A035

D-Fract Enhances Detection of Tumor-Derived cfDNA Fragments and Cancer Tissue Signal in Liquid Biopsy

Harbinger Health

Shiva Farashahi, PhD; Yifan Wu, MSc; Elie Massaad, PhD; Feras Hantash, PhD; Hutan Ashrafian, PhD; Dorna Kashef, PhD; Kieran I Chacko, PhD
Harbinger Health, Cambridge, MA

BACKGROUND

Cell-free DNA (cfDNA) methylation profiling enables non-invasive early cancer detection and tissue of origin (TOO) localization. In early-stage disease, circulating tumor DNA (ctDNA) comprises ~0.05% of the total cfDNA. Computationally filtering out non-tumor fragments from read-level sequencing data can improve tumor-specific signal, but distinguishing ctDNA from non-tumor methylation patterns is challenging due to biological variability and technical noise. Generative models, particularly denoising diffusion implicit models (DDIMs), are well-suited for cfDNA analysis due to their capacity to model complex data distributions perturbed by structured and stochastic noise. We developed D-Fract, a novel approach that adapts DDIMs to learn non-cancer cfDNA methylation patterns and filter fragment-level data, enhancing tumor-derived signal and TOO inference.

SAMPLES

Blood samples (NCT05435066) were collected from treatment-naive individuals with 9 different cancer types (*N*=401), and individuals with no reported cancer, evenly distributed across ages (23-75 years; *N*=100). Extracted cfDNA was analyzed using a custom targeted bisulfite sequencing hybrid capture assay (18.6 Mb). Tumor types were consolidated into 5 tumor categories and used to evaluate efficiency of D-Fract and TOO classification (**Table 1**).

Tumor categories	All samples	High TF (>0.1%)
Head and Neck	79	54
Liver + Biliary	23	17
Lung	153	78
Panc. + GB.[a]	79	46
Upper GI[b]	67	32
Total	401	227

[a] Panc. + GB.: Pancreas + Gallbladder
[b] Upper GI: Stomach + Esophagus + Gastroesophageal junction

Table 1. Breakdown of samples.

CONCLUSIONS

- We adapted discrete extensions of diffusion implicit models to learn methylation patterns in non-cancer cfDNA.
- We used reconstruction error and data-driven thresholds to identify and retain tumor-derived cfDNA fragments, resulting in substantial increase in estimated tumor fraction.
- We demonstrated that filtering cfDNA samples increases separability in methylation signal between cancer types and improves tissue of origin classification by 9% compared to models trained on unfiltered data, indicating D-Fract's potential to enhance current and future multi-cancer early detection (MCED) diagnostic performance.

D-FRACT DESIGN

Anomaly detection, which focuses on identifying atypical patterns relative to a defined norm, offers a promising approach for detection of tumor-derived signals. D-Fract enriches cancer signal by isolating these fragments using discrete extensions of diffusion implicit models (DDIMs) [1, 2]:

I. Each cfDNA fragment was encoded as a binary sequence representing CpG methylation states. To ensure uniform input length, sequence length was capped at 190 bps, reflecting typical maximum length of bisulfite-treated cfDNA fragments. A second binary channel was used to indicate positions originating from the actual fragment versus padded regions. These binary sequences were transformed into real-valued vectors for input into the model. The genomic start position of each fragment was added as an additional condition to preserve positional context (**Figure 1**).

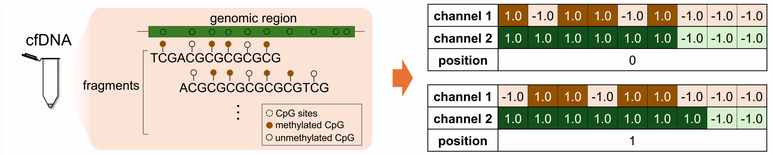

Figure 1. Encoding of cfDNA fragments

II. DDIMs were trained on cfDNA fragments within a single genomic region from non-cancer individuals (*N*=80, **Figure 2**). Continuous diffusion models were trained with 0-thresholding applied at the end to map the continuous outputs back to binary sequences.

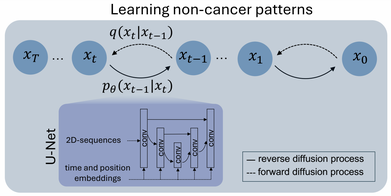

Figure 2. Learning non-cancer patterns in D-Fract.

III. To isolate tumor-derived fragments, each cfDNA fragment underwent partial reverse diffusion (a reduced number of steps, T_l) using the trained DDIM models. Reconstruction error between the original and DDIM-reconstructed fragment was calculated as the anomaly score (**Figure 3**). Tumor-derived fragments, being atypical relative to the learned non-cancer distribution, produce higher reconstruction errors. By applying a threshold derived from held-out non-cancer samples (e.g., 90th of their score distribution, *N*=20), we selectively retain fragments most likely derived from tumor.

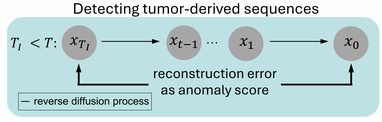

Figure 3. Detecting tumor fragments in D-Fract.

RESULTS

D-Fract increases tumor fraction and signal separability

Filtering cfDNA fragments using D-Fract significantly increased the estimated tumor fraction (TF) across cancer samples, indicating successful enrichment for tumor associated fragments (**Figure 4**). This enrichment was further supported by average 2.5x increase in Euclidean distance between methylation signal across tumor types (**Figure 5**).

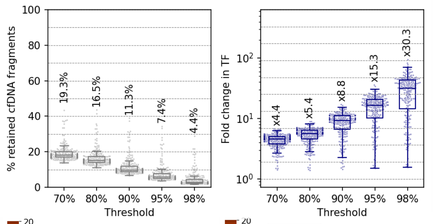

Figure 4. Percentage of retained cfDNA fragments (Left), and estimated TF (Right) of filtered cancer samples is plotted for different thresholds.

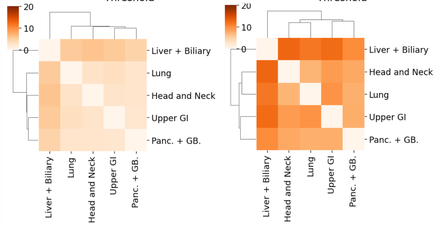

Figure 5. Pairwise Euclidean distances between average methylation across five tumor categories in unfiltered (Left) and D-Fract filtered samples using 90th percentile threshold (Right). The 90th percentile threshold was selected to retain enough cfDNA fragments while enhancing tumor signal.

D-Fract improves tissue of origin prediction

To further demonstrate the utility of fragment-level filtering, we evaluated TOO classification performance before and after applying D-Fract. A 4-layer feedforward neural network was trained on five regional methylation metrics, designed to quantify methylation per genomic region, and assessed via 10-fold cross-validation using plasma cfDNA cancer samples. Filtering input fragments using D-Fract (90th percentile threshold) improved balanced accuracy (BA) from 78% to 87% in samples with tumor fractions >0.1% (*N*=227) (**Figure 6**), and 60% to 66% across all samples (*N*=401). These results indicate that D-Fract enriches for tumor-derived methylation signals that enhance the resolution of TOO prediction.

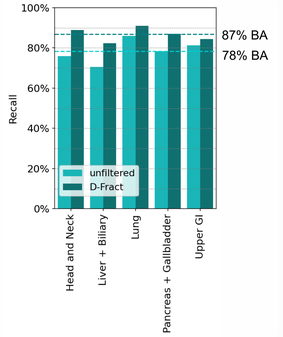

Figure 6. TOO BA and recall.

REFERENCES

1. J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv*:2010.02502, 2020.
2. T. Chen, R. Zhang, and G. Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv*:2208.04202, 2022.

ACKNOWLEDGEMENTS

Editorial support was provided by Sean Husick (Harbinger Health, Cambridge, MA).
We are thankful to the lab team at Harbinger Health for their dedication and expertise in managing the sample preparation.
We gratefully acknowledge all participants for their contributions, without whom this research would not have been possible.

www.harbinger-health.com