Integrated cfDNA methylation and genomic profiling enables tissue-of-origin prediction from a single liquid biopsy

Reilly O Cooper, Manan Chopra,
David M Huland, Wen Zhou, Allen M Chen

All authors are affiliated with BillionToOne, Inc., Menlo Park, CA
Contact: rcooper@billiontoone.com

NORTHSTAR by BILLIONTOONE

INTRODUCTION

- Patients with cancer of unknown primary (CUP) face limited treatment options and poor survival*.
- Standard workup fails to identify a primary site in a minority of cases, and tissue availability can be problematic in these situations.
- Cell-free DNA (cfDNA) carries tissue-specific methylation signatures, providing a basis for predicting a tumor's tissue of origin (TOO).
- We developed a 12-class machine-learning model trained on the per-locus methylation data; each sample receives a ranked prediction of TOO with calibrated confidence, tissue classes that cannot be excluded, and tissues of origin that can be excluded from consideration.
- Here, we characterize the model in a retrospective cohort of 1578 blood samples from advanced cancer patients.

* Spurgeon, L., Mitchell, C., Cook, N., Conway, A-M. British Journal of Cancer (2025) 133:733-742.

METHODS

- Methylation data from baseline time-point samples of patients with a Northstar Response result were analyzed retrospectively. Patients were split into a model training cohort (n = 4447) and a held-out test cohort (n = 1578).
- Training was restricted to 12 reportable tissue classes (biliary, bladder, breast, colorectal, gastroesophageal, gynecological, kidney, liver, lung, cutaneous, pancreatic, and prostate) and a non-cancer class. All samples, except those in the non-cancer class, had a total tumor methylation score (TMS) ≥ 25.
- Per-locus methylation values across the >2200 targeted regions in Northstar Response were normalized and used to train a regularized multinomial classifier over 12 tumor classes and a non-cancer class.
- Class probabilities were adjusted by a prior conditioned on patient sex. A set of tissue-specific somatic variants were used to further adjust class probabilities in the cases where these variants were identified in a paired Northstar Select result. Reported confidences were calibrated.
- Clinical diagnosis submitted with test requisition served as the ground truth for model classification accuracy. Diagnoses were mapped to tissue classes using a custom diagnosis ontology.
- Results were restricted to samples with TMS ≥ 25 or model confidence ≥ 40%. Additionally, if a sample's clinical diagnosis fell outside the 12 reportable tissue classes, it was excluded from analysis.

Download Poster

ACKNOWLEDGEMENTS

We thank Patrick Cherry and Laura Herron for their initial conceptual and model development work, BillionToOne Development Team members for collaborative scientific discussions, and Grace Lee for pathology report analysis.

RESULTS

High-confidence tissue-of-origin prediction achieves >96% accuracy

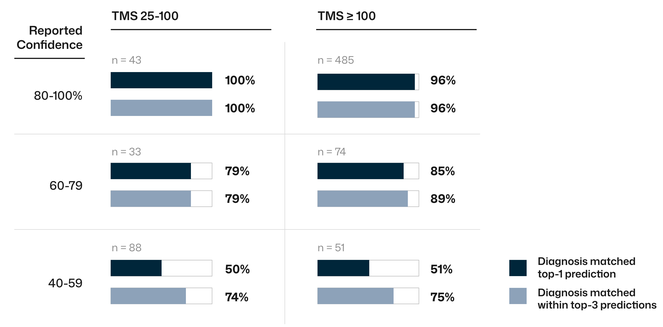
Table 1. Evaluated cohort characteristics

Tissue Class	n	Samples (Male / Female)	Relative TMS (log scale)
Breast	318	1 / 317	
Colorectal	257	136 / 121	
Lung	234	117 / 117	
Pancreatic	136	67 / 69	
Prostate	120	120 / 0	
Gynecologic	96	0 / 96	
Gastroesophageal	81	55 / 26	
Bladder	45	30 / 15	
Cutaneous	38	25 / 13	
Kidney	36	22 / 14	
Biliary	35	14 / 21	
Liver	22	20 / 2	
Untrained classes	160	90 / 70	
All patients	**1578**	**686 / 872**	

25 50 100 200 1000

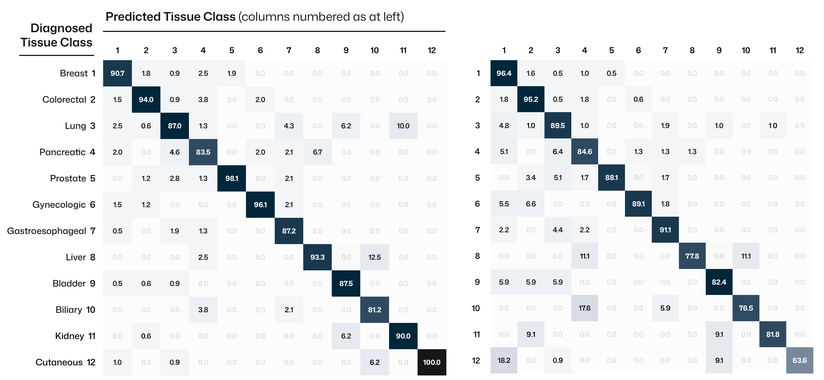
Among 1578 patients, 1418 were eligible for classification. The cohort included some samples outside of the reportable classes: sarcoma (62), head and neck (42), hematologic (15), neuroendocrine (12), SCC and other non-classifiable skin (17), anal (8), and testicular (4).

Figure 1. Model confidence and tumor methylation score (TMS)



Many predictions were high-confidence (>80%), indicating the model can identify tissue-specific methylation patterns across TMS (96.4% concordance across TMS bins). Bars show the share of samples in each bin whose diagnosis matched the top prediction (dark) and whose diagnosis fell within the top three predictions (light), with the number of samples in each bin.

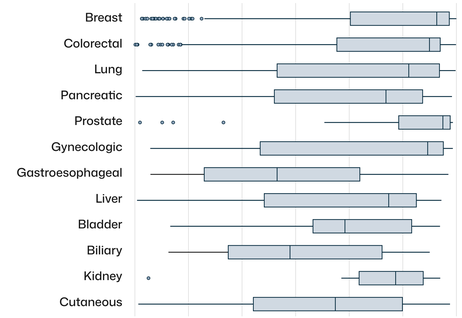
Figure 2. Concordance between predicted and diagnosed tissue of origin



(A) Positive predictive value, normalized by column. The model is precise across classes. Common tissue classes are well-characterized by the model, and incorrect assignment only occurs in rare cases among less well-represented classes that are biologically similar.

(B) Recall, normalized by row. In general, incorrect predictions occur within biologically related tissues; liver, biliary, and pancreatic tissues of origin are incorrectly assigned to each other at higher frequencies than other tissues.

Figure 3. Distribution of model prediction confidence across classes



Among reported cases, many predictions are high confidence. Commonly represented tissue classes have primarily high confidence predictions, while rarer classes range in confidence in a methylation-dependent manner.

DISCUSSION & CONCLUSIONS

- The distribution of methylation across >2200 selected loci allows for the classification of cancer tissue of origin. Our developed model correctly assigned the tissue of origin in the top three model predictions in 91% of cases, and as its top model prediction in 87% of eligible cases. Results are highly accurate (>96%) when the model is highly confident. Discordant predictions primarily occur in biologically related tissues, as methylation signals are similar.
- Confident tissue class rule-outs at 97.5% sensitivity provides additional benefit to patients, naming primary sites that are unlikely to be the true tissue of origin. Each report excluded a mean of 3.8 classes, while retaining the correct TOO in 98.4% of eligible samples.
- For patients whose primary site is unresolved after standard evaluation, this assay returns a ranked set of tissue classes and associated probabilities from a single blood draw. These findings may be useful to customize treatment for patients and to provide valuable reassurance to clinicians aiming to optimize care delivery.

Colors (• Primary)

Coral 1 HEX: #CD4D54 RGB: 205 77 84	**Coral 2 •** HEX: #FE706F RGB: 254 112 111	**Coral 3** HEX: #FEA9A9 RGB: 254 169 169	**Coral 4** HEX: #FADCDC RGB: 250 220 220	**Coral 5** HEX: #FAEBEB RGB: 250 235 235	**Coral 6** HEX: #FAF5F5 RGB: 250 245 245
Blue 1 (Midnight) • HEX: #00263a RGB: 0 38 58	**Blue 2** HEX: #2f4a60 RGB: 47 74 96	**Blue 3 (Powder) •** HEX: #8DA2B9 RGB: 141 162 185	**Blue 4** HEX: #D1D7E2 RGB: 209 215 226	**Blue 5** HEX: #E6EAF0 RGB: 230 233 240	**Blue 6** HEX: #F3F5F7 RGB: 244 245 249
Sand 1 HEX: #8E7F6D RGB: 142 127 109	**Sand 2** HEX: #B9AEA2 RGB: 185 174 162	**Sand 3 •** HEX: #D6CEC4 RGB: 214 206 196	**Sand 4** HEX: #F0ECE8 RGB: 240 236 232	**Sand 5** HEX: #F5F3F0 RGB: 245 243 240	**Sand 6** HEX: #F7F7F5 RGB: 247 247 245
Moss 1 HEX: #324832 RGB: 50 72 50	**Moss 2** HEX: #426043 RGB: 66 96 67	**Moss 3 •** HEX: #A7AC72 RGB: 167 172 114	**Moss 4** HEX: #D9DBC4 RGB: 217 219 196	**Moss 5** HEX: #E9EBDD RGB: 233 235 221	**Moss 6** HEX: #F5F5F2 RGB: 245 245 242
Amber 1 HEX: #9E7445 RGB: 158 116 69	**Amber 2** HEX: #E1A45B RGB: 225 164 91	**Amber 3 •** HEX: #E7B781 RGB: 231 183 129	**Amber 4** HEX: #EDE0D1 RGB: 237 224 209	**Amber 5** HEX: #F2EEE9 RGB: 242 238 233	**Amber 6** HEX: #F7F5F2 RGB: 247 245 242
Soft Black HEX: #171717 RGB: 23 23 23	**Neutral 1** HEX: #434343 RGB: 67 67 67	**Neutral 2** HEX: #606060 RGB: 96 96 96	**Neutral 3** HEX: #7d7d7d RGB: 125 125 125	**Neutral 4** HEX: #9a9a9a RGB: 154 154 154	**Neutral 5** HEX: #b3b3b3 RGB: 179 179 179
Neutral 6 HEX: #cccccc RGB: 204 204 204	**Neutral 7** HEX: #e5e5e5 RGB: 229 229 229	**Neutral 8** HEX: #f0f0f0 RGB: 240 240 240	**Neutral 9** HEX: #f5f5f5 RGB: 245 245 245	**Neutral 10** HEX: #fafafa RGB: 250 250 250	**White** HEX: #ffffff RGB: 255 255 255