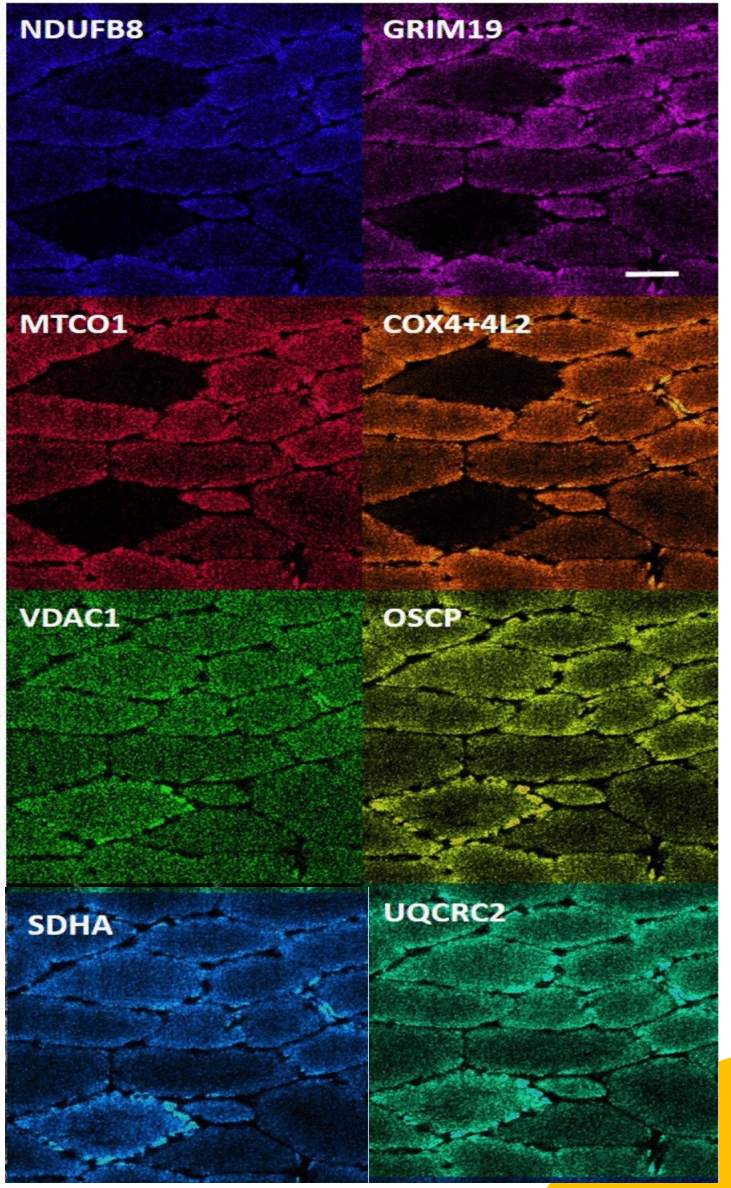
# CLASSIFICATION OF FIBRES FROM THE MITOCHONRIOAL DISEASE DATASET USING CLUSTERING TECHNIQUES

*MSc Data Science Project and dissertation 2020/21*
*by Frestie Ngongo*

# Background to the Project

- Clinical data collected from 13 anonymous patients – 10 with mitochondrial disease and 3 without the disease (as control)
- Dataset produced by Imaging Mass Cytometry (IMC) in a prior study by Warren et al., 2021
- They identified that a deficiency in oxidative phosphorylation proteins was a characteristic of mitochondrial disease.
- Classification as having the Mitochondrial Disease was based on the proteins that are part of the respiratory chain (RC), which is a step in oxidative phosphorylation.
- Looked at two proteins – NDUFA13 and NDUFB8 and their ability to be used to classify a fibre as RC deficient and non-RC deficient using unsupervised learning models:
    - K-means Clustering
    - Gaussian Mixture Model



# Proteins

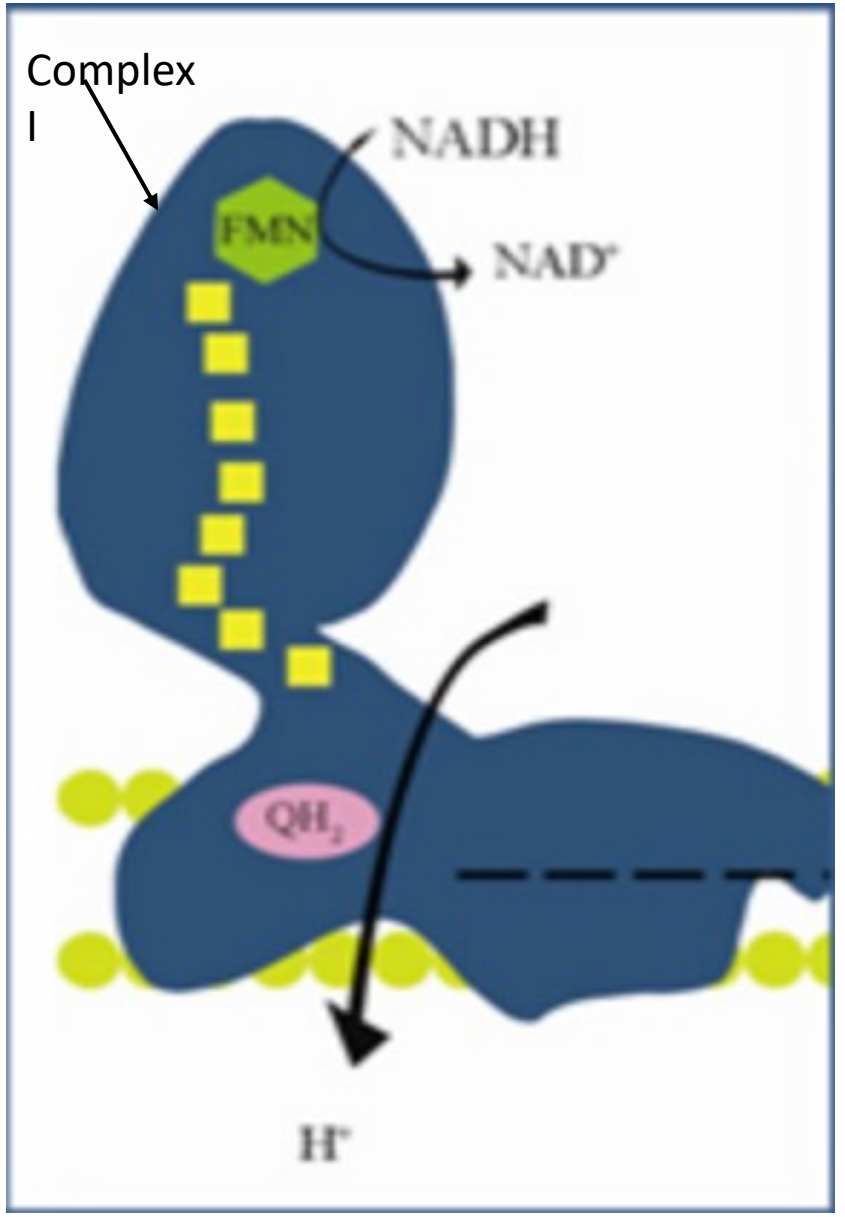
The 2 proteins investigated in this study were:

1. NDUFB8
2. NDUFA13

Both are part of Complex I (CI) of the respiratory chain and are used in the oxidative phosphorylation process during the transfer of electrons in the respiratory chain for the release of energy for the cell.

## **Why them?**

Using these proteins because complex I (CI) is known to be deficient in patients with mitochondrial diseases, which would correlate with deficiencies of these proteins, hence a deficiency in the respiratory chain protein.



# Work Undertaken

LOG(PROTEIN EXPRESSION) FOR THE 8 PROTEINS OF ALL THE 9000 FIBRES

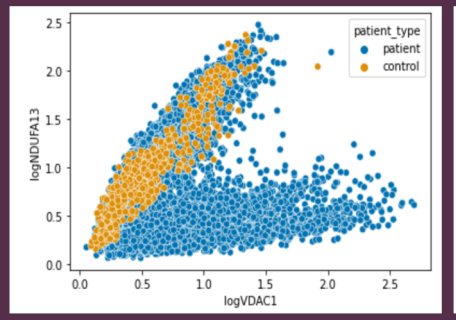
CREATED A DATA FRAME CONSISTING ONLY THE PROTEINS, PATIENT TYPE, PATIENT ID, CELL ID, SUBJECT GROUP

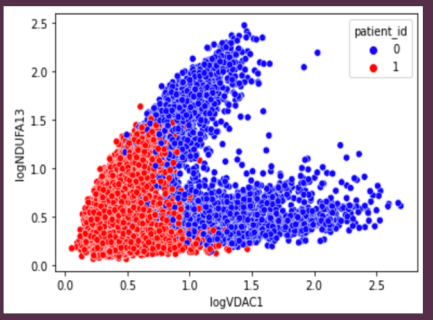
ANALYSED THE RAW DATA – PRESENTED AS A SCATTERPLOT OF LOG(PROTEIN EXPRESSION) VS THE LOG(VDAC1) PROTEIN EXPRESSION.

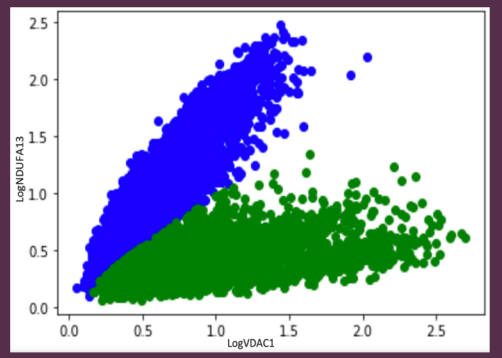
PREDICTED CLUSTERING OF ALL 9000 FIBRES USING K-MEANS AND GMM CLUSTERING MODELS

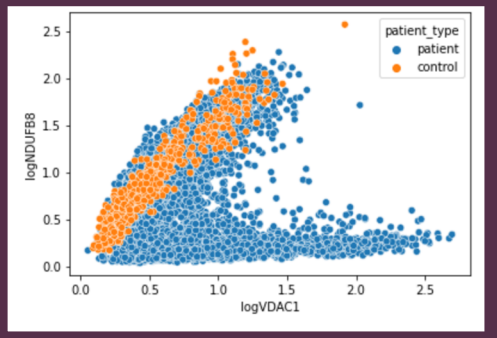
CALCULATED PROPORTIONS OF FIBRES CLASSIFIED AS RC (RESPIRATORY CHAIN) DEFICIENT FOR EACH PATIENT/CONTROL, FOLLOWED BY THE PROPORTION FOR EACH DISEASE TYPE.

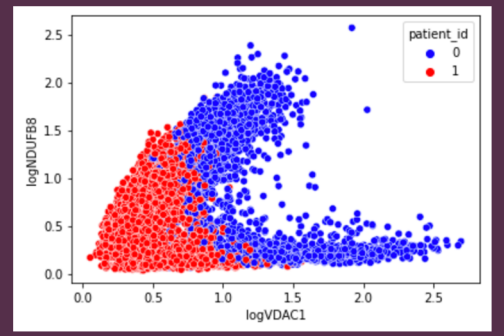
# Key Findings

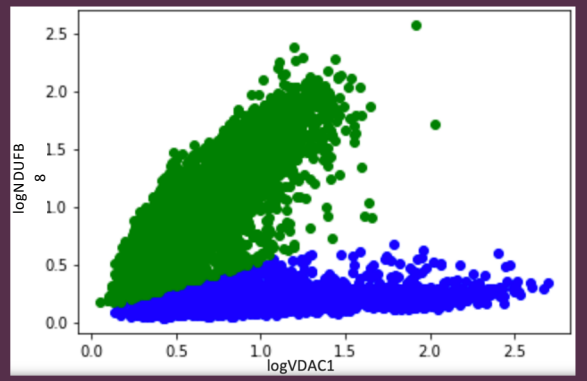



Comparing both the GMM and k-means plots of the two proteins, GMM produced a split that corresponds to the natural split of the data seen in the plot, whereas k-means did not, although it produced two clusters.

# K-means

- The control, especially C03 had a high proportion of RC-deficient fibres which was unexpected.
- CI had almost 100% of fibres as RC deficient so this met expectations
- New finding that deletion and MT-TL1 disease types had varying proportions for each patient which was surprising.

| INDIVIDUAL | PROPORTION OF RC DEFICIENT FIBRES (%) | DISEASE TYPE | PROPORTION OF RC DEFICIENT FIBRES (%) |
|---|---|---|---|
| C01 | 1.4% | | |
| C02 | 0% | Control | 19.0% |
| C03 | 80.9% | | |
| P01 | 97.6% | CI | 98.1% |
| P02 | 98.7% | | |
| P03 | 0.8% | Deletion | 33.8% |
| P04 | 84.8% | | |
| P05 | 20.1% | | |
| P06 | 76.4% | MT-TL1 | 39.8% |
| P07 | 49.5% | | |
| P08 | 54.8% | MT-TG | 54.8% |
| P09 | 50.2% | MT-TE | 50.2% |
| P10 | 80.9% | MT-TW | 80.9% |

# GMM

- Results from both proteins show that the GMM predicted the control as close to 0%.
- CI fibres were predicted to be approximately 100% RC deficient
- Both patients with the deletion variant were predicted to have a very low proportion of RC deficient fibres
- MT-TL1 is still showing each patient with varying RC deficient fibre proportions – likely that other factors play a role in the classification of this disease e.g. deficiency of other proteins produced by tRNA for oxidative phosphorylation
- The remaining diseases have a high proportion of RC deficient fibres

## NDUFB8

| Individual | Proportion of RC deficient Fibres (%) | Disease Type | Proportion of RC deficient Fibres (%) |
|---|---|---|---|
| C01 | 0.7% | | |
| C02 | 2.1% | Control | 1.4% |
| C03 | 0% | | |
| P01 | 99.7% | CI | 99.8% |
| P02 | 100% | | |
| P03 | 3.8% | Deletion | 4.0% |
| P04 | 4.3% | | |
| P05 | 23.7% | | |
| P06 | 11.3% | MT-TL1 | 30.7% |
| P07 | 68.9% | | |
| P08 | 87.3% | MT-TG | 87.3% |
| P09 | 42.5% | MT-TE | 42.5% |
| P10 | 83.6% | MT-TW | 83.6% |

## NDUFA13

| Individual | Proportion of RC deficient Fibres (%) | Disease Type | Proportion of RC deficient Fibres (%) |
|---|---|---|---|
| C01 | 0% | | |
| C02 | 0.3% | Control | 0.2% |
| C03 | 0% | | |
| P01 | 100% | CI | 100% |
| P02 | 100% | | |
| P03 | 2.6% | Deletion | 1.5% |
| P04 | 3.4% | | |
| P05 | 21.4% | | |
| P06 | 9.5 % | MT-TL1 | 28.2% |
| P07 | 65.3% | | |
| P08 | 83.6% | MT-TG | 83.6% |
| P09 | 37.7% | MT-TE | 37.7% |
| P10 | 80.9% | MT-TW | 80.9% |

## Conclusion

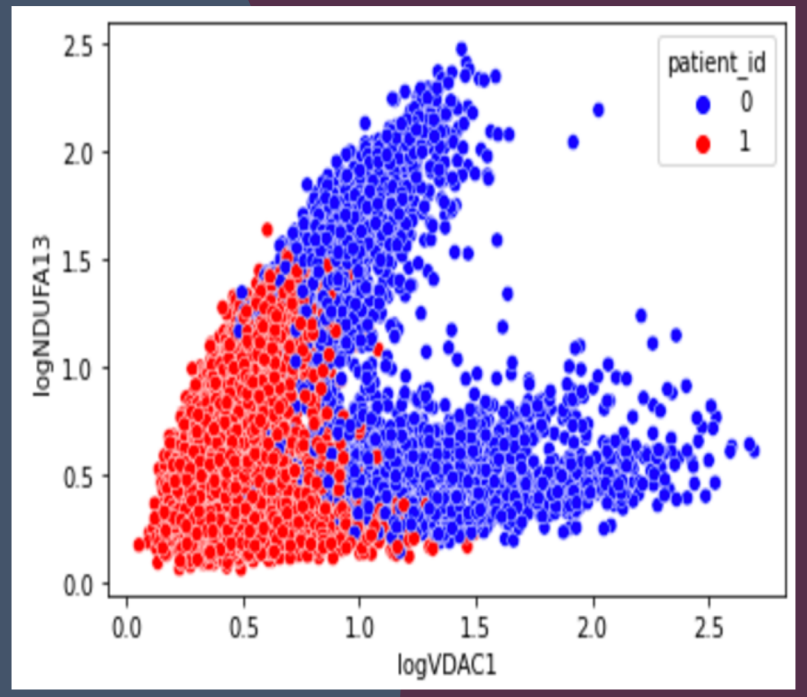
The k-means vs GMM models:

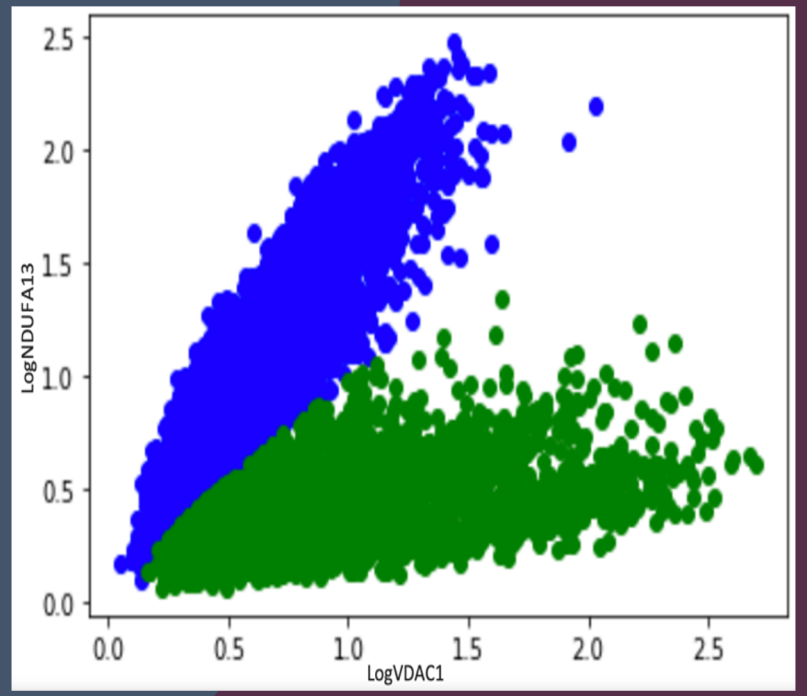
- Visual inspection of 2Dmito plots shows GMM captures the two forks of the V-shaped data whilst k-means does not

Quantitative inspection:

- Control expected to be 0%, GMM shows this but k-means does not
- CI expected to be 100%, both GMM and k-means show this
- Deletion expected to be low but the same for both patients which GMM shows. However, k-means shows varying results for each patient.

GitHub




THANK YOU