COL4A2 is associated with lacunar ischemic stroke and deep ICH: Meta-analyses among 21,500 cases and 40,600 controls

Title: *COL4A2* is associated with lacunar ischemic stroke and deep ICH: meta-analyses among 21,500 cases and 40,600 controls.

Kristiina Rannikmäe (MD, PhD)[1], Vhinoth Sivakumaran[2], Henry Millar[2], Rainer Malik (PhD)[3], Christopher D Anderson (MD, MMSc)[4-7], Mike Chong (MSc)[8], Tushar Dave (MS)[9], Guido J Falcone (MD, ScD, MPH)[10], Israel Fernandez-Cadenas (PhD)[11], Jordi Jimenez-Conde (MD, PhD)[12,13], Arne Lindgren (MD, PhD)[14,15], Joan Montaner (MD, PhD)[16], Martin O'Donnell (MD, PhD, MRCPI, BCh, BAO)[8,17], Guillaume Paré (MD, MSc, FRCPC)[8], Farid Radmanesh (MD, MPH)[4-7], Natalia S Rost (MD, MPH, FAAN, FAHA)[5], Agnieszka Slowik (MD, PhD)[18], Martin Söderholm (MD)[15,19], Matthew Traylor (PhD)[20], Sara L Pulit (PhD)[21], Sudha Seshadri (MD)[22,23], Brad B Worrall (MD, MSc)[24], Daniel Woo (MD)[25], Hugh S Markus (DM, FRCP)[20], Braxton D Mitchell (MPH, PhD)[9,26], Martin Dichgans (MD, PhD)[3,27], Jonathan Rosand (MD, MsC)[4-7], and Cathie LM Sudlow (BMBCh, MSc, DPhil, FRCPE)[1,28] on behalf of the Stroke Genetics Network (SiGN), METASTROKE Collaboration, and the International Stroke Genetics Consortium (ISGC).

1. Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

2. College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK

3. Institute for Stroke and Dementia Research, Klinikum der Universität München, Munich, Germany

4. Center for Human Genetic Research, Massachusetts General Hospital (MGH), Boston, MA, USA

5. J. Philip Kistler Stroke Research Center, Department of Neurology, MGH, Boston, MA, USA

6. Division of Neurocritical Care and Emergency Neurology, Department of Neurology, MGH, Boston, MA, USA

7. Program in Medical and Population Genetics, Broad Institute, Cambridge MA, USA

8. Population Health Research Institute, McMaster University, Hamilton Health Sciences Centre, Canada

9. Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA

10. Division of Neurocritical Care and Emergency Neurology, Department of Neurology, Yale University School of Medicine, New Haven, USA

11. Stroke pharmacogenomics and genetics, Fundació Docència i Recerca Mutua Terrassa, Mutua de Terrassa Hospital, Terrassa, Spain

12. Neurovascular Research Unit, Department of Neurology, Institut Municipal d'Investigacio´ Medica–Hospital del Mar, Universitat Autonoma de Barcelona, Barcelona, Spain

13. Program in Inflammation and Cardiovascular Disorders, Institut Municipal d'Investigacio´ Medica–Hospital del Mar, Universitat Autonoma de Barcelona, Barcelona, Spain

14. Department of Clinical Sciences Lund, Neurology, Lund University, Lund, Sweden

15. Department of Neurology and Rehabilitation Medicine, Neurology, Skåne University Hospital, Lund, Sweden

16. Neurovascular Research Laboratory and Neurovascular Unit, Institut de Recerca, Hospital Vall d'Hebron, Universitat Autonoma de Barcelona, Barcelona, Spain

17. HRB Clinical Research Facility, NUI Galway, and University Hospital Galway, Canada

18. Department of Neurology, Jagiellonian University Medical College, Krakow, Poland

19. Cardiovascular Epidemiology Research Group, Department of Clinical Sciences, Lund University, Malmö, Sweden

20. Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

21. Department of Neurology, Brain Centre Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

22. Boston University Schools of Medicine and Public Health, Boston, MA, USA

23. Framingham Heart Study, Framingham, MA, USA

24. Departments of Neurology and Public Health Sciences, University of Virginia, Charlottesville, Virginia, USA

25. Department of Neurology, University of Cincinnati College of Medicine, Cincinnati, OH, USA.

26. Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, Maryland, USA

27. Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

28. Institute for Genetics and Molecular Medicine, University of Edinburgh, UK

Corresponding author:

Professor Cathie Sudlow

Centre for Clinical Brain Sciences, Clinical Neurosciences, University of Edinburgh

Chancellor's Building

49 Little France Crescent

Edinburgh EH16 4SB

Telephone 0131 537 2622

Fax 0131 332 5150

Email cathie.sudlow@ed.ac.uk

Title character count: 117

Abstract word count: 211

Overall word count: 3000

Number of figures: 2

Number of tables: 3

Number of references: 37

Supplemental Data: 2 Supplemental tables (Table e-1 and Table e-2) and 1 Supplemental figure (Figure e-1).

Search terms: [2] All Cerebrovascular disease/Stroke; [6] Infarction; [7] Intracerebral haemorrhage; [91] All Genetics; [95] Association studies in genetics;

Author contributions: KR, RM and CLMS contributed to the conception and design of the study. KR, VS and HM conducted the meta-analyses. KR, VS, HM, RM and CLMS all contributed by drafting a significant portion of the manuscript. KR, RM, VS, HM, CDA, MC, JLP, SLP, GJF, TD, IFC, JJC, AL, JM, MO, GP, FR, AS, NSR, MS, MT, SS, BBW, DW, HSM, BDM, MD, JR and CLMS contributed to the acquisition and analysis of the data. All authors reviewed and approved the final version.

Study funding: This work was funded in part by NIH grants U01 NS069208 and P30 DK072488 (SiGN). MD and RM were supported by grants from the Deutsche Forschungsgemeinschaft (CRC 1123 [B3] and Munich Cluster for Systems Neurology [SyNergy]), the German Federal Ministry of Education and Research (BMBF, e:Med programme e:AtheroSysMed), the FP7/2007-2103

European Union project CVgenes@target (grant agreement No Health-F2-2013-601456), the European Union Horizon2020 projects SVDs@target (grant agreement No 66688) and CoSTREAM (grant agreement No 667375), the Fondation Leducq (Transatlantic Network of Excellence on the Pathogenesis of Small Vessel Disease of the Brain), the Vascular Dementia Research Foundation, and the Jackstaedt Foundation. GP and MC were supported by the Canadian Stroke Network, Canadian Institutes of Health Research and Heart & Stroke Foundation. NSR acknowledges the National Institute of Neurologic Disorders and Stroke K23NS064052, R01NS086905 and R01NS082285. JR acknowledges the National Institute of Neurologic Disorders and Stroke R01NS059727 and Keane Stroke Genetics Fund. HSM is supported by a NIHR Senior Investigator award and support from the Cambridge University Hospitals Comprehensive Biomedical Research Centre. CDA acknowledges funding from NIH K23 NS086873. AL is supported by Region Skåne, Lund University, the Swedish Heart and Lung Foundation, the Freemasons Lodge of Instructions EOS Lund, and the Swedish Stroke Association.

Disclosures: KR, VS, HM, RM, CDA, MC, TD, GJF, IFC, JJC, AL, JM, MO, GP, FR, NSR, AS, MS, MT, SLP, SS, BBW, DW, HSM, BDM, MD, JR, and CLMS do not report any disclosures.

Abstract

Objective: We aimed to determine whether common variants in familial cerebral small vessel disease (SVD) genes confer risk of sporadic cerebral SVD.

Methods: We meta-analyzed genotype data from individuals of European ancestry to determine associations of common single nucleotide polymorphisms (SNPs) in six familial cerebral SVD genes (*COL4A1, COL4A2, NOTCH3, HTRA1, TREX1 and CECR1*) with: ICH (deep, lobar, all; 1,878 cases, 2,830 controls) and IS (lacunar, cardioembolic, large vessel disease, all; 19,569 cases, 37,853 controls). We applied data quality filters, and set statistical significance thresholds accounting for linkage disequilibrium and multiple testing.

Results: A locus in *COL4A2* was associated (significance threshold $p<3.5x10^{-4}$) with both lacunar IS (lead SNP rs9515201: OR 1.17; 95%CI 1.11-1.24, $p=6.62x10^{-8}$) and deep ICH (lead SNP rs4771674: OR 1.28; 95%CI 1.13-1.44; $p=5.76x10^{-5}$). A SNP in *HTRA1* was associated (significance threshold $p<5.5x10^{-4}$) with lacunar IS (rs79043147: OR 1.23, 95%CI 1.10-1.37, $p=1.90x10^{-4}$), and less robustly with deep ICH. There was no clear evidence for association of common variants in: either *COL4A2* or *HTRA1* with non-SVD strokes; or in any of the other genes with any stroke phenotype.

Conclusions: These results provide evidence of shared genetic determinants and suggest common pathophysiological mechanisms of distinct ischemic and hemorrhagic cerebral SVD stroke phenotypes, offering new insights into the causal mechanisms of cerebral SVD.

Introduction

Small vessel diseases of the brain include a subtype that affects the small, deep, penetrating arteries and arterioles in the brain, hereafter referred to as "deep cerebral SVD". This deep cerebral SVD is thought to be responsible for most symptomatic lacunar ischemic strokes and deep intracerebral hemorrhages (ICHs), as well as for substantial cognitive and physical disabilities, and to be a major pathological substrate for brain MRI features, including white matter hyperintensities (WMH) and brain microbleeds.[1, 2] Increasing evidence supports a distinct vascular pathology for deep cerebral SVD, but our knowledge of the underlying genes and pathophysiological mechanisms is limited, with lack of specific treatment strategies.[1]

While the genetic determinants of common sporadic forms of cerebral SVD remain largely unknown, mutations in at least six genes (*COL4A1*, *COL4A2, HTRA1, CECR1, NOTCH3, TREX1*) are known to cause rare familial forms of deep cerebral SVD.[3, 4] Such genes may also contain variants conferring risk for sporadic deep cerebral SVD. We previously investigated associations of common variation in the *COL4A1* and *COL4A2* genes with cerebrovascular phenotypes in a collaborative meta-analysis, demonstrating an association between an intronic *COL4A2* locus and sporadic deep ICH, and a suggestive association with other deep cerebral SVD phenotypes (lacunar ischemic stroke and WMH).[5] The same genetic locus has since been shown to be associated with WMH at genome-wide association study (GWAS) levels of significance.[6]

We aimed to extend this promising candidate gene approach to assess associations of common variants in all currently known familial deep cerebral SVD genes with stroke and its subtypes, investigating the hypothesis that associations would be specific to the two key sporadic deep

cerebral SVD stroke phenotypes, lacunar ischemic stroke and deep ICH. We were able to take advantage of the increased sample sizes and more densely imputed genotype data now available through the International Stroke Genetics Consortium (ISGC) (http://www.strokegenetics.org/) and associated collaborative groups.

Methods

Identification of participating studies

We identified most currently available large GWASs of stroke and stroke subtypes in individuals of European ancestry, using a network of collaborations associated with the ISGC.[7-11] The entire dataset comprised: 20 case-control collections including 19,569 ischemic stroke cases and 37,853 controls, with information on Trial of Org 10172 in Acute Stroke Treatment [TOAST] subtypes (lacunar ischemic stroke, large vessel disease [LVD], cardio-embolic [CE])[12]; and five case-control collections including 1,878 ICH cases and 2,830 controls, with information on the main ICH subtypes (Table 1). For the majority of case collections, population matched controls were recruited from studies with existing genotyping data (details of case-control collections found in references 7-11).

Individual studies applied quality control measures prior to providing the data. All data were imputed with the 1000 Genomes Phase 1 reference dataset (or to a merged reference panel including the Genome of the Netherlands), using IMPUTE2 or MACH software[13] and provided with reference to Human Genome reference build 19.

Data collection

We collected genotype summary statistics from participating case-control collections for the *COL4A1*, *COL4A2*, *HTRA1, CECR1, NOTCH3* and *TREX1* genes (encompassing all known familial deep cerebral SVD genes), including a 10kbp up- and downstream flanking region for each gene (Table 2).

We focused on the lacunar ischemic stroke and deep ICH phenotypes, but also assessed LVD, CE and all ischemic stroke for ischemic stroke case-control collections, and lobar ICH and all ICH for ICH case-control collections, to show specificity of the association. For each of these phenotypes, we collected summary data from each collection for all directly genotyped or imputed SNPs in genes of interest: SNP reference number and position; allele frequencies; association effect size (ß-coefficient) and its standard error; association p-value; and imputation quality measure and value.

Data analysis

Setting the significance threshold

To allow for multiple testing while accounting for linkage disequilibrium (LD) between SNPs, we calculated significance p-values for each gene using a modified version of Nyholt's method (MeffLi), which controls accurately for error rate in evaluations of real and simulated data.[14-17] We used the 1000G CEU dataset[18] SNP genotype information to calculate p-values for five genomic regions, treating the *COL4A1* and *COL4A2* genes as one region since they are located in tandem on chromosome 13q34 and share a promoter (Table 2).

Pre meta-analyses data filtering

We further filtered the data to include only SNPs with the following attributes: (1) imputation quality ≥0.3 from MACH, IMPUTE2 or SNPTEST (since SNPs with very poor imputation quality may yield unreliable associations); (2) minor allele frequency ≥1% (since we were investigating common SNPs); (3) absolute beta value <100 000 (since higher beta values would generate implausible odds ratios, suggesting unreliable associations); (4) biallelic SNPs (since the meta-analyses program could not process multiallelic SNPs).

Meta-analyses of *COL4A1, COL4A2, HTRA1, CECR1, NOTCH3* and *TREX1* SNPs for each phenotype

We meta-analyzed genotype summary data from each contributing case-control collection. We assessed associations of *COL4A1, COL4A2, HTRA1, CECR1, NOTCH3* and *TREX1* SNPs with each of the stroke phenotypes available, both those assumed to represent deep cerebral SVD specifically (lacunar ischemic stroke, deep ICH) and others (LVD ischemic stroke, CE ischemic stroke, all ischemic stroke, lobar ICH, all ICH). Our hypothesis was that associations would be specific to (or at least strongest with) deep cerebral SVD phenotypes. We used a fixed effects inverse-variance-based model in METAL genetic meta-analysis software, weighting the ß-coefficients by their estimated standard errors and generating, for each SNP, the odds ratio (OR) per additional minor allele for being a case versus a control.[19]

Post meta-analyses data filtering

Following the meta-analyses, we considered SNPs to be associated with the respective phenotype if the relevant associations: (1) passed the relevant modified Nyholt corrected p-value threshold; (2) were based on data from ≥50% of cases contributing to the analyses; (3) did not demonstrate substantial heterogeneity (requiring I^2 <50% and p >0.001 from chi-squared test).

We chose our filtering thresholds based on those most commonly used and accepted[6-8,20] with the aim of ensuring that any associations deemed significant would be based on SNPs with sufficient, reliable and consistent data.

Further exploration of associated SNPs

Where more than one SNP in the same gene was associated with any given phenotype, we used the 1000 Genomes project CEU population data to investigate the LD between the lead SNP (with the lowest p-value) and all other associated SNPs for the relevant gene-phenotype association. SNPs in moderate or strong LD (defined respectively as r^2 and/or D' ≥0.6 or ≥0.8) with the lead SNP were considered likely to represent a signal from the same locus.[21]

We examined associations of all lead SNPs across all case-control collections included in the respective meta-analyses, and of all lead SNPs with all other phenotypes, comparing findings for deep cerebral SVD stroke phenotypes versus non-SVD stroke phenotypes.

Finally, we sought functional annotation data from the Haploreg v2 database (http://www.broadinstitute.org/mammals/haploreg/haploreg.php)[22], Genotype-Tissue Expression Portal eQTL browser (http://www.broadinstitute.org/gtex/) and the RegulomeDB database (http://regulome.stanford.edu/) for all associated SNPs.

Results

Meta-analyses of COL4A1, COL4A2, HTRA1, CECR1, NOTCH3 and *TREX1* SNPs *for each phenotype*

Modified Nyholt significance thresholds for the five genomic regions are shown in Table 2. Based on our preset data filtering criteria, we found associations of 18 SNPs in *COL4A2* with lacunar

ischemic stroke, nine SNPs in *COL4A2* with deep ICH*,* and one SNP in *HTRA1* with lacunar ischemic stroke (Table 3, Figure 1). Two of the SNPs in the *COL4A2* gene (rs4771674 and rs9515199) were associated with both lacunar ischemic stroke and deep ICH. There were no associations of common SNPs in *COL4A2* or *HTRA1* with any of the non-cerebral SVD or combined stroke phenotypes, or of common variants in *COL4A1, CECR1*, *NOTCH3* or *TREX1* with any of the stroke phenotypes.

Further exploration of associated SNPs

Linkage disequilibrium between *COL4A2* SNPs

The lead SNPs were rs9515201 for lacunar ischemic stroke (OR per additional A allele 1.17; 95% CI 1.11-1.24, $p=6.62x10^{-8}$), and rs4771674 for deep ICH (OR per additional A allele 1.28; 95% CI 1.13-1.44; $p=5.76x10^{-5}$). These two lead SNPs are in strong LD with each other (r^2=0.9/D´=1), suggesting that this represents the same genetic signal.

We investigated the LD between the lead SNP rs9515201 (most strongly associated SNP in the locus) and all other associated SNPs in *COL4A2*. Out of the other 24 SNPs in *COL4A2* associated with lacunar ischemic stroke and/or deep ICH, 19 were in moderate to strong LD with the lead SNP, suggesting that their signal may be from the same *COL4A2* locus. The remaining 4 SNPs showed minimal LD with the lead SNP, while data for one SNP was not available, suggesting that there might possibly be additional relevant loci (Table e-1).

Associations across individual case-control collections in the meta-analyses

The associations for *COL4A2* SNPs showed minimal to moderate heterogeneity across individual case-control collections in the lacunar ischemic stroke and deep ICH meta-analyses (I^2 0%-49%, heterogeneity p-values 0.7-0.01, and I^2 0%-42%, heterogeneity p-values 0.1-0.5 respectively), while the associations across individual collections for rs79043147 in *HTRA1* in the lacunar ischemic stroke meta-analysis showed only minimal heterogeneity (I^2=4%, heterogeneity p=0.41), suggesting consistent results (Figure e-1). All imputed SNPs showed a good imputation quality of >0.7.

Associations with other phenotypes of the lead *COL4A2* SNPs

Figure 2 shows association results for the lead *COL4A2* SNPs (rs9515201 and rs4771674) associated with lacunar ischemic stroke and deep ICH across all seven phenotypes assessed. Although rs9515201 was associated only with lacunar ischemic stroke (OR 1.17; 95% CI 1.11-1.24; $p=6.62x10^{-8}$), there was a suggestive association of similar magnitude with deep ICH (OR 1.21, 95% CI 1.07-1.37, $p=2.15x10^{-3}$). Rs4771674 was associated with both cerebral SVD phenotypes – lacunar ischemic stroke (OR 1.14; 95% CI 1.07-2.10; $p=1.6x10^{-5}$) and deep ICH (OR 1.28; 95% CI 1.13-1.44; $p=5.76x10^{-5}$).

There were no associations with non-SVD stroke, or combined SVD and non-SVD phenotypes. ORs for the all ischemic stroke and all ICH phenotypes were intermediate between those for SVD and non-SVD phenotypes, suggesting that associations with these combined phenotypes were driven by results for lacunar ischemic stroke and deep ICH.

Associations with other cerebrovascular phenotypes of the *HTRA1* SNP

Rs79043147 was associated only with lacunar ischemic stroke (OR 1.23, 95% CI 1.10-1.37, $p=1.90x10^{-4}$) but also showed a suggestive association with deep ICH (OR 1.56, 95% CI 1.24-1.97, $p=1.71x10^{-4}$) (Figure 2). In fact, the p-value for deep ICH passed the significance threshold, but the SNP did not pass our preset heterogeneity filter and was therefore not considered associated overall. There were no associations with non-SVD stroke, or combined SVD and non-SVD phenotypes.

Functional annotation

All *COL4A2* and *HTRA1* SNPs associated with lacunar ischemic stroke and/or deep ICH were intronic. The GTEx eQTL browser search revealed no significant eQTLs for any of these SNPs. However, the RegulomeDB database revealed that two *COL4A2* SNPs were in an area likely to affect binding, two *COL4A2* SNP were in an area less likely to affect binding, and 17 *COL4A2* SNP showed minimal binding evidence, suggesting that these SNPs are located in areas of the genome that may have regulatory functions (Table e-2).

Discussion

Our results demonstrate an association of an intronic, possibly regulatory locus in *COL4A2* with two distinct deep cerebral small vessel disease phenotypes - lacunar ischemic stroke and deep ICH. We also found an association of deep cerebral SVD with *HTRA1*, demonstrating an association with lacunar ischemic stroke and a suggestive association with deep ICH. Finding the same genetic signal associated with both ischemic and hemorrhagic sporadic stroke confirms the usefulness of a joint exploration of cerebrovascular phenotypes, and the potential for genetic studies to shed light on common underlying mechanisms.

Our findings for *COL4A2* are supported by previous work showing the relevance of this genomic region in sporadic deep cerebral SVD. A sequence analysis of *COL4A1/COL4A2* found missense mutations in sporadic ICH cases.[23,24] Also, our previous meta-analyses in a smaller, partly overlapping sample, already demonstrated an association of this *COL4A2* locus with deep ICH, and a suggestive association with other cerebral SVD phenotypes.[5] By increasing the sample size (by 40% for lacunar ischemic stroke and by 15% for deep ICH) and density of coverage in the current study, we have now established a substantial association of the same locus with lacunar ischemic stroke and confirmed the association with deep ICH. Furthermore, a recently published GWAS identified our lead SNP for lacunar ischemic stroke to be associated with another deep cerebral SVD phenotype, WMH.[6]

While the association with *COL4A2* is supported by previous data and a convincing signal for both ischemic and hemorrhagic phenotypes, the association with *HTRA1* is suggestive but less certain. Based on our p-value threshold, the *HTRA1* SNP was associated with both lacunar ischemic stroke and deep ICH, but there was significant heterogeneity in the deep ICH meta-analyses. We are also aware of one previous report suggesting an association of rare variation in *HTRA1* with more extreme sporadic deep cerebral SVD phenotypes.[25] Thus, this finding should be pursued in independent, large samples in order to replicate the association.

From a biological point of view, our strategy of investigating these familial genes jointly is supported by an emerging view that the resulting familial deep cerebral SVDs have similar disease mechanisms involving disruption of the cerebrovascular matrisome. Familial mutations leading to alterations of matrisome proteins and function could be a convergent pathway driving the

functional and structural alterations of small brain vessels and disease manifestations, and similar mechanisms could also play a role in sporadic disease.[26]

COL4A1 and *COL4A2* genes encode the collagen protein chains, a major component of the vascular basement membrane.[27] Their dominant missense mutations are associated with basement membrane defects and endoplasmic reticulum (ER) stress, and cause rare familial SVDs.[28-31] Recent data suggest that manipulation of ER stress (e.g. with 4-phenyl butyric acid) is a potential therapeutic option for collagen IV diseases including hemorrhagic stroke.[29] Mutations in *HTRA1* gene are associated with cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL).[32] While the majority of *HTRA1* mutations cause this autosomal recessive cerebral SVD, heterozygous *HTRA1* mutations associated with cerebral SVD have also recently been reported.[33] *HTRA1* encodes the HTRA1 enzyme, which through regulating transforming growth factor-beta signaling plays an important role in the formation of blood vessels.

Possible reasons for an apparent lack of an association with *COL4A1*, *CECR1*, *NOTCH3* and *TREX1* include: genuine lack of an association in our study population; weaker association not detected due to sample size; suboptimal diagnosis of SVD phenotypes in the original studies resulting in reduced power; and variability in the density and quality of genotyping across different genes. In addition, we used a 10kb flanking region to cover regulatory areas for all genes, therefore relatively more conservative p-values may have been derived for smaller genes (*TREX1*) after adjusting for the number of SNPs tested. Also, because we treated the *COL4A1/COL4A2* region

as one, more conservative p-values were derived for *COL4A1* gene than if we had treated it as a separate region.

Our study has several strengths. We investigated a specific, pre-specified hypothesis, clearly defining the phenotypes and candidate genes of interest, based on pre-existing supporting data. Through a network of collaborative groups we could include the majority of currently available data from stroke genetics studies of European ancestry individuals. We used appropriate methods to correct for multiple testing.

There were some limitations. Firstly, while we have shown that SNPs in *COL4A2* are associated with lacunar ischemic stroke and deep ICH through analyzing data for the specific candidate region, the associations did not reach GWAS significance ($p \leq 5\times10^{-8}$), most likely because of limited sample size. However, the lead lacunar ischemic stroke SNP (rs9515201) had a p-value of 6.6×10^{-8}, and it has been shown that a substantial proportion of SNPs with a p-value in this ´borderline´ GWAS significance range ($P>5\times10^{-8}$ and $P\leq1\times10^{-7}$) represent genuine, replicable associations.[34] Secondly, we did not adjust the statistical threshold for the number of genomic regions and phenotypes investigated, considering this overly conservative, since we were investigating a series of specific related hypotheses rather than one single hypothesis. However, had we further adjusted the *COL4A1*/*COL4A2* region p-value for the number of tests ($3.5\text{x}10^{-4}/25 = 1.4\times10^{-5}$), the association for the lead SNP with lacunar ischemic stroke would have remained significant. Thirdly, our analyses found a locus in *COL4A2* containing several SNPs associated with deep cerebral SVD, most (but not all) of which were in moderate to strong LD with the lead SNP. This suggests the association was likely driven by the lead SNP, but there remains a possibility that

independently significant signals in the locus may emerge.[35] Further investigation of this would require additional analyses adjusting for the lead SNP, requiring genome-wide genetic data which were not sought for this study, given its targeted hypothesis-driven approach. Fourth, the diagnostic work-up leading to TOAST subtype classification was study-specific which may introduce some heterogeneity. Fifth, not all studies controlled for age in their statistical analyses before inclusion in the meta-analyses, and this may decrease the study power. Finally, we were not able to include data for additional relevant phenotypes such as WMH and brain microbleeds in the current study.

While genetic studies of ischemic stroke and ICH have generally been pursued separately, these findings emphasize the mechanistic insights that can be gained from joint analyses of cerebrovascular phenotypes. We have shown that the same genetic signal is associated with clinically evident sporadic ischemic and hemorrhagic stroke, but the joint exploration approach is further supported by previous GWA studies showing a locus on chromosome 1q22 to be associated with both deep ICH and WMH[10,36]. In addition, it has recently been shown that a locus on chromosme 6p25, near the *FOXF2* gene (also associated with familial deep SVD), is associated with all stroke (likely driven by SVD stroke phenotypes) and suggestively with WMH.[37]

Follow-up studies should further explore potential common genetic signals for deep cerebrovascular ischemic and hemorrhagic SVD phenotypes in larger sample sizes, for additional relevant phenotypes such as WMH and brain microbleeds, and including non-European ethnic groups. Future studies could also assess the potential contribution of rare variants to common cerebral SVD phenotypes in these Mendelian genes. In addition, the robust findings for *COL4A2*

now merit further deep sequencing of the entire genomic region among sporadic deep cerebral SVD cases, with detailed functional studies of promising variants thus identified.

References:

1. Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. 2010;9:689-701.

2. Wardlaw JM, Smith EE, Biessels GJ, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 2013;12:822-838.

3. Tan RY & Markus HS. Monogenic causes of stroke: now and the future. J Neurol. 2015;262:2601-2616.

4. Haffner C, Malik R, Dichgans M. Genetic factors in cerebral small vessel disease and their impact on stroke and dementia. J Cereb Blood Flow Metab. 2015;36:158-71.

5. Rannikmäe K, Davies G, Thomson PA, et al. Common variation in COL4A1/COL4A2 is associated with sporadic cerebral small vessel disease. Neurology. 2015;84:918-926.

6. Traylor M, Zhang CR, Adib-Samii P, et al. Genome-wide meta-analysis of cerebral white matter hyperintensities in patients with stroke. Neurology. 2016;86:146-153.

7. Malik R, Traylor M, Pulit SL, et al. Low-frequency and common genetic variation in ischemic stroke: The METASTROKE collaboration. Neurology. 2016;86:1217-1226.

8. NINDS Stroke Genetics Network (SiGN) & International Stroke Genetics Consortium (ISGC). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Lancet Neurol. 2015;doi: 10.1016/S1474-4422(15)00338-5.

9. O'Donnell M, Xavier D, Diener C, et al. Rationale and design of INTERSTROKE: a global case-control study of risk factors for stroke. Neuroepidemiology. 2010;35:36-44.

10. Woo D, Falcone GJ, Devan WJ, et al. Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. Am J Hum Genet. 2014;94:511-521.

11. Woo D, Sauerbeck LR, Kissela BM, et al. Genetic and environmental risk factors for intracerebral hemorrhage: preliminary results of a population-based study. Stroke. 2002;33:1190-1195.

12. Adams HP, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24:35-41.

13. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499-511.

14. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb). 2005;95:221-227.

15. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004;74:765-769.

16. Nyholt DR. Evaluation of Nyholt's procedure for multiple testing correction - author's reply. Hum Hered. 2005;60:61-62.

17. Salyakina D, Seaman SR, Browning BL, et al. Evaluation of Nyholt's procedure for multiple testing correction. Hum Hered. 2005;60:19-25.

18. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. Nature. 2015;526:68-74.

19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190-2191.

20. Rutten-Jacobs LC, Traylor M, Adib-Samii P, et al. Common NOTCH3 Variants and Cerebral Small-Vessel Disease. Stroke. 2015;46:1482-1487.

21. Mueller JC. Linkage disequilibrium for different scales and applications. Brief Bioinform. 2004;5:355-364.

22. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930-934.

23. Jeanne M, Labelle-Dumais C, Jorgensen J, et al. COL4A2 mutations impair COL4A1 and COL4A2 secretion and cause hemorrhagic stroke. Am J Hum Genet 2012;90:91–101.

24. Weng YC, Sonni A, Labelle-Dumais C, et al. COL4A1 mutations in patients with sporadic late-onset intracerebral hemorrhage. Ann Neurol 2012;71:470–477.

25. Nozaki H, Kato T, Nihonmatsu M, et al. Distinct molecular mechanisms of HTRA1 mutants in manifesting heterozygotes with CARASIL. Neurology. 2016;86:1964-1974.

26. Joutel A, Haddad I, Ratelade J, et al. Perturbations of the cerebrovascular matrisome: a convergent mechanism in small vessel disease of the brain? J Cereb Blood Flow Metab. 2015;36:143-157.

27. Khoshnoodi J, Pedchenko V, Hudson BG. Mammalian collagen IV. Microsc Res Tech. 2008;71:357-370.

28. Jeanne M, Labelle-Dumais C, Jorgensen J, et al. COL4A2 mutations impair COL4A1 and COL4A2 secretion and cause hemorrhagic stroke. Am J Hum Genet. 2012;90:91-101.

29. Murray LS, Lu Y, Taggart A, et al. Chemical chaperone treatment reduces intracellular accumulation of mutant collagen IV and ameliorates the cellular phenotype of a COL4A2 mutation that causes haemorrhagic stroke. Hum Mol Genet. 2014;23:283-292.

30. Van Agtmael T, Schlötzer-Schrehardt U, McKie L, et al. Dominant mutations of Col4a1 result in basement membrane defects which lead to anterior segment dysgenesis and glomerulopathy. Hum Mol Genet. 2005;14:3161-3168.

31. Van Agtmael T, Bailey MA, Schlötzer-Schrehardt U, et al. Col4a1 mutation in mice causes defects in vascular function and low blood pressure associated with reduced red blood cell volume. Hum Mol Genet. 2010;19:1119-1128.

32. Hara K, Shiga A, Fukutake T, et al. Association of HTRA1 mutations and familial ischemic cerebral small-vessel disease. N Engl J Med. 2009;360:1729-1739.

33. Verdura E, Hervé D, Scharrer E, et al. Heterozygous HTRA1 mutations are associated with autosomal dominant cerebral small vessel disease. Brain. 2015;138:2347-2358.

34. Panagiotou OA, Ioannidis JP, Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol. 2012;41:273-286.

35. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44:369-375, S361-363.

36. Verhaaren BF, Debette S, Bis JC, et al. Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI. Circ Cardiovasc Genet. 2015;8:398-409.

37. Chauhan G, Arnold CR, Chu AY, et al. Identification of additional risk loci for stroke and small vessel disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2016;15:695-707.

Figure titles and legends.

Figure 1 title: *COL4A2* regional association plots for (A) lacunar ischemic stroke, (B) deep ICH; and (C) *HTRA1* regional association plot for lacunar ischemic stroke

Figure 1 legend: Only SNPs passing the post meta-analyses filters (heterogeneity I^2 <50%; p>0.001; ≥50% cases contributing data) have been displayed. Red dashed lines mark the relevant Nyholt significance p-value thresholds. Dots mark individual SNPs with respect to their chromosomal position (x axis) and p-value for association between each SNP and phenotype (left y axis). The SNP in purple is the most strongly associated (lead) SNP; linkage disequilibrium with this lead SNP determines the colors for other SNPs, as seen from the r^2 color coding on figure. Recombination rates (right y axis), shown by the continuous blue lines, are measured as frequency of exchange per unit physical distance (centimorgan [cM]/mega base pair [Mb]).

Figure 2 title: Associations of *COL4A2* and *HTRA1* SNPs across all phenotypes.

Figure 2 legend: OR: odds ratio; CI: confidence interval; IS: ischemic stroke; ICH: intracerebral hemorrhage; LVD: large vessel disease; CE: cardioembolic; N: number; Diamonds represent pooled odds ratios across all case-control collections for each phenotype, with the line through the diamond showing its 95% confidence interval. Associations significant at relevant Nyholt threshold are shown in red, non-significant associations with SVD phenotypes shown in black, non-significant associations with non-SVD phenotypes shown in gray.

Table 1. Participating case-control collections

ISCHEMIC STROKE								
Studies forming case-control collections[1]	Cases (n)	Controls (n)	CE cases (n)	LVD cases (n)	SVD cases (n)	Mean age cases	Mean age controls	Genotyping panel
ASGC	1162	1244	240	421	310	73	70	Illumina 610
BRAINS	371	2640	78	40	29	74	≥65	Illumina 610
GASROS illumina	296	377	106	68	24	67	48	Illumina
GASROS affy	485	3030	198	102	59	69	51	Affymetrix
GEOS	448	498	90	37	54	41	40	Illumina H Omni
HPS	588	571	-	-	-	65	59	Illumina 610
ISGS & SWISS	1014	1370	235	217	187	67	65	Illumina 550/610/660
Milano	366	407	64	73	25	57	51	Illumina 610/660
WHI	306	2170	42	31	81	59	52	Illumina 1M
VISP	1723	1047	-	-	-	61	63	Illumina 5M
WTCCC2 Germany	1174	797	346	330	106	67	63	Illumina 660
WTCCC2 UK	2374	5175	460	498	474	72	52[2]	Illumina 660
BRAINS, ISGS, GASROS, SWISS, HABC	754	1586	176	163	149	67	74	Illumina 650Q, 610, 1M
CIDR[3], HRS, OAI	3291	11514	598	423	761	69	67	Illumina 5M, 2.5M, HumanOmni, 5Exome-4v1
Krakow	878	716	407	173	36	69	56	Illumina 5M
LSGS	459	453	157	75	55	67	56	Illumina 5M
BASICMAR, ADHD, INMA	868	1218	408	184	276	75	-	Illumina 5M, 1M
Graz	607	815	166	85	74	69	65	Illumina 610, 5M
SAHLSIS, LSR, MDC	1579	1362	223	150	183	62	61	Illumina 5M, 610
Interstroke	826	863	208	188	243	68	66	Illumina Human Exome Chip, Cardiometabochip
TOTAL[4]	**19 569**	**37 853**	**4202**	**3258**	**3126**			-

INTRACEREBRAL HEMORRHAGE							
Case-control collections[1]	Cases (n)	Controls (n)	Lobar ICH[6] cases (n)	Deep ICH[7] cases (n)	Mean age cases	Mean age controls	Genotyping panel
GOCHA	387	387	210	167	72	72	Illumina HumanHap 610
ISGC	528	530	181	313	72	66	Illumina HumanHap 610
GERFHS	628	573	258	370	68	67	Affymetrix 6.0
MDC	199	372	76	95	62	58	Human OmniExpress Exome Bead Chip v 1.0
Cambridge	136	968	59	77	71	60	Illumina HumanCoreExome
TOTAL[5]	**1878**	**2830**	**784**	**1022**			-

[1]Case-control collections were analyzed as individual studies and/or as groups of studies, depending on how data were provided. Where the same study appears more than once in this table, non-overlapping sets of cases were included; [2]approximate age at genotyping of the 2738 controls from the 1958 Birth Cohort; age unavailable for the remaining controls; [3]CIDR (samples genotyped in the Center for Inherited Disease Research): BRAINS, GASROS, GCNKSS (Greater Cincinnati/Northern Kentucky Stroke Study), ISGS, MCISS (Middlesex County Ischemic Stroke Study), MIAMISR (Miami Stroke Registry and Biorepository), NHS (Nurses' Health Study), NOMAS (Northern Manhattan Study), REGARDS (Reasons for Geographic and Racial Differences in Stroke), SPS3 (Secondary Prevention of Small Subcortical Strokes), SWISS, WHI, WUSTL (Washington University St Louis); [4]Further information in references 7-9; [5]Further information in references 10-11; [6]Lobar ICH defined as involving predominantly the cortex and underlying white matter; [7]Deep ICH defined as ICH involving predominantly the basal ganglia, periventricular white matter, or internal capsule, and infratentorial ICH;

CE: cardioembolic; LVD: large vessel disease; ICH: intracerebral hemorrhage;

ASGC: Australian Stroke Genetics Collaborative; BRAINS: Bio-Repository of DNA in stroke; GASROS: The MGH Genes Affecting Stroke Risk and Outcome Study; GEOS: Genetics of Early-Onset Stroke; HPS: Heart Protection Study; ISGS/SWISS: The Ischemic Stroke Genetics Study/Sibling with Ischaemic Stroke Study; WHI: The Women's Health Initiative; VISP: The Vitamin Intervention for Stroke Prevention Trial; WTCCC2: The Wellcome Trust Case-Control Consortium; HABC: Health ABC; HRS: Health and Retirement Study; OAI: Osteoarthritis Initiative; LSGS: Leuven Stroke Genetics Study; BASICMAR: Base de Datos de Ictus del Hospital del Mar; ADHD: Attention-deficit Hyperactivity Disorder; INMA: Infancia y medio ambiente; SAHLSIS: Sahlgrenska Academy Study of Ischemic Stroke; LSR: Lund Stroke Register; MDC: Malmö Diet and Cancer Study; GOCHA: Genetics of Cerebral Hemorrhage with Anticoagulation; ISGC: International Stroke Genetics Consortium; GERFHS: Genetic and Environmental Risk Factors for Hemorrhagic Stroke;

Table 2. Six genes assessed: location, number of SNPs and Nyholt association p values regarded as significant

Gene/genomic region	Chromosome	Coordinates[1]	SNPs		Nyholt p value
			N_{all}	N_{eff}	
COL4A1/COL4A2	13	110,791,310 – 111,175,373	2555	147	$3.5x10^{-4}$
HTRA1	10	124,211,041 - 124,284,424	417	93	$5.5x10^{-4}$
CECR1	22	17,650,192 - 17,712,738	422	97	$5.3x10^{-4}$
NOTCH3	19	15,260,444 - 15,321,792	278	72	$7.1x10^{-4}$
TREX1	3	48,491,186 - 48,519,044	60	38	$1.3x10^{-3}$

[1]Based on the Human Genome reference build 19.

Table 3. All associated SNPs passing post meta-analyses filters.

SNP	Minor allele	Major allele	Minor allele freq.[1]	OR (95% CI)[2]	Association p-value	Direction of effect[3]	% cases[4]	I^2	Heterogeneity p-value
LACUNAR ISCHEMIC STROKE: COL4A2 (Nyholt significance p-value threshold $3.5x10^{-4}$)									
rs9515201	A	C	0.31	1.17 (1.11 to 1.24)	6.62×10^{-8}	+++++++-+++++++--+	100	48	0.01
rs4771674	A	G	0.39	1.14 (1.07 to 1.2)	1.60×10^{-5}	+++++++-++-+?++--+	92	49	0.01
rs113696651	T	C	0.02	1.61 (1.29 to 2.01)	2.25×10^{-5}	++--+++?-+++++?+++	98	29	0.13
rs9521729	A	G	0.31	0.88 (0.83 to 0.94)	5.42×10^{-5}	----+----+--?++---	92	11	0.32
rs9515199	C	T	0.39	1.13 (1.06 to 1.19)	5.47×10^{-5}	+++++++-++-+?++--+	92	39	0.05
rs9559771	G	A	0.12	0.84 (0.77 to 0.91)	6.56×10^{-5}	----+----------	100	20	0.21
rs7319311	G	A	0.32	0.88 (0.83 to 0.94)	7.48×10^{-5}	----+-------?++---	92	0	0.45
rs4502089	C	T	0.27	0.88 (0.83 to 0.94)	7.49×10^{-5}	-----+---+---+++--	100	17	0.25
rs9521768	T	G	0.27	1.13 (1.06 to 1.21)	9.09×10^{-5}	++++-++-+-++-+++++	100	34	0.08
rs67472641	D[5]	R[5]	0.37	1.16 (1.08 to 1.26)	9.86×10^{-5}	?????++-+++++?????	57	24	0.24
rs9521770	G	A	0.27	1.13 (1.06 to 1.2)	1.21×10^{-4}	++++-++-+-++-+++++	100	33	0.08
rs9583488	A	G	0.32	0.89 (0.83 to 0.94)	1.21×10^{-4}	----+-------?+++--	92	10	0.34
rs11619583	C	T	0.46	1.12 (1.06 to 1.18)	1.23×10^{-4}	+++++-+++++++-++++	100	36	0.06
rs7320755	C	G	0.32	0.89 (0.84 to 0.94)	1.53×10^{-4}	----+-------?0++--	86	8	0.36

SNP	Minor allele	Major allele	Minor allele freq.[1]	OR (95% CI)[2]	Association p-value	Direction of effect[3]	% cases[4]	I^2	Heterogeneity p-value
rs77104783	A	C	0.08	0.8 (0.71 to 0.9)	1.78×10^{-4}	-----+---+--?-----	92	0	0.66
rs9515198	T	C	0.28	0.89 (0.83 to 0.94)	2.00×10^{-4}	----+----+--?+++--	92	13	0.30
rs7140030	A	G	0.25	1.12 (1.06 to 1.2)	2.36×10^{-4}	++++++++++-+++-+-+	100	36	0.06
rs9588148	T	G	0.21	0.88 (0.83 to 0.94)	3.19×10^{-4}	----------++---	100	3	0.41
DEEP ICH: COL4A2 (Nyholt significance p-value threshold 3.5×10^{-4})									
rs4771674	A	G	0.4	1.28 (1.13 to 1.44)	5.76×10^{-5}	-++++	100	27	0.24
rs9521733	C	T	0.4	1.27 (1.13 to 1.42)	8.10×10^{-5}	++++-	100	0	0.48
rs9521735	C	G	0.41	1.27 (1.13 to 1.43)	8.91×10^{-5}	++++-	100	0	0.51
rs9515200	C	G	0.39	1.27 (1.13 to 1.43)	9.36×10^{-5}	++++-	100	26	0.25
rs9521732	A	C	0.41	1.26 (1.12 to 1.42)	9.89×10^{-5}	++++-	100	0	0.45
rs9521734	T	A	0.4	1.26 (1.12 to 1.42)	1.01×10^{-4}	++++-	100	0	0.52
rs1999013	G	A	0.41	1.27 (1.12 to 1.43)	1.07×10^{-4}	-+++-	100	38	0.17
rs61963197	A	G	0.37	1.28 (1.13 to 1.45)	1.15×10^{-4}	++++-	100	41	0.15
rs9515199	C	T	0.41	1.24 (1.11 to 1.4)	2.74×10^{-4}	-+++-	100	42	0.14
LACUNAR ISCHEMIC STROKE: HTRA1 (Nyholt significance p-value threshold 5.5×10^{-4})									
rs79043147	T	C	0.06	1.23 (1.10 to 1.37)	1.91×10^{-4}	+---++--+++++-++++	100	4	0.41

[1]freq: frequency; [2]OR: odds ratio for minor allele being the effect allele; [3]Direction of effect: shows direction of association in each case-control collection included in the meta-analyses: + (OR>1), - (OR<1), ? (no data), 0 (OR=1); [4]% cases: percent of overall cases contributing data to the meta-analysis; [5]D(deletion)/R(regular) SNP (D=G, R=GGCCTGAGAAGCGACAGGGCA); I^2: heterogeneity measure; CI: confidence interval;