

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

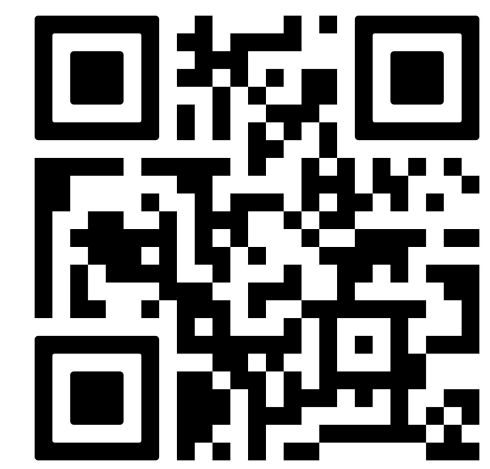
All authors are affiliated with BillionToOne, Inc., Menlo Park, CA
Contact: rcoper@billiontoone.com



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.



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.

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.

RESULTS

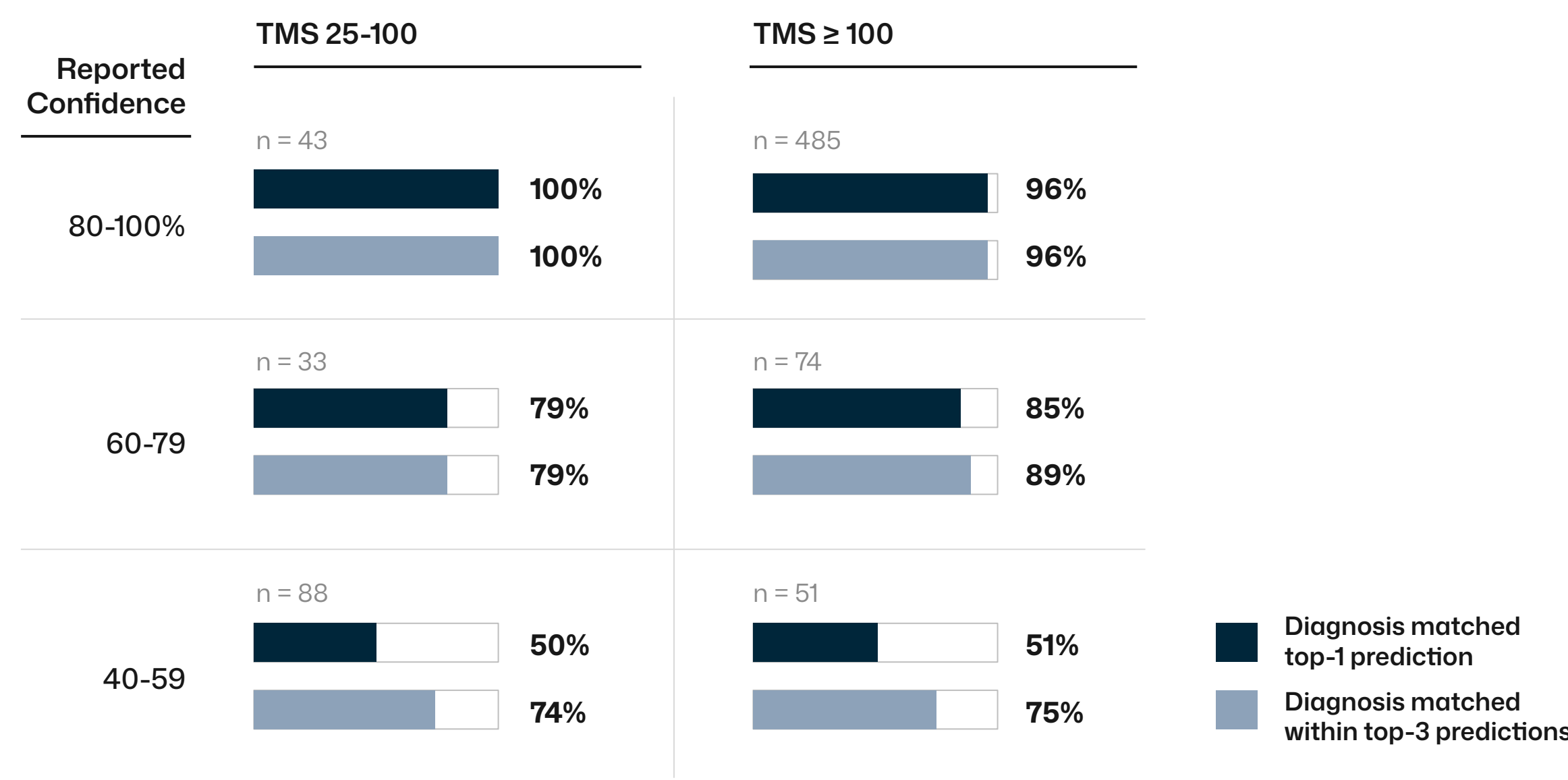
High-confidence tissue-of-origin prediction achieves >96% accuracy

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

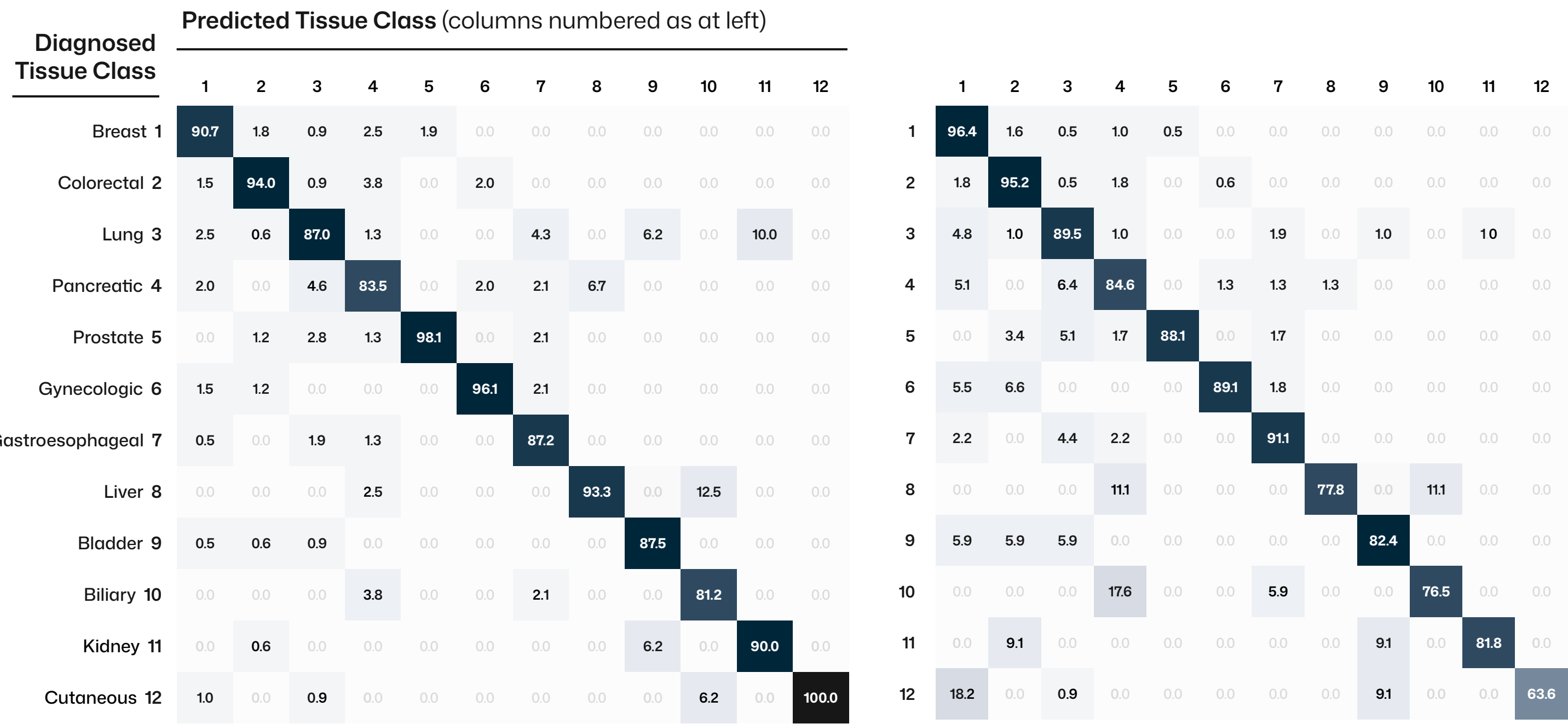
Among 1578 patients, 1418 were eligible for classification. The cohort included some samples outside of the reportable classes: sarcoma (62), head and neck (42), hematologic (15), neuroendocrine (12), SCC and other non-classifiable skin (17), anal (8), and testicular (4).

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.

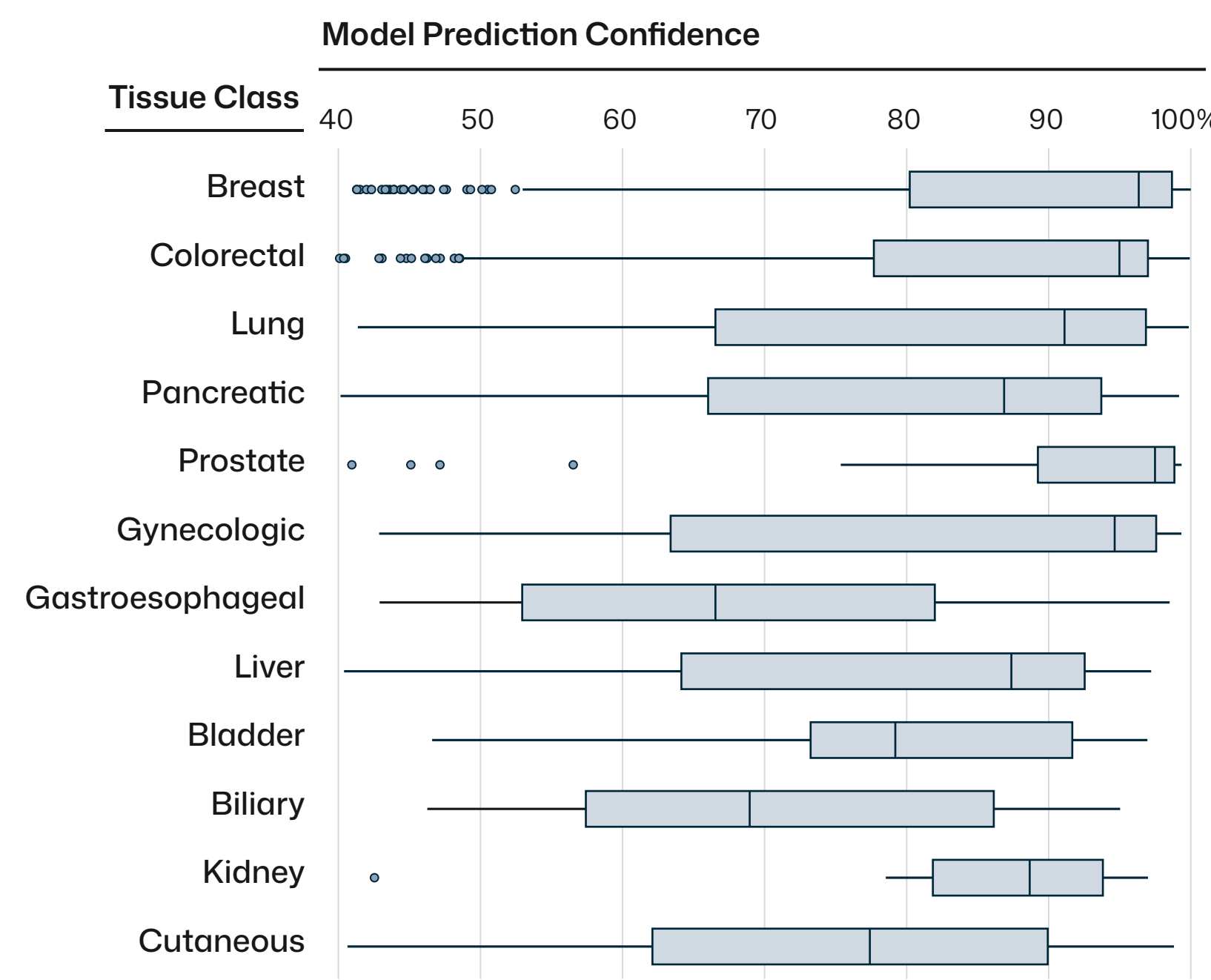
Figure 2. Concordance between predicted and diagnosed tissue of origin



(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.