Understanding the effect of pre-processing methods on fragmentomics analysis

Studying the effects of GC-correction and MAPQ filtering on fragmentomics analysis when using short/long ratios

Mirko Sander Boon[1]

Supervisors: Marcel Reinders[1]**, Bram Pronk**[1]**,**
Daan Hazelaar[2]**, Stavros Makrodimitris**[2]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2]**Erasmus Medical Center, 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: Mirko Sander Boon
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Daan Hazelaar, Johan Pouwelse

Abstract

Cancer is one of the leading causes of death. To reduce the amount of deaths caused by cancer, a number of different screening methods are used to detect cancer in an earlier stage, to improve survival rates when treating patients with cancer. Current screening methods are often invasive, costly and not very accurate. Therefore, new methods are being sought that aim to be cheaper, less invasive and provide more accurate results. One of these methods is fragmentomics. Multiple methods have been proposed to use fragmentomics analysis in the context of screening for cancer, including using the short/long ratio as well as investigating the nucleotides at the ends of the fragments. Across previous works using fragmentomics analysis to predict cancer, different pre-proccessing steps are used, with limited explanation why the pre-processing methods were chosen. Research into the effects of pre-processing steps used when using fragmentomics analysis is lacking. Two main pre-processing steps in the field are correcting GC-bias and filtering on MAPQ. Here we investigated the impact of three GC-correction methods by applying the correction method and then analyzing the resulting fragmentation profiles using short/long fragment ratios. Furthermore, three different MAPQ filtering thresholds were studied. This showed that Deeptools correction of the GC-bias lowered performance, with the accuracy dropping from 77.8% to 69.4%. Applying LOESS correction using all fragments at the same time resulted in an accuracy of 83.3%, while applying LOESS correction using the short and long fragments separately resulted in an accuracy of 91.7%. The impact of filtering the data based on mapping quality was determined by comparing the results of analysing all fragments, analyzing only fragments with mapping quality 5, 20 or 30. This showed that not filtering by mapping quality has a big impact on the profiles of cancer samples, with a KS-test statistic of 0.08 for MAPQ 5 and MAPQ 20 and larger differences in correlations between healthy and cancer samples. The performance of classification was much higher when not filtering, with an accuracy of 97.3%, which dropped whenever the filtering threshold was raised, bottoming out at 62.7% for a threshold of MAPQ 30. Due to limitations with the study, the combined pre-processing of not filtering on MAPQ and using the LOESS separate correction were not studied.

1 Introduction

Cancer is one of the leading causes of death, causing nearly one out of 6 deaths [1]. In an attempt to reduce the amount of people dying due to cancer, a number of tests to screen for cancer have been devised in the last decades, with varying amounts of succes. One example is a colorectal screening program in the Netherlands forecasted to lead to the prevention of 2900 deaths annually [2]. Most of these screening methods have certain drawbacks, which can be the cost or the invasiveness of the procedure or their accuracy. There is a constant search for new methods, which are less invasive, cheaper or more accurate. One of these potential new methods is fragmentomics analysis.

Fragmentomics is a rapidly evolving field studying fragmentation patterns in cell free DNA (cfDNA). cfDNA is released into the bloodstream through various mechanisms, the primary mechanism is believed to be cell death [3]. cfDNA carries both genetic and epigenetic information about its tissue of origin. In recent years evidence was found that these genetic and epigenetic characteristics are different depending on whether an individual is healthy or has cancer [4]. An example of this can be found in figure 1 showing the fragments lengths of a healthy individual and an individual with colorectal cancer.

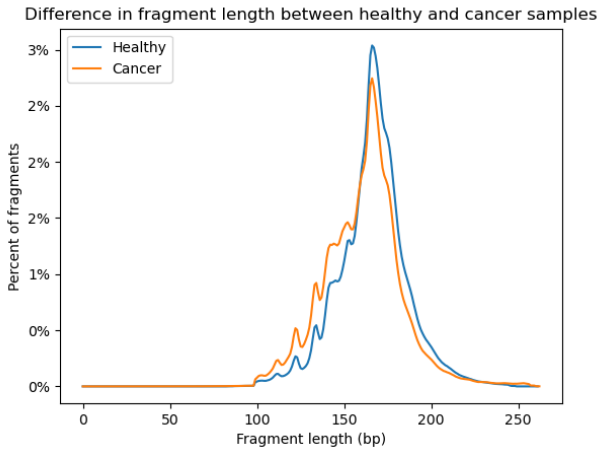


Figure 1: **Fragment lengths of healthy sample compared to a colorectal sample.**

The method works by drawing a blood sample from the subject. The cfDNA is then sequenced, and the position of the fragment compared to the reference genome is determined. This leads to a dataset containing the location as well as the genetic code of all the fragments found. Each processed fragment is called a read.

Fragmentomics analysis is showing promising results, a classifier trained using the ratio between short (150 base pairs or less) and long (151 base pairs or more) to screen for cancer had sensitivities up to 99% with a specificity of up to 98% [5]. Another recent method is using the last three base pairs. Multiple combinations where found which are more prevalent in people with cancer as well as patterns which were less frequent in people with cancer. Training a model with this feature showed a 72% detection at 95% specificity [6].

While many methods are being tested to determine the possibilities of using fragmentomics as a biomarker to predict cancer, research on the impact of pre-processing the data as well as the effects of pre-analytical values on the predictive qualities of machine learning models designed to screen for

cancer is currently lacking. This study will attempt to answer some questions with regards to the effects of pre-processing data used in fragmentomics analysis. The main question to be answered is: What is the impact of different pre-processing steps and pre-analytical values on fragmentomics analysis?

There are a large number of pre-processing steps used in fragmentomics, as well as a large number of pre-analytical values that might influence the fragmentomics analysis. Due to the limited scope, a selection of two pre-proccesing steps was made. The pre-processing steps chosen to be analyzed were the GC-correction and filtering based on mapping quality.

GC-Bias

Due to the nature of sequencing, fragment abundance may be inaccurate due to GC-bias [7]. To receive more accurate counts, multiple methods exist to correct this bias. One such method used in the DELFI paper is using LOESS regression to determine what amount of coverage is explained by the GC-content and then subtracting this explained coverage from the found coverage [5]. The DELFI paper does this separately for short and long fragments, loosely based on work done by Benjamani & Speed [8]. Another method to correct GC-bias is using the Deeptools libraries functions computeGCBias and correctGCBias. These methods implement the methods proposed in the Benjamani & Speed paper [8].

We evaluated the impact of correcting GC-bias using the Deeptools GC-correction algorithm, the LOESS method as described in the DELFI paper and the LOESS method applied to all fragments equally. This was done to answer the question: How do different GC-correction methods influence the downstream fragmentomics analysis using short/long ratios.

MAPQ

When aligning the fragments, there is a chance that the fragment is not aligned at the correct place on the reference genome. This uncertainty is expressed in a MAPQ score. This MAPQ score is calculated using the following formula:

$$-10log10(p)$$

Where p is the probability that the alignment is misaligned. A MAPQ of 30 thus corresponds a 99.9% certainty that the alignment is correct. A MAPQ of 20 corresponds to a certainty of 99%. And a MAPQ of 5 corresponds to a certainty of roughly 68%.

In the literature a number of different MAPQ values were found that are used in fragmentomics analysis for cancer screening. The DELFI [5] paper uses a MAPQ of 30, and the 4-MER[9] paper uses a MAPQ of 20 while the FREIA[6] paper uses a MAPQ of 5. To answer the question of the effect of different MAPQ values when filtering the data on the downstream fragmentomics analyisis these three MAPQ values were chosen to study.

2 Methodology

The data that will be used for all experiments is a subset of the data published by the DELFI paper [5]. This subset contains 50 samples of people who have breast cancer, 113 samples of healthy people, 26 samples of people with colorectal cancer and 75 samples of people with lung cancer. 19 lung samples were collected after the patient had undergone treatment, all other samples where drawn before the patient underwent treatment [5]. Due to problems with file corruption the actual amount of samples used differ between experiments. All the data used was anonymized.

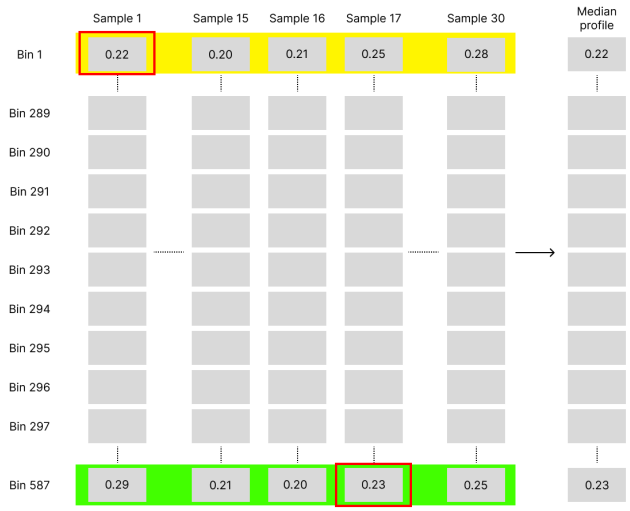


Figure 2: **Schematic showing creation of median profile.** First the median value for bin one gets found for samples 1-30. Then this median value is chosen for the median profile. In this case, the median value is 0.22 belonging to sample 1, so the value for bin 1 of the median profile is 0.22. This process is repeated for all bins. So for the last bin, the median value found is 0.23, belonging to sample 17. In this case the value from sample 17 is chosen for the median profile, thus the value of bin 587 of the median profile is 0.23.

For all sub questions we created a fragmentomics profile for each sample, where the short/long ratio was calculated for 5Mb windows. For all experiments, reads that were unpaired or were secondary aligned were filtered out. All reads that are not in the autosomal chromosomes were filtered out as well. We Z-score normalized all the ratios to reduce the effect of outliers. To create the median healthy profile 30 samples were taken at random. Then a median profile was created by taking the median of all samples per bin. A schematic of this process can be found in figure 2. Two median profiles were created, one for the experiments concerning GC-bias, and one for the MAPQ experiment. The profile consists of one column which is unprocessed, and three columns for the different methods used in case of the GC-bias profile, and also three columns for the different MAPQ filtering thresholds used.

We calculated the correlation between the healthy median profile and the sample using the Spearman correlation method, this was only done for samples which are not part of the 30 selected samples for the creation of the median profile. A KS-test was used to determine the difference between the profiles before and after processing. To evaluate the effects on predicting cancer, a simple 1-NN classifier was made, where the class assigned is equal to the closest profile according to the Spearman correlation using Euclidean distance. A train/test split was made, with a 70/30 split.

GC-Bias correction

All reads with a MAPQ ≤ 30 were filtered out. All chromosomes were divided up into bins of 100 kilobase (Kb). Per bin, the amount of short and long fragments was collected, as well as the average GC-content of that bin. If a bin had less than 10 short or long fragments, it was disregarded for future analysis. If the length of a bin was less than 100 Kb, due to it being the last bin of a chromosome, counts were scaled based on what percentage of 100 Kb was covered by the bin. I.e. if the bin only contained 50 Kb, both short and long count would be scaled by a factor of 2.

For the LOESS correction based on short and long at the same time, the short and long counts are summed to create a total count per bin. The bottom 1 percent and top 1 percent were excluded for creating the LOESS regression curve, to reduce the effect of outliers. A LOESS regression curve was created, with the counts on the y-axis and the average-gc content on the x-axis, with a span of 0.75. For all bins, the prediction of coverage explained by the GC-content obtained by the LOESS curve was subtracted from the coverage of that bin. To return the coverage to the original scale, the median count of all the bins was added. By dividing this by the original count a scale value was created. The same process was done for short and long counts as well.

This resulted in 3 scale values per 100 Kb bin, one to scale both short and long with at the same time, one separately for short and one separately for long. Spearman correlation between GC and short coverage as well as between GC and long coverage was calculated before the LOESS regression correction, as well as for the correction that was executed on both short and long fragments at the same time and for the separate corrections to determine the effectiveness of the GC-correction.

Afterwards, the chromosomes were divided up into 5 Mb bins. Short to long ratios were calculated per bin, one using just the raw counts, one where both short and long fragments were scaled according to the scale value found for their 100 Kb window using LOESS regression and one where short and long ratios were scaled separately using the scales found when doing the LOESS correction separately.

For the Deeptools correction, the computeGCBias and correctGCBias commands were used to create a GC-corrected BAM file. For this file the Spearman correlation between GC and short and long coverage were calculated. Furthermore the short/long ratios were calculated for 5 Mb windows. This was done for all samples to create fragmentation profiles.

MAPQ filtering

All chromosomes were divided up into bins of 5 megabase (Mb). Per bin, ratios of short to long were calculated for all reads, all reads with a MAPQ ≥ 5, all reads with a MAPQ ≥ 20 and all reads with a MAPQ ≥ 30. If the amount of short or long fragments was less than 10, the ratio for that bin is set to NaN to be disregarded in further analysis. This was done for all samples to create fragmentation profiles.

3 Results

GC-Bias correction

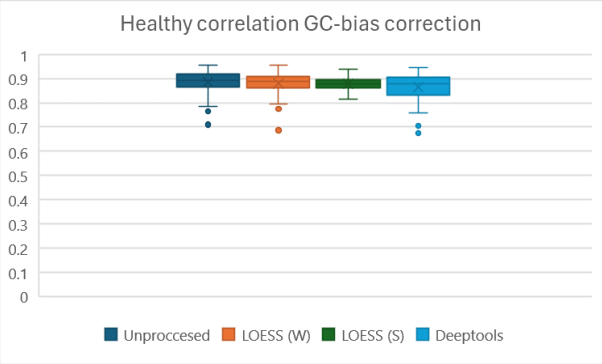
Figure 3a shows that for all GC-bias correction methods healthy samples correspond closely to the healthy median profile. With the correlation being slightly higher for the unproccesed data and the LOESS whole method compared to the LOESS seperate method and the Deeptools method. However, for all methods except the LOESS seperate method there were some profiles with very low correlation to the healthy median profile, which are classified as outliers in figure 3a. The minimum correlations for these methods was arround 0.70. The minimum correlation for the LOESS seperate was much higher, with a minimum correlation found of 0.83. Interestingly, the profile that was the minimum correlation found was the same when correcting using Deeptools, the LOESS whole method or not correcting at all. However, the method that outperforms the others in minimum correlation is the only method with a different profile that is the minimum.

Figure 3b shows that correlations between the healthy median profile and the cancer samples are lower compared to the correlations between the healthy median profile with the healthy samples found in 3a. A median correlation of 0.85 was found when no processing was done as well as when Deeptools processing was executed. A median correlation of 0.84 was found for both LOESS methods. As can be seen in figure 3b a lot more samples were classified as outliers and the inter quartile ranges are larger compared to the healthy samples. Correlations for the separate cancer types can be found in A.

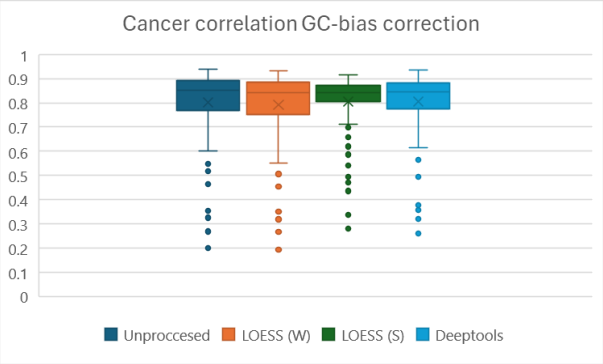
Table 1 shows that the LOESS method correcting both short and long fragments at the same time leads to a KS-statistic of 0.04 for both cancer and healthy samples. The KS-statistic is higher for both the LOESS seperate and the Deeptools method, with 0.13 and 0.09 being the statistic for the LOESS seperate method for healthy and cancer samples respectively and the Deeptools method has a KS-statistic of 0.06 for healthy samples and of 0.10 for cancer samples.

	LOESS (W)	LOESS (S)	Deeptools
Healthy (N=98)	0.04	0.13	0.06
Cancer (N=103)	0.04	0.09	0.08

Table 1: **Median KS-statistic between the methods and the unprocessed sample.**



(a) Healthy correlation with median profile



(b) Cancer correlation with median profile

Figure 3: **Spearman correlation for healthy samples 3a and cancer samples 3b with the median healthy profile for different GC-bias correction methods.** Box-whisker plot showing the Spearman correlation for 70 healthy samples and 145 cancer samples. The x mark inside the box shows the mean of the data, the line shows the median. The top whisker shows either 1.5x the distance of the IQR from the third quartile or the maximum value, while the bottom whisker shows either 1.5x the distance of the IQR from the 1st quartile or the minimum value. Points falling outside 1.5x the IQR are plotted as dots

Using 1-NN classifying based on the closest sample according to the Spearman correlation gives some interesting results. A test train split of 70/30 was used, leading to 69 healthy training samples and 101 cancer training samples as well as 29 healthy test samples and 43 cancer test samples. As can be seen in 2 the LOESS separate method has the highest accuracy and specificity at 91.7% and 90.7% respectively and is tied for the highest sensitivity at 93.1%. Applying the Deeptools correction actually worsens performance, with the accuracy dropping from 77.8% when not correcting for GC to 69.1% when using the Deeptools method.

	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 2: **1-NN results for different GC-correction methods.** Nearest neighbours were determined using the Spearman correlation between profiles. Results are from a test set containing 29 healthy samples and 43 cancer samples.

For some additional analysis, Spearman correlation between coverage and average GC-content was calculated before and after applying the methods. The results can be seen in figure 4. The figure shows that all methods are effective at reducing GC-bias to some extent, as the correlations are closer to 0 after correction compared to before. Correcting the GC-bias using Deeptools leaves more GC-bias than with the other methods. Correcting GC-bias for all fragments simulteanously using the LOESS method leads to good results for long fragments, but for short fragments there is still some bias left. Doing the LOESS correction separately for short and long fragments leads to low GC-bias for both short and long fragments.

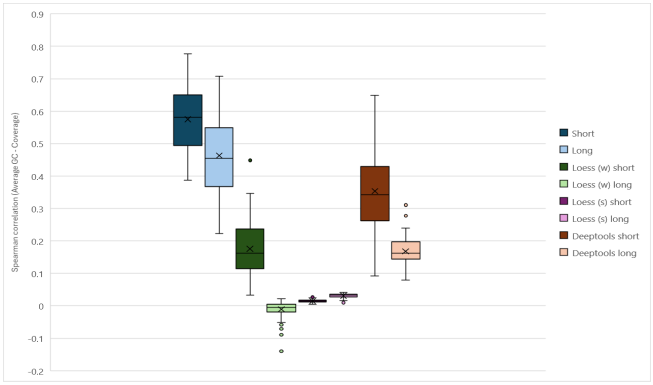


Figure 4: **Spearman correlation before and after the various methods for healthy patients (N=63).** Correlation is shown for both short and long fragments separately. The blue data shows correlations before processing, the green data shows correlations using the LOESS method applied to all fragments simultaneously, the purple fragments show the correlations after LOESS correction done separately and the brown fragments the correlations after Deeptools GC-correction.

MAPQ filtering

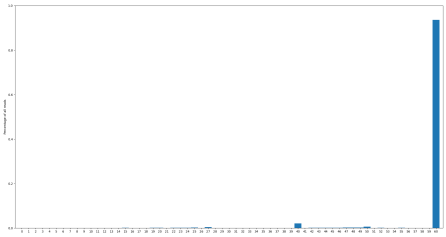
When not filtering on MAPQ, the median amount of fragments retained from the unfiltered sample was 100% for healthy samples and 94.2% for samples from patients with cancer. When filtering on MAPQ 20, this drops down to 99.4% and 93.3% respectively and when filtering on MAPQ 30, this additionally drops down to 98.3% and 92.2%.

To gain a better understanding of how MAPQ is distributed for cancer samples and healthy samples, histograms of the MAPQ distribution were plotted. This plot can be found in figure 5. Notably, the only apparent difference between healthy and cancer samples is the lack of fragments with a MAPQ below 5. The rest of the distributions appear to be similar. To gain further inside in the distribution of the low MAPQ fragments, an extra plot was made for the cancer sam-

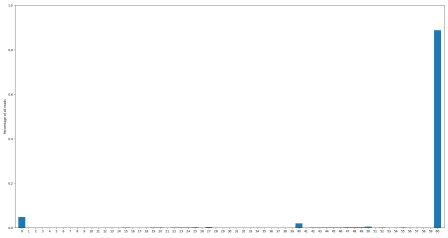
ples showing only the fragments below MAPQ 5. As can be seen in figure 5c, almost all fragments with a MAPQ < 5 have a MAPQ of 0. The average length of fragments with a MAPQ lower than 5 was shorter than the average length of the rest of the fragments for all cancer samples which had fragments with MAPQ < 5. The amount of reads with a MAPQ < 5 was

The location of these fragments with a MAPQ < 5 on the reference genome was investigated. As can be seen in figure 6a, the location where there are relatively more MAPQ < 5 fragments then MAPQ ≥ 5 are clustered together. The bin with the most of these low MAPQ fragments was located in the centre of chromosome 1, with 4.1% of fragments with a MAPQ < 5 being present in this single bin. Plotting just a single chromosome shows that most of the clusters that can be seen in figure 6a are around the centromere. An example of these chromosomes is plotted in figure 6b. The other chromosomes can be found in Appendix C

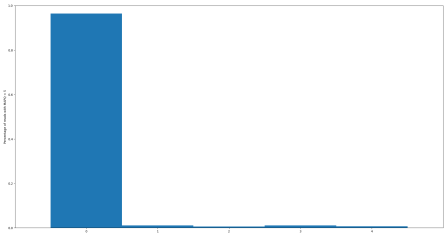
For 10 fragments below MAPQ 5 an evaluation of the origin of the read was done using the BLAST[10] tool. For all these fragments there was a 100% identity and 100% coverage match for at least one chromosome in the human genome.



(a) Healthy



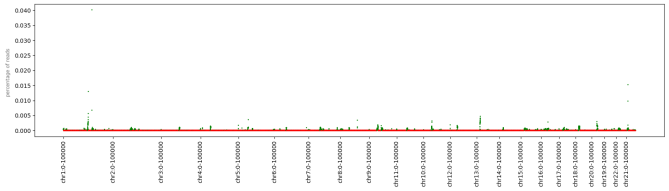
(b) Cancer



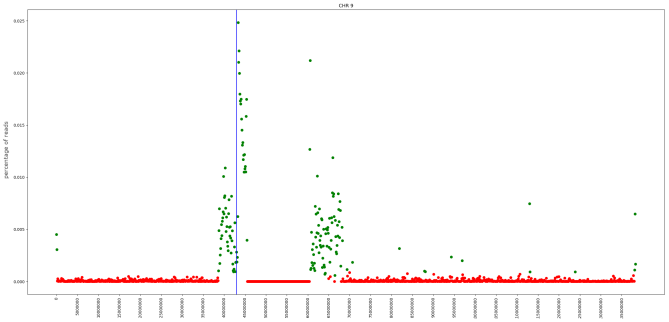
(c) Cancer for MAPQ < 5

Figure 5: **MAPQ distribution showing percentage of all reads in category per MAPQ value.** Maximum MAPQ value found was 60

Figure 7a shows that for all MAPQ values healthy samples correspond closely to the healthy median profile. The IQR of the correlations were also low, with an IQR of 0.05 for all MAPQ values. However, there were some profiles with very



(a) Fragments across the entire genome.



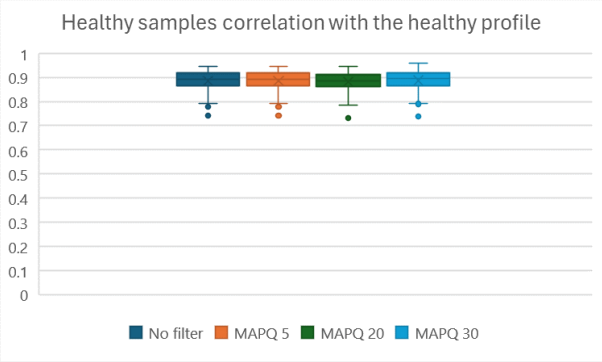
(b) Fragments on the 9th chromosome. The centromere is plotted in blue.

Figure 6: **Location of fragments with a MAPQ lower than five for a single breast cancer sample.** The percentage of reads with a MAPQ < 5 per 100Kb bin compared to the total amount of fragments with a MAPQ < 5 is shown. If the share of fragments with MAPQ < 5 is higher than the share of fragments with MAPQ ≥ 5, the bin is plotted green. If the share of MAPQ < 5 is smaller the bin is plotted in red.

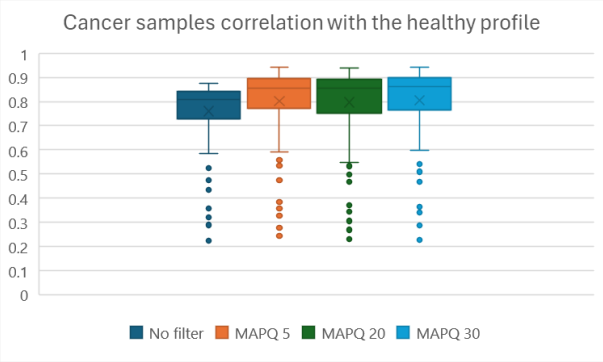
low correlation to the healthy median profile. There were little outliers. Figure 7b shows that correlations between the healthy median profile and the cancer samples are lower compared to the correlation with the healthy samples. A median correlation of 0.81 was found when no filtering was applied, and a median correlation of 0.86 was found for MAPQ 5, 20 and 30. The IQR of the cancer samples was much higher, varying between 0.11 to 0.13. As can be seen in figure 7b a lot of samples were classified as outliers. Correlations for the different cancer types can be found in appendix B.

As can be seen in table 3, median KS-statistics are low for all MAPQ values, ranging from 0.00 to 0.08. Cancerous samples have a higher KS-statistic, with KS-statistics ranging from 0.07 to 0.08, compared to the healthy samples which range between 0.00 and 0.03.

Using 1-NN classifying based on the closest sample according to the Spearman correlation gives some interesting results. A test train split of 70/30 was used, leading to 71 healthy training samples and 106 cancer training samples as well as 30 healthy test samples and 45 cancer test samples. Accuracy's range from 62.7% to 97.3% as can be seen in table 4. Increasing the MAPQ value on which is being filtered leads to lower accuracy, sensitivity and specificity.



(a) Healthy correlation with median profile (N=71)



(b) Cancer correlation with median profile (N=152)

Figure 7: **Spearman correlation for healthy samples (7a) and cancer samples (7b) with the median healthy profile for different MAPQ filtering thresholds**. The x mark inside the box shows the mean of the data, the line shows the median. The top whisker shows the maximum value, while the bottom whisker shows either 1.5x the distance of the IQR from the 1st quartile or the minimum value. Points falling outside 1.5x the IQR are plotted as dots

	MAPQ 5	MAPQ 20	MAPQ 30
Healthy	0.00	0.02	0.03
Cancer	0.08	0.08	0.07

Table 3: **Median KS-statistic between the different MAPQ values and the unfiltered data from 99 healthy and 151 cancer samples.**

	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 4: **1-NN results for different MAPQ filtering thresholds.** Nearest neighbours were determined using the Spearman correlation between profiles.

4 Responsible Research

Ethical considerations

DNA from human subjects is considered to be sensitive data, thus it is important to use it in a safe manner. The data used from the DELFI paper has been anonymized, making it impossible to trace back to the individuals who gave the samples [5]. Furthermore, due to nature of the data being short DNA fragments and not a fully sequenced genome, it is hard to match the DNA collected in this way to a DNA sample collected from a person at some other time. This means the risk of identifying someone with this data using existing DNA records is also small.

The DELFI data set contains data collected from hospitals in the Netherlands, Denmark and the United States of America. This means that data is sampled from a predominantly Caucasian population. A concern with this is that while the methods discussed in this paper may work for regions with similar demographics, it might not be as effective in populations with different demographics. A way to prevent this issue is by using a secondary dataset with a different demographic, however due to the limited scope of this paper and time constraints, this was not possible.

One more ethical risk to consider is the availability. While costs of sequencing are getting lower, it can still be expensive. This could potentially lead to a future were good cancer screening is available only to the rich, increasing the already existing gap in life expectancy caused by income.

Reproducibility

The DELFI data set used in this research is open to the public, allowing access to the data used in this research to anyone who wants to reproduce the results. The code used to obtain the results in this research paper is available at https://github.com/RainingBlue/ResearchProject. Lastly, the method is described extensively in the methodology chapter, including details such as settings used. All of these things together should make this research highly reproducible.

5 Discussion

GC-Bias correction

The median KS-test statistic found was much higher for the LOESS separate method and the Deeptools method compared to the KS-test statistic for the LOESS whole method. Implying that the LOESS method and the Deeptools method are more transformative of the data. The correlation for the healthy samples was slightly lower for the LOESS separete and Deeptools methods compared to the unproccessed samples and the LOESS whole corrected samples. Median correlations with the healthy profile were lower for all methods for the cancer samples, with them being the lowest for both LOESS methods. The difference in median correlation for cancer and healthy samples was largest in the LOESS whole method. This implies that healthy and cancerous samples are most different when using this method.

However the LOESS separate method out performs the LOESS whole method when doing 1-NN classifying. This can be explained by the fact that there were less outliers and that the IQR was smaller for the LOESS seperate method. This means that all healthy samples are relatively close to the median healthy profile. This can be seen due to the fact that while the sensitivity is tied between the LOESS whole and LOESS seperate methods, the LOESS separate method scores a lot higher on specificity. The LOESS separate model is better at identifying the healthy samples.

The smallest difference in median correlation is when using the Deeptools method, meaning that cancer samples and healthy samples closely resemble eachother after processing. This is confirmed by the Deeptools method performing the worst in specificity, sensitivity and accuracy in the 1-NN classfier.

The effect of the GC-bias correction methods on the amount of GC-bias remaining was also determined. This found that GC-bias was lower after applying all methods. GC-bias is close to zero for both short and long fragments when applying the LOESS separate method. LOESS whole performs well on the long fragments, but worse on the short fragments. This is presumably caused due to the fact that there are more long fragments than short fragments present, thus the correction is heavily biased towards correcting the GC-bias for long fragments. The Deeptools methods performs the worst of all GC-bias correction methods, only slighly beating the unprocessed data on the amount of bias.

Although GC-bias was lower in the Deeptools method compared to the unprocessed data, performance in the 1-NN classifier were worse. This suggests that just removing GC-bias will not automatically increase performance in classifying cancer. However, the LOESS methods both significantly out perform the unprocessed data, implying that reducing GC-bias can increase performance when done using certain methods.

MAPQ filtering

For all MAPQ values the KS-test statistcs were close to zero for the healthy samples. Profiles after proccessing are thus similair to the original profile, this makes sense, seeing as at every MAPQ value only a small amount of fragments is filtered out. Unsurprisingly the median correlations between the healthy profiles and the healthy median profile also stay roughly the same. With filtering on MAPQ 30 having a slightly higher median correlation than not filtering or using the other MAPQ values.

For cancer samples, the median KS-test statistic was much larger then it was for the healthy samples for all MAPQ thresholds. This is probably caused by the much higher amount of fragments that is filtered out when filtering on MAPQ. These much higher KS-test statistics also translate to a large change in correlation between the median healthy profile and the cancer samples. Correlation when not filtering is much lower for the cancer samples then it is after filtering using a MAPQ threshold. Filtering on MAPQ brings the cancerous samples thus closer to the healthy samples. This would imply that accuracy drops when predicting cancer when filtering on MAPQ, and indeed, this is the case. For the unfiltered data performance is the best, with accuracies dropping whenever more fragments are filtered out.

A small sample of reads with a MAPQ of less than 5 were checked to determine whether or not these reads were of hu-

man origin. All reads had at least one region of the human chromosome with which there was a 100% match according to BLAST [10]. For all cancer samples it was found that the average length of the fragments with a MAPQ < 5 was lower than for fragments with a MAPQ ≥ 5. Most of the fragments with a MAPQ below 5 had a MAPQ of 0. This might be caused by reads with multiple exact matches. If multiple exact matches occur, one read gets assigned at random. The corresponding MAPQ is then set to 0. Most of the fragments with a MAPQ < 5 were found near the centromeres of the chromosomes. A potential explanation for this is that the centromeres have repeating sequences, making them hard to place. This does not however explain why there were no low MAPQ fragments in healthy samples, however due to time constraints this is left for future research.

These results indicate that filtering out the large amount of fragments that are below a MAPQ of 5 when analyzing cancer samples makes the cancer samples more closely resemble a healthy profile. This in turn increases the difficulty of classifying the samples correctly.

Limitations

Some of the BAM files used were corrupted, and some of the files timed-out when applying Deeptools GC-bias correction. This lead to fewer files to be able to be used than anticipated. Due to only limited data being available, the test and train sets were relatively small with only about 30 test samples for healthy patients and about 45 test samples for patients with cancer. A recommendation would be to repeat this process with more data available.

Due to the limited scope of this paper, no validation of the found results was made using proper models to predict cancer such as linear regression or gradient boosting machine. The data used only contains samples of subjects from the Netherlands, Denmark and the USA, which could potentially lead to biases in the results found. Future research should aim to include data from different geographical regions.

Due to the limited scope of this research, only the normalized short/long ratios were used to determine the effects of pre-proccesing the data. This is only one of several metrics that are used in fragmentomics analysis.

6 Conclusions

Pre-processing data can have a heavy impact on the downstream fragmentomics analysis. Correcting for GC-bias using the LOESS seperate and LOESS whole methods improves accuracy when predicting cancer, while correcting GC-bias using Deeptools decreases this performance.

The difference in distributions does not appear to play a big role in determining the effectiveness of the method in both GC-correction as well as in predicting cancer. KS-test statistics of the LOESS whole method were the lowest, while the LOESS separate method were the highest, however, they both outperformed the Deeptools method that has KS-test statisics inbetween the two LOESS methods.

The difference between the healthy and cancer samples median correlation does not appear to be the only thing influencing classifying capability. The LOESS whole method had the largest difference between healthy and cancer samples, but was outperformed in specificity by the LOESS separate method. This could potentially be caused by the fact that the variability of the LOESS separate method is much lower, however, this should be studied in future research.

Filtering on MAPQ has an even bigger effect. Not filtering on MAPQ leads to a much higher accuracy then filtering with a MAPQ threshold of 30. Filtering on MAPQ only has a limited effect on the healthy samples, while there is a much larger difference in the cancer samples.

A Correlation of different cancer types for different GC-correction methods

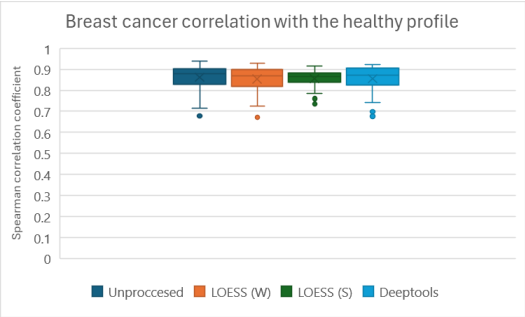


Figure 8: Spearman correlation between the healthy median profile and breast cancer samples for different GC-correction methods.

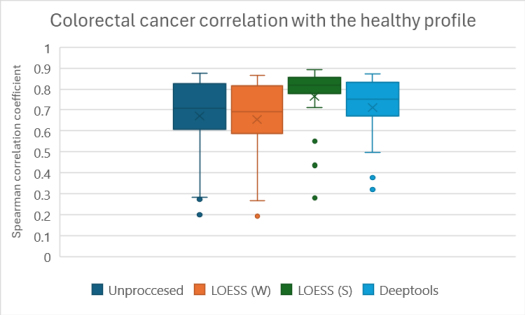


Figure 9: Spearman correlation between the healthy median profile and colorectal cancer samples for different GC-correction methods.

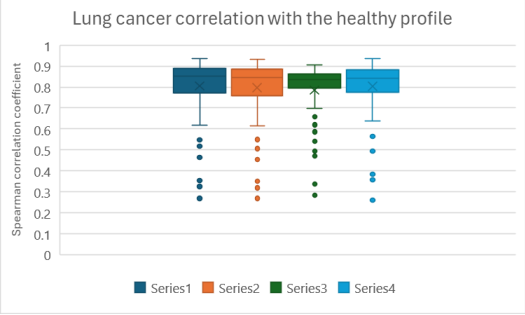


Figure 10: Spearman correlation between the healthy median profile and lung cancer samples for different GC-correction methods.

B Correlation of different cancer types for different MAPQ filtering thresholds

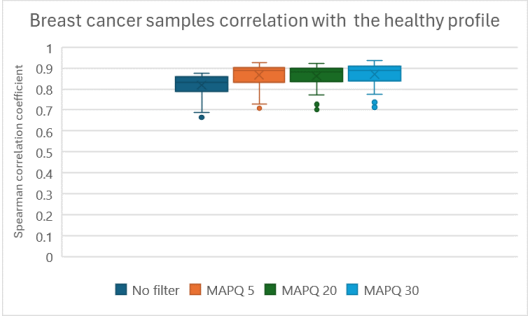


Figure 11: Spearman correlation between the healthy median profile and breast cancer samples for different MAPQ filtering thresholds

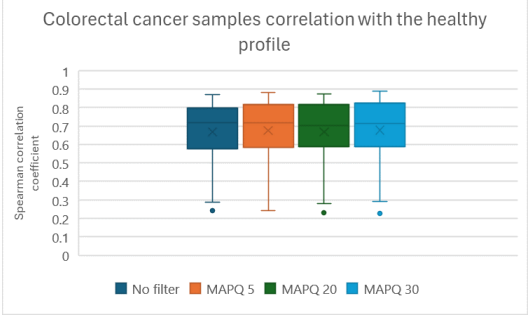


Figure 12: Spearman correlation between the healthy median profile and colorectal cancer samples for different MAPQ filtering thresholds

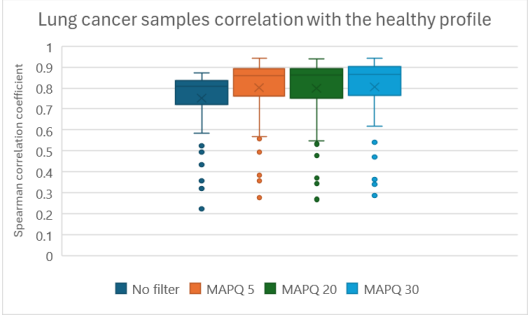


Figure 13: Spearman correlation between the healthy median profile and lung cancer samples for different MAPQ filtering thresholds

C MAPQ < 5 fragments locations

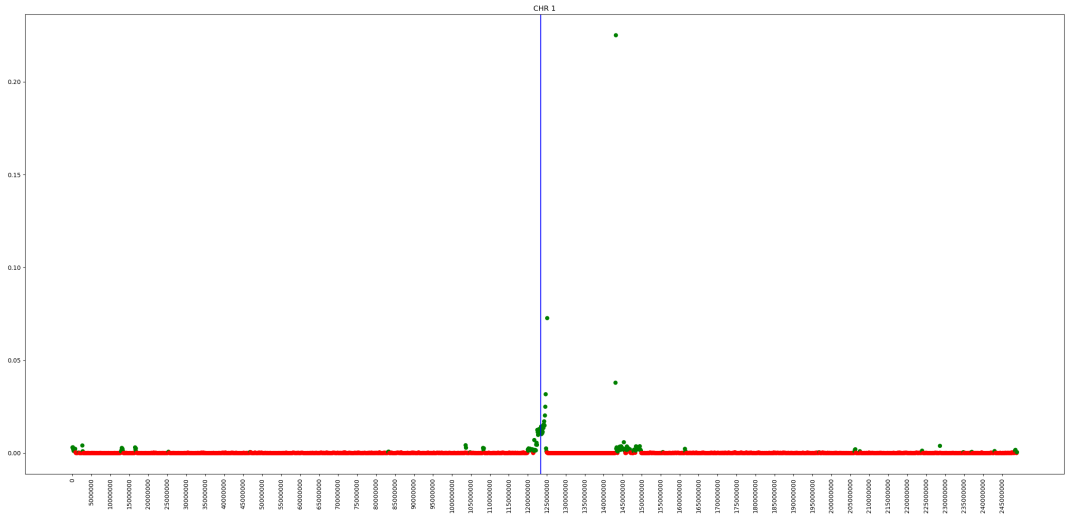


Figure 14: Fragments with MAPQ < 5 location on chromosome 1

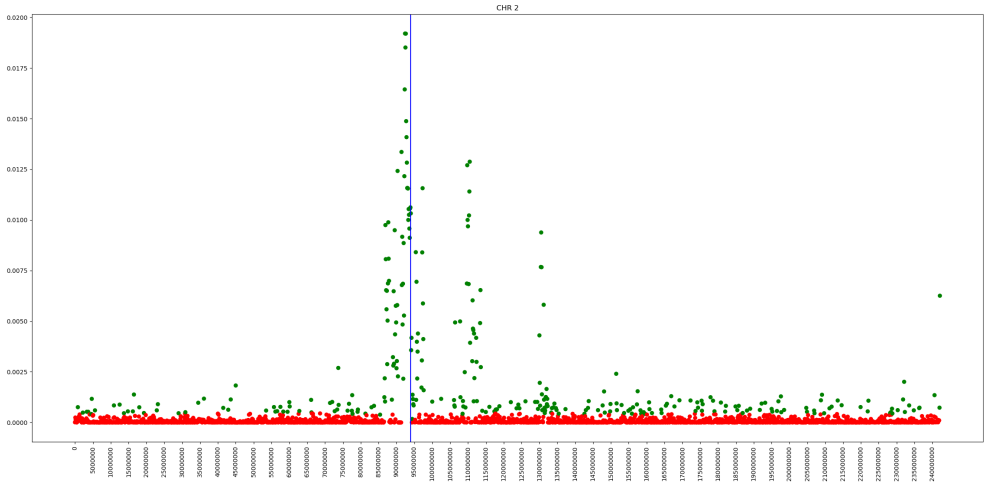


Figure 15: Fragments with MAPQ < 5 location on chromosome 2

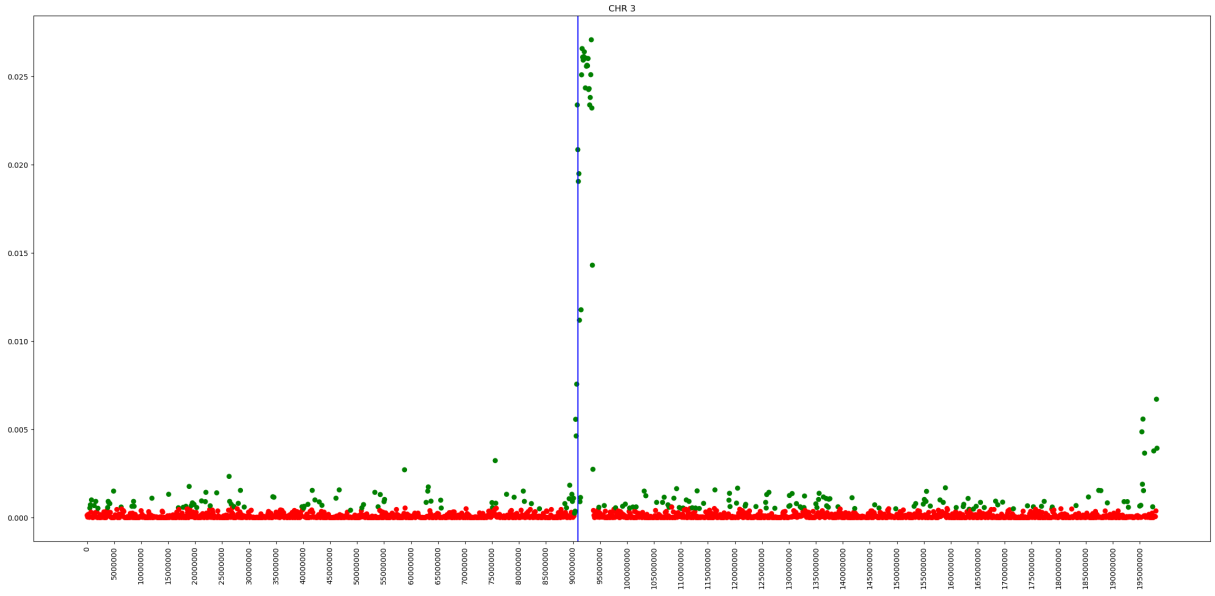


Figure 16: Fragments with MAPQ < 5 location on chromosome 3

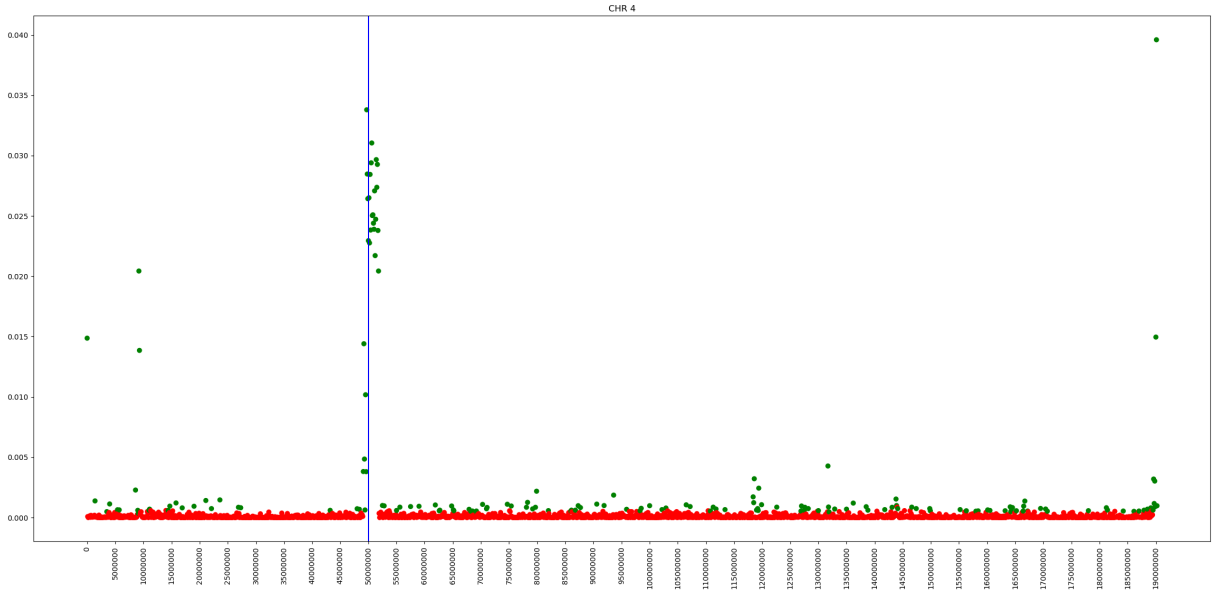


Figure 17: Fragments with MAPQ < 5 location on chromosome 4

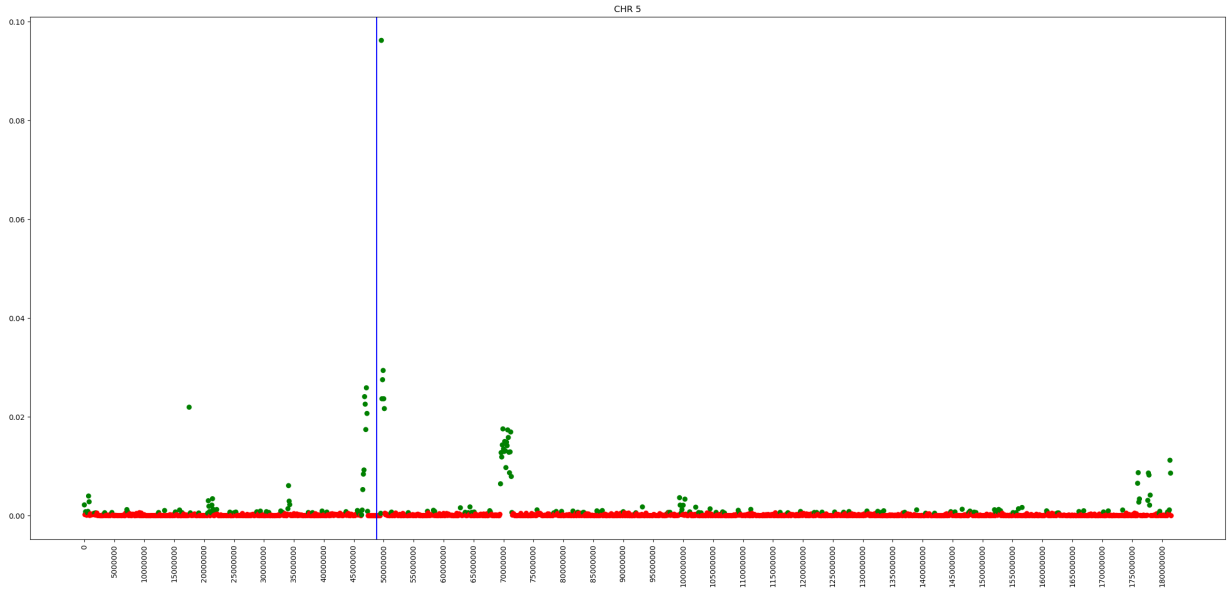


Figure 18: Fragments with MAPQ < 5 location on chromosome 5

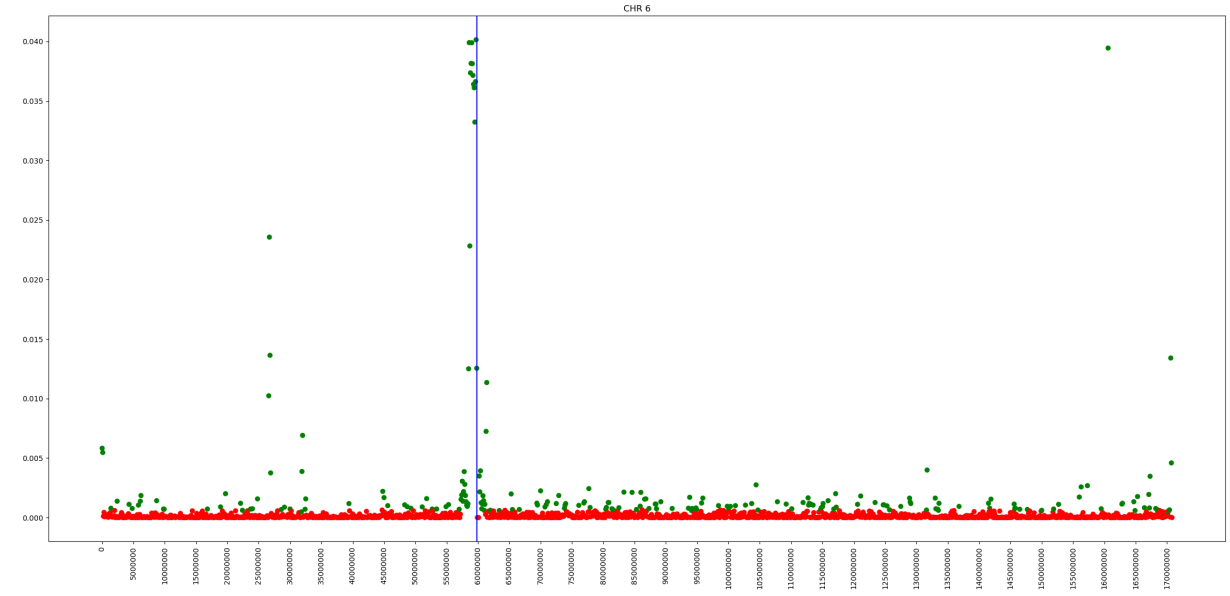


Figure 19: Fragments with MAPQ < 5 location on chromosome 6

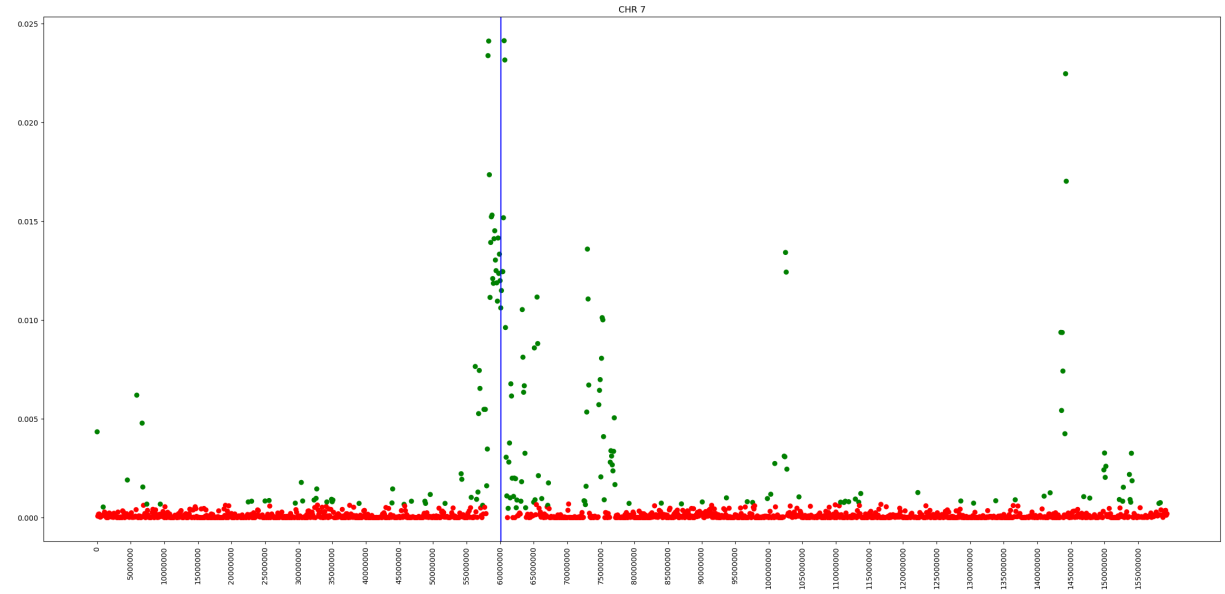


Figure 20: Fragments with MAPQ < 5 location on chromosome 7

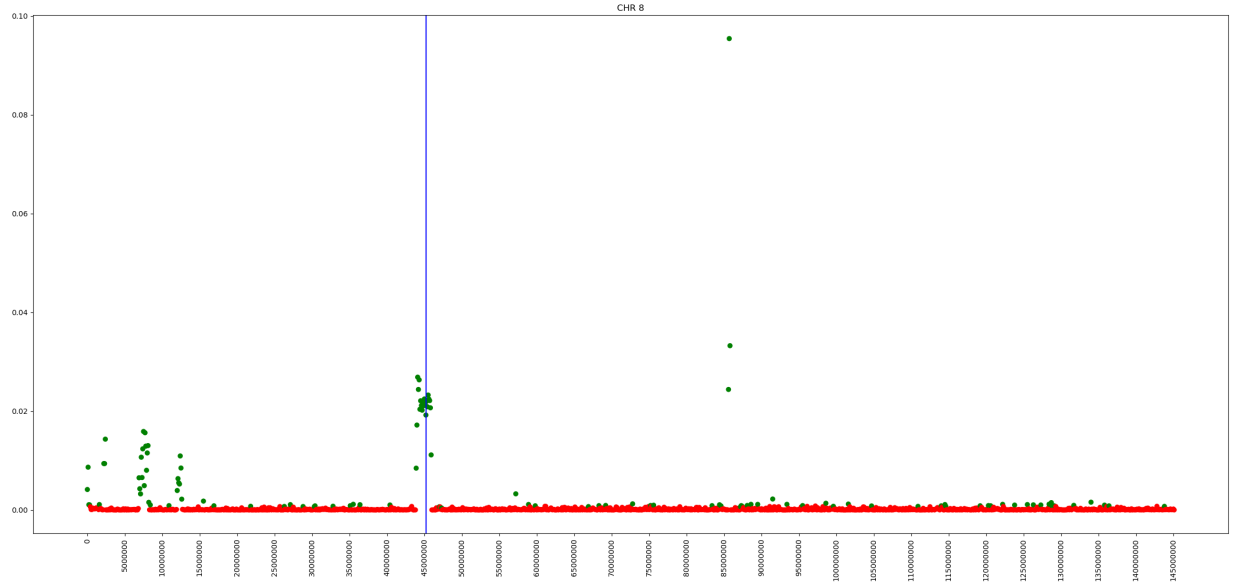


Figure 21: Fragments with MAPQ < 5 location on chromosome 8

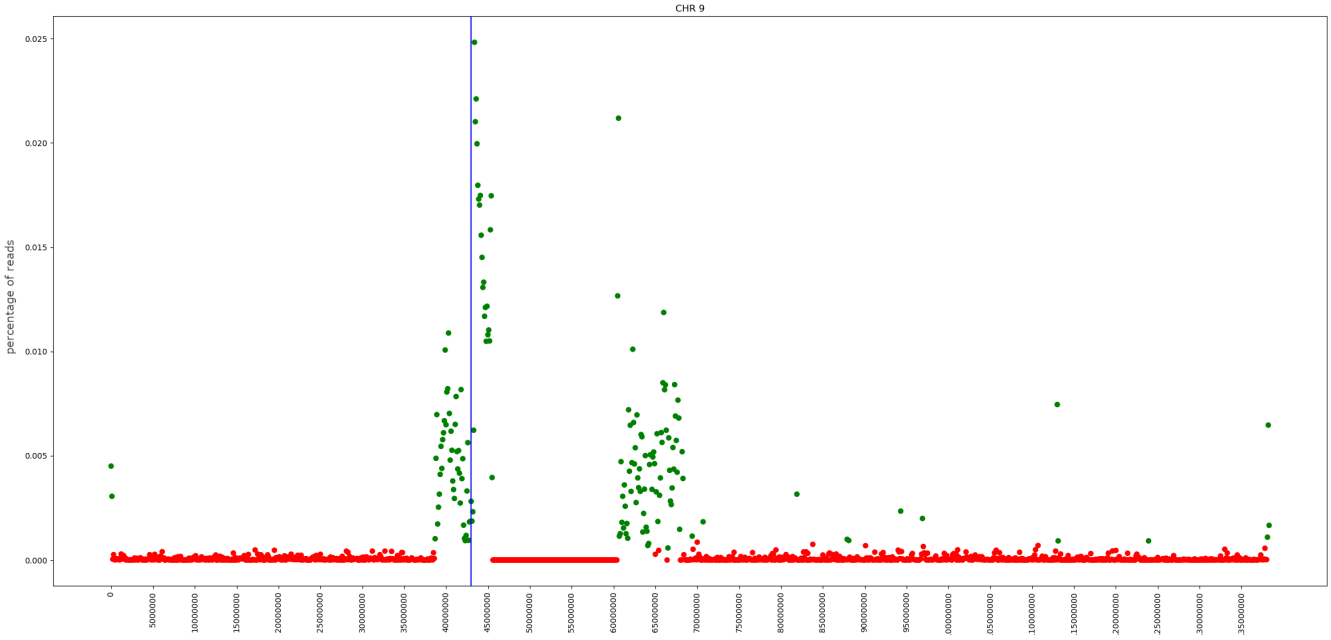


Figure 22: Fragments with MAPQ < 5 location on chromosome 9

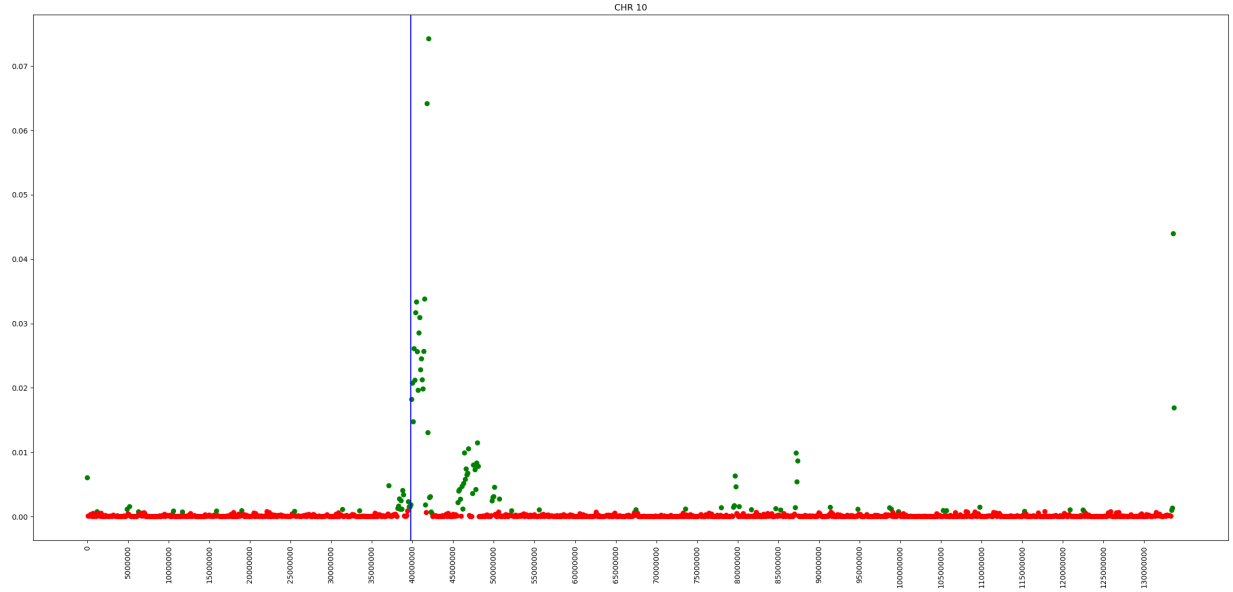


Figure 23: Fragments with MAPQ < 5 location on chromosome 10

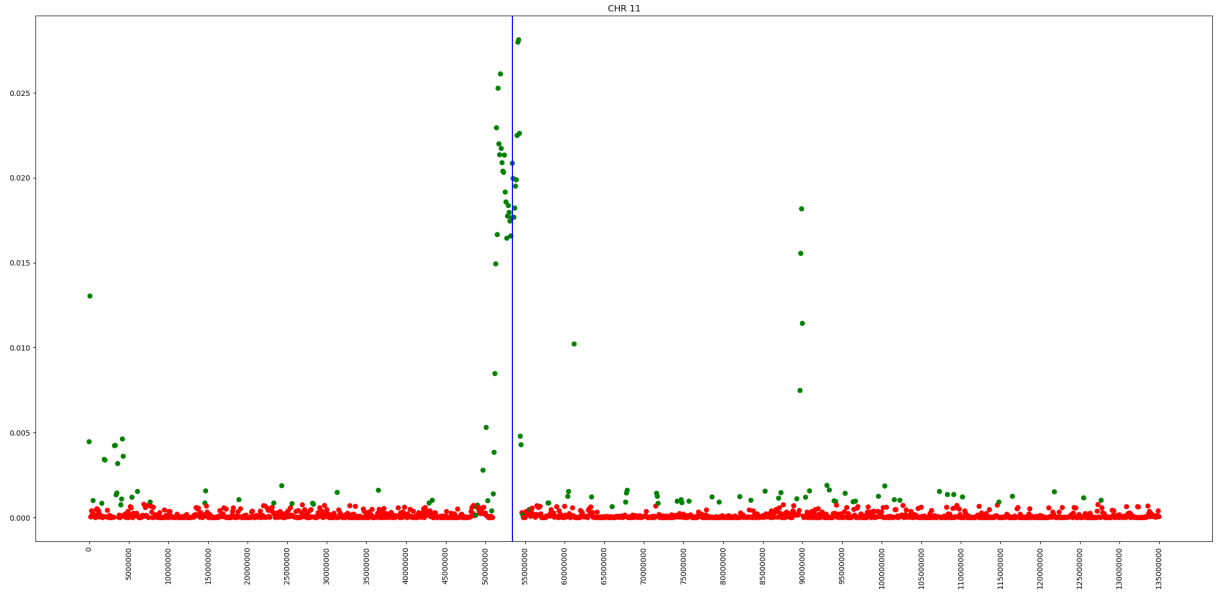


Figure 24: Fragments with MAPQ < 5 location on chromosome 11

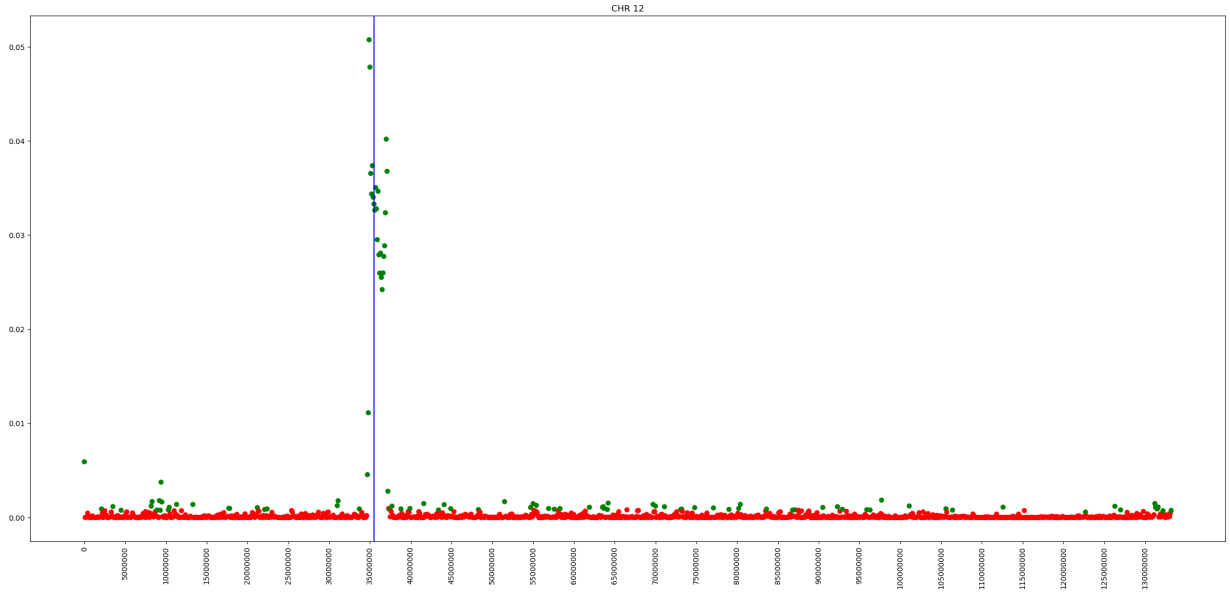


Figure 25: Fragments with MAPQ < 5 location on chromosome 12

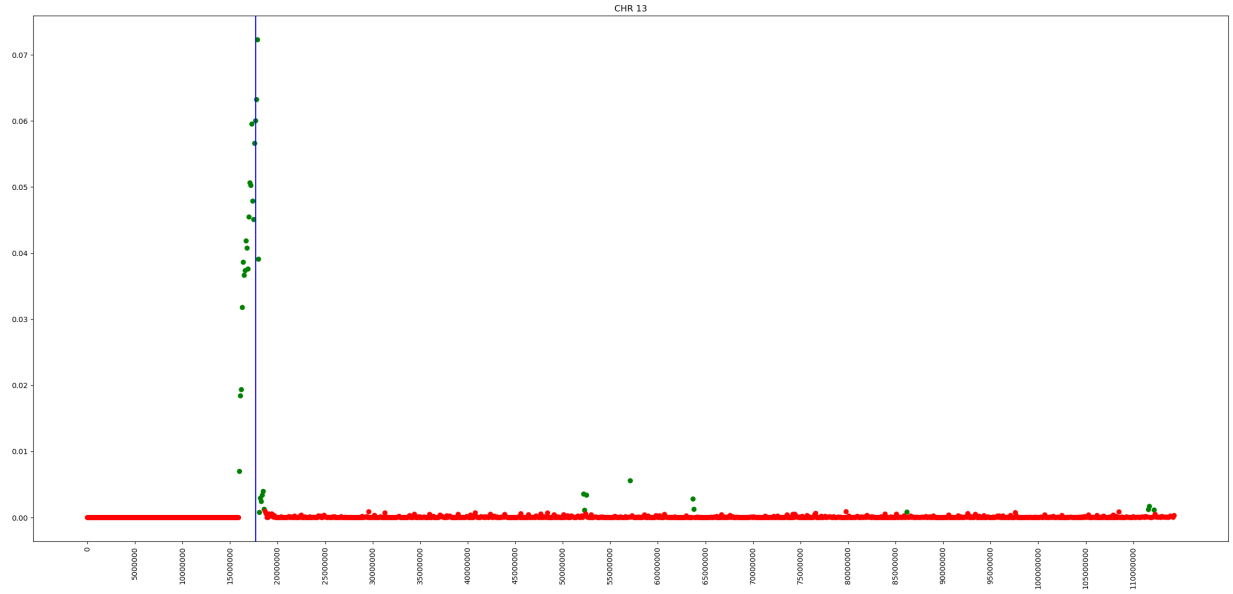


Figure 26: Fragments with MAPQ < 5 location on chromosome 13

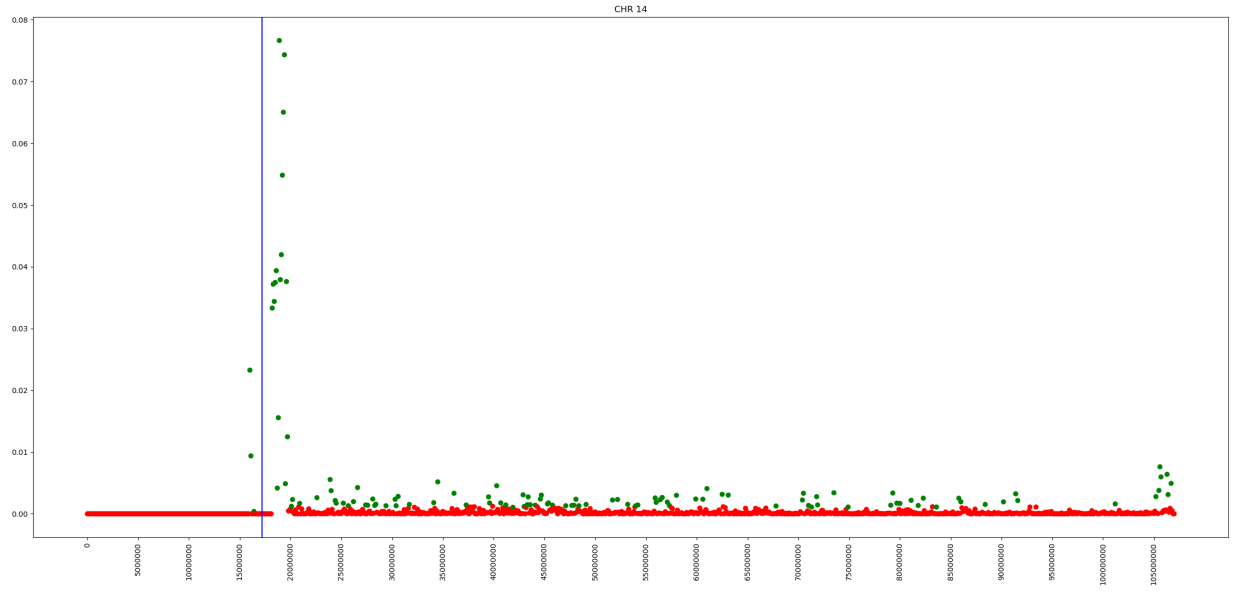


Figure 27: Fragments with MAPQ < 5 location on chromosome 14

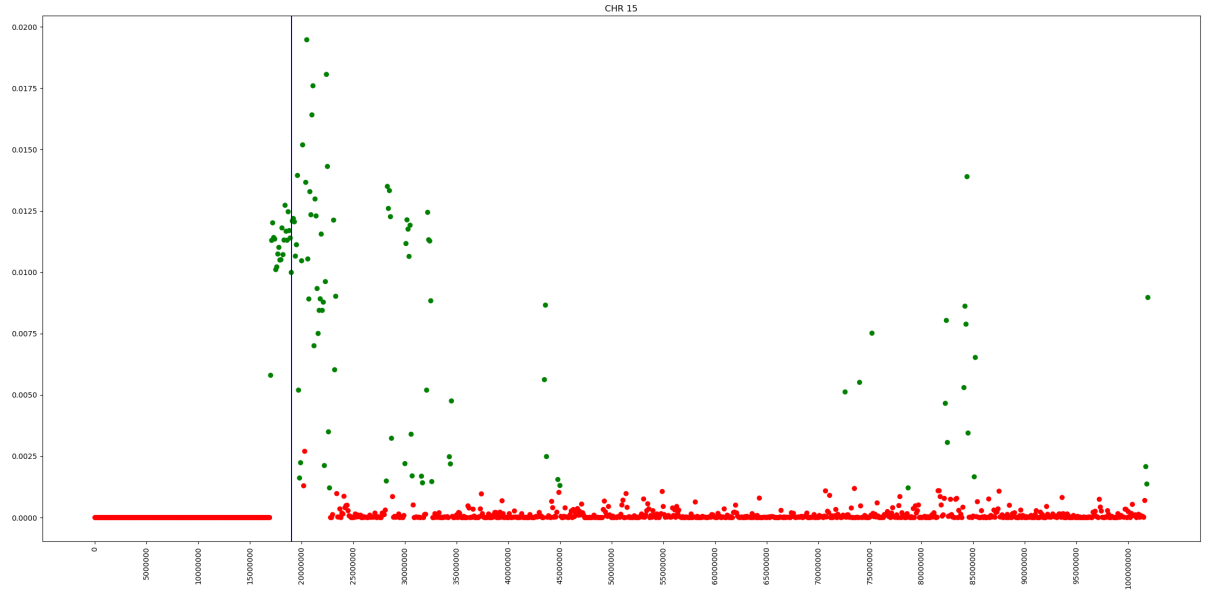


Figure 28: Fragments with MAPQ < 5 location on chromosome 15

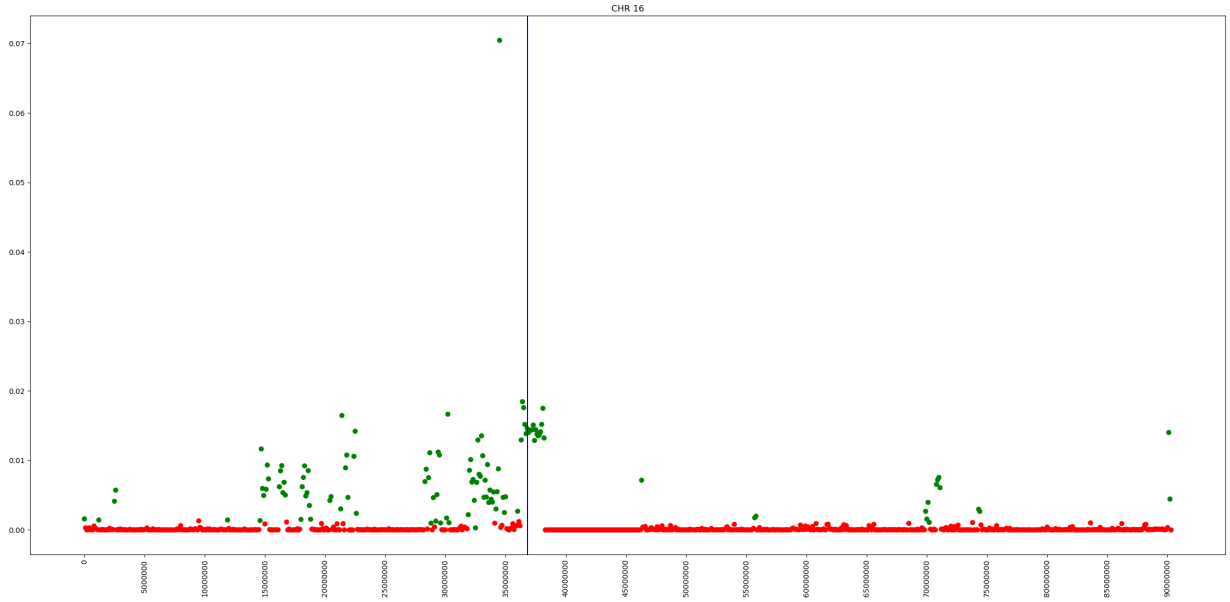


Figure 29: Fragments with MAPQ < 5 location on chromosome 16

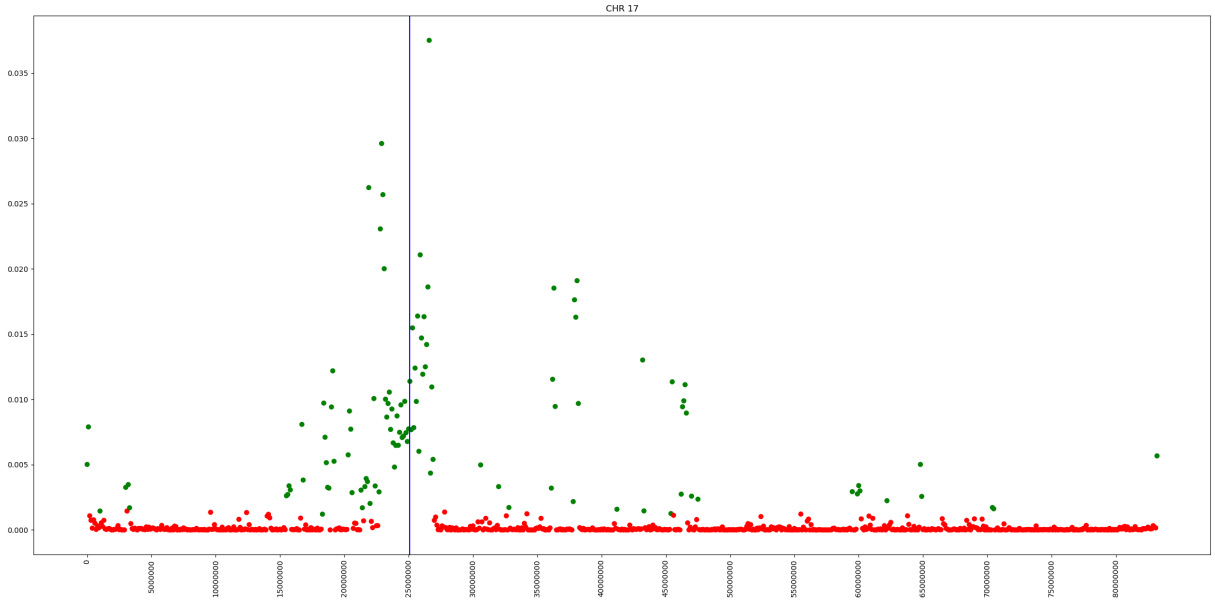


Figure 30: Fragments with MAPQ < 5 location on chromosome 17

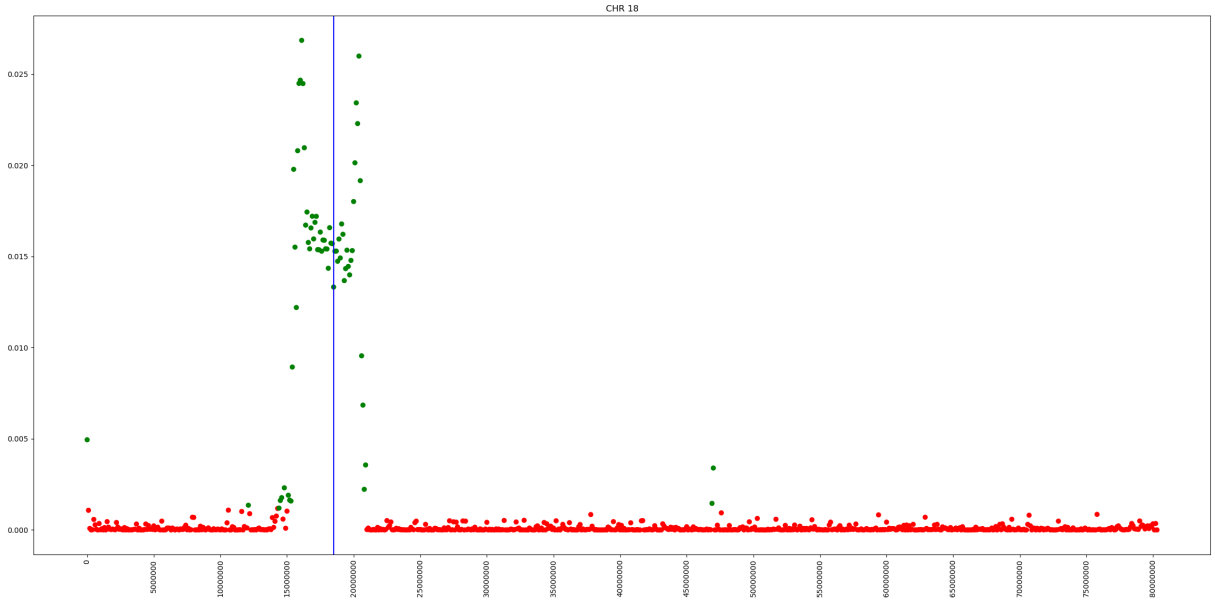


Figure 31: Fragments with MAPQ < 5 location on chromosome 18

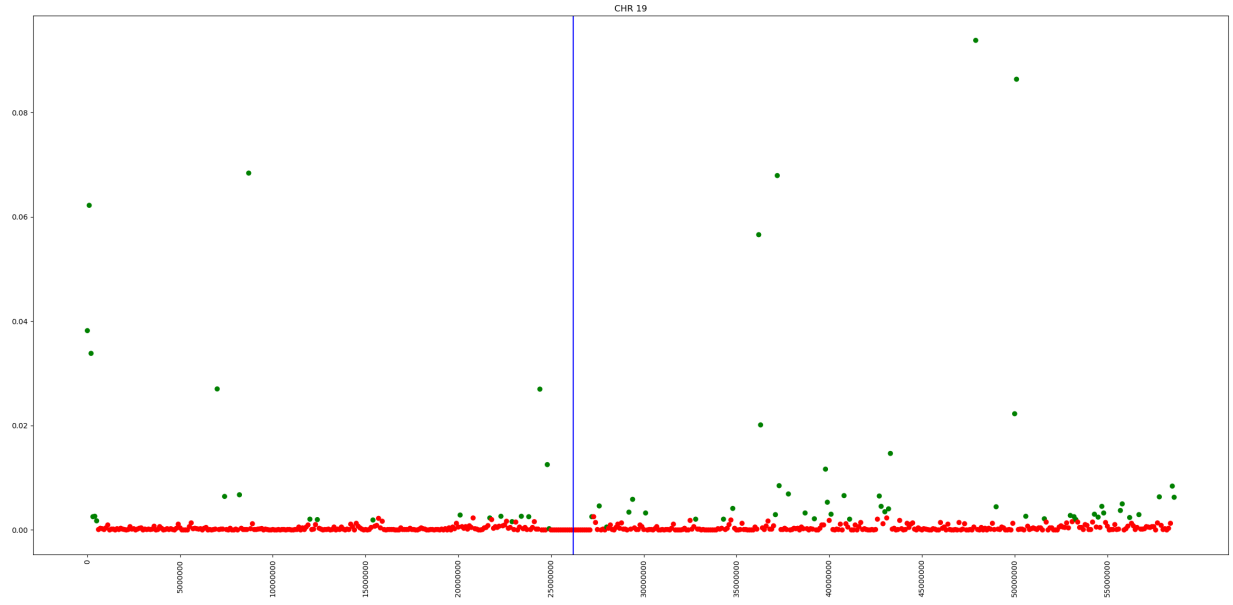


Figure 32: Fragments with MAPQ < 5 location on chromosome 19

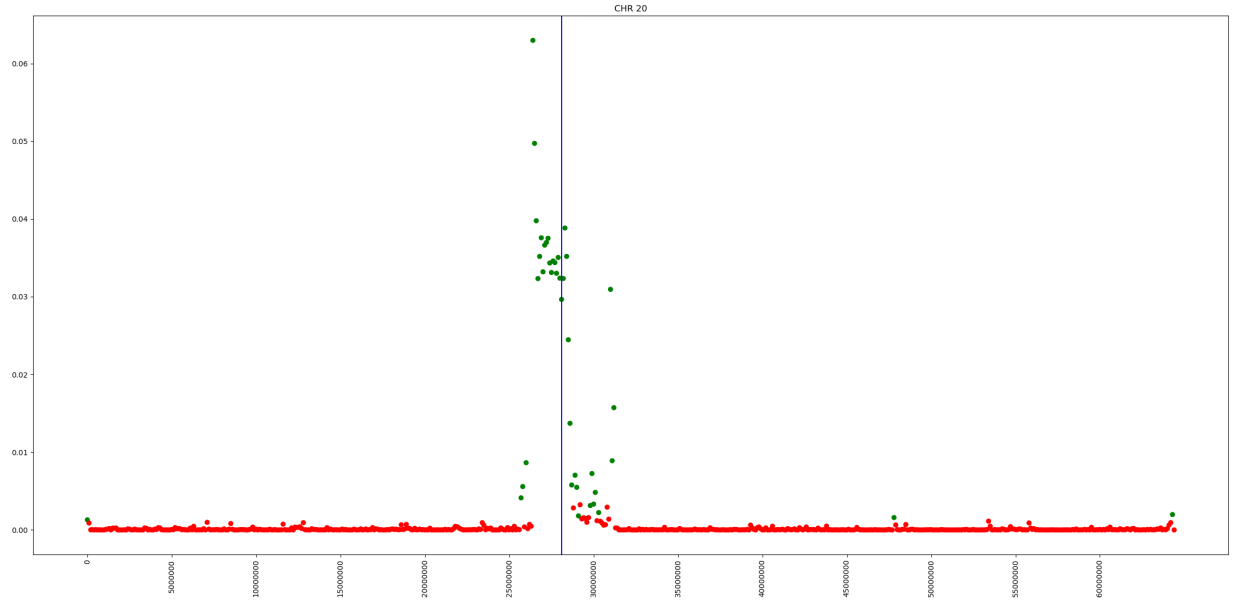


Figure 33: Fragments with MAPQ < 5 location on chromosome 20

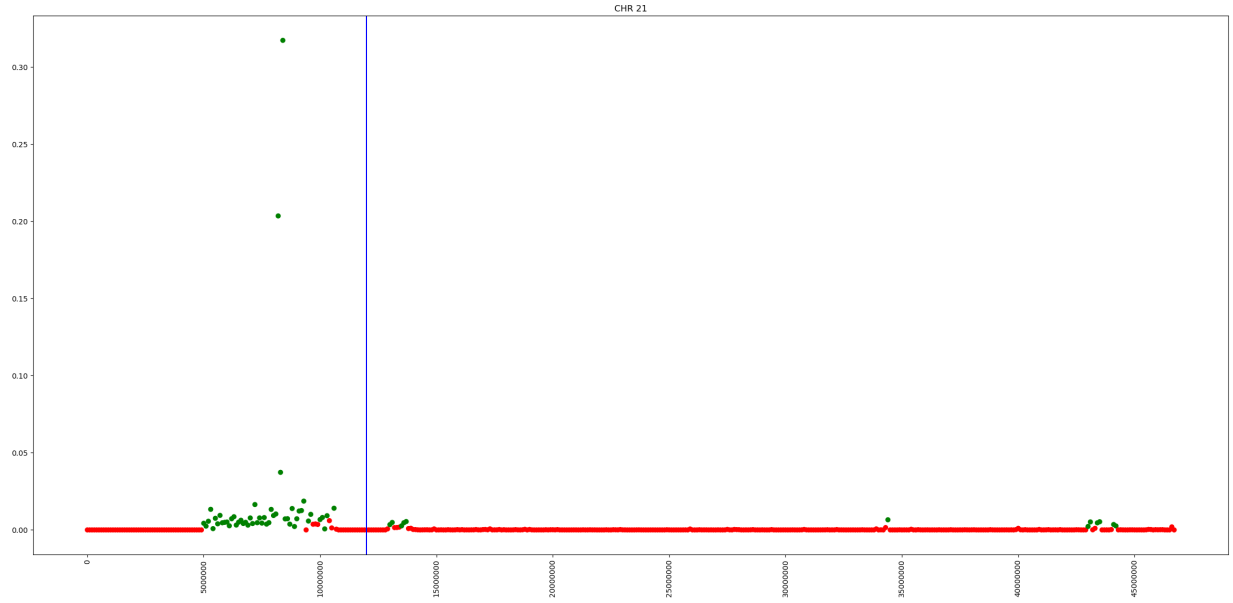


Figure 34: Fragments with MAPQ < 5 location on chromosome 21

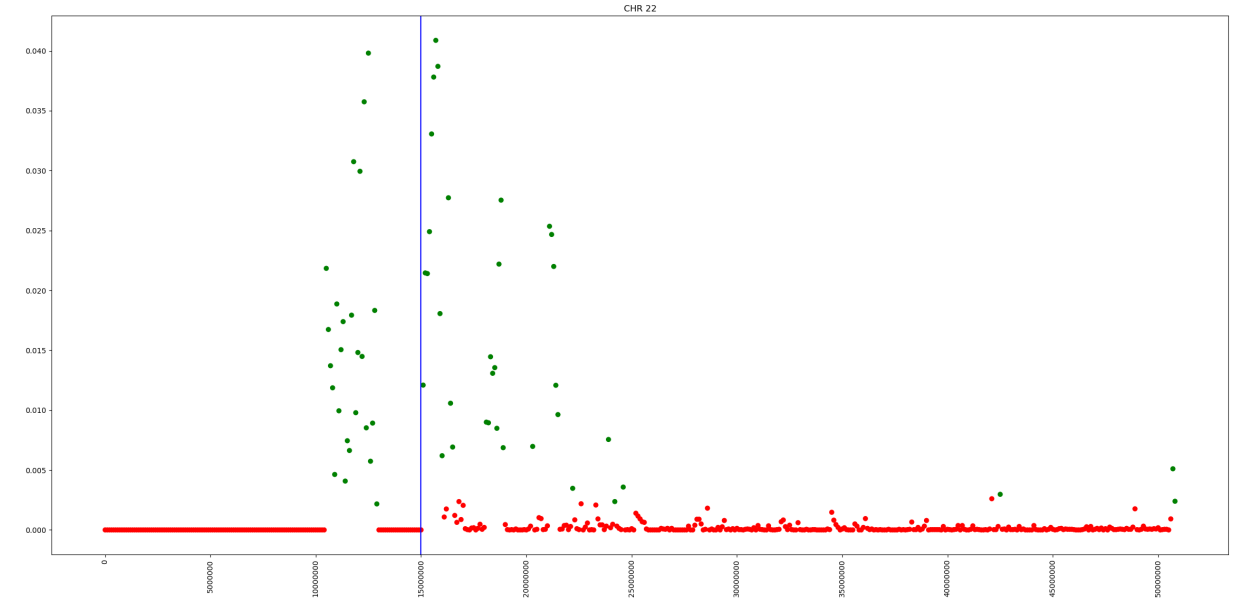


Figure 35: Fragments with MAPQ < 5 location on chromosome 22

References

[1] Feb. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] C. van Iersel, E. Toes-Zoutendijk, L. de Jonge, van, and H. Schootbrugge-Vandermeer, *National evaluation of the populationbased colorectal cancer screening programme in the netherlands 2018-2021*, Jan. 2023. [Online]. Available: https://www.rivm.nl/sites/default/files/2023-05/Evaluation%20of%20the%20Colorectal%20Cancer%20Screening%20Programme%202018-2021.pdf (visited on 06/10/2024).

[3] S. Volik, M. Alcaide, R. D. Morin, and C. Collins, "Cell-free DNA (cfDNA): Clinical significance and utility in cancer shaped by emerging technologies," *Molecular Cancer Research*, vol. 14, no. 10, pp. 898–908, Oct. 1, 2016, ISSN: 1541-7786, 1557-3125. DOI: 10.1158/1541-7786.MCR-16-0044. [Online]. Available: https://aacrjournals.org/mcr/article/14/10/898/135577/Cell-free-DNA-cfDNA-Clinical-Significance-and (visited on 06/10/2024).

[4] T. Qi, M. Pan, H. Shi, L. Wang, Y. Bai, and Q. Ge, "Cell-free DNA fragmentomics: The novel promising biomarker," *International Journal of Molecular Sciences*, vol. 24, no. 2, p. 1503, Jan. 12, 2023, ISSN: 1422-0067. DOI: 10.3390/ijms24021503. [Online]. Available: https://www.mdpi.com/1422-0067/24/2/1503 (visited on 04/26/2024).

[5] S. Cristiano, A. Leal, J. Phallen, *et al.*, "Genome-wide cell-free DNA fragmentation in patients with cancer," *Nature*, vol. 570, no. 7761, pp. 385–389, Jun. 2019, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1272-6. [Online]. Available: https://www.nature.com/articles/s41586-019-1272-6 (visited on 04/22/2024).

[6] N. Moldovan, Y. Van Der Pol, T. Van Den Ende, *et al.*, "Multi-modal cell-free DNA genomic and fragmentomic patterns enhance cancer survival and recurrence analysis," *Cell Reports Medicine*, vol. 5, no. 1, p. 101 349, Jan. 2024, ISSN: 26663791. DOI: 10.1016/j.xcrm.2023.101349. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666379123005669 (visited on 05/08/2024).

[7] P. D. Browne, T. K. Nielsen, W. Kot, *et al.*, "GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms," *GigaScience*, vol. 9, no. 2, giaa008, Feb. 1, 2020, ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa008. [Online]. Available: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa008/5735313 (visited on 04/26/2024).

[8] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Research*, vol. 40, no. 10, e72–e72, May 1, 2012, ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gks001. [Online]. Available: https://academic.oup.com/nar/article/40/10/e72/2411059 (visited on 05/13/2024).

[9] C. Jin, X. Liu, W. Zheng, *et al.*, "Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection," *Molecular Oncology*, vol. 15, no. 9, pp. 2377–2389, Sep. 2021, ISSN: 1574-7891, 1878-0261. DOI: 10.1002/1878-0261.13041. [Online]. Available: https://febs.onlinelibrary.wiley.com/doi/10.1002/1878-0261.13041 (visited on 05/24/2024).

[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990, ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602 (visited on 06/13/2024).