INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: https://www.ijpsjournal.com

OPEN ACCESS

Research Article

Identification And Computational Characterization Of Novel Genetic Variants Associated With Alzheimer's Disease Using Bioinformatics And Machine Learning

Seema Tabassum*

Biotecnika info & labs, India

ARTICLE INFO

Published: 22 Aug. 2026

Keywords:
Alzheimer's disease, amyloid-beta, microglia, astrocytes, genetic variants, RNA editing.

DOI:
10.5281/zenodo.22057543

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, amyloid-beta (Aβ) accumulation, neuroinflammation, oxidative stress, and neuronal dysfunction. Microglia–astrocyte interactions play an important role in modulating Aβ-associated cellular responses and neuroprotection. The present study aimed to computationally identify and characterize candidate genetic variants associated with AD using bioinformatics and machine-learning approaches. Sequencing data corresponding to accession SRR29819222, derived from a human induced pluripotent stem cell-based three-dimensional neurosphere model investigating microglia–astrocyte interactions and Aβ toxicity, were retrieved from the NCBI database. The paired-end sequencing dataset comprised 38,384,873 sequences, approximately 90 bp in length, with 45% GC content and approximately 3.4 Gbp of sequence data. Sequence reads were aligned to the reference genome using HISAT2, followed by SAM/BAM processing and statistical assessment using SAMtools. Genomic variants were identified using DeepVariant and subsequently annotated using Ensembl Variant Effect Predictor (VEP) and SnpEff. Machine-learning-based analysis was further applied to examine variant patterns and associated genes. Three notable variants were identified on chromosomes 7 and 10, comprising one intergenic variant on chromosome 7 and two intronic variants on chromosome 10. Gene-level analysis highlighted PSEN1, ABCA7, APP, NOS3, and APOE among the genes associated with the analyzed variant patterns. The findings provide candidate genomic regions and genes for further investigation of molecular mechanisms

***Corresponding Author:** Seema Tabassum
Address: *Biotecnika info & labs, India*
Email ✉: seema.sajidkhan@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

underlying AD and Aβ-associated cellular responses. However, the identified variants should be considered candidate genomic variants rather than confirmed RNA-editing events. Further RNA-editing-specific analyses, transcriptomic studies, experimental validation, and functional investigations are required to establish their biological significance and contribution to AD pathology

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and extensive molecular and cellular abnormalities within the central nervous system. Amyloid-beta (Aβ) accumulation is one of the major pathological features associated with AD and can contribute to neuroinflammation, oxidative stress, neuronal dysfunction, and neuronal loss.

The response of glial cells to Aβ accumulation is an important component of AD pathology. Astrocytes and microglia undergo characteristic changes in response to injury and disease, collectively referred to as gliosis. Microglia can interact with astrocytes and influence their responses to Aβ, thereby affecting inflammatory and neuroprotective pathways.

Human induced pluripotent stem cell (hiPSC)-derived three-dimensional neurosphere models provide an experimental platform for investigating interactions between neural and glial cells. The source study investigated Aβ-induced pathology in 3D neurospheres containing astrocytes and neurons and compared this with systems supplemented with hiPSC-derived microglia. In the presence of microglia, Aβ phagocytosis and neuroprotective responses were observed, suggesting that microglia–astrocyte interactions may influence Aβ-associated cellular injury.

In the present study, sequencing data associated with accession SRR29819222 were computationally analyzed to identify and characterize variants potentially associated with Alzheimer's disease. The study employed a workflow involving HISAT2, SAMtools, DeepVariant, Ensembl VEP, and SnpEff, followed by machine-learning-based analysis of the identified variants.

2. MATERIALS AND METHODS

2.1 Dataset Retrieval

The sequencing dataset analyzed in this study was obtained from the National Center for Biotechnology Information (NCBI) database using accession number SRR29819222. The dataset corresponds to a study investigating microglia–astrocyte interactions and Aβ toxicity in a human 3D neurosphere model of Alzheimer's disease.

2.2 Sequence Quality Assessment

The retrieved sequencing data were subjected to quality-control analysis. The dataset consisted of paired-end sequencing reads containing both forward and reverse reads. The reported total number of sequences was 38,384,873, with a sequence length of approximately 90 bases and a GC content of 45%. The dataset contained approximately 3.4 Gbp of total bases. No poor-quality sequences were reported in the analyzed dataset.

Sequence characteristics and nucleotide composition were assessed before downstream processing.

2.3 Sequence Alignment

The sequencing reads were aligned against the reference genome using HISAT2. The alignment workflow generated aligned reads suitable for downstream processing and variant analysis.

The source workflow also included processing phases involving SAM/BAM files, including

indexing, sorting, merging, reheadering, and phasing-related processing.

2.4 SAM/BAM Processing and Statistical Analysis

SAMtools was used for processing and assessment of the aligned sequencing data. Alignment statistics were evaluated using SAMtools stats and SAMtools idxstats. These analyses provided information regarding sequencing and alignment characteristics and facilitated downstream variant analysis.

2.5 Variant Calling

Genomic variants were identified using DeepVariant. The variant-calling workflow generated Variant Call Format (VCF) data for subsequent annotation and interpretation.

The source report indicates that DeepVariant generated the VCF file successfully and that the workflow subsequently proceeded to variant annotation.

2.6 Variant Annotation

The identified variants were annotated using Ensembl Variant Effect Predictor (VEP). VEP was used to determine the genomic location and predicted variant consequence. Functional annotation was additionally performed using SnpEff, which provides information regarding the potential effects of genetic variants on genes and proteins.

2.7 Machine-Learning-Based Variant Analysis

The annotated variant dataset was subjected to machine-learning-based analysis to identify patterns in the distribution of variants and associated genes. The analysis included examination of variant types and genes identified within the dataset.

According to the source analysis, genes including PSEN1, ABCA7, APP, NOS3, and APOE were identified among the analyzed variants.

3. RESULTS

3.1 Sequencing Data Characteristics

The analyzed SRR29819222 dataset contained both aligned and unaligned reads and represented paired-end sequencing data. A total of 38,384,873 sequences were reported. The sequence length was approximately 90 bases, the GC content was 45%, and the total amount of sequence data was approximately 3.4 Gbp.

Spots read:38,384,873.

```
spots read      : 38,384,873
reads read      : 76,769,746
reads written   : 76,769,746
```

The dataset passed the reported basic quality-control assessments. The sequencing data were subsequently processed through alignment, BAM processing, variant calling, and annotation pipelines.

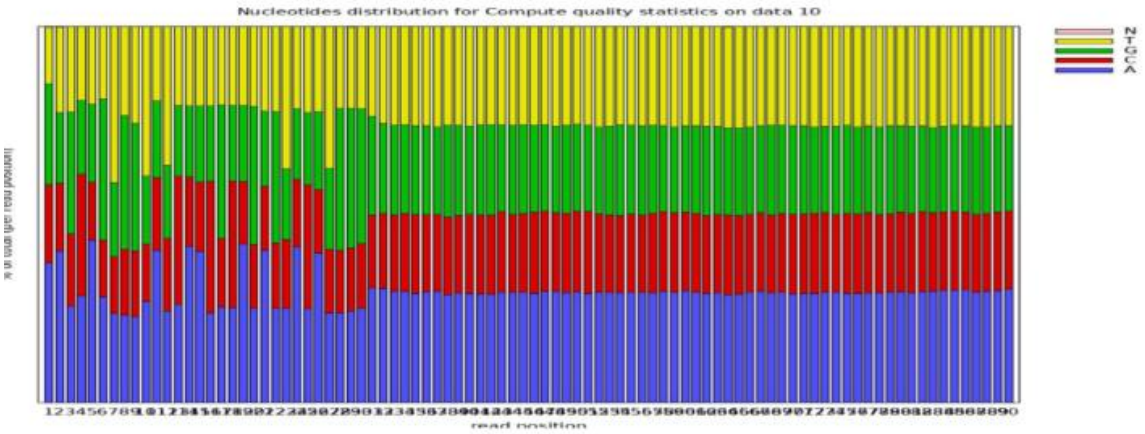


Figure 1. Spots read.

Format: html

```
##Falco 1.2.4
>>Basic Statistics      pass
#Measure        Value
Filename        SRR29819222_reverse
File type       Conventional base calls
Encoding        Sanger / Illumina 1.9
Total Sequences 38384873
Sequences flagged as poor quality       0
Sequence length 90
%GC     45
>>END_MODULE
>>Per base sequence quality     pass
#Base   Mean    Median  Lower Quartile  Upper Quartile  :
1       38.2485 40      40      40      40      40
2       39.0673 40      40      40      40      40
3       39.3862 40      40      40      40      40
4       39.3499 40      40      40      40      40
5       39.4844 40      40      40      40      40
6       39.5551 40      40      40      40      40
7       39.5414 40      40      40      40      40
8       39.5929 40      40      40      40      40
9       39.6032 40      40      40      40      40
10-11   39.582  40      40      40      40      40
12-13   39.6228 40      40      40      40      40
14-15   39.6082 40      40      40      40      40
16-17   39.6495 40      40      40      40      40
```

Figure 2. Format of html

3.2 Identification of Genomic Variants

Following alignment and variant calling, VEP identified three notable variants associated with chromosomes 7 and 10.

Chromosome 7

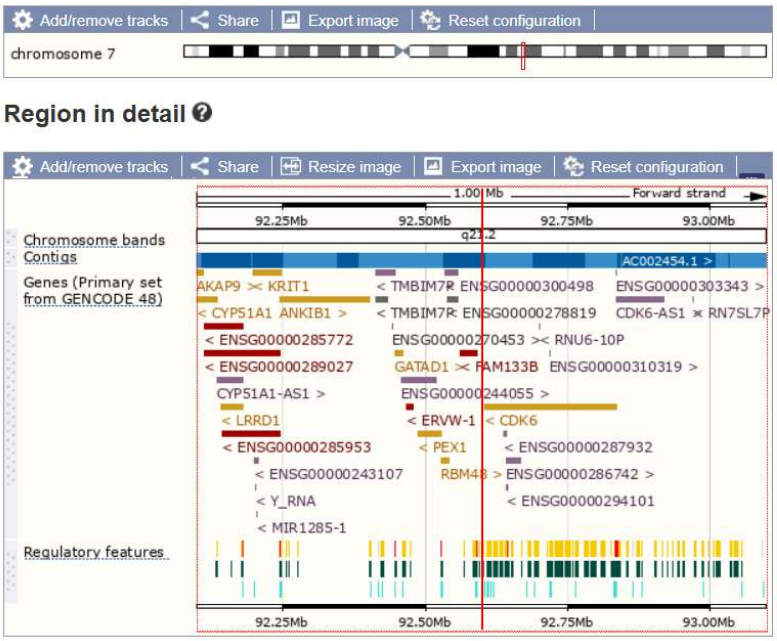


Figure 3. Chromosome no.7 Retrieved through VEP Ensembl

A variant was identified in the genomic region:

Chromosome 7: 92,597,676–92,597,776

The variant was classified as an intergenic variant. The source analysis indicates that the region is associated with long non-coding RNA (lncRNA)-related genomic regions. Variants located in intergenic regions may potentially influence regulatory elements or non-coding transcripts; however, functional effects require experimental validation.

Chromosome 10

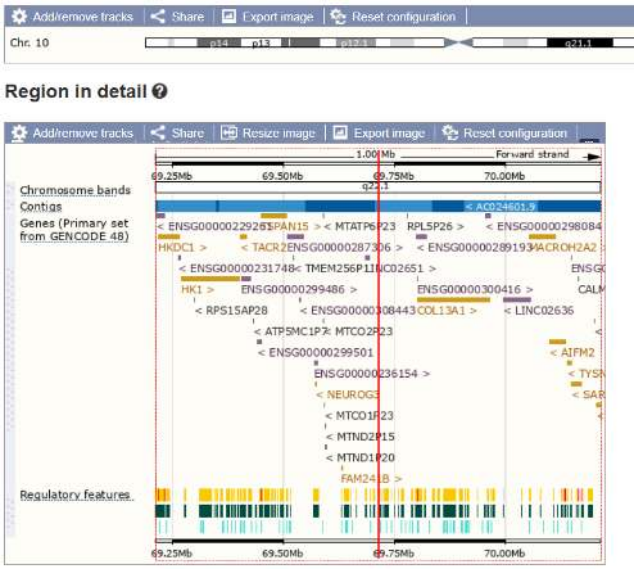


Figure 4. Chromosome no.10 Retrieved through VEP Ensembl

Two additional variants were identified on chromosome 10:

Chromosome 10: 49,703,047–49,703,147

This variant was classified as an intronic variant.

A second variant was identified at:

Chromosome 10: 112,676,367–112,676,467

This variant was also classified as an intronic variant.

3.3 Gene-Level Analysis

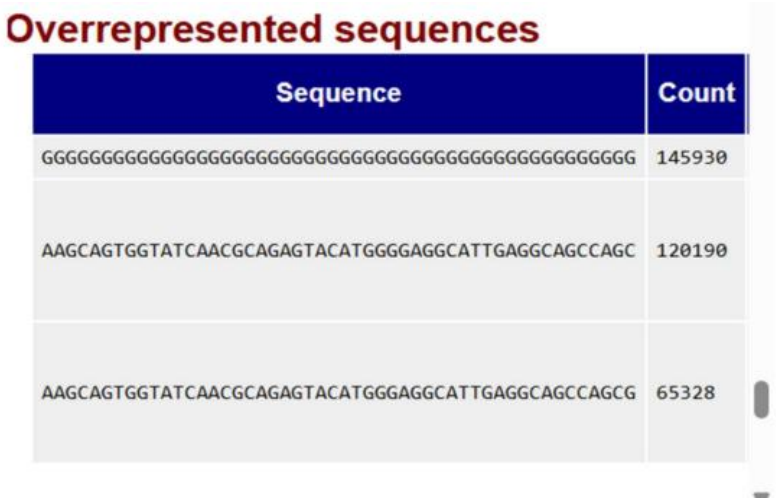
Machine-learning-based analysis identified variants involving several genes of potential relevance to Alzheimer's disease, including PSEN1, ABCA7, APP, NOS3, and APOE.

Basic Statistics

Measure	Value
Filename	SRR29819222_reverse.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	38384873
Total Bases	3.4 Gbp
Sequences flagged as poor quality	0
Sequence length	90
%GC	44

Figure 5. Basic Statistics

The source report describes two deletions involving PSEN1 and ABCA7, three single-nucleotide variants associated with PSEN1, and a cluster of single-nucleotide variants involving APP, NOS3, and APOE.



Overrepresented sequences

Sequence	Count
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	145930
AAGCAGTGGTATCAACGCAGAGTACATGGGGAGGCATTGAGGCAGCCAGC	120190
AAGCAGTGGTATCAACGCAGAGTACATGGGAGGCATTGAGGCAGCCAGCG	65328

Figure 6. Overrepresented sequences

Among these genes, APOE was highlighted because of its role in Aβ-related biological processes. The original analysis associated increased APOE expression with Aβ clearance and microglial responses.

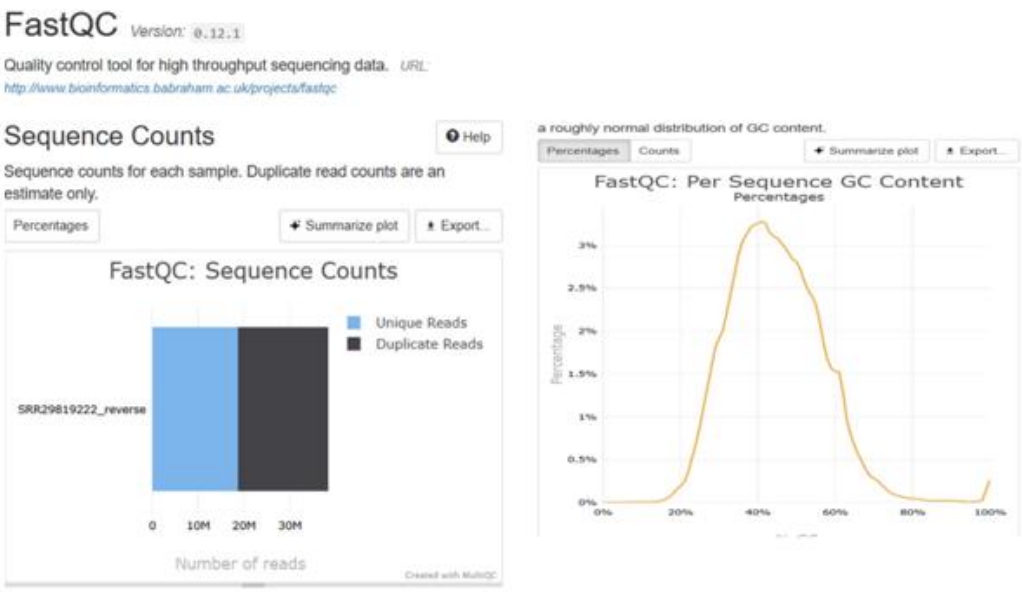


Figure 7. Fast QC & GC Content

DISCUSSION

The present computational analysis identified three notable genomic variants within the analyzed Alzheimer's disease-associated sequencing dataset. These variants were located on chromosomes 7 and 10 and were classified as one intergenic and two intronic variants.

The identification of an intergenic variant on chromosome 7 is potentially relevant because intergenic regions can contain regulatory elements and non-coding transcripts. The source report specifically associates this region with lncRNAs. However, the biological significance of the identified variant cannot be established solely from its genomic position and requires additional functional investigation.

The two chromosome 10 variants were classified as intronic variants. Intronic variants may influence gene regulation, transcript processing, splicing, or other molecular processes depending on their genomic context. Nevertheless, functional effects should not be inferred solely from the intronic classification.

The analysis also identified variants involving PSEN1, ABCA7, APP, NOS3, and APOE. These genes are biologically relevant to neurological disease mechanisms, making them potentially important candidates for further investigation. The source report particularly emphasizes APOE in relation to Aβ clearance and microglia–astrocyte interactions.

The findings are consistent with the broader concept that microglia can influence astrocytic responses to Aβ. The original experimental model described differences between neurospheres containing astrocytes and neurons and systems additionally containing microglia. Microglia were

reported to contribute to Aβ phagocytosis and neuroprotective responses.

Importantly, the present computational findings should be interpreted carefully. Identification of sequence variants from sequencing data does not by itself establish that the variants are RNA-editing events. RNA editing requires evidence demonstrating that the observed nucleotide alteration occurs at the RNA level and differs from the corresponding genomic sequence. Therefore, the variants identified in the present workflow are more appropriately described as candidate genomic variants associated with the analyzed dataset, unless additional RNA-editing-specific analyses were performed.

Similarly, the identification of variants in genes such as APOE, PSEN1, ABCA7, APP and NOS3 does demonstrate that incorporating these genes into medication would cure or eradicate Alzheimer's disease. Such therapeutic conclusions require extensive mechanistic, preclinical, and clinical evidence.

CONCLUSION

The present study demonstrates a computational workflow for the identification and characterization of sequence variants from an Alzheimer's disease-associated sequencing dataset. Using HISAT2, SAMtools, DeepVariant, Ensembl VEP, SnpEff, and machine-learning-based analysis, three notable variants were identified on chromosomes 7 and 10, including one intergenic variant and two intronic variants.

The gene-level analysis highlighted PSEN1, ABCA7, APP, NOS3, and APOE as genes associated with the identified variant patterns. The findings provide candidate genomic regions and genes for further investigation of molecular mechanisms associated with Aβ toxicity and microglia–astrocyte interactions.

The study supports the utility of integrated bioinformatics and machine-learning approaches for exploratory analysis of Alzheimer's disease sequencing data. However, additional RNA-editing-specific analyses, experimental validation, transcriptomic investigation, and functional studies are necessary to confirm the biological significance of the identified variants and to establish their contribution to Alzheimer's disease pathology.

REFERENCES

1. Hamed BA, Farghaly HM, Omar A, Abd El-Hafeez T. Identifying key genetic variants in Alzheimer's disease progression using Graph Convolutional Networks (GCN) and biological impact analysis. Journal of Big Data. 2025 Jul 16;12(1):171.
2. Jo T, Nho K, Bice P, Saykin AJ, Alzheimer's Disease Neuroimaging Initiative. Deep learning-based identification of genetic variants: application to Alzheimer's disease classification. Briefings in bioinformatics. 2022 Mar;23(2):bbac022.
3. Lundberg M, Sng LM, Szul P, Dunne R, Bayat A, Burnham SC, Bauer DC, Twine NA. Novel Alzheimer's disease genes and epistasis identified using machine learning GWAS platform. Scientific reports. 2023 Oct 17;13(1):17662.
4. Kim SH, Yang S, Lim KH, Ko E, Jang HJ, Kang M, Suh PG, Joo JY. Prediction of Alzheimer's disease-specific phospholipase c gamma-1 SNV by deep learning-based approach for high-throughput screening. Proceedings of the National Academy of Sciences. 2021 Jan 19;118(3):e2011250118.
5. Sharma A, Dey P. A machine learning approach to unmask novel gene signatures and prediction of Alzheimer's disease within

different brain regions. Genomics. 2021 Jul 1;113(4):1778-89.

6. Uppar AL, Patil CC, Namannavar S, Deshmane PA. Formulation and Evaluation of Caffeine-Loaded Cubosomes Hydrogel for Topical Delivery. Journal of Drug Delivery & Therapeutics. 2026 Apr 1;16(4):39.
7. Binder J, Ursu O, Bologa C, Jiang S, Maphis N, Dadras S, Chisholm D, Weick J, Myers O, Kumar P, Yang JJ. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Communications biology. 2022 Feb 11;5(1):125.
8. Khater T, Ansari S, Alatrany AS, Alaskar H, Mahmoud S, Turky A, Tawfik H, Almajali E, Hussain A. Explainable machine learning model for Alzheimer detection using genetic data: A genome-wide association study approach. IEEE Access. 2024 Jun 5;12:95091-105.
9. Orlenko A, Venkatesan M, Shen L, Ritchie MD, Wang ZP, Obafemi-Ajayi T, Moore JH. Biologically enhanced machine learning model to uncover novel gene-drug targets for Alzheimer's disease. InBiocomputing 2025: Proceedings of the Pacific Symposium 2024 (pp. 441-456).
10. Xie Z, Situ Y, Deng L, Liang M, Ding H, Guo Z, Xu Q, Liang Z, Shao Z. Identification of therapeutic targets for Alzheimer's Disease Treatment using bioinformatics and machine learning. Scientific Reports. 2025 Jan 31;15(1):3888.
11. Huang X, Liu H, Li X, Guan L, Li J, Tellier LC, Yang H, Wang J, Zhang J. Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. BMC neurology. 2018 Jan 10;18(1):5.
12. Abd El Hamid MM, Mabrouk MS, Omar YM. Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques. Biomedical Engineering: Applications, Basis and Communications. 2019 Oct 15;31(05):1950040.
13. Alatrany A, Hussain A, Mustafina J, Al-Jumeily D. A novel hybrid machine learning approach using deep learning for the prediction of alzheimer disease using genome data. InInternational Conference on Intelligent Computing 2021 Aug 9 (pp. 253-266). Cham: Springer International Publishing.
14. Arnal Segura M, Bini G, Krithara A, Paliouras G, Tartaglia GG. Machine learning methods for classifying multiple sclerosis and Alzheimer's disease using genomic data. International journal of molecular sciences. 2025 Feb 27;26(5):2085.

HOW TO CITE: Seema Tabassum*, Identification And Computational Characterization Of Novel Genetic Variants Associated With Alzheimer's Disease Using Bioinformatics And Machine Learning, Int. J. of Pharm. Sci., 2026, Vol 4, Issue 8, 3659-3668. https://doi.org/10.5281/zenodo.22057543