nature genetics

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

The International Stroke Genetics Consortium (ISGC)[1] & the Wellcome Trust Case Control Consortium 2 (WTCCC2)[1]

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 $P = 1.87 \times 10^{-11}$; 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.

Cerebrovascular disease (stroke) is one of the three most common causes of death and the major cause of adult chronic disability[1]. Stroke represents an increasing health problem throughout the world as the proportion of elderly individuals increases, and it is an important cause of dementia and age-related cognitive decline. Although conventional risk factors, such as hypertension, account for a substantial proportion of stroke risk, much remains unexplained[2]. Twin and family history studies suggest that genetic factors are responsible for some of this unexplained risk[3]. Stroke is a syndrome rather than a single disease, and different subtypes of stroke are caused by a number of distinct and specific disease processes. Approximately 80% of stroke is ischemic: the three most common ischemic stroke subtypes are large vessel, cardioembolic and small vessel (lacunar) stroke. Genetic epidemiological studies show heterogeneity between stroke subtypes, with the large vessel subtype being more strongly associated with family history[4]. SNPs associated with atrial fibrillation were found only to be significantly associated with cardioembolic stroke[5,6], and a variant on chromosome 9p21, initially associated with coronary artery disease and atherosclerosis, only associated with large vessel stroke[7]. These findings suggest that different genetic variants predispose to the different subtypes of ischemic stroke.

To date, there have been few GWAS for ischemic stroke, and few replicable associations have been identified[8]. 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 10172 in Acute Stroke Treatment (TOAST) classification[9], 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[5,6,10]. We also replicated a previously reported association between large vessel stroke and the 9p21 region[7]. As we, and others have already reported[11,12], we did not confirm the previously published association between all stroke cases and variants in the 12p13 region[13,14].

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. [1]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

		All strokes	LVD	CE	SVD	Controls
Discovery	Munich	1,174	346	330	106	797
	UK[a]	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
	Munich[b]	54	19	16	5	310
	UK[c]	1,749	306	303	490	2,578
	Total	3,863	561	974	814	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.

[a]The UK discovery cohort was composed of three UK cohorts from London, Oxford and Edinburgh and used the shared WTCCC2 controls. [b]The 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. [c]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.

samples, and we further investigated these loci in the European replication samples by genotyping 43 SNPs covering these loci as well as 7 SNPs covering the previously reported loci (**Supplementary Table 2**). Thirteen of the newly identified loci and all of the previously reported loci were taken forward to replication in the US samples, where we performed genotyping of 20 SNPs covering these regions (**Supplementary Table 3**). Most replication samples were genotyped using Sequenom assays; we used genotype imputation for those SNPs that were not directly genotyped (Online Methods and **Supplementary Tables 2** and **3**). A SNP on chromosome 7p21.1 (rs11984041) showed evidence of association with large vessel stroke in the discovery data ($P = 1.07 \times 10^{-5}$) and in the joint European and US replication data in the same direction (one-sided $P = 7.9 \times 10^{-5}$). As an additional verification, we investigated 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). The stage 3 data also showed evidence in the same direction (one-sided $P = 2.25 \times 10^{-4}$). Together, the combined discovery and three-stage replication data provide strong evidence for association ($P = 1.87 \times 10^{-11}$) and suggest that each copy of the A allele at rs11984041 increases the risk of large vessel stroke by approximately 1.4-fold (**Fig. 2** and **Table 2**). This SNP is located within the final intron of the *HDAC9* gene. The frequency of the risk allele (A) was 9.29% and 8.78% in the UK and German discovery controls, respectively.

Standard statistical tests determined that the association between rs11984041 and the individual sets of cardioembolic and small vessel stroke cases was not significant (discovery and two-stage replication $P = 0.12$; OR = 1.10, 95% CI = 0.98–1.23 and $P = 0.06$; OR = 1.13, 95% CI = 1.00–1.28, respectively). A nonsignificant result could simply be caused by a lack of power: lack of significance in itself cannot rule out an effect in these subtypes. We investigated this potential genetic heterogeneity further by formally comparing different statistical models for the effect of the SNP on the different stroke subtypes. The models we compared were (i) a model in which the variant has no effect on risk for any of the subtypes ('null' model), (ii) a model in which the SNP has the same effect on each subtype ('same effects' model), (iii) three models in each of which the SNP has an effect on one subtype and no effect for 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

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[a]	Candidate gene	Stroke subtype	Risk allele	RAF[b]	Discovery *P* value OR (95% CI)	Stages 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[c,d]	18998460	*HDAC9*	LVD	A	0.09	1.07×10^{-5}	7.90×10^{-5}	2.25×10^{-4}	1.87×10^{-11}
							1.50 (1.25–1.79)	1.38 (1.17–1.63)	1.39 (1.15–1.68)	1.42 (1.28–1.57)
4q25	rs2200733[d,e,f]	111929618	*PITX2*	CE	A	0.10	3.64×10^{-6}	3.99×10^{-4}	–	5.06×10^{-8}
							1.49 (1.26–1.77)	1.24 (1.09–1.41)	–	1.32 (1.20–1.46)
	rs1906599[e]	111932135			A	0.19	3.45×10^{-8}	3.16×10^{-4}	–	1.39×10^{-9}
							1.45 (1.27–1.66)	1.19 (1.08–1.32)	–	1.28 (1.18–1.39)
9p21.3	rs2383207[e]	22105959	*CDKN2A, CDKN2B*	LVD	G	0.51	2.35×10^{-3}	2.03×10^{-3}	–	2.93×10^{-5}
							1.18 (1.06–1.31)	1.16 (1.05–1.28)	–	1.17 (1.09–1.25)
12p13.33	rs11833579[e,g]	645460	*NINJ2*	All	G	0.75	9.65×10^{-1}	5.25×10^{-1}	–	9.81×10^{-1}
							1.00 (0.92–1.08)	1.00 (0.94–1.06)	–	1.00 (0.95–1.05)
16q22.3	rs7193343[e]	71586661	*ZFHX3*	CE	A	0.16	1.94×10^{-5}	–	–	–
							1.36 (1.18–1.57)	–	–	–
	rs12932445[d]	71627389			G	0.17	3.91×10^{-7}	4.84×10^{-2}	–	1.44×10^{-5}
							1.44 (1.25–1.66)	1.09 (0.98–1.21)	–	1.20 (1.11–1.31)

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 9p21 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 association[5,6,10,13]. A SNP in *PRKCH* that was associated with stroke in Japanese populations[23] 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.

[a]NCBI human genome build 36 coordinates. [b]Risk allele frequency (RAF) computed in the UK discovery control. [c]Krakow replication samples were not considered because the Hardy-Weinberg test *P* value was $<5 \times 10^{-4}$ in controls. [d]SNP imputed in GEOS replication samples. [e]SNP reported in the literature. [f]ISGS replication samples were not considered because the Hardy-Weinberg test *P* value was $<5 \times 10^{-4}$ in controls. [g]SNP imputed in the UK discovery samples and not genotyped or imputed in the discovery and replication German controls.

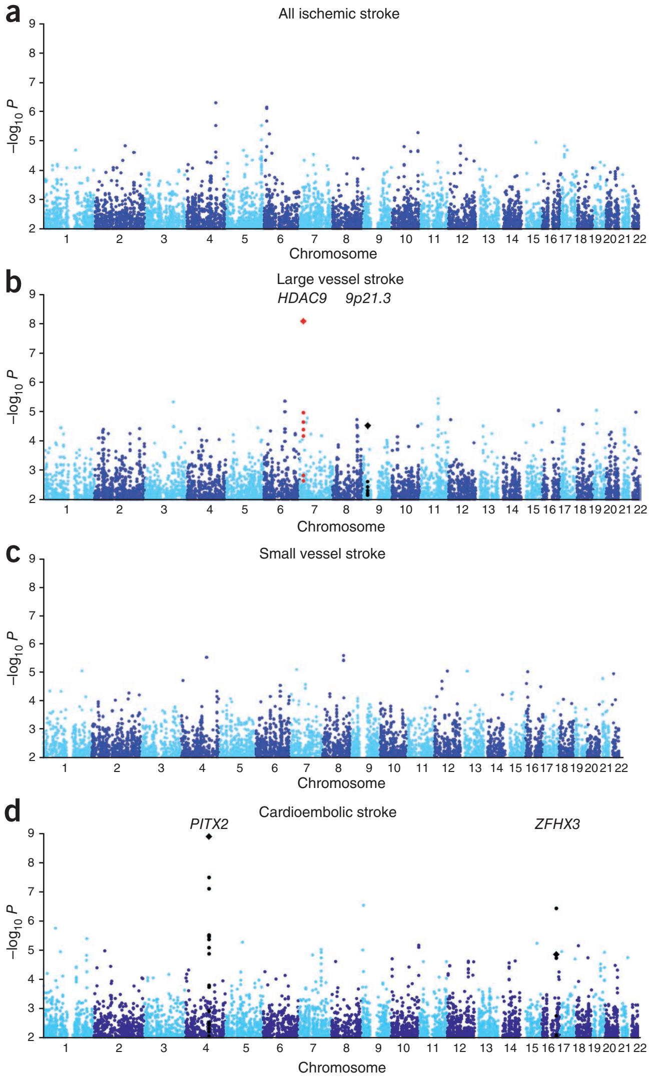

Figure 1 Genome-wide association results at autosomal SNPs in combined UK and German discovery samples. (**a**–**c**) Results are shown for all ischemic stroke (**a**), large vessel stroke (**b**), small vessel stroke (**c**) and cardioembolic stroke (**d**). Loci previously reported in the literature 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 region, a locus associated with heart disease and related phenotypes and previously associated with large vessel stroke. Most support was found for the model in which the effect sizes for the three stroke subtypes are correlated, but there was also substantial weight on the model in which there is only an effect for large vessel stroke. The same analyses in our data for the top SNPs in the regions previously associated with cardioembolic stroke (*PITX2* region, rs1906599, and *ZFHX3* region, rs12932445) showed strong support for the model in which these SNPs only affect risk for cardioembolic stroke. Together, these analyses provide compelling evidence for heterogeneity of genetic effects between stroke subtypes.

The association with rs11984041 in the *HDAC9* gene implicates a new locus in susceptibility to stroke. Any association with stroke could be mediated via associations with intermediate cardiovascular risk factors that themselves increase large vessel stroke risk. Our study design does not allow a direct assessment of this possibility, as data for these risk factors were not available. However, to date, no associations have been reported between rs11984041 or correlated SNPs and hypertension[15], hyperlipidemia[16] or diabetes[17] in large-scale GWAS of these risk factors.

All variants with an association signal in the region surrounding *HDAC9* resided within a peak between two recombination hotspots that encompasses the tail end of *HDAC9* (**Fig. 4**). The downstream *TWIST1* and *FERD3L* genes are physically relatively close to the identified peak and cannot be excluded as possible mechanisms through which genetic variants may exert *cis* effects on the large vessel stroke phenotype. *HDAC9* is a member of a large family of genes that encode proteins that deacetylate histones, thereby regulating chromatin structure and gene transcription[18]. *HDAC9* is ubiquitously expressed, with high levels of expression in cardiac tissue, muscle and brain[19]. Although known as histone deacetylases, these proteins also act on other substrates[20] and lead to both upregulation and downregulation of genes[21].

The mechanism by which variants in the *HDAC9* region increase large vessel stroke risk is not immediately clear. The specific association with this stroke subtype would be consistent with the association acting by accelerating atherosclerosis. The HDAC9 protein inhibits myogenesis and is involved in heart development[19], 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[22].

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

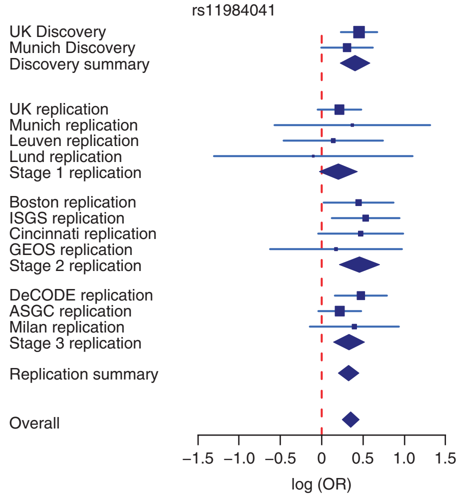

Figure 2 Forest plot for the associations between rs11984041 and large vessel stroke in discovery and replication collections. The blue lines show the 95% confidence intervals of the log (OR) for each cohort, with the area of each square proportional to the inverse of the standard error. The diamonds 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 ($P = 0.92$).

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; SNPTEST, 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/.

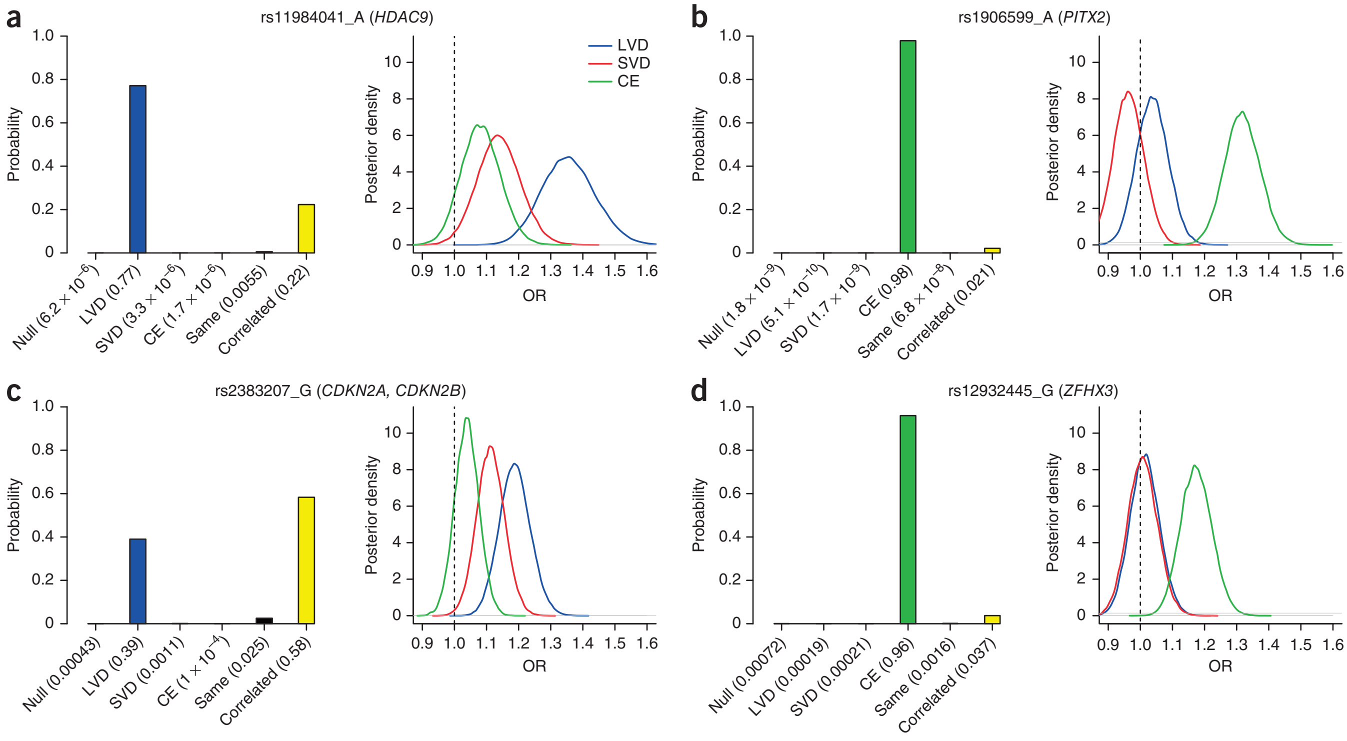

Figure 3 Genetic heterogeneity of different stroke subtypes for the four loci with significant associations. (**a**–**d**) Data are shown for *HDAC9* (**a**), *PITX2* (**b**), the 9p21 region (*CDKN2A*, *CDKN2B*) (**c**) and *ZFHX3* (**d**). Bar plots show the posterior probabilities on the models of 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).

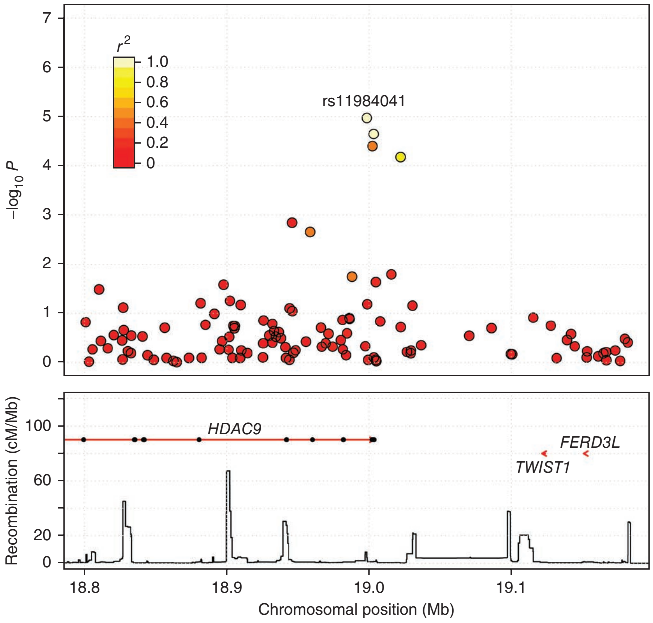

Figure 4 Plot of association signals around rs11984041 for large vessel stroke in the combined UK and German discovery samples. Top, SNPs are colored based on their correlation (r^2) with the labeled top SNP, which has the smallest *P* value in the region. r^2 is calculated from the WTCCC2 control data. 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 (068545/Z/02), and of the UK National Blood Service controls funded by the Wellcome Trust. We thank W. Bodmer and B. Winney for use of the People of the British Isles DNA collection, which was funded by the Wellcome Trust. We thank the following individuals who contributed to collection, phenotyping, sample processing and data management for the different cohorts: A. Burgess, A. Syed and N. Paul (Oxford Vascular Study); M. Dennis, P. Sandercock, C. Warlow, S. Hart, S. Keir, J. Wardlaw, A. Farrall, G. Potter, A. Hutchison and M. McDowall (Edinburgh Stroke Study); A. Pasdar and H. Clinkscale (Aberdeen); P. Higgins (Glasgow); T.G. Brott, R.D. Brown, S. Silliman, M. Frankel, D. Case, S. Rich, J. Hardy, A. Singleton (ISGS); M.J. Sparks, K. Ryan, J. Cole, M. Wozniak, B. Stern, R. Wityk, C. Johnson and D. Buchholz (GEOS); and J. Maguire, S. Koblar, J. Golledge, J. Surm, G. Hankey, J. Jannes, M. Lewis, R. Scott, L. Lincz; P. Moscato and R. Baker (Australian Stroke Genetics Collaborative membership). The principal funding for this study was provided by the Wellcome Trust as part of the WTCCC2 project (085475/B/08/Z, 085475/Z/08/Z and WT084724MA). For details of other funding support, see the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

S. Bevan, C.C.A.S., P.S., M.F., 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.J., T.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.V., J.S., K.F., H.D., S.G., G.T., U.T., K.S., G.B.B., E.A.P., J.A., E.H., C.L., M.-G.F., A.H., B.B.W., S.J.K., B.D.M., B.K., J.F.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., M.P., 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), L.P. (Deputy Chair), J.M.B., E.B., M.A.B., J.P.C., A.C., P. Deloukas, A.D., J.J., 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.

1. Department of Health. *Reducing Brain Damage: Faster Access to Better Stroke Care* (National Audit Office, London, 2005).
2. Sacco, R.L. *et al.* Infarcts of undetermined cause: the NINCDS Stroke Data Bank. *Ann. Neurol.* **25**, 382–390 (1989).
3. Dichgans, M. Genetics of ischaemic stroke. *Lancet Neurol.* **6**, 149–161 (2007).
4. Jerrard-Dunne, P., Cloud, G., Hassan, A. & Markus, H.S. Evaluating the genetic component of ischemic stroke subtypes: a family history study. *Stroke* **34**, 1364–1369 (2003).
5. Gretarsdottir, S. *et al.* Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann. Neurol.* **64**, 402–409 (2008).
6. Lemmens, R. *et al.* The association of the 4q25 susceptibility variant for atrial fibrillation with stroke is limited to stroke of cardioembolic etiology. *Stroke* **41**, 1850–1857 (2010).
7. Gschwendtner, A. *et al.* Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. *Ann. Neurol.* **65**, 531–539 (2009).
8. Markus, H.S. Genetics studies in ischaemic stroke. *Transl. Stroke Res.* **1**, 238–245 (2010).
9. 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).
10. Gudbjartsson, D.F. *et al.* A sequence variant in *ZFHX3* on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat. Genet.* **41**, 876–878 (2009).
11. International Stroke Genetics Consortium & Wellcome Trust Case-Control Consortium 2. Failure to validate association between 12p13 variants and ischemic stroke. *N. Engl. J. Med.* **22**, 1547–1550 (2010).
12. Olsson, S. *et al.* Genetic variant on chromosome 12p13 does mot show association to ischemic stroke in 3 Swedish case-control studies. *Stroke* **42**, 214–216 (2011).
13. Ikram, M.A. *et al.* Genomewide association studies of stroke. *N. Engl. J. Med.* **360**, 1718–1728 (2009).
14. Chen, K. *et al.* Strong association between the *NINJ2* gene polymorphism and the susceptibility of stroke in Chinese Han population in Fangshan district. *Beijing Da Xue Xue Bao* **42**, 498–502 (2010).
15. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* **41**, 666–676 (2009).
16. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **468**, 707–713 (2010).
17. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
18. Haberland, M., Montgomery, R.L. & Olson, E.N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat. Rev. Genet.* **10**, 32–42 (2009).
19. Chang, S. *et al.* Histone deacetylases 5 and 9 govern responsiveness of the heart to a subset of stress signals and play redundant roles in heart development. *Mol. Cell Biol* **24**, 8467–8476 (2004).
20. Kouzarides, T. Acetylation: a regulatory modification to rival phosphorylation? *EMBO J.* **19**, 1176–1179 (2000).
21. Glaser, K.B. *et al.* Gene expression profiling of multiple histone deacetylase (HDAC) inhibitors: defining a common gene set produced by HDAC inhibition in T24 and MDA carcinoma cell lines. *Mol. Cancer Ther.* **2**, 151–163 (2003).
22. Langley, B., Brochier, C. & Rivieccio, M.A. Targeting histone deacetylases as a multifaceted approach to treat the diverse outcomes of stroke. *Stroke* **40**, 2899–2905 (2009).
23. Kubo, M. *et al.* A nonsynonymous SNP in *PRKCH* (protein kinase C η) increases the risk of cerebral infarction. *Nat. Genet.* **39**, 212–217 (2007).

The authors of this paper are:
Céline Bellenguez[1,54], Steve Bevan[2,54], Andreas Gschwendtner[3], Chris C A Spencer[1], Annette I Burgess[4], Matti Pirinen[1], Caroline A Jackson[5], Matthew Traylor[2], Amy Strange[1], Zhan Su[1], Gavin Band[1], Paul D Syme[6], Rainer Malik[3], Joanna Pera[7], Bo Norrving[8,9], Robin Lemmens[10,11], Colin Freeman[1], Renata Schanz[12], Tom James[2], Deborah Poole[4], Lee Murphy[13], Helen Segal[4], Lynelle Cortellini[14,15], Yu-Ching Cheng[16,17], Daniel Woo[18], Michael A Nalls[19], Bertram Müller-Myhsok[20], Christa Meisinger[21], Udo Seedorf[22], Helen Ross-Adams[7], Steven Boonen[23], Dorota Wloch-Kopec[7], Valerie Valant[14,15], Julia Slark[12], Karen Furie[14], Hossein Delavaran[8,9], Cordelia Langford[24], Panos Deloukas[24], Sarah Edkins[24], Sarah Hunt[24], Emma Gray[24], Serge Dronov[24], Leena Peltonen[24,56], Solveig Gretarsdottir[25], Gudmar Thorleifsson[25], Unnur Thorsteinsdottir[25,26], Kari Stefansson[25,26], Giorgio B Boncoraglio[27], Eugenio A Parati[27], John Attia[28], Elizabeth Holliday[28], Chris Levi[28], Maria-Grazia Franzosi[29], Anuj Goel[1,30], Anna Helgadottir[1,25,30], Jenefer M Blackwell[31,32], Elvira Bramon[33], Matthew A Brown[34], Juan P Casas[35,36], Aiden Corvin[37], Audrey Duncanson[38], Janusz Jankowski[39,40], Christopher G Mathew[41], Colin N A Palmer[42], Robert Plomin[43], Anna Rautanen[1], Stephen J Sawcer[44], Richard C Trembath[41], Ananth C Viswanathan[45], Nicholas W Wood[46], Bradford B Worrall[47,48], Steven J Kittner[49,50], Braxton D Mitchell[16,17], Brett Kissela[18], James F Meschia[51], Vincent Thijs[10,11], Arne Lindgren[8,9], Mary Joan Macleod[6], Agnieszka Slowik[7], Matthew Walters[52], Jonathan Rosand[14,15], Pankaj Sharma[12], Martin Farrall[1,30], Cathie L M Sudlow[5], Peter M Rothwell[4], Martin Dichgans[3], Peter Donnelly[1,53,55] & Hugh S Markus[2,55]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Stroke and Dementia Research Group, St. George's University of London, London, UK. [3]Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians-Universität, Munich, Germany. [4]Stroke Prevention Research Unit, Nuffield Department of Clinical Neuroscience, University of Oxford, Oxford, UK. [5]Division of Clinical Neurosciences, University of Edinburgh, Edinburgh, UK. [6]Division of Applied Medicine, University of Aberdeen, Aberdeen, UK. [7]Department of Neurology, Jagiellonian University Medical College, Krakow, Poland. [8]Neurology, Department of Clinical Sciences, Lund University, Lund, Sweden. [9]Department of Neurology, Skåne University Hospital, Lund, Sweden. [10]Department of Neurology, University Hospitals Leuven, Leuven, Belgium. [11]Vesalius Research Center, Vlaams Instituut voor Biotechnologie (VIB), Leuven, Belgium. [12]Imperial College Cerebrovascular Research Unit (ICCRU), Imperial College London, London, UK. [13]Wellcome Trust Clinical Research Facility Genetics Core Laboratory, University of Edinburgh, Western General Hospital, Edinburgh, UK. [14]Department of Neurology, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [15]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [16]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA. [17]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA. [18]Department of Neurology, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA. [19]Laboratory of Neurogenetics, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA. [20]Max Planck Institute of Psychiatry, Munich, Germany. [21]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology II, Neuherberg, Germany. [22]Leibniz–Institut für Arterioskleroseforschung, Universität Münster, Münster, Germany. [23]Division of Geriatric Medicine, University Hospitals Leuven, Leuven, Belgium. [24]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [25]deCODE Genetics, Reykjavik, Iceland. [26]University of Iceland, Faculty of Medicine, Reykjavik, Iceland. [27]Fondazione Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Istituto Neurologico Carlo Besta, Milan, Italy. [28]Centre for Brain and Mental Health Research, University of Newcastle, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [29]Department of Cardiovascular Research, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy. [30]Department of Cardiovascular Medicine, University of Oxford, Oxford, UK. [31]Centre for Child Health Research, University of Western Australia, West Perth, Western Australia, Australia. [32]Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge, UK. [33]Division of Psychological Medicine and Psychiatry, Biomedical Research Centre for Mental Health, Institute of Psychiatry, King's College London, London, UK. [34]The University of Queensland Diamantina Institute, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [35]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [36]Department of Epidemiology and Public Health, University College London, London, UK. [37]Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland. [38]Molecular and Physiological Sciences, The Wellcome Trust, London, London, UK. [39]Centre for Gastroenterology, Barts and the London School of Medicine and Dentistry, London, UK. [40]Division of Clinical Pharmacology, University of Oxford, Oxford, UK. [41]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK. [42]Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee, UK. [43]Social, Genetic and Developmental Psychiatry Centre, King's College London Institute of Psychiatry, Denmark Hill, London, UK. [44]University of Cambridge Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, UK. [45]National 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. [46]Department of Molecular Neuroscience, Institute of Neurology, Queen Square, University College London, London, UK. [47]Department of Neurology, University of Virginia School of Medicine, Charlottesville, Virginia, USA. [48]Department of Public Health Science, University of Virginia School of Medicine, Charlottesville, Virginia, USA. [49]Department of Neurology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [50]Baltimore Geriatric Research, Education, and Clinical Center, Baltimore Veterans Affairs Medical Center, Baltimore, Maryland, USA. [51]Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA. [52]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. [53]Department of Statistics, University of Oxford, Oxford, UK. [54]These authors contributed equally to this work. [55]These authors jointly directed this work. [56]Deceased. 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[9], 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-Quad (a custom chip designed by WTCCC2 comprising Human550 SNPs and ~6,000 common copy-number variants (CNVs) from the Structural Variation Consortium[24]). Samples from UK control collections were genotyped on the Human1.2M-Duo (a WTCCC2 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 Illuminus[25]. 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[26,27]. 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[28] 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 with the genome-wide data. Our final discovery dataset consisted of 3,548 cases (2,374 UK and 1,174 German) and 5,972 controls (5,175 UK and 797 German) following sample quality control analysis (**Supplementary Table 4**). A full breakdown of samples by cohort and subtype is in provided **Table 1**. For quality control analysis of SNPs, a measure of (Fisher) information for allele frequency at each SNP was calculated using SNPTEST (see URLs). Autosomal SNPs were excluded if this information measure was below 0.98, if minor allele frequency was <0.01%, if the SNP had >5% missing data or if Hardy-Weinberg P value was $<1 \times 10^{-20}$ in the case or control collections. In the 58C, UKBS and case datasets, association between SNPs and the plate on which samples were genotyped was calculated, and SNPs with a plate effect P value of $<1 \times 10^{-6}$ were also excluded. An additional 45 SNPs were removed following visual inspection of cluster plots. A breakdown of the number of SNPs excluded is provided in **Supplementary Table 5**. Only SNPs genotyped in all the case and control collections were considered, leaving 495,851 autosomal SNPs after quality control. Hardy-Weinberg P values for the SNPs taken to replication are given in **Supplementary Table 6**.

Initial replication genotyping and **quality control.** Genotyping of European replication samples was carried out at the WTSI using Sequenom iPLEX Gold 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-Quad chips. Imputation to HapMap 3 using the BEAGLE software program[29] 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).

Association analysis. We performed single SNP 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 SNPTEST. 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[30]. 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

doi:10.1038/ng.1081

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 5′ 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).

doi:10.1038/ng.1081