Quantifying complementarity between different cfDNA features

Detection of cancer using blood

Amr Farooq
Supervisor(s): Dr.S.Makrodimitris, I.B. Pronk, D.M. Hazelaar
Responsible Professor: Prof.dr.ir.M.J.T. Reinders
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Amr Farooq
Final project course: CSE3000 Research Project
Thesis committee: Prof.dr.ir.M.J.T. Reinders, Dr.S. Makrodimitris, I.B. Pronk, D.M. Hazelaar, Dr.ir. J.A. Pouwelse

Abstract

Recent research has indicated attributes of cell-free DNA (cfDNA) called fragmentomics as a promising method for late stage cancer detection in a non-invasive manner. The primary objective of this research is to uncover hidden patterns and interactions that could enhance the accuracy and sensitivity of blood-based cancer diagnostics. This study explores the complementarity between three fragmentomics features; fragment length distribution, and nucleotide fragment end sequence diversity and nucleosome positioning for four different sample groups; breast cancer, colorectal cancer, lung cancer and healthy controls. Various machine learning techniques such as linear regression were employed to quantify any complementary relationships between the features.

1 Introduction

Accounting for approximately 20% of all deaths in 2020 [11], cancer is a complex disease affecting millions around the world. Unfortunately, once someone is diagnosed with late stage cancer, treatment options are limited. Options may focus on extending lifespan, controlling tumor growth, or simply alleviating symptoms to improve quality of life [2]. This is further worsened by the large financial burden placed on cancer patients and, their loved ones. In 2021, about 286 billion dollars were spent on getting treatment for cancer patients, a number which is expected to increase to 581 billion dollars in 2030 [11]. Given cancer's tricky nature, the best countermeasure is detecting it at earlier stages. Unfortunately, this presents many challenges in itself.

Conventional screening tests such as biopsies, where a sample of abnormal tissues is removed from the patient to confirm cancer [1]. This is extremely invasive and done with a single cancer type in mind, which could result in high false positives rates when used sequentially [11]. Fortunately, recent research has produced non-invasive diagnostic methods that are suitable for large-scale screening.

One such method are liquid biopsies where bodily fluids are extracted and analyzed for the presence of cancer in patients. Due to their minimally invasive nature, liquid biopsies have emerged as a promising avenue for cancer detection and monitoring that can be performed frequently with negligible burden on patients. The study of DNA fragments in blood (taken using liquid biopsies), is known as fragmentomics. According to Chao Li et al [13], research has indicated that fragmentation characteristics of cell-free DNA (cfDNA) differ in healthy and diseased individuals. As an example, patients with cancer had altered fragmentation profiles compared to healthy individuals whose profiles reflected nucleosomal patterns of white blood cells [6]. These findings highlight the potential of cfDNA fragmentomics as a novel biomarker for cancer detection and monitoring.

This study aims to explore the complementarity of various fragmentomics features. Feature complementarity or interaction between features [9], can provide vital insights into inner workings of features. The aim is not to bundle together multiple fragmentomics features in an ensemble model for classifying cancer versus healthy, but rather researching how machine learning can be leveraged to quantify the complementarity between features. This underlying information was previously neglected when trying to create a multi-feature model for classification, finding the best combination of features could aid in achieving a major goal in the field of cfDNA fragmentomics for cancer research; defining robust Multi-cancer early detection (MCED) tests

that delivers a screening approach with high sensitivity, specificity, and Tissue of Origin (TOO) identification accessible to the general public, providing better clinical outcomes and treatment opportunities [11].

Chapter 2 provides a comprehensive review of the most relevant existing works, explaining their contributions, and indicating what is still unanswered. Next Chapter 3, gives a detailed explanation of the methodology and experimental setup. It delves into the framework used for answering the research question; narrating the compartmentalization of the main research question into smaller, more manageable sub-problems, briefly sketching the algorithms, and models used to answer the sub-problems, and elaborate on key design decisions made throughout the research process. Based on the setup from Chapter 3, chapter 4 will discuss the results of the conducted experiment using the proposed approach(es) along with shortcomings and recommendations for future research. Chapter 5 will conclude the paper, with chapter 6 offering an insight into the ethical considerations of the research and discusses the reproducibility of the methods employed.

2 Synthesis of published research

Related works on the early diagnosis of cancer where cfDNA fragmentomics characteristics were integrated using multiple machine-learning models produced promising results [13]. One example of this is the multi-modal approach known as SPOT-MAS (Screening for the Presence of Tumor by DNA Methylation and Size) defined by Van Thien Chi Nguyen et al [11]. This approach was designed with the intent of performing analysis on methylomics, fragmentomics, DNA copy number, and end motifs of cfDNA and assess the combined potential of these fragmentomics features for a single, comprehensive cancer screening test capable of both detection and localization. SPOT-MAS was able to successfully detect five different types of cancer in their early stages and predict the tumor locations. In current studies on the use of cfDNA fragmentomics as a biomarker for the detection of cancer, much effort is placed on finding the best classification model. Studies like SPOT-MAS aimed to integrate multiple features to improve the cancer detection rate and identify tissue of origin.

A similar experiment by Halner et al [7] proposed the DEcancer framework, where various machine learning techniques were used to effectively select robust features from liquid biopsy samples to accurately detect cancer with minimal false positives or negatives. The objective was to streamline the process of detecting cancer by pinpointing the most essential set of features that best predict its presence. This involved developing a machine learning pipeline that utilizes feature selection methods and multiple data augmentation techniques to achieve feature selection and high cancer detection performance [7].

Whilst, the studies mentioned provide valuable insight into the future of cfDNA fragmentomics as a biomarker and highlight the potential of blood-based cancer detection tests to become a universal, simple, and cost-effective method for early multi-cancer detection in a large populace [11], they do not consider a very important aspect; feature complementarity. This paper aims to explore the complementarity of various fragmentomics features and how it can aid cancer detection.

3 Methodology and Experimental Setup

This chapter aims to provide a high-level overview of the methodology and experimental setup for this study. The research was split into three main phases: identification, processing, and evaluation. Throughout the three phases, the research question is compartmentalized into smaller sub-tasks to provide more structure in answering the original question.

3.1 Identification

In the identification phase, it was crucial to select which fragmentomics features such as the length distribution of cfDNA fragments [6] would be used for the experiment. These features had to be easily extracted from the raw data provided and available across the different sample types; Breast cancer (BRCA), Colorectal cancer (CRC), Lung cancer (LUAD), and healthy controls. Additionally, this phase involved determining the method for extracting these features from the data. Typically, fragmentomics features are calculated from patient reads by analyzing the entire genome or dividing it into non-overlapping bins.

For this experiment, three feature types were selected; The $\log_2(short/long)$ ratio of the fragment lengths [6], 5' trinucleotide fragment end sequence diversity [10], and nucleosome positioning patterns [14]

3.1.1 The short-long ratio of the fragment lengths

The fragment length ratio is calculated as:

$$\text{ratio} = \log_2\left(\frac{\text{short\_count}}{\text{long\_count}}\right)$$

short_count = number of short fragments (100-150 bp)
long_count = number of long fragments (151-220 bp)

The short to long ratio of cfDNA fragments defined by Cristiano et al [6], is a frequently applied fragmentomics feature used for enhancing ctDNA detection. This ratio effectively distinguishes tumor-derived fragments from those originating from healthy cells. For instance, Nguyen et al in their paper [12] demonstrated how short-long ratios aided in distinguishing cancer patients from healthy controls. Their research revealed a higher prevalence of shorter DNA fragments (<151 bp) in the plasma of cancer patients compared to healthy individuals, which is in line with other research done on this feature. All ratios were standardized using z-scores.

3.1.2 The 5' trinucleotide fragment end sequence diversity

As described by Moldovan et al [10], the 5' trinucleotide fragment end sequence diversity is calculated for every input sample as the Gini index using the formula:

$$G = 1 - \sum_{i=1}^{64} P_i^2$$

where P_i is the frequency of a specific i trinucleotide endings. DNA is comprised of four building blocks called nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). For this

feature, the frequency of different trinucleotide ending e.g. GTA, ATC, etc... found in the sample are counted and the Gini index is computed to quantify the diversity. CfDNA fragment end profiling (cfDNA-FEP), which can reveal cancer-specific fragment end sequences has shown promising results. For instance, Zhitnyuk et al [16] demonstrated that deep profiling of cfDNA fragment ends could aid in the detection of colorectal and renal cancers.

3.1.3 Nucleosome positioning patterns

Zhang et al [14] wrote, cfDNA is believed to derive predominately from apoptosis of normal cells of the hematopoietic lineage in healthy samples. However, in the case of cancer patients; cfDNA can also be released from tumor cells. The patterns of nucleosomes are spaced in cancer samples reveal additional contributions to cfDNA that correlate most strongly with non-blood tissues. tissues often matching the anatomical origin of the cancer. The nucleosome patterns can be observed using a numerical value known as The Windowed Protection Score (WPS). The WPS is defined as the difference between the number of DNA fragments completely spanning a 120 bp window centered at a given genomic coordinate and the number of fragments with an endpoint within that same window [14].
Synder et al [14] calculated the WPS for each base pair position across the genome. To maintain consistency between features, slight modifications were necessary to compute the WPS for each bin. After dividing the genome into 5Mb bins, a sliding window of 120 bp centered at each base pair within each bin was used. For each window, the number of fragments that completely span the window (start before and end after) and have an endpoint within the window are counted. The endpoint count is subtracted from the spanning count to get the WPS for that position. Finally, the WPS values are averaged across all positions within each bin to get one score per bin. A pseudo-code implementation can be seen in supplementary figure A2

3.2 Processing

Having identified which features to extract, the next phase is processing them. For each sample; BRCA(n=45), CRC(n=23), LUAD(n=75), and control(n=103), the genome was divided. Each chromosome was partitioned into bins of 5-megabase (Mb) windows to evaluate cfDNA fragmentation patterns [6]. Furthermore, reads were filtered to include only properly paired reads that are mapped, not secondary alignments, and have a mapping quality of 20 or higher, ensuring reliable feature extraction. This dissection of the genome into bins gives far more features per type for each sample. In supplementary figure A1, a pseudo code implementation is given to illustrate how for each sample the genome is divided and values for the feature (per bin) are saved. Values for chromosomes Y were excluded as they do not pertain to the use-case of this study, where patient data is being analyzed irrespective of sex, whilst chromosome X after the research supervisors advised due to the chromosome being difficult to map.

Having collected the features for each sample, it was vital to find the most appropriate manner to combine them for all patients for the same sample type to expand the feature space. To this end, for each feature type, four feature matrices (one for each sample type) were created with each column representing feature values for specific 5 million base pair segments across each chromosome from chromosome 1 to 22. Each feature value column corresponds to a specific genomic range within a chromosome, indicating the values for that segment, and the first column represents the identifiers of the sample in that sample group. A sample representation is provided in the table 1 with pseudo code available in appendix A3.

Sample Name	Chr1 0:5000000	Chr1 5000000:10000000	...	Chr22 ...
Sample 1	value 1	value 2	...	value n
Sample 2	value 3	value 4	...	value m
Sample 3	value 5	value 6	...	value o
...	...	...	...	...
Sample N	value x	value y	...	value z

Table 1: Representation of Feature Space Matrix

3.3 Evaluation

Finally, the evaluation phase involved selecting specific metrics from the processed data to assess the complementarity of the identified features.

3.3.1 Linear Regression

Linear regression was employeed to observe whether it was possible to predict one feature type from another. Using the R-squared score [8] to measure the strength of the relationship between two features, and their potential complementarity (a higher score indicating the two features are highly correlated, i.e. they carry similar information). Linear regression and the R-squared score provide an interesting insight into the computing the complementarity between different features. The R-squared score measures the overall model fit; a high score implies the model explains the data well, but it does not provide great insights about the relationship between two features. Hence, this metric can only serve as one part of the complementarity analysis and not its sole measure.

3.3.2 Multi-Omics Factor Analysis (MOFA)

Another algorithm is Multi-Omics Factor Analysis (MOFA+). MOFA+ is a framework designed for large-scale datasets with complex experimental designs that include multiple groups of features and multiple sample groups [5]. MOFA+ exploits the dependencies between the features to create a simplified representation of the larger dataset defined by multiple latent factors. These factors capture the global sources of variability in the data [4]. Each factor has weights that highlight how important each feature is in determining the factor's value. MOFA+ can use these factors to determine which features contribute to the same latent factor thus, indicating relationships like complementarity. MOFA+ requires the feature matrices in a specific format. As can be seen in table 2, MOFA+ allows for samples of the same group e.g. BRCA to be grouped together for both feature types. In total four MOFA+ objects will be created - one for each sample group.

4 Results

This chapter will present and discuss the results of the conducted experiment, the shortcomings that occurred and provide directions for future research.

Sample Name	Feature	Value	Group	View
Sample 1	Chr1 0:5000000	*****	BRCA	WPS
Sample 1	Chr1 0:5000000	*****	BRCA	Standardized Ratio
...	...	...	...	...
Sample N	Chr22 ...	*****	BRCA	WPS
Sample N	Chr22 ...	*****	BRCA	Standardized Ratio

Table 2: Representation of MOFA+ data frame

4.1 Short-long ratio of the fragment lengths and WPS

The first experiment run is checking for complementaritiy between the short-long ratio of the fragment lengths from 3.1.1 and WPS from 3.1.3. We begun with confirming the usefulness of the features types especially for the WPS, since we calculated it in a slightly different manner than in the original paper [14]. To that end, Manhattan plots showing the differences between healthy controls and each cancer case were generated. As shown in supplementary figures B1 and B3, both the short-long ratios and the WPS exhibit variations across different bins. This gave sufficient cause to continue the experiment.

We created four MOFA+ objects as mentioned in 3.3.2. Using these objects, we calculated the proportion of variance explained (i.e. the coefficient of determinations (R2)) by the MOFA+ factors across the two feature types. Tables 3 and 4 shows the amount of variation explained per factor, per sample group. Alongside the proportion of variance, we also computed the Pearson Correlation Coefficients and p-values for the factors as shown in tables 5 to 8 for each group. These statistics provide valuable insights into which factors contribute most to which feature type and which factors show correlation between the short-long ratios and the WPS. Using both, we eliminate factors that do not contribute to answering the research question. We continue with the analysis by examining each sample group independently.

Factor	BRCA	Control	CRC	LUAD
Factor 1	0.003962	0.475860	0.000000	0.797189
Factor 2	14.994697	63.559652	60.567389	28.096179
Factor 3	12.582171	11.757533	23.250430	18.140210
Factor 4	6.541745	4.677977	4.193566	11.922704
Factor 5	4.152841	4.954752	3.494248	7.697722
Factor 6	1.401962	1.192122	0.761268	2.605159
Factor 7	0.844254	0.707920	1.027932	2.399372
Factor 8	0.003578	0.626719	0.882350	2.157469
Factor 9	0.003293	0.266680	0.565307	2.056092
Factor 10	0.011781	0.145003	0.219798	1.233650

Table 3: Variance Explained by Factors for Standardized Ratio

4.1.1 BRCA

Table 5 suggests that most factors do not show a significant correlation between the two feature types. Factors 5, 6, and 9 exhibit a moderate negative correlation, whilst factor 10 shows a

Factor	BRCA	Control	CRC	LUAD
Factor 1	95.997220	95.924170	95.981025	95.926700
Factor 2	0.000091	0.000000	0.000000	1.400640
Factor 3	0.000076	0.039384	0.000000	0.028134
Factor 4	-0.002085	1.398381	0.030392	0.008740
Factor 5	-0.014069	0.000000	0.000000	0.069507
Factor 6	1.090837	0.000000	1.369869	0.096130
Factor 7	0.036310	0.079795	0.000000	0.030256
Factor 8	0.743937	0.000000	0.000000	0.000000
Factor 9	0.068830	0.059986	0.000000	0.000000
Factor 10	0.017368	0.018965	0.074195	0.000000

Table 4: Variance Explained by Factors for WPS

stronger positive correlation. This reduces the factors to evaluate down to four. We could have also chosen to keep factors such as 1 and 2 which demonstrate strong values for the short-long ratios and the WPS in tables 3 and 4 however, they are poor indicators of correlation between the two feature types and thus were omitted. Next we plotted heat-maps for each feature type using the four chosen factors as shown in figures 1 and 2 (Due to the small Pearson correlation coefficients, only the top feature for each factor is plotted).

Factor	Pearson Correlation Coefficient	p-value
Factor 1	0.00	9.21e-01
Factor 2	0.01	8.51e-01
Factor 3	-0.00	9.59e-01
Factor 4	0.05	2.01e-01
Factor 5	-0.14	6.02e-04
Factor 6	-0.14	9.30e-04
Factor 7	0.06	1.79e-01
Factor 8	-0.04	3.89e-01
Factor 9	-0.16	1.15e-04
Factor 10	0.22	5.50e-08

Table 5: Pearson Correlation Coefficients and p-values for BRCA

In this experiment bins are paired, unfortunately none of the top features for any of the chosen factors are shared between the two feature types in the BRCA dataset. This leads to a hypothesis that - Short-long ratio of the fragment lengths and the WPS are independent (for the BRCA samples). We justify this verdict by running a linear regression algorithm on the two features as described in section 3.3.1. Supplementary table 9, shows that per chromosome the R-Squared scores are very small further solidifying our verdict.

4.1.2 Control

For the healthy control samples, table 6 implies a moderate negative correlation exhibited by factors 4 and 10, while factors 7 and 9 show a stronger positive correlation. We plot heat-maps for each feature type for selected four factors as shown in figures 3 and 4.

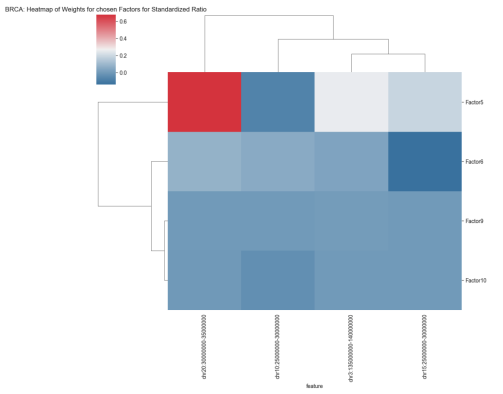


Figure 1: Heatmap of feature weights for factors 5, 6, 9, and 10 for the short-long ratios in the BRCA dataset

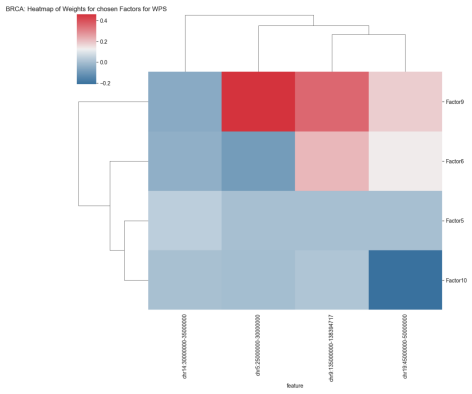


Figure 2: Heatmap of feature weights for factors 5, 6, 9, and 10 for the WPS in the BRCA dataset

Factor	Pearson Correlation Coefficient	p-value
1	-0.01	7.67e-01
2	0.03	5.06e-01
3	-0.03	4.38e-01
4	-0.10	1.17e-02
5	-0.04	3.54e-01
6	-0.09	2.98e-02
7	0.50	3.14e-38
8	-0.09	2.19e-02
9	0.26	1.55e-10
10	-0.10	1.80e-02

Table 6: Pearson Correlation Coefficients and p-values for Control

Similar as in the case of the BRCA dataset, none of the top features for any of the chosen factors are shared between the two feature types, implying independence between - Short-long ratio of the fragment lengths and the WPS for the healthy control dataset. As before, we attempt to justify this hypothesis by running a linear regression algorithm on the two features, and just like in the case of the BRCA dataset, supplementary table 10, shows that per chromosome the R-Squared scores are again quite insignificant validating our theory.

4.1.3 CRC

Table 7 provides a moderate negative correlation exhibited by factors 4 and 6, while factors 7 and 10 show a positive correlation for the CRC dataset. We proceed by plotting heat-maps for each feature type for selected four factors as shown in figures 5 and 6.
Akin to the BRCA and healthy control datasets, there are no shared features/bins for the chosen factors between the two feature types. We extend our analysis with linear regression through supplementary table 11, which shows that for more chromosomes on average the R-Squared scores are quite insignificant, however chromosomes 14 shows a moderate score. This makes us

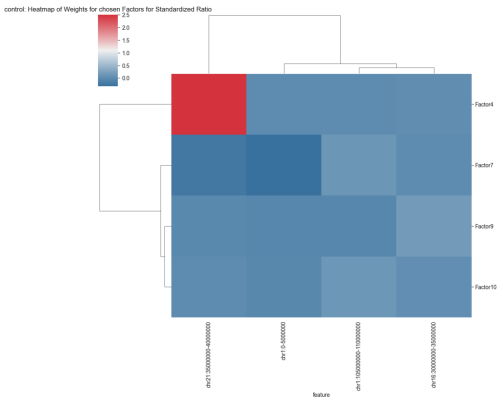


Figure 3: Heatmap of feature weights for factors 4, 7, 9, and 10 for the short-long ratios in the healthy control dataset

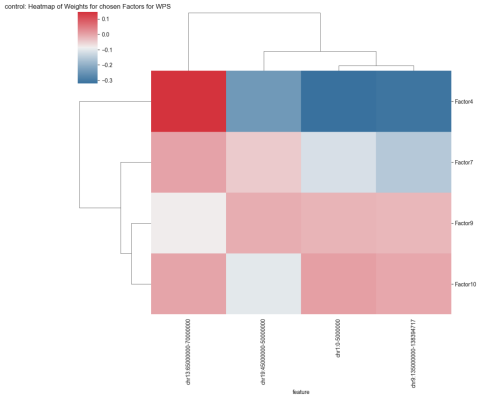


Figure 4: Heatmap of feature weights for factors 4, 7, 9, and 10 for the WPS in the healthy control dataset

Factor	Pearson Correlation Coefficient	p-value
1	-0.07	9.06e-02
2	0.03	5.41e-01
3	0.01	8.60e-01
4	-0.10	1.42e-02
5	-0.01	8.06e-01
6	-0.13	1.31e-03
7	0.12	3.46e-03
8	0.05	2.12e-01
9	-0.04	3.80e-01
10	0.14	7.14e-04

Table 7: Pearson Correlation Coefficients and p-values for CRC

turn to the heatmaps in figures 5 and 6 where in the heatmap for the short-long ratios we find a bin for chromosome 14.

Supplementary figure C33 shows that the slope of the regression line is negative which indicates a negative linear relationship between the two feature types for the bin chromosome 14: 105000000-107013718. This negative correlation implies that higher short-long ratios are associated with lower the WPS values for this particular chromosome region in the CRC dataset.

This linear regression plot indicates a moderate negative linear relationship in CRC data, suggesting that as the standardized ratio increases, the WPS decreases. The scatter of the data points around the regression line shows some variability but generally supports this negative trend. We estimate the degree of statistical significance by calculating the R-squared score. A score of 0.2233886433673239 means that approximately 22.3% of the variance in the WPS can be explained by the short-long ratios for this bin. This suggests a moderate fit of the model to the data. While it indicates some level of explanatory power, over 75% of the variance in the WPS is due to other factors. Leading to a conclusion, that despite showing some promise the

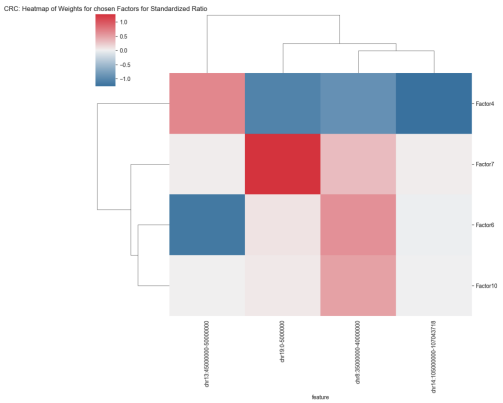


Figure 5: Heatmap of feature weights for factors 4, 6, 7, and 10 for the short-long ratios in the CRC dataset

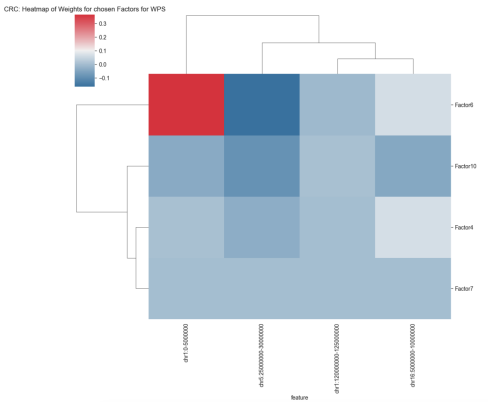


Figure 6: Heatmap of feature weights for factors 4, 6, 7, and 10 for the WPS in the CRC dataset

two feature types remain largely independent from each other for the CRC dataset.

4.1.4 LUAD

Factor	Pearson Correlation Coefficient	p-value
1	-0.04	2.96e-01
2	-0.04	3.11e-01
3	0.04	3.39e-01
4	-0.04	3.52e-01
5	-0.02	5.90e-01
6	-0.15	2.61e-04
7	0.08	4.64e-02
8	-0.03	4.77e-01
9	-0.01	7.93e-01
10	-0.07	7.51e-02

Table 8: Pearson Correlation Coefficients and p-values for LUAD

For LUAD dataset, only factor 6 shows any noticeable (negative) correlation as shown in table 8. Therefore we plot heat-maps for each feature type for factor 6. However, unlike before where the heatmaps where created for only the top feature per factor, here we plot for the top ten features as we only have one factor to investigate.

Figures 7 and 8 show that no features/bins are shared between the heatmaps. We further investigated this by running a linear regression algorithm on the LUAD dataset for the two feature types. As shown in supplementary table 12, the chromosomes have small r-squared scores implying that the short-long ratios do not explain the variation in the WPS (much) for this dataset. Leading to a verdict that for the LUAD, the short-long ratios and the WPS are (mostly) independent from each other.

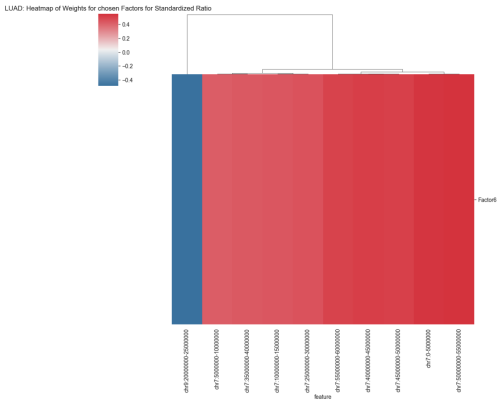


Figure 7: Heatmap of feature weights for factor 6 for the short-long ratios in the LUAD dataset

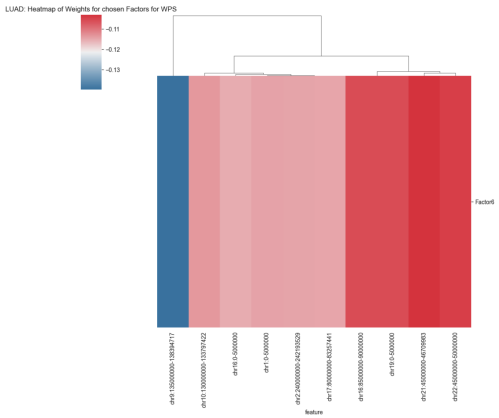


Figure 8: Heatmap of feature weights for factor 6 for the WPS in the LUAD dataset

4.1.5 Implications of Findings

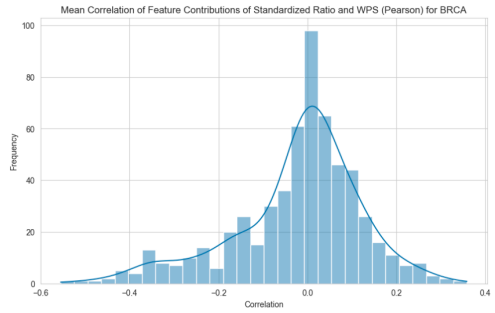


Figure 9: Distribution of Pearson correlation coefficients for BRCA samples

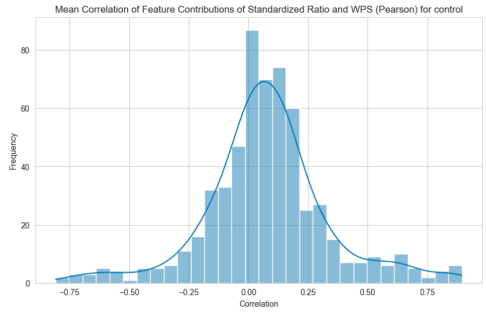


Figure 10: Distribution of Pearson correlation coefficients for control samples

We analyzed the relationships between the short-long ratios and the WPS for each sample group using both MOFA+ and linear regression. Across all four groups, our findings consistently indicated that the two feature types were largely independent from each other, suggesting that short-long ratios and the WPS do not significantly influence one another. This can be seen graphically in figures 9 to 12, where the histograms show that majority of the correlation values are centered around zero for all groups, suggesting a poor linear relationship between short-long ratios and the WPS. Therefore, we can ascertain that these two feature types exhibit a high degree of complementarity, as they provide unique and non-overlapping information.

4.2 Short-long ratio of the fragment lengths and 5' trinucleotide fragment end sequence diversity

The second experiment run is checking for complementaritiy between the short-long ratio of the fragment lengths from 3.1.1 and the Gini index from 3.1.2. As before, we began with confirming

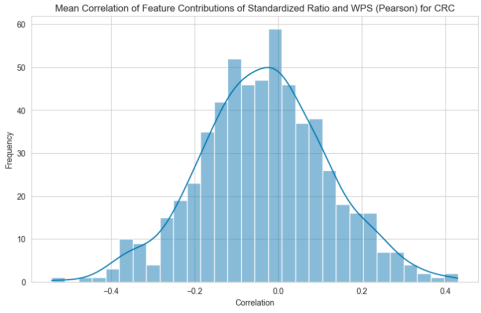


Figure 11: Distribution of Pearson correlation coefficients for CRC samples

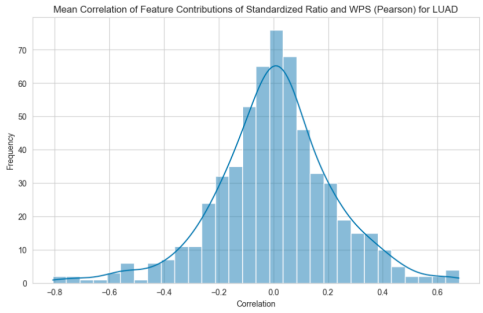


Figure 12: Distribution of Pearson correlation coefficients for LUAD samples

the usefulness of the features types. As can be seen in supplementary figures B1 and B2, there are stark differences between how the values are distributed. The plot for the short long ratios exhibit a noticeable deviation for each case versus control, providing results that are in line with previous literature [6] [11]. In comparison, supplementary figure B2 highlights a less varied distribution, demonstrating curious results prompting for more investigation in the feature.

To understand, why there is negligible variation in the values of the Gini index compared to the short-long rations, we plotted heatmaps for each sample group for the feature as shown in supplementary figures B4 - B6. Each cell represents the Gini Index for a specific chromosome in a particular sample, with color coding indicating the magnitude of the index. Despite showing a high diversity per sample group collectively, the indifference in the heatmaps implies that there are little to no distinct patterns that can be used to differentiate between the sample groups. This is in line with the results found in the original paper [10], where the values for the Gini Index were also quite similar between the different groups. However, questions still remain as to why the values for the Gini index are so uniform and ascertain if the trinucleotides vary at all.

To answer these questions additional probing into the 5' trinucleotide fragment end sequence diversity was needed. We charted the counts of the trinucleotide endings for one sample from each group for one bin. As seen in figures 13 to 16 the counts presented a noticeably similar distribution for each trinucleotide ending with some minor variations. We extended the analysis to calculate the average count for each ending per chromosome for all datasets.
Supplementary figures B8 - B29 the average counts per trinucleotide ending are close to identical per sample group for each chromosome, with endings such as AAA, and TTT regularly having large counts and TCG and CGA consistently showing low counts. As this research is conducted from a computer science perspective rather than a clinical biology one, it is challenging to draw conclusions why this feature is so uniform. Various factors such as PCR artifacts, preferential cleavage by DNASE13, or excessively large bins could have influenced the results. Nevertheless, the values are sufficient for our analysis, whereby we can hypothesize, the Gini Index as described in [10] does not capture unique fragmentomic features, and should not pursed as a biomarker for the detection of cancer.

To substantiate our hypothesis a linear regression is performed for each sample type per bin. We calculated the average R-Squared scores per chromosome for each sample group shown in the supplementary tables 13 - 16. For the BRCA and LUAD datasets we consistently saw low scores

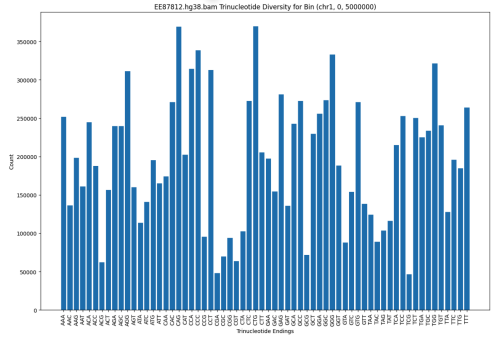


Figure 13: Distribution of Trinucleotide Counts for the bin chr1:0-5000000 for a BRCA sample

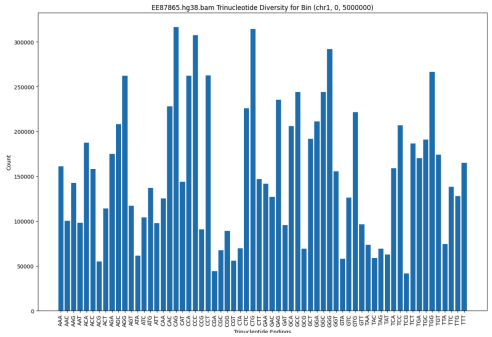


Figure 14: Distribution of Trinucleotide Counts for the bin chr1:0-5000000 for a CRC sample

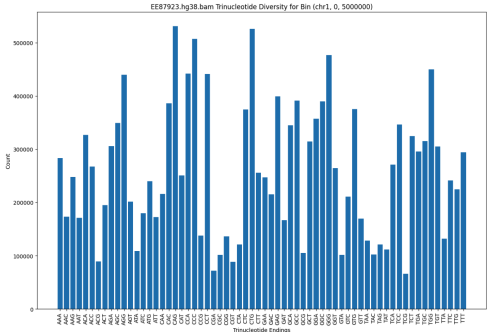


Figure 15: Distribution of Trinucleotide Counts for the bin chr1:0-5000000 for a healthy control

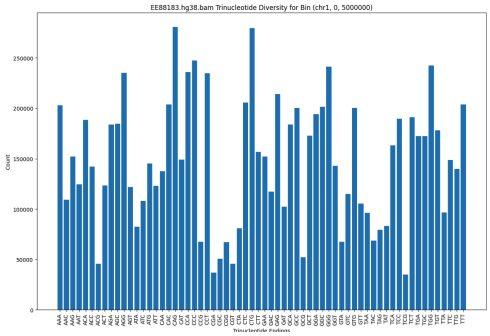


Figure 16: Distribution of Trinucleotide Counts for the bin chr1:0-5000000 for a LUAD sample

throughout indicating that the two features are very independent from each other (for BRCA and LUAD). However for the healthy control and CRC datasets some chromosomes had large scores, namely chromosomes twenty and fourteen being the highest respectively. We visualized linear regression plots for each bin in these chromosomes, as seen in supplementary figures C1 - C35. We observe a consistent negative correlation between the two feature types across both the CRC and healthy control datasets. However, the actual values for the Gini index are all closely clustered together, this relates back to the point made earlier that it is challenging to draw conclusions why this feature is so uniform from a purely computer science prospective. We therefore conclude our investigate and adjudicate our hypothesis.

4.3 Shortcomings

A major shortcoming in this research was the brief time allocated. A period of ten weeks was given and given that our background is in computer science, this which resulted in a severe lack of prerequisite biological knowledge needed to properly conduct this research. This resulted in some minor delays at the beginning as we attempted to bridge the gaps in knowledge. Secondly, the scope of the research was lessened as we were provided data which contained samples from

only four types of patients. Having data on other types of cancer could have impacted the final results due to the complex nature of cancer. These shortcomings were on a broader level, however other points of contention arose in the approaches taken in this experiment (defined in section 3). As mentioned in 3.1.1, all the short-long ratios were standardized using z-scores. However, when reviewing our methods we learnt that there were different ways on how we could have done this standardization. Our approach relied on calculating the z-scores for every sample separately. Therefore, the short-long ratios had a mean of 0 and a standard deviation of 1 on a per sample basis and when combined in the feature matrices from 3.2, the columns would not have this normalization. Secondly as mentioned in 3.1.3, our method of calculating the WPS was not a replication of the work done by Synder et al [14]. To assiduously replicate Snyderâs work would require making use of nucleosome peak calling and/or Fourier transformations and correlation with expression, unfortunately due to limited time constraints we would do this. A significant amount of time was spent on investigating the surprising results for the Gini Index. We researched if the score itself was flawed or the whether the trinucleotide endings themselves provide limited information.

4.4 Recommendations for future research

The research conducted in this study serves as a foundation for future investigations into the complimentarity of different blood-based cancer detection tests. Due to time constraints, the analysis was limited to three fragmentomics features. Features such as SNV detection, OCF analysis, and CNV analysis [15] could be explored. Using different features (alongside the three used in this study), further insights into the three features examined in this paper, particularly the Gini Index. For instance, could other features possibly explain the uniform distribution observed in the Gini Index?

This research relied on data from Western sources [6]. To ensure unbiased and universally applicable results, future works should strive to include data from a diverse population representing different racial and ethnic backgrounds. This would contribute to a more comprehensive understanding of the relationship between fragmentomics features.

Moreover, alternative methods beyond those used in this study should be explored. Methods like Angle-based Joint and Individual Variation Explained (AJIVE) could be employed to assess complementarity between features and be used to validate the universality of the results found in this study. Furthermore, given that we selected three feature types, our linear regression could be extended to a multiple regression model, which aims to predict a value based on two or more variables. Multiple regression can be useful to observe patterns between the three features i.e. can a combination of two features predict the third, making the latter irrelevant for further analysis when trying to create a multi-modal classifier.

Lastly, we recommend that all future works calculating the WPS in a manner similar to ours, to fully replicate Snyderâs work as previously mentioned. One approach would be to determine the distance between peaks without peak calling. This involves using the WPS signal of length 5M and determining its period using a Fast Fourier Transform (FFT). We advise visualizing smaller segments of the signal (1000-10000 base pairs) from different bins, displaying the original WPS signal, the mean, and the calculated periods. This visual inspection can help confirm the presence of a dominant frequency and verify the accuracy of the calculated periods.

5 Conclusion

In conclusion, the research done in this paper assessed the complementarity in different blood-based cancer detection tests. Three fragmentomics features were selected and analysed. We examined the relationships between the short-long ratios and the WPS for each sample group using both MOFA+ and linear regression. Our results consistently indicated that the two feature types were largely independent across the sample groups, suggesting they offer unique and non-overlapping information. This was evident in the correlation plots 9 - 12, where the majority of correlations centered around zero. Therefore, we can ascertain that these two feature types exhibit a high degree of complementarity.

We also examined the 5' trinucleotide fragment end sequence diversity (Gini index), finding consistently similar values within each sample group for every chromosome. This uniformity led us to inquire, the reasons behind such results. However, as this research was conducted from a computer science prospective and not a clinical biology one, explaining the uniformity of these values proved challenging. Consequently, we concluded our investigation and adjudicated that Gini Index does not capture unique fragmentomic features, and should not be pursed as a biomarker for the detection of cancer.

One must be very cautious about the interpretation of these results given that they were obtained from a computer science perspective, and not a biological one. It is highly recommended that other researchers who have a strong background in bioinformatics conduct a validation of the results obtained.

6 Responsible Research

This section aims to address the ethical aspects of conducting this research. The paper is primarily focused on the information entailed in the DNA fragments drawn from blood samples. It is paramount that one must be hyper-vigilant when working with such data.

6.1 Management of Data

The raw data used in this experiment is obtained from a research conducted by Cristiano et al [6], where samples from cancer patients and healthy controls were obtained from "ILSBio/Bioreclamation, Aarhus University, Herlev Hospital of the University of Copenhagen, Hvidovre Hospital, the University Medical Center of the University of Utrecht, the Academic Medical Center of the University of Amsterdam, the Netherlands Cancer Institute and the University of California, San Diego" [6]. In the paper, the authors mention that "all samples were obtained under Institutional Review Board approved protocols with informed consent from all participants for research use at participating institutions" [6].

As the data concerns real patients, personal information such as names and other parameters which reveal a participants identity must be redacted to protect their privacy. Research should also respect diversity and take full account of genetic factors such as ethnicity, gender, disability, age, and sexual orientation in its design, undertaking, and reporting. For instance, certain racial or ethnic groups will have higher proportions of slow metabolisers than others. One example of this is, "Japanese and Inuit populations have a high proportion of rapid acetylation metabolisers; European and African populations have an equal proportion of slow and rapid metabolisers" [3]. This highlights the importance of having data from a diverse group of participants. The data used in this experiment is taken from largely western sources, which could produce results that are not relevant to a large populace. This could have negative repercussions, as results from minorities could be construed as outliers and removed/ignored.

As this paper is part of Delft University of Technology's bachelor course, the task of procuring the raw data lies with the course coordinators and project supervisors. It is their duty to ensure the data was collected the required permissions and that, if necessary, an ethical review was conducted as the guidelines in Netherlands Code of Conduct for Research Integrity 2018 [1].

6.2 Use of large language models (LLMs)

LLMs and conversational agents such as ChatGPT[2] and Gemini[3] were used throughout the course of this research. They served to largely aid in debugging code and help decode unfamiliar biology idiom into a more regular tone for easier understanding, instead of regularly seeking out supervisors' assistance. An example of the use of LLMs, is shown in appendix D1, where a prompt was made asking for an explanation on what multi-omics data refers to.

As this was the first time conducting research in this domain, there was also a knowledge gap regarding the coding processes, given a background as a computer science student rather than a biologist or bioinformatician, conversational agents helped decode some of these processes,

[1] https://www.nwo.nl/en/netherlands-code-conduct-research-integrity
[2] https://chatgpt.com/
[3] https://gemini.google.com/

helping in understanding the data. An example of this is shown in appendix **??**, where a prompt is made asking for how to view end motifs using python.

References

[1] Tests and procedures used to diagnose cancer. Accessed: 22 April 2024.

[2] What is cancer? 2021. Accessed: 22 April 2024.

[3] Peter Allmark. Should research samples reflect the diversity of the population? *Journal of medical ethics*, 30(2):185–189, 2004.

[4] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124, 2018.

[5] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21:1–17, 2020.

[6] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.

[7] Andreas Halner, Luke Hankey, Zhu Liang, Francesco Pozzetti, Daniel A Szulc, Ella Mi, Geoffrey Liu, Benedikt M Kessler, Junetha Syed, and Peter Jianrui Liu. Decancer: Machine learning framework tailored to liquid biopsy based cancer detection and biomarker signature selection. *Iscience*, 26(5), 2023.

[8] Investopedia. R-squared definition, 2024. Accessed: 2024-05-24.

[9] Chao Li, Xiao Luo, Yanpeng Qi, Zhenbo Gao, and Xiaohui Lin. A new feature selection algorithm based on relevance, redundancy and complementarity. *Computers in Biology and Medicine*, 119:103667, 2020.

[10] Norbert Moldovan, Ymke van der Pol, Tom van den Ende, Dries Boers, Sandra Verkuijlen, Aafke Creemers, Jip Ramaker, Trang Vu, Sanne Bootsma, Kristiaan J Lenos, et al. Multi-modal cell-free dna genomic and fragmentomic patterns enhance cancer survival and recurrence analysis. *Cell Reports Medicine*, 5(1), 2024.

[11] Trong Hieu Nguyen, Nhu Nhat Tan Doan, Thi Mong Quynh Pham, Giang Thi Huong Nguyen, Thanh Dat Nguyen, Thuy Thi Thu Tran, Duy Long Vo, Thanh Hai Phan, Thanh Xuan Jasmine, Huu Thinh Nguyen, et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free dna for multi-cancer early detection and localization. *Elife*, 12:RP89083, 2023.

[12] Van-Chu Nguyen, Trong Hieu Nguyen, Thanh Hai Phan, Thanh-Huong Thi Tran, Thu Thuy Thi Pham, Tan Dat Ho, Hue Hanh Thi Nguyen, Minh-Long Duong, Cao Minh Nguyen, Que-Tran Bui Nguyen, et al. Fragment length profiles of cancer mutations enhance detection of circulating tumor dna in patients with early-stage hepatocellular carcinoma. *BMC cancer*, 23(1):233, 2023.

[13] Ting Qi, Min Pan, Huajuan Shi, Liangying Wang, Yunfei Bai, and Qinyu Ge. Cell-free dna fragmentomics: the novel promising biomarker. *International Journal of Molecular Sciences*, 24(2):1503, 2023.

[14] Matthew W Snyder, Martin Kircher, Andrew J Hill, Riza M Daza, and Jay Shendure. Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1):57–68, 2016.

[15] Wei Zhang, Lei Wei, Jiaqi Huang, Bixi Zhong, Jiaqi Li, Hanwen Xu, Shuying He, Yu Liu, Juhong Liu, Hairong Lv, and Xiaowo Wang. cfDNApipe: a comprehensive quality control and analysis pipeline for cell-free DNA high-throughput sequencing data. *Bioinformatics*, 37(22):4251–4252, 05 2021.

[16] Yulia V Zhitnyuk, Anastasia P Koval, Aleksandr A Alferov, Yanina A Shtykova, Ilgar Z Mamedov, Nikolay E Kushlinskii, Dmitriy M Chudakov, and Dmitry S Shcherbo. Deep cfdna fragment end profiling enables cancer detection. *Molecular Cancer*, 21(1):26, 2022.

A Pseudo Code

A.1 Feature Extraction

```
FUNCTION divide_the_genome_and_get_feature_values(sample):

    INITIALIZE featureValuesForSample as an empty dictionary

    FOR each chromosome and its length in sample:

        FOR each bin within the chromosome (with steps of 5000000):
            SET bin_end as the minimum of (bin_start + 5000000) and
                length
            SET bin_name as "chromosome:bin_start-bin_end"
            INITIALIZE feature_values as an empty list

            FOR each read in the bin range:

                IF read is valid (not unmapped, not secondary, high
                    quality, proper pair):
                    APPLY feature_function to read to get feature_value
                    ADD feature_value to feature_values

            IF chromosome not in featureValuesForSample:
                INITIALIZE data[chromosome] with MAX_POSITION as 0 and
                    BINS as an empty dictionary

            UPDATE featureValuesForSample[chromosome][MAX_POSITION] to
                the maximum of its current value and bin_end

            IF bin_name not in featureValuesForSample[chromosome][BINS]:
                INITIALIZE featureValuesForSample[chromosome][BINS][
                    bin_name] as an empty list

            ADD feature_values to featureValuesForSample[chromosome][BINS
                ][bin_name]

    CLOSE sample

    RETURN featureValuesForSample
```

Figure A1: Pseudo Code for Genome Division and Feature Extraction per sample

A.2 Calculate WPS per bin

```
FUNCTION calculate_wps(sample):
    INITIALIZE an empty list wps_data to store WPS results

    FOR each chromosome in sample:
        Determine the length of the chromosome

        FOR each 5Mb bin in the chromosome:
            INITIALIZE an empty list wps_values to store WPS for
                positions within the bin

            FOR each base pair position within the bin:
                INITIALIZE spanning_count and endpoint_count to 0

                Define a 120 bp sliding window centered at the current
                    position

                FOR each read within the 120 bp window:
                    if the read is properly mapped and is a proper pair:
                        Determine the fragment start and end positions

                        if the fragment completely spans the window:
                            Increment spanning_count

                        if the fragment has an endpoint within the window
                            :
                            Increment endpoint_count

                Calculate WPS for the current position as spanning_count
                    - endpoint_count
                Append the WPS value to wps_values

            Calculate the average WPS for the current bin from wps_values
            Store the chromosome, bin start, bin end, and average WPS in
                wps_data

    RETURN wps_data
```

Figure A2: Pseudo Code for calculating the WPS per bin

A.3 Feature Matrix Creation

```
FUNCTION create_feature_matrix(files, feature_type):
    FOR each sample IN files:
        sample_data = READ_SAMPLE(sample)

        # Select and transpose the feature column
        feature_column = SELECT(sample_data, feature_type)
        transposed_dataframe = TRANSPOSE(feature_column)

        # Construct new column names using CHROMOSOME, BIN_START, and
            BIN_END
        chromosomes = SELECT(sample_data, CHROMOSOME)
        bin_starts = SELECT(sample_data, BIN_START)
        bin_ends = SELECT(sample_data, BIN_END)

        new_columns = []
        FOR chrom, start, end IN ZIP(chromosomes, bin_starts, bin_ends):
            new_column_name = CONCATENATE(feature_type, "_", chrom, ":",
                start, "-", end)
            APPEND(new_columns, new_column_name)

        # Rename columns using the constructed names
        SET_COLUMNS(transposed_dataframe, new_columns)

        # Add the 'sample_name' column at the beginning
        sample_name_column = CREATE_DATAFRAME({SAMPLE_NAME: [sample_name
            ]})
        transposed_dataframe = HSTACK(sample_name_column,
            transposed_dataframe)

        APPEND(dataframes, transposed_dataframe)

    # Concatenate all single-row DataFrames
    feature_matrix = CONCAT(dataframes, how="vertical")

    RETURN feature_matrix
```

Figure A3: Pseudo Code for Feature Matrix Creation

B Plots

B.1 Manhattan Plots

Figure B1: Difference/ratio between cases and controls for short-long ratios

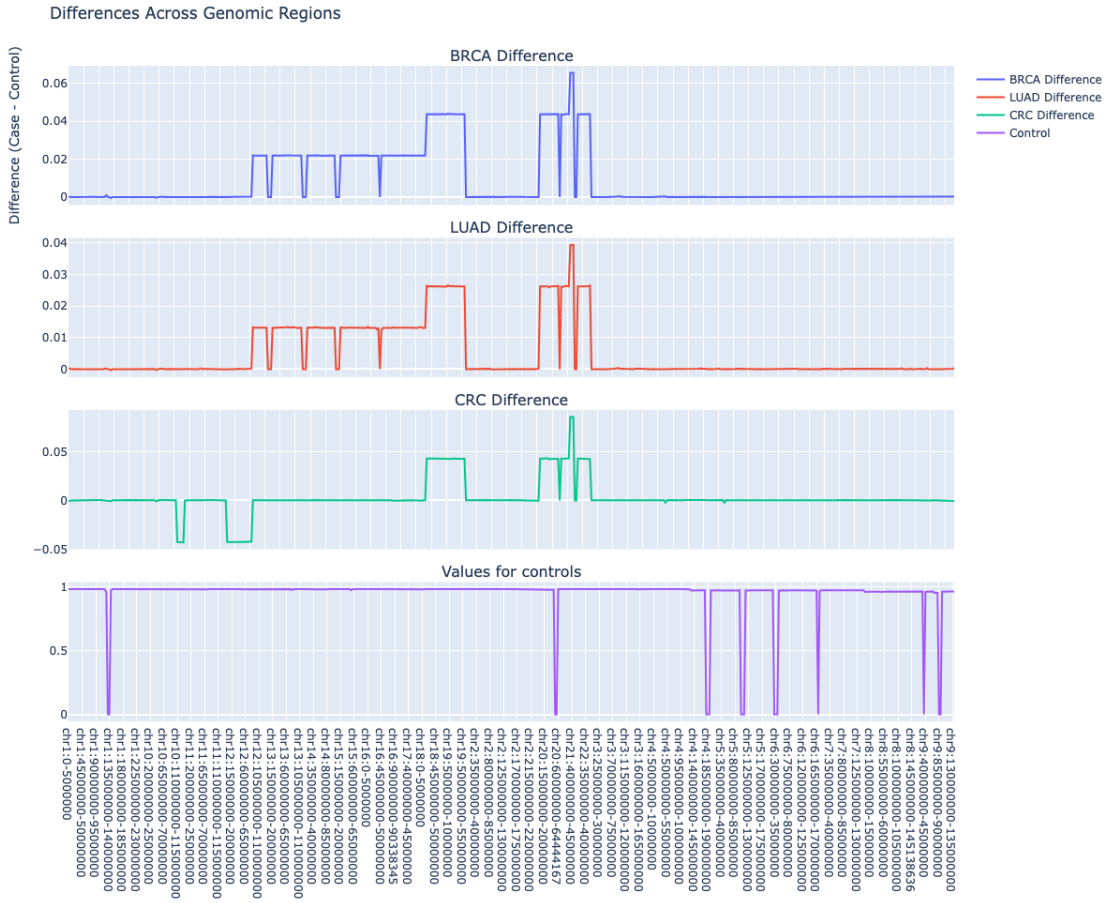


Figure B2: Difference/ratio between cases and controls for Gini Index

Figure B3: Difference/ratio between cases and controls for Window Protection Score

B.2 Heatmaps

Figure B4: Heatmap for the Gini Index for CRC Samples

Figure B5: Heatmap for the Gini Index for BRCA Samples

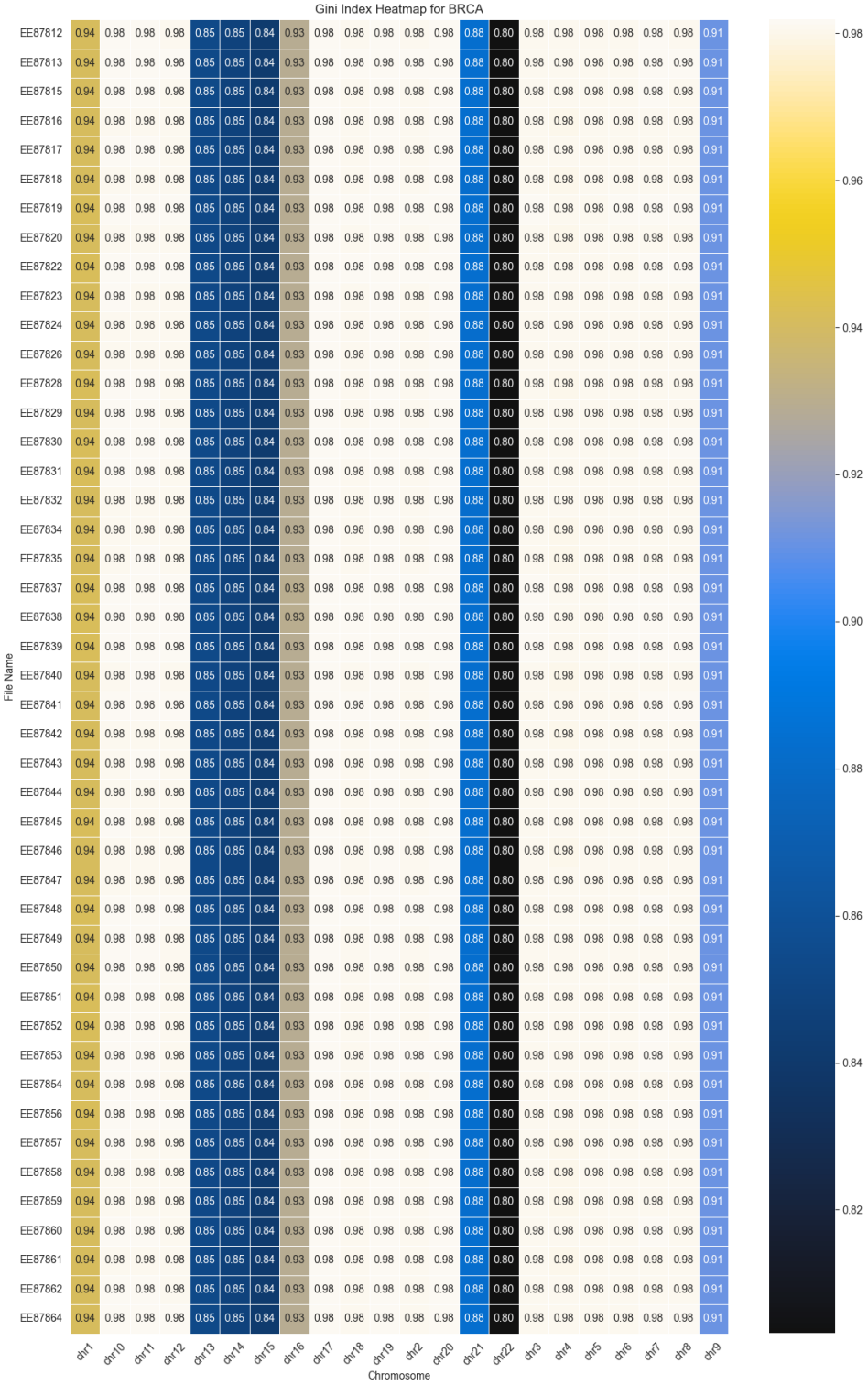


Figure B6: Heatmap for the Gini Index for healthy controls

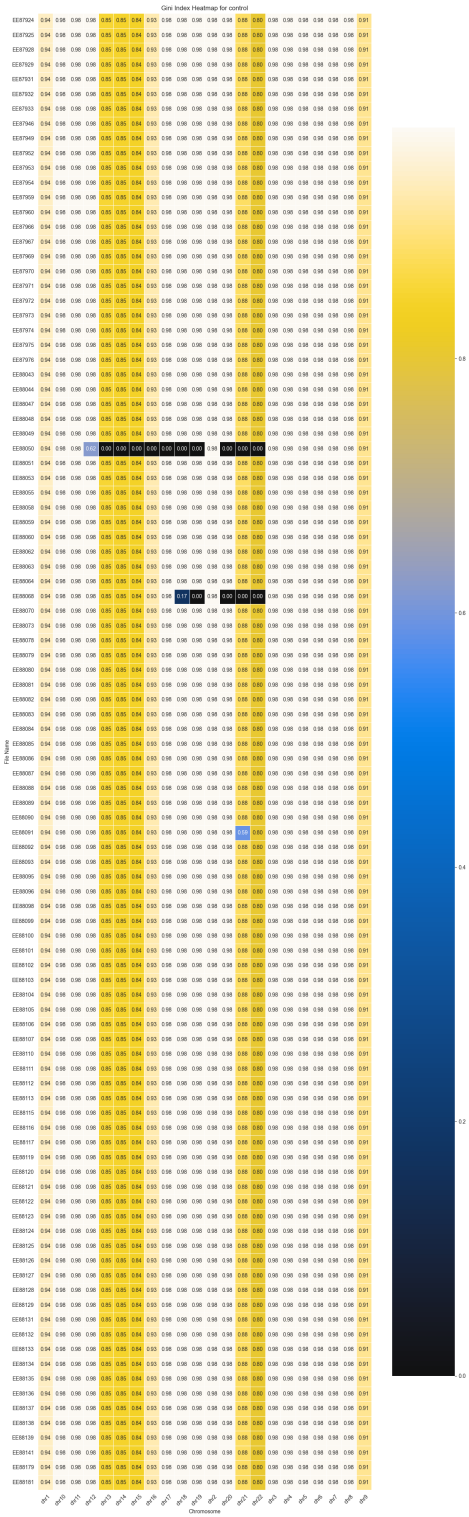
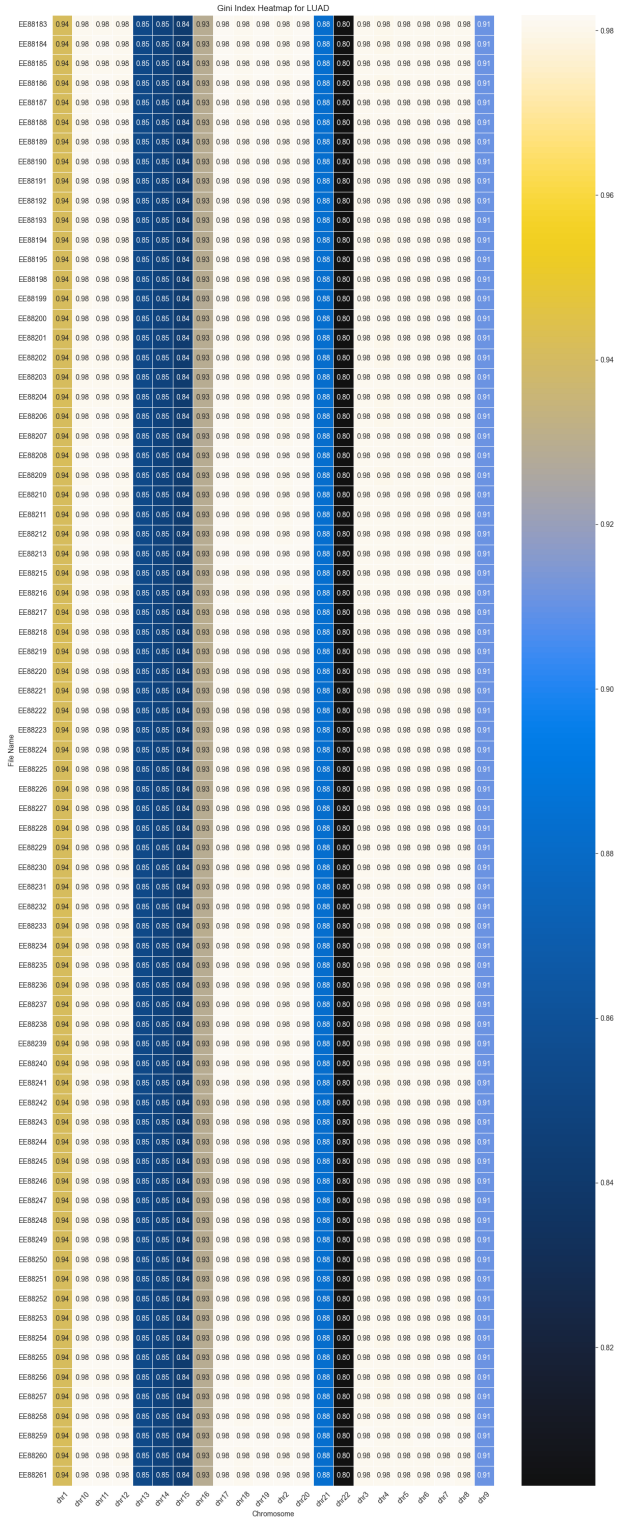


Figure B7: Heatmap for the Gini Index for LUAD Samples



B.3 Average Trinucleotide counts per bin

Figure B8: Average distribution of Trinucleotide Counts for Chromosome 1

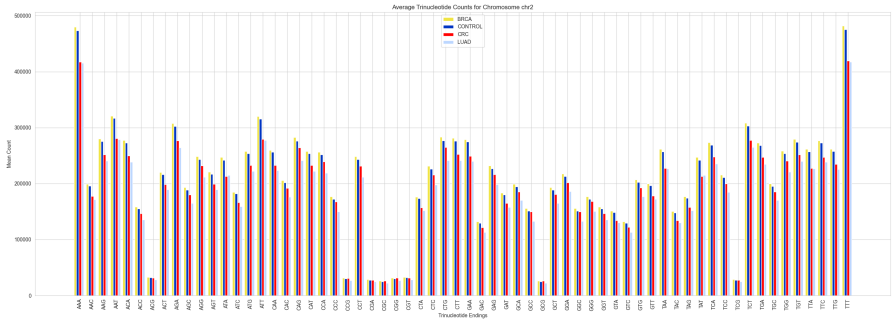


Figure B9: Average distribution of Trinucleotide Counts for Chromosome 2

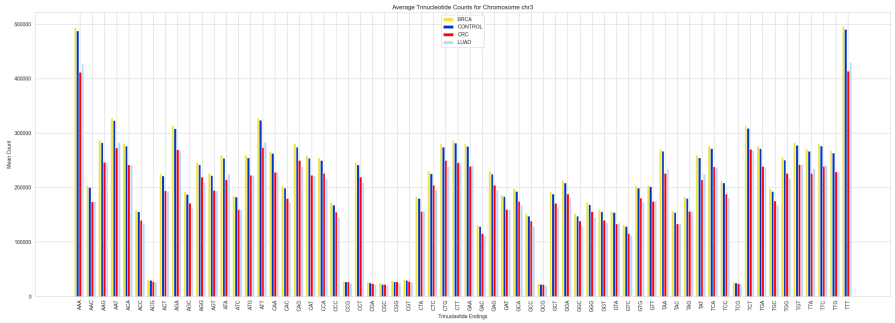


Figure B10: Average distribution of Trinucleotide Counts for Chromosome 3

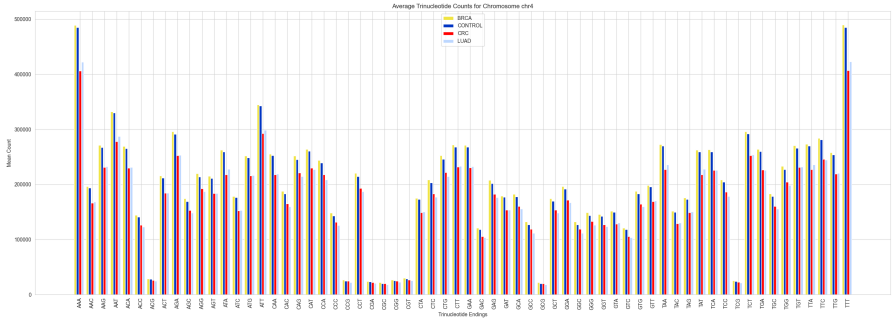


Figure B11: Average distribution of Trinucleotide Counts for Chromosome 4

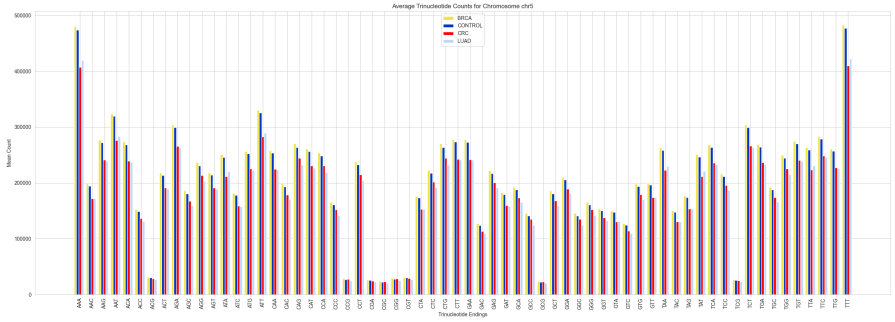


Figure B12: Average distribution of Trinucleotide Counts for Chromosome 5

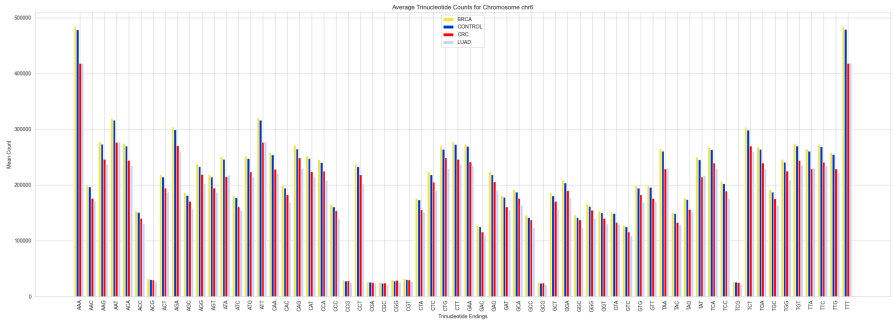


Figure B13: Average distribution of Trinucleotide Counts for Chromosome 6

Figure B14: Average distribution of Trinucleotide Counts for Chromosome 7

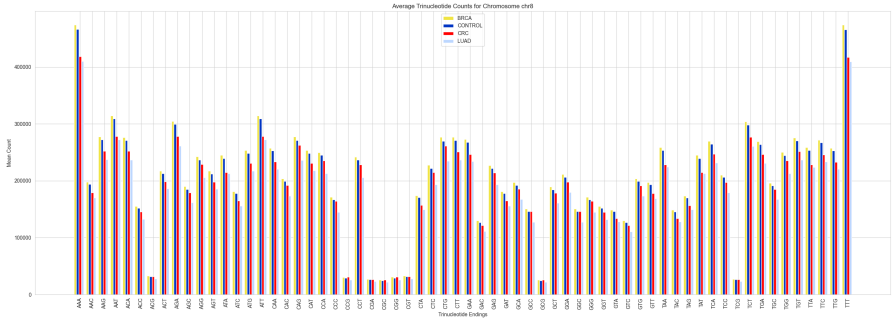


Figure B15: Average distribution of Trinucleotide Counts for Chromosome 8

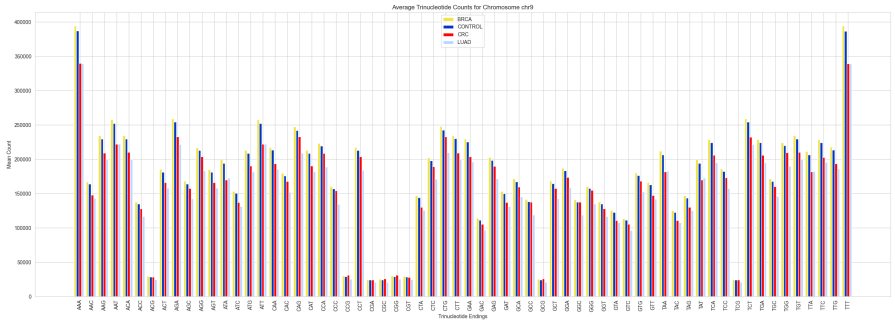


Figure B16: Average distribution of Trinucleotide Counts for Chromosome 9

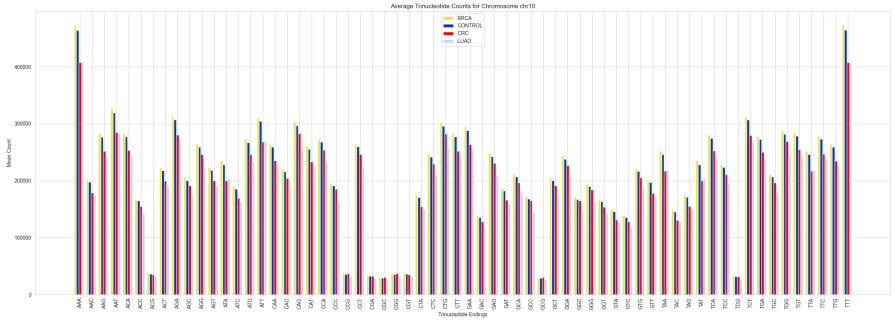


Figure B17: Average distribution of Trinucleotide Counts for Chromosome 10

Figure B18: Average distribution of Trinucleotide Counts for Chromosome 11

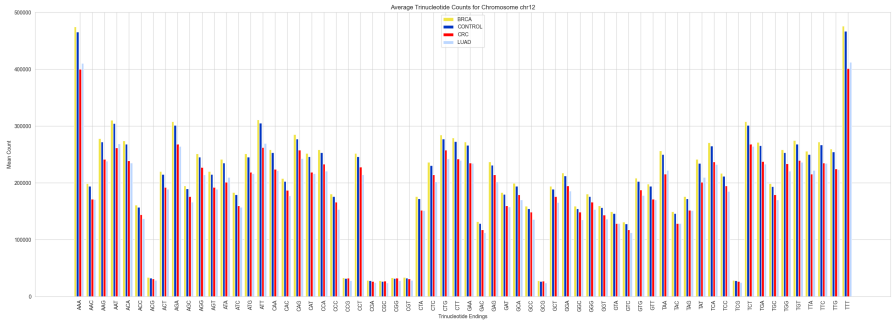


Figure B19: Average distribution of Trinucleotide Counts for Chromosome 12

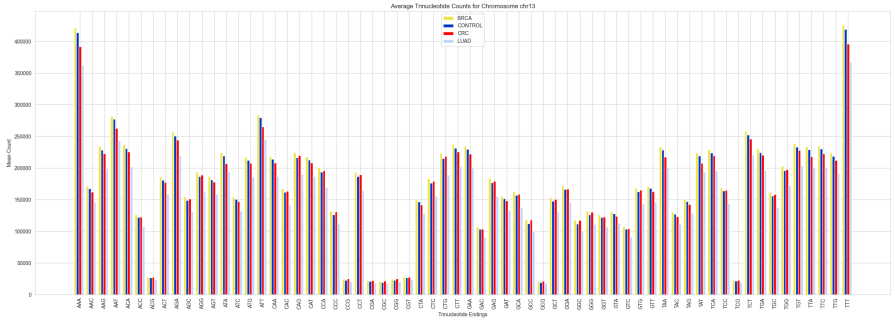


Figure B20: Average distribution of Trinucleotide Counts for Chromosome 13

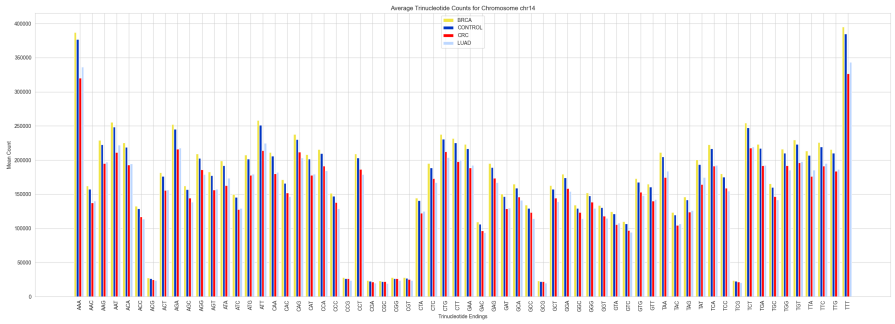


Figure B21: Average distribution of Trinucleotide Counts for Chromosome 14

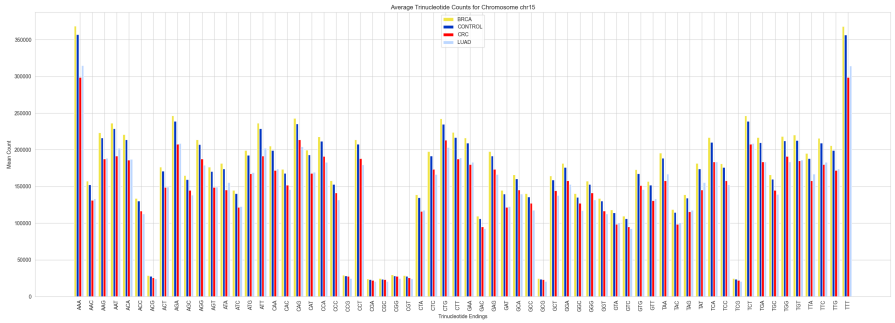


Figure B22: Average distribution of Trinucleotide Counts for Chromosome 15

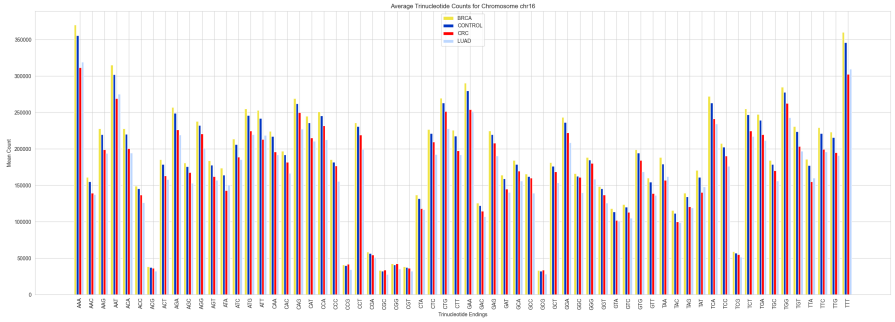


Figure B23: Average distribution of Trinucleotide Counts for Chromosome 16

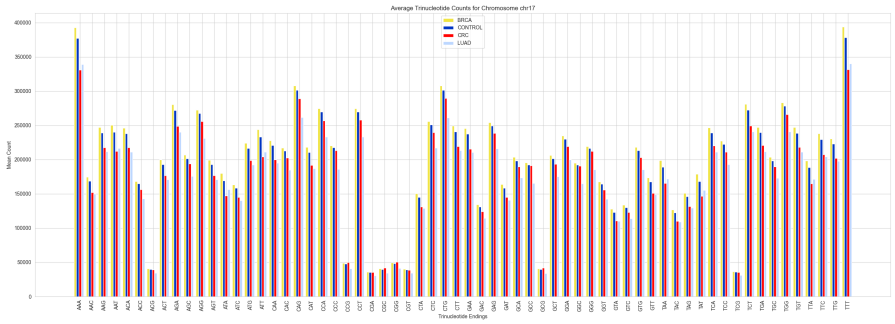


Figure B24: Average distribution of Trinucleotide Counts for Chromosome 17

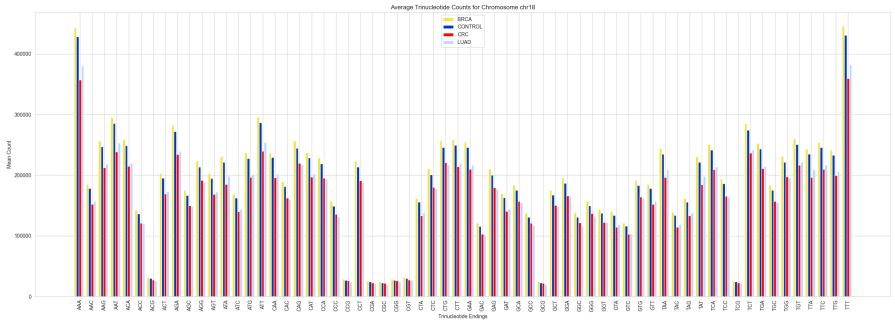


Figure B25: Average distribution of Trinucleotide Counts for Chromosome 18

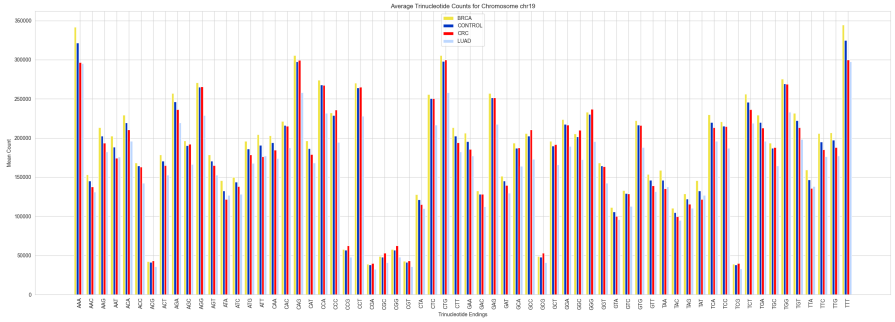


Figure B26: Average distribution of Trinucleotide Counts for Chromosome 19

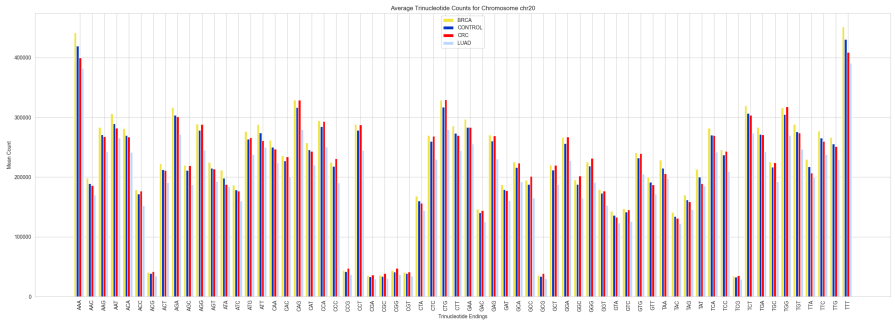


Figure B27: Average distribution of Trinucleotide Counts for Chromosome 20

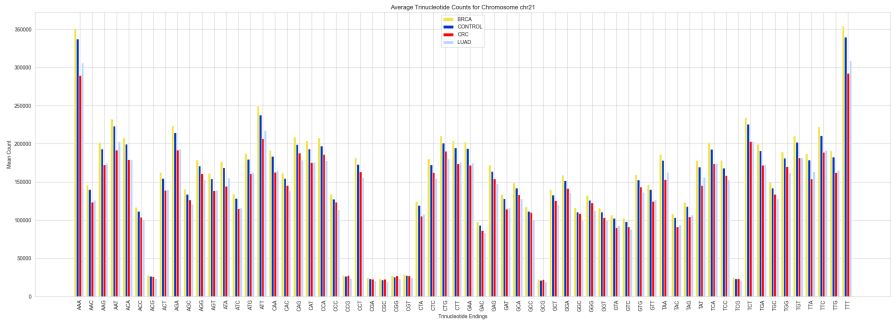


Figure B28: Average distribution of Trinucleotide Counts for Chromosome 21

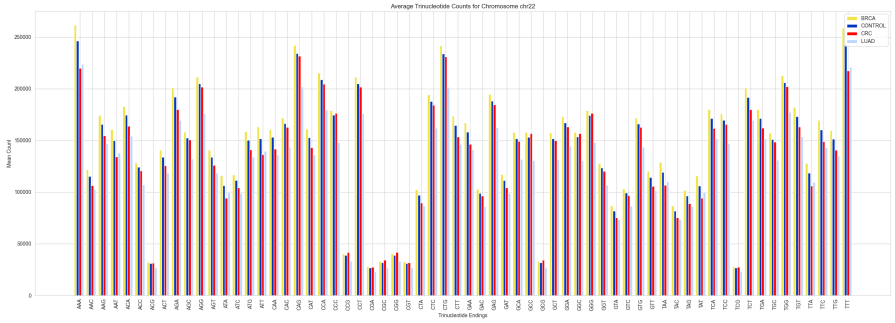


Figure B29: Average distribution of Trinucleotide Counts for Chromosome 22

C Linear Regression

C.1 Short-long ratio of the fragment lengths and WPS

Chromosome	R-Squared	Intercept	Coefficient
1	0.06880910577948958	2.881959603955279	0.2483373690562038
2	0.034116557443827245	2.8053845983713357	0.39105348818349644
3	0.03146048533794289	2.9434546213159605	1.0038784421819409
4	0.028519863343267116	2.9036583828553217	0.16721402864780308
5	0.029584923971439265	2.7760194105827223	0.32001177916890866
6	0.030499600062086803	2.778171314246935	0.6813183554301603
7	0.029413909395737306	2.7467503153467208	0.06322095281706308
8	0.032551750686750584	2.8751515997691746	0.5338264112086423
9	0.09689514136814315	2.613735466203455	0.5428058201839712
10	0.03470825439109727	2.9184346502521796	0.551252969033884
11	0.024780007039600035	2.9113108569903567	0.13095496461080708
12	0.04719749512939036	2.8594082936992726	1.0286195466001042
13	0.1509103355857079	2.5406197376610207	0.3311000565580337
14	0.1812033713105282	2.3457883516251608	0.7195610204175567
15	0.17741745677064588	2.449175023648279	0.5828360555504959
16	0.08238655932404017	2.5414384157261796	-0.8717740050014807
17	0.015646086286068606	2.722555403817797	0.034157834855014294
18	0.02151814574702578	2.7272619533714453	-0.3061472269921997
19	0.02404936100075528	2.4077945329612778	-0.6240055907954035
20	0.03459833476295544	2.8680957784834096	-0.9546131102059492
21	0.14200423475661947	2.4847305666213098	-0.6364230973953917
22	0.21347211987862405	2.0980275286994727	-0.8466082302724458

Table 9: BRCA Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.06880965157733453	2.500778293485195	0.4368147692174412
2	0.016027798665649035	2.5023090956026364	-0.47227528222775067
3	0.02044417485559008	2.4602602283140467	-0.631897890397452
4	0.03347014109084536	2.2149629985970147	-1.7511005079572448
5	0.022224388162725383	2.495174128113085	-0.5626800455169865
6	0.021645651390490333	2.285726958779289	-0.8494440984874739
7	0.019467875788638102	2.479969483020625	-0.11084591418253667
8	0.015793624649426376	2.429243385036345	-0.5102528541613808
9	0.08648068787339233	2.145949835074644	0.03846617510589336
10	0.014914674579999789	2.746373612819807	0.3374712469856886
11	0.019055353390373002	2.6525305565162682	0.08644344900784791
12	0.026026031229686903	2.337114000043204	-1.032257207707861
13	0.18974088132280761	1.9589812919881135	-1.3290635048194515
14	0.1773512170858593	2.005983843631658	-1.1110272342354677
15	0.17021698159616921	2.0112283102805146	-1.0127170510657302
16	0.07063440159312462	2.276186684488301	-0.7032930793886137
17	0.01938394023785322	2.5223071668612924	-0.8729307604479252
18	0.08024620381732978	2.230814083661842	-1.0667472086865102
19	0.0397350311700918	2.5343629053247247	-0.8147373884456323
20	0.06330814128048547	2.666301438653865	-1.065021754230189
21	0.18154293732741325	2.0107653774252223	-0.7546762907461757
22	0.21805684604889677	1.9558348082703	-0.6423363054168899

Table 10: Control Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.09953082840479588	2.4733857588094126	0.884575519475048
2	0.08680920420509365	2.6655010229612888	1.1179772092848785
3	0.05310850361421875	2.4917709032496327	0.6651459007933938
4	0.026699308862686382	2.328396796313251	0.2603414631109088
5	0.038697357637343895	2.4920909337558346	0.5992314395890268
6	0.14383875297637086	2.5860161890511293	1.350359534929962
7	0.39335909791757784	2.808769256078056	2.1610091924726826
8	0.18391200773039135	2.741211201427025	1.3674496468490704
9	0.21175937462548583	2.272751696525438	0.9726007107504663
10	0.1020779005549367	2.654944944777703	1.0428424257203344
11	0.09494567684026069	2.8349136852615753	1.118141486353342
12	0.2852615748948692	2.2920880447419	-1.3812088301997336
13	0.15186751655249414	2.1925148970241723	-0.1555000391443
14	0.43384348442634485	1.82238009513577	-1.1684361365679692
15	0.33621127843202225	1.7758534106833574	-0.8122199871306861
16	0.34855168865480624	2.167654341005522	-1.2752777947874347
17	0.19105033463861434	2.3645105162655025	-1.0553917586599257
18	0.22464442106176988	1.9404189011015418	-0.8568502972188815
19	0.18237966555634288	2.4119535341438945	-1.0385049044422552
20	0.07724166833614604	2.6838283515116816	-0.3392314573731183
21	0.3919561270961085	1.826169050321809	-1.0830169159340264
22	0.39424317081105914	1.76373965933062	-0.7927965785919475

Table 11: CRC Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.09391212626832916	2.5470834849171493	-0.29764700085311246
2	0.09293858962404417	2.2577449130106264	-1.339656067051109
3	0.053268468989810816	2.5198765172888766	0.18872857601447715
4	0.11673718682716301	2.945538500034519	1.130506595410072
5	0.08869672013335844	2.6836729243731305	0.963748535139818
6	0.0454524777960892	2.3082581173390824	-0.9224741111734517
7	0.05001579095757599	2.571811100485535	0.6988721853164802
8	0.04997654389508326	2.530915362737334	0.30534823644348236
9	0.10536839193136541	2.1811989446983646	-0.33414032584881326
10	0.050439384295361815	2.4466148419237017	-0.6763752383391719
11	0.06781490090861766	2.66931004287984	0.21430354995823933
12	0.07293916219693168	2.469967530094089	-0.562910870851863
13	0.21628116579050524	2.4140438694672874	0.9316568666971553
14	0.23673472233251144	2.2271523484092457	1.0455441150938742
15	0.17782725646952077	2.014423225107671	-0.5310674945002626
16	0.13174714396999365	2.4767162409181003	0.7720518029524754
17	0.06871803913860931	2.5352119193236704	-0.822820833238587
18	0.10928302712459	2.495601821872335	0.4187946711447668
19	0.09136494234403818	2.431929432536078	-0.9757574832968756
20	0.04999590872498218	2.932421754036853	-0.4750593575713319
21	0.17057569606547585	1.531402153555798	0.21727500979878767
22	0.1701927596268123	1.3743413315611286	-0.37327124279720364

Table 12: LUAD Chromosome Regression Summary

C.2 Short-long ratios and 5' trinucleotide fragment end sequence diversity

Chromosome	R-Squared	Intercept	Coefficient
1	0.10118850297045096	0.9413338545841274	0.000207476690473926
2	0.0900022490435446	0.9811262614963033	0.00032826133259176326
3	0.09690124914262692	0.9810906387402936	0.00063522417997337
4	0.07246245271080125	0.9807747795933535	0.0004067198118640718
5	0.08813867732606925	0.9810057008716676	0.0003980809770960306
6	0.06609400399120839	0.9810429541883993	0.00036609078271156284
7	0.07659887942399542	0.9811722538754053	0.00020026316348647813
8	0.06593376931201697	0.9811430560857393	0.0002525432480597266
9	0.14290260093719304	0.9110103952799441	0.00021875902518607142
10	0.04330211050694572	0.981248390272057	8.389275006648156e-06
11	0.07110339292945214	0.9813019878072042	0.00033584246408566783
12	0.10350836870664784	0.9812118435783903	0.0005271141446312222
13	0.17412812097993868	0.852879393084812	0.00021630873369620434
14	0.18230365461220754	0.8474139505740376	0.00030231468875717725
15	0.21081449035333724	0.8411660694215577	9.058493342304286e-05
16	0.1428012802073334	0.9293514314331548	-0.0008083611560107
17	0.1387279108887115	0.9817133600915757	-0.00011068061163351765
18	0.06810197736435207	0.9810752036192082	1.9595072363469442e-05
19	0.10442093333903696	0.9817787730637523	-0.000175342474024825
20	0.12313892942038668	0.9815205477845482	-0.0002724211413741615
21	0.15436169697312105	0.8827900819036802	-0.00010276879818965999
22	0.2872018764354058	0.8031944346616432	-9.332207862943459e-05

Table 13: BRCA Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.06527175478093009	0.9411890799916872	-0.00018994043189937237
2	0.013109632088614945	0.9809810991942092	-6.783807047386986e-05
3	0.033018600624163355	0.9808752734429464	-4.5578016044582075e-05
4	0.027574787201846102	0.9804915088017109	-0.0001519307547784541
5	0.026845155207299697	0.9808268632208454	-0.00012813521169324839
6	0.0197312036989467	0.9808918459295609	-3.167492361214109e-05
7	0.01796506327481154	0.9810828410018442	-1.972439960681104e-05
8	0.01818433570340187	0.9810044959862995	-8.520379514445488e-05
9	0.08372660863248309	0.9109019106852745	-1.0649035377485713e-05
10	0.013567463692356776	0.9811930158557236	-2.9582851107771648e-05
11	0.017381778198707912	0.9811732251130728	3.1995340434824755e-07
12	0.32634662562120975	0.9504348250333062	-0.1856534877128569
13	0.8639132280027175	0.7686347858795649	-0.4227004862127822
14	0.8815424479917194	0.7668693136963676	-0.4328660542934399
15	0.8938586555504503	0.7652011497903094	-0.43973528561722897
16	0.7693623026611691	0.882566061166241	-0.43726757871181104
17	0.7570172857227566	1.0028476750390707	-0.5091754789248927
18	0.9046879746094513	0.8712967421200025	-0.40313949570799484
19	0.7983829521848002	1.0782532492909491	-0.4288054504953056
20	0.9026883993839745	0.9277589903324771	-0.4044941571302132
21	0.8652389040029316	0.834132437078279	-0.30480395115012604
22	0.870565700201391	0.7946691692901685	-0.32497247678854796

Table 14: Control Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.16168137659249918	0.9413392019857468	-0.00020642626885021453
2	0.06416279682721042	0.9812135203362825	-0.00010954920329136473
3	0.04781097347989538	0.9811105247472778	-6.861232345204476e-05
4	0.048226805596974	0.9807231813074572	3.002720689110988e-05
5	0.04811410018873306	0.980953287901786	-8.310873863831408e-05
6	0.10870361618032452	0.9811197929231223	-0.00021564524986511714
7	0.047226388301785105	0.9813210847498582	-6.854460755119688e-05
8	0.037676678867051595	0.981264183544973	-4.130985895218728e-05
9	0.11925224739208008	0.9110964296012903	6.241069885831813e-05
10	0.1934386800293547	0.9515729715502167	-0.09045657343925437
11	0.09605725017903506	0.9813182933424266	-3.79063053956779e-05
12	0.8966241386794733	0.9053096922964053	-0.5860963166550336
13	0.5740303699839029	0.8256028545329578	-0.30883063227966945
14	0.9036512023808524	0.7508211270408677	-0.47885510596620456
15	0.750322000902095	0.7452815001269067	-0.37047454808111435
16	0.8496220699117453	0.8531539139318425	-0.5163326279348396
17	0.7377921118165396	0.9598724215767187	-0.5174859584743163
18	0.6984791887220496	0.8359448310958953	-0.40097818178291444
19	0.7382664357220282	1.0162663856292637	-0.5587335556095606
20	0.6250516742526494	0.9273323082381856	-0.41717467459859947
21	0.8446758503827281	0.7870243929029026	-0.4650735179482231
22	0.7377951894490934	0.7159244536299022	-0.34065497386294596

Table 15: CRC Chromosome Regression Summary

Chromosome	R-Squared	Intercept	Coefficient
1	0.0679855779129456	0.9412651349113854	6.35367417275149e-06
2	0.030978618966081534	0.9810312500960308	-8.510547050802037e-05
3	0.040359295825962135	0.9809412324774751	3.0359503627112174e-05
4	0.03321253906652119	0.9806721429936084	-7.941191951492845e-06
5	0.04298510760701074	0.980932305995518	9.509829598877972e-06
6	0.015138891726253153	0.9809621601794822	5.3705823935871585e-05
7	0.0156905878111175	0.9811138348223086	-3.173551527918333e-05
8	0.024112098167019945	0.9810846872471253	9.164495042697592e-05
9	0.10320647184683526	0.9109568093337074	1.858445263920263e-05
10	0.04180486324946227	0.9812198083943005	-9.814443728700546e-05
11	0.028850122692748788	0.9812435925987613	0.00014422794281104936
12	0.023455386177861897	0.9811237671927295	6.0514481721884574e-05
13	0.1493613819721836	0.8528125692017731	4.8868938632821736e-05
14	0.16394197643917094	0.8473513317200372	6.47211620877e-05
15	0.16783124123997992	0.8411717431846869	7.069026848149957e-05
16	0.08207332754942438	0.9292733907769029	-0.00017184479799938558
17	0.06835881681490934	0.9817268112114673	-5.642268903571737e-05
18	0.0201131839169317	0.9810540233135495	-1.1106002464527157e-05
19	0.10271224226061375	0.9817789215749898	1.4064600042193425e-06
20	0.07562875170074336	0.9815251348189385	-0.00018558975635644614
21	0.15295070616242756	0.8828066019533516	7.562568935538806e-05
22	0.24150450347156963	0.8032474948249007	5.809429709950909e-05

Table 16: LUAD Chromosome Regression Summary

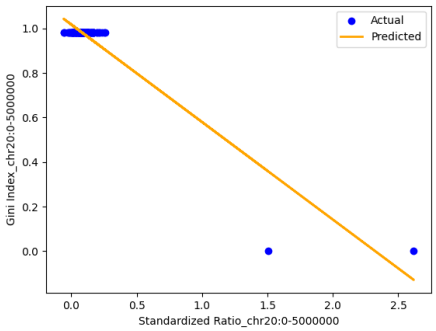


Figure C1: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:0-5000000 for control samples

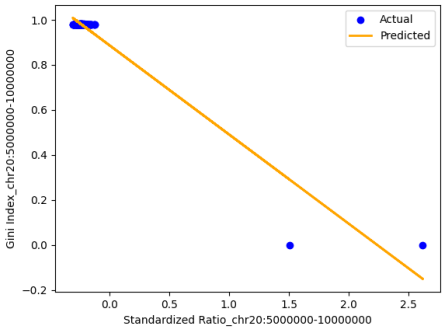


Figure C2: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:5000000-10000000 for control samples

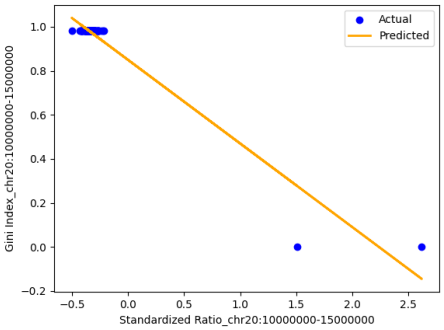


Figure C3: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:10000000-15000000 for control samples

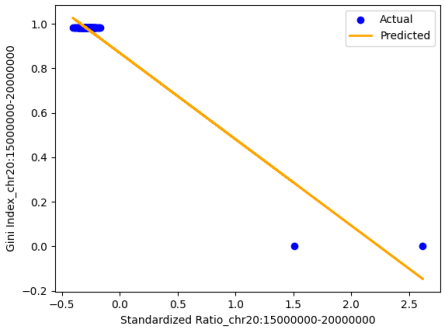


Figure C4: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:15000000-20000000 for control samples

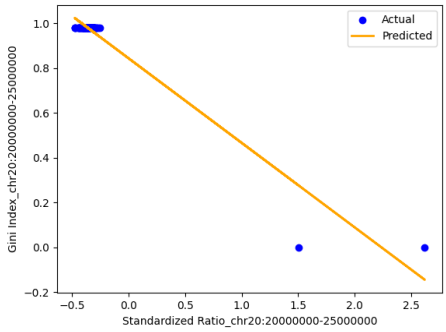


Figure C5: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:20000000-25000000 for control samples

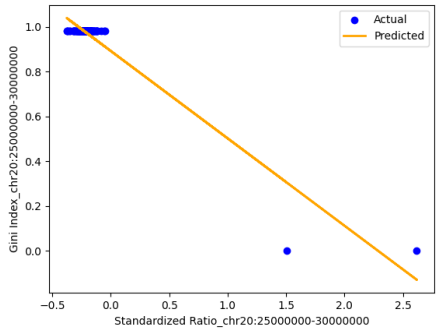


Figure C6: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:25000000-30000000 for control samples

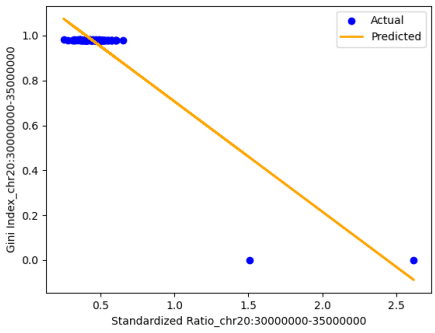


Figure C7: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:30000000-35000000 for control samples

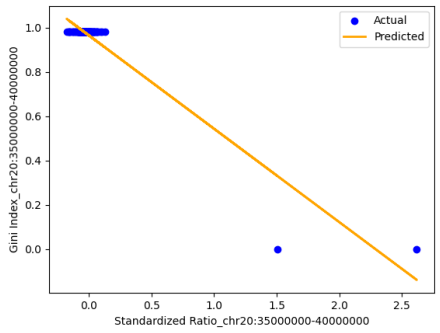


Figure C8: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:35000000-40000000 for control samples

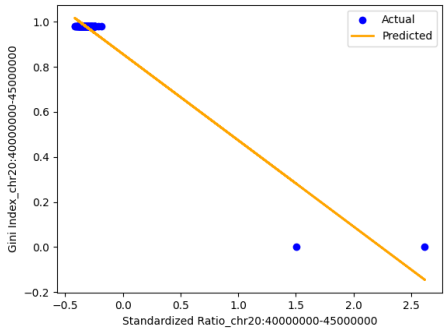


Figure C9: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:40000000-45000000 for control samples

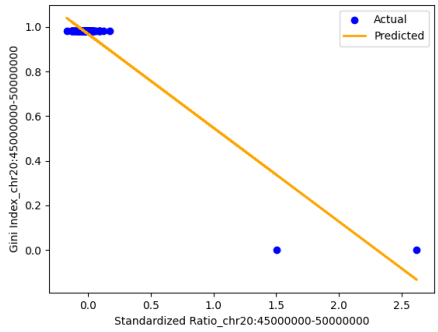


Figure C10: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:45000000-50000000 for control samples

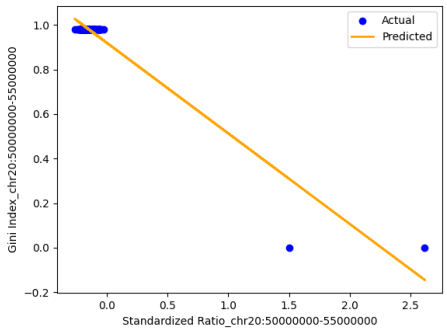


Figure C11: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:50000000-55000000 for control samples

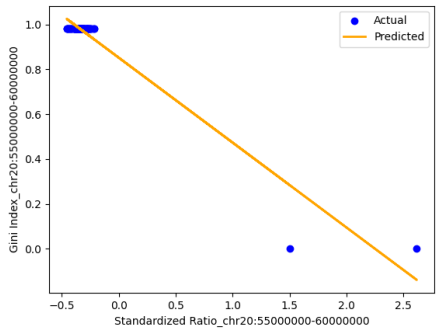


Figure C12: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:55000000-60000000 for control samples

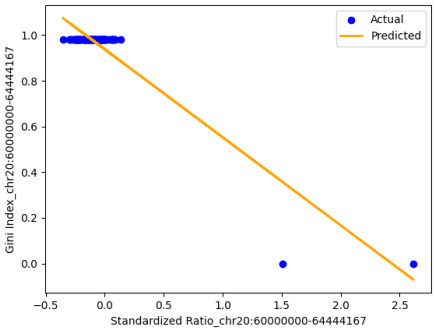


Figure C13: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr20:60000000-64444167 for control samples

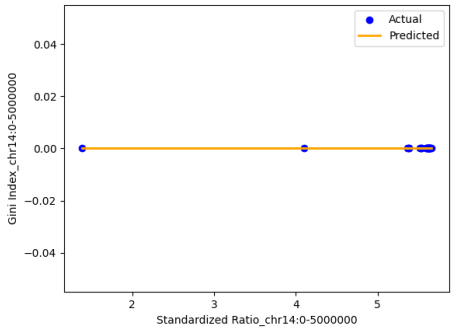


Figure C14: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:0-5000000 for CRC samples

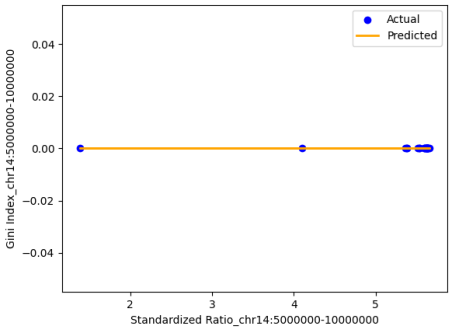


Figure C15: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:5000000-10000000 for CRC samples

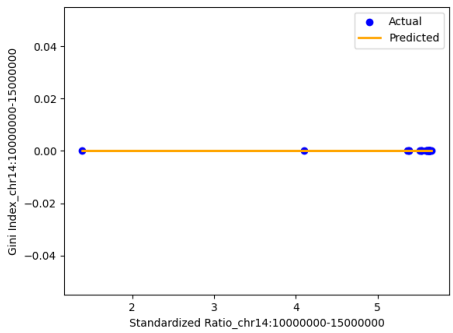


Figure C16: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:10000000-15000000 for CRC samples

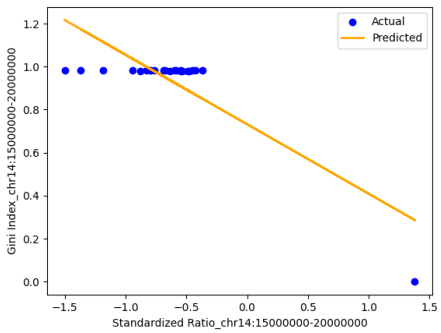


Figure C17: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:15000000-20000000 for CRC samples

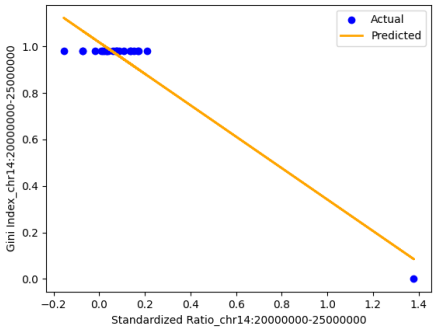


Figure C18: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:20000000-25000000 for CRC samples

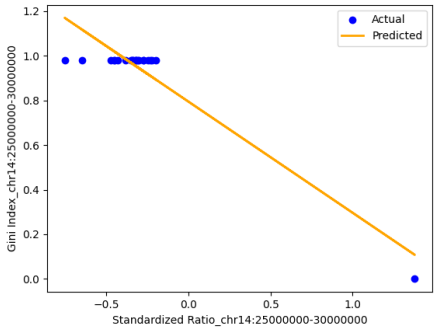


Figure C19: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:25000000-30000000 for CRC samples

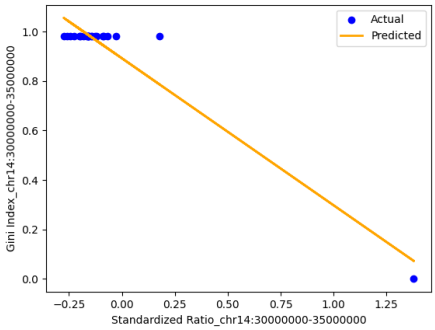


Figure C20: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:30000000-35000000 for CRC samples

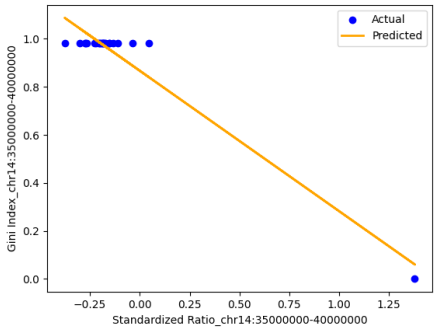


Figure C21: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:35000000-40000000 for CRC samples

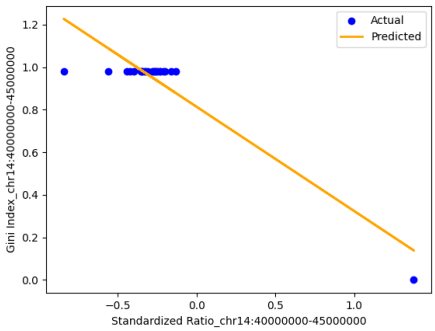


Figure C22: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:40000000-45000000 for CRC samples

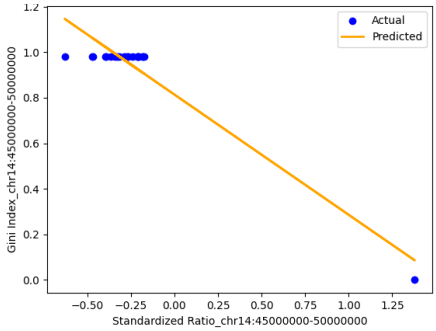


Figure C23: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:45000000-50000000 for CRC samples

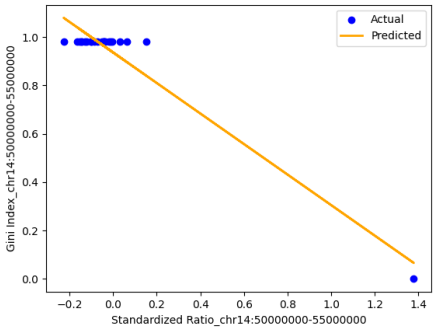


Figure C24: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:50000000-55000000 for CRC samples

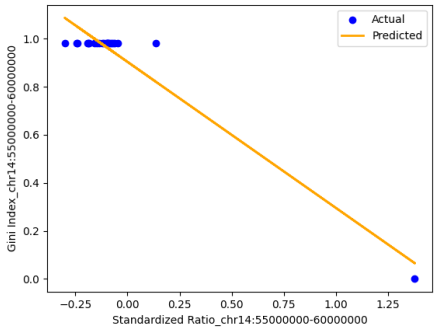


Figure C25: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:55000000-60000000 for CRC samples

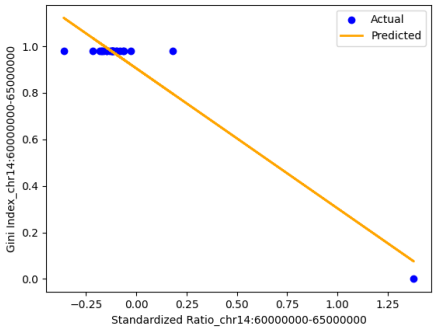


Figure C26: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:60000000-65000000 for CRC samples

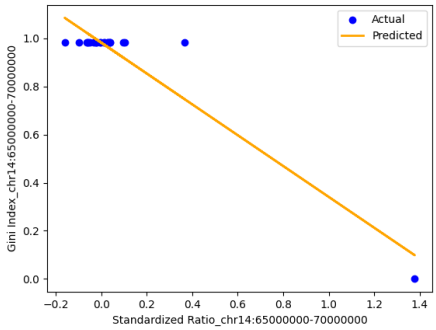


Figure C27: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:65000000-70000000 for CRC samples

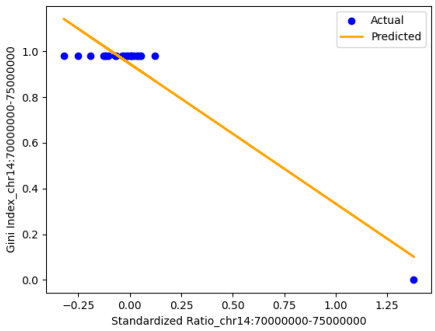


Figure C28: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:70000000-75000000 for CRC samples

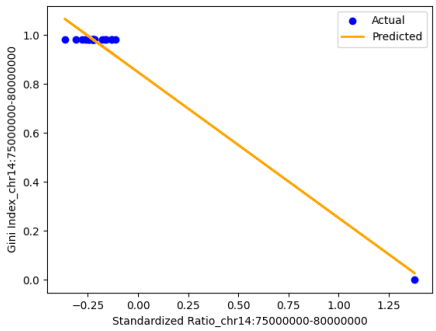


Figure C29: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:75000000-80000000 for CRC samples

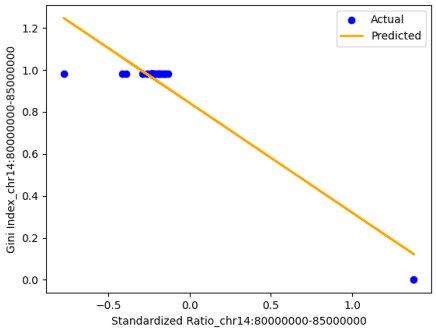


Figure C30: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:80000000-85000000 for CRC samples

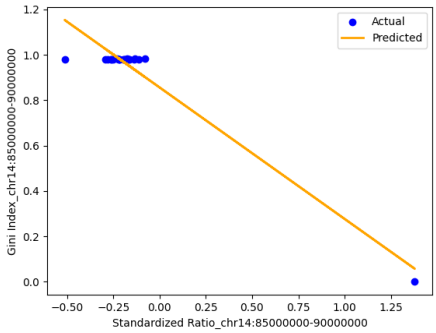


Figure C31: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:85000000-90000000 for CRC samples

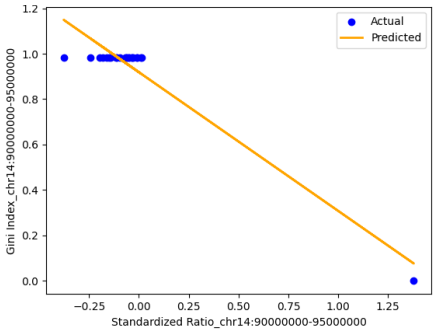


Figure C32: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:90000000-95000000 for CRC samples

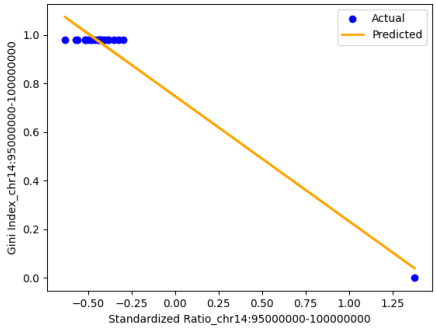


Figure C33: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:95000000-100000000 for CRC samples

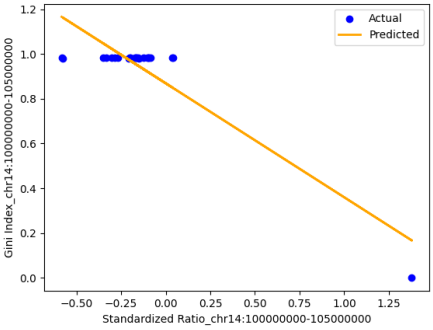


Figure C34: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:100000000-105000000 for CRC samples

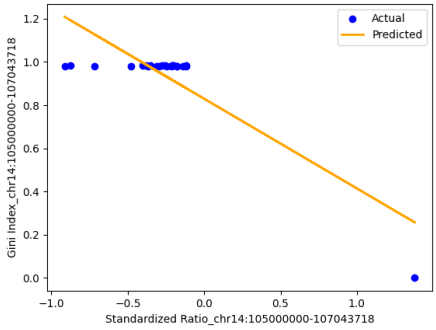


Figure C35: Linear regression plot showing the relationship between the short-long ratios and Gini Index for the bin chr14:105000000-107043718 for CRC samples

D AI Prompts

Multi-omics data refers to the integration and analysis of data from multiple "omes" within a biological system. These "omes" represent different layers of biological information, such as:

- **Genomics:** The complete set of an organism's DNA, including genes and their variations.
- **Transcriptomics:** The complete set of RNA molecules transcribed from DNA, reflecting gene expression levels.
- **Proteomics:** The complete set of proteins expressed by an organism, including their modifications and interactions.
- **Epigenomics:** Chemical modifications to DNA or histones that affect gene expression without altering the DNA sequence.
- **Metabolomics:** The complete set of small-molecule metabolites (e.g., sugars, lipids, amino acids) within an organism.
- **Microbiomics:** The study of the microorganisms (bacteria, viruses, fungi) living in a particular environment, like the human gut.

By combining data from multiple "omes," researchers can gain a more comprehensive understanding of the complex interactions and processes occurring within a biological system. This holistic approach is essential for understanding health and disease, as changes at one level (e.g., genetics) can often have ripple effects at other levels (e.g., protein expression or metabolism).

Applications of multi-omics data analysis include:

- **Disease diagnosis and prognosis:** Identifying biomarkers that can help diagnose diseases earlier or predict their progression.
- **Drug development:** Discovering new drug targets and understanding how drugs interact with multiple biological pathways.
- **Personalized medicine:** Tailoring treatments to individuals based on their unique genetic makeup and molecular profiles.
- **Agriculture:** Understanding how plants respond to environmental stressors and developing crops with improved traits.
- **Environmental monitoring:** Assessing the impact of pollutants on ecosystems by analyzing changes in microbial communities or gene expression.

Multi-omics data analysis is a rapidly growing field, and advancements in technology are making it increasingly feasible to collect and integrate large amounts of diverse biological data. This holds great promise for improving our understanding of complex biological systems and developing new therapies and diagnostics for a wide range of diseases.

Figure D1: What is multi–omics data?

Absolutely! Here's a basic workflow for how you can analyze cfDNA end motifs using Python. We'll focus on identifying simple sequence over-representation.

Assumptions

- You have a BAM file containing cfDNA sequencing reads.
- You're interested in basic motifs, like the prevalence of dinucleotides (e.g., "CG", "AT") at the start and end of fragments.

Outline of Python Script

Python

```python
import pysam

motif_length = 2  # Length of motifs to analyze
bam_filename = "your_cfDNA.bam"
bam_file = pysam.AlignmentFile(bam_filename, "rb")

end_motifs = {
    'start': {},
    'end': {}
}  # Dictionaries to store motif frequencies

for read in bam_file.fetch():
    if not read.is_paired or not read.is_proper_pair:
        continue  # Focus on paired-end reads

    start_motif = read.query_sequence[:motif_length]
    end_motif = read.query_sequence[-motif_length:]

    # Update start motif counts
    if start_motif in end_motifs['start']:
        end_motifs['start'][start_motif] += 1
    else:
        end_motifs['start'][start_motif] = 1

     # Update end motif counts (similar logic as above)
     if end_motif in end_motifs['end']:
        end_motifs['end'][end_motif] += 1
    else:
        end_motifs['end'][end_motif] = 1


bam_file.close()

# Print or further analyze the 'end_motifs' dictionary
print(end_motifs)
```

Use code with caution.

Explanation

1. **Setup:** Install pysam, set `motif_length` as desired, and provide your BAM file path.
2. **Dictionaries for Counting:** `end_motifs` will store the counts of each observed motif at the start and end positions of fragments.
3. **Paired-End Focus:** We filter for properly paired reads, as they provide more reliable fragment end inference.
4. **Extracting Motifs:** The script grabs the first/last `motif_length` bases from the read's sequence.
5. **Counting:** Motif frequencies are incremented in the dictionaries.
6. **Output:** The `end_motifs` dictionary now contains the counts of each unique dinucleotide (or longer motif) found at the beginning and end of your cfDNA fragments.

Figure D2: How to view end motifs using python?