Identification of Blood Based Bio-Marker for Huntington's disease using in-silico gene expression analysis.

Souvik Chakraborty

Department of Physiology

Bhairab Ganguly College

West Bengal,

India.

Abstract

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder with profound phenotypic characters. HD is at present incurable and there are several trials going on to find a cure. HD is caused when there is a mutation in the Huntingtin gene which is found to be associated with axonal transport. Diagnosis is based on the signs and symptoms of the patients but by that time the psychomotor problems have already reached the level from where reversing the disease is impossible. Blood based biomarkers can be used for the diagnosis of the disease at an early stage. In this study several gene expression study data were analyzed and there were 329 Differentially Expressed Genes (DEGs) in all the three chosen datasets. Protein protein interaction network was created using STRING and CytoHubba plug-in was used to identify top ten hub genes which are CXCL8, PSMC6, UBE2D1, UBE2D1, CD27, UBE2D3, SF3B1, CASP3, EIF4E, BIRC2 and PTEN. Online software Enrichr was used for Gene Ontology and KEGG pathway enrichment analysis to find out the biological process, molecular function, cellular component and the pathways that were enriched in HD. This study finds out that those genes which were present in all the three datasets namely FNDC3A, BCLAF1 and ALCAM were not the hub genes. So further studies are required for identifying a potential biomarker of HD.

Introduction

Huntington's disease (HD) is a neurodegenerative disease which primarily affects motor system of our body and is associated with many phenotypic characteristics such as athetosis, chorea etc.[1] HD affects approximately 1/20,000 people and the average age at the onset of the disease ranges from 30 – 50 years.[2] The neurodegenerative disorder affects a variety of regions within the human brain such as globus pallidus, cerebral cortex, substantia nigra, subthalamic nuclei, hypothalamus and also cerebellum in the advanced stage of the disease but the most devastating effect of this disease is seen in striatum which is a part of basal ganglia.[3] The symptoms ranges from uncontrolled spasmodic movements called Huntington's chorea or simply put chorea to gait disturbances. Some people also experience behavior alterations, cognitive impairment, dysphagia in addition to the above stated symptoms.[4] At present diagnosis of HD depends on the arrival of symptoms but by the time the disease has been detected, it has advanced to an irreversible stage. HD is a genetic disease and is diagnosed with the help of genetic screening test. Presently the treatment of HD is symptomatic which can delay the disease but cannot completely reverse the course of HD.[5]

Potential biomarkers for HD is the need of the hour. At present there are several biomarkers used for the detection of HD & they are -

(i) Clinical biomarkers – determined using Anti saccade error rate[6], Digitomotography[7].

(ii) Imaging – MRI[7], MNI-659 PET[8]

(iii) Immune biomarkers – IL8[9]

(iv) Neurodegeneration biomarker – NfL[10]

Blood based biomarkers are going to play a key role in early detection of the disease. Blood based biomarkers are going to play a key role in the early detection of HD because the starting material (in this case blood) is easier to collect and requires less expertise for the collection procedure.

Currently there is lack of exact blood-based biomarker for HD. Besides early detection of HD, blood biomarkers open up new window for designing novel therapeutic drugs in the future which can reverse HD completely.

Blood based biomarkers can be used profiled using many ways such as RNA-seq and microarray data analysis. Differential gene expression analysis from blood samples of HD

patients and normal controls can give us valuable information regarding biomarkers as well as the pathways associated with the disease. In one of the study using RNA-seq data it was found out that there were 5genes (PROK2, ZNF238, AQP9, CYSTM1 and ANXA3) that were significantly expressed more than other genes in HD samples.[11]

Microarray data were used for finding biomarker for disease like Alzheimer's disease.[12] Although gene expression profiling is an useful method but they have their own cons. Many incongruities occur during gene expression profiling which may occur due to designing of the experiment, storage method of blood samples and age of blood samples which directly effect the amount of RNA contained within the sample.[13] So there is high chance that the genes which are expressed significantly in one study may not be expressed to that extent in other and therefore it is highly required that multiple datasets are studied at the same time which reduces this risk.

In the present study, three microarray datasets were obtained from gene expression omnibus database and are then analyzed for the purpose of finding a potential biomarker for HD.

Methods

Dataset Selection

Gene Expression Omnibus (GEO) by National Center for Biotechnology Information (NCBI) is an internationally acclaimed online database for high-throughput sequencing data, micro array, hybridization array data.[14] GEO was searched for datasets in which keywords like "Huntington's", "blood", "expression profiling by array", "*Homo sapiens*" were present. There were three datasets which match the above criteria and their GEO accession numbers are GSE1751, GSE1767, GSE24250.[15,16] All the datasets which are used in this study are available online and no actual human experiments were performed by the author. In this study total 32 patients samples and 34 normal control subjects were present. All the three studies used microarrays for data collection. Microarray is a chip based technology in which 1000s of nucleic acids are bound to a surface and are used to measure relative concentration of nucleic acid sequences in a mixture.[17]

Datasets	Number of HD samples	Number of control samples
GSE1751	12	14
GSE1767	12	14
GSE24250	8	6
Total	32	34

Table 1- Dataset information for Microarray datasets obtained from GEO. There are in total 12 HD samples and 14 normal control samples in the dataset GSE1751, 12 HD samples and 14 normal control samples in the dataset GSE1767 and 8 HD samples and 6 normal control samples in the dataset GSE24250.

Differential Gene Expression Analysis

The above 3 datasets were analyzed using GEO2R tool (available at http://www.ncbi.nlm.nih.gov/geo/geo2r/) which is provided by the Gene Expression Omnibus.GEO2R is an online software that processes gene expression values and outputs a table of two differentially expressed genes (DEGs) between two user defined groups (HD and control in this case). After running GEO2R, several thousand genes were found to be significant. p values were found out using 't' test. Fold changes of each gene was calculated by the help of GEO2R tool. Fold change is represented by logFC values which tells us about the magnitude of gene expression change. So a value of 2.5 is 2 to the power 2.5 (2^2.5). This means that levels of gene expression for this gene are 5 times higher in patients than the normal subjects.

GEO2R tool was used for verifying a normal distribution of samples. No outliers were present in these datasets.

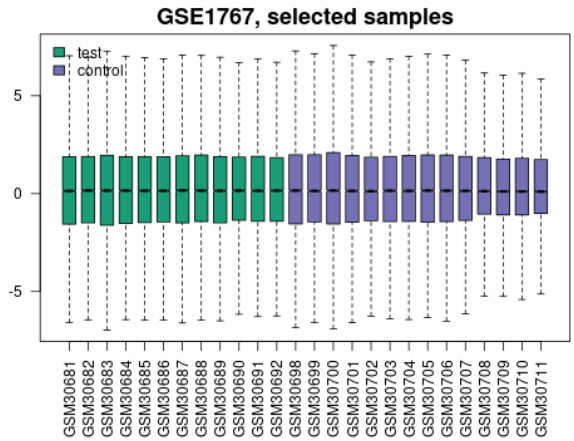


Figure 1- Gene expression value distribution for dataset GSE1767.There are no units for in y axis. Each box plot reperesents gene expression value of one patient sample.

After obtaining the list of DEGs, they were processed in Excel. DEGs with p value greater than 0.05 were removed and the rest of the DEGs were sorted into two categories, which are overexpressed and those which are underexpressed. Some of the genes were overexpressed in HD while some are underexpressed. In this study only those genes were selected which have a fold change value greater than 1.2.

Identification of Common Genes

Desktop based tool Fun Rich was used for creating Venn diagram in order to find out the common genes among all three datasets.[18]

Network Analysis

The final list of DEGs were analyzed by the help of STRING (available at https://string-db.org/) which is used to construct a network of proteins which interact with each other. STRING is a search tool for interacting proteins and also calculates statistically significant Gene Ontology processes and pathways.[19] A tab separated value file was obtained from STRING and was then imported to Cytoscape. Cytoscape is an open source desktop based software used for integrating biomolecular interaction networks with the high-throughput expression data.[20] Network analyzer, a Cytoscape plug-in CytoHubba was used for the identification of hub genes.

Enrichment Analysis of DEGs

Gene enrichment analysis was performed for DEGs using the online software called Enrichr (available at https://maayanlab.cloud/Enrichr/) for Gene Ontology(GO) function and also for KEGG pathway enrichment. GO terms were further subdivided into several terms such as Biological Process, Molecular Function and Cellular Component.

Results

Three gene expression profiles (GSE 1751, GSE 1767 and GSE 24250) were analyzed with the help of GEO2R tool and the list of DEGs were downloaded. p value was set at 0.05 and a log FC value of 1.2 was set as the cut off criteria. There were 903 genes in the dataset GSE 1751, 1168 genes in the GSE 1767 and 229 genes passed the above criteria in the dataset GSE 24250. Venn analysis was performed using the software Fun Rich and the DEGs were identified. 329 genes were found to be significantly differentially expressed among all the three above mentioned datasets. In these datasets there were three genes which are differentially expressed in all the datasets and the name of those genes are FNDC3A, BCLAF1 and ALCAM. FNDC3A (Fibronectin type-III domain-containing protein 3A) is a protein coding gene which produces a transmembrane protein present in human odontoblast and dental pulp nerves and to a much lesser extent in human brain, kidney, liver and lungs.[21] Not much is known with respect to the function of this protein. BCLAF1 (Bcl-2-associated transcription factor 1) is a transcription factor which is associated with several biological process such as apoptosis, negative as well as positive regulation of transcription, regulation of DNA template regulated transcription in response to stress.[22,23] ALCAM (Activated Leukocyte Cell Adhesion Molecule) encodes a protein called CD166 which is a transmembrane glycoprotein which is a cell adhesion molecule which mediates heterotypic cell adhesion, T cell activation, normal cell enfragment in the bone marrow.[24,25]

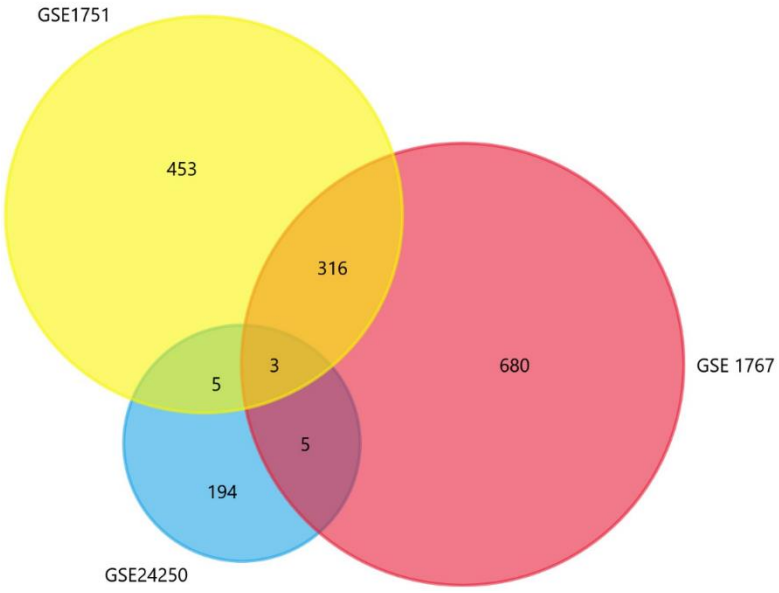


Figure 2-Venn diagram representing the DEGs in all the three datasets (GSE 1751, GSE 1767 and GSE 24250).

The 329 genes which passed the above cut-off criteria were submitted to the STRING tool. The disconnected genes were hidden and medium confidence option was selected in the minimum required interaction score option. Total of 326 nodes and 964 edges are present in the network(Figure 3). The online tool Enrichr was used for performing GO and KEGG pathway enrichment analysis. After submitting the list of DEGs the the ontology option in the Enrichr gave several options such as GO Biological Process (GO-BP), GO Molecular Function (GO -MF), GO Cellular Component (GO-CC). GO analysis showed that the DEGs were enriched in GO-BP such as negative regulation of programmed cell death, regulation of transcription from RNA polymeraseII promoter in response to hypoxia, negative regulation of apoptosis, toll like receptor signalling pathway, MYD88-independent toll like receptor signalling pathway,regulation of apoptotic process,regulation of cytokine mediated signaling pathway, activation of cysteine-type endopeptidase activity invoved in apoptotic signaling pathway(Fig 4A). GO-MF analysis showed that DEGs were specially enriched in cadherin bingding, RNA binding, kinase binding, protein heterodimerization activity, cysteine type endopeptidase inhibitor activity involved in apoptostic process, protein kinase binding, nucleoside triphosphatse activity, metal ion transmembrane activity, purine ribonucleoside triphosphate binding and annealing acitivity(Fig 4B). For GO-CF analysis, the DEGs were mainly enriched in nuclear body, nuclear speck, nucleolus, cytosolic part, cytosolic ribosome, cyclin/CDK positive transcription elongation factor complex, chromosome, cytosolic small ribosomal subunit, ribosome(Fig 4C). The KEGG pathway analysis showed that DEG are enriched in pathways which are associated with NOD-like receptor signalling pathway, Kaposi sarcoma-associated herpesvirus infection, Small cell lung cancer,Nf-kappa B signalling pathway, Apoptosis, paltelet activation, phosphatidyl inositol signalling system, several pathways in cancer and toxoplasmosis and also Human cytomegalovirus infection(Fig 4D). For finding out the hub gene, the genes with degree greater than equal to 15 were selected from the list of DEGs and was submitted to the Cytoscape offline tool.In the cytohubba plug-in it was found out that Phosphatse and tensin homolog has the highest connectivity with other genes (PTEN, degree = 18) followed by Interleukin 8 (CXCL8, degree = 18), 26s proteasome regulatory subunit 10B (PSMC6, degree = 15), Caspase 3 (CASP3, degree = 15), Ubiquitin Conjugating Enzyme E2D1 (UBE2D1, degree = 14), CD27 (CD27, degree = 14), Splicing factor 3b subunit 1 (SF3B1, degree = 14), Eukaryotic translation initiation factor 4E (EIF4E, degree = 14), Baculoviral IAP repeat-containing protein 2 (BIRC2, degree = 14) (Figure 5). These above stated genes are upregulated in HD.

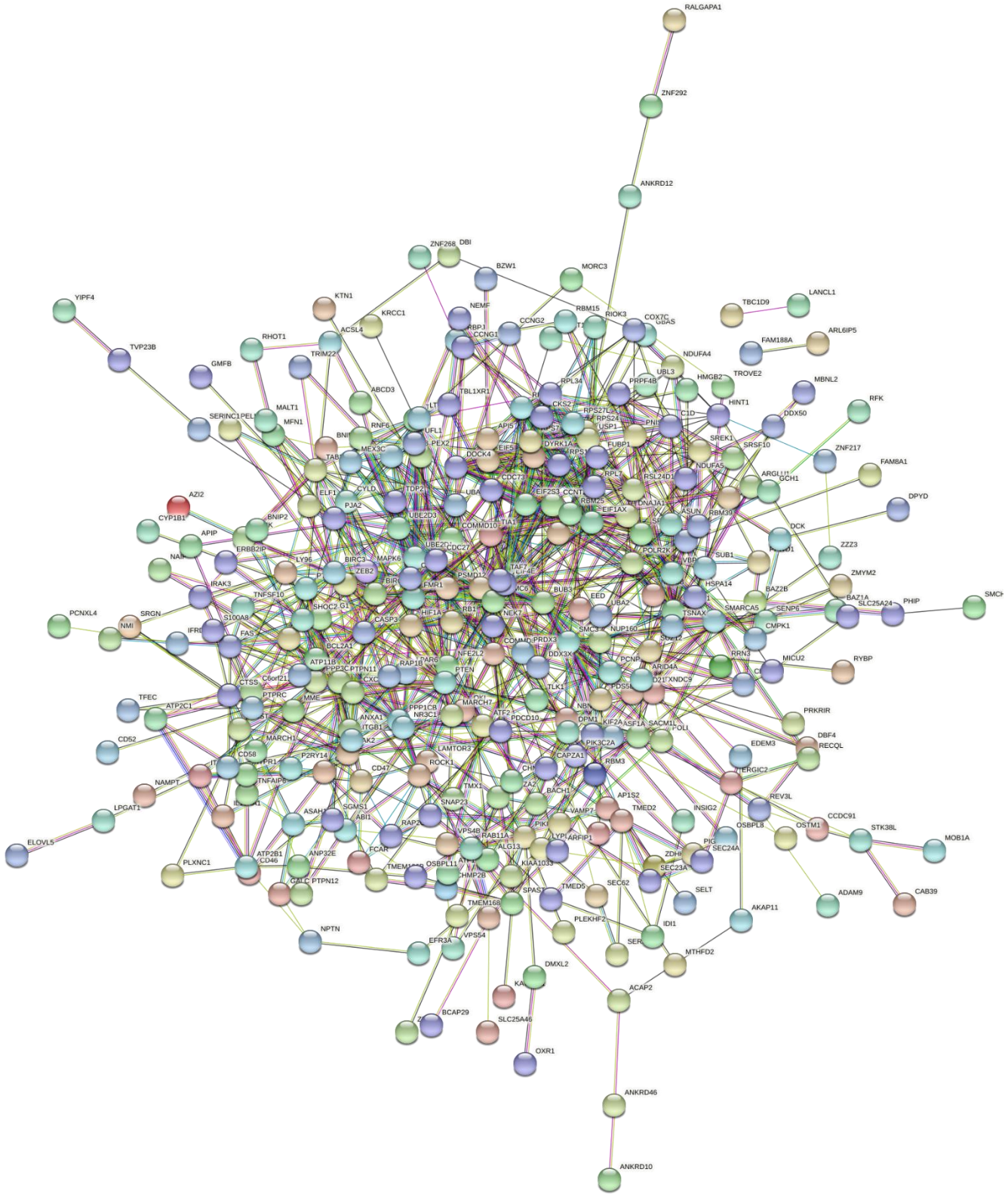
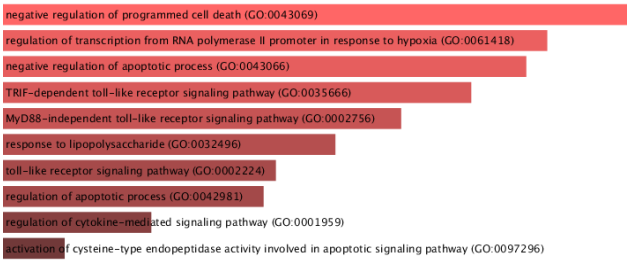
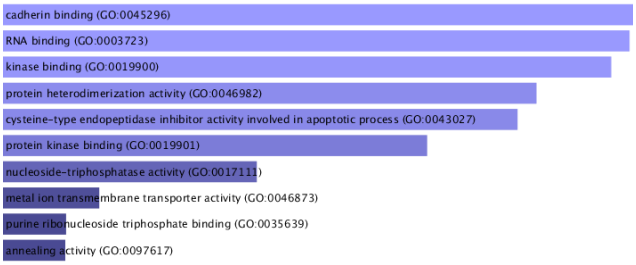


Figure 3 – STRING protein protein interaction network. In this network there are 326 nodes and 964 edges. Lines represents interaction and the circles represent genes. The colour of the lines represents the type of interaction.

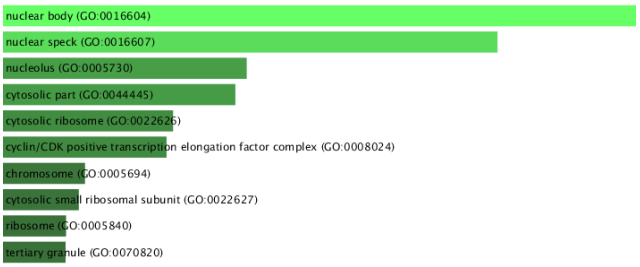
A



B



C



D

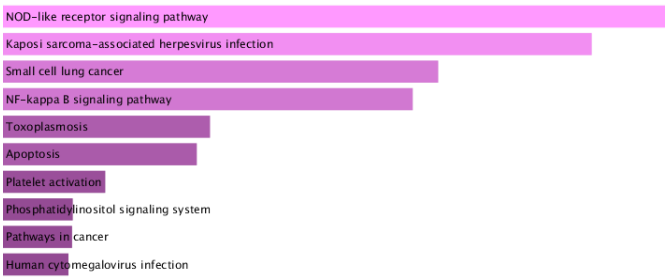


Figure 4 – GO and KEGG pathway enrichment analysis using Enrichr. A- Top 10 enriched biological processes in DEGs. The x axis represents number of genes and the y axis represents the biological process. B- Top 10 enriched molecular function in DEGs. The x axis represents number of genes and the y axis represents molecular function. C- Top 10 enriched cellular components in DEGs. The x axis represents the number of genes and the y axis represents cellular components. D – Top 10 enriched KEGG pathways for DEGs. The x axis represents the number of genes and the y axis represents the KEGG pathway names.

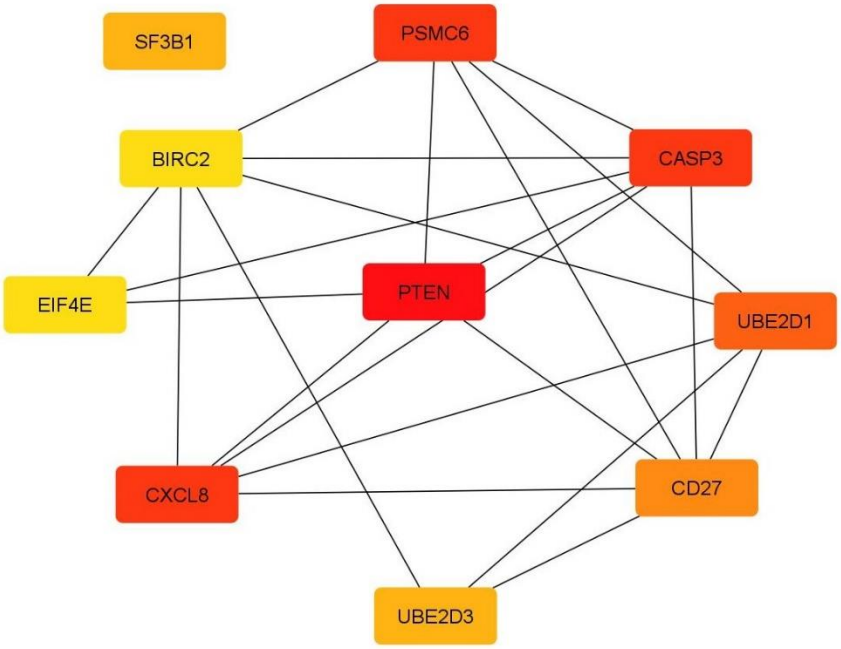


Figure 5 – Network of top ten hub genes using Cytoscape software. The colour represents the degree of connectivity, red colour represents the highest degree, the orange colour represent intermediate degree and the yellow represent the lowest degree.

Discussion

Huntington's disease is a neurodegenerative disease associated with many cellular changes and in this study some key genes were identified which were earlier not reported. Genes such as CXCL8, PSMC6, UBE2D1, UBE2D1, CD27, UBE2D3, SF3B1were previously not known to regulate HD. Similarly there were some genes which are already known to regulate HD such as CASP3, EIF4E, BIRC2.[26–28] Upregulation PTEN occurs in several neurodegenerative diseases.[29] Several pathways, biological functions associated with the above mentioned upregulated genes were also enriched in the KEGG and GO enrichment analysis. However the three genes namely FNDC3A, BCLAF1 and ALCAM which were present in all the three datasets were not the hub genes.

Limitations of this study includes that there was no data normalization in this study and no distinction has been made between male and female patients as well as normal controls where there can be significant difference in differentially expressed genes with respect to gender. Therefore further study on this subject is required so that new potential blood based biomarkers of HD can be found.

References

1. Walker FO. Huntington's disease. *The Lancet*. 2007;369(9557):218-228. doi:10.1016/S0140-6736(07)60111-1

2. Roos RA. Huntington's disease: a clinical review. *Orphanet J Rare Dis*. 2010;5(1):40. doi:10.1186/1750-1172-5-40

3. Eidelberg D, Surmeier DJ. Brain networks in Huntington disease. *J Clin Invest*. 2011;121(2):484-492. doi:10.1172/JCI45646

4. Myers RH. Huntington's disease genetics. *Neurotherapeutics*. 2004;1(2):255-262. doi:10.1602/neurorx.1.2.255

5. Ghosh R, Tabrizi SJ. Clinical Features of Huntington's Disease. In: Nóbrega C, Pereira de Almeida L, eds. *Polyglutamine Disorders*. Vol 1049. Advances in Experimental Medicine and Biology. Springer International Publishing; 2018:1-28. doi:10.1007/978-3-319-71779-1_1

6. Tabrizi SJ, Langbehn DR, Leavitt BR, et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *The Lancet Neurology*. 2009;8(9):791-801. doi:10.1016/S1474-4422(09)70170-X

7. Tabrizi SJ, Scahill RI, Owen G, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *The Lancet Neurology*. 2013;12(7):637-649. doi:10.1016/S1474-4422(13)70088-7

8. Russell DS, Barret O, Jennings DL, et al. The Phosphodiesterase 10 Positron Emission Tomography Tracer, [^{18} F]MNI-659, as a Novel Biomarker for Early Huntington Disease. *JAMA Neurol*. 2014;71(12):1520. doi:10.1001/jamaneurol.2014.1954

9. Björkqvist M, Wild EJ, Thiele J, et al. A novel pathogenic pathway of immune activation detectable before clinical onset in Huntington's disease. *Journal of Experimental Medicine*. 2008;205(8):1869-1877. doi:10.1084/jem.20080178

10. Rodrigues FB, Byrne LM, McColgan P, et al. Cerebrospinal Fluid Inflammatory Biomarkers Reflect Clinical Severity in Huntington's Disease. Blum D, ed. *PLoS ONE*. 2016;11(9):e0163479. doi:10.1371/journal.pone.0163479

11. Mastrokolias A, Ariyurek Y, Goeman JJ, et al. Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *Eur J Hum Genet*. 2015;23(10):1349-1356. doi:10.1038/ejhg.2014.281

12. Chen K-D, Chang P-T, Ping Y-H, Lee H-C, Yeh C-W, Wang P-N. Gene expression profiling of peripheral blood leukocytes identifies and validates ABCB1 as a novel biomarker for Alzheimer's disease. *Neurobiology of Disease*. 2011;43(3):698-705. doi:10.1016/j.nbd.2011.05.023

13. Chung J-Y, Braunschweig T, Williams R, et al. Factors in Tissue Handling and Processing That Impact RNA Obtained From Formalin-fixed, Paraffin-embedded Tissue. *J Histochem Cytochem*. 2008;56(11):1033-1042. doi:10.1369/jhc.2008.951863

14. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*. 2012;41(D1):D991-D995. doi:10.1093/nar/gks1193

15. Borovecki F, Lovrecic L, Zhou J, et al. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci U S A*. 2005;102(31):11023-11028. doi:10.1073/pnas.0504921102

16. Hu Y, Chopra V, Chopra R, et al. Transcriptional modulator H2A histone family, member Y (H2AFY) marks Huntington disease activity in man and mouse. *Proc Natl Acad Sci U S A*. 2011;108(41):17141-17146. doi:10.1073/pnas.1104409108

17. Bumgarner R. Overview of DNA Microarrays: Types, Applications, and Their Future. In: Ausubel FM, Brent R, Kingston RE, et al., eds. *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc.; 2013:mb2201s101. doi:10.1002/0471142727.mb2201s101

18. Pathan M, Keerthikumar S, Ang C-S, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics*. 2015;15(15):2597-2601.

19. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

20. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

21. Carrouel F, Couble M-L, Vanbelle C, Staquet M-J, Magloire H, Bleicher F. *HUGO* (*FNDC3A*): a New Gene Overexpressed in Human Odontoblasts. *J Dent Res*. 2008;87(2):131-136. doi:10.1177/154405910808700209

22. Kasof GM, Goyal L, White E. Btf, a Novel Death-Promoting Transcriptional Repressor That Interacts with Bcl-2-Related Proteins. *Mol Cell Biol*. 1999;19(6):4390-4404. doi:10.1128/MCB.19.6.4390

23. Liu H, Lu Z-G, Miki Y, Yoshida K. Protein Kinase C δ Induces Transcription of the TP53 Tumor Suppressor Gene by Controlling Death-Promoting Factor Btf in the Apoptotic Response to DNA Damage. *MCB*. 2007;27(24):8480-8491. doi:10.1128/MCB.01126-07

24. Hassan NJ, Barclay AN, Brown MH. Frontline: Optimal T cell activation requires the engagement of CD6 and CD166. *Eur J Immunol*. 2004;34(4):930-940. doi:10.1002/eji.200424856

25. Chitteti BR, Kobayashi M, Cheng Y, et al. CD166 regulates human and murine hematopoietic stem cells and the hematopoietic niche. *Blood*. 2014;124(4):519-529. doi:10.1182/blood-2014-03-565721

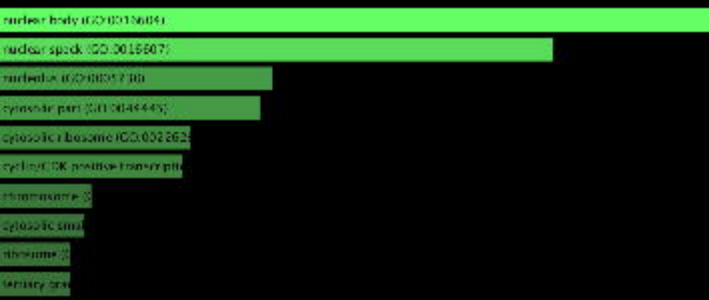
26. Zhang Y, Leavitt BR, van Raamsdonk JM, et al. Huntingtin inhibits caspase-3 activation. *EMBO J*. 2006;25(24):5896-5906. doi:10.1038/sj.emboj.7601445

27. Goffredo D, Rigamonti D, Zuccato C, Tartari M, Valenza M, Cattaneo E. Prevention of cytosolic IAPs degradation: a potential pharmacological target in Huntington's Disease. *Pharmacological Research*. 2005;52(2):140-150. doi:10.1016/j.phrs.2005.01.006

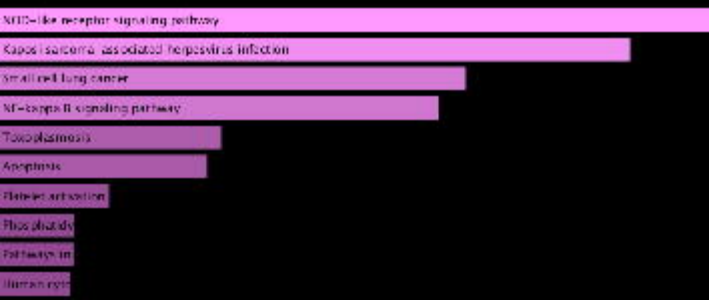
28. Creus-Muncunill J, Badillos-Rodríguez R, Garcia-Forn M, et al. Increased translation as a novel pathogenic mechanism in Huntington's disease. *Brain*. 2019;142(10):3158-3175. doi:10.1093/brain/awz230

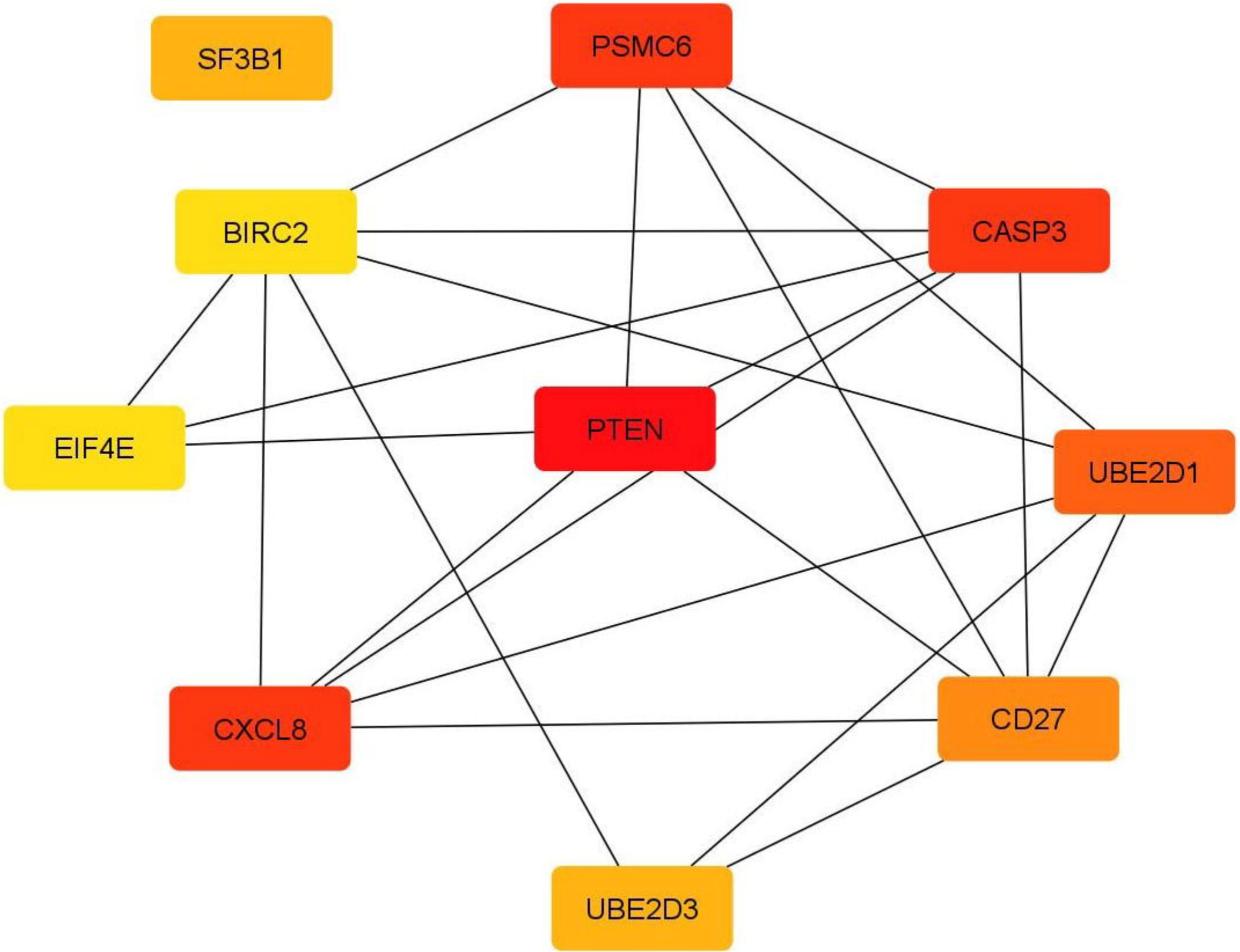
29. Ismail A, Ning K, Al-Hayani A, Sharrack B, Azzouz M. PTEN: A molecular target for neurodegenerative disorders. *Translational Neuroscience*. 2012;3(2). doi:10.2478/s13380-012-0018-9


negative regulation of programmed cell death (GO:0043069)
regulation of transcription from RNA polymerase II promoter in response to hypoxia (GO:0061418)
negative regulation of apoptotic process (GO:0043066)
TRIF-dependent toll-like receptor signaling pathway (GO:0035666)
MyD88- independent toll-like receptor signaling pathway (GO:0002756)
response to lipopolysaccharide (GO:0032496)
toll-like receptor signaling pathway (GO:0002224)
regulation of apoptotic process (GO:0042981)
regulation of cytokine-med
activation of


nuclear body (GO:0016604)
nuclear speck (GO:0016607)
nucleolus (GO:0005730)
cytosolic part (GO:0044445)
cytosolic ribosome (GO:0022626)
cytosolic small ribosomal subunit
ribosome
tertiary granule


cadherin binding (GO:0045296)
RNA binding (GO:0003723)
kinase binding (GO:0019900)
protein heterodimerization activity (GO:0046982)
cysteine-type endopeptidase inhibitor activity involved in apoptotic process (GO:0043027)
protein kinase binding (GO:0019901)
nucleoside-triphosphatase activity (GO:0017111)
metal ion transport
purine nucleotide
unfolded p


NOD-like receptor signaling pathway
Kaposi sarcoma-associated herpesvirus infection
Small cell lung cancer
NF-kappa B signaling pathway
Toxoplasmosis
Apoptosis
Platelet activation
Phosphatidy
Pathways in
Human cyto


SF3B1
PSMC6
BIRC2
CASP3
EIF4E
PTEN
UBE2D1
CXCL8
CD27
UBE2D3