Title

NOTCH3 p.Arg1231Cys is Markedly Enriched in South Asians and Associated with Stroke

Authors

Juan L. Rodriguez-Flores[1], Shareef Khalid[2] , Neelroop Parikshak[1], Asif Rasheed[3], Bin Ye[1], Manav Kapoor[1], Joshua Backman[1], Farshid Sepehrband[1], Silvio Alessandro DiGioia[4], Sahar Gelfman[1], Tanima De[1], Nilanjana Banerjee[1], Deepika Sharma[1], Hector Martinez[4], Sofia Castaneda[5], David D'Ambrosio[4], Xingmin A. Zhang[1], Pengcheng Xun[4], Ellen Tsai[6], I-Chun Tsai[4], Regeneron Genetics Center[1], Muhammad Jahanzaib[3], Maleeha Zaman Khan[3], Muhammad Rehan Mian[3], Muhammad Bilal Liaqat[3], Khalid Mahmood[7], Tanvir Us Salam[8], Muhammad Hussain[8], Javed Iqbal[10], Faizan Aslam[11], Michael N. Cantor[1], Gannie Tzoneva[1], John Overton[1], Jonathan Marchini[1], Jeff Reid[1], Aris Baras[1], Niek Verweij[1], Luca A. Lotta[1], Giovanni Coppola[1], Katia Karalis[1], Aris Economides[1], Sergio Fazio[4], Wolfgang Liedtke[4], John Danesh[12], Ayeesha Kamal[9], Philippe Frossard[3], Thomas Coleman[1], Alan R. Shuldiner[1], Danish Saleheen[2,3]

Affiliations

[1.] Regeneron Genetics Center, Regeneron Pharmaceuticals Inc, Tarrytown, NY, USA
[2.] Columbia University, New York, NY, USA
[3.] Center for Non-Communicable Diseases, Karachi, Pakistan
[4.] Health Economics and Outcome Research, Regeneron Pharmaceuticals Inc, Tarrytown, NY, USA
[5.] Rye Country Day School, Rye, NY, USA
[6.] University of California at Los Angeles, Los Angeles, CA, USA
[7.] Dow University of Health Sciences and Civil Hospital, Karachi, Pakistan
[8.] Lahore General Hospital, Lahore, Pakistan
[9.] Section of Neurology, Department of Medicine, Aga Khan University, Karachi, Pakistan
[10.] Department of Neurology, Allied Hospital, Faisalabad, Pakistan.
[11.] Department of Neurology, Aziz Fatima Hospital, Faisalabad, Pakistan.
[12.] Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
.

Correspondence

Danish Saleheen, danish.saleheen@cncdpk.com
Alan Shuldiner, alan.shuldiner@regeneron.com

Abstract

The genetic factors of stroke in South Asians are largely unexplored. Exome-wide sequencing and association analysis (ExWAS) in 75 K Pakistanis identified *NOTCH3* p.Arg1231Cys, enriched in South Asians (alternate allele frequency = 0.58% compared to 0.019% in Western Europeans), and associated with subcortical hemorrhagic stroke [odds ratio (OR) = 3.39, 95% confidence interval (CI) = [2.26, 5.10], p value = 3.87×10^{-9}), and all strokes (OR [CI] = 2.30 [3.01, 1.77], p value = 7.79×10^{-10}). *NOTCH3* p.Arg231Cys was strongly associated with white matter hyperintensity on MRI in United Kingdom Biobank (UKB) participants (effect [95%CI] in SD units = 1.1 [0.61, 1.5], p value = 3.0×10^{-6}). The variant is attributable for approximately 5.5% of hemorrhagic strokes and 1% of all strokes in South Asians. These findings highlight the value of diversity in genetic studies and have major implications for genomic medicine and therapeutic development in South Asian populations.

Introduction

Pakistan, a country in South Asia, comprises over 231 million inhabitants. It is the fifth most populous country in the world with diverse ancestral backgrounds from South and Central Asia, West Asia, and Africa. Pakistan, and in general South Asia, represents an understudied region in large-scale genetic studies [1], thus providing an opportunity for novel discoveries of the genetic basis of diseases, especially those with an elevated incidence relative to European ancestry populations. Recent surveys of stroke rates in Pakistan show enrichment of the disease relative to US and Western European populations, with a crude prevalence of 722 to 1,200 per 100 thousand (K) and crude incidence of 94.6 per 100 K. Hemorrhagic stroke incidence (52.4 per 100 K) is elevated relative to ischemic stroke (42.2 per 100 K) [2], in contrast to most European ancestry populations, where ischemic stroke has a markedly higher incidence than hemorrhagic stroke. In addition, reports suggest high variability of stroke incidence across Pakistani communities [3], suggesting the possibility of isolated endogamous founder populations with higher incidence of hemorrhagic stroke attributable to enrichment of genetic risk variants [4-6].

At least 9 rare monogenic disorders are characterized by increased stroke risk, such as cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) due to mutations in *NOTCH3* [7]. CADASIL is distinct from other hereditary stroke diseases because it is characterized by vascular smooth muscle cell (VSMC) degeneration in small arteries and accumulation of protein aggregates known as granular osmophilic material (GOM) that contain aggregates of misfolded *NOTCH3* extracellular domain (ECD). The more common forms of stroke are likely polygenic with substantial contributions from behavioral and environmental factors as well as age. Major risk factors include high systolic blood pressure, high body mass index, hyperlipidemia, elevated glucose, and smoking [8]. Recent genome-wide association studies (GWAS) identified single nucleotide variants (SNVs) in more than 28 loci associated with stroke [9]. These variants are common non-coding variants with small effect sizes and were identified in predominantly European ancestry populations.

The aim of this study was to identify protein coding deleterious missense or loss-of-function (LoF) variants associated with stroke phenotypes in the Pakistani population. We performed exome sequencing in a 31 K discovery cohort consisting of 5,135 stroke cases and 26,602 controls of Pakistani origin. Exome-wide association analysis (ExWAS) associated *NOTCH3* p.Arg1231Cys with an approximately three-fold increased risk of hemorrhagic stroke in heterozygotes. Follow-up meta-analysis of 61 K (including 160 cases and 30,239 controls in a replication cohort) also identified a significant association with all strokes (combined ischemic and hemorrhagic) for *NOTCH3* p.Arg1231Cys. This variant was present in approximately 1% of Pakistani and was markedly enriched with respect to Europeans in multiple South Asian (SAS) and West Asian (WAS) (also referred to as Greater Middle Eastern [10]) populations ranging from Turkey to India. The variant was estimated to explain up to 1% of strokes and 5% of hemorrhagic strokes in South Asia, a region having a population of > 2 billion people, thus having significant medical implications in these very large yet understudied populations and their global diaspora.

Results

ExWAS in the Pakistan Genomics Resource (PGR) discovery cohort identifies a markedly enriched missense variant in *NOTCH3* associated with stroke

Characteristics of the 5,135 stroke cases and 26,602 controls in the discovery cohort are summarized in **Table 1**. Compared to controls, stroke cases were modestly older and had a higher prevalence of known risk factors for vascular disease including hypertension, diabetes, myocardial infarction, and tobacco use (all $p < 0.0001$). As expected, (**Supplementary Figure 1, Supplementary Table 1**); in this cohort ascertained for stroke, most cases were ischemic strokes and most hemorrhagic strokes were subcortical (**Supplementary Figure 2**).

Case:control ExWAS for all stroke cases and the 4 stroke subtypes identified a genome-wide significant (p value $< 5.0\text{x}10^{-8}$) association for missense variant p.Arg1231Cys (rs201680145) in *NOTCH3* with subcortical hemorrhagic stroke (OR [95% CI] = 3.39 [2.26, 5.1], p value = $3.87\text{x}10^{-9}$; AAF = 0.58%) (**Figure 1B, Supplementary Table 1 and Supplementary Figure 3, Supplementary Figure 4**). p.Arg1231Cys also showed evidence for association with all strokes combined (OR [95% CI] = 2.18 [1.65, 2.89], p value = $4.44\text{x}10^{-8}$) and other sub-categories of stroke (**Supplementary Table 2,)**. No other variants in the locus were associated with stroke (**Figure 1C**). We did not observe an association between p.Arg1231Cys and history of hypertension, elevated systolic or diastolic blood pressure, or smoking, known major risk factors for stroke (**Supplementary Table 3, Supplementary Table 4**), and inclusion of these risk factors in regression analysis did not appreciably alter the effect of p.Arg1231Cys on stroke risk (**Supplementary Table 5**).

NOTCH3 encodes Notch Receptor 3, a transmembrane signaling protein and part of an evolutionarily conserved family that plays a pleiotropic role in cell-cell interaction and neural development [11]. The extra-cellular domain (ECD) of *NOTCH3* consists of 34 epidermal growth factor-like repeat (EGFR) domains [12], each containing 6 cysteine (Cys) residues that form three disulfide bonds (**Figure 2**). Adding or removing Cys residues in the first 6 EGFRs cause classical CADASIL, a highly penetrant rare autosomal dominant disease clinically characterized by migraine with aura, early-onset recurrent strokes, dementia, and behavioral changes [7]. The Cys-altering variant associated with stroke in this study, p.Arg1231Cys, occurs in the 31st EGFR [13] and is predicted deleterious (**Supplementary Table 6**). Stroke cases heterozygous for p.Arg1231Cys and stroke cases without the variant were similar with respect to age, age of stroke onset, type of stroke, and stroke risk factors, suggesting a milder form of CADASIL not obviously clinically distinguishable from common forms of stroke in this population, although a detailed history for migraine or other manifestations of CADASIL were not available (**Table 1**).

Replication and Meta Analysis in PGR

An additional 160 Pakistani self-reported stroke cases and 30,239 controls were recruited and whole-exome sequenced by the Broad Institute for meta-analysis as part of the follow-up cohort. Replication of the association was observed in this independent cohort (OR [95% CI] = 3.49

[1.56, 7.83], p value = 5.00 x 10^{-3}. Meta-analysis for all strokes in the combined 61 K cohort achieved a genome-wide significant p value (OR [95% CI]) = 2.30 [1.76, 2.99], p value = 7.08 x 10^{-10}) (**Figure 3**).

Recall by Genotype

A total of 12 homozygotes from 9 nuclear families were identified in the PGR, including 9 discovery cohort probands and 3 follow-up cohort relatives identified through a callback of proband relatives. Baseline characteristics of the 12 homozygotes are shown in **Supplementary Table 7**. Three of twelve (25%) homozygotes had a history of stroke; of note, all with stroke were >65 years of age while all without a history of stroke were <55 years of age. While there was no association between p.Arg1231Cys and hypertension in heterozygotes, 8 of 12 (66%) homozygotes had a history of hypertension.

Allele Frequency and Population Attributable Risk

The allele frequency of p.Arg1231Cys was 1.1% across the PGR 75 K, equivalent to a population prevalence of 1 in 46. After removing cases recruited for cardiovascular diseases (individuals enrolled at time of acute stroke, MI, and heart failure), the allele frequency of the variant was 0.51% equivalent to a population prevalence of 1 in 98. This frequency was orders of magnitude higher relative to exomes of European ancestry from UK Biobank (AAF = 0.019%), corresponding to a population prevalence of 1 in 2,614). The variant was enriched (AAF > 0.1%) in other South Asian and West Asian populations [14] both within and outside of Pakistan (**Supplementary Data, Supplementary Table 8, Supplementary Table 9, Supplementary Figure 5**).

We estimate that 5.5% [bootstrap 95% CI based on 10,000 resamples: 4.2% to 6.7%] of hemorrhagic strokes and 1.0% [bootstrap 95% CI based on 10,000 resamples: 0.6% to 1.4%] of all strokes in the Pakistani population are attributable to p.Arg1231Cys. Thus, this variant is a common cause of strokes in SAS and GME populations, a finding that has implications for global health as well as medical care in these populations.

Replication of known *LSP1* association with stroke

A common intronic variant (rs661348, AAF = 0.27) in lymphocyte specific protein *LSP1* previously associated with hypertension [15] exhibited a suggestive association (p value < $1x10^{-7}$) with intracerebral hemorrhage in PGR, a weak association with hypertension in PGR (p value = $2.61x10^{-2}$), and a strong association with hypertension in UKB (OR [95% CI] = 1.1 [1.0, 1.1], p value = $8.05x10^{-25}$) [16] (**Supplementary Tables 11, 12, 13, and 14**). Two common non-coding variants in the *LSP1* locus previously associated with stroke (rs569550 and rs1973765) [9], both were in linkage disequilibrium with rs661348 (r^2 > 0.4 in 10 K unrelated PGR participants). A test of association for rs661348 conditional these variants reduced the strength of the association for rs661348 to nominal($8.83x10^{-4}$) (**Supplemental Table 10**), suggesting that rs661348 represents the same known stroke risk locus.

p.Arg1231Cys *NOTCH3* is associated with stroke and CADASIL-like phenotypes in UK Biobank

To investigate stroke and CADASIL-related phenotypes in an independent cohort, UK Biobank data was reviewed for associations with *NOTCH3* p.Arg1231Cys. A total of 255 heterozygotes for *NOTCH3* p.Arg1231Cys were observed in 450 K exome-sequenced individuals from the predominantly European UKB cohort, with a markedly lower allele frequency in UKB Europeans (AAF = 0.019%) (**Supplementary Table 15**). Phenome-wide association (PheWAS) of 10,168 phenotypes revealed nominally significant association of p.Arg1231Cys with ischemic stroke (OR [95%CI] = 4.0 [1.9, 8.6]), p value = 4.1×10^{-4}), all strokes combined (OR [95% CI] = 1.9 [1.1, 3.5], p value = 0.031), hypertension (ICD 10 code I10) (OR [95% CI] = 1.5 [1.1, 2.2], p value = 0.019), and recurrent major depression (OR [95% CI] = 3.2 [1.5, 6.8], p value = 0.0031) (**Supplementary Table 16**). No association was observed for hemorrhagic stroke, migraine, dementia, mood changes, Alzheimer's disease, or urinary incontinence. The lack of association (p value = 0.066) with hemorrhagic stroke in UKB Europeans was likely due to low statistical power, given the lower variant allele frequency and lower hemorrhagic stroke prevalence in UKB compared to PGR. Nonetheless, the odds ratio (OR [95% CI] = 5.8 [0.88, 39.1]) was high.

In addition, *NOTCH3* p.Arg1231Cys was strongly associated with a cluster of brain MRI quantitative phenotypes, e.g., total volume of white matter hyperintensities (WMH) from T1 and T2 FLAIR images (effect [95% CI] in SD units = 1.1 [0.61, 1.5], p value = 3.0×10^{-6}) with carriers having 7.4 cm^3 more WMH volume than controls (**Supplementary Figure 6** and **Supplementary Table 17**). The most prominent alterations in WMH in p.Arg1231Cys carriers were observed in the centrum semiovale and periventricular white matter (**Supplementary Figure 7)**. Taken together, these results demonstrate *NOTCH3* p.Arg1231Cys carriers have increased risk of established markers of small vessel disease and clinical phenotypes observed in CADASIL [7].

Pathogenic burden of all Cys-altering variants within *NOTCH3* EGFR domains specifically associated with CADASIL phenotypes in UK Biobank

Burden test analysis allows for increased statistical power to detect association by combining signal across multiple rare variants. The combined population burden of over 90 Cys-altering variants in 34 *NOTCH3* EGFR-like domains (**Supplementary Table 18, Supplementary Table 19**) was strongly associated with ischemic stroke (OR [95% CI] = 2.86 [2.14, 3.82], p value = 6.29×10^{-10}; AAF = 0.01%) (**Table 2**). Despite low statistical power, EGFR Cys-altering variants were also nominally associated with hemorrhagic stroke [OR [95% CI] = 3.61 (1.39, 9.34), p value = 8.31×10^{-3}; AAF 0.025%)] (**Table 2**). In brain MRI data of 35,344 UKB individuals, Cys-altering variants in *NOTCH3* EGFRs were strongly associated with WMH volume ($P = 3.7 \times 10^{-13}$; with carriers having 5.4 cm^3 greater WMH volume than controls). These WMH differences were strongest in the centrum semiovale and periventricular white matter (**Supplementary Figure 7**). Additionally, we found strong WMH signal in the external capsule, which is known to be involved in CADASIL.

Prior studies suggest that pathogenic *NOTCH3* Cys-altering variants in classical CADASIL cluster in the first 6 EGFRs [13], and that patients with such variants have excess white matter

pathology [17], more granular osmophilic material (GOM) aggregates in blood vessels [17], and worse prognosis [13]. Indeed, we observed a much larger effects size for Cys-altering variants in EGFRs 1-6 (OR [95% CI] = 29.5 [10.4, 83.8], p value = 1.37×10^{-7}; AAF = 0.002%] compared to Cys-altering variants in EGFRs 7-34 (OR [95% CI] = 2.55 [1.87, 3.46], p value = 1.59×10^{-7}; AAF = 0.098%) (**Table 2, Supplementary Table 20**). Since EGFRs 10 and 11 are the primary EGFRs involved in ligand binding, these results suggest that deficient signaling may not be the primary driver CADASIL pathology.

Not all missense variants or Cys-altering variants in *NOTCH3* were associated with stroke. In contrast to Cys-altering variants within EGFR-like domains, Cys-altering variants outside of EGFRs were not associated with stroke (OR [95% CI] = 0.97 [0.46, 2.03], p value = 9.3×10^{-1}; AAF =0.039%). The most commonly altered (added or removed) amino acid in *NOTCH3* was serine (Ser), and Ser-altering burden test did not show any evidence of association with stroke (OR [95% CI] = 0.98 [(0.8, 1.2], p value = 0.084; AAF = 0.54%] (**Table 2**, **Supplementary Table 20**). Interestingly, a burden test limited only to predicted loss of function (pLoF) variants (frameshift, splice variant, stop gain) did not show any evidence for association with stroke (OR [95% CI] = 1.38 [0.50, 3.85], p value = 0.54; AAF = 0.019%) (**Table 2, Supplementary Table 21**). From -these results we conclude that only Cys-altering variants within *NOTCH3* EGFRs are associated with stroke, in contrast to other missense mutations or simple loss of function (haploinsufficiency) which are not, suggesting pathological mechanisms involving dysfunctional disulfide bridge formation and subsequent protein misfolding and aggregation, as is commonly observed in CADASIL [17].

Discussion

This report describes the largest ExWAS of stroke conducted thus far in a South Asian population and highlighting a Cys-altering missense variant in *NOTCH3* in the 31st EGFR binding domain associated with stroke at a genome-wide level of statistical significance. This is the first study to report a genome-wide-significant association between *NOTCH3* and stroke, a discovery enabled because *NOTCH3* p.Arg1231Cys is markedly enriched in Pakistanis compared to Western European and non-Eurasian populations.Harbored in ~1 percent of Pakistani, p.Arg1231Cys is associated with a ~3-fold increased risk of hemorrhagic stroke. While some regional variability in the allele frequency is observed, p.Arg1231Cys is enriched in populations ranging from Turkey in West Asia to India in South Asia, suggesting a substantial contribution to stroke risk in millions of individuals across South Asia and West Asia as well as their global diaspora.

NOTCH3 was not previously associated with stroke in large GWAS predominantly consisting of European-derived participants [9]. In contrast to prior studies, both the discovery and replication cohorts in this study was South Asian, hence avoiding the bias encountered in studies with a European discovery cohort. Given the much lower allele frequency of p.Arg1231Cys in European populations, we observed a nominal association between p.Arg1231Cys and stroke in the UK Biobank study, showing a similar effect size as in South Asians as well as phenotypes related to CADASIL such as hypertension, and depression. While brain images were not available for the Pakistani cohort, a strong association was observed between p.Arg1231Cys and quantitative brain MRI phenotypes in UKB data, such as white matter hyperintensity.

Cys-altering mutations in proximal EGFRs of *NOTCH3* are known to cause autosomal dominant CADASIL, a rare highly penetrant distinct syndrome that includes early onset recurrent subcortical strokes. In contrast to classical CADASIL, p.Arg1231Cys is in the 31st of 34 EGFR, appears to have more moderate penetrance, and is not obviously clinically distinguishable from more common multi-factorial forms of stroke in South Asians. The p.Arg1231Cys *NOTCH3* variant is currently classified in ClinVar and recent reviews [13] as a variant of uncertain significance [18]; however, based on our current findings, there is strong genetic, computational, and imaging evidence of pathogenicity for this variant despite reduced penetrance and severity compared to "classic" Cys-altering CADASIL pathogenic variants in EGFRs 1 to 6.

CADASIL is characterized by both ischemic and hemorrhagic strokes, although the factors that contribute to the manifestation of one versus the other stroke type awaits further clarification [7]. Hemorrhagic stroke appears to represent a larger proportion of strokes in South Asia than in Europe [2]. In this study, the p.Arg1231Cys association signal was stronger in PGR for hemorrhagic strokes than for ischemic strokes, despite nearly two-fold larger ischemic stroke case counts. In contrast, the UKB association signal appeared stronger for ischemic stroke, possibly due to low hemorrhagic stroke case count and thus statistical power in this cohort. Statistical power issues aside, differences in manifestation of p.Arg1231Cys in South Asians compared to Europeans may be attributable to differences in risk factors such as age, blood pressure, diabetes, air pollution, smoking, medications such as anti-platelet and anti-coagulants used to manage atherosclerotic disease, genetic background, or study-specific differences in criteria to categorize stroke sub-types. Further research is needed to better ascertain the mechanism behind cerebral arterial wall pathobiology and clinical presentation of ischemic versus hemorrhagic stroke in p.Arg1231Cys carriers.

Currently there are no known effective preventive or therapeutic interventions for CADASIL or less penetrant forms of *NOTCH3* related stroke. However, our analyses provide clues toward their development. First, in contrast to our analyses of EGFR Cys-altering missense variants in *NOTCH3* that were highly associated with stroke, predicted loss of function variants that would be expected to not produce a functional protein were not associated with stroke. In line with this finding, mouse models have successfully produced most of the known CADASIL pathophysiology in Cys-altering variant knock-in models [19], while knock-out models have been less successful [20]. These observations suggest that simple loss of function (haploinsufficiency) is not a major causative factor in CADASIL. Taken together with findings that deposition of globular osmophilic material (GOM) containing misfolded *NOTCH3* in brain is pathognomonic of CADASIL [21], our genetic findings suggest that therapeutic interventions that decrease expression of mutant protein, (such as siRNA, antisense oligonucleotides, and CRISPR), or modalities to accelerate removal of GOM may prove beneficial for prevention and/or treatment [22].

Our analyses also suggest that p.Arg1231Cys is modestly associated with hypertension, although p.Arg1231Cys association with stroke risk appears independent of hypertension or other stroke risk factors such as smoking, age and sex. Animal models of CADASIL show decreased vascular tone and contractility, most likely driven by loss of physiologic function and subsequent degeneration of vascular smooth muscle cells (VSMCs) [23]. These observations suggest that

while management of hypertension and smoking cessation are effective modalities for primary and secondary prevention of stroke, those with *NOTCH3* mutation related strokes will need additional therapeutic interventions, as existing hypertensive medications cannot restore VSMC function.

A limitation of this study is the lack of brain imaging analysis for the Pakistani carriers, such that specific brain regions affected by the ischemic and hemorrhagic strokes could be ascertained and compared. In addition, we lacked more detailed clinical data such as presence of migraines and longitudinal data of disease course including stroke recurrence, dementia, and depression. Further characterization of p.Arg1231Cys carriers will be necessary to obtain better estimates of penetrance as well as to identify distinguishing clinical or biomarker characteristics that may have utility in early diagnosis, prevention and treatment, and for recommendations for cascade screening in family members. Migraine symptoms typically precede stroke by 10+ years in CADASIL patients [7], thus the combination of migraine with aura, depression and family history of stroke could be sufficient evidence to prescribe *NOTCH3* genetic testing.

In conclusion, we identified a highly enriched Cys-altering variant in *NOTCH3* in South Asians that expands the phenotypic spectrum of CADASIL from rare and highly penetrant to common and moderately penetrant. Based on our estimates, this single variant may be responsible for ~1.0% of all strokes combined and ~5.5% of hemorrhagic strokes in South Asians. Among 1.9 billion South Asians there could be over 26 million carriers for the variant. Thus, this work has important implications for genetic screening and early identification of at-risk individuals, and the future opportunity for rationally targeted therapeutic interventions.

Online Methods

Details of methods are available in the Supplementary text. Briefly, 75 K individuals were recruited and consented in Pakistan for exome sequencing, including a stroke case:control discovery cohort of 31 K (5,135 cases and 26,602 controls) sequenced by the Regeneron Genetics Center and a follow-up cohort of 44 K, including 30 K with self-report stroke case:control status used for meta-analysis (160 cases and 30,239 controls). ExWAS was conducted for stroke and 4 overlapping stroke subtypes (intracerebral hemorrhage, subcortical intracerebral hemorrhage, ischemic stroke, and partial anterior circulating infarct) in the discovery cohort and combined in a meta-analysis of stroke with the replication cohort using both single-variant and gene burden test models [24]. Population attributable fraction of stroke for associated variants was calculated as described previously [25]. Consented callbacks were conducted in families of homozygotes for associated variants. For comparison and validation, analyses were conducted in UK Biobank data as described previously [16, 26-28], including association analysis of p.Arg1231Cys *NOTCH3* with stroke phenotypes in 370 K participants, and brain imaging phenotypes in 33 K participants (**Supplementary Methods**; **Figure 1A**).

Figure Legends

Figure 1. ExWAS Identifies *NOTCH3* p.Arg1231Cys Associated with Subcortical Hemorrhagic Stroke in Pakistan Genome Resource 31 K Discovery Cohort. A. Flow chart of the study described in this report. The discovery cohort consisted of a 31 K stroke case-control cohort from the Pakistan Genome Resource (PGR). A second PGR follow-up cohort of 44 K included 30 K participants with self-reported stroke case:control status for replication. UK Biobank data from 450 K sequenced participants was used for further analysis in a predominantly European ancestry population, 370 K of whom had stroke case:control status known, and 33 K of whom had brain MRI data (see Methods and Supplementary Methods). **B.** Manhattan plot of subcortical hemorrhagic stroke ExWAS in 31 K PGR discovery cohort participants with $-\log_{10}$ p-values (y-axis) across chromosomes and variants (x-axis). A single variant on chromosome 19 predicting a missense variant p.Arg1231Cys in *NOTCH3* exceeded the genome-wide significance threshold of $5x10^{-8}$ (red line). **C.** *NOTCH3* locus zoom plot of subcortical stroke ExWAS. The $-\log_{10}$ p-values for variants tested are shown on the y-axis. The p.Arg1231Cys variant is labeled as a diamond. Other variants (circles) are colored based on linkage disequilibrium with the reference variant in 1000 Genomes [29]. Gene exon (thick line) and intron (thin line) model shown below the graph.

Figure 2. NOTCH3 EGFR Domain Disruption by p.Arg1231Cys. Shown is *NOTCH3* p.Arg1231 in context of human *NOTCH3* protein domains and cross-species alignment of NOTCH3 amino acid sequences. **A.** Human NOTCH3 Protein Domains. Shown is the position of the associated variant in context of transcript exons (top, alternating blue and purple with numbering) and protein domains (bottom, color coded). NOTCH3 can be divided into four major regions, from left-to-right the signal peptide (light blue), the extra-cellular domain (ECD, brown), the transmembrane domain (orange), and the intra-cellular domain (ICD, blue). The majority of the ECD is composed of 34 Epidermal Growth Factor-like Repeat (EGFR) domains (in purple with white numbers). Domains involved in signaling are highlighted, including EGFRs 10 to 11 involved in ligand binding (light purple, numbered), three cleavage domains (S1 in yellow, S2 in yellow with diagonal black stripes, S3 in yellow with black horizontal stripes), and three Lin12/NOTCH repeats (light green, numbered). The ICD contains the Recombination signal-binding protein for Ig of κ region (RAM) domain for transcription factor interaction (green and white checkers), the Nuclear Localization Sequences (NLS, orange with black stripes), five Ankyrin repeats involved in signal transduction (green with white numbers), and the Proline, glutamic acid, serine, and threonine-rich (PEST) domain essential for degradation (green with white stripes). The p.Arg1231Cys variant (red line top to bottom) removes a disulfide-bridge-forming cysteine in the 31st EGFR of the ECD, coded by the 22nd exon. **B.** Cross-species Alignment of NOTCH3 (EGFR) Domain # 31 Amino Acid Sequences. Shown is an amino acid alignment of 31st EGFR of NOTCH3 (human sequence amino acids 1205 to 1244), including (top-to-bottom) human reference, human p.Arg1231Cys mutant, chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), western clawed frog (*Xenopus tropicalis*), and green sea turtle (*Chelonia mvdas*), indicating conservation of the arginine (R) at position 1231 in mammals. Highly-conserved cysteine (C) residues (normally 6 per EGFR) are highlighted in yellow.

Figure 3. Forest Plot Showing Replication of *NOTCH3* p.Arg1231Cys Association with Stroke Across 61 K Pakistan Genome Resource Meta-Analysis. Shown is the cohort name, trait, odds ratio with 95% confidence interval, p value, alternate allele frequency, case count, and control count for five stroke phenotypes in the PGR 31 K discovery cohort (top), and Inverse Variance Weighted (IVW) Meta-Analysis of stroke in 61 K PGR Cohort, including 31 K PGR discovery cohort and 30 K PGR replication cohort subset of the 44 K PGR follow-up cohort (bottom).

Tables

Table 1. Baseline Characteristics of PGR Stroke Case-control Discovery Cohort[1]

	Case (n = 5,135)		Control (n = 26,602)			Case p.Arg1231Cys Carrier (n=103)[3]		Case Non-Carrier (n=4,998)		
	mean / n	**SD / %**	**mean / n**	**SD / %**	**P value**	**mean / n**	**SD / %**	**mean / n**	**SD / %**	**P value**
Female, n (%)	2,277	44.3	9,330	35.1	< 0.01	50	48.5	2,211	44.2	0.44
Age at enrolment, years	58.9	13.1	52.8	11.4	< 0.01	58.7	12.1	58.9	13.2	0.87
BMI, kg/m^2	25	3.9	27.5	4.4	< 0.01	24.6	3.4	24.9	3.9	0.35
Cholesterol mg/dL	175.3	55.5	172.1	48.2	< 0.01	174.8	49.1	175.3	55.7	0.94
LDL-C mg/dL	111.1	45.8	101.2	38.2	< 0.01	110.8	41.1	111.1	45.9	0.94
HDL-C mg/dL	37.8	12.7	35	10.8	< 0.01	38.8	12.1	37.8	12.7	0.43
Triglyceride mg/dL	140.3	81.8	185.7	120.1	< 0.01	126.3	62.7	140.5	82.3	0.04
Glucose, mg/dL	148.7	75.5	143.4	84.9	< 0.01	145.2	74.7	148.6	75.3	0.68
HbA1c %	6.9	1.8	6.6	2	< 0.01	7.3	1.9	6.9	1.9	0.15
Creatinine mg/dl	1.2	0.8	0.9	0.5	< 0.01	1.1	0.7	1.2	0.8	0.39
Tobacco or other stimulant user[2]	1,850	36	9,242	34.7	< 0.01	32	31.1	0	0	0.34
Comorbidities										
Hypertension , n (%)	2,953	57.5	9,385	35.3	< 0.01	57	55.3	2,879	57.6	1
Diabetes, n (%)	1,195	23.3	5,766	21.7	< 0.01	29	28.2	1,157	23.1	0.29
Myocardial infarction, n (%)	371	6.2	622	2.3	< 0.01	5	4.9	215	4.3	0.98
Family history of										
Stroke, n (%)	324	6.3	0	0	< 0.01	14	13.6	306	6.1	< 0.01
Hypertension, n (%)	573	11.2	4,299	16.2	< 0.01	15	14.6	556	11.1	0.35
Diabetes, n (%)	349	6.8	5,199	19.5	< 0.01	11	10.7	334	6.7	0.16
Sudden death, n (%)	150	2.9	569	2.1	< 0.01	8	7.8	142	2.8	< 0.01

1. Shown are mean or total n and standard deviation or percentage for cases versus controls on the left and case p.Arg1231Cys *NOTCH3* carriers versus case non-carriers on the right. Comparison was conducted and p values are shown from chi-square test for categorical variables (Fisher's exact test if cell size was <5) and from T-test for continuous variables.
2. Tobacco or other stimulants include cigarettes, paan (chewed betel leaf and areca nut), naswar (snuff), gutka (chewing tobacco), huqqa (water pipe), chillum (hasish pipe).
3. Genotypes for the p.Arg1231Cys variant were not available for n=34.

Table 2. UKB Ischemic Stroke Association Across *NOTCH3* Variant Classes and Domains[1]

***NOTCH3* Variant Class**	**Effect [OR (95%CI)]**	**P value**	**AAF (%)**	**Cases (RR|RA|AA)**	**Cases (n)**	**Controls (RR|RA|AA)**	**Controls (n)**
p.Arg1231Cys	4.00 (1.9, 8.6)	$4.1x10^{-4}$	0.02	6873|10|0	6,883	353143|134|0	354,277
EGFR 1-34 Cys-altering	2.86 (2.14,3.82)	$6.3x10^{-10}$	0.01	9094|49|0	9,143	370693|709|1	371,403
EGFR 1-6 Cys-altering	29.51 (10.39,83.82)	$1.4x10^{-7}$	0.002	9137|6|0	9,143	371394|9|0	371,403
EGFR 7-34 Cys-altering	2.55 (1.87,3.46)	$1.6x10^{-7}$	0.098	9100|43|0	9,143	370702|700|1	371,403
Non-EGFR Cys-altering	0.97 (0.46,2.03)	$9.3x10^{-1}$	0.039	9136|7|0	9,143	371111|292|0	371,403
EGFR 1-34 Ser-altering	0.98 (0.8,1.2)	$8.4x10^{-1}$	0.54	9046|97|0	9,143	367355|4042|6	371,403
EGFR 1-6 Ser-altering	0.76 (0.23,2.56)	$6.6x10^{-1}$	0.018	9141|2|0	9,143	371271|132|0	371,403
EGFR 7-34 Ser-altering	0.99 (0.8,1.21)	$8.9x10^{-1}$	0.53	9048|95|0	9,143	367486|3911|6	371,403
Non-EGFR Ser-altering	0.84 (0.52,1.34)	$4.5x10^{-1}$	0.10	9128|15|0	9,143	370656|744|3	371,403
LoF variants	1.38 (0.50,3.85)	$5.4x10^{-1}$	0.019	9138|5|0	9,143	371261|142|0	371,403
LoF + any missense	1.09 (1.01,1.19	$3.3x10^{-2}$	3.30	8490|652|1	9,143	346648|24710|45	371,403

1. Burden test with age, age^2, gender, 10 PCs as covariates was conducted using REGENIE to compare the association signal and effect size across variant classes and domains for ischemic stroke. *NOTCH3* variant classes include the single variant association (p.Arg1231Cys) for reference at the top and burden tests limited to: Cys-altering variants in EGFRs 1 to 34; Cys-altering variants in EGFRs 1 to 6; Cys-altering variants in EGFRs 7 to 34; Cys-altering variants outside of EGFRs; Ser-altering variants in EGFRs 1 to 6; Ser-altering variants in EGFRs 7 to 34; Ser-altering variants outside of EGFRs; Loss of function (LoF) variants (defined as stop-gain, stop-loss, frameshift and splice-site variants with AAF <1%); and LoF and any missense variants (AAF < 1%). Abbreviations: EGFR, epidermal growth factor receptor domain; AAF, alternate allele frequency; RR, reference allele homozygote; RA, reference/alternate allele heterozygote; AA, alternate allele homozygote; 95% CI, 95% confidence interval

References

1. Mills, M.C. and C. Rahal, *The GWAS Diversity Monitor tracks diversity by disease in real time.* Nat Genet, 2020. **52**(3): p. 242-243.
2. Collaborators, G.B.D.S., *Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet Neurol, 2021. **20**(10): p. 795-820.
3. Sherin, A., et al., *Prevalence of stroke in Pakistan: Findings from Khyber Pakhtunkhwa integrated population health survey (KP-IPHS) 2016-17.* Pak J Med Sci, 2020. **36**(7): p. 1435-1440.
4. Wall, J.D., et al., *South Asian medical cohorts reveal strong founder effects and high rates of homozygosity.* Nat Commun, 2023. **14**(1): p. 3377.
5. Nakatsuka, N., et al., *The promise of discovering population-specific disease-associated genes in South Asia.* Nat Genet, 2017. **49**(9): p. 1403-1407.
6. Saleheen, D., et al., *Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity.* Nature, 2017. **544**(7649): p. 235-239.
7. Chabriat, H., et al., *Cadasil.* Lancet Neurol, 2009. **8**(7): p. 643-53.
8. Markidan, J., et al., *Smoking and Risk of Ischemic Stroke in Young Men.* Stroke, 2018. **49**(5): p. 1276-1278.
9. Mishra, A., et al., *Stroke genetics informs drug discovery and risk prediction across ancestries.* Nature, 2022. **611**(7934): p. 115-123.
10. Scott, E.M., et al., *Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery.* Nat Genet, 2016. **48**(9): p. 1071-6.
11. Wang, T., M. Baron, and D. Trump, *An overview of Notch3 function in vascular smooth muscle cells.* Prog Biophys Mol Biol, 2008. **96**(1-3): p. 499-509.
12. Duvaud, S., et al., *Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users.* Nucleic Acids Res, 2021. **49**(W1): p. W216-w227.
13. Rutten, J.W., et al., *Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank: CADASIL to nonpenetrance.* Neurology, 2020. **95**(13): p. e1835-e1843.
14. Rodriguez-Flores, J.L., et al., *The QChip1 knowledgebase and microarray for precision medicine in Qatar.* NPJ Genom Med, 2022. **7**(1): p. 3.
15. Hoffmann, T.J., et al., *Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation.* Nat Genet, 2017. **49**(1): p. 54-64.
16. Backman, J.D., et al., *Exome sequencing and analysis of 454,787 UK Biobank participants.* Nature, 2021. **599**(7886): p. 628-634.
17. Rutten, J.W., et al., *The effect of NOTCH3 pathogenic variant position on CADASIL disease severity: NOTCH3 EGFr 1-6 pathogenic variant are associated with a more severe phenotype and lower survival compared with EGFr 7-34 pathogenic variant.* Genet Med, 2019. **21**(3): p. 676-682.
18. ClinVar, N.C.f.B.I.
19. Joutel, A., et al., *Cerebrovascular dysfunction and microcirculation rarefaction precede white matter lesions in a mouse genetic model of cerebral ischemic small vessel disease.* J Clin Invest, 2010. **120**(2): p. 433-45.
20. Joutel, A., *Pathogenesis of CADASIL: transgenic and knock-out mice to probe function and dysfunction of the mutated gene, Notch3, in the cerebrovasculature.* Bioessays, 2011. **33**(1): p. 73-80.

21. Tikka, S., et al., *Congruence between NOTCH3 mutations and GOM in 131 CADASIL patients.* Brain, 2009. **132**(Pt 4): p. 933-9.
22. Ghezali, L., et al., *Notch3(ECD) immunotherapy improves cerebrovascular responses in CADASIL mice.* Ann Neurol, 2018. **84**(2): p. 246-259.
23. Belin de Chantemele, E.J., et al., *Notch3 is a major regulator of vascular tone in cerebral and tail resistance arteries.* Arterioscler Thromb Vasc Biol, 2008. **28**(12): p. 2216-24.
24. Mbatchou, J., et al., *Computationally efficient whole-genome regression for quantitative and binary traits.* Nat Genet, 2021. **53**(7): p. 1097-1103.
25. Greenland, S., *Applications of Stratified Analysis Methods*, in *Modern epidemiology: Third edition*. 2008, Lippincott Williams & Wilkins: Philadelphia. p. 295-297.
26. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological) 1995. **57**(1): p. 289-300.
27. Elliott, L.T., et al., *Genome-wide association studies of brain imaging phenotypes in UK Biobank.* Nature, 2018. **562**(7726): p. 210-216.
28. Griffanti, L., et al., *BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities.* Neuroimage, 2016. **141**: p. 191-205.
29. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

Acknowledgements

Supported by Regeneron Pharmaceuticals.

This research has been conducted using the UK Biobank Resource (project 26041). The authors thank everyone who made this work possible, particularly the UK Biobank team, their funders, the professionals from the member institutions who contributed to and supported this work, and most especially the UK Biobank participants, without whom this research would not be possible. The exome sequencing was funded by the UK Biobank Exome Sequencing Consortium (Bristol Myers Squibb, Regeneron, Biogen, Takeda, Abbvie, Alnylam, AstraZeneca and Pfizer). Ethical approval for the UK Biobank was previously obtained from the North West Centre for Research Ethics Committee (11/NW/0382).

Disclosure forms provided by the authors are available with the full text of this article.

Drs. Rodriguez-Flores, Khalid, Shuldiner and Saleheen contributed equally to this article.

Data Sharing

Data used in this manuscript not already available online can be requested from the authors.

Primary Individual Contributions

CONCEPTUALIZATION Juan L. Rodriguez-Flores (JLRF) Shareef Khalid (SK) Alan R. Shuldiner (ARS) Danish Saleheen (DS) DATA CURATION Nilanjana Banerjee (NB) Deepika Sharma (DeS) Michael Cantor (MC) John Overton (JO) Jeff Reid (JR) FORMAL ANALYSIS Bin Ye (BY) Manav Kapoor (MK) Joshua Backman (JB) Gannie Tzoneva (GT) Ellen Tsai (ET) Sahar Gelfman (SG) Tanima De (TD) Niek Verweij (NV) Luca A. Lotta (LAL) Aaron Zhang (AZ) Neelroop Parikshak (NP) Farshid Sepehrband (FS) Jonathan Marchini (JM) Giovanni Coppola (GC) Sofia Castaneda (SC) Pengcheng Xun (PX) Regeneron Genetics Center (RGC) FUNDING AQUISITION Danish Saleheen Alan R. Shuldiner Aris Baras (AB) Regeneron Genetics Center	INVESTIGATION Silvio Alessandro DiGioia (SADG) Hector Martinez (HM) I-Chun Tsai (IT) Katia Karalis (KK) Aris Economides (AE) David D'Ambrosio (DDA) Asif Rasheed (AR) METHODOLOGY Juan L. Rodriguez-Flores Shareef Khalid Regeneron Genetics Center PROJECT ADMINISTRATION Thomas Coleman (TC) Regeneron Genetics Center Asif Rasheed RESOURCES Muhammad Jahanzaib (MJ) Maleeha Zaman (MZ) Muhammad Rehan Mian (MRM) Muhammad Bilal Liaqat (MBL) Khalid Mahmood (KM) Tanvir-us-Salam (TUS) Muhammad Hussain (MH) Ayeesha Kamal (AK) Javed Iqbal (JI) Faizan Aslam (FA) SOFTWARE Regeneron Genetics Center Juan L. Rodriguez-Flores Shareef Khalid SUPERVISION Danish Saleheen Alan R. Shuldiner	VALIDATION Regeneron Genetics Center Shareef Khalid Juan L. Rodriguez-Flores VISUALIZATION Juan L. Rodriguez-Flores Neelroop Parikshak Farshid Sepehrband Jonathan Marchini Regeneron Genetics Center WRITING - ORIGINAL DRAFT Sergio Fazio (SF) Wolfgang Liedtke (WL) John Danesh (JD) Ayeesha Kamal (AK) Philippe Frossard (PF) WRITING - REVIEW AND EDITING

Competing Interests

Funding

Fieldwork for this study was funded by the Center for Non-Communicable Diseases, Pakistan. DNA sequencing was funded by Regeneron Pharmaceuticals Inc.

Employment

JLRF, ARS, NB, DeS, MC, JO, JR, BY, MK, JB, GT, SG, TD, NV, LAL, AZ, NP, FS, JM, GC, PX, AB, SADG, HM, IT, KK, AE, DDA, SF, WL, ET, SC, TC are or were employees of Regeneron Genetics Center LLC or Regeneron Pharmaceuticals Inc. and contributed to this manuscript as part of their regular duties as salaried employees.

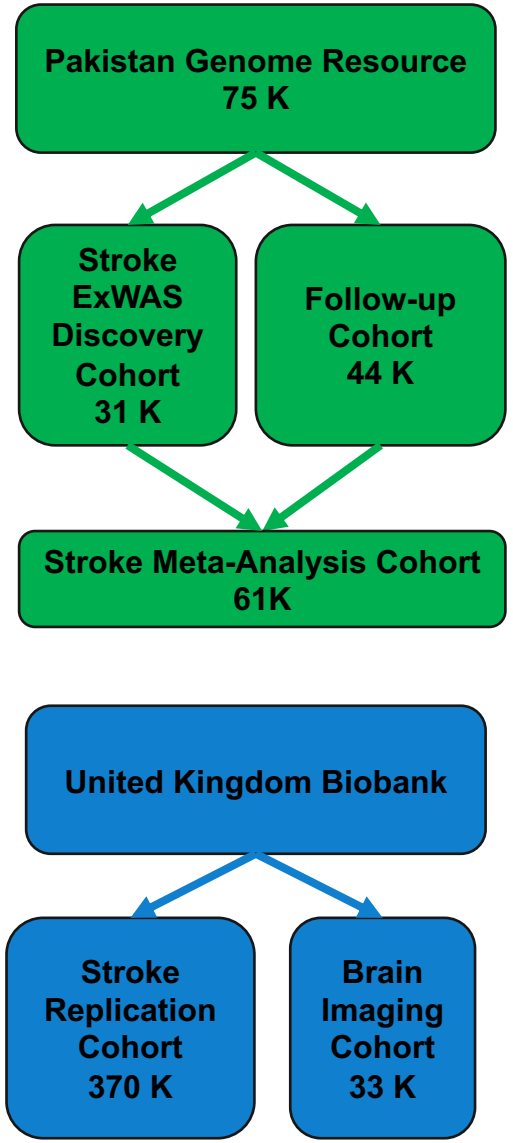
AR, MJ, MZK, MRM, MBL, PF, and DS and SK are employees of the Center for Non-Communicable Disease and received salaried compensation for their contribution to this manuscript.

Personal Financial Interests

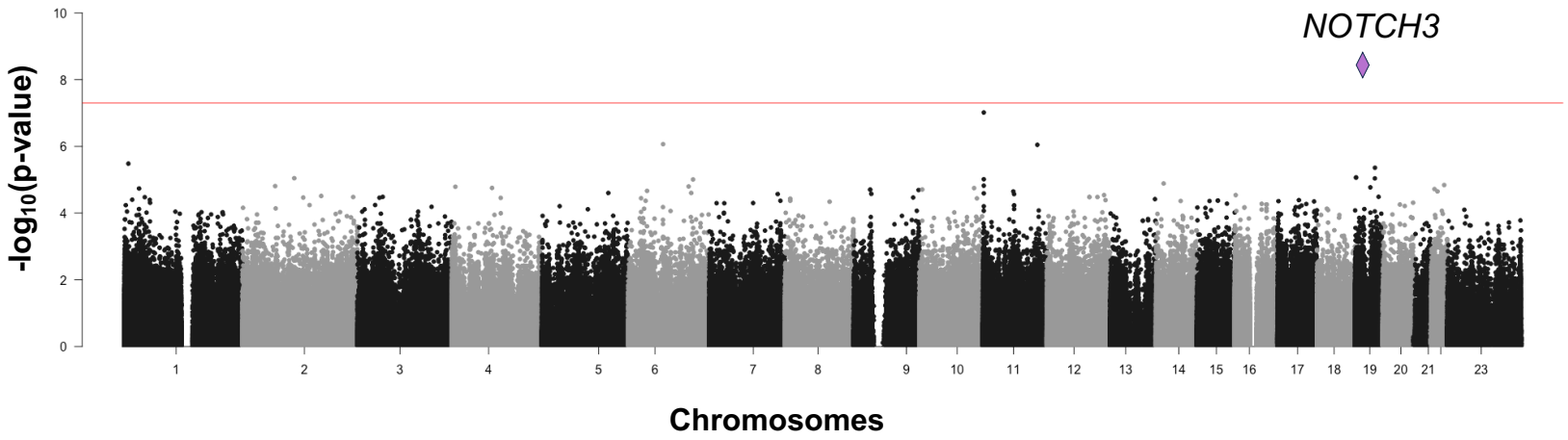
JLRF, ARS, NB, DeS, MC, JO, JR, BY, MK, JB, GT, SG, TD, NV, LAL, AZ, NP, FS, JM, GC, PX, AB, SADG, HM, IT, KK, AE, DDA, SF, WL, TC are or were employees of Regeneron Genetics Center LLC or Regeneron Pharmaceuticals Inc. and received stock and stock options as part of their compensation as employees.

JLRF, ARS, DS, AB, and SK are named inventors on patent pending US 20230000897A1 that discloses methods of treating subjects having a cerebrovascular disease by administering Neurogenic Locus Notch Homolog Protein 3 (NOTCH3) agents, and methods of identifying subjects having an increased risk of developing a cerebrovascular disease.

A.



B.



C.

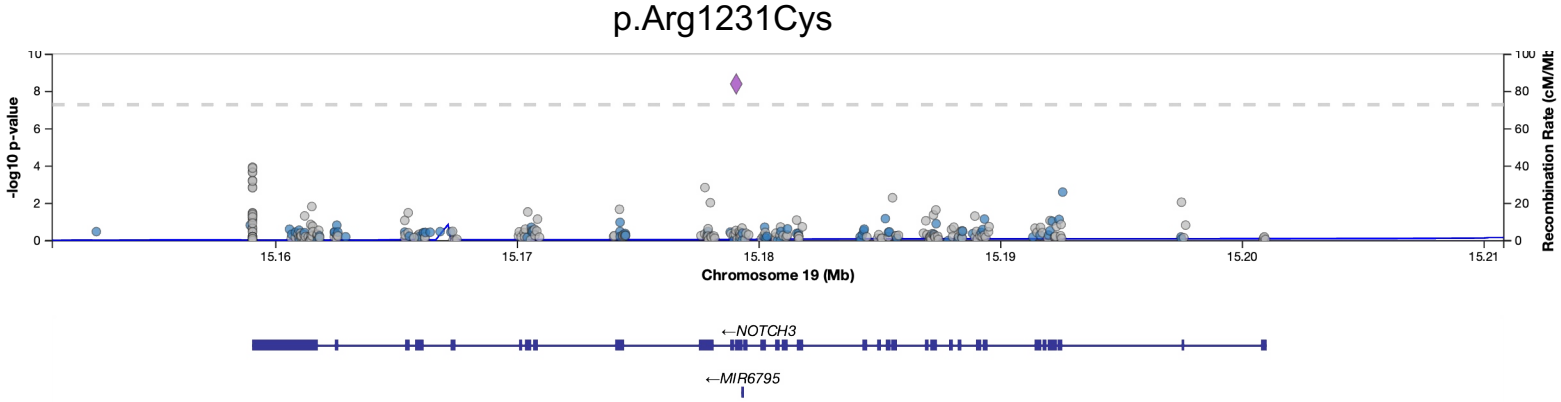
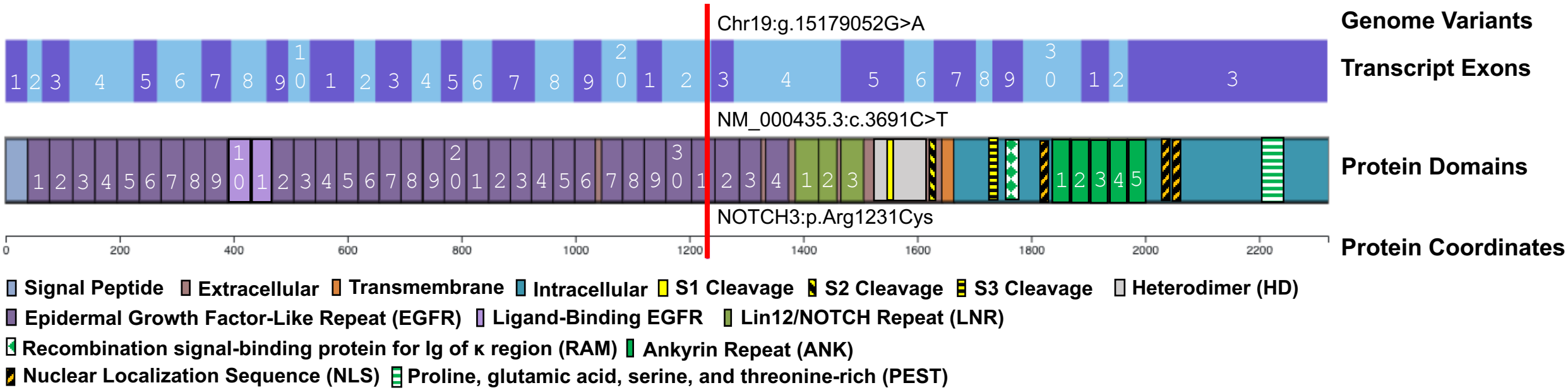


Figure 1

Regeneron - Internal Use Only

A.



B.

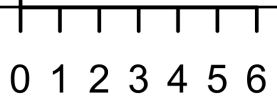
1205 1231 1244

```
INECRSGACHAAHTRDCLQDPGGGFRCLCHAGFSGPRCQT  Human 1231 Arg (reference)
INECRSGACHAAHTRDCLQDPGGGFCCLCHAGFSGPRCQT  Human 1231 Cys (mutant)
INECRSGACHAAHTRDCLQDPGGGFRCLCHAGFSGPRCQT  Chimpanzee
INECRPGACHAAHTRDCLQDPGGHFRCVCHPGFTGPRCQI  Mouse
INECLSNPCNPSNSLDCIQLPND-YQCVCKPGFTGRGCQS  Zebrafish
INECLSGPCHAQNTRHCVQLAND-YQCVCKSGYTGRRCQS  Western clawed frog
INECLAKPCLPQRTLDCVQGAND-FQCLCKPGYTGRRCQN  Green sea turtle
```

Figure 2

Regeneron - Internal Use Only

Cohort	Trait	OR (95% CI)	P value	AAF	Cases (RR|RA|AA)	Controls (RR|RA|AA)
Discovery	Stroke		4.92E-08	6.13E-03	4939|98|3	25619|271|2
	Ischemic stroke		2.49E-03	5.64E-03	2540|42|2	25619|271|2
	Intracerebral hemorrhage		1.79E-08	5.82E-03	1451|44|0	25619|271|2
	Subcortical intracerebral hemorrhage		3.87E-09	5.84E-03	1313|43|0	25619|271|2
	Partial anterior circulation infarcts		4.84E-03	5.56E-03	1227|25|1	25619|271|2



Cohort	Trait	OR (95% CI)	P value	AAF	Cases	Controls
Discovery	Stroke		4.92E-08	6.13E-03	4939|98|3	25619|271|2
Replication	Stroke		5.00E-03	0.00513	154|6|0	29935|302|2
IVW Meta-Analysis	Stroke		7.79E-10	0.00576	5093|104|3	55617|586|4

Figure 3

Regeneron - Internal Use Only