Introduction Advancement in the Diagnosis and Treatment of Alzheimer's Disease through Analysis High-throughput Sequencing Using R Packages

Fuyang Yu *

Bioengineering, School of Life Science and Engineering, Southwest Jiaotong University

* Corresponding Author Email: wsr1529n1y9r@my.swjtu.edu.cn

Abstract. Alzheimer's disease (AD), a common neurodegenerative disease, accounts for 80% in dementia, especially in the elderly people aged 65 years and above. Nowadays, with the acceleration of the population ageing, the incidence will continue to increase in the absence of new effective pathological interventional strategy, especially in developing countries. The high-throughput sequencing technologies and the integrated algorithms employed to discover more biomarkers and risk genes, such as APOE4, TREM2 and WWTR. This review mainly summarizes the new AD risk genes, marker molecules for diagnosis and targeting treatment using high-throughput sequencing technology and R language algorithm in recent years, which is conducive to the in-depth understanding of the pathogenesis and pathological progress of AD, and provides new ideas and directions for the early diagnosis, intervention and treatment of AD, eventually improves patient quality of life, reduces medical expenses and social burden.

Keywords: Alzheimer's disease; high-throughput sequencing; R language; R packages.

1. Introduction

Alzheimer's disease(AD), a common neurodegenerative disease, accounts for 80% in dementia, especially in the group of elderly people aged 65 years and above [1], affecting more than 50 million people worldwide. With the trend of the current population ageing, the incidence will continue to increase in the absence of new effective pathological interventional strategy, especially in developing countries, which is expected to reach more than 150 million by 2050. AD has become the fifth cause of the elderly death in the United States [2].

Therefore, the early and accurate diagnosis of those individuals with cognitive function abnormalities is very important. At present, the hallmark molecules of brain functional changes for AD patients include the accumulation of β amyloid outside neurons and the emergence of twisted chains of tau inside neurons, wherase these marker molecules accumulate in some cognitively normal people [2]. How to accurately distinguish between the two parts of the population become a challenge in this field. With the employment of innovative second-generation tau-PET tracer detection, it was able to detect tau pathology in the subregions of medial temporal lobe, which affected the development in the earliest stages of AD. Recently, new phosphorylated tau in cerebrospinal fluid and blood has been as a diagnostic marker using in clinic gradually [3], promoting diagnose AD more earlier.

With the development of high-throughput sequencing technologies and the integrated algorithms, more biomarker molecules and risk genes have been discovered. The data mining and detection of AD risk genes, such as *APOE4, TREM2* and *WWTR1*, are not only benefit for persons with these genes to start prevention earlier, but also contribute to promote the research on the pathogenesis of AD and select new targeted therapeutic drugs for these patients. This review mainly summarizes the new AD risk genes, marker molecules for diagnosis and targeting treatment using high-throughput sequencing technology and R language algorithm in recent years, which is conducive to the in-depth understanding of the pathogenesis, and provides new ideas and directions for the early diagnosis, intervention and treatment for AD patients.

2. Application of the R packages to analysis of high-throughput sequencing data

The techniques of high-throughput sequencing are divided into second-generation and third-generation sequencing. The aim of second-generation sequencing, developing in the early 21st century, is to resolve the drawback of detection only one single sequence at a time by the first-generation sequencing. Innovated nanotechniques allow for massively parallel sequencing of single DNA molecules with millions of fluxes, which greatly improves time efficiency and reduces labor costs. However, the gene length of the second generation sequencing, also known as next-generation sequencing (NGS), is limited to 250-800bp, thus resulting high ratio of sequencing errors for long gene sequences through overlapping method. The third generation sequencing technology arises with reading 10 kb length and improving accuracy of gene sequencing according to the progress of software and mathematical analysis [4]. In order to understand the pathological mechanisms and risk genes of diseases, sequencing technologies, such as RNA-seq and whole-exome sequencing in NGS, combined with other imaging and biochemical methods to promote research in microbial genetics, cancer genomics, and neurology, and help clinicians undestand diseases at the molecular level and track the site of variant gene that caused diseases [5].

R language, as a programming language, mainly used for statistical analysis and related graphics drawing, which was developed from the S language specialized for statistical computing. Nowadays, it is widely used in the data analysis of biostatistics, bioinformatics, omics and other disciplines. R packages can be set more parameters flexibly and have greater freedom to analyse customized data than SPSS. Compared with bioinformatics professional software such as SAMtools, R has wider applications and lower learning barriers. R can provide data structures such as data frame and matrix to manage a large amount of data, especially for analysing integrated data obtained from high-throughput sequencing. Importantly, R language supports parallel computing, which greatly improves the efficiency of data processing. For high-throughput sequencing data, R language of many installation packages such as Seurat, ClusterProfiler powerful, can do deep data mining and quality control, cell clustering, and difference expression genes (DEGs) analysis [6]. It's easy for researchers to find new target molecules and interaction pathway in AD patients through enrichment analysis related pathogenic risk genes .

3. Contribution to Diagnosis and Treatment of AD through Analyzing High-throughput Sequencing Data

3.1. Discovered Novel Pathogenesis in AD from High-throughput Sequencing Data

To investigate the effect of neurotoxic astrocytes on AD disease progression and find new molecular targets, YuW et al screened three AD scRNA-seq datasets from the PubMed database to integrate the sequencing datasets using the Harmony package. The R package Seurat "FindAllMarkers" divided whole cells and astrocytes into seven clusters and nine subclusters. The cluster 3 was remarkable importance in the following research using Seurat "AddModuleScore" calculating signature scores. The involved biological process of neurotoxic astrocytes about AD were revealed using GO enrichment and KEGG pathway analysis. DEGs of two isolated bulk RNA-sequencing datasets related to AD were identified by limma in R and 12 key genes were displayed, finding the novel markers of neurotoxic astrocytes. Similarly, the Spearman correlation analysis between the 12 key genes and 5 representative clinical pathologic features of AD from another independent AD snRNA-seq dataset, 6 key genes, such as *WWTR1*, *PHYHD1*, *ZFP36L1*, *AEBP1*, *DST* and *RASL12*, were significantly correlated with clinical severity of AD through influencing pathways of humoral immune response, complement activation, and cytokine metabolic processes [7]. This study indicated that these six key genes can affect immune regulation and cytokine signaling pathways in astrocytes, leading to increased neurotoxicity and emergence of pro-inflammatory phenotypes in A1-like astrocytes, facilitating an advanced understanding of the pathogenesis of AD.

In particular, the recent development of high-throughput sequencing technologies has provided more evidence for the association of *APOE4* risk genes with AD pathogenesis. The snRNA-seq was performed and the data were analyzed from dorsolateral frontal cortex cells of AD patients of different ethnic groups using Seurat 3.1 pipeline. Data quality control was expanded, then clustering analysis was done using 0.4 resolution as well as UMAP plots. The expression of *APOE4* showed significantly reduced in African native origin (ALA) compared to European native origin (ELA), which is responsible for the lower risk of AD. The following researchers explored the differential expression of *APOE4* between ELA and ALA. Celis et al. performed and analyzed single-nuclei ATAC-seq of frontal cortex cells of ELA and found that increased *APOE4* expression of ELA was associated with higher promoter chromatin accessibility. Rajabli et al. conducted whole genome sequencing of different individuals from Africa-American Ibadan and Puerto Rican, then found the A allele of rs10423769 significantly reduces the risk of AD in *APOE4* homozygotes in African ancestry using logistic regression model. Analysis of high-throughput sequencing data of African, European ancestry using R language and other mathematical models revealed the A allele of rs10423769 on *APOE4* regulated its promoter to inhibit AD, which further explained the association of the *APOE4* expression and the different AD incidence in African and European descent [8].

3.2. Developed Novel Diagnostic Methods in AD from High-throughput Sequencing Data

Developed novel diagnostic methods through analysis of high-throughput sequencing data is another application in AD. For instance, the analysis of data using R packages displayed more evidence about the TREM2 as a candidate molecular in AD diagnostic. The data of SnRNA-seq were quality control using Seurat and screened cells and cluster by FindClusters. Then, variability in TREM2 expression under different genotype conditions was analyzed through Seurat's MAST algorithm. Subsequent proteomics and phosphoproteomics manipulation and gene enrichment analysis of the resulting data demonstrated that disease-associated microglia (DAM) activation was dependent on TREM2 and that variation in the microglial receptor TREM2 with R62H and R47H increased AD risk. After the differential analysis of SnRNA-seq data in R language, other software such as GraphPad Prism (v.8) is used to verify the statistical results, which makes the information of high-throughput sequencing data mining more reliable [9]. Analyzing SnRNA-seq data can find DAM that affects the progression of AD pathology at the cellular level, and further evaluate the gene variants of TREM2 on the probability of AD disease in different types of TREM2 population. Importantly, as a receptor for microglia, TREM2 can serve as a new target molecule to block the development of AD.

In addition, innovative scRNA-seq analysis was performed on peripheral blood cells from AD patients to discover novel blood biomarkers. Unlike previous analyses of patients' post-mortem tissue, peripheral blood cells are more accessible and have more stable gene expression, and extensive scRNA-seq data can be obtained in a short time. Furthermore, it is possiblity to produce more accurate and effective biomarkers to promote the development of early diagnosis through combined the data of the blood and postmortem brain of AD patients [10].

The relationship of variant genes and AD can be assessed using R packages, which are more effective to apply genetic information. Tarozzi M et al. converted targeted next-generation sequencing data of patients with Inherited Creutzfeldt-Jakob disease (CJD) and Sporadic Alzheimer's disease (SAD) into binary using house python script, and analyzed these data using unsupervised and supervised machine learning respectivley, then balance them using R test Hardy-Weinberg to screen mutation and calculate the ratio. The unsupervised approach can identify related gene variants to cause the disease functionally according to Hardy-Weinberg equilibrium analysis, and the supervised machine learning divided sucessessfully CJD from SAD in 100% using the inheritance pattern of the variant genes [11], which further extended to the diagnosis of all neurodegenerative diseases through detection the position of variant genes.

3.3. Screened Novel Therapeutic Strategies in AD from High-throughput Sequencing Data

In addtion to understand the pathogenesis and diagnosis of AD , the R packages analyzed high-throughput sequencing data contributing the development of novel drugs. The scRNA-seq was performed on dormant neural stem cells (NSC) in the sub-ventricular region of adult mice, which gene transcriptome reads were mapped with picard-tools-1.123. Gene expression data were processed using the edge R package and principal component analysis was done using a custom R script based on the FactoMineR library. The pseudo-temporal differentiation trajectory of NSC was reconstructed with the Monocle package. This study found that NSC became activation after injury, which was significantly affected by interferon (IFN)-γsignaling pathway through promoting the activation and proliferation of immune cells, inhibiting pathogenic replication, and inducing repairment of damaged tissues [10]. Therefore, injection of recombinant IFN-γ activated dormant neural stem cells to repair the injured neurons, alleviating the decrement of memory and visual spatial caused by AD, which may develop into a potential therapeutic strategy for AD in the future.

In addition, the R package analysis of high-throughput sequencing data can also clarify the therapeutic mechanisms of AD drugs, providing a clearer direction for improvment of current treatment and development of novel drugs. For example, many studies have confirmed a positive effect of tetramethypyazine (TMP) in the treatment of AD without known therapeutic mechanisms completely. The gene expression in brain tissue cells of APP/PS1 transgenic mice before and after administration of TMP were compared using RNA-Seq, revealed a significant reduction in CUL4B expression through validation furtherly using RT-qPCR. CUL4B is the skeletal protein of the Cullin-RING ubiquitin ligase involved in the regulation of the β amyloid precursor protein, which is a biomarker of AD. Finally, the immunoprecipitation assay revealed that TMP downregulated the expression of CUL4B to decrease the ubiquitination of SSTR4 , thus contributed to the reduction of apoptosis, loss, and autophagy of neurons, and improved the cognition and hippocampal memory in AD patients. These data demonstrated R deeply mines transcriptome sequencing data of mouse brain cells and screens of differential genes provide a new aim in treatment of AD through reducing SSTR4 ubiquitination [12]. The table summarised the important application in this review (Table 1).

Table 1. Applications of R packages in analysis of high-throughput sequencing data.

Objectives	Applications	Examples
Pathogenesis	Analyze of the regulatory mechanisms at the cellular level Selection of risk genes	Key genes affecting astrocytes, such as *WWTR1*, *PHYHD1*, *ZFP36L1*, *AEBP1*, *DST* and *RASL12* Association between *APOE4* expression and risk of AD
Diagnosis	Development of new diagnostic biomarkers Accuration diagnosis	High risk TREM2 variants with R62H and R47H Classification of CJD and SAD
Treatment	Promotion of neuronal repair Allevation of neuronal damage	Injection of IFN-γ Decrement of SSTR4 ubiquitination

4. Conclusion

A large number of studies have combined brain MRI images and electroencephalogram to conduct intelligent diagnosis of AD, and explore the influence of other diseases on AD, such as diabetes and atherosclerosis, aiming to achieve accurate discrimination of AD and further stages classification. The development of AD drugs has started with the removement of β -amyloid from the brain or the increment of neurotransmitters. Although a number of drugs have been approved or pending clinical trials, there is still limitation of detailed research on AD pathology from the cell and gene level because of lacking technique to analysis and extract the information high dimension data effectivly. The analysis of high-throughput sequencing data using R packages, especially in single cell level, can

help to solve the above problems, exploring the mechanisms of expression changes and activity of regulated genes. The R language and its packages are suitable for a large number of statistical data processing, which can select the general data with quality control according to the research requirements, conduct cell clustering to study the effect of specific categories of cells on AD, and analyze the expression of differential genes in order to screen risk genes. The analysis of sequencing data using machine learning can develop intelligent AD diagnostic systems at the cellular and genetic levels. More than twenty risk genes have been discovered to explore the pathogenesis of AD as a polygenic disease through R deep mining of high-throughput sequencing data. In terms of disease diagnosis, more sensitive cerebrospinal fluid and blood biomarkers are constantly explored, and the diagnosis of AD does not stop at judging the disease, but can also be used to automatically diagnose the disease and assess its risk. In addition to the existing AD drugs designed to remove β -amyloid in the brain, the current treatment is also trying to reduce pathological process of neurons before or at the early stage of AD, and to find physiotherapy measures to promote neuronal repair in the middle and later stages of AD. Ultimately, it contributes to early diagnosis and intervention of disease, improve patient quality of life, reduce medical expenses and social burden.

Reference

[1] 2023 Alzheimer's disease facts and figures. (2023) Alzheimer's & dementia : the journal of the Alzheimer's Association, 19(4), 1598–1695.

[2] Zhang, X. X., Tian, Y., Wang, Z. T., Ma, Y. H., Tan, L., & Yu, J. T. (2021) The Epidemiology of Alzheimer's Disease Modifiable Risk Factors and Prevention. The journal of prevention of Alzheimer's disease, 8(3), 313–321.

[3] Ossenkoppele, R., van der Kant, R., & Hansson, O. (2022) Tau biomarkers in Alzheimer's disease: towards implementation in clinical practice and trials. The Lancet. Neurology, 21(8), 726–734.

[4] Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021) Next-generation sequencing technologies: An overview. Human immunology, 82(11), 801–811.

[5] Chandy T. (2023) Intervention of next-generation sequencing in diagnosis of Alzheimer's disease: challenges and future prospects. Dementia & neuropsychologia, 17, e20220025.

[6] Mañanes, D., Rivero-García, I., Relaño, C., Torres, M., Sancho, D., Jimenez-Carretero, D., Torroja, C., & Sánchez-Cabo, F. (2024) SpatialDDLS: an R package to deconvolute spatial transcriptomics data using neural networks. Bioinformatics (Oxford, England), 40(2), btae072.

[7] Yu, W., Li, Y., Zhong, F., Deng, Z., Wu, J., Yu, W., & Lü, Y. (2024) Disease-Associated Neurotoxic Astrocyte Markers in Alzheimer Disease Based on Integrative Single-Nucleus RNA Sequencing. Cellular and molecular neurobiology, 44(1), 20.

[8] Vance, J. M., Farrer, L. A., Huang, Y., Cruchaga, C., Hyman, B. T., Pericak-Vance, M. A., Goate, A. M., Greicius, M. D., Griswold, A. J., Haines, J. L., Tcw, J., Schellenberg, G. D., Tsai, L. H., Herz, J., & Holtzman, D. M. (2024) Report of the APOE4 National Institute on Aging/Alzheimer Disease Sequencing Project Consortium Working Group: Reducing APOE4 in Carriers is a Therapeutic Goal for Alzheimer's Disease. Annals of neurology, 10.1002/ana.26864. Advance online publication.

[9] Zhou, Y., Song, W. M., Andhey, P. S., Swain, A., Levy, T., Miller, K. R., Poliani, P. L., Cominelli, M., Grover, S., Gilfillan, S., Cella, M., Ulland, T. K., Zaitsev, K., Miyashita, A., Ikeuchi, T., Sainouchi, M., Kakita, A., Bennett, D. A., Schneider, J. A., Nichols, M. R., … Colonna, M. (2020) Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. Nature medicine, 26(1), 131–142.

[10] Wang, W., Li, T., Wang, Z., Yin, Y., Zhang, S., Wang, C., Hu, X., & Lu, S. (2023) Bibliometric analysis of research on neurodegenerative diseases and single-cell RNA sequencing: Opportunities and challenges. iScience, 26(10), 107833.

[11] Tarozzi, M., Bartoletti-Stella, A., Dall'Olio, D., Matteuzzi, T., Baiardi, S., Parchi, P., Castellani, G., & Capellari, S. (2022) Identification of recurrent genetic patterns from targeted sequencing panels with advanced data science: a case-study on sporadic and genetic neurodegenerative diseases. BMC medical genomics, 15(1), 26.

[12] Weng G, Zhou B, Liu T, Huang Z, Huang S. Tetramethylpyrazine Improves Cognitive Function of Alzheimer's Disease Mice by Regulating SSTR4 Ubiquitination. Drug Des Devel Ther. 2021 Jun 1;15:2385-2399. doi: 10.2147/DDDT.S290030. PMID: 34103899; PMCID: PMC8179737.