Bioinformatic analysis of DNA methylation patterns in cancer: potential applications in cancer detection

Meghan Sleeper

1. Explanation of work

This project will contribute to the rapidly developing fields investigating DNA methylation in cancer development, and non-invasive biomarkers for cancer detection. DNA contains segments called genes that serve as instructions for making proteins with specific functions. While cells throughout the human body contain the same DNA sequence, not all genes are expressed in each cell. As shown in **figure 1**, DNA methylation regulates which genes are expressed without changing the underlying DNA sequence.[11] Locations in DNA (loci) that are methylated vary by cell type, which drives cellular identity. Just as the sequence of the human genome has been mapped, studies have mapped the loci of DNA methylation by cell type.[2,3] Abnormal DNA methylation occurs in cancer cells and is thought to play a role in the development of cancer.[4] Comparative analysis of the methylated loci in genomes of patients with and without cancer can further our understanding of DNA methylation's role in cancer.

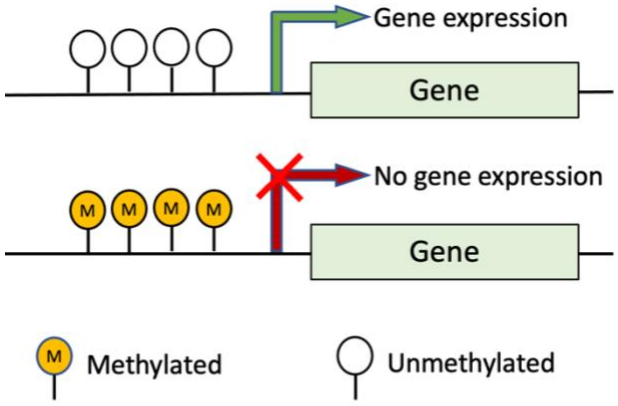

Figure 1. Simplified diagram of DNA methylation repressing gene expression. The solid line represents DNA.

In this project, I will identify genes within differentially methylated regions (DMRs) of DNA in individuals with cancer by analyzing DNA methylation of circulating cell-free DNA (cfDNA). Circulating cfDNA is fragmented DNA in blood plasma that has been released from cells after cellular death. Circulating cfDNA originates from cell types throughout the body including tumor cells in cancer patients.[1] cfDNA can be obtained through non-invasive blood sampling making it an ideal target for research, and discovery of cancer-related biomarkers. Various cfDNA methylation datasets are accessible through NCBI Sequence Read Archive (SRA). To accomplish the goals of this project, I will use Whole Genome Bisulfite Sequencing (WGBS) datasets to investigate the methylation of cfDNA sampled from multiple individuals with lung, colorectal, breast, pancreatic, or prostate cancer **(figure 2)**.

Cloud storage and computing services will be required to analyze WGBS datasets. Initially, I will test the functionality of the methods locally using small datasets. After I have verified that the methods function correctly, I will copy WGBS datasets from NCBI SRA to Amazon Web Services (AWS) Simple Storage Service (S3) buckets where more intensive computing will be carried out using S3 Object Lambda Computing services. To compress data into a workable format, I will convert the cfDNA WGBS datasets to beta-formatted files.[2] To

determine the originating cell type for each cfDNA fragment in a sample, I will reformat the beta files into compatible array-style data and perform deconvolution. Deconvolution is the process of sorting cfDNA by cell type.[2,3] Next, I will align cfDNA fragments originating from the cancer-associated cells with the healthy human methylation profile and check for differentially methylated regions (DMRs). For DMR analysis, I will first segment cfDNA sequence data into similarly methylated chunks and use an algorithm to check segments for DMRs.[2] If DMRs are statistically significant (p-value < 0.05), genes in the associated region will be identified.

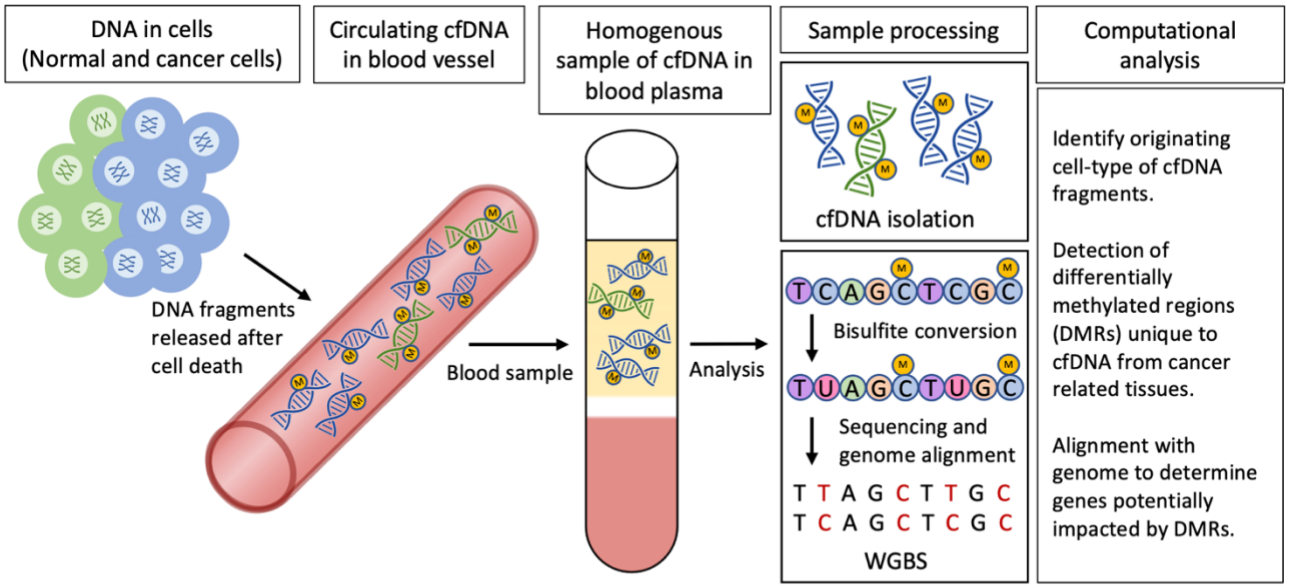

Figure 2. Overview of sampling, processing, and computational analysis of methylation in cell-free DNA (cfDNA) from individuals with cancer. Yellow circles denoted with M represent methylated loci in the DNA. **For this project, I will carry out the computational analysis portion of this figure on datasets from other studies.**

I will analyze data sampled from individuals with cancer and map their DMR loci to determine genes potentially impacted by abnormal methylation. I expect to find DMRs associated with genes already implicated in cancer formation and identify novel genes impacted by DMRs. **The results of this project will provide insight into the abnormal DNA methylation patterns that occur in cancer and the impact on gene regions.** This information could be applied to define new cfDNA biomarkers for non-invasive cancer detection.

2. Student contribution to project design and execution

I designed this project based on research in the field of DNA methylation and cancer. I find it fascinating that cfDNA methylation has been identified as a potential indicator of cancer and want to contribute to this field. The vast amounts of scientific data available for further analysis motivated me to plan a project utilizing data that has already been collected. This study involves comprehensive bioinformatic analysis of multiple datasets which will be carried out by me with guidance and feedback from my advisors.

3. Broader impact of the work

Cancer has a wide-reaching impact and is the second most common cause of death in the US. Changes in DNA methylation have been shown to play a role early in the development of cancer.[4] By investigating DNA methylation in cancer cases, scientists can better understand the role of DNA methylation in cancer formation. By compiling a list of abnormal methylation in various cancer cases, I hope to identify methylated loci that are unique to specific cancers. This knowledge could assist in the development of new biomarkers for detecting cancers or even new treatment methods.

4. Expected benefits of the award to the student

I want to pursue a career as a bioinformatic scientist working with genetic data. To qualify for jobs in this field, I will need to gain experience analyzing high-throughput sequence data. This task will require cloud storage and processing. This award will allow me to complete my thesis project and, in the process, learn how to manage large biological datasets in a cloud computing environment.

5. Itemized and detailed budget

Item	Monthly cost	10-month cost
Amazon S3 Service		
Storage for 3TB	$79.87	$798.70
S3 Object Lambda Computing	$3.00	$30
Total cost	$82.87	**$828.70**

It is necessary to use cloud storage and computing to accomplish the goals of this project due to the magnitude of the data being processed. SRA uses cold data storage requiring data transfer directly into cloud storage. The size of the data files transferred will range from 10-15Gb per replicate (Ex. SRA #SRR8117446) and up to 3Tb per collection of experimental replicates (Ex. BioProject #PRJNA494975).

Data transfer and processing will begin in February 2023 and continue until December 2023 for a total of 10 months. This will allow time for various new files to be transferred into storage and processed throughout the year. For efficient and timely computing, I will use S3 Object Lambda Computing services. Amazon S3 costs $0.039 per GB per month. I calculated the total estimated service cost using the AWS pricing calculator.

Literature cited

1. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).

2. Loyfer, N. *et al.* A human DNA methylation atlas reveals principles of cell type-specific methylation and identifies thousands of cell type-specific regulatory elements. 2022.01.24.477547 Preprint at https://doi.org/10.1101/2022.01.24.477547 (2022).

3. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* **9**, 5068 (2018).

4. Wajed, S. A., Laird, P. W. & DeMeester, T. R. DNA Methylation: An Alternative Pathway to Cancer. *Ann Surg* **234**, 10–20 (2001).