Genome-wide association study of ischemic stroke risk in Sickle Cell confirms ADAMTS2, CDK18, uncovers 12 novel loci.

Eric Jay Earley¹, Shannon Kelly²,³, Fang Fang¹, Cecilia Salete Alencar⁴, Daniela de Oliveira Werneck Rodrigues⁵, Dahra Teles Soares Cruz⁶, Ester Sabino⁷, Brian Custer³,⁸, Carla Dinardo⁷, Grier P. Page¹ on behalf of the International Component of the NHLBI Recipient Epidemiology and Donor Evaluation Study (REDS-III) and the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

¹GenOmnics, Bioinformatics, and Translational Research Center, RTI International, Research Triangle Park, Durham, NC, USA;
²Benioff Children’s Hospital, University of San Francisco, California, USA;
³Vitalant Research Institute, San Francisco, California, USA;
⁴Laboratório de Medicina Laboratorial LIM 03- HCFMUSP, São Paulo, Brazil;
⁵Fundação Hemominas Juiz de Fora, Brazil;
⁶Department of Hematology, Fundação de Hematologia e Hemoterapia de Pernambuco, HEMOPE, Pernambuco, Brazil;
⁷Instituto de Medicina Tropical, University of São Paulo, Brazil
⁸Department of Laboratory Medicine, University of California, San Francisco, USA.

Running Title: Ischemic Stroke GWAS in SCD Brazil cohort

Key Words: GWAS, sickle cell disease, Brazil, stroke, ischemic stroke

Corresponding author: Eric Jay Earley eearley@rti.org
3040 E. Cornwallis Rd., Research Triangle Park, NC 27709-2194

Word count: 5978 / 6,000

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ABSTRACT

Background: Ischemic stroke is a common complication of sickle cell disease (SCD) and without screening or intervention can affect 11% of children with SCD before the age of 20. This study sought to find genetic biomarkers for risk of stroke occurring at younger ages.

Methods: Within the Trans-Omics for Precision Medicine (TOPMed), a genome-wide association study (GWAS) of ischemic stroke was performed on 1,333 individuals with SCD from Brazil (178 cases, 1155 controls). Via a novel proportional hazards analysis approach, we searched for variants associated with strokes occurring at younger ages.

Results: Fourteen genomic regions were associated with early ischemic stroke at genome wide significance (P<5x10^{-8}). This included variants near two genes which have been previously linked to non-SCD early onset stroke (<65 years): ADAMTS2 (rs147625068, P=3.70 x 10^{-9}) and CDK18 (rs12144136, P=2.38 x 10^{-9}), respectively. Individuals harboring multiple risk alleles exhibited increasing rates of stroke at earlier timepoints (P < 0.001, Gehan-Wilcoxon) than those carrying only one. Enrichment tests suggest systemic dysregulation of gene expression in the hypothalamus (P = 0.03, FDR), substantia nigra (P = 0.03), spleen (P = 0.005) and coronary (P = 0.0005), tibial (P = 0.03) and aorta arteries (P = 0.03).

Conclusions: This findings from this study support a model of shared genetic architecture underlying ischemic stroke risk between SCD individuals and non-SCD individuals <65 years. In addition, results suggest an additive liability due to carrying multiple risk alleles.

INTRODUCTION

Ischemic stroke is one of the most common comorbidities of sickle cell disease (SCD), occurring 20 times more frequently than seen in non-SCD children\(^1,2\). Without early preventative therapy, 11% of children with SCD experience at least one overt ischemic stroke before the age of 20\(^3\), with recurrence ranging from 60-92%\(^4\) and mortality of roughly 5%\(^5\). Ischemic stroke is more common than hemorrhagic stroke in children, with hemorrhagic stroke more common in ages 20-35 years. The pathophysiology of ischemic stroke within SCD is complex\(^1\), and the current standard for risk assessment, transcranial doppler ultrasonography (TCD) of cerebral vessels, is difficult to perform on children under age 3. Also, some individuals with normal TCD still experience stroke\(^3,6,7\).

Genetic risk of stroke has been investigated in SCD and non-SCD populations. Within SCD, stroke risk is highest in HbSS homozygotes compared to other SCD genotypes, whereas the coinheritance of alpha-thalassemia is protective against stroke\(^1\). There is mounting evidence for the existence of other heritable genetic risk factors for stroke\(^8\text{-}15\); however, many of these are conflicting or non-replicated. A GWAS of 677 African Americans with SCD discovered two nonsynonymous variants in the genes GOLGB1 and ENPP1 associated with protection from stroke\(^13\); however, the same variant in ENPP1 showed higher risk for stroke within a different Brazilian SCD cohort and GOLGB1 was not replicated\(^15\).

Investigating genetic risk within non-SCD stroke may provide additional insight into the similarity or differences compared to SCD. Within non-SCD pediatric ischemic stroke studies, risk variants near genes ADAMTS2, ADAMTS12, and ADAMTS13 have been discovered, and both ADAMTS2 and ADAMTS12 have been replicated\(^16,17\). In separate studies, variants near CDK18\(^18\), HABP2\(^19\), were linked to early
onset ischemic stroke but have not been replicated. Gaps in our knowledge of the genetic contributors to ischemic stroke risk within SCD remain. To address this, we conducted the largest genome-wide association study of stroke in individuals with sickle cell disease (N=1,333). This was conducted within the Brazil Sickle Cell Disease Cohort Study as part of the Recipient Epidemiology Donor Evaluation Study III (REDS-III) \(^2\).

**METHODS**

Two novel approaches were implemented in this study. Variants were measured via whole genome sequencing, avoiding the need to impute on reference panels which may have little to no overlap in ancestry with the study cohort. In addition, a Cox proportional hazards model was used to estimate differences in time to stroke, allowing for the discovery of genetic variants associated with earlier stroke events.

**Study Population**

The Brazil SCD Cohort Study is part of the National Institutes of Health, National Heart Lung and Blood Institute (NHLBI) REDS-III program. \(^2\) The Brazil National Research Ethics Commission, local ethical committees at each participating center and the Institutional Review Boards at University of California, San Francisco (UCSF) and the REDS-III data coordinating center, Research Triangle Institute, International (RTI) all reviewed and approved the study.

2,793 sickle cell disease individuals were enrolled at 6 different centers in Brazil: Hemominas Belo Horizonte, Juiz de Fora, and Montes Claros, Hemorio, Rio de Janeiro, Instituto de Tratamento do Câncer Infantil, Sao Paulo, and Hemope, Recife. Enrollment included interviews, medical abstraction, and blood collection. Written informed consent was obtained from participants ≥18 years or from guardians of younger patients, and assent was obtained for children aged 7 to 17. Medical records were abstracted for clinical history using standardized definitions. \(^2\) This included a history of ischemic stroke defined as an acute neurological syndrome resulting from impaired cerebral blood flow without evidence of hemorrhage. Diagnostic criteria included either an MRI or CT scan showing an infarctive CNS event consistent with symptoms and signs or diagnosis based on examination and clinical history with neurologic symptoms/signs lasting >24 hours.

**DNA collection and sequencing**

Whole blood was collected in EDTA, and DNA was extracted from the buffy coat via alcohol precipitation, quantified with rtPCR, and normalized to 10 ng/ul. SCD genotypes were confirmed using allele-specific pyrosequencing (Qiagen, Hilden, Germany) \(^2\) and Sanger sequencing of exons 1 and 2 of \textit{HBB} if the pyrosequencing results conflicted with medical records. Additional sequencing of \textit{HBB} exon 3, introns, and promoter was performed for samples with unresolved genotypes at the Hemoglobinopathy Reference Laboratory at UCSF Benioff Children's Hospital Oakland.

Whole genome sequencing was conducted within the NHLBI Trans-Omics for Precision Medicine (TOPMed) program \(^2\). Average coverage was 38x, and variants were called jointly across roughly 140k samples, including the N=1,333 individuals in the current study, using the GotCloud pipeline \(^2\) on data freeze 6a. Variants were filtered to >1% frequency, resulting in a final list of 14.3M variant autosomal sites.
Exclusion Criteria

Individuals were excluded from this analysis for the following reasons: not having TOPMed DNA sequence data, being a duplicate sample, inability to link DNA to clinical records, outlier on PC1-3, missing age or stroke status, and not being an HbSS homozygote. Finally, we also excluded individuals with no history of stroke who were either undergoing CTT or had abnormal TCD (Suppl. Fig 1).

Statistical Analyses

A cox proportional hazards model was used to measure SNP associations. Age at stroke was used for cases, and age at enrollment was right censored for controls without a history of ischemic stroke. Principal components (PC) were calculated from LD pruned genotypes using the entire freeze6a TOPMed cohort of roughly 140k individuals and was then was subset to just the N=1,333 participants in the study. Regressions and statistical tests were performed using the R package survival with the following model:

\[
(Stroke, Age) \sim PC1 - 10 + SNP + Error
\]

Phased genotypes were converted to integers prior to populating models as co-variates – 0 = homozygous reference; 1 = heterozygous; 2 = homozygous alternate. Genome-wide significance was defined as \( P < 5 \times 10^{-8} \). Conditional regressions were performed on candidate 5Mb genomic regions surrounding the lead SNPs defined by lowest \( P \)-value. No independent loci within the candidate regions were observed. Code for this analysis is available online: https://github.com/earleyej/cox-gwas

All genotype-level analyses were performed in a HIPPA compliant environment on DNAnexus (https://www.dnanexus.com/). Annotated views of genomic candidate regions were created with LocusZoom\(^{27}\) using 1k genomes reference for linkage disequilibrium (LD). Tests on GWAS summary statistics, including gene-level association and enrichment tests were performed using the online portal Functional Mapping and Annotation of GWAS (FUMA\(^{28}\)). Gene-level association was performed on the full 14M p-values using Multi-marker Analysis of Genomc Annotation (MAGMA)\(^{29}\) using default settings. Gene set pathway analysis was performed on the Gene Ontology (GO) database (N=9,988). Colocalization analysis was performed with the R package coloc\(^{30}\) on 1Mb regions surrounding each independent genome-wide significant SNP with the lowest p-value for that region (aka ‘lead’ SNP). Additional gene set enrichment was performed on the GWAS Catalog of genes and associated conditions, as well as the Genotype-Tissue Expression (GTEx v8) databases and were performed using hypergeometric tests and corrected for multiple hypothesis testing using Benjamini-Hochberg (FDR). Candidate genes were defined as being <10kb of a lead SNP or were targets of a cis-eQTL <10kb of a lead SNP.

Multi-variant survival analysis

Individuals were put in categories reflecting the number of lead SNPs they carried. All categories were required to have at least 10 participants, for a total of 5 categories (0, 1, 2, 3, and ≥4). These were used as categorical variables in a univariate survival analysis. \( P \)-values were calculated using Gehan-Wilcoxon, a non-parametric test which has more sensitivity in detecting differences between groups where the hazard ratio is higher at earlier timepoints.
RESULTS

Cohort characteristics

Of the 2,793 SCD individuals enrolled in the REDS-III Brazil SCD Cohort, 1,460 were excluded, with a final set of 1,333 individuals (Suppl. Fig. 1). Younger age was strongly associated with ischemic stroke (OR=0.9; CI=0.89-0.92; P<0.001; Figure 1A) but not sex (OR=0.9; CI, 0.65-1.24) or Hemocenter (OR=0.98; CI, 0.89-1.08; P=0.7). Genome-wide association was conducted on HbSS homozygotes only (N=1,333). Within this set were 178 (13.4%) individuals who experienced ischemic stroke (Table 1) with a median age at stroke of 9.5 years (Table 1, Fig 1A).

Genetic architecture of ischemic stroke in Brazilian Sickle Cell

We observed 28 genome-wide significant variants (Suppl. Table 1) across 14 independent regions associated with stroke occurring at younger ages (Figure 1B, Table 2, Suppl. Fig 2). No genomic inflation in P-values was observed (Supplemental Figure 3). Most variants sites were intergenic; however, three regions were intronic to genes CUX1, KIAA1217, and CNNM2. Also included in this list of 14 lead SNPs were three regions previously linked to stroke – ADAMTS2, CDK18, FUT817-19.

Out of the 14 genome-wide lead variants, 5 were only observed in heterozygous state whereas 9 were observed in both heterozygous and homozygous variant state. Individuals possessing at least one of these 14 variants exhibited increased risk for ischemic stroke at earlier ages (P < 0.001 for all 14 variants, Gehan-Wilcoxon) (Suppl Figure 4).

Multiple variant survival analysis

Individuals with zero lead SNPs had a stroke rate of 4% (n=716), whereas those with at least one of the 14 lead variants had an increased rate at 14% (n=430), and this rate increased even further for individuals with two (36%; n=125), three (70%; n=50), or four or more variants (83%; n=12), and this trend was significant (P < 0.001, Gehan-Wilcoxon; Figure 2).

Cis-co-localization analysis of ischemic stroke using GTEx

SNPs within the 14 candidate regions were then searched against the GTEx database of cis-eQTL variant associations across different tissue types. The lead SNP on chromosome 1, rs12144136, is a cis-eQTL for the gene CDK18 within the Skeletal Muscle tissue category (Suppl. Table 2) which appears to upregulate this gene. No other SNPs at genome-wide significance showed evidence of cis-eQTL activity.

We assessed co-localization of cis-eQTL associations with 24 different tissue types relevant to stroke. Three of the 14 candidate regions showed strong evidence (PP > 0.9) of colocalization with a cis-eQTL (Suppl. Table 2). On chromosome 12, two cis-eQTLs were found for genes LUM in the frontal cortex (rs77217583) and NUDT4 (rs117881990) in the hypothalamus. On chromosome 14, one cis-eQTL was found for RAB15 in the substantia nigra; and on chromosome 15, one cis-eQTL was found for VPS13C in cervical spinal cord. Three other moderately strong signals (P > 0.5) were detected on chromosome 8—one cis-eQTL for WRN expression in heart left ventricle and two cis-eQTLs for PURG in the substantia nigra.

Gene-level Association
We performed a gene-level association on 19,021 protein coding genes using the 14M site summary statistics (Supplemental Figure 5; Supplementary Table 3). No genes exhibited genome-wide significant enrichment defined at the Bonferroni-corrected alpha of 2.6x10^{-6}; however, three genes exhibited some evidence of association (P<10^{-4}), including GPR15 (P=3.9x10^{-5}), MMP26 (P=4.6x10^{-5}), and GOLT1B (P=7.5x10^{-5}).

Gene-set enrichment analysis was performed to predict impact on biological functions using the Gene Ontology (GO) database (Supplemental Table 4). While no gene set achieved the Bonferroni-corrected p-value of 5x10^{-6}, top results at the P<0.001 threshold included the platelet dense tubular network (P = 3.63x10^{-5}), the hemoglobin complex (P = 0.0007), gas transport (P = 0.0006), and regulation of systemic arterial blood pressure by hormone (P = 0.0008).

**Candidate-gene Enrichment Analysis**

To assess potential tissue-specific expression, enrichment tests were performed on the GTEx catalog using 18 candidate genes selected using two criteria: located <10kb of a lead SNP; or target of an eQTL located <10kb of a lead SNP (Suppl. Table 5). Significant enrichment was observed in multiple tissue classes, including upregulated genes in the coronary (8 genes, P = 0.0005, FDR), tibial (9, P = 0.03, FDR) and aorta arteries (7, P = 0.03, FDR), as well as downregulated genes in the spleen (7, P = 0.005), pancreas (13 genes, P = 0.02), esophagus mucosa (8, P = 0.03), hypothalamus (10, P = 0.03), and substantia nigra (11, P = 0.03) (Figure 3A, Suppl. Table 6).

Gene enrichment test using the GWAS Catalog uncovered significant associations to disease categories of white matter lesion progression (CNNM2, NT5C2, P = 0.01) and coronary artery disease/large artery stroke (CNNM2, NT5C2, P = 0.04)(Figure 3B). A second enrichment test was performed on an expanded list of 35 genes (<100kb of a lead SNP), and this set exhibited significant overlap with genes linked to hypertension (5 genes, P = 7.8x10^{-7}), white matter lesion progression (3 genes, P = 4.67x10^{-7}), mean arterial pressure (8, P = 7.81x10^{-11}), microalbuminuria (4, P = 1.86x10^{-7}), coronary artery disease (7, P = 8.93x10^{-6}), and coronary artery disease/large artery stroke (2, P = 0.0007) (Suppl. Fig 6).

**Comparison to previous ischemic stroke studies in SCD and non-SCD**

Previous candidate SNP studies of stroke in SCD have evaluated 12 SNPs, as well as one additional SNP from a GWAS. None of these candidates SNPs reached genome-wide significance in the current study (Suppl. Table 7). Expanding the search to surrounding 1Mb region centered on the 13 candidate SNPs uncovered the lead SNP rs147625068 (P = 3.7x10^{-5}) near gene LTC4S on chromosome 9.

A similar search was performed for non-SCD early onset ischemic stroke candidate genes: ADAMTS2 (rs469568), ADAMTS12 (rs1364044), CDK18 (rs77571454), HABP2 (rs11196288) [16-19]. In the current study, the variant rs77571454 near gene CDK18 was not observed, and none of the other three variants reached genome-wide significance (Suppl Table 8). Searching within the 1Mb region centered on these four candidate SNPs uncovered the two lead SNPs rs12144136 (P = 2.38x10^{9}), near CDK18, and rs147625068 (P = 3.7x10^{-5}), near ADAMTS2, and one additional variant near ADAMTS12 (rs534089428, P = 2.43x10^{-5}).

**DISCUSSION**
We have discovered 14 genetic regions associated with ischemic stroke occurring at earlier ages in SCD individuals with homozygous HbSS genotype, encompassing 18 candidate genes. This study included 1,333 HbSS individuals, the largest genome-wide evaluation of ischemic stroke within SCD to date. We implemented a Cox proportional hazards model to estimate risk at earlier timepoints. Results included gene candidates previously linked with non-SCD early-onset stroke – *ADAMTS2, CDK18*; genes linked to cardiovascular disease risk factors – *BTG1, CNMM2, CUX1, EMARD, FUT8, KIAA1217, LUM, NT5C2, NUDT4, PKD1L1, WRN*; and eight novel genes – *SH3RF3, RAB15, RPS3AP52, VPS13C*, and *TLE6*. To our knowledge this is the first independent replication of *CDK18* and *FUT8*.

Previous candidate gene studies of stroke in SCD replicated five alleles within genes *TGFBR3, TEK, ANXA2, IL4R, and ADCY9*; however, the single published GWAS of ischemic stroke in SCD did not replicate these and instead proposed two novel genes – *GOLGB1* and *ENPP1*. We did not observe replication of the candidate genes nor the GWAS variants. The reasons for this are unclear, although we should note the previous GWAS enrolled African Americans whereas the current work enrolled Brazilians, each having distinct histories of admixture.

Results from this work replicate stroke risk genes for non-SCD early onset stroke: *ADAMTS2* and *CDK18*. Members of the *ADAMTS* gene family, in particular *ADAMTS2* and *ADAMTS12*, have been linked to pediatric stroke in two separate U.S. populations. *ADAMTS2* also appears to have a role in cerebral aneurysm and cardiac muscle hypertrophy possibly through regulation of the extracellular matrix. We did observe one variant, rs534089428, within 500kb of *ADAMTS12* at $P = 2 \times 10^{-5}$. *CDK18* has not been replicated until now.

High reticulocyte count is a reliable risk factor for ischemic stroke in sickle cell individuals. Candidate genes from the current study, *CUX1, CNNM2*, and *NUDT4*, have been previously associated with higher reticulocyte counts. *CNMM2* has also been linked to numerous stroke related traits – hypertension, coronary artery disease, and mean corpuscular hemoglobin. *CNMM2* gene is expressed throughout the body but is particularly high in the brain, kidney, and endocrine tissue and appears to play a role in renal magnesium uptake.

Pathway analysis also uncovered enrichment of variants in genes involved with the platelet dense tubular network. Platelets have been shown to play a role in SCD pathophysiology, and higher platelet aggregation and hypercoagulability has been observed in blood from non-SCD ischemic stroke patients. Results from the current work suggest a systemic dysregulation of platelet activation-sequestration within SCD children suffering from strokes, but further studies will be required to confirm this.

In conclusion, we have identified risk alleles associated with earlier ischemic stroke. Findings from this work have confirmed previously identified non-SCD early onset ischemic stroke genes – *ADAMTS2* and *CDK18*– and discovered 16 novel genes. Future studies will be needed to disentangle the potentially complex interaction among these alleles.

**Acknowledgements** Special thanks to patients and research staff at each hemocenter in Brazil participating in this study. We recognize the following people for their commitment and contribution to this project: Alfredo Mendrone Jr., Cesar de Almeida Neto, Roberta Calcucci, Erivanda Bezerra, Carolina Miranda, Franciane Mendes de Oliveira, Valquíria Reis, Nayara Durte, Barbara Malta, Jose Wilson Sales,
Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome Sequencing for "NHLBI TOPMed: REDS-III_Brazil" (phs001468.v3.p1) was performed at Baylor (HHSN268201500015C, HHSN268201600033I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

**Funding:**

The NHLBI REDS-III program was supported by NHLBI contracts HHSN2682011-00001l, -00002l, -00003l, -00004l, -00005l, -00006l, -00007l, -00008l, -00009l, 75N2019D00033, and R21 HL135367.

**REFERENCES**


**TABLES**

**Table 1. Demographics of HbSS Individuals within the REDS-III Brazil SCD Cohort**
<table>
<thead>
<tr>
<th>Total cohort size</th>
<th>N=1,333</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>720 (54.0)</td>
</tr>
<tr>
<td>Male</td>
<td>613 (46.0)</td>
</tr>
<tr>
<td><strong>Age at enrollment, mean (±SD), y</strong></td>
<td>20.6 (±13.1)</td>
</tr>
<tr>
<td><strong>Age at first stroke, mean (±SD), y</strong></td>
<td>9.5 (±9.4)</td>
</tr>
<tr>
<td><strong>Stroke, no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>cases</td>
<td>178 (13.4)</td>
</tr>
<tr>
<td>controls</td>
<td>1155 (86.6)</td>
</tr>
<tr>
<td><strong>Hemocenter, no. (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Hemominas BH</td>
<td>313 (23.5)</td>
</tr>
<tr>
<td>Hemominas JFO</td>
<td>149 (11.2)</td>
</tr>
<tr>
<td>Hemominas MOC</td>
<td>205 (15.4)</td>
</tr>
<tr>
<td>Hemope</td>
<td>285 (21.4)</td>
</tr>
<tr>
<td>Hemorio</td>
<td>332 (24.9)</td>
</tr>
<tr>
<td>ITACI SP</td>
<td>49 (0.04)</td>
</tr>
<tr>
<td>Chr</td>
<td>Position (b38)</td>
</tr>
<tr>
<td>-----</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>205,490,782</td>
</tr>
<tr>
<td>2</td>
<td>109,059,503</td>
</tr>
<tr>
<td>5</td>
<td>179,375,279</td>
</tr>
<tr>
<td>6</td>
<td>120,572,043</td>
</tr>
<tr>
<td>7</td>
<td>47,681,969</td>
</tr>
<tr>
<td>7</td>
<td>101,180,550</td>
</tr>
<tr>
<td>8</td>
<td>31,319,867</td>
</tr>
<tr>
<td>10</td>
<td>23,744,565</td>
</tr>
<tr>
<td>10</td>
<td>102,993,950</td>
</tr>
<tr>
<td>12</td>
<td>92,093,367</td>
</tr>
<tr>
<td>13</td>
<td>68,609,784</td>
</tr>
<tr>
<td>14</td>
<td>65,317,148</td>
</tr>
<tr>
<td>15</td>
<td>61,683,898</td>
</tr>
<tr>
<td>19</td>
<td>2,961,825</td>
</tr>
</tbody>
</table>

Chr = Chromosome; AF = Allele Frequency; HR = Hazard Ratio; 95% CI = 95% Confidence Interval; Genomic positions in GRCh38 build.
Figure 1. (A) Histogram of age at first ischemic stroke for the GWAS cohort. (B) Manhattan plot of results from proportional hazards GWAS P-values for ischemic stroke in HbSS SCD (N=1,333). Horizontal dotted line represents the genome-wide significance threshold of $P = 5 \times 10^{-8}$. The nearest gene to each lead genome-wide significant SNP is annotated with color representing confirmation of known genes linked to early onset stroke (red), genes linked to cardiovascular diseases and risk factors (black), and novel genes in blue.
Figure 2. Kaplan-Meier plot of individuals harboring multiple co-occurring genome-wide significant risk alleles for early stroke. Curves represent subsets of the study cohort who possessed zero (red, N=716), one (olive, N=430), two (green, N=125), three (blue, N=50), or four or more (purple, N=12) stroke risk alleles in any combination from the 14 lead SNPs. (+) represents right-censored data.
Figure 3. Gene set enrichment test results for 18 candidate genes. (A) Barplots of transformed P-values (FDR) from hypergeometric enrichment test of 18 genes located within 10kb of a lead SNP (15 genes) or were predicted targets of a cis-eQTL in the same region (3 genes). Red bars show tissues observed at $P < 0.05$. Left panel, general tissue classes; right panel, specific tissue types. (B) Horizontal barplots of significant enrichment results against the GWAS Catalog. Red, proportion of genes; blue, transformed p-values (FDR); yellow, genes.
**Known Stroke loci**
Linked to CVD & risk factors

**Novel**

- PKD1L1
- WRN
- CUX1
- ERMARD
- ADAMTS2
- CDK18
- KIAA1217
- CNNM2, NTSC2
- BTG1
- RPS3AP52
- FUT8
- VPS13C
- TLE6

---

CC-BY-NC-ND 4.0 International license

It is made available under a
is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

(which was not certified by peer review)

The copyright holder for this preprint this version posted August 23, 2022.

https://doi.org/10.1101/2022.08.22.22279082

doi: medRxiv preprint

---

**A**

Age at Ischemic Stroke (Years)

Number of Individuals

**B**

-log(P-value)

CDK18, SH3RF3, ADAMTS2, ERMARD, CUX1, PKD1L1, WRN, KIAA1217, CNNM2, NTSC2, BTG1, RPS3AP52, FUT8, VPS13C, TLE6
The graph shows the proportion of individuals not having a stroke across different age groups for varying numbers of co-occurring risk alleles. The x-axis represents age (years) ranging from 0 to 60, while the y-axis represents the proportion not having a stroke, ranging from 0.00 to 1.00. Different lines correspond to different numbers of risk alleles: 0, 1, 2, 3, and 4+. The data suggests a decrease in the proportion not having a stroke as the number of risk alleles increases, indicating a higher likelihood of stroke with more risk alleles.
A

B

Proportion of overlapping genes in gene sets

Enrichment P-value

overlapping genes

Proportion

-log10 adjusted P-value

White matter lesion progression
Schizophrenia
Coronary artery disease or large artery stroke

CC-BY-NC-ND 4.0 International license
It is made available under a
is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

(which was not certified by peer review)

The copyright holder for this preprint this version posted August 23, 2022.

; https://doi.org/10.1101/2022.08.22.22279082
doi: medRxiv preprint