Master's in Bioinformatics Thesis

Deep Self-Supervised Modeling of cfDNA Fragmentomics for Tumor Detection

Author:

Ruiqi Zhang

Supervisor:

Jakob Skou Pedersen

Computational Genomics & Transcriptomics Group

Department of Molecular Medicine

AARHUS UNIVERSITY

Jul 2026

Table of Contents

AARHUS UNIVERSITY

Thesis Title:

Deep Self-Supervised Modeling of cfDNA Fragmentomics for Tumor Detection

Thesis Period:

Spring Semester 2026

Author:

Ruiqi Zhang

Supervisor:

Jakob Skou Pedersen

Date of Completion:

June 15, 2026

Master's of Science in Bioinformatics

Bioinformatics Research Center (BiRC)
Aarhus University Denmark

In collaboration with:

Department of Molecular Medicine (MOMA)
Aarhus University Hospital, Denmark

Acknowledgements

First, I would like to extend my sincere gratitude to my supervisor Jakob Skou Pedersen, for providing me with the opportunity to start this thesis and work on this dataset, and for his critical and inspiring discussions with me throughout. I am also greatly thankful to Christian, Mathilde and Camous, for their constructive feedback and essential suggestions that enhanced the quality of my work.

My work is also credited to my classmate Mark with whom I had many meaningful and intuitive discussions. My appreciation goes to all the researchers and staff at the Department of Molecular Medicine and Bioinformatics Research Centre. The supportive environment and available resources have been invaluable to me.

Then, my deep thanks to all who inspired and encouraged me during this project. I would like to extend my innermost thanks to my family for their unconditional support and encouragement. To my parents, thanks is not enough for their unconditional love and for trusting me. I am also grateful to my friends here in Aarhus for their support and for keeping me updated.

Abbreviations

cfDNA cell-free DNA.

cfRNA cell-free RNA.

ctDNA circulating tumor DNA.

mRNA messenger RNA.

PBMC Peripheral Blood Mononuclear Cell.

NCBI National Center for Biotechnology Information

RNA-seq RNA-sequencing.

SGD Stochastic gradient descent.

TSS Transcription Start Site.

WGS Whole Genome Sequencing.

TCGA The Cancer Genome Atlas.

CRC Colorectal cancer.

AUC-ROC Area Under the Receiver Operating Characteristic.

MAE Masked Autoencoders.

NB negative binomial.

MSE Mean Squared Error.

MIL multi-instance learning**.**

SSL self-supervised learning.

Abstract

Cell-free DNA (cfDNA) fragmentomics provides a non-invasive means of studying chromatin organization and cancer-associated biological processes. However, the raw cfDNA fragments data are sparse and high-dimensional, making downstream tasks challenging. Meanwhile, fragmentation patterns surrounding transcription start sites have emerged as a promising source of information for inferring gene expression and chromatin activity. However, the extent to which these signals can be leveraged to learn sample-level representations for downstream tasks, such as cancer detection and tumor fraction estimation, is not yet well investigated.

In this thesis, a self-supervised learning framework inspired by a masked autoencoder was developed to learn representations of cfDNA fragmentomics features surrounding transcription start sites (TSSs) from whole-genome sequencing (WGS) data of a local colorectal cancer cohort. This thesis investigates how well the embedding learned is biologically informative, and can be leveraged for sample-level cancer detection and accurate tumor fraction estimation.

Results

The learned TSS embeddings exhibited clear associations with gene expression levels despite being trained solely on cfDNA fragmentomics features. Visualization of the embedding space revealed progressive transitions across gene expression deciles.

For downstream prediction tasks, sample-level embeddings generated through candidate-gene pooling outperformed embeddings pooled across all genes, with pooling top 50 candidate genes achieving the best tumor classification performance (AUC=0.865). Tumor fraction estimation achieved a Spearman correlation of 0.585, indicating that this approach is not sufficient to achieve accurate tumor fraction estimation.