Research Paper

Identifying Gene Signature in RNA Sequencing Multiple Sclerosis Data

Taiebe Kenarangi[1], Enayatolah Bakhshi[1], Kolsoum Inanloo Rahatloo[2], Akbar Biglarian[3*]

1. Department of Biostatistics and Epidemiology, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.
2. Department of Cell and Molecular Biology, School of Biology, College of Science, University of Tehran, Tehran, Iran.
3. Department of Biostatistics and Epidemiology, Social Determinants of Health Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

Use your device to scan and read the article online

Citation Kenarangi T, Bakhshi E, Inanloo Rahatloo K, Biglarian A. Identifying Gene Signature in RNA Sequencing Multiple Sclerosis Data. Iranian Rehabilitation Journal. 2022; 20(2):217-224. http://dx.doi.org/10.32598/irj.20.2.1606.1

http://dx.doi.org/10.32598/irj.20.2.1606.1

Article info:
Received: 26 Sep 2021
Accepted: 16 Apr 2022
Available Online: 01 Jun 2022

Keywords:
Multiple sclerosis, Gene signature, K-means, Pareto optimal clustering, RNA-seq

ABSTRACT

Objectives: Multiple Sclerosis (MS) is a complex central nervous system disease; it is the result of a combination of genetic predispositions and a nongenetic trigger. This study aims to find the gene signatures using a Pareto optimization algorithm for MS RNA sequencing (RNA-seq) data.

Methods: This case-control study involved 50 samples (25 MS patients and 25 age-matched healthy individuals) and their GSE profiles (GSE123496) were selected from the National Center for Biotechnology Information Gene Expression Omnibus database. We used Pareto-optimal cluster size identification to find the gene signatures in the RNA-seq data. After prefiltering and normalizing the data, we used the Limma package to find the differentially expressed genes (DEGs). The Pareto-optimal cluster size for these DEGs was then determined using the technique, multi-objective optimization for collecting the clusters alternatives. Afterward, the RNA-seq data were clustered via k-means with suitable cluster size. The best cluster, as a signature, was found by calculating the mean of the Spearman correlation coefficients (SCCs) of whole genes in the module in a pairwise manner. All analysis was performed in the R software, 4.1.1 package, under virtual space with 100 GB RAM.

Results: In total, 960 DEGs were identified by the Limma analysis. Among them, 720 were up-regulated genes and 240 were down-regulated genes. Meanwhile, 6 Pareto-optimal clusters were obtained. Two clusters that had the greatest average SCCs score (0.88 and 0.74, respectively) were chosen as the gene signatures.

Discussion: A total of 9 metabolic prognostic genes and 3 biological pathways were identified. These can provide more potent prognostic information for MS patients.

*** Corresponding Author:**
Akbar Biglarian, PhD.
Address: *Department of Biostatistics and Epidemiology, Social Determinants of Health Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.*
Tel: *+98 (21) 22180146*
E-mail: *abiglarian@uswr.ac.ir*

Highlights

- We identified 960 DEGs in MS (720 up-regulated and 240 down-regulated genes).
- We used an optimization technique, namely multi-objective optimization for collecting the clusters alternatives (in the R software) to find the optimal cluster for the DEGs.
- We identified a gene signature containing 9 genes for MS.

Plain Language Summary

MS is a chronic and vigorously disabling disease that affects the central nervous system. Researchers can find more effective therapies and cures for MS if they can detect its causes. Our findings could help researchers in treating and preventing MS. In our study, nine prognostic genes related to MS were recognized and also three biological pathways were identified. These can provide more potent prognostic information for MS patients and this could help researchers in treating and preventing MS.

1. Introduction

Multiple sclerosis (MS) is a complex central nervous system (CNS) disease with new aetiological elements being discovered all the time [1]. In this disease, myelin and axons are damaged to varying degrees [2]. There are reversible neurological impairments in most patients that are frequently followed by step-by-step neurological declination over time [3]. MS affects women twice as much as it affects men. It usually strikes people between the ages of 20 and 45. Clinical symptoms and supportive evidence from ancillary tests are used to diagnose the disease (CSF) [4]. In addition, MS is a heritable disease with a well-documented higher incidence in people with a family history of this disease [5]. The cause is unknown, although it appears to be the result of a combination of genetic predispositions and a nongenetic trigger, such as a virus, metabolism, or environmental factors, which results in a self-sustaining autoimmune illness with recurring immunological attacks on the CNS [6].

In recent years, deep sequencing has revolutionized biology and medicine, allowing us to understand nucleic acid sequences at single base level precision in a high throughput manner. RNA sequencing (RNA-seq) is now a widely used tool to analyze gene expression and discover new RNA species [7]. One of the applications of the clustering methods in medicine is to identify subgroups or classes of a disease type. Ascertaining the right number of clusters in a data file is a key question in partitioning clustering; however, there is no conclusive answer to this question. Consequently, the error rate may increase [8].

Cluster number estimation and also identifying the best cluster number is a multi-objective optimization problem. In this sense, cluster analysis can be considered an exploratory data mining technique that can be used. This study aims to use a multi-objective optimization strategy to identify the best cluster.

2. Materials and Methods

Data collection and gene expression analysis

In this study, mRNA expression datasets of MS patients and age-matched healthy individuals (control samples) were investigated. These expression datasets were searched by "Multiple Sclerosis," "mRNA," and "Count" in the gene expression omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo). Finally, the GSE123496 profiles were selected and analyzed. To decide the significance of differentially expressed genes (DEGs), the adjusted P value (with the threshold of 0.05) was used by the Limma package in the Bioconductor software. In addition, an expression matrix from up- and down-regulated genes (mural DEGs) was built for the next analysis.

Identifying gene signatures

After finding the collection of DEGs, we utilized the multi-objective optimization for collecting the clusters alternatives (MOCCA) package (in the R program) on the data from the DEGs to characterize the convenient number of clusters (up- and down-regulated genes). Additionally, a bootstrapping method based on various cluster validity indices was used to obtain the stable (Pareto-optimal) cluster numbers by the MOCCA ap-

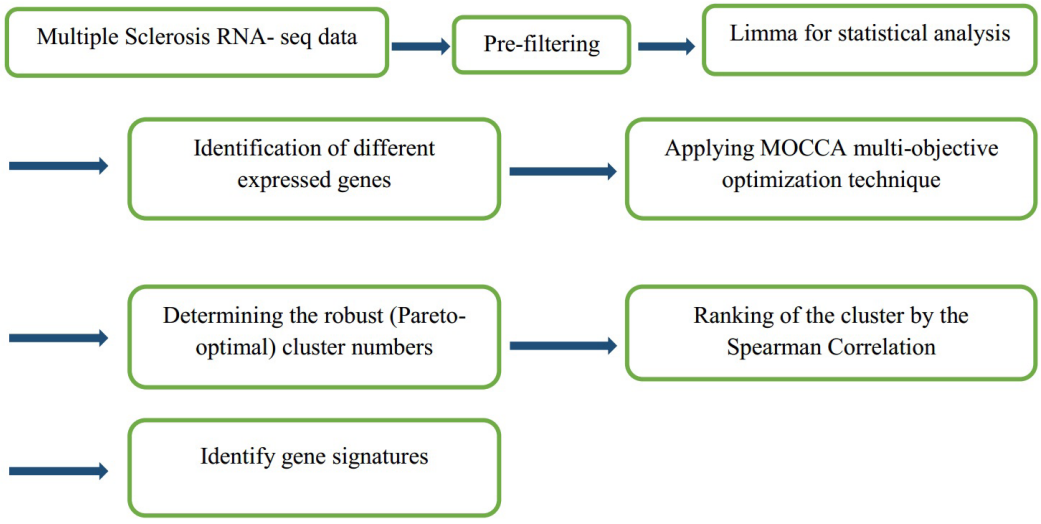

Figure 1. Flowchart of the Suggested framework to identify gene signature

Limma, linear models for microarray and RNA-Seq data.

MOCCA: multi-objective optimization for collecting the clusters alternatives; RNA-seq: RNA sequencing.

proach. Accordingly, for all the unique cluster numbers, such as k, 3 clustering approaches, namely k-means, single-linkage, and neural gas clustering were used. Subsequently, using a variety of cluster validation indices (MCA, Jaccard, FM, and CQS) were clustered. To achieve the optimum number of clusters, 12 objective functions (k-means.MCA, k-means.Jaccard, k-means. FM, k-means.neuralgas, CQS.MCA, neuralgas.Jaccard, neuralgas.FM, neuralgas.CQS, single.MCA, single. Jaccard, single.FM, single.CQS, single.MCA, single. Jaccard, single) were used.

After determining the ideal number of clusters, k-means clustering with the optimal cluster size was utilized to identify the cluster information of each gene. Subsequently, the Spearman correlation coefficient (SCC) value derived from the involved paired genes was used to get the mean SCC value of each cluster. The cluster with the highest mean SCC value was picked as the best. In this yield, the gene set of the best cluster was used as a gene signature (Figure 1).

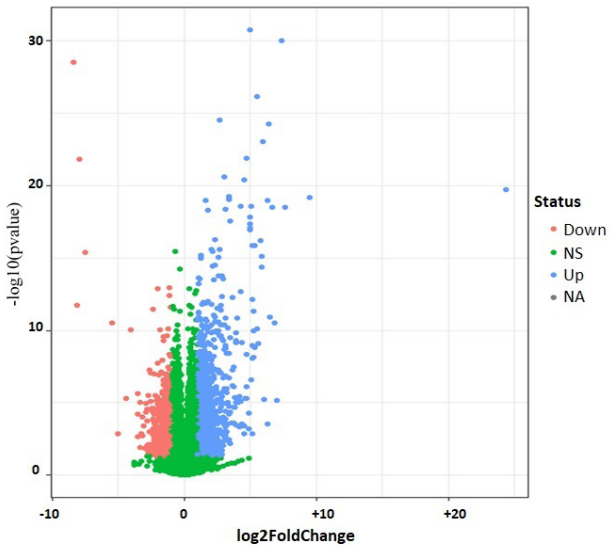

Figure 2. Identification of up- and down-regulated genes for multiple sclerosis dataset (volcano plot)

The DAVID database was used to conduct the KEGG pathway analyses for the signature's participating genes. Only the KEGG pathways with enrichment adjusted P value of less than 0.05 were included in this study.

Results

Data collection and DEGs identification

In this study, we used GSE123496. The RNA-seq dataset contained 50 samples (25 MS patients and 25 age-matched healthy individuals). By using the Limma package in the R software, differential expression analysis was carried out between patients and healthy individuals. In total, 960 DEGs (720 up-regulated and 240 down-regulated genes) were identified (supplementary file 1, Figure 2).

Determining gene signatures

As a result, the Pareto-optimal cluster size was discovered to be 6. The objective values for each of the 12 objective functions are provided in Table 1. After determining the appropriate number of clusters (n=6), k-means clustering with the optimal cluster size was utilized to collect cluster information for each participating gene. The SCC of the 6 clusters averaged 0.206, 0.879, 0.744, 0.477, 0.361, and 0.569, respectively. The gene signature was chosen as the second and third clusters with the highest mean SCC value (0.879, 0.744). The gene signature included 9 DEGs. The gene IDs and full names are provided in Table 2.

Pathway enrichment analysis

The DAVID online tool was utilized to find the enriched pathways with an adjusted P value of ≤ 0.05. In the GSEA analysis, 3 biological pathways were significantly enriched. The KEGG pathway analysis indicated that alternative complement activation, neutrophil degranulation, and activation of C3 and C5 were most prevalent in the final list (Table 3).

4. Discussion

MS is a central neuroinflammatory disease. Although the cause of MS is unknown, the most recent working model for disease pathogenesis argues that the interaction of genetic and environmental variables is required for MS development [9]. As a result, prognostic biomarker identification in MS is critical. In this study, we used a multi-objective optimization strategy to identify prognostic gene signatures in MS patients.

Table 1. Values of 12 objectives in multi-objective optimization for collecting the clusters alternatives from the MS RNA sequencing dataset

Objectives	Objective Value
k-means.MCA	0.6215
k-means.Jaccard	0.4386
k-means.FM	0.5983
k-means.CQS	0.9991
neuralgas.MCA	0.6580
neuralgas.Jaccard	0.4684
neuralgas.FM	0.6138
neuralgas.CQS	0.9885
single.MCA	0.6037
single.Jaccard	0.4118
single.FM	0.5849
single.CQS	0.9889

Table 2. Names of genes related to the multiple sclerosis gene signature

Gene Symbol	Full Name
CD74	CD74 molecule
VIM	Vimentin
C3	Complement C3
CHI3L1	Chitinase 3 like 1
SERPINA3	Serpin family A member 3
AL049839.2	AL049839.2
WNK1	WNK lysine deficient protein kinase 1
HSPB1	Heat shock protein family B (small) member 1
DNAJB1	DNAJ heat shock protein family (HSP40) member B1

Iranian Rehabilitation Journal

The protein produced by CD74 could have an intermediary role in survival pathways and cell proliferation [10]. According to a previous study, CD74 expression in B cells is linked to early MS disease activity. This study also looked at how CD74 regulation and downstream CD74 impact human B cell function [11].

A type III intermediate filament protein is encoded by Vimentin. This protein plays a role in neuritogenesis and cholesterol transport and also has a host cell role [12].

According to recent research, the 14-3-3 protein family found in the cerebrospinal fluid of patients, such as MS patients, indicates severe brain damage [13].

The component that complements C3 is crucial to activating the complement system. Both the traditional and the alternative complement activation routes require its activation. In human patients, mutations in this gene are linked to atypical hemolytic uremic syndrome and age-related macular degeneration [14].

Multiple types of demyelination and MS patients have synapse loss as a shared clinical characteristic [15]. The previous study was to see whether there was a link between C3 overexpression and synapse loss. As a result, complement C3 inhibition might be a potential treatment option for demyelinating disease [16].

Chitinases are enzymes that catalyze the hydrolysis of chitin, a glycopolymer prevalent in insect exoskeletons and fungal cell walls. Eight human members make up the glycoside hydrolase 18 families of chitinases. The protein encoded by this gene is hypothesized to be involved in the inflammatory and tissue remodeling processes [17].

Many different types of cells express CHI3L1, an extracellular monomeric single-chain glycoprotein. Patients with CNS inflammatory disorders have higher amounts of CHI3L1 in their cerebrospinal fluid [18]. In previous research, it was reported that the levels of CHI3L1 rose with age. This was supporting evidence that lower-grade inflammatory processes are produced in the aging brain [19].

Table 3. Differentially expressed pathways based on KEGG results

Pathway Name	Adjusted P	Genes
Alternative complement activation	0.002556	C3
Neutrophil degranulation	0.002919	CHI3L1, SERPINA3, C3
Activation of C3 and C5	0.004469	C3

Iranian Rehabilitation Journal

The protein encoded by SERPINA3 is serine protease inhibitors. Variations in the sequence of this protein have been linked to Alzheimer disease, and its shortage has been linked to liver illness [20].

In MS patients, SERPINA3 levels were considerably found. This may be has been linked to MS progression [21]. SERPINA3 has been related to hyperphosphorylation, which has been linked to neurodegeneration. These observations most probably indicate several functions of SERPINA3 that may play depending on the CNS's composition [22, 23].

The WNK protein kinases could play a considerable role in blood pressure control. Hereditary sensory neuropathy type II has been linked to mutations in this gene [24]. WNK1 is found in the CNS and is important in pathogenic nervous system signaling [25].

The encoded protein, related to HSP20, is essential for cell differentiation in a wide range of cell types, and its expression has been linked to poor clinical outcomes in a variety of human malignancies. Patients with distal hereditary motor neuropathy were reported to have mutations in this gene [26].

Small heat shock proteins (HSPB110) signaling molecules play important roles in neuroinflammation [27]. Gorter et al. showed that HSPBs are mostly upregulated in astrocytes in spinal cord MS lesions, similar to the brain, but that expression levels of numerous HSPBs are significantly greater in the spinal cord and significantly altered throughout lesion development [28].

DNAJ family members have rolled in a variety of physiological processes [29]. DNAJB1 is a member of the HSP family. This is differently expressed in all regions of the brain and, remarkably, it is always upregulated [30]. In addition, we detected an AL049839.2 domain-containing protein as a novel protein (ENSG00000273259). However, we did not evaluate the association between this novel protein function and MS.In addition, the predicted genes were predominantly implicated in several pathways, according to the KEGG pathway analysis, including alternative complement activation, neutrophil degranulation, and activation of C3 and C5.

The most important advantage of our study was that we developed the predictive gene signature from samples, an approach that has clinical applications. One limitation of this study was the unavailability of gene expression data, and we had to get it from public databases.

5. Conclusion

Understanding what causes MS will help researchers identify more effective treatments and cures for the disease, or even prevent it from occurring. This study recognizes 9 prognostic genes related to MS and also identified three biological pathways. These can provide more potent prognostic information for MS patients.

Ethical Considerations

Compliance with ethical guidelines

The Research Ethics Committee of the University of Social Welfare and Rehabilitation Sciences approved this study (Code: IR.USWR.REC.1399.071).

Funding

This study was financially supported by the University of Social Welfare and Rehabilitation Sciences.

Authors' contributions

Conceptualization and Supervision: Akbar Biglarian and Kolsoum Inanloo Rahatloo; Methodology: All authors; Investigation, Writing–original draft, and Writing–review & editing: All authors; Data collection: Taiebe Kenarangi; Data analysis: Taiebe Kenarangi and Akbar Biglarian; Funding acquisition and Resources: Taiebe Kenarangi, Akbar Biglarian, and Kolsoum Inanloo Rahatloo.

Conflict of interest

The authors declared no conflicts of interest.

Acknowledgments

The authors would like to appreciate the Deputy of Research of the University of Social Welfare and Rehabilitation Sciences for providing the research facilities.

References

[1] Pakpoor J, Ramagopalan S. Russell W Brain and the aetiology of multiple sclerosis: A historical perspective. QJM: An International Journal of Medicine. 2014; 107(6):423-7. [DOI:10.1093/qjmed/hcu001] [PMID]

[2] Haines JD, Inglese M, Casaccia P. Axonal damage in multiple sclerosis. The Mount Sinai journal of Medicine, New York. 2011; 78(2):231-43. [DOI:10.1002/msj.20246] [PMID] [PMCID]

[3] Goldenberg MM. Multiple sclerosis review. Pharmacy and Therapeutics. 2012; 37(3):175-84. [PMID] [PMCID]

[4] Ömerhoca S, Akkaş SY, İçen NK. Multiple sclerosis: Diagnosis and differential diagnosis. Archives of Neuropsychiatry. 2018; 55(Suppl 1):S1-9. [DOI:10.29399/npa.23418] [PMID] [PMCID]

[5] Patsopoulos NA, De Jager PL. Genetic and gene expression signatures in multiple sclerosis. Multiple Sclerosis Journal. 2020; 26(5):576-81. [DOI:10.1177/1352458519898332] [PMID]

[6] Olek M. Epidemiology, risk factors and clinical features of multiple sclerosis in adults. Epidemiology and Clinical Features of Multiple Sclerosis in Adults Accessed October. 2011; 31. [Link]

[7] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10(1):57-63. [DOI:10.1038/nrg2484] [PMID] [PMCID]

[8] Chernoff H. Cluster analysis for applications (Michael R. Anderberg). SIAM Review. 1975; 17(3):580-2. [DOI:10.1137/1017065]

[9] Didonna A, Oksenberg JR. The genetics of multiple sclerosis. In: Didonna A, Oksenberg JR, Zagon I S, McLaughlin P J, editors. Multiple sclerosis: Perspectives in treatment and pathogenesis. Brisbane: Codon Publications. 2017. [DOI:10.15586/codon.multiplesclerosis.2017.ch1] [PMID]

[10] Schröder B. The multifaceted roles of the invariant chain CD74-more than just a chaperone. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2016; 1863(6):1269-81. [DOI:10.1016/j.bbamcr.2016.03.026] [PMID]

[11] Rijvers L, Melief MJ, van der Vuurst de Vries RM, Stéphant M, van Langelaar J, Wierenga-Wolf AF, et al. The macrophage migration inhibitory factor pathway in human B cells is tightly controlled and dysregulated in multiple sclerosis. European Journal of Immunology. 2018; 48(11):1861-71. [DOI:10.1002/eji.201847623] [PMID] [PMCID]

[12] Kondo T, Takahashi M, Yamasaki G, Sugimoto M, Kuse A, Morichika M, et al. Immunohistochemical analysis of vimentin expression in myocardial tissue from autopsy cases of ischemic heart disease. Legal Medicine. 2022; 54:102003. [DOI:10.1016/j.legalmed.2021.102003]

[13] Satoh J-i, Yamamura T, Arima K. The 14-3-3 protein ε isoform expressed in reactive astrocytes in demyelinating lesions of multiple sclerosis binds to vimentin and glial fibrillary acidic protein in cultured human astrocytes. The American Journal of Pathology. 2004; 165(2):577-92. [DOI:10.1016/S0002-9440(10)63322-6]

[14] Zipfel PF, Skerka C, Chen Q, Wiech T, Goodship T, Johnson S, et al. The role of complement in C3 glomerulopathy. Molecular Immunology. 2015; 67(1):21-30. [DOI:10.1016/j.molimm.2015.03.012] [PMID]

[15] Werneburg S, Jung J, Kunjamma RB, Ha S-K, Luciano NJ, Willis CM, et al. Targeted complement inhibition at synapses prevents microglial synaptic engulfment and synapse loss in demyelinating disease. Immunity. 2020; 52(1):167-82. e7. [DOI:10.1016/j.immuni.2019.12.004] [PMID] [PMCID]

[16] Xin W, Chan JR. That wasn't a complement-too much C3 in demyelinating disease. Immunity. 2020; 52(1):11-3. [DOI:10.1016/j.immuni.2019.12.014] [PMID] [PMCID]

[17] Ohi K, Hashimoto R, Yasuda Y, Yoshida T, Takahashi H, Iike N, et al. The chitinase 3-like 1 gene and schizophrenia: Evidence from a multi-center case-control study and meta-analysis. Schizophrenia Research. 2010; 116(2-3):126-32. [DOI:10.1016/j.schres.2009.12.002] [PMID]

[18] Kušnierová P, Zeman D, Hradílek P, Zapletalová O, Stejskal D. Determination of chitinase 3-like 1 in cerebrospinal fluid in multiple sclerosis and other neurological diseases. Plos one. 2020; 15(5):e0233519. [DOI:10.1371/journal.pone.0233519] [PMID] [PMCID]

[19] Bonneh-Barkay D, Wang G, Starkey A, Hamilton RL, Wiley CA. In vivo CHI3L1 (YKL-40) expression in astrocytes in acute and chronic neurological diseases. Journal of Neuroinflammation. 2010; 7(1):1-8. [DOI:10.1186/1742-2094-7-34] [PMID] [PMCID]

[20] Yuan Q, Wang SQ, Zhang GT, He J, Liu ZD, Wang MR, et al. Highly expressed of SERPINA3 indicated poor prognosis and involved in immune suppression in glioma. Immunity, Inflammation and Disease. 2021; 9(4):1618-30. [DOI:10.1002/iid3.515] [PMID] [PMCID]

[21] Fissolo N, Matute-Blanch C, Osman M, Costa C, Pinteac R, Miró B, et al. CSF SERPINA3 levels are elevated in patients with progressive MS. Neurology-Neuroimmunology Neuroinflammation. 2021; 8(2):e941. [DOI:10.1212/NXI.0000000000000941] [PMID] [PMCID]

[22] Padmanabhan J, Levy M, Dickson DW, Potter H. Alpha1-antichymotrypsin, an inflammatory protein overexpressed in Alzheimer's disease brain, induces tau phosphorylation in neurons. Brain. 2006; 129(11):3020-34. [DOI:10.1093/brain/awl255] [PMID]

[23] Aslam MS, Yuan L. Serpina3n: Potential drug and challenges, mini review. Journal of Drug Targeting. 2020; 28(4):368-78. [DOI:10.1080/1061186X.2019.1693576] [PMID]

[24] Newhouse S, Farrall M, Wallace C, Hoti M, Burke B, Howard P, et al. Polymorphisms in the WNK1 gene are associated with blood pressure variation and urinary potassium excretion. Plos One. 2009; 4(4):e5003. [DOI:10.1371/journal.pone.0005003] [PMID] [PMCID]

[25] Krueger EM, Miranpuri GS, Resnick DK. Emerging role of WNK1 in pathologic central nervous system signaling. Annals of Neurosciences. 2011; 18(2):70-5. [DOI:10.5214/ans.0972.7531.1118212] [PMID] [PMCID]

[26] Houlden H, Laura M, Wavrant-De Vrièze F, Blake J, Wood N, Reilly M. Mutations in the HSP27 (HSPB1) gene cause dominant, recessive, and sporadic distal HMN/CMT type 2. Neurology. 2008 ;71(21):1660-8. [DOI:10.1212/01.wnl.0000319696.14225.67] [PMID]

[27] Dulle JE, Fort PE. Crystallins and neuroinflammation: The glial side of the story. Biochimica et Biophysica Acta (BBA)-General Subjects. 2016; 1860(1):278-86. [DOI:10.1016/j.bbagen.2015.05.023] [PMID] [PMCID]

[28] Gorter RP, Nutma E, Jahrei MC, de Jonge JC, Quinlan R, van der Valk P, et al. Heat shock proteins are differentially expressed in brain and spinal cord: Implications for multiple sclerosis. Clinical & Experimental Immunology. 2018; 194(2):137-52. [DOI:10.1111/cei.13186] [PMID] [PMCID]

[29] Hata M, Okumura K, Seto M, Ohtsuka K. Genomic cloning of a human heat shock protein 40 (Hsp40) gene (HSPF1) and its chromosomal localization to 19p13. 2. Genomics. 1996; 38(3):446-9. [DOI:10.1006/geno.1996.0653] [PMID]

[30] Chiricosta L, Gugliandolo A, Bramanti P, Mazzon E. Could the heat shock proteins 70 family members exacerbate the immune response in multiple sclerosis? An in silico study. Genes. 2020; 11(6):615. [DOI:10.3390/genes11060615] [PMID] [PMCID]