MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Discovery of novel biomarkers using unsupervised statistical learning

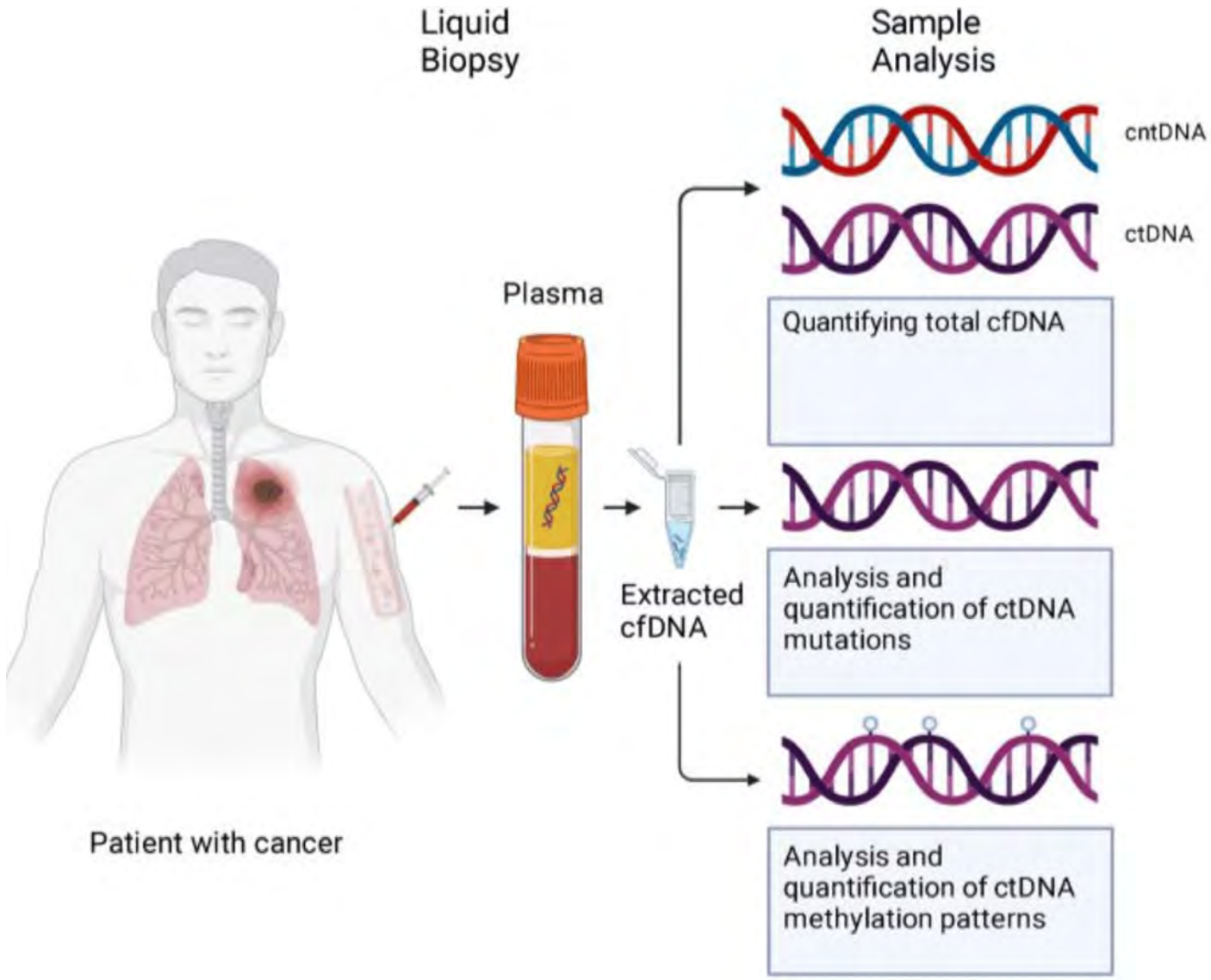
Scott Hislop
Supervisor: Solon Karapanagiotis

Contents

- Background
- Our data and goal
- Methods in the literature
- New methods
- Results
- Discussion

Background

- Want to detect and monitor cancerous tumors
- Tumor Biopsies – the standard method
- Liquid Biopsies – emerging method:
  - Earlier detection
  - Can monitor more easily, more often
  - Detects a wider range of mutations

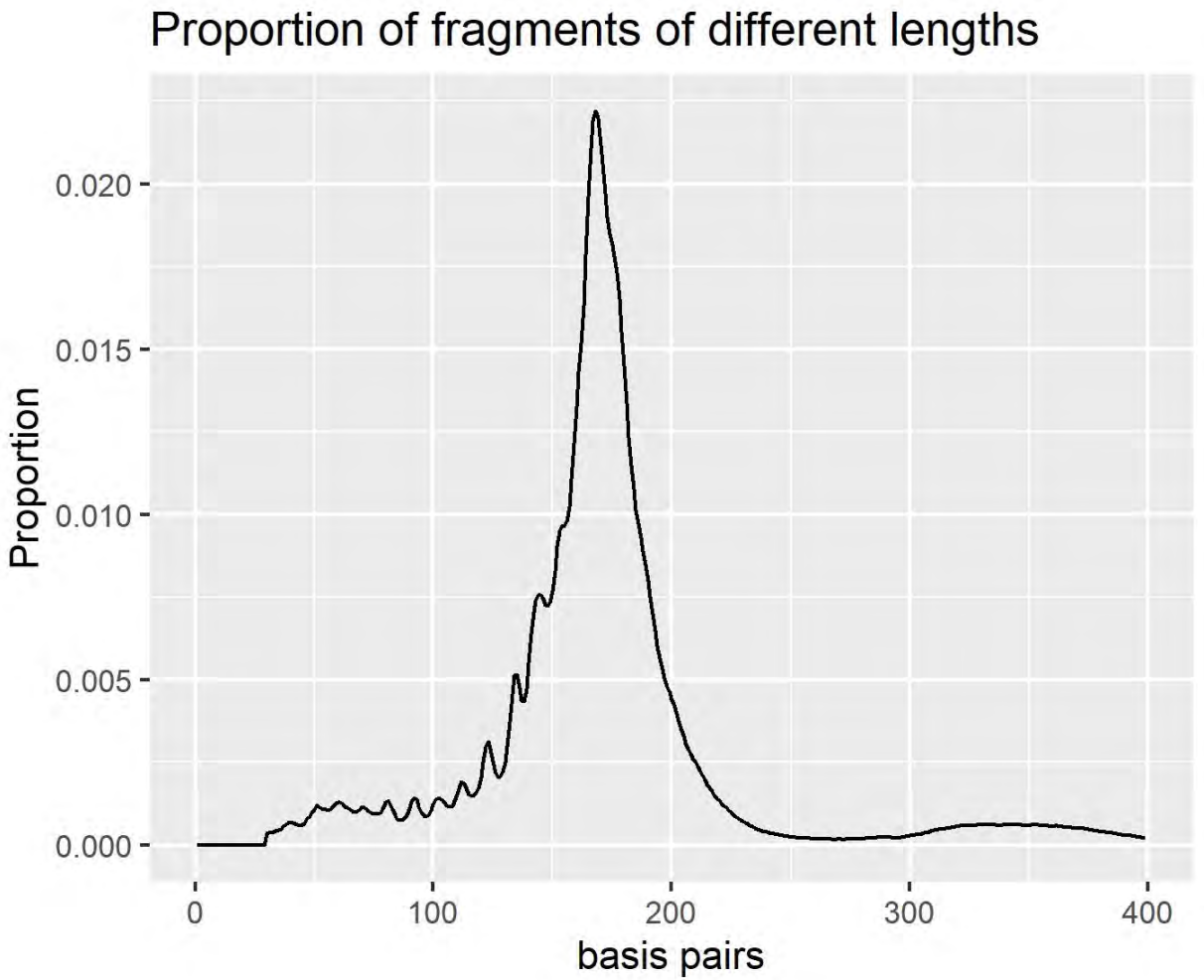


[1] - Image – (Dao, 2023)

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Our data



Form of data

- Lengths of each of the fragments measured (in basis pairs, e.g. ATGTC is 5 basis pairs)
- Usually around 10-100 million fragments collected per sample.
- Can convert these into proportions for each length

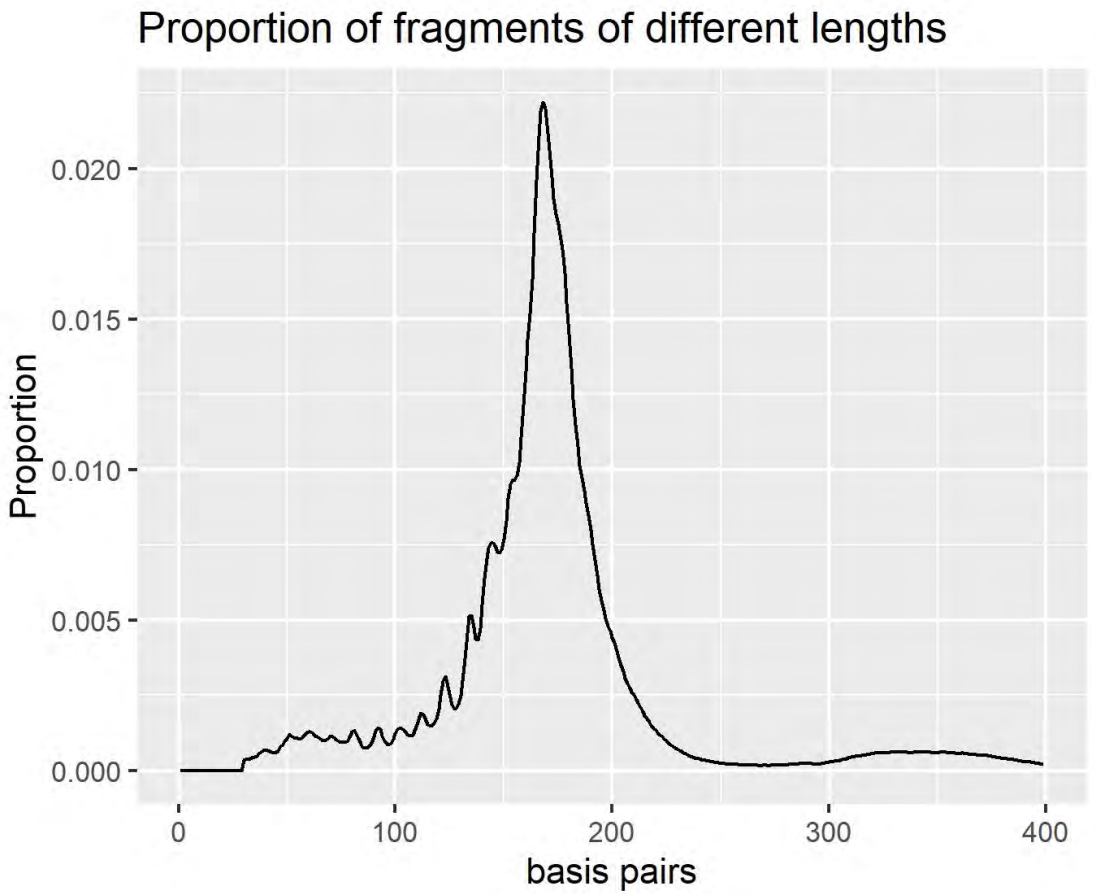
Labels

- Have CNA burden and Tumor fraction estimates for each sample – estimated with ichorCNA
- Define a sample as “healthy” if CNA burden is 0. This gives n=55 “healthy” samples, n=239 “unhealthy” samples.

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Goals



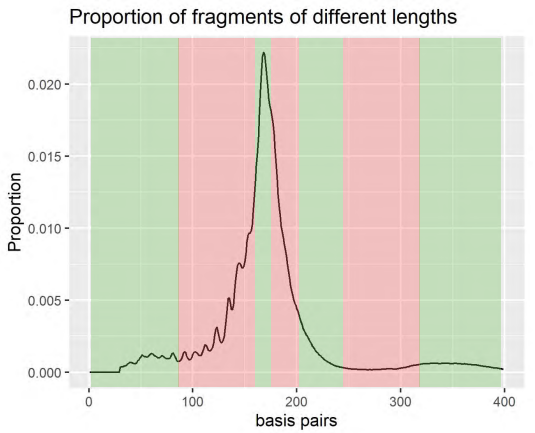
1 – Detect whether a sample has cancer

$$\mathbb{P}(sample\ has\ cancer|data) = 0.7$$

2 – Monitor changes in tumor burden

$$f(visit\ one) = 0.13$$
$$f(visit\ two) = 0.35$$
$$f(visit\ three) = 0.47$$

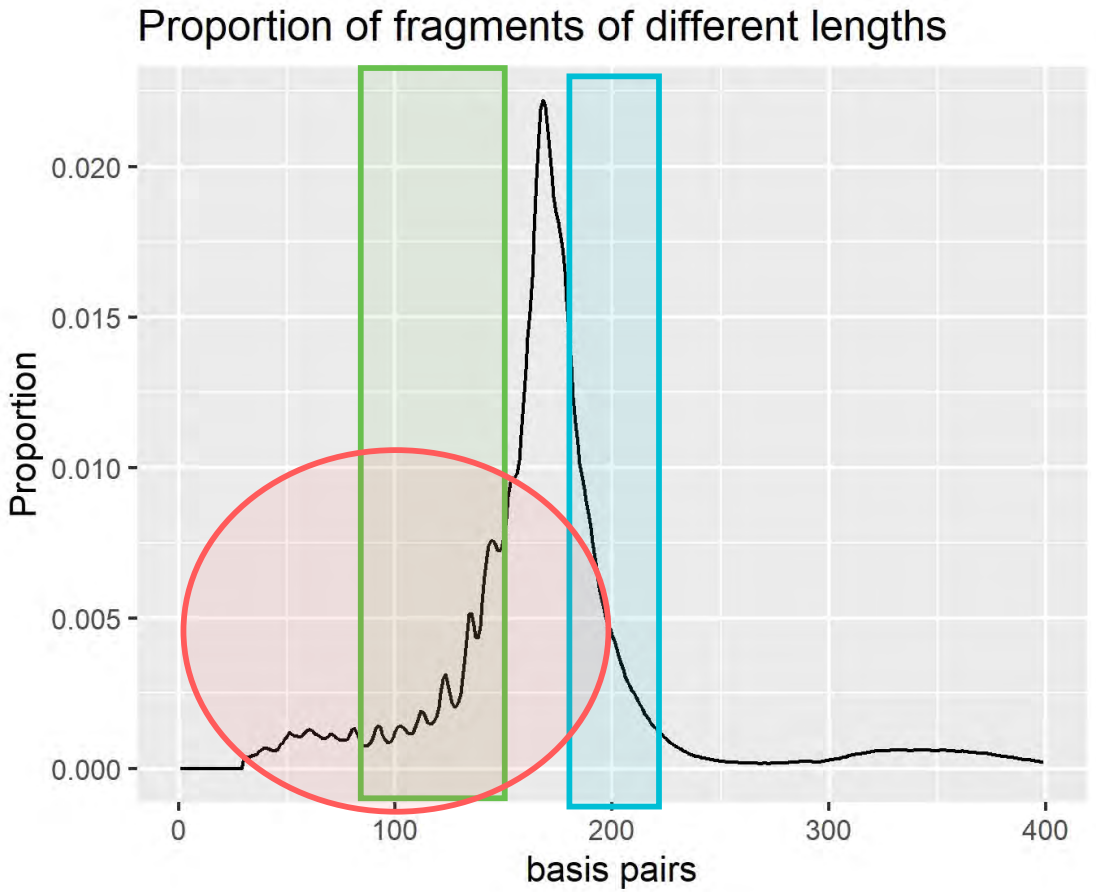
3 – Identify basis pair lengths with higher proportions of ctDNA



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Methods in the literature



ctDNA distribution

Some differences have been observed between cancer derived DNA (ctDNA) and non cancerous cfDNA - [2]:

- Higher in the green region
- Lower in the blue region
- Stronger periodicity in the red region

Measures used

- P[20,150], P[90,150], P[180,220] - [2]
- Test statistics or p-values of tests comparing the number of fragments in [110,135] against the number in [135,150] - [3]
- Using amplitude of oscillations with 10bp frequency in the red region -[2]

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

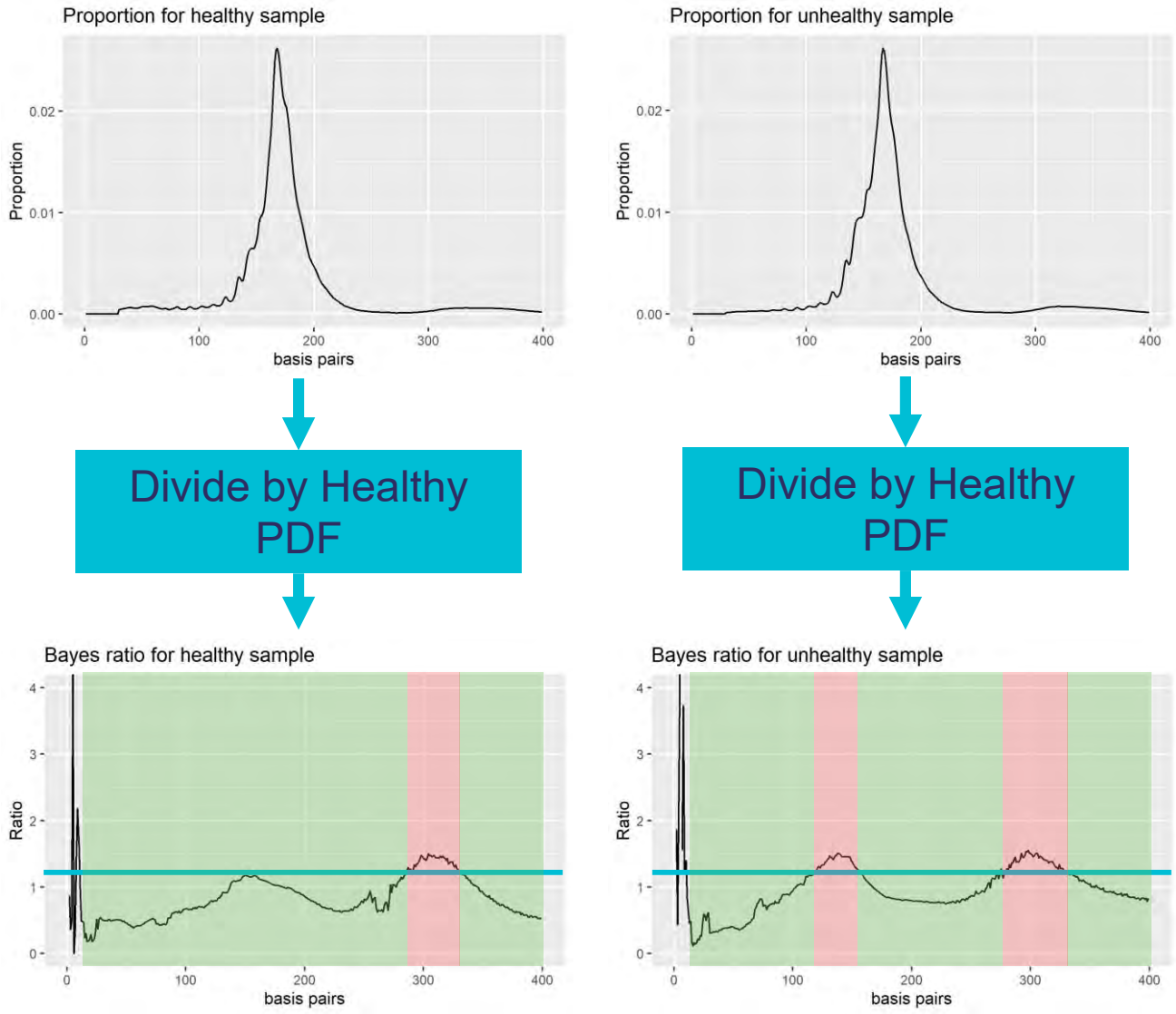
[2] – (Mouliere, 2018)
[3] – (Nguyen, 2023)

Bayes Classifier based method

Bayes classifier based approach

Basic idea

- Take an estimate of the PDF of the healthy distribution $f_{healthy}(x)$ for each fragment length x.
- For a given sample, compute the empirical PDF $f_{test}(x)$ for each fragment length x.
- Let $B(x) := \frac{f_{test}(x)}{f_{healthy}(x)}$
- Pick / determine some threshold T.
- If $B(x) \geq T$, we say that fragments of length x are likely cancerous
- If not, fragments of length x are likely healthy



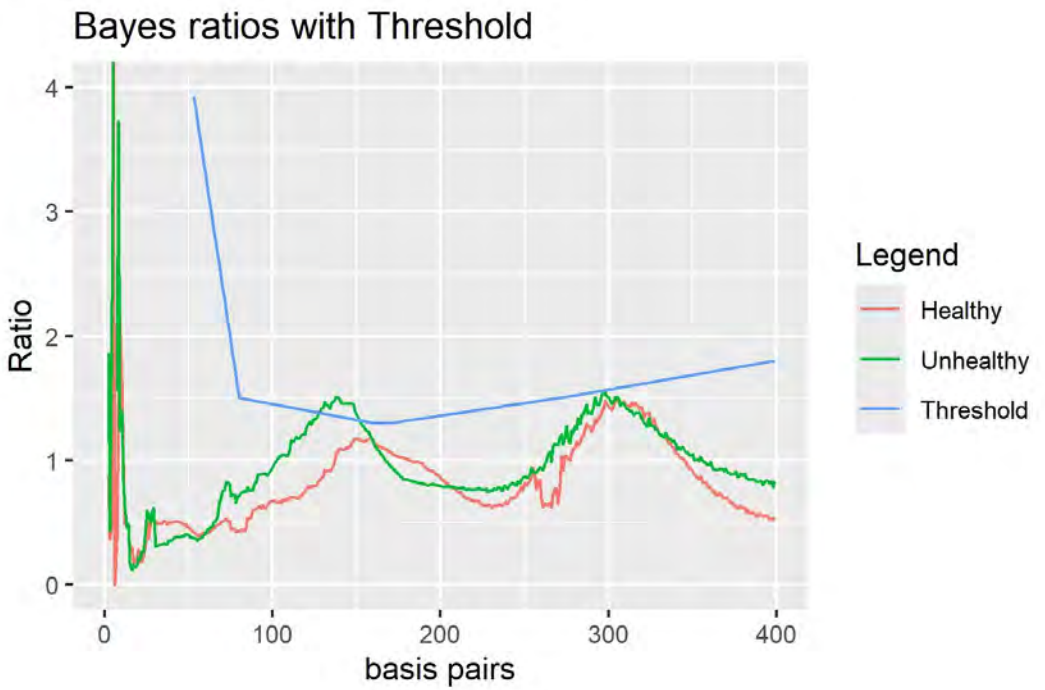
MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Bayes classifier – Thresholds

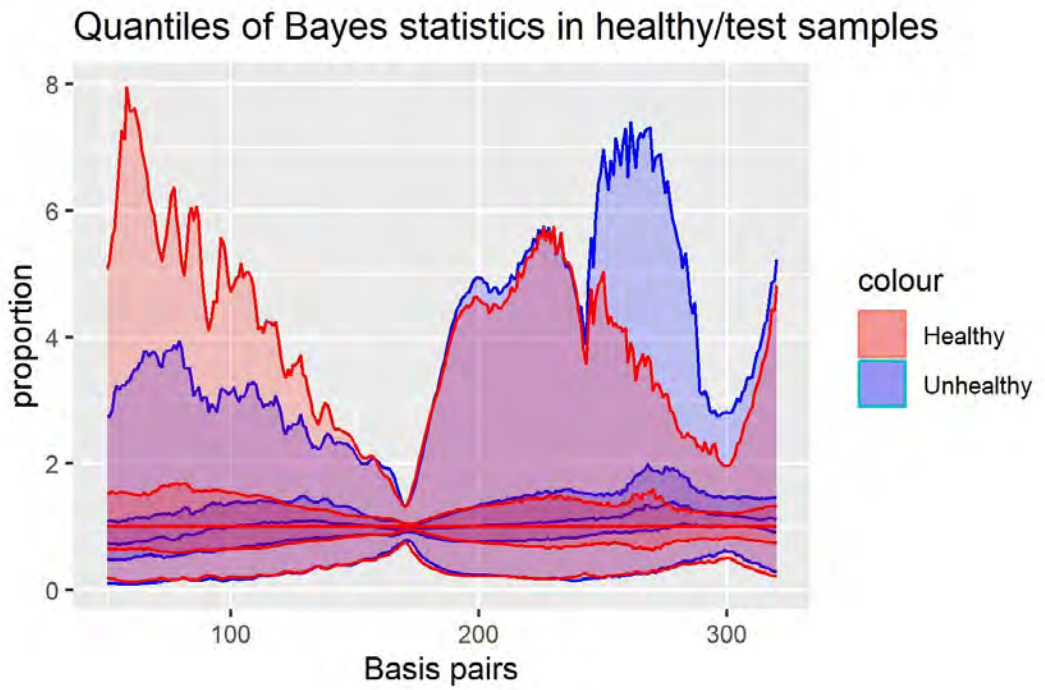
Variable thresholds

- Could set $T = T(x)$, so the threshold changes with the fragment length



Adaptive thresholds

- Could set $T(x)$ as a quantile of the ratios between healthy samples.



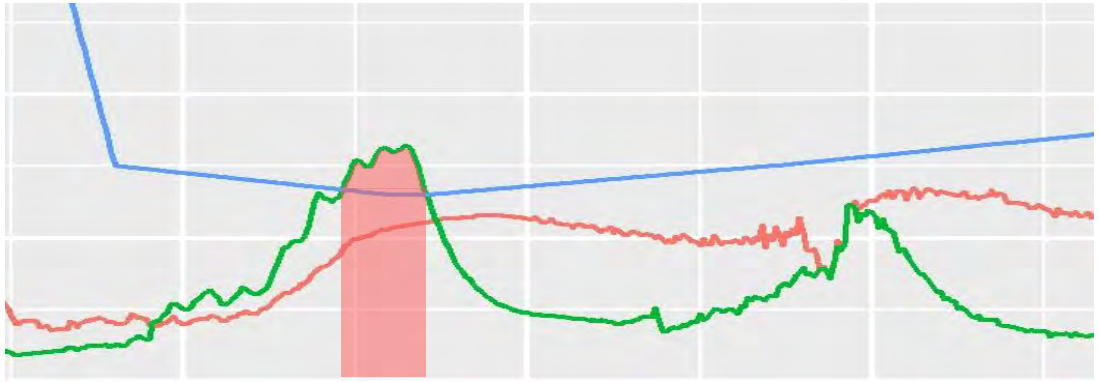
MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Bayes classifier output types

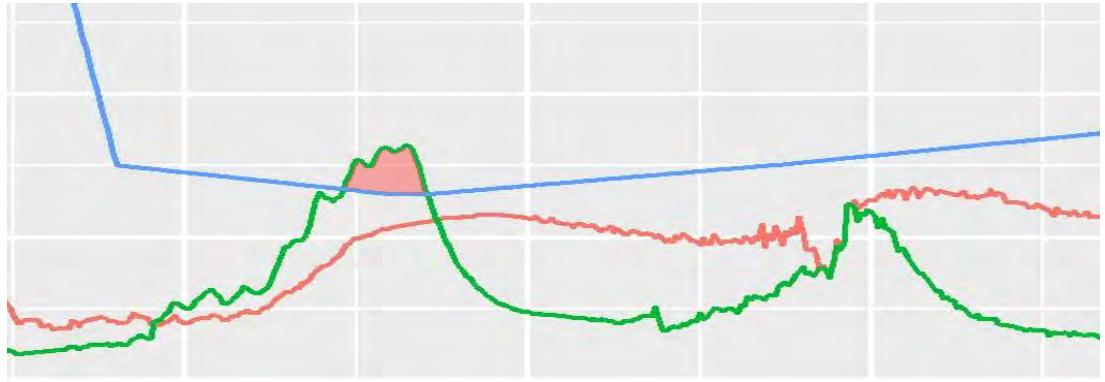
Type 1

The proportion labelled as likely cancerous



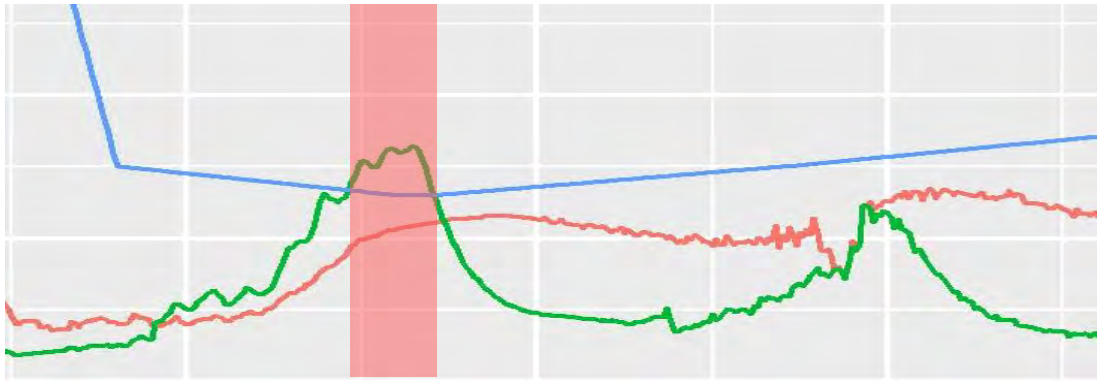
Type 2

The excess amount labelled as likely cancerous



Type 3

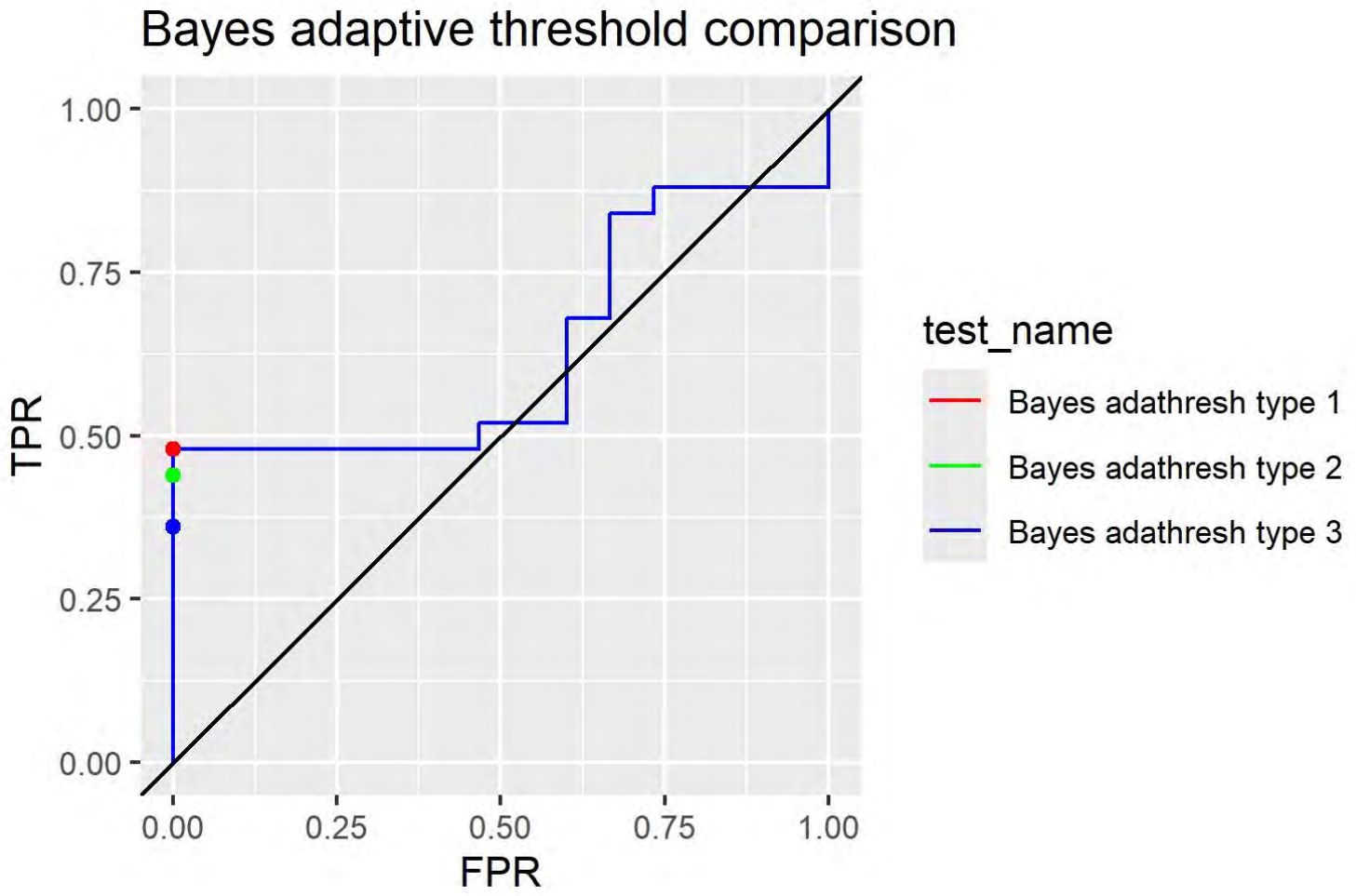
The number of fragment lengths labelled as likely cancerous



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Bayes classifier output types - ROC



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Mixture Models

Mixture Models applied to our data

- In our case, wish to determine the proportion of ctDNA (i.e. it's weight)
- Population has 2 classes, healthy cfDNA and ctDNA.
- Healthy class PDF can be estimated from healthy samples, we get some $f_{healthy}$
- ctDNA is modelled as a gaussian with PDF $g(x\,;\mu,\sigma^2\,)$
- PDF for the mixture model density is then:

$$f(x) = (1-w)\cdot f_{healthy}\,(x) + w\cdot g(x\,;\mu,\sigma^2\,)$$

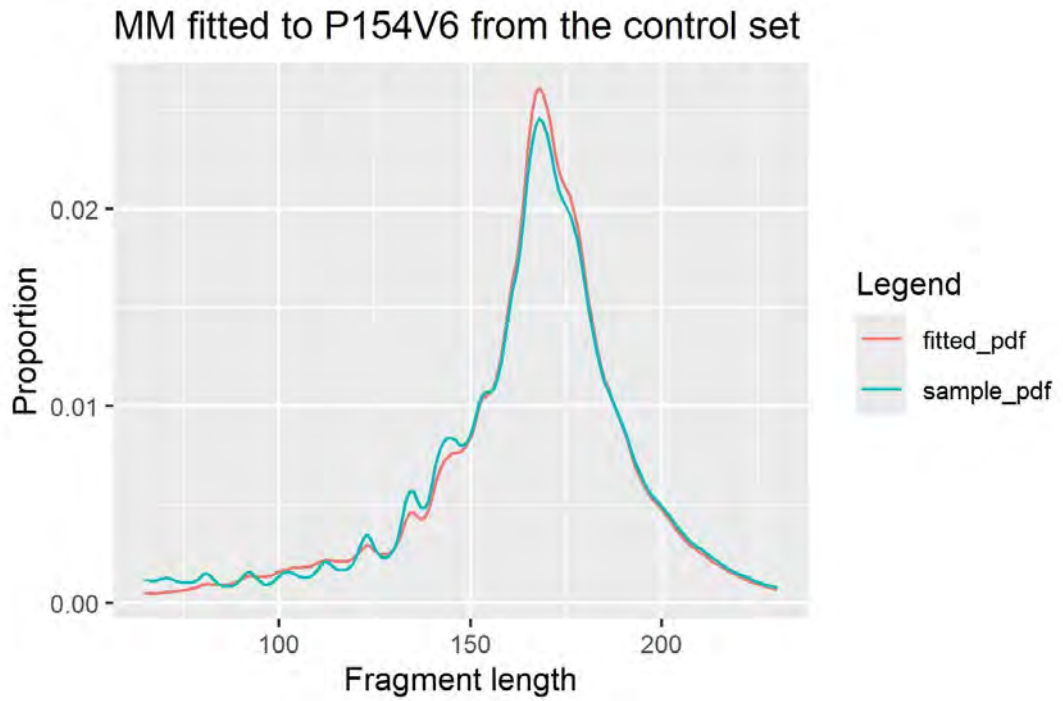
- We maximize the log-likelihood over w,μ,σ and return the weight on the gaussian, w.

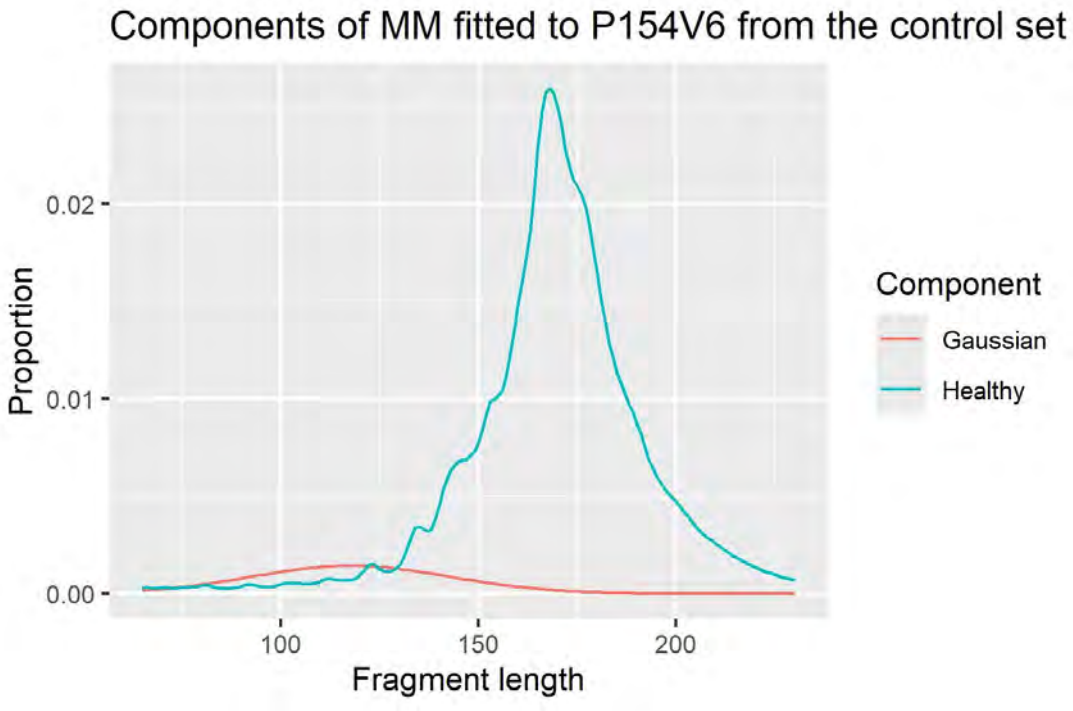
Fitting MMs - Healthy

In low ctDNA samples (or healthy samples):

- Proportions should look like PDF of a healthy sample
- Should put very little weight on the Gaussian

Healthy – 0.0896





MRC Biostatistics Unit

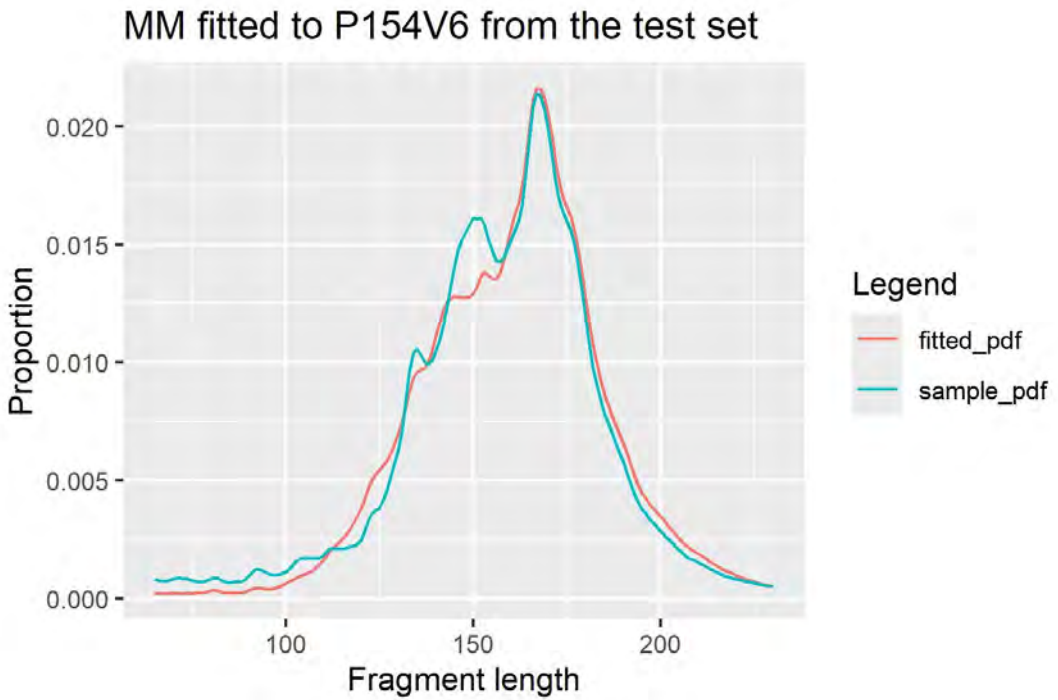
UNIVERSITY OF CAMBRIDGE

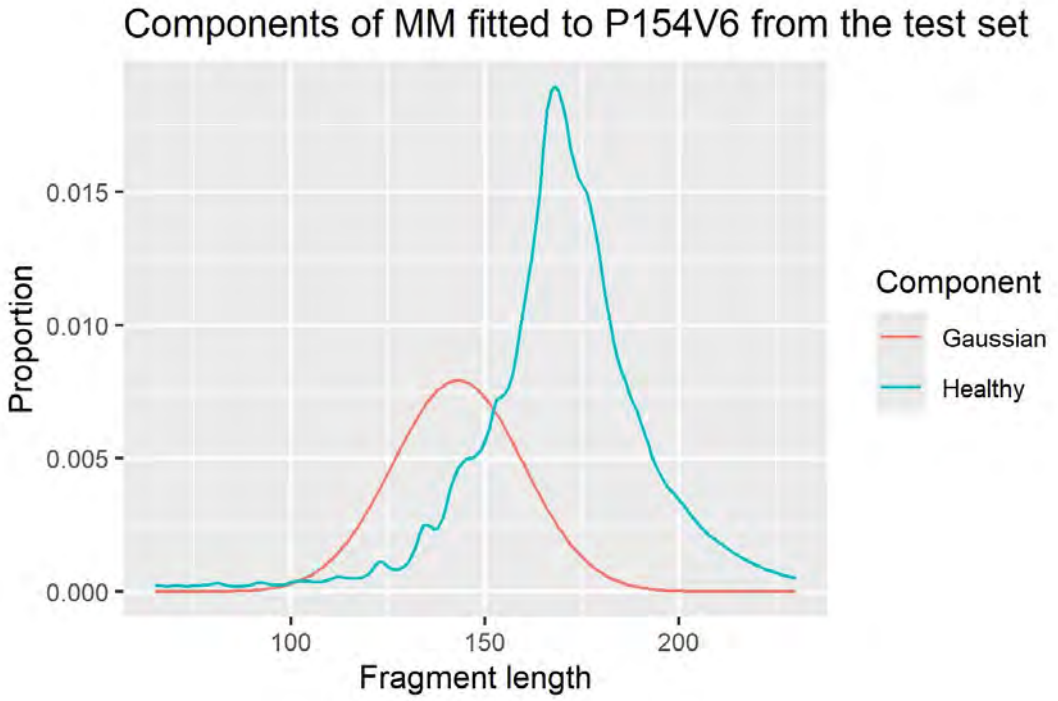
Fitting MMs - Patient

In high ctDNA samples:

- Proportions look very different to PDF of a healthy sample
- Puts lots of weight on the Gaussian

Patient – 0.3342

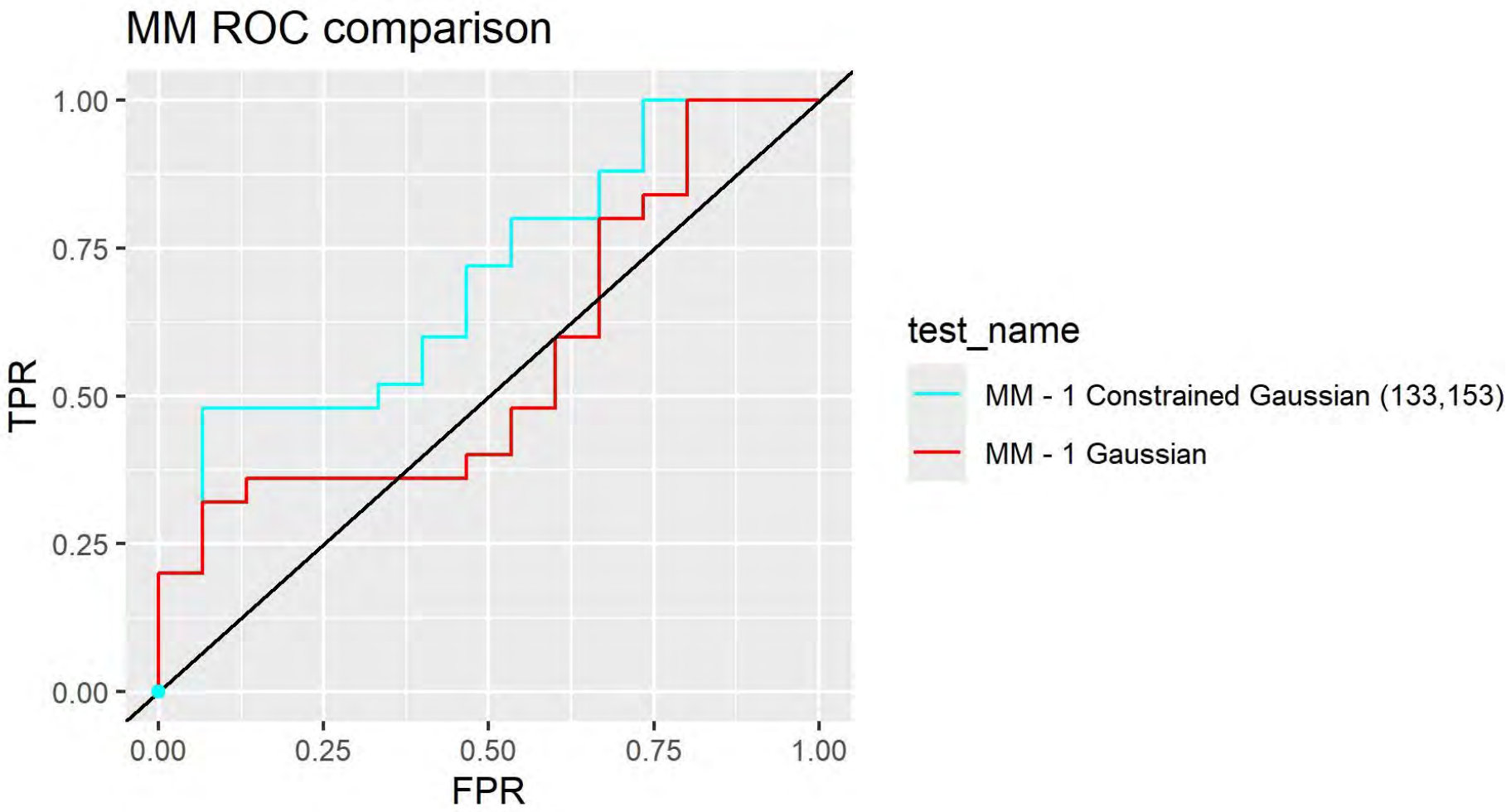




MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

MM – ROC Plot

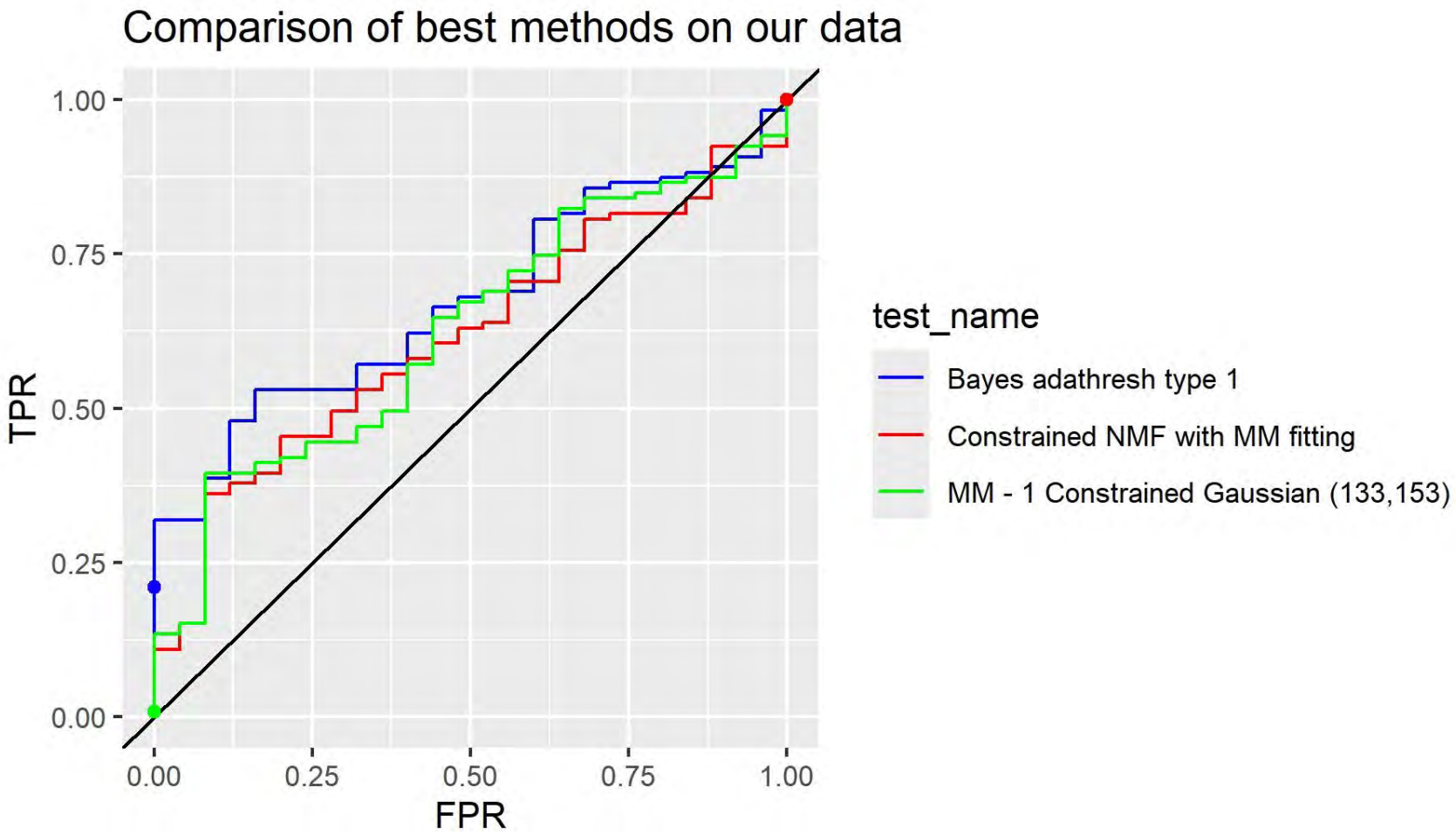


MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Comparing the methods

Overall comparison on the whole dataset



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Discussion

- Results indicate the bayes adaptive threshold performs best, gets reasonable results on our data.
- Most of our methods seemed to generalize to other samples well
- Performed poorly on low ctDNA samples, better on high ctDNA samples
  - Indicates these methods may not be overly useful for early detection.
- Models don’t show high correlation to other packages outputs, so they do provide a new signal.
  - Could combine this signal with others to generate better predictions, as other papers have done.
- Good performance of the bayes adaptive threshold method indicates it may be quite useful for labelling specific lengths as cancerous / non-cancerous and enhancing other methods performance.

Future research directions

- Using functional data analysis (or some other method) to run an 100+ dimensional difference in distributions test
- Investigating how results change over time for monitoring purposes.

UKRI MRC Biostatistics Unit
UNIVERSITY OF CAMBRIDGE

References

- [1] (Image) - Dao, J.; Conway, P.J.; Subramani, B.; Meyyappan, D.; Russell, S.; Mahadevan, D. Using cfDNA and ctDNA as Oncologic Markers: A Path to Clinical Validation. *Int. J. Mol. Sci.* **2023**, *24*, 13219. https://doi.org/10.3390/ijms241713219
- [2] - Mouliere F, Chandrananda D, Piskorz AM, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*. 2018;10(466):eaat4921. doi:10.1126/scitranslmed.aat4921. https://pubmed.ncbi.nlm.nih.gov/30404863/
- [3] - Nguyen, VC., Nguyen, T.H., Phan, T.H. *et al.* Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. *BMC Cancer* **23**, 233 (2023). https://doi.org/10.1186/s12885-023-10681-0
- [4] - Renaud G, Nørgaard M, Lindberg J, et al. Unsupervised detection of fragment length signatures of circulating tumor DNA using non-negative matrix factorization. Elife. 2022;11:e71569. Published 2022 Jul 27. doi:10.7554/eLife.71569. https://pubmed.ncbi.nlm.nih.gov/31271844/
- [5] (Image) - Zeng Z, Vo AH, Mao C, Clare SE, Khan SA, Luo Y. Cancer classification and pathway discovery using non-negative matrix factorization. *J Biomed Inform*. 2019;96:103247. doi:10.1016/j.jbi.2019.103247 https://pubmed.ncbi.nlm.nih.gov/31271844/

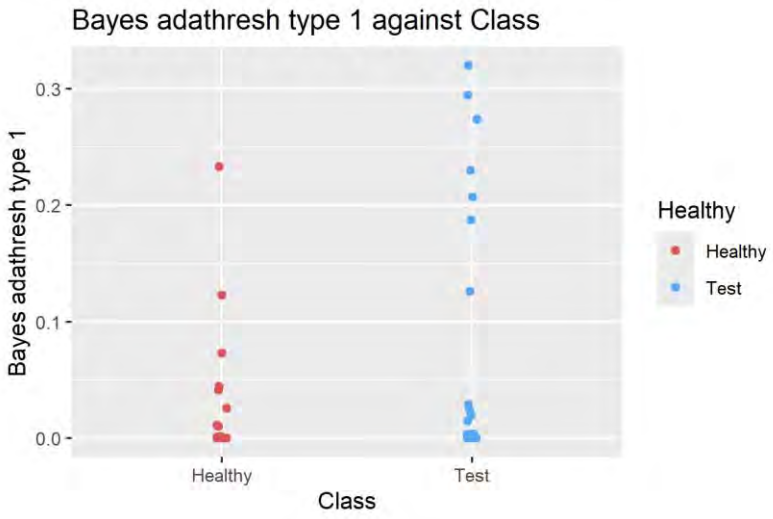
UNIVERSITY OF CAMBRIDGE

Supplementary slides

Bayes classifier output types - comparison

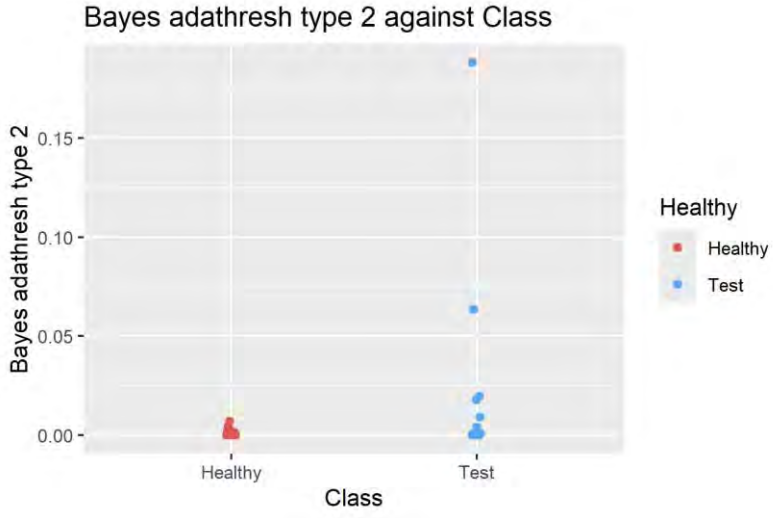
Type 1

The proportion labelled as likely cancerous



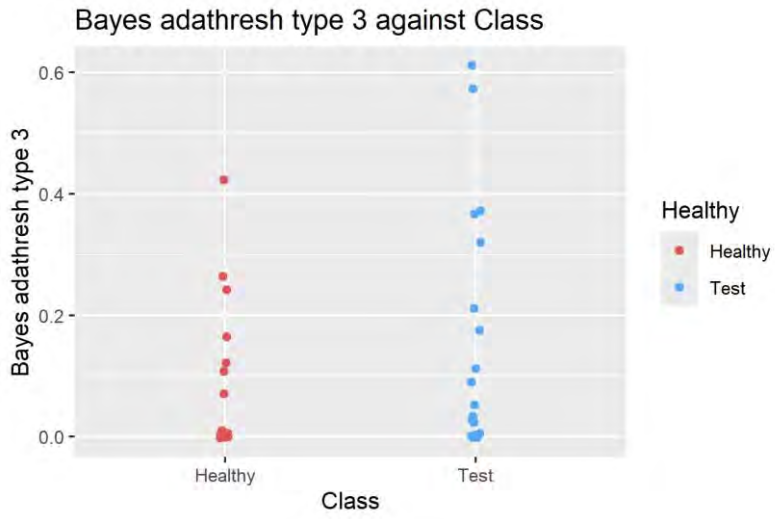
Type 2

The excess amount labelled as likely cancerous



Type 3

The number of fragment lengths labelled as likely cancerous

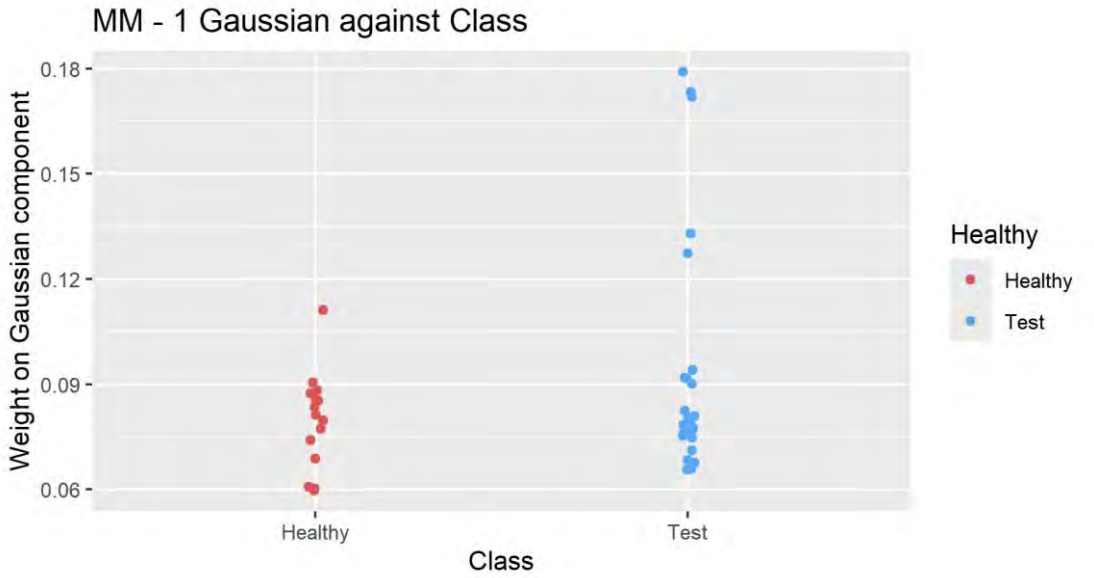


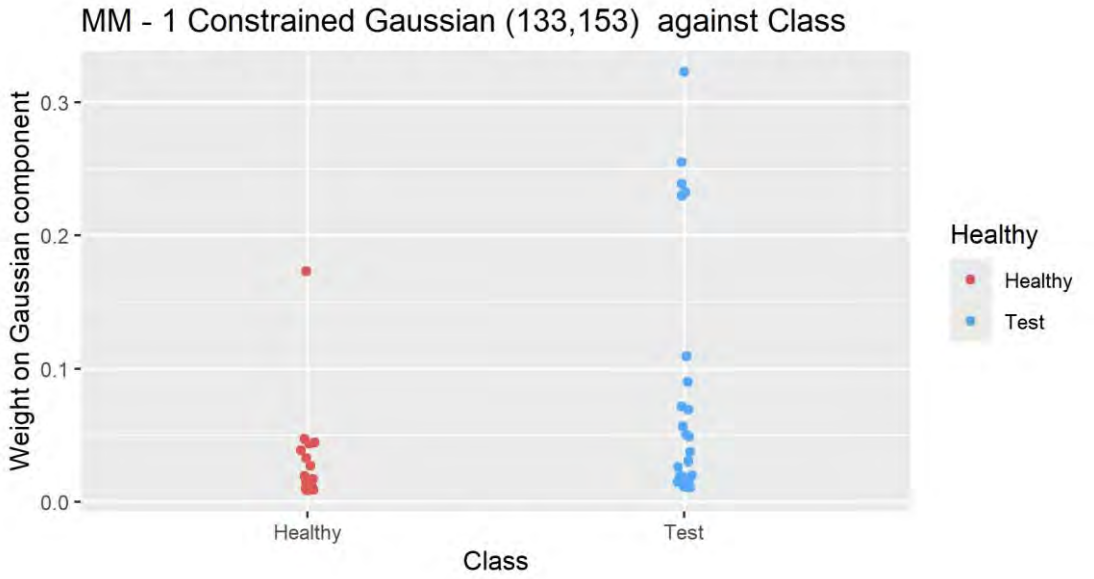
MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

MM Constraints

- In some cases, center of fitted gaussian was in a weird position with no observed link between that position and ctDNA.
- Tried constraining the mean to different ranges.
- Compared performance with and without constraints.





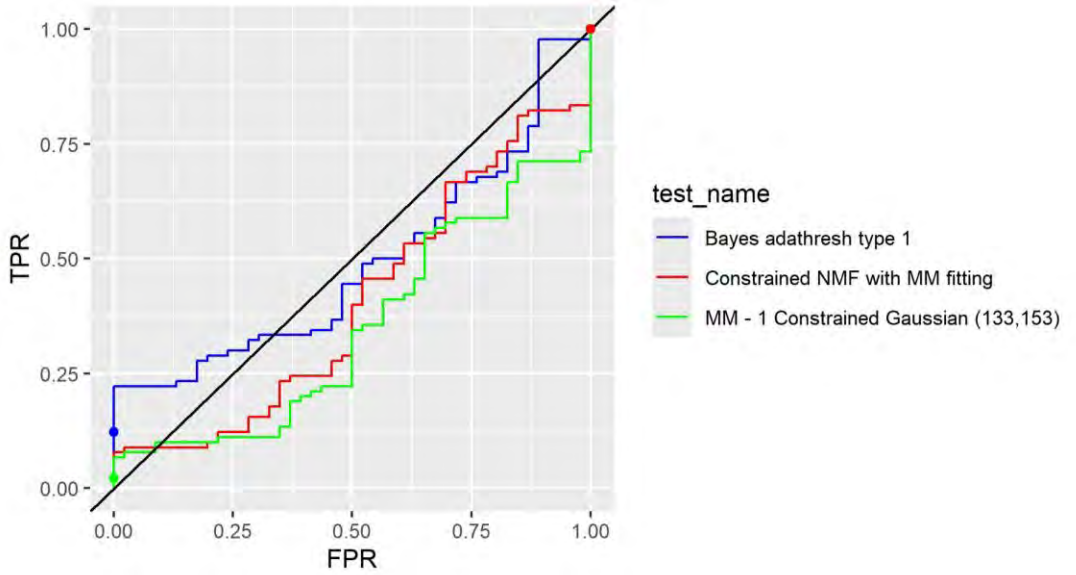
MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

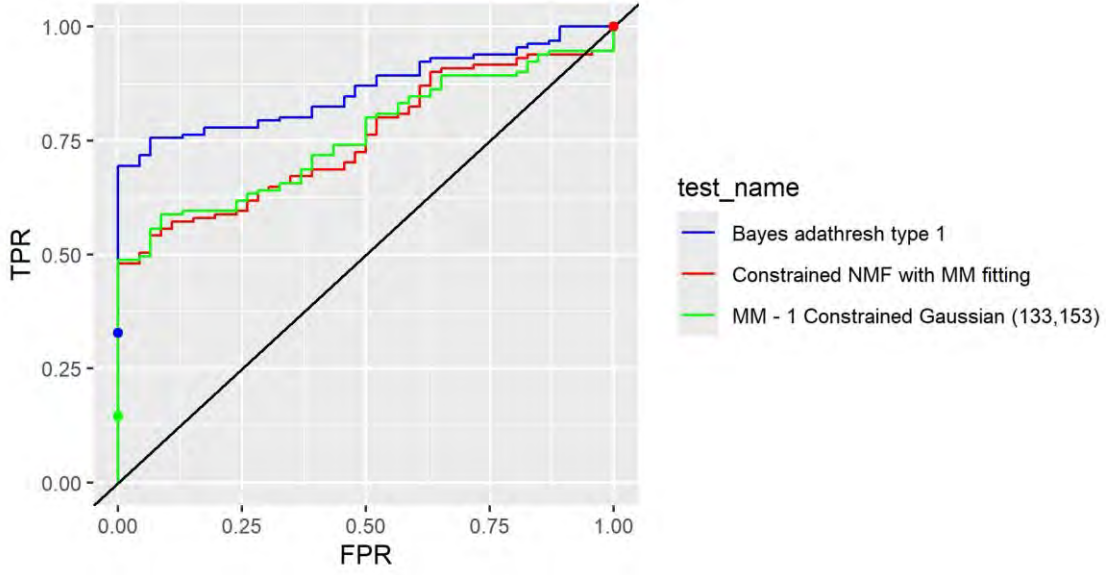
On another dataset

- Out of sample performance tested on second dataset (from [2]) with n=70 healthy samples and n=284 samples with cancer.
- Models were trained on the original dataset
- Tested on the Mouliere data.
- Data is broken up into three categories:
  - Healthy
  - Low ctDNA
  - High ctDNA

Healthy vs Low ctDNA



Healthy vs High ctDNA



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Non – Negative Matrix Factorisation

Non-negative matrix factorization (NMF)

- NMF was used to detect ctDNA in [4]

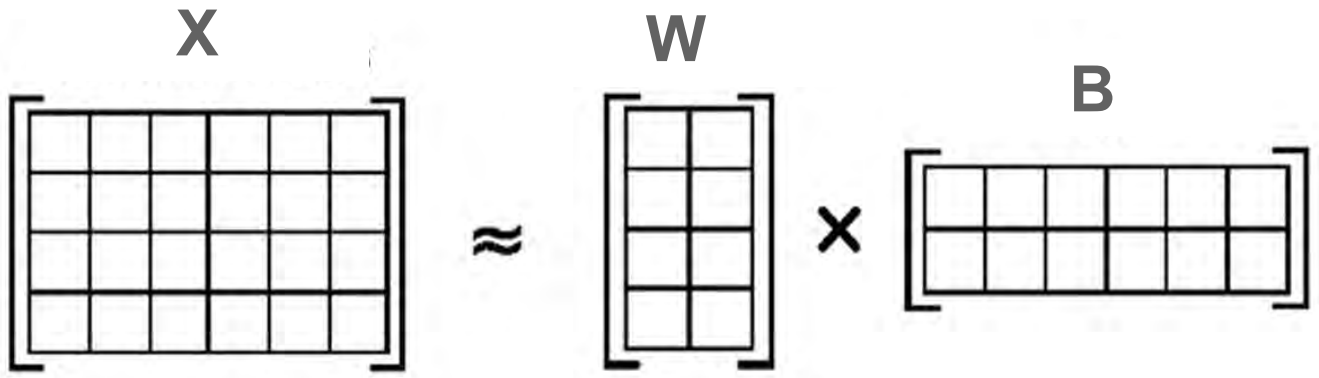
NMF Inputs:

- A matrix $X \in \mathbb{R}^{n \times p}$ - made of n samples of some p dimensional data
- A number of components r

NMF Outputs:

- $W \in \mathbb{R}^{p \times r}$ and $B \in \mathbb{R}^{r \times n}$ that minimize

$$||X - WB||_F^2 = \sum_{i,j} (X - WB)_{ij}^2$$



MRC Biostatistics Unit

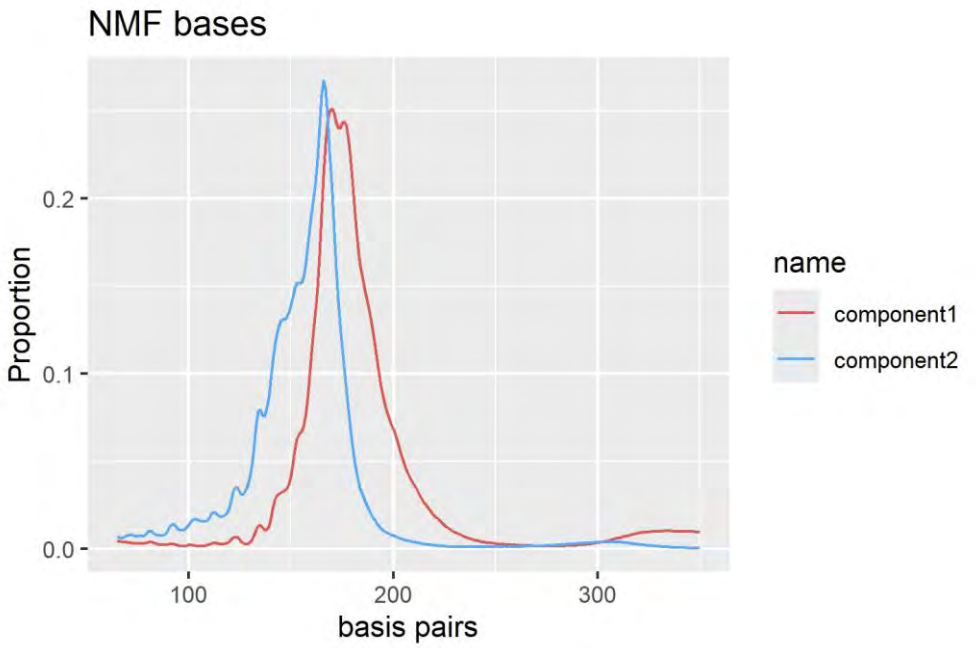
UNIVERSITY OF CAMBRIDGE

[4] – (Renaud, 2022)
[5] (Image) – (Zeng, 2019)

Applying NMF on our data

Components: Healthy cfDNA and the ctDNA. $r = 2$

$\mathrm{X} \in \mathbb{R}^{p \times n}$ - is a matrix with columns the proportions of fragment lengths (one column per sample).

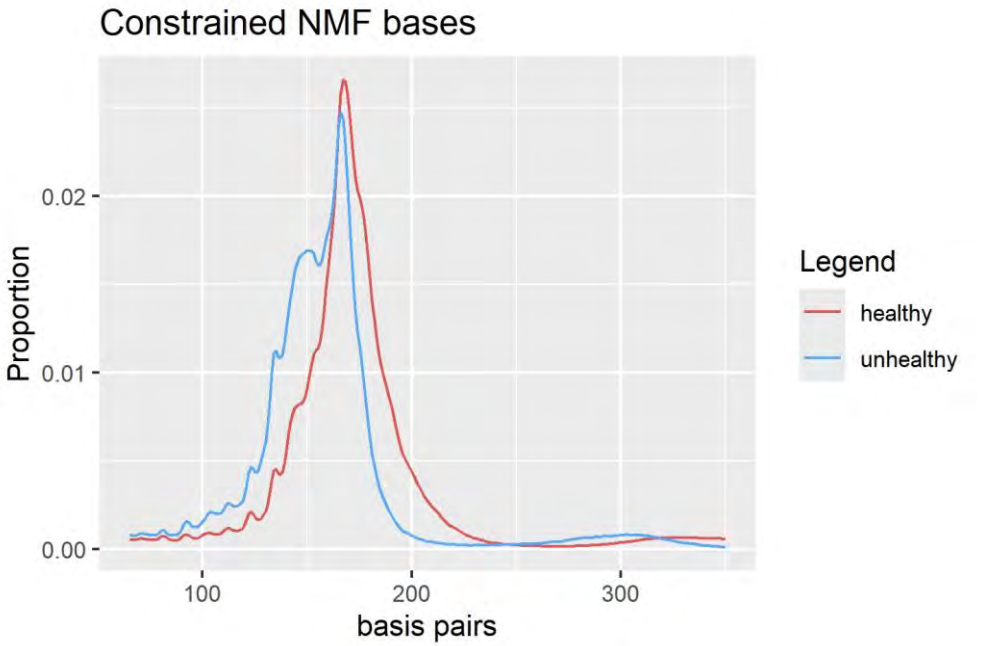


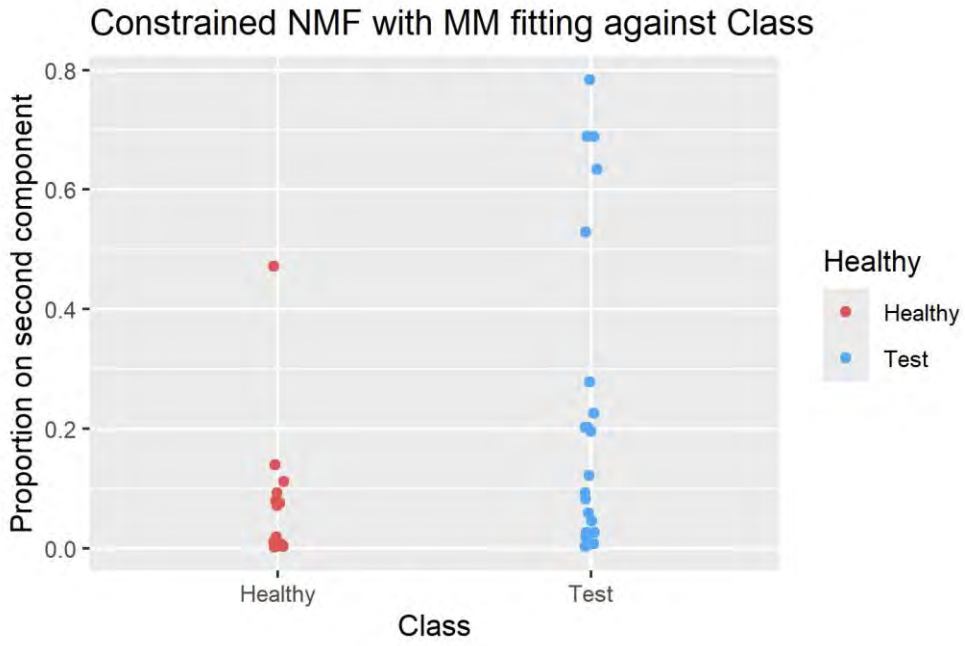
MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Constrained NMF on our data

In the train set we know which samples are healthy (and so have 0 ctDNA). We set the weight matrix W to have value 0 for the second component for all healthy samples (so the first component is the healthy distribution).

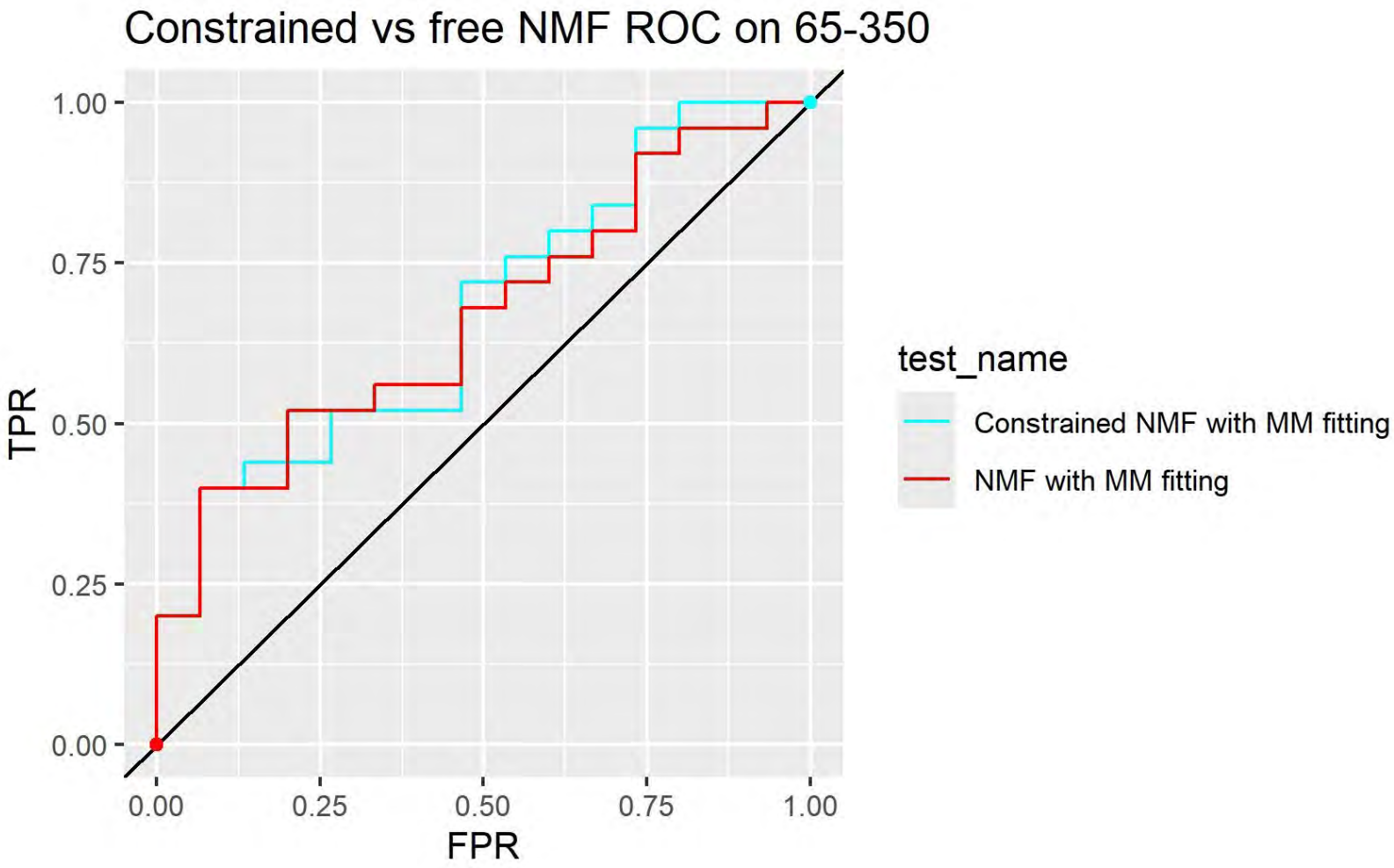




MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

NMF – ROC comparison



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Periodicity calculation

(From Mouliere, 2019):
10bp amplitude was determined as follows:

- Local maxima and minima in the range 75-150bp were identified (points s.t. y was max/min of [y-2,y+2]).
- Average of their positions across the samples in the train set was computed
  - Minima: 84, 96, 106, 116, 126, 137, 148
  - Maxima: 81, 92, 102, 112, 122, 134, 144
- Amplitudes computed with:
  - Sum of heights of the maxima – sum of heights of minima.
  - Height defined as proportion of fragments of that length.

Fitting mixture models

A single Gaussian

- Component 1 – PDF of a healthy sample
- Component 2 – Free Gaussian

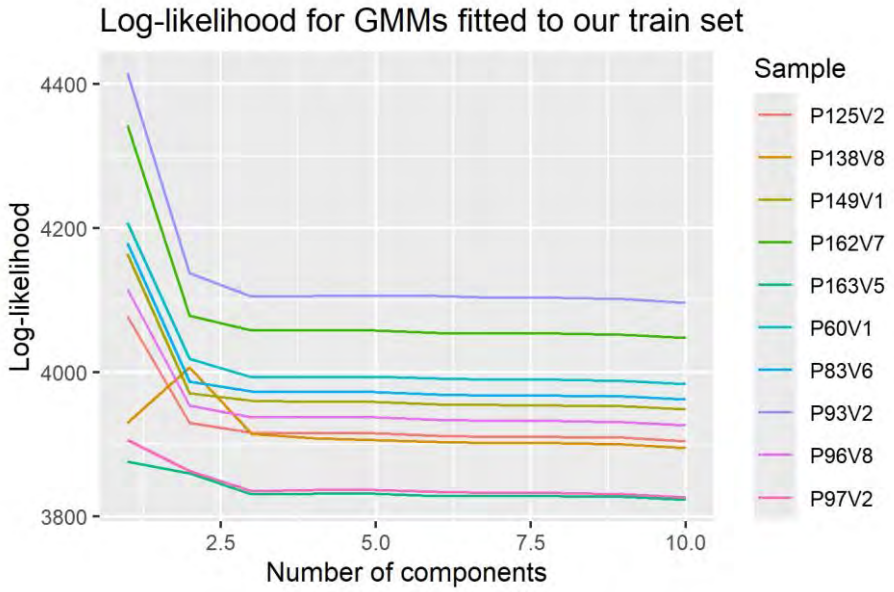
Adaptive thresholds

- Could set $T(x)$ as a quantile of the ratios between healthy samples.

Fitting GMM to healthy pdf

- We fit a GMM to our healthy PDF data.
- Need some number of components for the GMM
- Computed log-likelihood, BIC to determine the best number of components.

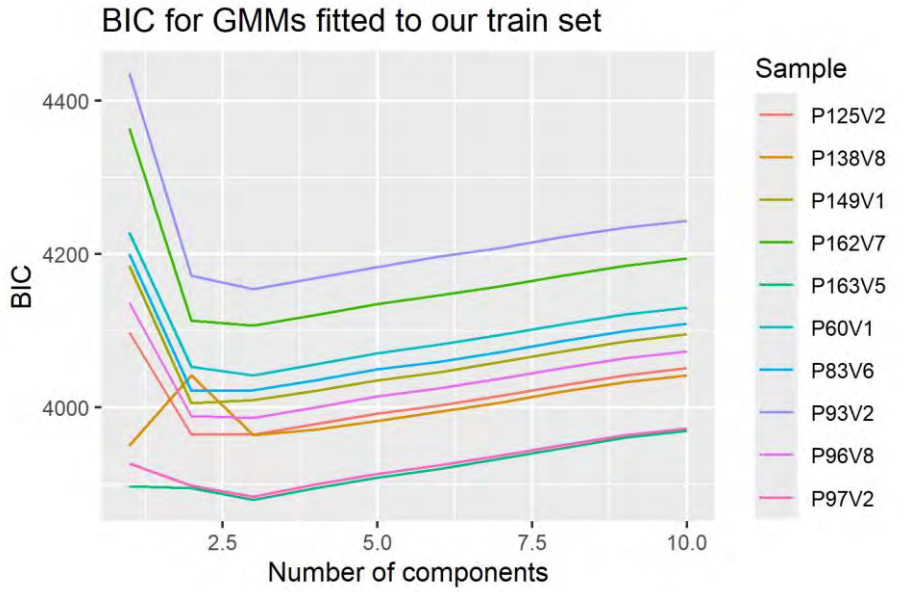
Negative Log-likelihood



Bayesian information criterion

- Can also compute the Bayesian information criterion
- We add a penalty term to the log-likelihood which penalizes having more dimensions in our model.

$$BIC = k\log(n) - 2\log(\hat{L})$$

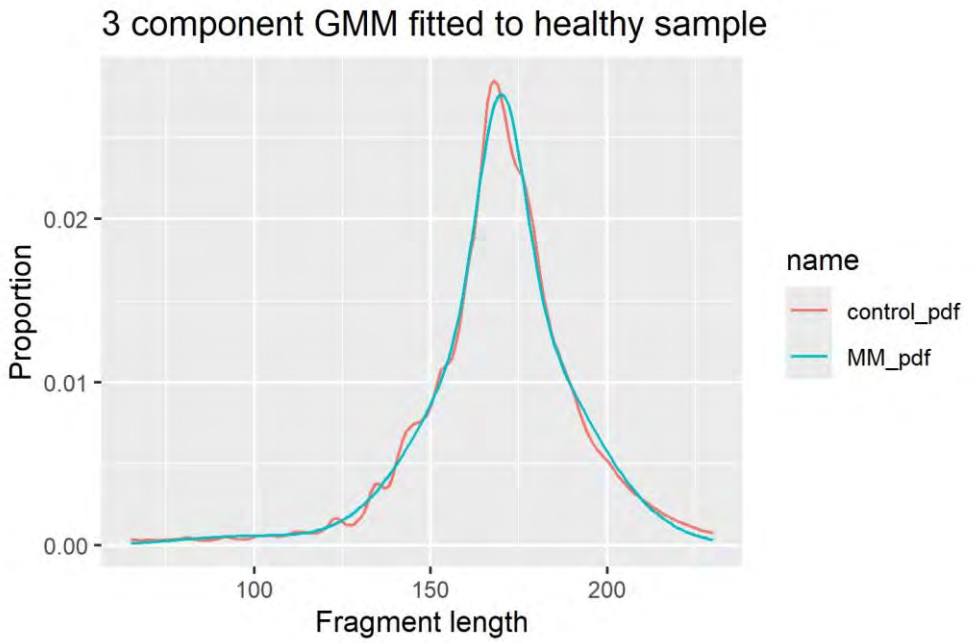


MRC Biostatistics Unit

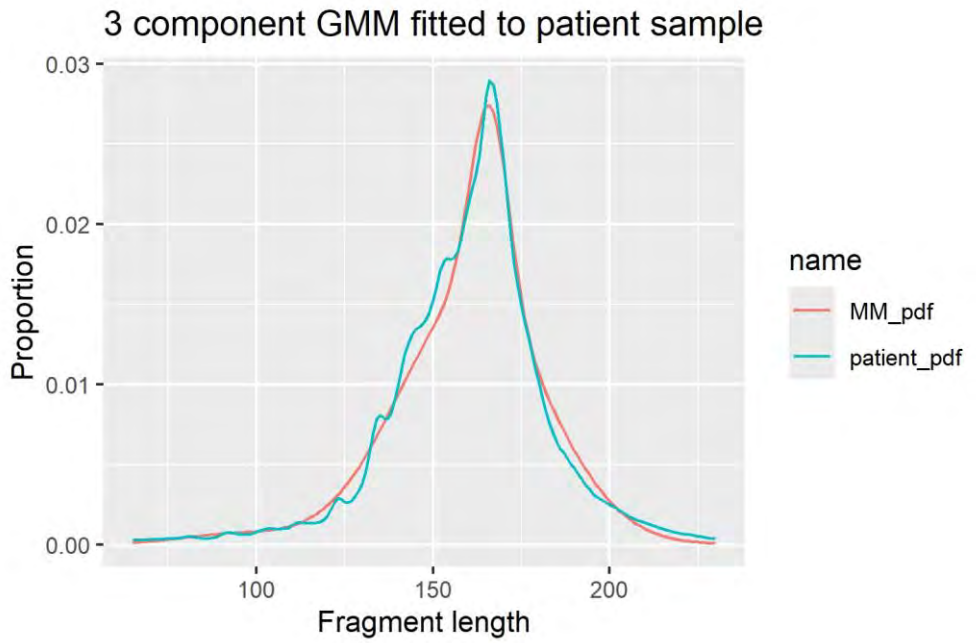
UNIVERSITY OF CAMBRIDGE

GMM fitted plots

Three component - healthy



Three component - patient



MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE