Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study

NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium (ISGC)*

Summary

Background The discovery of disease-associated loci through genome-wide association studies (GWAS) is the leading approach to the identification of novel biological pathways underlying diseases in humans. Until recently, GWAS in ischaemic stroke have been limited by small sample sizes and have yielded few loci associated with ischaemic stroke. We did a large-scale GWAS to identify additional susceptibility genes for stroke and its subtypes.

Methods To identify genetic loci associated with ischaemic stroke, we did a two-stage GWAS. In the first stage, we included 16851 cases with state-of-the-art phenotyping data and 32473 stroke-free controls. Cases were aged 16 to 104 years, recruited between 1989 and 2012, and subtypes of ischaemic stroke were recorded by centrally trained and certified investigators who used the web-based protocol, Causative Classification of Stroke (CCS). We constructed case-control strata by identifying samples that were genotyped on nearly identical arrays and were of similar genetic ancestral background. We cleaned and imputed data by use of dense imputation reference panels generated from whole-genome sequence data. We did genome-wide testing to identify stroke-associated loci within each stratum for each available phenotype, and we combined summary-level results using inverse variance-weighted fixed-effects meta-analysis. In the second stage, we did in-silico lookups of 1372 single nucleotide polymorphisms identified from the first stage GWAS in 20941 cases and 364736 unique stroke-free controls. The ischaemic stroke subtypes of these cases had previously been established with the Trial of Org 10172 in Acute Stroke Treatment (TOAST) classification system, in accordance with local standards. Results from the two stages were then jointly analysed in a final meta-analysis.

Findings We identified a novel locus (G allele at rs12122341) at 1p13.2 near TSPAN2 that was associated with large artery atherosclerosis-related stroke (first stage odds ratio [OR] 1.21, 95% CI 1.13–1.30, p=4.26×10⁻⁸; joint OR 1.24, 1.15–1.33, p=2.92×10⁻⁹) for large artery atherosclerosis stroke. The 12q24 locus near ALDH2, which has previously been associated with all ischaemic stroke but not with any specific subtype, exceeded genome-wide significance in the meta-analysis of small artery stroke (first stage OR 1.29, 1.18–1.42, p=3.50×10⁻⁸; joint OR 1.24, 1.15–1.33, p=4.52×10⁻⁸; joint OR 1.20, 1.12–1.28, p=6.30×10⁻⁹). Other loci associated with stroke in previous studies, including NINJ2, were not confirmed.

Interpretation Our results suggest that all ischaemic stroke-related loci previously implicated by GWAS are subtype specific. We identified a novel gene associated with large artery atherosclerosis stroke susceptibility. Follow-up studies will be necessary to establish whether the locus near TSPAN2 can be a target for a novel therapeutic approach to stroke prevention. In view of the subtype-specificity of the associations detected, the rich phenotyping data available in the Stroke Genetics Network (SiGN) are likely to be crucial for further genetic discoveries related to ischaemic stroke.

Introduction Worldwide, stroke is the second leading cause of death¹ and a major contributor to dementia and age-related cognitive decline. About 15 million people have a stroke each year.¹ Most survivors are left with a permanent disability, which makes stroke the world’s leading cause of adult incapacity.² Strokes result from the sudden occlusion or rupture of a blood vessel supplying the brain, and so are categorised accordingly as ischaemic (vessel occlusion) or haemorrhagic (vessel rupture) on the basis of neuroimaging results. Up to 85% of all strokes are ischaemic.

Although hypertension, atrial fibrillation, diabetes mellitus, and cigarette smoking are known risk factors for stroke,¹ a substantial proportion of the risk remains unexplained and might be attributable to inherited genetic variation. Discovery of genetic variants that predispose to stroke is a crucial first step toward the development of improved diagnostic tests for stroke and novel therapies that might reduce the disease burden. Genome-wide association studies (GWAS) have thus far identified only a few confirmed loci,³,⁴ which together account for a small proportion of the heritable risk.³ Ischaemic stroke occurs when the blood flow to a region of the brain is interrupted because of blockage of a blood vessel. Because vessel occlusion can occur through different mechanisms, ischaemic stroke can be subtyped.
Research in context

Evidence before this study
We searched PubMed with the search terms “stroke” and “genomic wide association study” for reports published before Oct 19, 2015. We only included peer-reviewed reports in English. Compared with the rapid pace of genetic discovery for other common disorders, only four loci (PITX2, HDAC9, ZFHX3, and 12q24.2) have been convincingly implicated by genome-wide association studies (GWAS) in ischaemic stroke. GWAS of stroke have been limited by small sample sizes and concerns about phenotypic heterogeneity.

Added value of this study
To our knowledge, the National Institute of Neurological Disorders and Stroke (NINDS)-Stroke Genetics Network (SiGN) project is the largest and most comprehensive study of ischaemic stroke so far. Discovery analyses were done in 16 851 cases and 32 473 controls and findings were followed up in an additional 20 941 cases and 364 736 controls. Furthermore, the project implemented the Causative Classification of Stroke (CCS) system to subtype cases and generated a rich phenotypic database. Trial of Org 10172 in Acute Stroke Treatment (TOAST)-based subtypes were also available, allowing for the first ever analysis of the genetic overlap between TOAST and CCS subtypes.

Implications of all the available evidence
Our data show that increasing sample size and applying a standardised subtyping method can reveal additional information about the underlying genetic architecture of stroke. Because we had access to phenotype information generated by two different subtyping methods, we also showed that there is moderate to strong genetic correlation between the CCS and TOAST subtyping methods, suggesting that future studies might benefit from liberal inclusion of cases, regardless of subtyping approach. Also, our results show that all discovered loci, including the 12q24.12 locus, which was previously implicated in all ischaemic stroke, are specific to a single subtype, suggesting that these subtypes will have at least partly distinct genetic signatures. Because of the subtype-specificity of genetic associations in stroke, substantially larger samples of stroke subtypes will probably be needed to expand the number of identified stroke loci to that of other common diseases.

on the basis of the presumed mechanism: large artery atherosclerosis, cardioembolism, or small artery occlusion. With one exception, all associations for ischaemic stroke detected in GWAS have been subtype-specific, suggesting the need for studies that are powered to detect subtype-specific associations. The National Institute of Neurological Disorders and Stroke (NINDS) Stroke Genetics Network (NINDS-SiGN) is the largest and most comprehensive GWAS of stroke and its subtypes to date. We sought to detect new associations of polymorphisms with risk of ischaemic stroke and its subtypes and to provide evidence for previously reported associations.

Methods

Study design
We did a two-stage joint association analysis of ischaemic stroke and its subtypes. The first stage consisted of a GWAS, and the second stage was an in-silico association analysis of the top single nucleotide polymorphisms (SNPs) identified in the first stage in a set of independent samples of cases and controls. We then analysed both stages together to identify loci that exceeded the threshold for genome-wide significance (1×10⁻⁸). Compared with separate discovery and replication analyses, this two-stage approach has been shown to improve the power for discovery without altering the type I error.¹⁰

Study sample
For the first stage, we assessed 31 existing collections that included cases of ischaemic stroke with either available genotypic data or DNA for genotyping, neuroimaging confirmation of stroke, and adequate clinical data to enable phenotypic classification. The cases of ischaemic stroke in the second stage met similar requirements, except that we used pre-existing Trial of Org 10172 in Acute Stroke Treatment (TOAST)¹⁰ subtyping data for the phenotypic classification. The appendix contains details about each collection, including their study design.

For each collection, approval for inclusion in the SiGN analysis complied with local ethical standards and with local institutional review board and ethics committee oversight. All people included as cases and controls provided written informed consent for genetic studies either directly or by a legally authorised representative.

Classification of stroke subtype
In the NINDS-SiGN, we used two subtyping systems: the Causative Classification of Stroke (CCS) system, which is a standardised web-based subtype classification system, and the more widely used TOAST subtype classification system.¹¹,¹² Both of these systems are based on a similar conceptual framework but are operationalised differently. The TOAST subtyping system is based on the application of written rules requiring clinician judgment; patients with conflicting potential causes are placed into an undetermined category. The CCS subtyping system uses two web-based algorithms that classify patients with conflicting potential causes. Causative (CCSc) categorisation uses historical examination and test data from each ischaemic stroke subject to assign the most probable cause in the presence of competing aetiologies, while phenotypic (CCSp) categorisation uses abnormal test findings to assign each case into one or more major groups without using rules to determine the most likely aetiology. In addition to the generation of both CCSc and CCSp subtype categories, the advantages of the CCS...
### Case-control group 1
- **Location of sample collection**: BRAINS UK 650Q, European 267, Controls...
- **Genotyping platform**: MGH-GASROS USA 610, European 111...
- **Ancestry groups**: ISGS USA 610, European 351...
- **Cases**: SWISS USA 610, European 25...
- **Controls**: HABIC USA 1M, European 1586...

### Case-control group 2
- **Edinburgh (UK)**: EDIN UK 660, European 566...
- **Munich (UK)**: MUNICH UK 660, European 1131...
- **Oxford (UK)**: OXVASC UK 660, European 457...
- **St George’s University**: STGERGE UK 660, European 418...

### Case-control group 3
- **GeoGenetics** (USA): GEOS USA 1M, African, European 843, Controls 880...

### Case-control group 4
- **Brains** (UK): BRAINS UK 5M, European, Hispanic 110...
- **MGH-GASROS** (USA): MGH-GASROS USA 5M, African, European, Hispanic 456...
- **GCNKS** (USA): GCNKS USA 5M, African, European, Hispanic 482...

### Case-control group 5
- **Kora** (Germany): KORA Germany 550, European 104...
- **Wits** (UK): WTCCC UK 660, European 258...

### Case-control group 6
- **GeoGenetics** (USA): GEOS USA 1M, African, European, Hispanic 294...
- **Nhsg** (US): NHG USA 5M, European, Hispanic 314...

### Case-control group 7
- **Spex** (Spain): SPE3 The Americas, Spain 5M, African, European, Hispanic 110...
- **Swiss** (USA): SWISS USA 5M, African, European, Hispanic 418...

### Case-control group 8
- **Hrs** (USA): HRS USA 2·5M, African, European, Hispanic 11174...
- **Oai** (USA): OAI USA 2·5M, African, European 3882...

### Case-control group 9
- **Hchs/Sol** (USA): HCHS/SOL USA 2·5M, Hispanic 1214...
- **Krakow** (Poland): Krakow Poland 5M, European, Hispanic 880...

### Case-control group 10
- **Leuven** (Belgium): Leuven Belgium 5M, European, Hispanic 460...

### Case-control group 11
- **Basicmar** (Spain): BASICMAR Spain 5M, European, Hispanic 890...
- **Adhd** (Spain): ADHD Spain 1M, European 411...
- **Inma** (Spain): INMA Spain 1M, European 807...

### Case-control group 12
- **Graz** (Austria): GRAZ Austria 610, European 915...
- **Graz** (Austria): GRAZ Austria 5M, European 609...

### Case-control group 13
- **Sahlgrenska** (Sweden): SAHLSISS Sweden 5M, European, Hispanic 783...
- **Lund** (Sweden): LUND Sweden 5M, European, Hispanic 613...

### Case-control group 14
- **Mdc** (Sweden): MDC Sweden 610, European, Hispanic 211, 1362...

### Case-control group 15
- **Asge** (Australia): ASGC Australia 610, European 1109, 1200...

### Case-control group 16
- **Visp** (USA, Canada, UK): VISP USA, Canada, UK 1M, African, European 1979...

(continued on next page)
Case-control strata then underwent extensive quality control (appendix). Finally, each stratum was phased and imputed. We imputed samples of European ancestry using a merged reference panel that included the 1000 Genomes Project Phase 1 and the Genome of the Netherlands. We imputed samples in the African and Hispanic groups using the 1000 Genomes Project Phase 1 reference panel only. We added summary-level imputed data from an additional cohort (Vitamin Intervention for Stroke Prevention) to the first stage meta-analysis.

First stage genome-wide association analysis
After quality control and imputation, 16851 cases and 32473 controls from 15 ancestry-specific groups were available for genome-wide testing (table 1, appendix). Within each stratum, we analysed all ischaemic stroke phenotypes and the four main subtypes (cardioembolism, large artery atherosclerosis, small artery occlusion, and undetermined) as established with CCSc, CCSp, and TOAST, which were available for 12612 (74.8%) cases. All GWAS were adjusted for sex and the top ten principal components; genome-wide testing was not corrected for age, because age information was missing for most of the controls.

After the GWAS, we removed SNPs with frequency of less than 1% because they showed excessive genomic inflation. We checked the frequencies of imputed SNPs for consistency with the continental populations represented in the 1000 Genomes Project Phase 1, and we removed SNPs with a difference in frequency of more than 30%. After quality control, 9-3 million to 15-4 million SNPs were available in the study strata for the meta-analysis. We did inverse variance-weighted fixed-effects meta-analysis across the 15 ancestry-specific strata using MANTEL in each of the 15 traits. The genomic inflation factor λ of the 15 meta-analyses for each trait ranged from 0-936 to 1-005 (appendix pp 5–8).

Second stage analysis
In the second stage, we performed in-silico lookups of association results in 18 independent studies that contained 20941 TOAST-subtyped cases and 364736 controls, using the nominally significant SNPs identified in the first stage (table 1 and appendix p 51). The SNPs selected for the second stage for each subtype were aggregated such that, for example, SNPs with p<1×10⁻⁶ from the three cardioembolism GWAS (CCSc, CCSp, and TOAST) were all selected for lookup in the independent TOAST cardioembolism cases and matched controls. This process was repeated for the other subtypes.

Joint analysis
We did a meta-analysis of the results from the in-silico lookups from the second stage and the results from the first stage. We set the threshold for genome-wide significance in the joint analysis at p<1×10⁻⁸, after correction for testing of the five phenotypes (all stroke, ischaemic stroke, and cerebroembolism) analyses because of overlap with CHARGE.

Table 1: Case and control cohorts in NINDS-SiGN

<table>
<thead>
<tr>
<th>Location of sample collection</th>
<th>Genotyping platform</th>
<th>Ancestry groups</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Continued from previous page)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**First stage meta-analysis**

**Second stage**

**Joint**

| Total | 38 333 | 400 315 |
includes all undetermined strokes; CCSc2 includes all incomplete and unclassified strokes; and CCSc3 includes all cryptogetic and cardioembolic minor strokes. The CCSc2 and CCSc3 classifications are mutually exclusive.

Acute Stroke Treatment classification system. GWAS=genome-wide association study.

CCS=Causative Classification of Stroke. CCSp=CCS phenotypic. TOAST=Trial of Org 10172 in Acute Stroke Treatment classification system. GWAS=genome-wide association study.

Figure 1: Genetic and phenotypic correlation between subtyping methods in the first stage analysis

All cases with an available CCS subtype were included in the first stage analyses. Genome-wide scores from the Z

Figure 1:

Genetic and phenotypic correlation between subtyping methods in the first stage analysis

All cases with an available CCS subtype were included in the first stage analyses. Genome-wide scores from the Z

Genetic correlation

Cardioembolic

Large artery atherosclerosis

Small artery occlusion

Undetermined

Phenotypic concordance

Genetic correlation

Genetic correlation

cardioembolic, large artery atherosclerosis, small artery occlusion, and undetermined). λ in the ischaemic stroke joint analysis was 1-005 and ranged from 0-936 to 0-998 in the subtype analyses (appendix pp 9–12).

Role of the funding source

The funder participated in the design of the study. The study investigators were solely responsible for the data collection, analysis, and interpretation. An employee of NINDS (KG) was a member of the writing committee. The analysis team had full access to all data included in the study. The steering committee had final responsibility for the decision to submit the report for publication.

Results

After data quality control (appendix p 4 and pp 114–26), we included 16851 stroke cases and 32473 controls in the first stage of our analyses. The first stage GWAS revealed 3172 SNPs in 268 loci associated with ischaemic stroke or a specific subtype in any of the CCS or TOAST traits at p<1×10⁻⁶. We included an additional independent set of 20941 cases and 364736 controls in the second stage, which enabled the joint analysis of 37893 cases and 397209 controls across five primary independent traits (ischaemic stroke and the four subtypes).

Genome-wide Z scores (SNP β values divided by their respective SE) from the CCSc, CCSp, and TOAST GWAS were checked for correlation (Pearson’s r) between each possible pair of traits. The analysis revealed moderate to strong genetic correlation (figure 1) between the standardised SNP effects in CCSc, CCSp, and TOAST, despite the modest phenotypic correlation noted previously.²¹ The moderate to strong genetic correlation between CCS and TOAST within subtype-specific clusters suggested that TOAST subtyping was appropriate for inclusion in the second stage of the analysis. Phenotypic correlations were also strong within subtype-specific clusters (figure 1).

In the joint analysis of CCS (first stage) and TOAST (second stage) results, SNPs in two novel loci exceeded genome-wide significance. Four common SNPs in linkage disequilibrium (r²>0-57 in the 1000 Genomes Project samples of European ancestry) near the TSPAN2 locus on chromosome 1 were associated at genome-wide significance with large artery atherosclerosis. The lead SNP in the associated locus was rs12122341 (odds ratio [OR] for the G allele 1-19, 95% CI 1-12–1-26, p=1-3×10⁻⁹; figure 2, table 2).

A second locus emerged as having a genome-wide significant association with ischaemic stroke, but only in samples of African ancestry. In view of the small sample size in which it was identified, the association must be interpreted with caution. rs74475935 in ABCCI on chromosome 16 was associated with the undetermined phenotype (table 2, appendix p 14), driven by a variant with rare frequency (minor allele frequency [MAF] about 0-01%) in European-ancestry samples and low frequency (MAF about 1-5% in African-ancestry samples).

We also identified associations for the previously reported loci PITX2 and ZFHX3 for cardioembolic stroke, and HDAC9 for large artery atherosclerotic stroke, all of which exceeded genome-wide significance in our samples (table 2). The 12q24.12 locus near ALDH2, previously reported to be associated with all ischaemic stroke, but not with any specific subtype, exceeded genome-wide significance in the joint analysis of all ischaemic stroke (OR for the T allele 1-07, 95% CI 1-01–1-14, p=4-2×10⁻⁹). However, the association was even stronger for small artery occlusion in the joint analysis of CCSp in the first stage and TOAST in the second stage (OR 1-17, 95% CI 1-11–1-23, p=2-9×10⁻⁷); the association was not genome-wide significant in the joint analysis of CCSc (first stage) and TOAST (second stage; OR 1-16, 95% CI 1-10–1-22, p=2-7×10⁻⁷). Evidence of associations with other subtypes was reduced in our study (OR<1-1 and p>4×10⁻³ for cardioembolic, large artery atherosclerosis, and undetermined in the combined CCSp and TOAST analysis; appendix p 15). Systematic testing that accounted for shared controls (appendix p 15) showed a significant difference in the magnitude of ORs between small artery occlusion and the combined non-small artery occlusion subtypes (p=0-048, appendix p 15), suggesting that the effect of 12q24.12 might be specific for small artery occlusion.

By contrast, we did not find any evidence for the previously reported association between ischaemic stroke and NINJ2 (rs34166160, OR for the A allele 1-20, 95% CI 0-96–1-48, p=0-106; table 2), even though our sample size had 100% power to detect an association (p<0-05) at this locus. In the full first stage analysis, evidence for association was weak for both the 6p212² and
CDKN2B-AS1 locus in large artery atherosclerosis, and for the ABO locus in all ischaemic stroke, large artery atherosclerosis, and cardioembolism (table 2). When we restricted our analysis to only the samples not used for the initial discovery (appendix p 52), CDKN2B-AS1 was associated with large artery atherosclerosis (OR for the G allele 1·09, 95% CI 1·02–1·17, p=0·009) and ABO was associated with all ischaemic stroke (OR for the C allele 1·05, 95% CI 1·07–1·24, p=2·5 × 10⁻⁴), large artery atherosclerosis (OR 1·15, 95% CI 1·07–1·24, p=2·5 × 10⁻⁴), and cardioembolism (OR 1·09, 95% CI 1·02–1·16, p=0·007). For 6p21, however, we detected no evidence for any association with large artery atherosclerosis (OR for the T allele 1·04, 95% CI 0·96–1·12, p=0·304).

Discussion

Our results show a novel association between a genetic locus and large artery atherosclerosis. The lead SNP, rs12122341, is located in an intergenic region 23·6 kb upstream of TSPAN2, the gene encoding tetraspanin-2 (figure 2) This SNP is in linkage disequilibrium with intronic and untranslated region variants in TSPAN2 (r²=0·03 in 1000 Genomes Project European samples), but is located in a DNA sequence immediately adjacent to TSPAN2 that can be bound by several transcription factor proteins, including CTCF. This sequence is a promoter and enhancer site that is marked by histone modification and DNase hypersensitivity according to experimental data from ENCODE and ROADMAP Epigenomics (appendix p 16), suggesting a potential role for rs12122341 in gene regulation. An intergenic SNP near rs12122341 has been reported to be associated with migraine, but the two SNPs are not in linkage disequilibrium (r²=0·03 in 1000 Genomes Project samples of European ancestry).

TSPAN2, the gene closest to rs12122341, is a member of the transmembrane 4 (tetraspanin) superfamily. This family of proteins can mediate signal transduction to regulate cell development, activation, growth, and motility. TSPAN2 knock-out mice have increased neuroinflammation, shown by activation of microglia and astrocytes with no effect on myelination and axon integrity. Notably, TSPAN2 is highly expressed in artery tissue and whole blood cells (appendix p 16), which accords with the association we detected between TSPAN2 with large artery atherosclerosis stroke. Whether the association of rs12122341 is caused by the locus’ regulation of TSPAN2 or other nearby genes will need further functional assessment.

The additional locus that we identified as being associated with undetermined stroke (rs74475935) is in a gene-rich region with linkage-disequilibrium-paired SNPs (r²>0·1 in 1000 Genomes Project samples of African ancestry) of up to 4 Mb. Because of the small sample size for rs74475935 (610 cases) and the shortage of samples from people with African ancestry, studies with large samples from people of African descent will be necessary to fully assess the robustness of this signal.

Figure 2: Forest plot (A) and regional association plot (B) of rs12122341
Plot of effect size of the association of rs12122341 with large artery atherosclerosis-related stroke across the case–control groups included in the first and second stage analyses (A). Association of rs12122341 and other SNPs in the region with large artery atherosclerosis-related stroke (B). Point shading shows linkage disequilibrium (r²) to rs12122341 as calculated in 1000 Genomes Project Phase I European samples. Purple lines show recombination rate. EUR=European ancestry. AFR=African ancestry. HIS=Hispanic samples. EAS=East Asian ancestry. SAS=south Asian ancestry.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Risk allele</th>
<th>Risk allele frequency (%)</th>
<th>Nearest gene</th>
<th>First stage</th>
<th>Second stage</th>
<th>Joint analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>European</td>
<td>African</td>
<td>The Americas</td>
<td>Subtyping system</td>
<td>Cases OR (95% CI)</td>
<td>p value</td>
</tr>
<tr>
<td>Novel loci</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large artery atherosclerosis</td>
<td>rs12122341</td>
<td>1 G</td>
<td>25·7</td>
<td>TSPAN2</td>
<td>CCSc 2454</td>
<td>1·20 (1·12–1·29)</td>
</tr>
<tr>
<td></td>
<td>rs12122341</td>
<td>1 G</td>
<td>25·7</td>
<td>TSPAN2</td>
<td>CCSp 2715</td>
<td>1·21 (1·13–1·30)</td>
</tr>
<tr>
<td></td>
<td>rs12122341</td>
<td>1 G</td>
<td>25·7</td>
<td>TSPAN2</td>
<td>TOAST 2346</td>
<td>1·15 (1·07–1·24)</td>
</tr>
<tr>
<td>Undetermined</td>
<td>rs74475935</td>
<td>16 G</td>
<td>0·2</td>
<td>ABCC1</td>
<td>CCSc 2392*</td>
<td>5·17 (2·99–8·92)</td>
</tr>
<tr>
<td></td>
<td>rs74475935</td>
<td>16 G</td>
<td>0·2</td>
<td>ABCC1</td>
<td>CCSp 1062*</td>
<td>8·68 (4·55–16·58)</td>
</tr>
<tr>
<td></td>
<td>rs74475935</td>
<td>16 G</td>
<td>0·2</td>
<td>ABCC1</td>
<td>TOAST 3593</td>
<td>2·18 (1·16–4·10)</td>
</tr>
<tr>
<td>Previously identified loci, first stage p&lt;1×10⁻⁶</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All ischaemic stroke</td>
<td>rs10744777</td>
<td>12 T</td>
<td>66·7</td>
<td>ALDH2</td>
<td>-</td>
<td>1·10 (1·06–1·13)</td>
</tr>
<tr>
<td></td>
<td>rs2634074</td>
<td>4 T</td>
<td>21</td>
<td>PITX2</td>
<td>-</td>
<td>1·10 (1·06–1·13)</td>
</tr>
<tr>
<td></td>
<td>rs2107595</td>
<td>7 A</td>
<td>15·7</td>
<td>HDAC9</td>
<td>-</td>
<td>1·10 (1·06–1·14)</td>
</tr>
<tr>
<td>Cardioembolism</td>
<td>rs2000733</td>
<td>4 T</td>
<td>12·0</td>
<td>PITX2</td>
<td>CCSc 3071</td>
<td>1·39 (1·28–1·50)</td>
</tr>
<tr>
<td></td>
<td>rs2000733</td>
<td>4 T</td>
<td>12·0</td>
<td>PITX2</td>
<td>CCSp 3695</td>
<td>1·39 (1·29–1·48)</td>
</tr>
<tr>
<td></td>
<td>rs2000733</td>
<td>4 T</td>
<td>12·0</td>
<td>PITX2</td>
<td>TOAST 3427</td>
<td>1·37 (1·27–1·48)</td>
</tr>
<tr>
<td></td>
<td>rs27193343</td>
<td>16 T</td>
<td>17·4</td>
<td>ZFHX3</td>
<td>CCSc 3071</td>
<td>1·37 (1·09–1·26)</td>
</tr>
</tbody>
</table>

(Table 2 continues on next page)
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Risk allele</th>
<th>Risk allele frequency (%)</th>
<th>Nearest gene</th>
<th>First stage</th>
<th>Second stage</th>
<th>Joint analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>European</td>
<td>African</td>
<td>The Americas</td>
<td>Subtyping</td>
<td>Subtyping</td>
<td>Subtyping</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>OR (95% CI)</td>
<td>Cases</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td></td>
<td>p value</td>
<td>p value</td>
<td>p value</td>
<td>p value</td>
</tr>
<tr>
<td>rs7193343</td>
<td>16</td>
<td>T</td>
<td>17.4</td>
<td>24</td>
<td>24</td>
<td>18.9</td>
</tr>
<tr>
<td></td>
<td>CCSp</td>
<td>3695</td>
<td>1.19</td>
<td>(1.11–1.27)</td>
<td>2.93×10⁻³</td>
<td>7.93×10⁻³</td>
</tr>
<tr>
<td></td>
<td>TOAST</td>
<td>3427</td>
<td>1.17</td>
<td>(1.09–1.25)</td>
<td>1.45×10⁻⁴</td>
<td>7.93×10⁻³</td>
</tr>
<tr>
<td>rs11984041</td>
<td>7</td>
<td>T</td>
<td>9.3</td>
<td>22</td>
<td>67</td>
<td>HDAC9</td>
</tr>
<tr>
<td></td>
<td>CCSc</td>
<td>2454</td>
<td>1.30</td>
<td>(1.18–1.42)</td>
<td>8.46×10⁻⁵</td>
<td>1.16×10⁻³</td>
</tr>
<tr>
<td></td>
<td>TOAST</td>
<td>2249</td>
<td>1.15</td>
<td>(1.03–1.29)</td>
<td>3.50×10⁻⁵</td>
<td>1.16×10⁻³</td>
</tr>
<tr>
<td>rs10744777</td>
<td>12</td>
<td>T</td>
<td>66.7</td>
<td>45</td>
<td>52</td>
<td>ALDH2</td>
</tr>
<tr>
<td></td>
<td>CCSc</td>
<td>2736</td>
<td>1.19</td>
<td>(1.11–1.27)</td>
<td>9.30×10⁻⁵</td>
<td>4.66×10⁻⁵</td>
</tr>
<tr>
<td></td>
<td>TOAST</td>
<td>2426</td>
<td>1.12</td>
<td>(1.03–1.21)</td>
<td>6.82×10⁻⁴</td>
<td>4.66×10⁻⁵</td>
</tr>
</tbody>
</table>

Previously identified loci, first stage p > 1×10⁻⁶

All ischaemic stroke

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Risk allele</th>
<th>Risk allele frequency (%)</th>
<th>Nearest gene</th>
<th>First stage</th>
<th>Second stage</th>
<th>Joint analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>European</td>
<td>African</td>
<td>The Americas</td>
<td>Subtyping</td>
<td>Subtyping</td>
<td>Subtyping</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>OR (95% CI)</td>
<td>Cases</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td></td>
<td>p value</td>
<td>p value</td>
<td>p value</td>
<td>p value</td>
</tr>
<tr>
<td>rs34166160</td>
<td>12</td>
<td>A</td>
<td>0.9</td>
<td>0.0</td>
<td>0.3</td>
<td>NINJ2</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>16 851</td>
<td>1.20</td>
<td>(0.96–1.48)</td>
<td>1.06×10⁻³</td>
<td>-</td>
</tr>
<tr>
<td>rs11833579</td>
<td>12</td>
<td>G</td>
<td>75.8</td>
<td>79.4</td>
<td>68.0</td>
<td>NINJ2</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>16 851</td>
<td>1.02</td>
<td>(0.95–1.01)</td>
<td>2.15×10⁻³</td>
<td>-</td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>16 851</td>
<td>1.07</td>
<td>(1.04–1.10)</td>
<td>2.03×10⁻³</td>
<td>-</td>
</tr>
</tbody>
</table>

(Table 2 continues on next page)
Table 2: Novel and previously identified loci implicated in ischaemic stroke and its subtypes through genome-wide testing

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Risk allele</th>
<th>Risk allele frequency (%)</th>
<th>Nearest gene</th>
<th>First stage</th>
<th>Second stage</th>
<th>Joint analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subtyping system</td>
<td>Cases</td>
<td>OR (95% CI)</td>
<td>p value</td>
<td>Subtyping system</td>
<td>Cases</td>
</tr>
<tr>
<td>Cardioembolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>Large artery atherosclerosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>rs505922</td>
<td>9</td>
<td>C</td>
<td>35.1</td>
<td>32.6</td>
<td>23.5</td>
<td>ABO</td>
</tr>
<tr>
<td>rs556621</td>
<td>6</td>
<td>T</td>
<td>29.1</td>
<td>8.1</td>
<td>40.7</td>
<td>6p21</td>
</tr>
<tr>
<td>rs556621</td>
<td>6</td>
<td>T</td>
<td>29.1</td>
<td>8.1</td>
<td>40.7</td>
<td>6p21</td>
</tr>
<tr>
<td>rs556621</td>
<td>6</td>
<td>T</td>
<td>29.1</td>
<td>8.1</td>
<td>40.7</td>
<td>6p21</td>
</tr>
<tr>
<td>rs2383207</td>
<td>9</td>
<td>G</td>
<td>49.9</td>
<td>45</td>
<td>41.3</td>
<td>CDKN2B-AS1</td>
</tr>
<tr>
<td>rs2383207</td>
<td>9</td>
<td>G</td>
<td>49.9</td>
<td>45</td>
<td>41.3</td>
<td>CDKN2B-AS1</td>
</tr>
<tr>
<td>rs2383207</td>
<td>9</td>
<td>G</td>
<td>49.9</td>
<td>45</td>
<td>41.3</td>
<td>CDKN2B-AS1</td>
</tr>
</tbody>
</table>

For subtype-specific loci, ORs and their corresponding p values are reported for the CCSc, CCSp, and TOAST subtypes. Risk allele frequency was calculated with 1000 Genomes (Phase I) European-ancestry samples, African-ancestry samples, and samples from the Americas. Association results were looked up in TOAST-subtyped cases and their matched controls and meta-analysed with the first stage results from CCSc, CCSp, and TOAST cases. CCS=Causes of Stroke Classification. CCSc=CCS causative. CCSp=CCS phenotypic. OR=odds ratio. *Results from the CCS cryptogenic phenotype.
So far, only four loci—PITX2, ZFHX3, HDAC9, and 1q24.12—have been repeatedly identified in GWAS of ischaemic stroke, all of which are subtype specific except for 1q24.12. Although the 1q locus association was originally identified for all ischaemic stroke, our analysis suggests that it is probably specific to small artery occlusion. These findings suggest that ischaemic stroke subtypes have distinct genetic signatures. Our analysis of genetic correlation across the traits also showed that the subtypes share subtle genetic associations (appendix p 17 and p 53). This finding is supported by the results of another study, which identified genetic overlap between the large artery atherosclerosis and small artery occlusion subtypes. Future efforts will help to clarify both the shared and unique genetic architectures within and between subtypes.

Until now, GWAS of ischaemic stroke subtypes have used far smaller sample sizes than studies of other complex traits. The SiGN study, the largest GWAS of ischaemic stroke so far, was well powered (75-1% to detect common SNP subtype-specific associations of larger effect (MAF 25% and OR 1-2 in 3000 cases and 30,000 controls) but was substantially less powered to identify lower frequency or lower effect SNPs (13-8% power for MAF 10% and OR 1-2; 1-1% power for MAF 25% and OR 1-1). Because of the almost linear relation that exists between sample size and discovered loci, and because large-scale GWAS in other complex traits have yielded hundreds of SNP-disease associations, studying ischaemic stroke subtypes in larger samples will probably yield additional associated common variants. Furthermore, the implementation of whole genome sequencing studies of stroke will begin to test whether rare alleles in the population account for a substantial proportion of disease heritability.

The SiGN study has several other limitations. First, sample inclusion was heavily biased towards individuals of European descent; inclusion of non-European populations will improve power for locus discovery and will be especially informative for future fine-mapping efforts. Second, the inclusion of TOAST-based classification for samples in the second stage probably added phenotypic heterogeneity (figure 1, appendix p 53), which potentially reduced power. Third, many of the participating studies within SiGN (especially the publicly available controls) had little or no stroke-specific risk factor data available. Such data are key to disentangling potential gene–environment interactions. Future genetic studies of stroke will continue to face challenges related to the disease phenotype, including high prevalence of the disease (lifetime risk about 20%), its late onset (mainly in individuals >65 years), the contribution of other cardiovascular diseases and environment as causative factors, and difficulties of subclassing (in SiGN 12·6–22·3% of all cases analysed were ultimately classified as undetermined by CCS or TOAST).

Our use of CCS enabled identification of candidate SNPs that were not significant for the second stage follow-up in TOAST, including those SNPs at the TSPAN2 locus. This refinement might represent a reduction in phenotypic heterogeneity that CCS introduces through its capture of clinical stroke features, completeness of diagnostic investigations, and, where possible, classification of cases with different potential causes into the most probable causes. The association signal of the TSPAN2 locus identified with CCS was, however, improved by the inclusion of TOAST-classified samples, suggesting that making use of the genetic correlation underlying the subtyping methods and allowing for broader inclusion of cases, regardless of subtyping system, can lead to the discovery of more susceptibility loci. Further studies will help to establish whether the rich repository of individual-level data created through the use of the CCS will help to uncover novel phenotypes and thus reveal biological mechanisms and broaden the understanding of the genetic architecture in patients with stroke.

**NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium (ISGC)**

Sara I. Pulit, Patrick F Mc Ardle, Quen na Wong, Rainer Malik, Katr ina Gwinn, Sel f an Ach terberg, Ale Algra, Philipp Amouyel, Christopher D Anderson, Donna K Arnett, Ethelm Murat Arsc w, John At tia, Hakan Ay, Traci M Bart z, Thomas Bat tye, Oscar R Bern avente, Steve Bevan, Alessandro Bif f, Joshua C Bi s, Susan H Blanton, Giorgio B Boncoraglio, Robert D Brown Jr, Annette I Burgess, C aty Carr er, Sherin a N Chapman Smith, Daniel I Chasman, G anc hao Chau han, Wei-Min Chen, Yu-Cheng Cheng, Michael Chong, Lisa K Clonna n, John W Cole, Joan a Goc tarnc u, Carlos Cruchaga, Elisa Cuadrado-God a, Tus har Dave, Jesse D wson, Stef enie Debet te, Hossein Delav a ran, Cameron A Dell, Martin D ich gan s, K imberly F Do heny, Chuan hui Dong, Da vid J Dugg an, Gunnar Engstr öm, M ich a el K Evans, Xavier Est ivil Pall eja, Jessic a D F aul, Is rael Fernández-C adenas, My ra mi F orn a ge, Philipp e M Frossard, Karen Furie, Dale M Gam ble, Christian Ge ger, Anne-K a tin Giese, Eva Gir alt-Steinh au ser, Hector M Gonz aléz, An Gor is, Solv eig Gretzardotti, Raji P Grewal, Ulrike Gtrittner, Stef an Gustafsson, Buhm Han, Graeme J Han key, Laura Heitsch, Peter Higgs, Marc C Hoch berg, Elizabeth Holliday, Jemma C Hopewell, Richard B H orenste in, George H ow ard, M Arfan Ikran, And rea Ill na, Erik Ingelshol m, Marguerite R Irving, Rebecca D Jack son, Christine J Jern, Jo rd i Jimé nez Conde, Julie A Johnson, K atarina J ood, Mu hamm ad S K ahd, Robert K aplan, J aap Kapp elle, Sharon I. R Kard ia, Keith I. K eene, Brett M Kissela, Dawn O Kleindor fer, Simon K obl ar, Daniel La bovitz, Lenore J Laun er, Cathy C L auri e, Cecelia A Laur ie, C u e Hyunkyu Lee, Jin-Moo Lee, Man uel Lehm, Rob in Lemm s, Christopher Levi, Didier Leys, Ar ne Lindgren, WT L Longstreth J r, Jane Maguire, An i Mariach ka ul, Hugh S Markus, Les li e A McClure, Ca trin W McDonn oh, Christ a Meising s, Olle Melan der, James F Meschia, Marina Mola-Caminal, Jo an Mont aner, Thomas H Mos ley, Mart ina Müll er-Nurasyid, Mike A Nalls, Jeffre y R O’Connell, Martin O’Donnell, Á ngel Ois, George J Papanicolaou, Guill aume Paré, Leema Reddy Peddareddy gari, Annie Pedersen, Joanna Pera, Annette Peters, Deborah Poole, Bruce M P tay, Raquel Rabo net, Miriam R Raffeld, Kristin na Ramming kae, Asif Rash eed, Petra Redefors, Alex P Reiner, Kath ry n Rezonde, Mar ta Ribası es, Stephen S Rich, W im Robbere cht, Ana Rodriguez-Campello, Arnld Rolfs, Jaume Ro quir, Lynda M Rose, Daniel Rosena um, Natalia S Rost, Peter M Rothwell, Tatjana Rund ek, Kathleen A Ryan, Ralph I. Sac co, Mich ele M Sale, Danish Saleehe n, Veikko Salo ma a, Cristina Sánchez-Mora, Carsten Oliver Schmid, Helena Schmidt, Reinhold Schmidt, Markus Schürk s, Rodney Scott, Helen C Segal, Stephan Seiler, S udha Seshad ri, Pankaj Sharm a, Alan R Shul d niner, Brian Silver, Ag nieszka Slowik, Jeni nefr a A Smith, Martin Söderholm, Carolina Soriano, Mary J Sparks, Tara St anne, K ari S tef annsson, O Colin Stone, Konstantin Strauch, Jonathan Sturm, Cathie LM Su d low, Sal iman M Taj addin, Robert I Talbert, Turgut Talhism a, Vincent Th ijs, Gudmar Thorleifsson, Umnur Thorst einot dott, Steffen Tiedt, Matthew Tray lor, Stella Trom pe l, Valérie Valant, Melanie Wäl denberger,

NINDS Stroke Genetics Network (SIGN) and International Stroke Genetics Consortium Writing committee Jonathan Rosand (chair), Braxton D Mitchell (co-chair), Hakan Ay, Paul IJ de Bakker, Katrina Gwinn, Steven J Knittner, Arne Lindgren, Hugh S Markus, Olle Melander, James F Meschia, Kathryn Rexrode, Jonathan Rosand, Peter M Rothwell, Tatjana Rundek, Ralph I. Sacco, Reinhold Schmidt, Pankaj Sharma, Agnieszka Sliwok, Cathie I M Sudlow, Vincent Thijs, Sylvia Wassermann-Smolker, Daniel Woo, Bradford B Worrell.

Contributors

JR, BDM, HA, PIWdB, SJ, AL, JFM, SLP, CLMS, VT, DW, and BBW contributed to data collection and provided critical review of the manuscript. JR, BDM, HA, PIWdB, KG, SJ, AL, SLP, CLMS, VT, DW, and BBW made critical decisions regarding study design and conduct. JR, BDM, HA, PIWdB, KG, SJ, AL, SLP, CLMS, VT, DW, and BBW participated in literature search and writing of the paper. BDM, PIWdB, and SLP did the statistical analysis and data interpretation.

Declaration of interests

KG is an employee of NINDS. The other members of the writing committee declare no competing interests.

Acknowledgments

The SiGN study was funded by a cooperative agreement grant from the US National Institutes of Neurological Disorders and Stroke, National Institutes of Health (U01 NS069208). The information about funding for each collection is reported in the appendix (pp 159–70).

References


