An Integrative Bioinformatics Approach in microRNA Data Analytics of Alzheimer's Disease

Jill Ann Chia[1], Mei Sze Tan[1], Siow-Wee Chang[1,2*]

[1]Bioinformatics Programme, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia.

[2]Centre of Research in System Biology, Structural, Bioinformatics and Human Digital Imaging (CRYSTAL), Universiti Malaya, Kuala Lumpur, Malaysia

*corresponding author

Siow-Wee Chang

siowwee@um.edu.my / changsiowwee@gmail.com

Abstract

Alzheimer's Disease (AD) is the most common type of dementia clinically recognised by cognitive function impairment. Lately, the blood-based biomarkers relating to AD are intensively investigated due to the minimum invasiveness and relatively low cost in the collection of blood samples compared to the cerebrospinal fluid in the brain. With regard to that, the study of the deregulation of microRNA (miRNA) levels in the blood of AD patients is on the rise too. In this study, data analysis was performed on the miRNA expression profiling dataset using an integrative bioinformatics approach. kNN imputation and quantile normalization were carried out as the data pre-processing step to remove outliers and reduce bias in the dataset. Differential expression analysis was performed to identify 10 significant dysregulated miRNAs using a cut-off at adjusted-p-value <0.05 and an absolute fold change of 1.6. Subsequently, 16 pathways were determined to be involved by the selected 10 miRNA signatures, and 7 genes were predicted as the common target genes, which are Cdc42, VEGFA, NTRK3, ESR1, SH3GL2, COX-2 and E2F1. The roles of these target genes in AD were proven by literature studies. Expansion of the current work on a bigger scale of data analysis is needed to further validate and understand the mechanism of miRNAs in AD development.

Keywords: Alzheimer's Disease, microRNA, differential expression analysis, data analytics, bioinformatics

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that is clinically recognised by the impairment of cognition (Dementia, 2007; Mucke, 2009). It is the most common type of dementia, a general term used to describe unusual changes in the brain. The prevalence of AD is often related to ageing, with a higher risk for senior individuals. The advancement in health care and medical technology has brought an increase in life expectancy worldwide resulting in the expansion of an ageing population globally. According to World Alzheimer Report 2015, it is estimated that 74.7 million and 131.5 million people will be living with dementia in the years 2030 and 2050 (Prince et al., 2015). It is reported that three persons will develop dementia every three seconds.

MicroRNAs (miRNAs) play a post-transcriptional role in the regulation of gene expression. Significant miRNAs are able to distinguish between AD and healthy controls, giving the potential to support AD diagnosis (Keller et al., 2016). However, the collection of cerebrospinal fluid is an invasive treatment with potential side effects. The blood gene expression data are useful in predicting the AD classification and are shown to be consistent with the observations from brain tissue-based studies (Lee & Lee, 2020).

Several studies have relate miRNA to the pathogenesis of neurodegenerative disease and reported that the miRNAs are essential for neuronal function and survival (Delay et al., 2012). For example, miR-9 is one of the most frequently altered miRNAs, which is downregulated in AD brains. Circulating miRNAs are among the most promising candidates for easily accessible and non-invasive biomarkers for AD diagnosis (Sturmberg et al., 2015). Over the years, it was found that miR-101, miR-20a, and miR-17 play important roles in AD pathogenesis (Kumar & Reddy, 2016; Liu et al., 2022). On the other hand, the level of miR-107 was found to be correlated with AD (Fransquet & Ryan, 2018; Kumar & Reddy, 2016; Liu et al., 2022; Takousis et al., 2019). Besides, the suppression of miR-203 was found to subsequently alleviating the cognitive function (Liu et al., 2022).

Technologies such as qRT-PCR, microarray and next generation sequencing are applied in the miRNA expression profiling (Roden et al., 2005). Differential expression analysis is performed on the expression profiles to study the differences of expression levels of miRNAs in the specific condition (Soneson & Delorenzi, 2013). The differential expression analysis can be done by statistical analysis or machine learning approaches.

Pathway analysis is then carried out to analyse and identify the relationship between the groups of gene as well as the biological role of the candidate gene.

In general, this study involves the data pre-processing, differential expression analysis and pathway analysis on the miRNA expression profiles. The raw read counts of miRNA expression were processed by *k*NN imputation and normalised using quantile normalization method. The differential expression analysis was performed using Welch's t-test, in order to identify the differentially expressed miRNAs based on statistical and fold change cut-offs. Next, pathway analysis was performed in this study using KEGG and GO analysis. By comparing two different resources on the pathway analysis, the common significant pathways and target genes which are related to AD were identified. The identified target genes may serve as potential biomarkers which could be beneficial in therapeutic approaches in AD.

Methods and Materials

Figure 1 illustrates the workflow proposed in this study. miRNA dataset was collected and followed by data pre-processing such as imputation and normalization. Next, differential expression analysis was performed to identify the differentially expressed miRNAs which were used in the pathway analysis. Finally, significant pathways and target genes which are related to AD were identified.

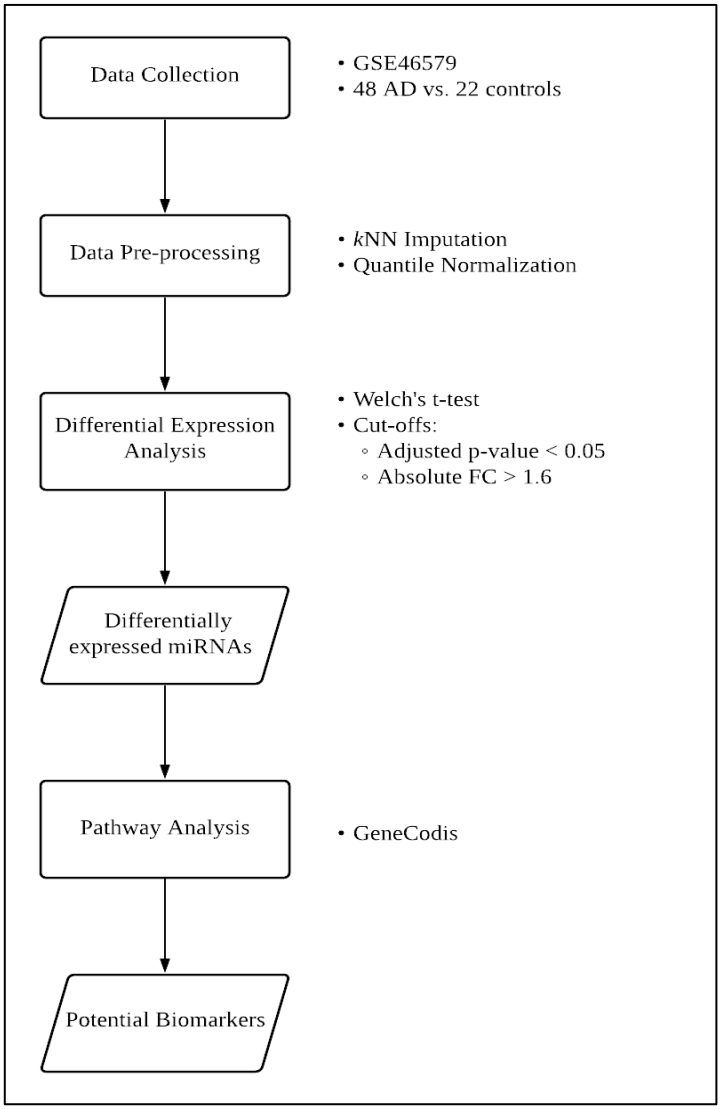

Figure 1: Proposed workflow

Dataset

The dataset used in the analysis was obtained from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) which is an open-source database repository that stores the array- and sequence-based gene expression data. The dataset used is with the reference series number GSE46579. The platform used to sequence the samples is GPL11154 Illumina HiSeq 2000. The dataset consists of 48 AD samples and 22 healthy controls, with 140 unique mature microRNAs. It was collected from blood samples by Next-Generation Sequencing (NGS).

Software and tools

All the pre-processing steps and differential expression analysis were conducted using Python programming language. Several Python libraries were used for the data pre-processing such as Pandas and Scikit-learn. AnnData and diffxpy libraries were used to

carry out the differential expression analysis. GENECODIS (https://genecodis.genyo.es/) was used in the pathway analysis as it searches for annotations that frequently co-occur in a set of genes from different sources such as KEGG pathways and Gene Ontology terms, and rank them by statistical significance (Carmona-Saez et al., 2007).

Data Pre-processing

The identifier for each miRNA is based on the database miRbase (Leidinger et al., 2013) in the raw miRNA dataset. The identifiers in that database are in the form of hsa-mir-121, while the first three letters represent the organism. If there are distinct precursor sequences and genomic loci that express identical mature sequences, those miRNAs will get the names in the form of hsa-mir-121-1 and hsa-mir-121-2. The mature miRNA is signified by miR-121 while mir-121 refers to the miRNA gene and the predicted stem-loop portion of the primary transcript. However, let-7 and lin-4 are exceptions to the naming scheme, which are retained for historical reasons (Griffiths-Jones et al., 2006). The dataset was first filtered out the microRNAs without following the naming scheme, such as brain-mir-192. Then, pre-processing of data was carried out.

First, *k*-nearest neighbour (*k*NN) was used to impute the missing values in the dataset, with k=5. It estimates k nearest group of miRNAs that are similar to the missing target miRNA, then average those miRNAs to impute the missing value of the target gene. Next, in the normalization step, quantile normalization method was applied to the dataset. Quantile normalization was initially developed for gene expression microarrays but nowadays it can be applied in various data types including RNA-Sequencing (Cloonan et al., 2008; Garmire et al., 2012). Quantile normalization is a global transformation method by assuming the statistical distribution for each sample is the same. It takes the average distribution which is obtained from the mean of each quantile across samples as the reference, and forces the observed distributions to be identical.

Differential Expression Analysis

Before the differential expression analysis is conducted, the normalised data were inputted as an annotated data matrix by using the AnnData library as required for differential expression analysis. The data was annotated into two groups, which are AD and controls. In this study, the diffxpy library was used to conduct the differential expression analysis. Various statistical tests are provided from the library and Welch's t-test was used in the study.

Welch's t-test is used when the variances of the two groups are not identical (Yaari et al., 2013). To model the difference in mean expression for miRNA *i* between two groups, treatments (T) and controls (C), we define:

$$t_i = \frac{\bar{E}_i^T - \bar{E}_i^C}{\sqrt{\frac{(s_i^T)^2}{N^T} + \frac{(s_i^C)^2}{N^C}}} \quad (1)$$

where $\bar{E}$ is the mean expression value of the miRNA, *s* is the standard deviation for the respective group on the miRNA, *N* is the total number of the samples that belong to the particular group.

Fold-change (FC) is an essential threshold that is used to identify the differentially expressed miRNAs. In this study, the differentially expressed miRNAs are selected by using the cut-offs of adjusted-*p* value < 0.05 and absolute arbitrary FC > 1.6, which is equivalent to $\log_2$FC > 0.678.

Pathway Analysis

GENECODIS was used for the pathway analysis. Two types of pathway annotations are selected, which are KEGG pathways and Gene Ontology Biological Process (GOBP). Those pathways that have an adjusted-*p* value of less than 0.05 from the hypergeometric test is considerably significant. From the filtered significant pathways, the pathways from KEGG and GO were compared to determine the common significant pathways. Lastly, the significant target genes can be identified from the common significant pathways.

Results

The differentially expressed miRNAs were selected as shown in Table 1 by using the cut-offs of adjusted-*p* value < 0.05 and absolute arbitrary FC > 1.6 (which is equivalent to $\log_2$FC > 0.678). A total of 11 differentially expressed miRNAs were selected.

Table 1: Significant miRNAs identified from the Welch's t-test with cut-offs of adjusted-p value < 0.05 and absolute FC > 1.6 (or log2FC > 0.678)

Precursor	Mature Sequence	pval	qval	log_2FC
hsa-mir-378e	hsa-miR-378e	0.002631	0.028485	-0.82127
hsa-mir-4781	hsa-miR-4781-3p	2.67E-07	6.02E-05	-0.88574
hsa-mir-5001	hsa-miR-5001-3p	7.69E-05	0.005796	-0.74152
hsa-mir-378b	hsa-miR-378b	0.000679	0.013395	-0.91671
hsa-mir-330	hsa-miR-330-3p	0.002401	0.02838	-0.68803
hsa-mir-3127	hsa-miR-3127-3p	6.23E-06	0.000939	-0.98285
hsa-mir-5701-1*	hsa-miR-5701	0.00354	0.034788	-0.68454
hsa-mir-5701-2*	hsa-miR-5701	0.00354	0.034788	-0.68454
hsa-mir-4659a	hsa-miR-4659a-3p	0.001111	0.018591	-0.71012
hsa-mir-26b	hsa-miR-26b-3p	2.11E-05	0.002385	-0.70693
hsa-mir-1468	hsa-miR-1468	3.88E-08	1.75E-05	-0.9363

Notes: pval – *p*-value; qval – adjusted-*p* value

* hsa-miR-5701-1 and hsa-miR-5701-2 are different precursors that produce the same mature miRNA sequence.

From the 11 identified differentially miRNAs, there are two identical mature miRNA sequences, hsa-miR-5701-1 and hsa-miR-5701-2, which are produced from different precursor. Therefore, only 10 unique miRNA mature sequence were used for the pathway analysis.

The identifiers of the 10 unique miRNAs were used as the input to the GENECODIS. The functional enrichment analysis was performed based on the KEGG pathways and GO BP. The significant pathways were identified by an adjusted-*p* value < 0.05 from the hypergeometric test. Figure 2 and Figure 3 show the top 10 significant pathways for KEGG and GO BP generated from GENECODIS respectively.

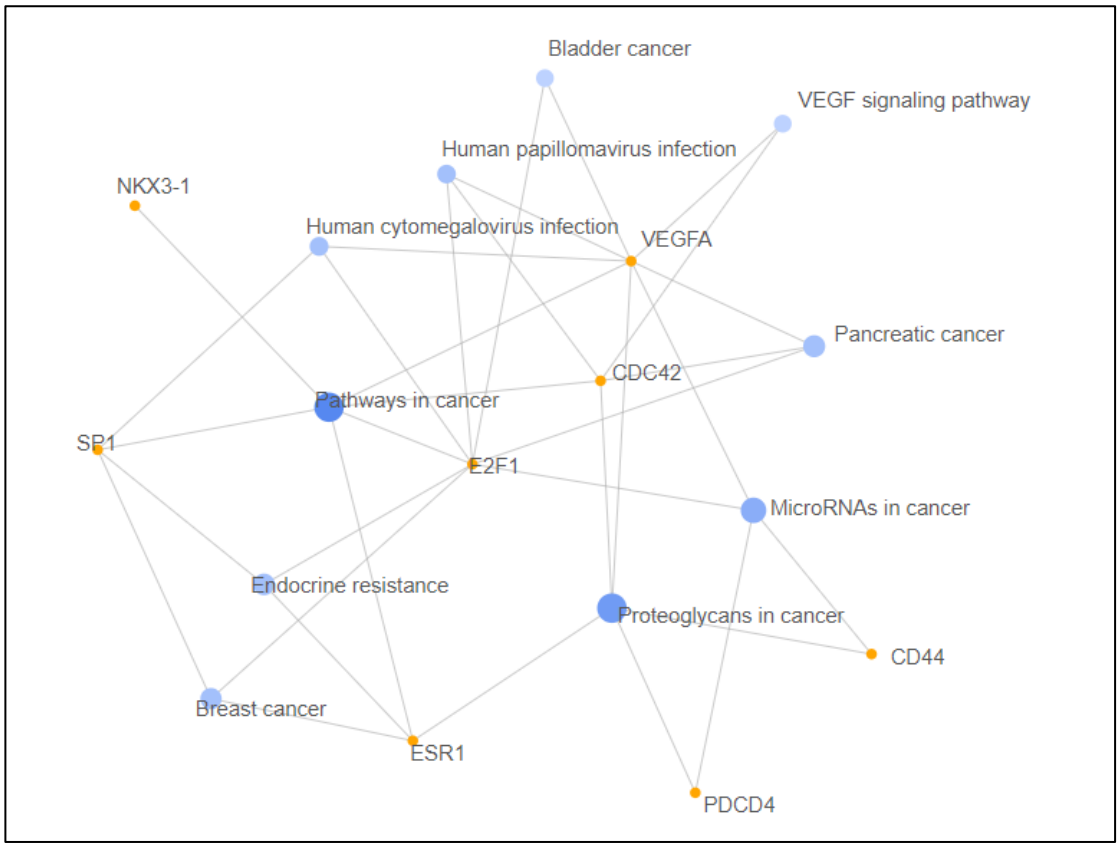

Figure 2: Top ten significant pathways for KEGG from GENECODIS

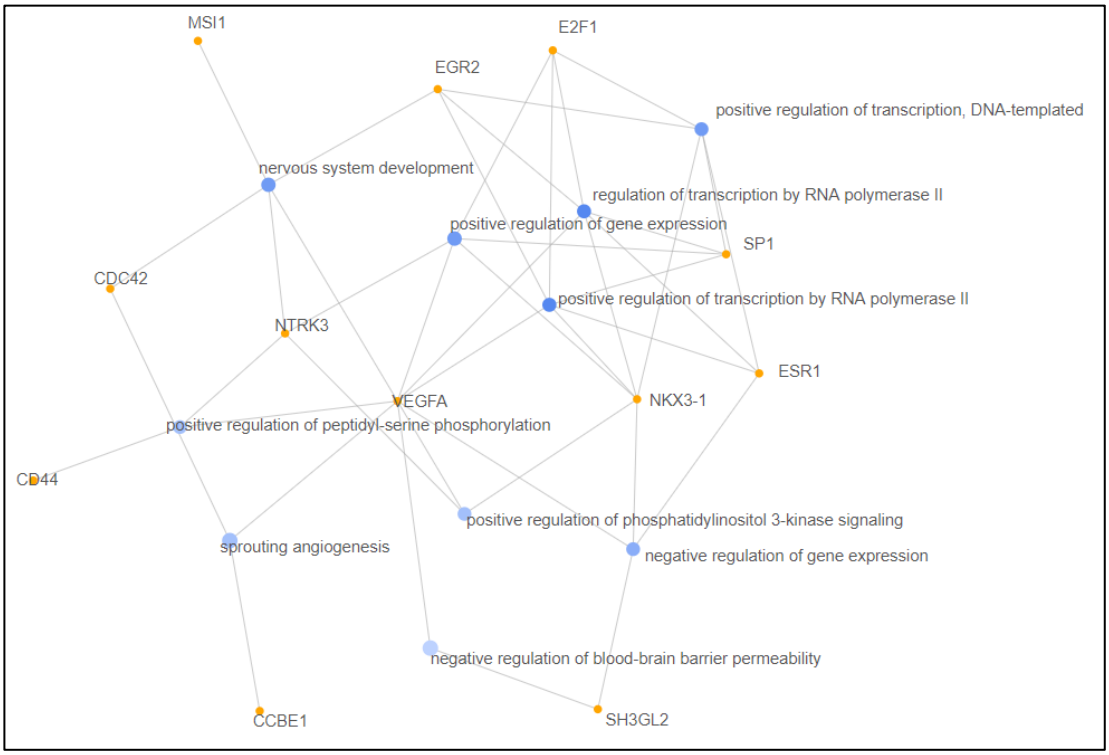

Figure 3: Top ten significant pathways for GOBP from GENECODIS

Two result files, each from KEGG and GOBP were downloaded in .tsv format and converted to .xls format. There are 78 significant pathways found in KEGG and 369

significant biological processes are identified in GOBP. Each significant pathway from KEGG as well as from GOBP were filtered by masking the cancer-related terms and excluding those without any supported findings that suggested it is related to AD. After the filtration step, there are 46 pathways and 162 pathways left for KEGG and GOBP respectively.

Next, the filtered pathways from KEGG and GOBP were compared to determine the common significant pathways. There are 16 common significant pathways identified from both KEGG and GOBP, which are AGE-RAGE signaling pathway, VEGF signaling pathway, neurotrophin signaling pathway, estrogen signaling pathway, endocytosis, focal adhension, MAPK signaling pathway, oxidative phosphorylation, adherens junction, glial cell proliferation, Fc gamma R-mediated phagocytosis, axon guidance, cell cycle, cellular senescence, cytokine production involved in inflammatory response and HIF-1 signaling pathway.

Next, the target genes for each pathway were identified from the common significant pathways that were determined earlier. A total of 11 target genes (9 target genes from KEGG and 9 target genes from GOBP) were identified from the sixteen common significant pathways listed above. The results of target genes identified are listed in Table 2. From these 11 target genes, 7 common target genes were identified as shown in the Venn diagram (Figure 4).

Table 2: Target genes of pathways from KEGG and GOBP

Terms	Target Genes	
	KEGG	GOBP
AGE-RAGE signaling pathway	CDC42, VEGFA	CDC42, VEGFA, NTRK3
VEGF signaling pathway	CDC42, VEGFA	CDC42, VEGFA
Neurotrophin signaling pathway	CDC42, NTRK3	NTRK3
Estrogen signaling pathway	SP1, ESR1	ESR1
Endocytosis	CDC42, SH3GL2	CDC42, SH3GL2
Focal adhension	CDC42, VEGFA	VEGFA
MAPK signaling pathway	CDC42, VEGFA	CDC42, VEGFA, NTRK3
Oxidative phosphorylation	COX2	COX2
Adherens junction	CDC42	CDC42
Glial cell proliferation	E2F1	E2F1
Fc gamma R-mediated phagocytosis	CDC42	CDC42
Axon guidance	CDC42	VEGFA
Cell cycle	E2F1	E2F1
Cellular senescence	E2F1	PDCD4
Cytokine production involved in inflammatory response	EGR2	PDCD4
HIF-1 signaling pathway	VEGFA	VEGFA, NKX3-1, E2F1

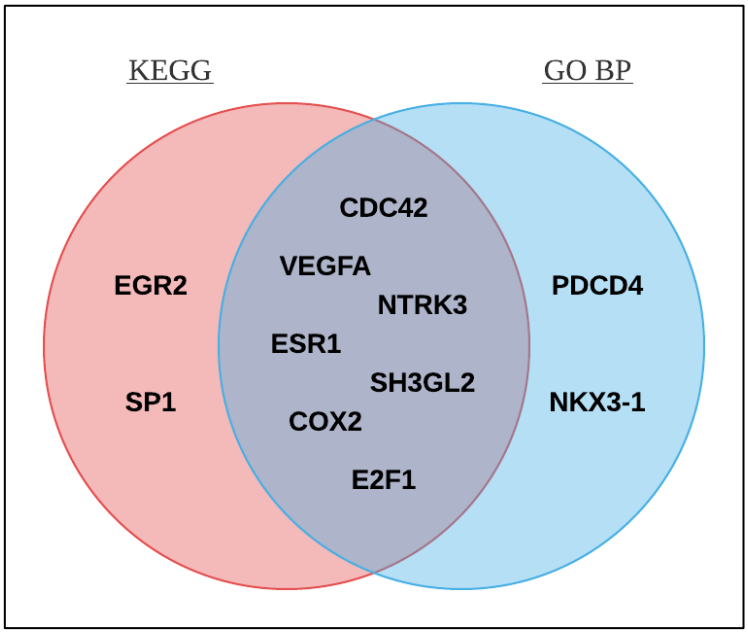

Figure 4: Venn diagram of target genes between KEGG and GOBP

Discussion

In this study, 10 differentially expressed miRNAs were identified from differential expression analysis. The significance of the 10 miRNAs identified are related to their functions in AD or other types of neurogenerative diseases. Table 3 lists the 10 identified miRNAs in with related functions.

Table 3: Significance of selected 10 miRNAs

miRNAs	Functions	References
hsa-miR-378e	Downregulated in ALS; its overexpression inhibits the glycolysis and promotes cell apoptosis which indicates its therapeutic effect in glioma.	Kovanda et al., 2018; Ding et al., 2019
hsa-miR-4781-3p	Upregulated in AD	Sproviero et al., 2021
hsa-miR-5001-3p	Upregulated in AD	Kumar & Reddy, 2016
hsa-miR-378b	Downregulated known miRNA in cerebrospinal fluid-derived exosomes	Hou et al., 2019
hsa-miR-330-3p	Exert protective effects on Aβ production, oxidative stress and mitochondrial dysfunction by targeting VAV1 via the MAPK signaling pathway	Zhou et al., 2018
hsa-miR-3127-3p	Upregulated in AD	Leidinger et al., 2013
hsa-miR-5701	Induces mitochondrial dysfunction, defect in autophagy flux in PD	Prajapati et al., 2018

hsa-miR-4659a-3p	Negatively regulate GNAQ, TMTC2 and BEND2 with multiple miRNAs in PD	Liu et al., 2019
hsa-miR-26b-3p	Deregulated early in AD brain, nearly 20 years before the onset of clinical symptoms (upregulated in brain while downregulated in blood)	Swarbrick et al., 2019
hsa-miR-1468	Upregulated in AD	Satoh et al., 2015

7 common target genes were identified from the results shown in Figure 4. The common target genes are Cdc42, VEGFA, NTRK3, ESR1, SH3GL2, COX2 and E2F1. Table 4 lists the functions related to AD for each of the target genes.

Table 4: Functions related AD in target genes

Genes	**Functions related to AD**
Cdc42	• Regulation of actin cytoskeleton dynamics and spine formation • Increased level Cdc42 in frontal cortex of AD
VEGFA	• Co-accumulated with beta-amyloid deposits in AD brains • Up- and down-regulated in brain, blood CSF of AD
NTRK3	• Activate neuronal survival pathways • Decreased NTRK3 expression found in AD, PD, Huntington's disease
ESR1	• Decrease tau hyperphosphorylation • High ESR1 expression in nucleus basalis of Meynert in AD
SH3GL2	• Increased endophilin A1 -> Increased JNK activation -> Neurons die
COX-2	• Regulate neurotoxicity • Increased COX-2 in AD brain
E2F1	• Increased immunoreactivity of E2F1 and ppRb in affected cortical brain region in AD

Conclusion

In this study, the miRNA expression profiles of Alzheimer's disease patients and healthy controls were analysed by an integrative bioinformatics data analysis approach. A total of 10 differentially expressed miRNAs were identified (hsa-miR-378e, hsa-miR-4781-3p, hsa-miR-5001-3p, hsa-miR-378b, hsa-miR-330-3p, hsa-miR-3127-3p, hsa-miR-5701, hsa-miR-4659a-3p, hsa-miR-26b-3p, hsa-miR-1468). Sixteen common significant pathways were identified

from the pathway analysis and their functions which related to AD and other neurogenerative diseases were discussed. Next, 7 common target genes were identified from the common significant pathways, including Cdc42, VEGFA, NTRK3, ESR1, SH3GL2, COX-2 and E2F1. All of them are shown to be involved in AD pathways which could be the potential biomarkers that would be beneficial towards therapeutic approaches in AD.

Acknowledgement

This work was supported in part by the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education Malaysia with the project number FRGS/1/2019/SKK06/UM/ 02/5 and UM International Collaboration Grant with the project number ST041-2022. The funders had no role in study, design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., & Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, *8*(1), 1-8.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., ... & Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, *5*(7), 613-619.

Delay, C., Mandemakers, W., & Hébert, S. S. (2012). MicroRNAs in Alzheimer's disease. *Neurobiology of disease*, *46*(2), 285-290.

Dementia, N.-S. (2007). The NICE-SCIE Guideline on supporting people with dementia and their carers in health and social care. *Leicester, London: British Psychological Society*.

Ding, C., Wu, Z., You, H., Ge, H., Zheng, S., Lin, Y., ... & Kang, D. (2019). CircNFIX promotes progression of glioma through regulating miR-378e/RPN2 axis. *Journal of Experimental & Clinical Cancer Research*, *38*(1), 1-12.

Fransquet, P. D., & Ryan, J. (2018). Micro RNA as a potential blood-based epigenetic biomarker for Alzheimer's disease. *Clinical Biochemistry*, *58*, 5-14.
Garmire, L. X., & Subramaniam, S. (2012). Evaluation of normalization methods in mammalian microRNA-Seq data. *Rna*, *18*(6), 1279-1288.

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, *34*(suppl_1), D140-D144.

Hou, X., Gong, X., Zhang, L., Li, T., Yuan, H., Xie, Y., ... & Jiang, H. (2019). Identification of a potential exosomal biomarker in spinocerebellar ataxia Type 3/Machado–Joseph disease. *Epigenomics*, *11*(9), 1037-1056.

Keller, A., Backes, C., Haas, J., Leidinger, P., Maetzler, W., Deuschle, C., ... & Stähler, C. (2016). Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimer's & Dementia*, *12*(5), 565-576.

Kovanda, A., Leonardis, L., Zidar, J., Koritnik, B., Dolenc-Groselj, L., Kovacic, S. R., ... & Rogelj, B. (2018). Differential expression of microRNAs and other small RNAs in muscle tissue of patients with ALS and healthy age-matched controls. *Scientific reports*, *8*(1), 1-15.

Kumar, S., & Reddy, P. H. (2016). Are circulating microRNAs peripheral biomarkers for Alzheimer's disease? *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1862*(9), 1617-1627.

Lee, T., & Lee, H. (2020). prediction of Alzheimer’s disease using blood gene expression data. *Scientific reports*, *10*(1), 1-13.

Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S. C., Frese, K., ... & Lang, C. J. (2013). A blood based 12-miRNA signature of Alzheimer disease patients. *Genome biology*, *14*(7), R78.

Liu, X., Erikson, C., & Brun, A. (1996). Cortical synaptic changes and gliosis in normal aging, Alzheimer's disease and frontal lobe degeneration. *Dementia and Geriatric Cognitive Disorders*, *7*(3), 128-134.

Mucke, L. (2009). Neuroscience Alzheimer's disease. In (Vol. 461, pp. 895-897): Nature Publishing Group, London, England.

Prajapati, P., Sripada, L., Singh, K., Roy, M., Bhatelia, K., Dalwadi, P., & Singh, R. (2018). Systemic analysis of miRNAs in PD stress condition: miR-5701 modulates mitochondrial–lysosomal cross talk to regulate neuronal death. *Molecular neurobiology*, *55*(6), 4689-4701.

Prince, M. J., Wimo, A., Guerchet, M. M., Ali, G. C., Wu, Y.-T., & Prina, M. (2015). World Alzheimer Report 2015-The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends.

Roden, C., Mastriano, S., Wang, N., & Lu, J. (2015). microRNA Expression Profiling: Technologies, Insights, and Prospects. *Advances in experimental medicine and biology*, *888*, 409–421.

Satoh, J. I., Kino, Y., & Niida, S. (2015). MicroRNA-Seq data analysis pipeline to identify blood biomarkers for Alzheimer's disease from public data. *Biomarker insights*, *10*, BMI-S25132.

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, *14*(1), 1-18.

Sproviero, D., Gagliardi, S., Zucca, S., Arigoni, M., Giannini, M., Garofalo, M., ... & Cereda, C. (2021). Different miRNA Profiles in Plasma Derived Small and Large Extracellular Vesicles from Patients with Neurodegenerative Diseases. *International Journal of Molecular Sciences*, *22*(5), 2737.

Sturmberg, J. P., Bennett, J. M., Picard, M., & Seely, A. J. (2015). The trajectory of life. Decreasing physiological network complexity through changing fractal patterns. *Frontiers in Physiology*, *6*, 169.

Swarbrick, S., Wragg, N., Ghosh, S., & Stolzing, A. (2019). Systematic review of miRNA as biomarkers in Alzheimer’s disease. *Molecular neurobiology*, *56*(9), 6156-6167.

Takousis, P., Sadlon, A., Schulz, J., Wohlers, I., Dobricic, V., Middleton, L., Lill, C. M., Perneczky, R., & Bertram, L. (2019). Differential expression of microRNAs in Alzheimer's disease brain, blood, and cerebrospinal fluid. *Alzheimer's & Dementia*, *15*(11), 1468-1477.

Yaari, G., Bolen, C. R., Thakar, J., & Kleinstein, S. H. (2013). Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic acids research*, *41*(18), e170-e170.

Zhou, Y., Wang, Z. F., Li, W., Hong, H., Chen, J., Tian, Y., & Liu, Z. Y. (2018). Protective effects of microRNA-330 on amyloid β-protein production, oxidative stress, and mitochondrial dysfunction in Alzheimer's disease by targeting VAV1 via the MAPK signaling pathway. *Journal of cellular biochemistry*, *119*(7), 5437-5448.