Common variants at 6p21.1 are associated with large artery atherosclerotic stroke

Elizabeth G. Holliday [1,2], Jane M. Maguire [3,4,5], Tiffany-Jane Evans [2,6], Simon Koblar [7,8], Jim Jannes [7,8], Jonathan W. Sturm [5,9,10], Graeme J. Hankey [11,12], Ross Baker [11,13] , Jonathan Golledge [14,15], Mark W. Parsons [4], Rainer Malik [16], Mark McEvoy [1,17], Erik Biros [14], Martin D. Lewis [7,18], Lisa F. Lincz [6,17,19], Roseanne Peel [1,17], Christopher Oldmeadow [17], Wayne Smith [17], Pablo Moscato [20], Simona Barlera [21], Steve Bevan [22], Joshua C. Bis [23], Eric Boerwinkle [24], Giorgio B. Boncoraglio [25], Thomas G. Brott [26], Robert D. Brown, Jr [27], Yu-Ching Cheng [28], John W. Cole [29], Ioana Cotlarciuc [30], William J. Devan [31,32], Myriam Fornage [24], Karen L. Furie [32], Sólveig Grétarsdóttir [33], Andreas Gschwendtner [16], Mohammad Arfan Ikram [34,35,36], W. T. Longstreth, Jr [37,38,39], James F. Meschia [26], Braxton D. Mitchell [28], Thomas H. Mosley [40], Michael A. Nalls [41], Eugenio A. Parati [25], Bruce M. Psaty [23,37,42,43], Pankaj Sharma [30], Kari Stefansson [33,44], Gudmar Thorleifsson [33], Unnur Thorsteinsdottir [33,44], Matthew Traylor [22], Benjamin F.J. Verhaaren [34,36], Kerri L. Wiggins [23], Bradford B. Worrall [45], The Australian Stroke Genetics Collaborative [46], The International Stroke Genetics Consortium [46], The Wellcome Trust Case Control Consortium 2 [46], Cathie Sudlow [47], Peter M. Rothwell [48], Martin Farrall [49,50], Martin Dichgans [16], Jonathan Rosand [31,32], Hugh S. Markus [22], Rodney J. Scott [2,6,51*], Christopher Levi [4*], John Attia [1,2*]

1 Centre for Clinical Epidemiology and Biostatistics, School of Medicine and Public Health, University of Newcastle, New South Wales, Australia
2 Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, Hunter Medical Research Institute, Newcastle, New South Wales, Australia
3 School of Nursing and Midwifery, University of Newcastle, New South Wales, Australia
4 Centre for Brain and Mental Health Research, University of Newcastle and Hunter Medical Research Institute, New South Wales, Australia
5 Department of Neurosciences, Gosford Hospital, Central Coast Area Health, New South Wales, Australia
6 School of Biomedical Sciences and Pharmacy, University of Newcastle, New South Wales, Australia
7 Stroke Research Program, School of Medicine, University of Adelaide, South Australia, Australia
8 Stroke Unit, Department of Neurology, Queen Elizabeth Hospital, Adelaide, Australia
9 Florey Neurosciences Research Institute , Melbourne, Australia
10 School of Medicine and Public Health, University of Newcastle, New South Wales, Australia
11 Royal Perth Hospital, Perth, Western Australia, Australia
12 School of Medicine and Pharmacology, University of Western Australia, Australia
13 Centre for Thrombosis and Haemophilia, Murdoch University, Perth
14 Vascular Biology Unit, School of Medicine and Dentistry, James Cook University, Townsville, Queensland, Australia
15 Department of Vascular Surgery, The Townsville Hospital, Townsville, Queensland, Australia
16 Institute for Stroke and Dementia Research [ISD], Medical Center, Klinikum der Universität München, Ludwig-Maximilians-University, Munich, Germany
17 Hunter Medical Research Institute and University of Newcastle, New South Wales, Australia
18 Discipline of Genetics, School of Molecular & Biomedical Sciences, University of Adelaide, South Australia
19 Hunter Haematology Research Group, Calvary Mater Newcastle Hospital, Newcastle, Australia

20 School of Electrical Engineering and Computer Science, University of Newcastle, New South Wales, Australia
21 Department of Cardiovascular Research, Istituto di Ricerche Farmacologiche “Mario Negri”, Milano, Italy
22 Stroke and Dementia Research Centre, St George's University of London, London, United Kingdom
23 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA
24 Institute of Molecular Medicine and Human Genetics Center, the University of Texas Health Science Center at Houston, Houston, Texas
25 Department of Cerebrovascular Diseases, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy
26 Department of Neurology, Mayo Clinic, Florida, USA
27 Mayo Clinic, Rochester, MN, USA
28 Department of Medicine, University of Maryland, Baltimore, MD, USA
29 Baltimore Veterans Affairs Medical Center and University of Maryland School of Medicine, MD, USA
30 Imperial College Cerebrovascular Research Unit (ICCRU), Imperial College, London, UK
31 Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
32 Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA and Department of Neurology, Harvard Medical School, Boston, MA, USA
33 deCODE Genetics, Sturlugata 8, IS-101 Reykjavik, Iceland
34 Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
35 Department of Neurology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
36 Department of Radiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
37 Department of Epidemiology, University of Washington, Seattle, WA, USA
38 Department of Medicine, University of Washington, Seattle, WA, USA
39 Department of Neurology, University of Washington, Seattle, WA, USA
40 University of Mississippi Medical Center, Department of Medicine, Jackson, Mississippi, USA
41 National Institute on Ageing, National Institute of Health, Bethesda, MD USA
42 Department of Health Services, University of Washington, Seattle, WA, USA
43 Group Health Research Institute, Group Health, Seattle, WA, USA
44 Faculty of Medicine, University of Iceland, Reykjavik, Iceland
45 Department of Neurology, University of Virginia, Charlottesville, USA
46 Members are listed in a Supplementary Note
47 Division of Clinical Neurosciences, University of Edinburgh, Edinburgh, United Kingdom
48 University Department of Clinical Neurology, John Radcliffe Hospital, Oxford, United Kingdom
49 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
50 Department of Cardiovascular Medicine, University of Oxford, Oxford, UK
51 Division of Genetics, Hunter Area Pathology Service, Newcastle, New South Wales, Australia

*These authors jointly directed this work

Corresponding (submitting) author:

Dr Elizabeth Holliday

Hunter Medical Research Institute

Locked Bag 1, HRMC NSW 2310, Australia

Tel: +61 (0)2 49138823

Fax: +61 (0)2 49138148

Email: Liz.Holliday@newcastle.edu.au

Genome-wide association studies (GWAS) have not consistently detected replicable genetic risk factors for ischemic stroke, potentially due to etiological heterogeneity of this trait. We performed GWAS of ischemic stroke and a major ischemic stroke subtype (large artery atherosclerosis: LAA) using 1162 ischemic stroke cases (including 421 LAA cases) and 1244 population controls from Australia. Evidence for a genetic influence on ischemic stroke risk was detected, but this was higher and more significant for the LAA subtype. We identified a novel LAA susceptibility locus on chromosome 6p21.1 (rs556621: OR=1.62, $P=3.9\times10^{-8}$) and replicated this in 1,715 LAA cases and 52,695 population controls from ten independent population cohorts (meta-analysis replication OR=1.15, $P=3.9\times10^{-4}$; discovery and replication combined OR=1.21, $P=4.7\times10^{-8}$). This study suggests a genetic risk locus for LAA and supports the analysis of etiological subtypes to better identify genetic risk alleles for ischemic stroke.

Stroke affects approximately 15 million persons worldwide each year [1] and is a leading cause of death and adult acquired disability [2,3]. The vast majority of strokes are ischemic, involving cerebral artery blockage by atherosclerotic plaque or embolus. While clinical risk factors for ischemic stroke are well established [4], the genetic risk alleles are incompletely identified. Genetic influences on stroke risk are supported, however, by higher concordance among monozygotic than dizygotic twins [5], increased risk among family members of affected individuals [6] and high heritability of intermediate predictors including carotid intima-media thickness (IMT: h^2≈30-60% [7,8]) and white matter lesions (h^2≈50-70% [9,10]).

With the exception of the 4q25 locus associated with atrial fibrillation and ischemic stroke [11,12], the 9p21 region associated with coronary artery disease and ischemic stroke [13,14], and a recently described 7p21.1 association with LAA [16], genome-wide association studies (GWAS) for ischemic stroke have identified few convincingly associated variants. Inability to replicate many reported associations may be attributable to phenotypic heterogeneity, a challenge that could be partly addressed by more complete subtyping of ischemic stroke etiology. At least three major ischemic stroke etiological types are commonly distinguished: 1) large artery atherosclerosis (LAA); 2) cardioembolism (CE) and; 3) small vessel occlusion (SVO) [15]. Genetic heterogeneity may contribute to this phenotypic diversity; a recent, well-powered GWAS of ischemic stroke detected heterogeneity of risk locus effects across stroke subtypes [16] and family studies have also identified differences in subtype heritability, owing perhaps to variable roles of heritable intermediate phenotypes such as hypertension and large vessel atherosclerosis [17]. The greatest familial risk has been associated with LAA, for which family history confers significant risk even beyond the seventh decade of life [6].

We conducted a GWAS of ischemic stroke in an Australian sample of European ancestry involving 1230 cases and 1280 population controls. The causal subtype of ischemic stroke was classified using TOAST criteria [15]. Demographic and clinical characteristics of the ASGC dataset are summarised in **Supplementary Table 1**.

After quality control of genotype data, data on 551, 514 SNPs from 1162 ischemic stroke cases and 1244 controls were used for genotype imputation and genetic analysis. Prior to GWAS, we assessed the genetic contribution to ischemic stroke and the LAA, CE and SVO subtypes using a recent method [18] which estimates the proportion of phenotypic variance (V_g/V_p) attributable to variation in genotyped SNPs. For ischemic stroke, the estimated genetic load was substantial (V_g/V_{p_IS} = 0.39), with SNPs explaining a significant proportion of phenotypic variation (P=4.5×10^{-4}). For cases with the LAA subtype, we observed a higher, more significant estimate of genetic load (V_g/V_{p_LAA} = 0.66; P=5.6×10^{-5}), consistent with previous reports of high familial risk for LAA [6]. Evidence for genetic contribution was less significant for the CE and SVO subtypes (V_g/V_{p_CE} = 0.6, P=0.0026 and V_g/V_{p_SVO} = 0.1, P=0.33, respectively: see **Table 1**).

We performed two primary genome-wide association analyses (GWAS) in the Australian discovery sample, comparing (i) all ischemic stroke cases (n=1162) and; (ii) LAA cases (n=421) with population controls (n=1244). GWAS of the CE and SVO subtypes, which had both fewer cases and a less significant V_g/V_p estimate, were performed as supplementary analyses (see **Supplementary Tables 2-3, Supplementary Figures 1-2**). Genotype effects were estimated using logistic regression models (1 degree of freedom additive trend test) adjusted for age and sex. Results were compared with a pre-specified significance threshold of 5x10^{-8}, corresponding to Bonferroni adjustment for 10^6 independent tests. Q-Q plots (**Supplementary Figure 3**) indicated excellent quality of the GWAS data and an absence of systematic bias by population sub-structure or other artefacts.

Analyses of ischemic stroke detected the strongest signals at several SNPs within the *SLC5A4* gene on chromosome 22q12.3 (see **Fig. 1, Supplementary Figure 4, Supplementary Table 4**). Peak association was detected at rs5998322 (P_{trend}=3.91×10^{-7}, OR=1.97, 95% CI 1.51 – 2.57) within exon 11. A strong signal was also detected 4 Mb downstream of this peak at a number of SNPs located within and upstream of the *APOL2* gene (peak association at rs4479522, P_{trend}=3.23×10^{-6}, OR=1.34, 95% CI 1.18 – 1.51). Analysis of rs5998322 adjusted for allele dosage at rs4479522 produced similar results to the unadjusted analysis (P_{trend}=4.47×10^{-7}), suggesting independence of the two associated 22q loci.

The GWAS of LAA detected two associated SNPs on chromosome 6p21.1 exceeding the pre-specified threshold for genome-wide significance (α=5×10^{-8}; see **Figures 1-2**). These variants, rs556621 (P_{trend}=3.92×10^{-8}, OR: A allele=1.62, 95% CI 1.36 – 1.93) and rs556512 (P_{trend}=4.25×10^{-8}, OR: A allele=1.62, 95% CI 1.36 – 1.93) are in perfect linkage disequilibrium (LD) in HapMap Phase II CEU data (r^2=1, D'=1)(see **Supplementary Table 5**), with a minor (A) allele population frequency of 0.33. SNP rs556621 was directly genotyped in our sample while rs556512 was imputed with excellent reliability (imputation r^2=0.99). Very similar effect sizes for rs556621 were estimated in logistic models further adjusted for the first ten ancestry principal components and several correlated clinical risk factors (see **Supplementary Table 6**), indicating a lack of confounding by

population substructure or clinically-related heritable traits. Consistent, but attenuated association of the 6p21.1 variants was observed for the broad ischemic stroke phenotype, with peak association also detected at rs556621 (P=5.6×10^{-5}, OR: A allele=1.29, 95% CI 1.14 – 1.47)(see **Table 2**). Supplementary analyses of CE and SVO revealed no association with rs556621 (P=0.73 and P=0.39, respectively). In addition to the 6p21.1 locus the LAA GWAS also detected clusters of suggestively associated SNPs (P<1×10^{-5}) at 14q32.33 and the second 22q12.3 locus detected in the GWAS of ischemic stroke (see **Supplementary Table 7, Supplementary Figure 5**).

In a subsequent LAA GWAS adjusted for rs556621 genotype, no SNP showed evidence of strong, independent association with LAA (peak P=5.6×10^{-6} for rs11625862 at 14q32.33). Haplotype association tests across the 6p21.1 region also failed to detect multi-marker haplotypes that were more strongly associated with LAA than the two index SNPs (results not shown).

The addition of rs556621 genotypes to a risk-prediction model containing various clinical traits associated with LAA occurrence produced a small, but significant increase in the area under the receiver-operator characteristic curve (ΔAUC=0.01; P=1.2×10^{-5} [see **Supplementary Table 8]**), although this ΔAUC estimate may be inflated by estimation in the discovery cohort. To further assess internal validity of the association at rs556621, the sample was randomly partitioned into training and test groups containing 2/3 and 1/3 of LAA cases and controls, respectively. Association with LAA was evaluated in the training set, with genotyped SNPs reaching P<1×10^{-4} (n=44) then assessed for association in the test set (remaining 1/3 of the sample). The index 6p21.1 SNP (rs556621) reached P= 5.69×10^{-5} in the training set and was the only SNP associated with LAA in the independent test set after permutation-based adjustment for testing 44 non-independent SNPs (family-wise adjusted P=6.74×10^{-3})(see **Supplementary Table 9**).

External validity of the observed association of rs556621 with LAA risk was assessed in a replication study involving ten (10) independent population cohorts contributing 1,715 LAA cases (1,323 European and 392 US) and 52,695 controls (39,509 European and 13,186 US) of confirmed European ancestry. Details of the individual cohorts are provided in the **Supplementary Note** and **Supplementary Table 10**. Association analyses for the index 6p21.1 SNP (rs556621) were performed separately within each of the 10 cohorts, with the results combined using fixed effects, inverse variance-weighted meta-analysis. Because association evidence was assessed for a single SNP in the independent replication study, no multiple testing adjustment was indicated and the result was compared with a pre-specified significance threshold of 0.05.

The replication study confirmed association of rs556621 with LAA (P_{trend}=3.9×10^{-4}, OR: A allele=1.15, 95% CI 1.06 – 1.24), with no evidence of between-study heterogeneity (P=0.50, I^2=0.0%) (see **Figure 3, Table 2** and **Supplementary Table 11**). The estimated population attributable risk for rs556621 in the replication study

was ~5%. When the discovery and replication cohorts were combined, meta-analyses yielded P_{trend}=4.7x10^{-8} (OR = 1.21, 95% CI 1.13 – 1.30). However the heterogeneity statistic for the combined analysis was moderately significant (P=0.02, I^2=43.4%,), indicating some inflation of the effect size in the discovery cohort ('Winner's Curse'). For this reason, the estimated effect in the independent replication study is likely a better estimate of the true population effect. Meta-analyses of rs556621 for overall ischemic stroke in the replication study showed no evidence for association, despite a greater than 5-fold increase in case numbers (9,552 cases, 52,695 controls: P_{trend}=0.29, OR: A allele=1.02, 95% CI 0.98 – 1.06)(see **Supplementary Figure 6**). These results support the existence of a common 6p21.1 risk variant of modest but genuine effect specific to the LAA stroke subtype. Neither this SNP, nor SNPs in high LD with rs556621, have previously been reported to be associated with coronary heart disease risk.

Genomically, the 6p21.1 SNPs are located in an intergenic region of moderate LD (see **Supplementary Figure 7**), ~200 kb upstream of the *SUPT3H* gene (forward strand) and ~180 kb upstream of *CDC5L* (reverse strand). SNPs rs556621 and rs556512 both lie within a small length of genomic sequence containing *BCL3* and *Pbx3* transcription factor binding motifs and enriched for enhancer/promoter-associated marks of histone protein modification. The associated SNPs or other, correlated variants may thus function in regulating gene expression, via altered responsiveness of key transcription factor binding sites [19]. A number of predicted microRNAs (miRNAs) also lie in the vicinity of rs556621 (see **Supplementary Table 12**), suggesting that variants in LD with rs556621 could also potentially regulate gene expression via regulatory miRNA sequence alteration. Queries of four public eQTL databases (see **Supplementary Note**) did not identify rs556621, or proxy SNPs in high LD with rs556621 as *cis* eQTL in the assayed tissue/cell types. Future, targeted investigations in atherosclerotic neurovascular tissue may help to elucidate the mechanisms by which the associated SNPs influence LAA risk.

Suggestive association with both ischemic stroke and LAA was also detected for variants in a chromosome 22q12.3 region containing the *APOL1-APOL4* gene cluster. These primate-specific genes are implicated in lipid metabolism and vascular biology [20,21], where their expression is strongly induced by pro-inflammatory cytokines [22-24]. *APOL2-APOL4* are thought to encode intracellular proteins; *APOL2*, across which association evidence was strongest, is almost exclusively expressed in the brain, with reduced expression in the heart [23].

This is one of the first reported GWAS for large artery atherosclerosis, a major subtype of ischemic stroke. We have detailed the discovery and replication of chromosome 6p21.1 variants that associate with LAA risk in European-ancestry individuals. We also report a locus within the *APOL1-APOL4* gene cluster that is suggestively associated with both LAA and broad ischemic stroke. The potential pathological function of these variants, and their contribution to stroke risk in non-European populations, remains to be determined.

URLs

MACH: http://www.sph.umich.edu/csg/yli/mach/index.html

Haploview: http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview

Unphased: http://homepages.lshtm.ac.uk/frankdudbridge/software/unphased/

LocusZoom: http://csg.sph.umich.edu/locuszoom/

Metal: http://www.sph.umich.edu/csg/abecasis/Metal/

Acknowledgments

A complete list of funding acknowledgments is included in the Supplementary Note. We are grateful to the participants with ischemic stroke and also their families for participating in this study. Australian population control data was derived from the Hunter Community Study. We also thank the University of Newcastle for funding and the men and women of the Hunter region who participated in this study. This research was funded by grants from the Australian National and Medical Health Research Council (NHMRC Project Grant ID: 569257), the Australian National Heart Foundation (NHF Project Grant ID: G 04S 1623), the University of Newcastle, the Gladys M Brawn Fellowship scheme and the Vincent Fairfax Family Foundation in Australia. EGH is supported by the Australian NHMRC Fellowship scheme. JG is supported by a Practitioner Fellowship from the NHMRC and a Senior Clinical Research Fellowship from the Australian Office of Health and Medical Research. The principal funding for the WTCCC2 ischemic stroke study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA). This work was also supported by the European Community Sixth Framework Program [LSHM-CT-2007-037273], the Wellcome Trust core award [090532/Z/09/Z] and AstraZeneca. MF is a member of the Oxford BHF Centre of Research Excellence. The Siblings with Ischemic Stroke Study (SWISS) and the Ischemic Stroke Genetics Study (ISGS) were funded by grants from the National Institute of Neurological Disorders and Stroke (US). Additional funding was provided by the U01NS069208 from the US National Institute of Neurological Disorders And Stroke. The Rotterdam Study received principal funding for this report from the Netherlands Heart Foundation (grant number 2009B102).

Author contributions

SK, JS, LL, PM, RJS, CL and JA designed the study. EGH performed statistical analyses in the discovery cohort, meta-analyses of replication data and wrote the first draft of the manuscript. TJE and RJS co-ordinated genotyping of the discovery cohort. JM, JG, JJ, GJH, RB, MWP, JWS, LL, CL, MM, RP, WS, and JA performed phenotype collection and data management in the Australian sample. Replication data were provided by SB, SB, JB, EB, GBB, TB, RDB, YC, JWC, IC, WJD, MF, KLF, SG, AG, MAI, WTL, RM, JFM, BDM, THM, MAN, EAP, BMP, PS, KS, GT, MT, UT, BFJV, KLW, BBW, CS, PMR, MD, JR and HM. EB, MDL and CO undertook bioinformatic analyses & searches. CO contributed to figure preparation. All authors critically reviewed the manuscript and gave advice on the contents of the paper.

Figure Legends

Figure 1 Genome-wide association results for a) ischemic stroke and b) LAA. The plots show $-\log_{10}$-transformed *P*-values for genotyped and imputed SNPs with respect to their physical position. The threshold for genome-wide significant association ($P=5\times10^{-8}$) is shown as the upper dashed line.

Figure 2 Regional association results for the chromosome 6p21.1 locus showing genome-wide significant association with LAA. The index, associated SNP is labelled (rs556621: $P=3.9\times10^{-8}$).

Figure 3 Forest plot showing association of rs556621 with large artery atherosclerotic stroke (LAA) across the ten replication cohorts. For each cohort, the square and horizontal line show the estimated odds ratio (OR) and 95% confidence interval, respectively, representing the effect of each additional copy of the risk (A) allele upon the odds of disease. The size of the square is inversely proportional to the standard error of the estimated allelic effect. An inverse variance-weighted fixed effects meta-analysis was used to combine association evidence across cohorts. There was no evidence of effect size heterogeneity across the ten cohorts (*P*-value=0.5).

Table 1 Proportion of case-control phenotypic variation explained by genome-wide SNP data[a] for all ischemic stroke, large artery atherosclerosis, small vessel occlusion and cardioembolism.

Phenotype	***cases***[b]	***controls***[b]	σ^2_g/σ^2_p **(s.e)**[c]	***LRT***[d]	***P-value***[e]
Ischemic stroke	1079	1172	0.39 (0.15)	536.95	4.5×10^{-4}
Large artery atherosclerosis (LAA)	400	1172	0.66 (0.21)	613.73	5.6×10^{-5}
Small vessel occlusion (SVO)	288	1172	0.10 (0.24)	653.58	3.3×10^{-1}
Cardioembolism (CE)	226	1172	0.60 (0.25)	808.62	2.6×10^{-3}

[a] Genetic relationships between individuals were estimated using 457,533 SNPs. [b]Smaller sample sizes compared with the GWAS owe to additional QC conducted prior to this analysis. [c]Estimated proportion (standard error) of variation in case-control status explained by all SNPs. [d]Likelihood ratio test statistic corresponding with a test of the null hypothesis that $\sigma^2_g = 0$. [e]*P*-values were calculated assuming that the LRT is distributed as a 50:50 mixture of a point mass at zero and $\chi^2_{(1)}$ under the null hypothesis.

Table 2 Association of rs556621 with large artery atherosclerosis (LAA) and overall ischemic stroke in discovery, replication and combined cohorts

SNP [minor allele] Chr[a].: position[a] genes[b]	Discovery					Replication			Combined discovery and replication		
	RAF[c]	Phenotype	*P* [e]	OR[f] (95% CI)	N_{ca} N_{co}[g]	*P*	OR (95% CI)	N_{ca} N_{co}	*P*	OR (95% CI)	N_{ca} N_{co}
rs556621 [A] 6p21.1: 44,702,137 *CDC5L*, *SUPT3H*	0.30	LAA[d]	3.9×10^{-8}	1.62 (1.36 - 1.93)	421, 1,244	3.9×10^{-4}	1.15 (1.06 - 1.24)	1,715, 52,695	4.7×10^{-8}	1.21 (1.13 - 1.30)	2,136, 53,939
		Ischemic stroke	5.6×10^{-5}	1.29 (1.14 - 1.47)	1,162, 1,244	0.29	1.02 (0.98 - 1.06)	9,552, 52,695	0.03	1.04 (1.00 - 1.08)	10,714, 53,939

[a]Chromosome and NCBI Human Genome Build 36.3 coordinates. [b]Genes located closest to the annotated SNP. [c]Risk allele frequency in controls. [d]LAA: large artery atherosclerotic stroke. [e]*P*-value from 1 d.f. trend test. [f]Odds ratio with 95% confidence interval for the effect of each additional copy of the minor allele, assuming an additive log-odds model. [g]Number of cases (N_{ca}) and controls (N_{co}).

Methods

Study participants: the Australian Stroke Genetics Collaborative (ASGC) discovery sample

ASGC stroke cases comprised European-ancestry stroke patients admitted to four clinical centres across Australia (The Neurosciences Department at Gosford Hospital, Gosford, New South Wales (NSW); the Neurology Department at John Hunter Hospital, Newcastle, NSW; The Queen Elizabeth Hospital, Adelaide ; and the Royal Perth Hospital, Perth) between 2003 and 2008. Stroke was defined by WHO criteria as a sudden focal neurologic deficit of vascular origin, lasting more than 24 hours and confirmed by imaging such as computerised tomography (CT) and/or magnetic resonance imaging (MRI) brain scan. Other investigative tests such as electrocardiogram, carotid doppler and trans-oesophageal echocardiogram were conducted to define ischemic stroke mechanism as clinically appropriate. Cases were excluded from participation if aged <18 years, diagnosed with haemorrhagic stroke or transient ischemic attack rather than ischemic stroke, or were unable to undergo baseline brain imaging. Based on these criteria, a total of 1230 ischemic stroke cases were included in the current study. Ischemic stroke subtypes were assigned using TOAST criteria, based on clinical, imaging and risk factor data [15]

ASGC controls were participants in the Hunter Community Study (HCS), a population-based cohort of individuals aged 55-85 years, predominantly of European ancestry and residing in the Hunter Region, NSW, Australia. Detailed recruitment methods for the HCS have been previously described [25]. Briefly, participants were randomly selected from the NSW State electoral roll and contacted by mail between 2004 and 2007. Consenting participants completed five detailed self-report questionnaires and attended the HCS data collection centre, at which time a series of clinical measures were obtained. A total of 1280 HCS participants were genotyped for the current study.

All study participants gave informed consent for participation in genetic studies. Approval for the individual studies was obtained from relevant institutional ethics committees.

Study participants: Replication cohorts

Replication data were contributed by a total of eleven (11) cohorts involved in the Metastroke and International Stroke Genetics Consortia (ISGC): the Atherosclerosis Risk in Communities Study (ARIC), the Bio-Repository of DNA in Stroke (BRAINS), deCODE Genetics, the Baltimore Genetics of Early Onset Stroke (GEOS) Study, the Heart and Vascular Health (HVH) Study, The Ischemic Stroke Genetics Study / Siblings With Ischemic Stroke Study (ISGS/SWISS), The MGH Genes Affecting Stroke Risk and Outcome Study (MGH-GASROS), the Milano stroke genetics study, the Rotterdam Study, the Wellcome Trust Case-Control

Consortium 2 – Munich (WTCCC2-Munich) and the Wellcome Trust Case-Control Consortium 2 – UK (WTCCC2-UK). All replication cohorts defined ischemic stroke and the LAA, CE and SVO subtypes using clinical criteria consistent with the ASGC discovery sample. Summary demographic data and clinical phenotyping details for these individual cohorts are provided in the Supplementary Methods and Supplementary Table 2.

Genomewide genotyping and quality control: ASGC Discovery sample

ASGC cases and controls were genotyped using the Illumina HumanHap610-Quad array. Quality control excluded SNPs with genotype call rate <0.95, deviation from Hardy-Weinberg equilibrium ($P<1\times10^{-6}$) or minor allele frequency <0.01. At the sample level, quality control excluded individuals with: (i) genotype call rate <95% (n=4); ii) genome-wide heterozygosity < 23.3% or > 27.2% (n=9); iii) inadequate clinical data or inconsistent clinical and genotypic gender (n=45) and; iv) an inferred first- or second-degree relative in the sample based on pair-wise allele sharing estimates (estimated genome proportion shared identical by descent (IBD): pi-hat >0.1875 : *n*=37). Following these exclusions, Eigenstrat principal components analysis (PCA) was performed, incorporating genotype data from Phase 3 HapMap populations (CEU, CHB, JPT, TSI, YRI). In eigenvector plots, the majority of ASGC samples clustered closely with European (CEU and TSI) reference populations. Eighteen samples (16 cases and 2 controls) showed prominent evidence of Asian ancestry and were removed. Principal component and IBD analyses were performed using a pruned subset of quasi-independent SNPs (~130,000 SNPs) to avoid confounding by linkage disequilibrium (LD). Following quality control, 1162 cases and 1244 controls were available for association analyses at 551,514 SNPs.

Genotype imputation in the filtered sample was performed using MACH v1.0.16 [30,31], based on HapMap Phase 2 (release #24) phased haplotypes for European-ancestry (CEU) samples. Subsequent quality control excluded imputed SNPs with MAF <0.01 or ratio of observed dosage variance to expected binomial variance of r^2 <0.3.

Genotyping and quality control: Replication cohorts

Each replication cohort performed genome-wide genotyping, quality control and imputation as part of their own primary study. The particular arrays and quality control filters used by the individual cohorts are described in the Supplementary methods. Of the eleven cohorts, six had directly genotyped rs556621 and five had imputed allelic dosages for this SNP. To ensure the accuracy of results, imputed data was only included if the quality of imputation was high, defined as a ratio of observed to expected binomial dosage

variance (r^2)>0.7. This resulted in the exclusion of one sample (HVH: r^2=0.64). All other samples had $r^2 \geq 0.95$ for rs556621.

Estimating the proportion of phenotypic variation attributable to genotyped SNPs

The proportion of case-control variation attributable to variation in genotyped SNPs was estimated in the Discovery sample using GCTA software [18, 26], which uses genome-wide SNP data to estimate additive genetic relationships (correlations) between essentially unrelated individuals, using a linear mixed model (LMM) to estimate the contribution of genotyped SNPs (and causal variants in LD with genotyped SNPs) to observed variation in case-control status. Prior to analysis, additional QC of genotype data was performed to reduce bias of variance estimates by the accrued effects of small genotyping errors [27]. We excluded SNPs with missingness >0.1% or Hardy-Weinberg *P*-value $<1\times10^{-4}$ and individuals with >0.1% missing genotype data or estimated relatedness >0.05 (approximately closer than second-cousins) [27]. Following QC, genotypes at 457,533 SNPs were available for estimating genetic effects for 1079 ischemic stroke cases, 400 large artery atherosclerosis cases, 288 small vessel occlusion cases and 226 cardioembolism cases. Each case group was evaluated in a separate analysis using a common control sample of 1172 individuals; all fitted LMM models were adjusted for age and sex. Genetic effects were first estimated in the ischemic stroke sample after permuting (shuffling) case-control labels. The estimated proportion of phenotypic variation attributable to genotyped SNPs in the permuted sample was 0.00, indicating an absence of bias due to genotyping or other artifacts. Heritability estimates shown in Table 1 relate to the observed (binary) risk scale and case-control proportions (see **Table 1**). We note that although these estimates do not represent heritability in the conventional sense, the test statistics and their associated significance levels are invariant under adjustment for ascertainment bias or liability scale [28].

Genome-wide association analyses in the Australian Discovery cohort

Genome-wide association analyses were performed using one-degree of freedom trend tests assuming an additive effect of allele dosage. Parameters were estimated using logistic regression models adjusted for age and sex. Analyses were not adjusted for principal components of population ancestry, as observed genomic inflation factors in unadjusted models (λ=1.031, λ_{1000}=1.026 for ischemic stroke, λ=1.007, λ_{1000}=1.011 for LAA) indicated an absence of bias due to population stratification. Meta-analysis genomic control inflation factors (λ) were calculated as previously described, as were standardised values for a sample of 1000 cases and 1000 controls (λ_{1000}) [29]. Secondary analyses of peak regions were adjusted for ancestry principal components and clinical traits including hypertension, hypercholesterolemia, diabetes mellitus, atrial

fibrillation, myocardial infarction and smoking status to investigate potential confounders of the observed genetic associations. Association tests were performed using maximum likelihood estimated dosages for imputed SNPs and observed integer dosages for genotyped SNPs. Logistic models were fitted using *mach2dat* software, which calculates significance levels for estimated parameters using a likelihood-ratio test [30,31]. The two secondary logistic analyses conditioned on rs4479522 and rs556621 genotypes were adjusted for age, sex and integer-valued dosage of the test allele at conditioned SNPs.

Pairwise linkage disequilibrium between SNPs was assessed and visualised using Haploview software [32] based on European (CEU) HapMap Phase 2 data. Haplotype analyses of the 6p21.1 region used genotyped data and maximum likelihood genotypes for SNPs imputed with high reliability (r^2>0.7). Sliding-window haplotypes incorporating from 2 to 6 adjacent SNPs were estimated and assessed for association with LAA case-control status using Unphased software [33]. Regional association plots were constructed using LocusZoom software [34].

Meta-analysis of rs556621 in replication cohorts

For rs556621, each replication sample performed logistic regression using a one-degree of freedom trend test relating the presence of stroke (LAA or overall ischemic stroke) to allelic dosage, assuming an additive effect of the test allele. The test allele, estimated beta coefficient, standard error and effective sample size were provided for the combined replication analysis. Fixed effects, inverse variance-weighted meta-analyses of the ten replication cohorts providing high quality data for rs556621 (see above) was performed using METAL software. Between-study heterogeneity was investigated using Cochran's Q statistic with its associated *P*-value and the I^2 metric, representing the percentage of between-study heterogeneity exceeding the value expected by chance. Population attributable risk (PAR%) was estimated for rs556621 using the formula:

$$PAR\% = \frac{100 \times p(OR-1)}{p(OR-1)+1}$$

where OR is the odds ratio estimated using independent replication data and *p* represents the prevalence of risk alleles in controls [35].

References

1. WHO. Atlas of Heart Disease and Stroke. World Health Organization, Geneva (2004).
2. Feigin, V.L., Lawes, C.M., Bennett, D.A., Barker-Collo, S.L. & Parag, V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurol* **8**, 355-69 (2009).
3. Strong, K., Mathers, C. & Bonita, R. Preventing stroke: saving lives around the world. *Lancet Neurol* **6**, 182-7 (2007).
4. O'Donnell, M.J. et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* **376**, 112-23 (2010).
5. Flossmann, E., Schulz, U.G. & Rothwell, P.M. Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. *Stroke* **35**, 212-27 (2004).
6. Jerrard-Dunne, P., Cloud, G., Hassan, A. & Markus, H.S. Evaluating the genetic component of ischemic stroke subtypes: a family history study. *Stroke* **34**, 1364-9 (2003).
7. Fox, C.S. et al. Genetic and environmental contributions to atherosclerosis phenotypes in men and women: heritability of carotid intima-media thickness in the Framingham Heart Study. *Stroke* **34**, 397-401 (2003).
8. Moskau, S. et al. Heritability of carotid artery atherosclerotic lesions: an ultrasound study in 154 families. *Stroke* **36**, 5-8 (2005).
9. Turner, S.T. et al. Heritability of leukoaraiosis in hypertensive sibships. *Hypertension* **43**, 483-7 (2004).
10. Carmelli, D. et al. Evidence for genetic variance in white matter hyperintensity volume in normal elderly male twins. *Stroke* **29**, 1177-81 (1998).
11. Kaab, S. et al. Large scale replication and meta-analysis of variants on chromosome 4q25 associated with atrial fibrillation. *Eur Heart J* **30**, 813-9 (2009).
12. Gretarsdottir, S. et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* **64**, 402-9 (2008).
13. Palomaki, G.E., Melillo, S. & Bradley, L.A. Association between 9p21 genomic markers and heart disease: a meta-analysis. *JAMA* **303**, 648-56 (2010).
14. Smith, J.G. et al. Common genetic variants on chromosome 9p21 confers risk of ischemic stroke: a large-scale genetic association study. *Circ Cardiovasc Genet* **2**, 159-64 (2009).
15. Adams, H.P., Jr. 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* **24**, 35-41 (1993).
16. Bellenguez, C. et al. Genome-wide association study identifies a variant in *HDAC9* associated with large vessel ischemic stroke. *Nature Genetics* **44**, 328-33 (2012).
17. Flossmann, E., Schulz, U.G. & Rothwell, P.M. Potential confounding by intermediate phenotypes in studies of the genetics of ischaemic stroke. *Cerebrovasc Dis* **19**, 1-10 (2005).
18. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
19. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).
20. Page, N.M., Butlin, D.J., Lomthaisong, K. & Lowry, P.J. The human apolipoprotein L gene cluster: identification, classification, and sites of distribution. *Genomics* **74**, 71-8 (2001).
21. Duchateau, P.N. et al. Plasma apolipoprotein L concentrations correlate with plasma triglycerides and cholesterol levels in normolipidemic, hyperlipidemic, and diabetic subjects. *J Lipid Res* **41**, 1231-6 (2000).
22. Horrevoets, A.J. et al. Vascular endothelial genes that are responsive to tumor necrosis factor-alpha in vitro are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes. *Blood* **93**, 3418-31 (1999).

23. Monajemi, H., Fontijn, R.D., Pannekoek, H. & Horrevoets, A.J. The apolipoprotein L gene cluster has emerged recently in evolution and is expressed in human vascular tissue. *Genomics* **79**, 539-46 (2002).
24. Sana, T.R., Janatpour, M.J., Sathe, M., McEvoy, L.M. & McClanahan, T.K. Microarray analysis of primary endothelial cells challenged with different inflammatory and immune cytokines. *Cytokine* **29**, 256-69 (2005).
25. McEvoy, M. et al. Cohort profile: The Hunter Community Study. *Int J Epidemiol* **39**, 1452-63 (2010).
26. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
27. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *Am J Hum Genet* **88**, 294-305 (2011).
28. Painter, J.N. et al. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat Genet* **43**, 51-4 (2011).
29. de Bakker, P.I. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17, R122-8 (2008).
30. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
31. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).
32. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
33. Dudbridge, F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* **66**, 87-98 (2008).
34. Pruim, R.J. et al. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
35. Zheng, S.L. et al. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* **358**, 910-9 (2008).