5465

AACR April 2026

Improving early cancer detection by training scalable deep neural networks to extract tumor signal from cell-free DNA

Harbinger Health

Section 2 | Board 1
Abstract 7064

Jackson A. Killian, PhD[1]; Kade Pettie, PhD[1]; Kyle Gowen[1]; Shiva Farashahi, PhD[1]; Esther Brown[1]; Feras Hantash, PhD[1]; Jocelyn Charlton, PhD[1]; Kieran I. Chacko, PhD[1]; Dorna Kashef, PhD[1]
[1]Harbinger Health, Cambridge, MA * Corresponding author: jkillian@harbinger-health.com

INTRODUCTION

Blood-based liquid biopsies offer potential for non-invasive cancer screening. However, detecting early-stage disease is complicated by low levels of circulating tumor biomarkers and background noise from normal cells. While existing liquid biopsy methods rely on region-level average methylation, which dilutes the sparse tumor signal present in early-stage disease, our approach operates at the resolution of individual DNA reads to preserve and directly exploit this signal. We developed ProbeCDN™, a deep-learning framework to extract cancer-associated methylation signal at single-read resolution. The framework leverages a novel, large-scale data generation technique that enables learning from limited input samples. Applied to bisulfite-converted cell-free DNA (cfDNA), our method improves stage I and II cancer sensitivity by 6.5 and 17.6 percentage points.

METHODS

We designed Probe cfDNA Network (**ProbeCDN™**), a parallel 2-D convolutional neural network architecture that differentiates cancer and non-cancer signal in cfDNA by learning local methylation patterns at thousands of genomic regions (CpG Islands). The model takes targeted bisulfite-converted NGS data as input (footprint of 18.6 Mb), encodes aligned sequences within genomic windows as images, and outputs informative feature vectors for classification (**Fig 1**).

Synthetic Data Generation

Training is complicated by real-world data limitations:

1) Disease samples contain an imbalanced mix of unlabeled fragments from normal and diseased cells
2) Acquiring sufficient early-stage disease data is costly, burdensome, and time-intensive
3) Batch effects + unintended confounders limit generalization

We address these via a novel data generation method[1] that combines tumor-derived data with cfDNA background yielding:

1) Positive region-level labels via in silico spike-in of tumor tissue reads at controlled tumor fractions (0.001%—25%)
2) Rich, large-scale datasets via controlled in silico mixing of non-cancer cfDNA reads
3) Cancer/non-cancer pairs with shared background to encourage invariance to batch-specific and biological noise

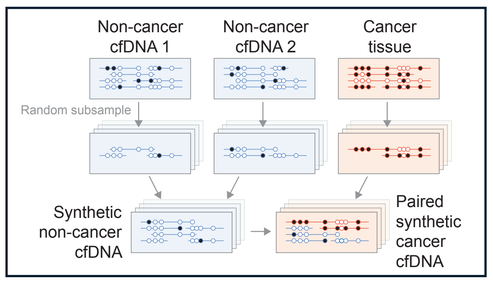

Figure 2. Schematic showing how synthetic training data is generated from a set of reference samples. CpGs are depicted as filled or unfilled circles to represent methylated and unmethylated states. Reads from two reference non-cancer cfDNA samples are mixed to create a synthetic non-cancer, which is then duplicated. Biopsy reads are spiked into one copy to create a matched cancer-non-cancer pair with identical background.

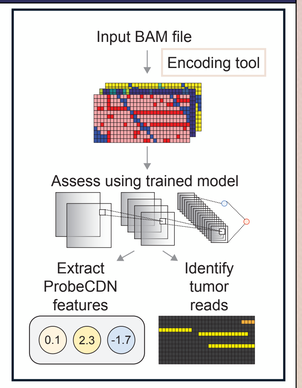

Figure 1. Schematic overview of ProbeCDN™ feature extraction for a single sample. Aligned reads from a BAM file are encoded as images which are then processed by the ProbeCDN™ network, producing output feature vectors and saliency maps (shown in Fig 3A) for inspecting reads the model identifies as tumor-associated.

METHODS

Architecture and Training

To balance performance and scalability, we trained multiple ProbeCDN™ models (sub-models), each specialized for a pre-specified cluster of genomic regions. Regions were grouped into 510 clusters based on pre-computed genomic features such as CpG density and region size. Each sub-model utilized a ResNet-18 backbone with a modified input head designed to process multiple channels of read-level data. After the adaptive average pooling layer, we attached multi-layer cancer yes/no classification head with dropout. Embeddings from the last hidden layer, together with the output layer values, were extracted and used as features for downstream classification (6 features per region). This architecture has key advantages: compact input encoding, interpretable saliency maps, and scalable parallel architecture. We trained the model on 1.3 billion synthetic training examples (see **Synthetic Data Generation**) generated using a held-out set of non-cancer plasma (N=174) and tumor tissue biopsies (N=505), amounting to 720TB of data across 10 million genomic bases. Since ProbeCDN™ sub-models operate independently across genomic regions, feature vectors from all 510 sub-models are concatenated and passed to a secondary binary classifier to enable genome-wide cancer signal detection. The secondary classifier applied Kernel PCA for nonlinear dimensionality reduction followed by logistic regression, evaluated using 10-fold cross-validation.

RESULTS

ProbeCDN™ Features Align with Biology

Saliency mapping on synthetic samples confirmed that ProbeCDN™ consistently prioritized spiked-in tumor reads based on their cancer-associated methylation patterns, while ignoring reads from the non-cancer background, including stochastic methylation patterns that may reflect sequencing errors, age-related changes or other confounders (**Fig 3A**). Applied to clinical samples from the CORE-HH study (NCT05435066, N=1229 non-cancers, N=1118 cancers across 20 cancer types, including N=599 Stage I/II), ProbeCDN™ scores showed strong separability of cancer and non-cancer samples (**Fig 3B, Fig 4A**) and correlated with tumor fraction (R^2 = 0.84, **Fig 4B**).

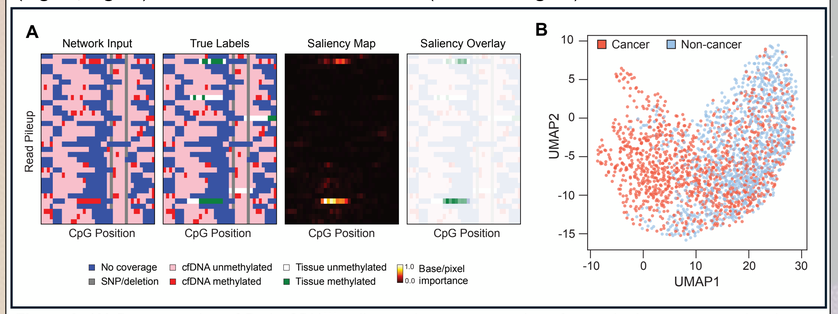

Figure 3. (A) Demonstration of ProbeCDN™ saliency plots on one region of a synthetic cancer sample where a contiguous row of non-blue pixels is one read. (Far left) input data of the methylation channel, as seen by ProbeCDN™. (Middle left) Input data of the methylation channel, showing which reads were from cancer tissue biopsy. (Middle right) saliency map produced by ProbeCDN™ after processing the image, brighter pixels indicate important pixels/bases. (Far right) The middle-left image with middle-right image overlaid as an alpha channel, showing that ProbeCDN™ has learned that highly methylated fragments are important cancer signal, while lowly or partially methylated fragments are not. **(B)** UMAP of all cancer and non-cancer samples using features extracted from ProbeCDN™.

RESULTS

Cancer Classification Performance

We evaluated the quality of ProbeCDN™ features for cancer classification performance on the CORE-HH cohort using the secondary classifier described in Methods. Performance was compared against a baseline classifier (same architecture) trained on region-level mean methylation values in place of ProbeCDN™ features. The classifier achieved an overall AUROC of 0.81 at a maximum FPR of 20% (**Fig 5A**). When targeting 98.5% specificity, the model achieved 98.4% specificity and improved overall sensitivity by 9.9 points (Stage I: +6.5 pts, II: +17.6 pts, III: +14.3 pts, IV: +9.5 pts), compared to the baseline (**Fig 5B**). ProbeCDN™ achieved an empirical limit of detection (LOD_{95}) of 0.078% tumor fraction, compared to 0.183% for the baseline model — a 2.3× improvement in detection sensitivity at the lowest tumor fractions (**Fig 5C**).

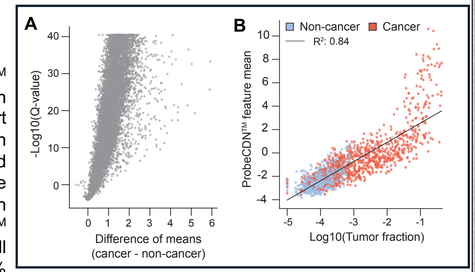

Figure 4. (A) Volcano plot of ProbeCDN™ classification-head feature across all regions. For each feature, the plot shows the mean cancer value minus the mean non-cancer value (x-axis) and the corresponding T-test Q-value (y-axis) for cfDNA samples. **(B)** For each sample, the tumor fraction is plotted against the ProbeCDN™ classification-head feature, averaged over regions. The line shows the fitted linear trend, and R^2 denotes the Pearson correlation coefficient.

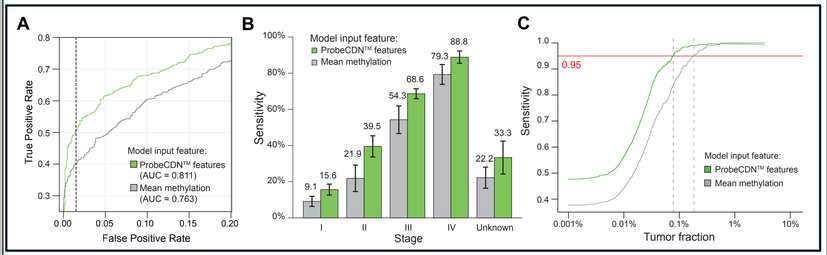

Figure 5. (A) ROC curves when training the binary classifier on ProbeCDN™ features versus region-level mean methylation. AUCs are computed across the displayed plotting area (FPR ≤ 20%). **(B)** Sensitivity by stage for classifiers trained using mean methylation or ProbeCDN™ features, at 98.5% achieved specificity. **(C)** Cumulative sensitivity by tumor fraction for samples evaluated using models trained with either mean methylation or ProbeCDN™ features. A red dashed line at 0.95 indicates 95% sensitivity and the empirical LOD_{95}.

CONCLUSIONS

ProbeCDN™ demonstrates that learning cancer-associated methylation patterns at single-read resolution improves early cancer detection, achieving a 2.3× improvement in limit of detection and up to 17.6 percentage points greater sensitivity in early-stage disease. Together with its scalable architecture and controlled synthetic data generation approach—which substantially reduces sample requirements—these results establish ProbeCDN™ as a powerful and generalizable framework for read-level deep learning in NGS-based liquid biopsy, with potential for extension to other molecular assays beyond methylation.

REFERENCES

[1] – K Pettie, S Farashahi, JA Killian, D Kashef, J Hubbell, F Hantash, J Charlton, KI Chacko. Fabrica™: A large-scale data simulation platform isolates tumor signal from cell-free DNA and improves tissue of origin prediction accuracy [abstract]. In: Proceedings of the AACR: Liquid Biopsy. Abstract nr B065.

DISCLOSURES

This study was sponsored by Harbinger Health. Cambridge, MA
JK, KP, SF, JC, KC, and DK are employees of Harbinger Health, Inc.
KG, EB and FH are former employees of Harbinger Health, Inc.

ACKNOWLEDGEMENTS

We gratefully acknowledge all participants for their contributions, without whom this research would not have been possible. We also thank all cross-functional teams for their dedicated efforts that made this work and the resulting data possible.

www.harbinger-health.com