THE IMPACT OF PRE-PROCESSING DATA ON FRAGMENTOMICS ANALYSIS USED IN CANCER SCREENING

Author: Mirko Boon <msboon@student.tudelft.nl>
Responsible professor: Marcel Reijnders
Supervisors: Daan Hazelaar, Bram Pronk, Stavros Makrodimitris

TU Delft

Introduction

- Cancer is one of the leading causes of death, causing nearly one out of 6 deaths [1].
- Screening the population for cancer is currently very difficult and very expensive.
- DNA fragments can end up in the bloodstream through a variety of biological mechanisms, the primary of which is cell death.
- The characteristics of these fragments differ between healthy people and people who suffer from cancer.
- This principle has been used in a number of different machine learning models to predict whether or not someone is suffering from cancer, with promising results (sensitivity up to 99%, specificity up to 98%).
- Although the method of fragmentomics analyses is showing promising results, the effects of the pre-processing steps used are not fully known yet.
- MAPQ thresholds found in literature range between 0 and 30.
- GC-bias correction in literature is done using LOESS correction or Deeptools correction

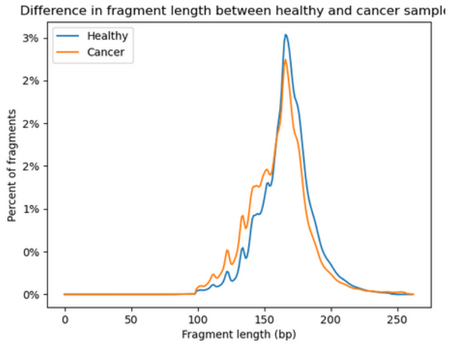

Fig 1. Fragmentation profile healthy and cancer sample
Graph showing fragmentation size profiles of healthy people compared to people who have cancer

Research Questions

The main research question is: "What is the impact of different
pre-processing steps and pre-analytical values on fragmentomics analysis?"

To answer this question, a set of sub-questions was created:

- How does changing the minimum mapping quality influence the fragmentomics analysis?
- How does GC-bias correction influence the fragmentomics analysis?

References

[1] World Health Organization, "Cancer," World Health Organization, 2022. https://www.who.int/news-room/fact-sheets/detail/cancer

Method

- MAPQ was tested for the values of 5, 20 and 30.
- GC-correction was done using Deeptools correction, LOESS correction applied to all fragments collectively and LOESS correction applied separately for short and long fragments.
- For all sub-questions, short/long ratios were calculated in 5Mb bins.
- These bins were compared to the unprocessed data using the KS-test.
- Furthermore, a median healthy profile was created using 30 healthy samples.
- Correlations between sampld and the median healthy profile were analyzed.
- A 1-NN classifier was used to predict whether samples are from patients with cancer.
- Percentage of reads for different MAPQ values will be determined, as well as the distribution of MAPQ in healthy samples and samples belonging to people with cancer.

Results (GC-bias correction)

- Correcting GC-bias using the LOESS whole and LOESS separate methods improves accuracy when predicting cancer using a 1-NN classifier based on Spearman correlation, while using the Deeptools method reduces the accuracy.
- The LOESS separate method produces the best results.
- IQR of the correlation between healthy samples and the healthy median profile is much lower when the LOESS separate method is used.
- Correlations between the healthy samples and the median profile are higher than the correlations between the cancer samples for all methods.
- IQR of the correlation with the median profile is in general much lower in healthy samples compared to cancer samples for all methods.
- KS-test statistics show that data is most transformed by applying the LOESS separate method and is least transformed by applying the LOESS whole method on it, with the Deeptools method lying in between.
- While difference in median correlation between healthy and cancer samples is highest for the LOESS whole method, the LOESS separate method outperforms it in classification.

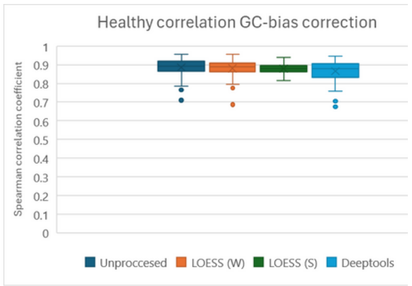

Fig 2. Correlation healthy profile for healthy samples
Boxplot showing correlation between healthy samples and the median healthy profile for different GC-bias correction methods.

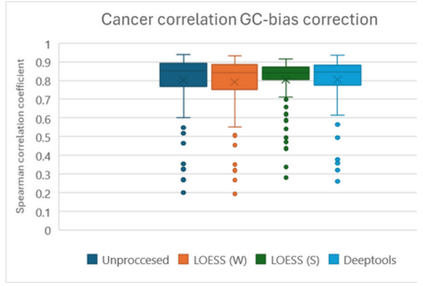

Fig 3. Correlation healthy profile for cancer samples
Boxplot showing correlation between cancer samples and the median healthy profile for different GC-bias correction methods.

	No processing	LOESS (W)	LOESS (S)	Deeptools
Accuracy	77.8%	83.3%	91.7%	69.4%
Specificity	74.4%	76.7%	90.7%	65.1%
Sensitivity	82.8%	93.1%	93.1%	75.9%

Table 1. 1-NN results for different GC-bias correction methods. Results are for a test set of 29 healthy samples and 43 cancer samples.

Results (MAPQ)

- Not filtering on MAPQ improves accuracy when predicting cancer using a 1-NN classifier based on Spearman correlation.
- Raising the MAPQ threshold used reduces both the specificity and sensitivity.
- Correlations between the healthy samples and the median profile are higher than the correlations between the cancer samples for all MAPQ values.
- IQR of the correlation with the median profile is in general much lower in healthy samples compared to cancer samples for all MAPQ values.
- KS-test statistics show that cancer samples are transformed much more compared to healthy samples when filtering on MAPQ.
- For healthy samples, almost no reads have a MAPQ lower than 5, while for cancer samples about 5% of the reads are below MAPQ 5.

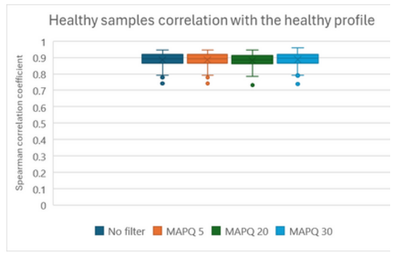

Fig 2. Correlation healthy profile for healthy samples. Boxplot showing correlation between healthy samples and the median healthy profile for different MAPQ thresholds.

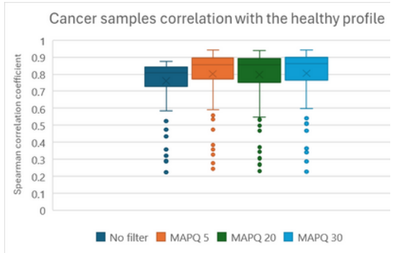

Fig 3. Correlation healthy profile for cancer samples. Boxplot showing correlation between cancer samples and the median healthy profile for different MAPQ thresholds.

	No filter	MAPQ 5	MAPQ 20	MAPQ 30
Accuracy	97.3%	69.3%	64.0%	62.7%
Specificity	95.6%	77.8%	71.1%	71.1%
Sensitivity	100.0%	56.7%	53.3%	50.0%

Table 2. 1-NN results for different MAPQ filtering thresholds. Results are for a test set of 30 healthy samples and 45 cancer samples.

Conclusion

- Pre-processing data can have a heavy impact on the framentomics analysis.
- Correcting GC-bias using one of either LOESS methods improves classification results, while results deteriorate when using Deeptools correction.
- Having a more similar distribution of ratios before and after processing does not seem to indicate a better classification performance.
- A larger difference in median correlation for healthy and cancer samples does not necessarily imply better classification results.
- Filtering on MAPQ changes the distribution of cancer samples more than it changes healthy samples.
- Healthy samples have no fragments with a MAPQ < 5, while about 5% of the fragments in cancers samples have a MAPQ < 5.
- Not filtering on MAPQ leads to significantly improved results for the classification task.