

Integrating Genomic and Fragmentomic Features for Tissue-of-Origin Prediction from Routine Targeted ctDNA Sequencing

Jaeryuk Kim

Samsung Medical Center, Sungkyunkwan University School of Medicine

You-Keun Ko

Samsung Medical Center, Sungkyunkwan University School of Medicine

Seo Yoon Jang

Samsung Medical Center, Sungkyunkwan University School of Medicine

Ji Eun Shin

Samsung Medical Center, Sungkyunkwan University School of Medicine

Junkyu Kim

Samsung Medical Center, Sungkyunkwan University School of Medicine

Junghoon Shin

Samsung Medical Center, Sungkyunkwan University School of Medicine

Sehhoon Park

Samsung Medical Center, Sungkyunkwan University School of Medicine

Minsuk Kwon

Samsung Medical Center, Sungkyunkwan University School of Medicine

Sung Hee Lim

Samsung Medical Center, Sungkyunkwan University School of Medicine

Hyun Ae Jung

Samsung Medical Center, Sungkyunkwan University School of Medicine

Minyong Kang

Samsung Medical Center, Sungkyunkwan University School of Medicine

Jung Yong Hong

Samsung Medical Center, Sungkyunkwan University School of Medicine

Seung Tae Kim

Samsung Medical Center, Sungkyunkwan University School of Medicine

Jong-Mu Sun

Samsung Medical Center, Sungkyunkwan University School of Medicine

Jeeyun Lee

Samsung Medical Center, Sungkyunkwan University School of Medicine

Sang-Won Um

Samsung Medical Center, Sungkyunkwan University School of Medicine

Se Hoon Park

Samsung Medical Center, Sungkyunkwan University School of Medicine

Yeon Hee Park

Samsung Medical Center, Sungkyunkwan University School of Medicine

Seong Il Seo

Samsung Medical Center, Sungkyunkwan University School of Medicine

Se-Hoon Lee

Samsung Medical Center, Sungkyunkwan University School of Medicine

Seong Soo Jeon

Samsung Medical Center, Sungkyunkwan University School of Medicine

Joon Oh Park

Samsung Medical Center, Sungkyunkwan University School of Medicine

Jin Seok Ahn

Samsung Medical Center, Sungkyunkwan University School of Medicine

Young Suk Park

Samsung Medical Center, Sungkyunkwan University School of Medicine

Young-gon Kim

`younggonn.kim@samsung.com`

Samsung Medical Center, Sungkyunkwan University School of Medicine

Article

Keywords: circulating tumor DNA, fragmentomics, transcription factor binding sites, tissue of origin, cancer of unknown primary, machine learning, liquid biopsy

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11257888/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Tissue-of-origin determination is important for treatment selection in cancer of unknown primary but remains challenging, with molecular classifiers often requiring dedicated assays beyond routine tumor genomic profiling. We asked whether this information could instead be recovered from the targeted circulating tumor DNA (ctDNA) sequencing that patients already receive. We developed Site-directed High-resolution Analysis of Positional Endpoints and Entropy (SHAPE), which exploits the high local sequencing depth of a routine 523-gene panel to resolve fragment-end positioning and fragment-length diversity at individual transcription factor binding sites, without whole-genome sequencing or dedicated regulatory-region capture. Across 1,889 plasma samples from patients with metastatic cancer spanning 11 primary types, profiled on two versions of a clinical assay, an ensemble integrating SHAPE with somatic mutations, broad copy-number alterations, all-exon depth, and exon 1 fragment-length entropy predicted the tissue of origin with macro-averaged areas under the receiver operating characteristic curve (AUROC) of 0.870 and 0.872, exceeding every individual signal; at estimated tumor fractions $\geq 10\%$, AUROC reached 0.94–0.96 with approximately 80% top-1 accuracy. Site-resolved SHAPE substantially outperformed conventional transcription factor–level aggregation (AUROC 0.816–0.818 versus 0.726–0.736). Feature contributions varied by cancer type and aligned with established tumor biology, and the framework generalized to an independent targeted cfDNA cohort (AUROC 0.944). Integrating complementary genomic and fragmentomic signals enables biologically interpretable tissue-of-origin prediction from existing ctDNA data, potentially extending the clinical utility of routine testing without additional sequencing.

Introduction

Identifying the tissue of origin helps guide treatment for metastatic cancer, yet the primary site remains unresolved in some patients despite imaging and histopathologic evaluation.^{1,2} Gene-expression and DNA-methylation classifiers can suggest an origin in cancer of unknown primary (CUP), but generally require dedicated assays beyond routine tumor genotyping.^{3–6} Recovering tissue-of-origin information from sequencing already performed for clinical care could therefore extend the diagnostic value of existing tests.

Targeted sequencing data generated from tumor tissue biopsies for routine genomic profiling have enabled cancer-type prediction using somatic mutations, copy number alterations, and mutational signatures.^{7,8} A tissue-trained classifier has also been evaluated in a small plasma cohort using genomic features alone.⁷ Beyond genomic alterations, plasma cell-free DNA (cfDNA) carries tissue-associated information in its fragmentation patterns.⁹ Fragment lengths, endpoint distributions, and regional coverage are shaped in part by nucleosomes and other DNA-binding proteins in the cells of origin.^{10–12} Whole-genome sequencing (WGS) studies have used fragmentation patterns to predict cancer origin¹³ and integrated genomic and chromatin-associated features for cancer detection and localization.^{14–16} Targeted promoter sequencing has also enabled tumor subtype classification by inferring gene

expression from cfDNA fragmentation.¹⁷ These studies demonstrate the diagnostic potential of cfDNA fragmentation, but rely on genome-wide sequencing or panels specifically designed to capture regulatory regions.

Targeted exon-panel sequencing of circulating tumor DNA (ctDNA) is increasingly used in routine care for advanced cancer to identify actionable alterations and guide treatment selection.

^{18–20} These data can also yield fragmentation features for cancer-type prediction. Helzer et al. first demonstrated this approach²¹ and subsequently evaluated individual fragmentomic features and their combination in a cohort of 431 samples spanning five primary cancer sites and healthy controls.²² Their analyses included normalized all-exon depth and fragmentation measures related to transcription factor binding sites (TFBSs), with all-exon depth performing best among individual measures. However, their combined models did not incorporate somatic mutations or separately derived copy-number features. Therefore, it remains unclear whether integrating genomic and fragmentomic signals from routine ctDNA sequencing improves tissue-of-origin prediction across diverse metastatic cancers. In addition, more detailed, biologically informed characterization of TFBS-associated fragmentation may further improve tissue-of-origin prediction.

Here, we investigated whether refining TFBS-associated fragmentation measures and integrating genomic and fragmentomic features could improve tissue-of-origin prediction from routine ctDNA sequencing data. We analyzed 1,889 plasma samples from patients with metastatic cancer spanning 11 known primary types, profiled using two versions of a clinical 523-gene assay. We developed Site-directed High-resolution Analysis of Positional Endpoints and Entropy (SHAPE), which uses the assay's high local sequencing depth to quantify fragment-end positioning and fragment-length diversity at TFBSs. We evaluated classifiers based on SHAPE, somatic mutations, broad copy-number alterations, all-exon depth, and exon 1 fragment-length entropy, individually and as an ensemble.

Results

Study cohort and multimodal feature extraction

We analyzed 1,889 plasma samples from patients with metastatic cancers of known primary sites using two versions of the error-corrected TruSight Oncology 500 ctDNA assay. Version 1 (v1) included 1,093 samples from 10 cancer types, and version 2 (v2) included 796 samples from eight cancer types; seven cancer types were shared (Fig. 1; Table 1). Both versions targeted the same 523 genes and had similar median exon depths (3,359× in v1 and 3,354× in v2). The median depth across targets was strongly correlated between the versions (Spearman's $\rho = 0.909$), although their target-level coverage profiles differed (**Figure S1**). Feature pre-processing, model development, and internal validation were performed separately for each version.

Five signal types were derived from the sequencing reads: somatic mutations, broad copy-number alterations, all-exon depth, exon 1 fragment length entropy, and TFBS SHAPE. We trained a classifier for each signal type and combined the predictions into a late-fusion ensemble (Fig. 1).

Table 1
Distribution of cancer types across the two assay-version cohorts (v1 and v2).

Cancer type	v1, n	v2, n
Lung cancer	391	257
Colorectal cancer	182	326
Gastric cancer	155	35
Pancreatic cancer	61	51
Biliary tract cancer	76	30
Melanoma	45	33
Prostate cancer	58	20
Liver cancer	67	—
Breast cancer	—	44
Sarcoma	31	—
Bladder cancer	27	—
Total	1093	796

—, cancer type not represented in the cohort.

Site-level TFBS features reflect local fragmentation patterns and predict tissue of origin

Intersecting assay targets with experimentally supported, motif-resolved sites in the UniBind catalog^{23,24} identified 26,845 candidate transcription factor binding sites (TFBSs; **Figure S2**). At sites supported by at least 300 fragment ends, SHAPE quantified the spatial distribution of fragment ends (endpoint Gini) and the diversity of fragment lengths (fragment length entropy) within motif-centered ± 40 -bp windows (Fig. 2A).

Fragments measuring 35–80 bp accounted for a median of only 0.28% of the sequenced fragments in v1 and 0.33% in v2 (**Figure S3A**). Therefore, we calculated the endpoint Gini coefficient using fragment ends across the available fragment length distribution. Sites in the highest endpoint-Gini quartile showed enrichment of ends at motif flanks and depletion near motif centers, whereas sites in the lowest quartile

showed more uniform endpoint distributions (Fig. 2B; **Figures S3B–D**). In both assay versions, classifiers using the observed endpoint Gini outperformed those using the sequence-predicted Gini from bias-matched fragment simulations (**Figure S3E**).

Sites with overlapping consensus hematopoietic DNase footprints had higher endpoint Gini and fragment-length entropy than sites without footprint evidence in both assay versions (Fig. 2C). Both features were positively associated with pan-hematopoietic chromatin accessibility, with a stronger association with fragment length entropy (Fig. 2D). Within samples, the correlation was weak (median Spearman's $\rho = -0.076$ in v1 and -0.036 in v2).

Endpoint Gini varied among sites assigned to the same transcription factor (**Figure S4B**). In the cross-validated tissue-of-origin prediction, site-level SHAPE outperformed features aggregated at the transcription factor level (Fig. 2E). Cancer type-specific feature rankings also included sites associated with lineage-relevant transcription factors (**Figure S4C**).

Integration of fragmentomic and genomic signals improves prediction

Combining the five signal types by late fusion improved tissue-of-origin prediction over any individual signal. All-exon depth and TFBS SHAPE were the strongest individual classifiers in both assay versions, with macro-averaged one-versus-rest AUROCs of 0.815 and 0.816 in v1 and 0.813 and 0.818 in v2, respectively. The ensemble achieved macro-AUROC of 0.870 and 0.872, exceeding the best individual classifier by 0.054 in each version (Fig. 3A; **Table S2**). It also achieved the highest or near-highest cancer-type-specific AUROCs (Fig. 3B) and the highest overall top-1 accuracy; ensemble confusion matrices are shown in **Figure S5**.

We next assessed each signal's average contribution across signal subsets and its conditional contribution beyond the remaining signals. Shapley decomposition assigned comparable contributions to all-exon depth, TFBS SHAPE, and somatic mutations, each accounting for 22–24% of the total AUROC gain above chance (**Table S3**). In leave-one-signal-out analyses, removing somatic mutations caused the largest decrease in ensemble macro-AUROC, by 0.033 in v1 and 0.028 in v2. Smaller decreases followed removal of all-exon depth, TFBS SHAPE, or broad CNA features. Conditional contributions exceeded permutation baselines for all signals except exon 1 entropy (Fig. 3C; **Table S3**). These differences are consistent with partial overlap in predictive information among the signals, with somatic mutations providing the largest incremental benefit when all other signals were included.

Conditional contributions also varied across cancer types and assay versions (Fig. 3D). Somatic mutations made the largest contributions in pancreatic, colorectal, and lung cancers in both versions, as well as in bladder cancer in v1. All-exon depth contributed most in prostate and breast cancers in v2, whereas TFBS SHAPE contributed most in biliary tract cancer in v1 and gastric cancer in v2. Exon 1

entropy and broad CNA features generally showed smaller conditional contributions. No single signal had the largest contribution across all cancer types.

To assess the incremental value of TFBS SHAPE beyond all-exon depth, we compared the depth-only classifier with an ensemble combining the two signals. Adding SHAPE increased macro-AUROC by 0.014 in v1 (95% CI, 0.010–0.020) and 0.029 in v2 (95% CI, 0.024–0.034; both $P < 0.001$; **Table S3**). The two signals also showed distinct patterns of cancer-type–specific contributions (**Figure S6**). These findings support additional tissue-of-origin information in TFBS SHAPE beyond that captured by all-exon depth.

We also examined whether informative features were consistent with established tissue- and cancer-associated biology. Exon 1 entropy was associated with cancer type–matched gene expression (**Figure S7**), all-exon depth features corresponded to characteristic focal copy-number alterations (**Figure S8**), broad CNA patterns included established cancer-associated chromosome-arm alterations (**Figure S9**), and informative mutation features corresponded to tissue-associated driver alterations (**Figure S10**).

Performance varies by estimated tumor fraction

We stratified samples by tumor fraction estimated from off-target reads (< 3%, 3–10%, 10–20%, and > 20%). Ensemble macro-AUROC increased from 0.812 in the < 3% stratum to 0.959 in the 10–20% stratum in v1, and from 0.811 to 0.945 in v2, but changed little in the > 20% stratum (0.949 and 0.941, respectively; Fig. 4A; **Table S4**). TFBS SHAPE and all-exon depth were the best-performing individual signals in the higher tumor-fraction strata (Fig. 4A).

In the Shapley analysis, TFBS SHAPE and all-exon depth contributed more in the > 20% than in the < 3% stratum, whereas somatic mutations contributed less in both assay versions (Fig. 4B). Top-1 accuracy was approximately 80% at estimated tumor fractions $\geq 10\%$ (Fig. 4C). Confusion matrices for samples with tumor fractions > 10% are shown in **Figure S11**.

Performance in an independent cohort and transfer across assay versions

We applied the analytical framework to an independent targeted cfDNA cohort generated on a different platform: 54 prostate cancer, 49 lung cancer, and 48 breast cancer samples together with 47 healthy controls. We derived all five signal types anew, trained, and evaluated the classifiers in this cohort. Across the four categories, the ensemble achieved a macro-AUROC of 0.944, compared to 0.942 for all-exon depth and 0.925 for TFBS SHAPE (Fig. 5A–D). The class-specific results and performance based on the estimated tumor fraction are shown in **Figure S12**.

Finally, we evaluated the transfer between the clinical assay versions using seven shared cancer types. The models trained in one version were applied to the other without refitting. Each individual signal retained a macro-AUROC above 0.5, in both directions, but gene-level somatic mutation and broad copy-

number features were transferred more consistently than local depth and fragmentation features (Fig. 5E–G).

Discussion

Here, we show that routine targeted ctDNA sequencing retains tissue-of-origin information in both conventional genomic alterations and chromatin-associated fragmentation patterns. We demonstrate that the high local depth of targeted sequencing can resolve fragmentation at individual, motif-defined TFBSs. TFBS SHAPE and all-exon depth were among the strongest individual signals, while integrating them with other genomic and fragmentomic features yielded the highest overall classification performance. Signal contributions varied across cancer types, suggesting that tissue identity is encoded by different combinations of genomic and fragmentation signals across tumor lineages.

Several observations in this study suggest that SHAPE reflects aspects of local chromatin organization. Motif-centered endpoint depletion with flanking enrichment was consistent with focal DNA protection, and both SHAPE measures were associated with external hematopoietic DNase-footprint and chromatin accessibility data. Their weak correlation suggests that endpoint positioning and fragment length diversity capture different aspects of local fragmentation. The observed endpoint Gini also outperformed its sequence-predicted counterpart. Together, these findings support a chromatin-associated signal that was not fully explained by sequence-dependent cleavage bias represented in our simulations.

Previous cfDNA studies have taken different approaches to TFBS analysis: some aggregated fragmentation signals across sites associated with each transcription factor,^{22,25} while others examined individual sites using whole-genome sequencing or panels designed to capture regulatory regions.^{26,27} Our approach used the high local depth of a routine targeted panel to quantify fragment end positioning and fragment length diversity without dedicated regulatory capture. Site-level SHAPE outperformed TF-level aggregation, suggesting that variation among sites associated with the same transcription factor contributes to tissue-of-origin prediction. This finding may inform the design of future clinical ctDNA panels to capture tissue-informative regulatory loci, as well as clinically relevant genomic alterations.

The gains from multimodal integration highlight the importance of both individual predictive strength and complementarity among signals. All-exon depth and TFBS SHAPE were the strongest individual predictors and, together with somatic mutations, made the largest average contributions across signal combinations. Somatic mutations nevertheless provided the greatest additional value when all other signals were included. This pattern suggests that some of the predictive information captured by depth and SHAPE is shared with other signals, whereas somatic mutations contribute information that is less readily captured by the remaining signals. Nonetheless, the conditional contribution of TFBS SHAPE exceeded its permutation-null baseline, and adding SHAPE to all-exon depth significantly increased macro-AUROC, indicating that TFBS SHAPE carries tissue-of-origin information not captured by coverage alone. Relative contributions also varied across cancer types and assay versions. Biological analyses linked informative mutation features to tissue-associated drivers,²⁸ depth and copy-number features to

characteristic genomic alterations, exon 1 entropy to cancer type–matched gene expression, and SHAPE sites to lineage-relevant transcription factors.^{29,30} These associations provide biological context for the cancer-specific contributions of individual signal types and help interpret their roles in the ensemble.

Although tissue-of-origin prediction generally improved with increasing estimated tumor fraction, higher fractions did not always translate into better performance. Top-1 accuracy was approximately 80% at estimated tumor fractions of 10–20%. Further gains in ensemble macro-AUROC were limited in the > 20% stratum in both assay versions. Biological heterogeneity, potentially including lineage plasticity in advanced cancers,^{31,32} may contribute to the residual classification errors, although its role was not directly assessed.

Independent cohort and cross-assay analyses further confirmed the applicability of this approach. Models trained within an independent targeted cfDNA cohort again recovered cancer type information. This supports the reproducibility of the analytical framework on another platform. Direct transfer between the two clinical assay versions showed that gene-level mutations and broad copy number features were more stable than local coverage and fragmentation features. These differences underscore the need for assay-specific calibration or harmonization when applying local fragmentomic features to other workflows.^{33,34}

The primary clinical cohorts in this study were retrospective, drawn from a single center, and limited to patients with known primary cancers. Prospective, multicenter studies involving patients with uncertain primary sites are needed to establish the clinical utility of our approach.

This study extends the diagnostic information obtained from routine ctDNA sequencing by resolving the fragmentation of individual TFBSs and integrating it with genomic and exon-based features. Relating the contributions of these signals to cancer types and established biology provides a more interpretable basis for tissue-of-origin prediction. Such tissue-of-origin predictions could complement imaging and pathology and help prioritize further diagnostic evaluations, extending the clinical value of existing ctDNA testing without additional specimen collection or sequencing.

Methods

Study cohorts

We retrospectively included 1,889 patients with metastatic cancer who underwent clinical targeted ctDNA sequencing using the TruSight Oncology 500 assay (Illumina) at Samsung Medical Center (Seoul, Republic of Korea) between 2023 and April 2026. Eligible patients had a documented primary cancer type and a median exon depth of $\geq 1,000\times$; one sample per patient was analyzed (the earliest when several were available). Cancer types with fewer than 20 patients in the assay version were excluded. The final cohort comprised 1,093 patients tested with assay version 1 (v1; 10 cancer types) and 796 patients tested with version 2 (v2; eight cancer types), of which seven types were shared (biliary tract,

colorectal, gastric, lung, melanoma, pancreatic, and prostate cancer). The median on-target depths were 3,359× (v1) and 3,354× (v2). As the versions differed in coverage characteristics and cohort composition, preprocessing, model development, and internal validation were performed separately for each version. The study was approved by the Institutional Review Board of the Samsung Medical Center (2026-04-209) in accordance with the Declaration of Helsinki, and the requirement for informed consent was waived.

The independent external cohort (GRAIL; European Genome-phenome Archive, EGAD00001005302) comprised 198 individuals with prostate cancer (n = 54), lung cancer (n = 49), breast cancer (n = 48), or no cancer (n = 47), each with matched plasma and leukocyte-targeted sequencing on a distinct deep targeted cfDNA platform (median on-target depths 5,236×).

Targeted ctDNA sequencing and fragment processing

cfDNA was analyzed using the TruSight Oncology 500 ctDNA assay, which targets coding and selected noncoding regions of 523 cancer-associated genes,³⁵ using the laboratory's validated clinical workflow with unique molecular identifier processing, alignment to the GRCh37 reference, and clinical variant calling (DRAGEN v3.8). For fragmentomic analyses, secondary, supplementary, and duplicate-marked alignments were excluded, properly paired reads with mapping quality ≥ 20 were retained, and fragment lengths and both endpoint coordinates were derived. Fragments longer than 1,000 bp were excluded from the analysis.

TFBS annotation and site selection

Experimentally supported direct transcription factor-DNA binding sites were obtained from UniBind (release 20220914)^{23,24} and converted into hg19 coordinates. Records with motif midpoints overlapping with a TruSight Oncology 500 capture interval were retained (113,571 records), and records sharing the same genomic midpoint were collapsed into a single locus, yielding 26,845 candidate TFBS loci, while retaining all associated transcription factor annotations. A ± 40 -bp window centered on each locus midpoint was used for feature extraction; TFBS SHAPE features were restricted to autosomal sites.

TFBS SHAPE: endpoint Gini and fragment-length entropy

TFBS SHAPE combines two complementary site-level features. Endpoint Gini quantified the spatial inequality of fragment-end positions. The ± 40 -bp window was divided into 16 5-bp bins, both ends of each qualifying fragment were counted, and inequality across bins was summarized as a Gini coefficient (0 = uniform, 1 = maximally concentrated). Fragment-length entropy quantified local fragment-length diversity: for fragments 20–500 bp overlapping the window by ≥ 1 bp, the Shannon entropy of the 1-bp-resolution length distribution was computed. To remove the dependence on local fragment support, each feature was residualized against the natural logarithm of the supporting count and converted within each sample to a robust z-score using the median and median absolute deviation across loci. Site-level values were retained only for loci supported by ≥ 300 fragment endpoints. Unsupported site-sample values were treated as missing and imputed using training-fold medians.

Short-fragment protection and sequence-bias controls

Short-fragment windowed protection scores were computed as described by Snyder and colleagues¹⁰ using fragments 35–80 bp and a 16-bp window. To assess sequence-dependent cleavage bias, we adapted their fragment-end bias model. Sequence-predicted endpoint Gini was computed from reference genome contexts using the same window and bins as the observed value, and observed, sequence-predicted, and bias-adjusted features were evaluated under identical cross-validation partitions.

Association of TFBS fragmentation with hematopoietic chromatin state

All samples were analyzed using UniBind TFBSs overlapping the targets, retaining sample-site pairs with > 300 informative endpoints. Consensus transcription factor footprints^{36–38} were obtained from 118 ENCODE DNase-seq experiments in primary hematopoietic cells (Table S1). For each sample, the difference in median endpoint Gini (and, separately, fragment-length entropy) between footprint-overlapping and non-overlapping sites was compared with a footprint-label-permuted null. Chromatin accessibility was obtained from a pan-hematopoietic assay for transposase-accessible chromatin using sequencing (ATAC-seq) reference³⁹, and the Spearman correlation between accessibility and each feature across eligible TFBSs was compared with an accessibility-permuted null. The observed and null distributions were compared using two-sided Wilcoxon rank-sum tests.

Conventional signal types

Exon 1 fragment-length entropy was computed as described by Helzer and colleagues²¹ from the fragment length distribution at one captured coding exon per gene. All-exon depth—a target-level measure of coverage that tracks focal copy-number dosage—was adapted from Helzer and colleagues²²: properly paired, primary, non-duplicate fragments were assigned to captured exons by midpoint, and per-exon counts were length-normalized, library-normalized, and scaled. Broad copy number alterations were derived from off-target read depth using assay-adapted ichorCNA v0.5.1 implementation⁴⁰, averaging GC- and mappability-corrected $\log_2$ ratios across bins within each chromosome arm. Somatic mutation features were binary gene-level nonsynonymous and truncating alterations and curated recurrent protein hotspots derived from clinically reportable variants classified as somatic. For all fragmentomic and copy-number signals, X and Y chromosomes were excluded to reduce sex-associated confounding factors.

Tumor-fraction estimation

The tumor fraction was estimated from off-target reads⁴¹ using the assay version-specific ichorCNA v0.5.1, and samples were assigned to four pre-specified strata (< 3%, 3% to < 10%, 10% to < 20%, and > 20%).

Model development and internal validation

For each signal type, we trained a multiclass tissue-of-origin classifier defined by the documented primary cancer type separately for each assay version. TFBS SHAPE, exon 1 entropy, and all-exon depth

were modeled using elastic-net multinomial logistic regression (a penalized linear classifier); broad copy-number alterations and somatic mutations were modeled using gradient-boosted decision trees (XGBoost). Performance was estimated by repeated nested cross-validation (repeated internal validation with strict separation of training and test data): five outer folds were repeated 20 times (100 held-out folds; identical partitions across signal types; seed 42), with hyperparameters tuned by three-fold inner cross-validation optimizing the macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUROC). All data-dependent preprocessing was confined to the training folds and was applied unchanged to the held-out data.

Multimodal integration

Signal-type class probabilities were combined into a single classifier using late fusion, a class-specific non-negative least-squares combiner fitted to cross-fitted training predictions, with per-class shrinkage toward a pooled solution to stabilize infrequent classes. Base models were refitted on the full outer-training set and applied to the held-out fold; combiner coefficients were applied unchanged, and class scores renormalized to sum to one.

Signal contribution analyses

We used exact Shapley decomposition⁴² to quantify each signal's average contribution to macro-AUROC across different signal combinations, based on all 32 possible signal subsets. Leave-one-signal-out analysis assessed each signal's additional value beyond the other four signals, measured as the decrease in macro-AUROC after its removal from the full ensemble. Cancer-specific contributions were calculated using the corresponding one-versus-rest AUROCs.

To assess whether TFBS SHAPE improved prediction beyond all-exon depth, we compared the depth-only classifier with an ensemble combining the two signals. For all analyses, the late-fusion combiner was refitted for each signal subset using cross-fitted training predictions, and performance was compared on identical held-out validation folds.

Cross-assay transferability

Cross-assay transferability was evaluated after restricting both cohorts to the seven shared cancer types and each signal to features common to both versions, training a class-balanced, L2-penalized multinomial logistic regression learner in one version and applying it without refitting the other. Directional retention was defined as the transfer AUROC divided by the corresponding within-assay AUROC, and cross-assay stability was defined as the mean retention across directions.

Independent replication

In the GRAIL cohort, reads that were deduplicated using duplex unique-molecular-identifier families were used to extract fragmentation and depth features, and somatic mutations were identified using matched leukocyte data. All five signals were derived de novo from GRAIL-captured regions, and the tumor fraction was estimated from off-probe reads. Models were trained and evaluated entirely within the

GRAIL cohort using the same nested cross-validation design, with all preprocessing confined to the training folds.

Biological interpretation and statistics

Exon 1 entropy was compared with cancer type-matched RNA-sequencing from The Cancer Genome Atlas (TCGA)⁴³ using Pearson correlation, and the all-exon depth, broad copy-number alterations, and somatic mutation models were interpreted using model coefficients or TreeSHAP⁴⁴ and compared with established focal and arm-level events and tissue-associated drivers. The performance was summarized as the macro-averaged one-versus-rest AUROC. Unless otherwise specified, correlations were Spearman's coefficient, tests were two-sided, and $P < 0.05$ was considered significant. Analyses were performed using Python (NumPy, scikit-learn, XGBoost, SHAP, and pysam) and R version 4.5.1.

Large language model use

A generative artificial intelligence tool (a large language model; Claude, Anthropic) was used to assist with language editing and formatting. The authors reviewed and approved all content and take full responsibility for the work.

Declarations

Competing interests

A patent application related to the SHAPE methodology described in this manuscript is pending. The authors otherwise declare no competing interests.

Author Contribution

Conceptualization, J.K. and Y.-g.K.; methodology, software, formal analysis, and visualization, J.K.; investigation, J.K., S.Y.J., J.E.S., Ju.K., J.S., Se.P., M.Kw., S.H.L., H.A.J., M.Ka., J.Y.H., S.T.K., J.-M.S., J.L., S.-W.U., S.H.P., Y.H.P., S.I.S., S.-H.L., S.S.J., J.O.P., J.S.A., Y.S.P., and Y.-g.K.; data curation, J.K. and Y.-K.K.; writing - original draft, J.K.; writing - review and editing, J.K. and Y.-g.K.; project administration and supervision, Y.-g.K. All authors have read and approved the final manuscript.

Acknowledgements

This study was supported by the Future Medicine 2030 Project of Samsung Medical Center (SM0126003). The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data availability

De-identified patient data and raw ctDNA sequencing data from the discovery cohorts are not publicly available because institutional protocols and patient-privacy requirements preclude unrestricted release; requests will be reviewed by Samsung Medical Center under a data-use agreement for academic, non-commercial research and should be directed to the corresponding author. Sequencing data from the independent GRAIL cohort are available from the European Genome-phenome Archive (accession EGAD00001005302).

Code availability

The custom code used in this study is not publicly available because of a pending patent application related to the SHAPE methodology, but is available from the corresponding author on reasonable request for academic, non-commercial research.

References

1. Rassy, E. & Pavlidis, N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat. Rev. Clin. Oncol.* 17, 541–554 (2020).
2. Kramer, A. et al. Cancer of unknown primary: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* 34, 228–246 (2023).
3. Hainsworth, J. D. et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon Research Institute. *J. Clin. Oncol.* 31, 217–223 (2013).
4. Moran, S. et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol.* 17, 1386–1395 (2016).
5. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* 31, 745–759 (2020).
6. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 563, 579–583 (2018).
7. Penson, A. et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol.* 6, 84–91 (2020).
8. Moon, I. et al. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nat. Med.* 29, 2057–2067 (2023).
9. Pereira, B. et al. Cell-free DNA captures tumor heterogeneity and driver alterations in rapid autopsies with pre-treated metastatic cancer. *Nat. Commun.* 12, 3199 (2021).
10. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164, 57–68 (2016).
11. Noe, M. et al. DNA methylation and gene expression as determinants of genome-wide cell-free DNA fragmentation. *Nat. Commun.* 15, 6690 (2024).

12. Kim, J. et al. Multidimensional fragmentomic profiling of cell-free DNA released from patient-derived organoids. *Hum. Genomics* 17, 96 (2023).
13. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 570, 385–389 (2019).
14. Bao, H. et al. Early detection of multiple cancer types using multidimensional cell-free DNA fragmentomics. *Nat. Med.* 31, 2737–2745 (2025).
15. Bae, M. et al. Integrative modeling of tumor genomes and epigenomes for enhanced cancer diagnosis by cell-free DNA. *Nat. Commun.* 14, 2017 (2023).
16. Peneder, P. et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat. Commun.* 12, 3230 (2021).
17. Esfahani, M. S. et al. Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* 40, 585–597 (2022).
18. Nakamura, Y. et al. Targeted therapy guided by circulating tumor DNA analysis in advanced gastrointestinal tumors. *Nat. Med.* 31, 165–175 (2025).
19. Assaf, Z. J. F. et al. A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. *Nat. Med.* 29, 859–868 (2023).
20. Rose Brannon, A. et al. Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. *Nat. Commun.* 12, 3770 (2021).
21. Helzer, K. T. et al. Fragmentomic analysis of circulating tumor DNA-targeted cancer panels. *Ann. Oncol.* 34, 813–825 (2023).
22. Helzer, K. T. et al. Analysis of cfDNA fragmentomics metrics and commercial targeted sequencing panels. *Nat. Commun.* 16, 9122 (2025).
23. Gheorghe, M. et al. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* 47, e21 (2019).
24. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics* 22, 482 (2021).
25. Ulz, P. et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.* 10, 4666 (2019).
26. Rao, S. et al. Transcription factor-nucleosome dynamics from plasma cfDNA identifies ER-driven states in breast cancer. *Sci. Adv.* 8, eabm4358 (2022).
27. Hiatt, J. B. et al. Molecular phenotyping of small cell lung cancer using targeted cfDNA profiling of transcriptional regulatory regions. *Sci. Adv.* 10, eadk2082 (2024).
28. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* 173, 321–337 (2018).
29. Joung, J. et al. A transcription factor atlas of directed differentiation. *Cell* 186, 209–229 (2023).
30. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* 370, eaba7612 (2020).

31. Cammareri, P. et al. Loss of colonic fidelity enables multilineage plasticity and metastasis. *Nature* 644, 547–556 (2025).
32. Yang, D. et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* 185, 1905–1923 (2022).
33. Passemiers, A. et al. DAGIP: alleviating cell-free DNA sequencing biases with optimal transport. *Genome Biol.* 26, 49 (2025).
34. Wang, H. et al. A standardized framework for robust fragmentomic feature extraction from cell-free DNA sequencing data. *Genome Biol.* 26, 141 (2025).
35. Harter, J. et al. Analytical performance evaluation of a 523-gene circulating tumor DNA assay for next-generation sequencing-based comprehensive tumor profiling in liquid biopsy samples. *J. Mol. Diagn.* 26, 61–72 (2024).
36. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90 (2012).
37. Bentsen, M. et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* 11, 4267 (2020).
38. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* 583, 729–736 (2020).
39. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* 48, 1193–1203 (2016).
40. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* 8, 1324 (2017).
41. Fonseca, N. M. et al. Prediction of plasma ctDNA fraction and prognostic implications of liquid biopsy in advanced prostate cancer. *Nat. Commun.* 15, 1828 (2024).
42. Rozemberczki, B. et al. The Shapley value in machine learning. In Proc. 31st International Joint Conference on Artificial Intelligence 5572–5579 (2022).
43. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120 (2013).
44. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 56–67 (2020).

Figures

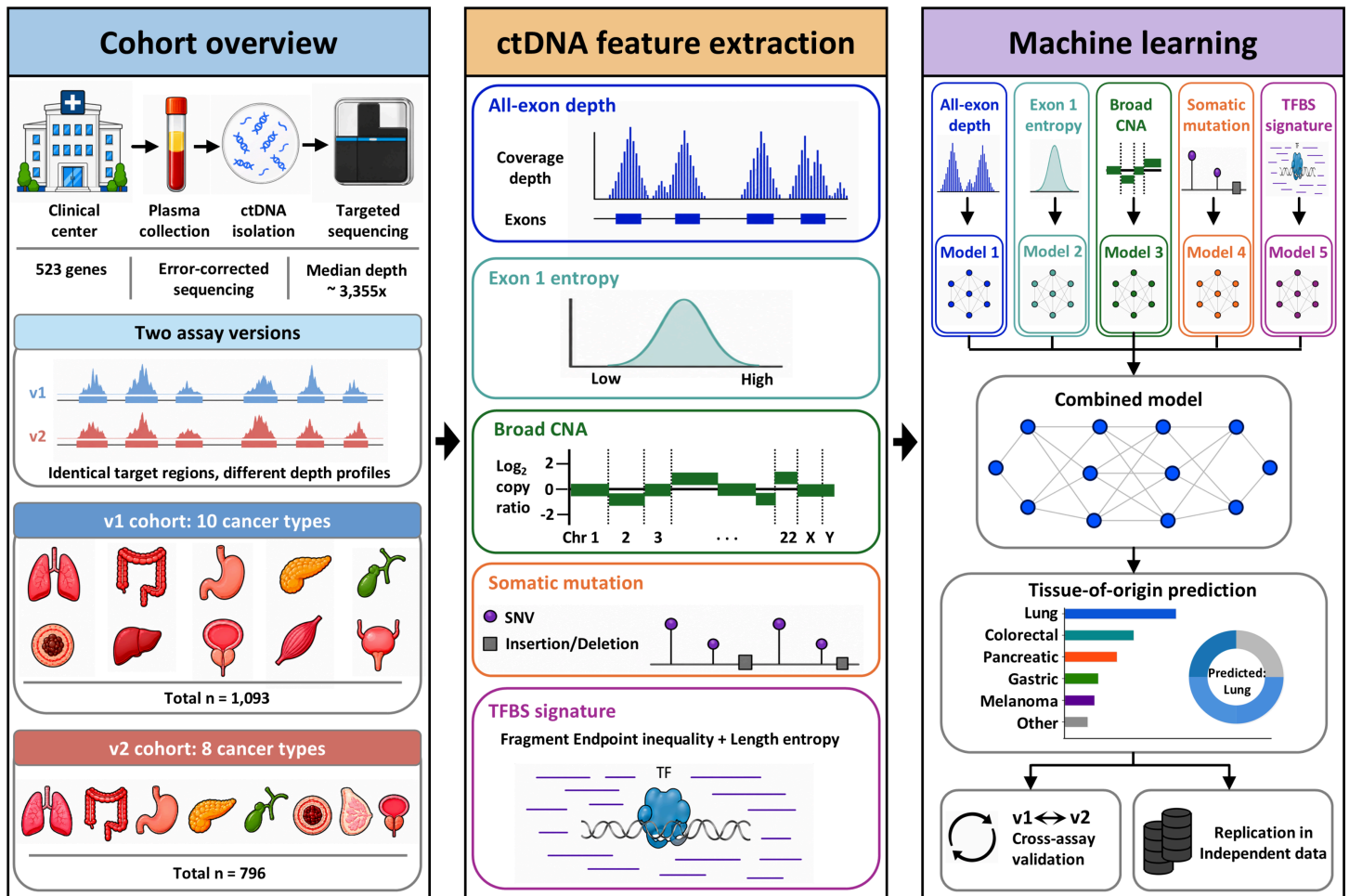


Figure 1

Study design and multimodal feature extraction from routine targeted ctDNA sequencing. Plasma collected during routine clinical care underwent error-corrected targeted sequencing of 523 cancer-associated genes on two versions of the TruSight Oncology 500 circulating tumor DNA (ctDNA) assay (v1 and v2). The cohort comprised 1,889 patients with metastatic cancer—1,093 profiled with v1 (10 cancer types) and 796 with v2 (8 cancer types), with 7 types shared. From the same sequencing data, each sample yielded five signal types: somatic mutations, broad copy-number alterations, all-exon depth, exon 1 fragment-length entropy, and transcription factor binding site (TFBS) SHAPE. A separate classifier was trained for each signal type, and the predictions were combined into a single tissue-of-origin classifier by late fusion. Preprocessing, model development, and internal validation were performed separately for each assay version; reproducibility was then evaluated in an independent cohort (GRAIL).

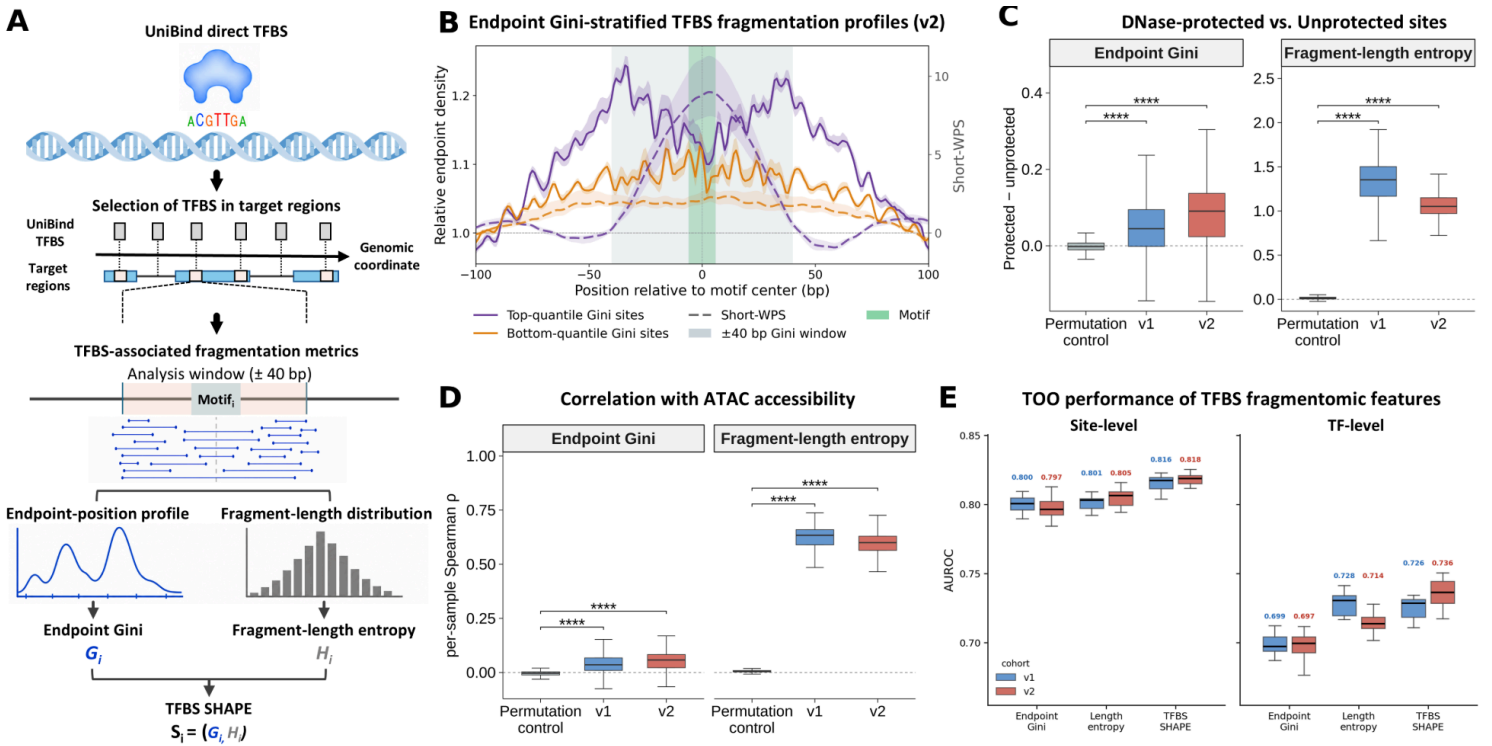


Figure 2

Site-resolved TFBS fragmentation captures chromatin-associated, tissue-informative signals. (A) SHAPE feature extraction. Within a motif-centered ± 40 -bp window at each TFBS, two complementary features are computed: endpoint Gini, the spatial inequality of fragment-end positions across sixteen 5-bp bins, and fragment-length entropy, the Shannon diversity of fragment lengths. **(B)** Aggregated fragment-endpoint density (top) and short-fragment windowed protection score (bottom) for TFBSs in the top versus bottom endpoint-Gini quartile, in the 10 samples with the highest proportion of 35-80-bp fragments. **(C)** Per-sample difference in median endpoint Gini and in fragment-length entropy between TFBSs overlapping versus not overlapping a consensus hematopoietic DNase footprint, each compared with a footprint-label-permuted null. **(D)** Per-sample Spearman correlation of each feature with pan-hematopoietic ATAC-seq accessibility, compared with an accessibility-permuted null. **(E)** Macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUROC) for tissue-of-origin classification using site-level endpoint Gini, fragment-length entropy, their combination (TFBS SHAPE), and transcription factor-level aggregation. Box plots show the median, interquartile range (IQR), and whiskers to $1.5 \times \text{IQR}$. **** $P < 0.0001$, two-sided Wilcoxon rank-sum test.

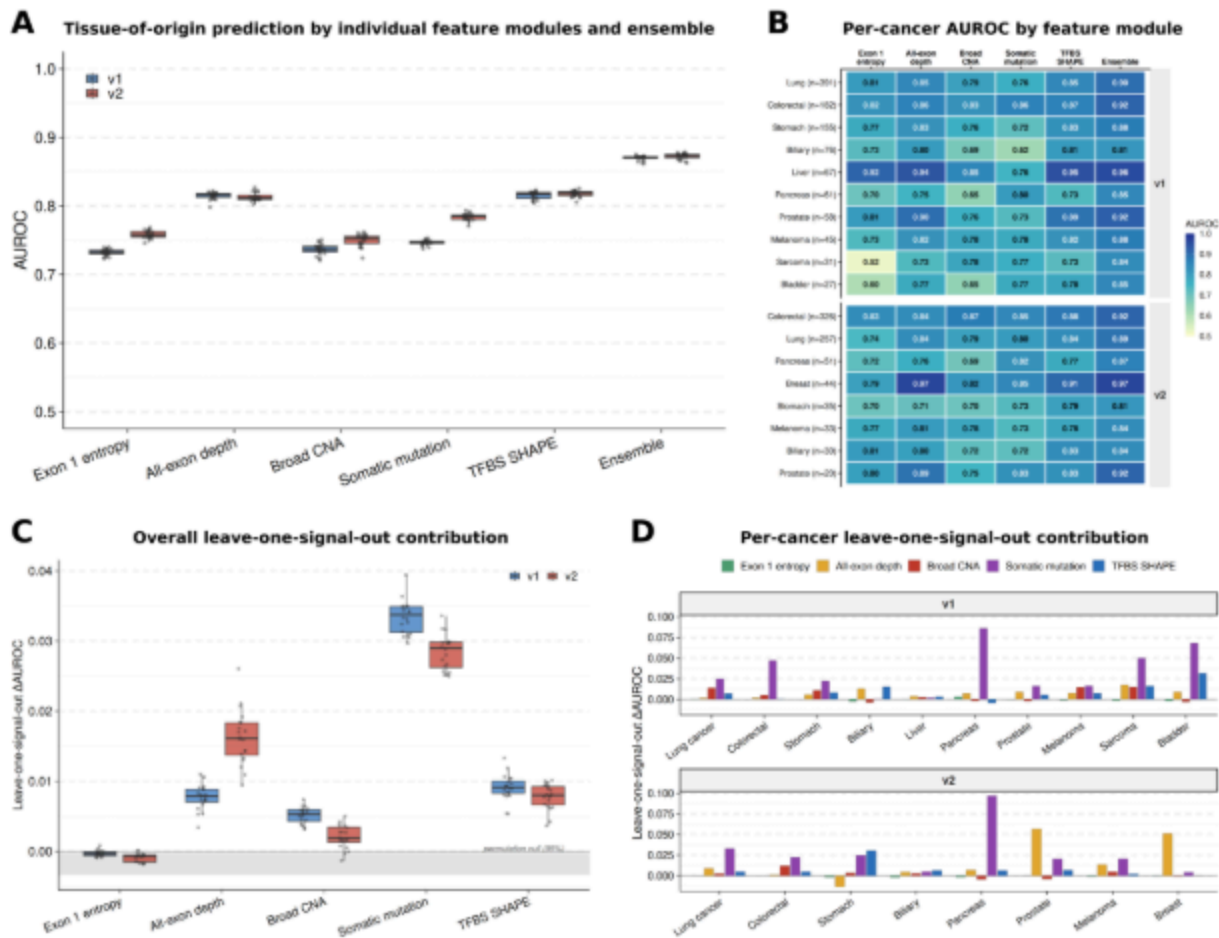


Figure 3

Complementary, non-redundant signals improve tissue-of-origin prediction. Performance and signal contributions in both cohorts (v1, 1,093 samples/10 cancer types; v2, 796/8). (A) Macro-averaged one-versus-rest AUROC of the five individual signal types and the late-fusion combined classifier, for each assay version. (B) Cancer type-specific one-versus-rest AUROC of each signal and the combined classifier. (C) Contribution of each signal beyond all the others, shown for each signal rather than ranked, with a shaded permutation-null band indicating the range expected by chance; bars above the band exceed chance. (D) Cancer type-specific leave-one-signal-out contribution of each signal, defined as the change in one-versus-rest AUROC when that signal was removed from the full ensemble, for each cancer type in v1 and v2. Box plots in (A) and (C) summarize the 100 held-out folds of repeated nested cross-validation (5 outer folds $\times$ 20 repeats; identical partitions across signal types). Contributions were quantified by exact Shapley decomposition and by leave-one-signal-out analysis; robust statistics are reported in **Table S3**.

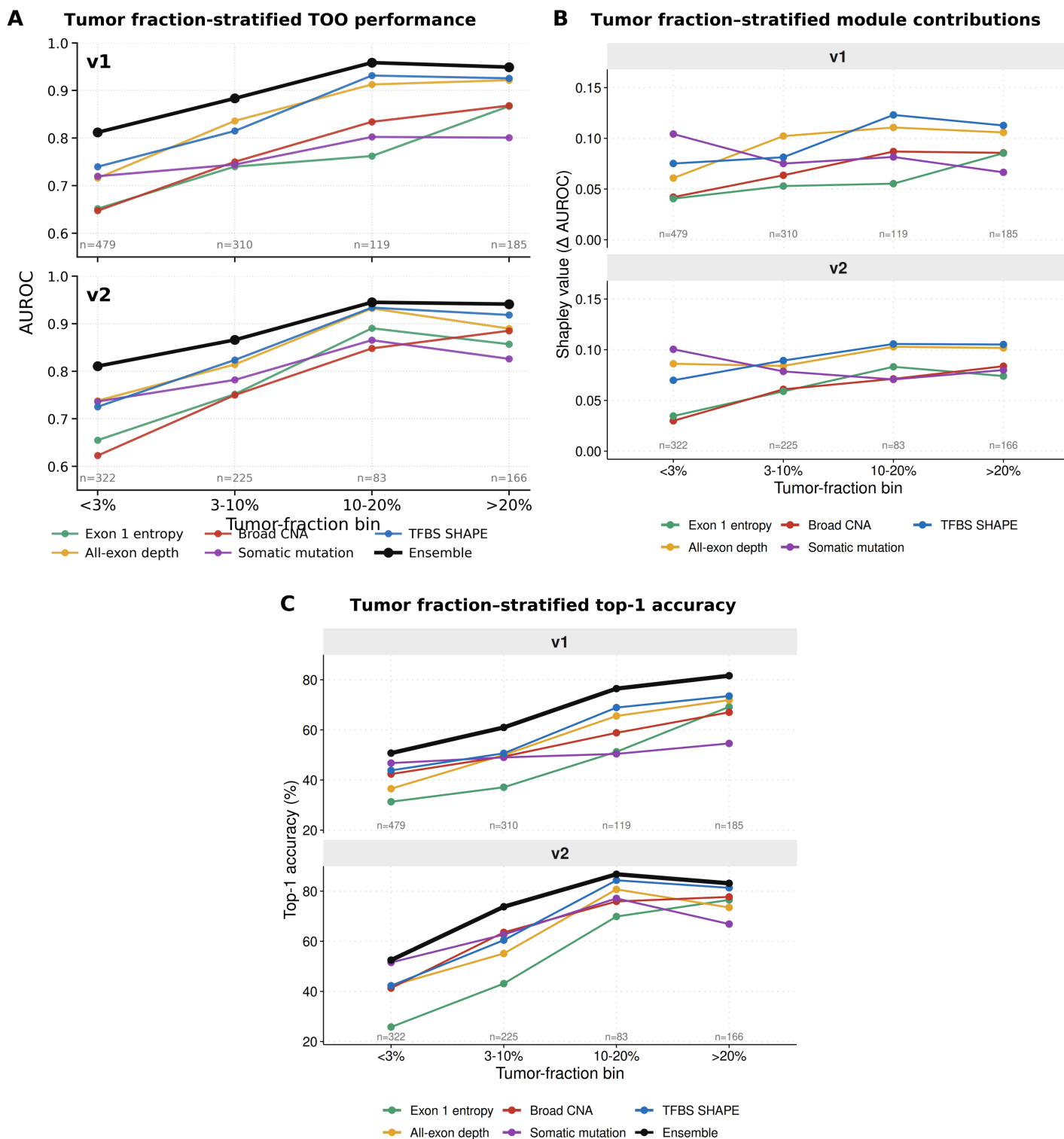


Figure 4

Tissue-of-origin performance and signal contributions across tumor-fraction strata. Held-out predictions from the primary nested cross-validation were stratified after prediction by ichorCNA-estimated tumor fraction into four prespecified groups (<3%, 3–10%, 10–20%, and >20%); models were not retrained within strata. **(A)** Macro-averaged one-versus-rest AUROC of the five individual signal types and the combined classifier across the four strata, for v1 and v2; the clinical operating point at tumor

fraction $\geq 10\%$ is highlighted. **(B)** Signal contributions across the same strata by exact Shapley decomposition. **(C)** Top-1 classification accuracy of the five signal types and the combined classifier across the same strata. Points in **(A)**–**(C)** show per-stratum estimates connected by lines.

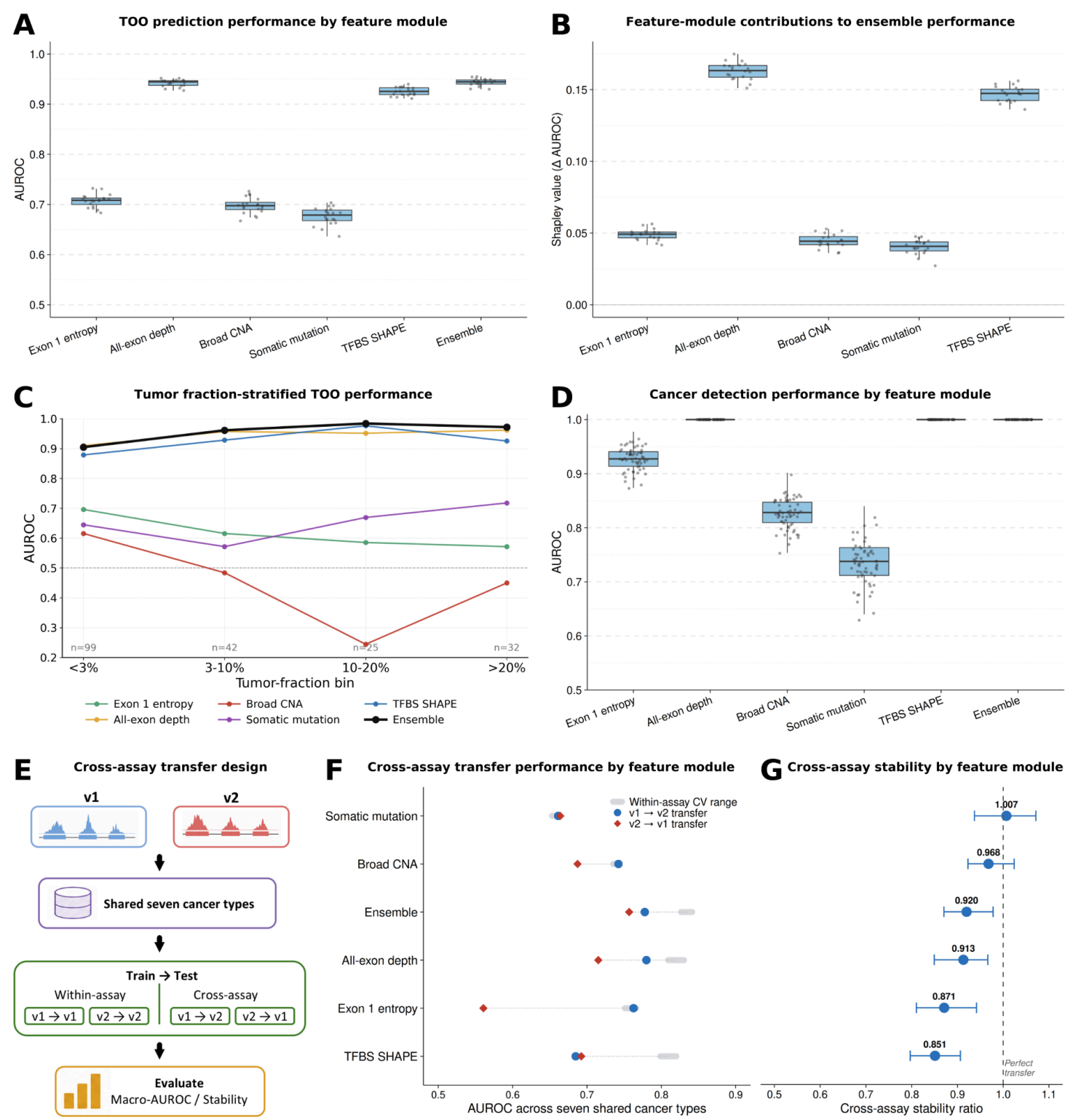


Figure 5

Independent evaluation in the GRAIL cohort and cross-assay transfer between clinical assay versions.

(A–D) Independent evaluation in the GRAIL cohort (198 individuals: prostate cancer, n=54; lung cancer, n=49; breast cancer, n=48; healthy controls, n=47); all five signal types were derived de novo and models were trained and evaluated entirely within this cohort. **(A)** Macro-averaged one-versus-rest AUROC of the five individual signal types and the combined classifier. **(B)** Contribution of each signal beyond all the others (leave-one-signal-out Δ AUROC with 95% confidence interval); only all-exon depth contributed robustly, whereas the marginal contributions of TFBS SHAPE and somatic mutations were small and not statistically robust in this cohort (Table S3). **(C)** Performance across ichorCNA-estimated tumor-fraction strata (per-stratum estimates connected by lines; dashed line, chance). **(D)** Exploratory cancer-versus-healthy discrimination; because the controls were non-cancer individuals, this case-control comparison should be interpreted with caution and separately from the multiclass results. **(E–G)** Cross-assay transfer between the two clinical assay versions (v1 and v2), restricted to the seven shared cancer types, with models trained on one version and applied to the other without refitting. **(E)** Schematic of the transfer analysis. **(F)** Cross-assay transfer performance by signal type (within-assay reference range and bidirectional transfer, $v1 \rightarrow v2$ and $v2 \rightarrow v1$). **(G)** Cross-assay stability (mean retention of AUROC on transfer) by signal type; gene-level somatic mutation and broad copy-number features were the most stable and TFBS SHAPE the least. Box plots show the median, IQR, and whiskers to $1.5 \times \text{IQR}$. GRAIL is presented as reproducibility of the analytical framework, not as independent additive proof for TFBS SHAPE.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [npjPOSupplementaryInformation.pdf](#)