Rare Missense Functional Variants at COL4A1 and COL4A2 in Sporadic Intracerebral Hemorrhage

Jaeyoon Chung, PhD, Graham Hamilton, PhD, Minsup Kim, PhD, Sandro Marini, MD, Bailey Montgomery, Jonathan Henry, Art E. Cho, PhD, Devin L. Brown, MD, Bradford B. Worrall, MD, James F. Meschia, MD, Scott L. Silliman, MD, Magdy Selim, MD, PhD, David L. Tirschwell, MD, MSc, Chelsea S. Kidwell, MD, Brett Kissela, MD, MS, Steven M. Greenberg, MD, PhD, Anand Viswanathan, MD, PhD, Joshua N. Goldstein, MD, PhD, Carl D. Langefeld, PhD, Kristiina Rannikmae, MD, Catherine L.M. Sudlow, BMBC, MSc, DPhil, FRCP, Neshika Samarasekera, PhD, MRCP, Mark Rodrigues, PhD, FRCR, Rustam Al-Shahi Salman, PhD, FRCP Edin, James G.D. Prendergast, PhD, Sarah E. Harris, PhD, Ian Deary, MBChB, PhD, FRCP, Daniel Woo, MD, MSc, Jonathan Rosand, MD, MSc, Tom Van Agtmael, PhD,* and Christopher D. Anderson, MD, MMSc*

Correspondence
Dr. Anderson
cdanderson@mgh.harvard.edu
or Dr. Van Agtmael
tom.vanagtmael@glasgow.ac.uk

Neurology® 2021;97:e236-e247. doi:10.1212/WNL.00000000000012227

Abstract

Objective
To test the genetic contribution of rare missense variants in COL4A1 and COL4A2 in which common variants are genetically associated with sporadic intracerebral hemorrhage (ICH), we performed rare variant analysis in multiple sequencing data for the risk for sporadic ICH.

Methods
We performed sequencing across 559 Kbp at 13q34 including COL4A1 and COL4A2 among 2,133 individuals (1,055 ICH cases; 1,078 controls) in United States–based and 1,381 individuals (192 ICH cases; 1,189 controls) from Scotland-based cohorts, followed by sequence annotation, functional impact prediction, genetic association testing, and in silico thermodynamic modeling.

Results
We identified 107 rare nonsynonymous variants in sporadic ICH, of which 2 missense variants, rs138269346 (COL4A1I110T) and rs201716258 (COL4A2H203L), were predicted to be highly functional and occurred in multiple ICH cases but not in controls from the United States–based cohort. The minor allele of rs201716258 was also present in Scottish patients with ICH, and rs138269346 was observed in 2 ICH-free controls with a history of hypertension and myocardial infarction. Rs138269346 was nominally associated with nonlobar ICH risk (p = 0.05), but not with lobar ICH (p = 0.08), while associations between rs201716258 and ICH subtypes were nonsignificant (p > 0.12). Both variants were considered pathogenic based on minor allele frequency (<0.00035 in European populations), predicted functional impact (deleterious or probably damaging), and in silico modeling studies (substantially altered physical length and thermal stability of collagen).

*These authors contributed equally to this work.
Intracerebral hemorrhage (ICH) accounts for 10%–15% of strokes but is the most fatal and least treatable stroke type. More than half of patients with ICH die within the first year after the disease, and most survivors have prolonged disability. The molecular pathways underlying ICH remain poorly understood, limiting therapeutic development.

Genome-wide association studies (GWAS) have identified genetic loci associated with ICH risk and outcome. We recently determined a genome-wide significant association for cerebral small vessel disease (CSVD) including nonlobar ICH and small vessel ischemic stroke (SVS) with COL4A1 and COL4A2 (collagen IV a chain 1 and 2) at 13q34.6 validating previous studies.

Rare mutations that most frequently affect Gly residues in the Gly-Xaa-Yaa repeat in COL4A1/A2 cause mendelian early-onset cerebrovascular disease, ocular dysgenesis, and myopathy. However, the full spectrum of effects of COL4A1/A2 mutations remain incompletely characterized. Sequencing of a few families or small number of patients (n < 100) suggested rare variants in COL4A1/A2 can contribute to sporadic ICH.

To identify rare coding variants in COL4A1/A2 that may underlie the aforementioned GWAS association with ICH and SVS, we performed targeted sequencing of the 13q34 region in United States–based studies (1,055 ICH cases and 1,078 ICH-free controls) as well as exonic sequencing of this locus in 192 ICH cases and whole genome sequencing (WGS) of 1,189 ICH-free controls from Scotland. We present annotation results, prediction of deleteriousness, and case/control segregation of rare variants across 13q34 from these datasets.

**Methods**

**Standard Protocol Approvals, Registrations, and Participant Consents**

The study protocols were approved for enrollment of the United States–based cohorts (Gene Discovery for Warfarin-Related Intracerebral Hemorrhage [GOCHA] and Ethnic/Racial Variation in Intracerebral Hemorrhage [ERICH]) and the Scotland-based cohorts (the Lothian study of Intracerebral Haemorrhage Pathology, Imaging and Neurologic Outcome [LINCHPIN] and Lothian Birth Cohorts [LBC]) by the institutional review boards of Massachusetts General Hospital and University of Edinburgh. Informed consent was obtained from participants or an appropriate legal surrogate according to all recruiting sites.

**Study Participants**

**United States–Based Participants**

We collected DNA and phenotyping data on 1,055 patients with ICH (534 lobar and 521 nonlobar ICH) and 1,078 ICH-free controls from GOCHA and ERICH. There is no overlap between the participants in the present study and those included in previous targeted sequencing studies of COL4A1/A2. A total of 145 participants (18 ICH and 176 controls) in this study appeared in the previous ICH GWAS.

**Scotland-Based Participants**

We collected DNA and phenotyping data on 192 ICH cases (40 lobar and 152 nonlobar) from LINCHPIN among the sample registered in the Edinburgh Lothian Audit of the Treatment of Cerebral Hemorrhage study.

As a control population for the LINCHPIN cohort, we selected clinical and sequencing data from 1,189 individuals from the LBC study as an ethnically and geographically matched stroke-free control population to cases. Self-reported ethnicity of the patients with ICH and controls was white British. The LBC study recruited people living in Edinburgh and the Lothians who were born either in 1921 or 1936. Individuals who self-reported stroke were removed for this study.

Demographic information for the 2,133 United States–based and 1,381 Scotland-based participants are presented in table 1.
Targeted Sequencing at 13q34 and Quality Control (GOCHA/ERICH and LINCHPIN)

We performed targeted sequencing across 559 Kbp at 13q34 (chr13:110,701,882–111,265,349) including COL4A1/A2 for 2,133 individuals (1,055 ICH cases and 1,078 controls) in GOCHA/ERICH on the 96plex Nimblegen SeqCap platform and for 192 ICH cases in LINCHPIN on the NextSeq 550 platform. The reads with high quality are aligned to the human genome (GRCh37) using BWA (version 0.7.10).20 We applied preprocessing steps including duplicate removal, local realignment, indel realignment using Genome Analysis Toolkit (GATK) IndelRealigner, and base quality recalibration using GATK Base Recalibration before the calling. Then GATK best practices pipeline was used for calling the variants from those 2 different platforms, separately. The variants were called and filtered using HaplotypeCaller.

Variants passing the GATK variant quality score recalibration metric were retained. We included only single nucleotide variants or indels with a depth of 10 or higher. We excluded variants with a call rate <0.98, case–control call rate difference >0.05, and differential missingness between cases and controls (p < 0.05). Samples were excluded if they had a low average call rate (<0.98), low mean sequence depth (<30), low mean genotype quality (<85), or differential missingness between cases and controls (p < 0.05) and Hardy-Weinberg equilibrium test (p value < 10^-6) on cases and controls separately.

Whole Genome Sequencing (LBC)

WGS of the LBC sample has been performed previously, with sequencing protocol and quality control (QC) metrics as reported.21 The variants selected from the GOCHA/ERICH and LINCHPIN studies were further evaluated in the LBC sequencing data as a control population.

Variant Selection and Annotation

We used the Ensembl variant effect predictor (VEP) software to annotate functional consequences of variants with the Ensembl annotation database on human genome assembly GRCh37.22 VEP provides various information of functional consequences of variants (splice acceptor variant, splice donor variant, start lost, stop lost, stop gained, frameshift variant, inframe insertion, inframe deletion, and missense variant) and functional impacts predicted by SIFT (deleterious or tolerate), PolyPhen (probably damaging, possibly damaging, benign, or unknown), combined annotation-dependent deletion (CADD), and LoFtool (0–1).26 We also used SnpEff, which categorizes the effects of variants by their genomic locations (e.g., introns, regulatory sites, splicing, missenses, nonsense) into high, moderate, low, or modifier.27 We selected nonsynonymous variants with high/moderate by SnpEff, deleterious by SIFT, probably/possibly damaging by PolyPhen, CADD scores >20.0, and LoFtool scores <0.1 in the GOCHA/ERICH and LINCHPIN datasets.

Single-Variant and Variant-Set Association Tests in GOCHA/ERICH

We used the CATS online calculator28 to estimate power to detect ICH association for single variant analysis. We performed single-variant association tests of those variants selected by annotation methods for the ICH risk by subtype (all-mixed, lobar, and nonlobar ICH) using logistic regression in R (r-project.org/). The association models were adjusted for age and sex. Significance thresholds for the single variant test was set as α = 0.05 with Bonferroni correction for the numbers of single variants (number of variants: 39; p < 1.28 × 10^-3).

Identification and Selection of Potential Causal Rare Variants in ICH (GOCHA/ERICH)

Because our sample size remained extremely underpowered for identifying rare functional variants by conventional association tests, we focused on rare nonsynonymous variants in COL4A1/A2, particularly those that occur in more than 1 ICH case but not in controls, or in more than 1 control but not in cases. Variants were first explored in GOCHA/ERICH as a discovery dataset and then examined in LINCHPIN and LBC as an external validation of the findings in an independent population.

We applied PLINK29 software to assess linkage disequilibrium (LD) and haplotype inferences at this locus, testing whether the common single nucleotide polymorphism (SNP), rs9515201, previously identified in a GWAS of CSVD risk, lies within a haplotype with our identified rare variants.

Because the genetic architectures of lobar and nonlobar ICH are known to differ,5 we tested ICH subtype-specific effects for identified variants using the BinomiRare exact test,30 which uses Poisson-binomial probabilities to calculate the association strength of variants based on the probabilities of diseased individuals carrying minor alleles of the variants under the null hypothesis that the variants are not associated with the disease. A classical logistic regression analysis was not applied for single-variant test because our selected variants occur only in 1 group (e.g., cases or controls), but not both.31 The BinomiRare regression models for ICH subtypes were adjusted for age and sex.

Our identified variants were evaluated using gnomAD32 for allele frequencies across populations and Geno2MP (geno2mp.gs.washington.edu/Geno2MP) for phenomic effect in carriers with variants or their family members.33 To further assess our selected variants, we also explored the recent whole exome sequencing (WES) of 200,000 individuals in UK Biobank (UKB).34,35 We leveraged 137,766 unrelated British White individuals based on kinship relatedness (removing at least 1 of a related pair of individuals) and predicted ancestries from principal components for population structure provided by the UKB.36 For ICH-related phenotypes, we used 2 traits in the electric health record including ICD-10 code for ICH (161; UKB field number 41,202) and “vascular/heart problems diagnosed by doctor” (UKB field number 6,150).
The age at onset for these conditions was not available in the UKB database. The differences in the numbers of carriers of our selected variants and noncarriers between cases and controls in the UKB was examined by a χ² test.

### Analysis of Previously Identified Variants on COL4A1/A2 in GOCHA/ERICH

We investigated coding variants in COL4A1/A2 previously identified by other sequencing studies in ICH (rs200786329, rs117412802, rs62621875, and rs201105747) or its related diseases including hereditary angiopathy, nephropathy, aneurysms, and cramps (rs113994104, rs113994105, and rs113994106), small vessel disease (rs121912857 and rs113994107), cerebral vasculopathic disease (rs672601346), and porencephaly (rs113994112, rs113994114). The information about their positions and amino acid changes are provided in table e-1 (data available from Dryad, doi.org/10.5061/dryad.z34tmpgcq). Previously reported variants that were also identified in our sequencing data were analyzed for association with ICH risk using BinomiRare.

### External Validation of Selected Variants in Scotland-Based Datasets

Selected nonsynonymous rare variants appearing only in cases or controls in the United States–based dataset were explored for validation in LINCHPIN and LBC to determine whether they segregated with cases or controls in these independent datasets. Because LINCHPIN and LBC utilized different recruitment methods and were sequenced using different approaches, discovery of candidate variants in these datasets for validation in the United States–based GOCHA/ERICH cases was not feasible due to concerns for bias.

### In Silico Modeling of Protein Structure and Thermal Stability of Identified Variants

We further assessed structural effects of selected variants on protein structures of COL4A1/A2 using molecular dynamics (MD) simulation. The α chains in COL4A1 and COL4A2 interact to form 1 triple-helical collagen type IV protomer, a triple-helical collagen (IV), that consists of 3 protein domains: N-terminal 7S, central triple-helical collagen, and C-terminal NC1 (figure 1B).

Thus far, the triple-helical collagen structure for a1α1a2(IV) has not been experimentally determined in the Protein Data Bank (PDB; rsb.org). Therefore, we predicted the tertiary heterotrimer structure of a1α1a2(IV) molecule using protein secondary/tertiary structure prediction tools including PSIPRED, PFAM, BLASTp (blast.ncbi.nlm.nih.gov/BLAST.cgi), and DESMOND (deshawresearch.com/resources_desmond.html) as a template and predicted 2 structure models for our selected variant in COL4A1 according to the number of copies of its mutations such as COL4A1sWT/mt/COL4A2WT and COL4A1smt/mt/COL4A2WT and 1 model for the variant in COL4A2 such as COL4A1sWT/WT/COL4A2mt (figure 2A). To compare structural effects of our variants on a1α1a2(IV), we also generated additional structure models for 2 previously reported variants including rs200786329 (COL4A1) and rs117412802 (COL4A2). These 2 variants were chosen for the comparison because rs200786329 and rs117412802 were most recently identified in targeted sequencing of sporadic ICH cases, other than familial early-onset type IV collagen-related diseases. Moreover, our selected variants were located in the same Gly-Xaa-Yaa pattern or the same functional domain. Furthermore, rs200786329 and rs117412802 showed strong functional changes in terms of intracellular accumulation and extracellular deficiency of COL4A1 or COL4A2 protein levels compared to other variants.

MD simulations were performed with the predicted tertiary structures of a1α1a2(IV) molecules containing our identified variants and the 2 previously reported variants using DESMOND (deshawresearch.com/resources_desmond.html) as described across 500 nanoseconds (ns). We included a modeling assumption that the atom positions of replaced amino acids stabilized after 100 ns from the start. Maestro (schrodinger.com/maestro; Schrödinger, LLC, 2016) was used to extract and analyze 800 sequential snapshots of simulated protein structures of COL4A1 and COL4A2 across 400 ns (100–500 ns).

In addition, the effect of the tripeptide composition (Gly-Xaa-Yaa) containing our identified variants on collagen triple helix stability was estimated using the collagen thermal stability calculator (compbio.cs.princeton.edu/csc), which predicts the melting temperature (Tm).
Highly functional nonsynonymous variants appearing in patients with intracerebral hemorrhage (ICH) predicted by various annotation methods and selected COL4A1 and COL4A2 variants.

(A) Highly functional nonsynonymous variants appearing in ICH cases predicted by annotation methods including SnpEff (high or moderate), SIFT (deleterious), PolyPhen (probably/possibly damaging), combined annotation-dependent deletion (CADD) (>20.0), and LoFtool (>=0.1). Green: N-terminal 7S domain; blue: triple helix collagen domain; brown: C-terminal NC domain. *Variants that exist in both datasets. (B) Selected COL4A1 and COL4A2 variants. ERICH = Ethnic/Racial Variation in Intracerebral Hemorrhage; GOCHA = Gene Discovery for Warfarin-Related Intracerebral Hemorrhage.
**Data Availability**

Sequencing data used in this study are available on dbGAP (ncbi.nlm.nih.gov/gap/; accession ID: phs000416.v2.p1). Additional data supporting these findings are available by the authors upon reasonable request.

**Results**

**Identification and Genetic Association Tests of Rare Variants in COL4A1/A2**

Following targeted sequencing and stringent QC of the 13q34 region in the United States–based GOCHA/ERICH studies, we retained 1,055 patients with ICH and 1,078 ICH-free controls with 11,815 variants. Among these variants, 0.91% were nonsynonymous (106 missense, 1 nonsense) and 0.70% were synonymous coding variants.

In the Scotland-based LINCHPIN study, 192 patients with ICH remained after QC with 1,380 variants at the 13q34 locus. From these, 1.01% were nonsynonymous (15 missense and 1 nonsense) and 1.88% were synonymous coding variants.

Selecting for nonsynonymous variants with high/moderate impact by SnpEff, deleterious predicted by SIFT, probably/possibly damaging predicted by PolyPhen, CADD scores >20.0, and LoFTool scores >0.1, we were left with 39 variants in GOCHA/ERICH and 4 variants in LINCHPIN, with 3 of these missense variants, including rs3404222, rs117412802, and rs12877501, common between studies. These nonsynonymous variants, predicted to be highly functional based on multiple annotation methods, are shown in figure 1A. None of these variants is in significant LD ($r^2 > 0.4$) with rs9515201, the leading common variant that we previously identified in a recent CSVD GWAS. Detailed information on these variants appears in table e-2 (data available from Dryad, doi.org/10.5061/dryad.34tmpgcq). Annotation information for all detected variants in both sequencing datasets is described in detail in table e-3 (GOCHA/ERICH) and table e-4 (LINCHPIN) (data available from Dryad, doi.org/10.5061/dryad.z34tmpgcq).

According to our statistical power calculation (figure e-1, doi.org/10.5061/dryad.z34tmpgcq), we have extremely limited power (less than 20%) for detecting genetic association of such a rare variant (minor allele frequency [MAF] <0.001) with risk of sporadic ICH. We therefore limited single variant tests to a rare variant (minor allele frequency [MAF] <0.001) with risk power (less than 20%) for detecting genetic association of such variants. We did not observe any nonsynonymous variants that occurred in at least 2 controls but not in cases, but we did identify 2 rare missense variants (rs138269346 and rs201716258) appearing in at least 2 ICH cases without occurrence in the controls (table 2). Rs138269346 (COL4A1 Ile110Thr) is in exon 5 of COL4A1 corresponding to the N-terminal 7S domain and rs201716258 (COL4A2 His203Leu) is in exon 10 of COL4A2 corresponding to the triple-helix collagen domain where it affects an X residue of the Gly-Xaa-Yaa repeat (figure 1B).

Functional Annotations of rs138269346 and rs201716258

Based on gnomAD, these 2 variants are rare in European populations (MAF <0.0005) and even rarer in other populations (table 2). Rs138269346 (COL4A1) was deemed pathogenic due to its annotation across multiple prediction tools: moderate by SnpEff, deleterious by SIFT, probably damaging by PolyPhen, and probably damaging by LoF- tool. It showed a very high CADD score of 23.8, suggesting that this variant is among the top 0.42% ($10^{-2.38}$) of most predicted deleterious variants in the human genome. The other variant, rs201716258 (COL4A2), was also considered to have a negative effect by SnpEff (moderate) and LoFtool (probably damaging), but not by SIFT (tolerated), PolyPhen (benign), and CADD (4.7).

Validation of rs138269346 and rs201716258 in Scotland-Based Participants

We set out to validate our identified variants using the independent LINCHPIN dataset. We detected 2 patients with ICH who carry the minor allele of rs201716258 (COL4A2) but no carriers for rs138269346 (COL4A1). For rs201716258, both individuals developed nonlobar ICH in the right ventricle (ICH epicenter was lentiform) with a low small vessel disease score. The ages of patients (1 male, 1 female) were 85 and 73 at the time of ICH, and there was no history of ischemic stroke.

Turning to the population-based LBC cohort, we identified 2 ICH-free controls who carry the minor allele of rs138269346 (COL4A1) but no carriers for rs201716258 (COL4A2). One of these 2 ICH-free controls carrying the rs138269346 (COL4A1) in LBC self-reported a history of hypertension and myocardial infarction.
Further Evaluation of rs138269346 and rs20176258 in Geno2MP and UKB

Within individuals contributing to the Geno2MP database, carriers of these variants expressed phenotypic abnormalities previously identified in patients with rare collagen IV mutation syndromes. In Geno2MP, we found 5 heterozygous rs138269346 (COL4A1) carriers and 6 heterozygous rs20176258 (COL4A2) carriers (table e-8, data available from Dryad, doi.org/10.5061/dryad.z3tmpgpcq).
One of the rs138269346\textsuperscript{COL4A1} carriers had multiple abnormalities in the eyes and the cardiovascular and nervous systems, and relatives of the other 4 carriers also had abnormalities in the eye, musculature, and cardiovascular and nervous systems. Two of rs201716258\textsuperscript{COL4A2} carriers had abnormalities in the ears and the cardiovascular system, respectively, and the relatives of the other 4 carriers also had abnormalities in the nervous system or musculature.

Among the 137,766 unrelated individuals in the 200K WES (table e-9), we found 1 rs138269346 carrier but none for rs201716258 out of 276 proxy patients with ICH (by ICD-10 code I61). Out of 41,032 individuals with vascular/heart problems, there are 33 carriers for rs138269346 and 63 carriers for rs201716258. According to our $\chi^2$ analysis, rs138269346 was nominally associated with the proxy ICH ($\chi^2 = 3.4; p = 0.063$) although there is only 1 proxy ICH case carrying rs138269346. We did not observe significant association in the rest of the $\chi^2$ tests (table e-10, data available from Dryad, doi.org/10.5061/dryad.z34tmpgcq).

**Genetic Association Tests With ICH Subtypes**

To determine which ICH subtype is relevant to these variants, we employed BinomiRare (table 3), which revealed that rs138269346\textsuperscript{COL4A1} is nominally associated with risk for nonlobar ICH ($p = 0.054$) and less associated with lobar ICH ($p = 0.077$) and all-mixed ICH ($p = 0.090$). However, rs201716258\textsuperscript{COL4A2} did not carry a significant association with any ICH subtype (all $p$ values $> 0.12$).

Among the previously reported variants in COL4A1/A2, we could detect 3 nonsynonymous variants in COL4A2, including rs117412802, rs62621875, and rs201105747, which were previously reported in only patients with ICH and not in controls.\textsuperscript{17} However, in the GOCHA/ERICH dataset, we found carriers of these variants in both ICH cases (carriers of rs117412802: 20; and rs62621875: 1) and controls (rs117412802: 24; and rs62621875: 1). We identified rs201105747 only in 1 control. From the LINCHPIN dataset, we found only rs117412802 among 4 patients with ICH (3 lobar and 1 nonlobar ICH). We found no significant associations of these observed variants with ICH risk from BinomiRare (table e-11, data available from Dryad, doi.org/10.5061/dryad.z34tmpgcq).

**In Silico Functional Analysis of rs138269346 and rs201716258**

To gain insight into any structural potential consequences of these 2 missense variants, we performed MD simulations. Based on a chain composition of $\alpha_1\alpha_1\alpha_2(IV)$, each protomer can contain 1 or 2 variant $\alpha_1(IV)$ molecules and 1 $\alpha_2(IV)$. We therefore constructed 2 structural models for rs138269346\textsuperscript{COL4A1}, COL4A1\textsuperscript{WT/110T}/COL4A2\textsuperscript{WT} and COL4A1\textsuperscript{I110T/110T}/COL4A2\textsuperscript{WT}. Next, for rs201716258\textsuperscript{COL4A2}, 1 structure model was predicted: COL4A1\textsuperscript{WT/WT}/COL4A2\textsuperscript{H203L} (figure 2A). From these simulations, we observed that 2 mutation structures of rs138269346\textsuperscript{COL4A1}, COL4A1\textsuperscript{WT/110T}/COL4A2\textsuperscript{WT} and COL4A1\textsuperscript{110T/110T}/COL4A2\textsuperscript{WT}, result in average lengths...
of 16.59 ångströms (Å; equal to $10^{-10}$m) (SD 5.58 Å) and 12.64 Å (SD 4.29 Å), respectively (figure 2B). These are 14.9% and 35.1% shorter than the average length of the wild type (average 19.49 Å; SD 6.41 Å). Likewise, we also found that the mutated structure for rs20176258COL4A2, COL4A1WT/WT/COL4A2H2O3L (average 19.13 Å; SD 6.03 Å), became 15.7% shorter than the wild type (average 22.70 Å; SD 5.94 Å).

For comparison, we also generated structural models of the previously reported COL4A1/A2 variants rs200786329COL4A1 and rs117412802COL4A2. Interestingly, we found the same pattern in the simulations of the structure models with mutations of the previously reported variants. The average lengths of those mutant forms for rs200786329COL4A1 and rs117412802COL4A2 are 9.7%–25.8% shorter than their wild types (figure 2B).

Table 3 Genetic Association Test Summary of the COL4A1/A2 Variants Using Binomirare

<table>
<thead>
<tr>
<th>Variant ID (effect allele)</th>
<th>All ICH</th>
<th>Lobar ICH</th>
<th>Nonlobar ICH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N in cases</td>
<td>Expected N in cases</td>
<td>p Value</td>
</tr>
<tr>
<td>rs138269346 (G)</td>
<td>4</td>
<td>2.05</td>
<td>0.09</td>
</tr>
<tr>
<td>rs201716258 (T)</td>
<td>2</td>
<td>0.98</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Abbreviation: ICH = intracerebral hemorrhage.

It is well established that the amino acid sequence of the collagenous domain and mutations in collagen influence the thermal stability and melting temperature of the triple helix.46 According to the collagen stability calculator,44 COL4A1I100T reduces predicted $T_m$ by 2.2°C compared to COL4A1wt, while COL4A2H2O3L increases the $T_m$ by 2.1–3.1°C compared to COL4A2wt (figure e-2, data available from Dryad, doi.org/10.5061/dryad.z34pmcg), further lending support to their functionality.

Discussion

We conducted a targeted sequencing study of 1,055 ICH cases and 1,078 controls, and an additional dataset (192 ICH cases) at the 13q34 locus, which we and others have previously identified in genetic association studies for ICH and related manifestations of CSVD.67 13q34 has been the target of prior sequencing studies in ICH, but the sample size in this current study (n = 2,325) is approximately 10 times larger than those prior efforts (n < 200).8,9,12,13,15,17,37 However, even with this boost in sample size, statistical power remains low for rare variants, so we focused on nonsynonymous variants appearing only in ICH cases but not in controls to restrict our search to the most likely pathogenic variants. Our study identified 2 novel missense variants in COL4A1/COL4A2, appearing in cases only in our discovery dataset that (1) are predicted to have damaging effects, (2) are not in LD with common SNP identified in the previous GWAS for CSVD, and (3) are predicted to substantially alter the physical length and the thermal stability of the type IV collagens. Furthermore, despite the limited statistical power of our sample size, we did observe a nominally significant association at rs138269346COL4A1 for nonlobar ICH risk.

In our attempt to externally validate these results in independent datasets, we detected 2 patients with ICH carrying rs201716258COL4A2 in LINCHPIN. No ICH case carriers for rs138269346COL4A1 were found, although 2 individuals with this variant were identified in the LBC control population. Given the size difference between the case populations in the 2 studies and the differences in recruitment between the case/control datasets and LBC, the significance of this replication failure at rs138269346COL4A1 is unclear.

The majority of mendelian disease–associated mutations in COL4A1/A2 identified to date affect the Gly residue in the Gly-Xaa-Yaa repeat pattern, which is necessary for stable triple-helix formation, and more mutations have been described in COL4A1 than COL4A2.14 However, the amino acids of our selected variants are located on the X residue in the pattern. Among the previously identified mutations for sporadic ICH, P352L3 rs200786329 (COL4A1)13 and E11223Grs117412802 (COL4A2)17 are also located on non-Gly residue in the tripeptide pattern, the Y and X residues, respectively. These previous variants (rs200786329 and rs117412802) significantly reduced the ratio of extracellular to intracellular COL4A1 or COL4A2 proteins compared to wild types in cultured cells, which is considered as a potential pathogenic mechanism underlying the type IV collagens related diseases.3,17 Interestingly, our MD simulation shows that these previous variants (rs200786329rs138269346 and rs20176258) substantially shorten the physical lengths of triple-helical structures, indicating that altered protein structures such as shortening of the helical domains could lead to changes in the protein functions including secretion of type IV collagen into extracellular matrix or secretion of mutant protein. In this regard, our thermal stability study predicted that COL4A1I100T decreases (−2.2°C) and COL4A2H2O3L increases (+2.1–3.1°C) $T_m$ of collagen structures. These changes are not small in magnitude compared to the 2°C reduction affected by mendelian disease–associated Col2a1 mutations.47 Therefore, it also may be possible that our identified variants exert their effects on type IV collagen...
function through affecting the thermal stability of collagen, analogous to other collagen mutations.

Our study has several limitations. First, due to limited statistical power, we focused on rare variants appearing only in ICH cases but not controls, which could be considered conservative and increase false-negative associations for variants with low penetrance. Furthermore, we lack an analogous case/control ICH dataset with deep sequencing for direct validation of our observed variants. Targeted sequencing of 13q34 in LINCHPIN allowed us to verify the existence of rs201716258 in ICH cases, but the relatively small sample size and substantial difference in proportions of ICH subtypes compared to the discovery dataset makes the lack of observation of rs138269346 inconclusive. The population-based subjects from LBC are a welcome adjunct to the LINCHPIN ICH cases, but the varying burden of ICH-relevant disease histories and comorbidities and prospective nature of the LBC study, which is not continually updated, make interpretation of the appearance of rs138269346 in ICH-free individuals challenging. Due to the lack of genome-wide genotyping on the vast majority of subjects, population structure could not be assessed and adjusted for in our single variant association tests. While interesting and supportive of known mechanisms related to associations at common variant studies of ICH, studies building on our observations may link the pathogenic processes that underlie monogenic and polygenic ICH risk at this locus.

**Study Funding**

Supported by US-NINDS R01NS103924, US NHLBI RS&G 224, UK MRC MR/R005567-1, G0900428, G1002605, and 203699/Z/16/Z. Targeted sequencing services were provided by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute (RS&G 224). T. Van Agtmael is supported by funding from the UK Medical Research Council (MRC; MR/R005567-1), Stroke Association (PFA 2016/02), and Heart Research UK (RG 2664/17/20). The LINCHPIN study was supported by an MRC/Stroke Association clinical research training fellowship (G0900428) awarded to Dr. Samarasekera, a Wellcome Trust clinical research training fellowship awarded to Dr. Rodrigues (203699/Z/16/Z), and an MRC senior clinical fellowship (G1002605) awarded to Prof. Al-Shahi Salman; the DNA sequencing was supported by a project grant from the MRC (MR/R005567/1). Phenotype collection in the Lothian Birth Cohort 1921 was supported by the United Kingdom’s Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society, and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Age UK (The Disconnected Mind project). Whole-genome sequencing of the Lothian Birth Cohorts was funded by the BBSRC.

**Disclosure**


**Publication History**

Received by Neurology November 17, 2020. Accepted in final form April 19, 2021.
## Appendix Authors

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaeyoon Chung, PhD</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; analysis or interpretation of data</td>
</tr>
<tr>
<td>Graham Hamilton, PhD</td>
<td>University of Glasgow, Bearsden, UK</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; analysis or interpretation of data</td>
</tr>
<tr>
<td>Minsup Kim, PhD</td>
<td>Broad Institute, Boston, MA</td>
<td>Analysis or interpretation of data</td>
</tr>
<tr>
<td>Sandro Marini, MD</td>
<td>Korea University, Sejong, South Korea</td>
<td>Analysis or interpretation of data</td>
</tr>
<tr>
<td>Bailey Montgomery</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Analysis or interpretation of data</td>
</tr>
<tr>
<td>Jonathan Henry</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Analysis or interpretation of data</td>
</tr>
<tr>
<td>Art E. Cho, PhD</td>
<td>Korea University, Sejong, South Korea</td>
<td>Analysis or interpretation of data</td>
</tr>
<tr>
<td>Devin L. Brown, MD</td>
<td>University of Michigan, Ann Arbor</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>Bradford B. Worrall, MD</td>
<td>University of Virginia Health System, Charlottesville</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>James F. Meschia, MD</td>
<td>Mayo Clinic, Jacksonville, FL</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>Scott L. Silliman, MD</td>
<td>University of Florida College of Medicine, Jacksonville</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>Magdy Selim, MD, PhD</td>
<td>Beth Israel Deaconess Medical Center, Boston, MA</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>David L. Tirschwell, MD, MSc</td>
<td>University of Washington, Seattle</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>Chelsea S. Kidwell, MD</td>
<td>The University of Arizona, Tucson</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content</td>
</tr>
<tr>
<td>Brett Kissela, MD, MS</td>
<td>University of Cincinnati, OH</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Steven M. Greenberg, MD, PhD</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Anand Viswanathan, MD, PhD</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Joshua N. Goldstein, MD, PhD</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Carl D. Langefeld, PhD</td>
<td>Wake Forest School of Medicine, Winston-Salem, NC</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Kristiina Rannikmae, MD</td>
<td>Wake Forest School of Medicine, Winston-Salem, NC</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Catherine L.M. Sudlow, MBMBCh, MSc, DPhil, FRCP Edin</td>
<td>Wake Forest School of Medicine, Winston-Salem, NC</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>Neshika Samarakseker, PhD, MRCP</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Mark Rodrigues, PhD, FRCP</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Rustam Al-Shahi Salman, PhD, FRCP Edin</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>James G.D. Prendergast, PhD</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Sarah E. Harris, PhD</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Ian Deary, MBChB, PhD, FRCP</td>
<td>University of Edinburgh, Scotland, UK</td>
<td>Major role in the acquisition of data</td>
</tr>
<tr>
<td>Daniel Woo, PhD</td>
<td>University of Cincinnati, OH</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>Jonathan Rosand, MD, MSc</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data</td>
</tr>
<tr>
<td>Tom Van Agtmael, PhD</td>
<td>University of Cincinnati, OH</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data</td>
</tr>
<tr>
<td>Christopher D. Anderson, MD, MMSc</td>
<td>Massachusetts General Hospital, Boston</td>
<td>Drafting/revision of the manuscript for content; including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data</td>
</tr>
</tbody>
</table>
Rare Missense Functional Variants at COL4A1 and COL4A2 in Sporadic Intracerebral Hemorrhage
Jaeyoon Chung, Graham Hamilton, Minsup Kim, et al.
Neurology 2021;97:e236-e247 Published Online before print May 24, 2021
DOI 10.1212/WNL.00000000000012227

This information is current as of May 24, 2021

Updated Information & Services
including high resolution figures, can be found at:
http://n.neurology.org/content/97/3/e236.full

References
This article cites 49 articles, 10 of which you can access for free at:
http://n.neurology.org/content/97/3/e236.full#ref-list-1

Subspecialty Collections
This article, along with others on similar topics, appears in the following collection(s):
All Cerebrovascular disease/Stroke
http://n.neurology.org/cgi/collection/all_cerebrovascular_disease_stroke
Association studies in genetics
http://n.neurology.org/cgi/collection/association_studies_in_genetics
Intracerebral hemorrhage
http://n.neurology.org/cgi/collection/intracerebral_hemorrhage

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
http://www.neurology.org/about/about_the_journal#permissions

Reprints
Information about ordering reprints can be found online:
http://n.neurology.org/subscribers/advertise