Association of SUMOlation Pathway Genes With Stroke in a Genome-wide Association Study in India

Amit Kumar[1*], Ph.D, Ganesh Chauhan[2*], Ph.D, Shriram Sharma[3],DM, Surekha Dabla[4], DM, P.N. Sylaja[5], DM, Neera Chaudhry[6], DM, Salil Gupta[7], DM, Chandra Sekhar Agrawal[8], DM, Kuljeet Singh Anand[9], DM, Achal Srivastava[1], DM, Deepti Vibha[1], DM, Ram Sagar[1],Msc, Ritesh Raj, MSc, Ph.D,Ankita Maheswari[1], MSc, Subbiah Vivekanandhan[1], Ph.D, Bhavna Kaul[6], DM, Samudrala Raghavan,[6] DM, Sankar Prasad Gorthi[7], DM, Dheeraj Mohania[8], Ph.D, Samander Kaushik[10], Ph.D, Rohtas Kanwar Yadav[11], MD, Anjali Hazarika[12], MBBS, Pankaj Sharma[13], MD, Ph.D, Kameshwar Prasad[1], DM**

1 Department of Neurology, All India Institute of Medical Sciences, New Delhi, India

[2] Centre for Brain Research, Indian Institute of Science, Bangalore, India

[3] Department of Neurology, North Eastern Indira Gandhi Regional Institute of Health and Medical Sciences, Shillong, Meghalaya, India

4 Department of Neurology, Pandit Bhagwat Dayal Sharma Post Graduate Institute of Medical Sciences, Rohtak, Haryana, India

5 Department of Neurology, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Kerala India

6 Department of Neurology, Vardhman Mahavir Medical College & Safdarjung Hospital, New Delhi, India

7 Department of Neurology, Army Research and Referral Hospital, New Delhi, India

[8] Department of Neurology, Sir Ganga Ram Hospital, New Delhi, India

[9]Department of Neurology, Ram Manohar Lohia Hospital, New Delhi, India

[10]Centre for Biotechnology, Pandit Bhagwat Dayal Sharma Post Graduate Institute of Medical Sciences, Rohtak, Haryana, India

[11]Department of Radiology, Pandit Bhagwat Dayal Sharma Post Graduate Institute of Medical Sciences, Rohtak, Haryana, India

[12] Department of Blood Bank, Cardio Neuro-Centre, All India Institute of Medical Sciences, New Delhi, India

[13]Institute of Cardiovascular Research, Royal Holloway, University of London

Corresponding Author

Kameshwar Prasad**

Professor, Department of Neurology

All India Institute of Medical Sciences, Ansari Nagar

New Delhi-110029, India

Ph: +91-11-26594013 (office), 9868350484 (mobile)

Email: drkameshwarprasad@gmail.com

Subtitle: Indian Stroke GWAS

Word counts for Abstract: 250 [max 250]

Word counts for Introduction: 246 [max 250]

Word counts in main text: 3,412 [max 4,500]

Character count in title: 127 [max 140 characters]

Number of References: 26 [max 50]

Supplementary data: Tables -17, figures-10 and text-1

Number of tables and figures: 3 tables and 3 figures [max 7 items in total]

Statistical Analyses: GWAS statistical data analyses were conducted by Dr. Amit Kumar, Scientist-D, Department of Neurology, All India Institute of Medical Sciences, New Delhi and Dr. Ganesh Chauhan, Assistant Professor, Centre for Brain Research, Indian Institute of Science, Bangalore, India. Epidemiological data analysis was completed by Dr. Amit Kumar, Scientist-D, Department of Neurology, AIIMS, New Delhi

Search terms: India, South Asians, stroke, genetics, GWAS

Disclosures: All authors declare no disclosures

Funding : Department of Biotechnology, Government of India

Abstract

Objective: To undertake a genome wide association study (GWAS) to identify genetic variants for stroke in South Indians.

Methods: In a hospital-based case-control study, eight teaching hospitals in India recruited 4,088 subjects including 1,609 stroke cases. Imputed genetic variants were tested for association with stroke subtypes using both single variant and gene-based tests. Association with vascular risk factors was performed using logistic regression. Various databases were searched for replication, functional annotation, and association with related traits for novel loci identified. Status of candidate genes previously reported in Indian population was checked.

Results: We report association of vascular risk factors with stroke, similar to previous reports, and show modifiable risk factors like hypertension, smoking, and alcohol consumption having the highest effect. Single marker-based association revealed two loci for cardioembolic stroke (1p21 and 16q24), two for small vessel disease stroke (3p26 and 16p13), and four for hemorrhagic stroke (3q24, 5q33, 6q13 and 19q13) at $P<5\times10^{-8}$. The index SNP of 1p21 is an eQTL ($P_{lowest}=1.74\times10^{-58}$)for *RWDD3* involved in SUMOlation and is associated with platelet distribution width (1.15×10^{-9}) and 18-carbon fatty acid metabolism ($P=7.36\times10^{-12}$). In gene-based analysis we identified three genes (*SLC17A2, FAM73A & OR52L1*) at $P<2.7\times10^{-6}$. 11 of 32 candidate gene loci studied in Indians replicated ($P<0.05$) and 21 of 32 loci identified through previous GWAS replicated based on directionality of effect.

Conclusions: This first GWAS of stroke in Indians identified novel loci and replicated previously known loci. For the first time genetic variants in the SUMOlation pathway which has been implicated in brain ischemia were identified.

Introduction

The burden of stroke as the second leading cause of death and the leading cause of long-term disability is being felt all over the world, especially in parts of South Asia including India, home to 20% of the worlds' stroke population.[1] The etiology of stroke is multifactorial and includes both genetic and environmental factors.[2] Understanding the genetic risk factors for stroke could give rise to new biological pathways important for new treatment and preventive options. To date GWAS is the most effective way for gene discovery when it comes to complex disorders like stroke.[3] A large number of GWAS and their meta-analysis has been conducted for stroke and revealed ~40 loci.[3] However, with the exception of a few studies, the majority of the subjects in stroke GWAS are of European origin. It has been suggested that such scenarios could lead to health disparities later as many of the findings based on Europeans are not transferable to populations of other ethnicities.[4] GWAS of stroke from India are completely lacking.[5] The Indian population of 1.35 billion people with a rich diversity of ethnic groups and ancestral populations with large founder effects and distinct genetic architecture in terms of allele frequency and linkage disequilibrium blocks offer tremendous opportunities for genetic studies.[6–9]

In this first large-scale GWAS of stroke from India, we also study association with vascular risk factors and check the status of previous stroke GWAS loci and candidate genes studied in Indians.

Materials & Method

Subject Recruitment

This study involved a multi-centric hospital-based case-control design. Eight teaching hospitals from different parts of India recruited phenotypically well-characterized stroke cases and ethnically matched controls (**figure e-1 and appendix e-1**). Stroke was diagnosed using guidelines set by the World Health Organization, by trained Neurologists, and was primarily of vascular origin. Eligibility criteria included age 18 to 85 year, both sexes with a history of stroke ever in life, Indian ancestry, and absences of any major neurological disorders (Epilepsy, Parkinson disorder, Alzheimer Disease, Multiple Sclerosis, Brain Tumor, etc.). In total 4088 subjects were recruited which included 1609 stroke cases and 2479 controls. Stroke free status was assessed by a well-validated questionnaire.[6] Ethnically matched subjects diagnosed as stroke free using this questionnaire and no history of any serious neurological disorder were recruited as controls. Each centre recruiting the stroke cases recruited its respective control subjects. Detailed demographic characteristics of the case and control subjects are presented in **table 1 and table e-1**. The study protocol was approved by the ethics committees of the respective participating institutions and signed informed consent was obtained from the subjects before enrolling them into the study.

Power of the study

We estimated the power of the study at GWAS significance levels of $P=5\times10^{-8}$ and assuming a population risk of 10% to detect an Odds Ratio(OR) in the range of 1.1 to 2.0 and allele

frequency ranges of 0.01 to 0.50 using Quanto version 1.2.4 (https://preventivemedicine.usc.edu/download-quanto/) (**figure e-2**).

Genome wide genotyping

Genome wide genotyping was performed on an Illumina platform using the genome screen array (GSA) version 2.0 (with additional multi-disease content, GSA-MD), which comprise of ~ 750,000 markers. Intensity data files (IDAT) were imported into Illumina's GenomeStudio software (https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html) for Intensity normalization, scaling, offset estimation followed by SNP clustering, genotype assignment and calling. Initial quality control was performed based on Illumina designed QC probes (sample independent and sample dependent) to assess experiment and sample quality. For association analysis cluster files generated from GenomeStudio were used to convert IDATfiles to variant call format (VCF) files using Illumina's GTCtoVCF tool (https://github.com/Illumina/GTCtoVCF). Individual VCF files were merged using bcftools (http://samtools.github.io/bcftools/bcftools.html) to create a single project level VCFfile. Emphasis was on getting a VCFfile rather than the usual PLINK format (https://www.cog-genomics.org/plink2) .ped and .map file, which is the default option in GenomeStudio, as allele coding in VCFfiles are as per the reference genome on the positive strand and allele coding are not omitted for variants were the alternate allele was not observed. With the availability of multiple tools designed to interrogate large VCFfiles being generated through large sequencing projects in a very short duration of time, VCFformat (https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format) has become a popular choice.

Genotype and sample quality checks

All marker and sample quality checks were carried out mainly using bcftools and plink2 unless mentioned otherwise. We implemented a stringent marker quality check and included only variants with genotype call rate >95%, deviation from Hardy Weinberg equilibrium (HWE) P-value $>1\times10^{-5}$), minor allele frequency (MAF) >0.01, biallelic markers, autosomal and X-chromosome markers. Both single nucleotide polymorphisms and short insertion and deletion were included for association analysis. After this first round of marker quality check, we were left with 399,541 markers. The number of marker failing quality checks in the various categories is mentioned in Supplementary Table 2, major loss of markers was due to the presence of non-polymorphic and very low-frequency markers (34.71%) in this Indian population due to the nature of the chip design to capture variants which could be specific to populations across the world (**table e-2**).

During sample quality check we excluded samples with call rate <95%, heterozygosity beyond ±3SD, genotype, and pedigree gender mismatch. Markers were pruned using the "--indep-pairwise 50 5 0.2" function of PLINK to retain only markers with r2<0.2 in a window size of 50 markers and slide of 5 markers per window. Pruned markers were used to ascertain cryptic relatedness using the "--genome" function in PLINK and among the pairs of duplicated samples "PI_HAT" values > 0.6, only one of the samples with the highest call rate was retained. These samples with LD pruned list of markers were used to perform principal components analysis (PCA) using FastPCA function implemented in EIGENSOFT version 7.2.1 (https://www.hsph.harvard.edu/alkes-price/software/) and the samples were projected upon the samples of the 1000 Genomes project using the first two PCA and PCA outliers were excluded. The sample quality check was blinded to the disease status of the subject. The number of

samples excluded at the various stages is listed in **table e-3**. Post sample quality checks, the marker quality check was repeated (round 2) due to a change in the number of subjects. Statistics like call rate, MAF, and HWE P-value were recalculated, which led to exclusion of 4,286 markers, leaving us in total 395,255 good quality genotyped markers (**table e-2**).

Imputation

Imputation was carried out using the TOPMed reference panel (Version R2 on GRC38). Whole-genome sequencing data of 97,256 individuals was available in this version of the reference panel. Genotypes were on the positive strand and allele checks were made along with other QC as suggested (https://topmedimpute.readthedocs.io/en/latest/) prior to submission of genotypes for imputation. Phasing and imputation were carried out on the TOPMed imputation server (https://imputation.biodatacatalyst.nhlbi.nih.gov/index.html#!). Analysis of the imputed genotypes was restricted to variants with MAF>0.01 and imputation quality >0.7.

Association analysis and Functional annotation

Association analyses with stroke and its subtypes were performed using the fast GWA function in GCTA tool that employs the generalized mixed model. The use of generalized mixed models has been shown to be useful to account for relatedness and residual population stratifications using PCs. We used the first 10 PCs as covariates to remove any residual population stratification. Age and gender were also used as covariates when performing association analysis. We calculated effective allele count as 2×MAF-cases×N-cases×imputation-quality and reported association analysis of only variants with effective allele count >10. Such an effective allele count restricts association analysis to only situations where at least 10 counts of good imputation quality minor allele are observed in cases, thereby ruling out false positives due to rare variants. We performed gene-based tests using the MAGMA tool as implemented in FUMA

(https://fuma.ctglab.nl/). For biological interpretation and better understating, the annotation was performed using Annovar, OMIM, KEGG, and GO database. We looked up previously published GWAS data of stroke and related phenotypes to check for replications and better biological insights using the following resources:

http://cerebrovascularportal.org/home/portalHome

http://www.type2diabetesgenetics.org/home/portalHome

http://geneatlas.roslin.ed.ac.uk/phewas/

The GTEx (https://gtexportal.org/home/) and Phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/) resources were accessed to identify eQTLs for the top associated SNPs.

Replication of previously reported GWAS loci of stroke and subtypes

We extracted the summary statistics of all previously reported loci based on GWAS studies of stroke (**appendix e-1**). This included GWAS studies on European and non-European subjects and their meta-analysis. The first replication was tested just based on the directionality of effect as the current study is not powered enough to replicate associations performed in large studies like those of MEGASTROKE[10] (>500,000 subjects). Later we also tested association based on both directionality of effect and $P<0.05$. Finally, we also report if any loci stood the multiple testing threshold after correcting for the number of independent loci being queried.

Candidate gene studies on Indian population

Details of the search strategy to identify candidate gene studies that investigated the association with stroke and its subtypes in the Indian population are presented in **appendix e-1**. Briefly, we

searched PubMed for studies investigating the association of genetic variants with stroke in the Indian population and extracted the association statistics from the publications. Later we extracted results of the same genetic variants in the current study and checked if they replicated at $P<0.05$ and if the direction of effect was the same. Multiple testing corrections were implemented after correcting for the number of independent loci.

Results:

Association with vascular risk factors

A total of 1255 cases and 2154 controls met the quality control criteria for the GWAS study and were considered for the association study. The study flow diagram adopted for this study is presented in **figure 1**. The stroke cases in this study were on average 6-10 years older than the controls and the majority were men (71% to 74%) but there was no significant gender difference between cases and controls (**table 1**). Hypertension status and measures of blood pressure were strongly associated with stroke, with the strongest association being for hemorrhagic stroke (SBP_{mean}=162.00; DBP_{mean}=93.50 and Hypertension status=59.2%). Tobacco smoking and alcohol consumption were also strongly associated with stroke. Among the risk factors, dyslipidemia and fasting glucose levels had mild to moderate effect except for large artery stroke which showed a strong association with dyslipidemia ($P=7\times10^{-5}$). Atrial Fibrillation was mainly associated with cardioembolic stroke ($P=3.3\times10^{-32}$). Disease comorbidities like diabetes (2.3×10^{-27}) and myocardial infarction ($p=0.003$) were also more prevalent in stroke cases than control subjects.

Association with genetic risk factors

Using 395,255 good quality genotyped markers we were able to impute 782,7597 variations using the TOPMed reference panel with MAF>0.01 and imputation quality>0.7. After excluding variants with effective allele count <10 we observed that eight independent loci reached genome-wide significance threshold ($P<5\times10^{-8}$) in three stroke subtypes (**figure 2 and figure e-3**). There was no major inflation observed in the test statistics as detected on the quantile-quantile plot and the median of chi-square value (lambda value) being close to 1 (**figure e-4**). Two loci at 1p21 (OR=1.265; $P=1.09\times10^{-9}$) and 16q24 ((OR=1.147; P-value=4.21×10^{-8}) were associated with cardioembolic stroke, (**figure 2 and table 2**). Two loci were identified for small vessel disease stroke at 3p26 (OR=1.226; $P=2.91\times10^{-8}$) and 16p13 (OR=1.053; $P=1.76\times10^{-8}$) (**figure 2 and table 2**). Rest of the four newly identified loci were for hemorrhagic stroke at 3q24 (OR=0.853; $P=8.95\times10^{-10}$), 5q33 (OR=1.192; $P=1.31\times10^{-8}$), 6q13 (OR=1.193; $P=3.16\times10^{-8}$) and 19q13 (OR=1.209; $P=1.14\times10^{-11}$) (**figure 2 and table 2**). The nearest gene(s) for these loci as annotated using ANNOVAR is presented in **table 2** along with the candidate gene in the region likely affecting the risk of stroke based on eQTL analysis and literature review. List of variants which were significant at $P<1\times10^{-5}$ in various stroke subtypes considered in the study are presented in **table e-4, table e-5, table e-6, table e-7, table e-8, table e-9**. In the gene-based analysis, we identified one gene for large vessel disease (SLC17A2, P-value=1.26×10-6) and two for small vessel disease (FAM73A, P-value=2.60×10-6, and OR52L1, P-value=1.38×10-6) that were significant after correcting for multiple testing P-value<2.701×10-6 (**figure e-5 and table e-10**). Apart from rs9924207(16p13.3) associated with small vessel disease stroke the rest of the seven variants are in the low-frequency range (**table 2**). The regional plots of the top 8 loci and the

variant showing the most significant association with all stroke are presented in **figure e-6, figure e-7, and figure e-8 and figure e-9**. Data for four out of eight of the top loci were available in MEGASTROKE and we observed that three of them (1p21.3, 16p13.3, 19q13.42) were associated with one of the stroke subtypes at P<0.05 (**table e-11**). The loci 1p21.3 and 16q24.2 also show a nominal association with stroke subtypes in other smaller studies like CADSIP and GERFHS III based on the cerebrovascular knowledge portal hosted by International Stroke Genetics Consortium (ISGC) (**table 3**).

Functional annotation and association with related traits

Functional annotation revealed that the index SNP of 1p21 (rs71654444) is an eQTL for the gene *RWDD3*($P_{lowest}=1.74\times10^{-58}$) (**table e-12**). Data from multiple tissues suggested that rs71654444is an eQTL of *RWDD3*(**table e-12**). The variant rs118126757 was identified as an eQTL for *LILRA3*($P_{lowest}=1.74\times10^{-58}$). Similarly, we identified rs9936995 as an eQTL for the gene *KLHDC4*(**table e-12**). Search for association with related traits revealed that the variant rs71654444-*RWDD3* was associated with platelet distribution width ($P=1.15\times10^{-9}$) in the UK Biobank study population (**figure e-10 andtable e-13**) and also affected metabolism of steric acid ($P=7.36\times10^{-12}$) and palmitic acid ($P=4.44\times10^{-7}$)(**table e-14**).In order to check association of these top 8 loci with related risk factors we also searched diabetes knowledge portal and the results are presented in **table e-15,** here the most significant association was ofrs71654444-*RWDD3* with extreme chronic kidney disease (P=0.000669).

Replication of loci identified by previous GWAS

Status of previously reported 32 genome-wide significant loci which replicated in the MEGASTROKE dataset is shown in **table e-16**. Based on similar directionality of effect we

were able to replicate 19-21 of these loci depending upon stroke subtypes (**table e-16**). Checking for both directionality of effect and nominal significance at P<0.05 we observed association of 6 loci (*KCNK3, LOC100505841, Chr9p21, LINC01492, ZFHX3, SMARCA4–LDLR*) with at least one of the stroke subtypes(**table e-16**). The strongest association was observed for the locus *ZFHX3* with cardioembolic stroke in the current study (P=0.0004) which was also primarily associated with cardioembolic stroke in MEGASTROKE analysis. The Association of *ZFHX3* remained significant after correcting for multiple testing (0.05/32=0.0016).

Status of candidate gene studies performed in Indian population

In total 32 independent loci had been investigated for association with stroke in 50 publications (**appendix e-1 and table e-17**). We observed that 11 loci were associated with at least one stroke subtype in this study (P<0.05) and the strongest association was observed for the variant rs1800610 of TNF (P=0.0032) with small vessel disease stroke. However, none of the associations of the candidate genes previously studied in the Indian population were significant after correcting for multiple testing (0.05/32=0.0016).

Discussion:

This first GWAS of stroke in Indians identified eleven novel loci (8 based on single variant and 3 based on gene-based test) and replicated previously known loci. Using functional annotation and association with stroke-related risk factors variants in genes for novel pathways not implicated previously by genetic studies for stroke were also identified.

The locus1p21.3 with the index SNP rs71654444 affecting the expression of the gene *RWDD3* was our most important finding. This locus which was associated with cardioembolic stroke in this study at the genome significance level was also found to replicate in MEGASTROKE, the largest genetic study of stroke till date with cardioembolic stroke. Strong association with platelet distribution width and 18 carbon fatty acid metabolism also provide evidence that this locus might be affecting known biological mechanisms that affect cardioembolic stroke. Platelet distribution width is an established marker for embolism and intriguingly long carbon fatty acid metabolism is specifically implicated in cardioembolic stroke and not with other stroke subtypes.[11,12] Considering together these findings strengthen the evidence of stroke subtype-specific association suggesting the potential influence of 1p.21.3 variant with the risk of CE stroke. Based on gene expression data obtained on multiple tissues (GTEx portal) it was observed that the index SNP rs71654444 of 1p21 is an eQTL for *RWDD3*. The gene *RWDD3* (RWD domain containing 3) is also known as *RSUME*(RWD-domain-containing sumoylation enhancer) and has a conserved sequence of 110 amino acids containing repeats (R) of Tryptophan (W) and Aspartic acid (D). This RWD domain-containing protein helps in SUMOylation, a post-translational modification similar to ubiquitination, but involves the addition of a small ubiquitin-like modifier (SUMO)instead of ubiquitin.[13,14] several studies have provided evidence that activation of SUMOylation pathway protects against brain ischemia.[15–19] The evidence in the literature suggests that SUMOylationhas emerged as a potential therapeutic target for neuroprotection in brain ischemic in both local and global brain ischemia.[20,21]

Another variant rs9936995 located on 16q24 also reached the GWAS significance level for association with cardioembolic stroke. This locus also showed evidence of nominal association with cardio-aortic embolism (P=0.0052), and all ischemic strokes (P=0.0031) in the smaller

CADISP 2015 dataset (cerebrovascular knowledge portal). This locus is located near to the gene *ZCCHC14* which has been previously implicated in all stroke (MEGASTROKE) and small vessel disease stroke including white matter hyperintensities.[10,22] However the index SNP at this locus in the current study did not show any correlation with SNPs previously reported to be associated with stroke at this region (r2=0.001 between rs9936995 and rs12445022).

The index SNP rs118126757 of 19q13.42 locus was found to affect the expression of its nearest gene *LILRB2* and a nearby gene *TSEN34*. The *LILRB2* gene has been implicated in a rare vascular disorder Takayasu arteritis, which leads to inflammation of the aorta and its branches and its orthologs have been implicated in platelet activation.[23,24] However more evidence is required to associate this locus with stroke.

Among the three genes which were identified through gene-based studies, *SLC17A2* has been shown to be associated with carotid intima-media thickness and ankle-brachial index two important mechanisms associated with risk of stroke.[25,26]

While checking for replication of established loci for stroke identified through GWAS studies we were able to replicate many of them, 22 in total if we were to consider only directionality of effects. The sample size in this study is comparatively very small to some of the large studies like the MEGASTROKE hence we do not expect to replicate most of the loci if we were to consider association based on P-values. Despite which we still observed a strong association of ZFHX3 with cardioembolic stroke which remained significant even after correcting for multiple testing.

A large number of studies from India have investigated the association of polymorphism in candidate genes for association with stroke. Our search revealed at least 50 such studies that had

investigated 38 independent loci. We were interested in checking the status of these candidate genes in our study as very few of such loci (like NOS2) have been confidently implicated in stroke, especially previous GWAS have had little evidence in their support. This could have been due to allele frequency and linkage disequilibrium differences between Indian genetic make-up and other populations in which GWAS was performed. But we were only able to replicate a few of them at nominal P values and none were significant post-correction for multiple testing. This could reflect bias due to small sizes or false positives due to population stratification, etc of the candidate gene studies. Even the current study is smaller in sample size considering the ischemic and hemorrhagic stroke subtypes. Hence, larger studies in the Indian population are required to better resolve these issues.

We also investigated the non-genetic factors like the vascular risk factors as this is one of the largest studies on stroke from India. We observed very similar observations of association for risk of stroke as has been previously reported in many studies investigating different ethnicities. The modifiable risk factors like hypertension, diabetes, the status of smoking, and alcohol intake appeared as having high effect estimates across stroke subtypes.

A big limitation of the study is small sizes for subtypes of ischemic stroke and lack of subtyping for hemorrhagic stroke which can limit the interpretation of low-frequency variants and another limitation was the lack of replication in an Indian sample set. However, this first GWAS of stoke in Indians was still able to identify robustly associated loci, especially the 1q21 locus of cardioembolic stroke given the multiple strong evidence for association with traits that are associated with stroke and nominal replication in independent data sets. This also led to first time reporting of variants in genes of the SUMOlation pathway which is being actively investigated as a therapeutic option for protection against brain ischemia.

Conflict of Interest : None

Funding : The present study was funded by Department of Biotechology, Ministry of Science and Technology, Government of India (grant number Ref. No. - BT/01/COE/06/02/09-II). The funding support for sample collection from South India and East India was provided by UKIERI (UK-India Education Research Initiative).

Acknowledgement :

We thank All India Institute of Medical Sciences (AIIMS), New Delhi, India for providing support and resources for successful completion of this study. We sincerely acknowledge UKIERI for providing technical support and funding for collection of samples from East and South India. Finally, we sincerely acknowledge the support of study participants for their participation in the present study.

References

1. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.***18**, 439–458 (2019).
2. Bevan, S. *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke***43**, 3161–7 (2012).
3. Chauhan, G. & Debette, S. Genetic Risk Factors for Ischemic and Hemorrhagic Stroke. *Curr Cardiol Rep***18**, 124 (2016).
4. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.***51**, 584–591 (2019).
5. Kumar, A. *et al.* Genetics of ischemic stroke: An Indian scenario. *Neurol. India***64**, 29–37 (2016).
6. Indian Genome Variation, C. The Indian Genome Variation database (IGVdb): a project overview. *Hum Genet***118**, 1–11 (2005).
7. Indian Genome Variation, C. Genetic landscape of the people of India: a canvas for disease gene exploration. *J Genet***87**, 3–20 (2008).
8. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature***461**, 489–94 (2009).
9. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated genes in South Asia. *Nat Genet* (2017) doi:10.1038/ng.3917.
10. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet***50**, 524–537 (2018).

11. Vagdatli, E. *et al.* Platelet distribution width: a simple, practical and specific marker of activation of coagulation. *Hippokratia***14**, 28–32 (2010).

12. Sun, D. *et al.* A prospective study of serum metabolites and risk of ischemic stroke. *Neurology***92**, e1890–e1898 (2019).

13. Alontaga, A. Y. *et al.* RWD Domain as an E2 (Ubc9)-Interaction Module. *J. Biol. Chem.***290**, 16550–16559 (2015).

14. Geiss-Friedlander, R. & Melchior, F. Concepts in sumoylation: a decade on. *Nat. Rev. Mol. Cell Biol.***8**, 947–956 (2007).

15. Anderson, D. B., Zanella, C. A., Henley, J. M. & Cimarosti, H. Sumoylation: Implications for Neurodegenerative Diseases. *Adv. Exp. Med. Biol.***963**, 261–281 (2017).

16. Peters, M., Wielsch, B. & Boltze, J. The role of SUMOylation in cerebral hypoxia and ischemia. *Neurochem. Int.***107**, 66–77 (2017).

17. Zhang, H., Huang, D., Zhou, J., Yue, Y. & Wang, X. SUMOylation participates in induction of ischemic tolerance in mice. *Brain Res. Bull.***147**, 159–164 (2019).

18. Lee, Y. & Hallenbeck, J. M. SUMO and ischemic tolerance. *Neuromolecular Med.***15**, 771–781 (2013).

19. Yang, W. *et al.* Small ubiquitin-like modifier 3-modified proteome regulated by brain ischemia in novel small ubiquitin-like modifier transgenic mice: putative protective proteins/pathways. *Stroke***45**, 1115–1122 (2014).

20. Bernstock, J. D. *et al.* SUMOylation in brain ischemia: Patterns, targets, and translational implications. *J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab.***38**, 5–16 (2018).

21. Yang, W., Sheng, H. & Wang, H. Targeting the SUMO pathway for neuroprotection in brain ischaemia. *Stroke Vasc. Neurol.***1**, 101–107 (2016).

22. Traylor, M. *et al.* Genetic variation at 16q24.2 is associated with small vessel stroke. *Ann. Neurol.***81**, (2017).

23. Terao, C. *et al.* Genetic determinants and an epistasis of LILRA3 and HLA-B*52 in Takayasu arteritis. *Proc. Natl. Acad. Sci. U. S. A.***115**, 13045–13050 (2018).

24. Fan, X. *et al.* Paired immunoglobulin-like receptor B regulates platelet activation. *Blood***124**, 2421–2430 (2014).

25. Arya, R. *et al.* A genetic association study of carotid intima-media thickness (CIMT) and plaque in Mexican Americans and European Americans with rheumatoid arthritis. *Atherosclerosis***271**, 92–101 (2018).

26. Kardia, S. L., Greene, M. T., Boerwinkle, E., Turner, S. T. & Kullo, I. J. Investigating the complex genetic architecture of ankle-brachial index, a measure of peripheral arterial disease, in non-Hispanic whites. *BMC Med. Genomics***1**, 16 (2008).

Figure 1: Study design for the Indian stroke GWAS

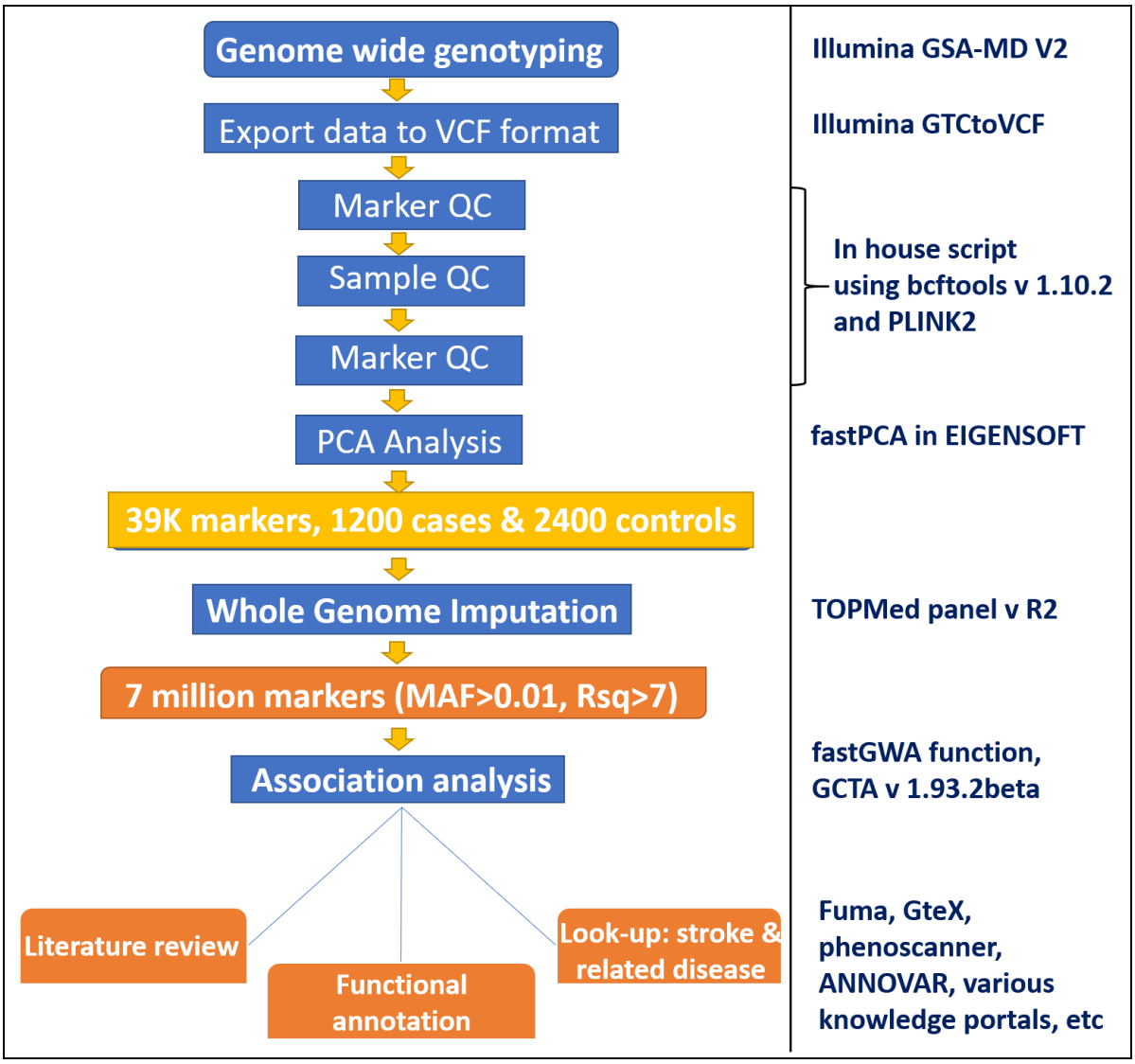

Figure 2: Manhattan plots for assication with stroke sub-types

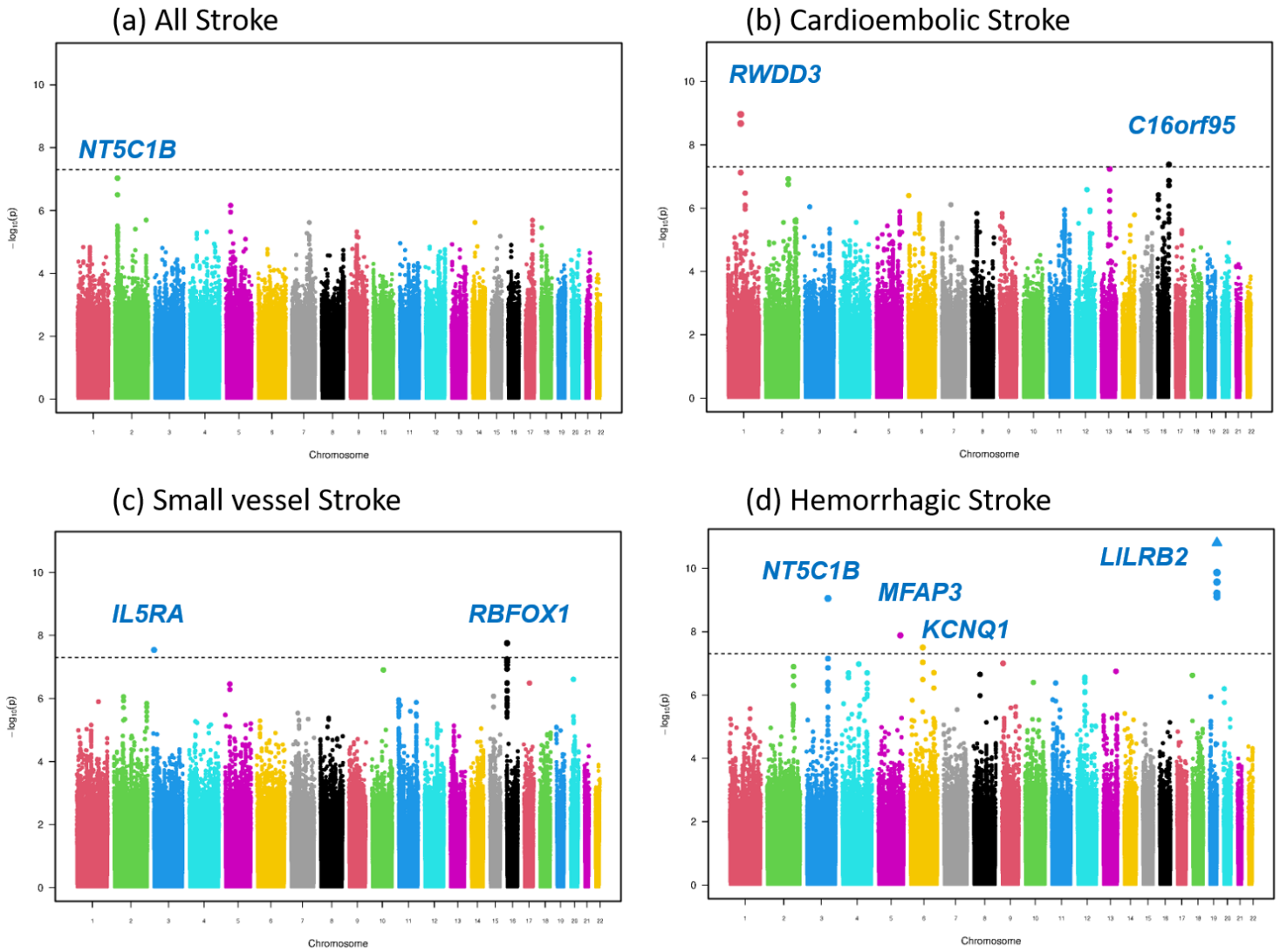

Figure 3: Regional plot of locus 1q21 associated with cardioembolic stroke

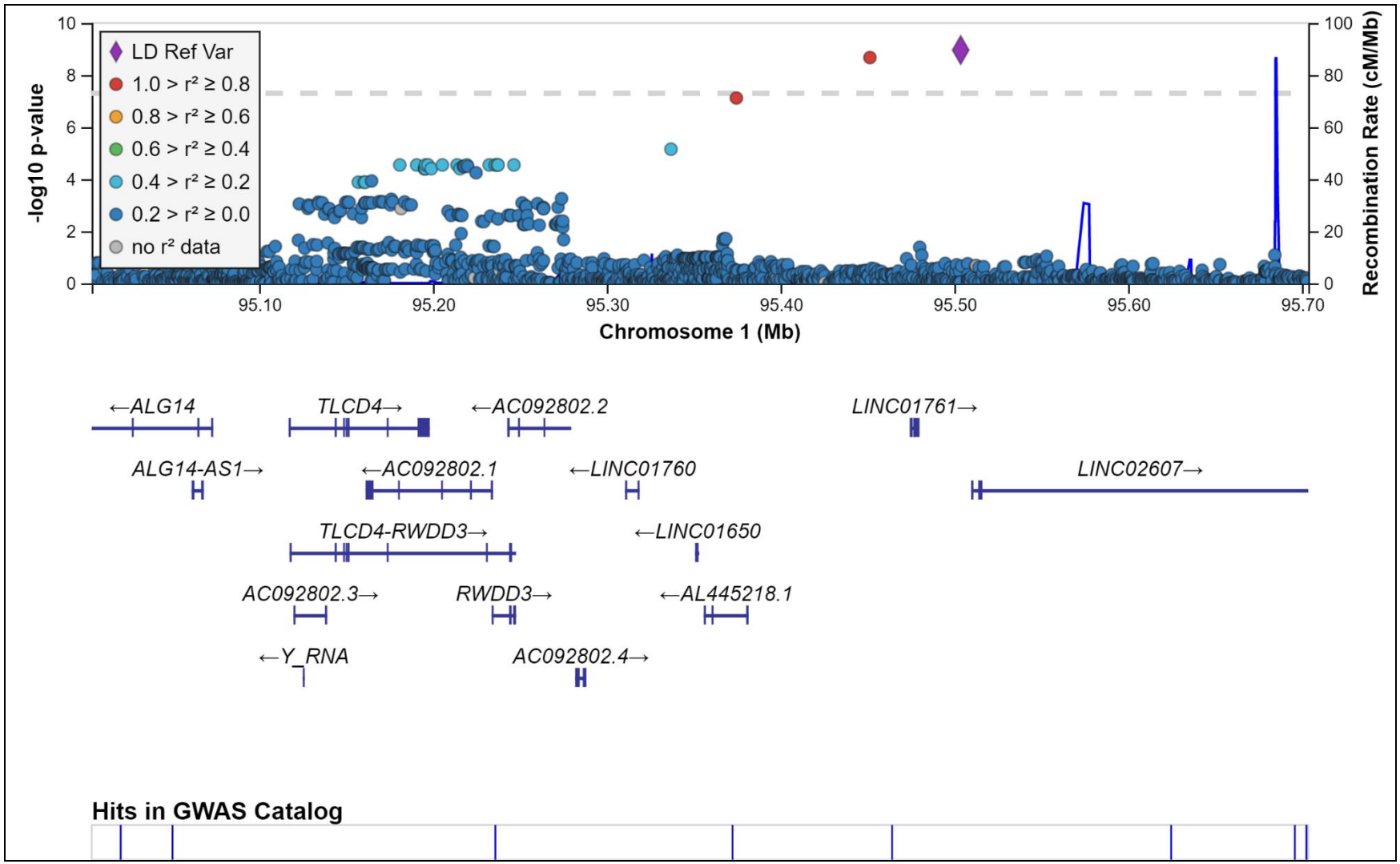

Table 1: Demographics characteristics and association with vascular risk factors

Variable	Control (n=2479)	AS (n=1609)	IS (n=1329)	LVD (n=439)	SVD (n=408)	CE (n=204)	HS (N=280)
Quantitative							
Age	46.46 ±15.39	54.24 ±13.86 P=3.9×10^{-54} *	54.50 ±14.31 P=1.6×10^{-50} *	55.24 ±15.39 P=2.9×10^{-28} *	56.42 ±13.70 P=6.9×10^{-34} *	55.35 ±14.95 P=2.8×10^{-15} *	52.74 ±11.10 P=4×10^{-11} *
SBP	128.49 ±15.49	143.32 ±26.69 P=1.5×10^{-89} *	139.13 ±23.80 P=3.7×10^{-52}*	139.50 ±25.53 P=7.1×10^{-31}*	140.65 ±23.58 P=3.5×10^{-37} *	137.36 ±21.59 P=2.2×10^{-13} *	162.96 ±30.52 P=4.8×10^{-165} *
DBP	81.11±9.64	86.39 ±14.83 P=1.5×10^{-36} *	84.87 ±14.76 P=1.9×10^{-18} *	85.05 ±15.87 P=1.9×10^{-11}*	85.82 ±14.76 P=1.2×10^{-15} *	84.67 ±13.46 P=2.09×10^{-6} *	93.50 ±12.98 P=8.4×10^{-76} *
Glucose	137.84 ±55.95	143.83 ±68.84 P=0.53	142.08 ±72.32 P=0.677	148.80 ±81.79 P=0.35	142.13 ±63.32 P=0.65	134.06 ±51.15 P=0.70	148.19 ±59.23 P=0.24
Dichotomous							
Male	1774 (71.50)	1176 (73.09) P=0.28	967 (72.70) P=0.43	324 (73.80) P=0.33	290 (71.08) P=0.84	151 (74.02) P=0.45	209 (74.60) P=0.27
Hypertension	463 (18.60)	786 (48.80) P=4.71×10^{-93}*	620 (46.60) P=2.4×10^{-74}*	186 (42.30) P=$3.7^{\times}10^{-28}$*	202 (49.51) P=9×10^{-43}*	101 (49.5) P=2.7×10^{-25}*	166 (59.20) P=3.3×10^{-53}*
Diabetes mellitus	82 (29.50)	348 (14.04) P=2.3×10^{-27}*	379 (28.50) P=2.3×10^{-27}*	138 (31.40) P=1.9×10^{-19}*	103 (25.20) P=7.5×10^{-09}*	56 (27.45) P=2.6×10^{-07}*	66 (23.57) P=2×10^{-5}*
Dyslipidemia	320 (12.90)	260 (16.10) P=0.004	231 (17.30) P=1×10^{-4} *	88 (20.05) P=7×10^{-5}*	57 (13.90) P=0.55	35 (17.10) P=0.08	29 (10.30) P=0.22
Atrial fibrillation	107 (4.32)	96 (5.97) P=0.017	93 (7.00) P=3×10^{-4} *	19 (4.30) P=0.98	11 (2.70) P=0.12	50 (24.50) P=3.3×10^{-32} *	3 (1.07) P=0.009
Myocardial infarction	92 (3.70)	91 (5.66) P=0.003	90 (6.78) P=2.3×10^{-5}*	19 (4.30) P=0.53	18 (4.42) P=0.48	43 (21.08) P=1.06×10^{-27} *	1 (0.36) P=0.001
Smoking status	149 (6.01)	316 (19.60) P=5.4×10^{-41}*	292 (21.90) P=9.8×10^{-49}*	106 (24.15) P=2.5×10^{-35}*	98 (24.02) P=1.9×10^{-33}*	32 (15.61) P=1.18×10^{-07} *	24 (8.54) P=0.094
Alcohol intake	76 (3.07)	245 (15.23) P=2.7×10^{-45}*	196 (14.75) P=1.31×10^{-40}*	63 (14.38) P=1.4×10^{-24} *	55 (13.40) P=7.5×10^{-21} *	19 (9.30) P=3.5×10^{-06} *	49 (17.50) P=3.4×10^{-28} *
Previous Stroke		104(6.40)	82(6.17)	31(7.08)	23(5.64)	11(5.39)	22(7.86)
FH stroke		111(8.40)	97(8.44)	33(8.87)	28(8.19)	11(6.40)	14(5.13

Abbreviations: IS, ischemic stroke; LVD, large vessel stroke; SVD, small vessel stroke; CE, cardiac embolic stroke; HS, hemorrhagic stroke; SBP, systolic blood pressure; DBP, diastolic blood pressure; FH stroke, family history of stroke. Data for quantitative traits is presented as mean ±SD with corresponding P-value. Data for dichotomous traits is presented as N (%) with corresponding P-value. The P-value in this table refers to the significance of association of the risk factor with stroke subtype as derived from a logistic regression using controls as the references.

*represents statistically significant P-value after correcting for multiple comparisons for performing association with 12 risk factors (0.05/12=0.0042)

Table 2: Association statistics of loci reaching genome wide significance in various stroke subtypes

Locus	**1p21.3**	**3p26.2**	**3q24**	**5q33.2**	**6q13**	**16p13.3**	**16q24.2**	**19q13.42**
Index SNP	rs71654444	rs114487517	rs146763130	rs118079585	rs80300510	rs9924207	rs9936995	rs118126757
Nearest-Gene(s)	*LINC01761; LINC02607*	*IL5RA; TRNT1*	*DIPK2A; LNCSRLR*	*MFAP3; GALNT10*	*KCNQ5*	*EEF2KMT; LINC01570*	*LOC101928708; LOC101928682*	*LILRB2*
Candidate Gene	*RWDD3*	*IL5RA*				*ZCCHC14*	*RBFOX1*	*LILRB2*
Chromosome	1	3	3	5	6	16	16	19
Position	95503558	3118595	144644768	154089521	73150642	5197852	87268693	54280214
Reference Allele	C	C	T	C	G	T	C	C
Effect Allele	A	T	C	A	A	A	T	G
All stroke	1.179 (1.055-1.317); $P=3.64\times10^{-3}$	1.102 (1.007-1.205); P=0.03	0.912 (0.840-0.991); P=0.03	1.090 (0.988-1.202); P=0.09	1.120(1.016-1.235); P=0.02	1.037 (1.015-1.059); $P=9.35\times10^{-4}$	1.113(1.036-1.197); $P=3.64\times10^{-3}$	1.157 (1.063-1.258); $P=7.11\times10^{-4}$
All ischemic stroke	1.197 (1.071-1.337); $P=1.52\times10^{-3}$	1.104 (1.008-1.209); P=0.03	0.979 (0.899-1.067); P=0.63	0.995 (0.886-1.117); P=0.93	1.052 (0.939-1.178); P=0.39	1.043 (1.021-1.065); $P=1.38\times10^{-4}$	1.102 (1.024-1.186); $P=9.29\times10^{-3}$	1.096 (0.998-1.205); P=0.05
Large vessel stroke	0.966 (0.864-1.081); P=0.55	1.010 (0.929-1.098); P=0.82	0.968 (0.900-1.040); P=0.37	0.972 (0.879-1.074); P=0.57	1.038(0.939-1.147); P=0.47	1.006 (0.988-1.025); P=0.51	1.058 (0.991-1.130); P=0.09	1.071 (0.987-1.163); P=0.10
Small vessel stroke	1.096 (0.991-1.213); P=0.07	1.226 (1.141-1.318); **$P=2.91\times10^{-8}$**	1.010 (0.942-1.083); P=0.78	1.039 (0.943-1.144); P=0.44	0.979 (0.882-1.088); P=0.69	1.053 (1.034-1.072); P=**1.76×10^{-8}**	1.015 (0.952-1.083); P=0.65	0.988 (0.907-1.076); P=0.78
Cardioembolic stroke	1.265 (1.173-1.365); **$P=1.09\times10^{-9}$**	0.967 (0.903-1.035); P=0.33	1.013 (0.958-1.070); P=0.66	0.942 (0.867-1.023); P=0.16	1.045(0.965-1.132); P=0.28	1.000 (0.985-1.014); P=0.95	1.147 (1.092-1.204); P=**4.21×10^{-8}**	1.088 (1.019-1.162); P=0.01
Hemorrhagic stroke	0.993 (0.914-1.079); P=0.87	1.022 (0.959-1.088); P=0.50	0.853 (0.811-0.898); **$P=8.95\times10^{-10}$**	1.192 (1.122-1.267); **$P=1.31\times10^{-8}$**	1.193 (1.121-1.270); **$P=3.16\times10^{-8}$**	0.997 (0.984-1.011); P=0.72	1.057 (1.006-1.112); P=0.03	1.209 (1.145-1.277); **$P=1.14\times10^{-11}$**

Index SNP,refers to the SNP with the lowest P-value in that locus; Nearest Gene, it refers to the gene on which the SNP is locate (if present in genic regions) or genes that flank the SNP (if present in intergenic region), this is as per positional annotation provided by

ANNOVAR. Candidate Gene, it refers to genes based on functional annotation like eQTL analysis or based on literature reference for association with stroke or related trait; Reference allele, it is the reference allele in the reference genome; Effect Allele is the alternate allele in the reference genome and the allele representing the direction of effect. Genome wide significant P-values ($P<\mathbf{5\times10^{-8}}$) are highlighted in bold font. Detailed association statistics of these variations in presented in Supplementary Table AAA

Table 3: Replication of variants in neurological and related traits based on Cerebrovascular Disease Knowledge Portal of ISGC

Locus	Index SNP	Primary traits, Indian stroke GWAS	Trait	dataset	P value	Direction of effect	OR	MAF	effect	samples
1p21.3	rs71654444	Cardioembolic stroke	TOAST cardio-aortic embolism	MEGASTROKE GWAS	0.0155	↑		0.074	1.09	521612
1p21.3	rs71654444	Cardioembolic stroke	All ICH, Dataset	GERFHS III 2017	0.0203	↑		0.0665		1201
1p21.3	rs71654444	Cardioembolic stroke	TOAST other undermined	CADISP 2015	0.0287	↑		0.0702		9814
16q24.2	rs9936995	Cardioembolic stroke	All ischemic stroke	CADISP 2015	0.00312	↑	1.5	0.0494		9814
16q24.2	rs9936995	Cardioembolic stroke	TOAST cardio-aortic embolism	CADISP 2015	0.00552	↑	1.75	0.0494		9814
16p13.3	rs9924207	Small vessel stroke	Body fat percentage	Body fat percentage GWAS	0.0148	↑		0.232	0.0529	100716
19q13.42	rs118126757	Hemorrhagic stroke	TOAST large artery atherosclerosis	MEGASTROKE GWAS	0.0166	↑		0.0521	1.2	521612
19q13.42	rs118126757	Hemorrhagic stroke	Triglycerides	GLGC GWAS	0.0373	↓			-0.0605	188577

Cerebrovascular Disease Knowledge Portal: http://cerebrovascularportal.org/home/portalHome