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)**

Lancet Neurol 2016; 15: 174–84

Published Online December 18, 2015 http://dx.doi.org/10.1016/S1474-4422(15)00338-5

This online publication has been corrected. The corrected version first appeared at thelancet.com/neurology on January 15, 2016

See Comment page 130

*Authors listed at end of the paper

Correspondence to: Dr Jonathan Rosand, Department of Neurology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA
jrosand@partners.org

Summary

Background The discovery of disease-associated loci through genome-wide association studies (GWAS) is the leading genetic 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 16 851 cases with state-of-the-art phenotyping data and 32 473 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 20 941 cases and 364 736 unique stroke-free controls. The ischaemic stroke subtypes of these cases had previously been established with the Trial of Org 10 172 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, $p=4·50\times10^{-8}$; joint OR 1·19, 1·12–1·26, $p=1·30\times10^{-9}$). Our results also supported robust associations with ischaemic stroke for four other loci that have been reported in previous studies, including *PITX2* (first stage OR 1·39, 1·29–1·49, $p=3·26\times10^{-19}$; joint OR 1·37, 1·30–1·45, $p=2·79\times10^{-32}$) and *ZFHX3* (first stage OR 1·19, 1·11–1·27, $p=2·93\times10^{-7}$; joint OR 1·17, 1·11–1·23, $p=2·29\times10^{-10}$) for cardioembolic stroke, and *HDAC9* (first stage OR 1·29, 1·18–1·42, $p=3·50\times10^{-8}$; joint OR 1·24, 1·15–1·33, $p=4·52\times10^{-9}$) 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 1·20, 1·12–1·28, $p=6·82\times10^{-8}$; joint OR 1·17, 1·11–1·23, $p=2·92\times10^{-9}$). 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.

Funding US National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Introduction

Worldwide, stroke is the second leading cause of death[1] 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,[3] 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,[4–7] which together account for a small proportion of the heritable risk.[8]

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 "genome 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 10 172 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)[9] 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 10 172 in Acute Stroke Treatment (TOAST)[11] subtyping data for the phenotypic classification. The appendix contains details about each collection, including their study design.

See Online for appendix

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,[9] we used two subtyping systems: the Causative Classification of Stroke (CCS) system, which is a standardised web-based subtype classification system,[12] and the more widely used TOAST subtype classification system.[11,13] 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

	Location of sample collection	Genotyping platform	Ancestry groups	Cases	Controls
First stage					
Case-control group 1					
BRAINS	UK	650Q	European	267	..
MGH-GASROS	USA	610	European	111	..
ISGS	USA	610	European	351	..
SWISS	USA	610	European	25	..
HABC	USA	1M	European	..	1586
Case-control group 2					
EDIN	UK	660	European	566	..
MUNICH	UK	660	European	1131	..
OXVASC	UK	660	European	457	..
STGEORGE	UK	660	European	418	..
KORA	Germany	550	European	..	804
WTCCC	UK	660	European	..	5150
Case-control group 3					
GEOS	USA	1M	African, European	843	880
Case-control group 4					
BRAINS	UK	5M	European, Hispanic	110	..
MGH-GASROS	USA	5M	African, European, Hispanic	456	..
GCNKSS	USA	5M	African, European, Hispanic	482	..
ISGS	USA	5M	African, European, Hispanic	178	..
MCISS	USA	5M	African, European, Hispanic	619	..
MIAMISR	USA	5M	African, European, Hispanic	294	..
NHS	USA	5M	European, Hispanic	314	..
NOMAS	USA	5M	African, European, Hispanic	358	..
REGARDS	USA	5M	African, European, Hispanic	304	..
SPS3	The Americas, Spain	5M	African, European, Hispanic	949	..
SWISS	USA	5M	African, European, Hispanic	181	..
WHI	USA	5M	African, European, Hispanic	454	..
WUSTL	USA	5M	African, European, Hispanic	449	..
HRS	USA	2·5M	African, European, Hispanic	..	11174
OAI	USA	2·5M	African, European	..	3882
HCHS/SOL	USA	2·5M	Hispanic	..	1214
Case-control group 5					
Krakow	Poland	5M	European, Hispanic	880	717
Case-control group 6					
Leuven	Belgium	5M	European, Hispanic	460	453
Case-control group 7					
BASICMAR	Spain	5M	European, Hispanic	890	..
ADHD	Spain	1M	European	..	411
INMA	Spain	1M	European	..	807
Case-control group 8					
GRAZ	Austria	610	European	..	815
GRAZ	Austria	5M	European	609	..
Case-control group 9					
SAHLSIS	Sweden	5M	European, Hispanic	783	..
LUND	Sweden	5M	European, Hispanic	613	..
MDC*	Sweden	610	European, Hispanic	211	1362
Case-control group 10					
ASGC	Australia	610	European	1109	1200
Case-control group 11					
VISP	USA, Canada, UK	1M	African, European	1979	..

(Table 1 continues on next page)

system are improved inter-observer and intra-observer reliability[12,14,15] and the ability to capture and store individual data elements from the clinical evaluation of the subject.

In the first stage of our study, each site assigned stroke subtypes with the CCS system (appendix). All of the CCS data were collected, subjected to quality control, and analysed centrally. Most sites had previously generated TOAST subtype classifications. In the second stage, we identified additional sites that had GWAS data for subtyped stroke cases. Because we included all available CCS-classified cases in the first stage, we used the corresponding subtype categories from TOAST in the second stage.

For both CCS and TOAST, each case was assigned to one of five ischaemic stroke subtypes: cardioembolic, large artery atherosclerosis, small artery occlusion, undetermined, and other. Although the classification of other was available, we did not analyse it because of low sample counts and insufficient power. In CCS, the classification undetermined was used to refer to cryptogenic cases in which no cause was identified after adequate assessment, whereas in TOAST, undetermined cases were those with incomplete assessment, more than one possible cause, and cryptogenic.

Quality control

Full details of the genotyping and quality control processes are provided in the appendix (p 4). Briefly, newly genotyped cases and about 1150 controls were genotyped on the Illumina 5M array (Illumina, San Diego, CA, USA) so we could include them in the analyses for the first stage. All other cases had been previously genotyped on various Illumina platforms (appendix). We selected publicly available external controls to match cases on the basis of ancestral background and genotyping array.

The cases and controls that were newly genotyped formed separate analysis groups (Krakow, Poland, and Leuven, Belgium; table 1). The remaining cases and controls were matched based on cohort, geographic region of the sample collection site, and genotyping platform (table 1). We assigned matched cases and controls into ancestry-specific analysis strata in two steps (appendix). We projected samples onto HapMap 3[16] data using principal component analysis to establish a group of European ancestry samples. Then, we implemented a hyper-ellipsoid clustering technique based on principal components within self-reported groups of non-Hispanic black and Asian participants. We used the hyper-ellipsoid analysis to establish a group of non-Hispanic black (African ancestry) participants and a group of participants of Asian ancestry. Samples that were not grouped as European, African, or Asian ancestry formed the Hispanic stratum. We excluded samples of Asian ancestry from further analysis because of the small number. After establishing the ancestry-based composite groups, we did principal component analysis again to confirm the ancestral homogeneity within each case-control stratum.

Case-control strata then underwent extensive quality control (appendix). Finally, each stratum was prephased[17] and imputed. We imputed samples of European ancestry using a merged reference panel that included the 1000 Genomes Project Phase I[18] and the Genome of the Netherlands.[19] We imputed samples in the African and Hispanic groups using the 1000 Genomes Project Phase I 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, 16 851 cases and 32 473 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 12 612 (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 I, 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[20] 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 20 941 TOAST-subtyped cases and 364 736 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 $p<1\times10^{-6}$ 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 $p<1\times10^{-8}$, after correction for testing of the five phenotypes (all stroke,

	Location of sample collection	Genotyping platform	Ancestry groups	Cases	Controls
(Continued from previous page)					
Melanoma Study	USA	1M	European	..	1047
HANDLs	USA	1M	African	..	971
Total	..	..	..	16 851	32 473
Second stage					
ARIC	USA	Affy 6.0	African	263	2466
CADISP†	Multi-cohort	Illumina 610	European	555	9259
CHARGE†	Multi-cohort	Multi-chip	European	3100	75 530
CHS	USA	Illumina Omni 1M	African	110	623
deCODE	Iceland	Multi-chip	European	5291	228 512
Glasgow	UK	ImmunoChip	European	599	1775
HVH	USA	Illumina 370CNV	European	577	1330
INTERSTROKE†	Multi-cohort	Cardio-metabochip	African, East Asian, European, Hispanic	1771	2103
LUND	Sweden	635	European	546	528
MDC	Sweden	5M	European	1304	3504
METASTROKE†	Multi-cohort	Multi-chip	European	1729	7925
RACE	Pakistan	660	South Asian	2385	5193
SAHLSIS	Sweden	750	European	299	596
SIFAP	Germany	2·5M	European	981	1825
SIGNET-REGARDS	USA	Affy 6.0	African	258	2094
SWISS/ISGS	USA	Illumina 610 or 660	African	173	389
UTRECHT	The Netherlands	ImmunoChip	European	556	1145
VHIR-FMT-BARCELONA	Spain	HumanCore and ExomeChip	European	545	320
WGHS‡	USA	Human Hap300 and custom array	European	440	22 725
Total	..	..	..	21 482	367 842
Joint					
Total	..	..	..	38 333	400 315

Case cohorts in the first stage were matched to external controls based on genotyping array, cohort, ancestry, and location of sample collection. Case-control groups were constructed for the first stage analyses from contributing cohorts, which were mainly case-only or control-only cohorts. Hispanic samples were an exception and are not shown as a separate group here, because the small number of samples required that we pool all available Hispanic samples into a single analysis stratum. The second stage consisted of in-silico SNP lookups of summary-level results in previously analysed case-control sets. Totals represent the number of unique samples, accounting for partial sample overlap between two sites (CHARGE and WGHS). NINDS-SiGN=National Institute of Neurological Disorders and Stroke Stroke Genetics Network. BRAINS=Biorepository of DNA in Stroke. MGH-GASROS=Massachusetts General Hospital—Genes Affecting Stroke Risk and Outcome Study. ISGS=Ischemic Stroke Genetics Study. SWISS=Siblings with Ischemic Stroke Study. HABC=Health ABC. EDIN=Edinburgh Stroke Stoke. OXVASC=Oxford Vascular Study. STGEORGE=St George's Hospital. KORA=MONICA/KORA Ausburg Study. WTCCC=Wellcome Trust Case Control Consortium. GEOS=Genetics of Early Onset Stroke. GCNKSS=Greater Cincinnati/Northern Kentucky Stroke Study. MCISS=Middlesex County Ischemic Stroke Study. MIAMISR=Miami Stroke Registry and Biorepository. NHS=Nurses' Health Study. NOMAS=Northern Manhattan Study. REGARDS=Reasons for Geographic and Racial Differences in Stroke. SPS3=Secondary Prevention of Small Subcortical Strokes. WHI=Women's Health Initiative. WUSTL=Washington University St Louis. HRS=Health and Retirement Study. OAI=Osteoarthritis Initiative. HCHS/SOL=The Hispanic Community Health Study/Study of Latinos. LEUVEN=Leuven Stroke Genetics Study. BASICMAR=Base de Datos de Ictus del Hospital del Mar. ADHD=Attention-deficit Hyperactivity Disorder. INMA=Infancia y medio ambiente. SAHLSIS=Sahlgrenska Academy Study of Ischemic Stroke. LUND=Lund Stroke Registry. MDC=Malmo Diet and Cancer Study. ASGC=Australian Stroke Genetics Collaborative. VISP=Vitamin Intervention for Stroke Prevention. HANDLs=Health/Aging in Neighborhoods of Diversity across the Lifespan Study. ARIC=Atherosclerosis Risk in Communities Study. CADISP=Cervical Artery Dissection and Ischemic Stroke Patients. CHARGE=Cohorts for Aging and Research in Genetic Epidemiology. CHS=Cardiovascular Health Study. HVH=Heart and Vascular Health Study. GLASGOW=Glasgow ImmunoChip Study. RACE=Risk Assessment of Cardiovascular Events. SIFAP=Stroke in Young Fabry Patients. SIGNET=The Sea Island Genetics Network. UTRECHT=Utrecht ImmunoChip Study/PROMISe Study. WGHS=Women's Genome Health Study. *Only TOAST subtypes available for the first stage. †Contributing cohorts are described in the appendix. ‡Not included in the ischaemic stroke and cerebroembolism analyses because of overlap with CHARGE.

Table 1: **Case and control cohorts in NINDS-SiGN**

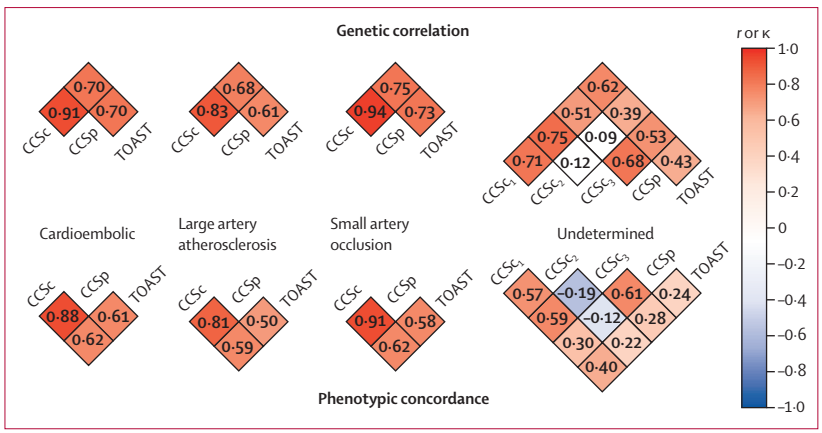

Figure 1: **Genetic and phenotypic correlation between subtyping methods in the first stage analysis**
All cases with an available CCS subtype were included in the first stage analyses. Genome-wide *Z* scores from the CCSc, CCSp, and TOAST GWAS were checked for correlation between each possible pair of traits. Pearson's *r* correlation coefficient (mathematically equivalent in this scenario to the Lin's concordance correlation coefficient) within each square shows genetic correlation. Cohen's κ within each square shows phenotypic agreement. $CCSc_1$ includes all undetermined strokes; $CCSc_2$ includes all incomplete and unclassified strokes; and $CCSc_3$ includes all cryptogenic and cardioembolic minor strokes. The $CCSc_2$ and $CCSc_3$ classifications are mutually exclusive. CCS=Causative Classification of Stroke. CCSc=CCS causative. CCSp=CCS phenotypic. TOAST=Trial of Org 10172 in Acute Stroke Treatment classification system. GWAS=genome-wide association study.

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 16 851 stroke cases and 32 473 controls in the first stage of our analyses. The first stage GWAS revealed 1372 SNPs in 268 loci associated with ischaemic stroke or a specific subtype in any of the CCS or TOAST traits at $p<1\times10^{-6}$. We included an additional independent set of 20 941 cases and 364 736 controls in the second stage, which enabled the joint analysis of 37 893 cases and 397 209 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 ($r^2>0\cdot57$ 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, $p=1\cdot3\times10^{-9}$; 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 *ABCC1* 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*[4] and *ZFHX3*[5] for cardioembolic stroke, and *HDAC9*[6] 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,[7] exceeded genome-wide significance in the joint analysis of all ischaemic stroke (OR for the T allele 1·07, 95% CI 1·5–1·09, $p=4\cdot20\times10^{-9}$). 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, $p=2\cdot92\times10^{-9}$); 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, $p=2\cdot77\times10^{-8}$). Evidence of associations with other subtypes was reduced in our study (OR<1·1 and $p>4\times10^{-3}$ for cardioembolism, 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 6p21[22] and

CDKN2B-AS1[23] loci in large artery atherosclerosis, and for the *ABO*[24] 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 1·07, 95% CI 1·03–1·10, p=$2·5\times10^{-4}$), large artery atherosclerosis (OR 1·15, 95% CI 1·07–1·24, p=$2·5\times10^{-4}$), 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* (r^2>0·3 in 1000 Genomes Project samples of European ancestry), 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 promotor and enhancer site that is marked by histone modification and DNase hypersensitivity according to experimental data from ENDCODE and ROADMAP Epigenomics (appendix p 16),[25,26] suggesting a potential role for rs12122341 in gene regulation. An intergenic SNP near rs12122341 has been reported to be associated with migraine,[27] but the two SNPs are not in linkage disequilibrium (r^2=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.[28] 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 (r^2>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.

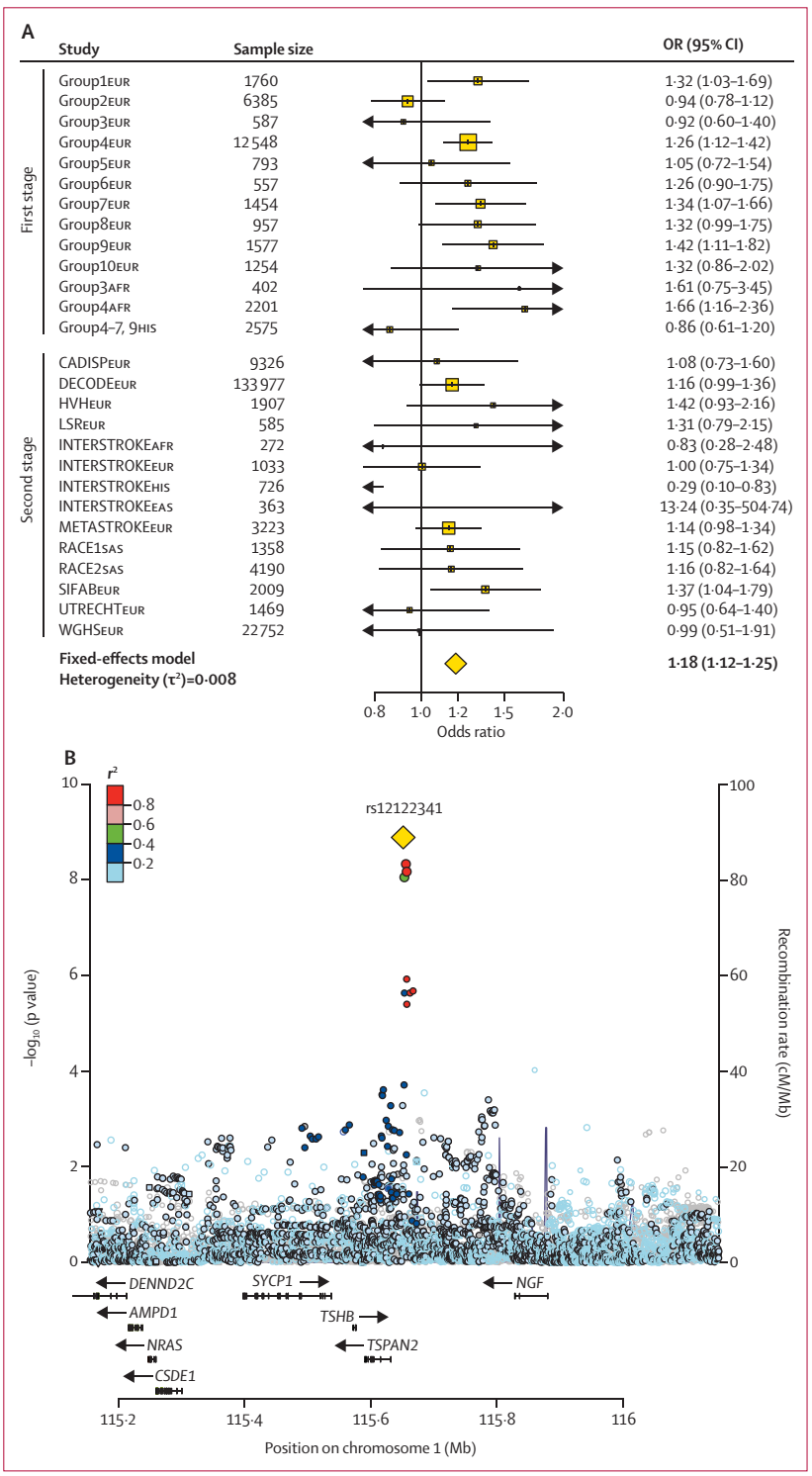

Figure 2: **Forest plot (A) and regional association plot (B) of rs12122341**
Plot of effect size of the association of rs12122341 with large artery atherosclerosis-related stroke across the case–control groups included in the fist and second stage analyses (A). Association of rs12122341 and other SNPs in the region with large artery atherosclerosis-related stroke (B). Point shading shows linkage disequilibrium (r^2) 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.

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
Novel loci																		
Large artery atherosclerosis																		
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	CCSc	2454	1·20 (1·12–1·29)	$3·38×10^{-7}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	CCSc	4703	1·18 (1·12–1·25)	$8·32×10^{-9}$
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	CCSp	2715	1·21 (1·13–1·30)	$4·50×10^{-8}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	CCSp	4964	1·19 (1·12–1·26)	$1·30×10^{-9}$
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	TOAST	2346	1·15 (1·07–1·24)	$1·61×10^{-4}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	TOAST	4595	1·15 (1·08–1·22)	$2·70×10^{-6}$
Undetermined																		
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	CCSc	2392*	5·17 (2·99–8·92)	$3·69×10^{-9}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	CCSc	5861	4·63 (2·77–7·72)	$4·70×10^{-11}$
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	CCSp	1062*	8·68 (4·55–16·58)	$5·94×10^{-11}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	CCSp	4531	6·89 (3·80–12·47)	$1·85×10^{-10}$
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	TOAST	3593	2·18 (1·16–4·10)	$1·58×10^{-2}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	TOAST	7062	2·11 (1·20–3·70)	$9·22×10^{-3}$
Previously identified loci, first stage $p<1×10^{-6}$																		
All ischaemic stroke																		
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	..	16 851	1·10 (1·06–1·13)	$3·07×10^{-8}$	..	21 042	1·05 (1·01–1·08)	$6·55×10^{-3}$		37 893	1·07 (1·5–1·09)	$4·20×10^{-9}$
rs2634074	4	T	2·1	4·8	4·1	*PITX2*	..	16 851	1·09 (1·06–1·13)	$2·56×10^{-7}$	..	21 042	1·10 (1·07–1·14)	$2·00×10^{-8}$		37 893	1·10 (1·07–1·12)	$2·68×10^{-14}$
rs2107595	7	A	15·7	2·2	2·2	*HDAC9*	..	16 851	1·10 (1·06–1·14)	$7·74×10^{-7}$	..	21 042	1·07 (1·03–1·11)	$1·70×10^{-4}$		37 893	1·09 (1·06–1·12)	$8·60×10^{-10}$
Cardioembolism																		
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	CCSc	3071	1·39 (1·28–1·50)	$1·24×10^{-16}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	CCSc	7062	1·37 (1·30–1·45)	$1·04×10^{-29}$
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	CCSp	3695	1·39 (1·29–1·49)	$3·26×10^{-19}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	CCSp	7686	1·37 (1·30–1·45)	$2·79×10^{-32}$
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	TOAST	3427	1·37 (1·27–1·48)	$1·02×10^{-16}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	TOAST	7418	1·36 (1·29–1·44)	$8·05×10^{-30}$
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	CCSc	3071	1·17 (1·09–1·26)	$1·12×10^{-5}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	CCSc	7062	1·17 (1·10–1·22)	$7·28×10^{-9}$

(Table 2 continues on next page)

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
(Continued from previous page)																		
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	CCSp	3695	1·19 (1·11–1·27)	$2·93×10^{-7}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	CCSp	7686	1·17 (1·11–1·23)	$2·29×10^{-10}$
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	TOAST	3427	1·17 (1·09–1·25)	$1·45×10^{-5}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	TOAST	7418	1·16 (1·10–1·22)	$8·88×10^{-9}$
Large artery atherosclerosis																		
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	CCSc	2454	1·30 (1·18–1·42)	$8·46×10^{-8}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	CCSc	4703	1·23 (1·15–1·33)	$1·10×10^{-8}$
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	CCSp	2715	1·29 (1·18–1·42)	$3·50×10^{-8}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	CCSp	4964	1·24 (1·15–1·33)	$4·52×10^{-9}$
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	TOAST	2346	1·30 (1·17–1·43)	$3·62×10^{-7}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	TOAST	4595	1·23 (1·14–1·33)	$4·48×10^{-8}$
Small artery occlusion																		
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	CCSc	2736	1·19 (1·11–1·27)	$9·10×10^{-7}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	CCSc	5162	1·16 (1·10–1·22)	$2·77×10^{-8}$
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	CCSp	2734	1·20 (1·12–1·28)	$6·82×10^{-8}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	CCSp	5160	1·17 (1·11–1·23)	$2·92×10^{-9}$
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	TOAST	3147	1·13 (1·06–1·21)	$1·05×10^{-4}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	TOAST	5573	1·13 (1·07–1·18)	$1·62×10^{-6}$
Previously identified loci, first stage $p>1×10^{-6}$																		
All ischaemic stroke																		
rs34166160	12	A	0·9	0·0	0·3	*NINJ2*	..	16 851	1·20 (0·96–1·48)	$1·06×10^{-1}$	..	..	..	..	..	..	..	..
rs11833579	12	G	75·8	79·4	68·0	*NINJ2*	..	16 851	1·02 (0·95–1·01)	$2·15×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	..	16 851	1·07 (1·04–1·10)	$2·03×10^{-5}$	..	..	..	..	..	..	..	..

(Table 2 continues on next page)

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
(Continued from previous page)																		
Cardioembolism																		
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSc	3071	1·04 (0·98–1·10)	$1·88×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSp	3695	1·04 (0·98–1·10)	$1·62×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	TOAST	3427	1·08 (1·02–1·15)	$5·66×10^{-3}$	..	..	..	..	..	..	..	..
Large artery atherosclerosis																		
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSc	2454	1·09 (1·02–1·17)	$6·93×10^{-3}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSp	2715	1·11 (1·04–1·18)	$1·29×10^{-3}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	TOAST	2346	1·14 (1·06–1·21)	$2·15×10^{-4}$	..	..	..	..	..	..	..	..
rs556621	6	T	29·1	8·1	40·7	6p21	CCSc	2454	1·04 (0·97–1·11)	$3·18×10^{-1}$	..	..	..	..	..	..	..	..
rs556621	6	T	29·1	8·1	40·7	6p21	CCSp	2715	1·02 (0·95–1·19)	$6·36×10^{-1}$	..							
rs556621	6	T	29·1	8·1	40·7	6p21	TOAST	2346	1·11 (1·04–1·19)	$2·55×10^{-3}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	CCSc	2454	1·12 (1·05–1·19)	$4·34×10^{-4}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	CCSp	2715	1·11 (1·05–1·19)	$7·93×10^{-4}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	TOAST	2346	1·09 (1·02–1·17)	$8·13×10^{-3}$	..	..	..	..	..	..	..	..

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=Causative Classification of Stroke. TOAST=Trial of Org 10172 in Acute Stroke Treatment classification system. CCSc=CCS causative. CCSp=CCS phenotypic. OR=odds ratio. *Results from the CCS cryptogenic phenotype.

Table 2: **Novel and previously identified loci implicated in ischaemic stroke and its subtypes through genome-wide testing**

So far, only four loci—*PITX2*,[4] *ZFHX3*,[5] *HDAC9*,[6] and 12q24.12[7]—have been repeatedly identified in GWAS of ischaemic stroke, all of which are subtype specific except for 12q24.12. Although the 12q 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.[29] 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,[30] and because large-scale GWAS in other complex traits have yielded hundreds of SNP-disease associations,[31–33] 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[34] and will be especially informative for future fine-mapping efforts.[35] 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.[21] 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 subtyping (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 L Pulit*, Patrick F McArdle*, Quenna Wong*, Rainer Malik*, Katrina Gwinn, Sefanja Achterberg, Ale Algra, Philippe Amouyel, Christopher D Anderson, Donna K Arnett, Ethem Murat Arsava, John Attia, Hakan Ay, Traci M Bartz, Thomas Battey, Oscar R Benavente, Steve Bevan, Alessandro Biffi, Joshua C Bis, Susan H Blanton, Giorgio B Boncoraglio, Robert D Brown Jr, Annette I Burgess, Caty Carrera, Sherita N Chapman Smith, Daniel I Chasman, Ganesh Chauhan, Wei-Min Chen, Yu-Ching Cheng, Michael Chong, Lisa K Cloonan, John W Cole, Ioana Cotlarciuc, Carlos Cruchaga, Elisa Cuadrado-Godia, Tushar Dave, Jesse Dawson, Stéphanie Debette, Hossein Delavaran, Cameron A Dell, Martin Dichgans, Kimberly F Doheny, Chuanhui Dong, David J Duggan, Gunnar Engström, Michele K Evans, Xavier Estivill Pallejà, Jessica D Faul, Israel Fernández-Cadenas, Myriam Fornage, Philippe M Frossard, Karen Furie, Dale M Gamble, Christian Gieger, Anne-Katrin Giese, Eva Giralt-Steinhauer, Hector M González, An Goris, Solveig Gretarsdottir, Raji P Grewal, Ulrike Grittner, Stefan Gustafsson, Buhm Han, Graeme J Hankey, Laura Heitsch, Peter Higgins, Marc C Hochberg, Elizabeth Holliday, Jemma C Hopewell, Richard B Horenstein, George Howard, M Arfan Ikram, Andreea Ilinca, Erik Ingelsson, Marguerite R Irvin, Rebecca D Jackson, Christina Jern, Jordi Jiménez Conde, Julie A Johnson, Katarina Jood, Muhammad S Kahn, Robert Kaplan, L Jaap Kappelle, Sharon L R Kardia, Keith L Keene, Brett M Kissela, Dawn O Kleindorfer, Simon Koblar, Daniel Labovitz, Lenore J Launer, Cathy C Laurie, Cecelia A Laurie, Cue Hyunkyu Lee, Jin-Moo Lee, Manuel Lehm, Robin Lemmens, Christopher Levi, Didier Leys, Arne Lindgren, W T Longstreth Jr, Jane Maguire, Ani Manichaikul, Hugh S Markus, Leslie A McClure, Caitrin W McDonough, Christa Meisinger, Olle Melander, James F Meschia, Marina Mola-Caminal, Joan Montaner, Thomas H Mosley, Martina Müller-Nurasyid, Mike A Nalls, Jeffrey R O'Connell, Martin O'Donnell, Ángel Ois, George J Papanicolaou, Guillaume Paré, Leema Reddy Peddareddygari, Annie Pedersén, Joanna Pera, Annette Peters, Deborah Poole, Bruce M Psaty, Raquel Rabionet, Miriam R Raffeld, Kristiina Rannikmäe, Asif Rasheed, Petra Redfors, Alex P Reiner, Kathryn Rexrode, Marta Ribasés, Stephen S Rich, Wim Robberecht, Ana Rodriguez-Campello, Arndt Rolfs, Jaume Roquer, Lynda M Rose, Daniel Rosenbaum, Natalia S Rost, Peter M Rothwell, Tatjana Rundek, Kathleen A Ryan, Ralph L Sacco, Michèle M Sale, Danish Saleheen, Veikko Salomaa, Cristina Sánchez-Mora, Carsten Oliver Schmidt, Helena Schmidt, Reinhold Schmidt, Markus Schürks, Rodney Scott, Helen C Segal, Stephan Seiler, Sudha Seshadri, Pankaj Sharma, Alan R Shuldiner, Brian Silver, Agnieszka Slowik, Jennifer A Smith, Martin Söderholm, Carolina Soriano, Mary J Sparks, Tara Stanne, Kari Stefansson, O Colin Stine, Konstantin Strauch, Jonathan Sturm, Cathie LM Sudlow, Salman M Tajuddin, Robert L Talbert, Turgut Tatlisumak, Vincent Thijs, Gudmar Thorleifsson, Unnur Thorsteindottir, Steffen Tiedt, Matthew Traylor, Stella Trompet, Valerie Valant, Melanie Waldenberger,

Matthew Walters, Liyong Wang, Sylvia Wassertheil-Smoller, David R Weir, Kerri L Wiggins, Stephen R Williams, Dorota Wloch-Kopec, Daniel Woo, Rebecca Woodfield, Ona Wu, Huichun Xu, Alan B Zonderman, Australian Stroke Genetics Consortium, Cervical Artery Dissection and Ischemic Stroke Patients (CADISP) study, Cohorts of Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium, Consortium of Minority Population genome-wide Association Studies of Stroke (COMPASS), METASTROKE consortium, Wellcome Trust Case-Control Consortium, Bradford B Worrall*, Paul IW de Bakker*, Steven J Kittner*, Braxton D Mitchell*, Jonathan Rosand*. *Contributed equally.
NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium Writing committee Jonathan Rosand (chair), Braxton D Mitchell (co-chair), Hakan Ay, Paul I W de Bakker, Katrina Gwinn, Steven J Kittner, Arne Lindgren, James F Meschia, Sara L Pulit, Cathie L M Sudlow, Vincent Thijs, Daniel Woo, Bradford B Worrall. *Steering committee* Donna K Arnett, Oscar Benavente, John W Cole, Martin Dichgans, Raji P Grewal, Christina Jern, Jordi Jiménez Conde, Julie A Johnson, Steven J Kittner, Jin-Moo Lee, Christopher Levi, Arne Lindgren, Hugh S Markus, Olle Melander, James F Meschia, Kathryn Rexrode, Jonathan Rosand, Peter M Rothwell, Tatjana Rundek, Ralph L Sacco, Reinhold Schmidt, Pankaj Sharma, Agnieszka Slowik, Cathie L M Sudlow, Vincent Thijs, Sylvia Wasssertheil-Smoller, Daniel Woo, Bradford B Worrall.

Contributors

JR, BDM, HA, PIWdB, SJK, AL, JFM, SLP, CLMS, VT, DW, and BBW contributed to data collection and provided critical review of the manuscript. JR, BDM, HA, PIWdB, KG, SJK, AL, SLP, CLMS, VT, DW, and BBW made critical decisions regarding study design and conduct. JR, BDM, HA, PIWdB, KG, SJK, 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 Institute 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

1 World Health Organization. The top 10 causes of death. May 2014. http://www.who.int/mediacentre/factsheets/fs310/en/ (accessed April 23, 2015).
2 Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation* 2014; **131:** e29–322.
3 Meschia JF, Bushnell C, Boden-Albala B, et al. Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2014; **45:** 3754–832.
4 Gretarsdottir S, Thorleifsson G, Manolescu A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* 2008; **64:** 402–09.
5 Gudbjartsson DF, Holm H, Gretarsdottir S, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet* 2009; **41:** 876–78.
6 Bellenguez C, Bevan S, Gschwendtner A, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* 2012; **44:** 328–33.
7 Kilarski LL, Achterberg S, Devan WJ, et al. Meta-analysis in more than 17,900 cases of ischemic stroke reveals a novel association at 12q24.12. *Neurology* 2014; **83:** 678–85.
8 Bevan S, Traylor M, Adib-Samii P, et al. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke* 2012; **43:** 3161–67.
9 Meschia JF, Arnett DK, Ay H, et al. Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke* 2013; **44:** 2694–702.
10 Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38:** 209–13.
11 Adams H, Bendixen B, Kappelle L, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* 1993; **24:** 35–41.
12 Ay H, Benner T, Arsava EM, et al. A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke* 2007; **38:** 2979–84.
13 Kolominsky-Rabas PL, Weber M, Gefeller O, Neundoerfer B, Heuschmann PU. Epidemiology of ischemic stroke subtypes according to TOAST criteria: incidence, recurrence, and long-term survival in ischemic stroke subtypes: a population-based study. *Stroke* 2001; **32:** 2735–40.
14 Arsava EM, Ballabio E, Benner T, et al. The causative classification of stroke system: an international reliability and optimization study. *Neurology* 2010; **75:** 1277–84.
15 Ay H, Arsava EM, Andsberg G, et al. Pathogenic ischemic stroke phenotypes in the NINDS-Stroke Genetics Network. *Stroke* 2014; **45:** 3589–96.
16 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38:** 904–09.
17 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012; **44:** 955–59.
18 Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491:** 56–65.
19 Francioli LC, Menelaou A, Pulit SL, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; **46:** 818–25.
20 de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17:** R122–28.
21 McArdle PF, Kittner SJ, Ay H, et al. Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. *Neurology* 2014; **83:** 1653–60.
22 Holliday EG, Maguire JM, Evans T-J, et al. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet* 2012; **44:** 1147–51.
23 Smith JG, Melander O, Lövkvist H, et al. Common genetic variants on chromosome 9p21 confers risk of ischemic stroke a large-scale genetic association study. *Circ Cardiovasc Genet* 2009; **2:** 159–64.
24 Williams FMK, Carter AM, Hysi PG, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. *Ann Neurol* 2013; **73:** 16–31.
25 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518:** 317–30.
26 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012; **40:** D930–34.
27 Esserlind AL, Christensen AF, Le H, et al. Replication and meta-analysis of common variants identifies a genome-wide significant locus in migraine. *Eur J Neurol* 2013; **20:** 765–72.
28 de Monasterio-Schrader P, Patzig J, Möbius W, et al. Uncoupling of neuroinflammation from axonal degeneration in mice lacking the myelin protein tetraspanin-2. *Glia* 2013; **61:** 1832–47.
29 Holliday EG, Traylor M, Malik R, et al. Genetic overlap between diagnostic subtypes of ischemic stroke. *Stroke* 2015; **46:** 615–19.
30 Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet* 2012; **90:** 7–24.
31 Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511:** 421–27.
32 Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46:** 1173–86.
33 Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; **45:** 1274–83.
34 Pulit SL, Voight BF, de Bakker PIW. Multiethnic genetic association studies improve power for locus discovery. *PLoS One* 2010; **5:** e12600.
35 Zaitlen N, Paşaniuc B, Gur T, Ziv E, Halperin E. Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet* 2010; **86:** 23–33.