Supplementary Methods

Dataset

We extracted information on genetic variants related to stroke from the data of 15,548 individuals obtained through the Pilot Project for National Bio Big Data of Korea at the Clinical & Omics Data Archive (https://coda.nih.go.kr/stats/selectRegList.do). This study includes whole-genome sequencing dataset from the following cohorts (dataset accession numbers): (1) 5,000 individuals from the Korea Epidemiology Study cohort (CODA_D22001, CODA_D23010); (2) 2,504 individuals from the Ulsan 10,000 Genome Project (CODA_D22006, CODA_D23009); (3) 7,722 individuals from families of rare disease patients' families (CODA_D22004, CODA_D23012); and (4) 322 individuals from colorectal patients' families (CODA_D22002, CODA_D23013). There were 7,401 men and 8,147 women, with a mean age of 46 years.

The analysis focused on 24 genes related to stroke: *ABCC6*, *ACVRL1*, *APP*, *CBS*, *CCM2*, *CECR1*, *COL3A1*, *COL4A1*, *COL4A2*, *COLGALT1*, *CST3*, *CTSA*, *ENG*, *FBN1*, *GLA*, *HBB*, *HTRA1*, *ITM2B*, *KRIT1*, *NF1*, *NOTCH3*, *PDCD10*, *RNF213*, and *TREX1*.

Variant annotation

The extracted variants were annotated using the ANNOVAR software (https://annovar.openbioinformatics.org/) with the following databases:

Item	Database	Version/date	Purpose
ANNOVAR	refGene	21.10.19	To identify variant type and information
	gnomAD (exome)	v2.1.1	For allele frequency (AF) filtering
	ClinVar	22.03.20	For pathogenic variant filtering
	dbNSFP	v4.2a	For CADD score filtering
	avsnp150	17.09.29	To confirm variant rsID (dbSNP)

Variant filtering

We filtered the identified variants to retain only those with a minor allele frequency (MAF) of less than 0.001, based on the gnomAD v2.1.1 (https://gnomad.broadinstitute.org/; exome, non-neuro) dataset. However, we included the *NOTCH3* variants (R544C, R1231C) and the *RNF213* variant (R4810K) in our list if they were known to be pathogenic, even if their MAF was above 0.001. For the *TREX1* gene, we included only the variants in the C-terminus, as only these variants were linked to strokes.

Analysis categories

From the filtered list, we performed analyses on the following three categories: (1) pathogenic clinical variant: variants classified as pathogenic or likely pathogenic in the ClinVar_20220320 database ([CLNSIG] entry); (2) CADD to predict variant effect: variants with a PHRED-scaled C-score ([CADD_phred]) over 20 in the dbNSFP v4.2 database (http://database.liulab.science/dbNSFP), indicating a higher likelihood of impacting protein function; (3) all nonsynonymous variants: variants classified as nonsynonymous, stopgain, nonframeshift insertion/deletion, or frameshift insertion/deletion in the refGene database ([ExonicFunc.refGene]), as these changes are likely to affect protein production.

Calculation of population frequency

For each variant that met the criteria, we calculated the number of individuals carrying the variant and determined the population frequency. The population frequency was calculated as follows: population frequency = number of individuals with the variant/total number of individuals (15,548).