

Supplementary Figures:

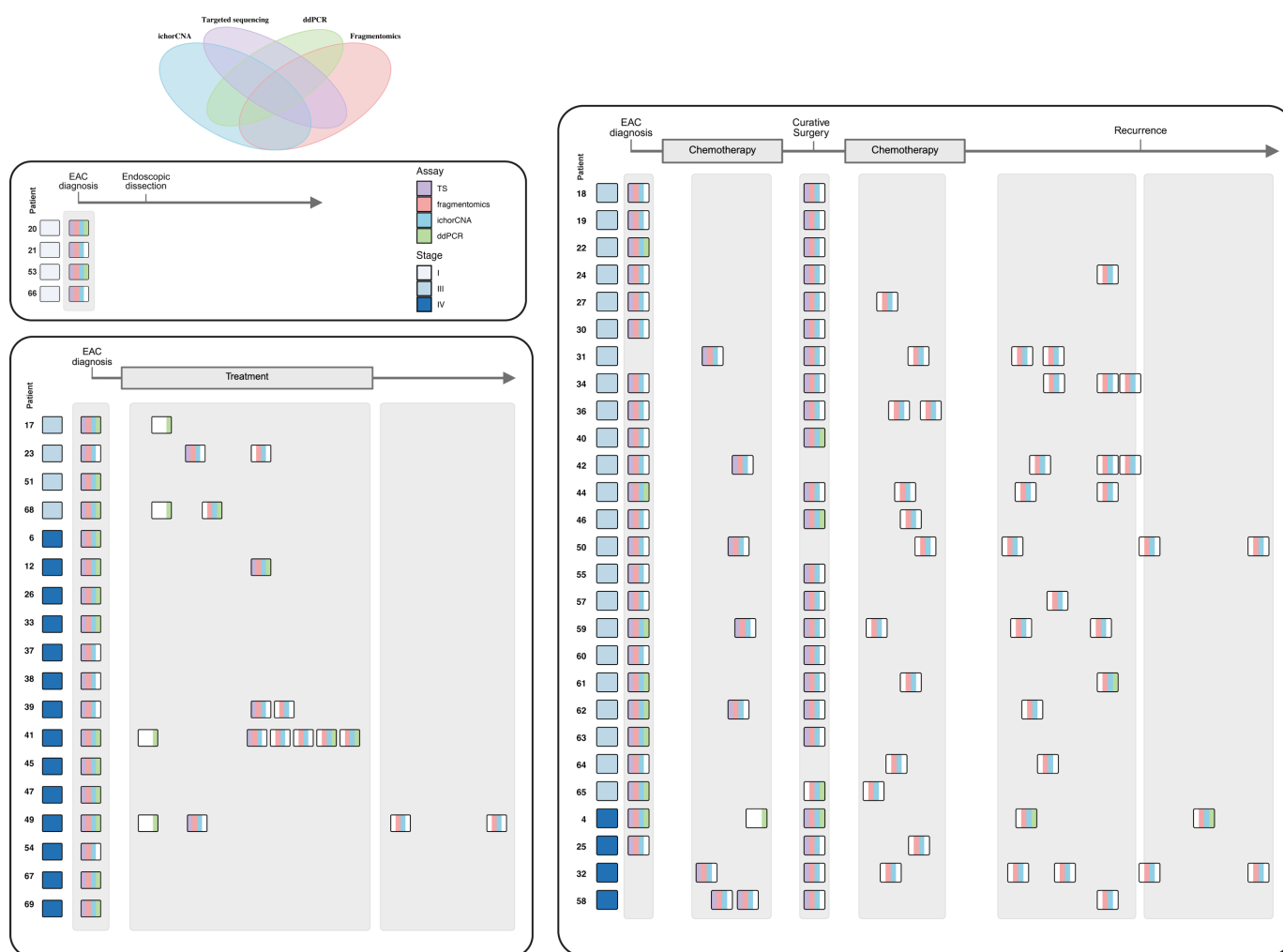


Figure S1. Longitudinal blood sampling and analyses performed. Overview of cfDNA sample availability across clinical timepoints, stratified by clinical group. Each row represents a patient, with samples plotted along the x-axis and colored by assay type. Panels show Stage I patients treated with endoscopic dissection (top left), Stage III/IV non-surgical patients (bottom left), and Stage III/IV surgical candidates (right).

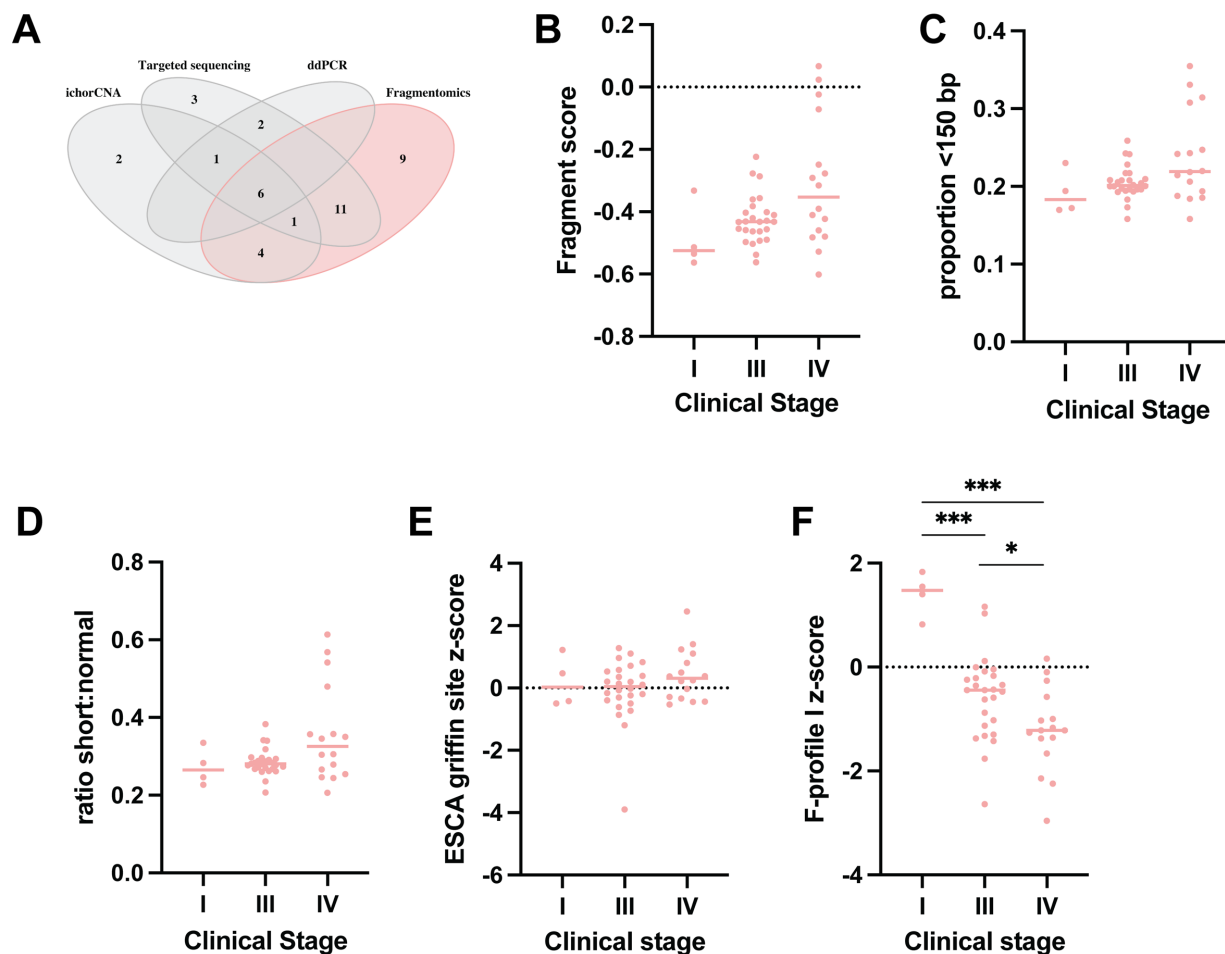


Figure S2. Baseline fragmentomic metrics correlate to clinical staging at diagnosis. (A)

Venn diagram showing the overlap of positivity in different multimodal analyses at baseline sampling in the cohort. **(B–F)** Baseline fragmentomic metrics stratified by clinical stage, including **(B)** fragment score, **(C)** proportion of cfDNA fragments <150 bp, **(D)** short-to-normal fragment ratio, **(E)** ESCA-specific Griffin chromatin accessibility Z-scores, and **(F)** Z-scores of F-profile I weights. Overall, advanced-stage patients (stage III–IV) exhibit increased fragment scores, higher proportions of short fragments, and elevated short-to-normal fragment ratios compared to stage I patients. In contrast, F-profile I motif frequencies are decreased in advanced disease, with significant differences between stages (Mann-Whitney U test: I vs III, $p = 0.0003$; I vs IV, $p = 0.0004$; III vs IV, $p = 0.0256$).

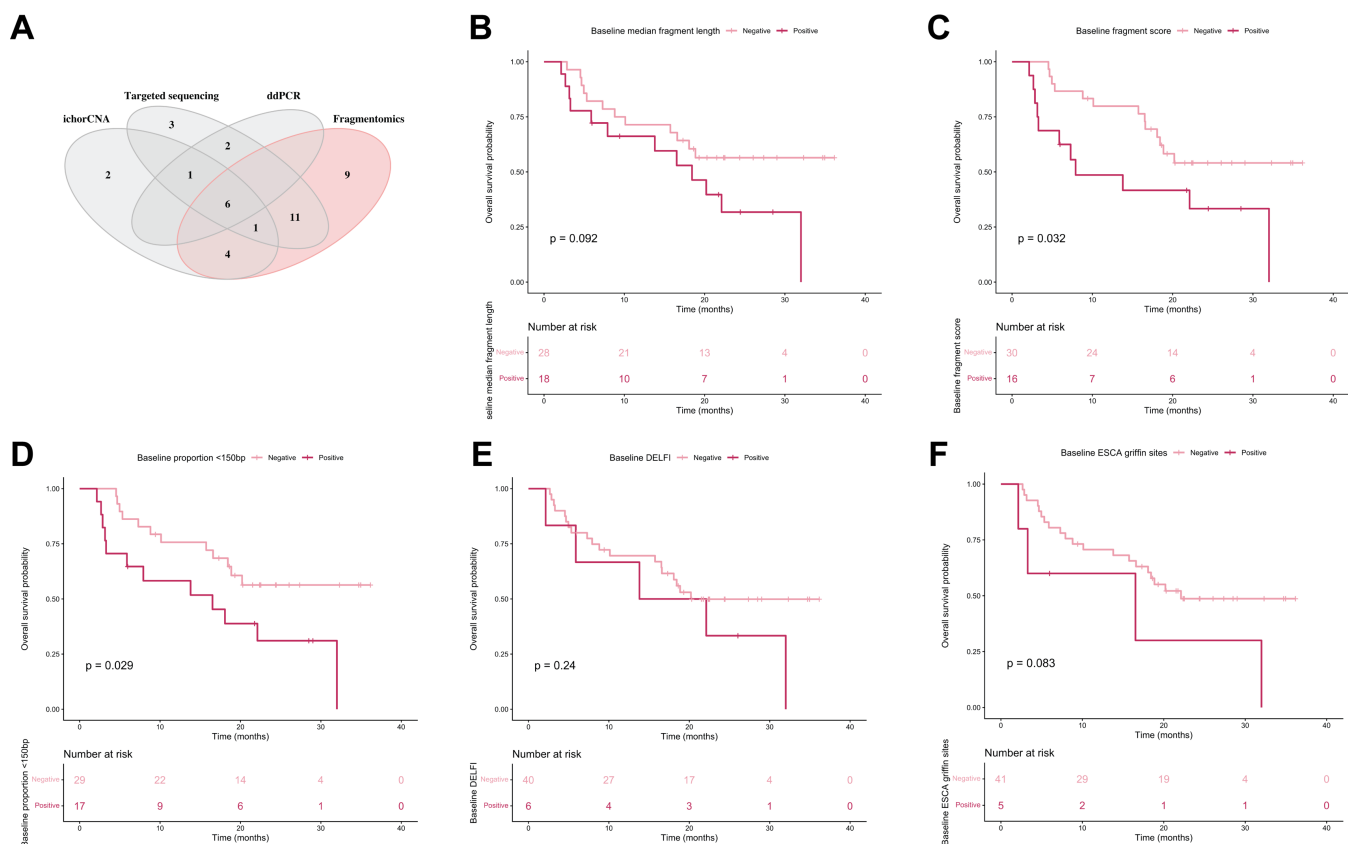


Figure S3. Baseline fragmentomics features are associated with overall survival (OS). (A) Venn diagram showing the overlap of positivity in different multimodal analyses at baseline sampling in the cohort. Kaplan-Meier curves of OS stratified by (B) baseline median fragment length (log-rank $p = 0.092$), (C) baseline fragment score (log-rank $p = 0.032$), (D) baseline proportion <150 bp (log-rank $p = 0.029$), (E) baseline DELFI (log-rank $p = 0.24$), and (F) baseline ESCA griffin score (log-rank $p = 0.083$).

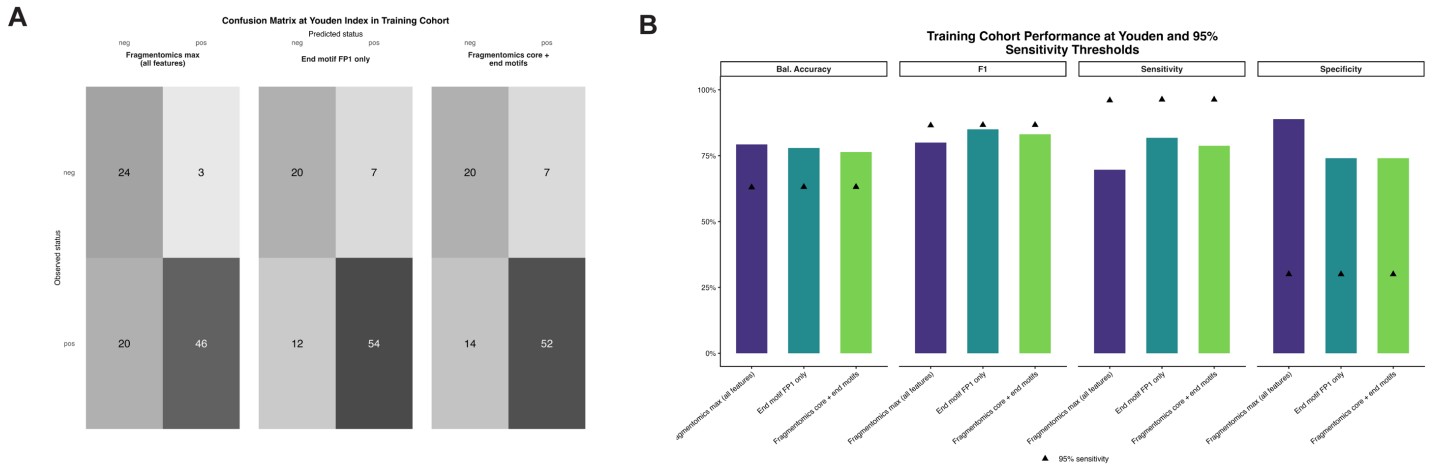


Figure S4. Training cohort performance of clinical disease status classifier. (A) Confusion matrices at the optimal Youden index threshold for different genome-wide feature combination in the training cohort. **(B)** Model performance across genome-wide feature combinations in the training cohort, including balanced accuracy, F1 score, sensitivity, and specificity at the Youden index threshold. Black triangles indicate model performance at a fixed 95% sensitivity operating point.

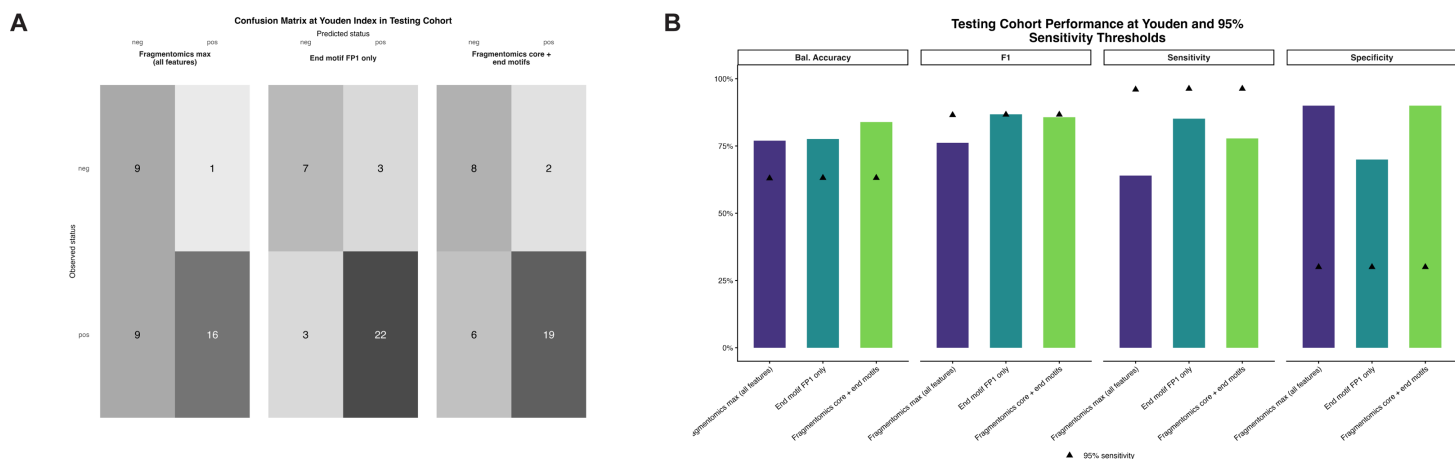


Figure S5. Testing cohort performance of clinical disease status classifier. (A) Confusion matrices at the optimal Youden index threshold for different genome-wide feature combination in the testing cohort. **(B)** Model performance across genome-wide feature combinations in the testing cohort, including balanced accuracy, F1 score, sensitivity, and specificity at the Youden index threshold. Black triangles indicate model performance at a fixed 95% sensitivity operating point.

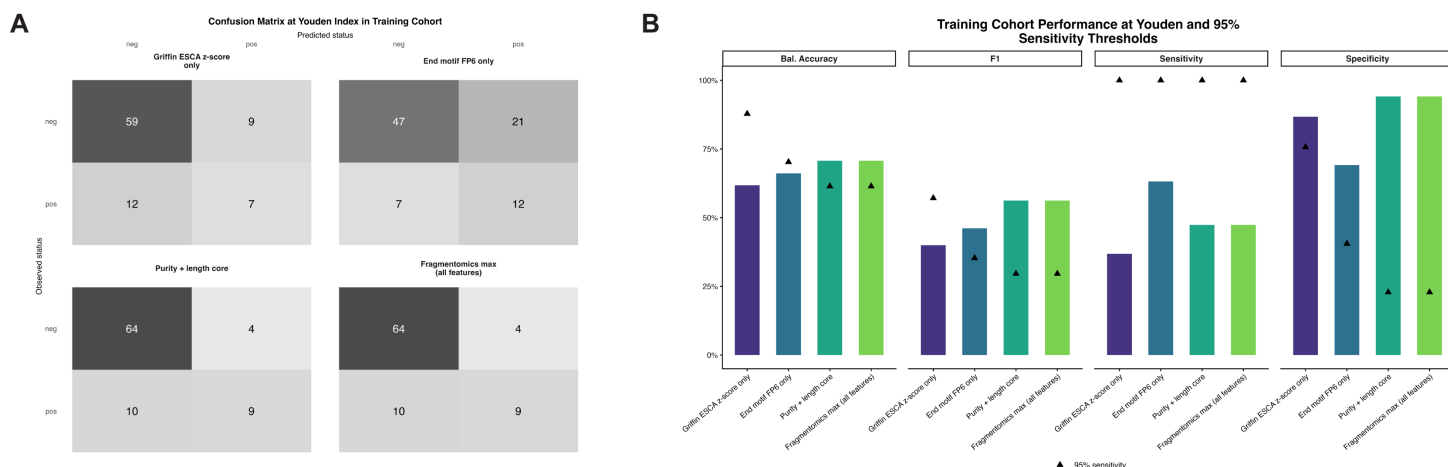


Figure S6. Training cohort performance of short-term mortality classifier. (A) Confusion matrices at the optimal Youden index threshold for different genome-wide feature combination in the training cohort. **(B)** Model performance across genome-wide feature combinations in the training cohort, including balanced accuracy, F1 score, sensitivity, and specificity at the Youden index threshold. Black triangles indicate model performance at a fixed 95% sensitivity operating point.

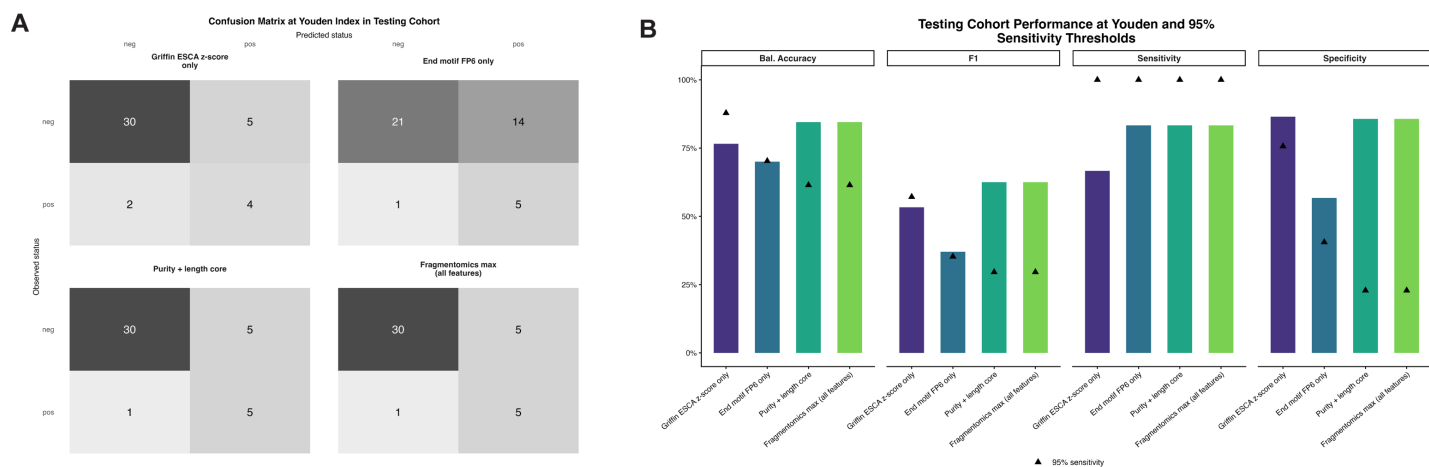


Figure S7. Testing cohort performance of short-term mortality classifier. (A) Confusion matrices at the optimal Youden index threshold for different genome-wide feature combination in the testing cohort. **(B)** Model performance across genome-wide feature combinations in the testing cohort, including balanced accuracy, F1 score, sensitivity, and specificity at the Youden index threshold. Black triangles indicate model performance at a fixed 95% sensitivity operating point.