# UC Davis
## UC Davis Previously Published Works

**Title**

Genome-wide scan in Hispanics highlights candidate loci for brain white matter hyperintensities

**Permalink**

**Journal**

**ISSN**

**Authors**

Beecham, Ashley
Dong, Chuanhui
Wright, Clinton B
et al.

**Publication Date**

**DOI**

# Genome-wide scan in Hispanics highlights candidate loci for brain white matter hyperintensities

Ashley Beecham, MS
Chuanhui Dong, PhD
Clinton B. Wright, MD
Nicole Dueker, PhD
Adam M. Brickman, PhD
Liyong Wang, PhD
Charles DeCarli, MD
Susan H. Blanton, PhD
Tatjana Rundek, MD, PhD
Richard Mayeux, MD, MS
Ralph L. Sacco, MD

Correspondence to
Dr. Sacco:
rsacco@med.miami.edu

## ABSTRACT

**Objective:** To investigate genetic variants influencing white matter hyperintensities (WMHs) in the understudied Hispanic population.

**Methods:** Using 6.8 million single nucleotide polymorphisms (SNPs), we conducted a genome-wide association study (GWAS) to identify SNPs associated with WMH volume (WMHV) in 922 Hispanics who underwent brain MRI as a cross-section of 2 community-based cohorts in the Northern Manhattan Study and the Washington Heights–Inwood Columbia Aging Project. Multiple linear modeling with PLINK was performed to examine the additive genetic effects on ln (WMHV) after controlling for age, sex, total intracranial volume, and principal components of ancestry. Gene-based tests of association were performed using VEGAS. Replication was performed in independent samples of Europeans, African Americans, and Asians.

**Results:** From the SNP analysis, a total of 17 independent SNPs in 7 genes had suggestive evidence of association with WMHV in Hispanics ($p < 1 \times 10^{-5}$) and 5 genes from the gene-based analysis with $p < 1 \times 10^{-3}$. One SNP (rs9957475 in *GATA6*) and 1 gene (*UBE2C*) demonstrated evidence of association ($p < 0.05$) in the African American sample. Four SNPs with $p < 1 \times 10^{-5}$ were shown to affect binding of *SPI1* using RegulomeDB.

**Conclusions:** This GWAS of 2 community-based Hispanic cohorts revealed several novel WMH-associated genetic variants. Further replication is needed in independent Hispanic samples to validate these suggestive associations, and fine mapping is needed to pinpoint causal variants. *Neurol Genet* 2017;3:e185; doi: 10.1212/NXG.000000000000185

## GLOSSARY

**GWAS** = genome-wide association study; **HDAC** = histone deacetylase; **LD** = linkage disequilibrium; **mRNA** = messenger RNA; **NOMAS** = Northern Manhattan Study; **SNP** = single nucleotide polymorphism; **WHICAP** = Washington Heights–Inwood Columbia Aging Project; **WMH** = white matter hyperintensity; **WMHV** = WMH volume.

White matter hyperintensities (WMHs) are frequently detected by MRI in the aging brain[1] and are associated with a range of negative health outcomes.[2–6] Prevalence ranges from 40% to 70% in the fifth decade[7] and increases with age. It is important that Hispanics and African Americans have shown more severe WMHs than ancestral Europeans.[8] Although heterogeneous in etiology, there is a consistent link between WMH burden and cerebrovascular risk factors, as well as with retinal microvascular abnormalities and vascular pathology.[9] Thus, WMHs could be considered a quantitative marker of small vessel injury.[10] Moreover, WMHs have a significant genetic component, with heritability estimates from 0.45 to 0.80.[10–13]

Identifying genetic determinants of WMHs has been a challenge. Linkage studies have yielded conflicting findings regarding the genetic loci influencing WMHs.[12–14] The most

consistently replicated locus identified for WMHs is on chromosome 17q25 and was first identified in ∼9,500 individuals of European descent.[15] This locus was later replicated and additional loci on chromosomes 10q24 and 2p21 identified.[16] A multiethnic meta-analysis, including ∼2,000 African Americans, ∼800 Hispanics, and ∼400 Asians, identified additional loci on chromosomes 1q22 and 2p16.[16]

Although Hispanics and African Americans show more severe WMHs than Europeans,[8] these populations have not been the focus of genetic discovery. To address this, we performed a genome-wide association scan to identify genetic variants influencing WMHs in Hispanics in 2 community-based cohorts from the Northern Manhattan Study (NOMAS) and the Washington Heights–Inwood Columbia Aging Project (WHICAP).

**METHODS** **Participants.** Hispanic participants from NOMAS and WHICAP with WMHs and genotype data available were included in this study. All participants provided written informed consent. The study was approved by the Institutional Review Boards of the University of Miami and Columbia University Medical Center. The research design and detailed ascertainment scheme for NOMAS and WHICAP were described in detail previously.[17,18] To be eligible for recruitment in the NOMAS, participants had to be at least 40 years of age, never been diagnosed with a stroke, and resided for at least 3 months in a household with a telephone in Northern Manhattan. A total of 3,298 participants were enrolled from 1993 to 2001, and 199 unrelated household members were recruited from 2003 to 2008. Data were collected at enrollment using standardized data collection instruments, medical records, and fasting blood samples.

For the WHICAP, participants were recruited in 2 waves (1992–1994 and 1999–2002) through the use of a probability sample of Medicare beneficiaries who were older than 65 years and residing in a defined area of Northern Manhattan. Each WHICAP participant underwent an in-person interview of general health and functional ability at the time of study entry, followed by a standardized assessment, including medical history, physical and neurologic examination, and a neuropsychological battery.

**Quantitative measurement of WMH volume.** Selected NOMAS and WHICAP participants were scanned on the same 1.5T Philips Intera scanner at Columbia University Medical Center using the same sequences. Participants were selected for MRI, on average 6 years after recruitment, if they were 55 years or older, were classified as nondemented at a previous study visit, had no contraindications for MRI, and signed informed consent.[17,18] Data sets were transferred electronically to UC Davis for morphometric analysis as previously described.[19] Briefly, nonbrain elements were manually removed from the image by operator-guided tracing of the dura mater within the cranial vault, including the middle cranial fossa but excluding the posterior fossa and cerebellum. The resulting measure of the cranial vault was defined as the total intracranial volume. After image intensity nonuniformities were removed from the image, the corrected image was modeled as a mixture of 2 Gaussian probability functions with the segmentation threshold determined at the minimum probability between these 2 distributions. Once brain matter segmentation was achieved, a single Gaussian distribution was fitted to image data, and a segmentation threshold for WMHs was determined a priori as 3.5 SDs in pixel intensity above the mean of the fitted distribution of the brain parenchyma.[20] For the NOMAS, high interrater reliabilities were documented for both the intracranial volume (0.97) and WMH (0.99).[2]

**Genotyping, imputation, and quality control.** DNA samples were obtained through whole blood extraction. For the NOMAS, genotyping of DNA samples was performed using Affymetrix Genome-Wide Human SNP Array 6.0 chips according to Affymetrix procedures at the Genotyping Core of the John P. Hussman Institute for Human Genomics at the University of Miami. Genotype calling was performed using Affymetrix Power Tools v.1.15.0. For the WHICAP, genotyping was performed in 2 batches on the Illumina HumanHap 650Y (WHICAP1) and Omni 1M (WHICAP2) chips, according to Illumina procedures at Columbia University. Genotype calling was performed separately for WHICAP1 and WHICAP2 using GenomeStudio v.1.0. Genome-wide genotype data were available on a total of 1,067 Hispanics meeting all inclusion criteria from either NOMAS (811) or WHICAP (191 WHICAP1 and 65 WHICAP2). Because both NOMAS and WHICAP were sampled from Northern Manhattan, we expected some participants to be enrolled in both studies. Therefore, we checked for unexpected relationships with identical-by-descent proportion >0.25 using PLINK[21] and found 34 duplicate and 15 related pairs between NOMAS and WHICAP1 or WHICAP2. After dropping at least 1 from each pair, typically the one genotyped on the smaller array, a total of 1,019 unrelated individuals remained for the final joint analysis.

Quality control was applied to both DNA samples and single nucleotide polymorphisms (SNPs). Specifically, samples were removed from further analysis if they had call rates below 95%, sex discrepancies, relatedness, autosomal heterozygosity beyond 4 SDs from the mean. In addition, a subset of ∼51,000 independent SNPs ($R^2 ≤ 0.2$) across the genome were used to infer principal components using Eigenstrat software.[22] Samples were eliminated if they were outliers beyond 6 SDs from the mean based on the first 10 principal components. An additional 50 participants from the WHICAP with a diagnosis of stroke were excluded from further analysis. This led to 760, 97, and 65 samples available for analysis from NOMAS, WHICAP1, and WHICAP2, respectively. SNPs were excluded if they were not in the Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$) or had a genotyping call rate less than 95%. This led to a total of 813,769; 624,026; and 970,293 SNPs available in NOMAS, WHICAP1, and WHICAP2, respectively. A total of 6%, 1%, and 11% of SNPs were dropped because of low call rate for NOMAS, WHICAP1, and WHICAP2, respectively. Only ∼100,000 SNPs overlapped between all 3 data sets, so imputation was performed on each data set using IMPUTE2[23] with the 1000 Genomes Phase I (interim) NCBI Build b37 (June 2011) reference panel. This reference panel included 1,094 individuals from across Europe, Africa, Asia, and the Americas. SNPs were subsequently removed if the imputation quality score was less than 0.8 or the minor allele frequency was less than 0.01 in any of the NOMAS, WHICAP1, or WHICAP2 samples, resulting in a total of 6,783,792 SNPs for analysis.

**Statistical analyses.** To compare the sample characteristics between the NOMAS and WHICAP Hispanic cohorts, a $\chi^2$ test was used to test for frequency differences for categorical variables and a student $t$ test was performed to test for mean or median differences for continuous variables using R.[24] Population substructure was assessed using ADMIXTURE,[25] including reference populations of Europeans (Utah residents with northern and western European ancestry) and Africans (Yoruba in Ibadan, Nigeria) from 1000 Genomes as well as Native Americans from the Human Genome Diversity Project.[26] Because of the skewed distribution, WMHs were log transformed. To evaluate the association of WMHs with each SNP, using PLINK,[21] an additive genetic model was fit by regressing log-transformed WMHs on genotype dosage (0–2 copies of the variant allele) after adjusting for age at MRI, sex, total intracranial volume, and the top 2 principal components of ancestry to account for population substructure. No adjustment for treatment was needed, as all patients were classified as nondemented. These analyses were performed separately in each of the NOMAS, WHICAP1, and WHICAP2 cohorts. An inverse-variance meta-analysis under a fixed-effects model was used in METAL (released March 27, 2011)[27] to summarize results across the 3 strata. SNPs with I² heterogeneity estimates greater than 50% were removed post hoc. VEGAS (v.01)[28] was used to assign SNPs to genes (within 50KB using NCBI Build b37) and produce gene-based test statistics and empirical $p$ values by simulation. A random sample of 200 Hispanics from the NOMAS was used to estimate linkage disequilibrium (LD) patterns for each gene.

To determine the relevance of our findings to other race and ethnic groups, summary statistics were obtained from CHARGE analyses of ~18,000 Europeans, ~2,000 African Americans, and ~400 Asians[16] for each of our top SNPs from the single SNP analysis (using an a priori threshold of $p < 1 \times 10^{-5}$ and independence with $R^2 \leq 0.2$ based on an LD-clumping procedure) and each SNP within our top genes (using an a priori threshold of $p < 1 \times 10^{-3}$). Although the CHARGE analyses[16] included ~800 Hispanics, these samples overlapped largely with NOMAS and WHICAP and therefore did not provide a suitable and independent replication of our study. A multiethnic meta-analysis was performed using a z-score–based fixed-effects model in METAL (released March 27, 2011). An inverse-variance meta-analysis was not possible, given the data format produced by the previous CHARGE z-score–based meta-analyses.

VEGAS2 (v.02)[29] was used to produce gene-based empirical $p$ values from the CHARGE summary statistics using 1000 Genomes reference panels of Europeans, Africans, and Asians to estimate LD patterns for each gene. While VEGAS allows input of a customized reference data set to estimate LD (necessary for our Hispanic cohorts), VEGAS2 does not allow this option. VEGAS2 does, however, provide more extensive reference panel data than VEGAS for the analysis of the European, African American, and Asian samples used in the CHARGE analyses. Nonetheless, the VEGAS and VEGAS2 algorithms are based on the same simulation strategy and produce similar results, given the same input data set. Sensitivity analyses were conducted using MAGMA (v.1.06),[30] which produced similar results (data not shown).

**RESULTS** The sample characteristics of the NOMAS and WHICAP Hispanic MRI cohorts are presented in table 1. Among the 922 Hispanic participants overall, the mean age was 71.1 ± 9.0 years, and 63.7% were women. Compared with NOMAS Hispanic participants, WHICAP Hispanic participants had a greater WMH volume (WMHV) and smaller intracranial volumes because of the older age distribution. The population substructure of each of the NOMAS and WHICAP Hispanic MRI cohorts is presented in figure 1. The overall cohort was on average 57% European, 31% African, and 12% Native American, expected, given the largely Dominican Hispanic population of Northern Manhattan.

In the meta-analysis of Hispanic samples, the quantile-quantile plot (figure e-1 at Neurology.org/ng) did not reveal an obvious departure from the distribution of expected $p$ values by chance (genomic inflation factor [λ] of 0.997), suggesting no significant inflation of probability values due to population substructure. Genome-wide association results are shown in figure 2. While no SNP reached genome-wide significance, 17 independent SNPs ($R^2 \leq 0.2$) were associated with WMHV at $p < 1 \times 10^{-5}$. Frequencies and effect estimates for these 17 SNPs are shown in table 2. The mean changes in ln (WMHV) for per copy increase in these risk alleles ranged from 0.14 to 0.55, corresponding to 14.4%–21.7% of the mean WMHV. Seven of the 17 SNPs were within a known gene. While none of the 17 SNPs demonstrated statistical significance at a multiple testing threshold of $2.9 \times 10^{-3}$ (0.05/17), one of

| Table 1 | WMH measures and associated risk factors | | | | | | |
|---|---|---|---|---|---|---|---|
| **Risk factors and phenotypes** | **Overall (N = 922)** | **NOMAS (N = 760)** | **WHICAP1 (N = 97)** | **WHICAP2 (N = 65)** | **NOMAS vs WHICAP1** | **NOMAS vs WHICAP2** | **WHICAP1 vs WHICAP2** |
| | **N (%)** | | | | **$p$ Value (Pearson χ²)** | | |
| **Female** | 587 (63.7) | 474 (62.4) | 68 (70.8) | 45 (69.2) | $1.69 \times 10^{-01}$ | $3.34 \times 10^{-01}$ | 1.00 |
| | **Mean ± SD** | | | | **$p$ Value (Student $t$ test)** | | |
| **Age at MRI** | 71.1 ± 9.0 | 69.1 ± 8.4 | 80.2 ± 5.3 | 81.0 ± 5.1 | $1.42 \times 10^{-40}$ | $9.92 \times 10^{-31}$ | $3.52 \times 10^{-01}$ |
| **WMHV, mL** | 8.0 ± 9.7 | 6.9 ± 8.7 | 11.3 ± 11.4 | 15.5 ± 13.2 | $3.88 \times 10^{-04}$ | $2.72 \times 10^{-06}$ | $4.13 \times 10^{-02}$ |
| **Intracranial volume, mL** | 1,133.7 ± 112.5 | 1,139.0 ± 112.9 | 1,114.9 ± 108.5 | 1,099.7 ± 105.3 | $4.23 \times 10^{-02}$ | $5.20 \times 10^{-03}$ | $3.74 \times 10^{-01}$ |

Abbreviations: NOMAS = Northern Manhattan Study; WHICAP = Washington Heights-Inwood Columbia Aging Project; WMH = white matter hyperintensity; WMHV = WMH volume.

**Global genetic ancestry**
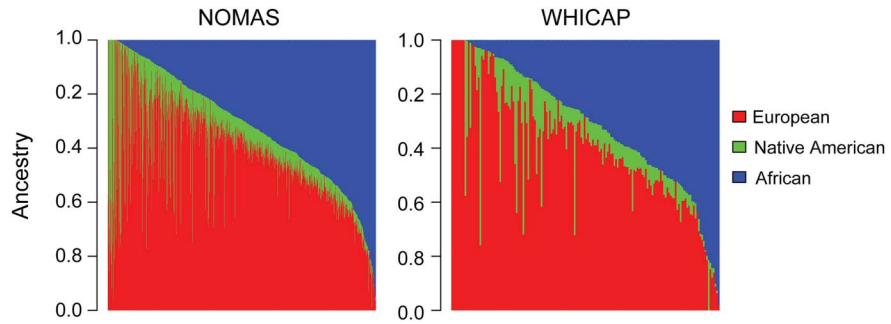


The plots depict the ancestral proportions observed within each cohort, where individual is represented on the x-axis and proportion ancestry on the y-axis. Global ancestry was calculated using 50,327 independent single nucleotide polymorphisms (R² ≤ 0.2), with reference populations including 46 Europeans (Utah residents with northern and western European ancestry from 1000 Genomes), 46 Africans (Yoruba in Ibadan, Nigeria from 1000 Genomes), and 46 Native Americans (Surui and Karitiana of Brazil and Maya and Pima of Mexico from the Human Genome Diversity Project). NOMAS = Northern Manhattan Study; WHICAP = Washington Heights–Inwood Columbia Aging Project.

these 17 SNPs, rs9957475 (risk allele = C, beta = 0.188, $p = 6.19 \times 10^{-6}$) located in GATA-binding protein 6 (*GATA6*), demonstrated evidence of association with $p < 0.05$ in an independent sample of African Americans (rs9957475, risk allele = C, $p = 0.020$). In the NOMAS Hispanic sample, the average WMHV for 0, 1, and 2 copies of the risk allele were 6.08, 8.58, and 8.95 mL, respectively. No SNPs reached genome-wide significance in the z-score–based or sample size–weighted, multiethnic meta-analysis, not surprising given the substantial weight (per the sizable sample) of the largely insignificant European cohort.

Using RegulomeDB,[31] we sought to determine whether any SNP associated with WMHV had regulatory function. Of the 95 SNPs with $p < 1 \times 10^{-5}$ (17 independent SNPs with R² ≤ 0.2), 5 were likely to affect protein binding: one of these variants, rs2670314, was located downstream of GDNF family receptor α 4 (*GFRA4*) on chromosome 20p13, while the remaining 4 variants (rs73306471, rs73306449, rs73306445, and rs73306424) were located on chromosome 20q13 within an LD block spanning 3 genes, including WAP four-disulfide core domain 3 (*WFDC3*), deoxynucleotidyltransferase terminal-interacting protein 1 (*DNTTIP1*), and

**Single nucleotide polymorphism association results for meta-analysis of white matter hyperintensity volume**
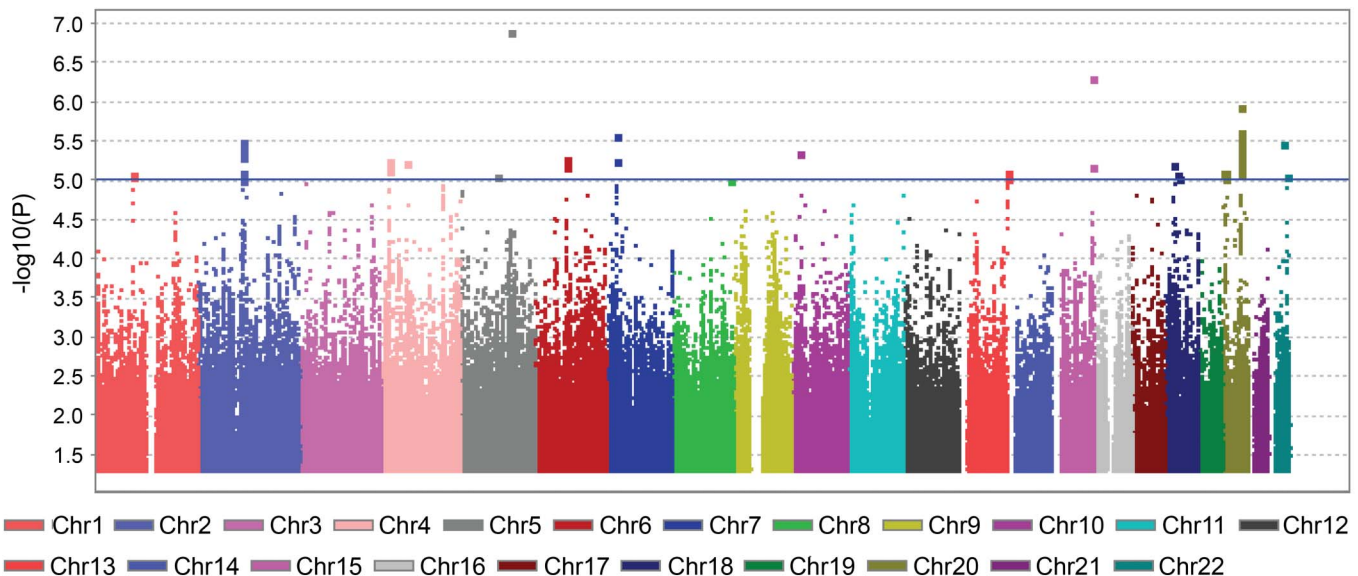


The Manhattan plot depicts the –log 10 *p* values by position from the Hispanic meta-analysis of Northern Manhattan Study and Washington Heights–Inwood Columbia Aging Project single single nucleotide polymorphism association results.

**Table 2** Single SNP association results (independent with $R^2 < 0.2$) with $p \le 1.0 \times 10^{-5}$ among Hispanics

| Chr:MB | rsID | Gene:Func | Hispanic cohorts | | | | | | | CHARGE | | | | | | Meta-$p$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | RA | N FRQ | W1 FRQ | W2 FRQ | β | $p$ | I² | AFR FRQ | AFR $p$ | ASN FRQ | ASN $p$ | EUR FRQ | EUR $p$ | |
| **1:92.0** | rs13447479 | CDC7:intron | G | 0.16 | 0.17 | 0.11 | 0.19 | $8.12 \times 10^{-06}$ | 42.6 | 0.10 | $9.40 \times 10^{-01}$ | 0.26 | $+3.53 \times 10^{-01}$ | 0.18 | $+8.22 \times 10^{-02}$ | $8.83 \times 10^{-03}$ |
| **2:107.4** | rs76479473 | ST6GAL2:UTR3 | A | 0.97 | 0.97 | 0.96 | 0.44 | $3.16 \times 10^{-06}$ | 0.0 | 0.98 | $9.70 \times 10^{-01}$ | 0.95 | $6.33 \times 10^{-01}$ | 0.95 | $+3.04 \times 10^{-01}$ | $6.61 \times 10^{-02}$ |
| **4:18.4** | rs6832924 | — | T | 0.93 | 0.95 | 0.90 | 0.28 | $5.62 \times 10^{-06}$ | 0.0 | 0.79 | $+5.39 \times 10^{-01}$ | NA | NA | NA | NA | $1.97 \times 10^{-03}$ |
| **4:58.4** | rs1372094 | — | T | 0.29 | 0.22 | 0.21 | 0.15 | $5.81 \times 10^{-06}$ | 0.0 | 0.29 | $1.12 \times 10^{-01}$ | 0.02 | $+2.80 \times 10^{-01}$ | 0.28 | $+6.57 \times 10^{-01}$ | $3.35 \times 10^{-01}$ |
| **5:87.4** | rs75027182 | — | T | 0.02 | 0.01 | 0.02 | 0.55 | $8.56 \times 10^{-06}$ | 0.0 | 0.06 | $+6.31 \times 10^{-01}$ | NA | NA | NA | NA | $4.44 \times 10^{-03}$ |
| **5:119.0** | rs162603 | — | G | 0.33 | 0.31 | 0.34 | 0.18 | $1.26 \times 10^{-07}$ | 31.3 | 0.33 | $+8.59 \times 10^{-01}$ | 0.33 | $+2.36 \times 10^{-01}$ | 0.31 | $5.72 \times 10^{-01}$ | $4.77 \times 10^{-01}$ |
| **6:72.4** | rs272210 | — | A | 0.03 | 0.02 | 0.08 | 0.42 | $5.23 \times 10^{-06}$ | 11.9 | 0.09 | $+3.11 \times 10^{-01}$ | NA | NA | NA | NA | $7.67 \times 10^{-04}$ |
| **8:136.1** | rs79780482 | — | A | 0.02 | 0.03 | 0.01 | 0.46 | $9.98 \times 10^{-06}$ | 0.0 | 0.07 | $8.46 \times 10^{-01}$ | 0.01 | $7.33 \times 10^{-01}$ | NA | NA | $1.16 \times 10^{-02}$ |
| **15:98.7** | rs62024995 | — | C | 0.75 | 0.75 | 0.79 | 0.18 | $4.78 \times 10^{-07}$ | 25.8 | 0.89 | $+7.41 \times 10^{-01}$ | 0.75 | $+3.67 \times 10^{-01}$ | 0.63 | $+5.98 \times 10^{-02}$ | $2.59 \times 10^{-03}$ |
| **18:19.8** | rs9957475 | GATA6:intron | C | 0.19 | 0.18 | 0.18 | 0.19 | $6.19 \times 10^{-06}$ | 0.0 | 0.45 | **$+1.99 \times 10^{-02}$** | 0.09 | $+4.98 \times 10^{-01}$ | 0.02 | $6.65 \times 10^{-01}$ | $1.12 \times 10^{-01}$ |
| **18:27.0** | rs9947408 | — | G | 0.10 | 0.07 | 0.10 | 0.26 | $8.51 \times 10^{-06}$ | 0.0 | 0.31 | $4.74 \times 10^{-01}$ | NA | NA | NA | NA | $7.27 \times 10^{-02}$ |
| **18:29.8** | rs10502586 | — | G | 0.11 | 0.13 | 0.08 | 0.23 | $9.13 \times 10^{-06}$ | 0.0 | 0.19 | $+9.40 \times 10^{-01}$ | 0.03 | $6.87 \times 10^{-03}$ | 0.07 | $9.24 \times 10^{-01}$ | $6.53 \times 10^{-01}$ |
| **20:3.6** | rs2670314 | GFRA4:ds | G | 0.61 | 0.58 | 0.53 | 0.14 | $7.74 \times 10^{-06}$ | 0.0 | 0.39 | $6.75 \times 10^{-01}$ | 0.54 | $+3.66 \times 10^{-01}$ | 0.71 | $+8.34 \times 10^{-01}$ | $2.70 \times 10^{-01}$ |
| **20:5.9** | rs73596147 | TRMT6:intron | T | 0.92 | 0.91 | 0.96 | 0.24 | $7.95 \times 10^{-06}$ | 0.0 | 0.89 | $7.02 \times 10^{-01}$ | 0.91 | $8.75 \times 10^{-01}$ | 0.98 | $+8.40 \times 10^{-01}$ | $3.27 \times 10^{-01}$ |
| **20:44.4** | rs6032525 | WFDC3:intron | G | 0.03 | 0.02 | 0.03 | 0.44 | $1.17 \times 10^{-06}$ | 18.7 | 0.06 | $+1.50 \times 10^{-01}$ | NA | NA | NA | NA | $1.02 \times 10^{-04}$ |
| **22:35.4** | rs5999656 | — | C | 0.85 | 0.87 | 0.87 | 0.20 | $3.34 \times 10^{-06}$ | 0.0 | 0.68 | $+4.39 \times 10^{-01}$ | 0.94 | $+9.22 \times 10^{-01}$ | 0.93 | $3.00 \times 10^{-01}$ | $8.18 \times 10^{-01}$ |
| **22:45.7** | rs6007590 | FAM118A:intron | G | 0.08 | 0.13 | 0.11 | 0.24 | $8.77 \times 10^{-06}$ | 0.0 | 0.09 | $+7.83 \times 10^{-01}$ | 0.31 | $+6.44 \times 10^{-01}$ | 0.11 | $+9.25 \times 10^{-01}$ | $2.59 \times 10^{-01}$ |

Abbreviations: NOMAS = Northern Manhattan Study; SNP = single nucleotide polymorphism; WHICAP = Washington Heights–Inwood Columbia Aging Project. N FRQ, W1 FRQ, W2 FRQ, AFR FRQ, ASN FRQ, and EUR FRQ - NOMAS, WHICAP1, WHICAP2, CHARGE African American, CHARGE Asian, and CHARGE European allele frequencies, respectively. NA - not applicable, indicating the variant was monomorphic or not well imputed (INFO < 0.4) in the replication sample. CHARGE results with p < 0.05 are indicated in bold text.

All BP are in HG19 coordinates in the form chromosome:megabase. I² = heterogeneity statistic describing the percentage of variation across NOMAS, WHICAP1, and WHICAP2 due to heterogeneity. CHARGE results are based on 17,936 Europeans, 1,943 Africans, and 405 Asians. SNPs which have not been analyzed in select populations are monomorphic.

A "+" sign indicates that the direction of effect is the same as that seen in Hispanics.

The Meta-$p$ was calculated using a z-score–based fixed-effect meta-analysis of Hispanic and CHARGE cohorts. No I² is available due to the z-score–based method.

ubiquitin-conjugating enzyme E2 C (*UBE2C*). In total, these 5 SNPs were predicted to influence the binding of 70 proteins. A list of the proteins affected by each SNP and associated RegulomeDB scores are shown in table e-1. Notably, Spi-1 proto-oncogene (*SPI1*) was the only protein influenced by rs2670314 on 20q13 and rs73306449 on 20p13. Two additional SNPs within the *WFDC3-DNTTIP1-UBE2C* LD block influenced the binding of *SPI1*. No other proteins were influenced by 4 or more SNPs. Of the 6,395 SNPs with $p < 1 \times 10^{-3}$, 110 were likely to affect the binding of 130 proteins (table e-1). The most commonly affected proteins were RNA polymerase II subunit A (*POLR2A*), which was influenced by 24 variants across 12 chromosomes, and CCCTC-binding factor (*CTCF*), which was influenced by 23 variants across 13 chromosomes (table e-2).

Table 3 indicates the results from our gene-based association tests ($p < 1.00 \times 10^{-3}$). The most strongly associated gene in our Hispanic sample, *UBE2C* ($p = 3.61 \times 10^{-4}$), demonstrated evidence of association in the African American sample ($p = 0.034$). The single SNP results of all variants within this gene region in the Hispanic and all replication samples are shown in table e-3. The majority of SNPs within 50KB of *UBE2C* with $p < 1.00 \times 10^{-5}$ in Hispanics were monomorphic in the European and Asian samples.

**DISCUSSION** In this genome-wide association study (GWAS) of 2 community-based Hispanic cohorts, 17 independent SNPs located within 7 genes showed suggestive associations with WMHV with $p < 1 \times 10^{-5}$. One SNP in *GATA6* demonstrated evidence of association in an independent sample of African Americans. A total of 110 unique variants with $p < 1 \times 10^{-3}$ were shown to influence the binding of 130 unique proteins, most commonly *CTCF* and *POLR2A*. Four of 5 of the 110 variants with $p < 1 \times$

$10^{-5}$, all on chromosomes 20q13 and 20p13, influenced the binding of *SPI1*. A gene-based analysis implicated *UBE2C* on chromosome 20q13 in our Hispanic sample, a finding which was supported in an independent sample of African Americans. This study focuses on a minority population as a means of discovery for novel genetic variants that influence white matter lesion burden.

*GATA6* encodes a member of the GATA family of transcription factors and has been shown to regulate angiogenesis and promote survival of endothelial cells.[32] Expression in early embryogenesis is crucial to heart development, and mutations in this gene have been shown to be associated with a number of congenital abnormalities including dilated cardiomyopathy[33] and atrial fibrillation.[34] In addition to *GATA6* displaying a single SNP association with WMHV in Hispanics and African Americans, 18 variants with $p < 1 \times 10^{-3}$ in the Hispanic analysis affect the binding of at least 1 of *GATA6*, *GATA1*, *GATA2*, and *GATA3*. Similar to *GATA6*, the other members of the GATA family of transcription factors are important in the vascular and lymphatic systems through regulation in the development of erythrocytes (*GATA1*) and lymphocytes (*GATA2* and *GATA3*).[35]

Other notable transcription factors affected by our top SNP associations include *CTCF* and *POLR2A*. *CTCF* is a regulator protein, which can bind histone acetyltransferase–containing complex or histone de-acetylase (HDAC)-containing complex, functioning as a transcriptional activator or repressor. Protein acetylation has been implicated in a host of neurologic diseases in recent years due primarily to HDAC-associated decrease in the level of a protein or widespread genomic transcriptional losses.[36] Thus, *CTCF* also plays a critical role in this pathway for neurodegeneration. *POLR2A* plays a role in messenger RNA (mRNA) processing through the transcription of pre-mRNA from the genomic DNA; thus, inhibition of *POLR2A* binding could

| | | | | | | CHARGE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | AFR | | ASN | | EUR | |
| Chr | Start (MB) | End (MB) | Gene | # SNP | p | # SNP | p | # SNP | p | # SNP | p |
| 20 | 44.39 | 44.50 | *UBE2C* | 285 | $3.61 \times 10^{-04}$ | 479 | $\mathbf{3.35 \times 10^{-02}}$ | 244 | $7.14 \times 10^{-01}$ | 327 | $8.34 \times 10^{-01}$ |
| 10 | 17.58 | 17.71 | *PTPLA* | 251 | $2.15 \times 10^{-04}$ | 775 | $3.75 \times 10^{-01}$ | 433 | $9.68 \times 10^{-01}$ | 535 | $8.20 \times 10^{-01}$ |
| 2 | 108.4 | 108.6 | *RGPD4* | 193 | $5.37 \times 10^{-04}$ | 316 | $8.49 \times 10^{-01}$ | 189 | $5.95 \times 10^{-01}$ | 284 | $7.52 \times 10^{-01}$ |
| 2 | 73.56 | 73.89 | *ALMS1* | 641 | $6.01 \times 10^{-04}$ | 1,251 | $9.79 \times 10^{-01}$ | 412 | $8.01 \times 10^{-01}$ | 673 | $4.85 \times 10^{-01}$ |
| 15 | 56.87 | 57.08 | *ZNF280D* | 265 | $3.70 \times 10^{-04}$ | 312 | $9.09 \times 10^{-02}$ | 258 | $7.15 \times 10^{-01}$ | 428 | $8.78 \times 10^{-01}$ |

Abbreviation: AFR, ASN, EUR - CHARGE African American, Asian, and European cohorts. SNP = single nucleotide polymorphism. All BP are in HG19 coordinates in the form chromosome:megabase. CHARGE results are based on 17,936 Europeans, 1,943 Africans and 405 Asians.
Gene start and end positions are ±50KB from the gene. # SNPs = # nonmonomorphic SNPs analyzed.

also lead to widespread RNA dysregulation, an integral element in a number of neurologic disease.[37]

Perhaps, one of the most interesting findings in this study is the implication of *UBE2C* in the gene-based test for both Hispanics and African Americans. This ubiquitin-conjugating enzyme is an important regulator in cell cycle progression, and overexpression has been associated with progression of malignant gliomas.[38] There is a substantial amount of LD in Hispanics between SNPs within 50KB of *UBE2C*, including SNPs in *WFDC3*, *DNTTIP1*, *TNNC2*, *SNX21*, *ACOT8*, and *ZSWIM3*. *WFDC3* functions as a protease inhibitor, and notably, both protein degradation though ubiquitination and protease inhibition are represented among neurodegenerative diseases.[39] Furthermore, *DNTTPI1* acts as a chromatin-binding module in complex with *HDAC1* and *HDAC2*. As previously mentioned, 3 SNPs with $p < 1 \times 10^{-5}$ within *WFDC3* and *DNTTIP1* on chromosome 20q13 affect protein binding of *SPI1*, with both genes also showing nominal significance in Hispanics with $p = 9.55 \times 10^{-3}$ and $p = 2.40 \times 10^{-3}$, respectively. This is a transcription factor which is involved in the activation of gene expression during myeloid and B-lymphocyte development and is critical for viability and function of human brain microglia, with a loss reducing the ability of microglia to clear debris by phagocytosis.[40] Outside chromosome 20q13, 1 additional SNP on chromosome 20p13 with $p < 1 \times 10^{-5}$ and 7 additional SNPs with $p < 1 \times 10^{-3}$ across 5 additional chromosomes affect binding of *SPI1*.

The CHARGE consortium identified loci for WMHV on chromosomes 17q25, 10q24 2p21, 1q22, and 2p16 in primarily individuals of European descent but also through multiethnic meta-analysis and replication, inclusive of ~2,000 African Americans, ~400 Asians, in addition to the ~18,000 Europeans (which we have now used as replication of our findings) and ~800 Hispanics (a majority subset of our NOMAS and WHICAP samples).[16] Only 2 of these loci (*PDCD11* on 10q24 and *PMF1* on 1q22) demonstrated a nominal level of significance ($p < 0.05$) in the Hispanic sample, with the same risk allele as seen in Europeans. This may be due to differences in allele frequency across populations and supports the importance of genetic discovery in minority populations. However, SNPs implicated by CHARGE within 17q25, 10q24, and 2p16 have been shown to influence binding of 30 proteins,[30] 29 of which are also influenced by SNPs with $p < 1 \times 10^{-3}$ in the Hispanic analysis, most notably *POLR2A*. This may indicate common pathways across populations.

A strength of the current study is that it focuses on discovery of novel variation influencing WMHV in a Hispanic sample. It includes analysis of 2 community-based cohorts which used the same MRI protocol and quantitative measurement of WMHV at the same medical center. However, our study also has limitations. First, with 922 Hispanic participants, we had limited power, at a genome-wide significance level, to uncover an association with small-to-moderate effect sizes. Second, as comprehensive genotype data are not available for Hispanic populations in the 1000 Genome database, imputations were based on pooled samples from populations of the Americas, Europe, Africa, and Asia. Bias in allele frequency estimates cannot be excluded for those SNPs with large allele differences across subpopulations, although high imputation quality was observed. Third, since this GWAS was performed in Hispanics, findings may not necessarily extend to other race-ethnic groups. For instance, we see varying allele frequency differences across notable SNPs such as the one found in *GATA6*. The majority of SNPs we see as associated in Hispanics within 50KB of *UBE2C* are in fact monomorphic in the European and Asian replication samples, findings which are underscored by the population substructure observed in our study sample. This is a weakness for generalizability but emphasizes the importance of genetic study in diverse samples.

This GWAS of 2 community-based Hispanic cohorts revealed several novel genetic loci for WMHV. Further replication is needed in an independent Hispanic sample as well as samples from additional race/ethnic groups. In addition, local ancestry computation may provide additional insight into the ancestral risk haplotypes. While fine mapping is needed to pinpoint causal variation for drug targets in relevant populations, this study represents progress toward elucidating the genetic underpinnings of increased WMHV potentially underlying cognitive impairment and vascular dementia.

## REFERENCES

1. de Leeuw FE, de Groot JC, Achten E, et al. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam Scan Study. J Neurol Neurosurg Psychiatry 2001;1:9–14.
2. Wright CB, Festa JR, Paik MC, et al. White matter hyper-intensities and subclinical infarction: associations with psychomotor speed and cognitive flexibility. Stroke 2008;3:800–805.
3. Prins ND, van Dijk EJ, den Heijer T, et al. Cerebral white matter lesions and the risk of dementia. Arch Neurol 2004;10:1531–1534.
4. Guttmann CR, Benson R, Warfield SK, et al. White matter abnormalities in mobility-impaired older persons. Neurology 2000;6:1277–1283.
5. Smith EE. Leukoaraiosis and stroke. Stroke 2010;41(10 suppl):S139–S143.
6. Arsava EM, Rahman R, Rosand J, et al. Severity of leukoaraiosis correlates with clinical outcome after ischemic stroke. Neurology 2009;16:1403–1410.
7. Launer LJ. Epidemiology of white matter lesions. Top Magn Reson Imaging 2004;6:365–367.
8. Brickman AM, Schupf N, Manly JJ, et al. Brain morphology in older African Americans, Caribbean Hispanics and Whites from Northern Manhattan. Arch Neurol 2008;8:1053–1061.
9. Jeerakathil T, Wolf PA, Beiser A, et al. Stroke risk profile predicts white matter hyperintensity volume: the Framingham Study. Stroke 2004;8:1857–1861.
10. Atwood LD, Wolf PA, Heard-Costa NL, et al. Genetic variation in white matter hyperintensity volume in the Framingham Study. Stroke 2004;7:1609–1613.
11. Carmelli D, DeCarli C, Swan GE, et al. Evidence for genetic variance in white matter hyperintensity volume in normal elderly male twins. Stroke 1998;6:1177–1181.
12. Kochunov P, Glahn D, Winkler A, et al. Analysis of genetic variability and whole genome linkage of whole-brain, subcortical, and ependymal hyperintense white matter volume. Stroke 2009;12:3685–3690.
13. Turner ST, Fornage M, Jack CR, et al. Genomic susceptibility Loci for brain atrophy, ventricular volume, and leukoaraiosis in hypertensive sibships. Arch Neurol 2009;7:847–857.
14. DeStefano AL, Atwood LD, Massaro JM, et al. Genome-wide scan for white matter hyperintensity: the Framingham Heart Study. Stroke 2006;1:77–81.
15. Fornage M, Debette S, Bis JC, et al. Genome-wide association studies of cerebral white matter lesion burden: the CHARGE consortium. Ann Neurol 2011;6:928–939.
16. Verhaaren BF, Debette S, Bis JC, et al. Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI. Circ Cardiovasc Genet 2015;2:398–409.
17. Warsch JRL, Rundek T, Paik MC, et al. Association of the Northern Manhattan study Global vascular risk score and Successful aging. J Am Geriatr Soc 2013;61:519–524.
18. Scarmeas N, Luchsinger JA, Stern Y, et al. Mediterranean diet and magnetic resonance imaging-assessed cerebrovascular disease. Ann Neurol 2011;69:257–268.
19. DeCarli C, Maisog J, Murphy DG, Teichberg D, Rapoport SI, Horwitz B. Method for quantification of brain, ventricular, and subarachnoid CSF volumes from MR images. J Comput Assist Tomogr 1992;2:274–284.
20. DeCarli C, Murphy DG, Teichberg D, Campbell G, Sobering GS. Local histogram correction of MRI spatially dependent image pixel intensity nonuniformity. J Magn Reson Imaging 1996;3:519–528.
21. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;3:559–575.
22. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;8:904–909.
23. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda) 2011;6:457–470.
24. R Core Team. R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013. Available at: http://www.R-project.org/. Accessed January, 2017.
25. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–1664.

26. Stanford University. Human Genome Diversity Project. 2007. Available at: http://www.hagsc.org/hgdp/. Accessed January 5, 2016.

27. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;17:2190–2191.

28. Liu JZ, McRae AF, Nyholt DR, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet 2010;1:139–145.

29. Mishra A, Macgregor S. VEGAS2: Software for more flexible gene-based testing. Twin Res Hum Genet 2015;1:86–91.

30. de Leeuw CA, Mooij JM, Heskes T, et al. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11:e1004219.

31. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;9:1790–1797.

32. Froese N, Kattih B, Breitbart A, et al. GATA6 promotes angiogenic function and survival in endothelial cells by suppression of autocrine transforming growth factor beta/activin receptor-like kinase 5 signaling. J Biol Chem 2011;7:5680–5690.

33. Xu L, Zhao L, Yuan F, et al. GATA6 loss-of-function mutations contribute to familial dilated cardiomyopathy. Int J Mol Med 2014;5:1315–1322.

34. Li J, Liu WD, Yang ZL, Yang YQ. Novel GATA6 loss-of-function mutation responsible for familial atrial fibrillation. Int J Mol Med 2012;4:783–790.

35. Gao J, Chen YH, and Peterson LC. GATA family transcriptional factors: emerging suspects in hematologic disorders. Exp Hematol Oncol 2015;4:28.

36. Didonna A, Opal P. The promise and perils of HDAC inhibitors in neurodegeneration. Ann Clin Transl Neurol 2015;1:79–101.

37. Walsh MJ, Cooper-Knock J, Dodd JE, et al. Invited review: decoding the pathophysiological mechanisms that underlie RNA dysregulation in neurodegenerative disorders: a review of the current state of the art. Neuropathol Appl Neurobiol 2015;2:109–134.

38. Ma R, Kang X, Zhang G, et al. High expression of UBE2C is associated with the aggressive progression and poor outcome of malignant glioma. Oncol Lett 2016;3: 2300–2304.

39. Simonsen AH, Hagnelius NO, Waldemar G, Nilsson TK, and McGuire J. Protein markers for the differential diagnosis of vascular dementia and Alzheimer's disease. Int J Proteomics 2012;2012:824024.

40. Smith AM, Gibbons HM, Oldfield RL, et al. The transcription factor PU.1 is critical for viability and function of human brain microglia. Glia 2013;6:929–942.