

Supplementary Figures

Quantitative gene-level inference of tumor transcription from plasma epigenomes

Gunsagar S. Gulati^{1-4,15,*}, Damien Vasseur^{1,2,4,5,15}, Rashad Nawfal^{1,2,4,15}, Shahabeddin Sotudian^{1-4,15}, Karl Semaan¹⁻⁴, Marc Eid^{1,2,4}, Ji-Heui Seo^{1,2,4}, Shin Saito^{1,4,13}, Noa Phillips^{1,2}, John J. Canniff^{1,2}, Hunter Savignano^{1,2}, Surya B. Chhetri¹⁻⁴, Zhenjie Jin¹⁻³, Yukun Ou¹⁻³, Mohamed-Amine Bani^{5,6}, Renchin Wu⁷, Maria-Anna Chrysovergi⁷, Gwo-Shu Mary Lee^{1,8}, Rachel Trowbridge^{1,8}, Ilana B. Epstein^{1,8}, Gabriella Rickards^{1,8}, Paulo Roberto Da Silva Cordeiro¹⁻⁴, Ze Zhang¹⁻⁴, Razane El Hajj Chehade¹⁻⁴, Brady James^{1,2,8}, Christophe Massard^{9,10}, Antoine Italiano^{9,10}, Antoine Hollebecque^{9,10}, Jean-Charles Soria^{10,11}, Fabrice André^{9,11}, Cécile Badoual^{5,6}, Joaquim Bellmunt^{1,4,12}, Harshabad Singh^{1,4,13}, Andrew J. Aguirre^{1,4,13}, Brian M. Wolpin^{1,4,13}, Jacob Sands^{1,4,14}, Matthew G. Oser^{1,4,14}, Toni K. Choueiri^{1,4,12}, Sylvan C. Baca^{1-4,12,16,*}, Matthew L. Freedman^{1,2,4,12,16,17,*}

¹Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

²Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA

³Division of Population Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

⁴Harvard Medical School, Boston, MA, USA

⁵Department of Medical Biology and Pathology, Gustave Roussy, Villejuif, France

⁶Paris Cité University, Paris, France

⁷Department of Pathology, Dana-Farber Cancer Institute, Boston, MA, USA

⁸Arthur and Linda Gelb Center for Translational Research, Dana-Farber Cancer Institute, Boston, MA, USA

⁹Department of Medical Oncology, Gustave Roussy, Villejuif, France

¹⁰Department of Therapeutic Innovation and Early Phase Trials, Gustave Roussy, Villejuif, France

¹¹Paris-Saclay University, Department of Cancer Medicine, Gustave Roussy, Villejuif, France

¹²Lank Center for Genitourinary Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

¹³Gastrointestinal Cancer Center, Dana-Farber Cancer Institute, Boston, MA, USA

¹⁴Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

¹⁵These authors contributed equally.

¹⁶Senior authors.

¹⁷Lead contact.

*Correspondence: gunsagar_gulati@dfci.harvard.edu (G.S.G.), sylvan_baca@dfci.harvard.edu (S.C.B.), matthew_freedman@dfci.harvard.edu (M.L.F.)

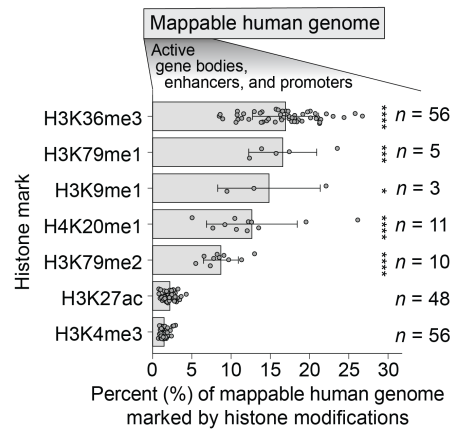
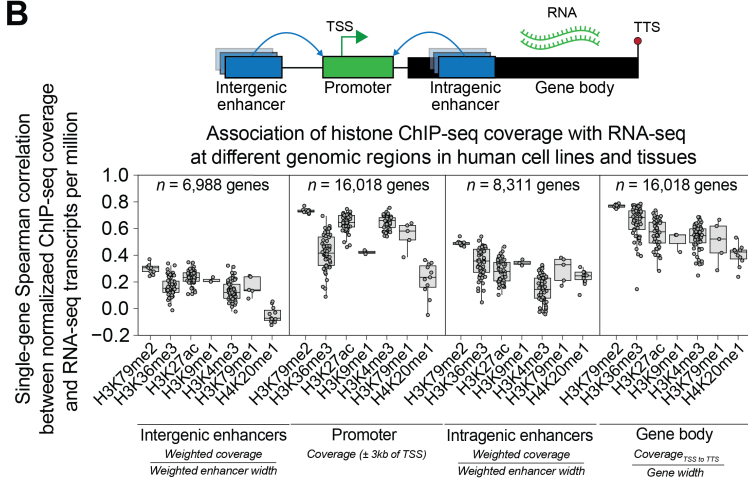
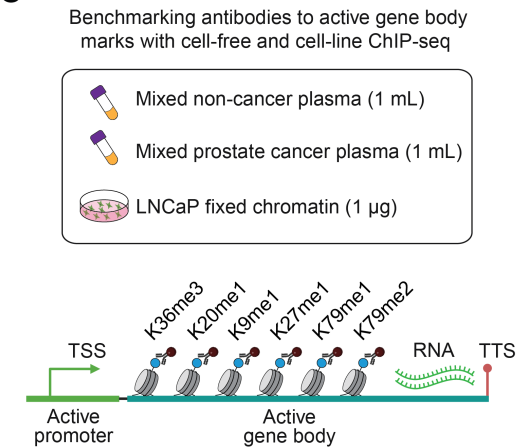
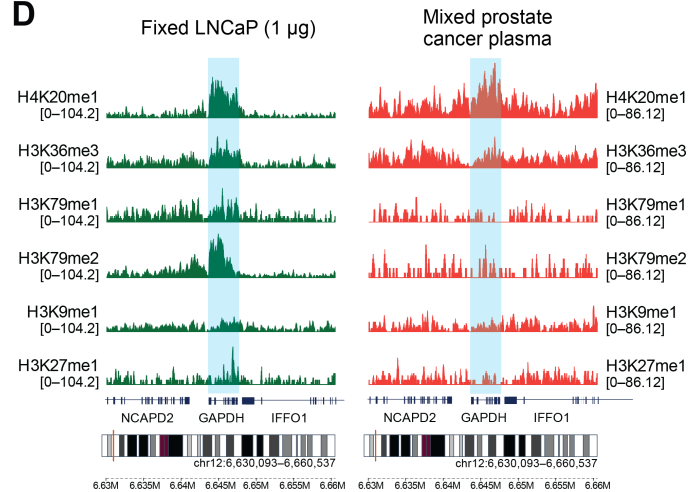
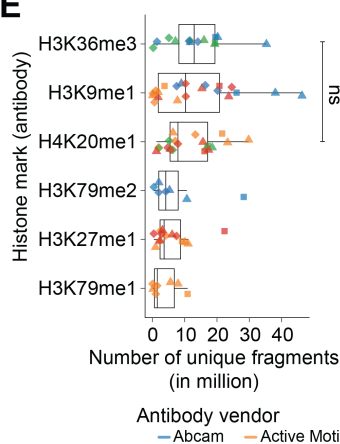
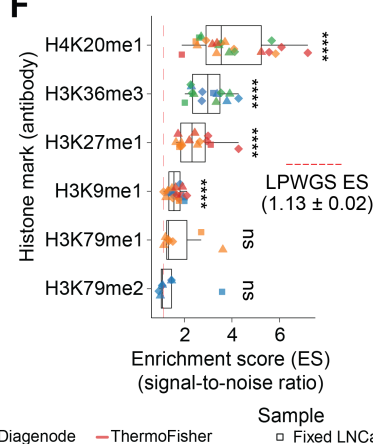
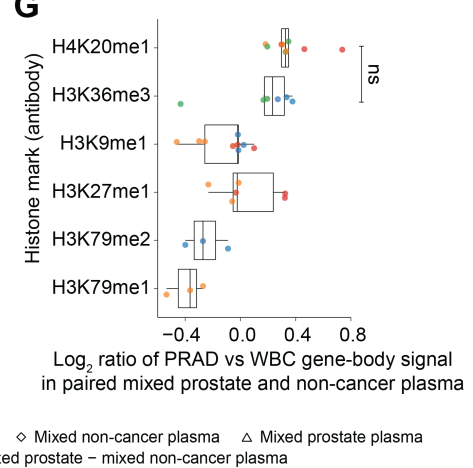
A**B****C****D****E****F****G**

Figure S1. Selection and benchmarking of antibodies to active gene body histone marks in plasma and cell-line ChIP-seq, related to Figure 1.

(A) Barplots showing mean (solid bars) $\pm$ standard deviation (error bars) of the fraction of the genome covered by significant peaks for seven active histone marks across multiple tissues using data obtained from Roadmap Epigenomic tissue ChIP-seq (sample numbers indicated at right). Peaks were called with MACS v2.0.10 using matched whole-cell extract controls, applying narrow-peak thresholds (H3K4me3, H3K27ac; Poisson $P = 0.01$) or broad-peak thresholds (H3K36me3, H3K79me1, H3K79me2, H3K9me1, H4K20me1; Poisson $P = 0.1$) tailored to each mark. Statistical significance between each gene body mark and the promoter-associated mark H3K4me3 was evaluated using an unpaired one-sided t -test. **(B)** Boxplots showing Spearman correlations between ChIP-seq coverage of seven active histone marks at distinct regulatory regions (intergenic enhancers, intragenic enhancers, promoters, and gene bodies) and RNA-seq transcript abundance. ChIP-seq fragments (tagAlign files) and RNA-seq matrices were obtained from the Roadmap Epigenomics database (sample numbers as in **A**). For enhancers, coverage was calculated as the confidence-weighted mean across all GeneHancer-associated enhancers per gene. Schematic (top) indicates genomic regions relative to gene structure; the number of evaluable genes per region is shown above each box. **(C)** Schematic of experimental design for benchmarking antibodies to active gene body marks using cell-free (1 mL of mixed non-cancer and prostate cancer plasma) and low-input (1 μ g) cell-line (LNCaP) ChIP-seq. **(D)** Signal tracks of ChIP-seq coverage for active gene body marks across an active housekeeping gene, *GAPDH*, in 0.8–1 μ g fixed LNCaP chromatin and 1 mL mixed prostate cancer plasma. Individual tracks show different modifications with y-axis coverage range indicated in brackets. **(E–G)**, Performance of ChIP-seq antibodies in fixed LNCaP chromatin and mixed plasma samples, grouped by vendor (color) and sample type (shape). **(E)**, Number of unique sequencing fragments (millions). **(F)** Enrichment score (signal-to-noise ratio), defined as coverage at gene-body regions divided by signal in regions >10 kb from gene bodies; dashed red line indicates threshold from low-pass whole-genome sequencing (LPWGS; non-immunoprecipitated negative control; $n = 15$). **(G)** Log₂ ratio of average gene-body signal across the top 500 prostate (PRAD) versus top 500 white blood cell (WBC) genes in paired mixed prostate cancer and non-cancer plasma. Antibodies are ordered top-to-bottom by median performance. For **(E)** and **(G)**, statistical significance was determined with a two-sided unpaired t -test between histone marks, while for **(F)**, statistical significance was determined with a two-sided unpaired t -test between histone marks and LPWGS ($n = 24$ cancer plasma samples); **** $P < 0.0001$, ns, not significant.

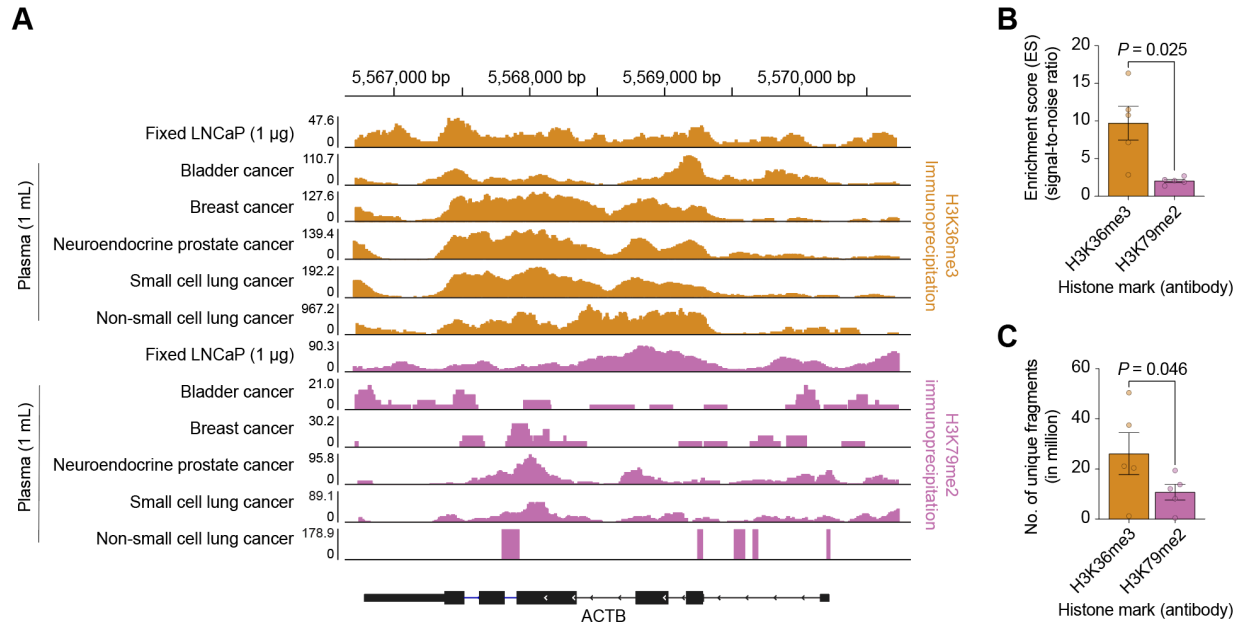


Figure S2. Comparison of H3K36me3 and H3K79me2 plasma cfChIP performance across different cancers, related to Figure 1. (A) Genome-browser tracks at the *ACTB* locus, a housekeeping gene, showing H3K36me3 (Cat. No. C15410192) and H3K79me2 (Cat. No. ab3594) immunoprecipitation signal from 1 μ g fixed LNCaP cells, used as a positive control, and from 1 mL plasma samples across five different cancer types. Tracks are independently scaled, with maximum values shown at left. **(B)** Enrichment scores for H3K36me3 and H3K79me2 plasma cfChIP libraries. Enrichment score was defined as the signal-to-noise ratio, calculated as coverage across gene-body regions divided by coverage in regions located >10 kb from gene bodies. **(C)** Number of unique fragments recovered from H3K36me3 and H3K79me2 plasma cfChIP libraries. Bars represent mean $\pm$ SEM; points represent individual libraries; P values were calculated using paired two-sided t -tests.

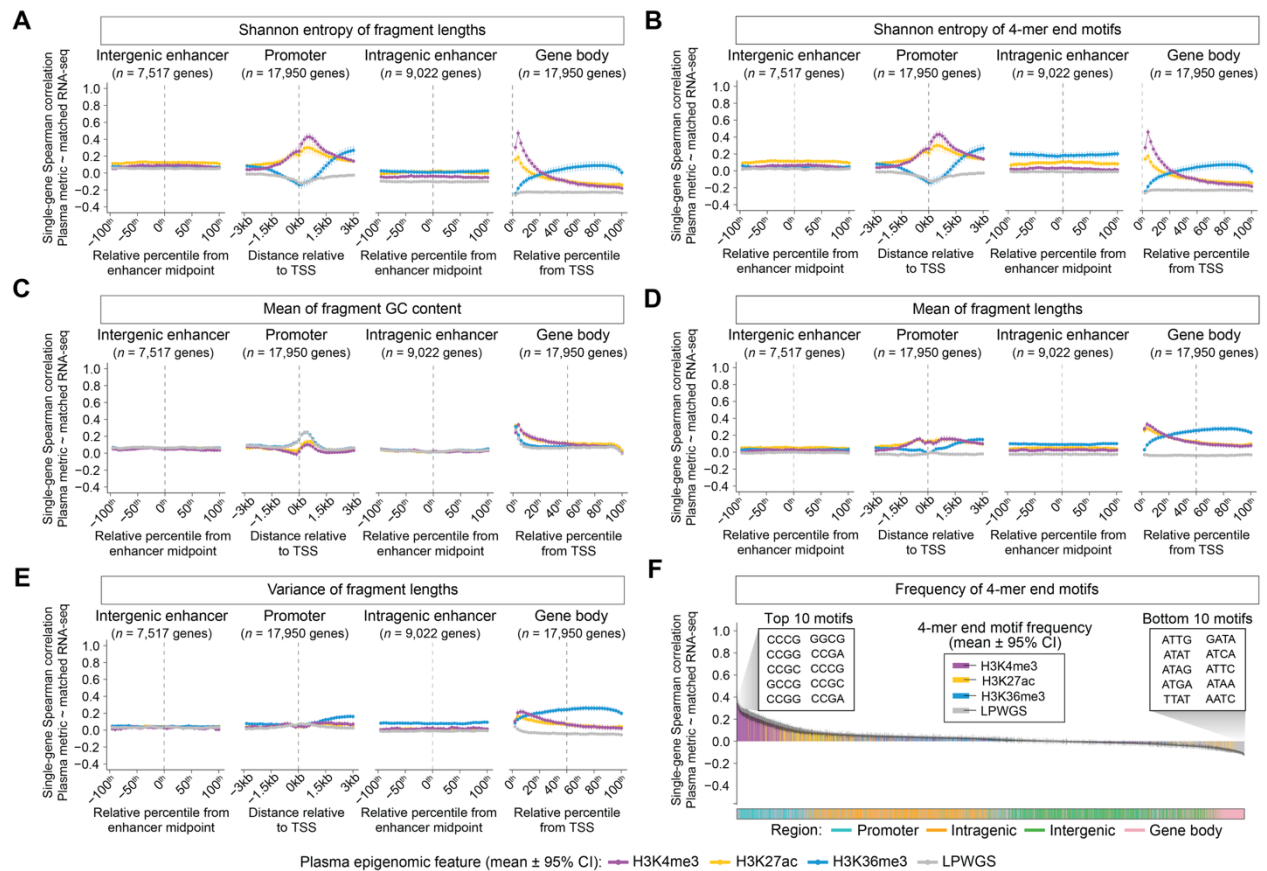


Figure S3. Association between plasma epigenomic features and matched tumor RNA-seq across regulatory regions, related to Figure 1. (A–E), Line plots show mean single-gene Spearman correlations (points) with 95% confidence intervals (vertical bars) between plasma-derived epigenomic features and matched tumor RNA-seq across protein-coding genes in 15 patients. Plasma-derived epigenomic features include **(A)** Shannon entropy of fragment lengths; **(B)** Shannon entropy of 4-mer end motifs; **(C)** mean fragment GC content; **(D)** mean of fragment lengths; and **(E)** variance of fragment lengths. These features were derived from H3K4me3 (purple), H3K27ac (yellow), and H3K36me3 (blue) cfChIP-seq and low-pass whole-genome sequencing (LPWGS; gray). Correlations were computed in 40 bins spanning intergenic enhancers, intragenic enhancers, promoters, and gene bodies. For enhancers, coverage was calculated as the confidence-weighted mean across all annotated enhancer peaks per gene. The vertical dotted line marks the transcription start site (TSS). Enhancer windows were normalized into 40 bins spanning -100 to 100 percentiles around the midpoint; promoter windows were divided into 40 bins across ± 3 kb from the transcription start site (TSS); and gene bodies were partitioned into 40 percentile bins from TSS to transcription termination site (TTS). **(F)** Barplots showing mean $\pm$ 95% confidence intervals for single-gene Spearman correlations between the frequency of each 4-mer end motif and matched tumor RNA-seq, ordered from most positively to most negatively correlated. Frequencies were calculated across promoter, intragenic enhancer, intergenic enhancer, and gene body regions (color-coded at bottom) for 15 patients in the training cohort. The ten 4-mer motifs most positively and negatively correlated with tumor RNA-seq are highlighted.

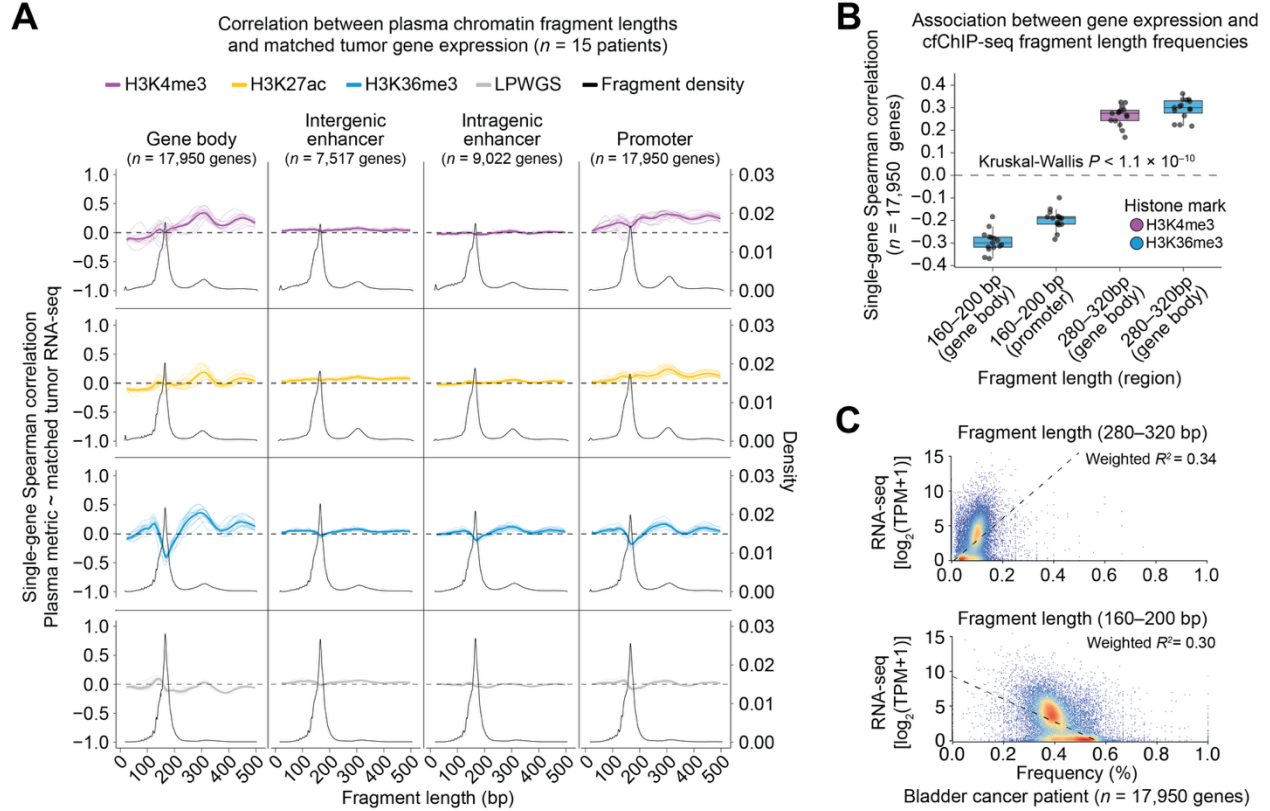


Figure S4. Association between plasma cell-free chromatin fragment length and matched tumor RNA-seq, related to Figure 1. (A) Line plots showing the relationship between fragment length (x-axis) and single-gene Spearman correlation with matched tumor RNA-seq (y-axis) for H3K4me3 (purple), H3K27ac (yellow), H3K36me3 (blue), and low-pass whole-genome sequencing (LPWGS; gray) across promoters (transcription start site [TSS] ± 3 kb), intergenic enhancers, intragenic enhancers, and gene bodies (TSS to transcription termination site [TTS]). Mean correlations across 15 individual samples are shown (dark lines), with per-sample correlations shown in lighter shades. Black lines indicate normalized fragment density profiles, and the horizontal dashed gray line denotes zero Spearman correlation. **(B)** Boxplots showing single-gene Spearman correlations between matched tumor RNA-seq and the two most positively and negatively correlated fragment length-related features identified in the in silico screen (corresponding to **Figure 1D**; $n = 15$ patients). Statistical significance was assessed using a Kruskal–Wallis test. **(C)** Representative scatter plots from a bladder cancer patient in the training cohort showing associations between RNA-seq gene expression and cfChIP-seq fragment frequencies within the 280–320 bp (top) and 160–200 bp (bottom) fragment length ranges. Dashed lines depict linear regression weighted by a 2D kernel density estimate (R package MASS, v7.3-65). R^2 values summarize the strength of association.

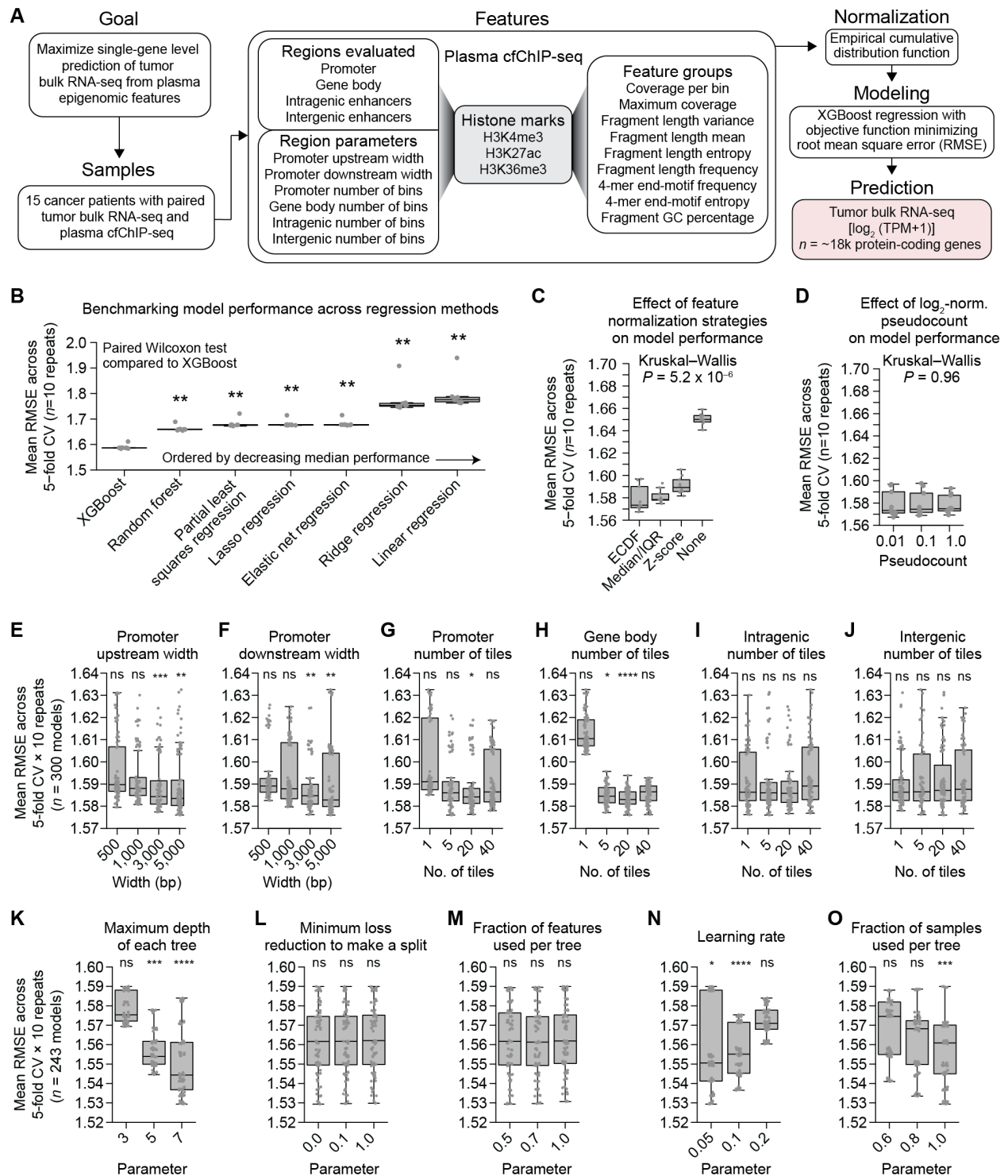


Figure S5. Optimization of APEX model parameters and input features, related to Figure 2. (A) Overview of the APEX modeling framework. Paired plasma cfChIP-seq and matched tumor bulk RNA-seq from 15 cancer patients were used to train a generalized model to predict $\log_2(\text{TPM}+1)$ expression for $\sim 18,000$ protein-coding genes. Input features included coverage metrics (per-bin and maximum), fragment-length statistics (mean, variance, entropy, and binned frequencies), 4-mer end-motif frequency and entropy, and fragment GC content. Six region-

specific parameters, including window size and number of tiles, were varied as shown in panels **E–J**. Models were constructed using extreme gradient boosting regression (XGBoost; panel **B**) following empirical cumulative distribution function (ECDF) normalization (panel **C**). Log₂ pseudocount settings are shown in panel **D**. XGBoost hyperparameters were varied as shown in panels **K–O**. **(B)**, Benchmarking regression models. Mean root-mean-square error (RMSE) across five-fold cross-validation × 10 repeats (~270k gene–sample pairs) for candidate models. Statistical comparisons were performed using paired two-sided Wilcoxon tests versus XGBoost. **(C)** Effect of ECDF normalization. RMSE across five-fold cross-validation × 10 repeats with and without ECDF normalization. Statistical comparison performed using a paired two-sided Wilcoxon test. **(D)** Effect of log₂ pseudocount choice. RMSE across five-fold cross-validation × 10 repeats using differing pseudocount values in the log₂ transformation. Statistical comparison was performed using a Kruskal–Wallis test. **(E–J)** Optimization of genomic tiling and promoter windowing parameters. RMSE across five-fold cross-validation × 10 repeats for models trained with different **(E)** promoter upstream window widths, **(F)** promoter downstream window widths, **(G)** numbers of promoter tiles, **(H)** numbers of gene-body tiles, **(I)** numbers of intragenic enhancer tiles, and **(J)** numbers of intergenic enhancer tiles. Statistical comparisons were performed using two-sided Wilcoxon rank-sum tests comparing each parameter setting versus all others (ns, not significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$). **(K–O)** Optimization of XGBoost hyperparameters. RMSE across five-fold cross-validation × 10 repeats for models trained with different **(K)** maximum tree depths, **(L)** minimum loss reduction (gamma) thresholds, **(M)** fractions of features per tree, **(N)** learning rates (eta), and **(O)** fractions of samples per tree (subsample). Statistical comparisons were performed using two-sided Wilcoxon rank-sum tests comparing each parameter setting versus all others (ns, not significant; * $P < 0.05$; *** $P < 0.001$; **** $P < 0.0001$).

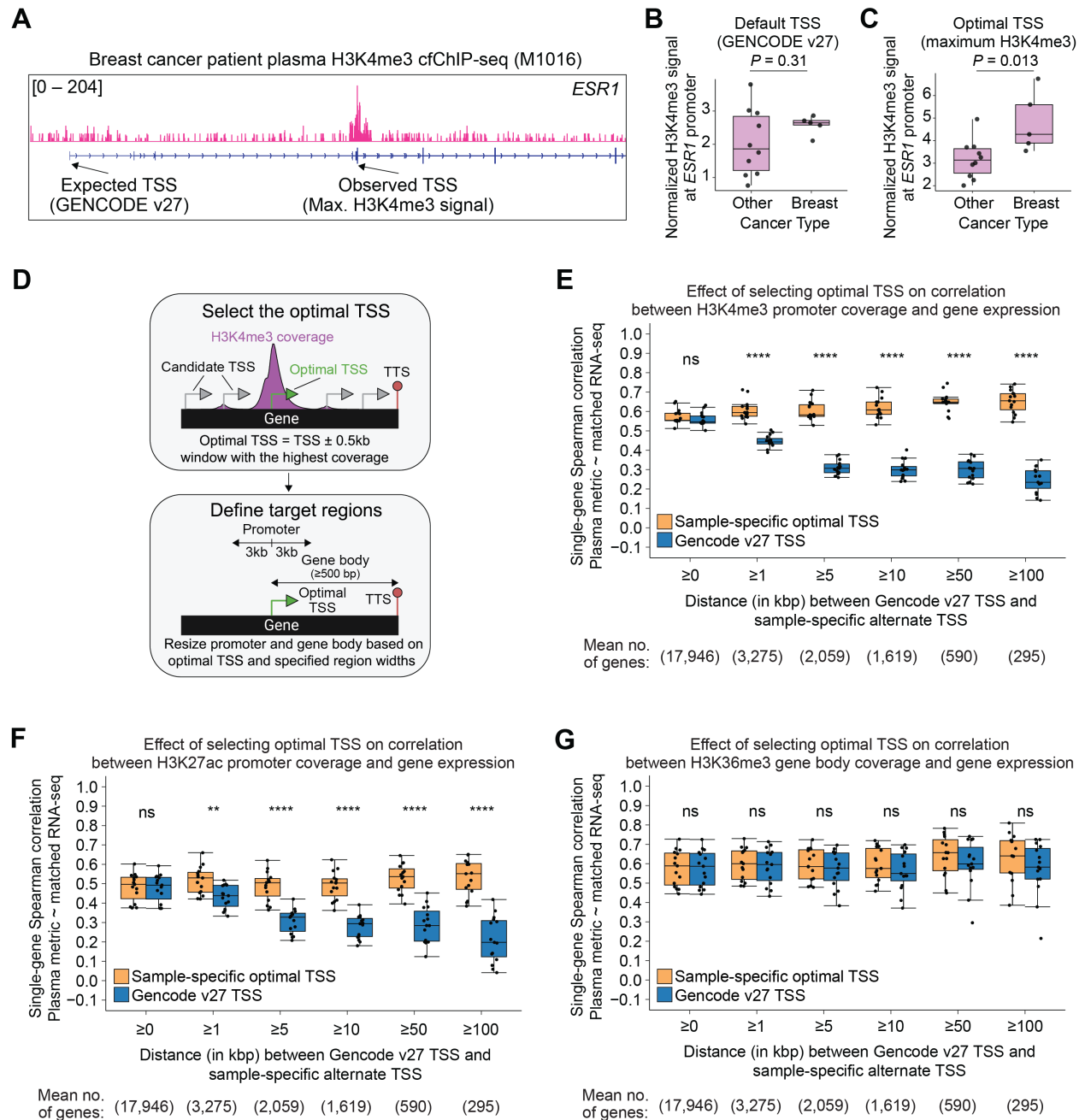


Figure S6. Selecting optimal transcription start sites (TSSs) improves cfChIP-seq correlation with tumor gene expression, related to Figure 2. (A) Signal tracks of cfChIP-seq coverage for H3K4me3 in plasma from a breast cancer patient, showing the expected TSS (GENCODE v27 annotation) and the sample-specific TSS corresponding to the maximum H3K4me3 signal at the *ESR1* locus. **(B and C)** Boxplots showing normalized H3K4me3 signal at the *ESR1* promoter using the default GENCODE v27 TSS **(B)** and the sample-specific optimal TSS **(C)**, comparing patients with breast cancer and other cancer types in the training cohort. Statistical significance was determined using a two-sided Wilcoxon test. **(D)** Schematic illustrating the method for selecting the optimal TSS and defining target regions. For each gene, the optimal TSS was identified as the ± 0.5 kb window with the highest H3K4me3 coverage among all annotated TSSs. Promoter (± 3 kb from TSS) and gene body (at least 500 bp) regions

were then resized based on the optimal TSS and the GENCODE v27 transcription termination sites (TTS). **(E–G)**, Boxplots showing the effect of using sample-specific optimal TSSs on correlations between plasma histone coverage and matched tumor RNA-seq in the training cohort ($n = 15$ patients) for **(E)** H3K4me3 promoter coverage, **(F)** H3K27ac promoter coverage, and **(G)** H3K36me3 gene-body coverage. Genes are stratified by the genomic distance between the GENCODE v27 TSS and the sample-specific optimal TSS. The mean number of evaluable genes across patients per group is shown below. Statistical significance was assessed using two-sided unpaired Wilcoxon tests (ns, not significant; ** $P < 0.01$; **** $P < 0.0001$).

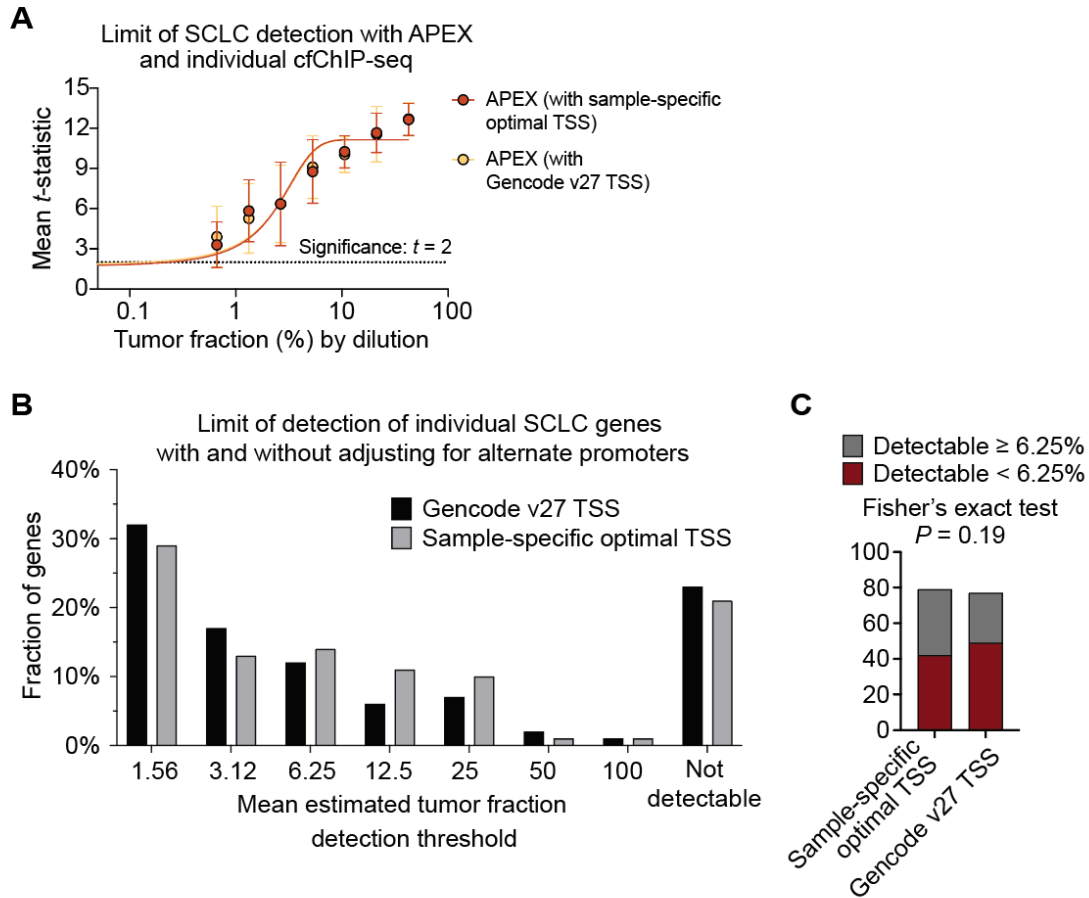


Figure S7. Effect of sample-specific promoter selection by tumor fraction, related to Figure 2. (A) Detection of a 100-gene SCLC signature across tumor fractions achieved by serial dilution when APEX is run with sample-specific optimal transcription start sites (TSSs) or fixed GENCODE v27 TSSs (related to **Figure 2G–I**). Points and error bars show mean $\pm$ SEM. **(B)** Distribution of per-gene detection thresholds for individual SCLC marker genes using GENCODE v27 TSSs or sample-specific optimal TSSs. **(C)** Fraction of SCLC marker genes detectable below or above 6.25% tumor fraction using each TSS strategy. Statistical significance was assessed by Fisher's exact test.

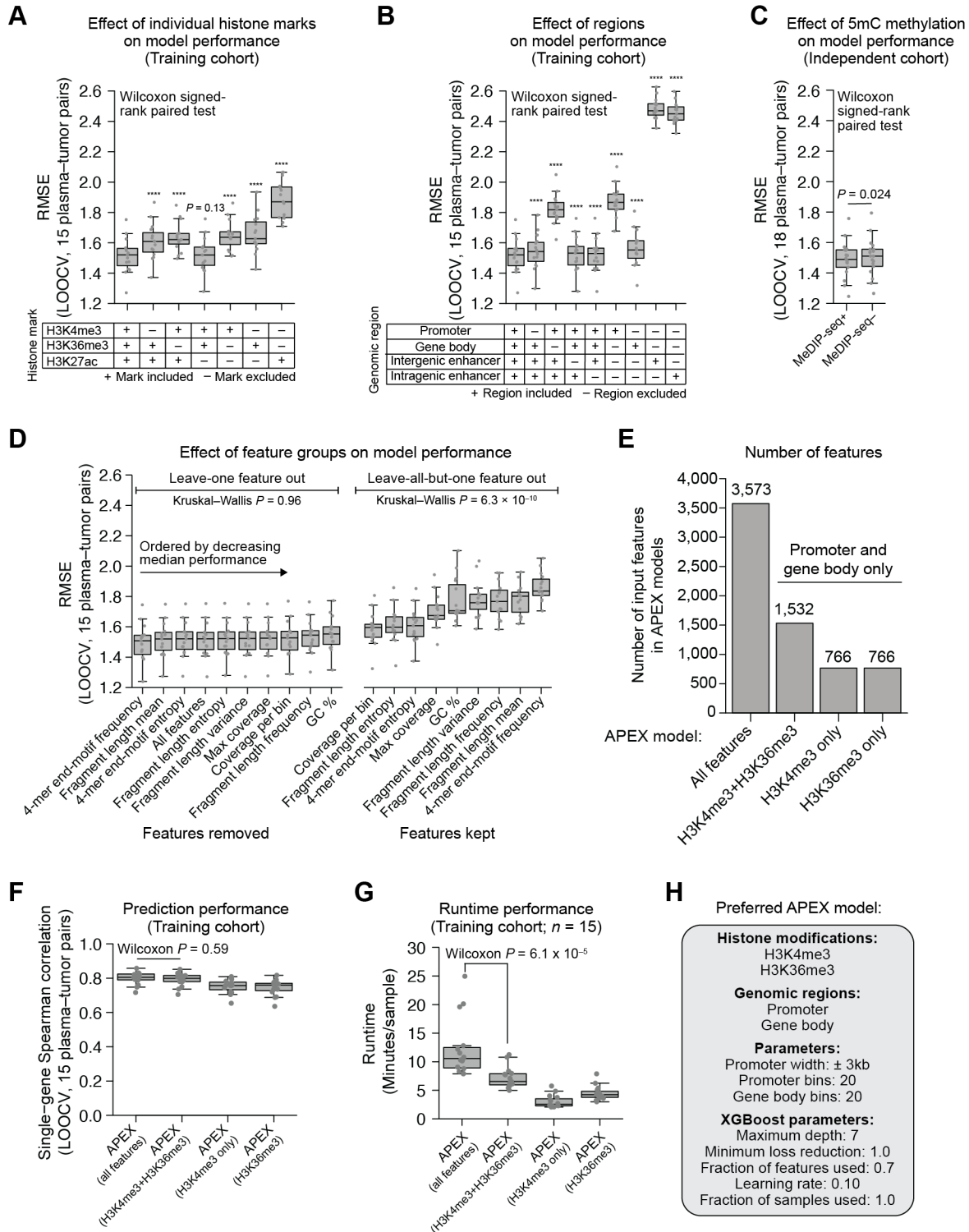


Figure S8. Determinants of APEX model performance and preferred feature configuration, related to Figure 2. (A) Effect of individual histone marks on model performance. Root mean square error (RMSE) under leave-one-out cross-validation (LOOCV) across 15 plasma-tumor pairs when H3K4me3, H3K27ac, or H3K36me3 features were included or excluded. Statistical comparisons were performed using paired two-sided Wilcoxon signed-rank tests relative to the model including all marks (**** $P < 0.0001$). **(B)** Effect of genomic

regions on model performance. RMSE under LOOCV across 15 plasma–tumor pairs, comparing models in which promoter, gene body, intergenic enhancer, or intragenic enhancer features were included or excluded. Statistical comparisons were performed using paired two-sided Wilcoxon signed-rank tests relative to the model including all regions (ns, not significant; **** $P < 0.0001$). **(C)** Effect of 5mC DNA methylation on model performance. RMSE under LOOCV across 18 plasma–tumor pairs using matched cfMeDIP-seq for methylation profiling (see **Methods**). Statistical comparison was performed using a paired two-sided Wilcoxon signed-rank test. **(D)** Effect of feature groups on model performance. RMSE under LOOCV across 15 plasma–tumor pairs in leave-one-feature-group-out and leave-all-but-one-feature-group-in analyses. Feature groups were compared using Kruskal–Wallis tests. **(E)** Number of input features across APEX model variants. Counts are shown for the full model, a refined model using H3K4me3 and H3K36me3 features from promoter and gene-body regions, and single-mark models using only H3K4me3 or H3K36me3 features. **(F)** Prediction performance across APEX model variants. Single-gene Spearman correlations between predicted and measured tumor RNA-seq expression across 15 plasma–tumor pairs. Statistical comparison between full and refined models was performed using paired two-sided Wilcoxon signed-rank tests. **(G)** Runtime per sample across APEX model variants. Runtime (minutes) measured on a system with 16 CPU cores and 64 GB RAM. Statistical comparison between full and refined models was performed using paired two-sided Wilcoxon signed-rank tests. **(H)** Preferred APEX model configuration. Summary of histone marks, genomic regions, and XGBoost parameters selected for the final model.

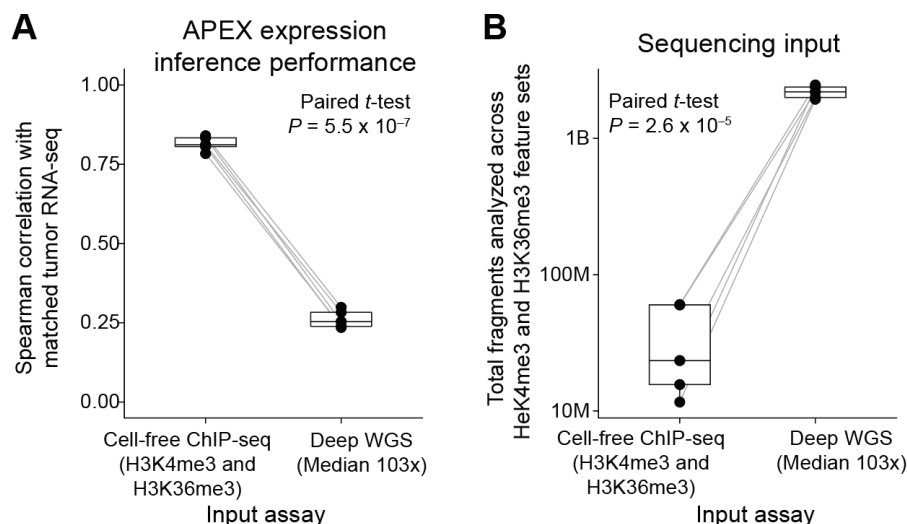
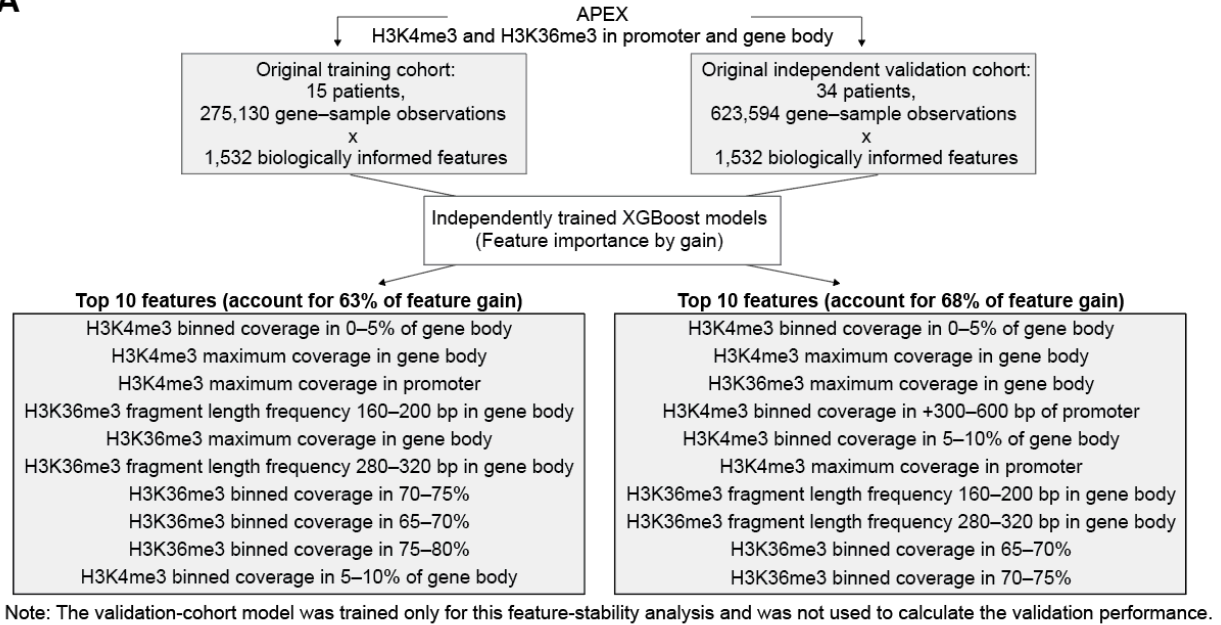


Figure S10. APEX performance using cfChIP-seq and deep whole-genome sequencing inputs, related to Figure 2. (A) APEX gene-expression inference performance using H3K4me3 and H3K36me3 cfChIP-seq data compared with deep whole-genome sequencing data from the same plasma samples. For both input types, predictions were generated using the same APEX model architecture, feature definitions, and learned feature weights derived from the cfChIP-seq-trained model. **(B)** Total fragment observations analyzed for the corresponding cell-free ChIP-seq and deep whole-genome sequencing datasets. Cell-free ChIP-seq values represent the combined H3K4me3 and H3K36me3 fragment counts. For whole-genome sequencing, the same fragment set was evaluated across the H3K4me3- and H3K36me3-derived feature spaces. Points represent individual patients, and lines connect paired measurements from the same patient. Boxplots show medians, interquartile ranges, and 1.5× IQR whiskers; $n = 5$ paired samples.

A



B

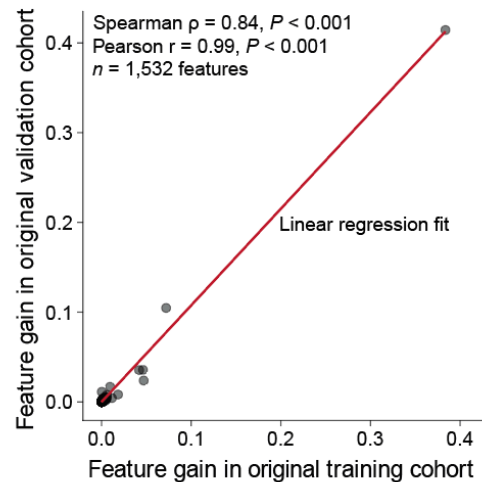


Figure S11. Reproducibility of APEX feature importance across independently trained patient cohorts, related to Figure 2. (A) Schematic of the feature-stability analysis. Separate XGBoost models were trained de novo in the original 15-patient training cohort and the independent 34-patient validation cohort using the same prespecified set of 1,532 H3K4me3- and H3K36me3-derived promoter and gene-body features. Feature importance was quantified using normalized XGBoost gain. Nine of the ten highest-ranked features were shared between the two models and collectively accounted for 63% and 68% of total feature gain in the training and validation cohorts, respectively. The validation-cohort model was trained solely for this feature-stability analysis and was not used to calculate the reported validation performance. **(B)** Concordance of feature gain across the 1,532 features in the independently trained models. Each point represents one feature. Feature-importance rankings were strongly concordant across cohorts (Spearman $\rho = 0.84$, $P < 0.001$; Pearson $r = 0.99$, $P < 0.001$). The red line indicates the linear regression fit.

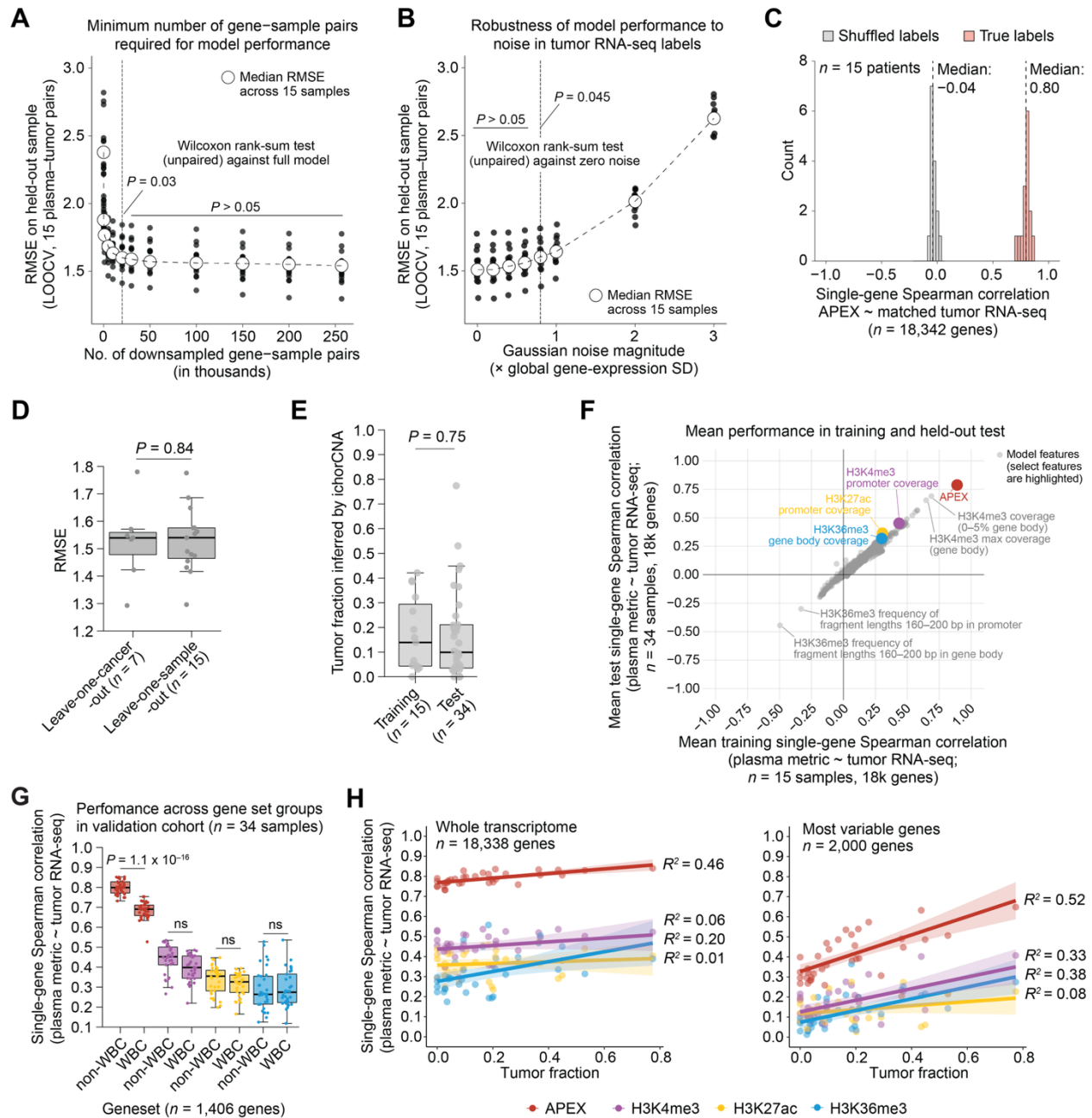


Figure S12. Robustness and generalization of the APEX model, related to Figure 2. (A) Sensitivity of APEX model performance to training data size. Root mean square error (RMSE) under leave-one-out cross-validation (LOOCV) across 15 plasma-tumor pairs is shown for models trained on progressively downsampled numbers of gene-sample pairs (x-axis). Dashed line indicates the minimum number of feature-gene pairs required to maintain comparable performance to the full model. Statistical significance was assessed using unpaired two-sided Wilcoxon rank-sum tests against the full model. **(B)** Robustness of model performance to noise in tumor RNA-seq labels. Gaussian noise (in units of the global standard deviation) was added to expression values prior to model training and RMSE was evaluated on unperturbed held-out samples using LOOCV. Statistical significance was assessed using unpaired two-sided Wilcoxon rank-sum tests against the zero-noise condition. **(C)** Distribution of single-gene

Spearman correlations between APEX-inferred gene expression and matched tumor RNA-seq across 18,342 genes in the training cohort ($n = 15$ patients), estimated using LOOCV. The null distribution was generated by shuffling RNA-seq labels. Median correlation values are indicated.

(D) Generalization of the APEX model to unseen cancer types via cross-validation. Comparison of RMSE between leave-one-sample-out cross-validation ($n = 15$; training on 14 samples and testing on one held-out sample) and leave-one-cancer-type-out cross-validation ($n = 7$; training on six cancer types and testing on a held-out cancer type). Statistical significance was assessed using a two-sided unpaired Wilcoxon rank-sum test.

(E) Tumor fraction estimated by ichorCNA in training ($n = 15$) and test ($n = 34$) cohorts. Boxplots show medians, quartiles, and $1.5 \times \text{IQR}$; individual samples are plotted as points. Statistical significance was assessed using a two-sided unpaired Wilcoxon rank-sum test.

(F) Mean single-gene Spearman correlation between plasma epigenomic metrics and matched tumor RNA-seq across the training (x-axis) and held-out test (y-axis) sets ($n = 34$ samples; $\sim 18\text{k}$ genes). Points represent individual feature types and colored points indicate select plasma-based metrics, including APEX-inferred gene expression and histone-mark coverage.

(G), Same as **Figure 2F**, but showing single-gene Spearman correlations across WBC and non-WBC genes (1,406 genes in each set). Statistical significance was calculated by paired Wilcoxon test between WBC and non-WBC gene sets for each approach (color legend below **H**).

(H) Scatter plots show single-gene Spearman correlations between plasma-derived metrics and matched tumor RNA-seq as a function of tumor fraction. (Left) Correlations computed across the whole transcriptome ($n = 18,338$ genes). (Right) Correlations computed using the 2,000 most variable genes. Lines indicate linear regression fits with 95% confidence intervals. Coefficients of determination (R^2) are shown for each metric.

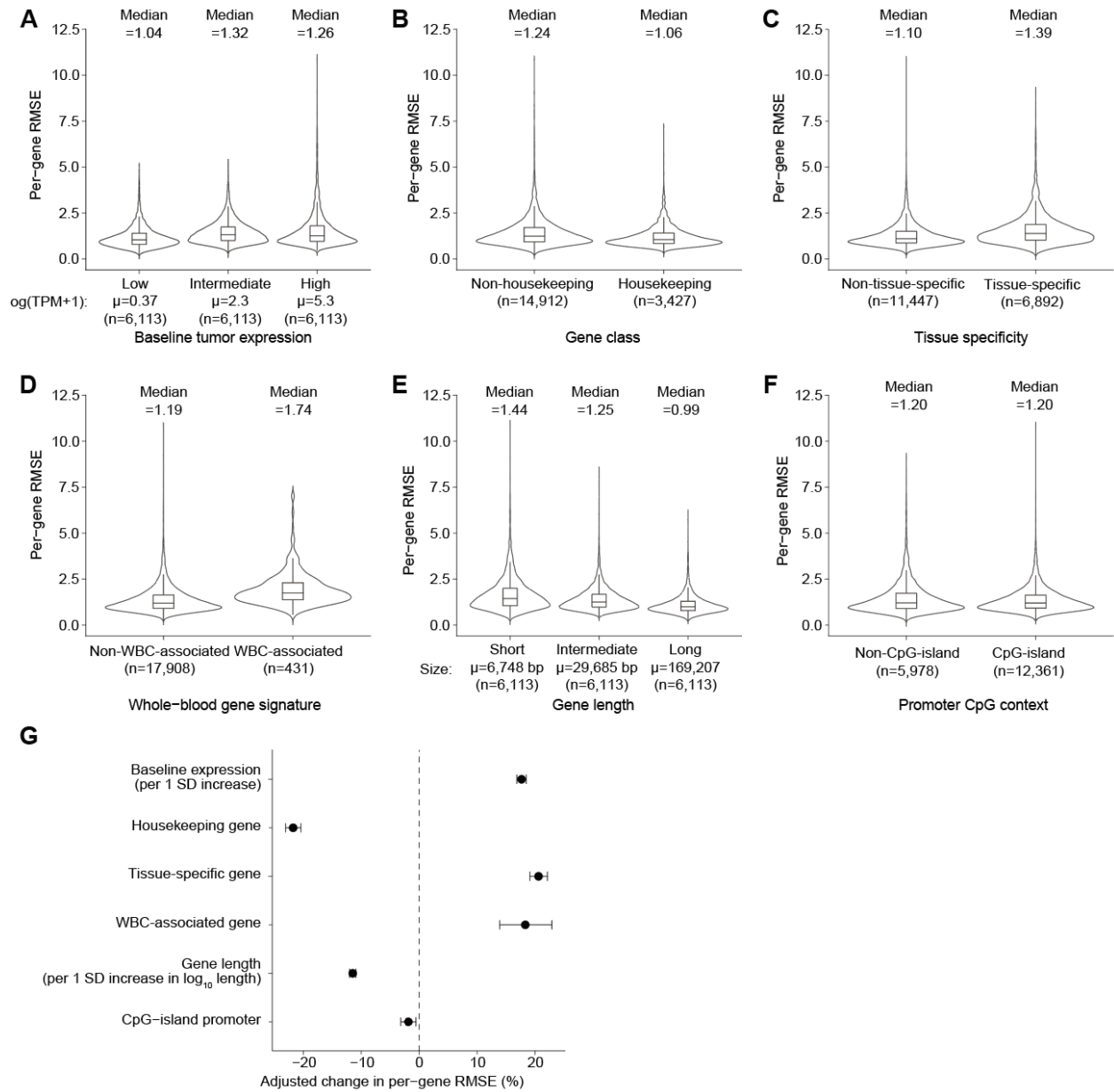


Figure S13. Gene-level determinants of APEX prediction accuracy in the independent validation cohort, related to Figure 2. Per-gene RMSE was calculated across 34 matched plasma–tumor pairs for 18,339 genes and stratified by **(A)** baseline tumor expression, **(B)** housekeeping-gene status, **(C)** tissue specificity, **(D)** whole-blood-associated gene status, **(E)** gene length, and **(F)** promoter CpG-island status. Violin plots show RMSE distributions, with embedded boxplots indicating medians and interquartile ranges; group medians and sample sizes are shown. **(G)** Multivariable associations between gene characteristics and prediction error. Points show the adjusted percentage change in per-gene RMSE with 95% confidence intervals from a linear model including all displayed gene characteristics. Continuous predictors were standardized; negative values indicate lower prediction error, and categorical estimates are relative to the corresponding reference groups.

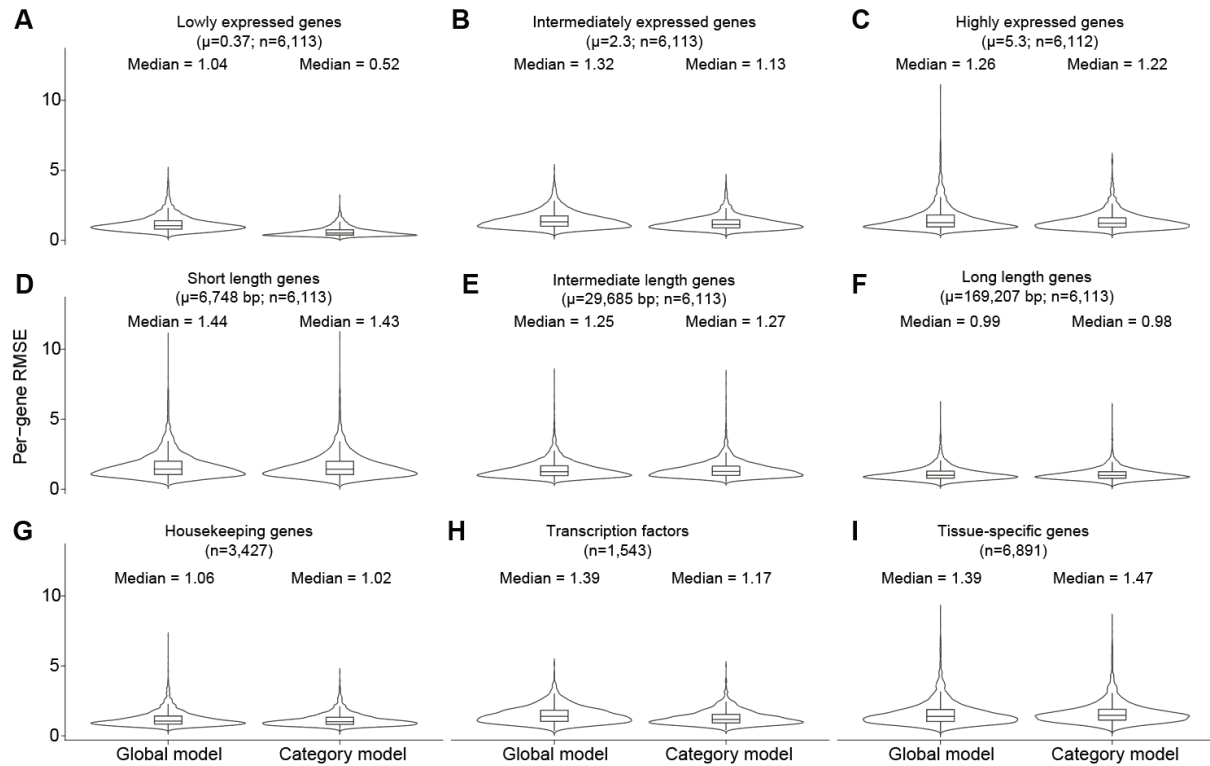


Figure S14. Performance of global and category-specific APEX models across gene classes, related to Figure 2. Violin and box plots show per-gene RMSE for the global APEX model and models trained separately within genes stratified by expression level (A–C), gene length (D–F), housekeeping status (G), transcription-factor identity (H), or tissue specificity (I). Medians are shown above each group; sample sizes indicate the number of genes in each category.

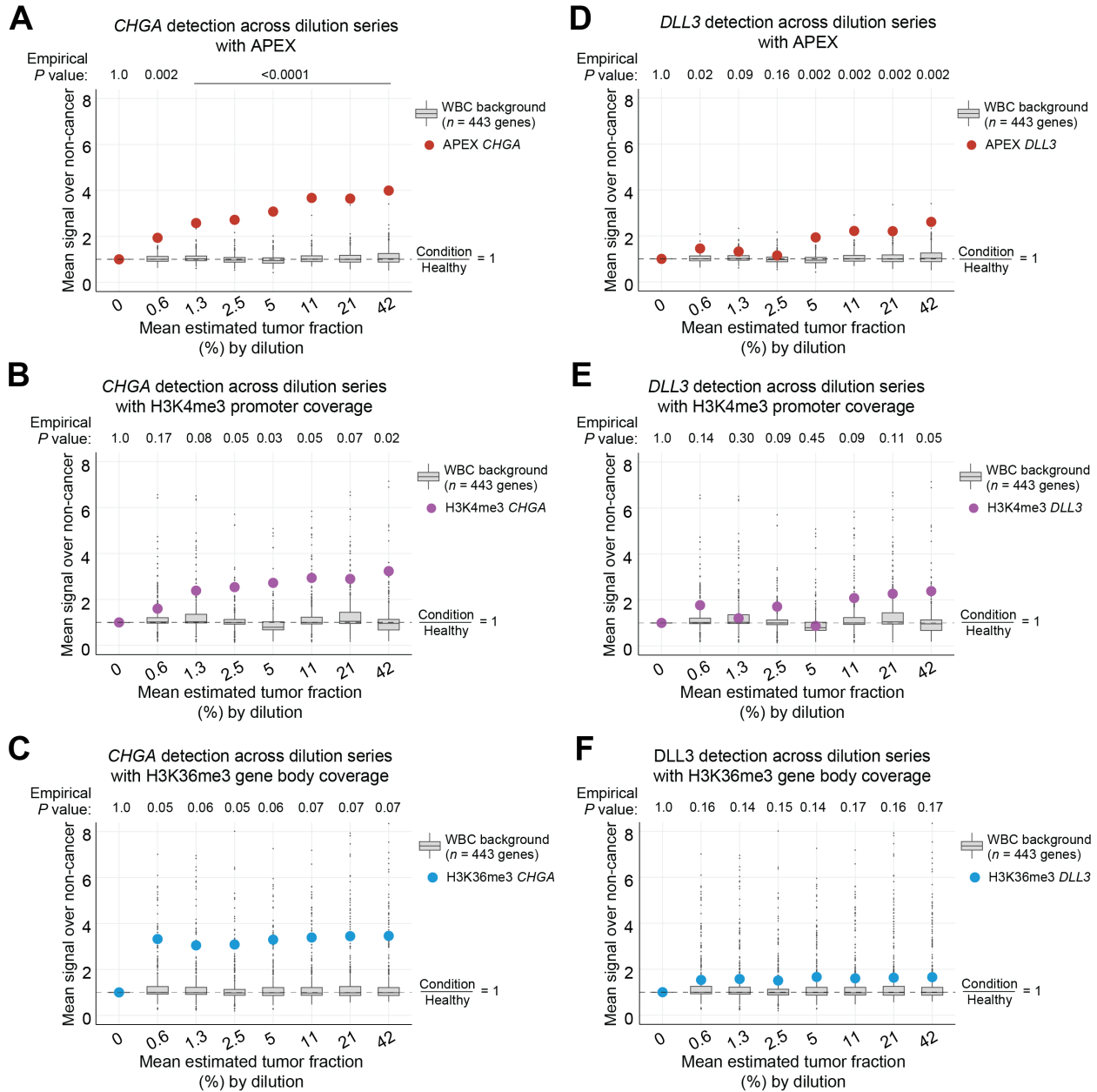


Figure S15. Single-gene detectability with APEX and individual cfChIP-seq across experimentally diluted tumor fractions, related to Figure 2. (A–F), Detection of *CHGA* (lineage marker; **A–C**) and *DLL3* (therapeutic target; **D–F**) across the cfChromatin dilution series (related to **Figures 2G–2I**) using APEX, H3K4me3 promoter coverage, and H3K36me3 gene-body coverage. For each dilution, the x-axis reflects the mean effective tumor fraction derived from the three starting tumor fractions. The y-axis reflects mean signal enrichment relative to non-cancer baseline, indicated by a colored circle for the target gene, and compared with a background distribution of enrichment across 443 WBC-signature genes (gray points and boxplots) to derive empirical *P*-values. Boxplots show medians, quartiles, and $1.5 \times \text{IQR}$.

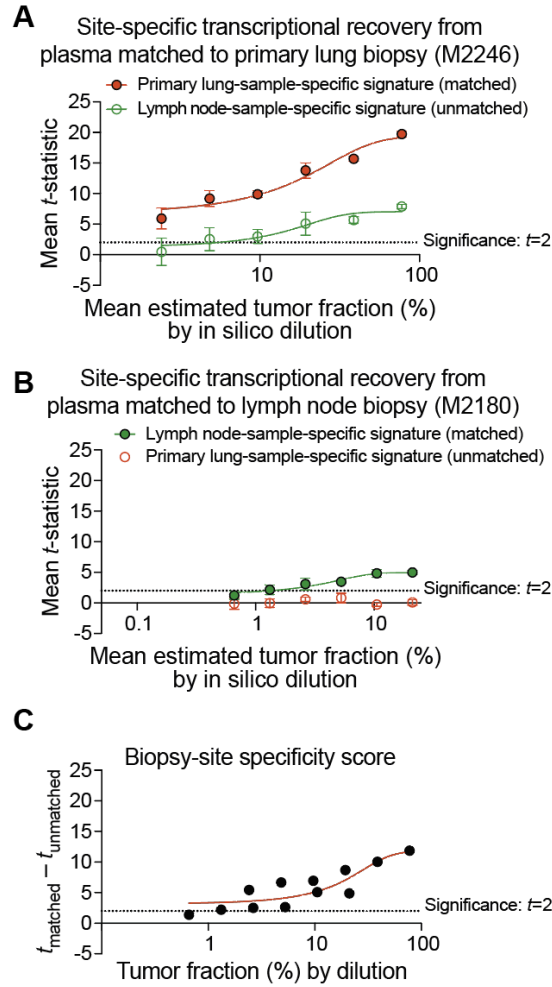


Figure S16. APEX retains biopsy-site-specific transcriptional signals across serial dilution, related to Figure 2. (A) Plasma from a patient with a matched primary lung adenocarcinoma biopsy (M2246). Site-specific expression signatures were defined using the top 100 genes enriched in either the primary lung or lymph node biopsy RNA-seq profile relative to the alternate biopsy site. Across serial in silico dilution, APEX preferentially recovered the matched primary lung-specific signature compared with the unmatched lymph node-specific signature. (B) Plasma from a patient with a matched lymph node biopsy (M2180). Across serial dilution, APEX preferentially recovered the matched lymph node-specific signature, whereas enrichment of the unmatched primary lung-specific signature remained minimal. (C) Biopsy-site specificity score, defined as the difference between enrichment of the matched and unmatched site-specific signatures ($t_{\text{matched}} - t_{\text{unmatched}}$), across all dilution points from both patients. Positive values indicate preferential recovery of the transcriptional program corresponding to the biopsy site matched to the plasma sample. The dashed line indicates the significance threshold ($t = 2$) used throughout the study.

		APEX predictions	
		Undetectable ctDNA ichorCNA ≤ 0.03	Detectable ctDNA ichorCNA > 0.03
		Classical Basal	Classical Basal
RNA-seq-predicted	Classical	9	2
	Basal	0	1
		Fisher's $P = 0.25$	Fisher's $P = 0.049$

Figure S17. Concordance between APEX- and tissue RNA-seq-based PDAC subtype classification according to plasma tumor fraction, related to Figure 4. Confusion matrices compare APEX and tissue RNA-seq PuriST classifications in patients with undetectable ctDNA by ichorCNA ($<3\%$; left) and detectable ctDNA ($\geq 3\%$; right). Rows indicate tissue RNA-seq classification and columns indicate APEX classification. Concordance was assessed using Fisher's exact test. Association between APEX and tissue subtype was significant among patients with detectable ctDNA ($P=0.049$), whereas the ctDNA-undetectable subgroup contained only one tissue-defined basal tumor and therefore provided limited power to assess subtype concordance ($P=0.25$).

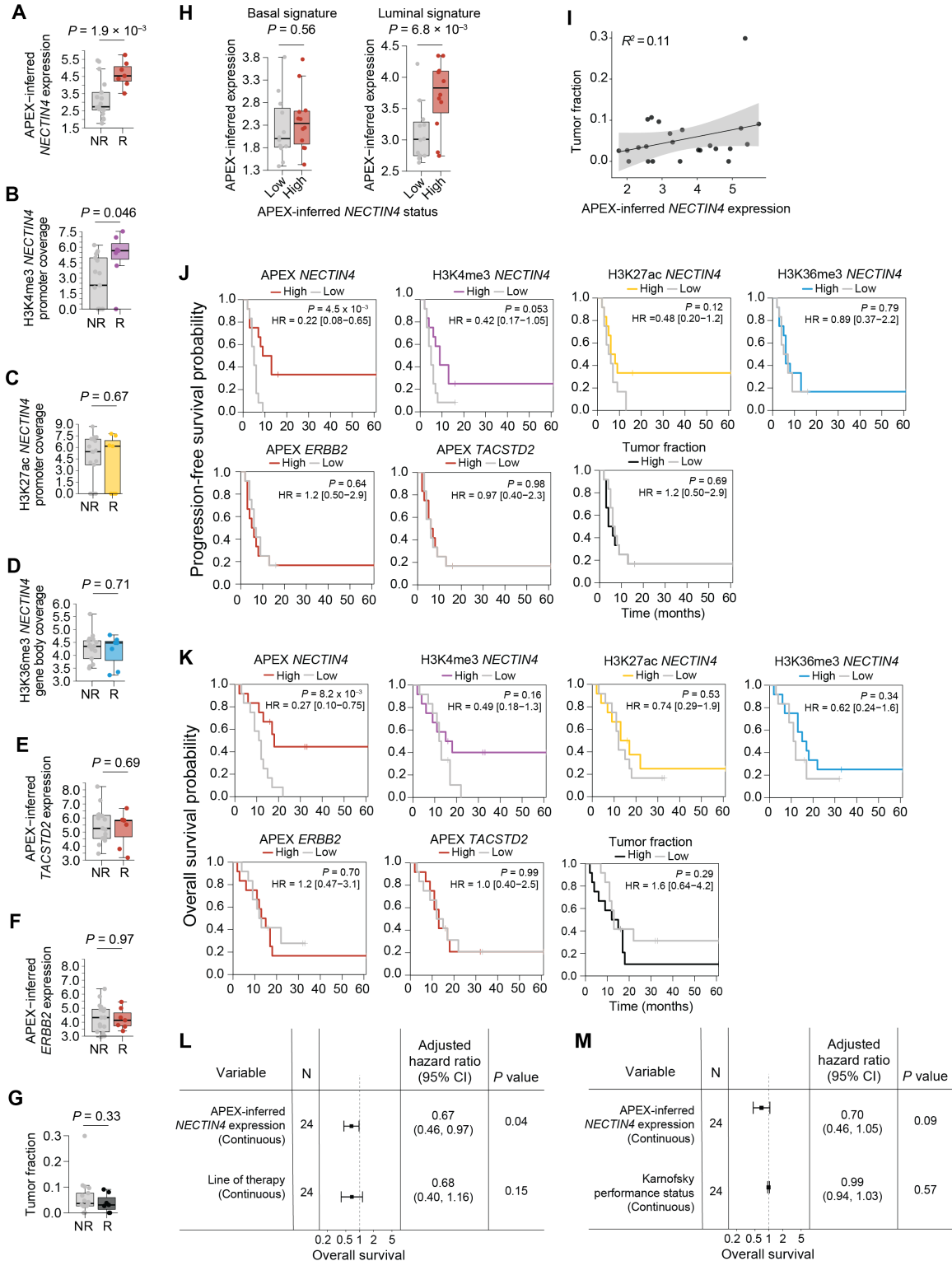


Figure S18. Baseline plasma-based epigenomic and genomic features and clinical outcomes in metastatic bladder cancer treated with enfortumab vedotin, related to Figure 6. (A–D), Comparison of baseline plasma-inferred *NECTIN4* expression between non-responders (NR; $n = 17$) and responders (R; $n = 7$) to enfortumab vedotin (EV). Responders were defined as patients with complete or partial response (CR/PR), and non-responders as

those with stable or progressive disease (SD/PD). *NECTIN4* expression was quantified using **(A)** APEX-inferred expression, **(B)** H3K4me3 promoter coverage, **(C)** H3K27ac promoter coverage, and **(D)** H3K36me3 gene-body coverage. **(E–G)**, Comparison of additional antibody-drug conjugate (ADC) targets and genomic features between NR and R, including **(E)** APEX-inferred *TACSTD2* (TROP2) expression, **(F)** APEX-inferred *ERBB2* expression, and **(G)** tumor fraction. **(H)** Association of APEX-inferred *NECTIN4* expression, dichotomized into high and low groups by the cohort median, with basal and luminal urothelial lineage signatures. **(I)** Correlation of APEX-inferred *NECTIN4* expression with plasma tumor fraction estimated using ichorCNA (v0.2.0). Linear regression fits are shown with coefficients of determination (R^2). Shaded regions denote 95% confidence intervals. **(J)** Progression-free survival and **(K)** overall survival stratified by high versus low levels of each indicated baseline plasma-derived feature, with high and low groups defined by median dichotomization of the corresponding feature across the cohort. Features shown include histone mark-specific cfChIP-seq coverage at the *NECTIN4* locus (H3K4me3 promoter, H3K27ac promoter, H3K36me3 gene body), APEX-inferred *NECTIN4* expression, APEX-inferred expression of additional ADC targets (*TACSTD2*, *ERBB2*), and plasma tumor fraction. Survival curves were compared using two-sided log-rank tests; hazard ratios (HR) and 95% confidence intervals are shown. Box plots indicate the median and interquartile range; whiskers denote $1.5 \times \text{IQR}$. For **A–H**, P -values were calculated using unpaired two-sided t -tests. **(L and M)**, Multivariable Cox model for overall survival including APEX-inferred *NECTIN4* expression and **(L)** line of therapy at EV initiation and **(M)** Karnofsky performance status at ADC start as continuous variables.

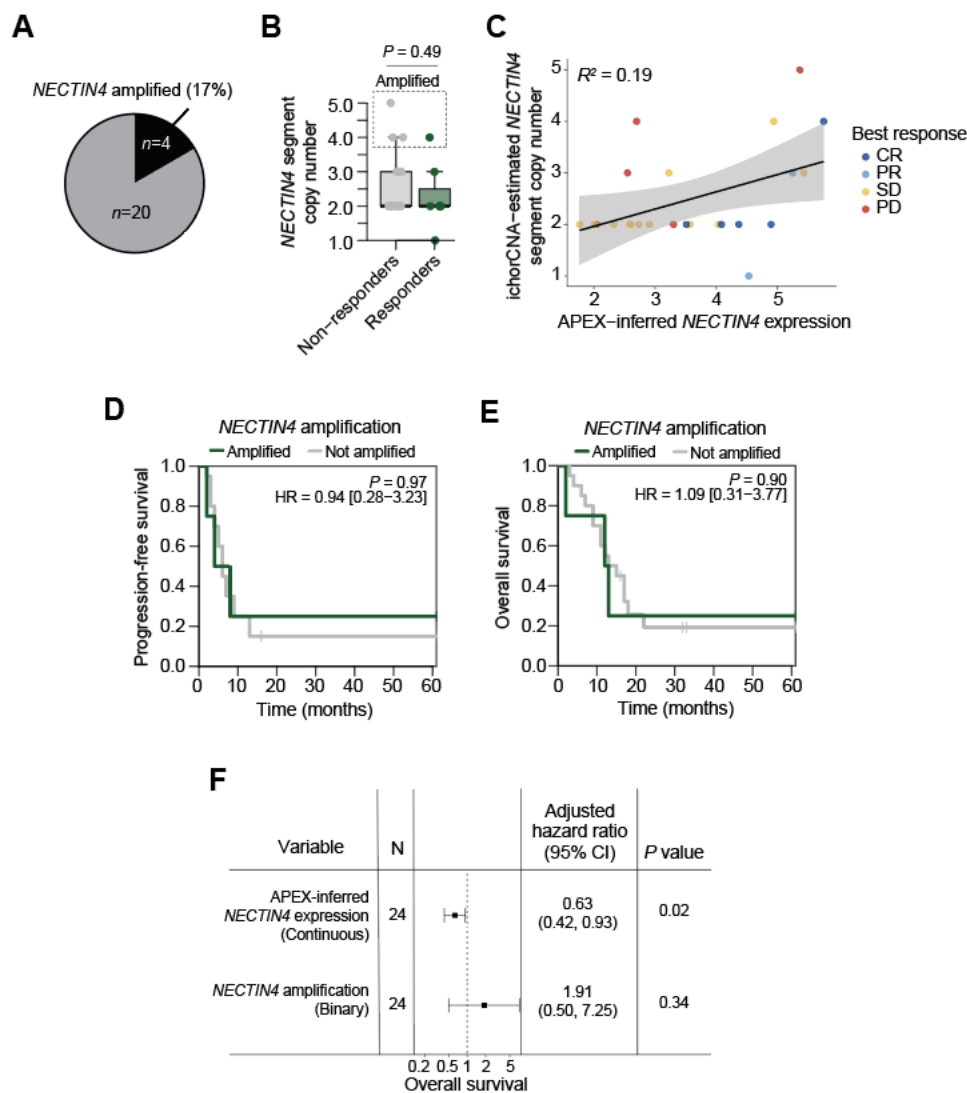


Figure S19. APEX-inferred *NECTIN4* expression is associated with outcome independently of *NECTIN4* amplification, related to Figure 6. (A) Prevalence of ichorCNA-defined *NECTIN4* amplification. **(B)** *NECTIN4* segment copy number in responders and nonresponders to enfortumab vedotin. **(C)** Association between APEX-inferred *NECTIN4* expression and ichorCNA-estimated *NECTIN4* segment copy number; points are colored by best response. **(D–E)** Progression-free and overall survival according to *NECTIN4* amplification status. **(F)** Multivariable Cox model for overall survival including continuous APEX-inferred *NECTIN4* expression and binary amplification status.