1. Title Page

Rare Variant Association Analysis Uncovers Involvement of *VNN2* in Stroke Outcome

Estefanía Alcaide-Consuegra, MSc[1,2,3]; Marina Mola-Caminal, PhD[4,5]; Georgia Escaramís, PhD[6,7]; Uxue Lazcano, MSc[4,8]; Isabel Fernández Pérez, MD[4]; Jaume Reig-Palou, MSc[1,2]; Joan Jimenez-Balado, PhD[4]; Eva Giralt-Steinhauer, MD PhD[4]; Elisa Cuadrado-Godia, MD PhD[4,9]; Angel Ois, MD PhD[4,9]; Ana Rodríguez-Campello, MD PhD[4,9]; Marta Vallverdú, PhD[4], Aina Medina-Dols, PhD[10]; Carmen Jiménez, MD[10,11]; Silvia Tur, MD[10,11]; Rosa M Díaz, MD[10,11]; Carlos David Bruque, PhD[12]; Nuria Andreu-Somavilla, PhD[13]; Irene González-Navarrete, PhD[13]; Ferran Casals, PhD[1,2,3]; Cristòfol Vives-Bauzà, PhD[10,14]; Israel Fernández-Cadenas, PhD[15]; Jordi Jiménez-Conde, MD PhD[4,9]; Susanna Balcells, PhD[1,2,3,16]; Raquel Rabionet, PhD[1,2,3,16]

[1] Department of Genetics, Microbiology and Statistics, Faculty of Biology, Universitat de Barcelona, Barcelona, Spain.

[2] Institute of Biomedicine of the Universitat de Barcelona (IBUB), Universitat de Barcelona, Barcelona, Spain.

[3] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Madrid, Spain.

[4] Neurovascular Research Group. Department of Neurology, Hospital del Mar Research Institute, Barcelona, Spain

[5] Current affiliation: Unit of Medical Epidemiology, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.

[6] Department of Biomedical Sciences, Institute of Neuroscience, Universitat de Barcelona, Barcelona, Spain.

[7] Centro de Investigación Biomédica en Red de Epidemiología y salud pública (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain.

[8] *Current affiliation: Center for Cooperative Research in Biosciences (CIC bioGUNE), Basque Research and Technology Alliance (BRTA), Bizkaia Technology Park, Biscaia, Spain.*

[9] *Medicine Department, DCEXS-Universitat Pompeu Fabra (UPF), Barcelona, Spain.*

[10] *Neurobiology Laboratory, Research Unit, Hospital Universitari Son Espases, Health Research Institute of Balearic Islands (IdISBa), Palma, Spain.*

[11] *Department of Neurology, Hospital Universitari Son Espases, Health Research Institute of Balearic Islands (IdISBa), Palma, Spain.*

[12] *Unidad de Conocimiento Traslacional Hospitalaria Patagónica - Hospital de Alta Complejidad El Calafate SAMIC, El Calafate, Santa Cruz, Argentina.*

[13] *Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.*

[14] *Department of Biology, University of Balearic Islands (UIB), Institut Universitari d'Investigacions en Ciències de la Salut (IUNICS), Palma, Spain.*

[15] *IFC: Stroke Pharmacogenomics and Genetics, Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain.*

[16] *Institut de Recerca Sant Joan de Déu (IRSJD), Esplugues de Llobregat, Barcelona, Spain.*

Short title (not to exceed 50 characters, including spaces)

Stroke outcome RVAS identifies *VNN2* association

Name, email address, and complete address of corresponding author

Correspondence to Dr. Raquel Rabionet, Department of Genetics, Microbiology and Statistics, IBUB, IRSJD, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalunya 08028, Spain; kelly.rabionet@ub.edu

Social media handles:

Instagram: @genetics.ub

Bluesky:@krabio.bsky.social @alcoestefania.bsky.social

@geneticsub.bsky.social

The total word count of the manuscript

6448 words

2. Abstract

BACKGROUND: A stroke's functional outcome presents vast variability among patients, which is influenced by age, sex, characteristics of the lesion, and genetic factors. However, there is very little knowledge about stroke recovery genetics. Recently, some GWAS (Genome-Wide Association Studies) have highlighted the involvement of common or low-frequency variants near or within *PATJ*, *PPP1R21, PTCH1, NTN4* and *TEK genes*, whereas the role of rare variants is still unclear. This study aims to identify the genetic contributions to differences in stroke outcome analyzing the effect of rare variants.

METHODS: We performed a pilot study analyzing 90 exomes of extreme good and bad recovery (modified Rankin Scale (mRS) at 3 months 0-1 vs 4-5) to select target genes involved in stroke recovery. To expand this study, 702 additional samples were sequenced by Targeted Next-Generation Sequencing capturing loci selected from the pilot study, GWAS studies and literature input. Here, we performed continuous (mRS 0-6) and dichotomous (mRS 0-1 vs 3-6) analyses, yielding one candidate gene. All samples were selected by a retrospective cohort study from incidental stroke cases collected at Spanish Hospitals between 2000-2018. The identified *VNN2* variants were assessed for protein structure and stability analysis, and an analysis of their effect on basal inflammation levels was performed using UKBiobank data.

RESULTS: Our work identified rare coding variants in *VNN2* associated with patients with a better stroke recovery (ΔDIC > 10, equivalent to p value < 0.001). Six rare variants were predicted to significantly affect protein stability ($\Delta\Delta$G > 1.6

kcal/mol), meanwhile, another variant, located in the active site, could affect the electrostatic surface.

CONCLUSIONS: We propose that *VNN2* might play a role in stroke outcome by modulating post-stroke inflammation. A potentially affected function would be neutrophil cell adhesion and migration.

KEY WORDS: ischemic stroke, genetics, rare variants, VNN2, inflammation, neutrophils, transendothelial migration.

Nonstandard abbreviations and Acronyms

- BATI: Bayesian rare variant Association Test using Integrated Nested Laplace Approximation
- CADD: Combined Annotation Dependent Depletion
- CE: Cardioembolism
- DAMPs: Damage-associated molecular patterns
- DIC: Deviance Information Criterion
- FS: Fisher Strand
- GATK: Genome Analysis Toolkit
- GWAS: Genome-Wide Association Study
- GPI: Glycosylphosphatidylinositol
- IS: Ischemic Stroke
- KBAC: Kernel Based Adaptive Clustering Method
- LAA: Large-Artery Atherosclerosis
- LMR: Lymphocyte-to-Monocyte Ratio
- MQRankSum: Mapping Quality Rank Sum Test
- mRS: modified Rankin Scale
- NGS: Next-Generation Sequencing
- NIHSS: National Institutes of Health Stroke Scale
- NLR: Neutrophil-to-Lymphocyte Ratio
- NMD: Nonsense Mediated Decay
- OD: Other Determined

- PCA: Principal Component Analysis
- QD: Quality by Depth
- QUAL: Quality
- RVAS: Rare variant association studies
- rvGWAS: Rare variant genome wide association framework
- ReadPosRankSum: Read Position Rank Sum Test
- RMSMQ: Root Mean Square (RMS) Mapping Quality (MQ)
- SII: Systemic Immune-Inflammation Index
- SKAT: Sequence Kernel Association Test
- SKAT-O: Optimal Sequence Kernel Association Test
- SOR: Strand Odds Ratio
- TOAST: Trial of Org 10172 in Acute Stroke Treatment
- WES: Whole Exome Sequencing

3. Introduction

Worldwide, stroke is the second leading cause of death and adult disability. Approximately 1.1 million people in Europe suffer a stroke every year, and its incidence and prevalence are expected to increase along with the aging of the population[1]. The outcome of a stroke is affected by many factors including gender, age, stroke severity, size, and location. However, even adjusting for these factors, significant clinical heterogeneity remains. This heterogeneity can be partially explained by genetic variation affecting proteins involved in the stroke recovery process[2].

Stroke recovery is a complex process that involves neuronal, vascular and immune responses[3]. Initially, after the ischemic event, in response to the injury, necrotic and dying neurons release damage-associated molecular patterns (DAMPs). These activate an innate immune response, including glial activation and infiltration of blood-borne immune cells into the brain. Microglia secrete matrix metallopeptidases that disrupt the integrity of the blood brain barrier,

136 facilitating the invasion of macrophages and neutrophils. While this immune
137 response is beneficial at stroke onset, if this pro-inflammatory response is
138 extended, as it happens in aged individuals, it contributes to the participation of
139 T-cells and a magnification of the immune response, and a worse stroke
140 outcome[4]. Once this first response has taken place, poststroke recovery
141 processes lead to restoration or compensation of function. These are based on
142 the induction of key biological processes such as angiogenesis, neurogenesis,
143 axonal sprouting, dendritic branching, synaptogenesis, and oligodendrogenesis[3].
144 An avenue to understand the relevance of all these processes is to identify key
145 genes involved in stroke outcome, and the biological pathways that mediate the
146 identified allele-outcome correlations.

147 Only a few studies have investigated the role of genome-wide genetic variation
148 in ischemic stroke (IS) outcome, focusing on the common (>1%) genetic variation
149 captured by classic GWAS (i.e. genotyping arrays)[5-8]. Studies looking at mid-term
150 outcome (60 to 190 days) have identified two genome-wide significant
151 associations, with variants near *PATJ*[5], a gene related to tight junction formation
152 and maintenance[9], and variants regulating *PPP1R21*[6], a gene involved in
153 learning, memory and neuronal plasticity[10-12]. They also found suggestive
154 association with *PTCH1*, *TEK*, and *NTN4*[6]. In addition, the analysis of the global
155 effect of copy number variation, or genomic imbalance, on stroke outcome,
156 yielded association of increased genomic imbalance with poorer stroke outcome[7].
157 Finally, another GWAS has uncovered associations of excitotoxicity related
158 genes with short-term (24h) stroke outcome[8], which is also correlated with 90-
159 days outcome[13]. Meanwhile, the potential effect of rare variants, tackled by
160 exome or whole genome sequencing, remains unexplored[2].

Importantly, there are few to none widely accepted neuroprotective or neuroreparative drugs, nor personalized approaches to guide therapies to mitigate ischemic brain injury or enhance recovery[2]. Therefore, advances in the identification of novel outcome related variants and genes, coupled with the understanding of the pathways through which these genes affect stroke outcome, should provide critical knowledge for drug selection and allow for personalized therapy approaches.

The improvement of Next-Generation Sequencing (NGS) methods has allowed the study of both common and rare variants in complex diseases. Some studies have shown that rare variants could have higher impact on the structure, stability, or function of proteins than common variants[14,15], explaining part of the heritability of complex traits[16]. However, association tests of individual rare variants require very large sample sizes, which are difficult to obtain for very specific phenotypes, such as a stroke's functional outcome. To overcome this limitation, various rare variant association studies (RVAS) aggregate the effects of variants affecting the same biological entity (e.g. genes), and test for association of the aggregated variants with the phenotype. Other RVAS go further and consider more complex scenarios, such as heterogeneity of the variant's effects (SKAT, SKAT-O), or variant-specific characteristics (BATI), providing increased statistical power[17-19].

In the present analysis we explore the effect of rare variants on stroke outcome by applying BATI, which also integrates patients- and variant-specific characteristics as covariates. We performed a two-phase study, including exome sequencing in a small (n=90) but carefully curated cohort of extreme outcome phenotypes to select genomic regions that were further explored in a targeted sequencing analysis in 702 additional stroke cases. One gene, *VNN2*, harbored

an excess of rare variants that potentially affect protein function in individuals with better outcome scores.

4. Methods

Data availability

Exome sequencing data from the pilot study is deposited at the European Genome Archive (EGA dataset ID: EGAD00001004808). Targeted resequencing data is available upon request.

Study Design

The study was divided into two phases: a first approach through a whole exome sequencing (WES) pilot study, and a follow-up by targeted resequencing analysis. The WES pilot study was performed on a selection of 90 cases matched by age, gender, and stroke type and location, divided into good (mRS 0-1) or poor (mRS 3-5) functional outcome. A second phase involved targeted resequencing analysis of 702 cases with a wider range of outcomes (mRS 0-6). Selected regions for resequencing included the top genes identified in the pilot WES study together with additional genes and regions selected based on published GWAS hits and a relaxed analysis of the Mola-Caminal *et al.* [5] discovery GWAS dataset combined with protein interaction networks and literature support (See Supplemental file S1).

Study Subjects

This was a retrospective cohort study using data and DNA samples from three sources: a cohort of patients with incidental stroke admitted to the Hospital del Mar in Barcelona between 2005-2018 (the BASICMAR[20] study); a cohort of

patients with incidental stroke admitted to one of 23 participating Spanish Hospitals between 2005 and 2009 (the GRECOS[21] study); and a cohort of patients with incidental stroke admitted to the Vall d'Hebron hospital or one of 5 other participating hospitals between 2000-2005 (the Geno-tPA[22] study). A subset of patients with European ancestry, a diagnosis of IS according to World Health Organization criteria, fulfilling inclusion and exclusion criteria, with available DNA samples, and including a spectrum of recovery outcomes were included in the current analysis.

The Pilot study included 69 patients from the BASICMAR study and 21 patients recruited at hospital Vall d'Hebron for the GRECOS and Geno-tPA studies. These patients were distributed in two subgroups based on their mRS scores at three months (0-1, "good outcome", 49 patients vs 3-4-5 "poor outcome", 41 patients), which were matched by age, gender, stroke location and type (Table 1). Inclusion and exclusion criteria are detailed in Supplementary methods. The targeted resequencing phase included 702 patients from the same cohorts (441 samples from BASICMAR study and 261 samples from GRECOS and Geno-tPA studies; Table 2 and Figure S1), fulfilling the same inclusion and exclusion criteria, except for the patients with unusual stroke ("other determined etiology" in TOAST). The study was approved by the Institutional Review Boards of the participant hospitals (CEIm-PSMAR (2008/3083/I); IRB00002850 (PR(AG)157/2011)) and the CEIC of Fundació Sant Joan de Déu (C.I: PIC-82-17), and all participants provided written informed consent to participate. The research was conducted in accordance with the Helsinki declaration.

Sequencing

For the pilot study, DNA obtained from peripheral blood mononucleated cells was fragmented using a covaris system. Libraries were generated with TruSeqTM DNA Sample Preparation v2 Kits (Illumina, Inc., San Diego, CA, USA) followed by exome capture with NimbleGen SeqCap EZ Library SR v3.0 (Roche, Inc., Madison, WI, USA) in pools of 5 samples, which were sequenced to a 30-40x coverage on an Illumina HiSeq2500 at the Spanish National Center for Genomic Analysis (CNAG).

For the targeted phase study, libraries were prepared from DNA from peripheral blood mononucleated cells and captured with a custom agilent sureselect capture kit, using an in-house sample preparation system. Libraries were pooled and sequenced on Illumina Hiseq 2500 at the CRG-CNAG, to a targeted coverage of 150x.

Bioinformatic Analysis of sequencing data

FASTQ raw sequences were processed following an in-house pipeline (Figure S2) including quality control with FASTQC, alignment with BWA-mem (http://bio-bwa.sourceforge.net/), duplicate marking with Picard (https://broadinstitute.github.io/picard/), Samtools (http://www.htslib.org/) processing, local realignment base recalibration and variant calling with the Genome Analysis Toolkit (GATK; https://software.broadinstitute.org/gatk/). Variants were called by HaplotypeCaller using standard parameters except for min-pruning, which was set at 5 in the targeted analysis. Genome build 37 (GRCh37/hg19) was used as the reference genome. For exome sequencing, variants were further filtered with VQSR, following standard recommendations,

while for the targeted resequencing, the resulting variants were hard-filtered following GATK recommendations (https://gatk.broadinstitute.org/) [QualbyDepth (QD) < 2.0; Quality (QUAL) < 30.0; StrandOddsRatio (SOR) > 3.0; FisherStrand (FS) > 60.0; RMSMappingQuality (MQ) < 40.0; MappingQualityRankSumTest (MQRankSum) < -12.5; ReadPosRankSumTest (ReadPosRankSum) < -8.0]. Multiallelic variants were filtered out with BCFtools (http://www.htslib.org/). The filtered VCF files were annotated with ANNOVAR (https://annovar.openbioinformatics.org/).

Quality Control

In the pilot phase, 90 samples passed all quality criteria. In the targeted phase, three samples were removed based on inclusion and exclusion criteria, and another five were discarded because of missing information. Quality control on the pilot phase samples and the remaining samples of the targeted phase was performed on the rvGWAS framework (available at https://github.com/hanasusak/rvGWAS). In both phases, variants with a genotyping call rate <95% were removed, as well as samples with <95% called variants, and outlier samples based on principal component analysis (PCA) and the number of called variants (Figure S3). This resulted in the removal of 25 samples in the targeted analysis, leaving a total of 677 samples for analysis. Finally, variants in top selected genes were manually curated by revision of variant calling files in IGV v. 2.8.0.

rvGWAS framework

For downstream statistical analysis, only rare (European Allele Frequency < 0.01), nonsynonymous, probably damaging (CADD[23] score > 20) exonic and

splice-site variants were considered in both the pilot and the follow-up phase. In the pilot phase, association with the 3-month mRS was considered under a dichotomous model only (outcome mRS 3m 0-1 vs 3-5, analyzed with BATI and SKAT-O). In the follow-up, we considered two models, a continuous model (outcome mRS 3m 0-6, all 677 samples) and a dichotomous model (outcome mRS 3m 0-1, 216 samples, vs 3-6, 335 samples, totaling 551 samples). In both cases, analysis was performed with the Burden test, SKAT-O, KBAC, and BATI. Analyses were adjusted for the first five principal components, age, sex, TOAST, and, to adjust for stroke severity, the initial NIHSS. For the BATI test, variant effect (classified as loss-of-function or missense) was included as a covariate. As BATI is a Bayesian test, to provide an analogy measure to the frequentist version of significance level, we obtained an empirical significance threshold for the follow-up test (ΔDIC; DIC: deviance information criterion, a metric used to compare Bayesian models) using a bootsrap strategy. We performed 1000 simulations randomizing the 3-month mRS score; in each simulation, genes were ranked by their ΔDIC values, and the highest score was extracted, selecting the thresholds at 1% and 0.1% significance level ($\Delta DIC_{0.01}$= 5.57, $\Delta DIC_{0.001}$= 10.3), as in Susak *et al*. [19]. We also considered a threshold of 10, a threshold commonly used as a rule of thumb in the context of Bayesian models ($\Delta DIC_{default} > 10$) [19].

In-silico Variant Functional Analysis

The effect of variants on protein structure (see supplementary methods) and stability was computed. Protein stability was calculated using FoldX (http://foldxsuite.crg.eu/)[24]. The structures were optimized to the FoldX force field command from the structure of VNN2 protein. The $\Delta\Delta G$ values were estimated as the difference between the energy of the wild-type and protein mutation (five

replicates for each point variation). Values above 1,6 kcal/mol (twice the standard deviation) were considered to significantly destabilize the protein[25,26]. Finally, we used UK Biobank data to assess if presence of these rare variants in VNN2 correlated with alterations in baseline inflammation biomarkers (see supplementary methods).

5. Results

Pilot study

The objective of the pilot study was to select approximately 100 candidate genes, potentially enriched in rare variants in either dependent or independent patients, to be followed up in a larger cohort by targeted resequencing. The final list of targeted regions which included regions selected based on the BATI and SKAT-O results and additional regions based on the literature, are provided in the Supplemental file S1.

Targeted NGS study highlights *VNN2* as a candidate gene

The analysis under the continuous outcome model did not yield any gene with significant enrichment of rare coding variants (Table S1 and S2). However, *VNN2*, the top gene in the BATI analysis, showed a ΔDIC value of 6.326, above the 1% empirical significance threshold ($\Delta DIC_{0.01} = 5.57$).

In the analysis under the dichotomous model, we compared patients with a 3-month mRS of 0 or 1 with those with a 3-month mRS score over 3, using the following test and corresponding thresholds: Burden, SKAT-O, KBAC (*p* value < 0.05) and BATI ($\Delta DIC_{0.001} < 10.3$; $\Delta DIC_{default} = 10$). Analysis with BATI showed the same gene as continuous analysis, *VNN2*, to be significantly enriched ($\Delta DIC_{VNN2} = 10.773$) for rare variants in patients with better outcome (Table 3).

While not significant, *VNN2* was also among the top genes identified in the other tests (Tables S3, S4, and S5).

VNN2 variants identified in IS patients affect protein stability

Six *VNN2* variants were identified in the follow-up cohort, five of which were identified in six patients with good recovery (mRs at 3 months between 0 and 1), and one in one case with poor recovery (mRs at 3 monts of 4). In addition, in the pilot study we had identified an additional two variants in cases with good recovery, while no variants were present in patients with poor recovery.

We explored the potential effect of the identified *VNN2* variants *in silico* (Table 4), and all of them were predicted to have a significant effect on the protein, either by affecting its stability, by altering its electrostatic surface or by truncating the protein.

Six out of eight variants are located in the CN hydrolase domain (Figure 1 and S4A). Two variants, p.(Ser46Phe) and p.(Leu53Pro), are in the first alpha helix. Both substitutions are expected to have a clear structural impact on the helix. In one case, by replacing a small proline for a bulky, aliphatic leucine; in the other, a very large, aromatic, non-polar amino acid (Phe) for a tiny, polar one (Ser). This substitution (p.(Ser46Phe)) is predicted to create a steric effect that would compromise its interaction with the Aspartic residue in position 49 (Figure S4B). On the other hand, variants p.(Val174Met) and p.(Arg205Ser), located in the beta-pleated sheets A and B, are predicted to affect the interaction of their corresponding residues with the threonine residue at 198 and the phenylalanine at 199, respectively, destabilizing the region between these beta sheets (Figure S4C). Variant p.(Ala265Thr) is also predicted to cause a steric shift destabilizing

the protein (Figure S4E), as it causes the substitution of a larger uncharged polar residue (Thr) for a small nonpolar aliphatic residue (Ala) in an internal hydrophobic region of the protein close to the active site. Finally, p.(Gly284Cys), the last variant in the CN-hydrolase domain, located at the end of the beta B fold sheet, would cause an effect on the adjacent loop affecting its mobility (Figure S4F).

The remaining two variants affect the same amino acid in the Vanin C domain (Figure 1). These variants are located in the exon 5 and may trigger nonsense-mediated decay (NMD). Even if NMD does not occur, the variant would still remove two thirds of this domain as well as the propeptide, including the loss of the GPI anchor. Moreover, the variant p.(Arg393Gln) is predicted to cause a change in the surface electrostatic charge of the protein. It is located at the access of the substrate to the active site, in a stretch of 4 consecutive arginines, conserved between species, that confer a positive charge to this region, which is affected by the substitution of an uncharged glutamine for the electropositive arginine (Figure S4D).

6. Discussion

After a stroke, the interaction of different environmental and genetic factors may define the functional outcome of the cerebrovascular accident. Although GWAS studies have successfully identified common variants involved in stroke recovery, rare variants have not been explored yet.

Here, for the first time, we performed an association study for rare variants involved in stroke recovery. This analysis involved a targeted analysis of 100 genomic regions in 702 patients, considering both a continuous (3-month mRS

0-6) and dichotomous (3-month mRS 0-1 vs 4-5) outcome variable model. Analysis was performed using SKAT-O and the BATI rare variant association test, which allows the integration of patient- and variant-specific features as covariates[19], and was proposed to have an improved power for the identification of the risk genes, especially in architectures with high genetic heterogeneity. BATI analysis yielded one gene, *VNN2,* significantly (ΔDIC > 10, equivalent to *p* value < 0.001) associated with functional independence at 3 months.

VNN2 (Vanin-2) [27] encodes a glycosylphosphatidylinositol (GPI)-anchored extracellular protein also known as GPI-80[28]. VNN2 has been identified to play a role in the cellular adhesion and transmigration processes of human neutrophil extravasation[28]. While the precise role of VNN2 in this multi-step process remains to be defined, it seems to take place in the transition from rolling to firm adhesion[28-30].

On the other hand, VNN2, together with VNN1 and VNN3, belong to the vanin protein family, characterized by their pantetheinase activity. Pantetheinase hydrolyzes pantetheine to pantothenic acid (vitamin B_5) and cysteamine, activating the stress pathway and inflammation. While this activity is stronger in VNN1, VNN2 also has this capability through its CN-Hydrolase domain[31,32].

We have identified eight rare variants in *VNN2*, seven of them present in patients with a good stroke outcome, while one variant was present in a single patient with poor outcome. *In silico* protein structure analysis predicted an effect in protein stability for the six missense variants located in the CN-Hydrolase domain (Table 4), indicating a potential effect of these variants through the modification of VNN2's enzymatic activity. While the remaining missense variant,

(p.(Arg393Gln)), is located in the vanin domain, and was not predicted to affect protein stability, it was predicted to affect the electrostatic surface charge, potentially limiting the access to the nearby active site. The last variant, p.(R393*), would lead to a truncated protein sequence lacking the propeptide sequence and the GPI anchor site. However, given the location of the nonsense mutation, it would be expected to lead to NMD and absence of protein. Therefore, all the heterozygous identified variants could lead to a reduction in VNN2 function, which could be associated with a lower inflammatory response. To explore this relationship, we analyzed different inflammatory biomarkers (Neutrophil-to-Lymphocyte Ratio (NLR), Lymphocyte-to-Monocyte Ratio (LMR), Systemic Inmune-Inflammation Index (SII) and C-reactive protein) in individuals from the UK Biobank who carried any of the identified *VNN2* variants However, no statistical differences were observed between carriers of these variants and a control subset of individuals (Table S6),), suggesting that these variants do not influence baseline inflammation levels. Given our hypothesis that VNN2 variants may affect neutrophil extravasation it is possible that subclinical variations in its function may not manifest as changes in baseline inflammatory markers.

The role of inflammation and immune mechanisms in the stroke recovery process is being increasingly recognized. Neutrophil extravasation and infiltration during inflammation are major contributors to poor ischemic outcomes[33]. In fact, two recent manuscripts have shown that a reduction in neutrophil infiltration correlates with better outcomes in animal models[34,35]. Neutrophil interaction with endothelial adhesion molecules, a process in which VNN2 could be involved, together with the ischemic environment, is reported to shift the neutrophil

phenotype from a protective anti-inflammatory (N2) to more damaging pro-inflammatory (N1) phenotype[33].

We speculate that the enrichment of *VNN2* rare variants in patients with better outcome score might be related to an effect of the variants on the role of VNN2 in neutrophils. Since it has previously been proposed that VNN2 plays a role in the neutrophil extravasation process, we hypothesize that variants affecting VNN2 might lead to reduced extravasation by either directly affecting the adhesion to the endothelium, or the shift to pro-inflammatory neutrophils, and reducing the infiltration of neutrophils into the brain parenchyma[33]. Alternatively, the variants might lead to a reduced pantetheinase activity, conferring higher resistance to oxidative stress and leading to reduced levels of inflammation. It would have been interesting to test for a correlation between the presence of variants in VNN2 with inflammation levels after the stroke event, however, data on inflammation levels at either acute stroke or at the 3 months timepoint was not collected.

To summarize, we have identified an excess of *VNN2* rare variants, predicted to affect protein stability or function, in stroke patients with a low mRs score at the 3-month evaluation, classified as a good functional outcome. We hypothesize that the loss or reduction of VNN2 activity could lead to lower inflammation in the context of tissue injury-induced inflammatory responses. It has been shown that while this inflammation process is a necessary event, excessive inflammation can be detrimental in the recovery process[4]. A lower neutrophil extravasation or protection against oxidative stress caused by reduced VNN2 activity could lead to a reduced inflammatory response and allow for a better outcome. Given that the literature on VNN2 and neutrophil biology is fairly limited, it would be

interesting to perform further studies to assess the role of VNN2 in neutrophil activation and extravasation, and the potential effect of these variants.

This study has potential limitations. Despite power calculations suggest that we have over 75% power (Table S7) to detect an association, a larger cohort would likely yield more robust results. Additionally, the functional impact of the identified variants in VNN2 has not been experimentally validated. Furthermore, it has not been possible to adjust the analysis by inflammation levels due to the absence of inflammation data from the patients.

In conclusion, we present the first study of rare variants involved in stroke recovery, highlighting a possible relationship between stroke outcome and rare variants in *VNN2*. This protein could act on oxidative stress response, or cell adhesion and migration of neutrophils contributing to a good outcome after stroke. Therefore, VNN2 might be a novel therapeutic target for stroke recovery. Nonetheless, to expand our knowledge of the rare variant architecture of IS recovery, it would be necessary to extend the sample size of the cohort and to increase the captured regions by using WES data. This would improve the detection of rare variants and the identification of novel genes, improving statistical power in both the continuous and dichotomous models.

7. Acknowledgments

We thank all the patients who suffered an ischemic stroke and allowed us to use their data in the study. This research has been conducted using the UK Biobank Resource under application number 119044.

8. Sources of Funding

This research was supported by Fundació Marató TV3 through the GENIUS study (307/U/2017) (RR), the GODs study (776/C/2011) (JJC) and *the Epigenesis study (188/C/2017)(IFC)*; by the Fondo de Investigaciones Sanitarias-Instituto de Salud Carlos III grant PI21/00890 (CV) and RICORS Ictus network (RD21/0006/0004, RD21/0006/0007) supported with Next Generation funds; by grants PID2022-141461OB-I00 (RR) and PID2021-125106OB-C32 (FC) funded by Spanish Ministerio de Ciencia, Innovación y Universidades MICIU/AEI/10.13039/501100011033 and by ERDF/EU; and by grants AGAUR-2021SGR-1093 (RR) and AGAUR-2021SGR-0656 (IFC) from Generalitat de Catalunya. EAC is supported by grant FPU2021/01324 from MICIU/AEI/10.13039/501100011033.

9. Disclosures

None.

10. Supplemental Material

- Supplemental Methods
- Tables S1-S7
- Figure S1-S4
- Supplementary file S1
- References 36-40

11. References

1. Bejot Y, Bailly H, Durier J, Giroud M. Epidemiology of stroke in Europe and trends for the 21st century. Presse Med. 2016;45:e391-e398. doi: 10.1016/j.lpm.2016.10.003
2. Lee JM, Fernandez-Cadenas I, Lindgren AG. Using Human Genetics to Understand Mechanisms in Ischemic Stroke Outcome: From Early Brain Injury to Long-Term Recovery. Stroke. 2021;52:3013-3024. doi: 10.1161/STROKEAHA.121.032622
3. Ma Y, Yang S, He Q, Zhang D, Chang J. The Role of Immune Cells in Post-Stroke Angiogenesis and Neuronal Remodeling: The Known and the Unknown. Front Immunol. 2021;12:784098. doi: 10.3389/fimmu.2021.784098
4. Finger CE, Moreno-Gonzalez I, Gutierrez A, Moruno-Manchon JF, McCullough LD. Age-related immune alterations and cerebrovascular inflammation. Mol Psychiatry. 2022;27:803-818. doi: 10.1038/s41380-021-01361-1
5. Mola-Caminal M, Carrera C, Soriano-Tarraga C, Giralt-Steinhauer E, Diaz-Navarro RM, Tur S, Jimenez C, Medina-Dols A, Cullell N, Torres-Aguila NP, et al. PATJ Low Frequency Variants Are Associated With Worse Ischemic Stroke Functional Outcome. Circ Res. 2019;124:114-120. doi: 10.1161/CIRCRESAHA.118.313533
6. Soderholm M, Pedersen A, Lorentzen E, Stanne TM, Bevan S, Olsson M, Cole JW, Fernandez-Cadenas I, Hankey GJ, Jimenez-Conde J, et al. Genome-wide association meta-analysis of functional outcome after

ischemic stroke. Neurology. 2019;92:e1271-e1283. doi: 10.1212/WNL.0000000000007138

7. Pfeiffer D, Chen B, Schlicht K, Ginsbach P, Abboud S, Bersano A, Bevan S, Brandt T, Caso V, Debette S, et al. Genetic Imbalance Is Associated With Functional Outcome After Ischemic Stroke. Stroke. 2019;50:298-304. doi: 10.1161/STROKEAHA.118.021856

8. Ibanez L, Heitsch L, Carrera C, Farias FHG, Del Aguila JL, Dhar R, Budde J, Bergmann K, Bradley J, Harari O, et al. Multi-ancestry GWAS reveals excitotoxicity associated with outcome after ischaemic stroke. Brain. 2022;145:2394-2406. doi: 10.1093/brain/awac080

9. Medina-Dols A, Canellas G, Capo T, Sole M, Mola-Caminal M, Cullell N, Jaume M, Nadal-Salas L, Llinas J, Gomez L, et al. Role of PATJ in stroke prognosis by modulating endothelial to mesenchymal transition through the Hippo/Notch/PI3K axis. Cell Death Discov. 2024;10:85. doi: 10.1038/s41420-024-01857-z

10. Giese KP, Mizuno K. The roles of protein kinases in learning and memory. Learn Mem. 2013;20:540-552. doi: 10.1101/lm.028449.112

11. Shioda N, Fukunaga K. Physiological and Pathological Roles of CaMKII-PP1 Signaling in the Brain. Int J Mol Sci. 2017;19. doi: 10.3390/ijms19010020

12. Heroes E, Lesage B, Gornemann J, Beullens M, Van Meervelt L, Bollen M. The PP1 binding code: a molecular-lego strategy that governs specificity. FEBS J. 2013;280:584-595. doi: 10.1111/j.1742-4658.2012.08547.x

13. Heitsch L, Ibanez L, Carrera C, Binkley MM, Strbian D, Tatlisumak T, Bustamante A, Ribo M, Molina C, Davalos A, et al. Early Neurological Change After Ischemic Stroke Is Associated With 90-Day Outcome. Stroke. 2021;52:132-141. doi: 10.1161/STROKEAHA.119.028687
14. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337:64-69. doi: 10.1126/science.1219240
15. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337:100-104. doi: 10.1126/science.1217876
16. Wainschtein P, Jain D, Zheng Z, Group TOAW, Consortium NT-OfPM, Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, Mitchell BD, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat Genet. 2022;54:263-273. doi: 10.1038/s41588-021-00997-7
17. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82-93. doi: 10.1016/j.ajhg.2011.05.029
18. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13:762-775. doi: 10.1093/biostatistics/kxs014
19. Susak H, Serra-Saurina L, Demidov G, Rabionet R, Domenech L, Bosio M, Muyas F, Estivill X, Escaramis G, Ossowski S. Efficient and flexible

570 Integration of variant characteristics in rare variant association studies
571 using integrated nested Laplace approximation. PLoS Comput Biol.
572 2021;17:e1007784. doi: 10.1371/journal.pcbi.1007784
573 20. Roquer J, Rodriguez-Campello A, Gomis M, Jimenez-Conde J, Cuadrado-
574 Godia E, Vivanco R, Giralt E, Sepulveda M, Pont-Sunyer C, Cucurella G,
575 et al. Acute stroke unit care and early neurological deterioration in
576 ischemic stroke. J Neurol. 2008;255:1012-1017. doi: 10.1007/s00415-
577 008-0820-z
578 21. Fernandez-Cadenas I, Mendioroz M, Giralt D, Nafria C, Garcia E, Carrera
579 C, Gallego-Fabrega C, Domingues-Montanari S, Delgado P, Ribo M, et al.
580 GRECOS Project (Genotyping Recurrence Risk of Stroke): The Use of
581 Genetics to Predict the Vascular Recurrence After Stroke. Stroke.
582 2017;48:1147-1153. doi: 10.1161/STROKEAHA.116.014322
583 22. Carrera C, Carcel-Marquez J, Cullell N, Torres-Aguila N, Muino E, Castillo
584 J, Sobrino T, Campos F, Rodriguez-Castro E, Llucia-Carol L, et al. Single
585 nucleotide variations in ZBTB46 are associated with post-thrombolytic
586 parenchymal haematoma. Brain. 2021;144:2416-2426. doi:
587 10.1093/brain/awab090
588 23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD:
589 predicting the deleteriousness of variants throughout the human genome.
590 Nucleic Acids Res. 2019;47:D886-D894. doi: 10.1093/nar/gky1016
591 24. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The
592 FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382-
593 388. doi: 10.1093/nar/gki387

25. Bruque CD, Delea M, Fernandez CS, Orza JV, Taboas M, Buzzalino N, Espeche LD, Solari A, Luccerini V, Alba L, et al. Structure-based activity prediction of CYP21A2 stability variants: A survey of available gene variations. Sci Rep. 2016;6:39082. doi: 10.1038/srep39082

26. Buonfiglio PI, Bruque CD, Lotersztein V, Luce L, Giliberto F, Menazzi S, Francipane L, Paoli B, Goldschmidt E, Elgoyhen AB, et al. Predicting pathogenicity for novel hearing loss mutations based on genetic and protein structure approaches. Sci Rep. 2022;12:301. doi: 10.1038/s41598-021-04081-2

27. Galland F, Malergue F, Bazin H, Mattei MG, Aurrand-Lions M, Theillet C, Naquet P. Two human genes related to murine vanin-1 are located on the long arm of human chromosome 6. Genomics. 1998;53:203-213. doi: 10.1006/geno.1998.5481

28. Suzuki K, Watanabe T, Sakurai S, Ohtake K, Kinoshita T, Araki A, Fujita T, Takei H, Takeda Y, Sato Y, et al. A novel glycosylphosphatidyl inositol-anchored protein on human leukocytes: a possible role for regulation of neutrophil adherence and migration. J Immunol. 1999;162:4277-4284.

29. Watanabe T, Sendo F. Physical association of beta 2 integrin with GPI-80, a novel glycosylphosphatidylinositol-anchored protein with potential for regulating adhesion and migration. Biochem Biophys Res Commun. 2002;294:692-694. doi: 10.1016/S0006-291X(02)00538-7

30. Yoshitake H, Takeda Y, Nitto T, Sendo F. Cross-linking of GPI-80, a possible regulatory molecule of cell adhesion, induces up-regulation of CD11b/CD18 expression on neutrophil surfaces and shedding of L-selectin. J Leukoc Biol. 2002;71:205-211.

31. Martin F, Malergue F, Pitari G, Philippe JM, Philips S, Chabret C, Granjeaud S, Mattei MG, Mungall AJ, Naquet P, et al. Vanin genes are clustered (human 6q22-24 and mouse 10A2B1) and encode isoforms of pantetheinase ectoenzymes. Immunogenetics. 2001;53:296-306. doi: 10.1007/s002510100327

32. Nitto T, Onodera K. Linkage between coenzyme a metabolism and inflammation: roles of pantetheinase. J Pharmacol Sci. 2013;123:1-8. doi: 10.1254/jphs.13r01cp

33. Bui TA, Jickling GC, Winship IR. Neutrophil dynamics and inflammaging in acute ischemic stroke: A transcriptomic review. Front Aging Neurosci. 2022;14:1041333. doi: 10.3389/fnagi.2022.1041333

34. Keuters MH, Keksa-Goldsteine V, Rolova T, Jaronen M, Kettunen P, Halkoluoto A, Goldsteins G, Koistinaho J, Dhungana H. Benserazide is neuroprotective and improves functional recovery after experimental ischemic stroke by altering the immune response. Sci Rep. 2024;14:17949. doi: 10.1038/s41598-024-68986-4

35. Jiang Y, Wang N, Liu J, Ren H, Jiang W, Lei Y, Fu X, Hao M, Lang X, Liu Y, et al. Intranasal delivery of hMSC-derived supernatant for treatment of ischemic stroke by inhibiting the pro-inflammatory polarization of neutrophils. Stem Cell Res Ther. 2025;16:43. doi: 10.1186/s13287-025-04172-1

36. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-589. doi: 10.1038/s41586-021-03819-2

644 37. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng
645 EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory
646 research and analysis. J Comput Chem. 2004;25:1605-1612. doi:
647 10.1002/jcc.20084
648 38. Shapovalov MV, Dunbrack RL, Jr. A smoothed backbone-dependent
649 rotamer library for proteins derived from adaptive kernel density estimates
650 and regressions. Structure. 2011;19:844-858. doi:
651 10.1016/j.str.2011.03.019
652 39. Kolomenski JE, Delea M, Simonetti L, Fabbro MC, Espeche LD, Taboas
653 M, Nadra AD, Bruque CD, Dain L. An update on genetic variants of the
654 NKX2-5. Hum Mutat. 2020;41:1187-1208. doi: 10.1002/humu.24030
655 40. Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A. The TOPCONS web
656 server for consensus prediction of membrane protein topology and signal
657 peptides. Nucleic Acids Res. 2015;43:W401-407. doi: 10.1093/nar/gkv485

12. Tables

Table 1. Characteristics of the pilot phase samples.

mRS (3 months)	Sex (female/ male)	Age (range)	Age (mean)	LAA	CE	OD	Undetermined	Incomplete
0/1	24/25	53-87	73.26	12	22	0	13	2
3/4/5	20/21	51-90	74.4	11	20	0	7	3

LAA, large-artery atherosclerosis; CE, cardioembolism; OD, other determined; Undetermined, two or more causes; Incomplete, incomplete evaluation.

663 **Table 2. Characteristics of the samples from the follow-up cohort**.

Characteristics	mRs at 3 months						
	0	1	2	3	4	5	6
Sex (female/male)	40/46	61/69	57/69	63/58	49/43	19/5	54/44
Age (range)	28-92	39-94	40-98	26-100	36-93	41-93	31-99
Age (mean)	70	72	74	76	77	77	81
LAA	13	23	29	23	18	4	9
CE	42	65	73	70	48	15	70
OD	18	31	19	17	11	2	8
Undetermined	3	4	2	5	9	1	10
Incomplete	10	7	3	6	6	2	1
Total	86	130	126	121	92	24	98

664 LAA, large-artery atherosclerosis; CE, cardioembolism; OD, other determined;
665 Undetermined, two or more causes; Incomplete, incomplete evaluation; NA, not
666 available.

667

Table 3. Top genes in the rare variant association study (dichotomous model) according to BATI. The results were performed in a dichotomous model using mRs at 3 months 0-1 vs 3-6.

Gene	Total variants	Variants in poor	Variants in good	Carriers in poor	Carriers in good	△DIC
VNN2	6	1	5	1	6	10.773*
FGD4	8	2	6	2	6	2.512
NEK10	9	7	2	10	2	2.409
RASAL1	6	6	4	6	6	2.371
ECE2;EEF1AK MT4-ECE2	11	4	7	4	6	2.264

*$\Delta DIC_{default} > 10$, $\Delta DIC_{0.001} > 10.3$ equivalent to *p* value < 0.001.

Table 4. VNN2 protein evaluation. CADD 1.3 [30]

Protein Variation	SNPs	Domain	mRs 3 months	Protein Effect	Protein Stability ΔΔG values	CADD (v 1.3)
p.(S46F)	rs139348170	CN hydrolase	0	Stability	1.99 ± 0.09	23.6
p.(L53P)	rs74854525	CN hydrolase	4	Stability	3.65 ± 0.03	25.8
p.(V174M)	rs142626186	CN hydrolase	1	Stability	1.7 ± 0.27	27
p.(R205S)	rs148308586	CN hydrolase	0	Stability	2.72 ± 0.09	22.7
p.(A265T)	-	CN hydrolase	1	Stability	5.9 ± 0.09	25
p.(G284C)	-	CN hydrolase	0	Stability	1.96 ± 0.30	23
p.(R393Q)	rs34856068	Vanin C	0, 1	Electrostatic Surface	-0.18 ± 0.06	20.8
p.(R393*)	rs34492437	Vanin C	0	NMD; GPI Anchor	-	34

13. Figure Legends

Figure 1. Lineal representation of protein domains of VNN2 and location of rare variants identified by WES and Targeted NGS analysis.