Detecting miRNA Biomarkers For Alzheimer's Disease Using Next Generation Sequencing Data

- Final Report -

Imalsha Dinuwanthi
Hasini Thilakarathna
Vidwa Sripadi

Department of Computer Engineering
University of Peradeniya

Final Year Project (courses CO421 & CO425) report submitted as a requirement of the degree of
B.Sc.Eng. in Computer Engineering

November 2020

Supervisors: Dr. Damayanthi Herath (University of Peradeniya) and Prof. Roshan Ragel (University of Peradeniya)

I would like to dedicate this thesis to my loving parents and “teachers” ...

Declaration

I/We hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my/our own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.

Imalsha Dinuwanthi
Hasini Thilakarathna
Vidwa Sripadi
November 2020

Acknowledgements

First of all we would like to acknowledge our supervisors Dr. Damayanthi Herath and Prof. Roshan Ragel for all the support given for us from the begin until the end of this project. You guided us in the correct path whenever we discussed the ideas in weekly meetings. So we are grateful for your guidance and motivation without which we would not have been come this far in the completion of this project.

We would also like to show our gratitude to all the lecturers in the department of computer engineering, who have been supporting us whenever we need to clarify anything regarding theoretical knowledge which indeed was a great help in the completion of this project.

Finally, we would like to acknowledge Prof. Sen-Lin Tang of Biodiversity Center in Taiwan, who directed us to the right path in the beginning of this project.

Abstract

Alzheimer's disease (AD) has been identified as one of the most common diseases found in people aged 65 or above by different researches done in past few years, which is still well known in medical field as a disease which has no efficient cure. We introduce a method for identifying AD patients using microRNAs (miRNAs). Different technologies such as microarray technology, Sanger sequencing and next generation sequencing been used by various researchers for gathering samples in previous researches. In this project, we used samples gathered from next generation sequencing technology. Out of many different sequencing techniques available in NGS like, Illumina sequencing, Roche sequencing and SoLiD sequencing, we considered a dataset which was collected with the Illumina sequencing technique. To find thw required samples we used the NCBI database. We used Galaxy tool for preprocessing the raw NGS data. Statistical methods like, Wilixcon Man Whitney test and Receiver operating characteristics area under curve (ROC AUC) values were used for obtaining the most significant miRNAs. We used machine learning approaches for the classification purposes.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AD	Alzheimer's Disease
AUC	Area Under Curve
HC	Healthy Controls
NCBI	National Centre for Biotechnology Information
NGS	Mild Cognitive Impairment
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
ROC	Receiver Operating Characterstics
SVM	Support Vector Machine
WMW	Wilixcon Man Whitney

Chapter 1

Introduction

1.1 What is Next-generation sequencing?

Next generation sequencing is a high throughput option introduced for gathering multiple DNA samples in parallel. Although different techniques have been introduced for next generation sequencing, all those techniques follow some common guidelines as sample preparation, sequencing machines and data output. To understand the next generation sequencing, one needs to get a clear idea on how these guidelines [1] are used in the process of next generation sequencing. For any sample to be sequenced, it is needed to be prepared into a sample library; which is collection of genomes with different sizes, either from genomic DNA or total RNA. When the samples are prepared into a library, the sequencing of the genomes begins providing a sequenced sample at the end of the process.

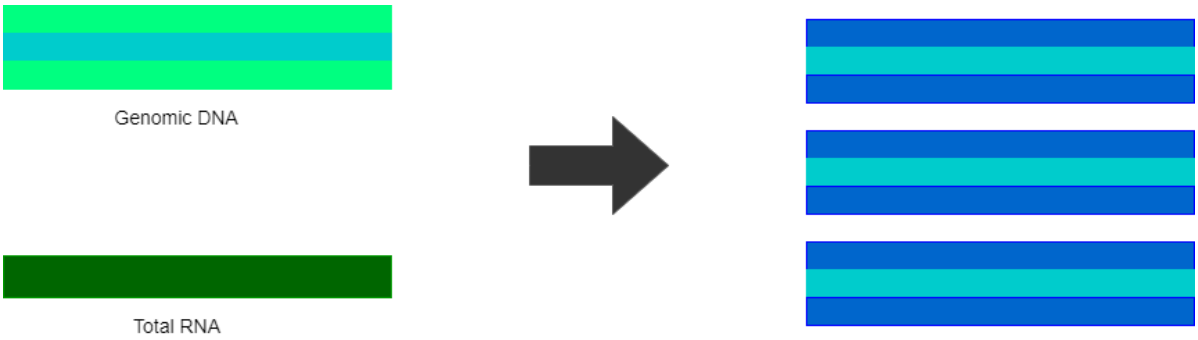

Fig. 1.1 Preparing library by genomic DNA or total RNA

The most popular technique used in diagnosis of human diseases is the Illumina technique, which is also the technique used in our project. In this technique, sequencing

by synthesis method is used for DNA sequencing.The process of sequencing by synthesis method is illustrated in the given Figure 1.2.

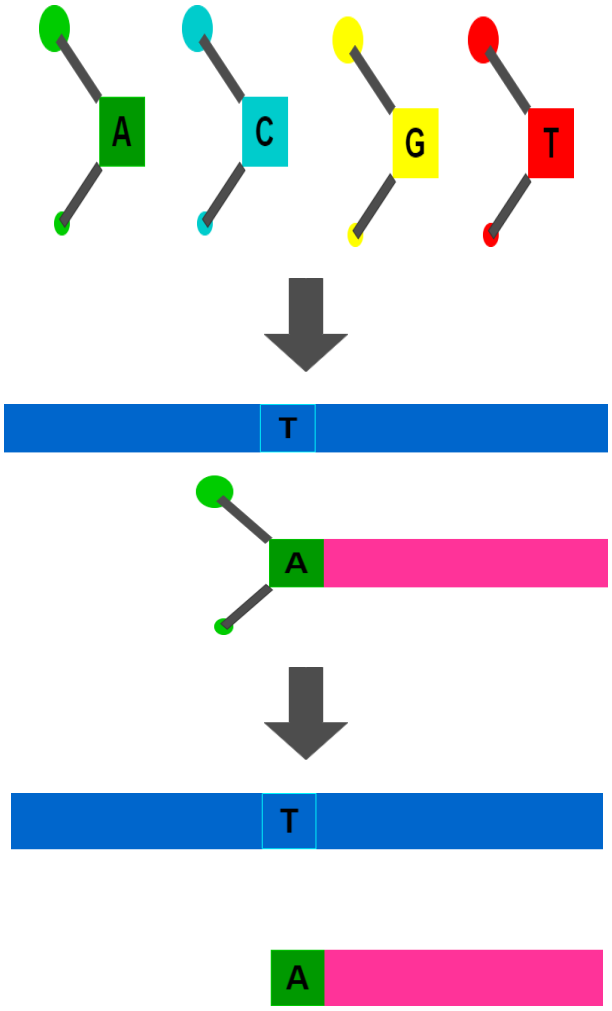

Fig. 1.2 Sequencing by synthesis process

1.2 Why detecting biomarker miRNAs for Alzheimer's disease using NGS?

1.2.1 Background

Alzheimer's disease is one of the diseases which still does not have any accurate cure. Many methods have been developed in order to diagnosis AD in past few years. Different technologies such as microarray technology, Sanger sequencing and next generation sequencing have been used by various researches for gathering samples. Out of theses, next generation sequencing has become more common now a day, as it is a powerful platform which enables to sequence thousands or millions of DNA molecules simultaneously. Samples are used in many forms as gene expressions, MicroRNAs and RNAs. Most of researches were done using gene expressions to find AD biomarkers. Thus, for going in a new direction and to develop a better method compared to the prevailing methods, in this project we used miRNAs samples for detecting biomarkers. miRNAs are small non coding RNAs which are most commonly used in finding many genomic diseases like cancers.

1.2.2 The problem

Alzheimer's disease has been identified as one of the most common diseases found in people aged 65 or above by different researches done in past few years, which is still well known in medical field as a disease which has no efficient cure. It is identified as a deadly disease which initially spreads in human body before 20 years when the symptoms are shown. Therefore, if a method to diagnosis AD at early stage can be found, then there's an opportunity for curing the disease at its early stage which will save many lives. Although many studies have carried out for diagnosis AD, none of them have been succeeded in addressing all the key features needed to be addresses when developing a method for finding biomarkers. The Alzheimer's disease spreads in the brain, which means that most significance features can be extracted from the brain samples. But, most of the researches done so far, were done using only blood sample probably because they are easily available. Only few have included brain samples in their data, which could give better result when finding candidate biomarkers for AD. For obtaining most significance miRNAs, analysis could be carried out using both statistical methods and machine learning approaches, which is a key point lacks in most of the previous studies. If there's a proper diagnosis method, then there won't be increase of AD patients day by

day as estimated by the Alzheimer's association. Therefore, clearly there's an opportunity for finding a more accurate method to identify AD patients.

1.2.3 The proposed solution

We are going to use a set of samples which are collected using next generation sequencing method. Preprocessing of the samples is going to be carried out using galaxy tool by adjusting suitable setting, in order to obtain a dataset with miRNA read counts for each sample. We are going to subject the prepared dataset to statistical analysis to filter out the most significant miRNAs. This can be done using Wilixcon Man Whitney test which is a test which gives p values for each miRNAs. miRNAs with p values less than the significance level (0.05), which are the most significant miRNAs, can be passed into the next level of analysis. For further identifying the most up regulated and the most down regulated miRNAs AUC values can be used. Machine learning approaches can be used for classification and analyzing the miRNAs filtered out by the statistical analysis step. Then we are going to use correlation values and significance values for identifying the most significant set of miRNAs. Following key points explains why our proposed solution could make a significant change in the field of medical when diagnosis AD patients.

- More accurate results since we are using both brain and blood samples
- Both statistical and machine learning approaches are going to be used for finding most significant miRNAs which will provide better results
- Validation will be done for the miRNAs selected from statistical analysis and machine learning approach

1.2.4 Deliverable and milestones

Following are the milestones and deliverables we are presenting for the phase 1 of our project.

- Milestone 01: Doing background study for understanding the topic and then narrowing down the project scope for a practically achievable one. For the selected scope, come up with a methodology for finding the biomarkers. Search for datasets to be used to initiate the proposed methodology. While doing the background search write a review paper with the project related articles.
- Milestone 02: Choosing a suitable tool for preprocessing the selected dataset and use the most effective one to preprocess the dataset. Study the dataset by visualizing

the its features using python. Using a suitable normalization technique normalize the dataset.

- Milestone 03: Analyze different statistical methods for finding p values and find the p values for each miRNA using the most suitable approach. Filter out the most significant features using significance level. Introduce the most dysregulated features by calculating receiver operating characteristics area under curve values. Classification of AD and control samples using machine learning approaches.

Gantt chart is given in the Figure 1.3 for showing the achievement of each task for phase 1.

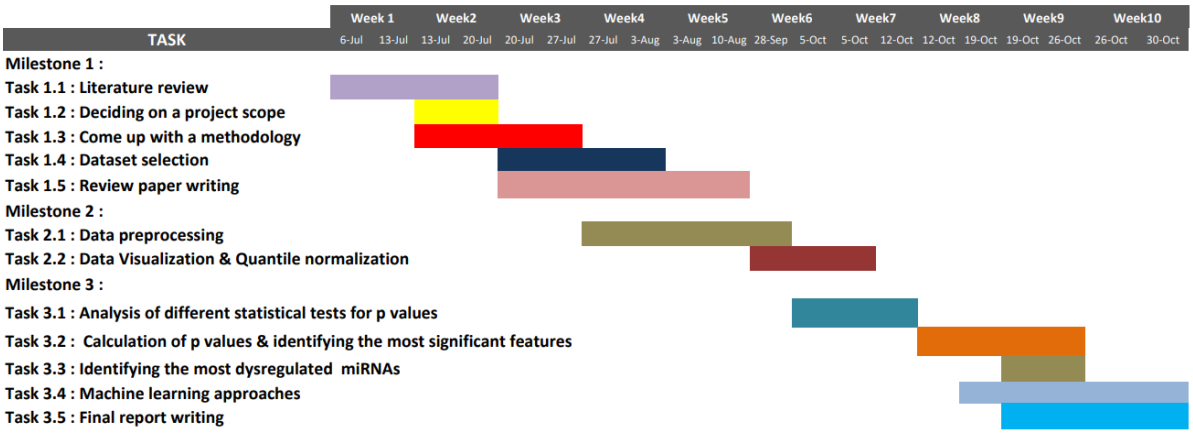

Fig. 1.3 Gantt chart for phase 1

1.2.5 Outline of the report

An introduction giving the basics of the research are discussed in this chapter. In the next chapter we are looking at the previous researches done for finding miRNA biomarkers using next generation sequencing. In chapter 3 and 4 we discuss the methodology used and how that methodology is implemented (respectively). Chapter 5 discusses the most important section in the report which is the results that we obtained. Conclusions and future works of this project are discussed in the final chapter

Chapter 2

Related work

2.1 Introduction

The diagnosis of AD patients has now become a huge challenge which is still not properly addressed. There are several researches which were carried out to detect biomarkers for Alzheimer's disease using different sequencing methods like microarray, Sanger sequencing and next generation sequencing [2]. Most of the researches have focused on next generation sequencing rather than microarray. Diverse researches have used different types of sequencing data for their investigations. miRNAs can be used to regulate AD-related proteins in the brain and it is considered as a potential novel biomarker which is mostly used in AD diagnosis [3, 4].

2.2 Sample selection

When detecting biomarkers for Alzheimer's disease, initially we have to select a sample for performing analysis. Mostly blood samples are used due to the high availability. Different types of blood samples including whole blood [5–9], serum [10–13] and plasma [14, 15] are used by many previous researchers where they have tried to find miRNA biomarkers. If we use brain samples it would give most accurate results than blood samples since AD is most prominently active in brain [16]. We would be able to give more accurate results if both blood and brain samples are used. Samples can be taken from participants generally as, AD and controls [6, 7, 10, 14] and also they can be taken considering the different stages as severe, moderate, mild AD and controls [12]. Another approach in collecting samples is taking then from participants with HC, MCI and AD [11]. The number of samples used when developing a diagnosis method can be identified as one of

the main factors which could affect the final results. Next generation sequencing platform is the most trending method used for gathering samples for various disease diagnosis researches. Many techniques like Illumina sequencing technique [5, 6, 8, 10, 12, 3] are introduced for working with NGS data. Preprocessing the raw sequence counts can be done using a bioinformatic pipeline, which gives the read counts for each miRNA as the final outcome.

2.3 Normalization

Normalization of sequencing read counts can be performed using several normalization methods. Quantile normalization is one way that we can do normalization when we are having a high dimensional dataset. It excludes selected samples to minimize noise [5, 6]. Mean normalized read counts also can be used to filter out the miRNAs. Also we can follow a stepwise procedure to do normalization as below [12].

- From all the samples, find sequences which are common.
- Build a reference dataset using those common sequences.
- Apply logarithmic transformation
- Calculate the logarithmic difference between each sample and reference dataset.
- Form a subset by taking sequences which has a difference<2.
- Perform linear regression.
- Calculate the mid value

Looking at the results obtained from the study which used the above normalization method, it can conclude that this type of step wise normalization method can be used for obtaining the best set of miRNAs. Data visualization can be used for selecting which normalization method is best suit for a given dataset.

2.4 Statistical Analysis

Initial detection of miRNA can be done by initially calculating a significance value(p value). P value is a value between 0 and 1, which shows the level of statistical significance. If a p value is less than the significance level (0.05), it is considered as a nominally significant p value and we can select those miRNA as the most impacting miRNAs.

WMW test [10, 13], Wald test and Fisher's exact test [7] can be used to calculate the p values and these p values can be adjusted for multiple testing using an approach like Benjamini-Hochberg approach [5, 6]. Other than that, t test and kruskal test can also be used to calculate significance values [14].

2.5 Validation of samples

Validation of the samples makes it easier for the next steps in the investigation and also it makes the final results more accurate. After the statistical analysis process, for validating the obtained samples, quantitative real time-polymerase chain reaction (qRT-PCR) method is used by many researchers. It analyzes the expression of single miRNAs by applying the method on previously used samples for sequencing [6, 8–13]. But in a previous study [6], they have additionally included patients with AD and also patients with other neurological disorders in the validation step, to analyze the the set of miRNAs they obtained in the previous step. After the validation is carried out, the miRNAs can be further filtered out to obtain the most significant miRNAs [12].

2.6 Receiver operating characteristic curves

Receiver operating characteristic curve analysis is used to evaluate the performance or accuracy of a classification model. ROC is a plot of sensitivity against specificity for selected samples. It is also used to initially detect the dysregulation of miRNAs and to discriminate between AD and NC sample groups. The area under the curve is the degree of separability. If the AUC is high, that means that particular miRNA is better to distinguish patients with AD and control.

2.7 Feature selection and Classification

If we use a classification model without using feature selection, it will take more run time due to the huge size with redundant features. Therefore it is required to apply some feature selection method to reduce those redundant features. Hierarchical clustering is a feature selection method which can be used to statistically analyze the dataset [5, 6, 8, 9, 11]. It will build clusters of miRNAs having similar patterns. Principal Component Analysis is another approach which can be used for the feature selection [5, 9, 11]. Machine learning classifier models are used to predict whether a sample belongs to AD or control. AdaboostM1, J48 decision tree, random forest and support vector

machines and radial basis SVM are some machine learning approaches that can be used for building prediction models. In a previously done study, they have built a separate model by performing 7-way cross validation using 7 randomly picked partitions of 5 positive and 5 negative samples each for the feature selection.

2.8 Summary

According to the review we have done, we identified how we can use miRNAs to diagnosis AD and what are the miRNA diagnostic biomarkers which can be found in AD patients. In each study, for filtering out the candidate miRNA, step wise procedures including initial detection and statistical analysis have performed. When consider about previous studies, there are several limitations. The most common limitation of most of the research is they used a limited number of the cohort to their experiments. It is hard to find a large number of Alzheimer's disease patients to do massive experiments. But we can obtain better results if we expand the cohort size. In many studies, samples with analyzed dementia and controls have used. But not discussing about the possibility to discover pre-clinical biomarkers for Alzheimer disease is a limitation of most of the previous studies. A model which was built in a one previously done study [11], does not develop to anticipate movement from HC to MCI or MCI to AD. Also, this model was incapable of applying for late-stage AD findings. In another study [13], they have mentioned that they were unable to recognize a mechanism to identify the variation of miRNAs in serum samples. Considering all the drawbacks, limitations and also the developments found in the previous studies, in this research, we are focusing on finding a more accurate solution for detecting AD biomarkers.

Chapter 3

Methodology

3.1 Proposed Methodology

This project is about identifying the best set of biomarker miRNAs for Alzheimer's disease using NGS data. It is important to identify the biomarkers to pre-identify the disease. To achieve this task we plan to follow the following procedure.

- Finding a suitable NGS dataset from the public repository
  - The dataset should contain data from NGS platform such as Illumina MiSeq, Illumina HiSeq, SOLiD, etc.
  - The dataset should contain blood or brain miRNAs from AD patients and healthy humans
- Preprocessing sequencing data
  - Trimming adapters, indexes and low quality sequences
  - Filtering short read sequences
  - Filtering remaining low quality reads
- Creating a summarized dataset by including miRNA read counts for each miRNAs
- Removing lowly abundant data
- Data visualization and normalization
- Statistical analysis of summarized data
  - Finding the most significant miRNAs using P value (significance value)

  - Calculating AUC value for ROC curve
  - Finding dysregulated miRNAs including upregulated miRNAs and downregulated miRNAs using AUC

- Classification using Machine Learning techniques
  - Support Vector Machine
- Validation of results
- Developing a method for clinical use

3.2 Design Methodology

3.2.1 Conceptual design

In Figure 3.1, the overall overview of the proposed methodology is shown.

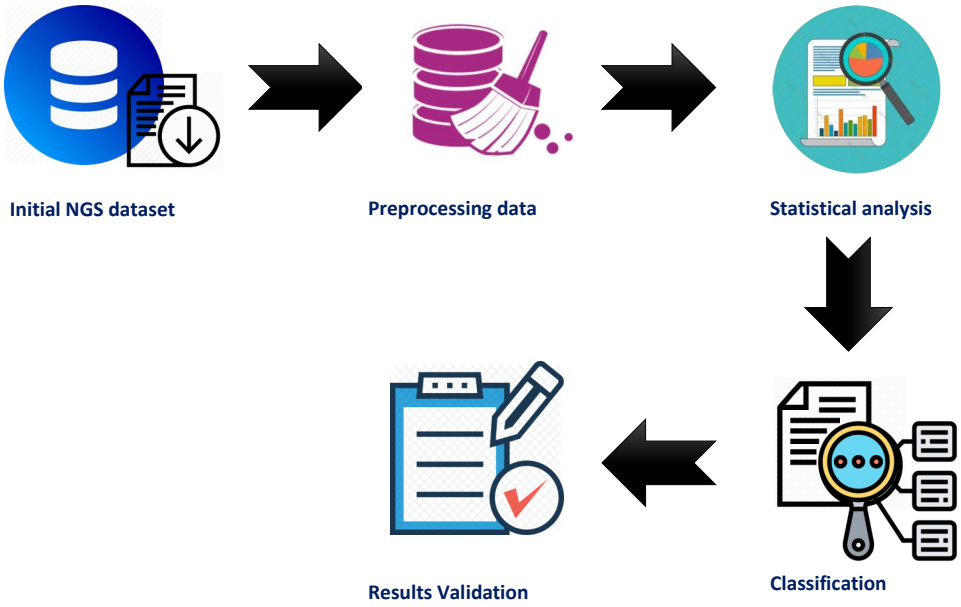

Fig. 3.1 Overview of the proposed methodology

3.2.2 Methodological approach

3.2.3 Sample Selection

We decided to include both brain and blood samples to be included in our dataset for obtaining more accurate final results. For selecting those samples, we used the NCBI database.

3.2.4 Next generation analysis

The raw sequencing data which was collected using the Illumina HiSeq 2000 platform was available in the NCBI GEO database under accession number GSE46579. The dataset contained samples from 48 AD patients (n = 48) and 22 healthy controls (n = 22). Sample data was downloaded and extracted to fastq type using sratoolkit. Then the data preprocessing and analysis were done by using the Galaxy platform. Galaxy is a web-based, open-source platform that is created for scientific data analysis. It provides a lot of bioinformatics tools for data preprocessing and analyzing. First, the quality report of sequencing data was generated using the FastQC tool. FastQC tool is used to quality checking in row sequencing data. Then, using the tool Trim Galore, data trimming was performed. The package Trim Galore allows both quality trimming and adapter trimming at once [17]. Adapters, Poly A tails, and low-quality reads were trimmed in the trimming procedure. Then, the data filtering procedure was performed using the Filter FASTQ tool. Short read sequences and low-quality sequences were removed in filtering. After that, the NGS reads were mapped against a reference genome (h38) using Bowtie2. Bowtie2 is memory efficient and ultrafast tool for aligning sequencing reads to the long reference sequencing reads. Then the reads were mapped against the hsa.gff3 miRNA precursor sequences from the miRBase database (v22) [18] and the read counts of each miRNA were founded using the htseq-count tool. This preprocessing procedure was done for every sample using the galaxy platform and then the summarized dataset was created by using miRNA read counts of each sample.

3.2.5 Data Visualization

All the features and also some random features were visualized in order to get an idea of the distribution. The method of normalization was chosen by studying the visualized distribution of dataset. In addition to that, data visualization techniques were used for studying the correlation of selected features.

3.2.6 Normalization

As mentioned in the previous section, when doing normalization data visualization was mostly used. We decided to use Quantile normalization method over general minmax normalization method. The quantile normalization method removed unwanted variations from noisy data.

3.2.7 Statistical Analysis

The summarized dataset from the NGS data analysis was used to perform statistical analysis. The statistical analysis was done by using python which has become a popular data analysis language in bioinformatics because of clean syntax, straightforward semantics, and third-party toolkits. As mentioned in the previous section, quantile normalization was done to make all distributions identical in statistical properties. We calculated the p values (significance values) for each miRNA using Wilcoxon-Mann-Whitney test and adjusted for multiple testing with the Benjamini-Hochberg adjustment technique. Fold change was caluculated for each miRNA in the dataset. Generally, it is a technique which is used to get an idea of how much change occurs going from one value to another. Here, we tried to get fold values for each miRNA, to check how each miRNA changes going from AD samples to control samples. P values were used to identify the most significant miRNAs. Addition to p values, we used fold change for each miRNA to get the most significant set of features. We used TAM (A tool for miRNA data analysis) for those selected features to get an idea of the most down regulated and the most up regulated features in the Alzheimer's disease. For the previously selected features using p values and fold changes, we calculated the receiver operator characteristics area under the curve values for each miRNA. The AUC values are used to identify the most upregulated and the most downregulated miRNAs.

3.2.8 Classification

We used machine learning approaches for the classification. Performances of different classification algorithms with our dataset were analyzed for obtaining the best suited algorithm. The classifier with the highest accuracy was then used for testing the selected features filtered out from statistical analysis step.

Chapter 4

Experimental Setup and Implementation

4.1 Technical Tools

- Sratoolkit : The toolkit was used to download SRA datasets from the NCBI database and convert them to fastq format.

The data preprocessing and analysis were done by using the Galaxy platform. Galaxy is a web-based, open-source platform that is created for scientific data analysis. It provides a lot of bioinformatics tools for data preprocessing and analyzing. The following tools were used to analyze data in the galaxy.

- FastQC : FastQC tool is used for quality checking in row sequencing data. It generates a quality report of sequencing data
- Trim Galore : Data trimming was performed using Trim Galore. The package Trim Galore allows both quality trimming and adapter trimming at once. Adapters, Poly A tails, and low-quality reads were trimmed using Trim Galore.
- Filter FASTQ : Using Filter FASTQ, short-read sequences and low-quality sequences were removed.
- Bowtie2 : The NGS reads were mapped against a reference genome (h38) using Bowtie2. Bowtie2 is memory efficient and ultrafast tool for aligning sequencing reads to the long reference sequencing reads.

- Htseq-count : The reads were mapped against the hsa.gff3 miRNA precursor sequences from the miRBase database (v22) [18] and the read counts of each miRNA were founded using the htseq-count tool.

The statistical analysis was done by using python which becomes a popular data analysis language in bioinformatics because of clean syntax, straightforward semantics, and third-party toolkits. Following python libraries were used to the procedure.

- Numpy : Numpy supports large matrices and arrays analysis
- Scipy : Scipy supports mathematics, science and engineering
- scikit learn : Provide an efficient Machine Learning environment for analysis and classification of data
- Matplotlib : Matplotlib is used to plotting graphs in python

These are the tools that we used to perform NGS data analysis and statistical analysis. Also we used a jupyter notebook to run python codes. It is an open-source web application that allows create and share documents with code, comments, results and visualizations.

4.2 Sample Selection

A set of samples which was collected using the Illumina HiSeq 2000 platform was used. It was available in the NCBI GEO database under accession number GSE46579. We used 69 of those samples with 502 unique features.

4.3 Data Manipulation and Testing

4.3.1 Data Visualization

As mentioned in the previous chapter, data visualization is mostly carried out to study the features in the dataset and their distributions. We visualized density plots as in Figure 4.1 and Figure 4.2 of the normalized data for getting an idea how the distribution is after Qunatile normalization and after minmax normalization. Since the distribution looked the same, we decided to go with the quantile normalization as in this kind of bioinformatic analysis, quantile normalization is most commonly used mostly considering the size of the datasets used. Figure 4.3 and Figure 4.4 shows the mean distribution and the box plot of selected features which we used to study the features of the dataset.

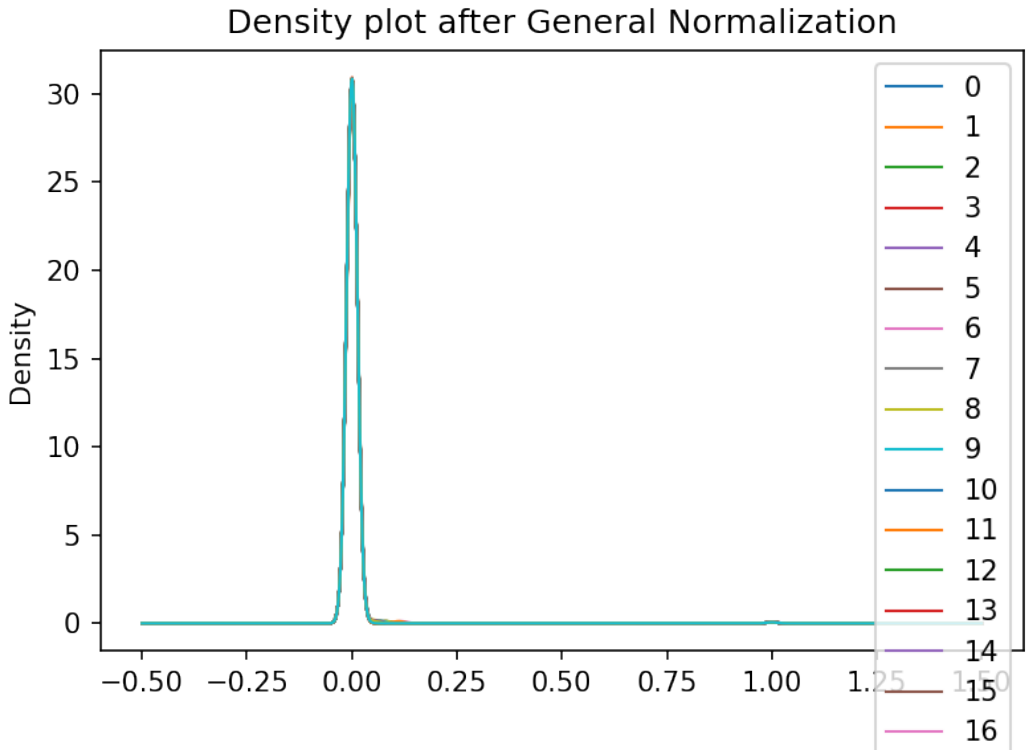

Fig. 4.1 Density plot after minmax normalization

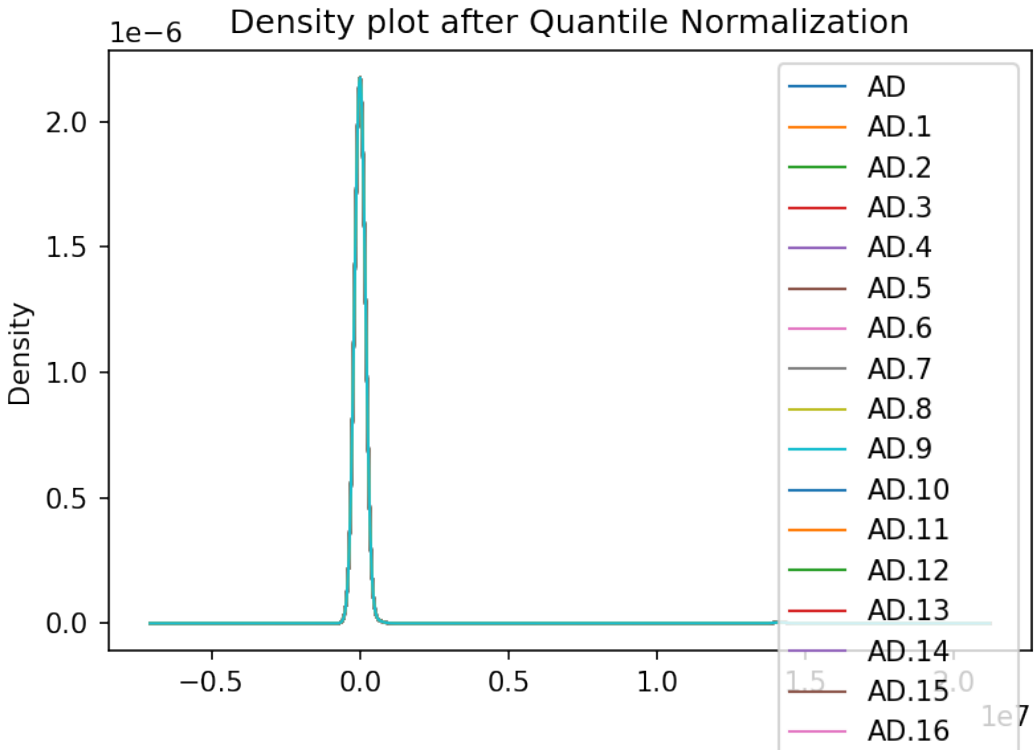

Fig. 4.2 Density plot after Quantile normalization

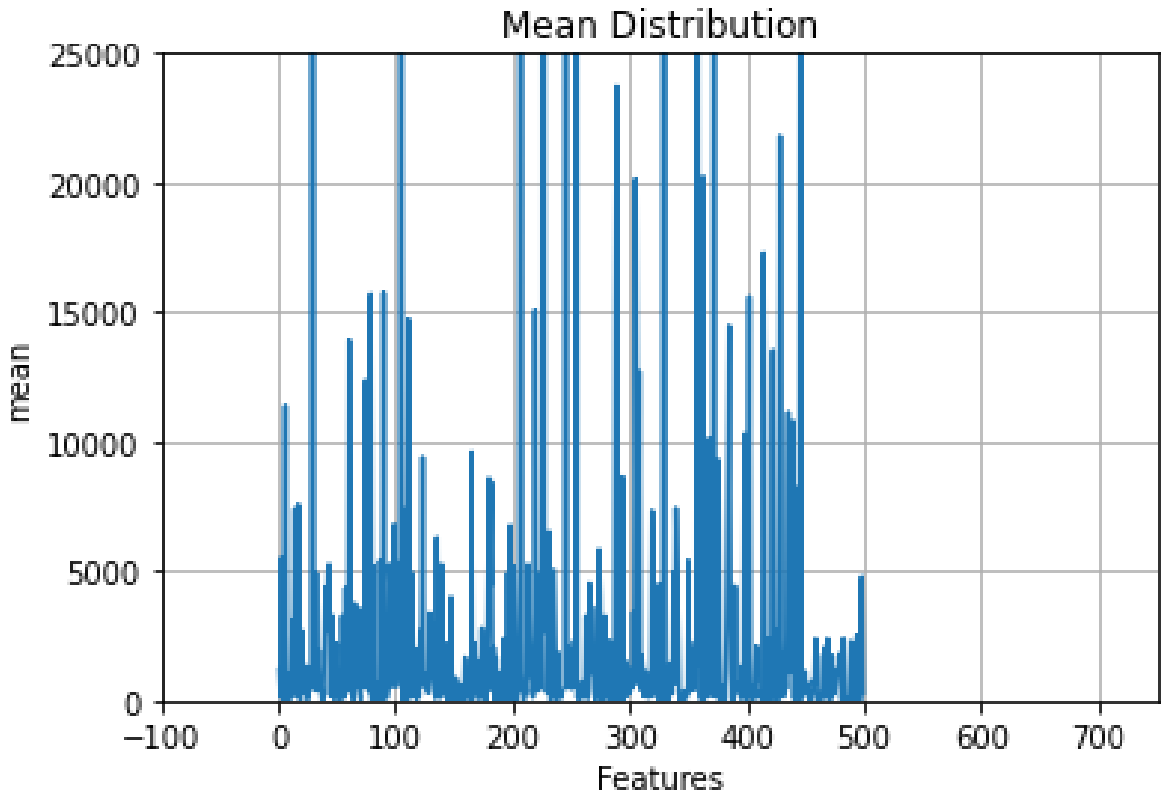

Fig. 4.3 Mean Distribution

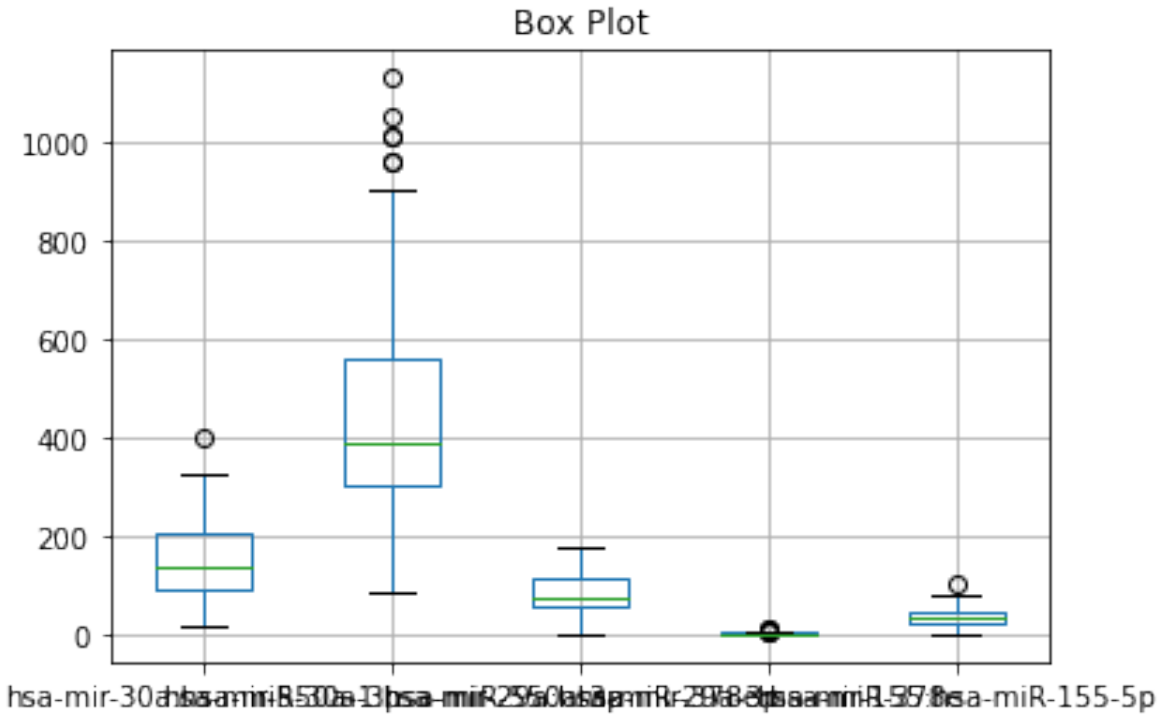

Fig. 4.4 Box plot for five features

4.3.2 Statistical analysis

Statistical analysis is mainly used for finding the most dysregulated miRNAs out of 503 miRNAs in the dataset. The most significant miRNA have the ability to identify strong distinction between AD and control sample. Therefore, using WMW test we calculated p values for each miRNA across AD and control samples in order to obtain those significant features. We used significance level of 0.05 to obtain 261 most significant set of miRNAs, as mentioned in the previous chapter.The null hypothesis is taken as that two distribution of AD and control samples for a particular miRNA is identical. This was rejected for those with significance value less than 0.05, making the alternative hypothesis, which states that two distributions for a miRNA is different, true. Using both the fold change values and the p values, we obtained 228 set of features with high significance values. For each of those significant miRNAs, AUC values were calculated for identifying the most up regulated and the most down regulated miRNAs. There, we identified 154 down regulated and 32 up regulated miRNAs. As mentioned in the previous chapter, we also used the TAM tool for analysing those selected significant features.

4.3.3 Classification

As mentioned in the previous chapter, for some selected classifiers, we fit our dataset with 503 miRNA to analyze the accuracies (Table 4.1) of each classifier. We used this analysis to select the best fit model for using in the classification of AD and control samples for the selected miRNAs.

Table 4.1 Accuracies of different algorithms modelled for the dataset

Model	Training	Testing
Logistis regression	0.82	0.71
Naive Bayes	0.73	0.50
KNN	0.78	0.50
Linear SVM	0.83	0.71
Random Forest	1	0.64
Gaussian SVM	0.73	0.5

4.3.4 Pitfalls and workarounds

When starting this project, the main challenge that we have to face was lack of background knowledge about the bioinformatics. Initially we didn't properly understand what kind of data that generated from next generation sequencing platforms. Also we hadn't knowledge to how to analyze those sequencing data. To overcome that issue we had to do lots of background searches. And we had to use some articles and tutorials to learn about next generation sequencing data analysis tools. Another problem that we had to face was finding a proper NGS miRNA Alzheimer's disease dataset for analysis. After searching a lot we found GSE46579 from NCBI public database. But the initial dataset was large (about 33GB) and there were fastq type datasets for each sample. So it was hard to download each dataset to a PC and perform analysis. So we used the galaxy data analysis platform for NGS data analysis. In the galaxy, there was a way to automate the analyzing procedure. Also, the data was used from the NCBI database without downloading it to the PC. Because of the lack of knowledge to find a new methodology at once we selected some articles and followed their methods to data analysis first. Because of the separated samples of data we had to create a summarized data set for further analysis after NGS analysis. When doing statistical analysis, we had to go through a lot of theories to understand the concepts since we were not familiar with most of those.

Chapter 5

Results and Analysis

5.1 Results

5.1.1 NGS analysis

There are 69 sample data files in the initial database. They contained NGS data in fastq format (Figure 5.1)

```
@SRR837486.1 HWI-ST937:130:D10R9ACXX:6:1101:1641:1978 length=50
TCCTGTACTGAGCTGCCCCGAGATGGAATTCTCGGGTGCCAAGGAACTCC
+SRR837486.1 HWI-ST937:130:D10R9ACXX:6:1101:1641:1978 length=50
CCCFFEEFHHHHHJJJJJIHHIFCCECHDHIIGIGH)8?BDH;B8CF<=F
```

Fig. 5.1 One read from fastq data file of sample SRR837486

As in Figure 5.1, the fastq sequencing data file contains the header, the sequence (+ denotes the end of the sequence), and the quality score of sequences for each read. For the sequencing data, FastQC generated a quality report about the sequences. It contains basic statistics about the file, per base, and per tile sequence quality, details about sequence length, and details about the adapter content of sequences.

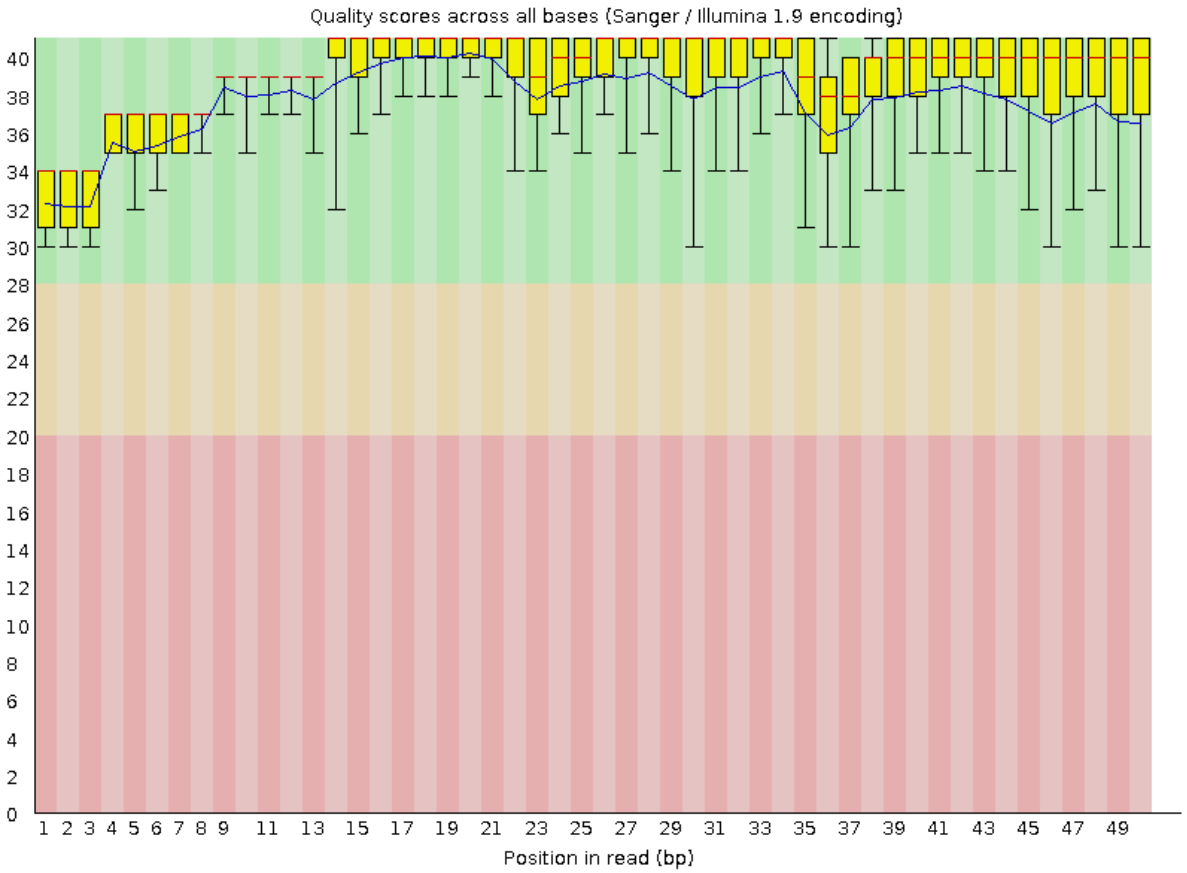

Fig. 5.2 Per base sequence quality for sample SRR837486

Figure 5.2 is a BoxWhisker type plot that shown an overview quality ranges of each sequence in fastq file of sample SRR837486. Yellow box represents 25 Adapters in sequences contain index sequences, primer binding sites, and the sites that allow fragments to connect with flow cells. Figure 5.3 shows the adapter contents of sequencing data in sample SRR837486. According to Figure 5.3, there was Illumina small RNA 3 adapter in row sequences.

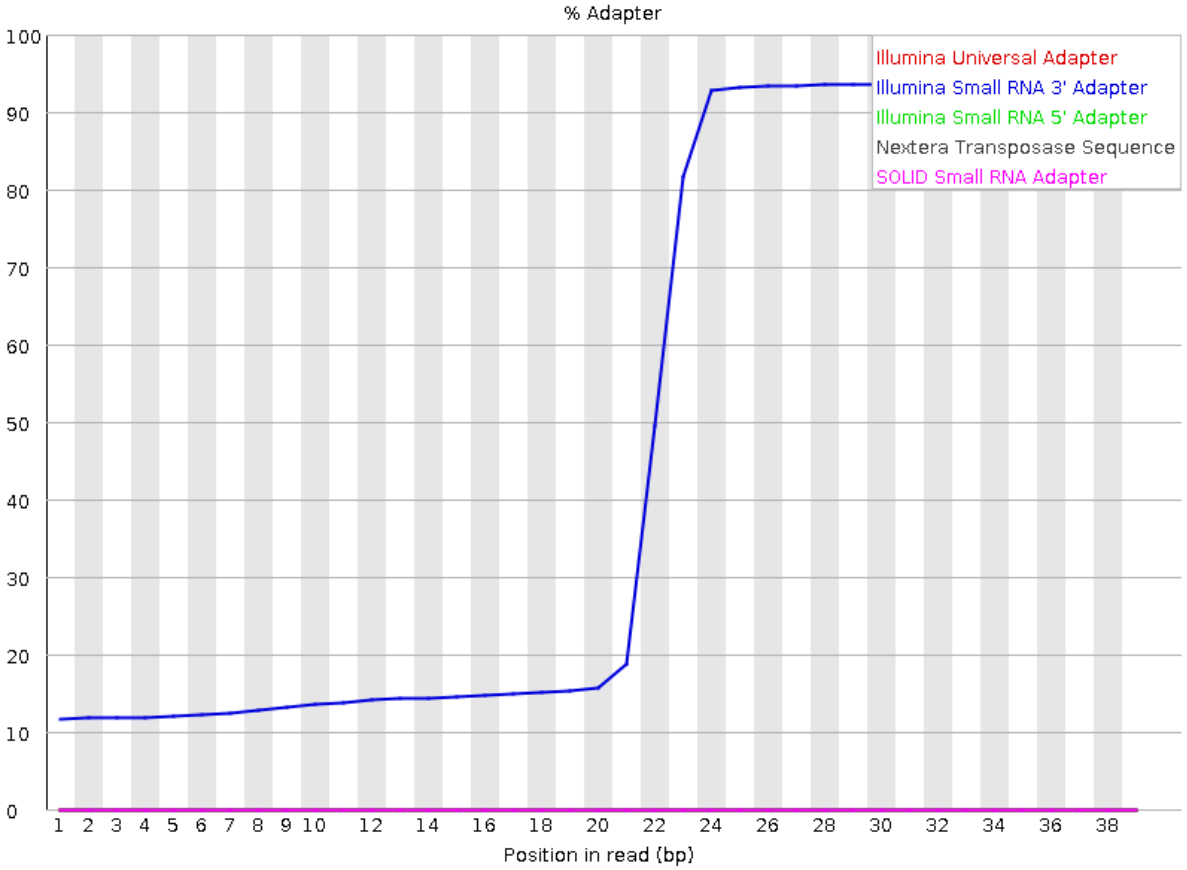

Fig. 5.3 Adapter contents of data

The NGS data preprocessing procedure contains trimming and filtering. Trimming and filtering were performed to improve the quality of data. In this procedure, the adapters and the low quality reads were removed.

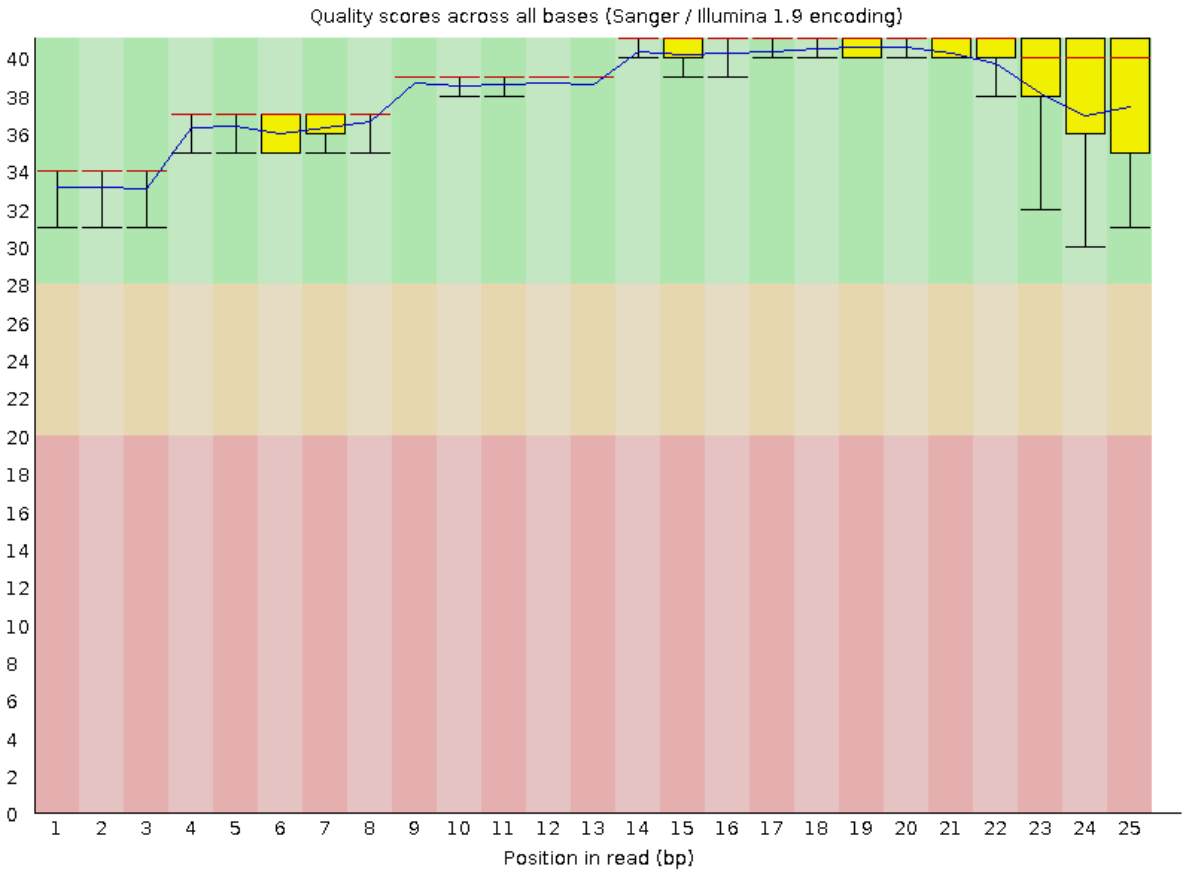

Fig. 5.4 Per base sequence quality for sample SRR837486 (After preprocessing)

Figure 5.4 shows the BoxWhisker plot for the sequence quality of sample data. As in Figure 5.4, the sequence quality was improved and the length of the sequences was reduced.

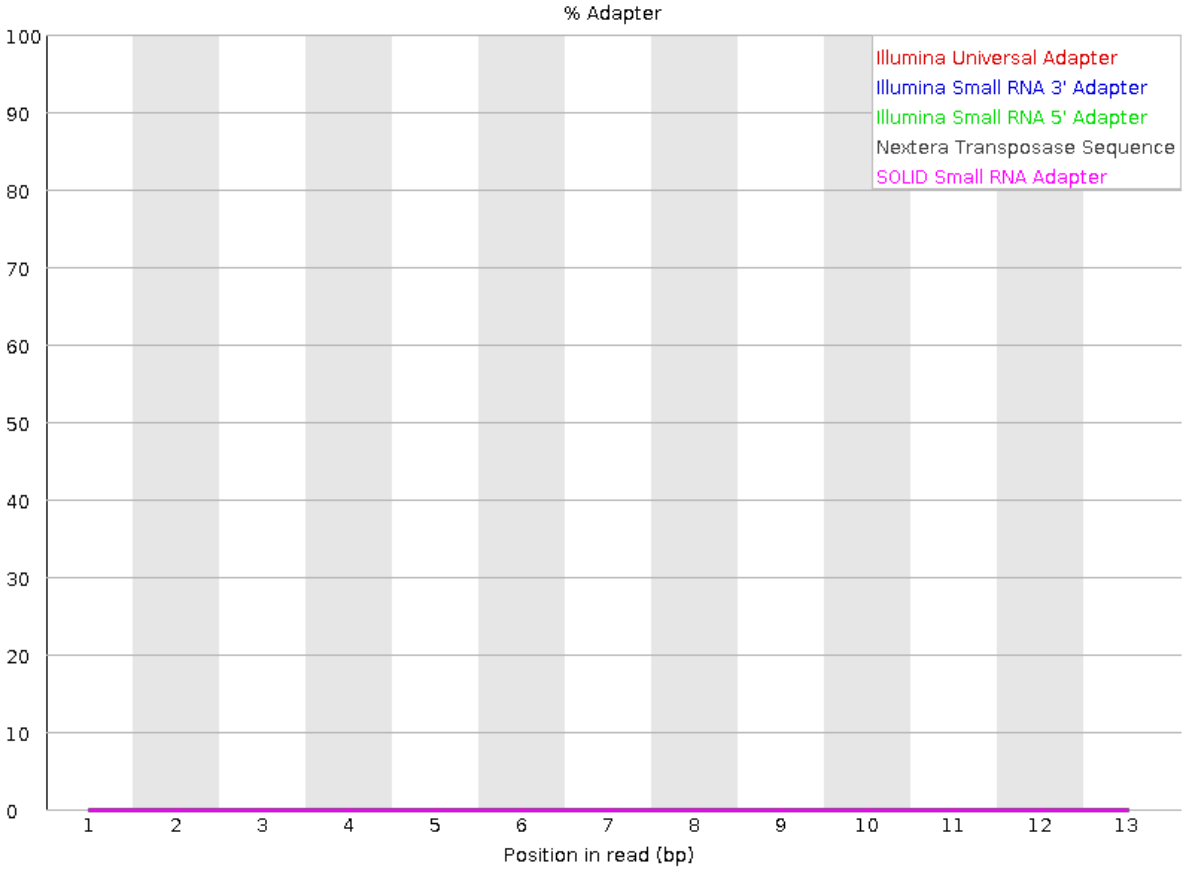

Fig. 5.5 Adapter contents of data (After preprocessing)

As show in Figure 5.5, all adapter contents were removed in preprocessing procedure. There were about 13 million reads in sample SRR837486 at beginning. After the preprocessing about 2 million reads were removed which are considered as low quality sequences or short read sequences. Then the mapping to the reference genome resulted in the BAM type file and using miRBase we identified the read counts of each miRNA of each sample (Figure 5.6).

Geneid	Bowtie2 on data 11: alignments
hsa-let-7a-2-3p	0
hsa-let-7a-3p	0
hsa-let-7a-5p	9996
hsa-let-7b-3p	45
hsa-let-7b-5p	4190
hsa-let-7c-3p	0
hsa-let-7c-5p	151
hsa-let-7d-3p	345
hsa-let-7d-5p	1096
hsa-let-7e-3p	0
hsa-let-7e-5p	21
hsa-let-7f-1-3p	0
hsa-let-7f-2-3p	0
hsa-let-7f-5p	3131
hsa-let-7g-3p	0
hsa-let-7g-5p	2088
hsa-let-7i-3p	41
hsa-let-7i-5p	2986
hsa-miR-1-3p	0
hsa-miR-1-5p	0
hsa-miR-100-3p	0
hsa-miR-100-5p	1197
hsa-miR-101-2-5p	0

Fig. 5.6 miRNA read counts of SRR837486

In this procedure we identified 2652 miRNAs. By removing <50 summed up read counts we created a summarized dataset for further analysis. After removing <50 read counts, there were only 503 miRNAs were left.

5.1.2 Feature selection

228 miRNAs were identified as the most statistically significant miRNAs using the significance level of 0.05 and the fold change of |1| (>|1|). Then, we obtained 192 features from the feature selection method done based on the univariate ROC AUC classification. Figure 5.7 shows the ROC AUC values for the previously selected 228 features.

According to the AUC values of each miRNA, we identified the the most down regulated miRNA and the most up regulated miRNA as hsa-mir-1294:hsa-miR-1294 (AUC = 0.875) and hsa-mir-548x:hsa-miR-548aj-5p (AUC = 0.2917)respectively.Figure 5.8 shows the distributions of those two miRNAs across AD and control samples.

As mentioned in the previous two chapters, the 228 features obtained using p value and fold change, were also subjected to TAM tool analysis. From that we obtained 9 down regulated and 14 up regulated miRNAs which can be used for further analysis purposes.

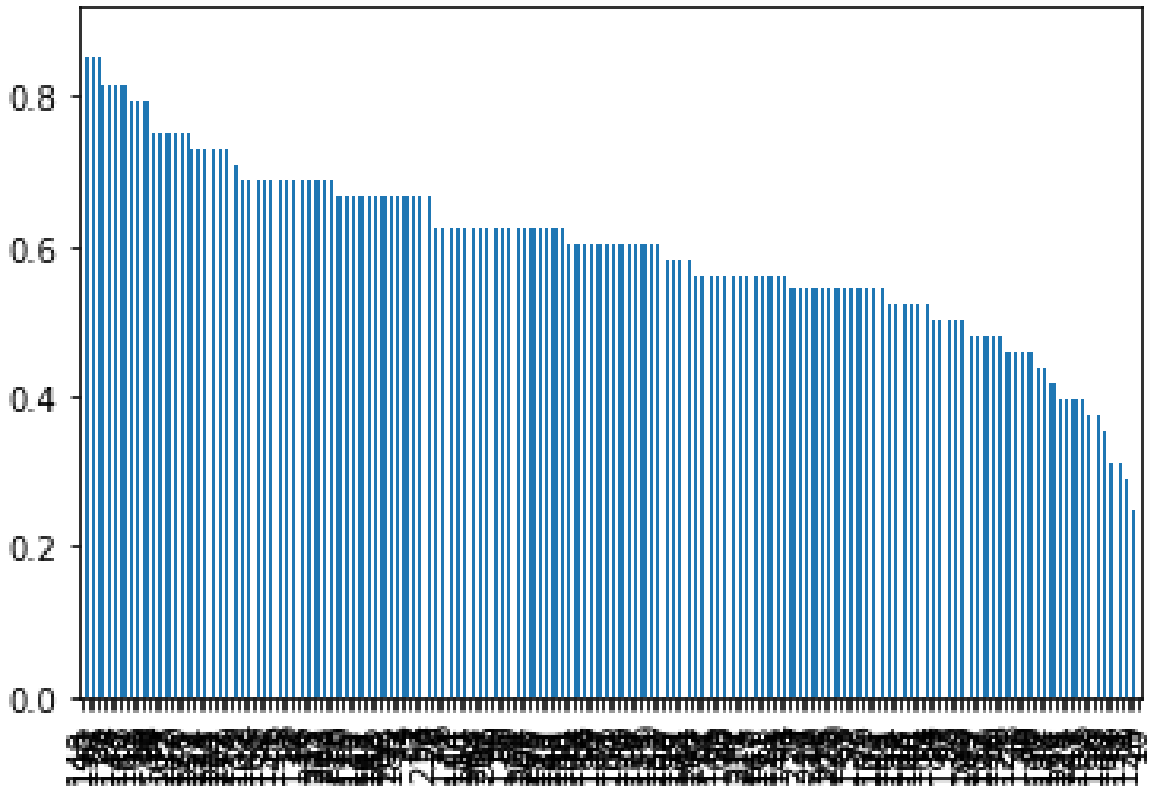

Fig. 5.7 ROC values of 228 features

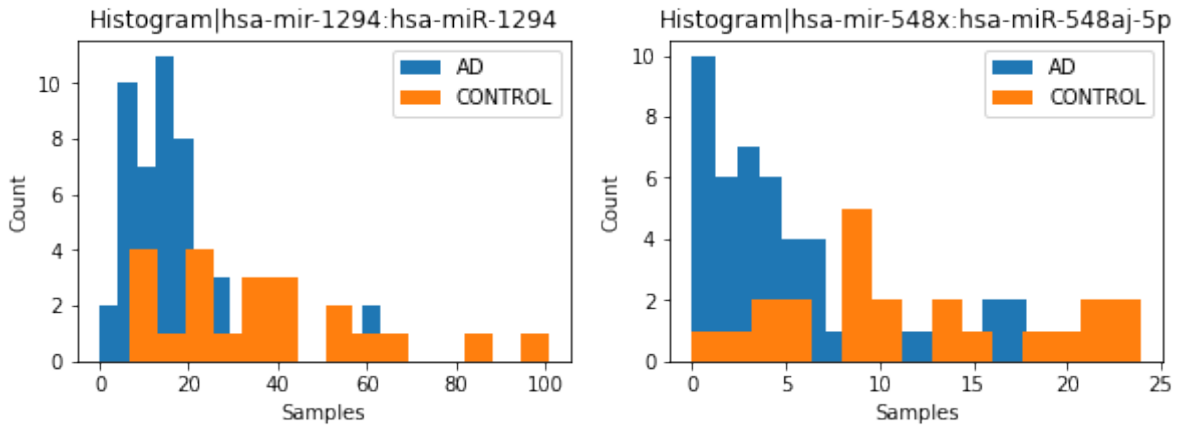

Fig. 5.8 Distributions of the most up regulated miRNA and the most down regulated miRNA

Chapter 6

Conclusions and Future Works

6.1 Conclusion

In this report we have discussed about how to detect miRNA biomarkers for Alzheimer's disease using next generation sequencing. Initially we have discussed about the need of a solution to identify Alzheimer's disease in the early stage. Then we have mentioned about the literature review we have done. When we were doing the literature review, we have identified several miRNA biomarkers in different studies which used NGS. In these studies there were some limitations. In our approach so far, initially we have taken samples from participants with AD and control. Then samples were preprocessed and statistically analyzed. Significance values were calculated using Wilcoxon-Mann-Whitney (WMW) test. Also we have used ROC analysis. Using this procedure, here we have calculated a set of the most significant biomarkers for AD. In the next phase we expect to validate the samples and make the final results more accurate and develop this methodology for clinical use.

6.2 Future Work

We have done selecting miRNA biomarkers by using most downregulated miRNAs. In the next phase, we are going to work on validating the same samples which have used previously for NGS. For validation we will use a new set of samples from AD patients and controls. Also, we are planning to create a web based tool to visualize the distributions of datasets, preprocessing, statistical analysis and classification. It will be useful for people who are working with genomic data science and also for the clinical use.

References

[1] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013.

[2] G. Stiglic, M. Bajgot, and P. Kokol, "Gene set enrichment meta-learning analysis: next-generation sequencing versus microarrays," *BMC bioinformatics*, vol. 11, no. 1, p. 176, 2010.

[3] M. Basavaraju and A. De Lencastre, "Alzheimer's disease: presence and role of micrornas," *Biomolecular concepts*, vol. 7, no. 4, pp. 241–252, 2016.

[4] C. Nie, Y. Sun, H. Zhen, M. Guo, J. Ye, Z. Liu, Y. Yang, and X. Zhang, "Differential expression of plasma exo-mirna in neurodegenerative diseases by next-generation sequencing," *Frontiers in Neuroscience*, vol. 14, p. 438, 2020.

[5] A. Keller, C. Backes, J. Haas, P. Leidinger, W. Maetzler, C. Deuschle, D. Berg, C. Ruschil, V. Galata, K. Ruprecht, *et al.*, "Validating alzheimer's disease micro rnas using next-generation sequencing," *Alzheimer's & Dementia*, vol. 12, no. 5, pp. 565–576, 2016.

[6] P. Leidinger, C. Backes, S. Deutscher, K. Schmitt, S. C. Mueller, K. Frese, J. Haas, K. Ruprecht, F. Paul, C. Stähler, *et al.*, "A blood based 12-mirna signature of alzheimer disease patients," *Genome biology*, vol. 14, no. 7, p. R78, 2013.

[7] H. Z. Y. Wu, A. Thalamuthu, L. Cheng, C. Fowler, C. L. Masters, P. Sachdev, K. A. Mather, A. I. Biomarkers, and L. F. S. of Ageing, "Differential blood mirna expression in brain amyloid imaging-defined alzheimer's disease and controls," *Alzheimer's Research & Therapy*, vol. 12, pp. 1–11, 2020.

[8] J.-i. Satoh, Y. Kino, and S. Niida, "Microrna-seq data analysis pipeline to identify blood biomarkers for alzheimer's disease from public data," *Biomarker insights*, vol. 10, pp. BMI–S25132, 2015.

[9] C. Backes, B. Meder, M. Hart, N. Ludwig, P. Leidinger, B. Vogel, V. Galata, P. Roth, J. Menegatti, F. Grässer, *et al.*, "Prioritizing and selecting likely novel mirnas from ngs data," *Nucleic acids research*, vol. 44, no. 6, pp. e53–e53, 2016.

[10] L. Tan, J.-T. Yu, M.-S. Tan, Q.-Y. Liu, H.-F. Wang, W. Zhang, T. Jiang, and L. Tan, "Genome-wide serum microrna expression profiling identifies serum biomarkers for alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 40, no. 4, pp. 1017–1027, 2014.

[11] L. Cheng, J. Doecke, R. Sharples, V. Villemagne, C. Fowler, A. Rembach, and B. Australian Imaging, "Lifestyle (aibl) research group.(2015). prognostic serum mirna biomarkers associated with alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment," *Molecular Psychiatry*, vol. 20, no. 10, pp. 1188–1196.

[12] R. Guo, G. Fan, J. Zhang, C. Wu, Y. Du, H. Ye, Z. Li, L. Wang, Z. Zhang, L. Zhang, *et al.*, "A 9-microrna signature in serum serves as a noninvasive biomarker in early diagnosis of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 60, no. 4, pp. 1365–1377, 2017.

[13] H. Dong, J. Li, L. Huang, X. Chen, D. Li, T. Wang, C. Hu, J. Xu, C. Zhang, K. Zen, *et al.*, "Serum microrna profiles serve as novel biomarkers for the diagnosis of alzheimer's disease," *Disease markers*, vol. 2015, 2015.

[14] G. Lugli, A. M. Cohen, D. A. Bennett, R. C. Shah, C. J. Fields, A. G. Hernandez, and N. R. Smalheiser, "Plasma exosomal mirnas in persons with and without alzheimer disease: altered expression and prospects for biomarkers," *PloS one*, vol. 10, no. 10, p. e0139233, 2015.

[15] A. Gámez-Valero, J. Campdelacreu, D. Vilas, L. Ispierto, R. Reñé, R. Álvarez, M. P. Armengol, F. E. Borràs, and K. Beyer, "Exploratory study on microrna profiles from plasma-derived extracellular vesicles in alzheimer's disease and dementia with lewy bodies," *Translational neurodegeneration*, vol. 8, no. 1, p. 31, 2019.

[16] N. Hara, M. Kikuchi, A. Miyashita, H. Hatsuta, Y. Saito, K. Kasuga, S. Murayama, T. Ikeuchi, and R. Kuwano, "Serum microrna mir-501-3p as a potential biomarker related to the progression of alzheimer's disease," *Acta neuropathologica communications*, vol. 5, no. 1, p. 10, 2017.

[17] S. V. Toshchakov, I. V. Kublanov, E. Messina, M. M. Yakimov, and P. N. Golyshin, "Genomic analysis of pure cultures and communities," in *Hydrocarbon and Lipid Microbiology Protocols*, pp. 5–27, Springer, 2015.

[18] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "mirbase: from microrna sequences to function," *Nucleic acids research*, vol. 47, no. D1, pp. D155–D162, 2019.