Using Machine Learning to identify microRNA biomarkers for predisposition to Juvenile Onset Huntington's Disease

Patel K[1], Sheridan C[1], Shanley DP*[1]

1. Campus for Ageing and Vitality, Biosciences Institute, Newcastle University, Newcastle upon-Tyne, NE4 5PL, UK.

Abstract

Background: Juvenile-Onset Huntington's disease (JOHD) is a rare form of Huntington's disease (HD) which leads to chronic neurodegeneration which begins prior to the age of 25. Like HD, JOHD is triggered by a large expansion of CAG nucleotides in the *HTT* gene which leads to neurotoxicity and aberrant gene expression. However, unlike HD, in JOHD the relationship between the length of CAG expansion and age of disease onset is non-linear. Thus, it would be of interest to identify molecular biomarkers which indicate predisposition to the development of JOHD, and as microRNAs (miRNAs) circulate in bio-fluids they would be particularly useful biomarkers.

Methods: We explored a large JOHD miRNA-mRNA expression dataset (GSE65776) to establish appropriate questions that could be addressed using Machine Learning (ML). We sought sets of features (mRNAs/ miRNAs) to predict JOHD or WT samples from aged or young mouse cortex samples, and we asked if a set of features could predict predisposition to JOHD or WT genotypes by training models on aged samples and testing the models on young samples. We used a k-best strategy which included oversampling for unbalanced classes, min_max scaling and 5-fold cross-validation. Several models were created using ADAboost, ExtraTrees, GaussianNB and Random Forest, and the best performing models were further analysed using AUC curves and PCA plots. Finally, genes used to train our miRNA-based predisposition model were compared to HD patient bio-fluid samples.

Results: Our testing accuracies were between 66-100% and AUC scores were between 31-100%. We generated several excellent models with testing accuracies >80% and AUC scores >90%. We also identified homologues of *mmu-miR-154-5p*, *mmu-miR-181a-5p* and *mmu-miR-212-3p, mmu-miR-378b, mmu-miR-382-5p* and *mmu-miR-770-5p* to be circulating in HD patient blood samples at p.values of <0.05.

Conclusions: We generated several age-based models which could differentiate between JOHD and WT samples, including an aged mRNA-based model with a 100% AUC score. We also identified several miRNAs used to train our miRNA-based predisposition model which were detectable in HD patient blood samples, which suggests they could be potential candidates for use as non-invasive biomarkers for JOHD/ HD research.

Keywords: Neurodegenerative Disease, Machine Learning, Bioinformatics, Juvenile-Onset Huntington's Disease, microRNAs, Transcriptomics, Biomarkers,

Correspondence: daryl.shanley@ncl.ac.uk, nkp68@ncl.ac.uk

Introduction

Huntington's Disease (HD) is a rare autosomal dominant disease, characterised by the progressive destruction of cortical and striatal neurons. HD patients have decreased motor skills, memory, and overall ability to maintain themselves without aid (Casella et al. 2020). HD affects 9.3-13.7 individuals per 100,000 in western populations (Evans et al. 2013; Ohlmeier et al. 2019). Another form of the neurodegenerative disorder is Juvenile Onset

Huntington's Disease (JOHD) which affects 1-9.6% of reported HD cases (Quarrell et al. 2012). CAG nucleotide repeats in the Huntingtin gene (*HTT*) leads to the translation of mutated HTT proteins with large a polyglutamine (poly-Q) repeat chain. Genetic screening of the CAG repeats is performed regularly as the length of the CAG repeats inversely correlates with adult-onset HD. Having >27 CAG repeats is indicative of no penetrance of HD, 27-35 CAG repeats indicate an intermediate chance of penetrance, and <36 CAG repeats will lead to full penetrance of HD (Semaka, Collins & Hayden 2010; Rosenblatt et al. 2012). Patients with at least 60 CAG repeats develop JOHD (Evans et al. 2013). HD patients show disease phenotypes during middle-aged life, and JOHD patients will show symptoms prior to age 25 (Casella et al. 2020). Understanding the underlying mechanisms of the HTT protein is important in determining how the neurotoxicity is triggered. HTT is found in the cytoplasm of striatal and cortical neurons (Fusco et al. 1999). Many HEAT binding motifs are found on HTT implying it has roles as a scaffolding protein and is involved in many protein interactions (Takano & Gusella 2002). However, a large CAG nucleotide expansion in the HTT protein halts its ability to exit the nucleus and thus make mutated HTT (mHTT) incapable of interacting with other scaffolding proteins: B-tubulins, microtubulins, dynein, dynactin (Hoffner, Kahlem & Djian 2002; Cornett et al. 2005; Caviston et al. 2007). Gene expression changes indicate that HTT does have roles as a transcriptomic regulator. A well characterised example is normal HTT proteins bind and sequester REST (repressor element-1 transcription factor) a negative regulator of BDNF (brain-derived neurotrophic factor) which is a neuron survival factor. Mutated HTT cannot sequester REST, leading to decreased BDNF levels, which has been observed in HD patients (Zuccato et al. 2003; McFarland et al. 2014).

Currently there are no therapies for HD or JOHD, and though age is a key risk factor for both diseases, further mechanistic knowledge is needed (Panas et al. 2008; Pan & Feigin 2021). This is shown by previous explorations of the length of mutations which cause JOHD showed non-linear correlations between age of onset of the disease and CAG repeats (Squitieri et al. 2006; Cronin, Rosser & Massey 2019; Schultz, Moser & Nopoulos 2020). Furthermore, twin studies showed individuals with the same genotype did not have uniform experiences HD (Georgiou et al. 1999; Panas et al. 2008). Further understanding of genetic and epigenetic factors could lead to novel insights for HD /JOHD research. Transcriptomic studies have shown mHTT changes expression profiles of neuronal of mRNAs and non-coding RNAs such as microRNAs (miRNAs) (Langfelder et al. 2016; Langfelder et al. 2018). miRNAs are small (18-22 nucleotides) single stranded RNAs which negatively regulate specific mRNAs via complementary binding (Bartel 2009). miRNAs contain a 6-8 nucleotides long tract called the seed site on their 5' end which complementary binds to target sites on the 3'-UTR of the mRNA (Doench & Sharp 2004). Biogenesis of a miRNA is a tightly regulated process which begins with transcription via RNA polymerase II (Lee et al. 2004). This produces a pri-miRNA (70-100 nucleotide long double stranded RNA hairpin) which will be processed by nuclear proteins DROSHA and DGCR8 (Han et al. 2004). This biogenesis step creates a pre-miRNA hairpin which are detected as cargo by EXP5 and transported to the cytoplasm (Lund & Dahlberg 2006). Cytoplasmic proteins DICER and TRBP perform further biogenesis and recruits the RISC complex which will attach itself to one of the mature miRNA strands which will act as the guide for the RISC complex (Chendrimada et al. 2005; Tsutsumi et al. 2010). The RISC complex comprises of sub-units such as AGO2 and a GW182 protein, and in mammals the GW182 protein is commonly responsible to mRNA silencing by orchestrating de-capping, de-adenylation and 5'-3' decay events (Figure 1) (Chendrimada et al. 2005; Rehwinkel et al. 2005; Yamashita et al. 2005; Zekri et al. 2009; Jonas & Izaurralde 2013). Both the seed site of the miRNA and the binding sites found on the

3'-UTR of target mRNAs are highly evolutionarily conserved (Friedman et al. 2009). A major interest for neurodegenerative research in miRNAs is their ability to be shuttled in bio-fluids as they could be used as non-invasive biomarkers which can be measured in blood plasma or cerebrospinal fluid (CSF) to assess the health of brains (Valadi et al. 2007; Kumar et al. 2017).

In this paper we used machine learning (ML) to identify biomarkers for JOHD in different age groups and we also used ML to predict predisposition of young samples to JOHD (Figure 2). A valuable dataset to explore potential biomarkers was stored in GSE65776. This dataset contained 168 RNAseq and miRNAseq samples from male and female mice cortexes, and the mice were sacrificed at 2, 6 and 10 months old. Most samples had knock-in mutations which would lead to JOHD in humans, and these were: 80, 92, 111, 140 and 175 CAG repeats in the *HTT* gene. This dataset also included WT samples and positive control samples which had 20 CAG repeat knock-ins (Langfelder et al. 2016; Langfelder et al. 2018). The initial data generation and analysis paper showed striatal neuron samples had greater gene expression changes than cortical neuron samples, however we were more interested in cortical samples as cortical loss has been seen as an early pathological event of HD and this would be useful for our predisposition detection queries (Rosas et al. 2002). Initial data exploration with differential expression (DE) analysis and the *TimiRGeN R* package identified a batch effect in the data, high homology, and some gender differences (Patel et al. 2021). Based on this, the data was corrected (batch effect removed, gender differences reduced) and re-labelled so that JOHD and WT samples could be differentiated using ML. While miRNAs and mRNAs were kept separated, three lines of enquiry were established: identification of post-symptomatic markers in aged (10m) samples, identification of pre-symptomatic markers in young (2m) samples, and identification of predisposition markers by training on aged samples and testing on young samples. Several high-quality models were produced as training and testing accuracies ranged between 73%-100% and 66-100% respectively. Finally, we validated the miRNAs selected to train the best performing JOHD predisposition model by contrasting results from miRNA-based HD patient bio-fluid samples.

Materials and Methods

Data download and raw data pre-processing

Fastq files were downloaded from GSE65776 (GSE65770 - mRNA and GSE657679 - miRNA) using *SRA-toolkit* (Leinonen, Sugawara & Shumway 2011). Quality of the fastq files were checked using *FASTQC* (Andrews & others 2010). For miRNA pre-processing, Mus_musculus.GRCm38.cdna.all.fa was used to create index files with *Bowtie* and mature miRNA count calling was performed with *miRDeep2* (Langmead 2010; Friedländer et al. 2012). mRNA samples were aligned into bam files using *Salmon*. mRNAs were converted into bam files and bam index files were created with *Samtools*. *Salmon* was used for alignment of the bam files to transcriptome Mus_musculus.GRCm38.cdna.all.fa (Patro et al. 2017). *Tximport* was then used to normalise mRNA data to gene count levels (Soneson, Love & Robinson 2015).

Initial data exploration

The mRNA and miRNA data consisted of 168 samples each. Seven genotypes were present: WT, 20 (positive control), 80, 92, 111, 140 and 175 CAG repeats. The data consisted of three age groups: 2-month-old (m), 6m, 10m. Male and female samples were evenly distributed for most genotypes. Outliers were detected using PCA and removed, and lowly expressed genes were removed if their value was <10 in at least 1/3rd of the total number of samples leaving 332 miRNAs and 13715 mRNAs. With *DESeq2* gender and age matched DE contrasts were

made between 20 CAG/WT, 80 CAG/WT, 92 CAG/WT, 111 CAG/WT, 140 CAG/WT, 175 CAG/WT (Love, Huber & Anders 2014). Genes with a Benjimini-Hochberg FDR of <0.05 were kept. Significantly expressed genes from the gender based DE analysis was taken forward for miRNA-mRNA integrated analysis with the *TimiRGeN R* package for pathway enrichment analysis (Patel et al. 2021).

Batch removal and removing variance from gender differences

A batch effect in the 6m for the miRNAs and mRNAs was suspected after DE analysis so the 6m was removed (Supplementary Figure 1). Variance from gender differences were reduced with *combat* (Leek et al. 2012). One 10m sample was removed by pca analysis, as it was not within 6* standard deviations from the mean.

Data preparation prior to ML

We found this data to be highly homogenous based on the lack of differentially expressed genes (DEGs) found when contrasting different mutation (Q) lengths with the WT samples, so we increased our sample power by labelling all Q80, Q92, Q111, Q140 and Q175 samples as "JOHD" samples, and the WT and Q20 samples were labelled as "WT" samples. Male and female samples from the same mutation categories were also labelled together after correcting for gender-based variance. For the miRNA and the mRNA data, genes were removed if their expression was <10 in at least half of the samples. Leaving 519 and 16432 genes for the miRNA and mRNA data respectively. The mRNA data was further filtered to only contain genes which showed a log2fc >0.2 or <-0.2 (after contrasting JOHD and WT samples with DE), leaving a total of 532 mRNAs. Normalised gene expression levels were extracted from *DESeq2* normalised counts.

Identified questions to investigate with ML

In line with data centric AI, we identified appropriate questions based on our data. We were interested in six questions: a) which set of miRNAs are the best predictors of JOHD after the onset of the symptoms (10m), b) which set of miRNAs are the best predictors of JOHD prior to the onset of the symptoms (2m), c) using miRNAs can we use the aged samples (10m) to predict predisposition to JOHD on young samples (2m), d) which set of mRNAs are the best predictors of JOHD after the onset of the symptoms, e) which set of mRNAs are the best predictors of JOHD prior to the onset of the symptoms, f) using mRNAs can we use the aged samples to predict predisposition to JOHD on young samples.

Data processing for ML and model performances

ML was performed with scikit-learn 2.2.4 (Pedregosa et al. 2011). miRNAs and mRNAs were analysed separately. 10M and 2M data were separated and split into training and testing sets at a 0.8:0.2 ratio, without shuffling. SMOTE was used to create random synthetic WT samples for each training step (but not for testing), and min_max scaling was performed on training and test data (Lemaître, Nogueira & Aridas 2017). 5-fold cross validations were performed during all model training steps. Recursive feature elimination (RFE) was performed using a linear SVM algorithm and a 5-fold cross-validation at an attempt to remove low variance genes from the training data. Eight popular classifiers were trained and tested, using the default settings for each algorithm, and these were: LinearSVM, PolynomialSVM, Guassian Process, Extra Trees (ET), Random Forest (RF), Neural Network, Adaboost (ADA), Gaussian Naïve Bayes (GNB). We found RFE led to overfitting. A K-best feature selection was instead utilised to identify if similar or better predictions could be generated with fewer genes. The four best performing classifiers (ADA, ET, GNB RF) for the

miRNA data from RFE were taken forward to determine the optimal features in a robust k-best selection approach. The 1-100 best features found by the k-best method (highest F-score) were sequentially used to train models using each of the four classifiers. This process was repeated 100 times, resulting in 100 x 100 x 4 (40,000) trained models per ML question. To limit the randomness of some classifiers, the best three training accuracies from each classifier were taken forward to create hyperparameter tuned models (Supplementary Table 1). For ADA number of estimators and learning rate were parametrised, for ET number of estimators and minimum splits were parameterised, for GNB the use of an automated variance smoother was contrasted against no variance smoother and if the results were the same – the scores from no variance smoother selected, and for RF number of estimators was parameterised. Training accuracies, testing accuracies, precision scores, recall scores, f1 scores and confusion matrices were calculated for each model (Supplementary Table 2). ROC_AUC and PCA plots were made for the best performing model from each of the six questions (Figure 3). While TP = true positives, TN = true negatives, FP = false positives and FN = false negatives, the equations of importance are displayed below.

Training/Testing accuracies = $\frac{TP+TN}{TP+TN+FP+FN}$

Precision = $\frac{TP}{TP+FP}$

Recall / True Positive Rate = $\frac{TP}{TP+FN}$

F1 = $\mathbf{2}.\left(\frac{Precision*Recall}{Precision+Recall}\right)$

False Positive Rate = $\frac{FP}{FP+TN}$

Systematic check in bio-fluid datasets

Processed samples were downloaded from GSE167630, GSE108395 and GSE108396 (Dong & Scherzer Clemens 2019; Ferraldeschi et al. 2021). Each were adult-onset HD vs adult control datasets which were either created using human miRNA specific microarrays or microRNA-seq. GSE167630 measured miRNA expression in blood samples, GSE108395 and GSE108396 were part of the same study (GSE108398) and respectively measured miRNAs in blood plasma and CSF. Datasets were checked if they were normalised and if there were visual outliers using MDS plots, and then analysed using standard methodology with either *Limma* or *DESeq2* (Love et al. 2014; Ritchie et al. 2015). Using MirBase and TargetScans we found and 53 of the 80 mouse miRNAs used for training the best performing predisposition model had known human homologues (Kozomara, Birgaoanu & Griffiths-Jones 2019; McGeary et al. 2019). The homologues were mined from the DE results of the three bio-fluid datasets.

Results

Initial data exploration found high homogeneity between WT and JOHD samples

Using DE on age and gender matched miRNA samples found between 0-6 DEGs from the 2m samples, between 0-17 DEGs from the 10m samples and surprisingly at least 230 DEGs from the 6m, including when contrasting the positive control 20 CAG samples with the WT samples. A similar trend was seen with the mRNAs as we found between 0-34 DEGs for the 2m samples, between 0-814 DEGs for the 10m samples and over 2000 DEGs in many of the 6m samples. This information made us suspect a batch effect in the 6m data, so it was removed from further analysis. The homogeneity of the data and the batch effect can be seen

in PCA plots (Supplementary Figure 1). DE also suggested that there may be some differences in how JOHD develops in different genders, and to further investigate for mechanistic differences between male and female samples we used the *TimiRGeN R* package. DEGs from: male 140 CAG 10m /male WT 10m samples were enriched in Triacyglyceride Synthesis, Sphingolipid Metabolism (integrated pathway) and Sphingolipid Metabolism (general overview), female 175 CAG 2m / female WT 2m samples were enriched in IL-9 Signaling Pathway and female 175 CAG 10m / female WT 10m were enriched in Myometrial Relaxation and Contraction Pathways, Calcium Regulation in the Cardiac Cell, Omega-9 FA synthesis, Glycosis and Gluconeogenesis and Cholestoral metabolism. This informed us that it would be best to reduce variance between the genders. The major conclusion was that this data had very low numbers of DEGs (after removing the suspected batch effect) which indicated high homogeneity between WT and JOHD samples, and so we devised an alternative approach to identify miRNA biomarkers from this data set. We opted to re-label all JOHD samples (> 60 CAG) as "JOHD" and all WT and positive control samples (20 CAG) as "WT" and trained ML models to decipher between JOHD and WT samples.

Machine Learning approach identified miRNAs and mRNAs which can classify WT and JOHD samples

The goal of this study was to identify potential miRNA and mRNA biomarkers in JOHD samples, and the initial data exploration helped set realistic queries. From our initial analysis we also found it important to separate the miRNAs and mRNAs, as the DE results showed there was greater variance seen in the mRNAs. The input data consisted of a young miRNAs (519 genes, 30 JOHD samples, 16 WT samples), aged miRNAs (519 genes, 29 JOHD samples, 16 WT samples), young mRNAs (532 genes, 30 JOHD, 16 WT samples) and aged mRNAs (532 gene, 29 JOHD samples, 16 WT samples).

When finding markers for aged or young samples, a 0.8:0.2 split was used, and when looking for markers which indicate predisposition, the aged samples were trained on, and the young samples were tested on. We initially used RFE, with a 5-fold cross-validation, to identify an optimal minimal number of features, however high numbers of features were identified during aged miRNA training (274), predisposed miRNA training (127), aged mRNA training (140) and young mRNA training (185). RFE selected features were trained using multiple classifiers and the training scores of the miRNAs and mRNAs were usually 1 and testing scores for the miRNAs and mRNAs were respectively between 0.5-0.9 and 0.69-1, indicating a tendency for overfitting (Figure 4A). Instead, we opted to use a robust k-best based feature selection strategy with a smaller number of classifiers (ADA, ET, GNB, RF). These classifiers were selected because they were the best performing classifiers from the RFE based miRNA models. Other classifiers such as Linear SVM and Naïve Bayes also performed well but were not taken forward because they were the worst performing classifiers for the miRNA predisposition question. The number of features found to be responsible for the highest three training scores for each of the four classifiers were taken forward for hyperparameter tuning (Supplementary Table 1). Some classifiers (mainly ET) found a range of best features, and in contrast to RFE based models, the k-best based models test accuracies were higher and used fewer features for training. Training accuracy, testing accuracy, recall score, precision score, f-score and confusion matrixes were recorded for each model (Figure 4B, Supplementary Table 2).

Best performing models for each question were selected on two criteria – high testing accuracies, and high precision scores (Table 1). Including good precision scores meant models had to be able to be able to discriminate between WT and JOHD cases. In cases where multiple classifiers led to the same scores, the classifier which used the least number of

features was selected. We also avoided using models built with ET due to high variability of scores seen during the training steps. With reference to the six questions we identified, the best features and classifiers were: a) the 11 miRNAs used to train aged miRNA samples with RF which tested at 85%, b) the 6 miRNAs used to train young miRNA samples with RF which tested at 66%, c) the 80 miRNAs used to train predisposed samples with RF which tested at 84%, d) the 5 mRNAs used to train aged mRNA samples with ADA which tested at 100%, e) the 87 mRNAs used to train young mRNA samples with ADA which tested at 90% and f) the 4 mRNAs used to train predisposed samples with GNB which tested at 83%.

Investigating model performances

AUC plots and PCA plots for the best performing models were created (Figure 5). Unsurprisingly, the mRNA-based models all outperformed their miRNA-based counterparts; most likely because mRNAs showed greater variance than miRNAs which was seen during DE analysis. The 2m and predisposition miRNA-based models were particularly difficult to train with AUC scores of 31% and 59% respectively, though the 10m miRNA-based model had an excellent AUC of 96%. In contrast the mRNA-based models performed well with AUC scores for the mRNA-based 10m, 2m and predisposition models to respectively be 100%, 86% and 92%. The PCA plots for the miRNA-based 2m and predisposition models showed that the model performance was poor when trying to correctly classify true WT samples, meanwhile the PCAs showed the aged miRNA model, and all mRNA-based models were better at classifying WT samples correctly.

Ratification using bio-fluid based transcriptomic datasets for the identified miRNAs

We downloaded and performed DE on three publicly available miRNA-based HD patient bio-fluid datasets and identified if homologues of the 80 miRNAs used to train the miRNA-based predisposition model were found to be DEGs in the datasets. Six miRNAs were identified to have p.values <0.05 from the blood based datasets. From GSE108395, *hsa-miR-382-5p* had a p.value of 0.0017 and a log2fc value of 0.56, *hsa-miR-770-5p* had a p.value of 0.012 and a log2fc of -0.48, *hsa-miR-154-5p* has a p.value of 0.02 and a log2fc value of 0.43, *hsa-miR-181a-5p* had a p.value of 0.034 and a log2fc value of 0.28 and, *hsa-miR-212-3p* had a p.value of 0.035 and a log2fc value of 0.38. Also, from GSE167630, *hsa-miR-378b* had a p.value of 0.097 and a log2fc value of -0.19. The cortex-based miRNAs needed to be checked if they could be detected while circulating in the bio-fluids to promote them as suitable candidates for use as non-invasive biomarkers for HD/ JOHD research.

Discussion and Conclusion

We analysed a publicly available dataset to train ML models with the aim of identifying potential biomarkers for predisposition to JOHD. Six questions were asked to achieve this aim: a) best features to train aged miRNA samples, b) best features to train young miRNA samples, c) best features to train predisposed miRNA samples, d) best features to train aged mRNA samples, e) best features to train young mRNA samples and f) best features to train predisposed mRNA samples. Initially we used a RFE method to remove poor training features, however this led to overfitting. Instead, our robust k-best feature selection method led to the creation several models which could differentiate JOHD and WT samples. For questions a, d, e, and f we generated models with prediction accuracies >80%, and even reached accuracies >90% for questions a, d and e. Models created for question b and c had decent test accuracies of 66% and 67% but poor AUC scores. All our selected mRNA-based models had AUC scores of at least 86%, and our aged miRNA-based model had an AUC of 96%. We saw that our selected miRNA-based models performed worse - though this was expected as our DE analysis showed the miRNAs vary less than mRNAs.

The predisposition models (questions c and f) were particularly ambitious as the training (10m) and testing (2m) datasets were almost the same size (respectively 55 and 56 samples). However, this was an important question as we know that clinically detectable alterations in the brain and behaviour of HD patients can be seen up to 15 years prior to any the classic HD related motor-loss (Djoussé et al. 2003). Indicating early sign of the neurodegenerative disease should be detectable at the gene expression level. For question f we created an excellent model using four mRNAs. *Enpp6* has been reported as a gene of interest in the hippocampus of HD mice and *Gng11* was associated with 4 of 10 Parkinson's disease related pathways in a GWAS study, which indicates both genes to be of interest to neurodegenerative research (Zhang et al. 2017; Skotte et al. 2018). *Gm5067* and *Gm6089* are currently predicted protein-coding genes which may have novel roles during HD progression, and both were also identified as features for the best performing model in question e, and the former was identified in question d. Though the model we presented for question c performed poorer, there is substantial interest in miRNAs circulating in bio-fluids as they could be non-invasive tools to measure the health of patients. Thus, DE results from three publicly available bio-fluids datasets were searched to find homologues of the 80 miRNAs selected to train the best performing miRNA-based predisposition model. We found *hsa-miR-154-5p*, *hsa-miR-181a-5p* and *hsa-miR-212-3p, hsa-miR-378b, hsa-miR-382-5p* and hsa-*miR-770-5p* to be DEGs in blood plasma with a p.value of <0.05. Interestingly, *mmu-miR-212-3p* was cited as a potential miRNA of interest in HD mice and hsa-*miR-212* has been linked to neuronal plasticity and cognition (2016; Fukuoka et al. 2018). Also, *mmu-miR-378b* was one of the miRNAs highlighted in the data generation study and *hsa-miR-382-5p* has been identified as a miRNA of interest in a Parkinson's Disease study and an Alzheimer's disease study, which may indicate it to be a useful miRNA in neurodegenerative disease research (Lau et al. 2013; Nair & Ge 2016; Langfelder et al. 2018). Further functional investigation of these miRNAs could be helpful for JOHD/ HD research and could supplement known HD related miRNAs of interest such as miR-9 and miR-9* which are involved in the regulation of neuron survival antagonist REST (Packer et al. 2008).

It is important to note that age may not have been the only variable of interest, as large systematic reviews of clinical studies of JOHD patients such as Predict-HD, Enroll-HD and kids-JOHD respectively found age to account for 26%, 59% and 86% of disease development (Squitieri et al. 2006; Cronin et al. 2019; Schultz et al. 2020). Thus, age is certainly an important factor in JOHD, but not the only one. Another important critique is to ask is how confident can we be that our mouse-based study is suitable for identifying biomarkers for the benefit of HD/ JOHD patient research. Mouse-Human HD progression does not correlate well, as genotypes which lead to severe neurodegenerative phenotypes via HD related neurotoxicity in humans may resolve as mildly affected/ WT mice, and *Donaldson et al (2021)* reviewed many cases where brain scans and behaviour of mice with human HD/ JOHD genotypes showed no clinical symptoms (Donaldson et al. 2021). However, it should be noted that many of the poly-glutamine repeat mice in the literature have a mix of CAG and CAA repeats, and CAA repeats have been linked to weaker onset of HD (Lee et al. 2019). The data we used was reported to be from mice that had undergone CAG knock-in experiments (Langfelder et al. 2016; Langfelder et al. 2018). Given this, when creating our ML models, we assumed the 2m mice to not have developed JOHD yet and the 10m mice to have developed JOHD symptoms. Overall, though age was predominantly reported as the major factor in HD research, further research is needed to understand the complexities of HD and JOHD (Gusella & MacDonald 2000; Djoussé et al. 2003).

To conclude, we created several good models for mRNA-based samples, including an aged model with 100% accuracy using five mRNAs and an 85% accuracy predisposed model

using only four mRNAs - two of which were novel genes. We also highlighted six potential miRNAs which could be markers of predisposition to developing JOHD/ HD. Importantly, these miRNAs were detected in circulating blood samples of HD patients. We believe this analysis study serves as a useful means of hypothesis generation to aid JOHD/ HD researchers and clinicians. The potential of detecting early predisposition biomarkers non-invasively could be of great benefit to neurodegenerative disease research and patient care.

Conflicts of Interest

None to report.

Author Contributions

KP conceptualised the project, performed all bioinformatics and data preparation. KP, CS generated the machine learning models. KP, DPS contributed equally to writing. KP, CS, DPS approved the submitted version.

Acknowledgements

We would like to thank a talented UG student Bethany Harley[1] for her work on gender differences between JOHD samples. We also thank Professor David Young[1] for supervisory support and Dr Jamie Soul[1], Dr Louise Peas[1] for technical advice during the project.

Funding

KP, CS, DPS were supported by Novo Nordisk Fonden Challenge Programme: Harnessing the Power of Big Data to Address the Societal Challenge of Aging [NNF17OC0027812].

Data availability

Preliminary and teaching scripts, plus some exploratory work which was used as a basis for much of the work performed here has been stored in Colleensdan/Predicting-Early-Onset-Huntington-s-Disease. Scripts (DE, *TimiRGeN*, plotting, ML and bio-fluid analysis) and associated data shown in this manuscript are available in Krutik6/Using-Machine-Learning-to-identify-microRNAs-as-biomarkers-for-pre-disposition-to-Juvenile-Onset-Hun.

References

Andrews, S. & others, 2010, *FastQC: a quality control tool for high throughput sequence data. 2010*, *Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*.

Bartel, D.P., 2009, *MicroRNAs: Target Recognition and Regulatory Functions*, *Cell*, 136(2).

Casella, C., Lipp, I., Rosser, A., Jones, D.K. & Metzler-Baddeley, C., 2020, *A Critical Review of White Matter Changes in Huntington's Disease*, *Movement Disorders*, 35(8).

Caviston, J.P., Ross, J.L., Antony, S.M., Tokito, M. & Holzbaur, E.L.F., 2007, "Huntingtin facilitates dynein/dynactin-mediated vesicle transport," *Proceedings of the National Academy of Sciences of the United States of America*, 104(24).

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K. & Shiekhattar, R., 2005, "TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing," *Nature*, 436(7051).

Cornett, J., Cao, F., Wang, C.E., Ross, C.A., Bates, G.P., Li, S.H. & Li, X.J., 2005, "Polyglutamine expansion of huntingtin impairs its nuclear export," *Nature Genetics*, 37(2).

Cronin, T., Rosser, A. & Massey, T., 2019, *Clinical Presentation and Features of Juvenile-Onset Huntington's Disease: A Systematic Review*, *Journal of Huntington's Disease*, 8(2).

Djoussé, L., Knowlton, B., Hayden, M., Almqvist, E.W., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., Morrison, P.J., Novelletto, A., Frontali, M., Trent, R.J.A., McCusker, E., Gómez-Tortosa, E., Mayo, D., Jones, R., Zanko, A., Nance, M., Abramson, R., Suchowersky, O., Paulsen, J., Harrison, M., Yang, Q., Cupples, L.A., Gusella, J.F., MacDonald, M.E. & Myers,

R.H., 2003, "Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease," *American Journal of Medical Genetics*, 119 A(3).

Doench, J.G. & Sharp, P.A., 2004, "Specificity of microRNA target selection in translational repression," *Genes and Development*, 18(5).

Donaldson, J., Powell, S., Rickards, N., Holmans, P. & Jones, L., 2021, *What is the Pathogenic CAG Expansion Length in Huntington's Disease*, *Journal of Huntington's Disease*, 10(1).

Dong, X. & Scherzer Clemens, 2019, "Differential expression analysis of miRNAs expressed in blood and CSF of Huntington's disease. (Unpublished Processed Data). Gene Expression Omnibus. Retrieved 07/05/2022, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108398. ."

Evans, S.J.W., Douglas, I., Rawlins, M.D., Wexler, N.S., Tabrizi, S.J. & Smeeth, L., 2013, "Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records," *Journal of Neurology, Neurosurgery and Psychiatry*, 84(10).

Ferraldeschi, M., Romano, S., Giglio, S., Romano, C., Morena, E., Mechelli, R., Annibali, V., Ubaldi, M., Buscarinu, M.C., Umeton, R., Sani, G., Vecchione, A., Salvetti, M. & Ristori, G., 2021, "Circulating hsa-miR-323b-3p in Huntington's Disease: A Pilot Study," *Frontiers in Neurology*, 12.

Friedländer, M.R., MacKowiak, S.D., Li, N., Chen, W. & Rajewsky, N., 2012, "MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades," *Nucleic Acids Research*, 40(1).

Friedman, R.C., Farh, K.K.H., Burge, C.B. & Bartel, D.P., 2009, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome Research*, 19(1).

Fukuoka, M., Takahashi, M., Fujita, H., Chiyo, T., Popiel, H.A., Watanabe, S., Furuya, H., Murata, M., Wada, K., Okada, T., Nagai, Y. & Hohjoh, H., 2018, "Supplemental Treatment for Huntington's Disease with miR-132 that Is Deficient in Huntington's Disease Brain," *Molecular Therapy - Nucleic Acids*, 11.

Fusco, F.R., Chen, Q., Lamoreaux, W.J., Figueredo-Cardenas, G., Jiao, Y., Coffman, J.A., Surmeier, D.J., Honig, M.G., Carlock, L.R. & Reiner, A., 1999, "Cellular localization of huntingtin in striatal and cortical neurons in rats: Lack of correlation with neuronal vulnerability in Huntington's disease," *Journal of Neuroscience*, 19(4).

Georgiou, N., Bradshaw, J.L., Chiu, E., Tudor, A., O'Gorman, L. & Phillips, J.G., 1999, "Differential clinical and motor control function in a pair of monozygotic twins with Huntington's disease," *Movement Disorders*, 14(2).

Gusella, J.F. & MacDonald, M.E., 2000, "Molecular genetics: Unmasking polyglutamine triggers in neurodegenerative disease," *Nature Reviews Neuroscience*, 1(2).

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H. & Kim, V.N., 2004, "The Drosha-DGCR8 complex in primary microRNA processing," *Genes and Development*, 18(24).

Hoffner, G., Kahlem, P. & Djian, P., 2002, "Perinuclear localization of huntingtin as a consequence of its binding to microtubules through an interaction with β-tubulin: Relevance to Huntington's disease," *Journal of Cell Science*, 115(5).

Jonas, S. & Izaurralde, E., 2013, *The role of disordered protein regions in the assembly of decapping complexes and RNP granules*, *Genes and Development*, 27(24).

Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S., 2019, "MiRBase: From microRNA sequences to function," *Nucleic Acids Research*, 47(D1).

Kumar, S., Vijayan, M., Bhatti, J.S. & Reddy, P.H., 2017, "MicroRNAs as Peripheral Biomarkers in Aging and Age-Related Diseases," *Progress in Molecular Biology and Translational Science*, vol. 146.

Langfelder, P., Cantle, J.P., Chatzopoulou, D., Wang, N., Gao, F., Al-Ramahi, I., Lu, X.H., Ramos, E.M., El-Zein, K., Zhao, Y., Deverasetty, S., Tebbe, A., Schaab, C., Lavery, D.J., Howland, D., Kwak, S., Botas, J., Aaronson, J.S., Rosinski, J., Coppola, G., Horvath, S. & Yang, X.W., 2016, "Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice," *Nature Neuroscience*, 19(4).

Langfelder, P., Gao, F., Wang, N., Howland, D., Kwak, S., Vogt, T.F., Aaronson, J.S., Rosinski, J., Coppola, G., Horvath, S. & Yang, X.W., 2018, "MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice," *PLoS ONE*, 13(1).

Langmead, B., 2010, "Aligning short sequencing reads with Bowtie," *Current Protocols in Bioinformatics*, (SUPP.32).

Lau, P., Bossers, K., Janky, R., Salta, E., Frigerio, C.S., Barbash, S., Rothman, R., Sierksma, A.S.R., Thathiah, A., Greenberg, D., Papadopoulou, A.S., Achsel, T., Ayoubi, T., Soreq, H., Verhaagen, J., Swaab, D.F., Aerts, S. & Strooper, B. de, 2013, "Alteration of the microRNA network during the progression of Alzheimer's disease," *EMBO Molecular Medicine*, 5(10).

Lee, J.M., Correia, K., Loupe, J., Kim, K.H., Barker, D., Hong, E.P., Chao, M.J., Long, J.D., Lucente, D., Vonsattel, J.P.G., Pinto, R.M., Abu Elneel, K., Ramos, E.M., Mysore, J.S., Gillis, T., Wheeler, V.C., MacDonald, M.E., Gusella, J.F., McAllister, B., Massey, T., Medway, C., Stone, T.C., Hall, L., Jones, L., Holmans, P., Kwak, S., Ehrhardt, A.G., Sampaio, C., Ciosi, M., Maxwell, A., Chatzi, A., Monckton, D.G., Orth, M., Landwehrmeyer, G.B., Paulsen, J.S., Dorsey, E.R., Shoulson, I. & Myers, R.H., 2019, "CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset," *Cell*, 178(4).

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. & Kim, V.N., 2004, *EMBO J*, *EMBO Journal*, 23(20).

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D., 2012, "The SVA package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, 28(6).

Leinonen, R., Sugawara, H. & Shumway, M., 2011, "The sequence read archive," *Nucleic Acids Research*, 39(SUPPL. 1).

Lemaître, G., Nogueira, F. & Aridas, C.K., 2017, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, 18.

Love, M.I., Huber, W. & Anders, S., 2014, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, 15(12).

Lund, E. & Dahlberg, J.E., 2006, *Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs*, *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 71.

McFarland, K.N., Huizenga, M.N., Darnell, S.B., Sangrey, G.R., Berezovska, O., Cha, J.H.J., Outeiro, T.F. & Sadri-Vakili, G., 2014, "MeCP2: A novel huntingtin interactor," *Human Molecular Genetics*, 23(4).

McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M. & Bartel, D.P., 2019, "The biochemical basis of microRNA targeting efficacy," *Science*, 366(6472).

Nair, V.D. & Ge, Y., 2016, "Alterations of miRNAs reveal a dysregulated molecular regulatory network in Parkinson's disease striatum," *Neuroscience Letters*, 629.

Ohlmeier, C., Saum, K.U., Galetzka, W., Beier, D. & Gothe, H., 2019, "Epidemiology and health care utilization of patients suffering from Huntington's disease in Germany: Real world evidence based on German claims data," *BMC Neurology*, 19(1).

Packer, A.N., Xing, Y., Harper, S.Q., Jones, L. & Davidson, B.L., 2008, "The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease," *Journal of Neuroscience*, 28(53).

Pan, L. & Feigin, A., 2021, *Huntington's Disease: New Frontiers in Therapeutics*, *Current Neurology and Neuroscience Reports*, 21(3).

Panas, M., Karadima, G., Markianos, M., Kalfakis, N. & Vassilopoulos, D., 2008, *Phenotypic discordance in a pair of monozygotic twins with Huntington's disease*, *Clinical Genetics*, 74(3).

Patel, K., Chandrasegaran, S., Clark, I.M., Proctor, C.J., Young, D.A. & Shanley, D.P., 2021, " TimiRGeN : R/Bioconductor package for time series microRNA–mRNA integration and analysis ," *Bioinformatics*.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C., 2017, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, 14(4).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É., 2011, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12.

Quarrell, O., O'Donovan, K.L., Bandmann, O. & Strong, M., 2012, *The prevalence of juvenile Huntington's disease: A review of the literature and meta-analysis*, *PLoS Currents*.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D. & Izaurralde, E., 2005, "A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing," *RNA*, 11(11).

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. & Smyth, G.K., 2015, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, 43(7).

Rosas, H.D., Liu, A.K., Hersch, S., Glessner, M., Ferrante, R.J., Salat, D.H., Kouwe, A. van der, Jenkins, B.G., Dale, A.M. & Fischl, B., 2002, "Regional and progressive thinning of the cortical ribbon in Huntington's disease," *Neurology*, 58(5).

Rosenblatt, A., Kumar, B. v., Mo, A., Welsh, C.S., Margolis, R.L. & Ross, C.A., 2012, "Age, CAG repeat length, and clinical progression in Huntington's disease," *Movement Disorders*, 27(2).

Schultz, J.L., Moser, A.D. & Nopoulos, P.C., 2020, "The association between cag repeat length and age of onset of juvenile-onset huntington's disease," *Brain Sciences*, 10(9).

Semaka, A., Collins, J.A. & Hayden, M.R., 2010, "Unstable familial transmissions of huntington disease alleles with 27-35 CAG repeats (intermediate alleles)," *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 153(1).

Skotte, N.H., Andersen, J. v., Santos, A., Aldana, B.I., Willert, C.W., Nørremølle, A., Waagepetersen, H.S. & Nielsen, M.L., 2018, "Integrative Characterization of the R6/2 Mouse Model of Huntington's Disease Reveals Dysfunctional Astrocyte Metabolism," *Cell Reports*, 23(7).

Soneson, C., Love, M.I. & Robinson, M.D., 2015, "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences," *F1000Research*, 4.

Squitieri, F., Frati, L., Ciarmiello, A., Lastoria, S. & Quarrell, O., 2006, *Juvenile Huntington's disease: Does a dosage-effect pathogenic mechanism differ from the classical adult disease?*, *Mechanisms of Ageing and Development*, vol. 127.

Takano, H. & Gusella, J.F., 2002, "The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor," *BMC Neuroscience*, 3.

Tsutsumi, A., Kawamata, T., Izumi, N., Seitz, H. & Tomari, Y., 2010, "Recognition of the pre-miRNA structure by Drosophila-Dicer-1," *Nature Structural and Molecular Biology*, 18(10).

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J.J. & Lötvall, J.O., 2007, "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells," *Nature Cell Biology*, 9(6).

Yamashita, A., Chang, T.C., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C.Y.A. & Shyu, A. bin, 2005, "Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover," *Nature Structural and Molecular Biology*, 12(12).

Zekri, L., Huntzinger, E., Heimstädt, S. & Izaurralde, E., 2009, "The Silencing Domain of GW182 Interacts with PABPC1 To Promote Translational Repression and Degradation of MicroRNA Targets and Is Required for Target Release," *Molecular and Cellular Biology*, 29(23).

Zhang, M., Mu, H., Shang, Z., Kang, K., Lv, H., Duan, L., Li, J., Chen, X., Teng, Y., Jiang, Y. & Zhang, R., 2017, "Genome-wide pathway-based association analysis identifies risk pathways associated with Parkinson's disease," *Neuroscience*, 340.

Zuccato, C., Tartari, M., Crotti, A., Goffredo, D., Valenza, M., Conti, L., Cataudella, T., Leavitt, B.R., Hayden, M.R., Timmusk, T., Rigamonti, D. & Cattaneo, E., 2003, "Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes," *Nature Genetics*, 35(1).

2016, "The miR-132/212 locus: a complex regulator of neuronal plasticity, gene expression and cognition," *RNA & DISEASE*.

Figures and Tables

Figure 1 | miRNA biogenesis. The Illustrations show canonical mammalian miRNA biogenesis steps. **(A)** Shows the initial transcription of the mature miRNA via RNA polymerase II and subsequent processing by DROSHA and DGCR8, which creates a pre-miRNA stem-loop structure which is exported into the cytoplasm by Exportin5. **(B)** Within the cytoplasm DICER and TRBP cut the loop of the pre-miRNA and the RISC complex is associated with the mature guide miRNA strand. The passenger strand is degraded. **(C)** The newly form miRISC complex will interfere with the eukaryotic initiation machinery via the GW182 protein binding to the PABPC protein. GW182 protein will then recruit deadenylation complexes, and this is followed by a decapping protein removing the CAP protein complex from the target mRNA. The naked mRNA structure is decayed by XRN1 and the miRISC complex may be recycled. Some mammalian miRNA-based mRNA degradation is based on exonuclease activity by AGO2. Image was made with BioRender.

Figure 2 | Simplified diagram of the JOHD ML project presented. The illustration shows the major steps of the project, including how the datasets used in this project were created, the separation of miRNAs and mRNAs, the data processing and ML steps. The image also shows the miRNA and mRNA biomarkers found through the predisposition models. The figure created using BioRender.

Figure 3 | Machine learning approach. This is an illustration our ML approach, including how the data was prepared (yellow arrows) for each of the six questions and how the data was split (green arrows). For the age matched classification questions, a simple 0.8:0.2 training to testing split was used, and for the more ambitious predisposition questions the 10m data was trained on and the 2m data was tested on. Following this, the RFE and k-best based feature selection (blue arrows and blue triangle) were performed. RFE was performed first and from its results (pink boxes) we were able to inform our k-best strategy by using classifiers which performed best on the miRNA-based ML questions (orange arrows, Supplementary Table 1). The three best performing models (orange arrows) from each classifier were hyperparameter tuned (grey arrow), and performance scores were measured (Supplementary Table 2). For each question, the best performing model had AUC-ROC curves and PCA plots made for them. For all model training steps, oversampling of WT samples (red circles), min_max scaling (yellow circles) and a 5-fold cross-validation strategy (dark blue circles) was used. Oversampling and 5-fold cross validation were only performed with training samples to limit leakage, and scaling was performed on training and test data.

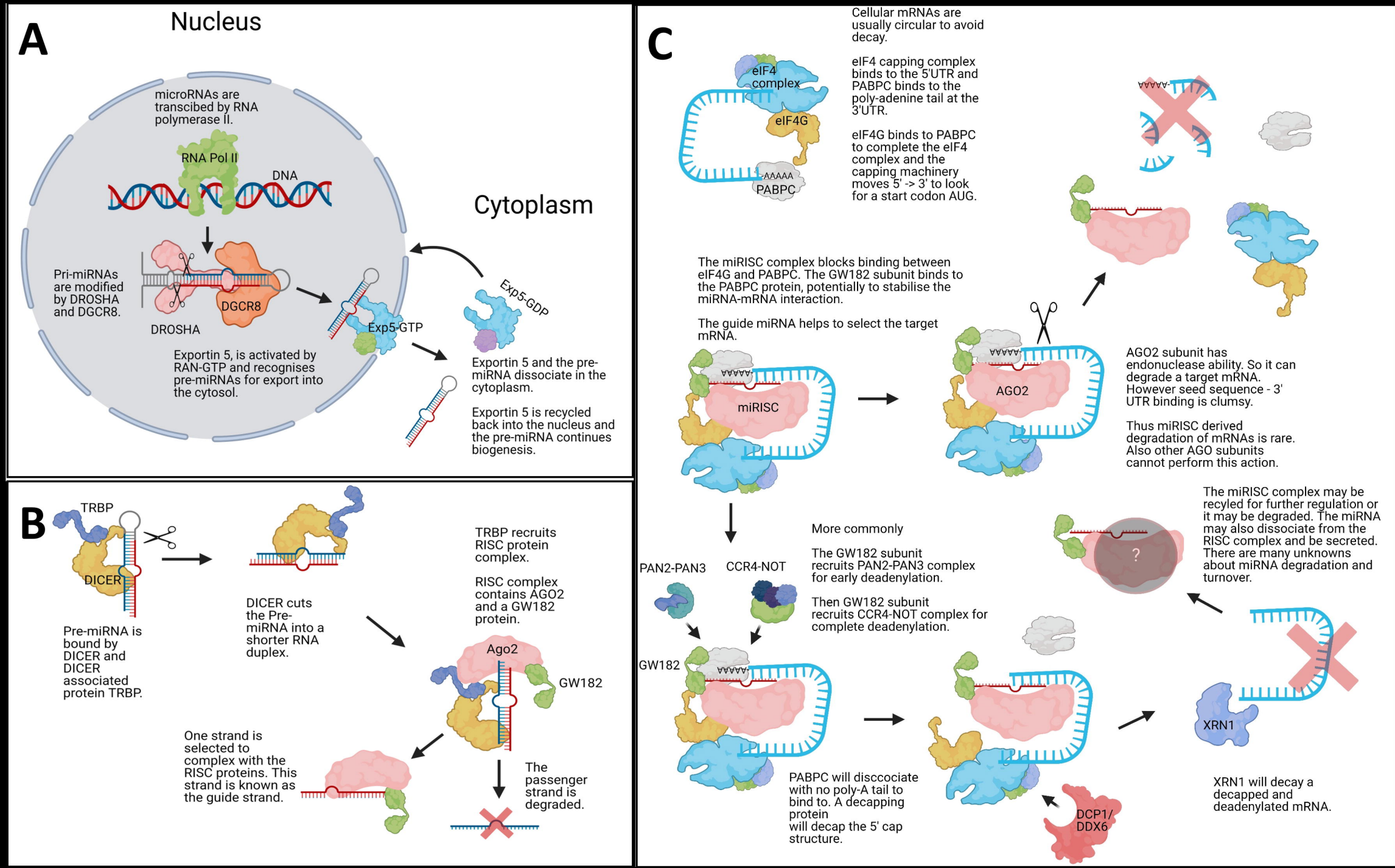
Figure 4 | Feature selection approaches. **(A)** RFE was used to identify the optimal features for each question. Eight classifiers were used. The bar chart shows the results from training accuracies (blue) and testing accuracies (beige). **(B)** From a robust k-best based feature identification method the best number of features were taken forward for hyperparameter tuning. Each classifier had each selected set of features hyperparameter tuned. The Training accuracies (purple) testing accuracies (blue), precision scores (yellow-green), recall scores (green) and f1 scores (orange) have been measured from each model.

Question	Genes/ Features used for training	Performance
a. Which set of miRNAs are the best predictors of JOHD after an assumed onset of the symptoms (10m)?	mmu.let.7g.3p, mmu.miR.135b.5p, mmu.miR.138.1.3p, mmu.miR.212.3p mmu.miR.221.3p, mmu.miR.3069.5p, mmu.miR.330.3p, mmu.miR.370.5p, mmu.miR.433.5p, mmu.miR.543.3p, mmu.miR.872.5p	Features: 11 Train: 0.85 Test: 0.9 Precision: 1 Recall: 0.87 F1: 0.93 CM: [[7 1] [0 3]]
b. Which set of miRNAs are the best predictors of JOHD prior to an assumed onset of the symptoms (2m)?	mmu.miR.151.3p, mmu.miR.191.5p, mmu.miR.218.1.3p, mmu.miR.376b.3p, mmu.miR.670.3p, mmu.miR.8114	Features: 6 Train: 0.85 Test: 0.66 Precision: 0.72 Recall: 0.88 F1: 0.8 CM: [[8 1] [3 0]]
c. Using miRNAs, can we use the aged samples (10m) to predict predisposition of JOHD in young samples (2m)?	mmu.let.7a.2.3p, mmu.let.7f.5p, mmu.let.7g.3p, mmu.let.7i.3p, mmu.let.7i.5p, mmu.miR.1198.5p, mmu.miR.124.3p, mmu.miR.1247.5p, mmu.miR.125b.1.3p, mmu.miR.126a.5p, mmu.miR.130b.5p, mmu.miR.132.3p, mmu.miR.134.5p, mmu.miR.135b.5p, mmu.miR.136.5p, mmu.miR.140.5p, mmu.miR.153.5p, mmu.miR.154.5p, mmu.miR.15a.5p, mmu.miR.181a.1.3p, mmu.miR.181a.2.3p, mmu.miR.181a.5p, mmu.miR.181b.1.3p, mmu.miR.181b.5p, mmu.miR.181c.5p, mmu.miR.1843b.3p, mmu.miR.212.3p, mmu.miR.212.5p, mmu.miR.218.5p, mmu.miR.221.3p, mmu.miR.221.5p, mmu.miR.223.3p, mmu.miR.24.1.5p, mmu.miR.300.3p, mmu.miR.301a.5p, mmu.miR.3061.3p, mmu.miR.3087.3p, mmu.miR.30d.3p, mmu.miR.3102.3p, mmu.miR.320.3p, mmu.miR.33.3p, mmu.miR.330.3p, mmu.miR.346.5p, mmu.miR.3475.3p, mmu.miR.3535, mmu.miR.370.5p, mmu.miR.378b, mmu.miR.382.5p, mmu.miR.409.5p, mmu.miR.410.3p, mmu.miR.412.5p, mmu.miR.431.5p, mmu.miR.433.5p, mmu.miR.467e.5p, mmu.miR.484, mmu.miR.488.3p, mmu.miR.488.5p, mmu.miR.496a.3p, mmu.miR.497a.5p, mmu.miR.543.3p, mmu.miR.6236, mmu.miR.6418.3p, mmu.miR.664.3p, mmu.miR.665.3p, mmu.miR.667.3p, mmu.miR.671.3p, mmu.miR.674.3p, mmu.miR.674.5p, mmu.miR.7047.3p, mmu.miR.708.5p, mmu.miR.770.5p, mmu.miR.8103, mmu.miR.8111, mmu.miR.8114, mmu.miR.872.5p, mmu.miR.92b.3p, mmu.miR.92b.5p, mmu.miR.935, mmu.miR.98.3p, mmu.miR.99b.5p	Features: 80 Train: 0.84 Test: 0.67 Precision: 0.75 Recall: 0.82 F1: 0.78 CM: [[33 7] [11 5]]
d. Which set of mRNAs are the best predictors of JOHD after an assumed onset of the symptoms?	Ccdc116, Gng11, Serpina3n, Enpp6, Gm6089	Features: 5 Train: 0.93 Test: 1 Precision: 1 Recall: 1 F1: 1 CM: [[8 0] [0 3]]
e. Which set of mRNAs are the best predictors of JOHD prior to an assumed onset of the symptoms?	Abca8a, Ada, Allc, Alx3, Amy2a4, Anln, Apol7e, Arhgap36, Arl4d, Armh1, Arrdc2, Atf3, Baiap3, Car5a, Cd209c, Cd28, Chdh, Chrna3, Chrnb4, Cnksr1, Ctla2b, Depp1, Derl3, Dlk1, Dmrtb1, Dsp, Ecel1, Enpp6, ENSMUSG00000098369, ENSMUSG00000102049, ENSMUSG00000115893, Gbx1, Gch1, Gck, Gdf10, Gimap5, Gm11331, Gm11627, Gm11847, Gm17084, Gm21168, Gm5067, Gm6089, Gm7361, Gm8206, Gpr153, Hcn4, Il1r1, Irag2, Kcnh4, Klhl1, Lbhd2, Lilrb4a, Mid1.ps1, Mslnl, Nat8f6, Ngfr, Nphs2, Omp, P4ha3, Pglyrp1, Pkhd1l1, Plxnb3, Poteb, Ptger2, Rpl17.ps5, Rps15a.ps4, Scara5, Serpina3n, Sfrp5, Slc27a3, Slc4a11, Smpdl3b, Spint1, Sspo, Stoml3, Strc, Sync, Tmc3, Tmem40, Trdn, Trh, Txnip, Ube2n.ps1, Vma21.ps, Vmn2r11, Wnt3	Features 87 Train: 0.93 Test: 0.9 Precision: 1 Recall: 0.85 F1: 0.92 CM: [[6 1] [0 4]]
f. Using mRNAs, can we use the aged samples to predict predisposition of JOHD in young samples?	Gm6089, Gng11, Gm5067, Enpp6	Features: 4 Train: 0.93 Test: 0.83 Precision: 0.89 Recall: 0.87 F1: 0.88 CM: [[34 5] [4 12]]

Table 1 | miRNAs/ mRNAs found to be the best features for training models to answer each question. The six questions asked using this dataset are displayed from a-f and the features used for training the best performing models for each question are displayed in alphabetical

order. Model scores are also provided, which state the number of features, training accuracies, testing accuracies, precision, recall, f1 and confusion matrices (CM).

Figure 5 | AUC plots and PCA plots of the best performing models. For each of the questions of interest the best performing model was selected for further analysis. The ROC-AUC curves show the performance of the models and PCA plots show if the samples were correctly (triangle) or incorrectly (circle) predicted to be JOHD (red) or WT (blue).


A
Nucleus
microRNAs are transcibed by RNA polymerase II.
RNA Pol II
DNA
Pri-miRNAs are modified by DROSHA and DGCR8.
DROSHA
DGCR8
Exportin 5, is activated by RAN-GTP and recognises pre-miRNAs for export into the cytosol.
Exp5-GTP
Exp5-GDP
Cytoplasm
Exportin 5 and the pre-miRNA dissociate in the cytoplasm.
Exportin 5 is recycled back into the nucleus and the pre-miRNA continues biogenesis.
B
TRBP
DICER
Pre-miRNA is bound by DICER and DICER associated protein TRBP.
DICER cuts the Pre-miRNA into a shorter RNA duplex.
TRBP recruits RISC protein complex.
RISC complex contains AGO2 and a GW182 protein.
Ago2
GW182
One strand is selected to complex with the RISC proteins. This strand is known as the guide strand.
The passenger strand is degraded.
C
Cellular mRNAs are usually circular to avoid decay.
eIF4 complex
eIF4G
AAAAA
PABPC
eIF4 capping complex binds to the 5'UTR and PABPC binds to the poly-adenine tail at the 3'UTR.
eIF4G binds to PABPC to complete the eIF4 complex and the capping machinery moves 5' -> 3' to look for a start codon AUG.
The miRISC complex blocks binding between eIF4G and PABPC. The GW182 subunit binds to the PABPC protein, potentially to stabilise the miRNA-mRNA interaction.
The guide miRNA helps to select the target mRNA.
miRISC
AGO2
AGO2 subunit has endonuclease ability. So it can degrade a target mRNA. However seed sequence - 3' UTR binding is clumsy.
Thus miRISC derived degradation of mRNAs is rare. Also other AGO subunits cannot perform this action.
PAN2-PAN3
CCR4-NOT
More commonly
The GW182 subunit recruits PAN2-PAN3 complex for early deadenylation.
Then GW182 subunit recruits CCR4-NOT complex for complete deadenylation.
GW182
PABPC will disccociate with no poly-A tail to bind to. A decapping protein will decap the 5' cap structure.
DCP1/ DDX6
XRN1
XRN1 will decay a decapped and deadenylated mRNA.
?
The miRISC complex may be recyled for further regulation or it may be degraded. The miRNA may also dissociate from the RISC complex and be secreted. There are many unknowns about miRNA degradation and turnover.

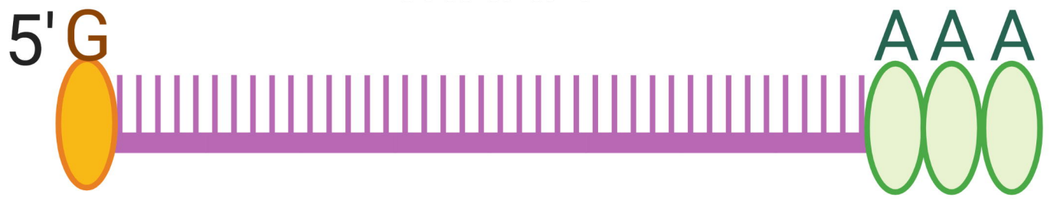
Young Mice

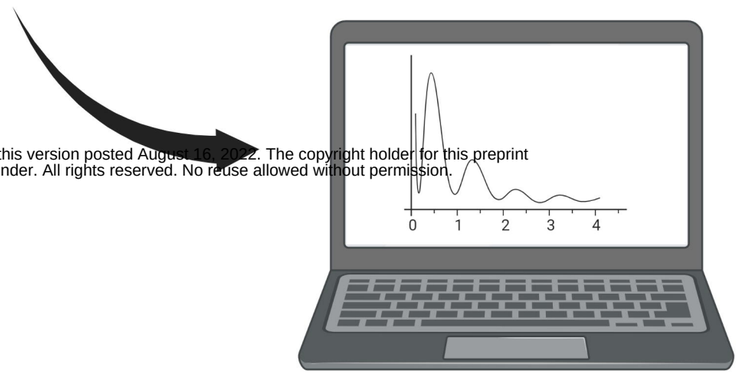
Aged Mice

Cortex

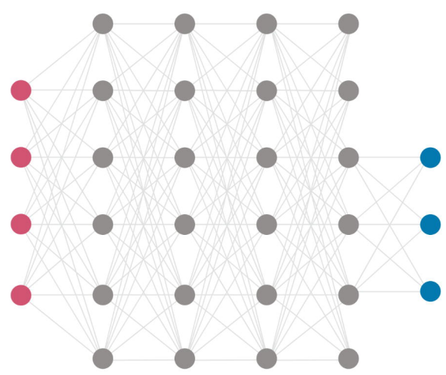
microRNA

mRNA



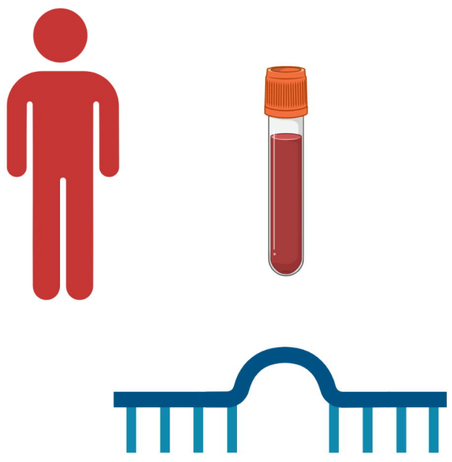


Data Preperation



Machine Learning Models

Patient microRNA bio-fluids



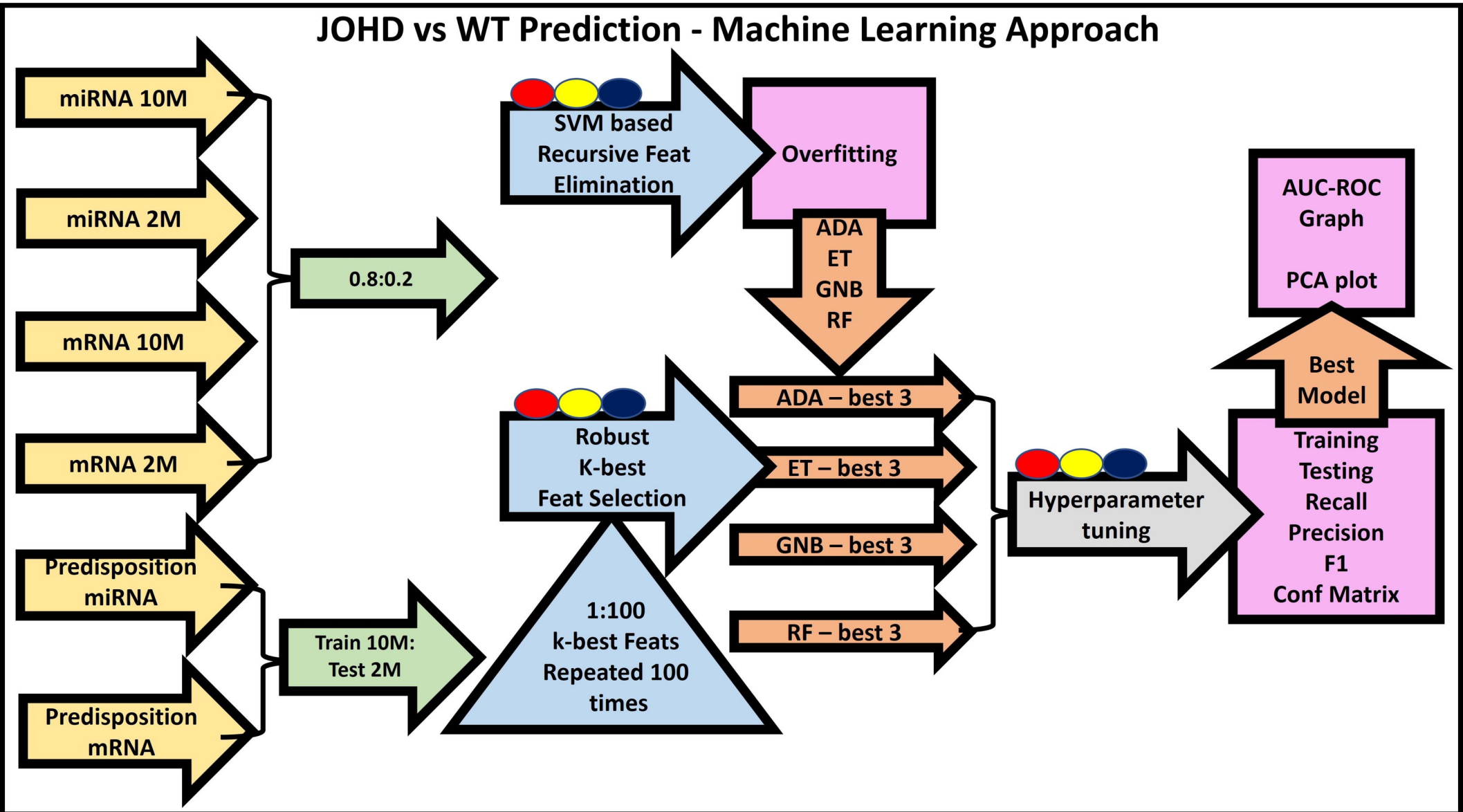
hsa-miR-154-5p
hsa-miR-181a-5p
hsa-miR-212-3p
has-miR-378b
hsa-miR-382-5p
has-miR-770-5p

Gm6089
Gng11
Gm5067
Enpp6

Potential biomarkers!

JOHD?

WT?

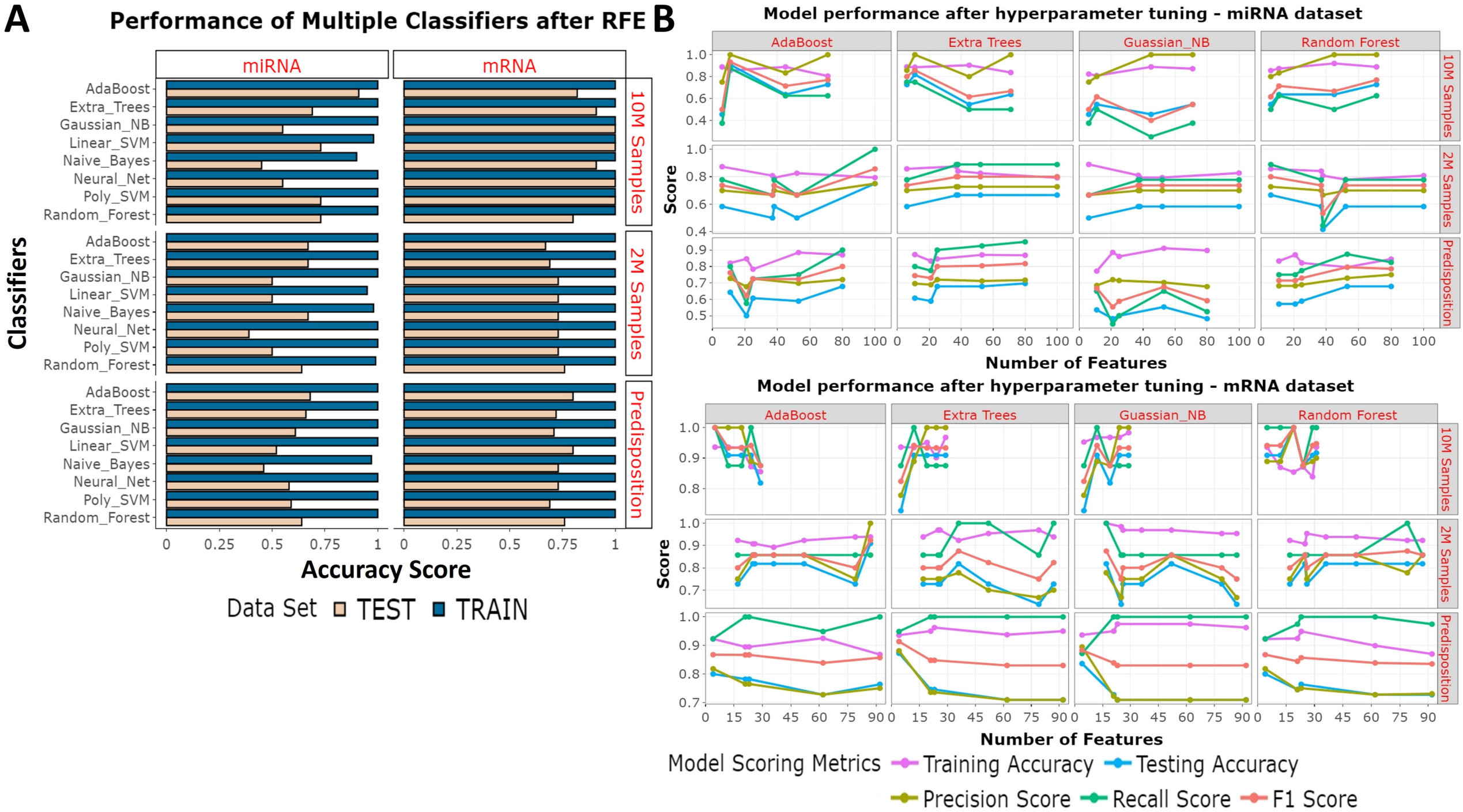


Data Prep	Train:Test Split	Feature Selection	Model Selection	Hyper Tuning	Result

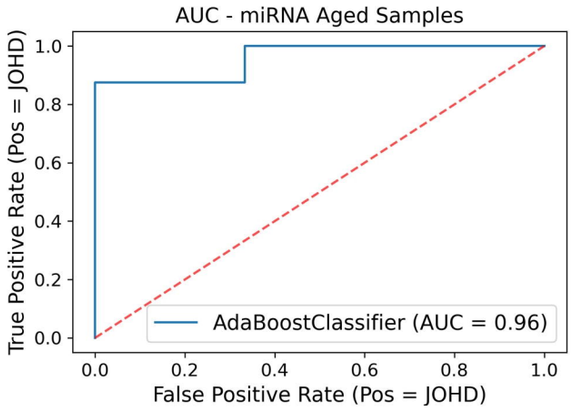
= Oversample WT samples

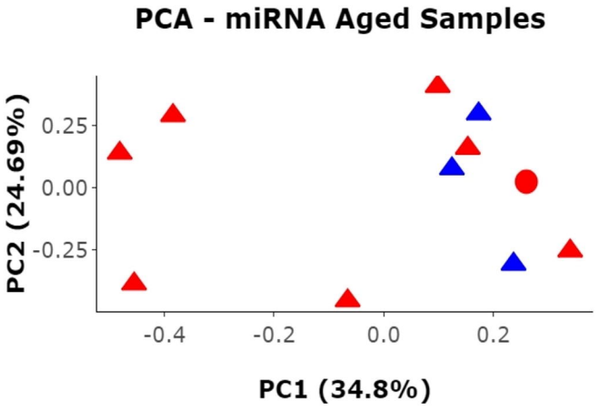
= Min_max scaling

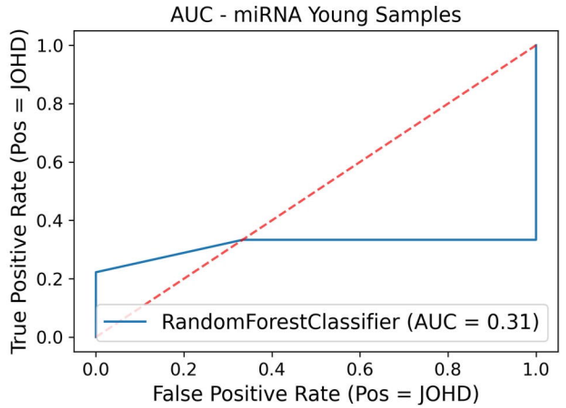
= 5-fold Cross-Validation

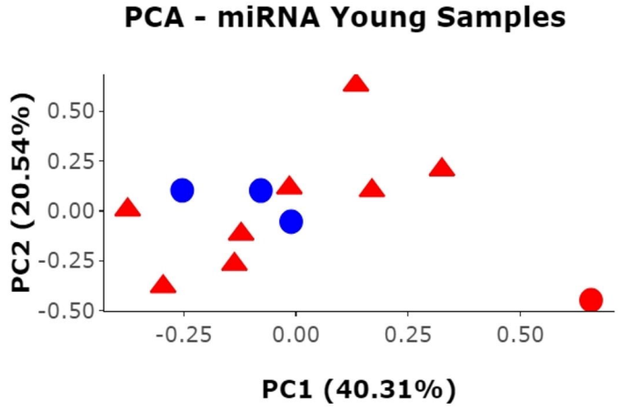

A
Performance of Multiple Classifiers after RFE
miRNA
mRNA
AdaBoost
Extra_Trees
Gaussian_NB
Linear_SVM
Naive_Bayes
Neural_Net
Poly_SVM
Random_Forest
10M Samples
2M Samples
Predisposition
Classifiers
0
0.25
0.5
0.75
1
Accuracy Score
Data Set
TEST
TRAIN
B
Model performance after hyperparameter tuning - miRNA dataset
AdaBoost
Extra Trees
Guassian_NB
Random Forest
Score
Number of Features
Model performance after hyperparameter tuning - mRNA dataset
Model Scoring Metrics
Training Accuracy
Testing Accuracy
Precision Score
Recall Score
F1 Score

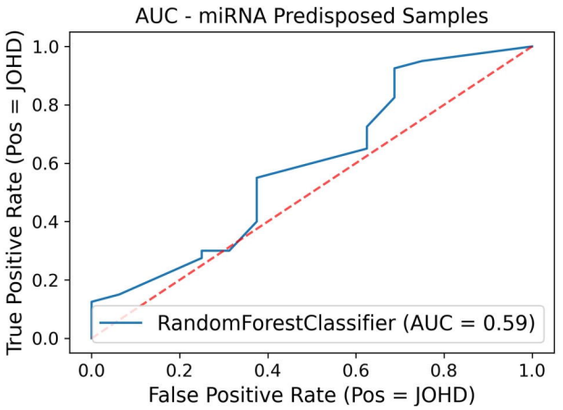
miRNA

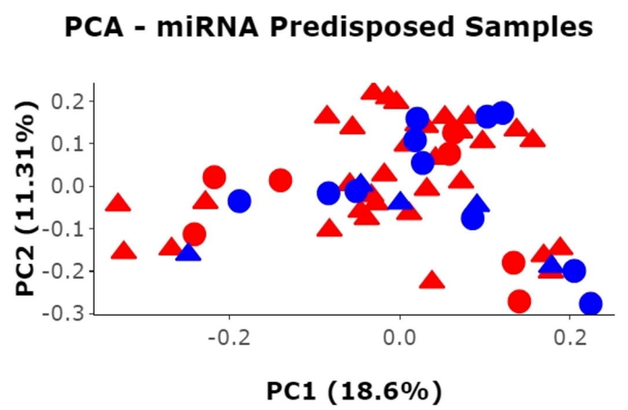




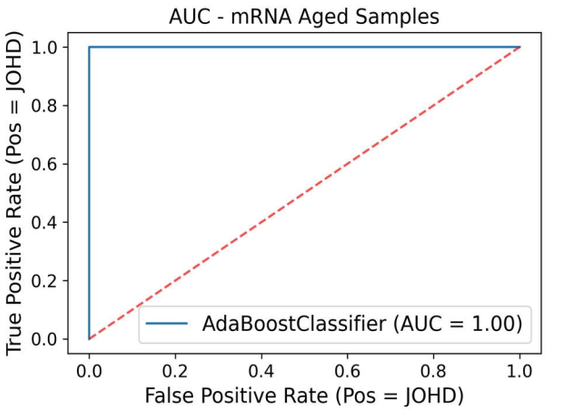


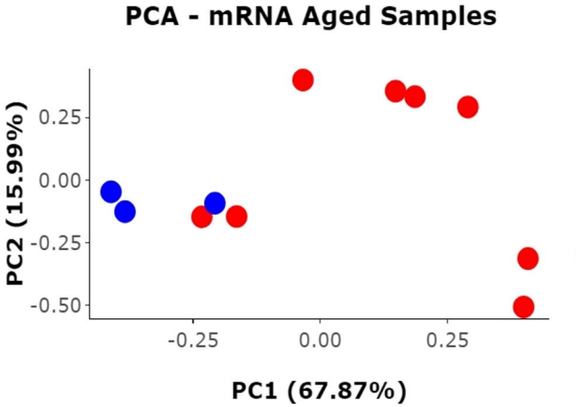


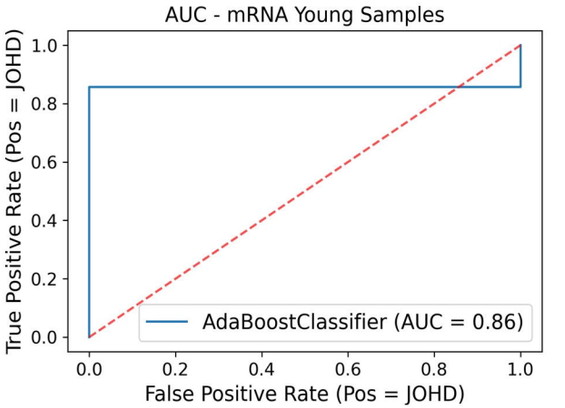


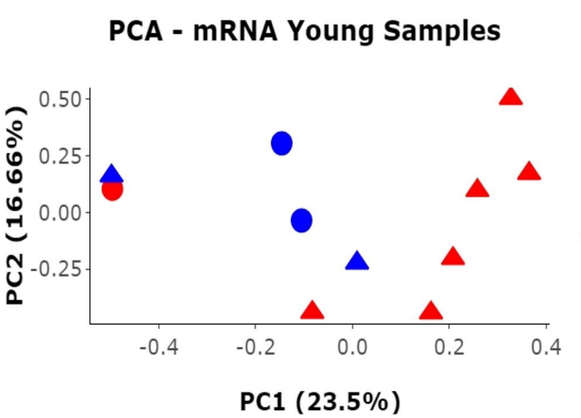


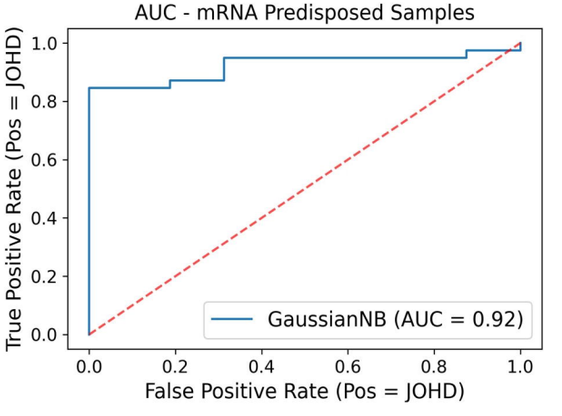
mRNA

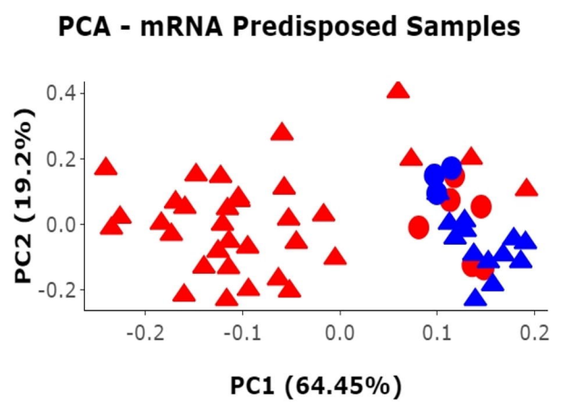












10m Samples

2m Samples

Predisposition

(Sample label, Was the sample predicted correctly?) (JOHD,FALSE) (JOHD,TRUE) (WT,FALSE) (WT,TRUE)