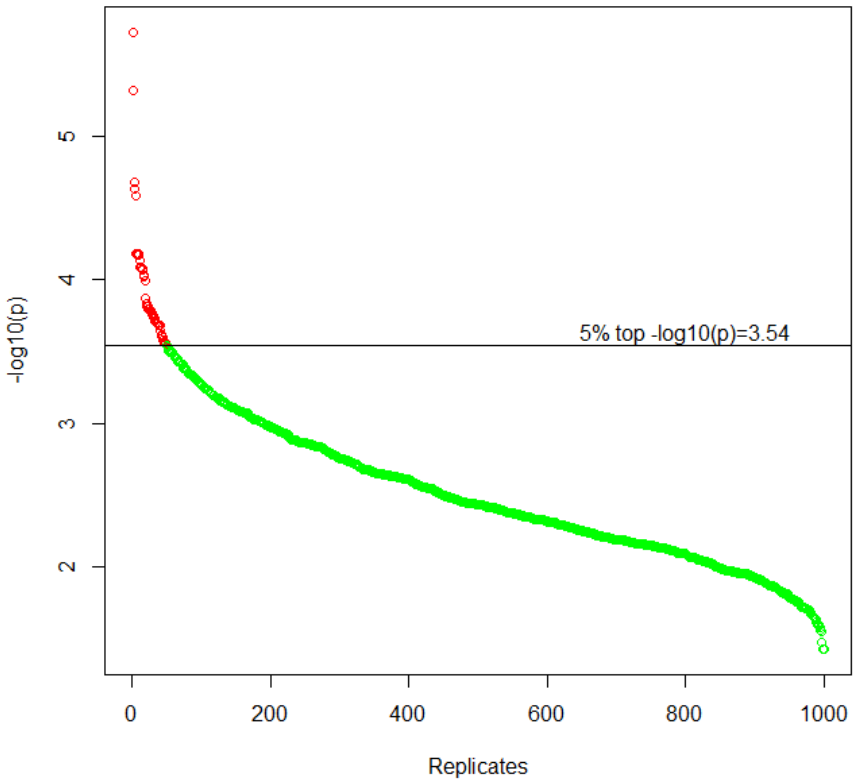
# Supplementary Text:

## Appendix A: Multiple testing correction and power calculations

A) The permutation strategy to derive significance threshold:

We used the permutation approach (1000 replicates) to derive the significance threshold correcting for multiple association tests performed for 389 common and low frequency variants.(minor allele frequency > 0.01) accounting for the linkage disequilibrium. Using the 'sample' command in R we generated 1000 random binary phenotypes. We used plink1.90 software[1] to perform association test between 389 common and low frequency variants and 1000 randomly generated discrete phenotypes. This permutation strategy identified $p=2.89\times10^{-4}$ as the 95% empirical significance threshold correcting for multiple association tests performed for 389 common and low frequency variants which accounts for the linkage disequilibrium. The distribution of absolute log10 transformed minimum p-value for 1000 replicates are shown in following figure:

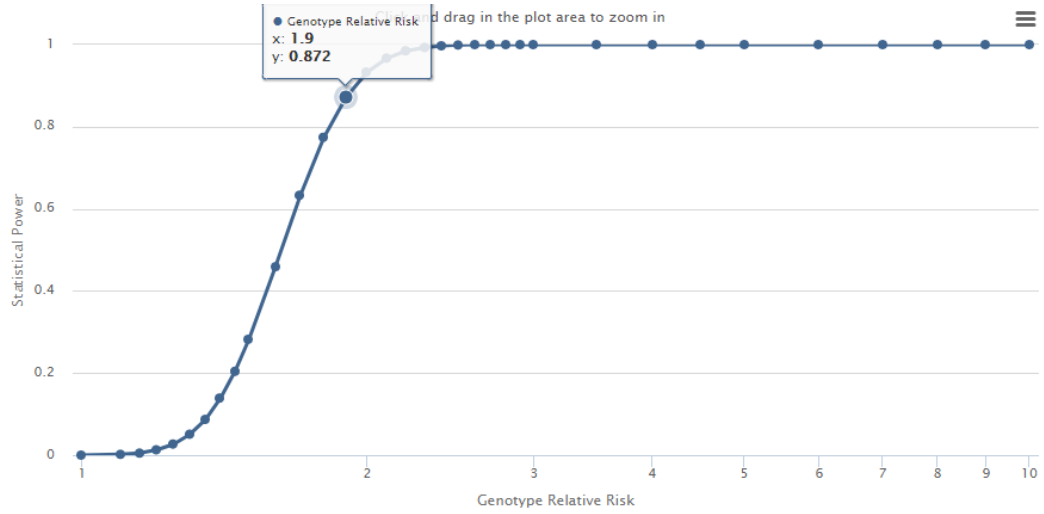


B) Power calculation for single variant test:

Following plots were created using the web interface of the Genetic Association Study (GAS) Power Calculator software[2].

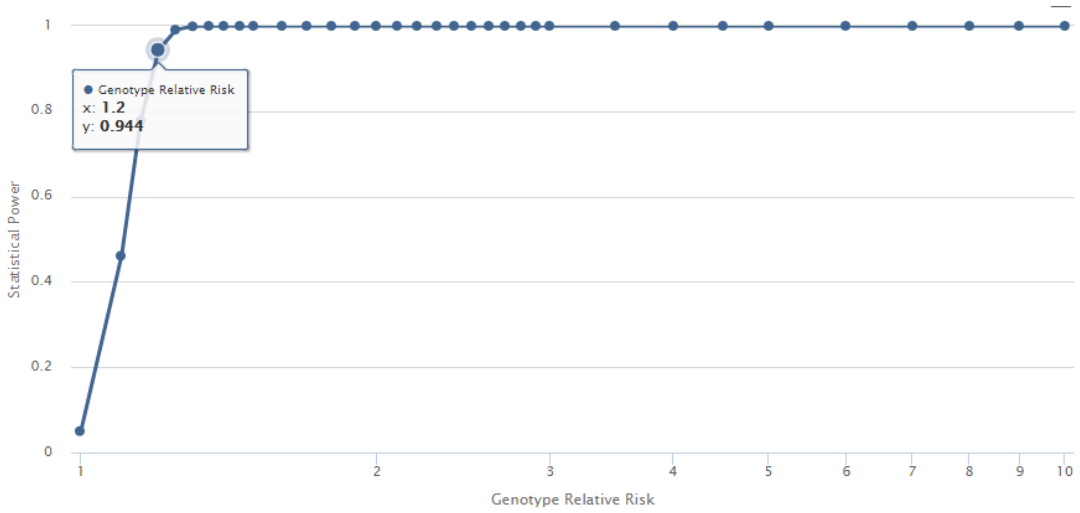
*1. Power verses genotype relative risk plot at permutation-derived significance threshold of $2.89\times10^{-4}$*

GAS Parameters: Significance threshold=$2.89\times10^{-4}$; prevalence of disease=0.17; allele frequency= 0.20; sample size= 250 cases and 250 controls.



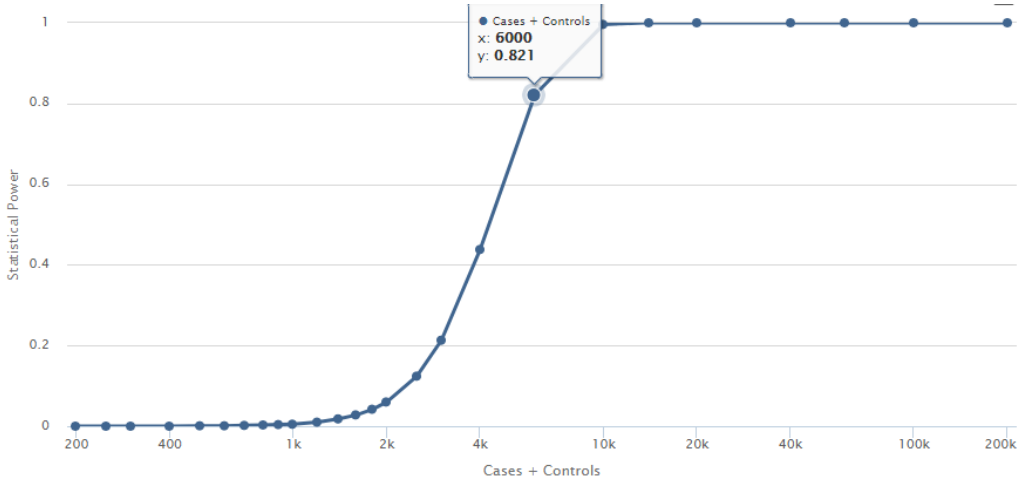

*2. Power verses genotype relative risk plot at replication threshold of 0.05*

GAS Parameters: Significance threshold=0.05; prevalence of disease=0.17; allele frequency= 0.20; sample size= 1650 cases and 1650 controls.



*3. Power verses sample size plot at genome-wide significance threshold*

GAS Parameters: Significance threshold of $5\times10^{-8}$; Prevalence of disease=0.17; Allele frequency= 0.20; Genotype relative risk= 1.25



C. Power calculation for SKAT-O gene-based test:

Power calculations for SKAT-O gene-based tests were performed using the 'Power_Logistic_R' command of the R package skat[3].

Constant parameters: N=512, N=512, Maximum OR=3, Causal MAF Cutoff=0.05, Percent of causal variants with negative effect=10, Case-Control Proportion=0.5, Number of simulations=500, remaining default parameters.

Following tables report power for SKAT-O gene-based tests:

At prevalence=0.01

| Causal variant percentage | Alpha=0.05 | Alpha=0.017 | Alpha=0.01 |
|---|---|---|---|
| 100 | 0.85 | 0.76 | 0.72 |
| 70 | 0.62 | 0.50 | 0.46 |
| 50 | 0.45 | 0.34 | 0.30 |
| 30 | 0.24 | 0.15 | 0.12 |

At prevalence=0.17

| Causal variant percentage | Alpha=0.05 | Alpha=0.017 | Alpha=0.01 |
|---|---|---|---|
| 100 | 0.75 | 0.63 | 0.59 |
| 70 | 0.55 | 0.42 | 0.38 |
| 50 | 0.34 | 0.23 | 0.20 |
| 30 | 0.21 | 0.12 | 0.10 |

## Appendix B: Description of replication studies

### A) Description

1. The Atherosclerosis Risk in Communities Study (ARIC)

The ARIC study is a prospective population-based study of 4 United States communities (Suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina; and Jackson, Mississippi) for studying atherosclerosis and clinical atherosclerotic diseases. During its inception (1987-1989) 15,792 men and women, including 11,478 white participants were recruited. Participants were between ages 45 and 64 years at their baseline examination in 1987 to 1989. Blood was drawn at baseline or at later visits, and DNA was extracted for participants who consented to genetic testing.3 Vascular risk factors and outcomes, including transient ischemic attack, stroke, and dementia, were determined in a standard fashion.4 During the first 2 years (1993–1994) of the third ARIC examination, participants aged 55 and older from the Forsyth County and Jackson sites were invited to undergo cranial MRI. This subgroup of individuals with MRI scanning represents a random sample of the full cohort because examination dates were allocated at baseline through randomly selected induction cycles. Only White participants have been included in the published MRI-marker GWAS meta-analyses. Following table describes the MRI measurements in ARIC study:

| **Variable** | **Description** |
|---|---|
| MRI-scanner (Tesla) | General Electric (General Electric Medical Systems) or Picker (Picker Medical Systems) 1.5-T scanners were used for the MRI examination. |
| MRI sequences used | sagittal T1-weighted scans and axial proton-density, T2-weighted, and T1-weighted scans with 5-mm thickness and no interslice gaps. |
| Definition used for MRI-defined brain infarcts | MR scans were independently evaluated by two trained neuroradiologists for the presence of large (> 3 mm) "infarctlike" lesions. Details of the protocol are given in Bryan et al.[4] For infarcts, inter-reader agreement was 79% |

| | |
|---|---|
| | with a kappa statistic of 0.52, and intra-reader agreement was 82% with a kappa statistic of 0.78. |
| Definition used for lacunes | Lacunes were more than 3 mm and less than 20 mm in diameter and exclusively localized within the subcortical region. |
| Method for quantifying white matter hyperintensity burden | Images were interpreted directly from a PDS-4 digital workstation consisting of four 1024×1024-pixel monitors capable of displaying all 96 images simultaneously. WMHs were estimated as the relative total volume of periventricular and subcortical white matter signal abnormality on proton density-weighted axial images by visual comparison with eight templates that successively increased from barely detectable white matter changes (Grade 1) to extensive, confluent changes (Grade 8). Individuals with no white matter changes received Grade 0, and those with changes worse than Grade 8 received Grade 9. |

2. The Cardiovascular Health Study (CHS)

The CHS is a population-based cohort study of risk factors for vascular disease in adults 65 years or older conducted across 4 field centers in the United States: Sacramento County, California; Washington County, Maryland; Forsyth County, North Carolina; and Pittsburgh, Allegheny County, Pennsylvania. The original predominantly white cohort of 5201 persons was recruited in 1989 to 1990 from a random sample of people on Medicare eligibility lists. An additional 687 blacks were enrolled in 1992 to 1993, for a total sample of 5,888. Vascular risk factors and outcomes, including transient ischemic attack, stroke, and dementia, were determined in a standard fashion. DNA was extracted from blood samples drawn from all participants who consented to genetic testing at their baseline examination in 1989 to 1990 or 1992 to 1993. In 2007 to 2008, genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai on 3980 CHS participants of European

ancestry who were free of cardiovascular disease at baseline and who had DNA available for genotyping.

Following table describes the MRI measurements in the CHS:

| **Variable** | **Description** |
|---|---|
| MRI-scanner (Tesla) | MRI was performed on General Electric or Picker 1.5-T scanners at three field centers and on a 0.35-T Toshiba instrument at the fourth. |
| MRI sequences used | The scanning protocol included a series of axial spin density and T2-weighted scans angled parallel to the anterior-posterior commissure line from vertex to skull base with the following parameters: repetition time, 3000 milliseconds; echo time, 30 and 100 milliseconds; compensated flow; 5 mm thickness; 0 gap; 256×192 matrix; and 1/2 nex (1 nex on a 0.35-T scanner). A series of axial T1-weighted scans was performed, also angled parallel to the anterior-posterior commissure line from vertex to skull base with the following parameters: repetition time, 500 milliseconds; echo time, 20 milliseconds; 5 mm thickness; 0 gap; 256×192 matrix; and 1 nex (2 nex on a 0.35-T scanner). |
| Definition used for MRI-defined brain infarcts | Infarcts were defined as lesions with abnormal signal in a vascular distribution and no mass effect. Infarcts of the cortical gray matter and deep nuclear regions and capsule were defined as lesions bright on spin-density and T2-weighted images compared with normal gray matter and isodense or hypodense on T1-weighted images. Infarcts in the white matter were also bright on spin-density and T2-weighted images but in addition were hypointense on T1-weighted images, approximating the intensity of |

| | |
|---|---|
| | cerebrospinal fluid. |
| Definition used for lacunes | Lacunes were less than 20 mm in all dimensions and exclusively subcortical. The requirement for hyperintensity on spin-density images was intended to distinguish small deep nuclear region infarcts (those in the caudate nucleus, lentiform nucleus, internal capsule, external capsule, extreme capsule, and thalamus) from dilated perivascular spaces. |
| Method for quantifying white matter hyperintensity burden | Because of concerns that specific findings such as rims or halos would be difficult to identify and quantify reliably, the decision was made to consider the total volume of white matter change rather than trying to grade specific findings or to grade separately changes in the periventricular and subcortical regions. Accordingly, neuroradiologists at the reading center estimated the total volume of periventricular and subcortical white matter signal abnormalities on spin density-weighted axial images by comparing the findings on any particular scan with sets of complete scans that demonstrated successively increasing changes from barely detectable (grade 1) to extensive and confluent (grade 8). A text description of the white matter grades was supplied to the neuroradiologists but was not used as much as matching visual patterns to the template images. Studies with no white matter findings were graded 0, and those with findings more remarkable than grade 8 were scored 9. As per the current analysis plan, in CHS with 10-point scale (grades 0-9), high burden of WMH will be grades strictly above age-specific median (by 5-year age-categories). |

3. Framingham Heart Study (FHS) and Gen3 of FHS

The FHS is a 3-generation, single-site, community-based, prospective cohort study that was initiated in 1948 to investigate risk factors for cardiovascular disease including stroke. It now comprises 3 generations of participants: the original cohort followed-up since 1948 (original) 12 their offspring and spouses of the offspring, followed-up since 1971 (offspring),13 and children from the largest offspring families enrolled in 2000 (Gen 3) 14. The original cohort enrolled 5209 men and women who comprised two-thirds of the adult population then residing in Framingham, Massachusetts. Survivors continue to receive biennial examinations. The offspring cohort comprises 5124 persons (including 3514 biological offspring) who have been examined approximately once every 4 years. Participants in the first 2 generations were invited to undergo an initial brain MRI in 1999 to 2005. Brain MRI in Gen 3 only began in 2009 and is not included in these analyses. The population of Framingham was virtually entirely white in 1948, when the original cohort was recruited. Vascular risk factors and outcomes, including transient ischemic attack, stroke, and dementia, were identified prospectively since 1948 through an ongoing system of FHS clinic and local hospital surveillance 15, 16 Participants had DNA extracted and provided consent for genotyping in the 1990s. Genotyping was performed at Affymetrix (Santa Clara, Calif) through an NHLBI-funded SNP-Health Association Resource (SHARe) project.

Following table describes the MRI measurements in the FHS:

| **Variable** | **Description** |
|---|---|
| MRI-scanner (Tesla) | 1.0 or 1.5 Tesla (Siemens Avanto Scanner) |
| MRI sequences used | 3-dimensional T1-weighted coronal spoiled gradient-recalled echo (SPGR), T2-weighted double spin-echo coronal images acquired in 4-mm contiguous slices, and fluid attenuated inversion recovery (FLAIR) sequences. |
| Definition used for MRI-defined brain infarcts | The presence of MRI infarction was determined from the size, location and imaging characteristics of the lesion.[5] The image analysis system allowed for superimposition of the subtraction image, the proton density image and the T2 weighted image at three times magnified view to assist in |

| | |
|---|---|
| | interpretation of lesion characteristics. Signal void, best seen the T2 weighted image was interpreted to indicate a vessel. Lesions 3mm or larger qualified for consideration as cerebral infarcts. |
| Definition used for lacunes | Lesions between 3-10 mm in diameter located in subcortical areas qualified for consideration as small cerebral infarcts. Other necessary imaging characteristics included (1) CSF density on the subtraction image and (2) if the stroke was in the basal ganglia area, distinct separation from the circle of Willis vessels. |
| Method for quantifying white matter hyperintensity burden | Segmentation and quantification of WMH was performed using a semi-automated procedure based on FLAIR sequences using a previously described algorithm.[6,7] After affine co-registration of the FLAIR image to the high-resolution T1 image, WMH voxels were used to correct intensity changes in the T1 image to reduce any adverse impact of the WMH voxel values on the accuracy of the nonlinear warping algorithm. A segmentation threshold for WMH was determined as 3.5 standard deviations in pixel intensity above the mean of the fitted distribution of brain parenchyma. These methods have been shown to have high inter- and intra- rater reliabilities.<br><br>WMH were defined as extreme if the log-transformed WMH volume was one SD unit above the age-adjusted mean volume for the total sample. |

4. Rotterdam Study (RS)

The Rotterdam Study is a prospective, population-based cohort study among inhabitants of a well-defined district of Rotterdam (Ommoord), The Netherlands.[8] This study aims to examine the determinants of disease and health in the elderly with a focus on neurogeriatric, cardiovascular, bone, and eye diseases.[8] The cohort was initially defined in 1990 among 7983 persons, aged 55 years and older, who underwent a home interview and extensive physical examination at the baseline and during follow-up rounds every 3-4 years (Rotterdam Study I). The cohort was extended in 2000/2001 with 3011 persons aged 55 years and older (Rotterdam Study II) and 2006/2008 with 3,932 persons aged 45 and older (Rotterdam Study III). All participants had DNA extracted at their first visit. Genotyping was performed at the Human Genotyping Facility, Genetic Laboratory Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. Initially, in 1995 to 1996, random subsamples of Rotterdam Study participants underwent neuroimaging, whereas from 2005 onwards MRI has been implemented as the core protocol of the Rotterdam Study.[9]

Following table describes the MRI measurements in the Rotterdam studies:

| **Variable** | **Description** |
|---|---|
| MRI-scanner (Tesla) | 1.5T MRI unit (General Electric Healthcare, Milwaukee, USA, software version 119)[9] |
| MRI sequences used | T1, T2, FLAIR |
| Definition used for MRI-defined brain infarcts | Trained readers rated infarcts using T1, T2, and FLAIR sequences. Lesions between 3-15mm were considered lacunar infarcts, >15mm as subcortical infarcts, and as cortical infarcts if the cortical grey matter was affected.[10] |
| Definition used for lacunes | Lacunes were distinguished from VR-spaces based on their irregular shape, presence of a hyperintense rim and non-vascular appearance.VR-spaces were rated on T2, T1 and FLAIR by the same investigator (H.-A.). [11] |

| | |
|---|---|
| Method for quantifying white matter hyperintensity burden | In the brain tissue segmentation, possible WMLs are misclassified as GM with a ring of WM voxels. In the FLAIR image the WMLs are hyperintense. We therefore process the histogram from the FLAIR image intensities of all voxels that are classified as GM, to estimate the mean and standard deviation of true GM voxels. Subsequently, WML voxels are extracted by intensity thresholding, where the threshold depends on the estimated GM distribution. False positives are removed by excluding voxels which are not sufficiently connected to the white matter. The different parameters (intensity threshold, and quantitative definition of not being sufficiently connected) have been optimized on large reference dataset.[12] |

## B) Quality control and imputation of genome-wide genotype data

Following tables details genotyping and imputation protocols of individual studies:

| Parameter | ARIC-EA | CHS-EA | FHS | Rotterdam studies I, II and III |
|---|---|---|---|---|
| Ancestry | European | European | European | European |
| GWAS genotyping chip/platform | Affymetrix SNP Array 6.0 | Illumina Human 370CNV Duo BeadChip® + ITMAT-Broad-CARe (IBC) Illumina iSelect chip | Affymetrix 500K (250K Nsp & 250K Sty), MIPS 50K | The Illumina 550K (RS-I, II; single + duo array format) and 610K (RS-III; quattro array format) |
| Genotype calling algorithm | Birdseed | Illumina Bead Studio | Affymetrix BRLMM | Illumina GeneCall |
| Pre-imputation sample filters/Sample QC (exclusion criteria) | Call rate <95%, First degree relatives, Ancestry outliers, Duplicates, Sex discrepancies | Call rate ≤95%, Discordance between genotyped and recorded sex, Limited to participants of European Ancestry | Call rate <97%, Heterozygosity >5 SD away from the mean, Large Mendelian error rate | Call rate <95%, Discordance between genotyped and recorded sex, Excess inter/intra heterozygosity, Non-European Ancestry |
| Pre-imputation SNP filters/SNP QC (exclusion | Call rate <95%, Chromosome of | Call rate <97%, Heterozygote frequency = 0, | Call rate <98%, MAF <1%, HWE P | Call rate <97.5%, MAF <1%, HWE P $<1\times10^{-6}$ |

| | | | | |
|---|---|---|---|---|
| criteria) | zero, Monomorphic, HWE $p<1\times10^{-6}$ | HWE P $<1\times10^{-5}$, >2 duplicate errors or Mendelian inconsistencies (among reference trios) | $<1\times10^{-6}$ | |
| Imputation panel | HRC v1.1 | HRC v1.1 | HRC v1.1 | HRC v1.1 |
| Imputation software | Michigan server | Michigan server | Michigan server | Michigan server |
| Covariates used for analysis | Age, sex, study center, PC1-2 | Age, sex, study center, PC1-5 | Age, sex, PC1-8 and family structure | Age and sex |
| Association analysis software | R | R | R and Perl | R |

| Parameter | ARIC-AA | CHS-AA |
|---|---|---|
| Ancestry | African | African |
| GWAS genotyping chip/platform | Affymetrix SNP Array 6.0 | Illumina HumanOmni1-Quad_v1 BeadChip |
| Genotype calling algorithm | Birdseed | Illumina GenomeStudio |

| | | |
|---|---|---|
| Pre-imputation sample filters/Sample QC (exclusion criteria) | Call rate <95%<br>First degree relatives, Ancestry outliers | Call rate ≤95%, Discordance between genotyped and recorded sex |
| Pre-imputation SNP filters/SNP QC (exclusion criteria) | Call rate <95%, MAF<1%, HWE $p<1\times10^{-5}$ | Call rate <97%, Heterozygote frequency = 0, HWE $P<1\times10^{-5}$, >1 duplicate errors or Mendelian inconsistency (among reference trios) |
| Imputation panel | 1KG p1 v3 | HRC v1.1 |
| Imputation software | SHAPEIT and IMPUTE2 | Michigan Server |
| Covariates used for analysis | Age, sex, study center, PC1-4 | Age, sex, study center, and PC1-5 |
| Association analysis software | R | R |

## C) Quality control of whole exome sequencing data

1) Joint calling of ARIC-EA, ARIC-AA, CHS-EA and FHS

***Exome Sequencing and Variant Calling***

For CHARGE Freeze 5, DNA samples were constructed into Illumina paired-end pre-capture libraries according to the manufacturer's protocol. The complete protocol and oligonucleotide sequences are accessible from the Baylor College of Medicine Human Genome Sequencing Center (HGSC) website (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). Two, four or six pre-capture libraries were pooled together and then hybridized to the HGSC VCRome 2.1 design[13] (42Mb, NimbleGen) and sequenced in paired-end mode in a single lane on the Illumina HiSeq 2000 or the HiSeq 2500 platform. Illumina sequence analysis was performed using the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/content/mercury). Pooled samples were de-multiplexed using the Consensus assessment of sequence and variation (CASAVA) software. Reads were mapped to the Genome Reference Consortium Human Build 37 (GRCh37) human reference sequence (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) using Burrows-Wheeler Alignment (BWA[14], http://bio-bwa.sourceforge.net/) producing Binary Alignment/Map (BAM[15]) files. Aligned reads were then recalibrated using Genome Analysis ToolKit (GATK[16], http://www.broadinstitute.org/gatk/) along with BAM sorting, duplicate read marking, and realignment near insertions or deletions (indels). The Atlas2[17] suite was used to call single nucleotide variants (SNVs) and insertion-deletions (indels) and produce high-quality variant call files (VCF[18]).

***Quality Control***

Each SNV call was filtered based on the following criteria to produce a high-quality variant list: low SNV posterior probability (<0.95), low variant read count (<3), variant read ratio <0.25 or >0.75, strand-bias of more than 99% variant reads in a single strand direction, or total coverage less than 10-fold. All variant calls filtered by these criteria, and reference calls with less than 10-fold coverage, were set to missing. The variant call filters were the same for indels except a total coverage less than 30-fold was used for variant sites.

Variant-level quality control steps excluded variants outside the exon capture regions (VCRrome 2.1), monomorphic sites, missing rate >20%, mappability score <0.8, and mean depth of coverage >500-fold. Variants not meeting Hardy-Weinberg equilibrium expectations ($P<5x10^{-6}$)

in ancestry-specific groups were also excluded. Sample-level quality control metrics were calculated by cohort and ancestry group. A sample was excluded for missingness >20%, or if compared to the other samples it fell less than 6 standard deviations (SD) for mean depth, more than 6 SD for singleton count, or outside of 6 SD for heterozygote to homozygote ratio or Ti/Tv ratio.

The final sample for CHARGE contained 11263 EA individuals (1751 for CHS, 7810 for ARIC, and 1702 for FHS) and 3180 AA from ARIC. In total, there were 2,556,859 SNVs and 76,133 indels after quality control. The mean depth of coverage was 78X.

#### *Annotation of Whole Exome Sequence*

To facilitate meta-analysis between CHARGE and other exome sequencing projects (e.g., the NHLBI Exome Sequencing Project) we created a combined variant annotation file including all quality-controlled variant sites observed in either study. Variants were annotated using ANNOVAR[8] and dbNSFP v2.0 (https://sites.google.com/site/jpopgen/dbNSFP) according to the reference genome GRCh37 and National Center for Biotechnology Information RefSeq. Coding variants were annotated to a unique gene and functional category. A file was created that merged the annotated variant lists between CHARGE and the other studies to ensure that a variant that was present in both studies had the same reference allele and functional annotation. This multiple study-combined SNPinfo file was used as a component of the seqMeta R package (http://cran.r-project.org/web/packages/seqMeta/index.html).

2) Rotterdam study 1 whole exome sequencing

Exomes of randomly selected individuals from the RS-I were sequenced at an average depth of 54X using the Nimblegen SeqCap EZ V2 capture kit on an Illumina Hiseq2000 sequencer using the TrueSeq Version 3 protocol.[19,20] Sequencing was performed at the Human Genotyping facility of the Department of Internal Medicine, Erasmus MC, The Netherlands. Sequence reads were aligned to human genome build 19 using Burrows–Wheeler Aligner[14] and subsequently processed further using Picard's MarkDuplicates, SAMtools[15] and the Indel Realignment and Base Quality Score Recalibration tools from Genome Analysis Toolkit.[21] Genetic variants were called using the HaplotypeCaller from Genome Analysis Toolkit.[19]. Sample-level quality control steps excluded samples with low concordance to genotyping array (< 95%), or that differed 4 s.d.

from the mean on either the number of detected variants per sample, transition to transversion ratio or high heterozygote to homozygote ratio and low call rate (< 90%).[19]. Variant-level quality control steps excluded variants with a low call rate (< 90%) and out of Hardy–Weinberg equilibrium (P-value $<10^{-8}$).[19] The final data set consisted of 600,806 SNVs in 2,356 individuals.

## References


1. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
2. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209-13 (2006).
3. Lee, S.S. & Lee, M.S.S. Package 'SKAT'. (2017).
4. Bryan, R.N. *et al.* A method for using MR to evaluate the effects of cardiovascular disease on the brain: the cardiovascular health study. *AJNR Am J Neuroradiol* **15**, 1625-33 (1994).
5. DeCarli, C. *et al.* Predictors of brain morphology for the men of the NHLBI twin study. *Stroke* **30**, 529-36 (1999).
6. DeCarli, C. *et al.* Method for quantification of brain, ventricular, and subarachnoid CSF volumes from MR images. *J Comput Assist Tomogr* **16**, 274-84 (1992).
7. DeCarli, C. *et al.* The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. *Neurology* **45**, 2077-84 (1995).
8. Ikram, M.A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* **32**, 807-850 (2017).
9. Ikram, M.A. *et al.* The Rotterdam Scan Study: design update 2016 and main findings. *Eur J Epidemiol* **30**, 1299-315 (2015).
10. Vermeer, S.E. *et al.* Silent brain infarcts and the risk of dementia and cognitive decline. *N Engl J Med* **348**, 1215-22 (2003).
11. Adams, H.H. *et al.* Rating method for dilated Virchow-Robin spaces on magnetic resonance imaging. *Stroke* **44**, 1732-5 (2013).
12. de Boer, R. *et al.* White matter lesion extension to automatic brain tissue segmentation on MRI. *Neuroimage* **45**, 1151-61 (2009).
13. Bainbridge, M.N. *et al.* Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol* **12**, R68 (2011).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
15. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
16. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).

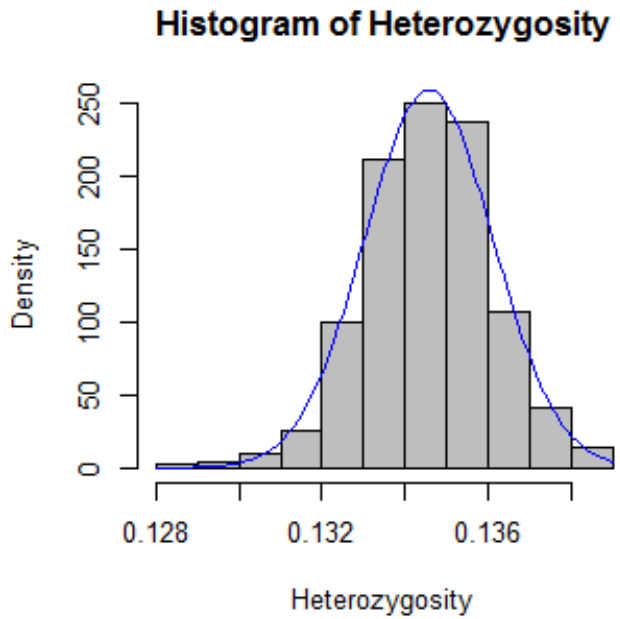
17. Challis, D. *et al.* An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* **13**, 8 (2012).

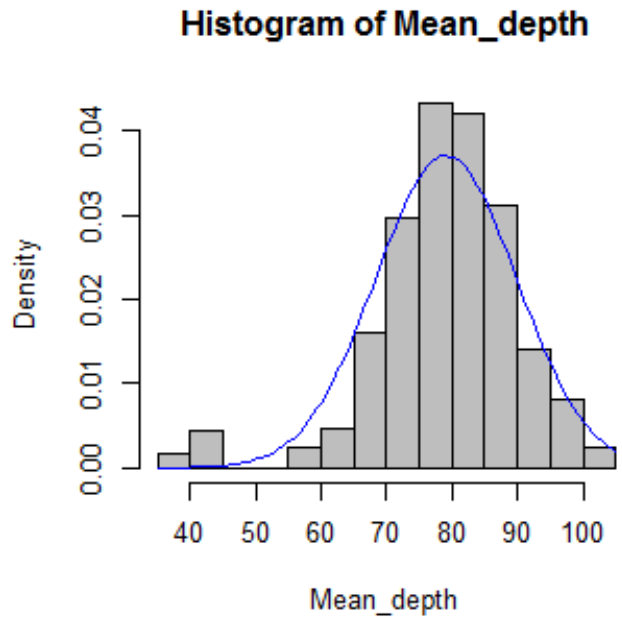
18. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).

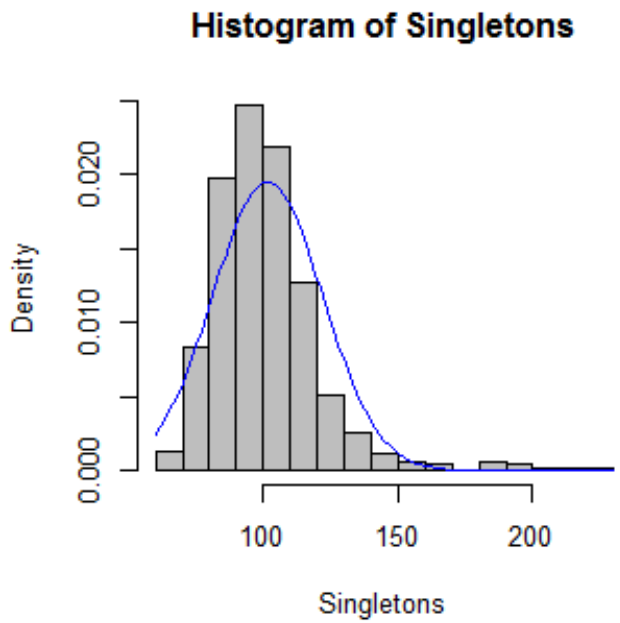
19. Amin, N. *et al.* Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol Psychiatry* **22**, 537-543 (2017).

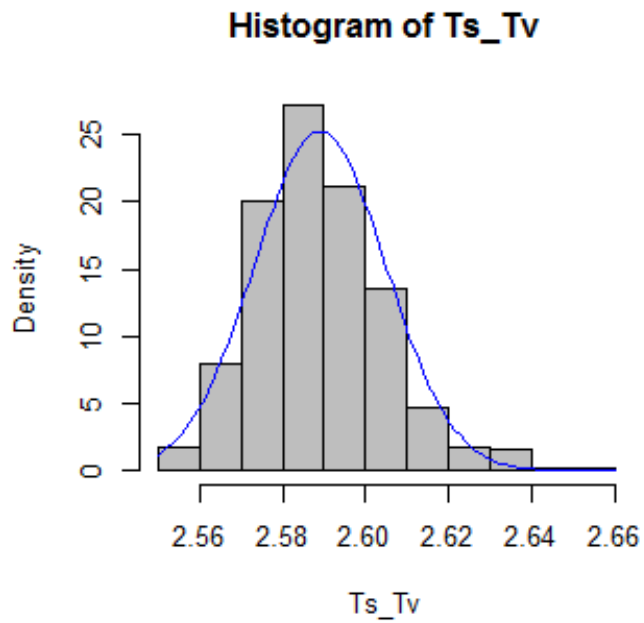
20. van Rooij, J.G.J. *et al.* Population-specific genetic variation in large sequencing data sets: why more data is still better. *Eur J Hum Genet* **25**, 1173-1175 (2017).

21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).



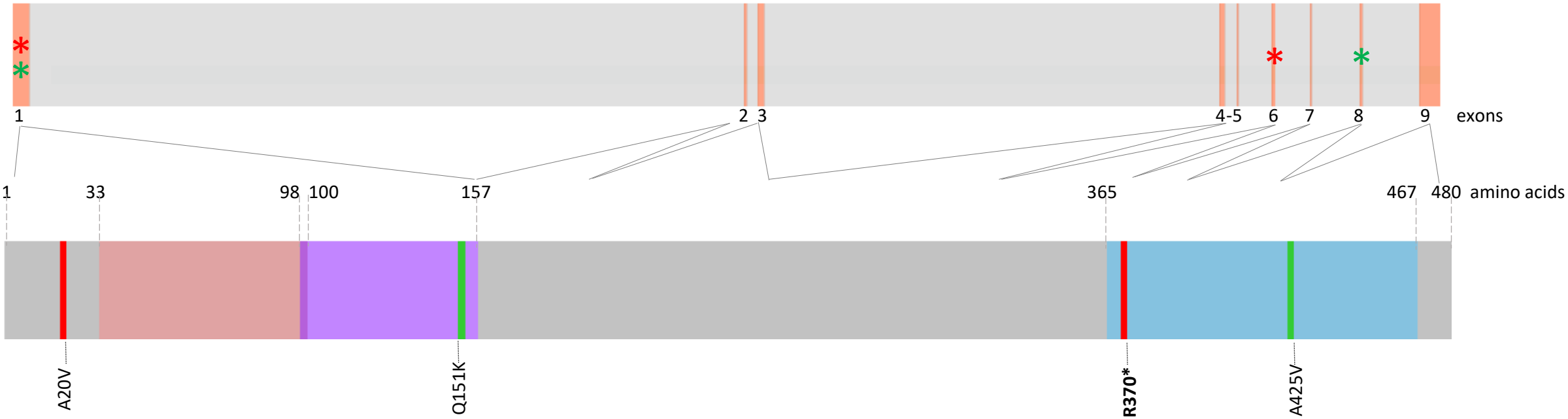




Supplementary figure 1: Histograms for quality control of whole exome sequencing of 3C-Dijon cohort

# Supplementary figure 2: HTRA1 protein-modifying rare and low frequency variants observed in the 3C-Dijon extreme-CSVD cohort




Gene composition

- Exons
- Introns
- Minor allele observed only in extensive SVD participants
- Minor allele observed only in minimal SVD participants

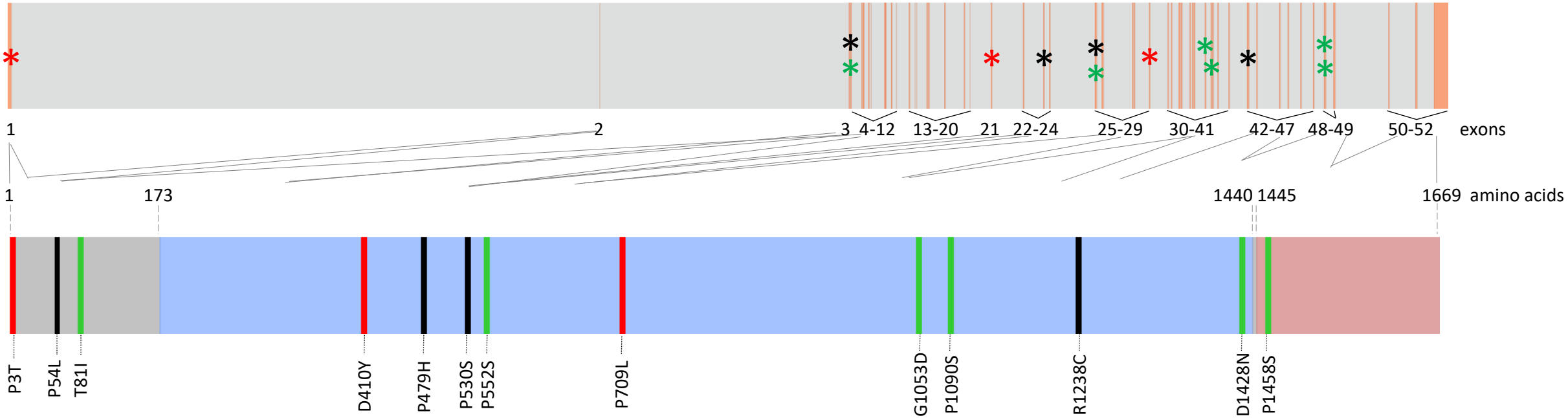
Protein composition

- IGFBP N-terminal
- Kazal Like
- No specific region
- PDZ domain

- Protein-modifying variants only in extensive SVD participants
- Protein-modifying variants only in minimal SVD participants

In bold: previously reported CARASIL causing variant

# Supplementary figure 3: COL4A1 protein-modifying rare and low frequency variants observed in the 3C-Dijon extreme-CSVD cohort




Gene composition

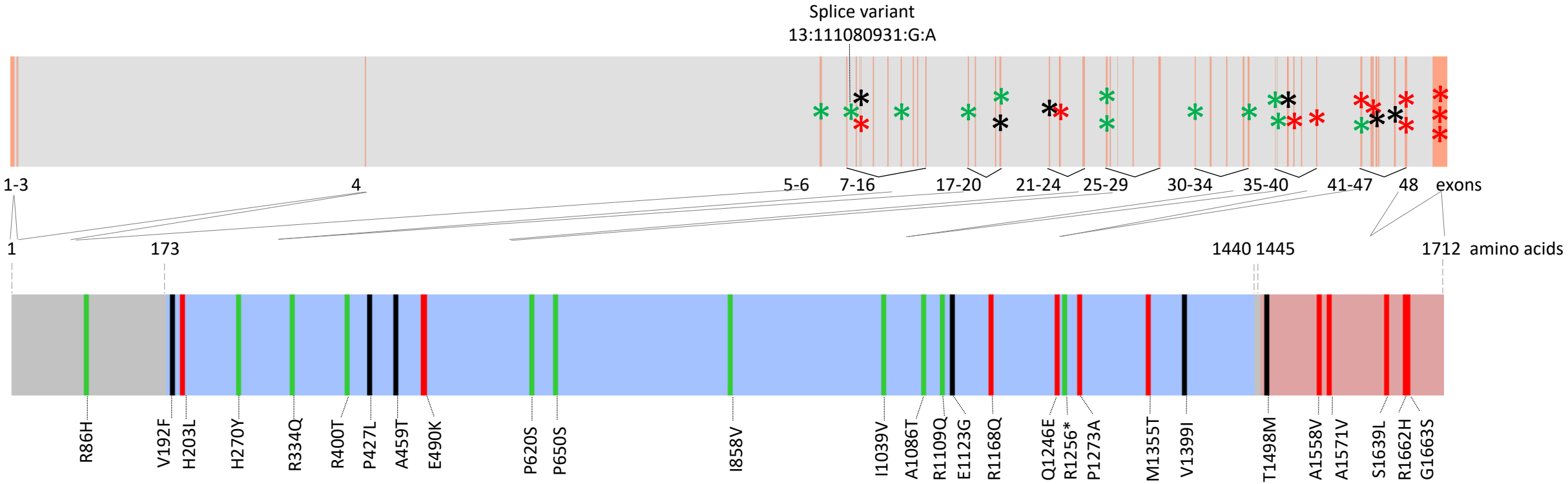
- Exons
- Introns
- Minor allele observed only in cases (extensive SVD)
- Minor allele observed only in controls (minimal SVD)
- Minor allele observed in both extensive and minimal SVD groups

Protein composition

- No specific region
- Triple-helical region
- Collagen IV NC1

- Protein-modifying variants only in extensive SVD participants
- Protein-modifying variants only in minimal SVD participants

# Supplementary figure 4: COL4A2 protein-modifying rare and low frequency variants observed in the 3C-Dijon extreme-CSVD cohort




Gene composition

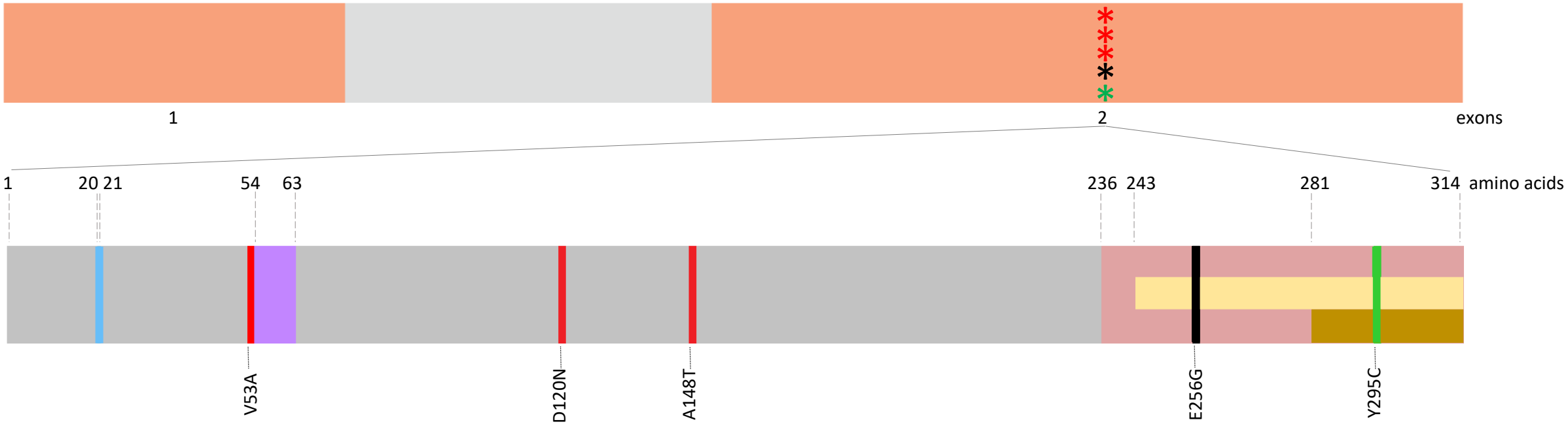
- Exons
- Introns
- Minor allele observed only in cases (extensive SVD)
- Minor allele observed only in controls (minimal SVD)
- Minor allele observed in both extensive and minimal SVD groups

Protein composition

- No specific region
- Triple-helical region
- Collagen IV NC1

- Protein-modifying variants only in extensive SVD participants
- Protein-modifying variants only in minimal SVD participants

# Supplementary figure 5: TREX1 protein-modifying rare and low frequency variants observed in the 3C-Dijon extreme-CSVD cohort




Gene composition

- Exons
- Introns
- Minor allele observed only in cases (extensive SVD)
- Minor allele observed only in controls (minimal SVD)
- Minor allele observed in both extensive and minimal SVD groups

Protein composition

- No specific region
- Substrate binding
- Proline-rich region
- Necessary for endoplasmic reticulum localization
- Interaction with UBQLN1
- Necessary for reticulum retention

- Protein-modifying variants only in extensive SVD participants
- Protein-modifying variants only in minimal SVD participants

# Supplementary Tables:

## Supplementary Table 1: Population characteristics of genome-wide genotype extreme bSVD cohorts of ARIC, CHS, FHS and Rotterdam studies

| | ARIC-EA | | | CHS-EA | | | FHS | | |
|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** |
| **N** | 131 | 131 | NA | 335 | 335 | NA | 539 | 539 | NA |
| **WMH (grades[0-9] or $mm^3$) mean (s.d)** | 2.95 (1.05) | 0.443 (0.50) | $1.41\times10^{-60}$ | 4.31 (1.36) | 0.78 (0.45) | $4.57\times10^{-160}$ | 7.68 (11.22) | 0.81 (0.59) | $4.17\times10^{-39}$ |
| **Lacunes N (%)** | 26 (0.20) | 0 | NA | 179 (0.53) | 0 | NA | 69 (0.13) | 0 | NA |
| **Age mean (s.d)** | 62.62 (4.62) | 64.56 (5.15) | $1.51\times10^{-3}$ | 75.21 (4.93) | 76.31 (4.57) | $2.91\times10^{-3}$ | 57.33 (13.77) | 57.28 (12.96) | 0.95 |
| **Females N (%)** | 81 (0.62) | 73 (0.56) | 0.38 | 207 (0.62) | 209 (0.62) | 0.94 | 286 (0.53) | 286 (0.53) | 1 |
| **Hypertension status N (%)** | 48 (0.37) | 41 (0.31) | 0.43 | 224 (0.67) | 163 (0.49) | $2.52\times10^{-6}$ | 244 (0.45) | 220 (0.41) | 0.16 |
| **Systolic blood pressure mean (s.d)** | 124.16 (22.05) | 124.35 (18.35) | 0.94 | 138.16 (21.74) | 131.16 (20.13) | $1.83\times10^{-5}$ | 122.84 (17.27) | 120.09 (16.52) | $7.68\times10^{-3}$ |
| **Anti-hypertensive medication status N (%)** | 37 (0.28) | 26 (0.20) | 0.15 | 164 (0.49) | 113 (0.34) | $8.48\times10^{-5}$ | 181 (0.34) | 173 (0.32) | 0.65 |
| **Fasting Glucose mean (s.d)** | NA | NA | NA | 105.63 (32.21) | 100.59 (18.70) | 0.01 | 101.63 (23.94) | 100.23 (18.71) | 0.29 |
| **Diabetes status N (%)** | 13 (0.10) | 12 (0.09) | 1 | 44 (0.13) | 32 (0.10) | 0.18 | 58 (0.11) | 35 (0.07) | 0.02 |
| **HDL mean (s.d)** | 56.25 (19.74) | 52.55 (19.24) | 0.13 | 53.71 (14.17) | 54.43 (14.45) | 0.52 | 59.21 (18.25) | 58.74 (17.73) | 0.67 |
| **LDL mean (s.d)** | 126.95 (31.84) | 125.91 (30.48) | 0.79 | 127.06 (33.29) | 125.09 (32.32) | 0.44 | 105.07 (29.94) | 104.24 (29.78) | 0.65 |
| **Triglyceride level mean (s.d)** | 1.80 (1.54) | 1.61 (0.87) | 0.23 | 150.05 (102.27) | 139.23 (79.1) | 0.13 | 115.93 (95.34) | 111.19 (70.34) | 0.35 |
| **Lipid lowering medication N (%)** | 32 (0.24) | 33 (0.25) | 1 | 25 (0.08) | 16 (0.05) | 0.20 | 151 (0.28) | 161 (0.30) | 0.55 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Body mass index mean (s.d)** | 26.32 (4.69) | 25.99 (4.16) | 0.56 | 26.41 (3.92) | 25.95 (4.46) | 0.16 | 27.82 (5.12) | 28.08 (5.57) | 0.44 |
| **Cardiovascular disease status_N (%)** | NA | NA | NA | 16 (0.05) | 21 (0.06) | 0.50 | 48 (0.09) | 41 (0.08) | 0.51 |

| | **Rotterdam study I** | | | **Rotterdam study II** | | | **Rotterdam study III** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Extensive-bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** |
| **N** | 148 | 148 | NA | 109 | 109 | NA | 392 | 392 | NA |
| **WMH (grades[0-9] or $mm^3$) mean (s.d)** | 37.92 (21.30) | 2.55 (1.10) | $2.71\times10^{-44}$ | 18.97 (17.29) | 1.71 (1.06) | $5.09\times10^{-18}$ | 9.13 (8.88) | 1.07 (0.54) | $5.60\times10^{-53}$ |
| **Lacunes N (%)** | 58 (0.39) | 0 | NA | 18 (0.17) | 0 | NA | 33 (0.08) | 0 | NA |
| **Age mean (s.d)** | 79.16 (4.39) | 79.42 (4.96) | 0.64 | 68.51 (6.05) | 68.61 (6.15) | 0.91 | 58.33 (7.36) | 58.85 (5.89) | 0.27 |
| **Females N (%)** | 76 (0.51) | 78 (0.53) | 0.91 | 54 (0.50) | 54 (0.50) | 1 | 213 (0.54) | 231 (0.59) | 0.22 |
| **Hypertension status N (%)** | 136 (0.92) | 124 (0.84) | 0.05 | 87 (0.80) | 65 (0.60) | $1.85\times10^{-3}$ | 236 (0.60) | 156 (0.40) | $1.50\times10^{-8}$ |
| **Systolic blood pressure mean (s.d)** | 157.20 (23.08) | 150.85 (21.46) | 0.02 | 149.02 (19.44) | 142.74 (17.23) | 0.01 | 137.13 (20.46) | 130.38 (18.47) | $1.57\times10^{-6}$ |
| **Anti-hypertensive medication status N (%)** | 94 (0.64) | 77 (0.52) | 0.06 | 38 (0.35) | 31 (0.28) | 0.38 | 122 (0.31) | 72 (0.18) | $4.67\times10^{-5}$ |
| **Fasting Glucose mean (s.d)** | 5.91 (1.16) | 5.82 (1.22) | 0.55 | 5.77 (1.20) | 5.67 (1.01) | 0.53 | NA | NA | NA |
| **Diabetes status N (%)** | 12 (0.08) | 11 (0.07) | 1 | 17 (0.16) | 6 (0.06) | 0.03 | 37 (0.09) | 36 (0.09) | 1 |
| **HDL mean (s.d)** | 1.42 (0.38) | 1.45 (0.38) | 0.53 | 1.40 (0.36) | 1.45 (0.38) | 0.37 | 1.41 (0.42) | 1.47 (0.43) | 0.07 |
| **LDL mean (s.d)** | 3.17 (1.01) | 3.288 (0.94) | 0.32 | NA | NA | NA | NA | NA | NA |
| **Triglyceride level mean (s.d)** | 1.42 (0.67) | 1.27 (0.50) | 0.03 | NA | NA | NA | NA | NA | NA |
| **Lipid lowering medication N (%)** | 54 (0.37) | 34 (0.23) | 0.02 | 23 (0.21) | 26 (0.24) | 0.63 | 78 (0.20) | 81 (0.21) | 0.86 |
| **Body mass index mean (s.d)** | 27.30 (4.20) | 27.56 (3.68) | 0.57 | 27.85 (3.70) | 28.08 (4.32) | 0.67 | 27.71 (4.24) | 27.09 (4.40) | 0.05 |
| **Cardiovascular disease status_N (%)** | 22 (0.15) | 12 (0.09) | 0.10 | 16 (0.15) | 5 (0.05) | 0.02 | 10 (0.03) | 5 (0.01) | 0.30 |

| | ARIC-AA | | | CHS-AA | | |
|---|---|---|---|---|---|---|
| **Variables** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** |
| **N** | 122 | 122 | NA | 73 | 73 | NA |
| **WMH (grades[0-9] or $mm^3$) mean (s.d)** | 3.48 (1.37) | 0 (0) | $6.57\times10^{-55}$ | 4.51 (1.67) | 0.70 (0.49) | $8.81\times10^{-32}$ |
| **Lacunar Brain Infract N (%)** | 55 (0.45) | 0 | NA | 43 (0.59) | 0 | NA |
| **Age mean (s.d)** | 62.21 (4.82) | 60.84 (3.81) | 0.02 | 73.56 (5.80) | 74.51 (4.85) | 0.29 |
| **Females N (%)** | 74 (0.61) | 78 (0.64) | 0.69 | 44 (0.60) | 44 (0.60) | 1 |
| **Hypertension status N (%)** | 101 (0.83) | 69 (0.57) | $1.27\times10^{-5}$ | 65 (0.89) | 51 (0.70) | $7.07\times10^{-3}$ |
| **Systolic blood pressure mean (s.d)** | 143.26 (28.05) | 127.94 (16.66) | $5.29\times10^{-7}$ | 146.22 (22.50) | 138.10 (20.78) | 0.03 |
| **Anti-hypertensive medication status N (%)** | 84 (0.69) | 54 (0.44) | $1.68\times10^{-4}$ | 49 (0.67) | 41 (0.56) | 0.23 |
| **Fasting Glucose mean (s.d)** | NA | NA | NA | 105.19 (21.47) | 113.48 (45.56) | 0.17 |
| **Diabetes status N (%)** | 39 (0.32) | 19 (0.16) | $2.60\times10^{-3}$ | 12 (0.16) | 12 (0.16) | 1 |
| **HDL mean (s.d)** | 55.03 (19.13) | 58.98 (18.56) | 0.10 | 57.25 (15.20) | 56.17 (13.44) | 0.65 |
| **LDL mean (s.d)** | 124.87 (42.35) | 133.16 (33.70) | 0.09 | 122.36 (35.22) | 130.70 (36.20) | 0.17 |
| **Triglyceride level mean (s.d)** | 1.33 (0.76) | 1.26 (0.64) | 0.42 | 117.18 (67.94) | 116.72 (53.06) | 0.96 |
| **Lipid lowering medication N (%)** | 71 (0.58) | 40 (0.33) | $1.00\times10^{-4}$ | 1 (0.01) | 7 (0.10) | 0.06 |
| **Body mass index mean (s.d)** | 29.17 (5.55) | 30.11 (5.80) | 0.20 | 28.05 (5.58) | 28.45 (4.63) | 0.64 |
| **Cardiovascular disease status N (%)** | NA | NA | NA | 12 (0.16) | 20 (0.27) | 0.16 |

## Supplementary Table 2: Population characteristics of Whole exome sequencing extreme bSVD cohorts of ARIC, CHS, FHS and Rotterdam studies

| | ARIC | | | CHS | | | FHS | | |
|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** |
| **N** | 108 | 108 | NA | 194 | 194 | NA | 116 | 113 | NA |
| **WMH (grades[0-9] or $mm^3$) mean (s.d)** | 2.75 (1.00) | 0.41 (0.49) | $1.67\times10^{-49}$ | 4.19 (1.25) | 0.76 (0.44) | $1.03\times10^{-98}$ | 12.64 (13.41) | 1.00 (0.83) | $8.66\times10^{-16}$ |
| **Lacunar Brain Infract N (%)** | 21 (0.19) | 0 | NA | 108 (0.56) | 0 | NA | 26 (0.22) | 0 | NA |
| **Age mean (s.d)** | 61.95 (4.47) | 64.16 (5.22) | $1.02\times10^{-3}$ | 73.89 (4.22) | 74.99 (4.50) | $1.40\times10^{-2}$ | 67.15 (10.49) | 67.89 (9.44) | 0.58 |
| **Females N (%)** | 60 (0.56) | 53 (0.49) | 0.41 | 115 (0.59) | 132 (0.68) | 0.09 | 52 (0.45) | 58 (0.51) | 0.36 |
| **Hypertension status N (%)** | 37 (0.34) | 36 (0.33) | 1 | 138 (0.71) | 100 (0.52) | $1.08\times10^{-4}$ | 81 (0.71) | 66 (0.58) | 0.05 |
| **Systolic blood pressure mean (s.d)** | 123 (20.43) | 123.03 (16.91) | 0.99 | 137.65 (21.68) | 129.12 (18.68) | $4.06\times10^{-5}$ | 131.75 (19.53) | 123.73 (18.30) | $1.52\times10^{-3}$ |
| **Anti-hypertensive medication status N (%)** | 28 (0.26) | 23 (0.21) | 0.52 | 109 (0.56) | 78 (0.40) | $2.26\times10^{-3}$ | 57 (0.49) | 55 (0.49) | 0.90 |
| **Fasting Glucose mean (s.d)** | NA | NA | NA | 106.20 (36.51) | 99.47 (16.50) | 0.02 | 107.55 (27.07) | 105.48 (21.53) | 0.53 |
| **Diabetes status N (%)** | 11 (0.10) | 9 (0.08) | 0.82 | 22 (0.11) | 20 (0.10) | 0.87 | 21 (0.18) | 13 (0.12) | 0.19 |
| **HDL mean (s.d)** | 54.94 (19.88) | 53.53 (20.16) | 0.61 | 53.26 (14.70) | 55.047 (14.49) | 0.229 | 55.37 (16.81) | 56.03 (18.84) | 0.78 |
| **LDL mean (s.d)** | 125.83 (31.35) | 124.06 (29.95) | 0.68 | 125.08 (32.03) | 126.43 (30.58) | 0.673 | 105.15 (29.30) | 106.80 (31.66) | 0.69 |
| **Triglyceride level mean (s.d)** | 1.85 (1.64) | 1.54 (0.84) | 0.09 | 148.36 (102.98) | 140.08 (82.62) | 0.38 | 121.74 (68.84) | 107.12 (48.60) | 0.07 |
| **Lipid lowering medication N (%)** | 28 (0.26) | 29 (0.27) | 1 | 24 (0.12) | 11 (0.06) | 0.03 | 55 (0.47) | 50 (0.44) | 0.69 |
| **Body mass index mean (s.d)** | 26.55 (4.67) | 25.73 (3.91) | 0.17 | 26.74 (3.90) | 26.12 (4.67) | 0.16 | 28.67 (5.52) | 27.65 (4.67) | 0.13 |
| **Cardiovascular disease status_N (%)** | NA | NA | NA | 33 (0.17) | 26 (0.13) | 0.40 | 18 (0.16) | 15 (0.13) | 0.71 |

| | Rotterdam study I | | | ARIC-AA | | |
|---|---|---|---|---|---|---|
| **Variables** | **Extensive bSVD** | **Minimal bSVD** | **p-value** | **Extensive bSVD** | **Minimal bSVD** | **p-value** |
| **N** | 62 | 62 | NA | 121 | 121 | NA |
| **WMH (grades[0-9] or $mm^3$) mean (s.d)** | 37.29 (20.66) | 2.77 (1.27) | $1.71\times10^{-19}$ | 2.72 (1.04) | 0 (0) | $6.62\times10^{-56}$ |
| **Lacunar Brain Infract N (%)** | 26 (0.42) | 0 | NA | 41 (0.34) | 0 | NA |
| **Age mean (s.d)** | 79.2 (4.70) | 79.10 (4.84) | 0.90 | 62.57 (4.33) | 61.09 (3.71) | $4.72\times10^{-3}$ |
| **Females N (%)** | 31 (0.50) | 28 (0.45) | 0.72 | 69 (0.57) | 74 (0.61) | 0.60 |
| **Hypertension status N (%)** | 55 (0.89) | 54 (0.87) | 1 | 95 (0.79) | 69 (0.57) | $5.41\times10^{-4}$ |
| **Systolic blood pressure mean (s.d)** | 157.97 (23.24) | 150.15 (17.20) | 0.04 | 141.90 (23.65) | 128.96 (16.33) | $1.49\times10^{-6}$ |
| **Anti-hypertensive medication status N (%)** | 33 (0.53) | 34 (0.55) | 1 | 76 (0.63) | 55 (0.46) | $9.73\times10^{-3}$ |
| **Fasting Glucose mean (s.d)** | 5.87 (1.19) | 5.89 (1.47) | 0.86 | NA | NA | NA |
| **Diabetes status N (%)** | 6 (0.10) | 3 (0.05) | 0.49 | 36 (0.30) | 18 (0.15) | $5.27\times10^{-3}$ |
| **HDL mean (s.d)** | 1.54 (0.36) | 1.477 (0.40) | 0.362 | 54.18 (21.72) | 57.84 (19.63) | 0.17 |
| **LDL mean (s.d)** | 3.14 (0.96) | 3.214 (0.96) | 0.665 | 127.57 (40.00) | 131.76 (35.40) | 0.39 |
| **Triglyceride level mean (s.d)** | 1.29 (0.62) | 1.255 (0.50) | 0.706 | 1.33 (0.75) | 1.34 (0.79) | 0.87 |
| **Lipid lowering medication N (%)** | 23 (0.37) | 16 (0.26) | 0.246 | 68 (0.56) | 41 (0.34) | $7.78\times10^{-4}$ |
| **Body mass index mean (s.d)** | 27.02 (4.41) | 27.402 (3.31) | 0.589 | 29.26 (4.86) | 29.8 (5.43) | 0.42 |
| **Cardiovascular disease status_N (%)** | 6 (0.10) | 5 (0.08) | 1 | NA | NA | NA |

## Supplementary table 3: Association of rs2293871 variant within *HTRA1* gene with extreme bSVD in individual cohorts of European and African ancestries

| Study | N extremes | SNP | RA/OA | RA frequency | OR (95% CI) | p-value |
|---|---|---|---|---|---|---|
| *European ancestry* | | | | | | |
| **3C-Dijon** | 512 | rs2293871 | T/C | 0.19 | 1.92 (1.39-2.65) | $8.21\times10^{-5}$ |
| | | | | | | |
| **ARIC-EA** | 262 | rs2293871 | T/C | 0.19 | 1.01 (0.64-1.58) | 0.97 |
| **CHS-EA** | 670 | rs2293871 | T/C | 0.17 | 1.53 (1.12-2.09) | $7.68\times10^{-3}$ |
| **FHS** | 1078 | rs2293871 | T/C | 0.19 | 1.13 (0.90-1.42) | 0.28 |
| **RS1** | 296 | rs2293871 | T/C | 0.19 | 1.64 (1.06-2.55) | 0.03 |
| **RS2** | 218 | rs2293871 | T/C | 0.19 | 0.91 (0.54-1.55) | 0.74 |
| **RS3** | 784 | rs2293871 | T/C | 0.19 | 1.13 (0.87-1.48) | 0.36 |
| **Combined (ARIC-EA, CHS-EA, FHS, RS1-3)** | 3308 | rs2293871 | T/C | 0.19 | 1.21 (1.06-1.38) | $5.25\times10^{-3}$ |
| **Combined (3C-Dijon, ARIC-EA, CHS-EA, FHS, RS1-3)** | 3802 | rs2293871 | T/C | 0.19 | 1.29 (1.14-1.46) | $4.72\times10^{-5}$ |
| *African ancestry* | | | | | | |
| **ARIC-AA** | 244 | rs2293871 | T/C | 0.13 | 0.75 (0.43-1.33) | 0.33 |
| **CHS-AA** | 146 | rs2293871 | T/C | 0.15 | 1.02 (0.50-2.07) | 0.96 |
| **Combined (ARIC-AA, CHS-AA)** | 390 | rs2293871 | T/C | 0.14 | 0.85 (0.54-1.32) | 0.47 |

## Supplementary table 4: Functional consequences of rs2293871 and variants in LD ($r^2$>0.60) in 1000 Genomes phase 1 European ancestry reference panel

| rsID | Chr pos. (hg38) | ref/alt alleles | Freq-AFR, AMR, ASN, EUR) | $R^2$ | D' | Chromatin Marks | DNAse | Proteins | eQTL | Motifs |
|---|---|---|---|---|---|---|---|---|---|---|
| rs2293871 | 10:122514155 | C/T | 0.14, 0.2, 0.46, 0.21 | 1 | 1 | E013,H3K4me1_Enh;E023,H3K4me1_Enh;E025,H3K4me1_Enh;E028,H3K4me1_Enh;E034,H3K4me1_Enh;E037,H3K4me1_Enh;E038,H3K4me1_Enh;E040,H3K4me1_Enh;E041,H3K4me1_Enh;E042,H3K4me1_Enh;E043,H3K4me1_Enh;E044,H3K4me1_Enh;E045,H3K4me1_Enh;E046,H3K4me1_Enh;E047,H3K4me1_Enh;E048,H3K4me1_Enh;E049,H3K4me1_Enh;E052,H3K4me1_Enh;E055,H3K4me1_Enh;E056,H3K4me1_Enh;E058,H3K4me1_Enh;E063,H3K4me1_Enh;E066,H3K4me1_Enh;E067,H3K4me1_Enh;E068,H3K4me1_Enh;E069,H3K4me1_Enh;E071,H3K4me1_Enh;E072,H3K4me1_Enh;E073,H3K4me1_Enh;E074,H3K4me1_Enh;E077,H3K4me1_Enh;E079,H3K4me1_Enh;E095,H3K4me1_Enh;E099,H3K4me1_Enh;E102,H3K4me1_Enh;E103,H3K4me1_Enh;E104,H3K4me1_Enh;E107,H3K4me1_Enh;E108,H3K4me1_Enh;E111,H3K4me1_Enh;E113,H3K4me1_Enh;E115,H3K4me1_Enh;E119,H3K4me1_Enh;E120,H3K4me1_Enh;E125,H3K4me1_Enh;E126,H3K4me1_Enh;E129,H3K4me1_Enh;E023,H3K9ac_Pro;E067,H3K9ac_Pro;E073,H3K9ac_Pro;E074,H3K9ac_Pro;E103,H3K9ac_Pro;E049,H3K27ac_Enh;E055,H3K27ac_Enh;E063,H3K27ac_Enh;E066,H3K27ac_Enh;E067,H3K27ac_Enh;E068,H3K27ac_Enh;E069,H3K27ac_Enh;E071,H3K27ac_Enh;E072,H3K27ac_Enh;E073,H3K27ac_Enh;E074,H3K27ac_Enh;E075,H3K27ac_Enh;E076,H3K27ac_Enh;E078,H3K27ac_Enh;E092,H3K27ac_Enh;E095,H3K27ac_Enh;E105,H3K27ac_Enh;E108,H3K27ac_Enh;E109,H3K27ac_Enh;E111,H3K27ac_Enh;E113,H3K27ac_Enh;E129,H3K27ac_Enh;E074,H3K4me3_Pro;E094,H3K4me3_Pro | . | . | . | DMRT4;Irf_known10 |
| rs2736928 | 10:122517501 | C/T | 0.72, 0.78, 0.55, 0.73 | 0.75 | -1 | E006,H3K4me1_Enh;E007,H3K4me1_Enh;E027,H3K4me1_Enh;E028,H3K4me1_Enh;E034,H3K4me1_Enh;E037,H3K4me1_Enh;E038,H3K4me1_Enh;E040,H3K4me1_Enh;E041,H3K4me1_Enh;E042,H3K4me1_Enh;E043,H3K4me1_Enh;E045,H3K4me1_Enh;E046,H3K4me1_Enh;E047,H3K4me1_Enh;E048,H3K4me1_Enh;E068,H3K4me1_Enh;E072,H3K4me1_Enh;E073,H3K4me1_Enh;E074,H3K4me1_Enh;E075,H3K4me1_Enh;E077,H3K4me1_Enh;E079,H3K4me1_Enh;E089,H3K4me1_Enh;E090,H3K4me1_Enh;E092,H3K4me1_Enh;E095,H3K4me1_Enh;E096,H3K4me1_Enh;E101,H3K4me1_Enh;E105,H3K4me1_Enh;E113,H3K4me1_Enh;E115,H3K4me1_Enh;E119,H3K4me1_Enh;E120,H3K4me1_Enh;E125,H3K4me1_Enh;E007,H3K9ac_Pro;E047,H3K9ac_Pro;E068,H3K9ac_Pro;E069,H3K9ac_Pro;E072,H3K9ac_Pro;E073,H3K9ac_Pro;E103,H3K9ac_Pro;E111,H3K9ac_Pro;E022,H3K4me3_Pro;E037,H3K4me3_Pro;E038,H3K4me3_Pro;E045,H3K4me3_Pro;E113,H3K4me3_Pro;E045,H3K27ac_Enh;E055,H3K27ac_Enh;E066,H3K27ac_Enh;E067,H3K2 | . | . | Westra2013, Whole_Blood, HTRA1, 9.62E-5 | Arid5b;HNF1_6;HNF1_7 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 7ac_Enh;E068,H3K27ac_Enh;E069,H3K27ac_Enh;E071,H3K27ac_Enh;E073,H3K27ac_Enh;E074,H3K27ac_Enh;E096,H3K27ac_Enh;E105,H3K27ac_Enh;E113,H3K27ac_Enh;E120,H3K27ac_Enh | | | | |
| rs714989 | 10:122521209 | A/G | 0.16, 0.19, 0.45, 0.21 | 0.97 | 0.98 | E006,H3K4me1_Enh;E013,H3K4me1_Enh;E015,H3K4me1_Enh;E017,H3K4me1_Enh;E025,H3K4me1_Enh;E027,H3K4me1_Enh;E052,H3K4me1_Enh;E053,H3K4me1_Enh;E054,H3K4me1_Enh;E055,H3K4me1_Enh;E056,H3K4me1_Enh;E068,H3K4me1_Enh;E069,H3K4me1_Enh;E071,H3K4me1_Enh;E075,H3K4me1_Enh;E079,H3K4me1_Enh;E080,H3K4me1_Enh;E084,H3K4me1_Enh;E085,H3K4me1_Enh;E089,H3K4me1_Enh;E092,H3K4me1_Enh;E094,H3K4me1_Enh;E097,H3K4me1_Enh;E098,H3K4me1_Enh;E111,H3K4me1_Enh;E119,H3K4me1_Enh;E120,H3K4me1_Enh;E122,H3K4me1_Enh;E125,H3K4me1_Enh;E129,H3K4me1_Enh;E007,H3K9ac_Pro;E018,H3K9ac_Pro;E068,H3K9ac_Pro;E072,H3K9ac_Pro;E067,H3K27ac_Enh;E069,H3K27ac_Enh;E071,H3K27ac_Enh;E072,H3K27ac_Enh;E073,H3K27ac_Enh;E074,H3K27ac_Enh;E097,H3K27ac_Enh;E120,H3K27ac_Enh;E094,H3K4me3_Pro | E055; E056; E080; E081; E082; E124; E125; E126; E128 | . | Lappalainen2013, Lymphoblastoid_EUR_exonlevel, ENSG00000179988.8_124742249_124742540, 3.75E-06 | CACD_1; GLI; Glis2; ZBTB7A_known2; Zic_1; Zic_2; Zic_3 |
| rs4279944 | 10:122538121 | C/T | 0.22, 0.19, 0.28, 0.19 | 0.84 | 0.97 | . | . | . | . | Ets_disc1; GATA_disc4 |
| rs2300431 | 10:122483301 | G/A | 0.33, 0.28, 0.46, 0.29 | 0.62 | 0.97 | E013,H3K27ac_Enh;E049,H3K27ac_Enh;E058,H3K27ac_Enh;E067,H3K27ac_Enh;E068,H3K27ac_Enh;E069,H3K27ac_Enh;E071,H3K27ac_Enh;E072,H3K27ac_Enh;E073,H3K27ac_Enh;E074,H3K27ac_Enh;E076,H3K27ac_Enh;E103,H3K27ac_Enh;E013,H3K4me1_Enh;E023,H3K4me1_Enh;E025,H3K4me1_Enh;E049,H3K4me1_Enh;E063,H3K4me1_Enh;E067,H3K4me1_Enh;E068,H3K4me1_Enh;E069,H3K4me1_Enh;E071,H3K4me1_Enh;E072,H3K4me1_Enh;E073,H3K4me1_Enh;E074,H3K4me1_Enh;E076,H3K4me1_Enh;E078,H3K4me1_Enh;E083,H3K4me1_Enh;E092,H3K4me1_Enh;E103,H3K4me1_Enh;E108,H3K4me1_Enh;E110,H3K4me1_Enh;E025,H3K9ac_Pro;E049,H3K9ac_Pro;E067,H3K9ac_Pro;E069,H3K9ac_Pro;E072,H3K9ac_Pro;E073,H3K9ac_Pro;E074,H3K9ac_Pro;E076,H3K9ac_Pro;E088,H3K9ac_Pro | E083; E089; E090 | . | Westra2013, Whole_Blood, HTRA1, 4.66E-4 | Cdx2_1; Cdx2_2; Hoxd8; Lhx3_2; Ncx_2; Sox_17 |
| rs876790 | 10:122504019 | C/T | 0.6, 0.74, 0.54, 0.74 | 0.75 | -0.99 | E002,H3K4me1_Enh;E005,H3K4me1_Enh;E006,H3K4me1_Enh;E012,H3K4me1_Enh;E022,H3K4me1_Enh;E023,H3K4me1_Enh;E025,H3K4me1_Enh;E026,H3K4me1_Enh;E027,H3K4me1_Enh;E028,H3K4me1_Enh;E049,H3K4me1_Enh;E052,H3K4me1_Enh;E055,H3K4me1_Enh;E056,H3K4me1_Enh;E057,H3K4me1_Enh;E058,H3K4me1_Enh;E063,H3K4me1_Enh;E065,H3K4me1_Enh;E067,H3K4me1_Enh;E068,H3K4me1_Enh;E069,H3K4me1_Enh;E070,H3K4me1_Enh;E071,H3K4me1_Enh;E072,H3K4me1_Enh;E073,H3K4me1_Enh;E074,H3K4me1_Enh;E075,H3K4me1_Enh;E076,H3K4me1_Enh;E077,H3K4me1_Enh;E078,H3K4me1_Enh;E079,H3K4me1_Enh;E081,H3K4me1_Enh;E083,H3K4me1_Enh;E084,H3K4me1_Enh;E085,H3K4me1_Enh;E088,H3K4me1_Enh;E089,H3K4me1_Enh;E090,H3K4me1_Enh;E092,H3K4me1_Enh;E093,H3K4me1_Enh;E094,H3K4me1_Enh;E095,H3K4me1_Enh;E096,H3K4me1_E | E089; E090 | GM12878, PAX5C20, HudsonAlpha, None | Westra2013, Whole_Blood, HTRA1, 6.92E-5 | Myc_disc2; Myc_known5; YY1_disc1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | nh;E097,H3K4me1_Enh;E098,H3K4me1_Enh;E102,H3K4me1_Enh;E103,H3K4me1_Enh;E105,H3K4me1_Enh;E107,H3K4me1_Enh;E108,H3K4me1_Enh;E110,H3K4me1_Enh;E111,H3K4me1_Enh;E113,H3K4me1_Enh;E114,H3K4me1_Enh;E117,H3K4me1_Enh;E119,H3K4me1_Enh;E120,H3K4me1_Enh;E121,H3K4me1_Enh;E125,H3K4me1_Enh;E126,H3K4me1_Enh;E127,H3K4me1_Enh;E002,H3K4me3_Pro;E005,H3K4me3_Pro;E022,H3K4me3_Pro;E055,H3K4me3_Pro;E073,H3K4me3_Pro;E098,H3K4me3_Pro;E111,H3K4me3_Pro;E113,H3K4me3_Pro;E005,H3K27ac_Enh;E006,H3K27ac_Enh;E011,H3K27ac_Enh;E013,H3K27ac_Enh;E026,H3K27ac_Enh;E049,H3K27ac_Enh;E055,H3K27ac_Enh;E056,H3K27ac_Enh;E058,H3K27ac_Enh;E065,H3K27ac_Enh;E067,H3K27ac_Enh;E068,H3K27ac_Enh;E069,H3K27ac_Enh;E071,H3K27ac_Enh;E072,H3K27ac_Enh;E073,H3K27ac_Enh;E074,H3K27ac_Enh;E076,H3K27ac_Enh;E090,H3K27ac_Enh;E092,H3K27ac_Enh;E097,H3K27ac_Enh;E103,H3K27ac_Enh;E105,H3K27ac_Enh;E108,H3K27ac_Enh;E111,H3K27ac_Enh;E112,H3K27ac_Enh;E113,H3K27ac_Enh;E121,H3K27ac_Enh;E122,H3K27ac_Enh;E125,H3K27ac_Enh;E126,H3K27ac_Enh;E128,H3K27ac_Enh;E129,H3K27ac_Enh;E007,H3K9ac_Pro;E017,H3K9ac_Pro;E023,H3K9ac_Pro;E025,H3K9ac_Pro;E027,H3K9ac_Pro;E052,H3K9ac_Pro;E062,H3K9ac_Pro;E067,H3K9ac_Pro;E068,H3K9ac_Pro;E069,H3K9ac_Pro;E072,H3K9ac_Pro;E073,H3K9ac_Pro;E074,H3K9ac_Pro;E075,H3K9ac_Pro;E076,H3K9ac_Pro;E083,H3K9ac_Pro;E088,H3K9ac_Pro;E107,H3K9ac_Pro;E111,H3K9ac_Pro;E117,H3K9ac_Pro;E125,H3K9ac_Pro | | | | |

## Supplementary table 5: Number of protein-modifying rare alleles and their cumulative frequencies in candidate genes in 3C-Dijon WES extreme bSVD cohort.

| Candidate gene | Number of protein-modifying rare and low frequency variants | Cumulative frequency of protein-modifying minor alleles |
|---|---|---|
| *COL4A1* | 13 | 0.02 |
| *COL4A2* | 29 | 0.09 |
| *HTRA1* | 4 | $3.91\times10^{-3}$ |
| *NOTCH3* | 31 | 0.10 |
| *TREX1* | 5 | $5.86\times10^{-3}$ |

## Supplementary table 6: Computationally predicated mucin type GalNAc O-glycosylation sites in NOTCH3 EGF like domain (Amino acids 40 to 1373).

The GalNAc O-glycosylation sites with high prediction score (>0.50) that are near protein-modifying variant observed in the 3C-Dijon cohort are highlighted in italics.

| Sequence Name | Source | Feature | Start (aa position) | End (aa position) | Score | Near (<6aa) protein-modifying variant in 3C-Dijon extreme-BSVD cohort |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 47 | 47 | 0.49 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 56 | 56 | 0.46 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *60* | *60* | *0.54* | *Yes (R61W)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 84 | 84 | 0.57 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *95* | *95* | *0.63* | *Yes (V98A)* |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *96* | *96* | *0.65* | *Yes (V98A)* |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *101* | *101* | *0.57* | *Yes (V98A)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 105 | 105 | 0.74 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 118 | 118 | 0.57 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 125 | 125 | 0.47 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 126 | 126 | 0.5 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 135 | 135 | 0.66 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 145 | 145 | 0.57 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 154 | 154 | 0.63 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 157 | 157 | 0.65 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *173* | *173* | *0.65* | *Yes (H170R)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 177 | 177 | 0.57 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 180 | 180 | 0.18 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 190 | 190 | 0.22 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 204 | 204 | 0.62 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 211 | 211 | 0.63 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 215 | 215 | 0.35 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 219 | 219 | 0.07 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 250 | 250 | 0.59 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 257 | 257 | 0.09 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 267 | 267 | 0.17 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 272 | 272 | 0.39 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 290 | 290 | 0.26 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 294 | 294 | 0.27 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 299 | 299 | 0.22 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 307 | 307 | 0.18 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 310 | 310 | 0.48 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 312 | 312 | 0.40 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 320 | 320 | 0.44 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *328* | *328* | *0.82* | *Yes (T328I)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 335 | 335 | 0.29 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 345 | 345 | 0.10 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 357 | 357 | 0.63 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 368 | 368 | 0.42 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 378 | 378 | 0.66 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 384 | 384 | 0.67 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 396 | 396 | 0.25 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 411 | 411 | 0.36 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 414 | 414 | 0.07 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 424 | 424 | 0.55 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 430 | 430 | 0.49 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 437 | 437 | 0.39 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 445 | 445 | 0.42 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 454 | 454 | 0.41 | Yes (Q452R) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 462 | 462 | 0.10 | Yes (A459T) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 464 | 464 | 0.05 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 475 | 475 | 0.35 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 476 | 476 | 0.20 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 492 | 492 | 0.42 | Yes (N489S, P496L, S497L) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 494 | 494 | 0.38 | Yes (N489S, P496L, S497L) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 497 | 497 | 0.34 | Yes (P496L, S497L, S502F) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 500 | 500 | 0.36 | Yes (P496L, S497L, S502F) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 502 | 502 | 0.18 | Yes (S497L, S502F) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 503 | 503 | 0.35 | Yes (S502F) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 513 | 513 | 0.77 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 514 | 514 | 0.59 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 540 | 540 | 0.06 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 550 | 550 | 0.64 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 565 | 565 | 0.19 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 567 | 567 | 0.38 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 575 | 575 | 0.32 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 577 | 577 | 0.10 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 581 | 581 | 0.71 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 588 | 588 | 0.62 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 610 | 610 | 0.63 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 612 | 612 | 0.47 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 613 | 613 | 0.30 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 626 | 626 | 0.44 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 630 | 630 | 0.33 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 650 | 650 | 0.31 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 663 | 663 | 0.48 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 664 | 664 | 0.40 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 671 | 671 | 0.45 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 687 | 687 | 0.33 | Yes (G686A) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 696 | 696 | 0.67 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 705 | 705 | 0.34 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 725 | 725 | 0.5 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 730 | 730 | 0.69 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 732 | 732 | 0.69 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 740 | 740 | 0.76 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 748 | 748 | 0.38 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 750 | 750 | 0.61 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 751 | 751 | 0.53 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *759* | *759* | *0.55* | *Yes (T759S)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 773 | 773 | 0.48 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 776 | 776 | 0.79 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *788* | *788* | *0.61* | *Yes (R785C)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 797 | 797 | 0.49 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 827 | 827 | 0.36 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 832 | 832 | 0.32 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 834 | 834 | 0.21 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 836 | 836 | 0.29 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 842 | 842 | 0.41 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 845 | 845 | 0.45 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 863 | 863 | 0.17 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 870 | 870 | 0.07 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 872 | 872 | 0.22 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 874 | 874 | 0.51 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 893 | 893 | 0.59 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 900 | 900 | 0.22 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 902 | 902 | 0.55 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 907 | 907 | 0.38 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 909 | 909 | 0.33 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 911 | 911 | 0.64 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 929 | 929 | 0.49 | Yes (S931G) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 931 | 931 | 0.43 | Yes (S931G) |

| | | | | | | |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 932 | 932 | 0.19 | Yes (S931G) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 938 | 938 | 0.31 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 945 | 945 | 0.12 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 947 | 947 | 0.27 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 955 | 955 | 0.32 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 968 | 968 | 0.43 | |
| *SP_Q9UM47_NOTC3_HUMAN* | *netOGlyc-4.0.0.13* | *CARBOHYD* | *978* | *978* | *0.65* | *Yes (H981Y)* |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 987 | 987 | 0.23 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 991 | 991 | 0.38 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 993 | 993 | 0.26 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 999 | 999 | 0.59 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1005 | 1005 | 0.70 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1018 | 1018 | 0.25 | Yes (A1020P) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1029 | 1029 | 0.21 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1037 | 1037 | 0.35 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1066 | 1066 | 0.08 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1067 | 1067 | 0.17 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1077 | 1077 | 0.28 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1079 | 1079 | 0.10 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1098 | 1098 | 0.54 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1128 | 1128 | 0.45 | Yes (H1133Q) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1136 | 1136 | 0.38 | Yes (H1133Q) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1147 | 1147 | 0.49 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1152 | 1152 | 0.14 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1172 | 1172 | 0.70 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1181 | 1181 | 0.47 | Yes (V1183M, V1186L) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1192 | 1192 | 0.74 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1198 | 1198 | 0.40 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1211 | 1211 | 0.82 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1219 | 1219 | 0.36 | Yes (A1217T) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1239 | 1239 | 0.29 | Yes (H1235L) |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1245 | 1245 | 0.74 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1248 | 1248 | 0.36 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1252 | 1252 | 0.81 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1264 | 1264 | 0.92 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1272 | 1272 | 0.15 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1274 | 1274 | 0.64 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1292 | 1292 | 0.52 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1307 | 1307 | 0.62 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1320 | 1320 | 0.44 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1323 | 1323 | 0.78 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1326 | 1326 | 0.83 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1330 | 1330 | 0.81 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1335 | 1335 | 0.72 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1338 | 1338 | 0.66 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1349 | 1349 | 0.60 | |
| SP_Q9UM47_NOTC3_HUMAN | netOGlyc-4.0.0.13 | CARBOHYD | 1368 | 1368 | 0.41 | |

## Supplementary Table 7: The ClinVar pathogenic and likely pathogenic mutations for small vessel disease of brain (27th February 2017)

| Gene | ClinVar Disease terms | RefSeq ID | cDNA modification | Protein modification |
|---|---|---|---|---|
| *NOTCH3* | 1. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy<br>2. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy Recurrent subcortical infarcts<br>3. Infantile myofibromatosis 1 Infantile myofibromatosis 2 | NM:000435.2 | c.187G>A | p.Ala63Thr |
| | | | c.213G>T | p.Trp71Cys |
| | | | c.397C>T | p.Arg133Cys |
| | | | c.457C>T | p.Arg153Cys |
| | | | c.505C>T | p.Arg169Cys |
| | | | c.544C>T | p.Arg182Cys |
| | | | c.714_758del45 | p.Asp239_Asp253del |
| | | | c.994C>T | p.Arg332Cys |
| | | | c.1187C>G | p.Ser396Cys |
| | | | c.1282T>A | p.Cys428Ser |
| | | | c.1363T>C | p.Cys455Arg |
| | | | c.2411_2566del156 | Not Applicable |
| *HTRA1* | 1. Cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy<br>2. Cerebral arteriopathy autosomal dominant with subcortical infarcts and leukoencephalopathy type 2 | NM:002775.4 | c.126delG | p.Glu42Aspfs |
| | | | c.497G>T | p.Arg166Leu |
| | | | c.517G>C | p.Ala173Pro |
| | | | c.754G>A | p.Ala252Thr |
| | | | c.821G>A | p.Arg274Gln |
| | | | c.852C>A | p.Ser284Arg |
| | | | c.883G>A | p.Gly295Arg |
| | | | c.889G>A | p.Val297Met |
| | | | c.904C>T | p.Arg302Ter |
| | | | c.961G>A | p.Ala321Thr |
| | | | c.973_1005del33 | Not Applicable |
| | | | c.1091T>C | p.Leu364Pro |
| | | | c.1108C>T | p.Arg370Ter |
| *COL4A1* | 1. Brain small vessel disease with hemorrhage<br>2. Brain small vessel disease with | NM:001845.5 | c.1A>T | p.Met1Leu |
| | | | c.1685G>A | p.Gly562Glu |

| | | | | |
|---|---|---|---|---|
| | hemorrhage not provided<br>3. Brain small vessel disease with hemorrhage Porencephaly-1<br>4. Porencephaly-1<br>5. SCHIZENCEPHALY | | c.1769G>A | p.Gly590Glu |
| | | | c.2085delC | p.Gly696Alafs |
| | | | c.2086G>A | p.Gly696Ser |
| | | | c.2122G>A | p.Gly708Arg |
| | | | c.2194_1G>A | Not Applicable |
| | | | c.2159G>A | p.Gly720Asp |
| | | | c.2245G>A | p.Gly749Ser |
| | | | c.2263G>A | p.Gly755Arg |
| | | | c.2317G>C | p.Gly773Arg |
| | | | c.2662G>A | p.Gly888Arg |
| | | | c.3389G>A | p.Gly1130Asp |
| | | | c.3555A>G | p.Lys1185 |
| | | | c.3706G>A | p.Gly1236Arg |
| | | | c.3976G>A | p.Gly1326Arg |
| | | | c.4267G>C | p.Gly1423Arg |
| | | | c.4738G>C | p.Gly1580Arg |
| | | | c.4881C>G | p.Asn1627Lys |
| *COL4A2* | 1. Porencephaly-2 | NM:001846.3 | c.3110G>A | p.Gly1037Glu |
| | | | c.3455G>A | p.Gly1152Asp |
| *TREX1* | 1. Aicardi Goutieres syndrome 1<br>2. Aicardi Goutieres syndrome 1 Aicardi Goutieres syndrome 1 autosomal dominant<br>3. Aicardi Goutieres syndrome 1 Aicardi Goutieres syndrome 1 autosomal dominant Chilblain lupus 1 not provided | NM:016381.5 | c.52G>A | p.Asp18Asn |
| | | | c.223dupG | p.Glu75Glyfs |
| | | | c.309dupC | p.Thr104Hisfs |
| | | | c.317_318delAG | p.Gln106Argfs |
| | | | c.340C>T | p.Arg114Cys |
| | | | c.377_378dupTG | p.Ala127Trpfs |
| | | | c.377_378delTG | p.Val126Glyfs |
| | | | c.530T>C | p.Val177Ala |
| | | | c.531_533dupGGC | p.Ala178_His179insAla |
| | | | c.558_573dup16 | p.Glu192Profs |

| | | | | |
|---|---|---|---|---|
| | | | c.562delC | p.Leu188Cysfs |
| | | | c.665delG | p.Ser222Thrfs |
| | | | c.764_766dupATG | p.Asp255dup |
| | | | c.767T>A | p.Val256Asp |
| | | | c.763G>A | p.Asp255Asn |
| | | | c.774_827dup54 | p.Ala276_His277insLeuLeuSerIleCysGlnTrpArg ProGlnAlaLeuLeuArgTrpValAspAla |
| | | | c.790_793dupCAGT | p.Trp265Serfs |
| | | | c.794G>A | p.Trp265Ter |
| | | | c.1033_1050del18 | p.Pro345_Ala350del |
| | | | c.1072A>C | p.Thr358Pro |