UCLA

Alignment-free colorectal colon cancer classification approach using cell-free DNA

PRAVEENA RATNAVEL[1], Lily Zello[1], Fei-man Hsu[2], Matteo Pellegrini[2]

1. BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2. Department of Molecular Cell and Developmental Biology, UCLA

ABSTRACT

Previous research has shown that DNA methylation signatures in cell-free DNA (cfDNA) can effectively classify cancer patients with high specificity. The conventional method first maps sequencing reads to a reference genome and then applies comprehensive bioinformatics analyses. However, this pipeline is computationally demanding. Here we use a publicly available colorectal cancer MeDIP-Seq dataset to assess an alignment-free classification technique utilizing k-mer counting[2] and compare it to the traditional alignment-based method. Our findings suggest that the alignment-free approach reduces computation time and resource usage while maintaining accuracy. These results indicate that k-mer counting could be more feasible and enable quicker diagnosis in healthcare settings.

BACKGROUND

DNA methylation signatures in cell-free DNA have promising clinical applications to classify tumors as either cancerous or non-cancerous through an alignment pipeline using MeDIP-Seq data[1,3] (Figure 1). MeDIP-Seq is derived from cell-free DNA; the plasma provides the clinical advantage of being easily accessible when tissue samples are insufficient or unavailable. Plasma cfDNA is also particularly relevant to studying distinctive DNA features of tumors, as the rapid cell growth and death that accompanies most cancers will increase the plasma levels of cfDNA. MeDIP-Seq involves immunoprecipitation of CpG methylated DNA, which are then typically aligned to the genome.

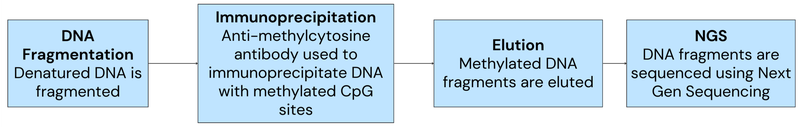

Figure 1: MeDIP-Seq Flowchart

The MeDIP-Seq data that we used included **colorectal cancer patients (CRC, n=30)** and **non-cancer control (NC, n=33)**. The patients are of varying adult ages, are distributed by gender, and the CRC patients have various stages of cancer (Figure 2).

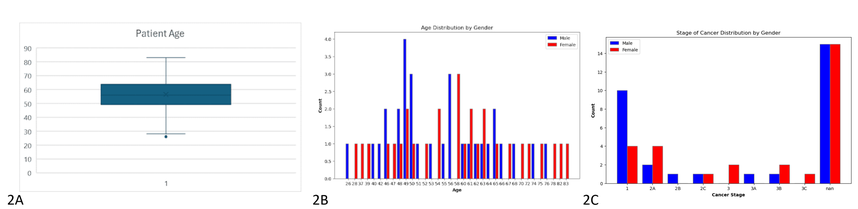

Figure 2: Characteristics of all 63 CRC and NC patients. 2a) Boxplot of patient age. 2b) Distribution of age by gender. 2c) Distribution of stage of cancer by gender

APPROACH & METHODOLOGY

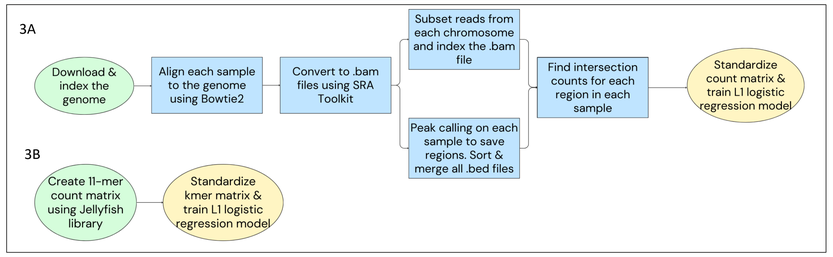

Figure 3: Methodologies flowcharts for 3a) alignment and 3b) alignment-free method

The CRC and NC reads aligned to the human genome with relatively high mappability scores (Figure 4), indicating that the reads were of high quality and reliability. The alignment pipeline (Figure 3a) involves aligning the sequenced reads to the human genome before applying peak calling to compute genomic region counts for each sample. The alignment-free pipeline (Figure 3b) incorporated the Jellyfish library in order to create a matrix of 11-mer counts. In order to visualize the matrices and portray any clustering, we used dimensionality reduction to produce UMAPs of both the alignment and alignment-free data (Figure 4).

RESULTS

Sequencing data pre-processing

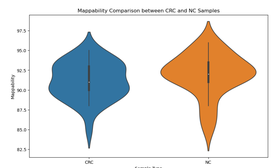

Figure 4: Violin plot of mappability scores from aligned MeDIP-seq data.

Dimensionality Reduction

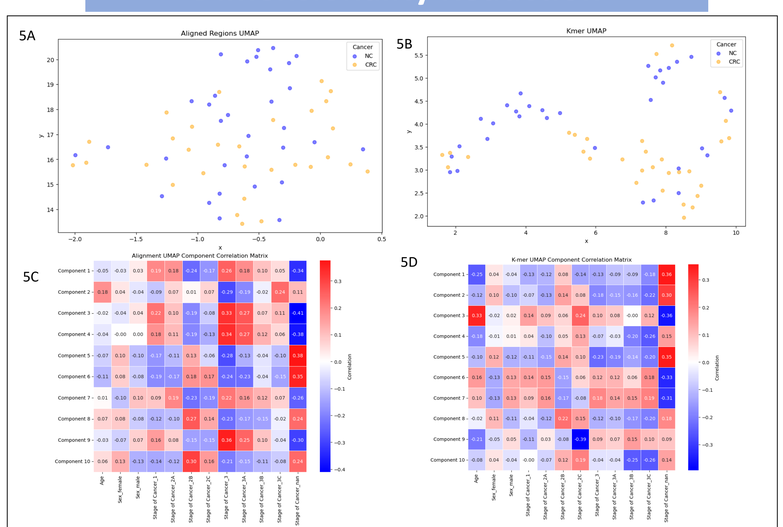

Figure 5: UMAP of NC and CRC samples from 5a) standardized aligned region counts and 5b) standardized k-mer counts.
Correlation matrix of top 10 UMAP components and patient characteristics from 5c) aligned data and 5d) alignment-free data.

Computing Resource Comparison

Alignment Pipeline

Step	Time (hours)
Download & Index Genome	2
Bowtie2 align reads to genome & convert to .bam	1.75/sample → 110.25 total
Extract chromosome reads & index .bam	1
Peak calling & merge .bed files	4.5
Intersect counts	2
Standardize data & train model	5
Total	124.75

Alignment-Free Pipeline

Step	Time (hours)
Use Jellyfish to create k-mer count matrix	0.75
Standardize counts & train model	4.3
Total	5.05

Figure 6: Alignment and alignment-free pipeline timings

Classification Model Comparison

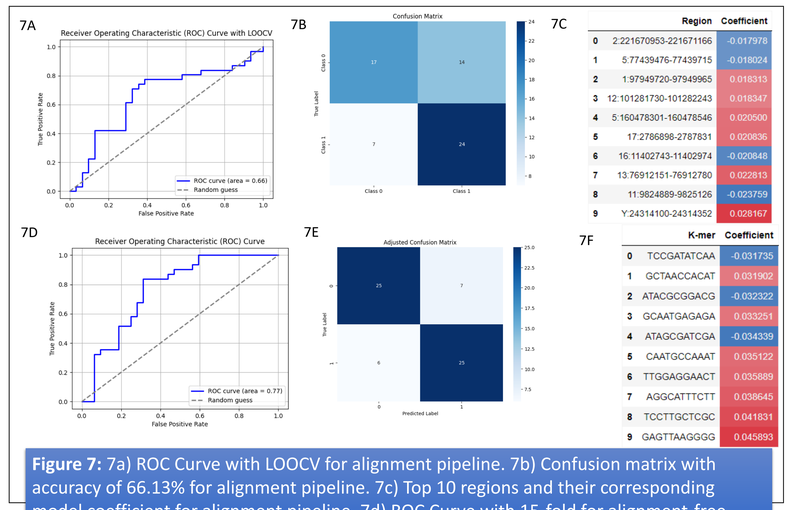

Figure 7: 7a) ROC Curve with LOOCV for alignment pipeline. 7b) Confusion matrix with accuracy of 66.13% for alignment pipeline. 7c) Top 10 regions and their corresponding model coefficient for alignment pipeline. 7d) ROC Curve with 15-fold for alignment-free pipeline. 7e) Confusion matrix with accuracy of 79.36% for alignment-free pipeline. 7f) Top 10 k-mers and their corresponding coefficient for alignment-free pipeline.

CONCLUSIONS

- The alignment-free method **increased accuracy by a factor of 20.00%, or 13.23%** and **reduced computation time by 96%, or 119.7 hours**.
- Estimating LOOCV in the alignment-free method would result in a **reduction of 105.94 hours, or 85%**.
- Fine-tuning of the L1 regression model, as well as exploration of various other types of classification techniques, will have monumental implications in clinical settings, where cfDNA can be used in conjunction with these machine learning models to assess risk and diagnose patients with colorectal colon cancer.

REFERENCES

1. Ayub, A. L. P., Perestrelo, B. de O., Pessoa, G. C., & Jasiulionis, M. G. (2022, August 19). *Useful methods to study epigenetic marks: DNA methylation, histone modifications, chromatin structure, and noncoding RNAS*. Useful methods to study epigenetic marks: DNA methylation, histone modifications, chromatin structure, and noncoding RNAs. https://www.sciencedirect.com/science/article/abs/pii/B9780323910811000121
2. Thomas, A., Barriere, S., Broseus, L., Brooke, J., Lorenzi, C., Villemin, J.-P., Beurier, G., Sabatier, R., Reynes, C., Mancheron, A., & Ritchie, W. (2019, June 20). *Gecko is a genetic algorithm to classify and explore high throughput sequencing data*. Nature News. https://www.nature.com/articles/s42003-019-0456-9
3. Yong, W.-S., Hsu, F.-M., & Chen, P.-Y. (2016, June 29). *Profiling genome-wide DNA methylation - epigenetics & Chromatin*. BioMed Central. https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-016-0075-3