Early detection of Huntington's disease with ML

Shanley lab meeting: 29th June

Krutik Patel

Newcastle University

29th June 2021

Project Preamble

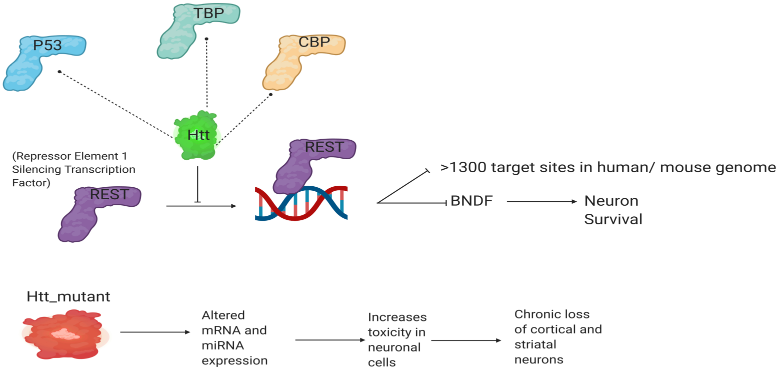
- I began a ML project earlier in the year. I have had to drop work on it due to other commitments and my PhDs impending end.
- Colleen (student) will join us for a few months and pick up where I left off.
- Today I will go over what I had done.
- Overall goal: Inform the group on what Colleen is getting herself involved in.

Presentation structure

- Background
- Data
- ML
- Further work

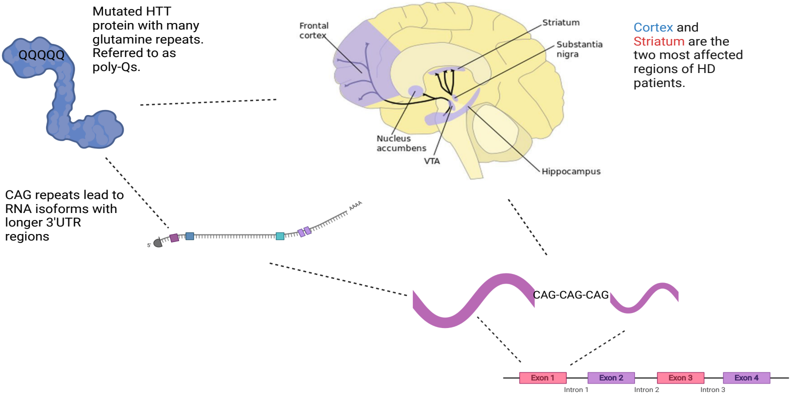
-Background

Huntington's disease leads to altered gene expression


TBP
P53
CBP
Htt
(Repressor Element 1 Silencing Transcription Factor)
REST
REST
>1300 target sites in human/ mouse genome
BNDF
Neuron Survival
Htt_mutant
Altered mRNA and miRNA expression
Increases toxicity in neuronal cells
Chronic loss of cortical and striatal neurons

- Mutant Htt protein leads to altered miRNA and mRNA expression.
- Could some of these miRNAs be biomarkers for HD?

Huntington's disease is caused by too many Qs



Mutated HTT protein with many glutamine repeats. Referred to as poly-Qs.

CAG repeats lead to RNA isoforms with longer 3'UTR regions

Cortex and Striatum are the two most affected regions of HD patients.

- What do the different numbers of Qs mean?

Different Qs can lead to alternate phenotypes

No onset: =<20Q

Middle aged onset:20-~75Q

Juvenile aged onset: >~75Q

- And are there differences between genders?

Some gender differences were found

- A recent project student, Bethany used this dataset to investigate male-female differences.
- She identified cholestoral synthesis to be differentially enriched between male-female mouse samples at some time-points.
- Though overall there were not huge differences; and this can justify treating samples from different genders as part of the catagory during ML research. – further discussed in Data section.

-Data

We have many mouse cortex samples

- We have data from 168 mouse cortex
- 5 outliers removed, so 163 mice
- RNAseq + miRNAseq was performed on each cortex, thus 326 individual data files
- The 163 mice can be divided by gender, age and Q mutation.

Data division by gender, age and Q

- The mice are 2, 6 or 10 months old at the time of sacrafice

- The mice can range from the following seven Q conditions: WT, Q20, Q80, Q92, Q111, Q140, Q175

- The mice are either male or female. The total number of males and females of each age and Q condition adds up to eight.

- To increase our number of samples per condition the genders were ignored.

Data division by age and Q

Age	Condition	Mice
2M	WT	7
	Q20	8
	Q80	8
	Q92	7
	Q111	8
	Q140	8
	Q175	8

Age	Condition	Mice
6M	WT	7
	Q20	8
	Q80	8
	Q92	8
	Q111	8
	Q140	8
	Q175	8

Age	Condition	Mice
10M	WT	8
	Q20	7
	Q80	8
	Q92	8
	Q111	8
	Q140	8
	Q175	7

Rephrased ML quesiton due to spread of data

- 7-8 samples size is small for ML classification. So I decided to make a HD or WT ML question.

- 2M = pups, 6M = young but breeding, 10M = fully formed skeletons and breeding.

- Thus we could further rephrase the question for early detection of HD or WT mice if we used the 2M data as a validation set and the 6M and 10M data for the training set.

Data division by training and validation

Data	WT	HD
Training	30	79
Validation	15	39

Feature selection by differential expression

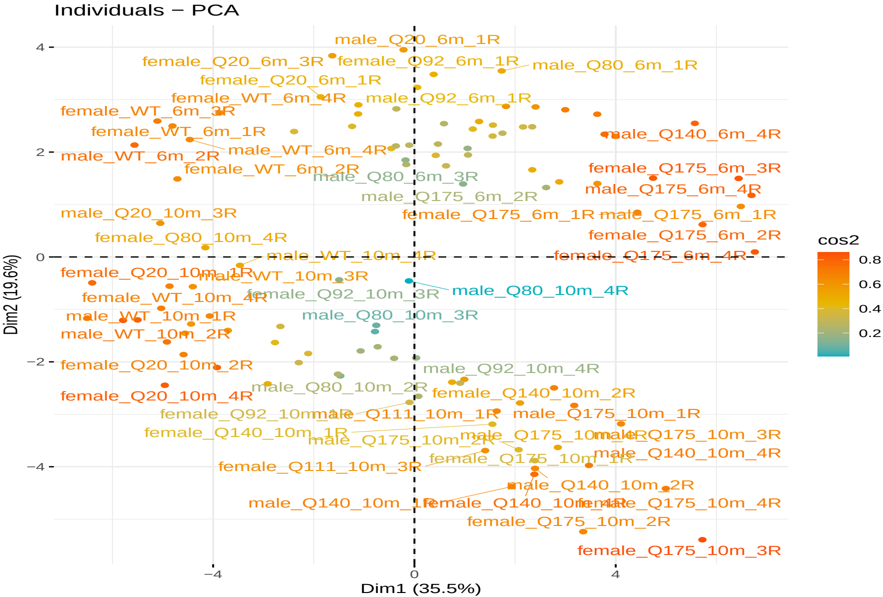
- To find genes to train, the 6 and 10 month were put through differential expression.

- 6M_HD/6M_WT and 10M_HD/10M_WT $(HD ==> Q20|WT == WT + Q20)$

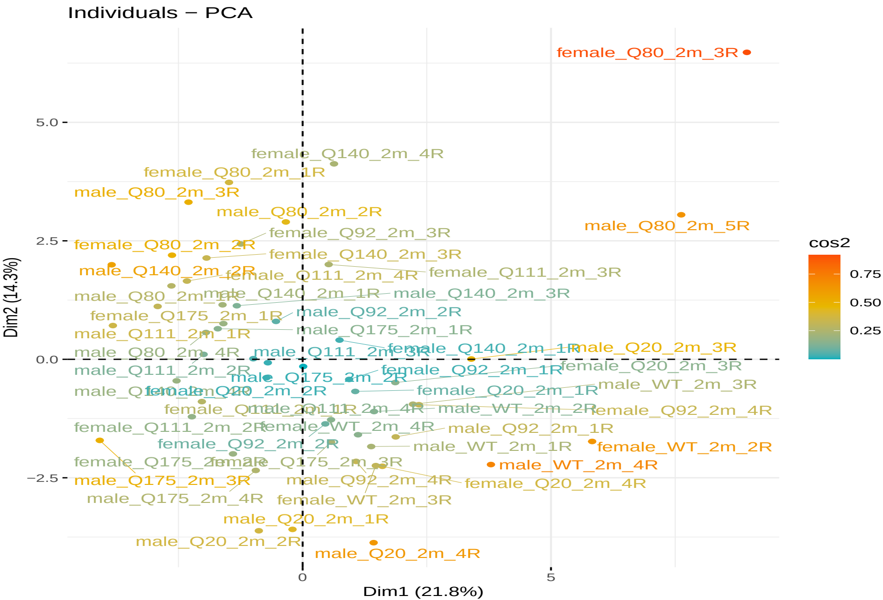
- Genes found to be significantly differentially expressed in both 6M and 10M analysis were taken forward for ML.

- These genes were also extracted from the validation set (2M)

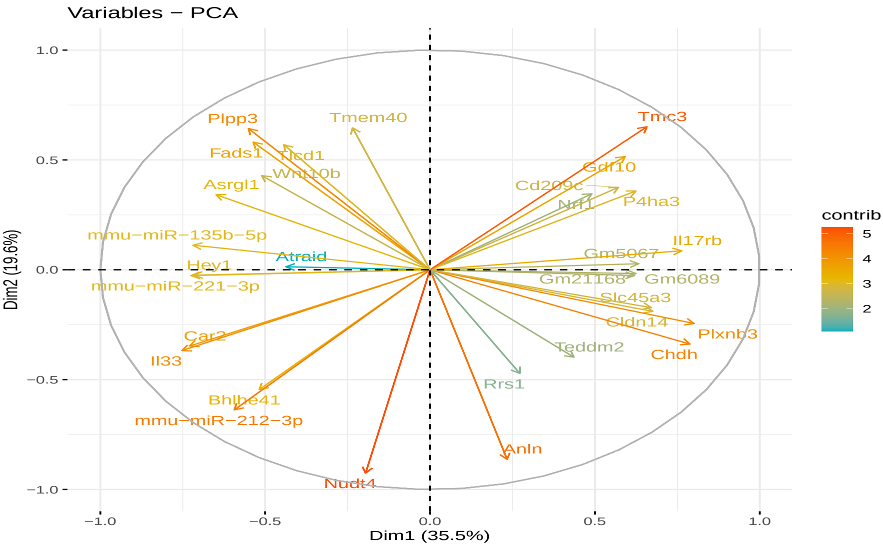
training data - spread of samples



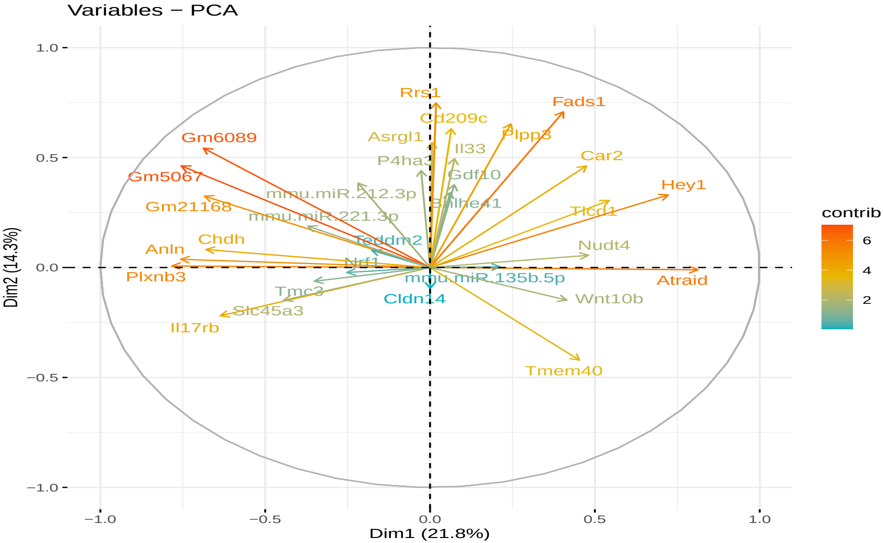
validation data - spread of samples



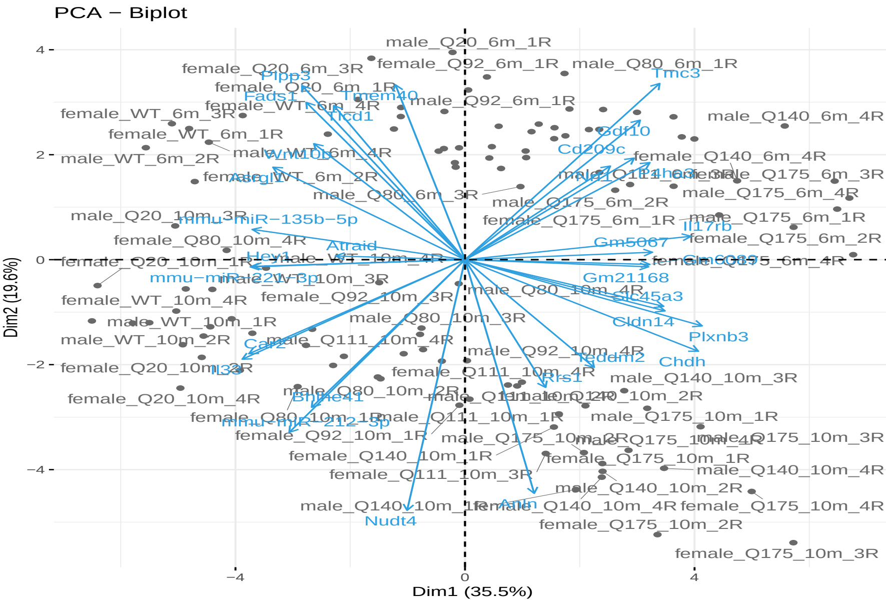
training data - spread of genes



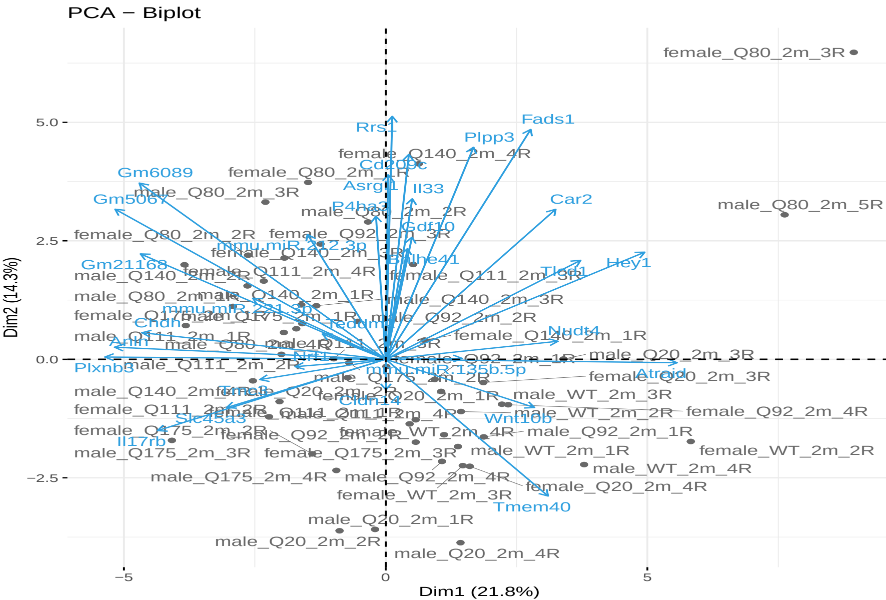
validation data - spread of genes



training data - spread of samples + genes



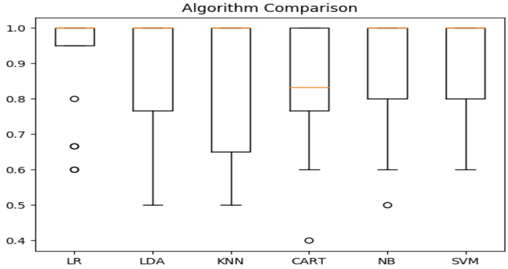
validation data - spread of samples + genes



-ML

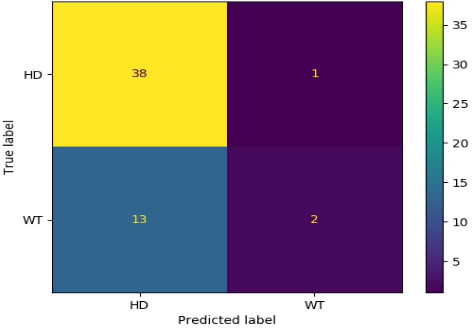
Followed a straight forward ML approach from popular resources

- split training (t) and validation (v) into x (values) and y (samples)
- scaled tx and vx data
- performed cross-validation of tx using several algorithms



Confusion matrix showed some mis-matches

- used LogisticRegression to train a model from tx which scored 91% accuracy
- shuffled cross-validation was used to on tx
- trained model was used to predict if the samples in vx were labelled as HD or WT, and this resulted in a 81% accuracy



-Further work

Are the miRNAs we found relevant in blood datasets?

- Ideally the miRNAs we find which can aid in early detection of HD will also be found in HD blood datasets.
- I have found a few of these datasets, I will be extracting their data to see which miRNAs are overexpressed during HD.
- May also download the associated striatum dataset to check which miRNAs are differentially expressed there.

Project students work

- Use standard ML feature selection method to find a different set of genes to train.

- Based on my attempts, perhaps use more genes for the task. Will be up to them how they do this.

- Overall goal: Try to get over 81% accuracy.