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

Group:
E/15/081 Dinuwanthi S.L.I
E/15/362 Thilakarathna K.A.H.P
E/15/345 Sripadi D.V.

Supervisors:
Dr. Damayanthi Herath
Prof. Roshan Ragel

Background

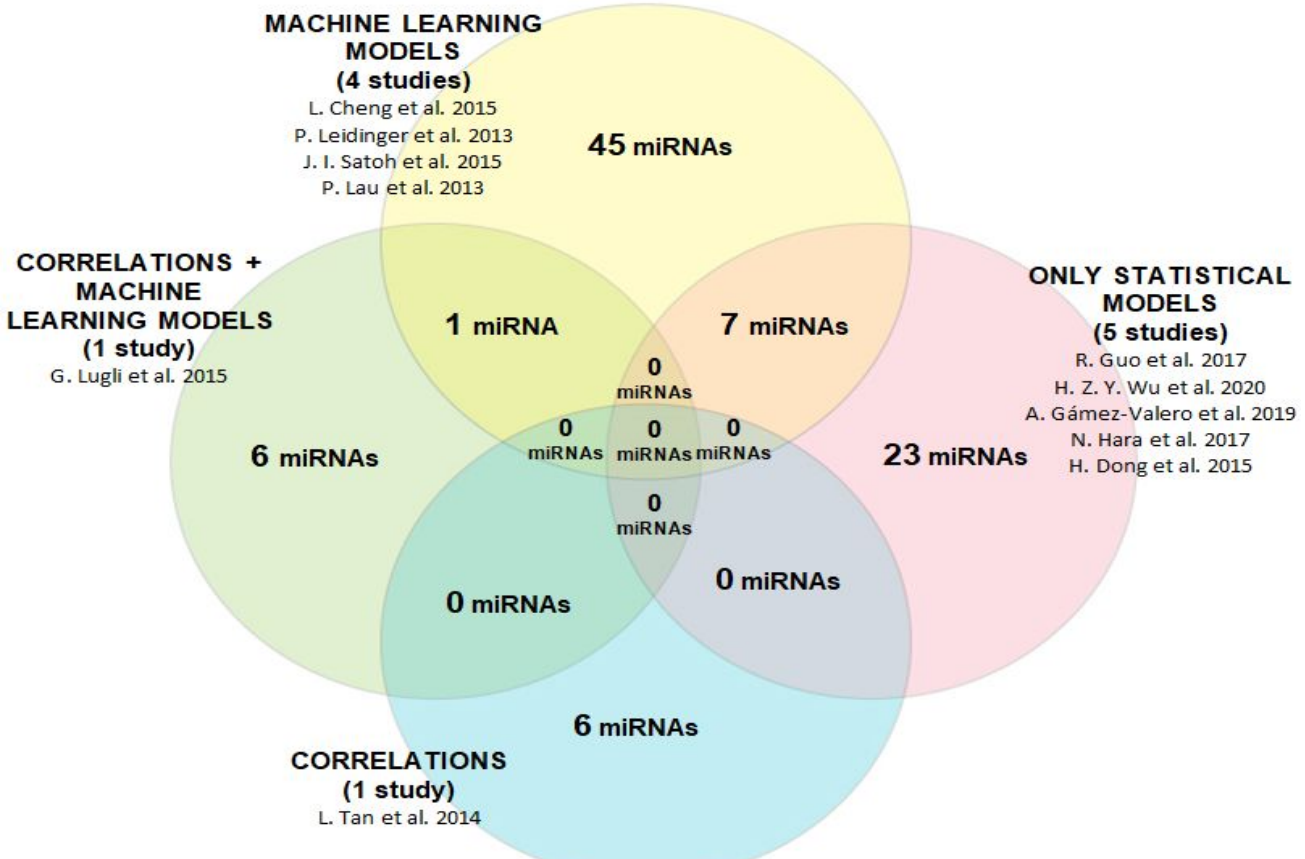
- What is Next Generation Sequencing ?
    - Sample preparation
    - Sequencing by machines
    - Data output
- What are miRNAs ?

Problem definition

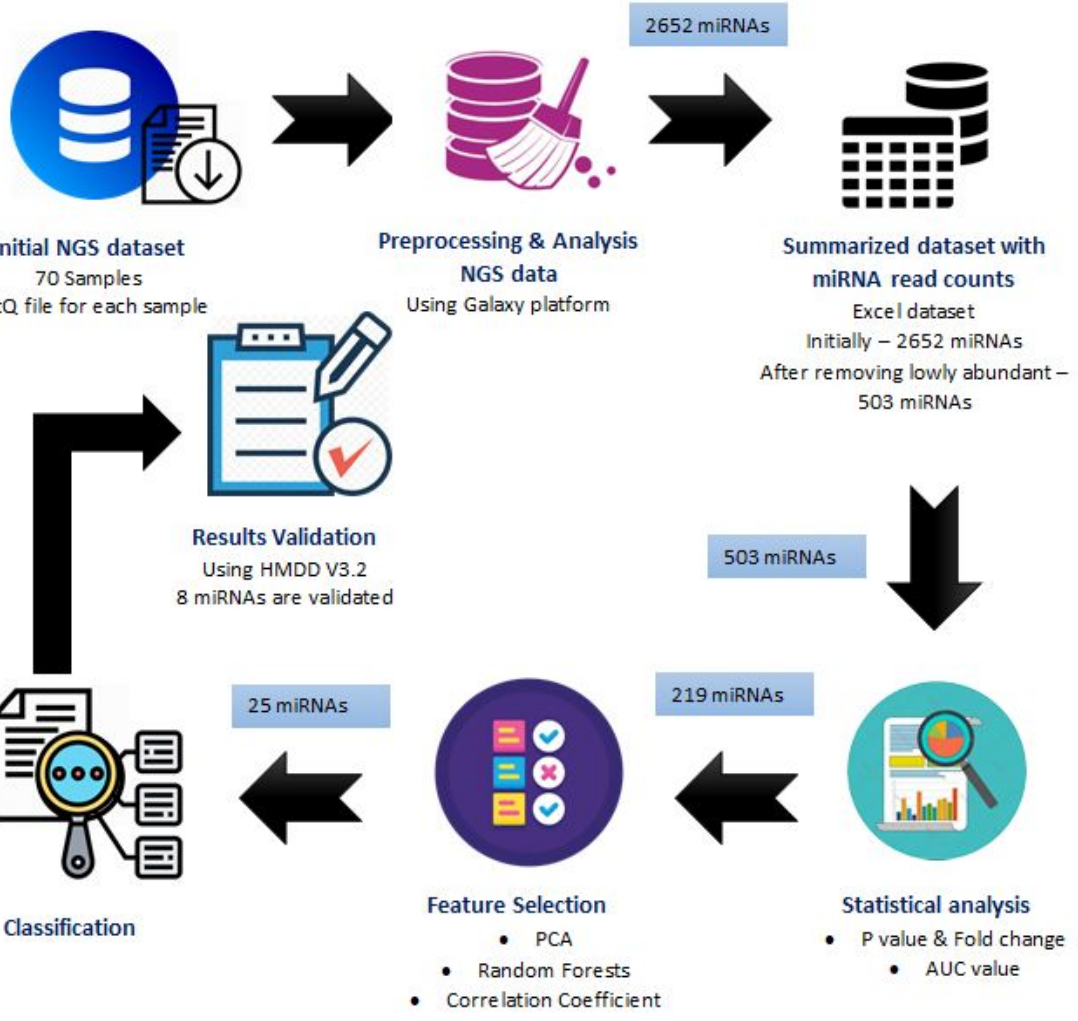
- Finding a method for identifying miRNA biomarkers for Alzheimer’s Disease

- Need of new method because of the drawbacks of previously introduced methods

Related works



Methodology



Implementation choices: Tools

NGS data analysis

- Galaxy platform - web based, open source platform
- Galaxy tools
  - FastQC - Quality check
  - Trim Galore - remove adapters and low quality reads
  - Filter FASTQ - filter short read sequences
  - Bowtie2 - map sequence against reference genome
  - Htseq-count - identify read counts

Implementation choices: Tools

Statistical Analysis

- Python - clean syntax, straightforward semantics & Third-party toolkits
- Python libraries
  - Numpy
  - Pandas
  - Scipy
  - Scikit learn
  - Matplotlib

Implementation choices: Models

Classification

- Machine Learning models
  - Logistic regression
  - Linear SVM
  - Gaussian SVM
  - Naive Bayes
  - K Nearest Neighbour
  - Random Forests

Implementation choices: Methodologies

Preprocessing

- Trimming sequence data - remove adapters, indexes, low quality reads
- Filtering - remove short read sequences

Statistical analysis

- P value & Fold change - identify most significance miRNAs
- AUC - identify dysregulated miRNAs

Implementation choices: Methodologies contd.

Feature selection

- PCA
- Random Forest
- Correlation coefficient

Classification

- Classification - Support vector machine, Logistic regression and Random Forest

NGS data analysis

TCCTGTACTGAGCTGCCCCGAGATGGAATTCTCGGGTGCCAAGGAACTCC

TCCTGTACTGAGCTGCCCCGAGGGGAATTCTGGGGTGCCAAGGAACCCCA

TCCTGTACTGAGCTGCCCCGAGTGGAATTCTCGGGTGCCAAGGAACTCCA

First 3 sequences from fastQ file

Trimming using Trim Galore

TCCTGTACTGAGCTGCCCCGAGA

TCCTGTACTGAGCTGCCCCGAGGGGAATTCTGGGG

TCCTGTACTGAGCTGCCCCGAG

After trimming

Filtering using Filter FASTQ

TCCTGTACTGAGCTGCCCCGAGA

TCCTGTACTGAGCTGCCCCGAG

After filtering

Second sequence was removed

Mapping against reference genome

Reference genome SN:chr1

..TGCCCCGAGTCCTGTACTGAGCTGCCCCGAGATCCTGTACTGA...

TCCTGTACTGAGCTGCCCCGAGA Seq 1 SN:chr1

Mapping against miRBase

miRBase sequence (From has.gtf file) Name=hsa-mir-200a

..TGCCCCGAGTCGTTCCGTCCTGTACTGAGCTGCCCCGAGATCCT...

TCCTGTACTGAGCTGCCCCGAGA Seq 1

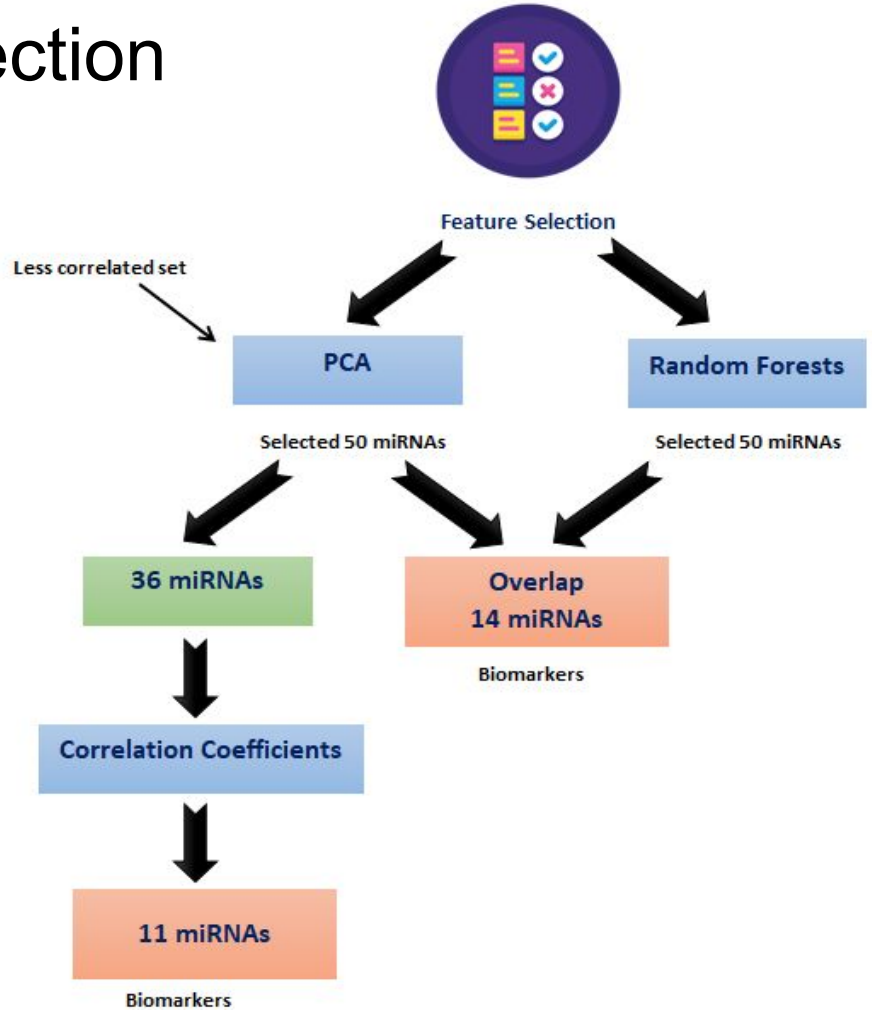
Seq 1 identified as hsa-mir-200a

Count reads of each miRNA

miRNA name	Read counts
hsa-miR-106b-3p	2622
hsa-miR-107	594
hsa-miR-10a-3p	1
hsa-mir-200a	45
hsa-miR-1269a	0

After counting reads for all miRNA

Feature Selection



Results Obtained From Each Stage of Analysis

- Filtered miRNAs Using Significance Values and Fold Change
  - 228 miRNAs

```
Filtered miRNAs using significance value and fold change are: ['hsa-mir-30a:hsa-miR-30a-3p', 'hsa-mir-550a-1:hsa-miR-550a-3p',
'hsa-mir-29a:hsa-miR-29a-3p', 'hsa-mir-628:hsa-miR-628-3p', 'hsa-mir-26a-2:hsa-miR-26a-5p', 'hsa-mir-106b:hsa-miR-106b-5p', 'hs
a-mir-4781:hsa-miR-4781-3p', 'hsa-mir-10b:hsa-miR-10b-5p', 'hsa-mir-215:hsa-miR-215', 'hsa-mir-548aj-2:hsa-miR-548g-5p', 'hsa-m
ir-181a-1:hsa-miR-181a-3p', 'hsa-mir-548x:hsa-miR-548ar-5p', 'hsa-mir-548k:hsa-miR-548av-5p', 'hsa-mir-199a-1:hsa-miR-199a-3p',
'hsa-mir-30e:hsa-miR-30e-3p', 'hsa-mir-4508:hsa-miR-4508', 'hsa-mir-548aj-2:hsa-miR-548x-5p', 'hsa-mir-371b:hsa-miR-371b-5p',
'hsa-mir-5001:hsa-miR-5001-3p', 'hsa-mir-16-2:hsa-miR-16-2-3p', 'hsa-mir-128-2:hsa-miR-128', 'hsa-mir-486:hsa-miR-486-3p', 'hsa
-mir-4482-1:hsa-miR-4482-3p', 'hsa-mir-941-4:hsa-miR-941', 'hsa-mir-550a-1:hsa-miR-550a-5p', 'hsa-mir-199a-2:hsa-miR-199b-3p',
'hsa-mir-144:hsa-miR-144-5p', 'hsa-let-7f-2:hsa-let-7f-5p', 'hsa-mir-126:hsa-miR-126-5p', 'hsa-mir-191:hsa-miR-191-3p', 'hsa-mi
r-10a:hsa-miR-10a-5p', 'hsa-mir-98:hsa-miR-98', 'hsa-mir-548x:hsa-miR-548x-5p', 'hsa-mir-363:hsa-miR-363-3p', 'hsa-mir-548h-1:h
sa-miR-548h-5p', 'hsa-mir-223:hsa-miR-223-3p', 'hsa-mir-5690:hsa-miR-5690', 'hsa-mir-199b:hsa-miR-199b-3p', 'hsa-mir-3200:hsa-m
iR-3200-3p', 'hsa-mir-424:hsa-miR-424-3p', 'hsa-mir-644b:hsa-miR-644b-3p', 'hsa-mir-548h-5:hsa-miR-548h-5p', 'hsa-mir-18a:hsa-m
iR-18a-5p', 'hsa-mir-548g:hsa-miR-548x-5p', 'hsa-mir-548g:hsa-miR-548g-5p', 'hsa-mir-21:hsa-miR-21-5p', 'hsa-mir-99b:hsa-miR-99
b-5p', 'hsa-mir-25:hsa-miR-25-3p', 'hsa-mir-937:hsa-miR-937', 'hsa-mir-1180:hsa-miR-1180', 'hsa-mir-30c-1:hsa-miR-30c-5p', 'hsa
-let-7a-1:hsa-let-7a-5p', 'hsa-mir-941-1:hsa-miR-941', 'hsa-mir-660:hsa-miR-660-5p', 'hsa-mir-421:hsa-miR-421', 'hsa-mir-374a:h
sa-miR-374a-5p', 'hsa-mir-328:hsa-miR-328', 'hsa-mir-151a:hsa-miR-151a-5p', 'hsa-mir-548x:hsa-miR-548aj-5p', 'hsa-mir-101-2:hsa
-miR-101-3p', 'hsa-mir-28:hsa-miR-28-3p', 'hsa-mir-139:hsa-miR-139-5p', 'hsa-mir-2110:hsa-miR-2110', 'hsa-let-7g:hsa-let-7g-5
p', 'hsa-mir-550a-3:hsa-miR-550a-3-5p', 'hsa-mir-548aj-2:hsa-miR-548ar-5p', 'hsa-mir-144:hsa-miR-144-3p', 'hsa-mir-548e:hsa-miR
```

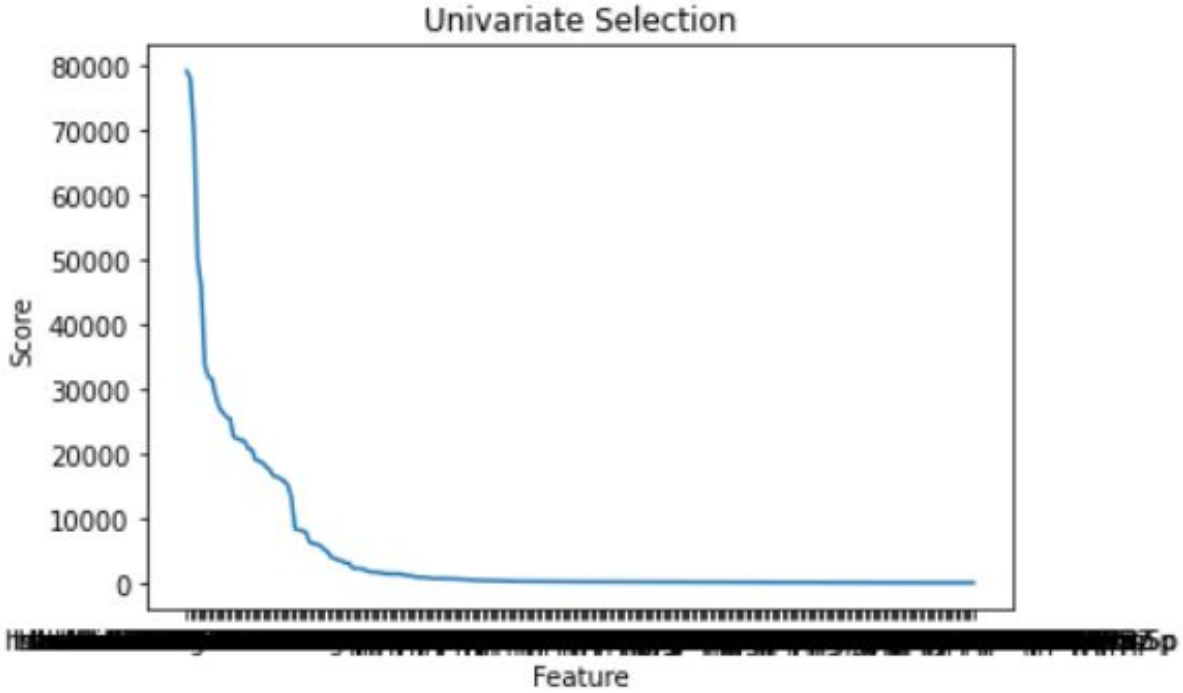
Filtered miRNAs Using AUC Analysis

- 219 miRNAs
  - 179 down-regulated miRNAs
  - 40 up-regulated miRNAs

Selected miRNAs using AUC analysys: ['hsa-mir-30a:hsa-miR-30a-3p', 'hsa-mir-550a-1:hsa-miR-550a-3p', 'hsa-mir-29a:hsa-miR-29a-3p', 'hsa-mir-628:hsa-miR-628-3p', 'hsa-mir-26a-2:hsa-miR-26a-5p', 'hsa-mir-106b:hsa-miR-106b-5p', 'hsa-mir-4781:hsa-miR-4781-3p', 'hsa-mir-10b:hsa-miR-10b-5p', 'hsa-mir-215:hsa-miR-215', 'hsa-mir-548aj-2:hsa-miR-548g-5p', 'hsa-mir-181a-1:hsa-miR-181a-3p', 'hsa-mir-548x:hsa-miR-548ar-5p', 'hsa-mir-548k:hsa-miR-548av-5p', 'hsa-mir-199a-1:hsa-miR-199a-3p', 'hsa-mir-30e:hsa-miR-30e-3p', 'hsa-mir-4508:hsa-miR-4508', 'hsa-mir-548aj-2:hsa-miR-548x-5p', 'hsa-mir-371b:hsa-miR-371b-5p', 'hsa-mir-5001:hsa-miR-5001-3p', 'hsa-mir-16-2:hsa-miR-16-2-3p', 'hsa-mir-128-2:hsa-miR-128', 'hsa-mir-486:hsa-miR-486-3p', 'hsa-mir-4482-1:hsa-miR-4482-3p', 'hsa-mir-550a-1:hsa-miR-550a-5p', 'hsa-mir-199a-2:hsa-miR-199b-3p', 'hsa-mir-144:hsa-miR-144-5p', 'hsa-let-7f-2:hsa-let-7f-5p', 'hsa-mir-126:hsa-miR-126-5p', 'hsa-mir-191:hsa-miR-191-3p', 'hsa-mir-10a:hsa-miR-10a-5p', 'hsa-mir-98:hsa-miR-98', 'hsa-mir-548x:hsa-miR-548x-5p', 'hsa-mir-363:hsa-miR-363-3p', 'hsa-mir-548h-1:hsa-miR-548h-5p', 'hsa-mir-223:hsa-miR-223-3p', 'hsa-mir-5690:hsa-miR-5690', 'hsa-mir-199b:hsa-miR-199b-3p', 'hsa-mir-3200:hsa-miR-3200-3p', 'hsa-mir-424:hsa-miR-424-3p', 'hsa-mir-644b:hsa-miR-644b-3p', 'hsa-mir-548h-5:hsa-miR-548h-5p', 'hsa-mir-18a:hsa-miR-18a-5p', 'hsa-mir-548g:hsa-miR-548x-5p', 'hsa-mir-548g:hsa-miR-548g-5p', 'hsa-mir-21:hsa-miR-21-5p', 'hsa-mir-99b:hsa-miR-99b-5p', 'hsa-mir-25:hsa-miR-25-3p', 'hsa-mir-937:hsa-miR-937', 'hsa-mir-1180:hsa-miR-1180', 'hsa-mir-30c-1:hsa-miR-30c-5p', 'hsa-let-7a-1:hsa-let-7a-5p', 'hsa-mir-660:hsa-miR-660-5p', 'hsa-mir-421:hsa-miR-421', 'hsa-mir-374a:hsa-miR-374a-5p', 'hsa-mir-151a:hsa-miR-151a-5p', 'hsa-mir-548x:hsa-miR-548aj-5p', 'hsa-mir-101-2:hsa-miR-101-3p', 'hsa-mir-28:hsa-miR-28-3p', 'hsa-mir-139:hsa-miR-139-5p', 'hsa-mir-2110:hsa-miR-2110', 'hsa-let-7g:hsa-let-7g-5p', 'hsa-mir-550a-3:hsa-miR-550a-3-5p', 'hsa-mir-548aj-2:hsa-miR-548ar-5p', 'hsa-mir-144:hsa-miR-144-3p', 'hsa-mir-548e:hsa-miR-548e', 'hsa-mir-3074:hsa-miR-3074-5p', 'hsa-mir-1294:hsa-miR-1294', 'hsa-mir-19a:hsa-miR-19a-3p', 'hsa-mir-199a-1:hsa-miR-199b-3p', 'hsa-mir-17:hsa-miR-17-3p', 'hsa-mir-340:hsa-miR-340-3p', 'hsa-mir-3158-2:hsa-miR-3158-3p', 'hsa-mir-548x:hsa-miR-548g-5p', 'hsa-mir-331:hsa-miR-331-3p', 'hsa-mir-4742:hsa-miR-4742-3p', 'hsa-mir-4482-2:hsa-miR-4482-3p', 'hsa-mir-548h-2:hsa-miR-548h-5p', 'hsa-mir-324:hsa-miR-324-5p', 'hsa-mir-18b:hsa-miR-18b-5p', 'hsa-mir-659:hsa-miR-659-5p', 'hsa-mir-532:hsa-miR-532-5p', 'hsa-mir-3158-1:hsa-miR-3158-3p', 'hsa-mir-671:hsa-miR-671-3p', 'hsa-let-7d:hsa-let-7d-3p', 'hsa-mir-29b-2:hsa-miR-29b-3p', 'hsa-mir-548h-4:hsa-miR-548h-5p', 'hsa-mir-3615:hsa-miR-3615', 'hsa-mir-4746:hsa-miR-4746-5p', 'hsa-mir-210:hsa-miR-210', 'hsa-mir-3127:hsa-miR-3127-3p', 'hsa-mir-130b:hsa-miR-130b-3p', 'hsa-mir-550a-2:hsa-miR-550a-5p', 'hsa-let-7a-2:hsa-let-7a-5p', 'hsa-mir-148a:hsa-miR-148a-3p', 'hsa-mir-190a:hsa-miR-190a', 'hsa-mir-1304:hsa-miR-1304-3p', 'hsa-mir-4792:hsa-miR-4792', 'hsa-mir-106b:hsa-miR-106b-3p', 'hsa-mir-29b-1:hsa-miR-29b-3p', 'hsa-mir-30b:hsa-miR-30b-5p', 'hsa-mir-378d-2:hsa-miR-378d', 'hsa-mir-20a:hsa-miR-20a-5p', 'hsa-mir-323b:hsa-miR-323b-3p', 'hsa-mir-548ar:hsa-miR-548ar-5p', 'hsa-mir-378i:hsa-miR-378

Univariate selection to identify most related miRNAs

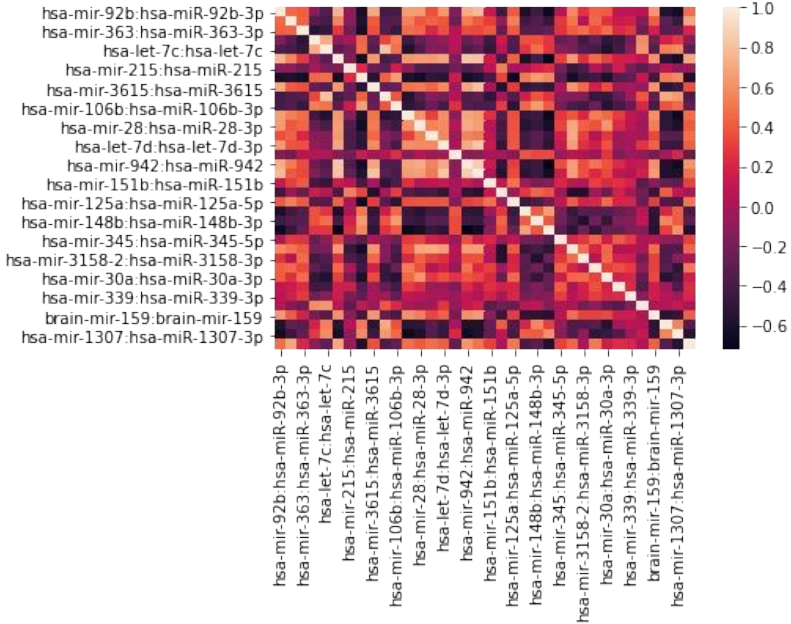
- 50 miRNAs



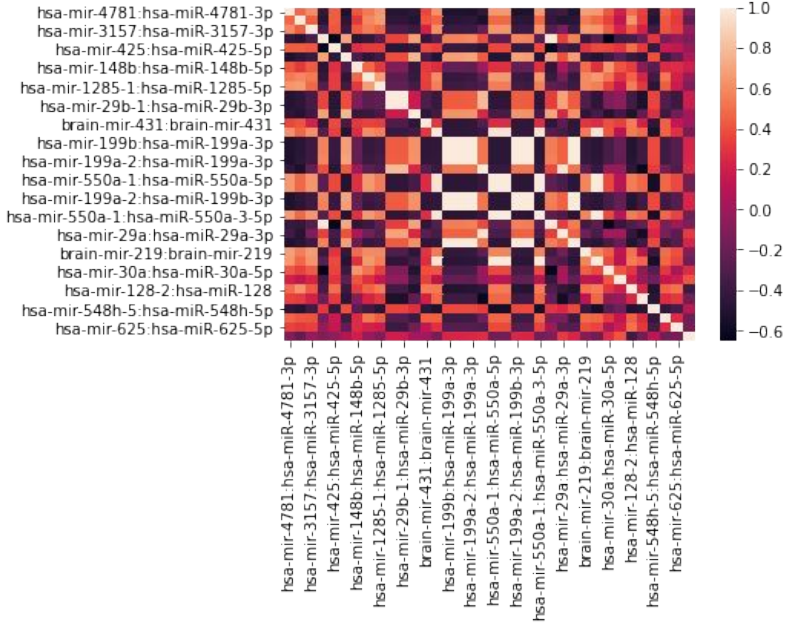
- Visualization of the overlap between set of miRNAs selected from PCA and Random Forest

Correlation heatmaps

PCA

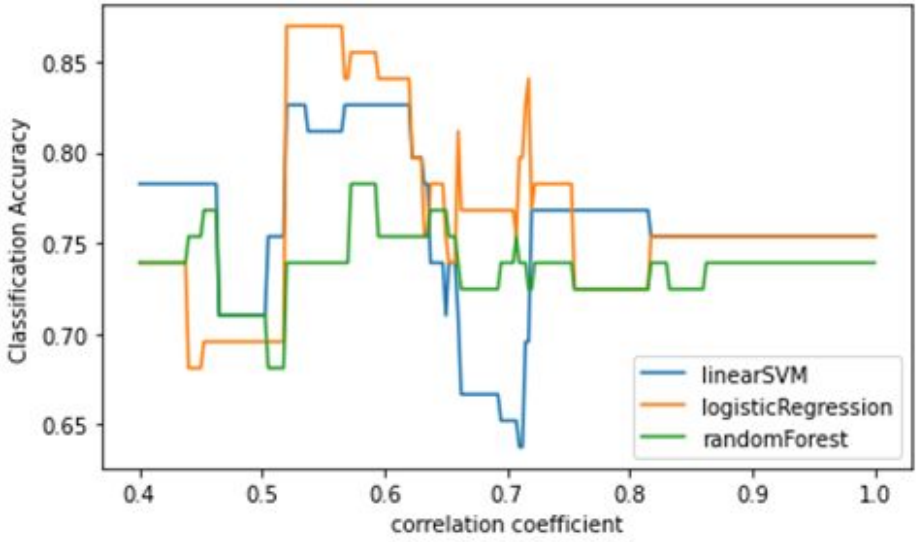


Random Forest

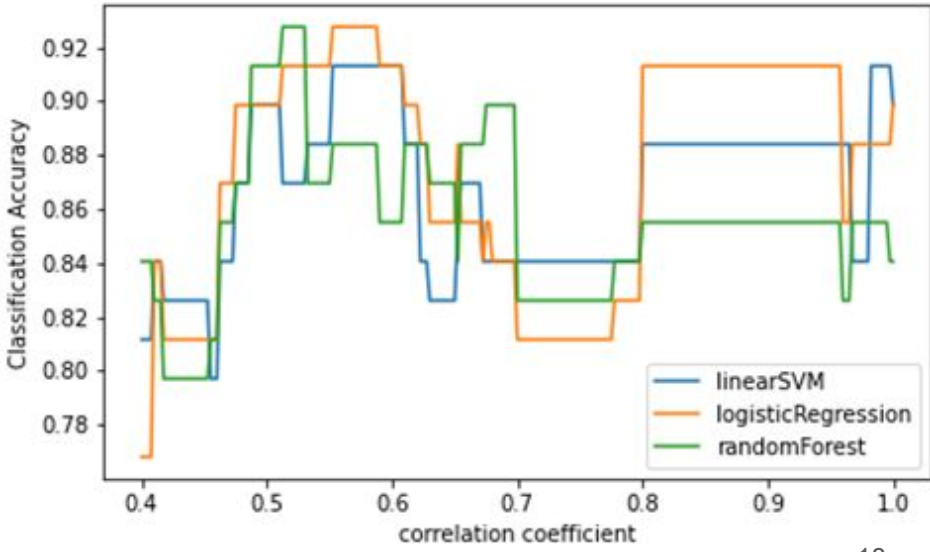


- Accuracy with correlation coefficients

PCA



Random Forest



Selected miRNAs using correlations

- 11 miRNAs

Correlation
hsa-mir-4781:hsa-miR-4781-3p
brain-mir-112:brain-mir-112
hsa-let-7a-3:hsa-let-7a-5p
hsa-mir-148b:hsa-miR-148b-5p
hsa-mir-29b-2:hsa-miR-29b-3p
brain-mir-431:brain-mir-431
hsa-mir-378a:hsa-miR-378a-5p
hsa-mir-548h-5:hsa-miR-548h-5p
hsa-mir-3909:hsa-miR-3909
hsa-mir-625:hsa-miR-625-5p
hsa-mir-24-1:hsa-miR-24-3p

Selected miRNAs using Overlap

- 14 miRNAs

overlap
hsa-mir-186:hsa-miR-186-5p
hsa-mir-144:hsa-miR-144-3p
hsa-mir-151a:hsa-miR-151a-3p
hsa-mir-99b:hsa-miR-99b-5p
hsa-mir-98:hsa-miR-98
hsa-mir-148a:hsa-miR-148a-3p
hsa-let-7g:hsa-let-7g-5p
hsa-let-7f-2:hsa-let-7f-5p
hsa-let-7a-1:hsa-let-7a-5p
hsa-mir-30d:hsa-miR-30d-5p
hsa-mir-15a:hsa-miR-15a-5p
hsa-mir-589:hsa-miR-589-5p
hsa-mir-144:hsa-miR-144-5p
hsa-let-7f-1:hsa-let-7f-5p

Accuracies

	Both PCA & RF (overlap)	**Overlap + Correlation**
Linear SVM	80.95	85.71
Logistic Regression	95.24	95.24
Random Forest	90.48	95.24

Validation of the results

- HMDD v3.2 was used with the final set of 25 miRNAs
  - 10 miRNAs were identified
    - 2 from less correlated set
    - 8 from overlapped data from both PCA & RF

```
corr_validate
```

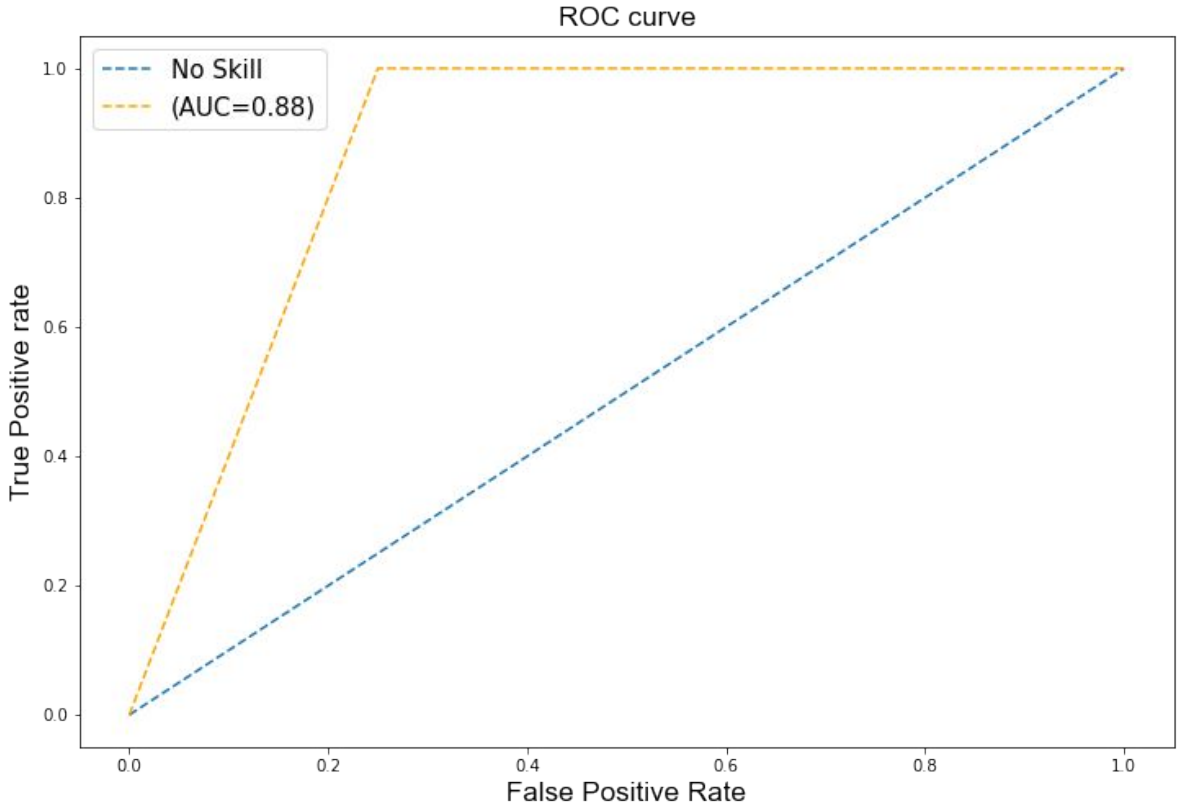
```
['hsa-mir-148b', 'hsa-mir-29b-2']
```

```
overlap_validate
```

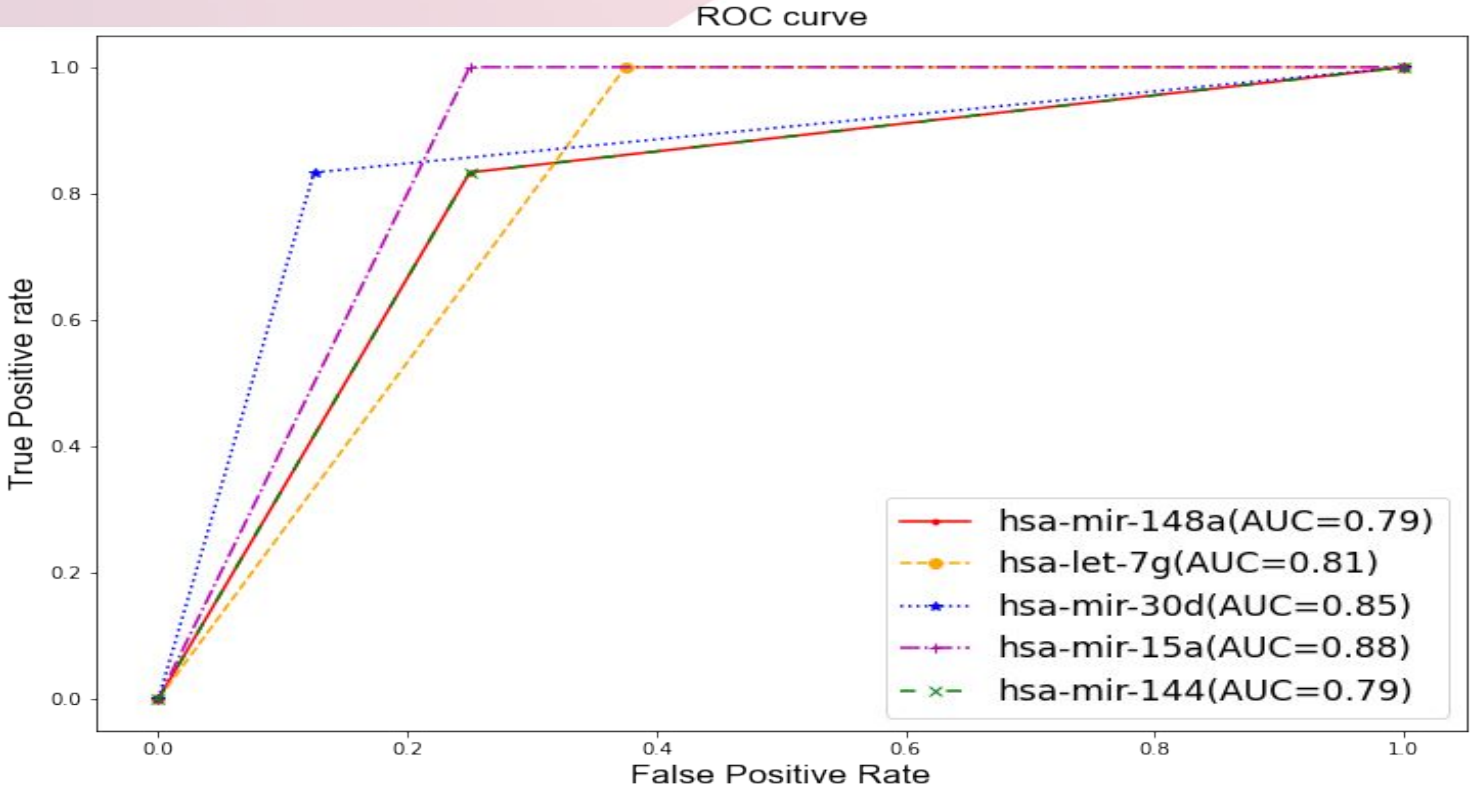
```
['hsa-mir-186',
 'hsa-mir-144',
 'hsa-mir-151a',
 'hsa-mir-98',
 'hsa-mir-148a',
 'hsa-let-7g',
 'hsa-mir-15a',
 'hsa-mir-144']
```

Evaluation of the results

- ROC curve for obtained 25 miRNAs



- ROC curves for 5 miRNAs with the highest AUC values



Evaluation of methodology

- Analyzed another dataset using the same methodology
- Used the data set under the accession number GSE147218 in NCBI database
- Obtained results were,
  - Overlap

```
['hsa-miR-193b-5p']
```

  - Correlation Coefficient

hsa-miR-1287-3p
hsa-let-7f-1-3p
hsa-miR-1283
hsa-miR-4703-3p
hsa-miR-155-3p
hsa-miR-1254
hsa-miR-3131
hsa-miR-26b-5p

Significance of our methodology

- Both statistical and machine learning approaches were used
- Feature selection was done using more than one method
  - Random Forest
  - PCA
  - Correlation coefficient values
- Higher accuracy compared to the study done using the same data set
  - Study by Leidinger et al. : 93.3%
  - Our Study : 95.24%

Deliverables Addressed in Phase 1

Milestone 01 : Background study

: Dataset selection

- Data set from NCBI database under the accession number GSE46579

Milestone 02: Preprocessing dataset (Galaxy tool)

: Data visualization and normalization

Milestone 03: Statistical analysis

- Using significance values, fold change and AUC values

Deliverables Addressed in Phase 2

- **Milestone 01** : Identify Biomarker miRNA set
  - Using,
    - PCA
    - Random Forests
    - Correlation coefficient
- **Milestone 02** : Result Validation
  - Using HMDD v3.2
- **Milestone 03** : Analyzing another data set using the same methodology
  - Use GSE147218 dataset

: Evaluating the obtained results

Phase 1 Gantt Chart

TASK	Week 1		Week2		Week3		Week4		Week5		Week6		Week7		Week8		Week9		Week10	
	6-Jul	13-Jul	13-Jul	20-Jul	20-Jul	27-Jul	27-Jul	3-Aug	3-Aug	10-Aug	28-Sep	5-Oct	5-Oct	12-Oct	12-Oct	19-Oct	19-Oct	26-Oct	26-Oct	30-Oct
Milestone 1 :																				
Task 1.1 : Literature review																				
Task 1.2 : Deciding on a project scope																				
Task 1.3 : Come up with a methodology																				
Task 1.4 : Dataset selection																				
Task 1.5 : Review paper writing																				
Milestone 2 :																				
Task 2.1 : Data preprocessing																				
Task 2.2 : Data Visualization & Quantile normalization																				
Milestone 3 :																				
Task 3.1 : Analysis of different statistical tests for p values																				
Task 3.2 : Calculation of p values & identifying the most significant features																				
Task 3.3 : Identifying the most dysregulated miRNAs																				
Task 3.4 : Machine learning approaches																				
Task 3.5 : Final report writing																				

Phase 2 Gantt Chart

TASK	Week 1 & 2		Week 3 & 4		Week 5 & 6		Week 7 & 8		Week 9 & 10		Week 11 & 12		Week 13 & 14		Week 15	
	9-Nov	20-Nov	23-Nov	4-Dec	7-Dec	18-Dec	4-Jan	15-Jan	18-Jan	29-Jan	1-Feb	12-Feb	15-Feb	26-Feb	1-Mar	5-Mar
Deciding milestones for sem 8																
Milestone 1 :																
Task 1.1 : Feature selection using PCA & RF																
Task 1.2 : Comparing with previous studies																
Task 1.3 : Feature selection using correlations																
Milestone 2 :																
Task 2.1 : Searching validation datasets																
Task 2.2 : Validation of the result																
Milestone 3 :																
Task 3.1 : Analysis of GSE147218 datset																
Task 3.2 : Compare the result with previous studies																
Task 3.3 : Result evaluation using ROC curve analysis																
Report writing																
Research article writing																
Creating web & github																

Future plan

- Developing a web based tool
  - Visualizing the distributions of datasets, preprocessing, statistical analysis and classification effortlessly.

Thank you

Q & A