INTERNATIONAL RESEARCH JOURNAL OF PHARMACY

www.irjponline.com

ISSN 2230 – 8407

Research Article

IDENTIFICATION OF NOVEL SNP PATTERNS IN ALZHEIMER'S DISEASE: NGS METADATA ANALYSIS

Ramaraj Kannan, Ramesh Kumar Gopal *

Bioinformatics Lab, AU-KBC Research centre, Anna University, Chennai, Tamilnadu, India

*Corresponding Author Email: grameshpub@au-kbc.org

Article Received on: 11/09/17 Approved for publication: 02/10/17

DOI: 10.7897/2230-8407.0810195

ABSTRACT

Alzheimer's disease (AD) is a neurodegenerative disorder, which is common among the elderly. The hallmark of AD is the slow progressive loss of neurons in Central Nervous System (CNS) which recognized as no curative or preventive treatment. By developing the personalized medicine, upon identification of SNPs present in the differentially expressed genes, the disease can be halted. The objective of the study is exploring the potent variants present in the healthy control and AD patient samples. The public data was retrieved from EBI. The preliminary comparative analysis of the six datasets is expected to get SNPs which are responsible for AD. The total of 1,45,410 SNPs present in the healthy control sample and 1,42,508 SNPs present in the Alzheimer's disease patient. GSK3β, AXIN1, AXIN2 and APOE are the genes get prioritized over other genes. The variants which were present only in the Alzheimer's disease patient sample are T788C in GSK3β, A1952G in AXIN1, G1524T, A2185T in AXIN2 and C526T, C604T in APOE. The SNPs identified in this study provides a resource for genetic studies in Alzheimer's disease and shall contribute to the development of personalized medicine. The wet lab validation and dynamic data analysis are required for this preliminary study.

Keywords: Alzheimer's disease, Next Generation Sequencing, Variant analysis, GSK3β, APOE

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder[1], which is common among the elderly[2]. Due to cognitive decline the AD becomes the global health crisis[3]. The symptoms of AD include memory loss, decline in thinking, impairment of language and bewilderment[4]. World Alzheimer Report 2010 estimated that 0.5% (36 million) of the world's population was suffered from Alzheimer's disease. In 2016, it gets increased to 0.8% (48 million) of the world's population. It is projected that, by the end of the year 2050, the number of people with dementia may increase to 2.5% of the global. The majority of the Alzheimer's cases occurs in the developed and developing countries. In United States, AD is one of the leading causes of death. In every 66 seconds, someone develops AD across the country. Over the past decade, the deaths resulting from AD increased to 71%. Since Alzheimer's disease has no known treatment to cure and the available therapies can only halt the progression of the disease[5], which consumes the cost of $605 billion (~1% of world's gross product) in the year of 2016.

Early detection of AD is the only way to slow the progression of the disease. But the changes in the brain occurs more than 20 years before the symptoms appear[6, 7]. The hallmark AD is the slow progressive loss of neurons in Central Nervous System (CNS) leads to the impairment of memory and cognitive decline[8-10] and recognized as no curative or preventive treatment[5] with both genetic[11] and nongenetic causes[12]. The pathological changes associated with the accumulation of insoluble fibrous protein[13] consist of extracellular amyloid plaques[14, 15] and intraneuronal neurofibrillary tangles[16], they are the prime reason for the neuronal cell loss and vascular damage.

The Amyloid β (Aβ) Peptide is the key component of the amyloid plaque formation. Aβ is a -4 kDa peptide composed of 39 to 43 residues[17, 18] produced by single copy gene which is the internal peptide within Amyloid Precursor Protein (APP)[19]. The APP is encoded by the gene APP which is located in trisomy region of Chromosome 21 . The amyloid protease sequentially cleaves the APP to yield subunits of APP. The enzymes are also known as secretases. Cleavage by α-secretase or β-secretase yields large soluble APP derivatives, they are membrane tethered α- or β-carboxyl terminal fragments[20], this action followed by γ-secretase. If α-secretase and γ-secretase cleaves APP there is no formation of $A\beta_{42}$, also they are not amyloidgenic. When β-secretase and γ-secretase cleave APP, it forms either $A\beta_{40}$ (soluble and innocuous[21]) or $A\beta_{42}$ (more hydrophobic and sticky [22]). $A\beta_{42}$ is reported to cause Alzheimer's disease[23]. In most of the cases, soluble $A\beta_{40}$ secreted to the Cerebrospinal fluid in high amount (90%) and low amount of $A\beta_{42}$ (<10%). Other species of Aβ also secret at very low level[24]. The Aβ peptide with 42 amino acid begins 99 residues from the carboxyl terminus of APP and clumps together to form insoluble amyloid plaques[23]. The amyloid cascade hypothesis states that the Alzheimer's pathology is the outcome of two successive events. The first event is cleavage of Aβ Peptide or APP results $A\beta_{42}$ formation and the second event is it must form plaques and neuronal death[25].

In the current study, SNPs were analyzed in six different transcriptome profiles (total brain, temporal and frontal lobe of healthy and Alzheimer's disease patient post-mortem tissue) processed by RNA Seq technology, these following data procured from Natalie Twine et. al., 2011[26]. Reads were pre-processed by quality checking and filtering of low quality reads. The processed reads were mapped to reference genome hg19. Hg19 is the major update of UCSC genes track[27], that includes non-coding transcripts. Reads that mapped to reference were detached and the

replica reads were removed to knock out the false SNPs. SNP calling was implemented with default parameters and it was annotated to predict the functional impact of variants. Genes that showed distinctive expression by differential expression, gene analysis and splice variants were reported in these samples[26]. The variants existing in the samples and SNPs which are expedient to cause Alzheimer's disease were identified. This is only a preliminary analysis and further in-depth data analysis and wet lab experiments would help for further validation.

MATERIALS AND METHODS

Data retrieval

The next generation sequencing single-end data were retrieved from the NCBI-SRA database. The data were freely accessible with the accession number of SRR085471, SRR085473, SRR085474, SRR085725, SRR085726 and SRR087416 for the six different categories. The details of the samples were shown in Table 1. Natalie Twine et. al., 2011 used Illumina sequencing technology to prepare the mRNA-Seq for the total human brain after post-mortem. And their study stated that different promoter regions control the transcriptional isoforms of apolipoprotein E and the level of expression of splice variants differing each other[26].

Table 1. Sample information

S.No	Accession Number	Sample	Gender	Age
1.	SRR085471	Normal Temporal Lobe	5 male	23 – 29
2.	SRR085473	AD Temporal Lobe	1 male	80
3.	SRR085474	Normal Frontal Lobe	5 male	23 – 29
4.	SRR085726	AD Frontal Lode	1 male	87
5.	SRR085725	Normal Total Brain	13 male; 10 Female	23 – 86
6.	SRR087416	AD Total Brain	1 male	87

Pre-processing

The sequence data might contain the sequence artifacts such as base calling errors, small insertions/deletions (read errors), poor quality reads and primer/adapter contamination. For variant analysis, the quality of the data plays significant role. The quality control check was executed with Trimmomatic to assess the quality of the data, examine the distribution of nucleotide and find the low quality reads based on sequence constitution[28]. Trimmomatic is more flexible and efficient preprocessing tool.

Alignment/mapping

The short reads of DNA sequence were efficiently aligned with the reference genome hg19 using Burrows-Wheeler Aligner. Based on the Burrows Wheeler Transform (BWT) algorithm, the short reads align with the large reference sequence efficiently. BWA is faster than Mapping and Assembly with Quality (MAQ) tool by 10 to 20 times with the higher accuracy[29].

During library construction and amplification bias using PCR, the duplicate reads also generate. To reduce the error rate and processing time during further downstream process, SAMtools was used to remove the PCR duplicates present in the data[30]. The aligned reads were sorted using SAM Tools sort option.

Variant calling

To identify the existence of single nucleotide variants Mpileup option in SAMtools was implemented position based on the Freud-scale quality[30]. The variant calling process pinpoint the SNP and structural divergent regions between six different categories of samples and reference genome. To detect somatic mutation, multiple tissue sample with single individual was aligned. The 'call' command of BCFtools is used to call the variants[30].

Variant annotation

The functional impact of variants were annotated using wANNOVAR[31]. Based on gene and its location and type of variation, ANNOVAR illustrate the functional genetic variations in the residues. SIFT and PolyPhen were used to detect the amino acid substitution on the protein function[32] and impact of structural changes intern how it is affecting the protein function by divergent sequence and homology based phyologenetic methods respectively[33].

Table 2. Number of reads survived and dropped after Primary processing

Sample Type	Site	Raw sequence	After trimming			
			Survived		Dropped	
			Read Count	Percentage	Read Count	Percentage
Normal	Temporal Lobe	15256752	14859586	97.72 %	397166	2.60 %
	Frontal Lobe	15772947	15087587	95.65 %	685360	4.35 %
	Total Brain	13442077	13111655	97.54 %	330422	2.62 %
Alzheimer's Disease Patient	Temporal Lobe	14227702	13587685	95.50 %	640017	4.50 %
	Frontal Lobe	15228832	14601148	95.88 %	627684	4.12 %
	Total Brain	14720816	14302335	97.14 %	418481	2.84 %

Table 3: Number of reads survived after secondary processing

Sample Type	Site	Processed Reads	Secondary Processing	
			Filtered reads	Removed Duplicates
Normal	Temporal Lobe	14859586	14188227	6742303 (47.52 %)
	Frontal Lobe	15087587	14621117	6711868 (45.91 %)
	Total Brain	13111655	12621596	6516200 (51.63 %)
Alzheimer's Disease Patient	Temporal Lobe	13587685	13569773	9795269 (72.18 %)
	Frontal Lobe	14601148	14296444	7549266 (52.81 %)
	Total Brain	14302335	13994783	7300642 (52.17 %)

Table 4: Novel variants present in each sample and number of genes prioritized

Sample Type	Site	Total variants	Exonic variants	Known variants	Novel variants	Genes prioritized
Normal	Temporal Lobe	56678	6568	34110	22568	537
	Frontal Lobe	56387	6974	31441	24946	553
	Total Brain	32345	11448	509	31836	892
Alzheimer's Disease Patient	Temporal Lobe	34629	3341	19306	15323	344
	Frontal Lobe	49482	5616	27913	21569	527
	Total Brain	58397	4689	35488	22909	408

Table 5: Number of SNPs present in the differentially expressed genes

Gene	Normal			Alzheimer's disease patient		
	Temporal Lobe	Frontal Lobe	Total Brain	Temporal Lobe	Frontal Lobe	Total Brain
APOE	1	-	3	-	1	2
APP	14	17	10	6	12	16
PSEN1	6	9	5	5	4	7
PSEN2	1	3	1	2	2	4
NCSTN	2	1	-	1	1	1
APH1	1	3	-	-	3	2
GSK3β	10	11	4	6	6	6
AXIN	6	7	1	9	9	13

Table 6: Plausible SNPs in the prioritized gene

S.No	Gene symbol	Exon number	SNP site	
			Nucleotide	Amino acid
1.	GSK3β	7	T788C	V263A
2.	AXIN1	7	A1952G	H651R
3.	AXIN2	6, 9	G1524T, A2185T	K508N, T729S
4.	APOE	4	C526T, C604T	R176C, R202C

RESULTS AND DISCUSSIONS

A pipeline, shown in Figure 1. was followed to find the effect of variations of the healthy and patient sample.

Primary processing – Quality Check and trimming

The quality of the sequence was analysed using the FASTQC tool. Based on the quality statistics, Trimmomatic was used to cut the primer/adapter region and other Illumina specific sequence from the read, discard the bases below quality score 30 and drop the reads below specified length of 35. The single-end read after the quality check was used for the further secondary processing. The dropped read counts were shown in Table 2.

Secondary Processing – mapping, alignment, sorting, filter and remove duplicates

The survived reads after the primary processing (Trimmomatic) were mapped to the annotated human reference genome Hg19, using BWA. The mapped reads were sorted using the Sort option in the SAMTools. Filter SAM separated the mapped reads from the unmapped reads. Due to sequencing flaws, uneven quality of sample preparations, physical gap of the reference and the defined mapping criteria are the major reasons behind the unmapped reads. Finally, the Rmdup option was used to remove the PCR duplicates present in the sequence. Number of reads survived after secondary processing were shown in Table 3.

Tertiary Processing – Variant calling and Variant Annotation

The variant discovery suit developed by the SAMTools, mPileup was used to identify SNPs. BCFTools was used to call the variants and it tends to call more SNPs and lower the false detection. The diploid variants with the depth of 8000 was set as the default parameter. ANNOVAR was used to annotate the variants against dbSNP of reference genome hg19. The number of exonic variants and novel variants annotated by ANNOVAR is shown in Table 4.

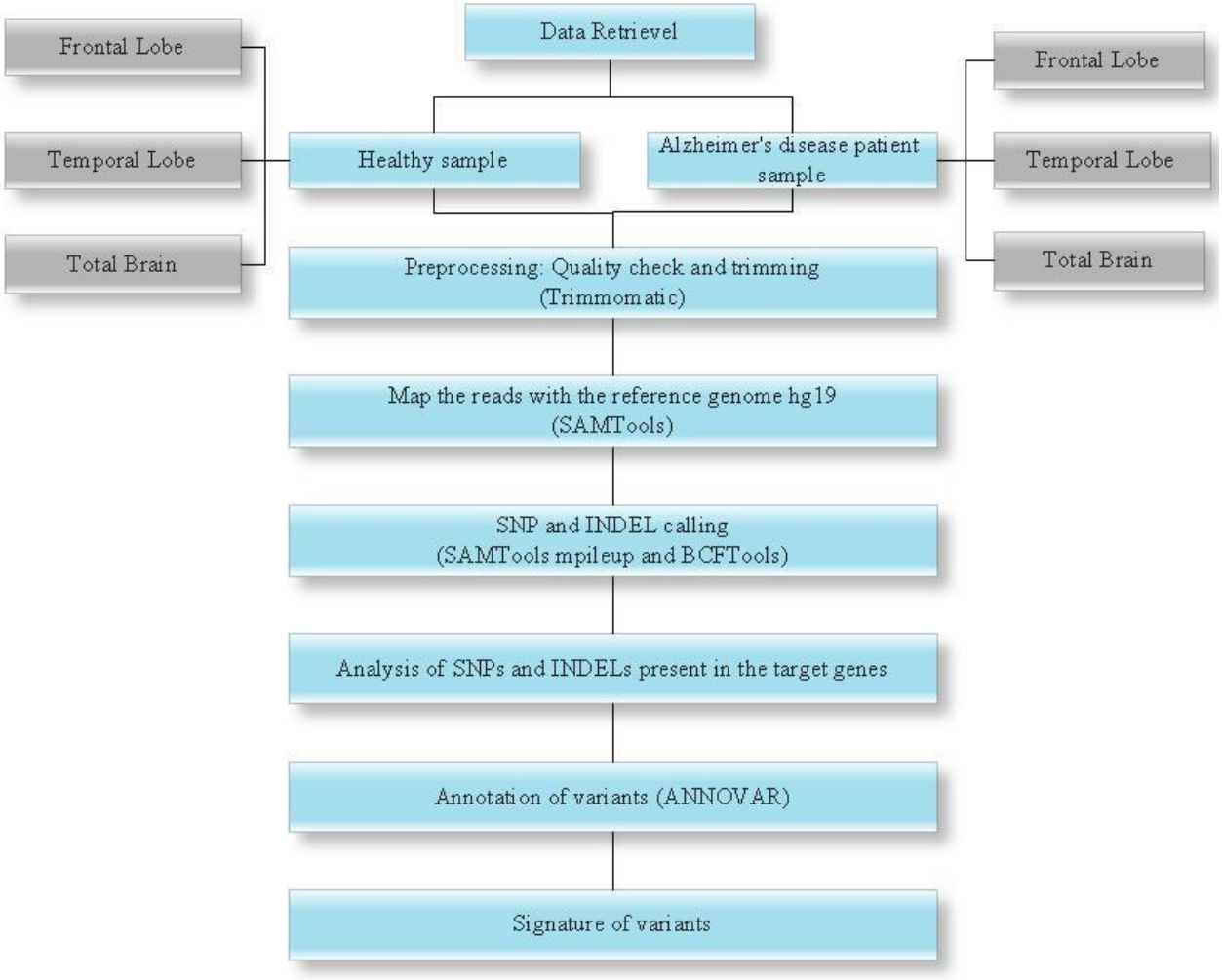

Figure 1: Variant analysis work-flow

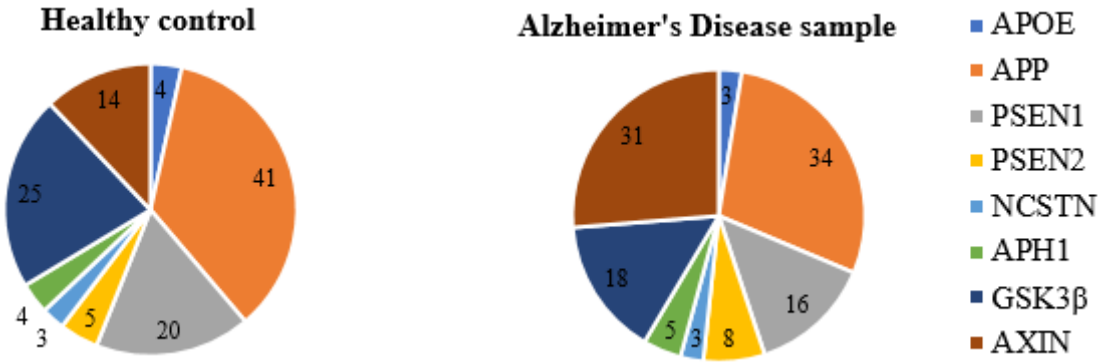

Figure 2: Total SNPs present in the differentially expressed genes of a) healthy control sample and b) Alzheimer's disease sample

Differentially expressed genes

Numerous research reports stated that both early and late onset Alzheimer's disease interlinked with several genes with differential expression. The SNPs were found in the following genes such as APOE, APP, PSEN1, PSEN2, NCSTN, PEN2, GSK-3β and AXIN. APOE metabolize the β-peptide and its isoforms play important role in late onset Alzheimer's disease. The mutation in one the three genes which are APP, PSEN1 and PSEN2 leads to the early onset Alzheimer's disease. NCSTN exhibits the same methylation pattern as APP but it related with late-onset of AD. γ-secretase complex which cleaves APP consist of the four genes such as NCSTN, APH1, PEN2 and PSEN. The proteolytic activity depends upon these four proteins. GSK3 hypothesis of Alzheimer's disease stated that the over-activity of GSK3 leads to the increased β-amyloid production, tau-hyper phosphorylation and memory impairment GSK-3β and AXIN are express in wnt signaling pathway and activate the misrecognition

events and onset the Alzheimer's disease. The number of SNPs present in the differentially expressed genes were shown in Table 5.

There are 116 SNPs present in the differentially expressed genes of normal sample, shown in Figure 2. Out of 116 SNPs, 28 SNPs were non-synonymous exonic variants and 16 SNPs were synonymous exonic variants and the other SNPs occur in the intronic, intergenic and UTR3' regions. The differentially expressed genes in the Alzheimer's disease patient sample had 118 variants. Among 118, 4 SNPs fall in the splicing site, 13 in the exonic region also comes under synonymous and the rest in the intronic and intergenic regions. AXIN is the gene which shows the potent difference between the patient and healthy sample. Also in the patient sample, SNPs occurred in the gene were exonic (synonymous) and splicing site.

Gene Prioritization

Based on the research report, statistical data analysis, integrated computational approach to data sets such as expression data, functional annotation and sequence information the candidate gene related to Alzheimer's disease were prioritized using Phenolyzer. GSK3β and AXIN were the two genes, which get prioritized among the other genes present in the Alzheimer's disease sample. The biologically plausible SNPs in the candidate genes were listed in Table 6.

The variants in the prioritized genes are not observed in the healthy control sample. Also, the changes noted in the regulatory and functional region that may affect the gene. When evaluating the changes among all the differentially expressed genes in all the samples, some variants are present in the Alzheimer's disease patient sample possess some unique variants. Especially the Alzheimer's disease sample from temporal lobe shows the potent variant in the GSK3β, which is reported for the increased β-amyloid production.

GSK3 genes is responsible for several cellular processes. It exists in two forms such as GSK3α and GSK3β. Wnt signaling pathway regulates the activity of GSK3 complex. GSK3β and β-catenin are the two transducers which along with the presenilin protein in the Notch signal transduction cascade involves in the development of brain. Dishevelled protein acts as the connector. The variants in the regulatory region of GSK3β or AXIN leads to the functional loss of Wnt signaling pathway. It would cause the misrecognition event and end with Alzheimer's disease development.

CONCLUSION

Next generation sequencing Transcriptome data analysis is the most knowledgeable method for identification of SNPs. In the current study, we explored the variants existing in the differentially expressed genes. We found four genes GSK3β, AXIN1, AXIN2 and APOE had the variants differ which present only in Alzheimer's disease patient sample. Hence, the Alzheimer's disease patient having this set of SNPs is a preliminary signature for the disease pathogenesis. To maximize drug efficacy and minimize adverse side effects, the genetic makeup of the individual is the key factor.

REFERENCES

1. Strassnig M, Ganguli M. About a peculiar disease of the cerebral cortex: Alzheimer's original case revisited. Psychiatry (Edgmont) 2005; 2(9):30-33.
2. Kay DWK, Beamish P, Roth M. Old age mental disorders in Newcastle upon Tyne. The British Journal of Psychiatry 1964; 110(468):668-682.
3. Wilson RS, Segawa E, Boyle PA, Anagnos SE, Hizel LP, Bennett DA. The natural history of cognitive decline in Alzheimer's disease. Psychology and aging 2012; 27(4):1008
4. Geldmacher DS, Whitehouse PJ. Evaluation of dementia. New England Journal of Medicine 1996; 335(5):330-336
5. Dunkin JJ, Anderson-Hanley C. Dementia caregiver burden A review of the literature and guidelines for assessment and intervention. Neurology 1998; 51(1 Suppl 1):S53-S60.
6. Bateman RJ, Xiong C, Benzinger TLS, Fagan AM, Goate A, Fox NC, *et al.* Clinical and biomarker changes in dominantly inherited Alzheimer's disease. New England Journal of Medicine 2012; 367(9):795-804
7. Jack CR, Lowe VJ, Weigand SD, Wiste HJ, Senjem ML, Knopman DS, *et al.* Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. Brain 2009; awp062
8. Merriam AE, Aronson MK, Gaston P, Wey SL, Katz I. The psychiatric symptoms of Alzheimer's disease. Journal of the American Geriatrics Society 1988; 36(1):7-22
9. Knesevich JW, Berg L, Danziger W. Clinical and Research Reports. Psychiatry 1983; 140:233-235.
10. Cummings JL, Miller B, Hill MA, Neshkes R. Neuropsychiatric aspects of multi-infarct dementia and dementia of the Alzheimer type. Archives of Neurology 1987; 44(4):389-393.
11. Janssen JC, Beck JA, Campbell TA, Dickinson A, Fox NC, Harvey RJ, *et al.* Early onset familial Alzheimer's disease Mutation frequency in 31 families. Neurology 2003; 60(2):235-239
12. Mandelkow E-M, Mandelkow E. Tau in Alzheimer's disease. Trends in cell biology 1998; 8(11):425-427.
13. Braak H, Braak E. Evolution of the neuropathology of Alzheimer's disease. Acta Neurologica Scandinavica 1996; 94(S165):3-12.
14. Braak H, Braak E, Ohm T, Bohl J. Alzheimer's disease: mismatch between amyloid plaques and neuritic plaques. Neuroscience letters 1989; 103(1):24-28.
15. Yamaguchi H, Hirai S, Morimatsu M, Shoji M, Ihara Y. A variety of cerebral amyloid deposits in the brains of the Alzheimer-type dementia demonstrated byβ protein immunostaining. Acta neuropathologica 1988; 76(6):541-549.
16. Kalus P, Braak H, Braak E. The presubicular region in Alzheimer's disease: topography of amyloid deposits and neurofibrillary changes. Brain research 1989; 494(1):198-203.
17. Masters CL, Simms G, Weinman NA, Multhaup G, McDonald BL, Beyreuther K. Amyloid plaque core protein in Alzheimer disease and Down syndrome. Proceedings of the National Academy of Sciences 1985; 82(12):4245-4249.
18. Prelli F, Castano E, Glenner GG, Frangione B. Differences between vascular and plaque core amyloid in Alzheimer's disease. Journal of neurochemistry 1988; 51(2):648-651.
19. Du Yan S, Chen X, Fu J, Chen M, Zhu H, Roher A, *et al.* RAGE and amyloid-β peptide neurotoxicity in Alzheimer's disease. Nature 1996; 382(6593):685-691.
20. Yan R, Bienkowski MJ, Shuck ME, Miao H, Tory MC, Pauley AM, *et al.* Membrane-anchored aspartyl protease with Alzheimer's disease β-secretase activity. Nature 1999; 402(6761):533-537.
21. Shinkai Y, Yoshimura M, Ito Y, Odaka A, Suzuki N, Yanagisawa K, *et al.* Amyloid β-proteins 1—40 and 1—42

(43) in the soluble fraction of extra-and intracranial blood vessels. Annals of neurology 1995; 38(3):421-428
22. Cai X-D, Golde TE, Younkin SG. Release of Excess Amyloid Protein from a Mutant Amyloid Protein Precursor. Science-new york then washington 1993; 259:514-516.
23. Younkin SG. The role of Aβ42 in Alzheimer's disease. Journal of Physiology-Paris 1998; 92(3):289-292.
24. Vigo-Pelfrey C, Lee D, Keim P, Lieberburg I, Schenk DB. Rapid Communication: Characterization of β-Amyloid Peptide from Human Cerebrospinal Fluid. Journal of neurochemistry 1993; 61(5):1965-1968.
25. Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. Science 1992; 256(5054):184-185.
26. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PloS one 2011; 6(1):e16266
27. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics 2006; 22(9):1036-1046.
28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; btu170.
29. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 2008; 18(11):1851-1858.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25(16):2078-2079.
31. Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. Journal of medical genetics 2012; 49(7):433-436.
32. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research 2001; 11(5):863-874.
33. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics 2013; 1-7.

Cite this article as:

Ramaraj Kannan and Ramesh Kumar Gopal. Identification of novel SNP patterns in Alzheimer's disease: NGS metadata analysis. Int. Res. J. Pharm. 2017;8(10):132-137 http://dx.doi.org/10.7897/2230-8407.0810195

Source of support: Nil, Conflict of interest: None Declared