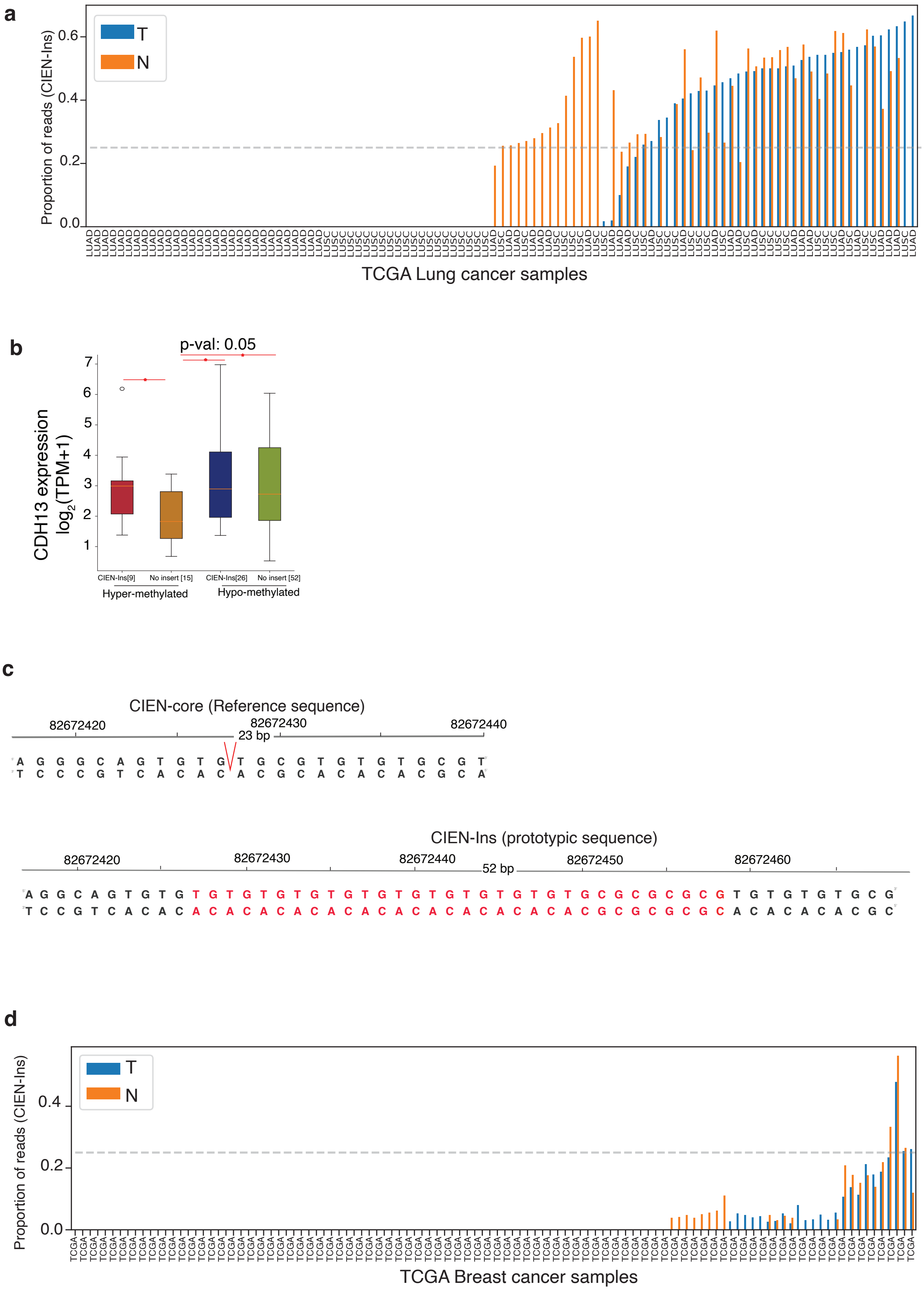
Supplementary Figure S4. Enhancer target gene pairing.

a) Number of target genes per enhancer. The bar-plot shows the frequency (log10 count on the Y-axis) of enhancers grouped by the number of target genes (X-axis) they are associated to. **b)** Number of enhancers per target gene. The bar-plot shows the frequency (count on the Y-axis) of genes grouped by the number of associated enhancers (X-axis). **c)** Tissue specific gene expression. Box-plot shows the expression in TPM in GTEX tissues (X-axis labels) for genes with at least 25 lung-specific enhancers (blue) to a set of background genes with fewer enhancers (grey). The red star indicates the significant difference between the two groups of genes (Mann–Whitney U test p-value < 0.05). The horizontal orange bars denote the median. (see methods)



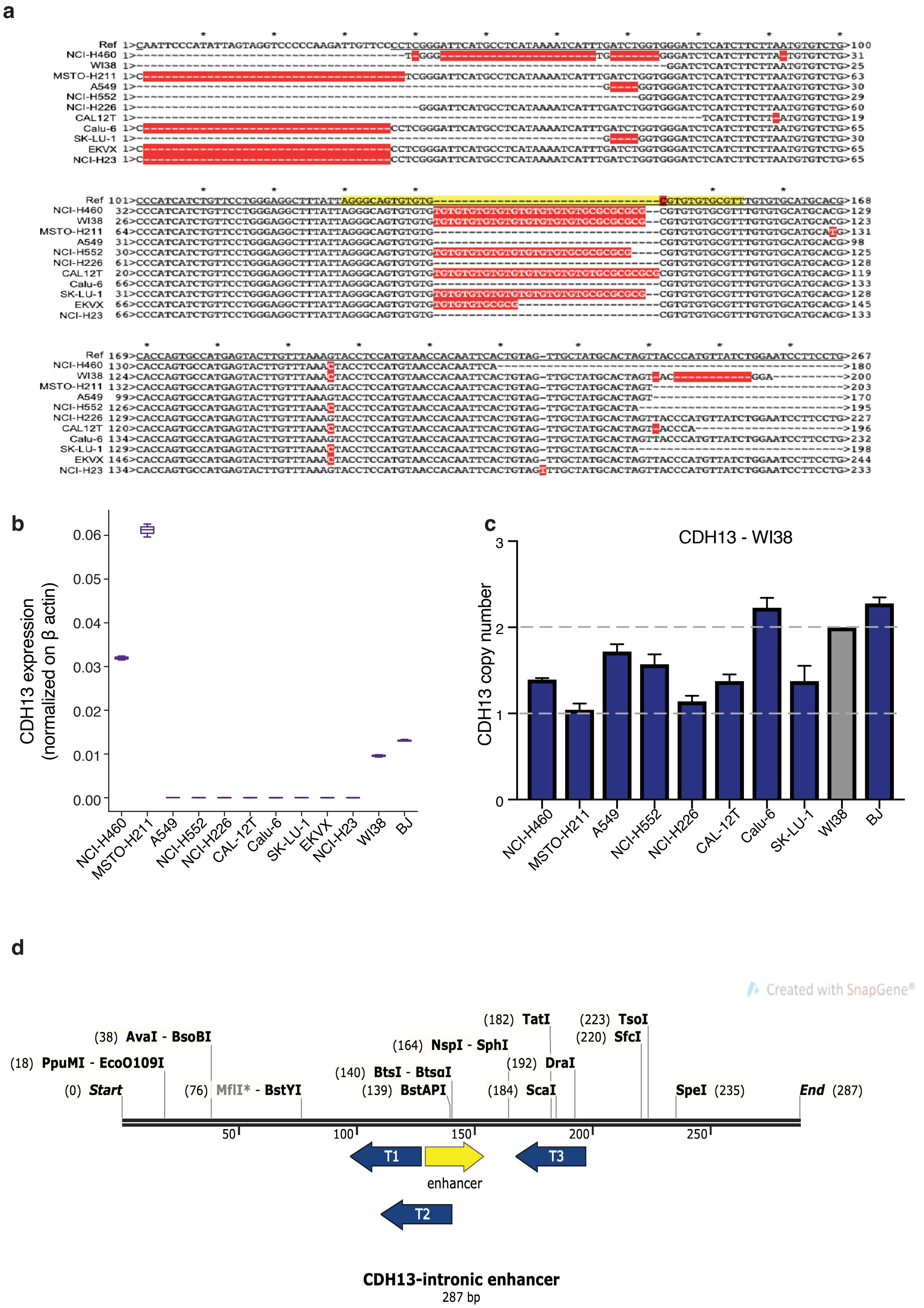
Supplementary Figure S5. Pathway level enrichment of enhancer mutations.
Scatter plot shows the over-representation of TGEM in **a)** MSigDB curated gene set X-axis represents the ratio of the overlapping TGEM to the total number of genes in the pathway. The size of the circle denotes the number of TGEM in overlap and the colour shows the negative logarithmic adjusted p-value. **b)** Recurrently mutated enhancer. Line plot depicts the number of recurrently mutated enhancers (Y-axis) and the number of recurrence (X-axis). **c)** Significantly altered gene sets (MSigDB – Hallmark). Violin plot shows the gene sets that are impacted by TGEM.

d) Recurrently mutated enhancer cores in lung cancer cohort. Co-mutation plot shows the top 5 recurrently mutated enhancer cores. Predicted target genes of the enhancers are mentioned in brackets. Dark blue bar indicates the presence of a mutation in the enhancer core in the sample (X-axis), samples are grouped into lung cancer subtypes.



Supplementary Figure S6. Exploration of CIEN-Ins.

a) CIEN-Ins in TCGA lung cancer. Bar-plot shows the proportion of reads corresponding to CIEN-Ins in tumour (blue) and normal (orange) WGS data in TCGA lung cancer cohort. Grey horizontal line marks the threshold used for classification of presence of variant. **b**) Effect of CIEN-Ins and methylation of CDH13 promoter. Box-plot showing the CDH13 gene expression in TCGA samples stratified based on the insertion mutation in tumor (CIEN-Ins) and methylation at the promoter of CDH13 gene (beta value). **c)** Transcription factor motif alteration at CIEN-core. The top panel represents the TF motifs observed at the CDH13 enhancer core with the reference sequence. Red V in the reference sequence represents the location of CIEN-Ins. The bottom panel represents the TF motifs present at CDH13-Ins (red section within the sequence). The square's position indicates the start of the motif, and the colour represents the transcription factor, as indicated in the legend. **d**) CIEN-Ins in breast cancer cohort. Bar-plot shows the proportion of reads corresponding to CIEN-Ins in tumour (blue) and normal (orange) WGS data of breast cancer cohort.



Supplementary Figure S6. Cell line screening for experimental validation

a) Sanger sequence of CIEN-core in lung cell lines. The sequence of CIEN- core in 11 cell lines with respect to reference genome is depicted in three consecutive group of rows. The region of CIEN-Ins is highlighted in yellow in the reference genome. CIEN-Ins if present in cell lines is highlighted in red. **b**) CDH13 gene expression normalised with beta-actin expression in lung cell lines. **c**) CDH13 gene copy number in lung cell lines normalised with GAPDH copy number with respect to WI38 (a normal lung fibroblast cell line). **d**) Map of the CIEN-core locus with CRISPR Cas9 shRNA guides. Yellow arrow indicates the enhancer core region, blue arrows indicate the Cas9 shRNA guides. Figure generated using SnapGene.