

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

Jaeryuk Kim¹, You-Keun Ko¹, Seo Yoon Jang², Ji Eun Shin², Junkyu Kim², Junghoon Shin², Sehhoon Park², Minsuk Kwon², Sung Hee Lim², Hyun Ae Jung², Minyong Kang³, Jung Yong Hong², Seung Tae Kim², Jong-Mu Sun², Jeeyun Lee², Sang-Won Um⁴, Se Hoon Park², Yeon Hee Park², Seong Il Seo³, Se-Hoon Lee², Seong Soo Jeon³, Joon Oh Park², Jin Seok Ahn², Young Suk Park², and Young-gon Kim^{1*}

¹ Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

² Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

³ Department of Urology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

⁴ Division of Pulmonary and Critical Care Medicine, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

**Corresponding author: Young-gon Kim, Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Republic of Korea. E-mail: younggonn.kim@samsung.com.*

Supplementary Figures

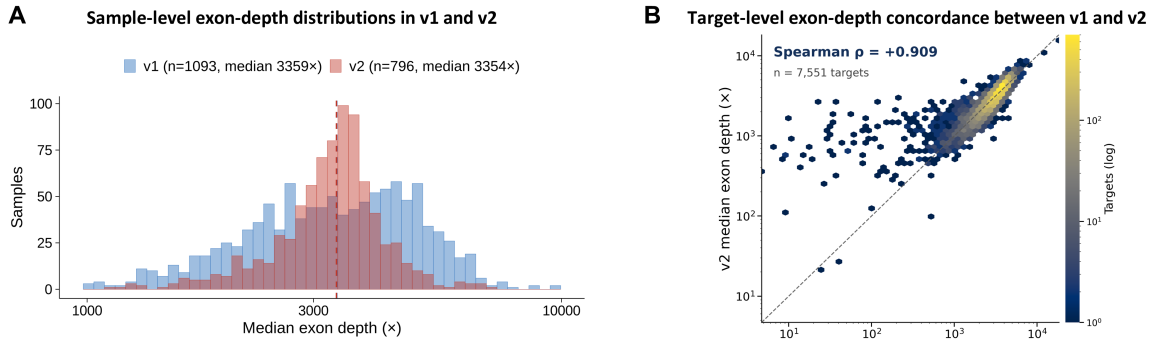


Figure S1. Sequencing-depth characteristics of the two assay versions.

(A) Distribution of sample-level median exon depth for v1 ($n = 1,093$; cohort median, $3,359\times$) and v2 ($n = 796$; cohort median, $3,354\times$). Dashed vertical lines indicate the corresponding cohort medians. The horizontal axis is logarithmically scaled.

(B) Concordance of target-level median exon depth between v1 and v2 across 7,551 target intervals interrogated by both assay versions. Each target is represented by its median depth across samples within the corresponding assay version. Density shading indicates the number of targets per bin on a logarithmic scale. Both axes are logarithmically scaled, and the dashed diagonal line denotes the line of identity. Spearman's correlation coefficient ($\rho = 0.909$) is shown.

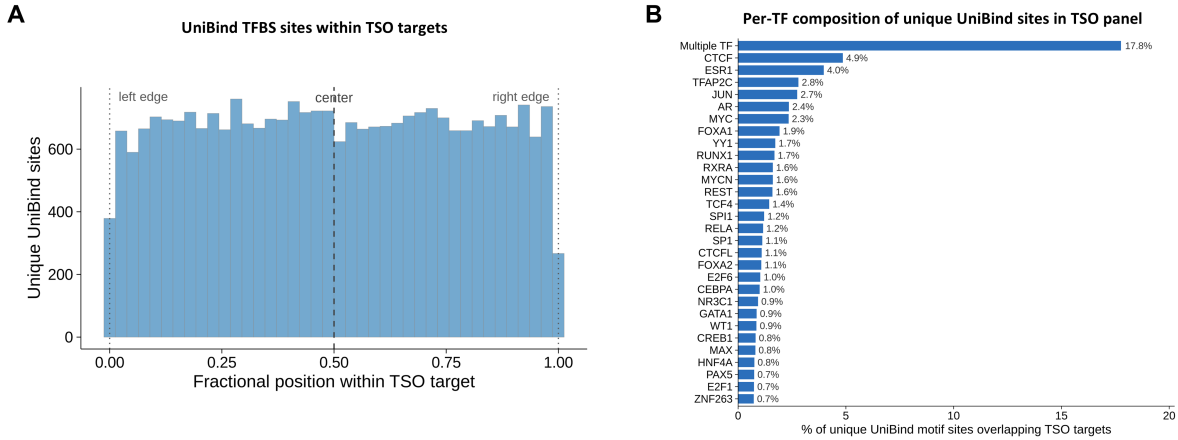


Figure S2. Genomic distribution and transcription factor composition of UniBind sites within TruSight Oncology 500 target regions.

(A) Distribution of 26,845 unique candidate TFBSs according to their normalized positions within TruSight Oncology 500 target intervals. For each site, a fractional position of 0 or 1 denotes the left or right boundary of the corresponding target interval, respectively, and 0.5 denotes its centre. Candidate sites were distributed broadly across the captured regions, with fewer sites immediately adjacent to target boundaries.

(B) TF composition of the candidate TFBSs. Bars show the percentage of unique sites associated with each of the most frequently represented TFs. Sites annotated to more than one TF were classified as ‘Multiple TF’.

TF, transcription factor; TFBS, transcription factor binding site.

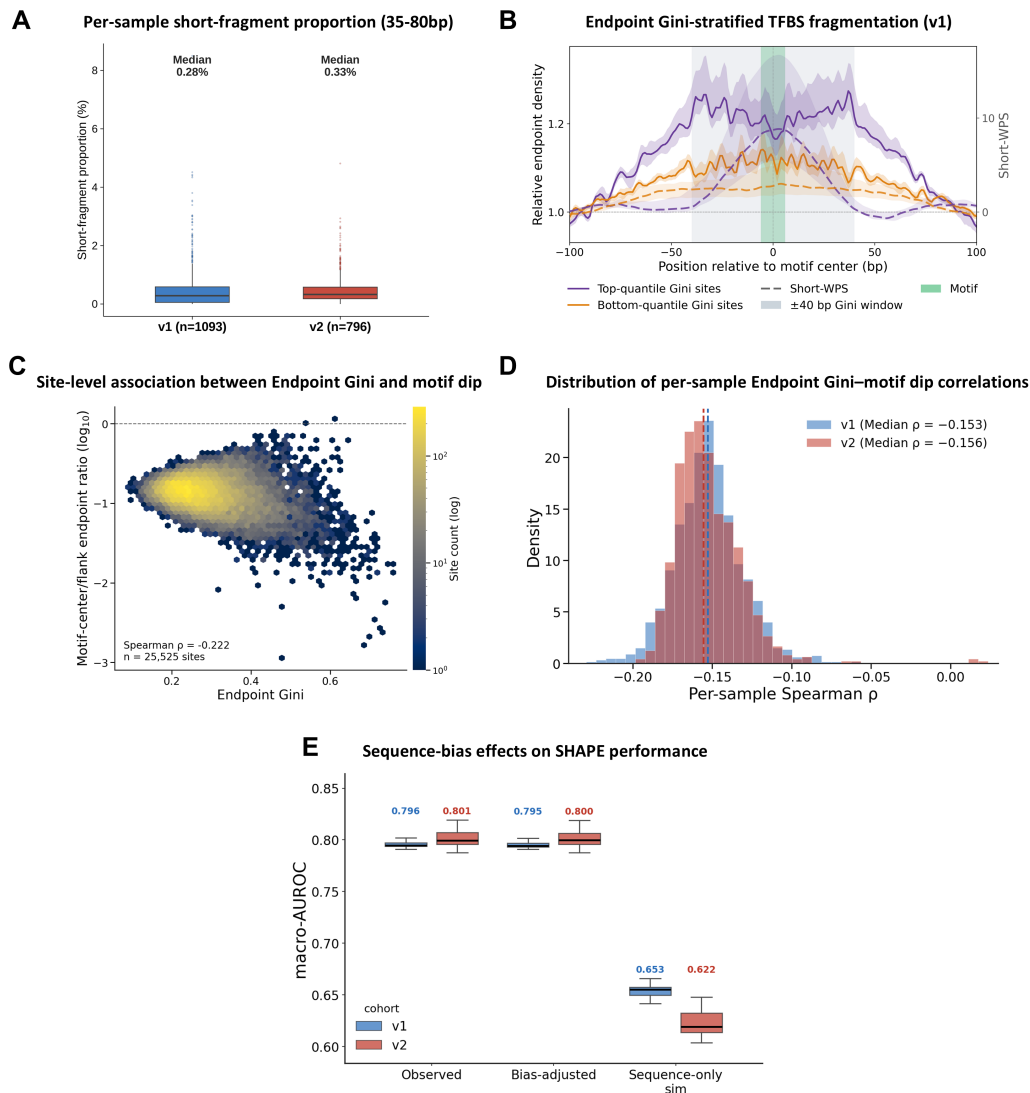


Figure S3. Short-fragment abundance and validation of endpoint Gini as a TFBS fragmentation measure.

(A) Per-sample proportions of short fragments, defined as fragments 35–80 bp in length, in v1 ($n = 1,093$) and v2 ($n = 796$). Cohort medians are shown above the distributions.

(B) Motif-centred fragmentation profiles of TFBSs in the highest and lowest quartiles of endpoint Gini in v1. Profiles were calculated across the ten v1 samples with the highest proportions of 35–80-bp fragments. Solid lines show the position-wise median relative fragment-endpoint density, and dashed lines show the corresponding median short-fragment windowed protection scores (WPSs). Shaded ribbons indicate position-wise bootstrap 95% confidence intervals for the median across the ten samples. The grey region indicates the ± 40 -bp window used for endpoint Gini calculation, and the green region indicates the motif interval.

(C) Site-level association between endpoint Gini and motif dip, defined as the ratio of motif-centred to flanking endpoint counts, in a representative sample. Point shading indicates local TFBS density on a

logarithmic scale. The horizontal dashed line denotes equal motif-centred and flanking endpoint counts ($\log_{10}\log_{10}$ ratio = 0). Spearman's correlation coefficient and the number of evaluated sites are shown.

(D) Distributions of within-sample Spearman correlations between endpoint Gini and motif dip in v1 and v2. Correlations were calculated across TFBSs within each sample, and dashed vertical lines indicate the corresponding cohort medians.

(E) Effects of local sequence-dependent cleavage bias on tissue-of-origin classification. Sequence-predicted endpoint Gini values were generated using bias-matched fragment simulations that preserved the observed fragment support and fragment-length distribution at each TFBS. Bias-adjusted features were obtained by removing the sequence-predicted component from the observed endpoint Gini.

AUROC, area under the receiver operating characteristic curve; SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TFBS, transcription factor binding site; WPS, windowed protection score.

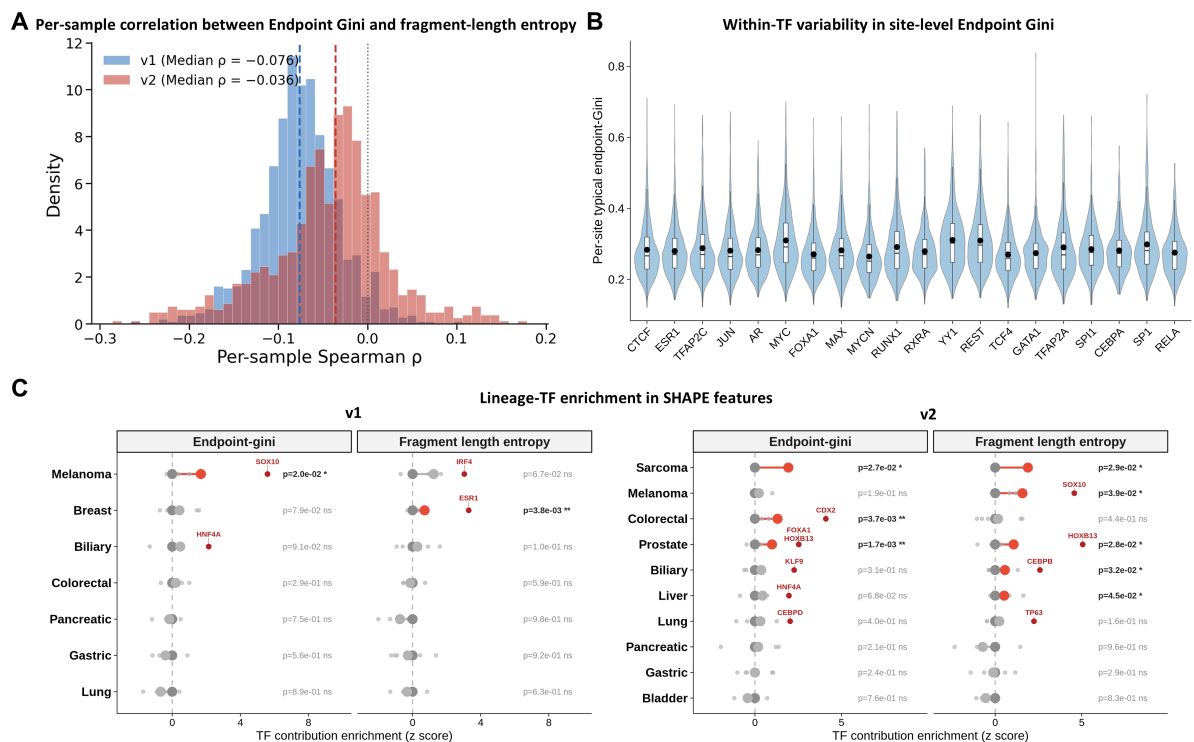


Figure S4. Within-TF heterogeneity and lineage-associated interpretation of SHAPE features.

(A) Within-sample association between endpoint Gini and fragment-length entropy. Spearman's correlations were calculated across TFBSs within each sample. Histograms show the distributions of per-sample correlation coefficients in v1 and v2; dashed vertical lines indicate the corresponding cohort medians, and the dotted line indicates $\rho = 0$ zero correlation.

(B) Within-TF variability in site-level endpoint Gini. Violin plots show the distributions of per-site robust z-normalized endpoint Gini values for highly represented TFs, illustrating substantial variation among binding sites associated with the same TF. Embedded box plots indicate the median and interquartile range.

(C) Enrichment of lineage-associated TF contributions among SHAPE features in v1 and v2. TF contribution-enrichment z scores are shown separately for endpoint Gini and fragment-length entropy across cancer types. Gray points represent individual TFs, and selected lineage-associated TFs are highlighted and labeled in red. Nominal P values for enrichment of lineage-associated TFs are shown for each cancer type and feature; ns, not significant.

SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TF, transcription factor; TFBS, transcription factor binding site.

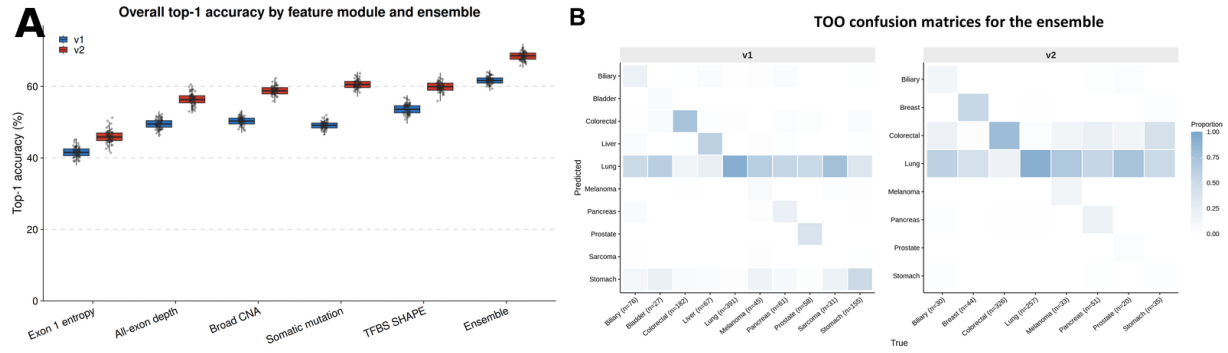


Figure S5. Top-1 tissue-of-origin classification performance of individual signal types and the ensemble.

(A) Overall top-1 accuracy of exon 1 fragment-length entropy, all-exon depth, broad copy-number alterations (CNAs), somatic mutations, TFBS SHAPE and the late-fusion ensemble in v1 and v2. Box plots show top-1 accuracy across repeated outer cross-validation evaluations.

(B) Confusion matrices for ensemble predictions in v1 and v2. Columns denote the true cancer type and rows denote the predicted cancer type. Color intensity indicates the proportion of samples within each true cancer type assigned to each predicted class.

CNA, copy-number alteration; SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TFBS, transcription factor binding site; TOO, tissue of origin.

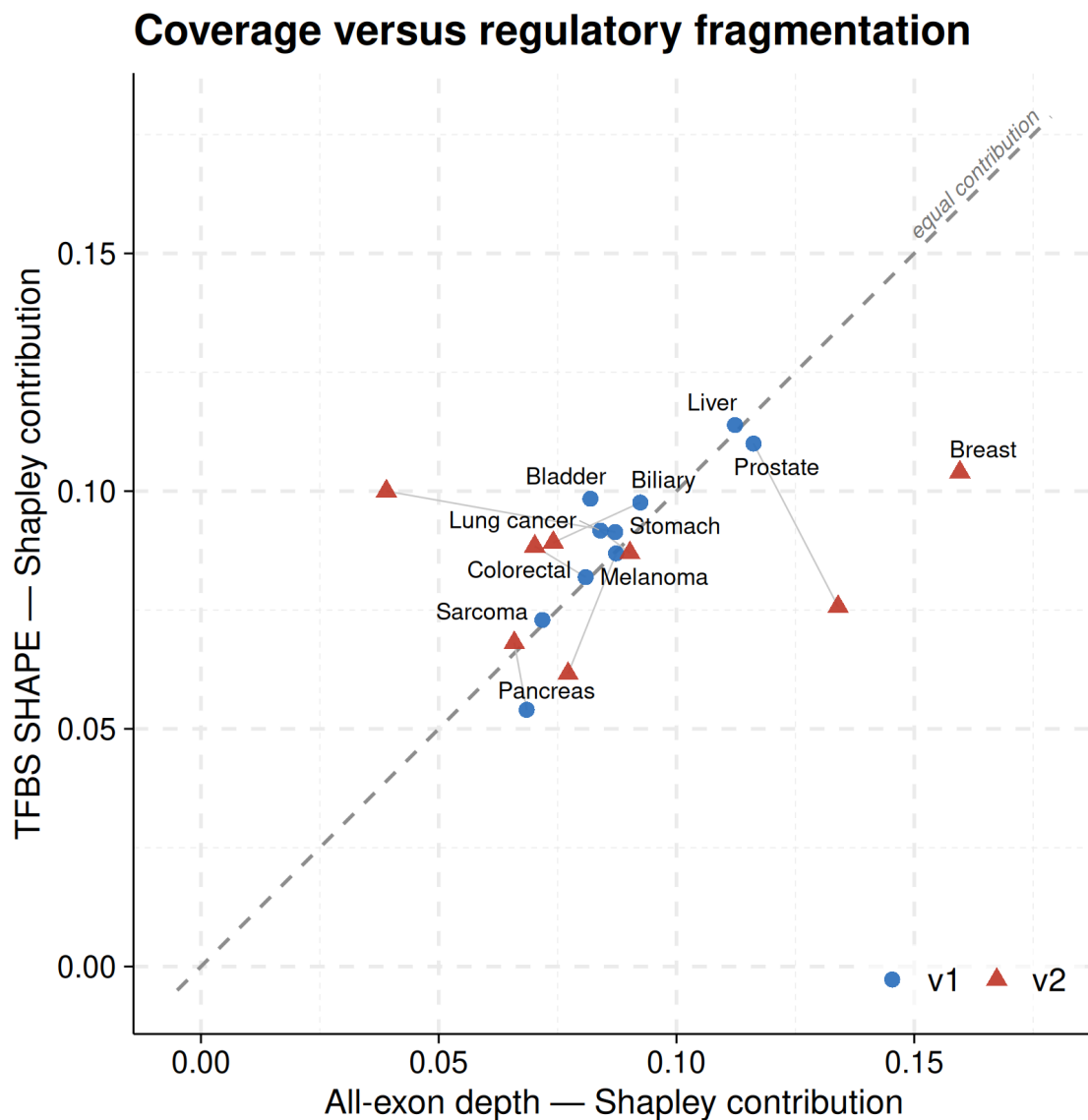


Figure S6. Cancer type-specific coverage-versus-fragmentation contribution.

For each cancer type, the Shapley contribution of all-exon depth (x-axis) is plotted against that of TFBS SHAPE (y-axis) in v1 (circles) and v2 (triangles); grey lines link the same cancer type across assay versions and the dashed diagonal marks equal contribution. Tumors separate along the two axes, indicating that coverage (all-exon depth) and regulatory fragmentation (TFBS SHAPE) make their largest contributions to largely different cancer types.

SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TFBS, transcription factor binding site.

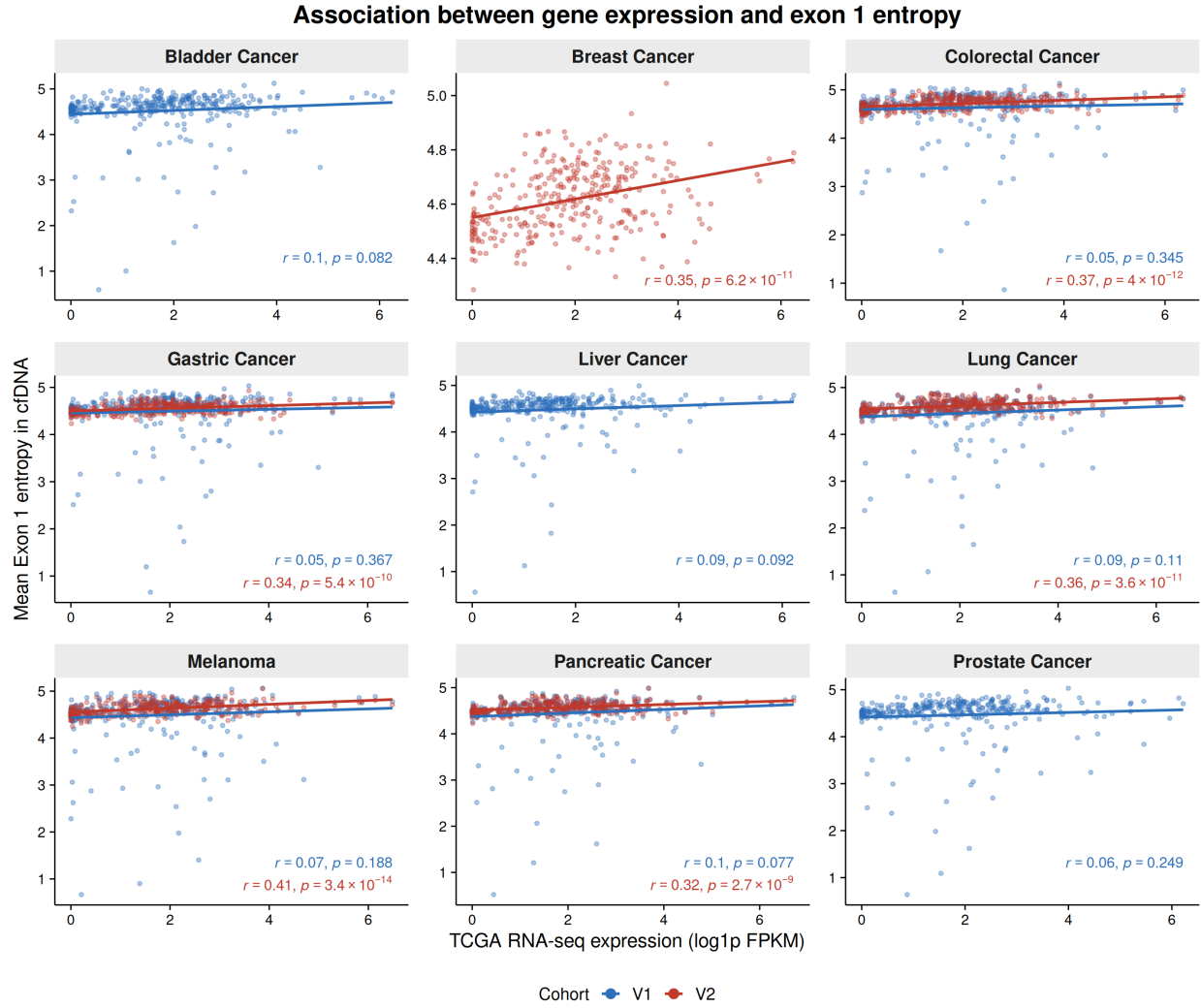


Figure S7. Association between cancer type–matched gene expression and cfDNA exon 1 fragment-length entropy.

Each panel shows the association between TCGA gene expression and exon 1 fragment-length entropy for the indicated cancer type. Each point represents a gene with an exon 1 target included in the assay. The horizontal axis shows cancer type–matched TCGA RNA-seq expression, expressed as $\log_{1p}$ -transformed FPKM, and the vertical axis shows the mean exon 1 fragment-length entropy across cfDNA samples of the corresponding cancer type. Blue and red points and fitted linear regression lines denote v1 and v2, respectively. Pearson correlation coefficients (r) and two-sided PP values are shown within each panel.

cfDNA, cell-free DNA; FPKM, fragments per kilobase of transcript per million mapped reads; TCGA, The Cancer Genome Atlas.

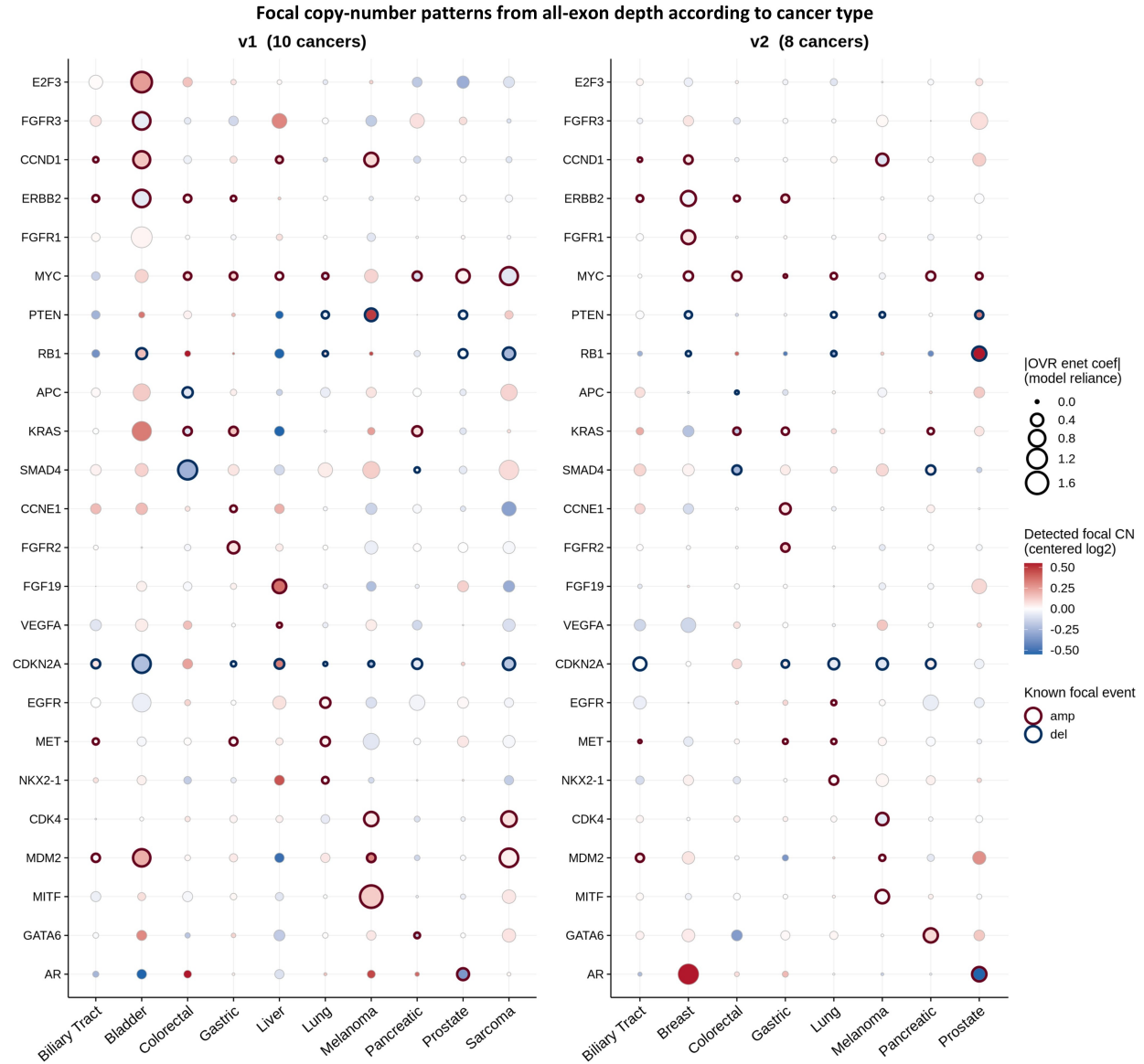


Figure S8. Cancer type–specific focal copy-number patterns captured by the all-exon depth signal. Dot plots summarize focal copy-number patterns derived from all-exon depth in v1 and v2. Rows denote cancer-associated genes and columns denote cancer types. Circle fill indicates the cancer type–specific deviation in normalized cfDNA depth from the cross-cancer average, with red and blue indicating relative gains and losses, respectively. Circle size represents the absolute one-versus-rest net coefficient. Colored outlines indicate curated focal amplifications and deletions. cfDNA, cell-free DNA; CNA, copy-number alteration.

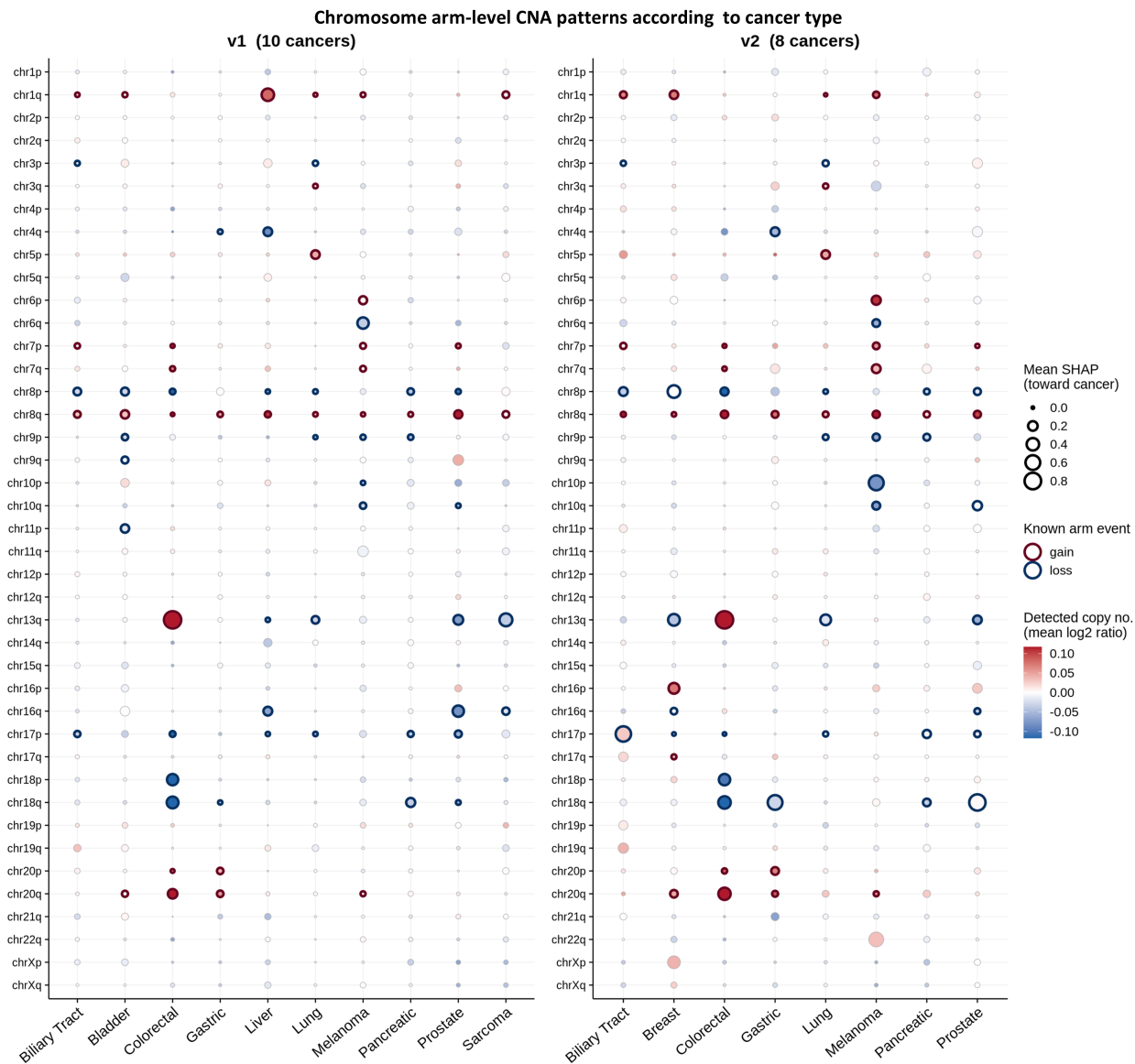


Figure S9. Cancer type–specific chromosome-arm copy-number patterns captured by the broad CNA signal.

Dot plots summarize chromosome-arm copy-number patterns in v1 and v2. Rows denote chromosome arms and columns denote cancer types. Circle fill indicates the mean arm-level log₂ copy ratio within each cancer type, with red and blue indicating copy-number gains and losses, respectively. Circle size represents the mean absolute SHAP value toward the corresponding cancer type. Colored outlines indicate curated cancer-associated arm-level gains and losses.

CNA, copy-number alteration; SHAP, SHapley Additive exPlanations.

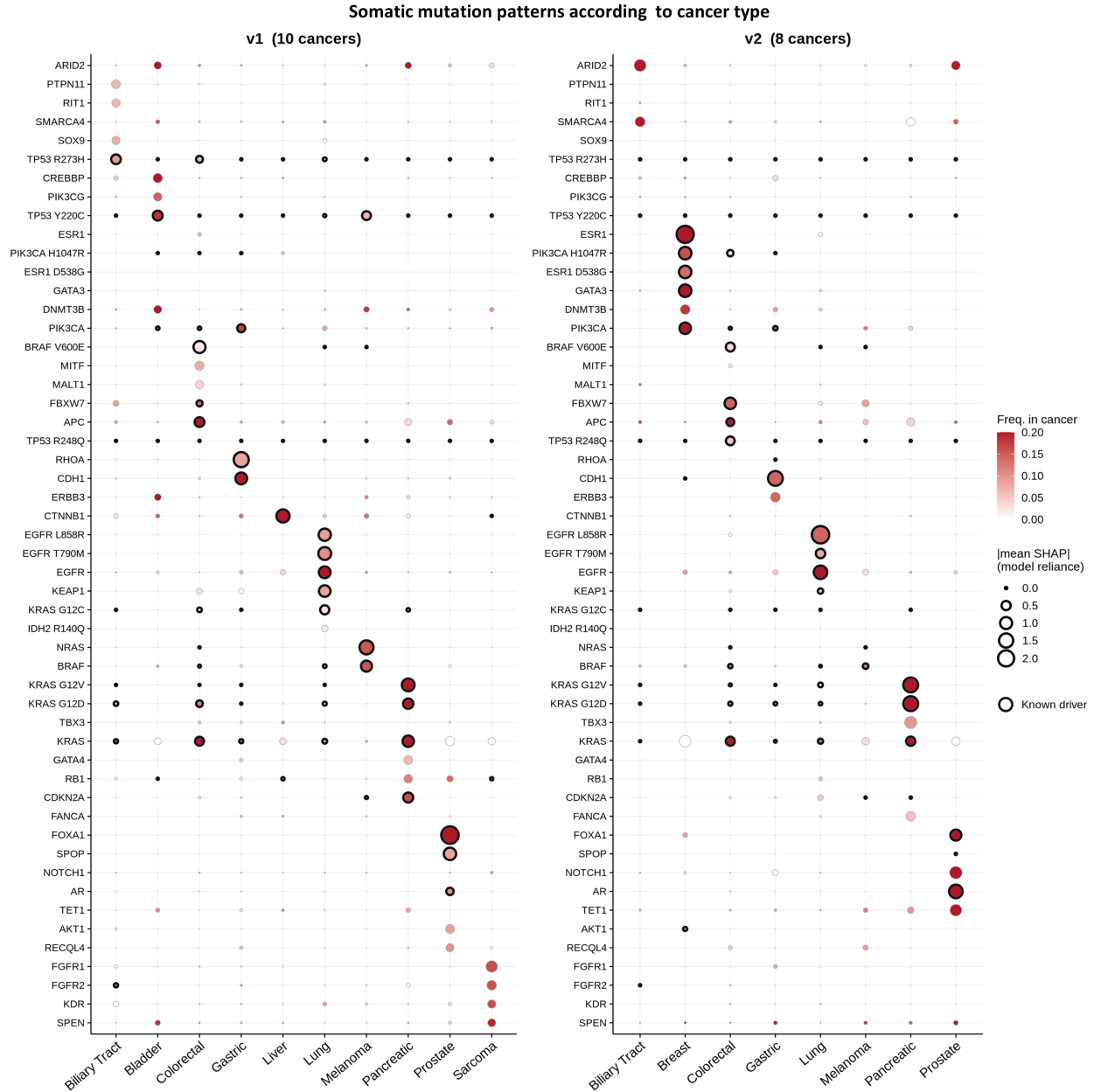


Figure S10. Cancer type-specific somatic mutation patterns and model contributions.

Dot plots summarize somatic mutation features in v1 and v2. Rows denote gene- or variant-level mutation features and columns denote cancer types. Circle fill indicates the proportion of samples within each cancer type harboring the corresponding alteration. Circle size represents the absolute mean class-directed SHAP value and therefore reflects the contribution of each mutation feature toward prediction of the corresponding cancer type. Outlined circles indicate curated cancer-associated driver alterations. SHAP, SHapley Additive exPlanations.

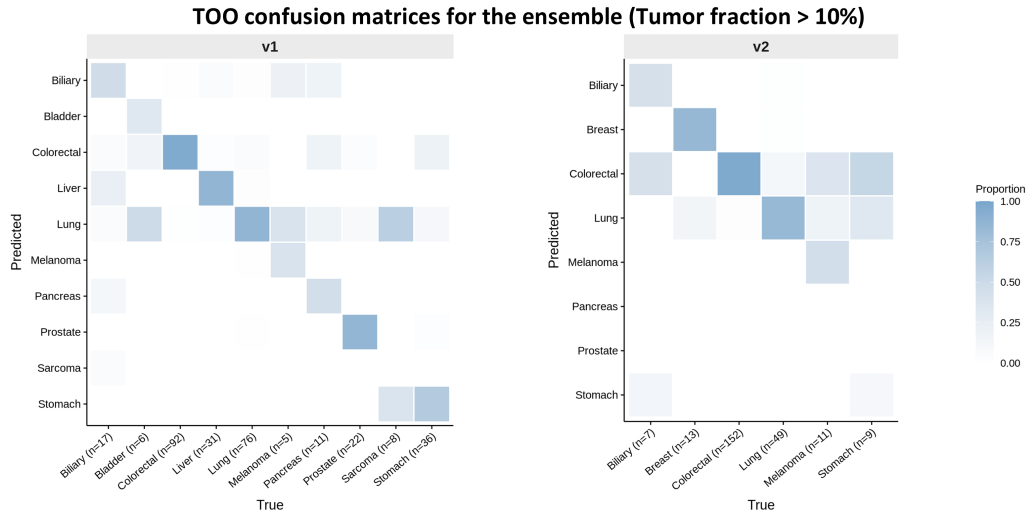


Figure S11. Ensemble tissue-of-origin confusion matrices for samples with tumor fraction above 10%.

Column-normalized confusion matrices show out-of-fold predictions from the late-fusion ensemble for samples with estimated tumor fraction >10% in v1 and v2. Columns denote the true cancer type and rows denote the predicted cancer type. Color intensity indicates the proportion of samples within each true class assigned to each predicted class; diagonal values therefore represent cancer type-specific sensitivity. Blank cells indicate zero or negligible proportions.

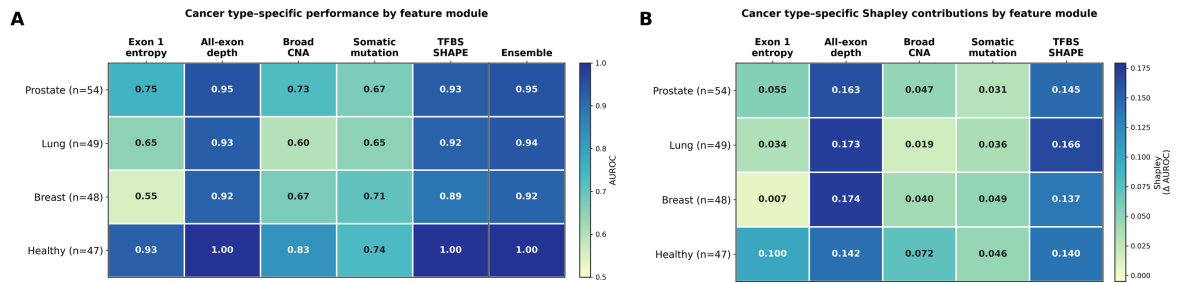


Figure S12. Class-specific prediction performance and feature-signal contributions in the GRAIL cohort.

(A) Heatmap showing the one-versus-rest AUROC of each signal type and the ensemble for individual prediction classes. The ensemble integrates all five signal types.

(B) Class-specific Shapley contributions of individual signal types to ensemble performance, expressed in AUROC units (Δ AUROC).

Values within cells represent estimates averaged across repeated cross-validation runs, and color intensity indicates the magnitude of the corresponding AUROC or Shapley contribution. CNA, copy-number alteration; TFBS, transcription factor binding site; SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; AUROC, area under the receiver operating characteristic curve.

Supplementary Tables

Table S1. ENCODE primary-hematopoietic DNase-seq footprint experiments used to define consensus transcription-factor footprints.

118 experiments across 6 hematopoietic lineages; a UniBind site was classified as footprint-overlapping ('protected') if at least one experiment called a footprint overlapping it (6,805 protected sites). Footprints were called on GRCh38. Source: ENCODE (<https://www.encodeproject.org>).

No	Cell type / lineage	ENCODE experiment accession	Footprint file accession
1	B cell	ENCSR000EII	ENCFF835IKT
2	B cell	ENCSR000EMJ	ENCFF379FLA
3	B cell	ENCSR247IUJ	ENCFF647APV
4	B cell	ENCSR381PXW	ENCFF570QXA
5	B cell	ENCSR468AKF	ENCFF425DHS
6	B cell	ENCSR891VOV	ENCFF589RPD
7	CD14+ monocyte	ENCSR000ELE	ENCFF434ICV
8	CD14+ monocyte	ENCSR000EPK	ENCFF410CBB
9	CD14+ monocyte	ENCSR073CPU	ENCFF451QLX
10	CD14+ monocyte	ENCSR083DZB	ENCFF892TQQ
11	CD14+ monocyte	ENCSR223HHH	ENCFF091CSP
12	CD14+ monocyte	ENCSR241BNZ	ENCFF449YDY
13	CD14+ monocyte	ENCSR277ZHX	ENCFF122VNI
14	CD14+ monocyte	ENCSR452DCM	ENCFF082WCI
15	CD14+ monocyte	ENCSR464ETX	ENCFF951CXU
16	CD14+ monocyte	ENCSR491DPW	ENCFF657IFU

17	CD14+ monocyte	ENCSR539ZYY	ENCFF048XTP
18	CD14+ monocyte	ENCSR629GFF	ENCFF714CUU
19	CD14+ monocyte	ENCSR674JIL	ENCFF249EJQ
20	CD14+ monocyte	ENCSR695AUY	ENCFF453NKC
21	CD14+ monocyte	ENCSR889CFS	ENCFF616QAR
22	CD14+ monocyte	ENCSR913VNJ	ENCFF922WCB
23	CD34+ common myeloid progenitor	ENCSR122VUW	ENCFF813EIE
24	CD34+ common myeloid progenitor	ENCSR191EII	ENCFF721EGT
25	CD34+ common myeloid progenitor	ENCSR271QSV	ENCFF010KVD
26	CD34+ common myeloid progenitor	ENCSR281HBB	ENCFF014VAZ
27	CD34+ common myeloid progenitor	ENCSR320TUI	ENCFF715OIT
28	CD34+ common myeloid progenitor	ENCSR453EVC	ENCFF267BQN
29	CD34+ common myeloid progenitor	ENCSR468ZXN	ENCFF197VDZ
30	CD34+ common myeloid progenitor	ENCSR495INQ	ENCFF989FQW
31	CD34+ common myeloid progenitor	ENCSR564TUY	ENCFF696NLO
32	CD34+ common myeloid progenitor	ENCSR639ZJI	ENCFF636AGQ
33	CD34+ common myeloid progenitor	ENCSR735BIM	ENCFF784OIL
34	CD34+ common myeloid progenitor	ENCSR772HUG	ENCFF757RIU
35	CD4+ T cell	ENCSR036ELY	ENCFF588CMG
36	CD4+ T cell	ENCSR060ZXQ	ENCFF487GDQ
37	CD4+ T cell	ENCSR098YWP	ENCFF810YJM
38	CD4+ T cell	ENCSR111WTE	ENCFF447TQH
39	CD4+ T cell	ENCSR126EVX	ENCFF108DVJ

40	CD4+ T cell	ENCSR129GCL	ENCFF656ZFW
41	CD4+ T cell	ENCSR134BSY	ENCFF535VQD
42	CD4+ T cell	ENCSR152FWD	ENCFF875LHK
43	CD4+ T cell	ENCSR155NBS	ENCFF523ZXE
44	CD4+ T cell	ENCSR163QXS	ENCFF471BUE
45	CD4+ T cell	ENCSR167JFX	ENCFF047JPF
46	CD4+ T cell	ENCSR206KJN	ENCFF663NFE
47	CD4+ T cell	ENCSR240TZG	ENCFF146CXP
48	CD4+ T cell	ENCSR250GDW	ENCFF288YPT
49	CD4+ T cell	ENCSR259YKO	ENCFF249KEF
50	CD4+ T cell	ENCSR289GZQ	ENCFF786WVQ
51	CD4+ T cell	ENCSR314IOV	ENCFF216RPT
52	CD4+ T cell	ENCSR324OTS	ENCFF165VZT
53	CD4+ T cell	ENCSR336QAO	ENCFF138PAU
54	CD4+ T cell	ENCSR346GIP	ENCFF122EYT
55	CD4+ T cell	ENCSR363LUK	ENCFF050EQS
56	CD4+ T cell	ENCSR366KPR	ENCFF828ZEN
57	CD4+ T cell	ENCSR376DAE	ENCFF460CTI
58	CD4+ T cell	ENCSR388NOI	ENCFF688WOM
59	CD4+ T cell	ENCSR389BWT	ENCFF769IAH
60	CD4+ T cell	ENCSR408GYQ	ENCFF892CSP
61	CD4+ T cell	ENCSR424MWH	ENCFF976BJJ
62	CD4+ T cell	ENCSR425MRQ	ENCFF146SAZ

63	CD4+ T cell	ENCSR452RAX	ENCFF241TRK
64	CD4+ T cell	ENCSR483PGP	ENCFF258KOH
65	CD4+ T cell	ENCSR498FCF	ENCFF445TFR
66	CD4+ T cell	ENCSR515DVK	ENCFF747KJZ
67	CD4+ T cell	ENCSR546FUC	ENCFF835LDD
68	CD4+ T cell	ENCSR566XSD	ENCFF812FXA
69	CD4+ T cell	ENCSR567IWT	ENCFF168WJZ
70	CD4+ T cell	ENCSR569ATD	ENCFF268EYE
71	CD4+ T cell	ENCSR572FSS	ENCFF082DWX
72	CD4+ T cell	ENCSR601QTE	ENCFF065LLR
73	CD4+ T cell	ENCSR635VQU	ENCFF581KTI
74	CD4+ T cell	ENCSR642BXP	ENCFF554WDQ
75	CD4+ T cell	ENCSR643GHI	ENCFF777VRB
76	CD4+ T cell	ENCSR654WXX	ENCFF931RWB
77	CD4+ T cell	ENCSR670KEJ	ENCFF908FCD
78	CD4+ T cell	ENCSR683QJJ	ENCFF227LMO
79	CD4+ T cell	ENCSR697FME	ENCFF521TUD
80	CD4+ T cell	ENCSR753LFV	ENCFF179LMU
81	CD4+ T cell	ENCSR769OIG	ENCFF334ZTO
82	CD4+ T cell	ENCSR824YLD	ENCFF608BXP
83	CD4+ T cell	ENCSR832IQN	ENCFF213YPL
84	CD4+ T cell	ENCSR883WDW	ENCFF071QNV
85	CD4+ T cell	ENCSR904GDU	ENCFF853DVV

86	CD4+ T cell	ENCSR930USX	ENCFF550TQL
87	CD4+ T cell	ENCSR940UEK	ENCFF088TUW
88	CD4+ T cell	ENCSR941MHZ	ENCFF604WIV
89	CD8+ T cell	ENCSR020LUD	ENCFF561QAJ
90	CD8+ T cell	ENCSR086FTP	ENCFF321QVN
91	CD8+ T cell	ENCSR097AMI	ENCFF206TIK
92	CD8+ T cell	ENCSR129HDK	ENCFF427IWI
93	CD8+ T cell	ENCSR133QZB	ENCFF323SLA
94	CD8+ T cell	ENCSR162JNV	ENCFF429SRP
95	CD8+ T cell	ENCSR235OZX	ENCFF899ZOW
96	CD8+ T cell	ENCSR254BCA	ENCFF030YRO
97	CD8+ T cell	ENCSR316UDN	ENCFF781LVL
98	CD8+ T cell	ENCSR321KUC	ENCFF187FJN
99	CD8+ T cell	ENCSR363BTB	ENCFF318HFV
100	CD8+ T cell	ENCSR424JPH	ENCFF204UBV
101	CD8+ T cell	ENCSR559QCM	ENCFF184SMU
102	CD8+ T cell	ENCSR609DDQ	ENCFF932JQB
103	CD8+ T cell	ENCSR649XRC	ENCFF308IFV
104	CD8+ T cell	ENCSR662IRS	ENCFF510VZE
105	CD8+ T cell	ENCSR744UQK	ENCFF139ABY
106	CD8+ T cell	ENCSR840TVG	ENCFF443CXY
107	CD8+ T cell	ENCSR842VTJ	ENCFF061GLN
108	CD8+ T cell	ENCSR937NLH	ENCFF095PVT

109	CD8+ T cell	ENCSR946BPA	ENCFF052GSM
110	CD8+ T cell	ENCSR984HIY	ENCFF166PFO
111	NK cell	ENCSR407WGG	ENCFF195NIF
112	NK cell	ENCSR409ZBD	ENCFF382ZEO
113	NK cell	ENCSR618TIP	ENCFF418HGN
114	NK cell	ENCSR704HNG	ENCFF797WOU
115	NK cell	ENCSR718AAB	ENCFF536YLG
116	NK cell	ENCSR723JLG	ENCFF250HRG
117	NK cell	ENCSR794TLA	ENCFF282XCP
118	NK cell	ENCSR975LGK	ENCFF449GCM

Table S2. Tissue-of-origin classification performance by cancer type and signal type.

Cohort	Cancer type	n	Exon 1 entropy	All-exon depth	Broad CNA	Somatic mutation	TFBS SHAPE	Ensemble
v1	Overall	1093	0.732	0.815	0.737	0.746	0.816	0.87
v1	Biliary tract	76	0.726	0.802	0.686	0.624	0.811	0.815
v1	Bladder cancer	27	0.604	0.772	0.652	0.77	0.778	0.849
v1	Colorectal cancer	182	0.824	0.86	0.827	0.864	0.866	0.919
v1	Gastric cancer	155	0.774	0.826	0.758	0.717	0.834	0.878
v1	Liver cancer	67	0.919	0.944	0.855	0.755	0.951	0.959
v1	Lung cancer	391	0.807	0.849	0.793	0.76	0.852	0.901
v1	Melanoma	45	0.734	0.824	0.782	0.775	0.825	0.878
v1	Pancreatic cancer	61	0.7	0.75	0.65	0.803	0.733	0.848
v1	Prostate cancer	58	0.81	0.902	0.759	0.73	0.893	0.921
v1	Sarcoma	31	0.522	0.733	0.782	0.765	0.726	0.844
v2	Overall	796	0.759	0.813	0.75	0.784	0.818	0.872
v2	Biliary tract cancer	30	0.808	0.797	0.719	0.718	0.835	0.842
v2	Breast cancer	44	0.793	0.97	0.819	0.846	0.911	0.969
v2	Colorectal cancer	326	0.835	0.837	0.874	0.851	0.88	0.917
v2	Gastric cancer	35	0.7	0.708	0.703	0.728	0.794	0.814
v2	Lung cancer	257	0.74	0.841	0.79	0.803	0.836	0.894
v2	Melanoma	33	0.77	0.809	0.783	0.735	0.778	0.843

v2	Pancreatic cancer	51	0.72	0.757	0.693	0.825	0.767	0.869
v2	Prostate cancer	20	0.799	0.891	0.747	0.831	0.828	0.918

Overall performance is the macro-averaged one-versus-rest AUROC, reported as the mean across cross-validation repeats (corresponding to the distributions in Figure 3A); cancer type-specific performance is the one-versus-rest AUROC computed on pooled held-out predictions (Figure 3B). Because these use different aggregations, the overall value is not the arithmetic mean of the cancer type-specific values. AUROC, area under the receiver operating characteristic curve; CNA, copy-number alteration; SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TFBS, transcription factor binding site.

Table S3. Statistical support for complementary, coverage-independent signal contributions across cohorts.

A. Contribution beyond all other signals (leave-one-out Δ AUROC, 95% CI)

Signal	v1	v2	GRAIL
Somatic mutations	+0.033 (0.023–0.045)	+0.028 (0.017–0.040)	+0.001 (–0.001–0.002), ns
All-exon depth	+0.007 (0.002–0.012)	+0.016 (0.005–0.028)	+0.015 (0.005–0.026)
TFBS SHAPE	+0.007 (0.003–0.011)	+0.008 (0.004–0.013)	+0.002 (–0.004–0.008), ns

B. TFBS SHAPE added to all-exon depth (Δ AUROC, 95% CI; *P*)

Signal	v1	v2	GRAIL
TFBS SHAPE on top of coverage	+0.014 (0.010–0.020; <i>P</i> <0.001)	+0.029 (0.024–0.034; <i>P</i> <0.001)	+0.002, ns

Δ AUROC, change in macro-averaged one-versus-rest AUROC (0.5 = chance); positive values indicate improved tissue-of-origin discrimination. Leave-one-out, accuracy lost when a single signal is removed from the full combined classifier; 95% CI from patient-level bootstrap (2,000 resamples). B compares a coverage-only classifier (all-exon depth) with coverage plus TFBS SHAPE, using a paired test across identical validation folds. ns, 95% CI includes zero.

Table S4. Tissue-of-origin classification performance across tumor-fraction strata.

Cohort	Signal	<3%	3–10%	10–20%	>20%
v1	Sample size, n	479	310	119	185
v1	Somatic mutation	0.72	0.744	0.802	0.801
v1	Broad CNA	0.647	0.75	0.834	0.868
v1	All-exon depth	0.716	0.836	0.913	0.922
v1	Exon 1 entropy	0.651	0.74	0.762	0.867
v1	TFBS SHAPE	0.74	0.815	0.931	0.926
v1	Ensemble	0.812	0.883	0.959	0.949
v2	Sample size, n	322	225	83	166
v2	Somatic mutation	0.736	0.782	0.865	0.826
v2	Broad CNA	0.623	0.75	0.848	0.885
v2	All-exon depth	0.738	0.814	0.932	0.89
v2	Exon 1 entropy	0.655	0.752	0.89	0.857
v2	TFBS SHAPE	0.725	0.824	0.934	0.918
v2	Ensemble	0.811	0.866	0.945	0.941

Values indicate macro-AUROC for tissue-of-origin classification within strata defined by tumor fraction estimated using ichorCNA. AUROC, area under the receiver operating characteristic curve; CNA, copy-number alteration; SHAPE, Site-directed High-resolution Analysis of Positional Endpoints and Entropy; TFBS, transcription factor binding site.