Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke

The International Stroke Genetics Consortium (ISGC) & the Wellcome Trust Case Control Consortium 2 (WTCCC2)

Genetic factors have been implicated in stroke risk, but few replicated associations have been reported. We conducted a genome-wide association study (GWAS) for ischemic stroke and its subtypes in 3,548 affected individuals and 5,972 controls, all of European ancestry. Replication of potential signals was performed in 5,859 affected individuals and 6,281 controls. We replicated previous associations for cardioembolic stroke near PITX2 and ZFHX3 and for large vessel stroke at a 9p21 locus. We identified a new association for large vessel stroke within HDAC9 (encoding histone deacetylase 9) on chromosome 7p21.1 (including further replication in an additional 735 affected individuals and 28,583 controls) (rs11984041; combined P = 1.87 x 10^-11; odds ratio (OR) = 1.42, 95% confidence interval (CI) = 1.28–1.57). All four loci exhibited evidence for heterogeneity of effect across the stroke subtypes, with some and possibly all affecting risk for only one subtype. This suggests distinct genetic architectures for different stroke subtypes.

To date, there have been few GWAS for ischemic stroke, and few replicative associations have been identified. To further understand the genetic basis of ischemic stroke, we undertook a GWAS as part of the Wellcome Trust Case Control Consortium 2 (WTCCC2). We hypothesized that associations might be present only with specific stroke subtypes. To investigate this, individuals that had suffered a stroke (cases) were classified into stroke subtypes according to pathophysiological Trial of Organization of Stroke Treatment (TOAST) classification, using clinical assessment as well as brain and vascular imaging where available (see Online Methods). Association analyses were performed on all ischemic stroke cases combined (including individuals not further classified by stroke subtype) and on the three major stroke subtypes: large vessel, small vessel and cardioembolic stroke. Discovery samples were of European ancestry and were genotyped on Illumina arrays (see Online Methods). Following quality control analysis, the discovery set consisted of 3,548 cases (2,374 UK and 1,174 German) and 5,972 controls (5,175 UK WTCCC2 common controls and 797 German controls) genotyped on an overlapping set of 495,851 autosomal SNPs (Table 1 and Online Methods). Within the individual UK and German datasets, cases and controls were well matched for ancestry (Online Methods and Supplementary Fig. 1). We therefore performed association analysis separately in the two groups and combined them using a fixed-effects meta-analysis approach. A two-stage replication study was performed in 5,859 cases (3,863 European and 1,996 US) and 6,281 controls (4,554 European and 1,727 US), all of self-reported European ancestry (Table 1 and Online Methods). Full details of the cohorts are provided in the Supplementary Note and in Supplementary Table 1.

We show the results at previously reported loci (Table 2) and the association analysis results across the autosomes (Fig. 1). We replicated an association between cardioembolic stroke and variants close to the PITX2 gene and also a SNP in the ZFHX3 gene, both of which were initially associated with atrial fibrillation, a well-recognized risk factor for stroke. We also replicated a previously reported association between large vessel stroke and the 9p21 region. As we, and others have already reported, we did not confirm the previously published association between all stroke cases and variants in the 12p13 region.

Thirty-eight previously unreported loci showed potential association for all stroke cases or for one of the stroke subtypes in the discovery

A full list of authors and affiliations appears at the end of the paper. A full list of members is provided in the Supplementary Note.

Received 26 May 2011; accepted 15 December 2011; published online 5 February 2012; doi:10.1038/ng.1081
Table 1 Breakdown of cases and controls by cohort and ischemic stroke subtype after quality control analyses

| Stage | Cohort   | All strokes | LVD | CE | SVD | Controls |
|-------|----------|-------------|-----|----|-----|----------|
| Discovery | Munich   | 1,174       | 346 | 330| 106 | 797      |
|        | UK       | 2,374       | 498 | 460| 474 | 5,175    |
|        | Total    | 3,548       | 844 | 790| 580 | 5,972    |
| Stage 1 replication – European | Krakow   | 1,214       | 152 | 362| 170 | 551      |
|        | Leuven   | 418         | 63  | 154| 52  | 650      |
|        | Lund     | 428         | 21  | 139| 97  | 465      |
|        | Munichb  | 54          | 19  | 16 | 5   | 310      |
|        | UKc      | 1,749       | 306 | 303| 490 | 2,578    |
|        | Total    | 3,863       | 561 | 674| 1,174| 4,554    |
| Stage 2 replication – US | Boston   | 533         | 150 | 206| 56  | 522      |
|        | Cincinnati | 438       | 67  | 106| 90  | 257      |
|        | GEOs     | 419         | 37  | 90 | 54  | 498      |
|        | ISGS     | 606         | 121 | 156| 111 | 450      |
|        | Total    | 1,996       | 375 | 558| 311 | 1,727    |
| Stage 1 and stage 2 replication | Total   | 5,859       | 936 | 1,532| 1,125 | 6,281   |
| Discovery and replication | Total  | 9,407       | 1,780 | 2,322| 1,705 | 12,253   |

All, all ischemic stroke; LVD, large vessel stroke; SVD, small vessel stroke; CE, cardioembolic stroke. Note that not all strokes are classified into a subtype.

aThe UK discovery cohort was composed of three UK cohorts from London, Oxford and Edinburgh and used the shared WTCCC2 controls.

bThe Munich replication samples comprised some samples planned for the discovery GWAS where there was insufficient DNA for GWAS but sufficient amounts for replication. It used controls from a German cohort enrolled in the PROCARDIS trial. The UK replication cohorts included samples from Aberdeen, Glasgow and Imperial as well as some samples planned for the discovery GWAS where there was insufficient DNA for GWAS but sufficient amounts for replication (see Online Methods). The UK replication cohorts used shared POBI controls genotyped as part of the WTCCC2.

together with the other two subtypes (‘LVD’, ‘SVD’ and ‘CE’ models, respectively, for the effect only in large vessel, small vessel and cardioembolic stroke) and (iv) a ‘correlated effects’ model allowing different but correlated effects for each subtype. We undertook model comparison in a Bayesian statistical framework for this SNP in three other collections of large vessel cases and matched controls (735 cases and 28,583 controls in total), which we refer to as stage 3 replication (see Online Methods for details).

Table 2 Association signals at the newly associated locus and at loci previously reported to be associated with stroke or a stroke subtype

| Chr. | SNP       | Position | Candidate gene | Stroke subtype | Risk allele | RAFe | Discovery P value OR (95% CI) | Stage 1 and 2 P value (one sided) OR (95% CI) | Stage 3 P value (one sided) OR (95% CI) | Combined P value OR (95% CI) |
|------|-----------|----------|----------------|---------------|-------------|------|-----------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| 7p21.1 | rs11984041<sup>c,d</sup> | 18998460 | HDAC9          | LVD           | A            | 0.09 | 1.07 x 10<sup>-5</sup>       | 7.90 x 10<sup>-5</sup>                           | 2.25 x 10<sup>-4</sup>                         | 1.87 x 10<sup>-11</sup>                      |
| 4q25 | rs200733e<sup>f</sup>| 111929618 | PITX2          | CE            | A            | 0.10 | 3.64 x 10<sup>-6</sup>       | 3.99 x 10<sup>-4</sup>                           | –                                         | 5.06 x 10<sup>-8</sup>                      |
| 9q21.3 | rs2383207<sup>e</sup> | 22105959 | CDKN2A, CDKN2B | LVD           | G             | 0.51 | 2.35 x 10<sup>-5</sup>       | 2.03 x 10<sup>-3</sup>                           | –                                         | 2.93 x 10<sup>-5</sup>                      |
| 12p13.33 | rs11833579<sup>a</sup> | 645460 | NINJ2          | All           | G             | 0.75 | 9.65 x 10<sup>-5</sup>       | 5.25 x 10<sup>-1</sup>                           | –                                         | 9.81 x 10<sup>-1</sup>                      |
| 16q22.3 | rs1292344<sup>d</sup> | 71586661 | ZFHX3          | CE            | A             | 0.16 | 1.94 x 10<sup>-5</sup>       | –                                           | –                                         | –                                             |
|       |           |           |                |               |               |      |                             |                                              |                                              |                                              |

Chr., chromosome; all, all ischemic stroke; LVD, large vessel stroke; CE, cardioembolic stroke. For PITX2 and ZFHX3, results are given for one SNP reported in the literature and the SNP that showed the strongest association signal in the discovery cohort. For the 7p21 region, the SNP reported in the literature is also the one showing the strongest association signal in the discovery cohort. There is some overlap between samples in this study and those in previously published studies of associations<sup>6,10,13</sup>. A SNP in PRKCH that was associated with stroke in Japanese populations<sup>23</sup> was very rare (RAF < 0.5%) in our population of European ancestry, so we had no power to perform an analysis of association with this SNP.

aNCBI human genome build 36 coordinates.

bRisk allele frequency (RAF) computed in the UK discovery control.

cKrakow replication samples were not considered because the Hardy-Weinberg test P value was <5 x 10<sup>-4</sup> in controls.

dSNP imputed in GEOs replication samples. SNP reported in the literature.

© 2012 Nature America, Inc. All rights reserved.
Figure 1  Genome-wide association results at autosomal SNPs in combined UK and German discovery samples. (a-c) Results are shown for all ischemic stroke (a), large vessel stroke (b), small vessel stroke (c) and cardioembolic stroke (d). Loci previously reported in the literature for particular stroke subtypes (Table 2) are shown in black, with the new HDAC9 locus shown in red. The combined P values for the discovery study and stages 1 and 2 of replication at the top SNPs for these loci are marked with diamonds.

(see Online Methods for details) for our new association near HDAC9, as well as for the previously reported associations that we confirmed (those listed in Table 2). The results, based on the discovery stage and the first two stages of replication, are shown in Figure 3.

For rs11984041 at HDAC9, there was very strong evidence against the null model and both the SVD and CE models (which was not unexpected, given that we ascertained this SNP on the basis of evidence for an effect in LVD), and there was also strong evidence against the model in which the SNP has the same effect in each stroke subtype, thereby showing genetic heterogeneity across stroke subtypes at this SNP. The greatest posterior weight rested on the model in which there is only an effect for large vessel disease, with some weight on the correlated effects model in which the posterior distributions on effect size for SVD and CE were concentrated on much smaller effect sizes than for LVD.

In our data, heterogeneity was also seen at rs2383207 in the 9p21
tissue, muscle and brain
ubiquitously expressed, with high levels of expression in cardiac
lating chromatin structure and gene transcription
genes that encode proteins that deacetylate histones, thereby regu
identified peak and cannot be excluded as possible mechanisms
TWIST1
that encompasses the tail end of
HDAC9
and cardioembolic stroke (see Online Methods for details) for our new association near HDAC9
myogenesis and is involved in heart development, although deleterious effects on systemic arteries have not yet been reported. Alternatively, it could increase risk by altering brain ischemic responses and might therefore have effects on neuronal survival. The HDAC9 protein has been shown to protect neurons from apoptosis, both by inhibiting JUN phosphorylation by MAPK10 and by repressing JUN transcription. HDAC inhibitors have been postulated as a treatment for stroke.

It is not uninformative that a large GWAS (~3,500 cases and ~6,000 controls) failed to find any new associations for the combined phenotype of ischemic stroke. It may be that the genetic architecture of the disease involves fewer variants of more moderate effect than many other diseases and/or these variants may not be well tagged by the Illumina Human660W-Quad chip used in the study. On the other hand, as our data show, all the known loci exhibit genetic heterogeneity across the stroke subtypes, with at least some and possibly all affecting only a single subtype. This supports the possibility that distinct subtypes of the disease have differing genetic architecture. However, this hypothesis is based on the results from only four loci and does not exclude the possibility that future loci associated with stroke may predispose to all ischemic stroke. Clinical classification of disease into subtypes is not perfect. As errors in classification would reduce the power to detect heterogeneity, our findings of homogeneity...
within classes indirectly reinforces the value of current classification methods. Because GWAS studies to date, including the one reported here, have had relatively small sample sizes for each disease subtype (and hence are underpowered for common variants of small effect), it remains possible and indeed is likely a priori that the range of effect sizes for each subtype will be similar to those for other common diseases. This suggests that future genetic studies should include adequate sample sizes for particular subtypes of ischemic stroke rather than for the disease as a whole.

In summary, in this largest GWAS study of ischemic stroke conducted to date, we identified a new association with the HDAC9 gene region in large vessel stroke with an estimated effect size that is at the larger end for GWAS loci (OR = 1.38, 95% CI = 1.22–1.57, from replication data). We also replicated known associations with three other loci and showed genetic heterogeneity across subtypes of the disease for all four stroke loci. This genetic heterogeneity seems likely to reflect heterogeneity in the underlying pathogenic mechanisms and reinforces the need for the consideration of stroke subtypes separately in research and clinical contexts.

URLs. Precocious Coronary Artery Disease (PROCARDIS), www.procardis.org; SNPTTEST, https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html; National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics DNA and Cell Line Repository, http://ccr.coriell.org/ninds; Biowulf Linux cluster at the US National Institutes of Health, http://biowulf.nih.gov/.

Figure 2 Forest plot for the associations between rs11984041 and large vessel stroke in discovery and replication collections. The blue lines show the 95% confidence intervals of the log (OR) for each cohort, with the area of each square proportional to the inverse of the standard error. The diamonds indicate the 95% confidence interval for the discovery summary (combined UK and German discovery collections), combined across collections within each of the three replication stages, the replication summary (combined across all three replication stages) and the overall summary (all discovery and replication collections combined). Evidence was combined across collections via an inverse variance weighted fixed-effect meta-analysis. There was no evidence of heterogeneity of effect across collections (P = 0.92).

Figure 3 Genetic heterogeneity of different stroke subtypes for the four loci with significant associations. (a–d) Data are shown for HDAC9 (a), PITX2 (b), the 9p21 region (CDKN2A, CDKN2B) (c) and ZFHX3 (d). Bar plots show the posterior probabilities on the models of association: no effect in any subtype (null), same effect in all subtypes, correlated effects across subtypes or subtype-specific effects. Models are a priori assumed to be equally likely. Bayes factors, which compare the evidence (marginal likelihood) between any pair of models, can be calculated as the ratio of the posterior probability assigned to each model as reported under each bar of the plot. Accompanying density plots show the marginal posterior distribution on the OR of the risk allele for each stroke subtype assuming a model of correlated effects (see Online Methods for specification of priors). These analyses were performed using both discovery and replication samples (stages 1 and 2).
Twelve type 2 diabetes susceptibility loci identified through The association of the 4q25 susceptibility variant for atrial fibrillation on chromosome 4q25

Figure 4 Plot of association signals around rs11984041 for large vessel stroke in the combined UK and German discovery samples. Top: SNPs are colored based on their correlation ($r^2$) with the labeled top SNP, which has the smallest $P$ value in the region. $r^2$ is calculated from the WTCCC2 control data. Bottom, the fine-scale recombination rates estimated from hapMap data, with genes marked by horizontal red lines. Arrows on the horizontal red lines show the direction of transcription, and black rectangles are exons.

METHODS
Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/. Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank S. Bertrand, J. Bryant, S.L. Clark, J.S. Conquer, T. Dibling, J.C. Eldred, S. Gamble, C. Hind, M.L. Perez, C.R. Stribling, S. Taylor and A. Wilk of the Wellcome Trust Sanger Institute’s Sample and Genotyping Facilities for technical assistance. We acknowledge use of the British 1958 Birth Cohort DNA collection, which is funded by the Medical Research Council (G0000934) and the Wellcome Trust Sanger Institute’s Sample and Genotyping Facilities for technical assistance. We acknowledge use of the British 1958 Birth Cohort DNA collection, which is funded by the Medical Research Council (G0000934) and the Wellcome Trust Sanger Institute’s Sample and Genotyping Facilities for technical assistance. We acknowledge use of the British 1958 Birth Cohort DNA collection, which is funded by the Medical Research Council (G0000934) and the Wellcome Trust Sanger Institute’s Sample and Genotyping Facilities for technical assistance. We acknowledge use of the British 1958 Birth Cohort DNA collection, which is funded by the Medical Research Council (G0000934) and the Wellcome Trust Sanger Institute’s Sample and Genotyping Facilities for technical assistance.

AUTHOR CONTRIBUTIONS
S. Bevan, C.C.A.S., P.S., M.E., C.L.M.S., P.M.R., M.D., P. Donnelly and H.S.M. designed the experiment. S. Bevan, A.G., A.I.B., C.A.I.J., D.P., L.M., H.S., C.L.M.S., P.M.R., M.D. and H.S.M. were responsible for collecting and phenotyping discovery samples. Replication sample or replication data were provided by P.D.S., J.P., B.N., R.L., R.S., L.C., Y.-C.C., D.W., M.A.N., U.S., H.R.-A., S. Boonen, D.W.-K., V.T., K.I., H.D., S.G., G.T., U.T., K.S., G.B.R., E.A.P., J.A., E.H., C.L., M.-F.E., A.H., B.R.W., S.I.K., R.D.M., B.K., J.E.M., V.T., A.L., M.J.M., A.S., M.W., J.R. and P.S.

Genotyping, quality control and informatics were conducted by C.B., S. Bevan, C.C.A.S., P.S., M.E., M.T., A.S., Z.S., G.B., C.F., R.M., B.M.-M., C.M., C.L., S.E., S.H., E.G., S.D., A.G., M.F., P. Donnelly and H.S.M. Genetic and statistical analysis was performed by C.B., S.B., C.C.A.S., M.P., A.S., Z.S., G.B., C.F., M.T., R.M., A.H., M.F. and P. Donnelly. The WTCCC2 management committee (P. Donnelly [Chair], J.P. [Deputy Chair], J.M.B., E.B., M.A.B., I.P.C., A.C., P. Deloukas, A.D., J.H., H.S.M., C.G.M., C.N.A.P., R.P., A.R., S.J.S., R.C.T., A.C.V. and N.W.W.) monitored the execution of the study. C.B., S. Bevan, C.C.A.S., M.P., M.F., P. Donnelly and H.S.M. contributed to writing the first draft of the manuscript. All authors reviewed and commented on the final manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.
The authors of this paper are:
Céline Bellenguez1,4,5, Steve Bevan1,4, Andrea Gschwendtner3, Chris CA Spencer1, Annette I Burgess4, Matti Pirinen5, Caroline A Jackson6, Matthew Traylor2, Amy Strange1, Zhan Su1, Gavin Band1, Paul D Syme6, Rainer Malik3, Joanna Pera7, Bo Norrving,8,9, Robin Lennm910,11, Colin Freeman1, Renata Schan12, Tom James2, Deborah Poole4, Lee Murphy13, Helen Segal3, Lynelle Cortellini14,15, Yu-Ching Cheng16,17, Daniel Woo18, Michael A Nalls19, Bertram Müller-Myllysk20, Christa Meisinger21, Udo Seedorf22, Helen Ross-Adams7, Steven Boonen23, Dorota Wloch-Kopec7, Valerie Valant4,15, Julia Slark12, Karen Furie14, Hossein Delavaran8,9, Cordelia Langford24, Panos Deloukas24, Sarah Edkins24, Sarah Hunt24, Emma Gray24, Serge Dronov24, Leena Peltonen24,25, Solveig Gretsbdottir25, Guthar Thorleifsson25, Unnur Thorsteinsdottir25,26, Kari Stefansson25,26, Giorgio B Boncoraglio27, Eugenio A Parati27, John Attia28, Elizabeth Holliday28, Chris Lev28, Maria-Grazia Franzoi9, Anuj Goel1,30, Anna Helgadottir1,25,30, Jenefer M Blackwell31,32, Elvira Bramon33, Matthew A Brown34, Juan P Casas35,36, Aiden Corvin37, Audrey Duncanson38, Janusz Jankowski39,40, Christopher G Mathew41, Colin NA Palmer43, Robert Plomin43, Anna Rautanen1, Stephen J Sawyer44, Richard C Trembath41, Ananth C Viswanathan45, Nicholas W Wood46, Bradford RJorrell47,48, Steven J Kottner49,50, Braxton D Mitchell16,17, Brett Kissela18, James F Meschia31, Vincent Thijs10,11, Arne Lindgren8,9, Mary Joan Macleod6, Agnieszka Slowik7, Matthew Walters52, Jonathan Rosand14,15, Pankaj Sharma12, Martin Farrall1,30, Cathie LM Sudlow5, Peter M Rothwell4, Martin Dichgans3, Peter Donnelly1,53,55 & Hugh S Markus2,55.

1Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. 2Stroke and Dementia Research Group, St. George’s University of London, London, UK. 3Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians-Universität, Munich, Germany. 4Stroke Prevention Research Unit, Nuffield Department of Clinical Neuroscience, University of Oxford, Oxford, UK. 5Division of Clinical Neurosciences, University of Edinburgh, Edinburgh, UK. 6Division of Applied Medicine, University of Aberdeen, UK. 7Department of Neurology, Jagiellonian University Medical College, Krakow, Poland. 8Neurology, Department of Clinical Sciences, Lund University, Lund, Sweden. 9Department of Neurology, Skåne University Hospital, Lund, Sweden. 10Department of Neurology, University Hospitals Leuven, Leuven, Belgium. 11Vesalius Research Center, Vlaams Instituut voor Biotechnologie (VIB), Leuven, Belgium. 12Imperial College Cerebrovascular Research Unit (ICCRU), Imperial College London, London, UK. 13Wellcome Trust Clinical Research Facility Genetics Core Laboratory, University of Edinburgh, Western General Hospital, Edinburgh, UK. 14Department of Neurology, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. 15Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. 16Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA. 17Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA. 18Department of Neurology, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA. 19Laboratory of Neurogenetics, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA. 20Max Planck Institute of Psychiatry, Munich, Germany. 21Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology II, Neuherberg, Germany. 22Leibniz-Institut für Arteriosklerosforshung, Universität Münster, Münster, Germany. 23Division of Geriatric Medicine, University Hospitals Leuven, Leuven, Belgium. 24Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. 25deCODE Genetics, Reykjavik, Iceland. 26University of Iceland, Faculty of Medicine, Reykjavik, Iceland. 27Fondazione Istituto di Ricoero e Cura a Carattere Scientifico (IRCCS) Istituto Neurologico Carlo Besta, Milan, Italy. 28Centre for Brain and Mental Health Research, University of Newcastle, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. 29Department of Cardiovascular Research, Istituto di Ricerca Farmacologiche Mario Negri, Milan, Italy. 30Department of Cardiovascular Medicine, University of Oxford, Oxford, UK. 31Centre for Child Health Research, University of Western Australia, West Perth, Western Australia, Australia. 32Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge, UK. 33Division of Psychological Medicine and Psychiatry, Biomedical Research Centre for Mental Health, Institute of Psychiatry, King’s College London, London, UK. 34The University of Queensland Diamantina Institute, Princess Alexander Hospital, University of Queensland, Brisbane, Queensland, Australia. 35Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. 36Department of Epidemiology and Public Health, University College London, London, UK. 37Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland. 38Molecular and Physiological Sciences, The Wellcome Trust, London, London, UK. 39Centre for Gastroenterology, Barts and the London School of Medicine and Dentistry, London, UK. 40Division of Clinical Pharmacology, University of Oxford, Oxford, UK. 41Department of Medical and Molecular Genetics, King’s College London School of Medicine, Guy’s Hospital, London, UK. 42Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, UK. 43Social, Genetic and Developmental Psychiatry Centre, King’s College London Institute of Psychiatry, Denmark Hill, London, UK. 44University of Cambridge Department of Clinical Neurosciences, Addenbrooke’s Hospital, Cambridge, UK. 45National Institute for Health Research (NIHR) Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital National Health Service (NHS) Foundation Trust and University College London Institute of Ophthalmology, London, UK. 46Department of Molecular Neuroscience, Institute of Neurology, Queen Square, University College London, London, UK. 47Department of Neurology, University of Virginia School of Medicine, Charlottesville, Virginia, USA. 48Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, USA. 49Department of Neurology, University of Maryland School of Medicine, Baltimore, Maryland, USA. 50Baltimore Geriatric Research, Education, and Clinical Center, Baltimore Veterans Affairs Medical Center, Baltimore, Maryland, USA. 51Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA. 52Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. 53Department of Statistics, University of Oxford, Oxford, UK. 54These authors contributed equally to this work. 55These authors jointly directed this work. 56Deceased. Correspondence should be addressed to P. Donnelly (peter.donnelly@well.ox.ac.uk) or H.S.M. (hmarkus@sgul.ac.uk).
ONLINE METHODS

Study subjects. All subjects were of self-reported European ancestry. Cases were classified into mutually exclusive etiologic subtypes according to TOAST classification\(^2\), which was performed in all stroke cases. The TOAST system has a category of ‘etiology unknown’ that includes cases in which no cause has been found as a result of insufficient investigation, as well as cases where no cause is found despite full investigation. This unknown group was not analyzed in subtype analyses described in this manuscript, which focused only on those cases for whom there were appropriate investigations to assign one of three subtypes: large vessel disease, cardioembolic and small vessel disease. The unknown cases were only included in the analyses of all ischemic stroke, which did not take into account subtype.

Our main analyses were of associations with all ischemic stroke and with the three main subtypes. We performed additional analyses in the discovery populations with young stroke (age <70 years at first stroke) and with the presence of large vessel stenosis and, separately, the presence of cardioembolic source, irrespective of assigned subtype. These last two analyses allowed inclusion of subjects whose data were excluded from individual subtype analysis because they had more than one potential stroke subtype. Details of individual populations are given in Table 1 and in the Supplementary Note. For all cohorts, recruitment of cases was approved by the relevant local ethics committee, and all participants gave informed consent.

GWAS genotyping. Samples from the cases were genotyped at the Wellcome Trust Sanger Institute (WTSI) on the Human660W-Qaud (a custom chip designed by WTTCC2 comprising Human550 SNPs and ~6,000 common copy-number variants (CNVs) from the Structural Variation Consortium\(^23\)). Samples from UK control collections were genotyped on the Human1.2M-Duo (a WTCC2 custom array comprising Human1M-Duo SNPs and the CNV content described above). Bead intensity data were processed and normalized in BeadStudio (Illumina); data for successfully genotyped samples were extracted and genotypes called within collections using Illuminuc25. German controls were genotyped on the Illumina Human550k platform, and intensity data were processed and normalized for each sample in GenomeStudio (Illumina) using the Illumina cluster file HumanHap550v3.

GWAS quality control. For samples, quality control analysis was performed as previously described\(^16,27\). We removed samples whose genome-wide patterns of diversity differed from those of the collection at large, as these differences were likely to be caused by biases or artifacts. We used a Bayesian clustering approach\(^27\) to infer outlying individuals on the basis of call rate, heterozygosity, ancestry and average probe intensity. To obtain a set of putatively unrelated individuals, we used a hidden Markov model (HMM) to infer identity by descent and then iteratively removed individuals to obtain a set with pairwise identity by descent of <5%. To guard against sample mishandling, we removed samples if their inferred gender was discordant with recorded gender or if <90% of the SNPs genotyped by Sequenom at the beginning of sample analysis agreed in the case or control collections. In the 58C, UKBS samples, UKBS cases had similar ancestry to German stroke cases (data not shown). The PROCARDIS controls were genotyped with the Illumina HumanHap610-Quad chip. Principal-component analysis (PCA) with HapMap 2 reference population data allowed exclusion of individuals with non-European ancestry. Subsequent PCA with HapMap 3 data on German stroke samples with GWAS data and additional European reference population data showed German PROCARDIS controls had similar ancestry to German stroke cases (data not shown).

Association analysis. We performed single GWAS analysis separately in the UK and German discovery datasets under an additive model (on the log (OR)) using missing data likelihood score tests as implemented in SNPTTEST. We conducted a fixed-effects meta-analysis in R to combine the evidence of association, averaging the estimated effect size parameters associated with genotype risk across the two datasets and weighting the effect size estimates by the inverse of the square of corresponding standard errors. \(P\) values were calculated assuming the combined data \(z\) score to be normally distributed. The UK and German cohorts had an inflation factor ranging from 1.014 to 1.058 and from 1.011 to 1.044, respectively, depending on the stroke subtype considered (Supplementary Fig. 1). This analysis was also performed separately in males and females.

We also conducted a genome-wide analysis based on a Bayesian model that allows each stroke subtype to have its own effect and models relationships between these effects using a hierarchical prior specification. The same effects were assumed for the corresponding stroke subtype in both the UK and German populations (Supplementary Table 2).

Finally, we performed a genome-wide scan using GENECLUSTER\(^20\). This estimates the genealogical tree of a case-control series at a position of interest selected on the basis of the genealogy of the reference panel (HapMap 2 Utah residents of Northern and Western European ancestry (CEU) in our study) by simultaneously phasing and clustering the case and control haplotypes to the tips of the reference genealogy. The method detects signals of association in the form of differential clustering of cases and controls underneath a branch or a number of branches in the estimated genealogy, which is equivalent to associations due to haplotype effects or allelic heterogeneity (Supplementary Table 2).

Stages 1 and 2 replication. Replication of potential associations found in the GWAS of the discovery cohorts was conducted in two stages in independent European and US samples. We investigated in the European replication cohorts 50 SNPs, that either were in loci reported in the literature from previous GWAS or showed potential associations (\(P < 1 \times 10^{-5}\)) with all stroke or one of the stroke subtypes in analysis of the discovery dataset with consistent direction of effect in both the UK and German cohorts (Supplementary Table 2). This threshold was chosen on the basis of resources available for replication. After analysis of the combined results of the discovery and European replication populations, 20 of these SNPs were taken forward to a second stage of replication in the US samples (Supplementary Table 3).

Association analysis was performed in each replication cohort separately via a logistic regression assuming an additive genetic model. Evidence of association across the replication data were combined using a fixed-effects meta-analysis. Data on the presence or absence of a cardioembolic source or large vessel stenosis (irrespective of assigned TOAST subtype) were not available in all replication cohorts. For replication of SNPs identified because of association with cardioembolic source or large vessel stenosis in the discovery cohorts, we assessed association in the replication cohorts with the cardioembolic or large vessel stroke subtypes, respectively.

Stage 3 replication of rs11984041. For the deCODE cases and controls, genotyping was performed on Illumina 317k or 370k chips. The rs11984041 SNP assays, and genotyping of the US samples was performed at the Broad Institute using the Sequenom platform, with the exception of the GEOS study, for which genotyping was carried out using Illumina HumanOmni1-Qaud chips. Imputation to HapMap 3 using the BEAGLE software program\(^26\) was performed. Individual samples were excluded from analysis if they had call rates of <80% or if reported gender was discordant with gender-specific markers. We removed pairs of samples showing concordance indicative of being duplicates. The PoBI samples were genotyped on the custom Human1.2M-Duo array using the Illumina Infinium platform and subjected to similar quality control as described above. For each SNP used in replication the cluster plot was visually inspected.

The PROCARDIS controls were genotyped with the Illumina HumanHap610-Quad chip. Principal-component analysis (PCA) with HapMap 2 reference population data allowed exclusion of individuals with non-European ancestry. Subsequent PCA with HapMap 3 data on German stroke samples with GWAS data and additional European reference population data showed German PROCARDIS controls had similar ancestry to German stroke cases (data not shown).
was imputed using HapMap. ASGC cases and control samples were genotyped on the Illumina HumanHap610-Quad chip, and the rs11984041 SNP was directly genotyped. Milan cases were genotyped using Illumina Human610-Quad v1_B or Human660W-Quad v1_A chips; both include the rs11984041 SNP. Milan controls were genotyped with the Illumina HumanHap610-Quad chip. PCA with HapMap 3 on the Italian cases showed that Italian PROCARDIS controls had similar ancestry to the cases.

24. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712 (2010).
25. Teo, Y.Y. et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23, 2741–2746 (2007).
26. Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat. Genet. 42, 985–990 (2010).
27. The UK Parkinson's Disease Consortium & The Wellcome Trust Case Control Consortium 2. Dissection of the genetics of Parkinson's disease identifies an additional association S' of SNCA and multiple associated haplotypes at 17q21. Hum. Mol. Genet. 20, 345–353 (2011).
28. Bellenguez, C. et al. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics 28, 134–135 (2012).
29. Browning, B.L. & Browning, S. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84, 210–223 (2009).
30. Su, Z., Cardin, N., Donnelly, P. & Marchini, J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. Stat. Sci. 24, 430–450 (2009).