nature genetics

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

Elizabeth G Holliday[1,2], Jane M Maguire[3–5], Tiffany-Jane Evans[2,6], Simon A Koblar[7,8], Jim Jannes[7,8], Jonathan W Sturm[5,9], Graeme J Hankey[10,11], Ross Baker[12,13], Jonathan Golledge[14,15], Mark W Parsons[4], Rainer Malik[16], Mark McEvoy[1,9,17], Erik Biros[14], Martin D Lewis[7,18], Lisa F Lincz[4,6,19], Roseanne Peel[1,9,17], Christopher Oldmeadow[9,20], Wayne Smith[9,17], Pablo Moscato[2,21], Simona Barlera[22], Steve Bevan[23], Joshua C Bis[24], Eric Boerwinkle[25,26], Giorgio B Boncoraglio[27], Thomas G Brott[28], Robert D Brown Jr[29], Yu-Ching Cheng[30], John W Cole[31,32], Ioana Cotlarciuc[33], William J Devan[34–36], Myriam Fornage[25,26], Karen L Furie[35,36], Sólveig Grétarsdóttir[37], Andreas Gschwendtner[16], M Arfan Ikram[38–40], W T Longstreth Jr[41–43], James F Meschia[28], Braxton D Mitchell[30], Thomas H Mosley[44], Michael A Nalls[45], Eugenio A Parati[27], Bruce M Psaty[24,41,46,47], Pankaj Sharma[33], Kari Stefansson[37,48], Gudmar Thorleifsson[37], Unnur Thorsteinsdottir[37,48], Matthew Traylor[23], Benjamin F J Verhaaren[38,40], Kerri L Wiggins[24], Bradford B Worrall[49], The Australian Stroke Genetics Collaborative[50], The International Stroke Genetics Consortium[50], The Wellcome Trust Case Control Consortium 2[50], Cathie Sudlow[51,52], Peter M Rothwell[53], Martin Farrall[54,55], Martin Dichgans[16], Jonathan Rosand[34–36], Hugh S Markus[23], Rodney J Scott[2,6,56,57], Christopher Levi[4,57] & John Attia[1,2,57]

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 1,162 ischemic stroke cases (including 421 LAA cases) and 1,244 population controls from Australia. Evidence for a genetic influence on ischemic stroke risk was detected, but this influence was higher and more significant for the LAA subtype. We identified a new LAA susceptibility locus on chromosome 6p21.1 (rs556621: odds ratio (OR) = 1.62, $P = 3.9 \times 10^{-8}$) and replicated this association in 1,715 LAA cases and 52,695 population controls from 10 independent population cohorts (meta-analysis replication OR = 1.15, $P = 3.9 \times 10^{-4}$; discovery and replication combined OR = 1.21, $P = 4.7 \times 10^{-8}$). This study identifies a genetic risk locus for LAA and shows how analyzing etiological subtypes may 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. Although 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: $h^2 \approx 30–60\%$)[7,8] and white matter lesions ($h^2 \approx 50–70\%$)[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[15], 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: (i) large artery atherosclerosis (LAA); (ii) cardioembolism (CE); and (iii) small vessel occlusion (SVO)[16]. 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[15], 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 1,230 cases and 1,280 population controls. The causal subtype of ischemic stroke was classified using TOAST criteria[16]. Demographic and clinical characteristics of the Australian Stroke Genetics Collaborative (ASGC) data set are summarized in **Supplementary Table 1**.

A full list of affiliations appears at the end of the paper.

Received 18 January; accepted 9 August; published online 2 September 2012; doi:10.1038/ng.2397

Table 1 Proportion of case-control phenotypic variation explained by genome-wide SNP data for all ischemic stroke, LAA, SVO and CE

Phenotype	Cases[a]	Controls[a]	σ_g^2/σ_p^2 (s.e.)[b]	LRT[c]	*P* value[d]
Ischemic stroke	1,079	1,172	0.39 (0.15)	11.04	4.5×10^{-4}
LAA	400	1,172	0.66 (0.21)	14.94	5.6×10^{-5}
SVO	288	1,172	0.10 (0.24)	0.20	3.3×10^{-1}
CE	226	1,172	0.60 (0.25)	7.78	2.6×10^{-3}

Genetic relationships between individuals were estimated using 457,533 SNPs. [a]Smaller sample sizes compared with the GWAS because of additional quality control filtering conducted before this analysis. [b]Estimated proportion (standard error, s.e.) of variation in case-control status explained by all SNPs. [c]Likelihood ratio test (LRT) statistic. [d]*P* values were calculated assuming that the LRT is distributed as a 50:50 mixture of a point mass at zero and χ_1^2 under the null hypothesis.

After quality control filtering of genotype data, data on 551, 514 SNPs from 1,162 ischemic stroke cases and 1,244 controls were used for genotype imputation and genetic analysis. Before performing genome-wide association analyses, we assessed the genetic contribution to ischemic stroke and the LAA, CE and SVO subtypes using a recent method[18] that estimates the proportion of phenotypic variance (V_g/V_p) attributable to variation in genotyped SNPs, where V_g is the component of phenotypic variance attributable to variation in genotyped SNPs and V_p is the total observed phenotypic variance. For ischemic stroke, the estimated genetic load was substantial ($V_g/V_{p\,\mathrm{IS}} = 0.39$), with SNPs explaining a significant proportion of phenotypic variation ($P = 4.5 \times 10^{-4}$). For cases classified in the LAA subtype, we observed a higher, more significant estimate of genetic load ($V_g/V_{p\,\mathrm{LAA}} = 0.66$; $P = 5.6 \times 10^{-5}$), consistent with previous reports of high familial risk for LAA[6]. Evidence for genetic contribution was less significant for the CE and SVO subtypes ($V_g/V_{p\,\mathrm{CE}} = 0.6$, $P = 0.0026$ and $V_g/V_{p\,\mathrm{SVO}} = 0.1$, $P = 0.33$, respectively; **Table 1**).

We performed two primary GWAS in the Australian discovery sample, comparing (i) all ischemic stroke cases (n = 1,162) and (ii) LAA cases (n = 421) with population controls (n = 1,244). GWAS of the CE and SVO subtypes, which both had fewer cases and a less significant V_g/V_p estimate, were performed as supplementary analyses (**Supplementary Figs. 1** and **2** and **Supplementary Tables 2** and **3**). Genotypic effects were estimated using logistic regression models (1-degree-of-freedom additive trend tests) adjusted for age and sex. Results were compared with a prespecified significance threshold of 5×10^{-8}, corresponding to Bonferroni adjustment for 1×10^6 independent tests. Quantile-quantile plots indicated excellent quality of the GWAS data and an absence of systematic bias caused by population substructure or other artifacts (**Supplementary Fig. 3**).

Analyses of ischemic stroke detected the strongest signals at several SNPs within the *SLC5A4* gene on chromosome 22q12.3 (**Fig. 1**, **Supplementary Fig. 4** and **Supplementary Table 4**). Peak association was detected at rs5998322 ($P_{\mathrm{trend}} = 3.91 \times 10^{-7}$; OR = 1.97, 95% confidence interval (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: $P_{\mathrm{trend}} = 3.23 \times 10^{-6}$; 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 ($P_{\mathrm{trend}} = 4.47 \times 10^{-7}$), suggesting independence of the two associated loci at 22q.

The GWAS of LAA detected two associated SNPs on chromosome 6p21.1 exceeding the prespecified threshold for genome-wide significance ($\alpha = 5 \times 10^{-8}$; **Figs. 1** and **2**). These variants, rs556621 ($P_{\mathrm{trend}} = 3.92 \times 10^{-8}$; OR (A allele) = 1.62, 95% CI = 1.36–1.93) and rs556512 ($P_{\mathrm{trend}} = 4.25 \times 10^{-8}$; OR (A allele) = 1.62, 95% CI = 1.36–1.93) were in perfect linkage disequilibrium (LD) in HapMap Phase 2 Utah residents of Northern and Western European ancestry (CEU) data ($r^2 = 1$, $D' = 1$; **Supplementary Table 5**), with a minor (A) allele population frequency of 0.33. The rs556621 SNP was directly genotyped in our sample, whereas 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 (**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

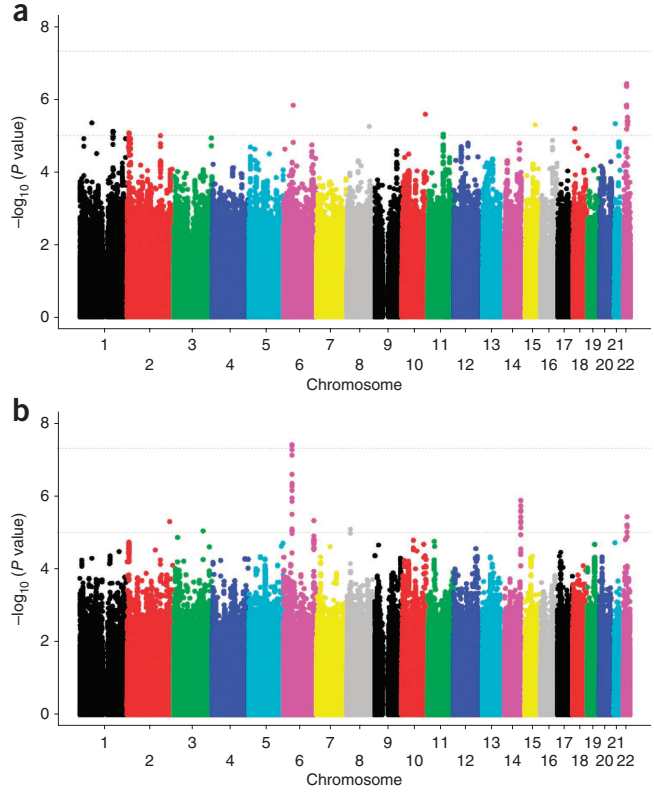

Figure 1 Genome-wide association results. (**a**,**b**) Data are shown for ischemic stroke (**a**) and LAA (**b**). The plots show $-\log_{10}$-transformed *P* values for genotyped and imputed SNPs with respect to their physical positions. The threshold for association at genome-wide significance ($P = 5 \times 10^{-8}$) is shown by the upper dashed line, and the lower dashed line corresponds to $P = 1 \times 10^{-5}$.

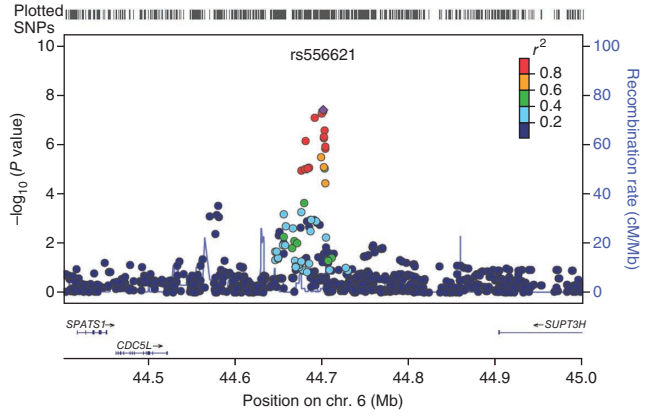

Figure 2 Regional association results for the chromosome 6p21.1 locus showing association at genome-wide significance with LAA. The index associated SNP is labeled (rs556621: $P = 3.9 \times 10^{-8}$).

Table 2 Association of rs556621 with LAA and overall ischemic stroke in discovery, replication and combined cohorts

	Discovery			Replication			Combined discovery and replication		
Phenotype	P^a	OR[b] (95% CI)	N_{cases}; $N_{controls}$	P	OR (95% CI)	N_{cases}; $N_{controls}$	P	OR (95% CI)	N_{cases}; $N_{controls}$
LAA	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

rs556621 is located at 6p21.1 (44,702,137 bp) according to chromosome and NCBI Human Genome Build 36.3 coordinates, and *CDC5L* and *SUPT3H* are the closest genes. The minor allele A has a risk allele frequency (RAF) in controls of 0.30.
[a]*P* value from 1-degree-of-freedom trend test. [b]Odds ratio with 95% confidence interval for the effect of each additional copy of the minor allele, assuming an additive effect on the log-odds scale.

broad ischemic stroke phenotype, with peak association also detected at rs556621 ($P = 5.6 \times 10^{-5}$; OR (A allele) = 1.29, 95% CI = 1.14–1.47) (**Table 2**). Supplementary analyses of CE and SVO subtypes revealed no association with rs556621 ($P = 0.73$ and 0.39, respectively; data not shown). In addition to the 6p21.1 locus, the LAA GWAS also detected clusters of suggestively associated SNPs ($P < 1 \times 10^{-5}$) at 14q32.33 and the second 22q12.3 locus detected in the GWAS of ischemic stroke (**Supplementary Fig. 5** and **Supplementary Table 7**).

In a subsequent LAA GWAS adjusted for rs556621 genotype, no SNP showed evidence of strong independent association with LAA (peak $P = 5.6 \times 10^{-6}$ for rs11625862 at 14q32.33). Haplotype association tests across the 6p21.1 region also did not detect multi-marker haplotypes that were more strongly associated with LAA than the two index SNPs (data 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 (ROC) curve (ΔAUC = 0.01; $P = 1.2 \times 10^{-5}$; **Supplementary Table 8**), although this ΔAUC estimate may be inflated by estimation in the discovery cohort. To further assess the internal validity of the association at rs556621, the sample was randomly partitioned into training and test groups containing two-thirds and one-third of the LAA cases and controls, respectively. Association with LAA was evaluated in the training set, with genotyped SNPs reaching $P < 1 \times 10^{-4}$ ($n = 44$) then assessed in the test set (the remaining third of the sample). The index SNP at 6p21.1 (rs556621) reached $P = 5.69 \times 10^{-5}$ in the training set and was the only SNP associated with LAA in the independent test set after permutation-based adjustment for the testing of 44 non-independent SNPs (familywise adjusted $P = 6.74 \times 10^{-3}$; **Supplementary Table 9**).

External validity of the observed association of rs556621 with LAA risk was assessed in a replication study involving 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 **Supplementary Table 10** and the **Supplementary Note**. Association analyses for the index SNP at 6p21.1 (rs556621) were performed separately within each of the ten 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 prespecified significance threshold of 0.05.

The replication study confirmed association of rs556621 with LAA ($P_{trend} = 3.9 \times 10^{-4}$; OR (A allele) = 1.15, 95% CI = 1.06–1.24), with no evidence of between-study heterogeneity ($P = 0.50$, $I^2 = 0.0\%$) (**Fig. 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 $P_{trend} = 4.7 \times 10^{-8}$ for the association (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 five-fold increase in case numbers (9,552 cases and 52,695 controls; $P_{trend} = 0.29$; OR (A allele) = 1.02, 95% CI = 0.98–1.06; **Supplementary Fig. 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.

The 6p21.1 SNPs are located in an intergenic region of moderate LD (**Supplementary Fig. 7**), ~200 kb upstream of the *SUPT3H* gene (forward strand) and ~180 kb upstream of *CDC5L* (reverse strand). rs556621 and rs556512 both lie within a small length of genomic sequence that contains BCL3 and PBX3 transcription factor–binding motifs and enriched for enhancer- and/or 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 (**Supplementary Table 12**), suggesting that variants in LD with rs556621 could also potentially regulate gene expression through alteration of regulatory miRNA sequences. Queries of four public expression quantitative trait locus (eQTL) databases did not identify rs556621 or proxy SNPs in high LD with rs556621 as *cis* eQTLs in the assayed tissue or cell types. Future targeted investigations in atherosclerotic neurovascular tissue may help to elucidate the mechanisms by which the associated SNPs influence LAA risk.

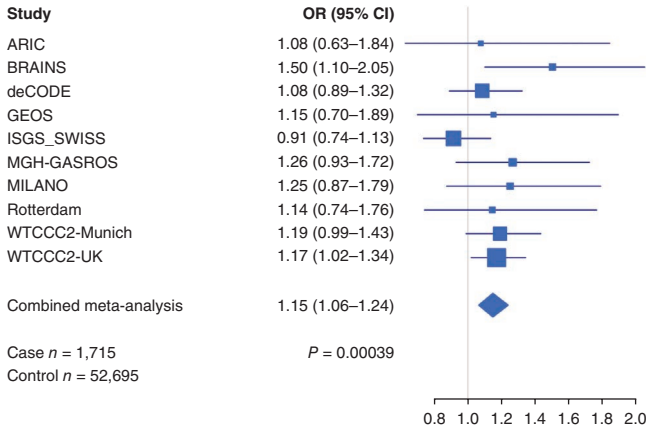

Figure 3 Forest plot showing association of rs556621 with LAA across the ten replication cohorts. For each cohort, the square and horizontal line show the estimated OR and 95% CI, respectively, representing the effect of each additional copy of the risk (A) allele on the odds of disease. The size of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

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 proinflammatory cytokines[22–24]. *APOL2*, *APOL3* and *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 report the identification of variants at 6p21.1 that associate with LAA risk in individuals of European ancestry. 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 contributions to stroke risk in non-European populations remain 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://unphased.sourceforge.net/; LocusZoom, http://csg.sph.umich.edu/locuszoom/; METAL, http://www.sph.umich.edu/csg/abecasis/Metal/; SNAP, http://www.broadinstitute.org/mpg/snap/; SCAN–SNP and CNV Annotation Database, http://scan.bsd.uchicago.edu/newinterface/about.html; NCBI GTEx (Genotype-Tissue Expression) eQTL Browser, http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi; Pritchard laboratory UChicago eQTL browser, http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/; mRNA by SNP Browser, http://www.sph.umich.edu/csg/liang/asthma/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

A complete list of funding acknowledgments is included in the **Supplementary Note**. We are grateful to the participants with ischemic stroke and also to their families for participating in this study. Australian population control data were 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 Health and Medical Research Council (NHMRC; project grant 569257), the Australian National Heart Foundation (NHF; project grant G 04S 1623), the University of Newcastle, the Gladys M Brawn Fellowship scheme and the Vincent Fairfax Family Foundation in Australia. E.G.H. is supported by the Australian NHMRC Fellowship scheme. J.G. 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 Wellcome Trust Case Control Consortium 2 (WTCCC2) ischemic stroke study was provided by the Wellcome Trust, as part of the WTCCC2 project (085475/B/08/Z, 085475/Z/08/Z and WT084724MA). This work was also supported by the European Community's Sixth Framework Programme (LSHM-CT-2007-037273), the Wellcome Trust core award (090532/Z/09/Z) and AstraZeneca. M. Farrall is a member of the Oxford British Heart Foundation (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 US National Institute of Neurological Disorders and Stroke. Additional funding was provided by the US National Institute of Neurological Disorders and Stroke (U01NS069208). The Rotterdam Study received principal funding for this report from the Netherlands Heart Foundation (grant 2009B102).

AUTHOR CONTRIBUTIONS

S.A.K., J.W.S., L.F.L., P.M., R.J.S., C.L. and J.A. designed the study. E.G.H. performed statistical analyses in the discovery cohort, meta-analyses of replication data and wrote the first draft of the manuscript. T.-J.E. and R.J.S. coordinated genotyping of the discovery cohort. J.M.M., J.G., J.J., G.J.H., R.B., M.W.P., J.W.S., L.F.L., C.L., M.M., R.P., W.S. and J.A. performed phenotype collection and data management in the Australian sample. E. Biros, M.D.L. and C.O. performed bioinformatic analyses. Replication data were provided by S. Barlera, S. Bevan, J.C.B., E. Boerwinkle, G.B.B., T.G.B., R.D.B., Y.-C.C., J.W.C., I.C., W.J.D., M. Fornage, K.L.F., S.G., A.G., M.A.I., W.T.L., R.M., J.F.M., B.D.M., T.H.M., M.A.N., E.A.P., B.M.P., P.S., K.S., G.T., M.T., U.T., B.F.J.V., K.L.W., B.B.W., C.S., P.M.R., M. Farrall, M.D., J.R. and H.S.M. All authors critically reviewed the manuscript and gave advice on the contents of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2397.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. World Health Organization. *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**, 355–369 (2009).
3. Strong, K., Mathers, C. & Bonita, R. Preventing stroke: saving lives around the world. *Lancet Neurol.* **6**, 182–187 (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**, 112–123 (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**, 212–227 (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**, 1364–1369 (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**, 483–487 (2004).
10. Carmelli, D. *et al.* Evidence for genetic variance in white matter hyperintensity volume in normal elderly male twins. *Stroke* **29**, 1177–1181 (1998).
11. Kääb, S. *et al.* Large scale replication and meta-analysis of variants on chromosome 4q25 associated with atrial fibrillation. *Eur. Heart J.* **30**, 813–819 (2009).
12. Gretarsdottir, S. *et al.* Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann. Neurol.* **64**, 402–409 (2008).
13. Palomaki, G.E., Melillo, S. & Bradley, L.A. Association between 9p21 genomic markers and heart disease: a meta-analysis. *J. Am. Med. Assoc.* **303**, 648–656 (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**, 159–164 (2009).
15. Bellenguez, C. *et al.* Genome-wide association study identifies a variant in *HDAC9* associated with large vessel ischemic stroke. *Nat. Genet.* **44**, 328–333 (2012).
16. 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).
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**, 108–112 (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**, 71–78 (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**, 1231–1236 (2000).
22. Horrevoets, A.J. *et al.* Vascular endothelial genes that are responsive to tumor necrosis factor-α *in vitro* are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes. *Blood* **93**, 3418–3431 (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**, 539–546 (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**, 256–269 (2005).

[1]Centre for Clinical Epidemiology and Biostatistics, School of Medicine and Public Health, University of Newcastle, 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, Newcastle, New South Wales, Australia. [4]Centre for Brain and Mental Health Research, University of Newcastle and Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [5]Department of Neurosciences, Gosford Hospital, Central Coast Area Health, Gosford, New South Wales, Australia. [6]School of Biomedical Sciences and Pharmacy, University of Newcastle, Newcastle, New South Wales, Australia. [7]Stroke Research Program, School of Medicine, University of Adelaide, Adelaide, South Australia, Australia. [8]Stroke Unit, Department of Neurology, Queen Elizabeth Hospital, Adelaide, South Australia, Australia. [9]School of Medicine and Public Health, University of Newcastle, Newcastle, New South Wales, Australia. [10]Department of Neurology, Royal Perth Hospital, Perth, Western Australia, Australia. [11]School of Medicine and Pharmacology, University of Western Australia, Perth, Western Australia, Australia. [12]Department of Haematology, Royal Perth Hospital, Perth, Western Australia, Australia. [13]Centre for Thrombosis and Haemophilia, Murdoch University, Perth, Western Australia, Australia. [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]Public Health Research Program, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [18]Discipline of Genetics, School of Molecular & Biomedical Sciences, University of Adelaide, Adelaide, South Australia, Australia. [19]Hunter Haematology Research Group, Calvary Mater Newcastle Hospital, Newcastle, New South Wales, Australia. [20]Clinical Research Design, IT and Statistical Support Unit, Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [21]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle, New South Wales, Australia. [22]Department of Cardiovascular Research, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy. [23]Stroke and Dementia Research Centre, St. George's University of London, London, UK. [24]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, Washington, USA. [25]Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, Texas, USA. [26]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA. [27]Department of Cerebrovascular Diseases, Fondazione Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Istituto Neurologico Carlo Besta, Milan, Italy. [28]Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA. [29]Department of Neurology, Mayo Clinic, Rochester, Minnesota, USA. [30]Department of Medicine, University of Maryland, Baltimore, Maryland, USA. [31]Baltimore Veterans Affairs Medical Center, Baltimore, Maryland, USA. [32]School of Medicine, University of Maryland, Baltimore, Maryland, USA. [33]Imperial College Cerebrovascular Research Unit (ICCRU), Imperial College London, London, UK. [34]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [35]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [36]Department of Neurology, Harvard Medical School, Boston, Massachusetts, USA. [37]deCODE Genetics, Reykjavik, Iceland. [38]Department of Epidemiology, Erasmus Medical Center (MC)–University Medical Center, Rotterdam, The Netherlands. [39]Department of Neurology, Erasmus MC–University Medical Center, Rotterdam, The Netherlands. [40]Department of Radiology, Erasmus MC–University Medical Center, Rotterdam, The Netherlands. [41]Department of Epidemiology, University of Washington, Seattle, Washington, USA. [42]Department of Medicine, University of Washington, Seattle, Washington, USA. [43]Department of Neurology, University of Washington, Seattle, Washington, USA. [44]Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA. [45]Laboratory of Neurogenetics, National Institute on Aging, US National Institutes of Health, Bethesda, Maryland, USA. [46]Department of Health Services, University of Washington, Seattle, Washington, USA. [47]Group Health Research Institute, Group Health, Seattle, Washington, USA. [48]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [49]Department of Neurology, University of Virginia, Charlottesville, Virginia, USA. [50]A full list of members is provided in the **Supplementary Note**. [51]Division of Clinical Neurosciences, University of Edinburgh, Edinburgh, UK. [52]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [53]Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK. [54]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [55]Department of Cardiovascular Medicine, University of Oxford, Oxford, UK. [56]Division of Genetics, Hunter Area Pathology Service, Newcastle, New South Wales, Australia. [57]These authors jointly directed this work. Correspondence should be addressed to E.G.H. (liz.holliday@newcastle.edu.au).

ONLINE METHODS

Study participants: the ASGC discovery sample. ASGC stroke cases comprised stroke patients of European ancestry who were admitted to four clinical centers across Australia (The Neurosciences Department at Gosford Hospital, Gosford; the Neurology Department at John Hunter Hospital, Newcastle; The Queen Elizabeth Hospital, Adelaide; and the Royal Perth Hospital, Perth) between 2003 and 2008. Stroke was defined by World Health Organization criteria as a sudden focal neurological deficit of vascular origin, lasting more than 24 h and confirmed by imaging, such as computerized tomography (CT) and/or magnetic resonance imaging (MRI) brain scan. Other investigative tests such as electrocardiogram, carotid doppler and trans-esophageal echocardiogram were conducted to define ischemic stroke mechanism as clinically appropriate. Cases were excluded from participation if they were aged <18 years, were diagnosed with hemorrhagic stroke or had transient ischemic attack rather than ischemic stroke or if they were unable to undergo baseline brain imaging. On the basis of these criteria, a total of 1,230 ischemic stroke cases were included in the current study. Ischemic stroke subtypes were assigned using TOAST criteria on the basis of clinical, imaging and risk factor data[16].

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 in New South Wales, Australia. Detailed recruitment methods for the HCS have been previously described[25]. Briefly, participants were randomly selected from the New South Wales State electoral roll and were contacted by mail between 2004 and 2007. Consenting participants completed five detailed self-report questionnaires and attended the HCS data collection center, at which time a series of clinical measures were obtained. A total of 1,280 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 the relevant institutional ethics committees.

Study participants: replication cohorts. Replication data were contributed by a total of 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 Massachusetts General Hospital 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 those used for the ASGC discovery sample. Summary demographic data and clinical phenotyping details for these individual cohorts are provided in **Supplementary Table 2** and the **Supplementary Note**.

Genome-wide 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 of <0.95, deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) or minor allele frequency of <0.01. At the sample level, quality control excluded individuals with (i) genotype call rate of <95% ($n = 4$); (ii) genome-wide heterozygosity of <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 identified on the basis of pairwise allele sharing estimates (estimated genome proportion shared identical by descent (IBD); $\hat{\pi} > 0.1875$; $n = 37$). After these exclusions, Eigenstrat principal-components analysis (PCA) was performed, incorporating genotype data from Phase 3 HapMap populations (CEU, Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT), Toscani in Italia (TSI) and Yoruba from Ibadan, Nigeria (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 LD. After quality control, 1,162 cases and 1,244 controls were available for association analyses at 551,514 SNPs.

Genotype imputation in the filtered sample was performed using MACH v1.0.16 on the basis of HapMap Phase 2 (release 24) phased haplotypes for samples of European ancestry (CEU). Subsequent quality control excluded imputed SNPs with minor allele frequency of <0.01 or ratio of observed dosage variance to expected binomial variance (r^2) of <0.3.

Genotyping and quality control: replication cohorts. Each replication cohort performed genome-wide genotyping, quality control and imputation as part of its own primary study. The particular arrays and quality control filters used by the individual cohorts are described in the **Supplementary Note**. Of the 11 cohorts, 6 directly genotyped rs556621, and 5 imputed allelic dosages for this SNP. To ensure the accuracy of results, imputed data were only included if the quality of imputation was high, defined as a ratio of observed to expected binomial dosage variance (r^2) of >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 with 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. Before analysis, additional quality control of genotype data was performed to reduce bias in variance estimates from the accrued effects of small genotyping errors[27]. We excluded SNPs with missingness of >0.1% or Hardy-Weinberg equilibrium P value of $< 1 \times 10^{-4}$ and individuals with >0.1% missing genotype data or estimated relatedness of >0.05 (approximately closer than second cousins)[27]. After quality control, genotypes at 457,533 SNPs were available for estimating genetic effects for 1,079 ischemic stroke cases, 400 LAA cases, 288 SVO cases and 226 CE cases. Each case group was evaluated in a separate analysis using a common control sample of 1,172 individuals; all fitted LMMs were adjusted for age and sex. Heritability estimates shown in **Table 1** relate to the observed (binary) risk scale and case-control proportions. 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 1-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 ($\lambda = 1.031$, $\lambda_{1,000} = 1.026$ for ischemic stroke; $\lambda = 1.007$, $\lambda_{1,000} = 1.011$ 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 standardized values for a sample of 1,000 cases and 1,000 controls ($\lambda_{1,000}$)[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 LD between SNPs was assessed and visualized using Haploview software[32] on the basis of 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 two to six 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].

doi:10.1038/ng.2397

Meta-analysis of rs556621 in replication cohorts. For rs556621, each replication sample performed logistic regression using a 1-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 β 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 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

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

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

Predictive modeling using ROC curves. Predictive models incorporating clinical and genetic risk factors were evaluated for their ability to discriminate between case and control participants by calculating the area under the receiver operator characteristic (ROC) curve (AUC). ROC curves show the relationship between sensitivity (true positive rate) and 1-specificity (false negative rate) for all possible cut-points of a diagnostic test. For specified covariates, the ROC curve was fitted and the AUC calculated using Stata software[36], on the basis of parameter estimates from logistic regression models. Likelihood-ratio tests were used to assess the significance of changes in model fit.

eQTL analyses. For the lead SNP at 6p21.1 (rs556621), proxy SNPs with r^2 of >0.8 were identified from HapMap CEU Phases 1 and 2 (release 22) and 3 data (release 2) using SNAP (v2.2). Four publicly available eQTL databases were searched to determine whether genotypes of the lead or proxy SNPs have been previously associated with gene expression in *cis* in a range of tissue and cell types. We defined potential *cis* eQTLs as candidate SNPs associated with expression of a gene transcript mapping to a genomic region within 1 Mb[37] at a nominal significance level of 1×10^{-3}. The databases searched were (i) SCAN–SNP and CNV Annotation Database; (ii) the NCBI GTEx (Genotype-Tissue Expression) eQTL Browser; (iii) the Pritchard laboratory UChicago eQTL browser; and (iv) mRNA by SNP Browser v1.0.1. The tissue and cell types assessed in these databases include liver, brain, lymphoblastoid cell lines (LCLs), monocytes, fibroblasts and T cells.

25. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int. J. Epidemiol.* **39**, 1452–1463 (2010).
26. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (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**, 51–54 (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–R128 (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**, 816–834 (2010).
32. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (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**, 2336–2337 (2010).
35. Zheng, S.L. *et al.* Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* **358**, 910–919 (2008).
36. StataCorp. *Stata: Release 11. Statistical Software.* (StataCorp LP, College Station, Texas, 2009).
37. Webster, J.A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* **84**, 445–458 (2009).