Potential Key Genes Associated with Stroke types and its subtypes: A Computational Approach

Gourab Das[1], Pradeep Kumar[2*]

[1]Department of Pediatrics, Army Hospital Research & Referral, New Delhi-110010, India
[2]Department of Neurology, All India Institute of Medical Sciences, New Delhi-110029, India

Author Contributions:
GD was involved in data extraction, computational work and manuscript writing and PK was involved in conceptualization and writing manuscript upto finalize version of it.

Conflict of Interest
No potential conflict of interest

Data Availability
All datasets generated for this study are included in the manuscript and/or the Supplementary Files.

Funding Source
None

Corresponding Author:

Dr. Pradeep Kumar
Senior Research Officer
Department of Neurology
All India Institute of Medical Sciences
New Delhi-110029, India
E-mail: pradeepguptaneuro@gmail.com
pradeepguptaneuro@aiims.edu
Mobile No-+91-9910128469

Potential Key Genes Associated with Stroke types and its subtypes: A Computational Approach

Abstract

To investigate prospective key genes and pathways associated with the pathogenesis and prognosis of stroke types along with subtypes. Human genes using genome assembly build 38 patch release 13 with known gene symbols through NCBI gene database (https://www.ncbi.nlm.nih.gov/gene) were fetched. PubMed advanced queries were constructed using stroke-related keywords and associations were calculated using Normalized pointwise mutual information (nPMI) between each gene symbol and queries. Genes related with stroke risk within their types and subtypes were investigated in order to discover genetic markers to predict individuals who are at the risk of developing stroke with their subtypes. A total of 2,785 (9.4%) genes were found to be linked to the risk of stroke. Based on stroke types, 1,287 (46.2%) and 376 (13.5%) genes were found to be related with IS and HS respectively. Further stratification of IS based on TOAST classification, 86 (6.6%) genes were confined to Large artery atherosclerosis; 131 (10.1%) and 130 (10%) genes were related with the risk of small vessel disease and Cardioembolism subtypes of IS. Besides, a prognostic panel of 9 genes signature consisting of CYP4A11, ALOX5P, NOTCH, NINJ2, FGB, MTHFR, PDE4D, HDAC9, and ZHFX3 can be treated as a diagnostic marker to predict individuals who are at the risk of developing stroke with their subtypes.

Keywords: Stroke; Ischemic Stroke; Intracerebral Hemorrhage; Subtypes; TOAST Classification; Cerebrovascular Disease

Introduction

Stroke is a complex heterogeneous disorder that occurs due to the interaction between environmental and genetic risk factors.[1, 2] It is one of the main important causes of mortality and long-term disability worldwide.[3] About 85% of stroke cases are ischemic stroke (IS), whereas 15% are hemorrhagic stroke (HS).[4] According to the Trial of Org 10172 in Acute Stroke Treatment (TOAST) classification; IS has been categorized according to the presumed etiological mechanism into five groups: large artery atherosclerosis (LAA), small vessel disease (SVD), cardio-embolic disease (CE), other determined etiology (ODE), and undetermined etiology (UDE).[5] Furthermore, HS is categorized into intracerebral hemorrhage (ICH) and Subarachnoid hemorrhage (SAH). Despite recent advancements in treatment modality, very few are known regarding the essential pathophysiology of stroke, and further research is still warranted to elucidate mechanisms in order to identify stroke occurrences.

Several established risk factors including diabetes, hypertension, dyslipidemia, smoking, atrial fibrillation, and obesity have been a link to the happening of stroke.[6] The fraction of strokes of undermined or rare causes is greater for young adults as compared to elders, and in many cases, underlying causes are genetic related. More than hundreds of genes have been described to be linked with the risk of stroke.[7, 8] Unravelling the genetic causes that play an important role in IS and HS is very challenging, as the genetic part of it is multifaceted.[9] In most cases, numerous genes are likely involved in the pathogenesis of stroke performing on a broad variety of candidate pathways, such as inflammatory, haemostatic, renin-angiotensin-aldosterone, and homocysteine metabolisms.[10, 11]

The genetic constituent is more predominant in LAA subtypes of IS than in SVD or cryptogenic IS and in patients younger than 70 years of age.[12] Previously published multicentric genetic studies using genome-wide data estimated that 40% for LAA, 33% for CE, 16% for SVD, and 38% for combined (Determined plus undetermined) etiology comprises the heritability of IS.[7, 13, 14] with the illustration that some genetic variants may serve as causal markers for stroke. To recognize the role of particular risk elements in regulating the pathophysiology of stroke, the hereditary basis of every risk factor is desirable to be examined and integrated, in context to their biological role and pathway interactions. To date, there are no well-established genetic markers that may discriminate the stroke types as well as their subtypes.

Identifying novel diagnostic and prognostic genetic markers has become an urgent demand. But its experimental determination remains a costly and time-consuming process. Hence, novel computational methods are needed to fulfil this requirement. But, very few *in silico* methods were developed in this regard including gene expression-based models [18], machine learning-based classifiers,[19] genetic algorithm-based models [20], and a relational database named SigCS base (http://sysbio.kribb.re.kr/sigcs) [21] which documented genes, variants, and pathways related to cerebral stroke. Unfortunately, this rich resource was discontinued as of February 2021. So, there was a huge scope for the development of a computational algorithm for the prediction of genes associated with stroke types and their sub-types. Our computational approach was aimed to recognize possible important genes and the pathways linked with the pathogenesis and prediction of stroke types along with their subtypes.

Methods

Advanced query building and searching PubMed

We fetched Human genes using genome assembly build 38 patch release 13 with known (status "Active") gene symbols through the NCBI gene database (https://www.ncbi.nlm.nih.gov/gene). PubMed advanced queries were constructed using stroke-related keywords and associations were calculated using Normalized pointwise mutual information (nPMI) between each gene symbol and queries.[22] To reduce the false hits, only titles and abstracts were searched from the articles published till 31st August 2020.[23, 24] A list of sample queries used for searching has been provided in **Table-1** with the number of hits observed.

Model development

Document frequency (DF) related to a query is defined as the number of hits fetched by searching the database. DF can be easily normalized by the total no. of entries in the database. Similarly, pointwise mutual information (PMI) is another important metric often employed to find the association between two random variables (RV). A normalized form of PMI (nPMI) was derived by Bouma et al.[22] In several published works, nPMI were employed to estimate the association between entities like genomic repeats, stress, virulence, computational tools, drug-discovery related keywords, etc.[25, 26] Following a similar approach of association mining, individual and co-occurrences of each gene symbol (RV1) and stroke-related keywords (RV2) in PubMed titles and abstracts was calculated using normalized DF (nDF) which was further utilized to compute nPMI. This nPMI value represents the strength of association between the gene (genotype) and stroke (phenotype).

Performance evaluation

Performance of the model was assessed using receiver operating characteristic (ROC) and precision-recall (PR) curve analysis on a cumulative dataset of human housekeeping (negative) [27] and already published stroke-related genes (positive). A positive gene set was constructed by compiling gene lists provided in the stroke-related research articles published during the last decade. [7, 18, 28] Different DF and nPMI value pairs were used to find the best model for stroke-associated gene identification.

Pathway analysis

To achieve insight into the biological roles and pathological mechanisms of stroke and its etiologies, we examined the biological ways that significantly overlapped with the curated stroke and etiology gene sets. For this, we calculated common genes related to stroke etiology gene sets and the genes wiki-pathways [29] and executed statistical testing to measure the significance of the overlaps. To achieve insight into the biological functions and pathological mechanisms of stroke and its etiologies, we recognized the biological pathways that significantly enriched with the curated stroke and etiology gene sets. Using cluster Profiler R package,[30] enrichment analysis was performed on three major pathway databases namely KEGG (https://www.genome.jp/kegg/) (release 96.0),[31] WikiPathways (https://www.wikipathways.org/) (release September 2020)[29] and Reactome (https://reactome.org/) (version 75)[32] which are curated, comprehensive and rich data sources on human metabolic pathways. To initiate the analysis, probable stoke associated (PSA) gene symbols were first converted to ENTREZ ids and then pathway enrichment was done with a *p-value* cutoff of 0.05 and was adjusted by the Bonferroni method.[33]

Results

Genes associated with stroke and its types

PubMed advanced searched using stroke-related keywords as mentioned in the **Table-1** and associations were calculated using nPMI between each gene symbol and queries. To reduce the false hits, only titles and abstracts were searched from the articles published till 31st August 2020.

A total of 2,785 (9.4%) genes were found to be linked to stroke risk. Based on stroke types, 1,287 (46.2%) and 376 (13.5%) genes were found to be associated with the risk of IS and HS respectively. Further stratification of IS based on TOAST classification, it was found that 86 (6.6%) genes were confined to Large artery atherosclerosis (LAA); 131 (10.1%) and 130 (10%) genes were related with the risk of small vessel disease (SVD) and Cardioembolism (CE) subtypes of IS. Circos diagram for the identified genes associated with stroke types and subtypes are represented in **Figure-1**.

Total 28,281 human gene symbols (**Supplementary Table-T1**) were extracted from the NCBI gene database with the status tag "Active" and used for calculation of nPMI with query no. 2 from **Table-1**. 2,785 (9.8%) symbols were found to be associated (having positive or negative nPMI values) with stroke (set A). To determine the stroke subtypes, nPMI was computed with query no. 3 and query no. 4 (**Table-1**) for these 2,785 gene symbols (**Supplementary Figure-S1**) resulting 1,294 (46.5%) and 376 (13.5%) genes in association with IS and HS respectively. The rest of the symbols were marked as "Unclassified". Further filtering using DF values (>5) and removing symbols (#11) like CAT, IMPACT, SET, etc. which are common English words, were listed 441 PSA genes along with their types (**Figure-1**). Gene symbols that were found to be in association with both types (HS and IS) of strokes were tagged as “Both Types” (**Figure-1**). Mining manually curated TRRUST database v2 (transcriptional regulatory relationships unraveled by sentence-based text-mining) (https://www.grnpedia.org/trrust/)[34] catalogued all the transcription factors and their target genes along with existing interaction types (activation, repression, or unknown) (**Figure-1**). A prognostic panel of 9 genes signature consisting of CYP4A11, ALOX5P, NOTCH, NINJ2, FGB, MTHFR, PDE4D, HDAC9, and ZHFX3 can be treated as a diagnostic marker to predict individuals who are at the risk of developing stroke with their subtypes

Genes associated with stroke sub-types

Queries no. 5-9 from **Table-1** have been used in nPMI model for stratification of genes associated with IS subtypes as per TOAST classification. 131 (10.1%) genes were confined to Small Vessel Disease (SVD) followed by 130 (10%) Cardioembolism (CE), 86 (6.6%) Large artery atherosclerosis (LAA), 30 (2.3%) Undetermined etiology (UDE) and 7 (0.5%) Other determined etiology (ODE) (**Supplementary Figure-S2**). While classifying HS sub-types using queries no. 10-11, 292 (77.8%), and 132 (35.2%) were predicted as Intracerebral hemorrhage (ICH) and Subarachnoid hemorrhage (SAH) respectively (**Supplementary Figure-S3**). A subset of PSA (#140) genes associated with different stroke subtypes was represented in **Figure-2**. Complete lists along with DF and nPMI values can be found in **Supplementary Table T2-T10**.

Evaluation of nPMI model

For performance evaluation of the developed nPMI model, a list of already published 7,431 genes consisting of 2,168 (29%) stroke-related (positive set) and 5,263 (71%) human housekeeping genes (negative set) were constructed. An intersection of 1,144 genes was found between set A and published gene sets (positive: 581, negative: 563) which was used for evaluation. Multiple models were built using different DF (ranging from 1 to 5) and nPMI (ranging from 0.05 to 0.5) cut-offs and the best model having Accuracy: 0.64, Sensitivity: 0.63, Specificity: 0.65, and Precision: 0.66 was reported at DF cut-off 2 and nPMI cut-off 0.1. Performance evaluation by receiver operating characteristic (ROC) and Precision-Recall (PR) curves analysis resulted Area under curve (AUC) values of 0.64 and 0.63 respectively (**Figure-3**).

Stroke can influence many pathways

To reduce false-positive hits, pathway enrichment analysis was done using 190 genes (**Supplementary Table-T11**), a subset of PSA gene symbols that were manually curated and already known to be associated with stroke. Analysis with Reactome, WikiPathways, and KEGG resulting 53, 32, and 35 unique pathways enriched with aforementioned stroke-associated genes (**Supplementary Table-T12, Table-T13, and Table-T14**). However manual curation of these lists of pathways showed promising results with WikiPathways and Reactome which were presented in **Figure-4** and **Supplementary Figure-S4**. Findings for the genes associated with biological Network pathways including Kegg and Reactome are reported in **Figure-4**.

Discussion

Emerging evidences from published meta-analysis and GWAS studies suggests that several genetic variants (MTHFR, MMP9, PDE4D, CYP4A11, ALOX5P, NOTCH, NINJ2, FGB, eNOS, PITX2, ZFHX3, HDAC9, ABO, etc) have been identified, even though the extent of the effect of each variant is regarded as inter-varying within different populations including Asian, Caucasian, African.[1, 1, 7, 14, 35–39, 39–44] The findings, however, are often unclear and hard to interpret. Genetic association studies in diverse stroke populations negate matters of the restricted patient population by the a priori choice of a functionally relevant gene and its relation with a specific phenotype. Improving patient outcomes in stroke requires a rapid and accurate prediction of stroke and its subtypes. The genetic signature could help it to distinguish or calculate the incidence of hemorrhagic and ischemic stroke along with its subtypes and may also help in the prognosis of further risk of stroke recurrence.

Our findings are confined specifically to predict the associated genes for the predisposition of ischemic or hemorrhagic stroke. A prognostic panel of 9 genes signature consisting of CYP4A11, ALOX5P, NOTCH, NINJ2, FGB, MTHFR, PDE4D, HDAC9, and ZHFX3 can be designed and treated as a diagnostic marker to predict individuals who are at the risk of developing stroke with their subtypes. Developing this genetic markers panel seems to offer hope of significantly better sensitivity and specificity may provide a quick and reliable assessment with revolutionize stroke management. It will also reduce cost and timing for preventing the stroke incidence in susceptible individuals having a chance of developing stroke with LVD, SVD and CE subtypes and HS. These genetic markers could also have the potential to enter into a routine clinical use despite their obvious promise by only single validation of our findings.

Limitations

The current study has certain limitations in spite of these interesting results. Meanwhile the datasets fused were from the published studies in patients with stroke, misdiagnosis or misclassification of stroke subtypes could have potentially influenced our findings. Also, we could not access the original SNP genotype data, we had to use summary data from published stroke GWAS and candidate gene studies, which prohibited us to address the common genetics of complex traits and could have affected our results. Moreover, as the various testing corrections we used in our statistical analyses may be inadequate to clarify all biases, permutation testing should be used to adjust the results at the single SNP level. Furthermore, we needed transcriptomic and epigenetic data, which may contribute to the identification of additional potential causal mechanisms and links. False positives are highly expected since the method was predicting the genes using PubMed title and abstracts. However, false positives in our findings were rectified using manual curation.

Conclusion

A novel text data-driven method was developed to identify genes associated with stroke types and their subtypes. Our findings might offer certain directive implications for further exploring the diagnostic and prognostic genetic markers to empower the molecular targeting treatment for stroke prevention.

References

1. Boehme AK, Esenwa C, Elkind MSV (2017) Stroke Risk Factors, Genetics, and Prevention. Circ Res 120:472–495. https://doi.org/10.1161/CIRCRESAHA.116.308398

2. Chauhan G, Debette S (2016) Genetic Risk Factors for Ischemic and Hemorrhagic Stroke. Curr Cardiol Rep 18:. https://doi.org/10.1007/s11886-016-0804-z

3. Feigin VL, Forouzanfar MH, Krishnamurthi R, et al (2014) Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. Lancet 383:245–254

4. Feigin VL (2005) Stroke epidemiology in the developing world. Lancet 365:2160–2161. https://doi.org/10.1016/S0140-6736(05)66755-4

5. Adams HP, Bendixen BH, Kappelle LJ, et al (1993) Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 24:35–41. https://doi.org/10.1161/01.str.24.1.35

6. O’Donnell MJ, Xavier D, Liu L, et al (2010) Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet 376:112–123. https://doi.org/10.1016/S0140-6736(10)60834-3

7. Malik R, Chauhan G, Traylor M, et al (2018) Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 50:524–537. https://doi.org/10.1038/s41588-018-0058-3

8. Wassertheil-Smoller S, Qi Q, Dave T, et al (2018) Polygenic Risk for Depression Increases Risk of Ischemic Stroke: from the Stroke Genetics Network (SiGN) Study. Stroke 49:543–548. https://doi.org/10.1161/STROKEAHA.117.018857

9. Tonk M, Haan J (2007) A review of genetic causes of ischemic and hemorrhagic stroke. J Neurol Sci 257:273–279. https://doi.org/10.1016/j.jns.2007.01.037

10. Dichgans M (2007) Genetics of ischaemic stroke. The Lancet Neurology 6:149–161. https://doi.org/10.1016/S1474-4422(07)70028-5

11. Bersano A, Ballabio E, Bresolin N, Candelise L (2008) Genetic polymorphisms for the study of multifactorial stroke. Hum Mutat 29:776–795. https://doi.org/10.1002/humu.20666

12. Jood Katarina, Ladenvall Per, Tjärnlund-Wolf Anna, et al (2005) Fibrinolytic Gene Polymorphism and Ischemic Stroke. Stroke 36:2077–2081. https://doi.org/10.1161/01.STR.0000183617.54752.69

13. Bevan Steve, Traylor Matthew, Adib-Samii Poneh, et al (2012) Genetic Heritability of Ischemic Stroke and the Contribution of Previously Reported Candidate Gene and Genomewide Associations. Stroke 43:3161–3167. https://doi.org/10.1161/STROKEAHA.112.665760

14. Lindgren A (2014) Stroke Genetics: A Review and Update. J Stroke 16:114–123. https://doi.org/10.5853/jos.2014.16.3.114

15. Fornage M (2009) Genetics of stroke. Curr Atheroscler Rep 11:167–174. https://doi.org/10.1007/s11883-009-0027-5

16. Matarin M, Singleton A, Hardy J, Meschia J (2010) The genetics of ischaemic stroke. J Intern Med 267:139–155. https://doi.org/10.1111/j.1365-2796.2009.02202.x

17. Carty CL, Keene KL, Cheng Y-C, et al (2015) Meta-analysis of genome-wide association studies identifies genetic risk factors for stroke in African-Americans. Stroke 46:2063–2068. https://doi.org/10.1161/STROKEAHA.115.009044

18. Theofilatos K, Korfiati A, Mavroudi S, et al (2019) Discovery of stroke-related blood biomarkers from gene expression network models. BMC Med Genomics 12:118. https://doi.org/10.1186/s12920-019-0566-8

19. O’Connell GC, Petrone AB, Treadway MB, et al (2016) Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. NPJ Genom Med 1:16038. https://doi.org/10.1038/npjgenmed.2016.38

20. Alawad DM, Mishra A, Hoque MT (2020) AIBH: Accurate Identification of Brain Hemorrhage Using Genetic Algorithm Based Feature Selection and Stacking. Machine Learning and Knowledge Extraction 2:56–77. https://doi.org/10.3390/make2020005

21. Park Y-K, Bang OS, Cha M-H, et al (2011) SigCS base: an integrated genetic information resource for human cerebral stroke. BMC Syst Biol 5 Suppl 2:S10. https://doi.org/10.1186/1752-0509-5-S2-S10

22. Bouma G (2009) Normalized (pointwise) mutual information in collocation extraction. In: undefined. /paper/Normalized-(pointwise)-mutual-information-in-Bouma/15218d9c029cbb903ae7c729b2c644c24994c201. Accessed 9 Mar 2021

23. Entrez Direct: E-utilities on the Unix Command Line - Entrez Programming Utilities Help - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK179288/. Accessed 9 Mar 2021

24. Azam N, Yao J (2012) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications 39:4760–4768. https://doi.org/10.1016/j.eswa.2011.09.160

25. Das G, Das S, Dutta S, Ghosh I (2018) In silico identification and characterization of stress and virulence associated repeats in Salmonella. Genomics 110:23–34. https://doi.org/10.1016/j.ygeno.2017.08.002

26. Kumar P, Das G, Ghosh I (2017) Critical assessment of contribution from indian publications: the role of in silico designing methods leading to drugs or drug-like compounds using text based mining and association. 8:133–148

27. Human housekeeping genes, revisited - PubMed. https://pubmed.ncbi.nlm.nih.gov/23810203/. Accessed 8 Mar 2021

28. Fang F, Xu Z, Suo Y, et al (2020) Gene panel for Mendelian strokes. Stroke Vasc Neurol 5:416–421. https://doi.org/10.1136/svn-2020-000352

29. WikiPathways: connecting communities - PubMed. https://pubmed.ncbi.nlm.nih.gov/33211851/. Accessed 8 Mar 2021

30. Yu G, Wang L-G, Han Y, He Q-Y (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS 16:284–287. https://doi.org/10.1089/omi.2011.0118

31. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30

32. The reactome pathway knowledgebase. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7145712/. Accessed 9 Mar 2021

33. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310:170

34. Han H, Cho J-W, Lee S, et al (2017) TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic acids research 46:. https://doi.org/10.1093/nar/gkx1013

35. Kumar A, Sharma R, Misra S, et al (2020) Relationship between methylenetetrahydrofolate reductase (MTHFR) gene (A1298C) polymorphism with the risk of stroke: A systematic review and meta-analysis. Neurol Res 42:913–922. https://doi.org/10.1080/01616412.2020.1798107

36. Alhazzani AA, Kumar A, Selim M (2018) Association between Factor V Gene Polymorphism and Risk of Ischemic Stroke: An Updated Meta-Analysis. J Stroke Cerebrovasc Dis 27:1252–1261. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.12.006

37. Kumar A, Kumar P, Prasad M, et al (2015) Association of C677T polymorphism in the methylenetetrahydrofolate reductase gene (MTHFR gene) with ischemic stroke: a meta-analysis. Neurol Res 37:568–577. https://doi.org/10.1179/1743132815Y.0000000008

38. Zhou X, Guan T, Li S, et al (2017) The association between HDAC9 gene polymorphisms and stroke risk in the Chinese population: A meta-analysis. Sci Rep 7:41538. https://doi.org/10.1038/srep41538

39. Kumar A, Misra S, Kumar P, et al (2017) Association between endothelial nitric oxide synthase gene polymorphisms and risk of ischemic stroke: A meta-analysis. Neurol India 65:22–34. https://doi.org/10.4103/0028-3886.198170

40. Misra S, Talwar P, Kumar A, et al (2018) Association between matrix metalloproteinase family gene polymorphisms and risk of ischemic stroke: A systematic review and meta-analysis of 29 studies. Gene 672:180–194. https://doi.org/10.1016/j.gene.2018.06.027

41. Kumar P, Misra S, Kumar Yadav A, et al (2016) Relationship between Interleukin-6 (-174G/C and -572C/G) Promoter Gene Polymorphisms and Risk of Intracerebral Hemorrhage: A Meta-Analysis. Pulse (Basel) 4:61–68. https://doi.org/10.1159/000447677

42. Keat Wei L, Griffiths LR, Irene L, Kooi CW (2019) Association of NOTCH3 Gene Polymorphisms with Ischemic Stroke and its Subtypes: A Meta-Analysis. Medicina (Kaunas) 55:. https://doi.org/10.3390/medicina55070351

43. Zhang XF, Luo TY (2015) Association between the FGB gene polymorphism and ischemic stroke: a meta-analysis. Genet Mol Res 14:1741–1747. https://doi.org/10.4238/2015.March.6.21

44. Chauhan G, Arnold CR, Chu AY, et al (2016) Identification of additional risk loci for stroke and small vessel disease: a meta-analysis of genome-wide association studies. The Lancet Neurology 15:695–707. https://doi.org/10.1016/S1474-4422(16)00102-2

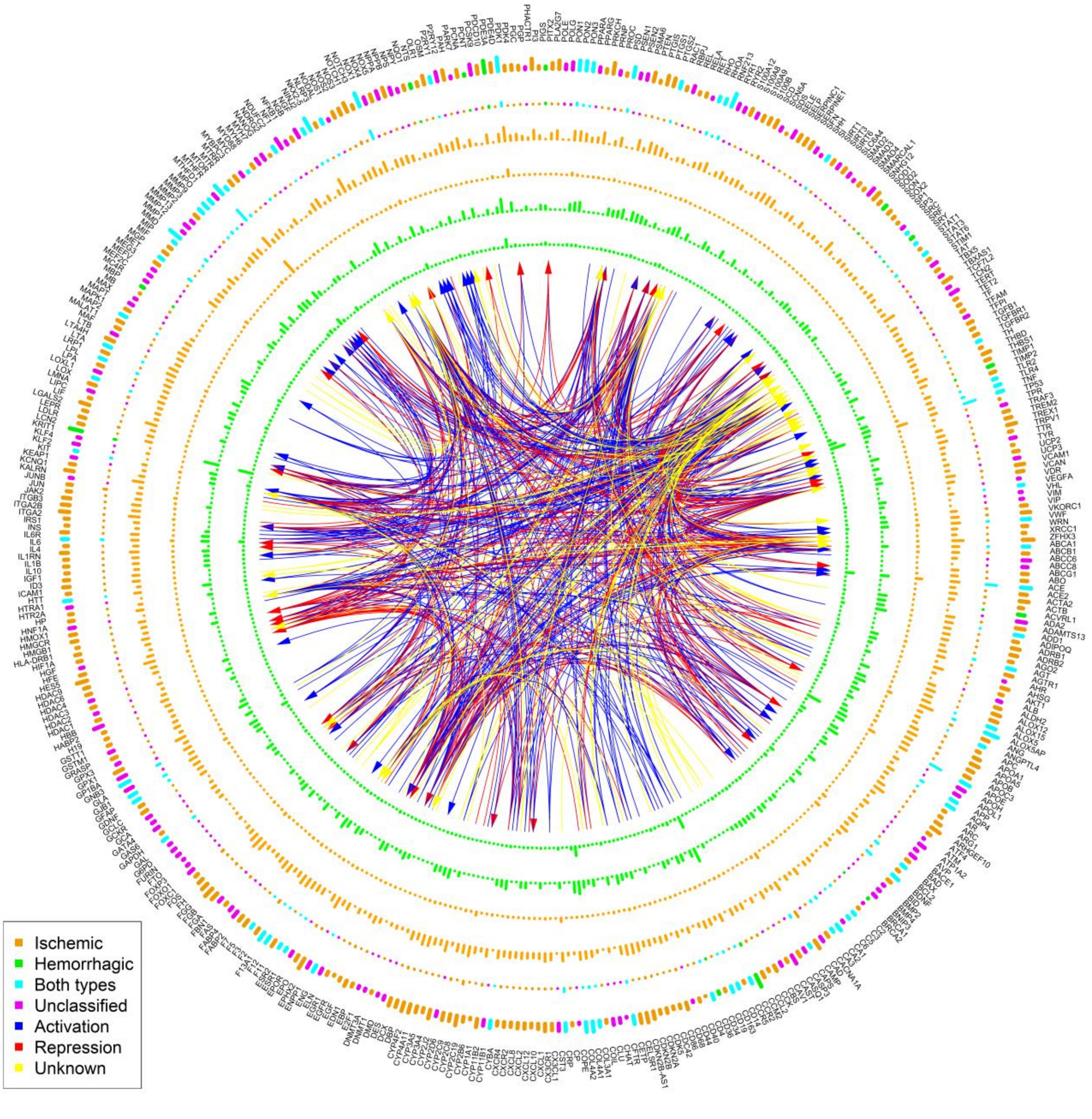

Figure 1: Prediction of the genes associated with stroke and its sub-types using normalized pointwise mutual information (nPMI) and document frequency (nDF) calculated from PubMed database. The outermost track of the circus plot represents the selected human gene symbols (#440) based on the nDF cut-off. The next two inner tracks with different colours show the nPMI and nDF values respectively; the colour codes are as follows: symbols associated with ischemic stroke: orange, hemorrhagic stroke: green, both stroke types: cyan and unclassified: magenta.

nPMI and nDF values of the genes related with ischemic stroke have been presented in the next two orange inner tracks. Similarly, the two innermost green tracks displays nPMI and nDF values of the genes related with hemorrhagic stroke. The height of the bars indicates the nPMI and nDF values. The transcriptional regulatory links have been created between the transcription factors and pointing towards their target genes in different colours with codes: blue: activation, red: repression, yellow: unknown based on the manually curated database TRRUST (v2).

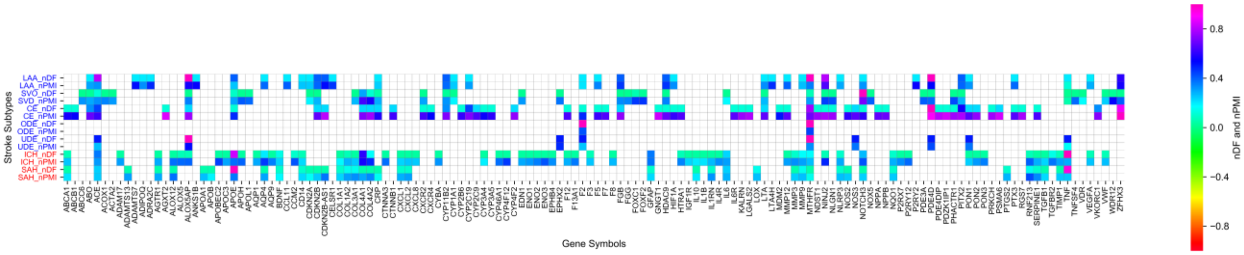

Figure 2: Prediction of genes associated with stroke types and subtypes. X-axis represents human gene symbols and Y-axis represents different stroke types and subtypes. Sub-types of Ischemic stroke (IS) are in blue and Hemorrhagic stroke (HS) are in red. Values of normalized Document frequency (nDF) (ranging 0 to 1) and normalized pointwise mutual information (nPMI) (ranging -1 to 1) are shown using color bar. Abbreviation of different stroke subtypes are as follows: SVD-Small Vessel Disease, CE-Cardioembolism, LAA-Large artery atherosclerosis, UDE-Undetermined etiology, ODE-Other determined etiology, ICH-Intracerebral hemorrhage SAH-Subarchanoid hemorrhage.

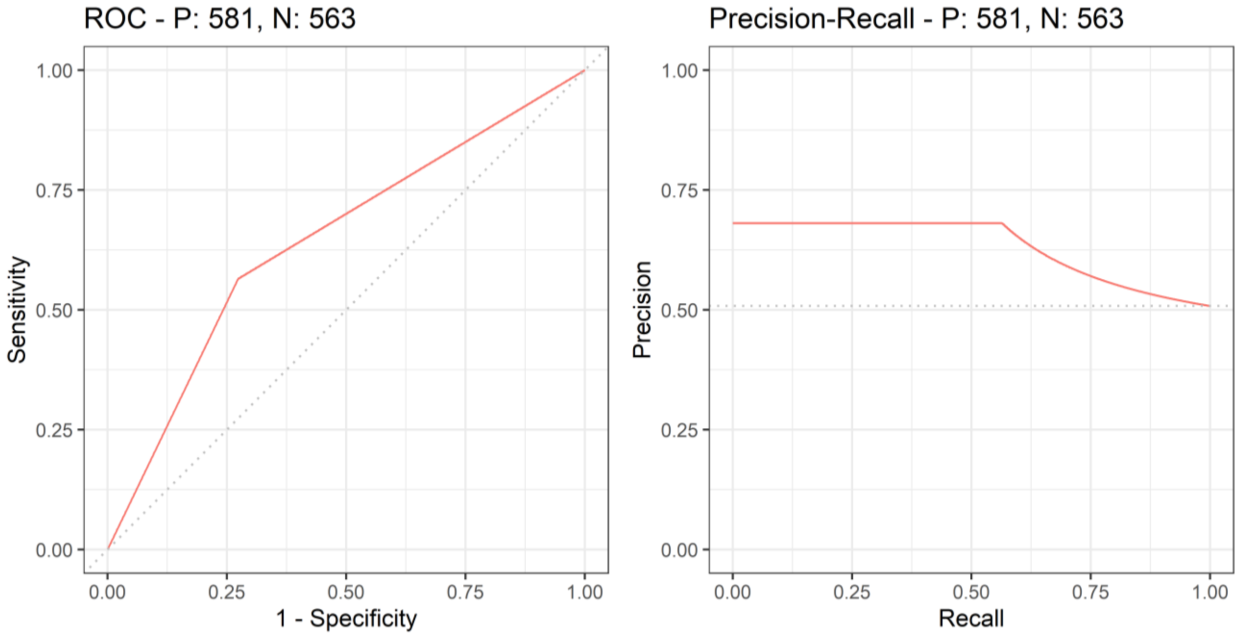

Figure 3: Receiver operating characteristic (ROC) and Precision-Recall (PR) curve analysis for the evaluation of developed normalized pointwise mutual information (nPMI) based model for classification of stroke related genes. Using different normalized document frequency (nDF) and nPMI cut-offs ROC and PR curves have been plotted. Area under curve (AUC) values for ROC and PR curves are 0.64 and 0.63 respectively.

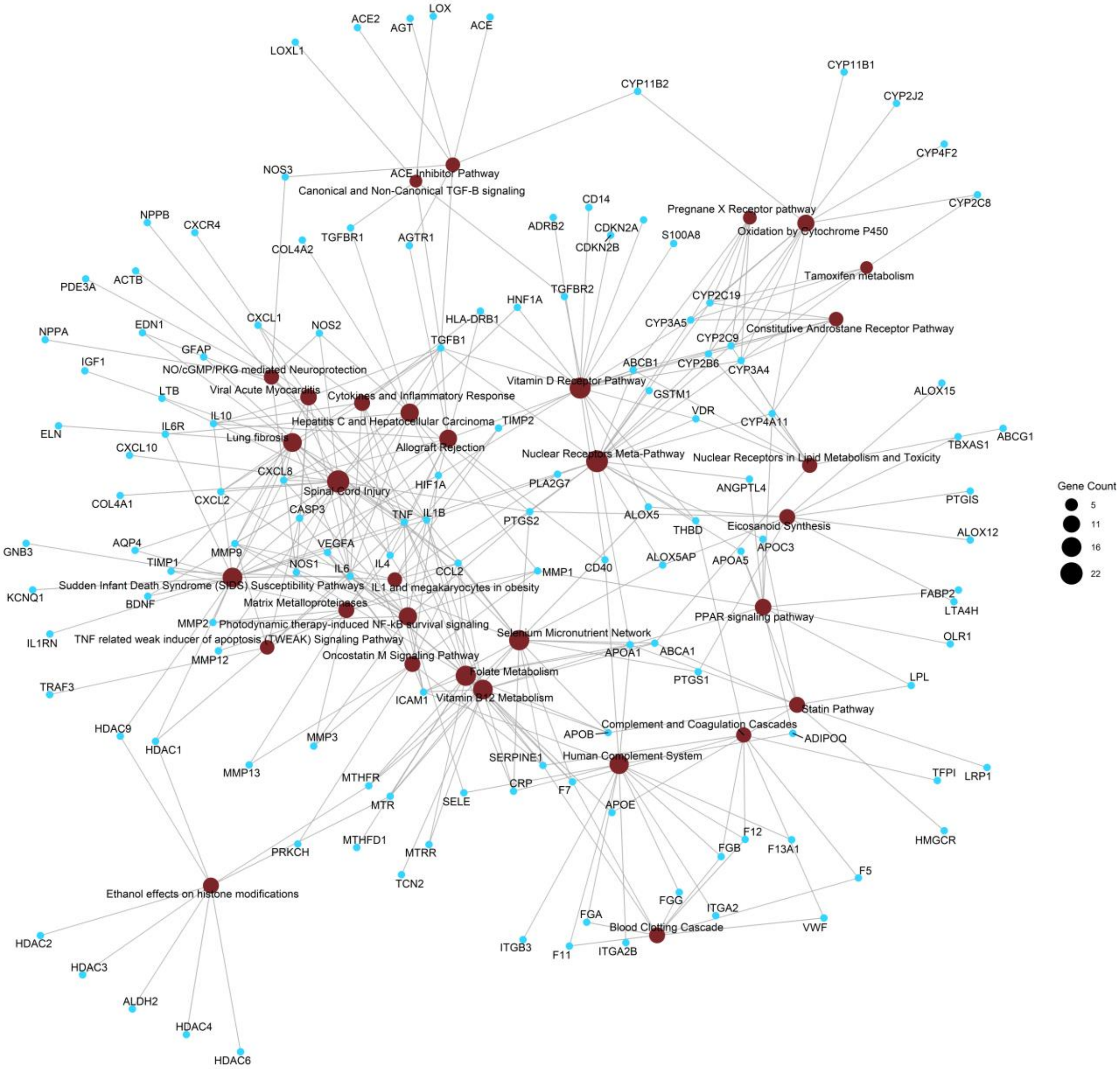

Figure 4: Enrichment analysis for finding the important pathways associated with the predicted stroke genes. Analysis has been performed using WikiPathways database. Brown and cyan nodes are representing pathways and genes respectively. No. of genes associated with a pathway is represented by size of the brown node. Connecting links are in grey color.

Table 1: PubMed sample queries used in the present study and respective no. of hits obtained till August 2020.

No.	Topic	PubMed Advanced queries	Total no. of records in PubMed
1	All abstracts	all[sb] AND hasabstract[text]	20816630
2	Stroke related abstracts	(stroke[TIAB] OR Cerebrovascular[TIAB]) AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	9626
3	Hemorrhagic Stroke related abstracts	("Intracerebral hemorrhage"[TIAB] OR "Hemorrhagic Stroke"[TIAB] OR "Subarchanoid hemorrhage"[TIAB]) AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	585
4	Ischemic Stroke related abstracts	("Ischemic Stroke"[TIAB] OR (TOAST[TIAB] AND (Classification[TIAB] OR Subtypes[TIAB]))) AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	2738
5	Large artery atherosclerosis	"Large artery atherosclerosis"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	103
6	Small vessel disease	"Small vessel disease"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	288
7	Cardioembolic disease	"Cardioembolic disease"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	117
8	Other determined etiology	"Other determined etiology"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	3
9	Undetermined etiology	"Undetermined etiology"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	45
10	Intracerebral hemorrhage	"Intracerebral hemorrhage"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	424
11	Subarchanoid hemorrhage	"Subarchanoid hemorrhage"[TIAB] AND (gene[TIAB] OR genes[TIAB]) AND hasabstract[text]	434