****

**MBI6013**

**Research proposal**

**Develop the deep learning model for diagnosing cancer through cfDNA analysis.**

Faculty of Medicine, Chinese University of Hong Kong, Shenzhen

Author: <<\* Li Yixiao \*>>

Student ID: <<\* 223050025 \*>>

Advisor: Professor Hao Sun

**Background**

Plasma Cell-free DNA (cfDNA) molecules are DNA found in body fluids such as blood, first discovered in 19811. cfDNA does not contain DNA within a complete cellular structure; it originates from various processes such as cell apoptosis, necrosis, embryonic fluid, extracellular vesicles and nucleic acid fragments, with programmed cell death (apoptosis) and damage-induced necrosis (necrosis) being primary sources. Viral DNA in the blood of virus carriers is also falls into the category of cfDNA.

In healthy people, most cfDNA originates from the apoptosis of normal cells2. Cells during apoptosis will decompose and release fragmented DNA into the surrounding environment, which is mediated by the enzymatic action of cellular nucleases such as CAD. The length of these DNA fragments is uniform, typically ranging between 185bp and 200bp. Some of these fragments enter the bloodstream, becoming cfDNA. However, in cancer patients, cfDNA comes not only from apoptotic cells but also from DNA actively released by tumor cells and fragments released from necrotic tumor cells3. This is due to increased cellular metabolism and cell death rates surpassing macrophage clearance during tumorigenesis, significantly increase in cfDNA levels.

Macrophages in the blood of healthy individuals can clear large DNA molecules resulting from cell fragmentation or cell death. Therefore, cfDNA fragments that are significantly longer or shorter are likely derived from tumor cells. These cfDNA released by tumor cells are called circulating tumor DNA (ctDNA), make up about 0.1% of total cfDNA. Molecular analysis of cfDNA from blood is rapidly becoming an important tool and indicator for cancer diagnosis, detection and prognosis. With an expected 40% increase in cancer cases over the next 20 years, leading to 30 million new cases annually by 2040, mastering cancer prevention and early diagnosis is vital4. Therefore, it is crucial to stay on top of cancer prevention and early diagnosis. Traditional cancer screening methods, such as colonoscopy, have certain limitations and risks in the early detection and identification of cancer. cfDNA early cancer screening has high sensitivity and specificity and can detect even small tumors or early stages. to ctDNA. Compared to traditional tissue biopsies or imaging, cfDNA screening is also more cost-effective.

Because cfDNA can provide a lot of information about an individual's health status, cfDNA analysis has become an important tool for non-invasive diagnosis. In addition to cancer detection, cfDNA is also used for non-invasive prenatal testing (NIPT). Initially, cfDNA research focused on analyzing plasma from pregnant women since fetal cfDNA can be detected in maternal blood. This facilitates non-invasive screening for fetal aneuploidies, including trisomy 13, 18, 21, and sex chromosome abnormalities.

**Significance**

For cfDNA from different sources, Next-Generation Sequencing (NGS) and nanopore sequencing, combined with bioinformatics analysis, have different clinical application directions. Despite remarkable achievements in translational medicine, factors like psychosocial elements, physical activity, and stress conditions can affect cfDNA release5, and its origin and molecular characteristics are not fully understood. And because the amount of ctDNA released by tumors into the blood is very small, the detection of early tumors is still difficult. Whole-genome sequencing (WGS) of cfDNA can identify chromosomal abnormalities in cancer patients, but detection of such changes is challenging due to the limited number of abnormal chromosomal changes. A 2018 study showed that size selection of small cfDNA can enrich ctDNA6, and that transcription start sites, nucleosome positions, and cfDNA terminus motifs may be altered in cancer, but the sequencing required for these assays is difficult for routine analysis. It is also costly7. Obtaining ample samples for traditional sequencing analysis is difficult for researchers, with many datasets being confidential or ethically sensitive. Hence, there's a need for a new pipeline to classify cfDNA for cancer diagnosis.

Detecting more mutations within the genome may improve the sensitivity of detecting ctDNA in the blood circulation system8. Therefore, the possibility of using deep learning models to classify cfDNA for cancer diagnosis has become a hot research topic in recent years. In particular, deep learning algorithms represented by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are very good at identifying patterns from complex clinical data, and these patterns are often specific to ctDNA. In theory, CNN is good at proximal feature representation, and RNN can find long distance dependencies. By combining these technologies with third-generation sequencing technology, the deep learning model can use the short-range and long-range information hidden in the original electrical signals of nanopore sequencing to perform effective classification9. Deep learning models can integrate data from different sources, such as mutation information from cfDNA, DNA methylation patterns, copy number variations (CNVs), and other epigenetic information, to improve the accuracy of cancer diagnosis. Through training with a large amount of data, deep learning models can reduce the false positive rate while maintaining high sensitivity, which is especially important for early cancer screening8. Therefore, this project aims to develop a deep learning model for diagnosing cancer through cfDNA analysis.

**Research target**

This project focuses on developing a pipeline for classifying cfDNA to diagnose cancer based on deep learning models. After building and training a deep learning classifier with high accuracy, low false positives, high sensitivity, high recall, and good versatility for various cancers, deploy the model code and all necessary dependency packages to Docker image. This image is then uploaded to a container registry for deployment.

Users can pull the Docker image onto a cluster to launch container instances containing the classifier, thereby distributing and running the model cross multiple nodes of the cluster to complete the cancer classification task of their own clinical instances. This approach improves the deployment efficiency of the designed toolkit, ensures environmental consistency for large-scale model application, and simplifies the running and management process on the cluster. The ultimate goal is to guide users to upload data to the GitHub website to run the image of the classifier. The classified results will also be automatically stored on the GitHub website for users to download, which will not occupy the user's local server computing power. Using different sequencing depths and sequencing methods will result in significantly different sizes of cfDNA sequencing files. For example, using whole-exome sequencing (WGS) with about 40x coverage and 75bp read length, the BAW file of one sample will be approximately 5.7GB, which requires substantial computational server capacity for multi-sample analysis, hence the obvious advantages of distributed computing.

**Data collection**

As previously mentioned, classifying cfDNA to diagnose cancer has become a research hotspot in recent years. This project will collect data from published papers in advance. Papers published in journals will store data according to unique IDs to major database websites, such as dbGap, Gene Expression Omnibus (GEO), National Genomics Data Center (GSA) and Encylopedia of DNA Elements project (ENCODE). These public databases can be accessed via the National Library of Medicine or the National Center for Biotechnology Information (NCBI) by entering accession numbers to download original data. The code of these papers will also be uploaded to the GitHub website simultaneously, and some preprocessed data can also be found on the GitHub website.

For public cfDNA sequencing data downloaded from GEO, the original data is stored in SRA, and the processed data is stored in GEO. This project uses SRA Toolkit software to download the original paired-end sequencing files. The official version of this software can be downloaded from NCBI's SRA Toolkit website. For the downloaded original file, convert it into FASTQ format and FAST5 format (ONT sequencing). The FAST5 file includes the raw electrical signal data of sequencing, sequencing event data and other metadata about the sequencing process, and FAST5 supports real-time sequencing analysis and Long-read sequencing technology therefore has important application value in fields such as genomic variant identification and can be used for DNA methylation analysis. For the WGBS data set of cfDNA downloaded from GSA, the key information of DNA methylation level is recorded10, and the SRA ToolKit software can also be used for download operation. For data downloaded from dbGap, you need to register an account to access some restricted data sets (such as sensitive personal health information) and apply for access permissions for downloading, which will take a lot of time11. Finally, the steering group of this project also has some FAST5 sequencing files of cfDNA, including samples from healthy people and samples from cancer patients.

**Data Preprocessing**

**The Fragmentation Pattern Classifier**

Align the binary sequence alignment/mapping file (BAM file) of the sequencing file to the reference sequence and filter the data by excluding duplicates, secondary and unmapped alignments, and save the read paired sequence alignment objects into a file in RDS format. Only the data of chromosomes 1 to 22 are retained, and fragments with extreme lengths are deleted. Calculate the GC content of each fragment and correct for GC bias. Filter A/B partition data to match chromosome arms removing centrioles and telomeres and delete overlapping regions. Each sample is grouped by chromosome arms, and the location of the arms determines how to merge the segments to form larger 5mb segments (bins). Calculate multiple statistics for each 5mb segment, such as short and long coverage, corrected coverage, hic feature value, GC content, etc.11.

**The DNA Methylation Classifier**

The DNA methylation classifier divides the genome into 500bp bins, filtering out areas in the training samples with fewer than 25 reads aligned to each bin. Integrating the WGBS sequencing data of cfDNA and methylation information, for each read, the first 5 bp of the 5' end was removed to prevent the influence of sequencing adapters, and then all reads were taken to the same length from the 3' end.

**Research Design and Methods**

Cancer cells can be identified based on the unique fragmentation pattern of cfDNA, as the size and number of DNA fragments in different regions vary significantly. Different regions of the genome of healthy cells are carefully placed into the nucleus so that the cfDNA fragments appear to be relatively evenly distributed in length. The genome of cancer cells is disordered in packaging. When cancer cells die, DNA is released into the blood in a chaotic manner11.

The reference information of the sample is compared with the preprocessed 5mb chromosome segment data, and the median of different coverages and ratios is calculated, as well as the median of the corrected coverage and ratios. Calculate the correlation between the cancer sample and the median of the healthy sample. Then calculate the coverage, pattern size, mean size, median size, 25% and 75% quantile size and the number of high-quality bases analyzed and coverage for each sample.

A machine learning classification model of gradient boosting machine (GBM) is constructed to distinguish between two types of samples: healthy and cancer. GBM is an ensemble learning method. Convert the continuous 5mb interval data to wide format, calculate the number of corrected fragments and the corrected number of short fragments for each sample, and then perform normalization processing. Create a matrix with total segment length and short segment data as features and use repeated cross-validation method to train multiple GBM models, including models using only specific features and all features respectively. An evaluation of the model classification results is performed, and specific specificity thresholds are calculated.

DNA methylation acts as a regulator of cfDNA length, and different methylation levels will determine the different sites where DNA is cleaved in nucleosomes10. Therefore, it is also possible to diagnose cancer based on DNA methylation information using deep learning models. Calculate the methylation rate of all DNA fragments in a certain region, obtain the methylation pattern in cfDNA of cancer samples and healthy samples, and screen the WGBS data of cfDNA for reads in differentially methylated regions. Compare the maximum and minimum methylation rates of the two methylation patterns. Construct a deep learning model to predict the possibility of reads coming from tumor tissue13. The modified DanQ model is used as the core of the deep learning model. This is because the DanQ model is a model specially designed to automatically learn large-scale functional prediction from DNA sequences. The DanQ model first identifies basic patterns in DNA sequences through CNN. BiLSTM then parses the long-distance relationships between these patterns, thereby improving the recognition and interpretation capabilities of sequence features14.

**Reference**

1. Mandel, P. & Metais, P. Les acides nucléiques du plasma sanguin chez l’homme. C. R. Seances Soc. Biol. Fil. 142, 241–243 (1948).
2. Suzuki N, Kamataki A, Yamaki J, et al. Characterization of circulating DNA in healthy human plasma[J]. Clin Chim Acta, 2008, 387(1-2): 55-58.
3. Li CN, Hsu HL, Wu TL, et al. Cell-free DNA is released from tumor cells upon cell death: a study of tissue cultures of tumor cell lines[J]. J Clin Lab Anal, 2003, 17(4): 103-107.
4. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71:209–49.
5. Hummel, E. M. et al. Cell-free DNA release under psychosocial and physical stress conditions. Transl. Psychiatry 8, 236 (2018).
6. Mouliere F et al. Enhanced detection of circulating tumor DNA by fragment size analysis. SciTransl Med 10, doi:10.1126/scitranslmed. aat4921 (2018).
7. Jiang P et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci U S A, doi:10.1073/pnas. 1814616115 (2018).
8. Zhitynyuk, Y. V., Koval, A. P., Alferov, A. A., Shtykova, Y. A., Mamedov, I. Z., Kushlinskii, N. E., Chudakov, D. M., & Shcherbo, D. S. (2022). Deep cfDNA fragment end profiling enables cancer detection. Molecular Cancer, 21(1), Article 26. <https://doi.org/10.1186/s12943-022-01462-5>
9. Danilevsky, A., Polsky, A. L., Alferov, A. A., Shtykova, Y. A., Mamedov, I. Z., Kushlinskii, N. E., Chudakov, D. M., & Shcherbo, D. S. (2022). Adaptive sequencing using nanopores and deep learning of mitochondrial DNA. Briefings in Bioinformatics, 23(4), Article 26. <https://doi.org/10.1093/bib/bbac251>
10. An, Y., Zhao, X., Zhang, Z., Xia, Z., Yang, M., Ma, L., ... & Sun, K. (2023). DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. Nature Communications. <https://doi.org/10.1038/s41467-023-35959-6>
11. Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., ... & Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 570(7761), 385–389. <https://doi.org/10.1038/s41586-019-1272-6>
12. Zhang, W., Wei, L., Huang, J., Zhong, B., Li, J., Xu, H., He, S., Liu, Y., Liu, J., Lv, H., & Wang, X. (2021). cfDNApipe: A comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data. Bioinformatics, 37(22), 4251–4252. <https://doi.org/10.1093/bioinformatics/btab413>
13. Li, J., Wei, L., Zhang, X., Zhang, W., Wang, H., Zhong, B., Xie, Z., Lv, H., & Wang, X. (2021). DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads. Briefings in Bioinformatics, 22(6), bbab250. <https://doi.org/10.1093/bib/bbab250>
14. Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 44(11), e107. <https://doi.org/10.1093/nar/gkw226>