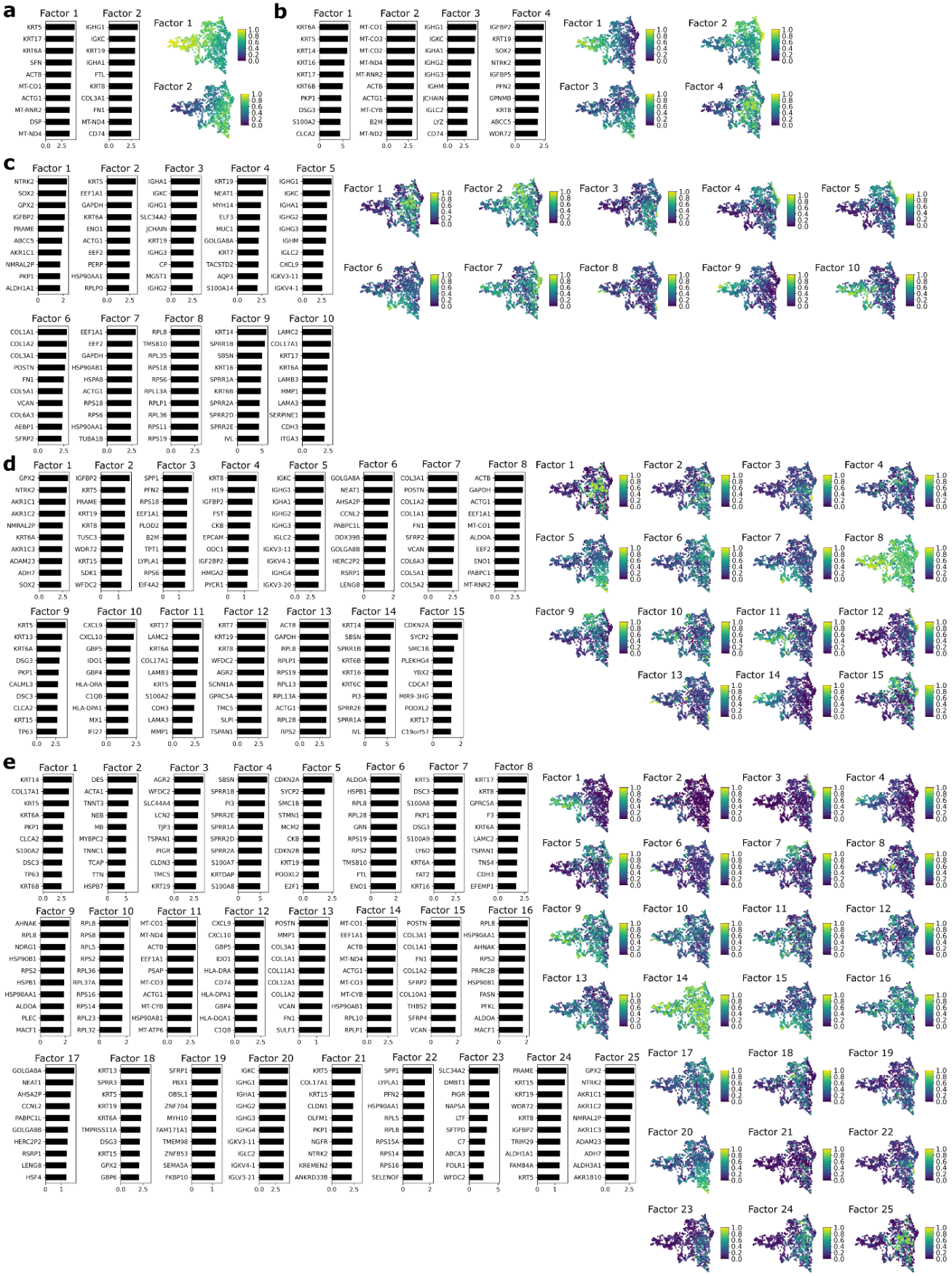

Figure S1: Selection of the number of CLFs. a, For a CoGAPS model with k=2 CLFs, for each factor, a bar chart shows weights (x axis) of top-weighted genes (y axis), and a UMAP (as in Fig. 1c), created from TCGA RNA-seq data, shows samples colored by CLF weight. b–e, Analogous plots for models with varying k: k=4 (**b**), k=10 (**c**), k=15 (**d**), and k=25 (**e**).

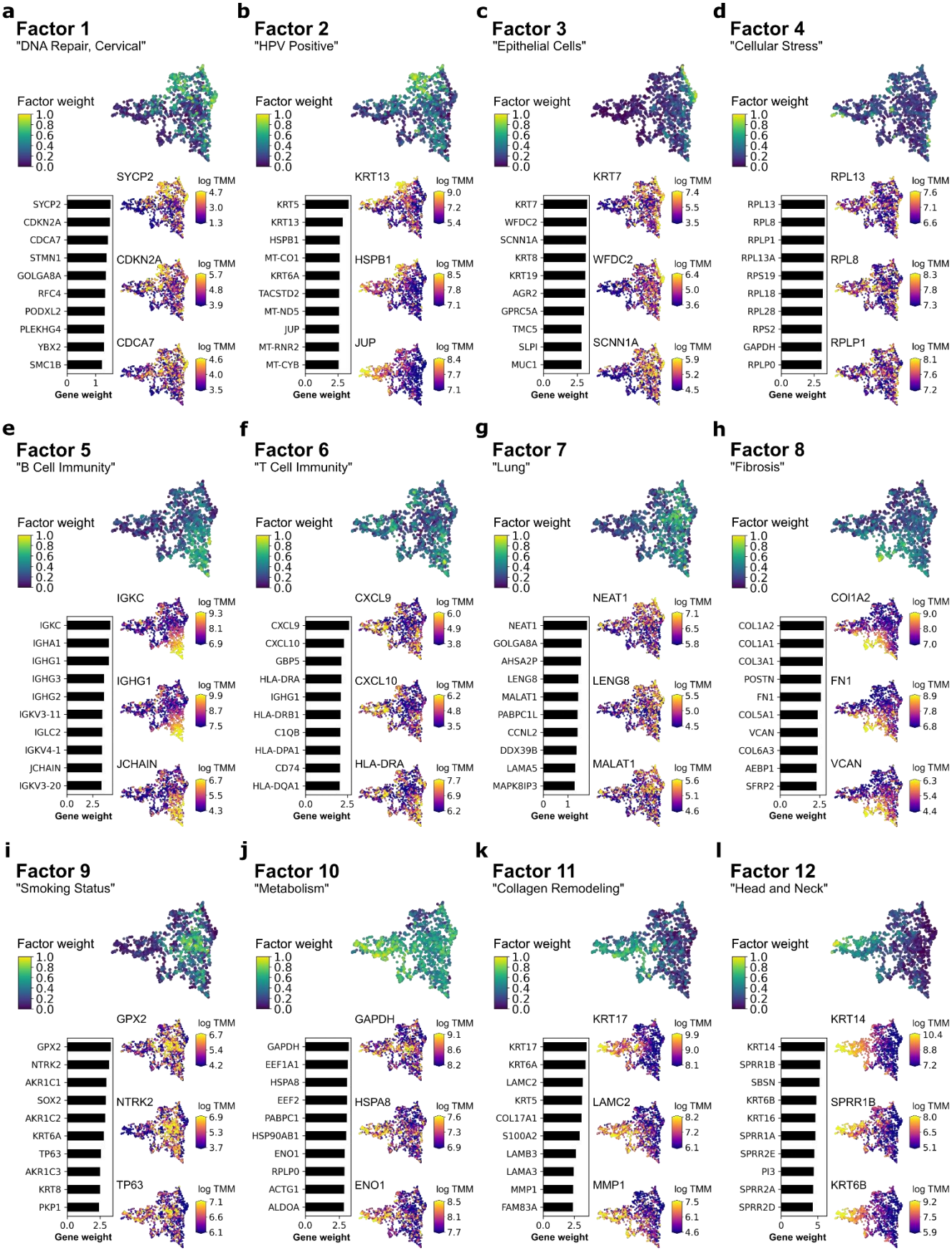

Figure S2: CLF genes are associated with salient features of SCC biology. a, For Factor 1, UMAPs (as in Supp. Fig 1) show patient samples (dots) colored by factor weight (top) or by expression (log TMM) of 3 genes with high CLF weights (right column); bar plot (left) shows top 10 genes (y axis) by factor weight (x axis). **b–l**, Analogous plots are shown for Factors 2–12, respectively.

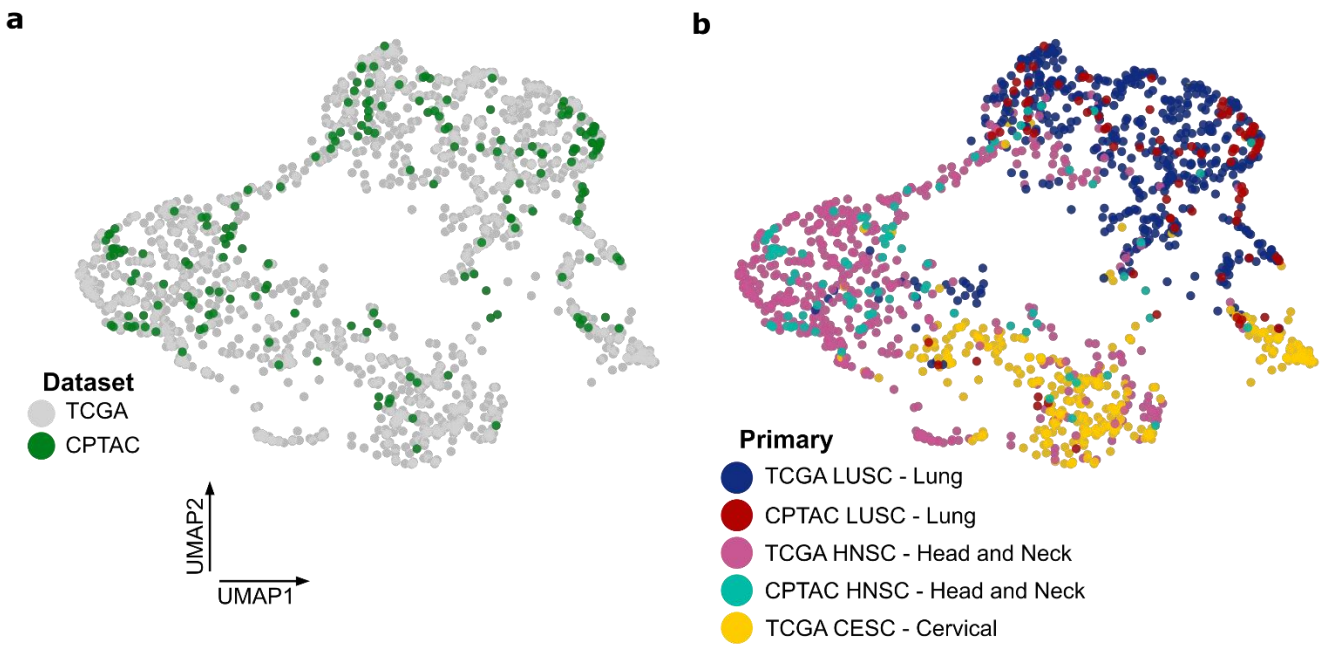

Figure S3: CPTAC data were meaningfully projected into the latent factor space inferred using TCGA data. a,b, A UMAP embedding was created from the CLF coordinates of TCGA patient samples, as inferred by CoGAPS, and of CPTAC patients, as obtained by projecting their RNA-seq profiles into the TCGA-defined CLF space. Samples (dots) are colored by dataset of origin (**a**), or by both primary cancer type and dataset (**b**).

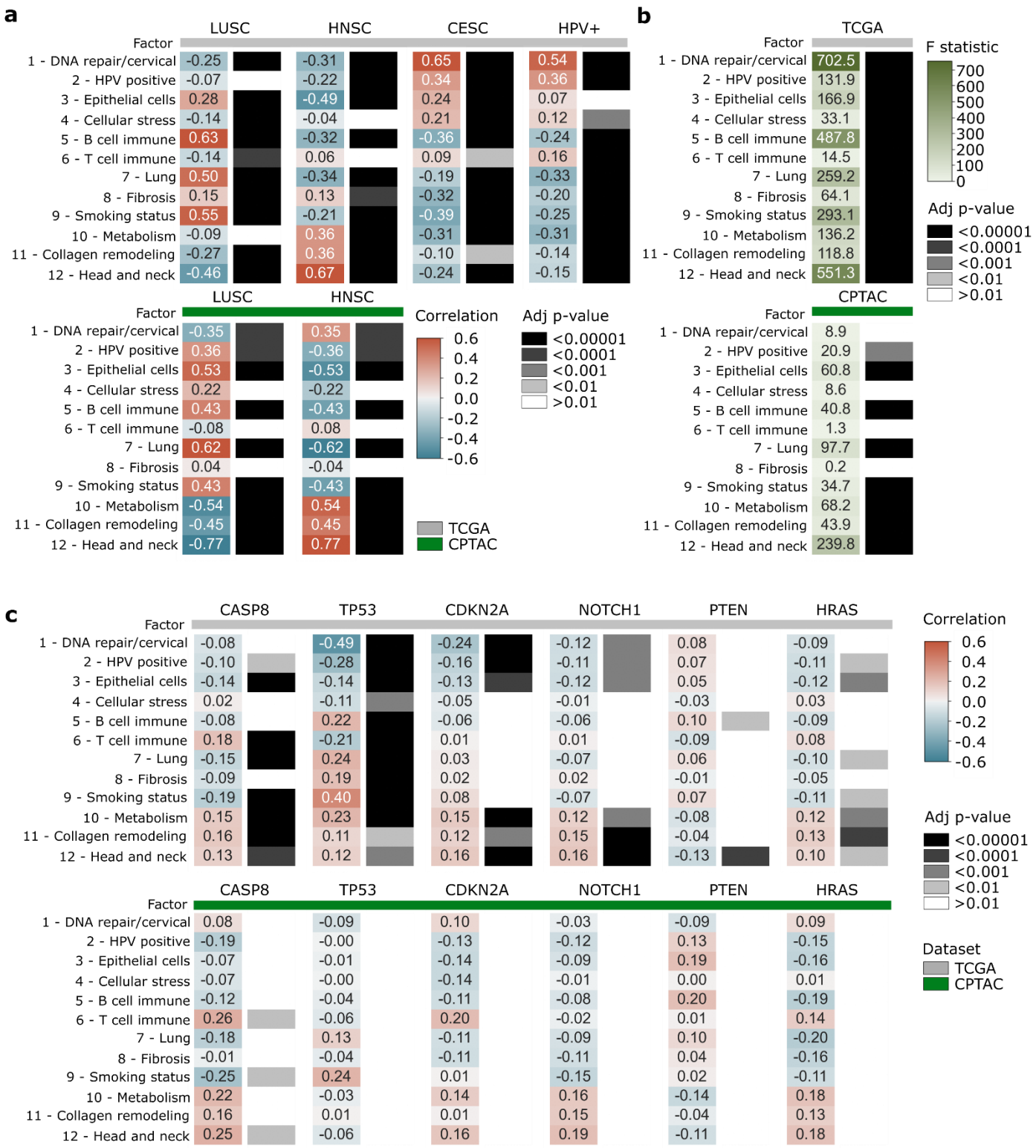

Figure S4: Transcriptomic latent space factorization highlighted primary-specific and shared patterns of variation. a, Heatmap shows Spearman's rank correlation coefficient (color) between sample weights for each latent factor (row) and primary-cancer or HPV-status group (headings annotating column pairs), paired with Bonferroni-corrected p-values (gray scale), in TCGA (top) and CPTAC (bottom) datasets. **b,** Heatmap shows the variation explained (ANOVA F statistic, color) by primary cancer type in the weights of each latent factor (row), paired with Bonferroni-corrected p-values (gray scale); darker color indicates more tissue-specific variation. **c,** Heatmap shows Spearman's rank correlation coefficient (color), paired with Bonferroni-corrected p-value (gray scale), between sample weights of each latent factor (row) and the mutation status for various genes (headings annotating column pairs) in TCGA (top) and CPTAC (bottom) datasets.

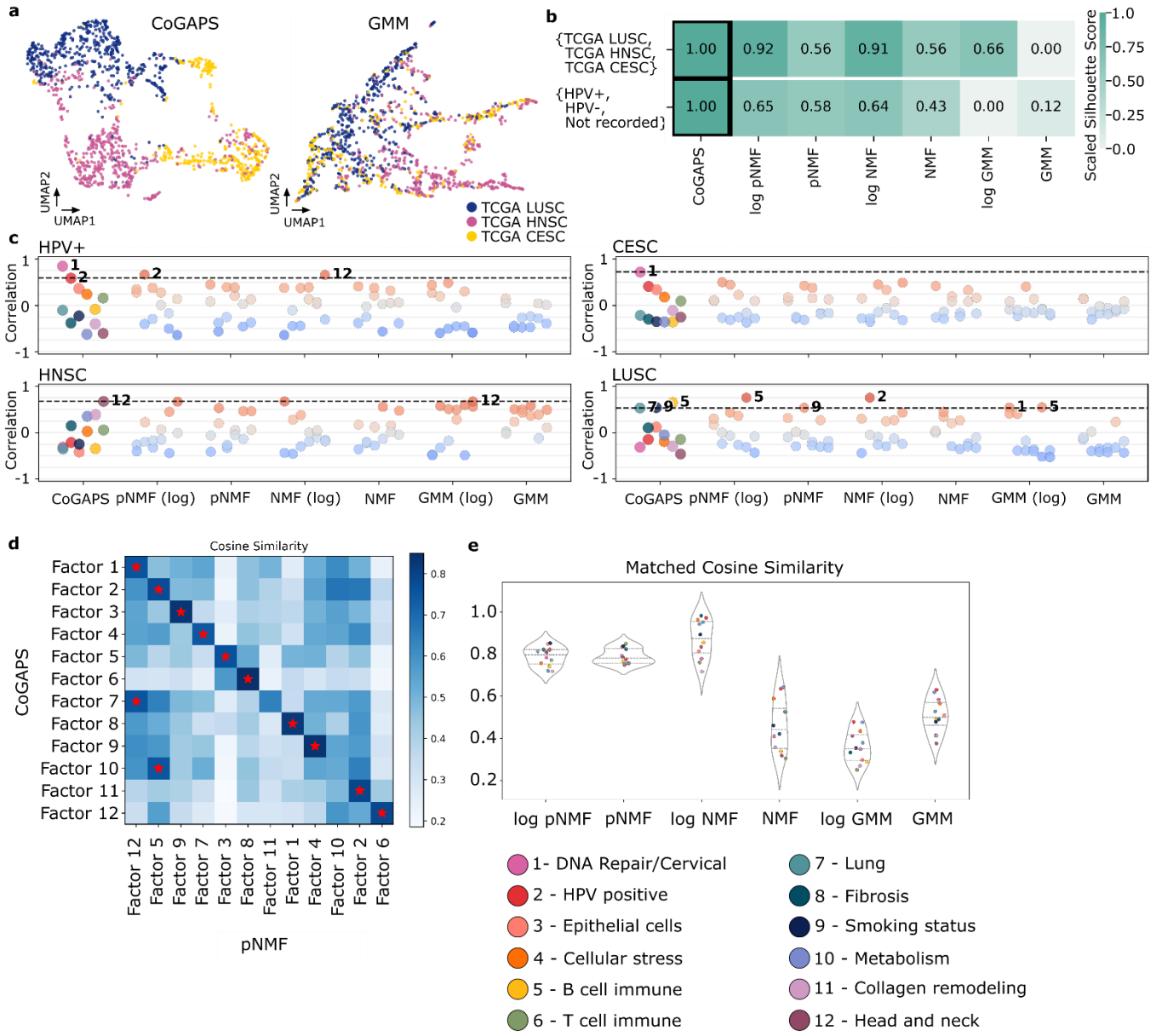

Figure S5: Comparison in SCCs of CoGAPS with alternative soft clustering methods. a, UMAP plots of CoGAPS and GMM embeddings, colored by organ of origin. **b**, Heatmap of row-scaled silhouette scores for organ of origin (top) and HPV status (bottom) as cluster labels. Black box: highest silhouette score per row. **c**, Spearman's rank correlation coefficient (y axis) of HPV status or organ-of-origin label (panel title) with the weights of each latent factor (dots) for each method (x axis). Color indicates factor identity for CoGAPS and correlation strength otherwise. Black dashed line indicates the correlation for the CLF with the associated biological function; all factors with at least this correlation value are labeled by index. **d**, Heatmap shows cosine similarity (color) between each CLF (row) and Poisson NMF (pNMF) factor (column). Red stars indicate highest-similarity match for each CLF. Multiple red stars in one column indicate a pNMF factor with high similarity to multiple CLFs. **e**, Violin plots show the distribution of cosine similarities between each CLF and its best matched factors from alternative methods (x axis) pNMF, NMF, and Gaussian mixture models (GMM), each applied to normalized or log-scaled normalized (indicated by "log" prefix) data.

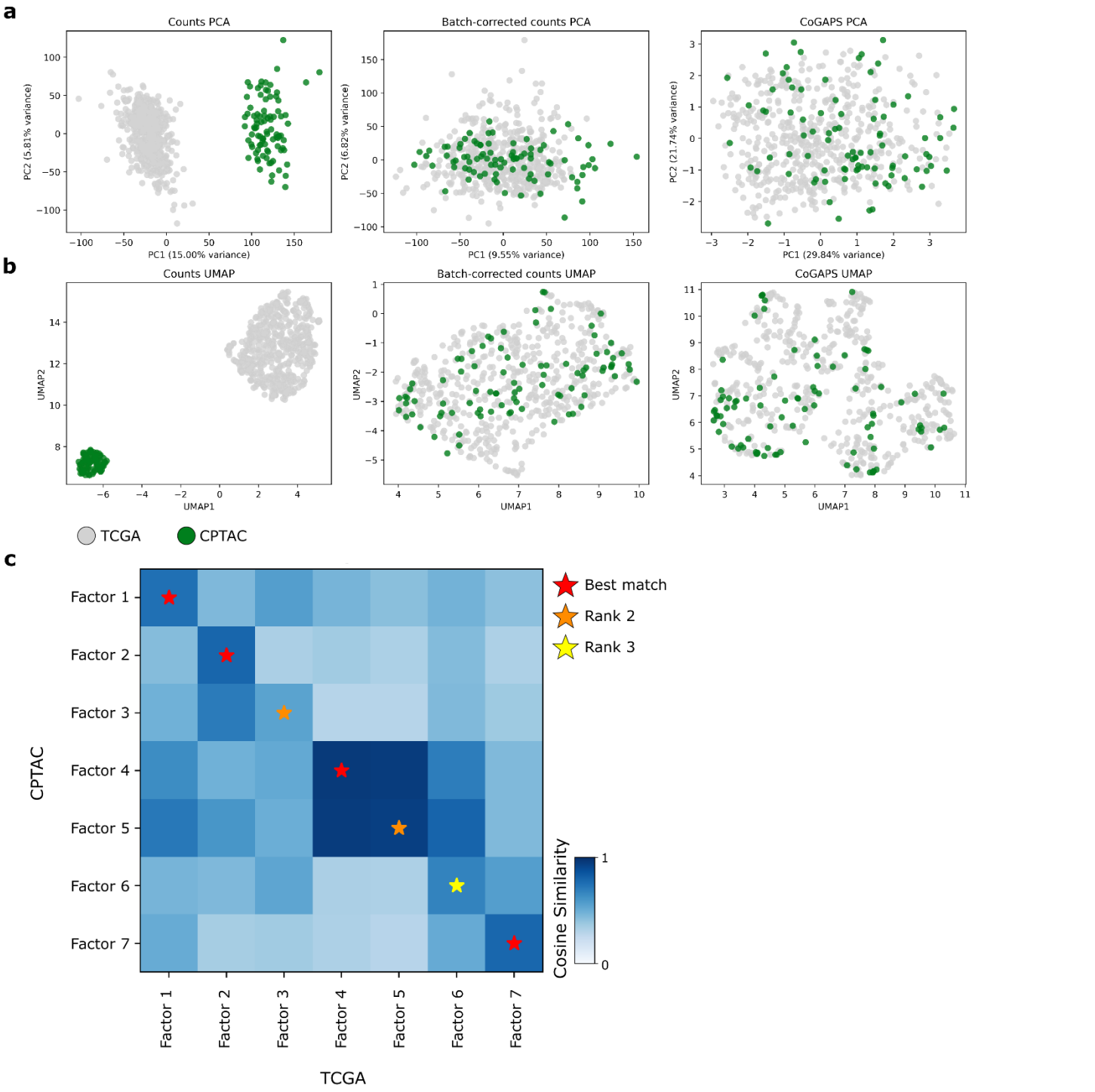

Figure S6: Independent TCGA and CPTAC analyses reveal similar patterns in LUSC data. a,b, PCA (**a**) and UMAP (**b**) embeddings computed from (left to right) normalized counts, batch-corrected counts, and sample weights for matched CoGAPS factors. For the matched factorizations, CPTAC factors were reordered to correspond to their best-matching TCGA counterparts based on cosine similarity, as shown in panel c. **c**, Cosine similarity heatmap between TCGA and CPTAC CoGAPS latent factors, which were computed independently on each dataset using 7 factors. The matching process first identified primary matches (red stars) by selecting TCGA factors with highest cosine similarity to each CPTAC factor. When a TCGA factor was initially matched to multiple CPTAC factors, the match with lower cosine similarity was discarded and the next-highest similarity match was chosen (orange stars). This process was repeated until achieving unique one-to-one factor pairs, with yellow stars indicating tertiary matches if needed.

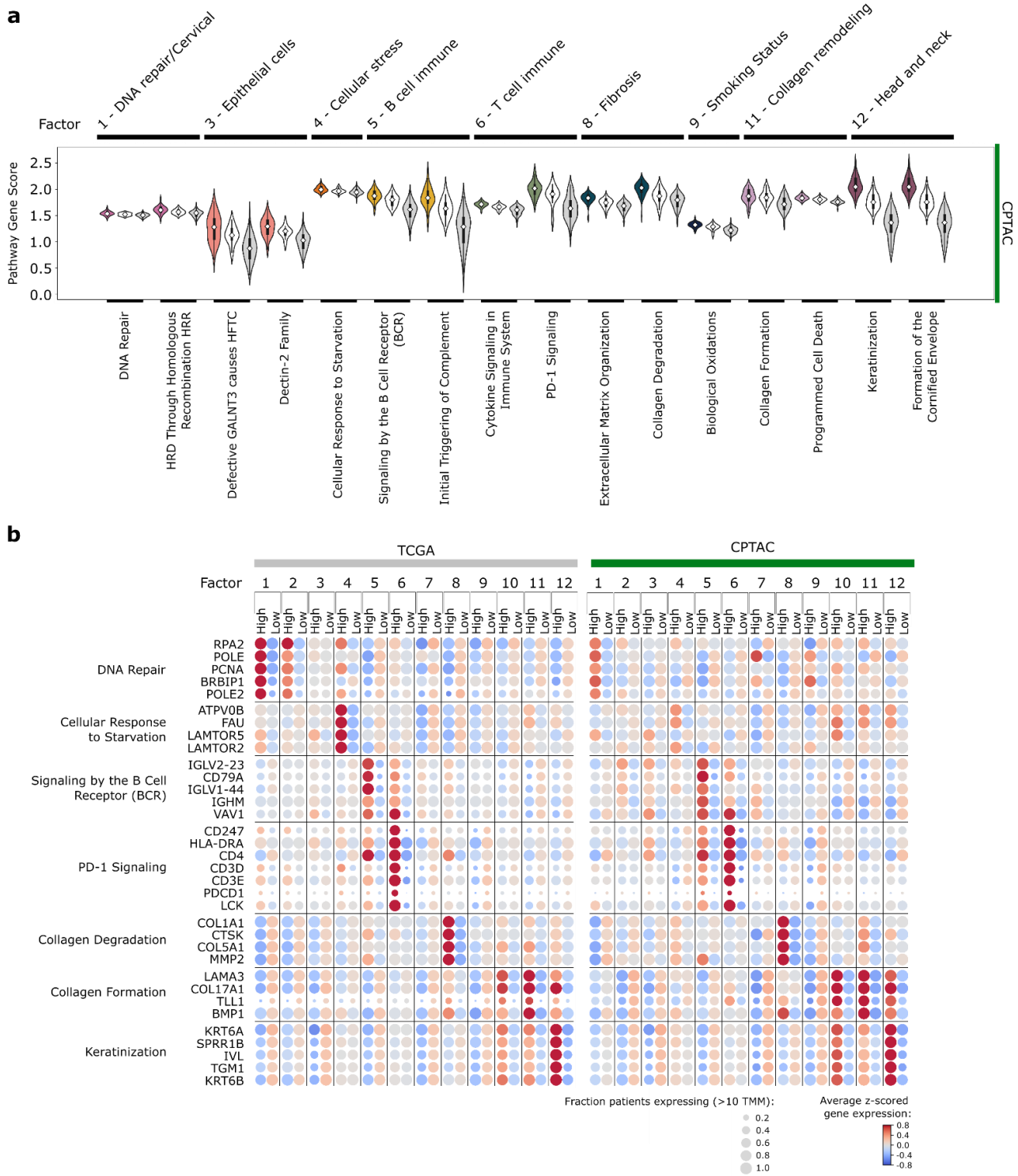

Figure S7: Latent factor genes are enriched in distinct, cancer-associated pathways. a, For each CLF (top annotation) and pathway (x axis, bottom annotation) (as in Fig. 2), which were manually selected from the pathways that were significantly enriched in genes upweighted in the indicated CLF, a trio of violin plots show the distributions of pathway gene scores (y axis), computed as the average log expression of all genes in a pathway, in TCGA (gray, top) and CPTAC (green, bottom) patient samples grouped by "high" (>75th percentile, color), "mid" (60th–75th percentile, white), and "low" (<60th percentile, gray) factor weights. Boxplots inside violins shows the median (white diamond), first and third quartiles (ends of thicker black line), and minimum and maximum (ends of thin black line). For each pathway shown, distributions for "high" vs. "low" groups were significantly different (*t*-test, Bonferroni-corrected *P* <0.0001). **b,** For individual genes (y axis) in significantly enriched pathways (annotations, left), and for patient samples grouped by "high" or "low" factor weights for each CLF (columns, annotations above) in each dataset (TCGA: left, gray bar; CPTAC: right, green bar), the dot plot shows the average z-scored gene expression (dot color) and fraction of samples with nonzero expression (dot size) in each group, where z-scoring is performed separately for each CLF and dataset, before averaging.

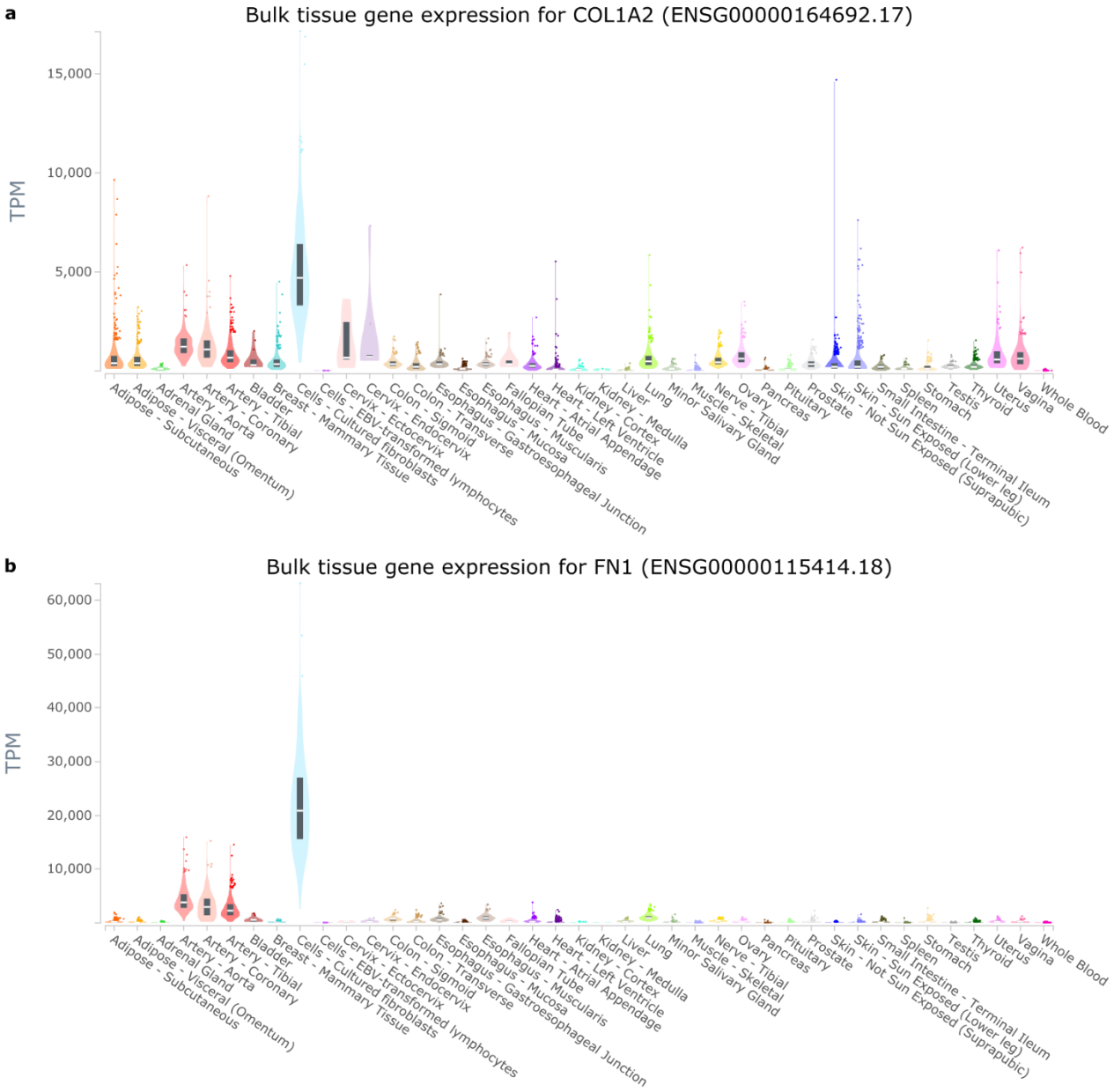

Figure S8: Genes upweighted in CLF 8 are highly expressed in fibroblasts. a,b, Violin plots show the distributions of expression values (TMM, y axis) in normal tissues (x axis) of the genes COL1A2 (**a**) and FN1 (**b**), both highly weighted in CLF 8, according to the genotype-tissue expression **(**GTEx) database. Boxplots in the violins show the median (white line), and first and third quartiles (ends of thick black line).

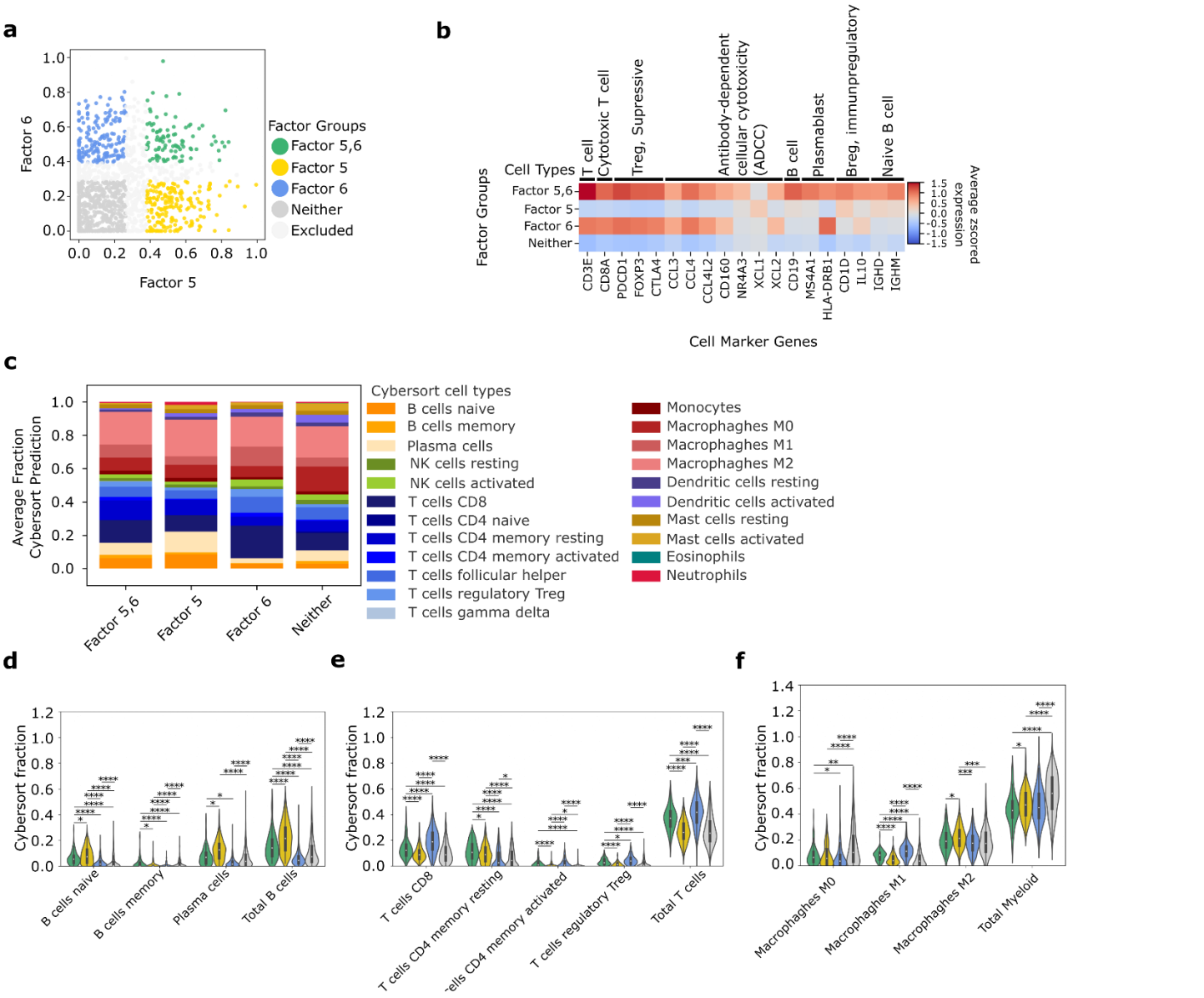

Figure S9: CLFs 5 and 6 identify both distinct and shared immune features of SCCs. a, Scatterplot shows patients (dots) by their weights for "Factor 5 – B cell immune" (x axis) and "Factor 6 – T cell immune" (y axis), colored by their status as "high" or "low" for one, both, or neither factor. **b,** Heatmap shows the average of z-scored gene expression (color) of individual immune genes (x axis, annotation above) in patient subsets (y axis, as in panel a). **c,** Bar charts show the average fraction (y axis), in patient subsets (x axis, as in panel a), of each sample that is inferred by CIBERSORT to belong to various immune-cell populations (color). d–e**,** Violin plots, colored by patient subset (as in panel a), show the distributions of CIBERSORT-inferred fractions of B-cell and plasma-cell populations (**d**), T-cell populations (**e**), and myeloid populations (**f**), with statistically significant differences indicated (Mann-Whitney U test, * P <0.05, ** P <0.001, *** P <0.0001, **** P <0.00001). Boxplots inside the violins show the median (white diamond), first and third quartiles (ends of thicker black line), and the minimum and maximum (ends of thin black line).

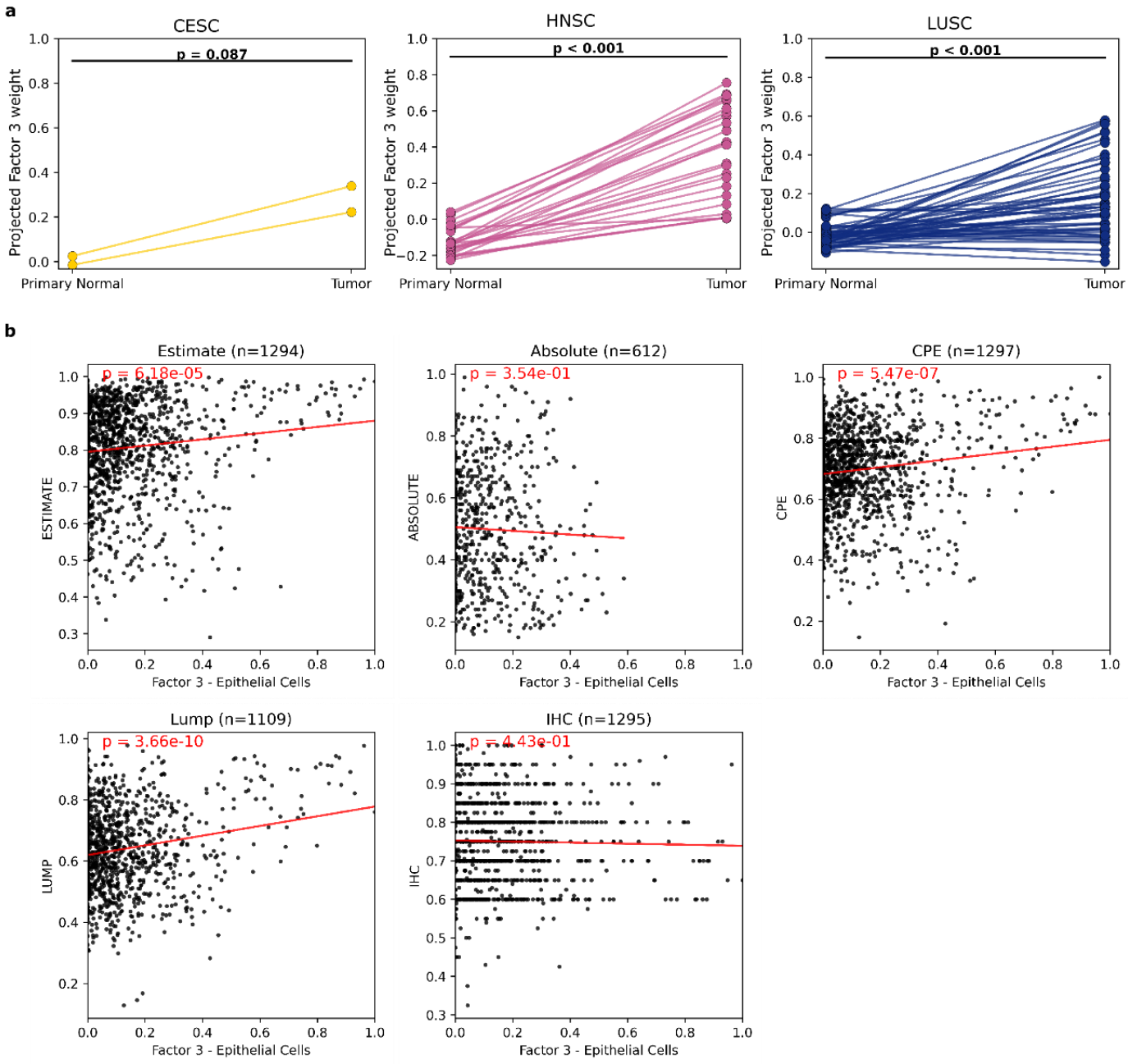

Figure S10: CLF3 correlates with tumor purity. a, Plots show the weight (y axis) of CLF 3, “Factor 3 - Epithelial cells”, in paired normal and tumor samples (x axis), computed by TMM-normalizing gene expression data from all paired samples and then projecting the results into the previously computed CoGAPS embedding space. For each primary cancer type (panel title), CLF 3 weight is significantly higher in tumor versus normal samples (paired *t*-test, p-value annotated in panel). **b**, Scatter plots depict the relationship between CLF 3 factor weights (x axis) in the original CoGAPS embedding and tumor purity estimates (y axis), computed according to several published methods (panels). The red line represents the linear regression fit (significance of correlation indicated by p-value annotation).

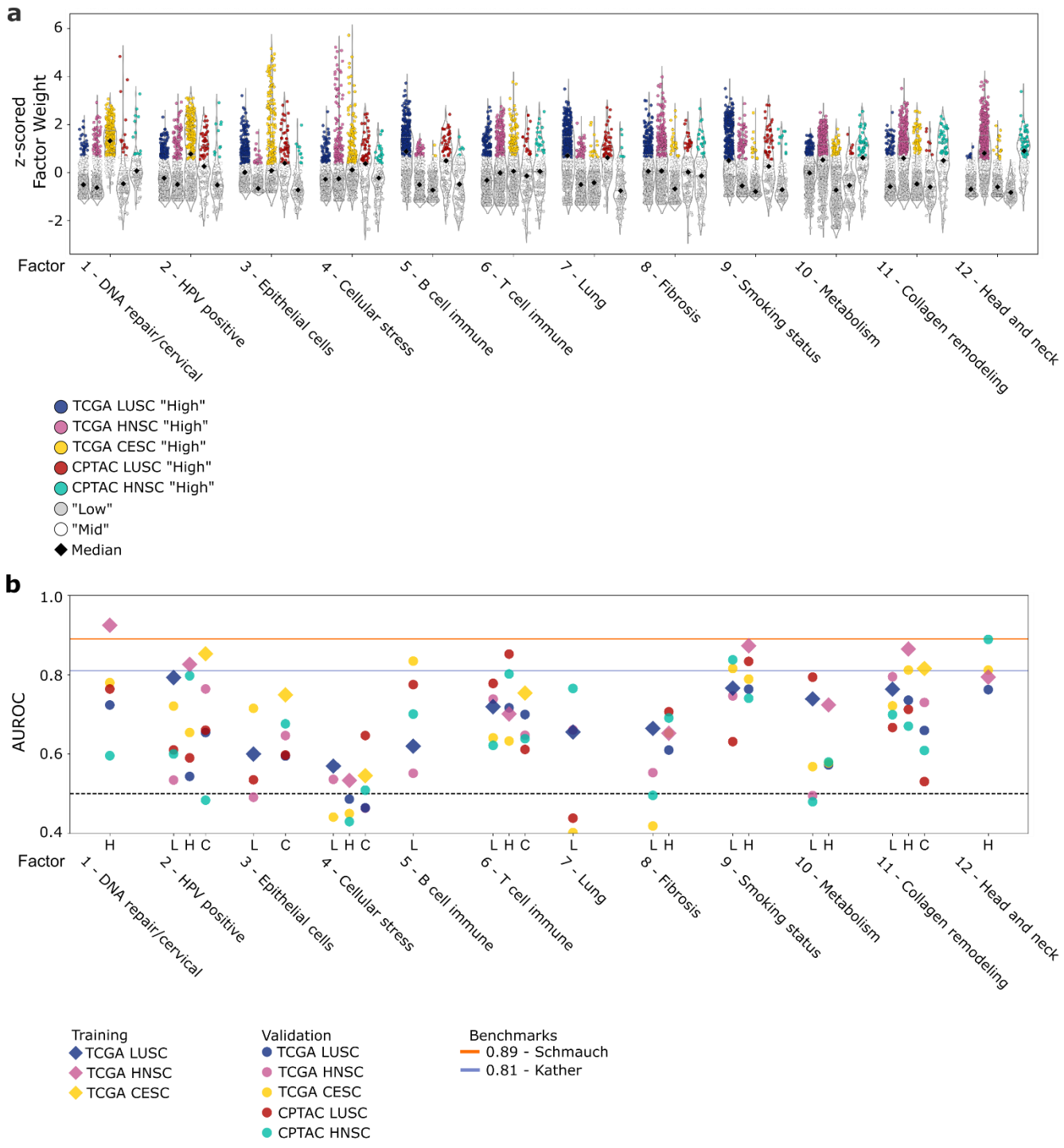

Figure 11: CLF prediction models were validated across primary cancer types and datasets, with variable performance. a, For each CLF (x axis labels), individual violin plots show the distributions within each primary cancer type and dataset of the z-scored CLF sample weights (y axis; z-scores computed across all patients, separately for each CLF and dataset), binned (color) into the "high" (> 75th percentile, colored by primary cancer type and dataset), "mid" (60–75th percentile, white) and "low" (<60th percentile, gray) status categories used for training ("mid" samples were censored for training). Median: black diamond, thick dashed line; first and third quartiles: thinner dashed lines. **b,** Results (AUROC, y axis) of latent factor status predictions from tumor histology. For each CLF (x axis groups, bottom annotation) and primary type with sufficient training data (x axis labels; L: lung; H: head and neck; C: cervical), a separate model was trained. Training results (diamonds, colored by primary type used in training) represent the average AUROC, where ROC curves were created by varying the threshold for tile-level calls, and results were averaged across 3 k-fold cross-validation rounds for a model with optimized hyperparameter values. Trained models were validated on the other primary types and datasets; validation results (circles, colored by validation primary type and dataset) represent the AUROC of the model

performance. Reference lines show best results for gene expression-based T and B cell detection in liver hepatocellular carcinoma from Schmauch et al. (14) (orange solid line), gene signature-based proliferation detection in breast cancer from Kather et al. (13) (blue solid line), and expected results for random guessing (black dashed line).

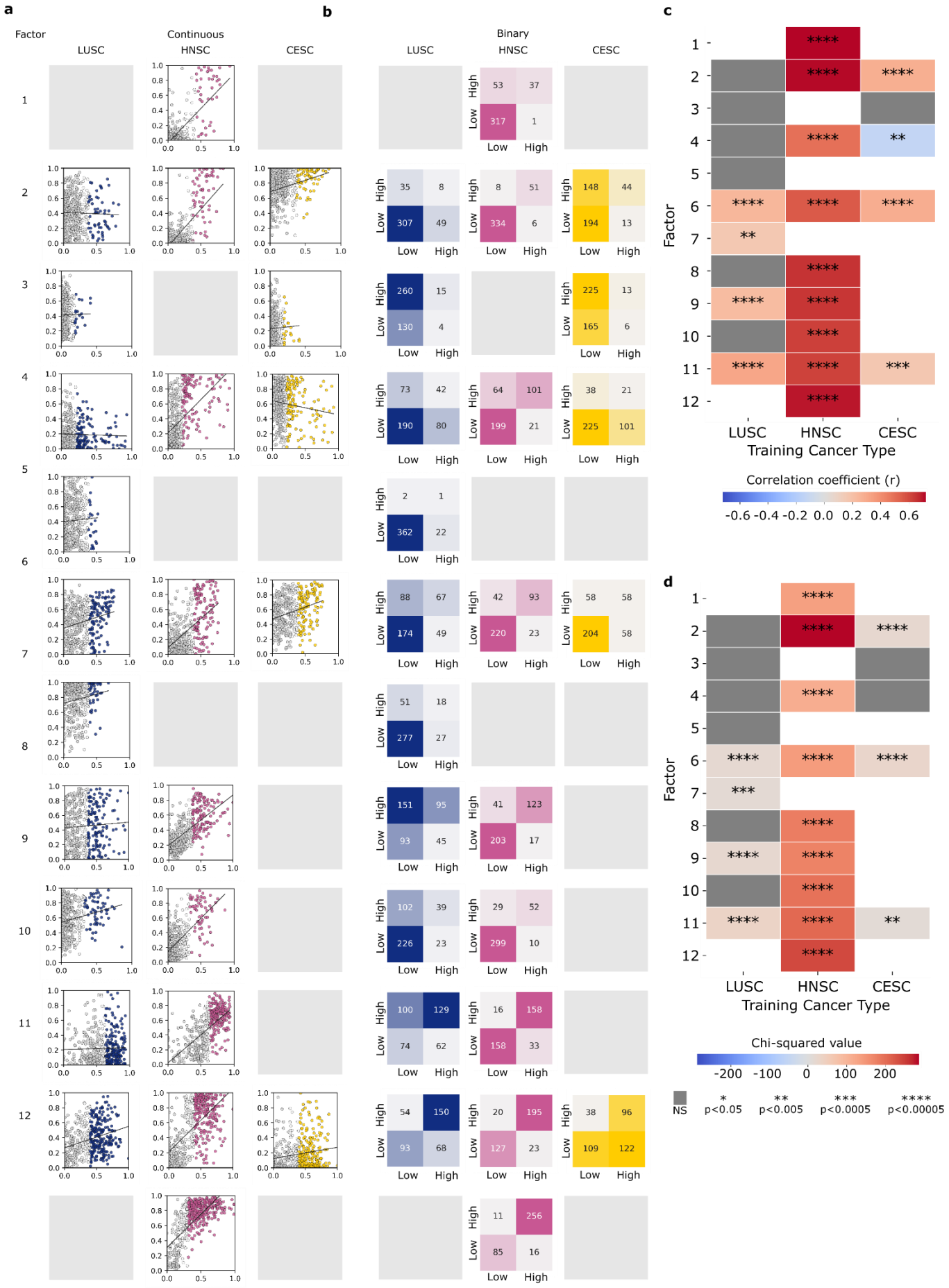
a
Factor
Continuous
LUSC
HNSC
CESC
b
Binary
LUSC
HNSC
CESC
High
Low
c
Factor
Training Cancer Type
LUSC
HNSC
CESC
Correlation coefficient (r)
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
d
Chi-squared value
-200 -100 0 100 200
NS
* p<0.05
** p<0.005
*** p<0.0005
**** p<0.00005

Figure S12: Neural networks trained to detect "high" versus "low" CLFs learn a continuum. a, Scatter plots show TCGA HNSC samples (dots) by CLF weight (x axis, "True Weight") versus the percent of image tiles per patient classified as positive (y axis, "Prediction") by neural network models trained on TCGA HNSC (left), LUSC (middle), and CESC (right). Point color indicates CLF weights: "high" (>75th percentile, colored), "mid" (60th–75th percentile, white) or "low" (<60th percentile, grey). Blank gray box indicates no model was trained for specific CLF and primary combination. **b**, Confusion matrices show results for predictions binarized using Youden's J statistic. c-d, Heatmaps show correlation (**c**) and chi-squared (**d**) statistics corresponding to continuous and binarized model results, respectively (* $p<0.05$, ** $p<0.005$, *** $p<0.0005$, **** $p<0.00005$; gray indicates non-significant results). Blank white indicates no model was trained for a specific CLF and primary combination.

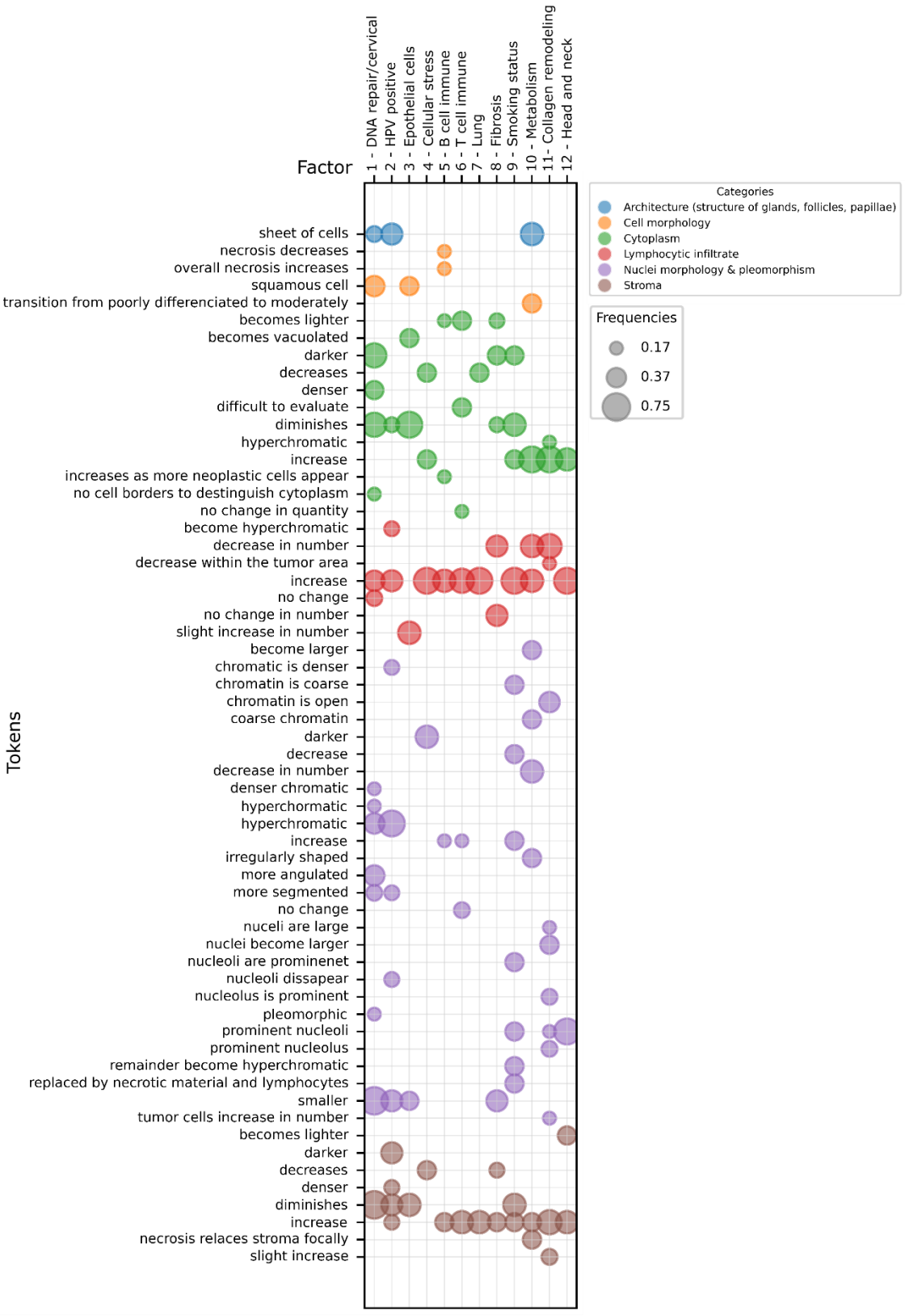

Figure S13: Pathologist annotations revealed histologic features that differentiate synthetic images corresponding to CLF status. A pathologist characterized the changes between “high” and “low” CLF synthetic images, noting differences in six categories: architecture (blue), stroma (brown), nuclei (purple), lymphocytic infiltrate (red), cytoplasm (green) and cell morphology (orange). The unique descriptive phrases (y-axis) used by the pathologist were tokenized and quantified for each CLF (x-axis). The size of each dot represents the fraction of images within a given category that received that specific annotation for each CLF. Only terms that appeared in multiple image pairs (“high” and “low”) for a given CLF were included; terms observed in only a single pair (seed) were excluded to minimize noise.

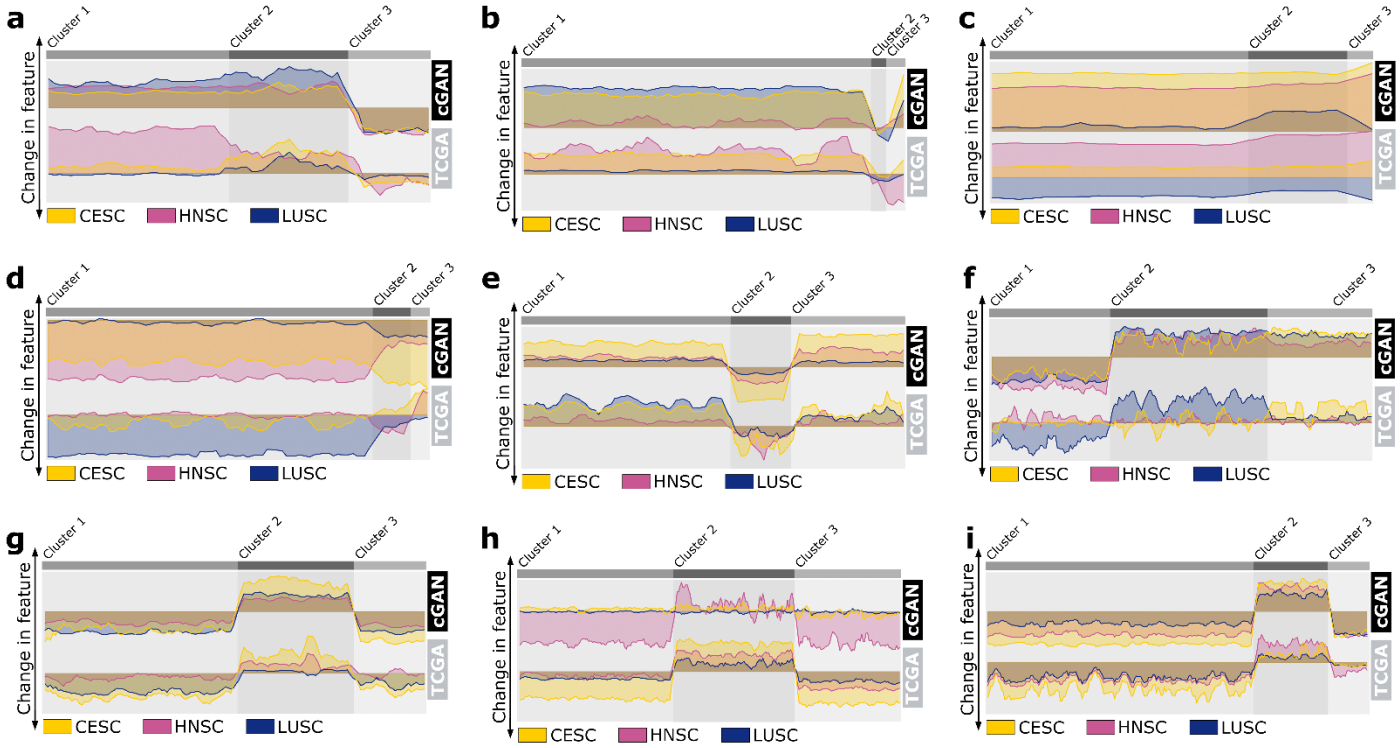

Figure S14: CellProfiler revealed standardized image features that differentiate synthetic and real images corresponding to CLF status. a, For all distinct CellProfiler-identified features (x axis) from cGAN and TCGA images that significantly differentiate CLF2 “high” versus “low” factor status, plots show the difference between factor status groups (“high” – “low”) of the z-scored, average feature value (y axis) in each cancer primary type (color), with features ordered by clusters (top annotation, gray shading) of similar patterns. **b–i,** Analogous plots for CLFs 3–8, 10, and 12, respectively.