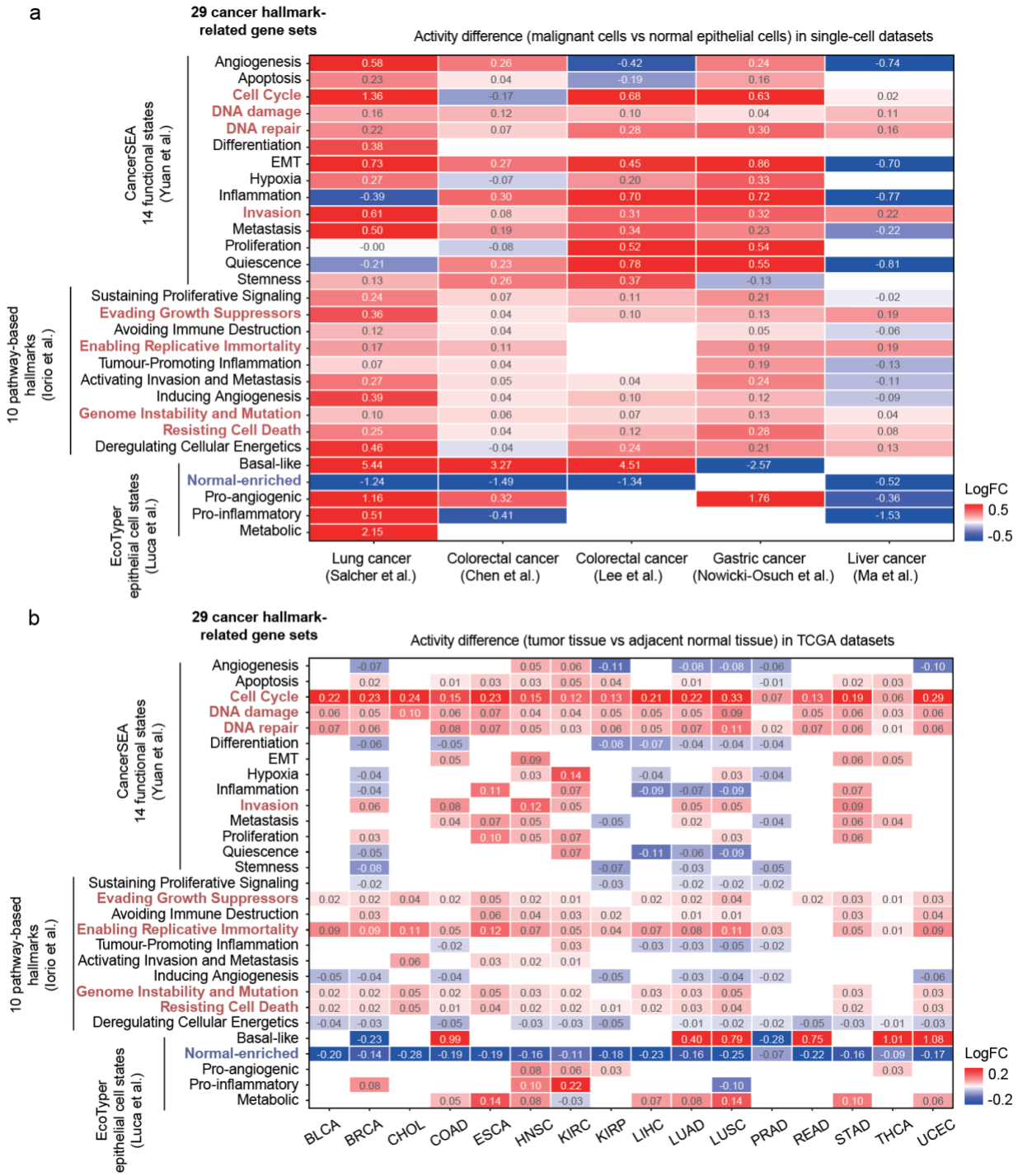
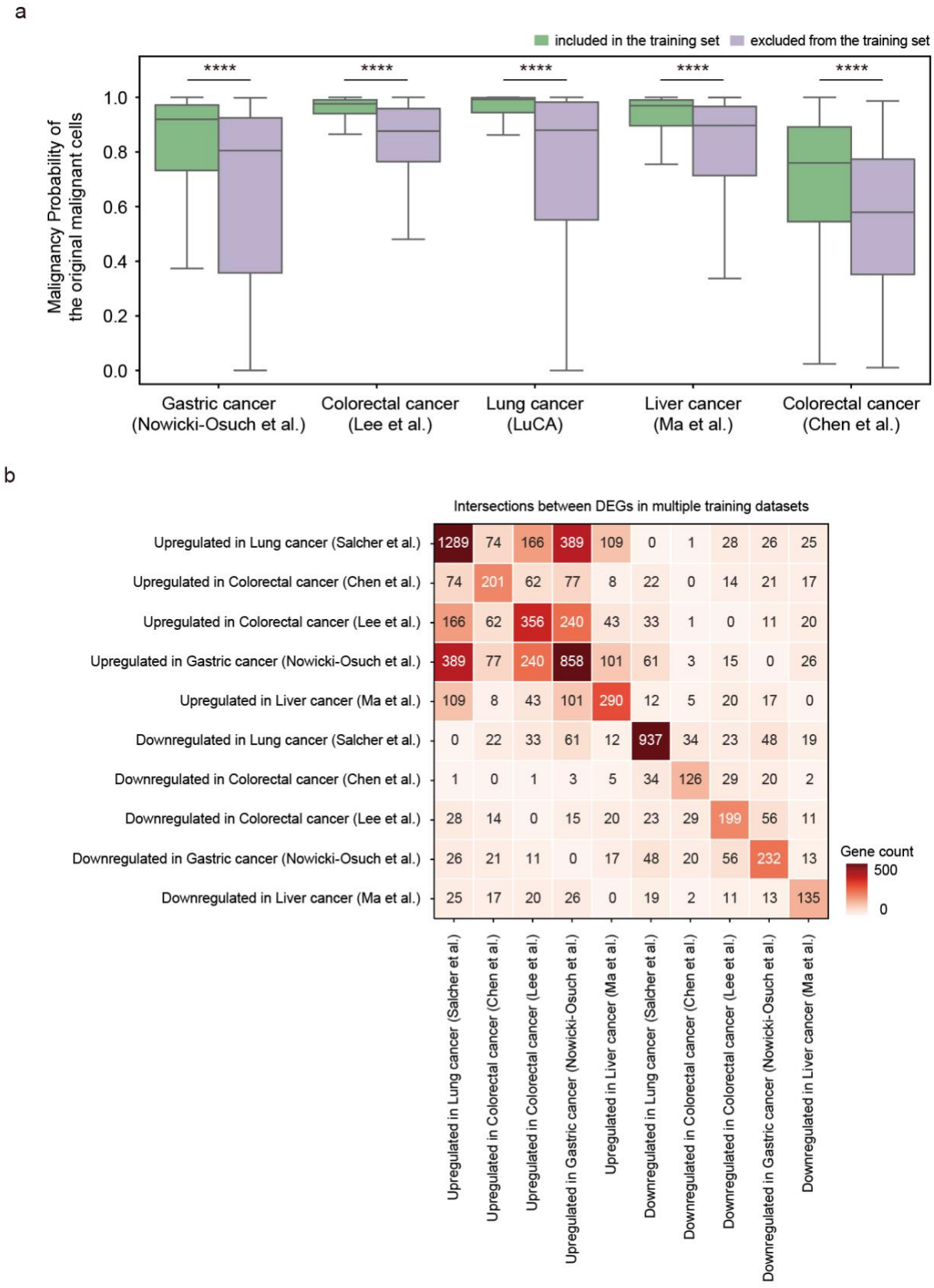
Supplementary Figure legends



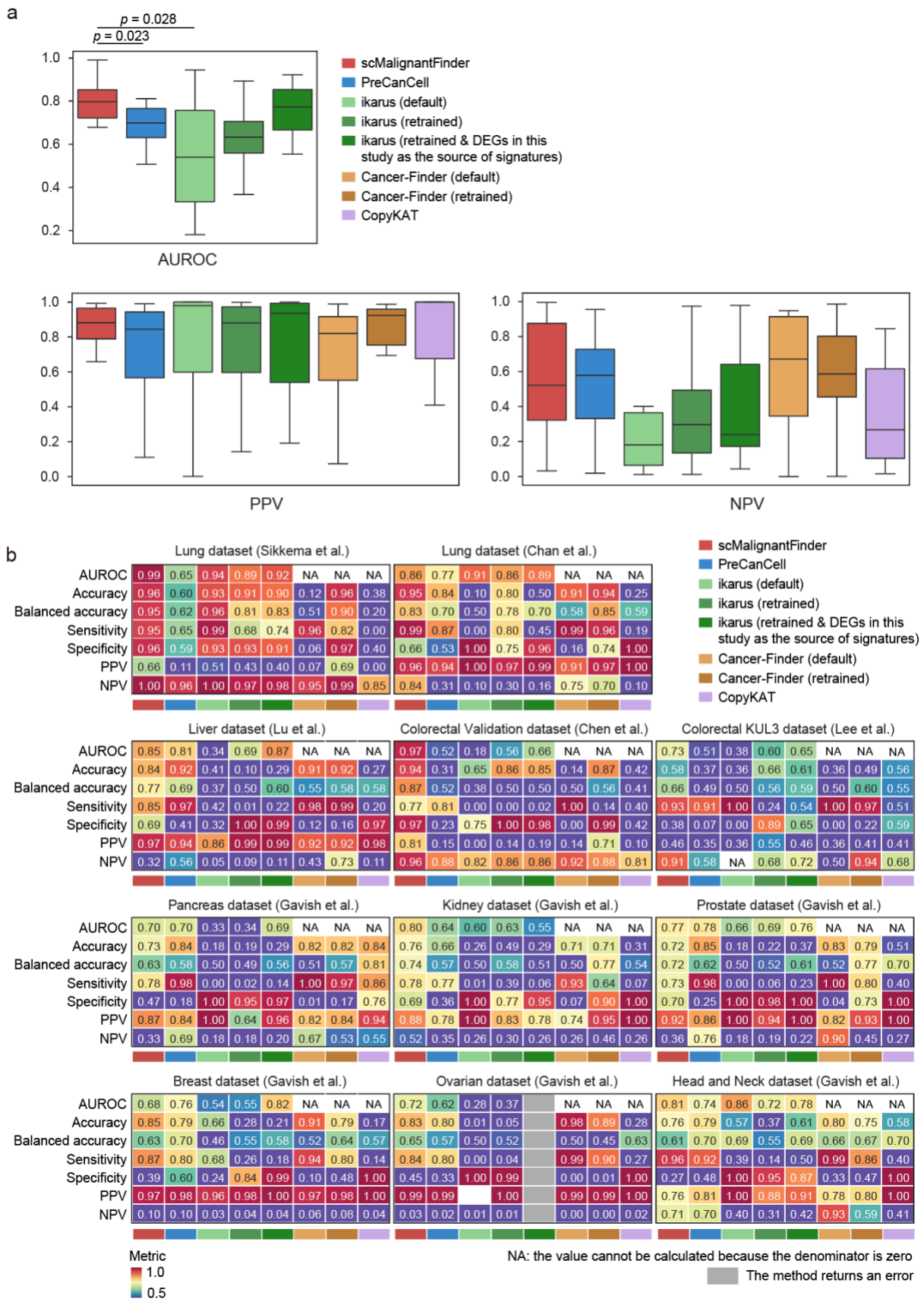
Supplementary Fig. 1 Activity of 29 cancer gene signatures in the collected single-cell datasets and the TCGA datasets. a-b, Log2 fold change of cancer gene signature activity between

malignant and normal epithelial cells in five scRNA-seq datasets (**a**), and between tumor and adjacent normal tissues in TCGA bulk RNA-seq data from 16 cancer types (**b**). Color-labeled gene signatures show consistent activity changes at both single-cell and bulk levels, with eight upregulated in malignant cells and tumors (red) and one downregulated (blue).



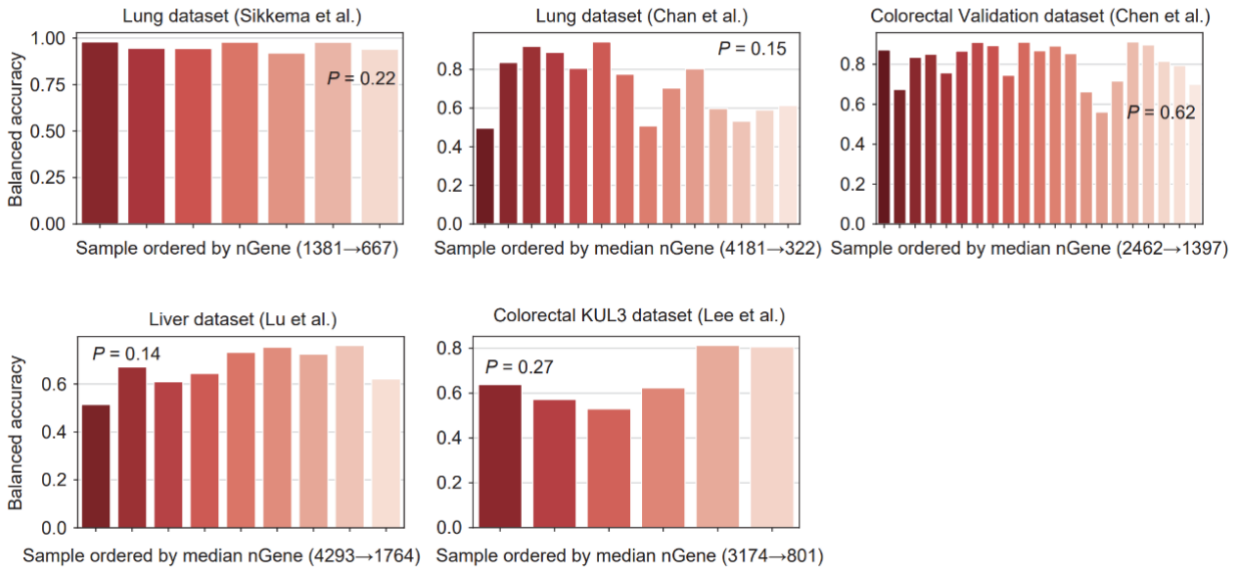
Supplementary Fig. 2 Training set calibration and DEG comparison. a, Boxplot showing

malignancy probability for malignant cells included or excluded in the training set. Statistical significance was determined by a two-sided Wilcoxon rank-sum test (****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b**, The intersection of identified DEGs among the five datasets in the training set. The values in the heatmap represent the number of overlapping DEGs between each pair of datasets.

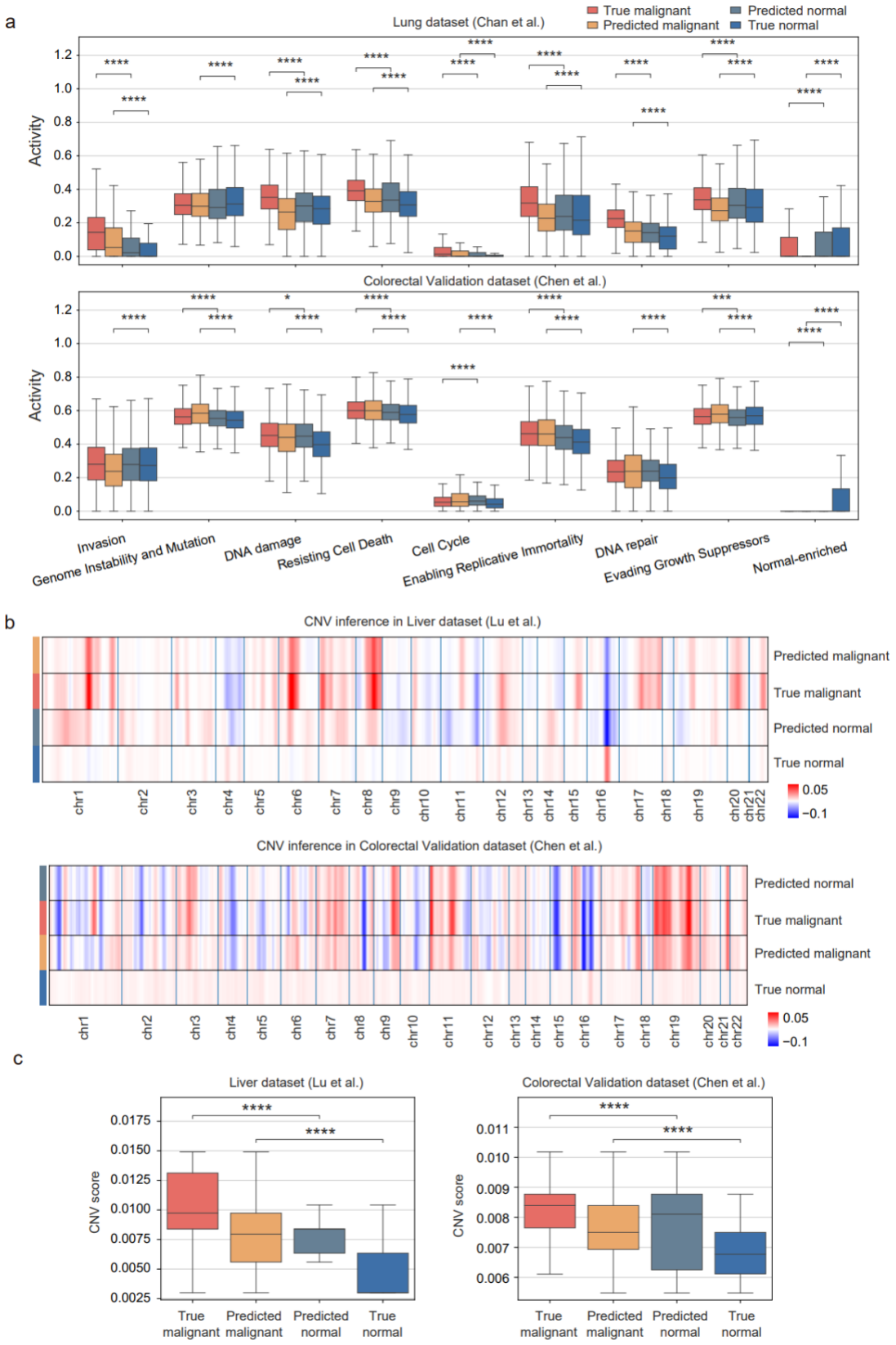


Supplementary Fig. 3 Performance comparison of various methods on multiple independent

test sets. a, Boxplot showing the three metrics of each method on 13 scRNA-seq datasets (Supplementary Data 1). Statistical significance was determined by a two-sided Wilcoxon rank-sum test. n = 13 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b**, Heatmap showing seven metrics from 11 test datasets derived from cancer patients. The NA meants the value cannot be calculated because the denominator is zero. Gray means the method returned an error. Source data are provided as a Source Data file (Supplementary Data 9).

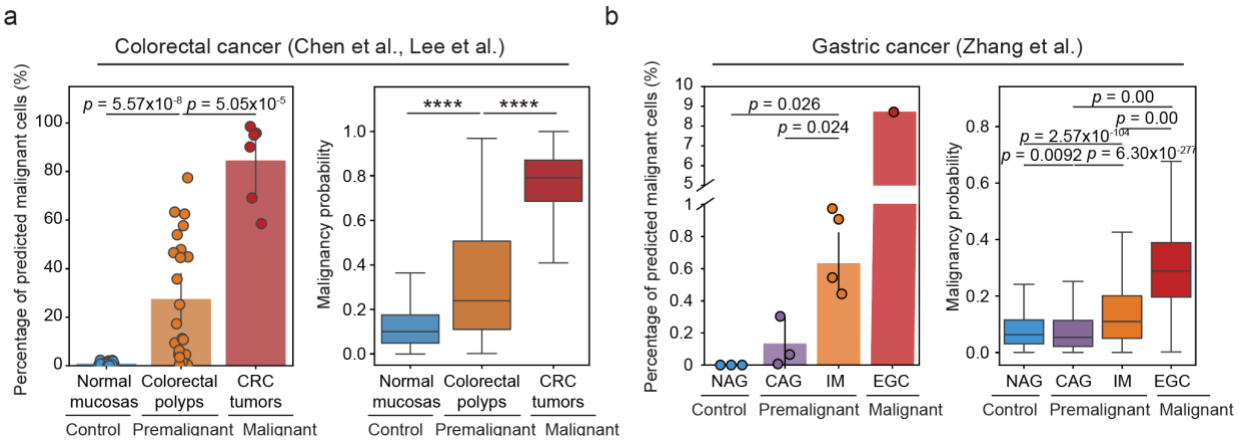


Supplementary Fig. 4 Balanced accuracy of scMalignantFinder when applied to samples with varying sequencing depths in the test set. For each dataset, the Spearman correlation coefficient between the median number of genes and the balanced accuracy at the sample level was annotated on the figure.

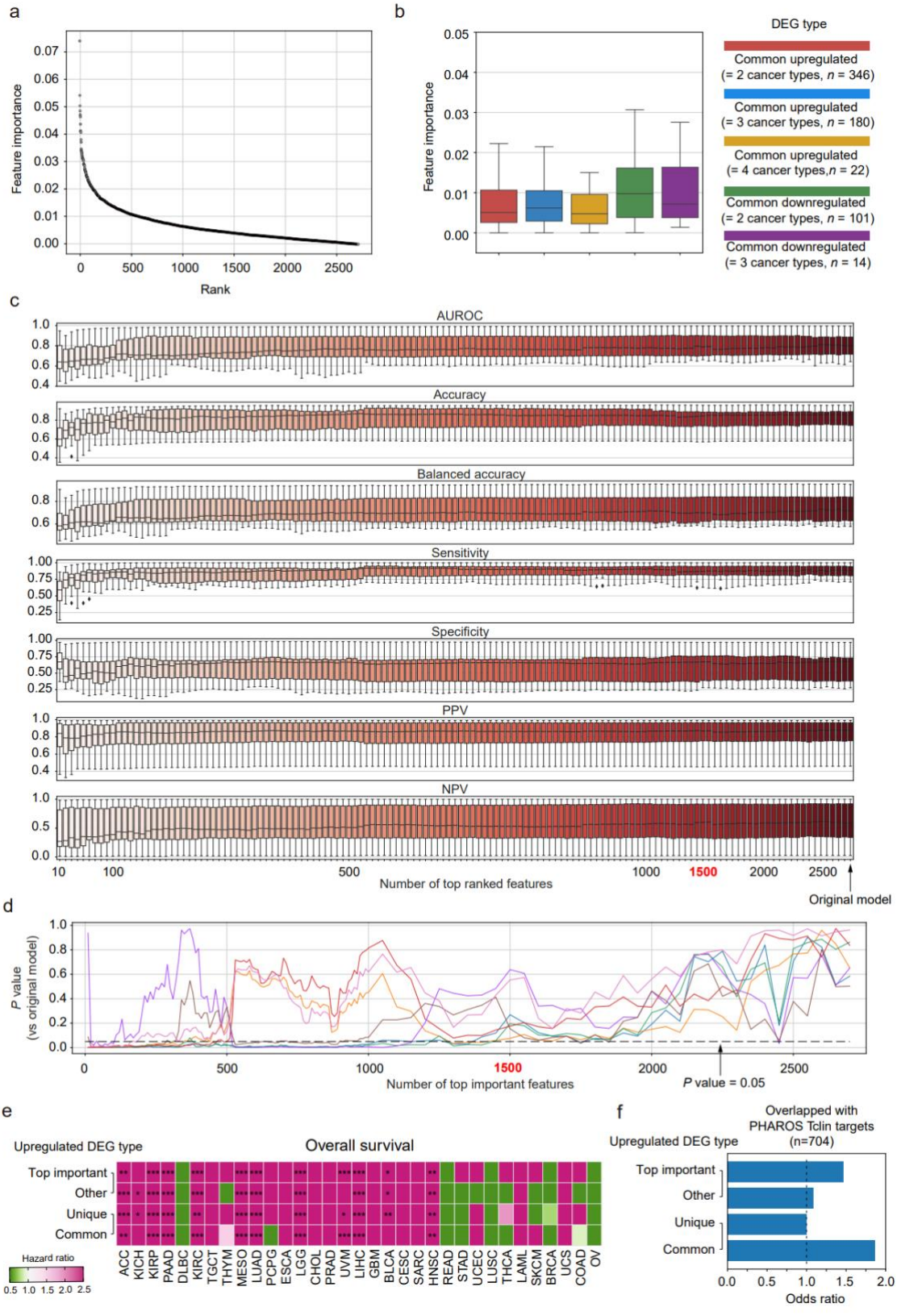


Supplementary Fig. 5 Transcriptional and CNV characteristics in misclassified cells. a,

Boxplot showing activity of curated cancer gene signatures in four groups of cells (true malignant, predicted malignant, predicted normal, true normal) across two test sets. Statistical significance was determined by a two-sided Student's t test (*: $P < 0.05$, ***: $P < 0.001$, ****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b**, Heatmap showing CNV inference in four groups across two test sets. **c**, Boxplot showing CNV scores in four groups. Statistical significance was determined by a two-sided Student's t test (****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 10).

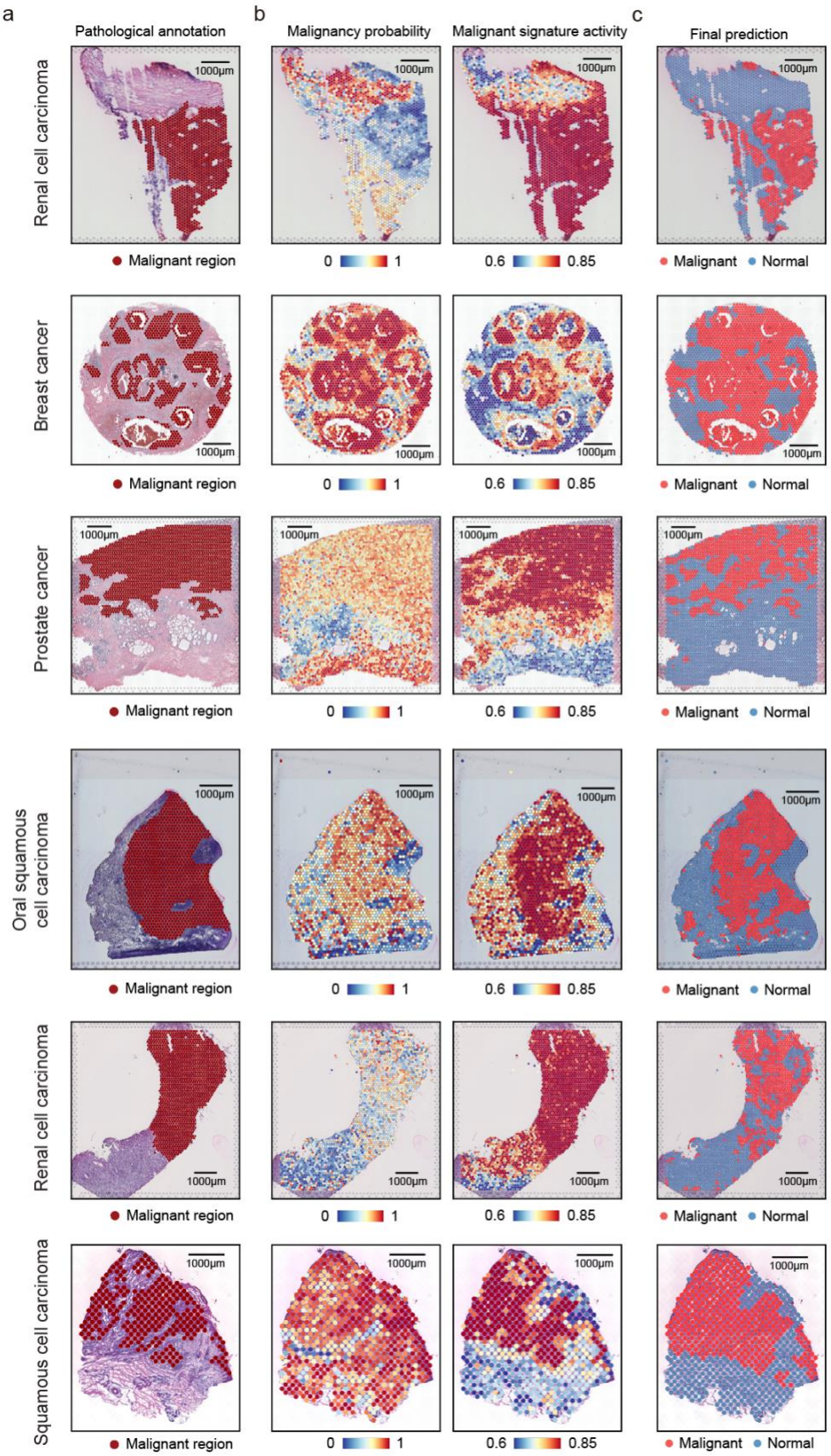


Supplementary Fig. 6 Prediction of malignant cells during carcinogenesis in colorectal cancer (a) and gastric cancer (b) using scMalignantFinder. For both cancers, the percentage of predicted malignant cells (left panels) and malignancy probability (right panels) were calculated for samples representing control, premalignant, and malignant stages. Statistical significance for the percentage of predicted malignant cells was assessed using the Mann-Whitney U test, and Student's t-test was applied for malignancy probability. Left: bar height represents the mean value; whiskers extend to the mean ± 95% confidence intervals; right: center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 11).



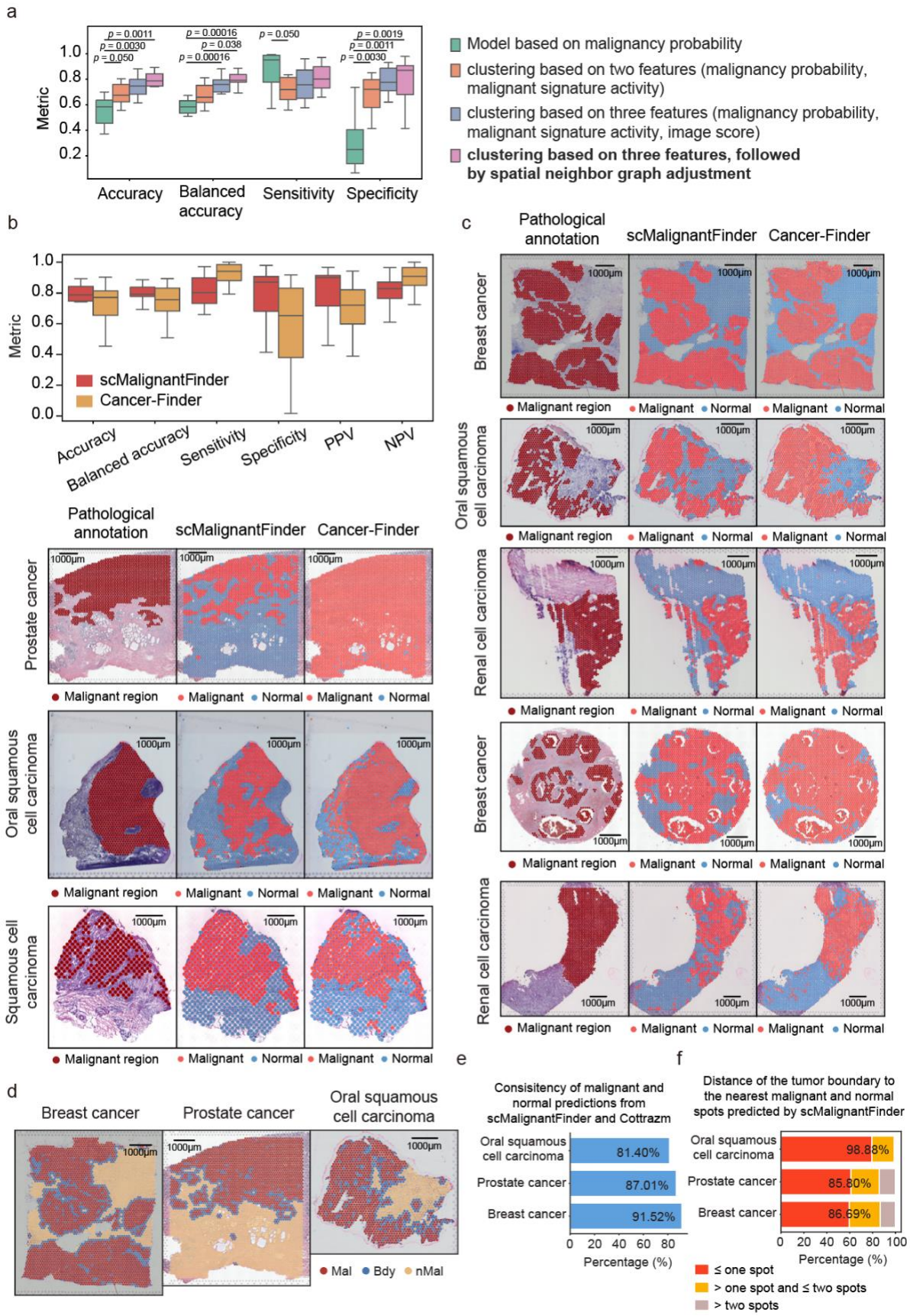
Supplementary Fig. 7 Feature importance analysis of DEGs. a, Ranking of DEGs based on

their feature importance. **b**, Boxplot showing the feature importance of common and unique DEGs. Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **c**, Boxplot illustrating the performance of models using top important DEGs. n = 13 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **d**, Statistical significance analysis of performance metrics between the original model and models built with top-ranked DEGs. Dashed lines represent a p-value of 0.05. The red mark highlights 1,500 DEGs (defined as top important DEGs), where the performance of the model becomes comparable to the original model. Statistical significance was determined by a two-sided Student's t test. **e**, Heatmap showing the hazard ratios of upregulated DEG set activity relation to overall survival across multiple cancer types. Statistical significance was determined by cox regression and likelihood ratio test (*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$). **f**, Odds ratio for the association between targets of approved drugs with known mechanism in the PHAROS database (n=704) and specific DEG sets. The dotted line indicates an odds ratio value of 1. Source data are provided as a Source Data file (Supplementary Data 12).



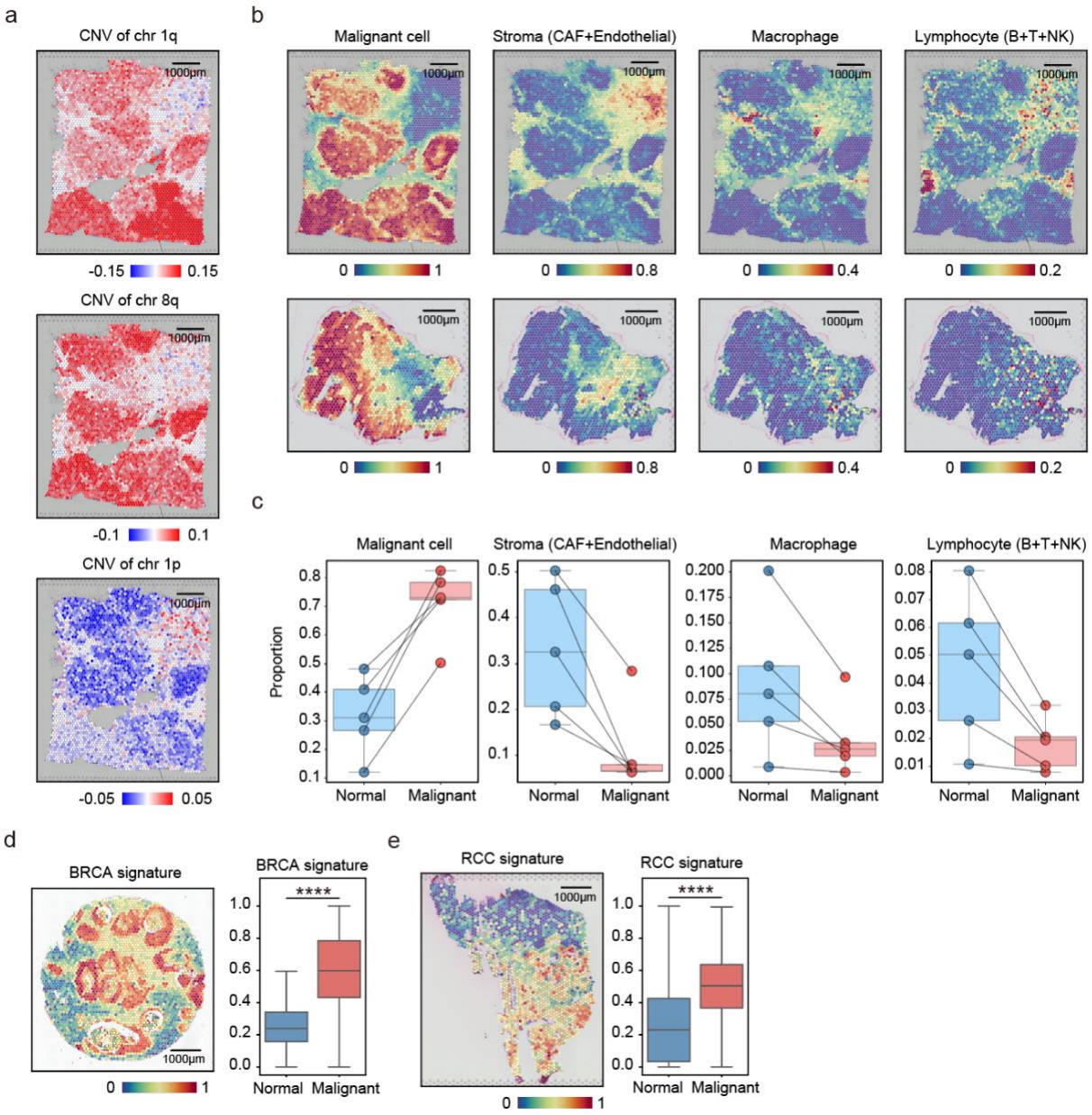
Supplementary Fig. 8 Expansion of scMalignantFinder to spatial transcriptomic (ST)

annotation. a-c, Annotations of ST spots by different methods: annotation by pathologists (**a**), determination of malignancy probability and malignant signature activity (**b**), and final classification (**c**). Source data are provided as a Source Data file (Supplementary Data 13).



Supplementary Fig. 9 Comparison of scMalignantFinder with other ST analysis tools. a, Box

plots show the performance metrics for four models in malignant region identification. The four models were (1) the default single-cell model relying solely on malignancy probability; (2) hierarchical clustering based on two features (malignancy probability and malignant signature activity); (3) hierarchical clustering incorporating three features (malignancy probability, malignant signature activity, and image feature); and (4) hierarchical clustering with three features followed by spatial neighbor graph adjustment. Statistical significance was determined by a two-sided Wilcoxon rank-sum test. n = 8 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **b**, Box plots comparing the performance metrics of scMalignantFinder and Cancer-Finder. **c**, Spatial visualization of malignant region identification between pathological annotations, scMalignantFinder, and Cancer-Finder across eight ST datasets. n = 8 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **d**, ST slices from breast cancer, prostate cancer, and oral squamous cell carcinoma showing regions classified as malignant (Mal), boundary (Bdy), and normal (nMal) by Cottrazm. **e**, Consistency percentages between scMalignantFinder and Cottrazm for malignant and normal spots across the three datasets. **f**, Classification of tumor boundary spots identified by Cottrazm based on the distance to the nearest malignant and normal spots predicted by scMalignantFinder, with distances categorized as ≤ one spot, > one spot and ≤ two spots, or > two spots. Source data are provided as a Source Data file (Supplementary Data 14).



Supplementary Fig. 10 Molecular characterization of predicted malignant and normal regions. a, Spatial distribution of copy number gains on chromosomes 1q and 8 and copy number losses on chromosome 1p in an ST slice of breast cancer. **b**, Spatial distribution of malignant cells, stromal cells, macrophages, and lymphocytes in ST slices of breast and oral squamous cell carcinoma. **c**, Boxplots comparing the proportions of malignant cells, stromal cells, macrophages, and lymphocytes between predicted malignant and normal regions. n = 5 datasets; center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. **d-e**, Tumor signature activity (left) and boxplot comparison (right)

between predicted malignant and normal regions (right) in ST slices of breast cancer (**d**) and renal cell carcinoma (**e**). Statistical significance was determined by a two-sided Student's t test (****: $P < 0.0001$). Center line represents the median value; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range. Source data are provided as a Source Data file (Supplementary Data 15).