Universitat Autònoma de Barcelona

STUDY OF GENETIC RISK FACTORS ASSOCIATED WITH ISCHEMIC STROKE

Jara Cárcel Márquez

Universitat Autònoma de Barcelona

Doctoral Thesis

STUDY OF GENETIC RISK FACTORS ASSOCIATED WITH ISCHEMIC STROKE

Author: Jara Cárcel Márquez

Director: Israel Fernández Cadenas

Tutor: Joan Martí Fàbregas

Doctoral Program in Medicine

Departament of Medicine

Universitat Autònoma de Barcelona

Barcelona, 2022

List of abbreviations

A

AF: Atrial Fibrillation

AUC: Area Under the roc Curve

AUPRC: Area Under the Precision-Recall C

B

B: Beta

BP: Base Pairs

C

caQTL: chromatin accessibility Quantitative Trait Loci

CADASIL: Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy

CARASIL: Cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy

CCS: Causative Classification of Stroke System

CE: Cardioembolic Stroke

CHR: Chromosome

CI: Confidence Interval

CONIC: CONtrol ICtus study

CT: Computerized Tomography

D

DNA: Deoxyribonucleic acid

E

EA: Effect Allele

ESUS: Embolic Stroke of Undetermined Source

eQTL: expression Quantitative Trait Loci

F

FDR: False Discovery Rate

G

GENESIS: Genetics of Early Neurological Instability After Ischemic Stroke

GenoTPA: Genetic Study in Ischemic Stroke Patients treated with tPA

GO: Gene Ontology

GODS: Genetic contribution to Functional Outcome and Disability after Stroke

GRECOS: Genotyping Recurrence Risk of Stroke

GSEA: Gene-Set Enrichment Analysis

GWAS: Genome-Wide Association Studies

GWASseq: sequencing-based GWAS

H

haQTL: histone acetylation Quantitative Trait Loci

Hg19: Human Genome version 19

Hi-C: assay that investigates the 3D organization of the genome

I

IDI: Integrated Discrimination Index

IGF: Insulin-like Growth Factor

IL-5: interleukin 5

ipaQTL: intronic alternative polyadenylation Quantitative Trait Loci

IS: Ischemic Stroke

ISSYS: Investigating Silent Stroke in hYpertensives: A magnetic resonance imaging Study

IVW: Inverse Variance Weighted

L

LAA: Large-Artery Atherosclerosis

M

MAF: Minor Allele Frequency

MCA: Middle Cerebral Artery

meQTL: methylation Quantitative Trait Loci

MMP: Matrix Metalloproteinase

MR: Mendelian Randomization

mRS: modified Rankin Scale

MTAG: Multitrait Analysis of GWAS

N

N: sample size

NIHSS: National Institutes of Health Stroke Scale

NGS: Next Generation Sequencing

NRI: Net Reclassification Index

O

OA: Other Allele

OR: Odds Ratio

P

PCA: Principal Component Analysis

pQTL: protein Quantitative Trait Loci

P: P-value

PRS: Polygenic Risk Score

PADMAL: pontine autosomal dominant microangiopathy with leukoencephalopathy

PWM: Penalized Weighted Median

Q

QTL: Quantitative Trait Loci

QQ: Quantile-Quantile

S

SE: Standard Error

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variant

sQTL: splicing Quantitative Trait Loci

SVO: Small Vessel Occlusion

T

TCD: Transcranial Doppler ultrasonography

TOAST: Trial of ORG 10172 in Acute Stroke Treatment

TSS: Transcription Start Site

W

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

WM: Weighted Median

WMode: Weighted Mode

Z

Z: Z-score

Others

3D: Tridimensional

λ: Genomic inflation factor

Table of contents

Summary

This thesis widens our knowledge of key aspects of genomics approaches that can be implemented in the study of ischemic stroke. The main objective of the present thesis dissertation is to dig into the genetic risk factors associated with ischemic stroke and use them to find novel associations. To achieve this, genome-wide association studies (GWAS) of ischemic stroke risk and subtypes in the Spanish population, multitrait analysis of GWAS of cardioembolism (CE) and atrial fibrillation (AF), development of a clinical polygenic risk score to predict cardioembolic stroke and perform mendelian randomization analyses to evaluate causality between some metalloproteinase levels in plasma and risk of ischemic stroke and poor functional outcome post-stroke.

Data obtained from this thesis revealed the genomic of ischemic stroke risk and subtypes in Spain (GENERACION cohort) for the first time. Two loci have been replicated in an international cohort, locus 5p15.2 was found significantly associated with small vessel occlusion (SVO), and locus 8p11.22 was found significantly associated with females' SVO. Pathways' analyses revealed the involvement of different interleukins, as well as amyloid clearance of the brain as potential players of this genetic susceptibility.

Findings on the multitrait study of CE and AF revealed different novel loci associated with CE that were replicated in the GENERACION cohort. Also, it revealed several genomic loci that show an association with AF independent of CE that could be used to understand why some AF patients do not develop CE. Additionally, the implementation of a polygenic risk score proved to be associated with CE in the GENERACION cohort and independent of age, sex, and hypertension.

Lastly, using a mendelian randomization approach MMP-12 lower plasma levels where found causally associated with ischemic stroke risk. MMP-1 and MMP-12 lower plasma levels were suggested to play a causal role in the risk of large-artery atherosclerosis risk, and higher serum levels of MMP-8 and the risk of small vessel occlusion. Suggesting the importance of targeting these proteins to prevent stroke occurrence.

Resumen

Esta tesis amplía nuestro conocimiento de los aspectos clave de los enfoques genómicos que pueden aplicarse en el estudio del ictus isquémico. El objetivo principal de la presente tesis doctoral es profundizar en los factores de riesgo genéticos asociados al ictus isquémico y usarlos para encontrar nuevas asociaciones. Para ello, se han utilizado estudios de asociación de todo el genoma de riesgo ictus isquémico y subtipos en la población española, análisis multitrait de GWAS de cardioembolismo (CE) y fibrilación auricular (FA), desarrollo de una puntuación de riesgo clínico-poligénico para predecir el ictus cardioembólico y realizar análisis de aleatorización mendeliana para evaluar la causalidad entre los niveles plasmáticos de ciertas metaloproteinasas y el riesgo de ictus isquémico y el estado funcional post-ictus.

Los datos obtenidos en esta tesis revelaron por primera vez la genómica del riesgo de ictus isquémico y sus subtipos en España (cohorte GENERACION). Dos loci han sido replicados en una cohorte internacional, el locus 5p15.2 se encontró significativamente asociado a ictus lacunar, y el locus 8p11.22 se asociado a ictus lacunar en mujeres. Los análisis de las vías revelaron la implicación de diferentes interleucinas, así como del aclaramiento de amiloide del cerebro como posibles protagonistas de esta susceptibilidad genética.

Los resultados del estudio multifenotipo de ictus cardioembólico y fibrilación auricular revelaron diferentes loci novedosos asociados a CE que se replicaron en la cohorte GENERACION. Asimismo, se revelaron varios loci que muestran una asociación con FA independiente del riesgo de CE y que podrían utilizarse para entender por qué algunos pacientes con FA no desarrollan CE. Además, la aplicación de una puntuación de riesgo poligénica demostró estar asociada con CE en la cohorte GENERACION e independiente de edad, sexo e hipertensión.

Por último, utilizando un enfoque de aleatorización mendeliana los niveles plasmáticos más bajos de MMP-12 se asociaron causalmente con el riesgo de ictus isquémico. Se sugirió que los niveles plasmáticos más bajos de MMP-1 y MMP-12 desempeñan un papel causal en el riesgo de ictus de gran isquemia, y los niveles séricos más altos de MMP-8 y el riesgo de ictus lacunar. Lo que sugiere la importancia de dirigirse a estas proteínas para prevenir la aparición de ictus.

1. Introduction

1. Introduction

1.1 Stroke: A multifactorial disease

Stroke is a heterogeneous group of disorders characterized by a permanent or transient neurological deficit caused by a sudden and focal interruption of blood flow to the brain (1). When this occurs, the brain region that does not receive blood is under stress by the lack of oxygen and nutrients. If the interruption is prolonged over time, the affected brain cells may be damaged or die.

1.1.1 Epidemiology

Regarding the epidemiology of stroke incidence, for the first time in documented history, a negative trend is observed in a recent metanalysis of 28 studies (2). This optimistic result is probably due to the efforts and advances in primary and secondary prevention of the scientific community (2). Despite this, numbers are still high, only in Spain every year there are more than 71,000 incident strokes. Besides, only in 2017, almost 17,000 deaths in Spain were caused by a stroke. Being the second cause of mortality in Spain, first in women and third in men. It is also important to mention that two of every three people who survive a stroke have some type of sequelae, in many cases disabling (3).

1.1.2. Classification

Given the cause of the stroke, two main categories are described (Figure 1):

- Hemorrhagic stroke, in which the interruption of the blood flow is due to the rupture of a blood vessel, causing hemorrhage into the parenchymal tissue. This category represents approximately 13% of all stroke cases.
- Ischemic stroke, in which there is an obstruction of the blood flow to the brain by a blood clot. This subgroup represents most of the cases, approximately 87% (4). The present thesis dissertation is focused on this type of stroke.

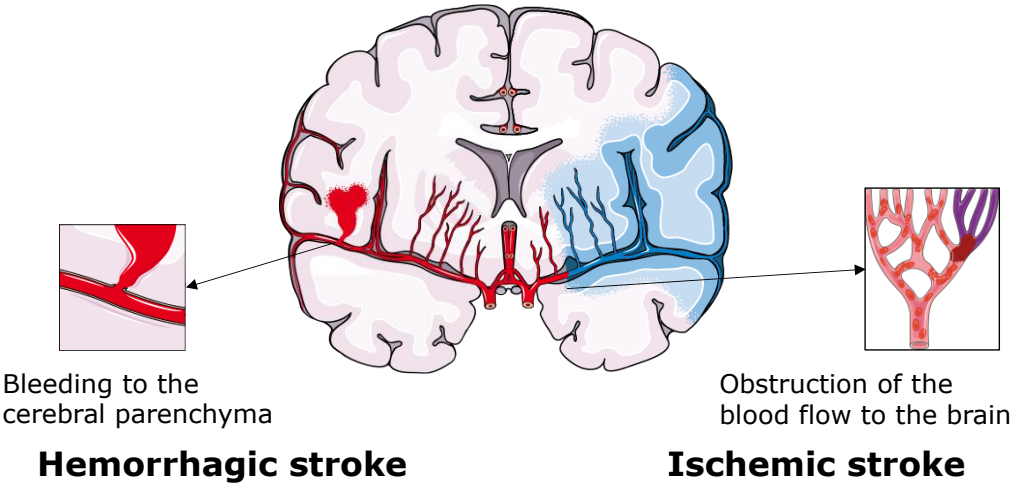


FIGURE 1. Representation of stroke classification.

Different classifications have been done about ischemic stroke etiologies or subtypes, the more extended one is the Trial of ORG 10172 in Acute Stroke Treatment (TOAST), with five different subcategories: large-artery atherosclerosis (LAA), cardioembolism (CE), small-vessel occlusion (SVO), other of determined causes (OE) and strokes of undetermined etiology (UE) (Table 1), based on clinical features and different evaluations such as brain and cardiac imaging, duplex imaging extracranial arteries, laboratory results for the prothrombotic state, and arteriography (5).

Later, the Causative Classification of Stroke System (CCS) classification was created, which is a computerized, evidence-based algorithm that is especially interesting because adds the degree of uncertainty, classifying it as evident, probable, and possible. It also discerns between causative and phenotypic etiologies of stroke (6).

TABLE 1. TOAST classification of ischemic stroke subtypes (5)

Subtypes of stroke in TOAST Classification
Large-artery atherosclerosis
Cardioembolism
Small-vessel occlusion (lacune)
Stroke of other determined etiology
Stroke of undetermined etiology a. Two or more causes identified b. Negative evaluation c. Incomplete evaluation

Large-artery atherosclerosis stroke etiology correspond to patients with confirmation of stenosis of more than 50% or occlusion of a major brain artery, presumably due to an atheromatous plaque. **Cardioembolism** or cardioembolic strokes are those in which the embolus causing the occlusion of blood flow to the brain has been formed inside of the cardiac chambers, and a potential cardiac source must be confirmed. **Small-vessel occlusion** stroke includes patients with lacune infarcts, defined lesions of less than 1.5 cm of diameter, in the subcortical hemispheric due to occlusion of perforating arterioles. In the **other determined etiology** classification, there are less frequent causes of strokes that are differentiated from LAA, CE, and SVO. Examples of these causes are Moya-Moya disease and carotid artery dissection. Finally, in the **undetermined category**, all the patients do not have a defined category, due to the coexistence of two or more potential causes, a negative evaluation of all the evaluated causes, or an incomplete study because of *exitus*, capability of the hospital or consent of the patient.

Additionally, a different entity was defined in 2014, the embolic stroke of undetermined source (ESUS) (7). This category aimed to differentiate a subgroup of patients inside of the undetermined TOAST category that is relevant from the therapeutical point of view. It corresponds to approximately 17% of ischemic stroke patients. They are defined as those non-lacunar brain infarcts without proximal arterial stenosis or cardioembolic sources (7).

1.1.3 Risk factors

Several risk factors have been described to be associated with stroke incidence. The identification and study of these are especially difficult by the fact that stroke is composed of different entities (8). Risk factors for ischemic and hemorrhagic stroke are very similar, but there are important differences. Hypertension is especially important concerning hemorrhagic stroke but, it is also a risk factor for atherosclerosis and therefore it can also lead to ischemic stroke (8).

Risk factors can be divided into modifiable and non-modifiable factors (Table 2). Non-modifiable risk factors include age, sex, ethnicity, and genetics. Stroke has always been considered a disease of aging; it is estimated that every decade after the age of 55, the risk of suffering a stroke doubles. (9). Regarding the relation between sex and stroke, globally more women suffer from stroke probably due to the longer life expectancy compared to men. But this relation strongly depends on age, at younger ages women are at higher risk of suffering a stroke, likely due to a relation with stages like pregnancy and post-partum. Although, at more advanced ages, the relative risk is moderately higher for men. Regarding age and sex interaction, a less stablished risk factor is biological age evaluated from DNA methylation in blood. This variable has been linked to ischemic stroke risk and to sex-differences between patients(10).

Several associations between different racial ethnicities and stroke have been described. African Americans have twice the risk of stroke compared to whites, in addition, Hispanic-Latinos also have a higher risk of stroke in some cohorts. These associations are thought to be an intersection of genetics and differences in socioeconomic status (11).

Genetics has always been considered a non-modifiable risk factor although nowadays it is starting to be considered at the intersection of modifiable and non-modifiable, due to advances in the field of genetic engineering as well as the indirect modification of our genetics through the alteration of gene-environment interactions (11).

Among modifiable risk factors are well-established risk factors like hypertension, with a lineal, positive relation between blood pressure and the risk of suffering a stroke (12). Other risk factors include diabetes, hyperlipidemia, cardiac sources,

and different proteins suggested as risk factors such as apolipoprotein B to A1 levels in the blood (8). As mentioned above, risk factors contribute differently to stroke etiologies, in the case of cardiac sources such as atrial fibrillation, this is a pervasive factor for cardioembolic stroke, and defines the reason for classifying a stroke as cardioembolic.

Other less established modifiable risk factors include other proteins such as matrix metalloproteinases (MMPs). MMPs are a diverse group of endopeptidases. They are known to mediate the degradation or remodeling of the extracellular matrix. Observational studies have found a correlation between MMP levels and the risk of atheromatous plaque instability and ischemic stroke (13). Several studies have also suggested that MMPs may play a key role in the outcome post-stroke (14).

TABLE 2. Risk factors of stroke risk (12).

	Non-modifiable	**Modifiable**
Ischemic Stroke	Age Sex Race/Ethnicity Genetics	Hypertension Current Smoking Waist-to-hip ratio Diet Physical Inactivity Hyperlipidemia Diabetes Alcohol Consumption Cardiac Causes Apolipoprotein B to A1
Hemorrhagic Stroke	Age Sex Race/Ethnicity Genetics	Hypertension Current Smoking Waist-to-hip ratio Alcohol Consumption Diet

1.1.4 Treatment

Concerning the treatment of patients, we need to consider primary prevention, acute phase and secondary prevention.

Primary prevention involve patients at high risk of disease debut, this is evaluated by targeting well stablished risk factors such as diabetes, obesity, atrial fibrillation, hypertension, lifestyle and diet. Therefore, in the last guidelines the recommendation involve medication to control blood pressure and lipids, anticoagulants drugs in patients with confirmed atrial fibrillation, stop cigarette smoking, healthier diet, exercise among other interventions. (15).

When in the acute phase of ischemic stroke, there are two main options: thrombolytic therapy or endovascular treatment. Thrombolytic treatment includes drugs such as alteplase, which is a thrombolytic drug that targets the fibrin fibers that are present in clots. If this treatment does not work properly and the blocked artery is a large vessel, the patient would be a candidate for mechanical thrombectomy. This consists of the direct removal of the clot that is disrupting the blood flow with a stent retriever (16). Alternative treatments are currently under study with successful results in clinical trials like Tenecteplase a more attractive thrombolytic agent because of the simplicity of administration in a single bolus(17), staphylokinase, a novel thrombolytic obtained by protein engineering (18), and even a not pharmacologic treatment like remote ischemic conditioning (19).

Regarding the secondary prevention of stroke, it is very important to identify the precise cause of the first event because this is going to guide the clinical practice. Nowadays, patients are prescribed antiplatelet or anticoagulant drugs. Both treatments target clot formation but affect different clot etiologies and therefore alter different pathways. Antiplatelet agents have strong clinical evidence supporting their use for the prevention of stroke of non-cardioembolic etiology; in contrast, anticoagulants are strongly recommended for the prevention of cardioembolic stroke (20).

The gold standard for stroke care for each patient would be to individualize treatment, based on clinical, genetic, and other characteristics to establish

personalized medicine that minimizes the risk of adverse effects and enables patients to have a better functional status after stroke.

1.1.5 Short- and long-term outcome

To evaluate the neurological state of the patient at hospital admission and during follow-up, the National Institutes of Health Stroke Scale (NIHSS) is widely used. NIHSS consists of a scale that evaluates different domains, such as eye movement, level of consciousness, muscle strength, language, and coordination, among others. The NIHSS is a reliable, valid, and sensitive tool for measuring stroke severity; it is useful in both clinical practice and research (21). Interestingly, changes in short-term outcome evaluated with NIHSS during 24h post-stroke have been found associated with long-term outcome, suggesting that the first hours after stroke are crutial (22).

For the long-term outcome post-stroke, the focus is to evaluate the global disability of the patient. In clinical practice, this is performed using the modified Rankin Scale (mRS) which is a clinician-reported measure of global disability. It consists of 7 levels ranging from no symptoms to death. It is usually evaluated 3 months after the event. This scale is broadly used in clinical trials to evaluate the efficacy of treatments (23).

1.2. Genomics: a tool to understand disease

Genomics is defined as the study of all the genes of an organism, as well as their regulation of them. It was in 1977 with the discovery of the first genotyping technique by Frederick Sanger (24) when the genomics era was about to start. It encouraged the development of the Human Genome Project in 1990 (25). A project that sought to unravel the genome of humankind, by analyzing a composite of different individuals. The project was completed in 2003 and confirmed that humans have about 25,000 genes (26).

1.2.1 Genetic variation

Genetic variation is defined as the difference in the DNA sequence among individuals or populations (27). This diversity is key to making possible evolution by natural selection. Although only variation that arises in germ cells can be inherited from one individual to another and so affect population dynamics (28). The major sources of genetic variation are given by errors during DNA replication, known as mutations, and recombination between paternal and maternal chromosomes during meiosis (29).

Genetic variations range from 1 nucleotide to megabases of them or even can involve structural changes. Different types of genetic variations have been defined, although there is not a final classification of them, they can be grouped into (Figure 2):

- **Single nucleotide changes**.
    - *Single Nucleotide Polymorphism or Single Nucleotide Substitution*. It has been a substitution of a single nucleotide. Single nucleotide polymorphisms are those that are present in a population at a frequency higher than 1%, if not, they are considered mutations (27). Strictly, variants, where population frequency is yet to be confirmed in population-based studies, should be labeled as Single Nucleotide Variant (SNV) to avoid confusion (30).
    - *Single Nucleotide Insertion*. A nucleotide is inserted into the reference DNA sequence.
    - *Single Nucleotide Deletion*. A nucleotide is removed from the reference DNA sequence.

- **Tandem repeats**.
  - *Short tandem repeats*. Usually repetitions of 8 nucleotides or less.
  - *A variable number of tandem repeats*. Repetitions of sequences of more than 8 nucleotides.
- **Indels**
  - *Short indels*. Insertion/deletion of a 6-nucleotide sequence or less.
  - *Indels***.** Insertion/deletion of a sequence ranging from 100 bp to 1 kb.
- **Structural variations**.
  - *Copy number variation*. Deletion or duplication of a DNA segment larger than 1kb.
  - *Copy neutral variation***.** These are also considered rearrangements of the genome sequence. Inversions involved a flip of the sequence and translocation an exchange of sequence between chromosomes (27).

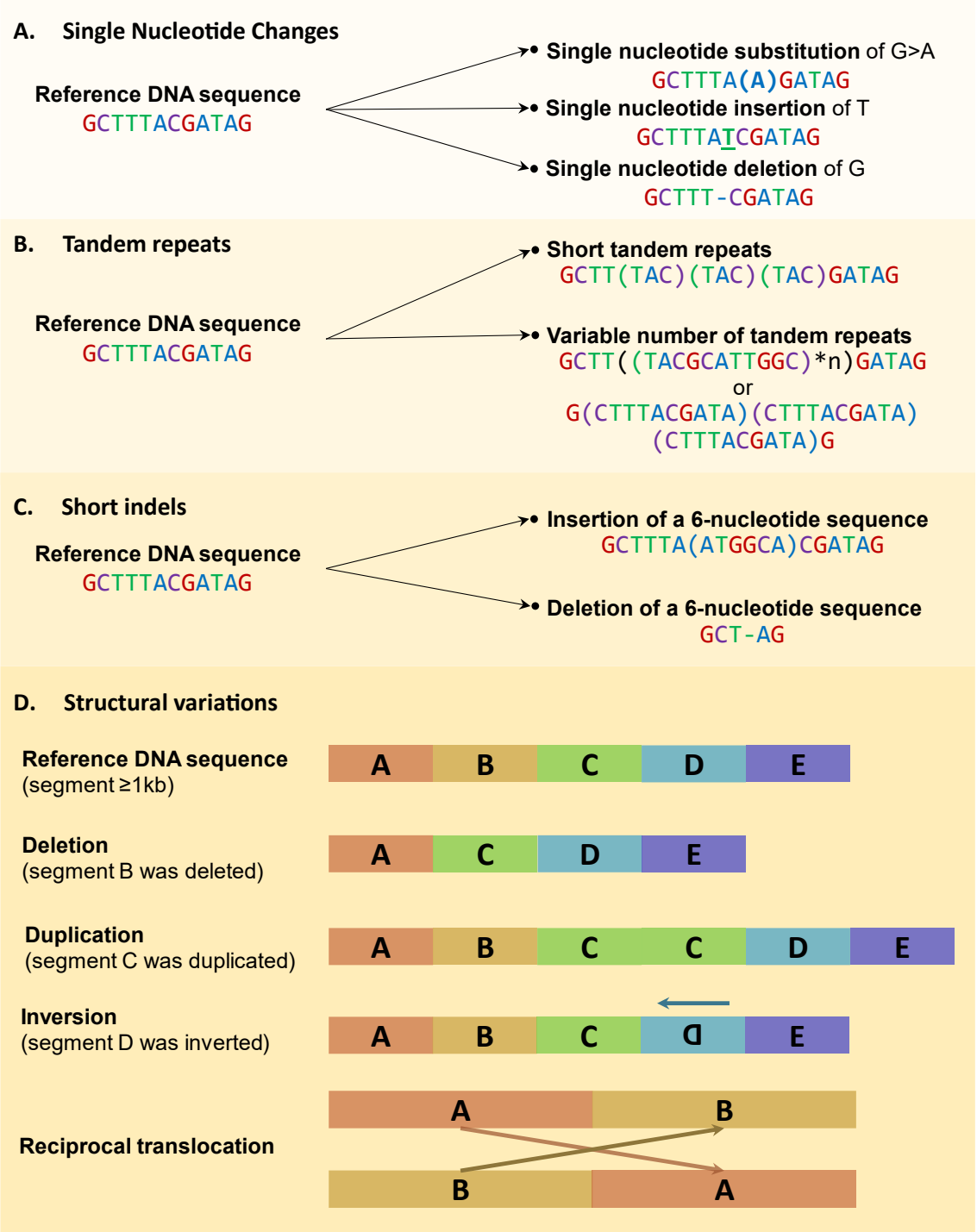


FIGURE 2. Schematic illustration of genetic variants. Based on figure 1 of Ku *et al*, 2010 (27).

1.2.2 Single Nucleotide Polymorphism

Single Nucleotide Polymorphisms (SNP) are the most common type of genetic variation, 84.7 million of them have been described in the 1000 genomes project (31). Given their position SNPs are classified as intergenic if the SNPs does not locate in any gene sequence. Among the SNPs that do locate in gene regions, they can be intronic or exonic if they are in the region that will be translated into

protein. Regarding the consequence of an SNP, they can be classified as synonymous if the protein sequence is not altered or nonsynonymous if it is altered. This alteration can be in form of an amino acid substitution (missense) or the shortening of the protein due to the apparition of an early stop codon (nonsense) (32). Other consequences of SNPs include the modulation of quantitative factors, known as quantitative trait loci (QTL). Multiple types are described: expression QTL (eQTL(33)), methylation QTL (meQTL(34)), chromatin accessibility QTL (caQTL(35)), protein QTL (pQTL (36)), intronic alternative polyadenylation QTL (ipaQTL(37)), histone acetylation (haQTL(38)), splicing QTL (sQTL(39)), and probably more would be described in the near future as far as technologies are evolving. The description of these is essential because they can fill the gap of a causal relationship between SNPs and disease risk.

1.2.3 Genome-Wide Association Studies

Genome-wide association study (GWAS) is a statistical analysis in which hundreds of thousands or even millions of genetic variants across the genome are tested to find those statistically associated with a specific disease or trait. The main strength of the GWAS strategy is that it is an unbiased tool for finding genetic susceptibility loci, as it does not start from any prior assumptions about the genes involved in the disease or trait (40).

The basis behind GWAS is in general an array genotyping technique in which hundreds of thousands or millions of SNPs are addressed. In the Illumina® genotyping Bead Array technology, the DNA sample extracted from patient tissue (e.g., blood, saliva…) in introduced to a BeadChip. In this chip, there are different probes for each variant to be analyzed. In each probe, the sample is bonded to a complementary DNA sequence, but the DNA probe lacks a position, the locus of interest. After that, labeled nucleotides are added so one of the four will be bonded to the sequence. Finally, a laser is used to evaluate the nucleotide signal and translate this intensity into allelic information for the genetic position (41).

In addition, an essential part of GWAS analysis is the imputation of SNPs. Imputation is commonly done in GWAS analysis and consists of increasing the number of variants that will be analyzed by inferring non-genotyped variants thanks to haplotypes, and using a representative reference panel. The quality of

these panels has increased drastically in recent years, in 2016 the Haplotype Reference Consortium (HRC) released a combined reference panel of 64,976 haplotypes of primarily European ancestry. HRC panel allows us to perform accurate imputation of variants with a minor allele frequency (MAF) of 0.1% (42).

An alternative to classical GWAS is sequencing-based GWASs (GWASseq) (43). These are based on next-generation sequencing (NGS), this technology consists of the massive analysis of DNA sequences employing multiplexed parallel analysis. For this, DNA sequences are fragmented to construct a sequencing library that will be genotyped. Using this technology two main analyses are possible, whole exome sequencing (WES) analysis and whole genome sequencing (WGS), the difference is that in the first one, the DNA analyzed is mainly limited to the coding sequence of the genome (43). This technology has the advantage of deep coverage of the genome, but it is highly expensive compared with array-based genotyping techniques. In 2019, the cost of NGS was 30 times higher than that of genotyping arrays, and the added costs to perform GWASseq, such as those required for processing and data storage, research staff, and computational infrastructure, are still very high (40,44).

GWAS summary statistics or individual-level data have made possible not only the identification of variants associated with disease but also the determination of population substructure (45), estimation of SNP heritability of complex traits (46), the estimate of the genetic correlation between traits (47), polygenic risk scores (48), Mendelian randomization studies (49), among different applications (40).

1.2.4 Multi-trait analysis

The standard approach in genome-wide association studies is to analyze one trait at a time. Although this is the ideal strategy, it is highly dependent on sample size and is not informative about the possible pleiotropic loci with related phenotypes. A solution to this is to perform multi- or cross-trait analysis. This approach will allow us to increase the sample size by using genetically correlated traits, thereby enhancing the statistical power to detect new signals. This strategy is particularly interesting in the context of a complex diseases, where multiple intermediate phenotypes may be playing an important role.

In recent years, multi-trait analysis in GWASs has evolved to improve the performance of these methods. One major improvement was the possibility to use summary statistics instead of individual-level data, which has significantly enhanced the power of multi-trait analyses given the availability of thousands of summary statistics in different repositories such as GWAS Catalog (50), Common Metabolic Diseases Knowledge portal (cmdkp.org) and IEU OpenGWAS project (51), among others. Another major advance in the field was the possibility of using GWAS summary statistics with known or unknown sample overlap, which can be the case for multiple studies in different biobanks (52).

1.2.5 Polygenic risk scores

In a polygenic disorder, a single polymorphism is not informative enough for assessing disease risk. Instead, a summation of the genetic composite of phenotype-associated SNPs would be necessary to obtain the full spectrum of individuals at increased risk (53). Polygenic risk scores (PRSs) are estimates of individual genetic liability to a trait or disease, this is calculated using the individual genotypes and information on the relation of these genotypes with disease/trait obtained by GWAS. The classical approach consists of computing the sum of risk alleles weighted by the effect sizes estimated in the GWAS of that phenotype (54).

PRS is particularly interesting because it allows us to stratify patients at different levels of risk of disease or adverse effects of a drug. Also, they have been used to evaluate the genetic association between phenotypes (55) and also to combine them to develop clinical-genetic scores to perform more accurate predictions (56).

1.2.6 Mendelian Randomization Analysis

Inferring causality from observational studies is challenging due to the high possibility of bias. Randomized controlled trials are the gold-standard study design for determining the causal status of risk factors. However, as this approach has some limitations (i.e., time-consuming, and expensive), alternative approaches are required. Mendelian Randomization (MR) is a statistical method that uses genetic variants to determine and quantify causal relationships between the effect of exposure on disease outcomes (57).

MR studies use genetic variants to form subgroups analogous to those in a randomized control trial, although in this scenario the subgroups differ only in terms of exposure and not in any other factor, except for those that are causally linked with that exposure. Given the correct MR assumptions, a given genetic polymorphism that is strongly associated with a risk factor can be used to estimate the relationship between the risk factor and an outcome, this can be a disease, trait, expression of a transcript, or others. An association between genetic polymorphism and the outcome will only be possible if the risk factor is causally associated with the targeted outcome.

1.2.7 Drug target discovery

GWAS results can also be used to identify potential drug targets for drug development or drug repurposing. This is especially interesting because it has been estimated that drugs that are supported by genome-wide studies in humans are likely to get to phase III of clinical trials, therefore reducing the costs of testing multiple drugs not supported by human genetics studies (40,58).

This added value of GWAS is due to post-GWAS analysis. Tools have been developed to identify potential drug targets based on gene-based analysis from GWAS data. These strategies have been based on the location of the SNPs or using eQTL information to predict the modulation of transcripts by the SNPs (59,60).

1.3. Genetic behind ischemic stroke

1.3.1 Monogenic strokes

Thanks to advances in sequencing technologies, it has been possible to find the cause of some Mendelian disorders that are associated with and increased risk of stroke, especially at younger ages. Some of these disorders are, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) caused by mutations in *NOTCH3* and associated with an early onset of lacunar stroke; pontine autosomal dominant microangiopathy with leukoencephalopathy (PADMAL) caused by mutations in the 3' untranslated region of *COL4A1*; cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL) caused by recessive mutations in *HTRA1*, and also related to lacunar stroke at younger

ages, 10-30 years; Vascular Ehlers-Danlos syndrome caused my mutations in *COL3A1*, associated with stroke before 40 years (44).

1.3.2 Complex genetic strokes

Ischemic stroke is estimated to have a genetic compound among other factors that contribute to stroke onset. This genetic compound is estimated in the form of heritability. Ischemic stroke is considered to have a heritability of 37.9%, but it differs depending on the subtype, the largest heritability has been determined for large-artery atherosclerosis stroke with 40.3%, 32.6% for cardioembolic stroke, and the lower for lacunar stroke 16.1% (61).

In the last years, several GWAS have been performed trying to uncover the full genomic spectrum of the disease. The first IS GWAS was published in 2007 and consisted of an analysis of about 500 ischemic stroke cases and healthy controls (62). Although this study did not find any significant signal, probably due to the small sample size signal, it was nevertheless the start of subsequent very successful GWAS analyses in this field. In 2016 the METASTROKE collaboration (63), which involved a GWAS meta-analysis comprising up to 10,307 cases and 19,326 controls, revealed four loci and subtypes associated with IS risk (*ABO*, *HDAC9*, *PITX2,* and *ZFHX3*). The *ABO* locus was found to be associated with IS and has previously been shown to be genome-wide associated with circulating levels of von Willebrand factor and factor VIII (64). The *HDAC9* locus showed significance for large artery atherosclerosis stroke (LAA) and has also been linked to inflammation and atherosclerosis (65). Finally, two loci associated with cardioembolic stroke (CE) risk, namely the priority genes *PITX2* and *ZFHX3*, are already known to be signals associated with atrial fibrillation risk (66), the most prevalent risk factor for CE. In the same year, and with a sample size of 16,851 cases and 32,473 non-stroke controls, the NINDS Stroke Genetics Network (SiGN) (67) found a new locus associated with LAA close to *TSPAN2* and confirmed the previous loci *ABO*, *HDAC9*, *PITX2,* and *ZFHX3*. In 2018, the MEGASTROKE Consortium published a GWAS meta-analysis with 67,162 cases and 454,450 controls (68). In this analysis, 32 loci associated with stroke and its subtypes were found, confirming previous loci, and revealing different novel loci (Figure 3). The loci found in the MEGASTROKE study showed previous associations with traits related to stroke risk, especially white matter

hyperintensities in the brain, atrial fibrillation, intima-media thickness, blood pressure, coronary artery disease, lipid levels, and venous thromboembolism. Remarkably, 16 (11%) of the 149 genes located in the 32 stroke risk loci were found to be targeted for currently approved drugs, especially antithrombotic therapies such as alteplase, tenecteplase, and cilostazol (68).

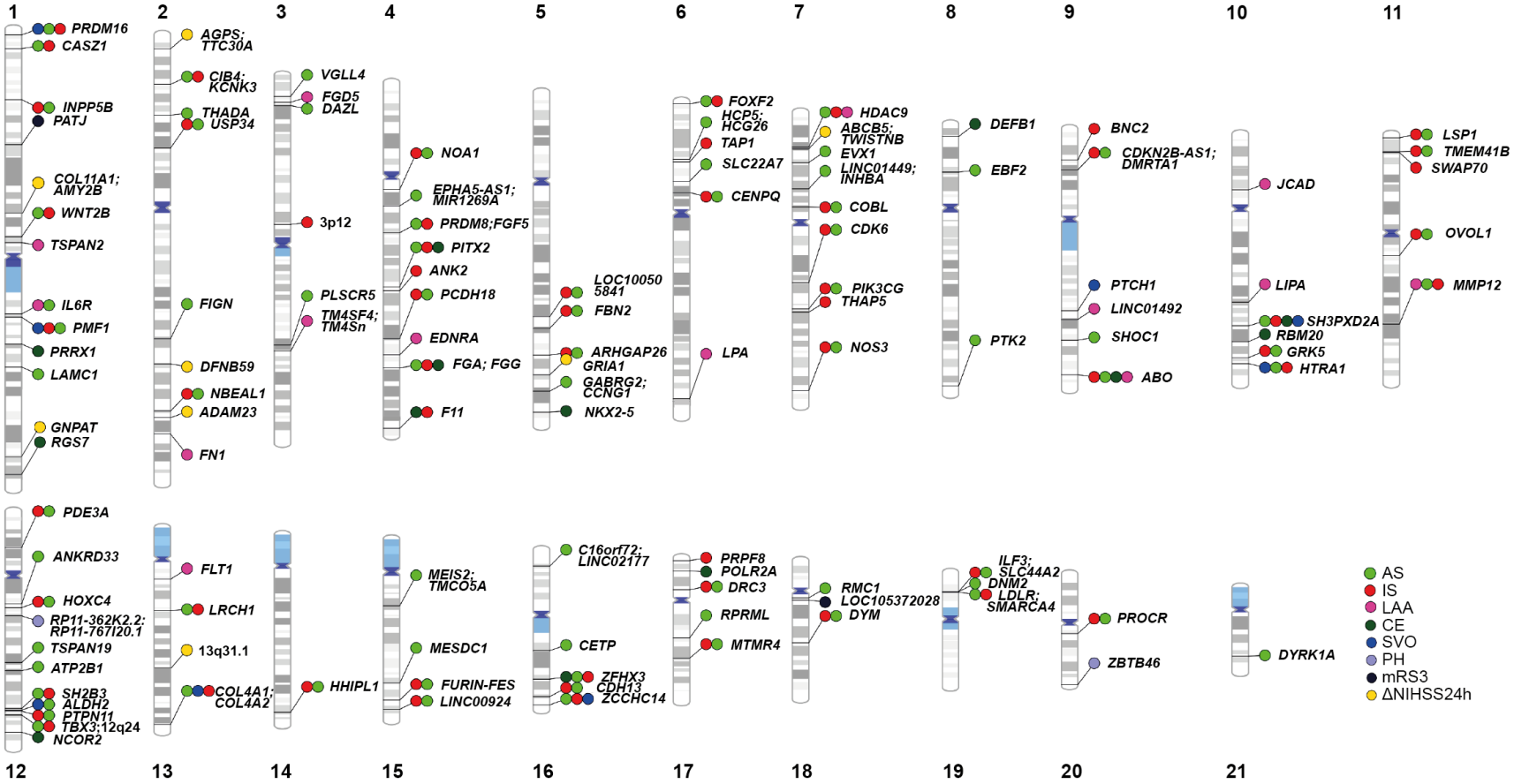


FIGURE 3. Genetic loci influencing stroke phenotypes. Ideogram of genomic regions influencing stroke phenotypes; colored circles represent genome-wide significant loci in published studies. Colors correspond to associated stroke phenotypes (green, AS: All Strokes; red, IS: Ischemic Strokes; pink, LAA: Large-artery atherosclerosis Stroke; dark green, CE: Cardioembolic Strokes; blue, SVO: Small Vessel Occlusion; purple, PH: Parenchymal Hematoma; black, mRS3: modified Rankin Scale three months after stroke; yellow, ΔNIHSS24h: the difference between NIH stroke scale (NIHSS) within six hours of stroke onset and NIHSS at 24 h. Prioritized genes in the original publications are displayed.

In the same year, a joint GWAS meta-analysis between MEGASTROKE and The United Kingdom Biobank (UKB) (up to 72,147 cases and 823,869 controls) revealed three new loci associated with stroke (Figure 3)(69). Of these, the *NOS3* gene reached significance only when considering the European population, whereas *COL4A1* and *DYRK1A* were identified in the multiethnic GWAS meta-analysis. These findings highlight the role of *COL4A1* in stroke. Indeed, this gene has previously been associated with small vessel disease (70). In addition, the rs720470 variant identified is an eQTL of *DYRK1A*, which is involved in angiogenic responses in vascular endothelial cells (71). Concerning *NOS3*, functional variants in this and other nitric oxide synthases (NOS)/nitric oxide (NO) pathway genes have been associated with hypertension (72). The authors of that study demonstrated how genetic variation in the NOS/NO pathway partly affects stroke risk because of variations in blood pressure by using the Mendelian randomization approach (69).

A multi-ethnic GWAS meta-analysis by The Global Biobank Meta-Analysis Initiative (GBMI), including 19 biobanks and more than 2.1 million people, has been published this year (73). In this study 14 diseases, including stroke, were analyzed (74). By combining both fixed-effects and inverse variance-based meta-analysis, the authors identified seven loci previously reported in stroke along with nine new loci (Figure 3). For the *CENPQ* and *ALDH2* genes, a protein-coding variant was the most significant variant. Additionally, it was found that the *MEIS2*/*TMCO5A* gene locus was driven by variants that are more frequent in African ancestry. Furthermore, in a female-only meta-analysis, the previously reported *CETP* locus met the genome-wide significance threshold but did not reach significance in the combined sex meta-analysis (74).

However, the largest step forward was taken by the GIGASTROKE project, which involved the largest multiethnic meta-analysis GWAS in stroke (110,182 stroke patients and 1,503,898 controls) (75) and in which the International Stroke Genetics Consortium (ISGC) collaborated. The results represent the most comprehensive description of stroke-risk genetic variants in Europeans, East Asians, South Asians, African Americans, and Hispanics to date. Thus, 89 loci, 61 of which are novel, were identified for stroke and stroke subtypes (Figure 3). The results suggest substantial shared susceptibility to stroke across

populations. Cross-ancestry fine-mapping, in silico mutagenesis analysis with a machine learning approach, and transcriptome and proteome-wide association analyses were used to reveal putative causal genes and variants, such as *SH3PXD2A*, *FURIN,* and *NOS3*. Furthermore, the authors used a three-pronged approach to identify the putative drug targets for the prevention or treatment of stroke. Indeed, drugs targeting *F11* and *PROC* are currently being explored in clinical trials (76). Finally, polygenic risk scores integrating cross-ancestry and ancestry-specific stroke genetic risk variants with vascular risk factors were performed and enabled the prediction of ischemic stroke in 52,600 participants with cardiometabolic disease.

In addition, several GWASs have been carried out to understand ischemic stroke outcomes leading to the identification of genome-wide significant loci (Figure 3). These have been focused on studying the functional recovery of patients focused in the modified Rankin Scale at three months (77), neurological deterioration after stroke as ΔNIHSS (78), and even evaluating adverse effects in the form of parenchymal hematoma after rtPA treatment (79).

Regarding multi-trait analysis in GWASs in stroke, the subtype most studied has been lacunar stroke subtype. Analysis has been carried due to the phenotypic relationship known to exist between this subtype and the presence of white matter hyperintensities in magnetic resonance imaging (MRI). A multitrait analysis revealed seven additional loci to those found only in the single-trait study. The prioritized genes and associated loci were *SLC25A44-PMF1-BGLAP*, *LOX-ZNF474-LOC100505841*, *FOXF2-FOXQ1*, *VTA1-GPR126*, *SH3PXD2A*, *HTRA1-ARMS2*, *COL4A2*. Interestingly, two of the loci identified contain genes (*COL4A2* and *HTRA1*) that are implicated in monogenic lacunar stroke, thereby highlighting the power of this multitrait approach (80). Another study also used lacunar stroke risk information, but in this case to potentiate genomic loci associated with intracerebral hemorrhage, as these two phenotypes are two diverse manifestations of cerebral small vessel disease (cSVD). In this case, two new loci associated with non-lobar intracerebral hemorrhage were found, and a previous locus was confirmed (81).

Using the PRS approach has also explored the relationship between stroke genetics and related phenotypes. An example of this is the significant association of genetics of stroke and cognitive ability and genetics of stroke with depression (82,83). Besides, as commented before an important application of PRS is the joint of PRS with clinical data to obtain more accurate predictions. In this sense, a clinical-genetic score was developed to predict hemorrhagic transformation after thrombolytic treatment with rtPA (56).

All these genomic techniques, GWAS, multi-trait analysis, drug targeting, and PRS, are adding valuable information to be potentially used in the future in the clinical practice.

1.4. Mendelian Randomization and stroke

Regarding the search for the consequence/causal relationship of established and candidate risk factors for stroke, a wide variety of MR studies have been carried out. In general, the most well-established risk factors such as atrial fibrillation, diastolic and systolic blood pressure, smoking, type 2 diabetes, obesity, etc. have been demonstrated via MR as causal players in ischemic stroke risk and subtypes (84–94). Regarding lifestyle factors, such as physical activity, sedentary lifestyle, sleeping habits (95–100), and dietary habits (101–104), there is a general lack of evidence about their causal effect on IS. The exceptions are alcohol consumption (95,101,102) and insomnia (100), which are causally associated with IS and LAA, respectively, in addition following a Mediterranean diet seems to protect against IS (105), and education level (95,106–108) and tea consumption are causally associated with a lower risk of SVO (109).

Perhaps the most interesting studies are those assessing causality using blood biomolecules as these are easy biomarkers to capture for stroke risk assessment but can also be potential drug targets. Not all the studies to date have observed causal associations with stroke risk, in either direction (inflammatory biomarkers (110–113), circulating cytokines (114), vitamins (115–120), and many polyunsaturated fatty acids (121)). In a recent study a total of 653 blood circulating proteins were interrogated revealing ABO, the cluster of differentiation 40, apolipoprotein(a), and matrix metalloproteinase-12 causally associated with an elevated risk of large-artery atherosclerosis stroke, and proteins SCARA5 and

TNFSF12 were suggested to play a causal role in cardioembolic stroke risk protection (49). Genetically determined levels of some hemostatic factors have also been associated with the risk of IS (122,123). Iron factors are causally associated with an increased risk of IS and CE subtype, except transferrin, which is protective against IS and CE (124). Among the cytokines studied, monocyte chemoattractant protein-1 is the only one that was associated with an increased risk of IS, LAA, and CE (114).

2. Hypothesis

2. Hypothesis

Three main hypotheses were considered for the present thesis dissertation:

- The study of stroke genetics in the Spanish population will allow us to find specific genetic risk factors.
- The joint analysis of atrial fibrillation and cardioembolic stroke genetics will permit us to better understand the occurrence of stroke in patients with atrial fibrillation.
- Blood levels of metalloproteinases may have a causal relationship to stroke risk and outcome after stroke.

3. Objectives

3. Objectives

The main objective of the present thesis dissertation is:

- Study of genetic risk factors associated with ischemic stroke.

Other secondary objectives include:

- Conduct genome-wide association analysis of ischemic stroke and subtypes in a Spanish population.
- Conduct a multitrait analysis of genome-wide association analyses of cardioembolic stroke and atrial fibrillation.
- Develop a clinical-genetic score to predict cardioembolic etiology.
- Evaluate the causality of Metalloproteinases blood levels and the risk of ischemic stroke and post-stroke recovery using Mendelian Randomization.

4. Methods

4. Methods

4.1. Genome-Wide association Studies in the GENERACION cohort.

4.1.1 Cohort description

The GENERACION cohort was formed by 9,111 individuals, with ischemic strokes, and controls recruited in Spain. IS patients over 18 years were recruited via hospital-based studies, between 2003 and 2020 in Spain, if they had a measurable neurologic deficit on the NIHSS within 6 hours of the last known asymptomatic status, had been diagnosed with stroke by an experienced neurologist and confirmed by neuroimaging. Controls were subjects over 18 years recruited in Spain, without a history of IS, who declared they were free of neurovascular diseases before enrollment. An Institutional Review Board or Ethics Committee approved the study at each participating site. All patients or their relatives provided written informed consent.

Participants of the GENERACION project were part of the Genetics of Early Neurological Instability After Ischemic Stroke (GENISIS), Genetic contribution to Functional Outcome and Disability after Stroke (GODS), the Genetic Study in Ischemic Stroke Patients treated with tPA (GenoTPA), the CONtrol ICtus (CONIC), and SEDMAN studies. Controls were subjects without a history of ischemic stroke, aged over 18 years, who declared they were free of neurovascular diseases before recruitment. The control cohort was collected in blood donation at primary care centers in Barcelona and in hospitals throughout Spain as a part of the GCAT, CONtrol ICtus (CONIC), Investigating Silent Stroke in hYpertensives: A magnetic resonance imaging Study (ISSYS) and the Genotyping Recurrence Risk of Stroke (GRECOS) projects.

IS patients were recruited as part of the GENISIS(125), GODS(126), and CONIC(127) projects.

Controls were subjects without a history of ischemic stroke, aged over 18 years, who declared they were free of neurovascular diseases before recruitment. The control cohort was collected in blood donation and primary care centers in Barcelona and hospitals throughout Spain as a part of the GCAT(128), CONIC(129), GRECOS(130), and ISSYS(131) projects. The exact number of

ischemic strokes and controls from each participant hospital and project can be observed in Tables 3-4.

TABLE 3. The detailed number of participants (IS and controls) included in each project and array type.

Name	Array	IS	Controls	Total
GCAT	Illumina® Infimum Multi-Ethnic Global consortium		4893	4893
GENISIS	Illumina® Human Core Exome	2217		2217
GODS	Illumina® Human Core Exome	640		640
CONIC	Illumina® Human Core Exome	182	191	373
ISSYS	Illumina® Human Core Exome		316	316
SEDMAN	Illumina® Human Core Exome	110		110
	Axiom™ Spain Biobank	206		206
GRECOS	Illumina® Human Core Exome		192	192
GENOTPA	Illumina® Omni 2.5M	164		164
TOTAL		3519	5592	9111

IS: ischemic stroke patients.

TABLE 4. The detailed number of individuals included form each participant Hospital on the GENERACION cohort.

Hospitals (City)	IS	Controls	Total
Hosp. Universitari Vall d'Hebron (Barcelona)	1101	609	1710
Hosp. Clínico Universitario de Santiago (Galicia)	585	0	585
Hosp. del Mar (Barcelona)	548	6	554
Hosp. Universitari Germans Tries i Pujol (Badalona)	345	13	358
Hosp. Universitari Son Espases (Palma de Mallorca)	253	0	253
Hosp. Universitari Mútua de Terrassa (Terrassa)	208	35	243
Complejo Hospitalario Universitario de Albacete (Albacete)	191	0	191
Hosp. Clinic de Barcelona (Barcelona)	110	0	110
Hosp. Universitario Virgen del Rocio Y Virgen Macarena (Sevilla)	65	6	71
Hosp. de la Santa Creu i Sant Pau (Barcelona)	61	2	63
Hosp. Universitario Río Hortega (Valladolid)	26	7	33
Hosp. Universitario de Basurto (Bilbao)	22	11	33
Hosp. Universitario Doctor Josep Trueta (Girona)	1	10	11
Hosp. Arnau de Vilanova (Lleida)	3	0	3
GCAT Project		4893	4893
TOTAL	3519	5592	9111

Hosp: Hospital; IS: ischemic stroke patients.

Description of the cohorts included:

GENISIS cohort.—Genetics of Early Neurological Instability after Ischemic Stroke (GENISIS)(132) is an international study currently recruiting patients from four different locations: the United States, Finland, Poland, and Spain. The inclusion

criteria for the GENISIS study are IS patients (age ≥ 18 years) Collected from 2003 to 2016 with a measurable neurologic deficit on the NIHSS within 6 hours of the last known normal. Patients who received endovascular thrombectomy, or for whom consent and/or a blood sample could not be obtained were excluded. For our study, we only include Spanish patients. Genotyping was performed with the Human Core Exome chip (Illumina®).

GODS cohort.—Genetic contribution to functional Outcome and Disability after Stroke (GODS)(77) project is a study that aims to find genetic factors associated with stroke outcome. All participants met the following criteria: (1) European descent, aged >18 years, diagnosis of IS in the anterior vascular territory; (2) assessed by a neurologist during the acute phase of stroke; (3) initial stroke severity >4, according to the National Institutes of Health Stroke Scale (NIHSS); (4) information on post-stroke functional status at 3 months (or between 3-6 months); (5) evidence of acute IS in a neuroimaging study; (6) lack of concomitant pathology. Individuals with stroke recurrence during the follow-up period were excluded, as well as posterior vascular territory and lacunar strokes. Samples were genotyped at the Genetic and Molecular Epidemiology Laboratory of McMaster University (David Braley Research Institute) in Ontario, Canada, with the Human Core Exome chip (Illumina®).

CONIC cohort.—CONtrol ICtus (CONIC) study(129) is a national study that recruited control and IS case participants in Vall d'Hebron Hospital between 2007 and 2008. All controls were older than 65 years of age and declared free of dementia, neurovascular, and/or cardiovascular disease, as evaluated by self-description during a direct interview before recruitment. Subjects with a history of first and/or second-degree neurovascular disorder were also excluded from the study.

The IS cases were admitted to the emergency department of a university hospital who had a documented middle cerebral artery (MCA) occlusion on transcranial Doppler ultrasonography (TCD) and received tPA in a standard 0.9-mg/kg dose (10% bolus, 90% continuous infusion for 1 hour) within 3 hours of symptom onset following National Institute of Neurological Disorders and Stroke (NINDS)

recommendations. Cases and controls were genotyped with the Human Core Exome chip (Illumina®).

GRECOS cohort.—Genotyping RECurrence Risk Of Stroke (GRECOS)(133) project is a national study that aims to find genetic factors associated with the recurrence after stroke. Control participants were selected from relatives of patients (wife or husband, without any consanguinity among cases and controls) and healthy volunteers visiting the same hospital for routine testing. They were >65 years of age and classified as free of neurovascular and cardiovascular history and familial history of stroke by direct interview before recruitment. All samples were genotyped with the Human Core Exome chip (Illumina®).

ISSYS cohort.—Investigating Silent Stroke in hYpertensives: A magnetic resonance imaging Study (ISSYS)(134) is an observational prospective study in hypertensive participants to determine the prevalence of silent or MRI–defined brain infarcts and cognitive impairment. This cohort comprises 1000 non-demented individuals, aged 50 to 70 years old, and diagnosed with essential hypertension at least one year before inclusion in the ISSYS study. Those individuals were genotyped with the Human Core Exome chip (Illumina®).

GCAT cohort.—GCAT health databank is a collection of health data and samples from participants of the “GCAT/Genomes for Life. Cohort Study of the Genomes of Catalonia Study”(135). The GCAT project aimed to study the genetic and environmental factors that lead to the appearance of chronic diseases in the general population. The study is conducted in several waves of data gathering, namely GCAT1, the baseline Survey from 2014-2017, and GCAT2, the GCAT follow-up in the second year. Data collection is done with web-based self-questionnaires, direct interviews, clinical data, and analyses of DNA blood-derived samples. Genome-wide genotypes have been generated using Illumina Infinium SNV-bead array technology using the Multi-Ethnic Global (MEGAEX, V.2) consortium array. We used only GCAT1 genotyped patients, and we exclude individuals with heart infarct or heart diseases or with non-Caucasian ancestry.

genotPA—Consecutive Caucasian patients with acute ischemic stroke who were admitted to the emergency room and received recombinant tissue-type plasminogen activator (r-tPA) within 4.5 hours of symptom onset were recruited.

Patients were enrolled at Spanish hospitals (Vall d'Hebron University Hospital, Hospital Clinic, Hospital Universitari de Girona Doctor Josep Trueta, Hospital de la Santa Creu i Sant Pau, Hospital Universitari Germans Trias I Pujol, Hospital Universitari del Mar, Hospital de Basurto) between 2002 to 2012. The study protocol was approved by the Ethics Committee of each center, and all patients or relatives signed the informed consent.

Patients were identified by medical evaluation at emergency room arrival; stroke diagnosis was performed by trained neurologists and confirmed by neuroimaging. There were no exclusion criteria regarding age, sex, or ethnicity. Follow-up computerized tomography (CT) scans 24 hours after the stroke onset of symptoms or if neurological worsening occurred were performed and were classified according to European Cooperative Acute Stroke Study (ECASS) (136).

SEDMAN cohort. - patients ≥ 18 years old, treated with acenocoumarol or dabigatran for stroke or systemic embolism prevention following the local recommendations. All patients had a stroke or transient ischemic attack (TIA) during the previous 14 days before the initiation of anticoagulation treatment and had a diagnostic of non-valvular atrial fibrillation. Only patients with mild to moderate stroke (less than 2/3 of the vascular territory) with initial Alberta Stroke Program Early CT Score (ASPECTS) in the first CT/MRI > 6 and National Institute of Health Stroke Scale (NIHSS) < 25 were included. All patients had a general condition which allowed 12 months of follow-up. Only patients with established stroke were used in this analysis.

4.1.2 Quality control and imputation

DNA samples were genotyped on commercial arrays from Illumina (San Diego, CA) and using the Axiom Spain Biobank Array (Table 3). Quality controls were performed using PLINK v1.9 and KING v2.1.3 software for the different arrays analyzed and by two batches of Human Core Exome samples. For all datasets, samples were excluded if there was a mismatch between the genetic and reported sex, genotype call rate lower than 95%, excess or loss of heterozygosity, non-European detected as outliers of 1000 Genomes Project Phase 3 dataset (1000G), duplicate samples or relatedness at a PI-HAT>0.20. SNPs were excluded if the call rate was lower than 95%, located in non-autosomes, non-

biallelic, strand ambiguous, monomorphic or were deviated from Hardy-Weinberg equilibrium (p-value $<1\times10^{-6}$ in controls, $<1\times10^{-10}$ in IS).

Imputation was performed in the Michigan Imputation Server (137) using Minimac4. HRC r1.1 2016 (GRCh37/hg19) was the reference panel used, with the European population, and for phasing Eagle v2.4 was used. After imputation, we removed SNV with an imputation score $r^2<0.6$, MAF <1%, and variants that deviated from Hardy-Weinberg equilibrium in controls (p-value $< 1\times10^{-6}$). SNPs that were not present in at least 95% of the individuals were removed. Additionally, we removed duplicate samples or relatedness at a PI-HAT > 0.2.

4.1.3 GWAS analysis

We performed different case-control GWAS evaluating ischemic strokes vs controls, and different etiologies: large artery-atherosclerosis strokes vs controls, cardioembolic strokes vs controls, small vessel strokes vs controls, and strokes of undetermined cause vs controls. Additionally, as a positive control of our results, we performed an analysis of atrial fibrillation confirmed in cases and controls vs non-atrial fibrillation strokes and controls. We performed logistic regression considering an additive genetic effect of the variants using fastGWA from GCTA (138). Age, sex (if not stratified analysis), and the first ten genetic principal components were used as covariates. Nine individuals were missing for the age variable and were excluded from the subsequent analyses. We considered significant variants when reaching a p-value threshold of 5×10^{-8}. Additionally, after the analysis, we removed variants with minor allele count in cases and controls less than 6.

Replication

We evaluated significant replication of the obtained variant in MEGASTROKE (78) and GIGASTROKE (75) European analyses.

4.1.4 Gene-based analyses

We used MAGMA (139), to aggregate SNP-level associations to gene-level signals using the resulting GWAS summary statistics from fastGWA.

4.1.5 In silico proteome analyses

We used the results of the INTERVAL study (36) in which 3,622 plasma proteins were evaluated with the Somalogic array and genetics were measured in 3,301 healthy participants.

We selected proteins modulated by the candidate SNPs selected from GWAS analyses that were replicated in the international cohorts.

4.1.6 Pathway's analysis

Webgstal tool (140) was selected to perform an Overrepresentation analysis of Gene Ontology Biological processes by the significantly (p-value < 0.05) modulated proteins. Background genome was selected for all the proteins evaluated in Somalogic by the INTERVAL study, to avoid bias due to the protein selection of the array. Additionally, we performed Gene-Set Enrichment Analysis (GSEA) for biological processes altered using the information of the effect sizes upon all the INTERVAL evaluated proteins.

4.1.7 Evaluation of previously reported loci for stroke and subtypes

The lead variants of the 87 loci from GIGASTROKE (75) analyses were selected to evaluate replicability in our cohort. The first replication was considered using only the directionality effect, given that the current GENERACION cohort is not powered enough to replicate signals that have been obtained from very larger sample sizes as in GIGASTROKE (> 1 million individuals). We also tested if variants were significant at a p-value threshold of less than 0.05 and finally also if they were significant after Bonferroni correction for multiple comparisons. This replication methodology was followed previously (141).

4.2. Multitrait analysis of GWAS of cardioembolic stroke

4.2.1 Cohorts' description

The summary statistics for CE were obtained from the MEGASTROKE analysis (MEGASTROKE-CE) through the Cerebrovascular Disease Knowledge Portal (http://cerebrovascularportal.org). This cohort was composed of 7,193 CE patients and 355,468 controls of European ancestry. The summary statistics for AF were obtained from the Atrial Fibrillation 2018 analysis (AF-2018) through the GWAS catalog portal (https://www.ebi.ac.uk/gwas/). The AF-2018 cohort was composed of 60,620 AF cases and 970,216 controls.

MEGASTROKE-CE

For the European ancestry analysis of the MEGASTROKE consortium, 16 different cohorts were analyzed, comprising up to 34,217 cases of ischemic stroke and 405,111 healthy controls. Stroke was defined according to the World Health Organization (WHO) as rapidly developing signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin. Strokes were defined as ischemic stroke (IS) or intracerebral hemorrhage (ICH) based on clinical and imaging criteria. IS was further subdivided into the following categories mostly using the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria): i) large vessel ischemic stroke; ii) cardioembolic ischemic stroke; iii) small vessel ischemic stroke. Specifically, for the European analysis of CE, they analyzed 7,193 cases of CE and 355,468 healthy controls. Further details of the cohorts are in the original publication (68).

AF-2018 study

A total of 60,620 cases of AF and 970,216 controls were analyzed, these patients were recruited as part of cohorts:

HUNT.—The Nord-Trøndelag Health Study (HUNT) is a population-based health survey conducted in the county of Nord-Trøndelag, Norway from 1984 to 2009 (142). They used a combination of hospital, out-patient, and emergency room discharge diagnoses (ICD-9 and ICD-10) to identify 6,493 atrial fibrillation cases and 63,142 atrial fibrillation-free controls with genotype data. Participation in the HUNT Study is based on informed consent, and the study was approved by the

Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway.

deCODE.—The Icelandic atrial fibrillation population consisted of all patients diagnosed with atrial fibrillation (ICD-10 code I48 and ICD-9 code 427.3) at Landspitali, The National University Hospital, in Reykjavik, and Akureyri Hospital (the two largest hospitals in Iceland) from 1987 to 2015. All atrial fibrillation cases, a total of 13,471, were included. Controls were 358,161 Icelanders recruited through different genetic research projects at deCODE genetics, excluding those in the atrial fibrillation cohort. The study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee of Iceland (no. VSNb2015030021).

MGI.—MGI is a hospital-based cohort collected at Michigan Medicine, USA. Atrial fibrillation cases (n = 1,226) were defined as patients with ICD-9 billing code 427.31, and controls were individuals without atrial fibrillation, atrial flutter, or related phenotypes (ICD-9 426–427.99). MGI was reviewed and approved by the Institutional Review Board of the University of Michigan Medical School.

DiscovEHR.—The DiscovEHR collaboration cohort is a hospital-based cohort including 58,124 genotyped individuals of European ancestry from the ongoing MyCode Community Health Initiative of the Geisinger Health System, USA(143). Atrial fibrillation cases (n = 6,679) were defined as DiscovEHR participants with at least one electronic health record problem list entry or at least two diagnosis code entries for two separate clinical encounters on separate calendar days for ICD-10 I48: atrial fibrillation and flutter. Corresponding controls (n = 41,803) were defined as individuals with no electronic health record diagnosis code entries (problem list or encounter codes) for ICD-10 I48. The study was approved by the Geisinger Institutional Review Board.

UK Biobank.—The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom(144). Cases of atrial fibrillation were selected using ICD-9 and ICD-10 codes for atrial fibrillation or atrial flutter (ICD-9 427.3 and ICD-10 I48). Controls were participants without any ICD-9 or ICD-10 codes specific for atrial fibrillation, atrial flutter, other cardiac arrhythmias, or conduction disorders.

AFGen Consortium.—Published atrial fibrillation association summary statistics from 31 cohorts representing 17,931 atrial fibrillation cases and 115,142 controls were obtained from the authors.

4.2.2 Single-Nucleotide Polymorphism Quality Controls

A series of standard quality controls (QC) was applied to select the single nucleotide variants for the analysis. Variant exclusion criteria: 1) Not common to the summary statistics of the traits, 2) Minor allele frequency lower or equal to 0.01, 3) (140) values, 4) Negative standard error or not a number value, 5) p-value of 0, 6) Not SNPs, 7) Duplicated SNPs, 8) Strand ambiguity and 9) Inconsistent allele pairs. Locus 15q21.3, prioritized genes *GCOM1* and *MYZAP,* from AF-2018 was not evaluated due to the absence of the significant SNPs of AF-2018 in the MEGASTROKE-CE analysis.

4.2.3 Multitrait Analysis of GWAS

We applied MTAG (145) of MEGASTROKE-CE and AF-2018 summary statistics. We considered loci to be significantly associated with the trait of interest when the p-value was $<5\times10^{-8}$ in the MTAG result and the p-value <0.05 in the original GWAS. We considered replicating the SNPs with a p-value <0.05 in the GWAS of our independent cohort.

To avoid an increase in the type I error rate due to the presence of SNPs that are not associated with CE but with AF or vice versa, we used GWAS-pairwise (146). This is a Bayesian pleiotropy association test to identify genetic variants that influence pairs of traits (146). We used it to ensure that the leading SNP of a significant locus belongs to a genomic region influenced by both traits evaluated (146), since SNPs that are not associated with one trait but are associated with the other one, could bias effect-size estimates for the first trait and increase false positive rate (145). The posterior probability for model-3 (PPA-3) >0.6 suggests that a specific genomic region is associated with both traits. A PPA-1 >0.6 will suggest that the genomic region is associated only with CE, and a PPA-2 >0.6 is only associated with AF. Genomic inflation was estimated as lambda.

4.2.4 Identification of Independent and Novel Loci Associated With CE

Independent loci were defined as those >1 megabase (Mb) apart in the physical distance among SNPs with a genome-wide significance threshold of p-

value<5×10^{-8} (66). Loci were defined as novel when SNPs had an r^2<0.1 compared with the index SNPs of the loci: *PITX2* (68), *ZFHX3* (68), *NKX2-5* (68), *RGS7* (68), *ABO* (68) (147), *PHF20* (148), *GNAO1* (148) and 5q22.3 (148) that were GWAS significant in previous studies.

4.2.5 Replication Stage in an Independent European Cohort

We performed GWAS analyses of the GENERACION cohort in an independent cohort of 9,105 individuals with genotyped data, ischemic stroke patients, controls, and data for age and sex. (GENERACION cohort: 3,479 ischemic stroke patients and 5,625 controls). Further information on the GENERACION cohort is contained in subsection 5.1.2.1.

GWAS analysis

We performed two different GWAS in the same cohort (for the two different traits here studied), with an additive genetic model using fastGWA from GCTA (138). We studied the association with CE (CE = 1,515; controls = 5,626) and AF (AF patients = 1,110; controls = 7,791). Age, sex, and the first ten principal components were used as covariates.

The results of these two GWAS were used to evaluate replicability. We studied those index variants from significant loci with a p-value <5×10^{-8} in the MTAG, a p-value <0.05 in the original GWAS used for performing the MTAG, and PPA-3 >0.6 suggesting that the genomic region is associated with CE, and AF. We considered replicating the SNPs with a p-value <0.05 and a consistent direction of the effect in this analysis.

4.2.6 Functional Annotation and Gene Prioritization

Gene prioritization was performed for the novel loci using the Variant-to-Gene tool from Open targets Genetics Version 7 (149). This tool integrates biological evidence of four main data types: 1) Molecular phenotype quantitative trait loci experiments (QTLs), 2) Chromatin interaction experiments, e.g., Promoter Capture Hi-C (PCHi-C), 3) *In silico* functional predictions, e.g. Variant Effect Predictor (VEP) from Ensembl and 4) Distance between the variant and each gene's canonical transcription start site (TSS). Additionally, we used the HaploReg database to determine the functional annotation of the most strongly associated SNPs per locus. For the missense SNPs, we determined the

likelihood that amino acid substitution has a deleterious effect on protein function using SIFT score.

4.2.7 Gene Set Analysis

We conducted a WebGestalt Overrepresentation Analysis of the selected prioritized genes associated with MTAG-CE. Gene Ontology (GO) of biological processes was performed, as well as a Benjamini Hochberg correction of the association p-value. We defined a biological process with a q-value<0.05 as statistically significant.

4.2.8 Polygenic Risk Score Development

A PRS was conducted through PRSice-2 software version 2.3.3 (48), where estimation is based on the risk alleles of having a CE and their effect size extracted from the regions with PPA-3 >0.6 of the MTAG summary statistics here created.

The GENERACION cohort was randomly split into training and test sets in 80:20 proportion. Best score threshold selection was performed based on the major variance explained by the score (PRS r^2) in the training set. The evaluation of this score was performed in the independent test set.

We used R version 4.1.3 and Bioconductor packages to evaluate the clinical relevance of this PRS. We calculated three models: model-1 including only the PRS, model-2 including statistically significant clinical variables with <10% missing values, since a high rate of missing values might bias the results of subsequent statistical analyses (150), and model-3 adding the PRS to model-2. Model discrimination was assessed with the area under the roc curve (AUC) and the area under the precision-recall curve (AUPRC). We used DeLong's test for two correlated ROC curves to find out whether there are significant differences between the discrimination of the models. The net reclassification index (NRI) and integrated discrimination index (IDI) were performed to evaluate model-2 and model-3. Additionally, we estimated the AUC and AUPRC for each predictor.

4.3. Metalloproteinases levels and ischemic stroke

4.3.1 SNP Selection and Data Sources

A literature search was conducted in PubMed in May 2020. The keywords used were: “matrix metalloproteinase”, “serum” or “plasma”, “levels” or “concentration”, and “GWAS”. Genome-wide significant (p-value < 5×10^{-8}) and independent (r^2 < 0.2) single nucleotide polymorphisms were used as instruments for MR analysis. To evaluate a causal effect on risk of stroke we used summary-level data from the European analysis of MEGASTROKE (68) (IS subjects = 34,217, controls = 406,111) for IS and its subtypes: LAA (n = 4,373), cardioembolism (CE) (n = 7,193) and small-vessel occlusion (SVO) (n = 5,386). To evaluate post-stroke functional outcome with the modified Rankin scale (mRS) at three months, we extracted summary-level data from the GODS (Genetic contribution to functional Outcome and Disability after Stroke) project in which mRS was analyzed as a continuous variable (n = 1,791) (126).

For each cohort, all aspects of the studies were approved by the local institutional review board and ethics committee. All the participants included, or their approved representatives provided written informed consent for participation.

4.3.2 Cohort’s description

MEGASTROKE consortium GWAS - For the European ancestry analysis of the MEGASTROKE consortium, 16 different cohorts were analyzed, comprising 34,217 cases of ischemic stroke and 406,111 healthy controls. Stroke was defined according to the World Health Organization (WHO) definition, i.e. rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer, or leading to death, with no apparent cause other than of vascular origin. Strokes were defined as ischemic stroke (IS) or intracerebral hemorrhage (ICH) based on clinical and imaging criteria. IS was further subdivided into the following categories, mostly using the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria: i) large vessel ischemic stroke; ii) cardioembolic ischemic stroke; iii) small vessel ischemic stroke.

GODS GWAS - European ancestry patients with a diagnosis of IS according to World Health Organization criteria were selected from the Spanish Stroke Genetics Consortium (GeneStroke) and the International Stroke Genetics Consortium (ISGC). All participants met the following criteria: (1) European

descent, aged >18 years, diagnosis of IS in the anterior vascular territory; (2) assessed by a neurologist during the acute phase of stroke; (3) initial stroke severity >4, according to the National Institutes of Health Stroke Scale (NIHSS); (4) information on post-stroke functional status at 3 months (or between 3 and 6 months); (5) evidence of acute IS in a neuroimaging study; (6) lack of concomitant pathology. Individuals with stroke recurrence during the follow-up period were excluded. Posterior vascular territory and lacunar strokes were also excluded, as these locations show a poor correlation between infarct size and clinical symptoms, and thus, functional outcome. In these cases, the recovery process could be masked by the random location effect and lead to imprecision in measuring the genetic contribution to the degree of recovery.

Stringent analysis was the analysis used in this study. They only included cohorts that fulfilled the above criteria and had complete information for the following variables: (1) 3-month post-stroke functional status; (2) NIHSS score at hospital discharge; and (3) previous functional independence (mRS<3) before the stroke.

MMP-1 levels GWAS - For the MMP-1 levels GWAS, they used the Heredity and Phenotype Intervention (HAPI) Heart Study that was initiated in 2002. Participants in this study comprised adults from the Old Order Amish community of Lancaster County, PA, USA, who were recruited over three years. Study participants were aged 20 years and older and relatively healthy based on a variety of exclusion criteria, including severe hypertension (blood pressure > 180/105 mm Hg), among others. The study aims, recruitment procedures, and ascertainment criteria have been described previously(151). The study included 868 participants, 792 of whom had available DNA and serum MMP-1 measurements. Fourteen additional individuals were excluded from the final analysis due to genotyping issues, leaving 778 subjects(152).

MMP-8 levels GWAS - For this study, the Corogene and FINRISK 1997 cohorts were used:

The Corogene study included 5295 Finnish patients who were assigned for coronary angiography in the Helsinki University Central Hospital, Finland, between 2006 and 2008(153). Blood samples were drawn from the arterial line during the coronary angiography into serum and citrate plasma vacuum tubes.

The samples were handled according to the laboratory standards of Helsinki University Central Hospital (accredited laboratory).

FINRISK is a Finnish national population-based study(154) that has been conducted at 5-year intervals since 1972. In 1992, 1997, 2002, and 2007, the surveys were performed in 5 geographical areas of Finland, and they included a clinical examination, questionnaire, and laboratory analyses. Blood samples were collected during the health examination. Serum samples were taken without coagulation activators. Cases with prevalent coronary heart disease (CHD) or CVD at baseline were identified using (1) the questionnaire as a doctor-diagnosed disease, (2) the disease-associated drug reimbursement records from the Social Insurance Institution of Finland, including purchased medications and entitled reimbursements, and (3) the National Hospital Discharge register for hospitalizations. The study includes up to 20 years of follow-up data. The incident CHD events (acute myocardial infarction [AMI], bypass surgery, and angioplasty), CVD events (CHD events and stroke), and all-cause deaths during follow-up were identified using (1) the drug reimbursement records from the Social Insurance Institution of Finland, (2) the National Hospital Discharge register for hospitalizations, and (3) the National Causes-of-Death Register.

In total, genotype data and serum concentrations of MMP-8 were available for 2203 subjects in the Corogene cohort and 3846 subjects in the FINRISK 1997 cohort.

MMP-12 levels GWAS - For this study, they used the IMPROVE cohort, which consists of a multicenter, longitudinal, observational study that recruited 3711 individuals, aged 54-79 years, with at least three cardiovascular risk factors but who were asymptomatic for cardiovascular disease. Subjects were considered to be exposed to a vascular risk factor when one of the following criteria was met: male sex or at least 5 years after menopause for women; hypercholesterolemia (mean calculated LDL-C blood levels > 160 mg/dL or treatment with lipid-lowering drugs); hypertriglyceridemia (triglycerides levels > 200 mg/dL after diet or treatment with triglyceride-lowering drugs); hypoalphalipoproteinemia (HDL-C < 40 mg/dL); hypertension (diastolic blood pressure > 90 mmHg and/or systolic blood pressure >140 mmHg or treatment with anti-hypertensive drugs); diabetes

or impaired fasting glucose (blood glucose level > 110 mg/dL or treatment with insulin or oral hypoglycemic drugs); smoking habit (at least 10 cigarettes/day for at least thirty months); family history of cardiovascular disease. Exclusion criteria were: age under 54 or over 79 years; abnormal anatomical configuration of neck and muscles, marked tortuosity and/or depth of the carotid vessels and/or uncommon location of arterial branches; personal history of myocardial infarction (MI), angina pectoris, stroke, transient ischemic attack, aortic aneurysm, intermittent claudication, surgical revascularization in carotid, coronary or peripheral arterial territories, congestive heart failure (III-IV NYHA Class); and history of serious medical conditions that might limit longevity (e.g. cancer)(155).

4.3.3 Evaluation of metalloproteinase levels

The protocols of the three MMPs studied were:

Serum levels of MMP-1, measured in the fasting state, were determined by an enzyme-linked immunosorbent assay (ELISA) (R&D Systems, Minneapolis, MN, USA) in the University of Maryland Cytokine Core Laboratory in Baltimore, Maryland, USA. MMP-1 levels were measured in duplicate, and the mean values of duplicates were used for data analysis. Because study participants were recruited over three years, MMP-1 levels were measured in five different batches throughout the study period, and the batch effect was included in the regression model using dummy variables. The detection range of MMP-1 values was between 0.16 and 10 ng/ml, and the intra-assay coefficient of variation was 7.5%. MMP-1 values that were above (n = 33) and below (n = 2) the detection range were assigned the maximum and minimum values of the detection range, respectively.

MMP-8 was measured from serum samples of the Corogene and FINRISK 1997 subjects with a time-resolved immunofluorometric assay (Medix Biochemica, Kauniainen, Finland) according to the manufacturer's instructions. The MMP-8 concentrations were log-transformed before analysis because they were not normally distributed.

Plasma concentrations of MMP12 were measured at baseline using the Olink ProSeek CVD I array (Olink Proteomics, Uppsala, Sweden), according to the

standard protocol (156). The interplate coefficient of variation for the MMP12 assay was 14%, as estimated using a pooled plasma control across all plates. The measurements were carried out in 3394 IMPROVE participants in whom genotype information post-quality control (QC) was also present. Plasma MMP12 activity was measured in duplicates in 20 plasma samples from the IMPROVE cohort. Ten samples were randomly chosen from the 75th to 95th percentile of plasma MMP12 concentration and ten from the 5th to 25th percentile. The activity was measured with the Fluorimetric Sensolyte 520 MMP12 Assay Kit (Anaspec), which specifically detects elastin degradation products (desmosine).

4.3.4 Statistical Analysis

The main MR method was inverse variance weighted(157). We applied the Benjamini-Hochberg procedure to control the false positive rate. Horizontal pleiotropy was assessed using Egger regression(158) and heterogeneity was analyzed with Cochran's Q statistic. Complementary MR approaches were applied: MR-Egger, weighted median, penalized weighted median, and weighted mode approaches (159). Finally, when significant MR results were found, we used the MR-PRESSO outlier test(160) and the leave-one-out analysis to explore the presence of outliers that could bias the results and performed the analysis extracting them.

Statistical analysis was conducted in R using TwoSampleMR, version 0.4.22(161). The methodology followed in this study was under the latest guidelines for Mendelian randomization studies(162). A checklist of questions to consider for MR studies can be found in Table 5 as a sensitivity analysis of the methodology.

TABLE 5. Checklist extracted from Burgess S *et al* 2020 (162)**.**

Checklist of questions to consider for Mendelian randomization studies (162)
1. What is the primary hypothesis of interest?
Serum/plasma levels of MMPs are causally associated with ischemic stroke risk and long-term outcome.
2. Data sources
We performed two-sample MR analyses. Summary-level GWAS data from European ancestry studies evaluating MMP-1(152) serum levels, MMP-8(163) serum levels, MMP-12(155) plasma levels, modified Rankin scale 3 months post-stroke(126) and ischemic stroke risk and subtypes (68) were used. No sample overlap is expected after the evaluation of the cohorts described in each study.
3. Selection of genetic variants
Genetic variants were selected reaching a p-value < 5×10^{-8} at a clump of r^2 < 0.2 from summary-level GWAS data. When pleiotropy was detected MR-PRESSO outliers test was performed to detect and remove the pleiotropic variants.
4. Variant harmonization
The correct orientation of the variants' alleles was performed for all studies.
5. Primary analysis
The main MR method was inverse variance weighted (IVW). We applied the Benjamini-Hochberg procedure to control the false positive rate.
6. and 7. Supplementary and sensitivity analyses
Horizontal pleiotropy was assessed using Egger regression (158) and heterogeneity was analyzed with Cochran's Q statistic. Complementary MR approaches were applied: MR-Egger, weighted median, penalized weighted median, and weighted mode approaches (158). Finally, when significant MR results were found, we used the MR-PRESSO outlier test(160) and the leave-one-out analysis to explore the presence of outliers that could bias the results and performed the analysis extracting them.
8. Data presentation
The results of the IVW MR primary analysis are presented in a table. For the significant and robust results, we performed scatterplots shown in the figure of the results section. Appendix material and sensitivity analyses are presented in the appendix subsection in the form of tables but leave-one-out tests that are shown in forest plots.
9. Interpretation
A significant causal association was considered to exist when the adjusted p-value (q-value) was lower than 0.05, complementary methods showed the same direction of the association, no pleiotropy or heterogeneity was detected and no single SNP was driving the causal association.

5. Results

5. Results

5.1. Genome-Wide association Studies in the GENERACION cohort

5.1.1 Introduction

Although several GWAS analyses have been performed in different populations to uncover genomic risk factors of ischemic stroke risk, none of them have been performed in the Spanish population. This is important because there is space to test if there are specific risk loci associated with stroke risk and subtypes in the Spanish population.

Following a similar premise, in 2012, it was conducted a GWAS analysis(164) in the Australian population, although the ancestry of these individuals corresponds to European, the fact that is a more homogenous population made possible the revealed of a locus associated with LAA that could be replicated in a more heterogenous European population such as GIGASTROKE. We aimed to find new genetic risk factors for IS and subtypes following a strategy of analysis focused on homogeneous specific populations and replication in large international cohorts. As well as, to evaluate the known genomic risk factors in the Spanish population.

Additionally, in our study, we want to shed light on the study of stroke genetics by sex stratification, which is essential to understand the differences according to sex that unfortunately has not been explored yet. In a recent metanalysis(165), it has been described the sex disparities in cSVD severity, with more males having moderate to severe cSVD. In this study, authors highlight the unavailability of data stratified by sex, that could have been used to perform stratified metanalyses to better understand the nature of cSVD in males and females. We hypothesize that some differences between ischemic stroke male and female patients could be explained by differences in genomic risk factors.

5.1.2 Results

5.1.2.1 The GENERACION dataset

After quality controls, a total of 9,111 individuals were retained (Figure 4), 3519 IS patients and 5592 controls. According to TOAST IS etiologies: 1530 CE, 553 LAA, 228 SVO, 985 UE, and 99 OE. The clinical characteristics of the cohort are described in Table 6.

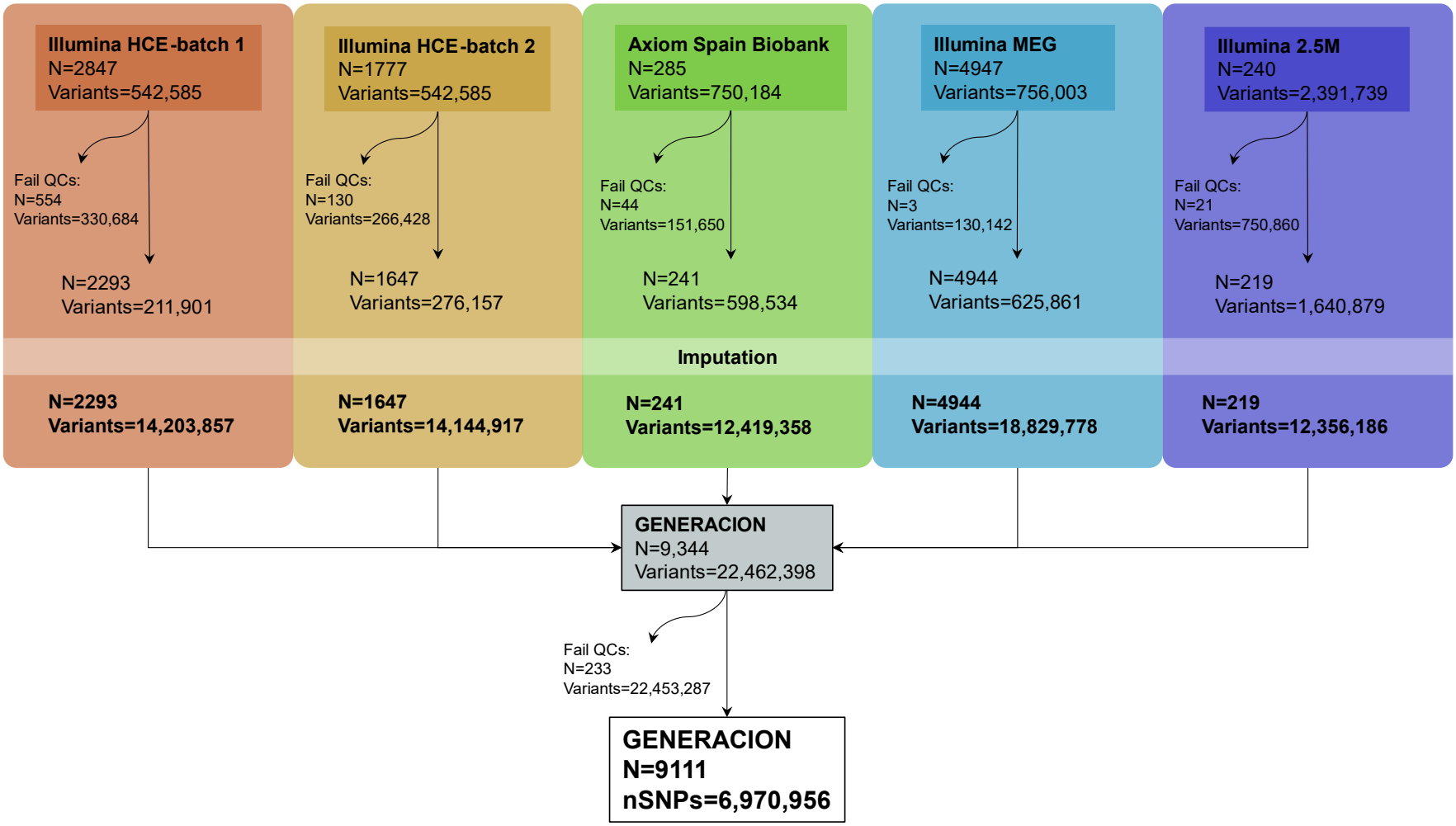


FIGURE 4. Flowchart of sample processing and quality control for variants and samples of the GENERACION cohort.

TABLE 6. Clinical findings and univariate analysis of the GENERACION cohort.

	Total (n=9111)	**Controls** (n=5592)	**IS** (n=3519)	**P**
Sex (% female)	4676 (51)	3091 (55)	1585 (45)	$< 2.2\times10^{-16}$
DM (%)	1036 (25)	112 (16)	924 (26)	1.13×10^{-08}
HTN (%)	3610 (40)	1261 (23)	2349 (67)	$< 2.2\times10^{-16}$
DL (%)	1013 (43)	263 (50)	750 (41)	3.47×10^{-04}
AF (%)	1120 (13)	25 (0)	1095 (31)	$< 2.2\times10^{-16}$
TOAST(%)				
CE			1530 (45)	
LAA			553 (16)	
SVO			228 (7)	
UE			985 (29)	
OE			99(3)	
Age (Years. IQR)	58 (49-72)	52 (46-59)	75 (66-82)	$< 2.2\times10^{-16}$

DM: Diabetes Mellitus; DL: Dyslipidemia; AF: Atrial Fibrillation; CE: Cardioembolic Stroke; LAA: Large Artery-Atherosclerosis Stroke; SVO: Small Vessel Occlusion; UE: Undetermined Etiology; OE: Other Etiology; TOAST: Trial of Org 10172 in Acute Stroke Treatment.

5.1.2.2 Genetic analysis and annotation

Twelve different loci were found significantly associated in the GWAS of stroke analyses (Table 7). V2G prioritization revealed 9 potentially regulated genes: *RSU1*, *XPC*, *TIPARP*, *FGFR1*, *SERPINE3*, *PTPRT*, *STPG2*, *LSM12*, and *OLFML3*. Additionally, one locus was found significant in the AF absent vs presence analyses, accounting for the *PITX2* locus.

Manhattan plots and QQ plots for all the analyses can be observed in Figures 5-10. No remarkable genomic inflation was detected in each analysis with λ values of approximately 1 in all cases.

TABLE 7. Genome-wide significant loci in GENERACION.

Locus	Analysis	ID	Nearest Gene	Gene V2G	CHR	BP	EA	OA	AF	BETA	SE	P
10p13	IS	rs7092141	*C1QL3*	*RSU1*	10	16711372	C	T	0.033	0.067	0.012	2.85×10^{-08}
3p25.1	IS	rs184652834	*NR2C2*	*XPC*	3	14982761	G	T	0.014	0.121	0.019	8.75×10^{-11}
3p25.1	IS only females	rs184652834	*NR2C2*	*XPC*	3	14982761	G	T	0.015	0.142	0.023	6.61×10^{-10}
13p21.1	LAA	rs7328431	*PCDH17*	-	13	58824830	C	T	0.281	0.025	0.005	2.35×10^{-08}
3p25.1	CE only females	rs184652834	*NR2C2*	*XPC*	3	14982761	G	T	0.014	0.132	0.024	2.88×10^{-08}
5p15.2	SVO	rs59970332	*CTNND2*	-	5	12639845	T	C	0.013	0.079	0.014	1.99×10^{-08}
3q25.31	SVO	rs4680292	*TIPARP*	*TIPARP*	3	156353040	G	A	0.020	0.062	0.011	4.52×10^{-08}
8p11.22	SVO only females	rs146966463	*TACC1*	*FGFR1*	8	38459179	C	T	0.015	0.085	0.015	6.60×10^{-09}
5p14.1	SVO only females	rs184186520	*CDH9*	-	5	28866208	T	C	0.013	0.089	0.016	9.48×10^{-09}
13q14.3	SVO only females	rs117761762	*SERPINE3*	*SERPINE3*	13	51900684	G	C	0.028	0.060	0.011	1.19×10^{-08}
20q13.11	SVO only females	rs78101684	*PTPRT*	*PTPRT*	20	41750826	A	G	0.012	0.089	0.016	2.47×10^{-08}
4q22.3	SVO only females	rs147605147	*STPG2*	*STPG2*	4	98274010	G	A	0.010	0.098	0.018	3.27×10^{-08}
17q21.31	SVO only females	rs78501768	*LSM12*	*LSM12*	17	42130551	A	G	0.017	0.075	0.014	3.83×10^{-08}
1p13.2	SVO only females	rs181211009	*SYT6*	*OLFML3*	1	114715420	T	C	0.030	0.057	0.010	4.65×10^{-08}
4q25	AF	rs59788391	*PITX2*	*PITX2*	4	111701433	G	A	0.144	0.040	0.006	2.03×10^{-10}
4q25	AF only males	rs13143308	*PITX2*	*PITX2*	4	111714419	T	G	0.234	0.046	0.007	5.06×10^{-10}

V2G: Variant to Gene; CHR: Chromosome; BP: Base Pairs; EA: Effect Allele; OA: Other Allele.

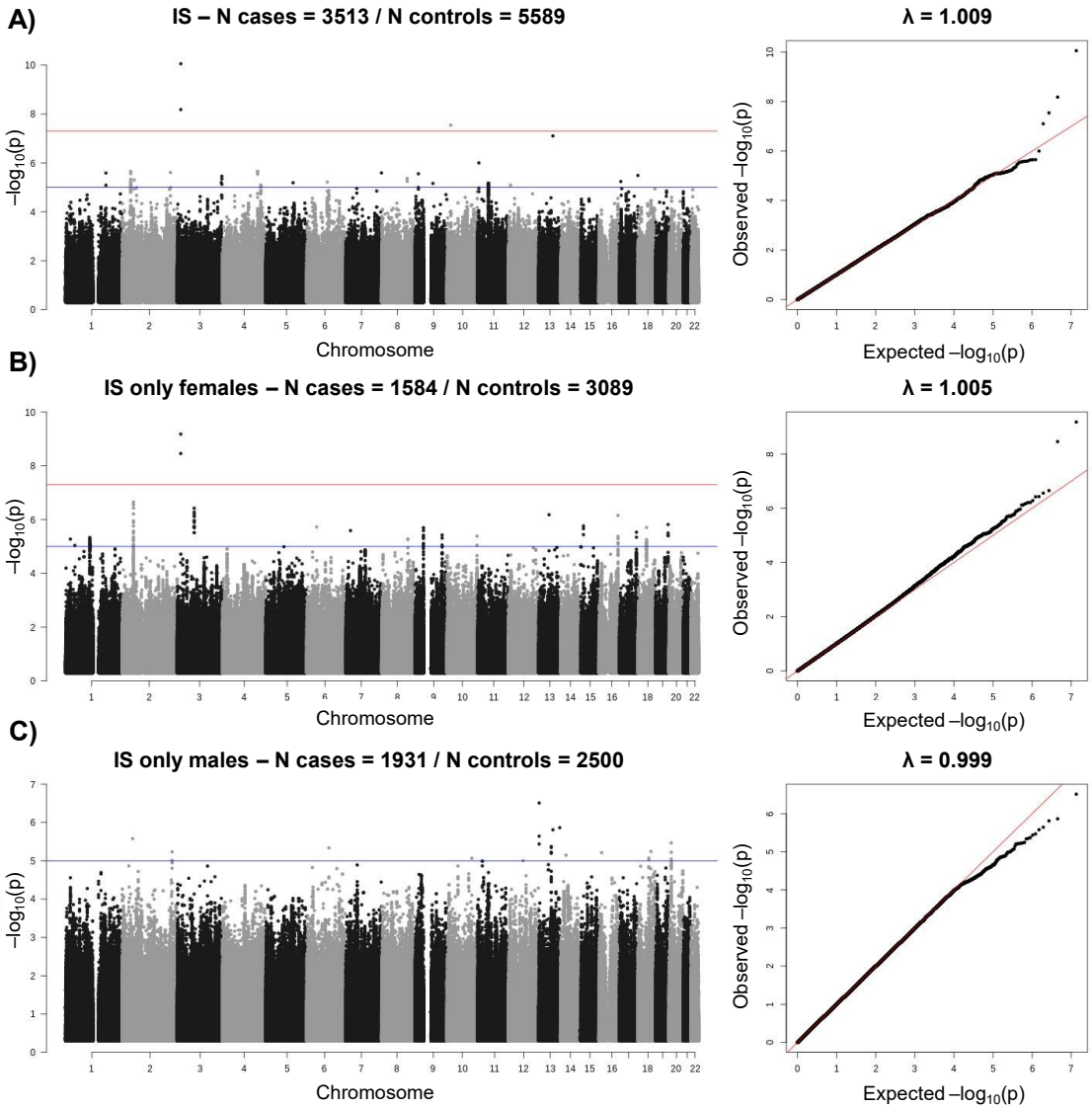


FIGURE 5. Manhattan and QQ plots of ischemic stroke analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. IS: Ischemic Stroke.

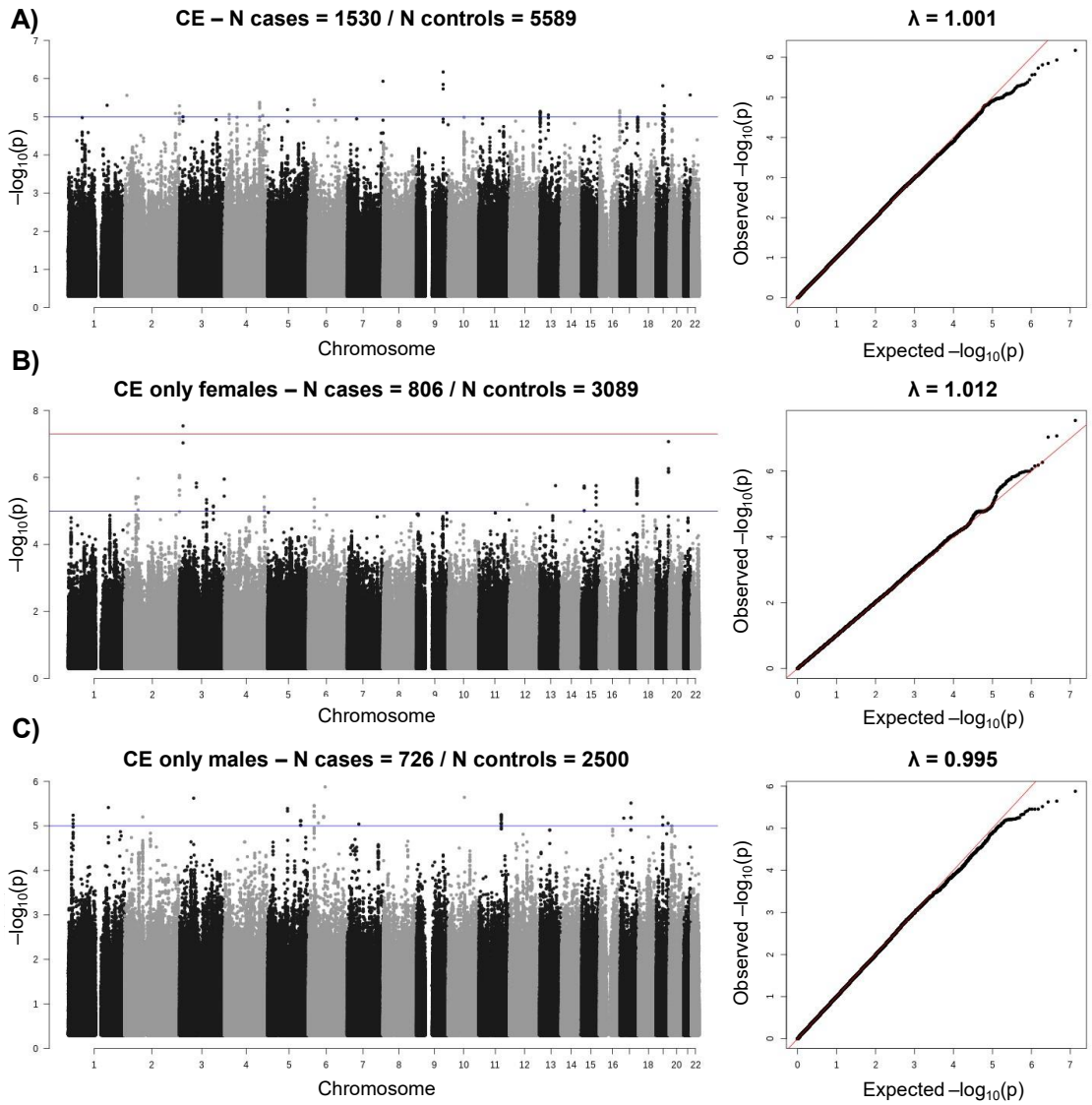


FIGURE 6. Manhattan and QQ plots of cardioembolic stroke analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. CE: Cardioembolic Stroke.

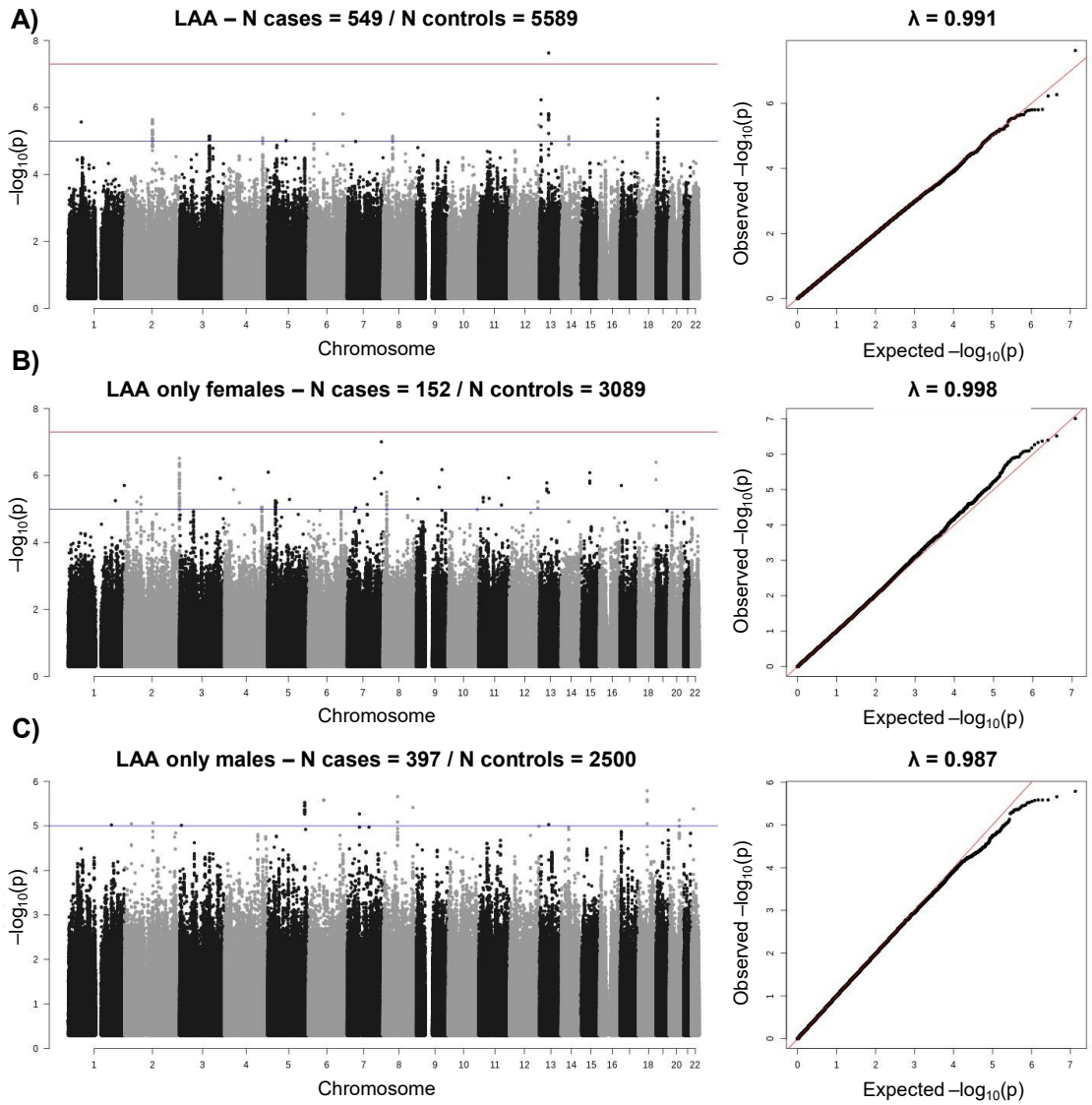


FIGURE 7. Manhattan and QQ plots of large-artery atherosclerosis stroke analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. LAA: Large-Artery Atherosclerosis Stroke.

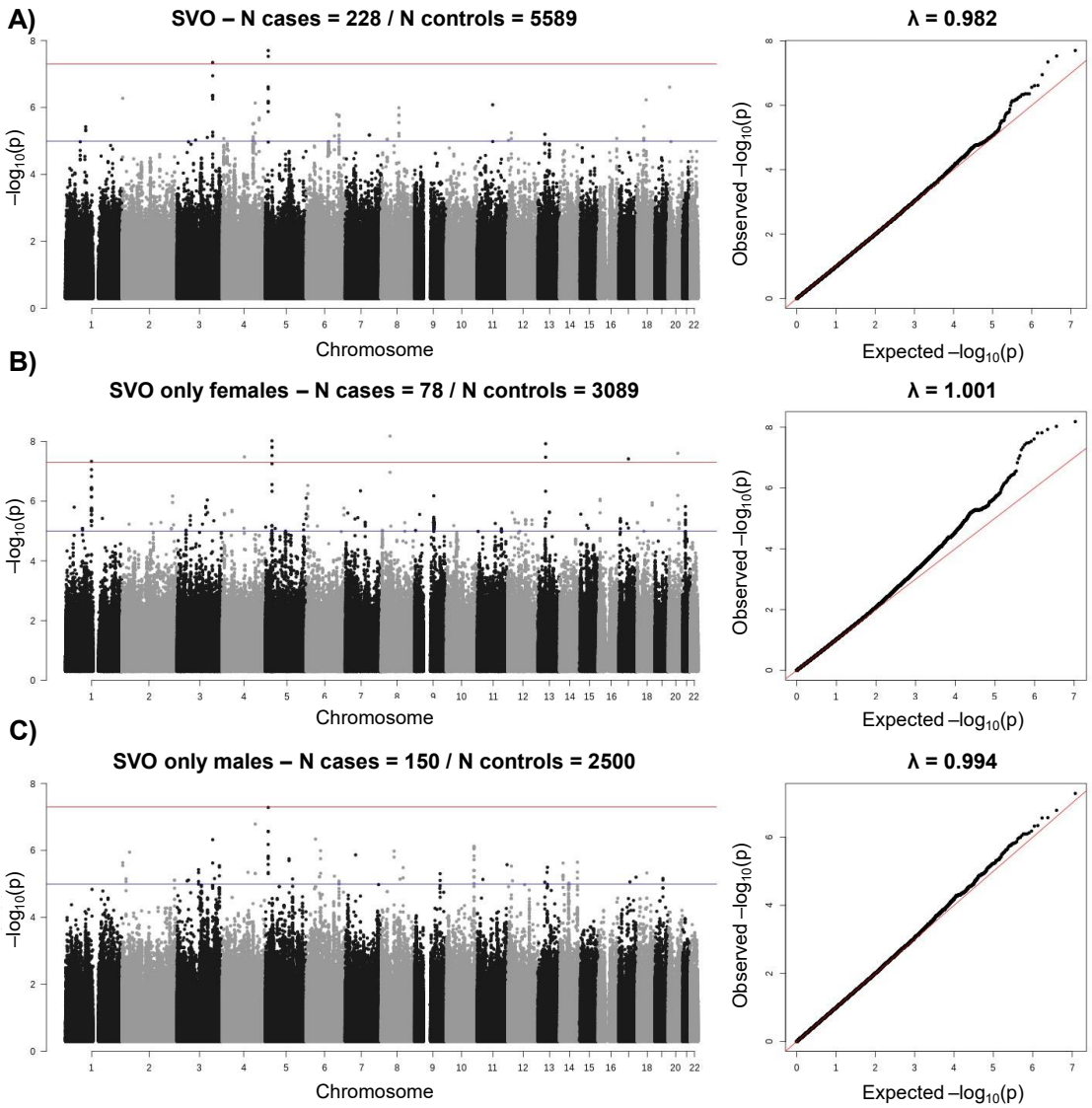


FIGURE 8. Manhattan and QQ plots of small vessel occlusion analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. SVO: Small Vessel Occlusion.

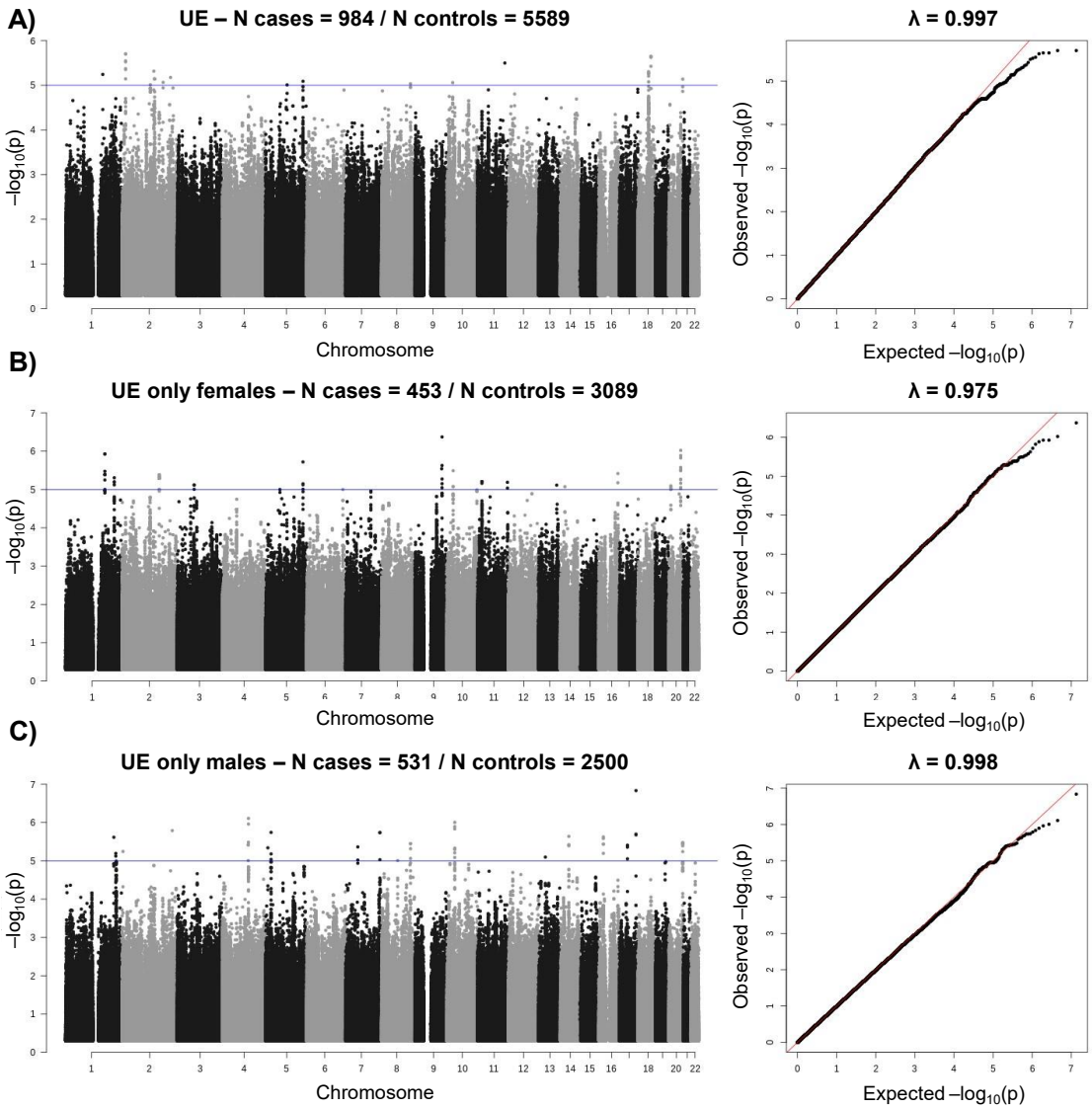


FIGURE 9. Manhattan and QQ plots of undetermined etiology analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. UE: Undetermined Etiology.

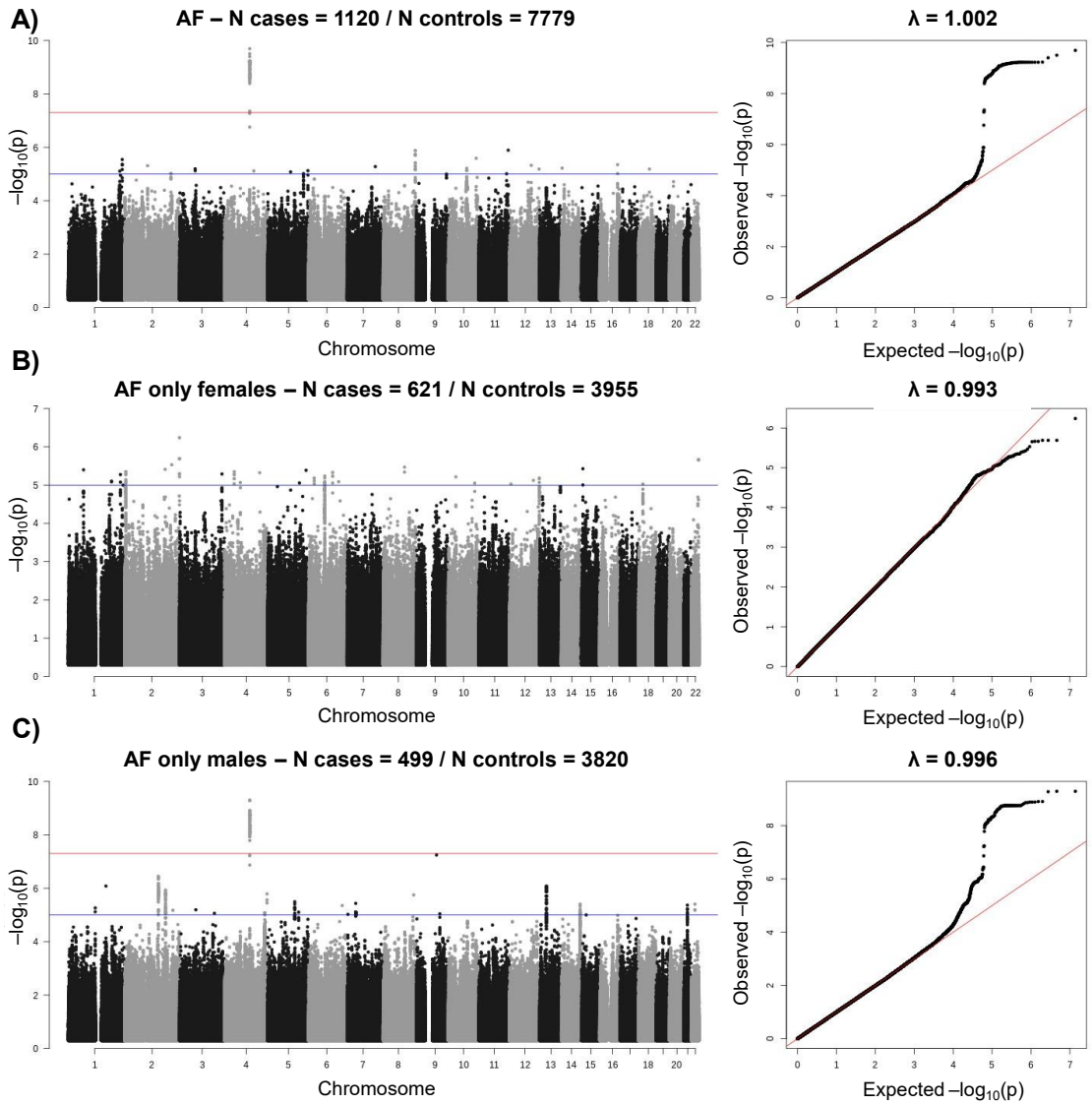


FIGURE 10. Manhattan and QQ plots of atrial fibrillation analysis. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. AF: Atrial Fibrillation.

5.1.2.3 Gene-based analyses

A different number of genes were analyzed for each phenotype, ranging from 18,206 to 18,509 genes. No genes were found significant after Bonferroni correction (p-value < 2×10^{-6}). The top 10 genes are mapped in the Manhattan plots of gene-based analyses (Figures 11-15).

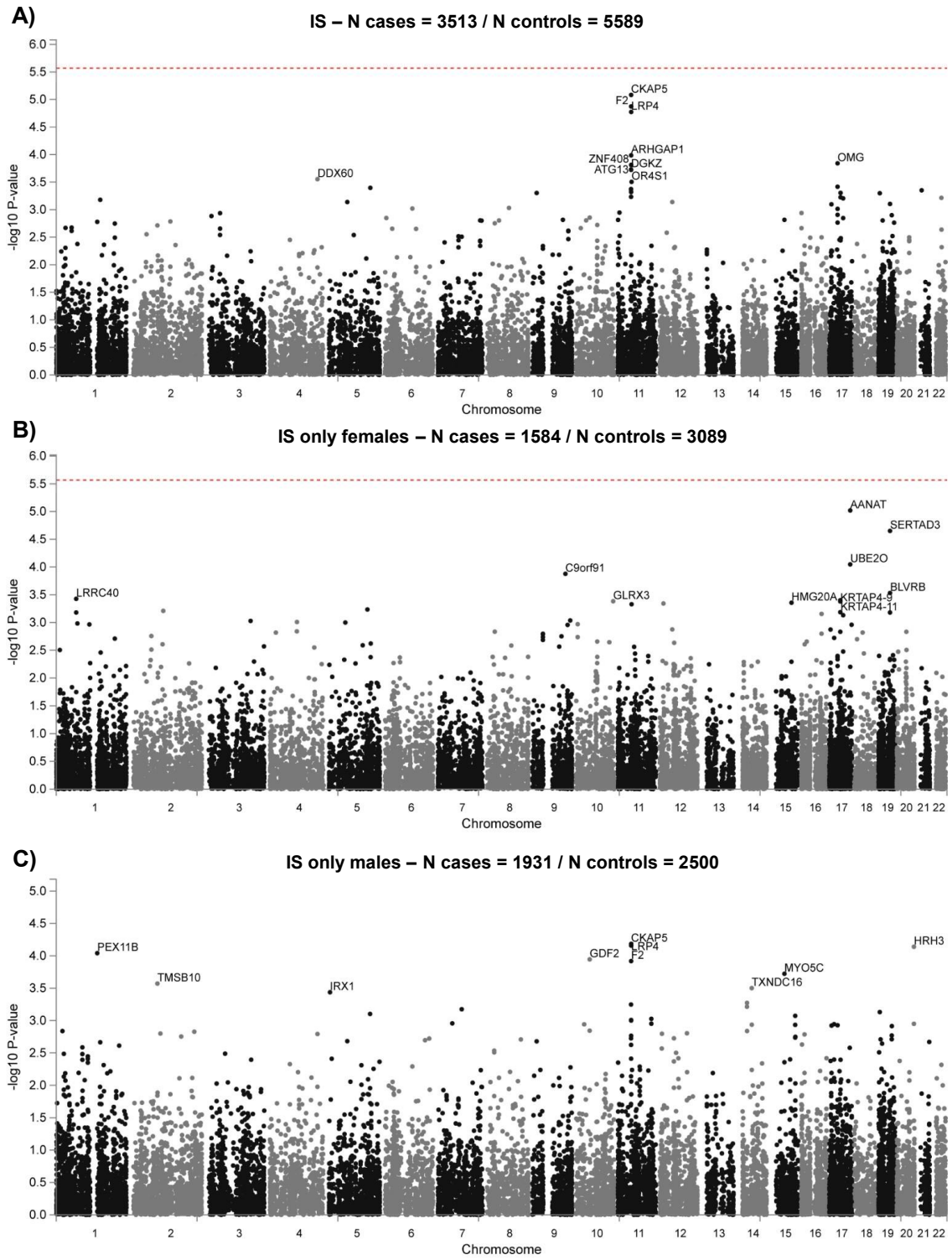


FIGURE 11. Manhattan plot of gene-based analysis of ischemic stroke. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. IS: Ischemic Stroke. The top ten significant genes are marked by gene symbol names. The Red dashed line represents the threshold of p-value = 2×10^{-6}.

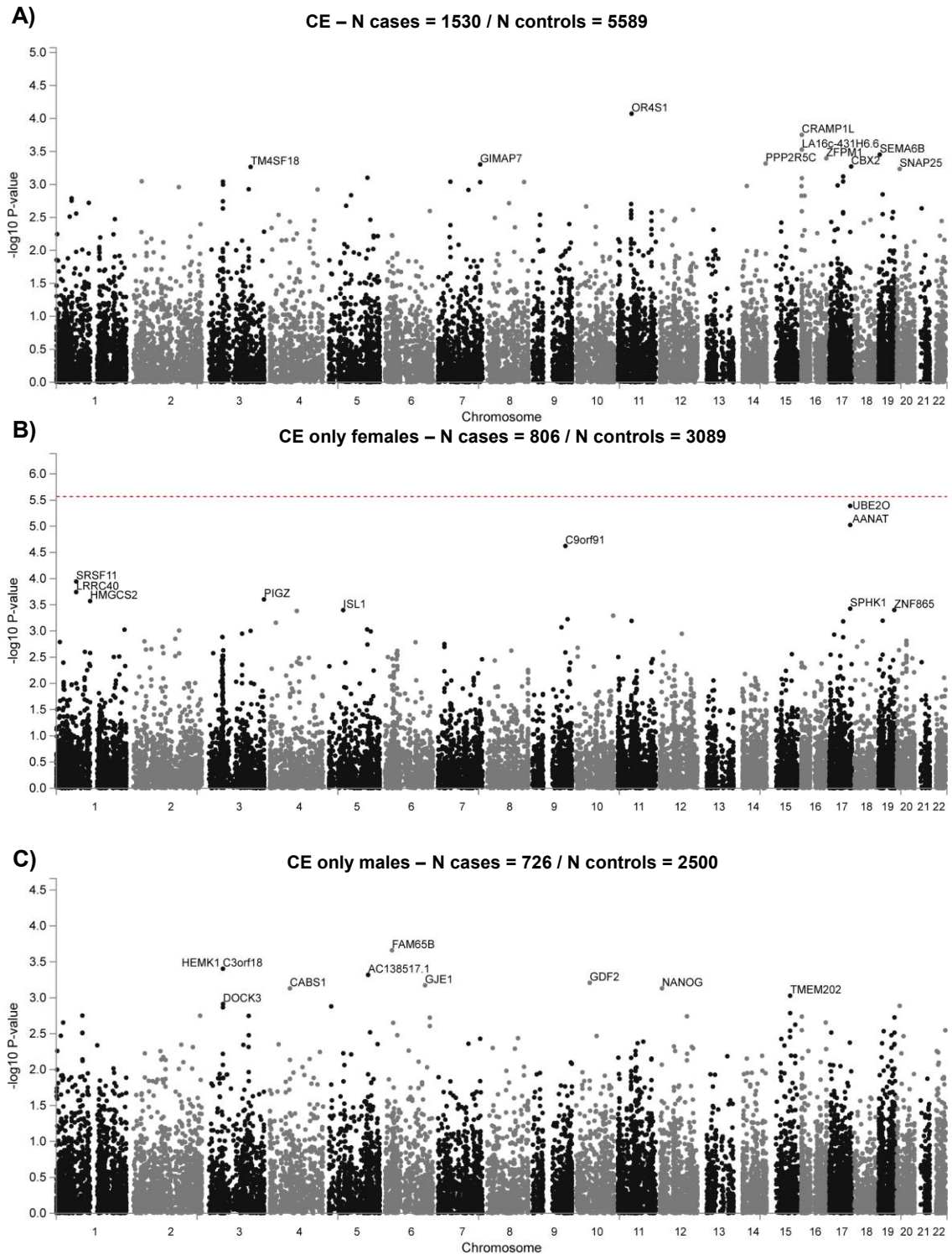


FIGURE 12. Manhattan plot of gene-based analysis of cardioembolic stroke. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. CE: Cardioembolic stroke. The top ten significant genes are marked by gene symbol names. The Red dashed line represents the threshold of p-value = 2×10^{-6}.

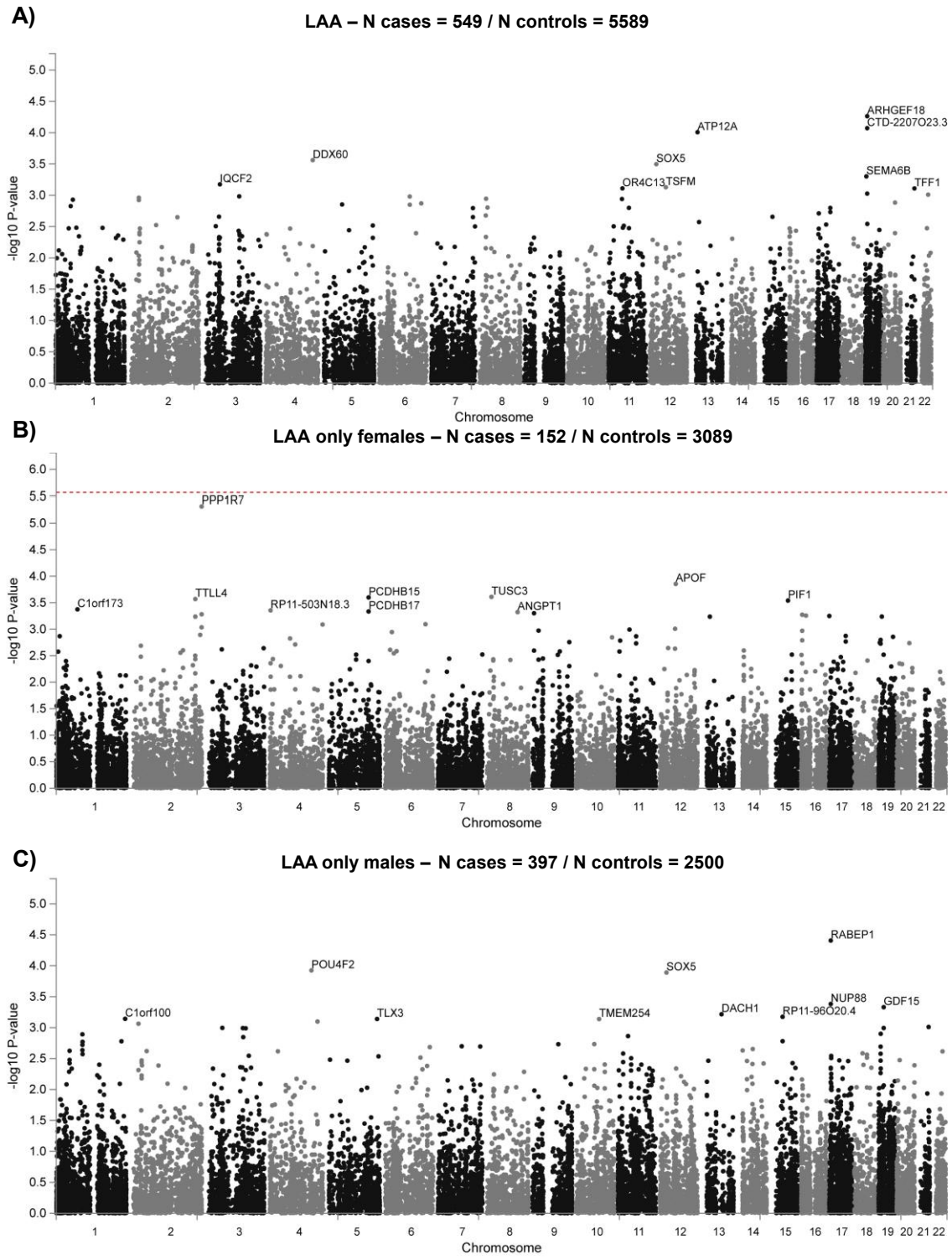


FIGURE 13. Manhattan plot of gene-based analysis of large-artery atherosclerosis stroke. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. LAA: Large-Artery Atherosclerosis stroke. The top ten significant genes are marked by gene symbol names. The Red dashed line represents the threshold of p-value = 2×10^{-6}.

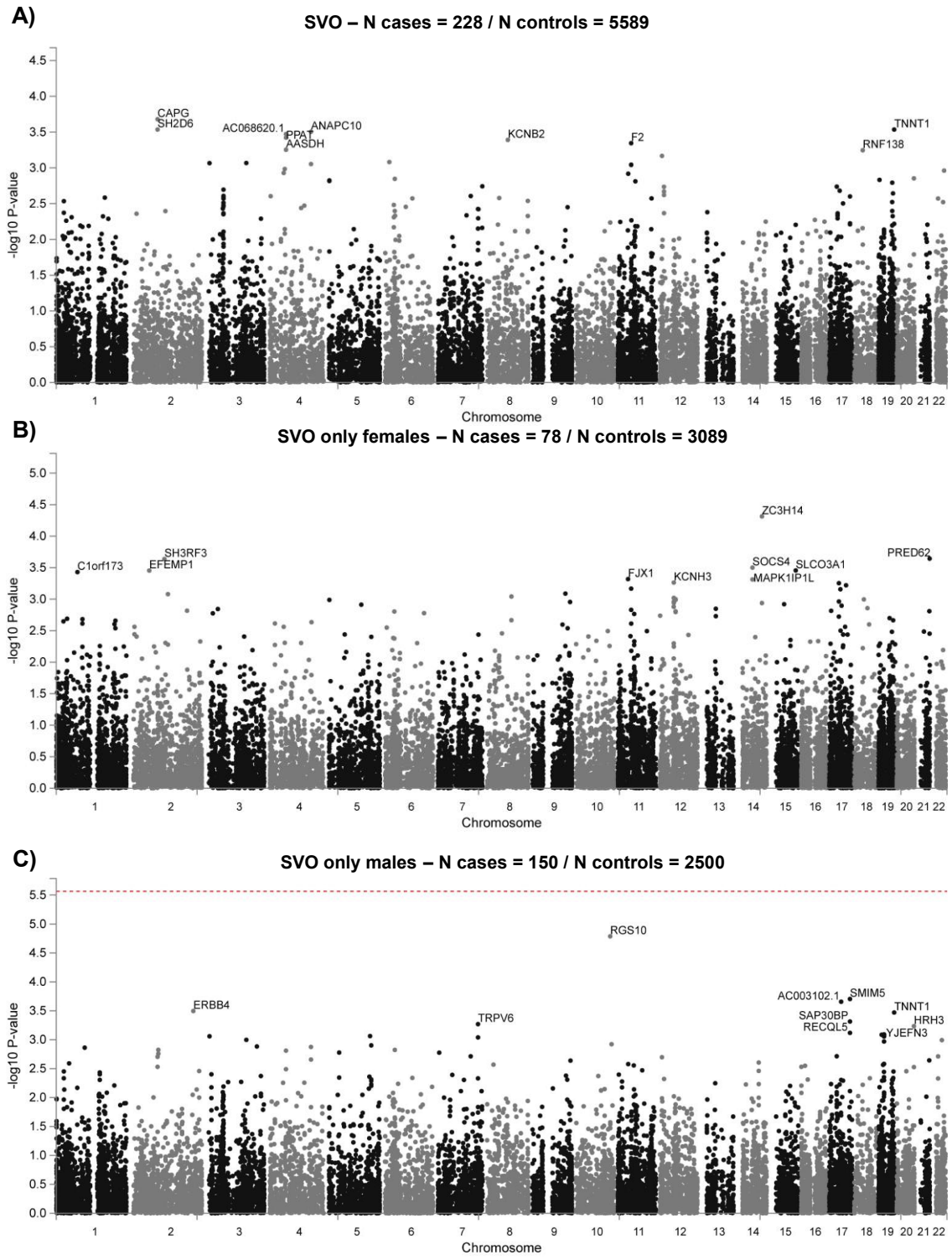


FIGURE 14. Manhattan plot of gene-based analysis of small-vessel occlusion stroke. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. SVO: Small-Vessel Occlusion stroke. The top ten significant genes are marked by gene symbol names. The Red dashed line represents the threshold of p-value = 2×10^{-6}.

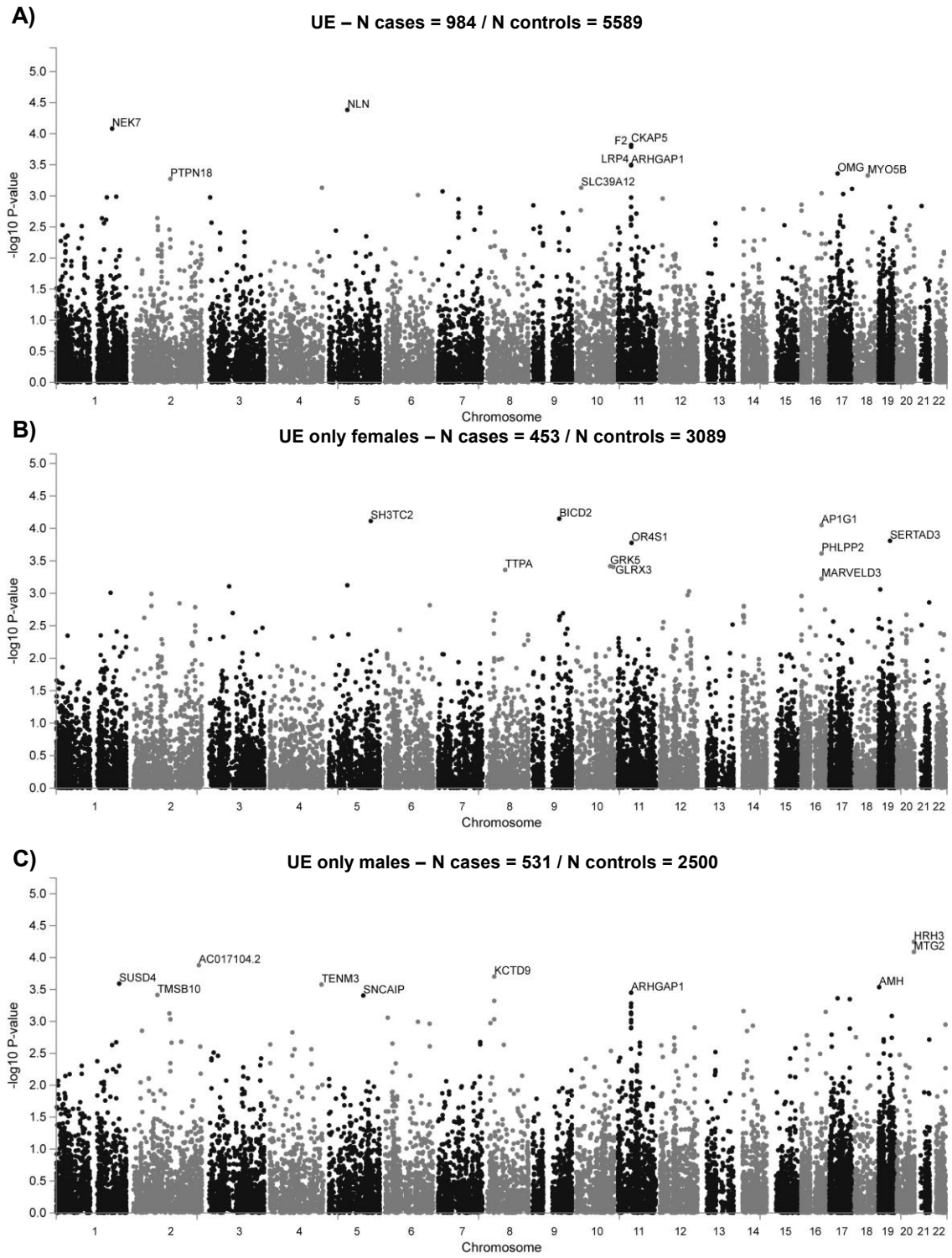


FIGURE 15. Manhattan plot of gene-based analysis of undetermined etiology stroke. A) Both sexes analysis, B) Only females' analysis, C) Only males' analysis. UE: Undetermined Etiology stroke. The top ten significant genes are marked by gene symbol names.

5.1.2.4 Replication of significant loci

Two loci were found replicated (p-value < 0.05) in MEGASTROKE-SVO data (Table 8), accounting for locus 5p15.2 found significant in SVO both sexes analysis closest gene *CTNND2* (Table 7), and locus 8p11.22 found significant in SVO only females' analysis prioritized gene *FGFR1* (Table 7). With 5p15.2 locus, we found that the strength of the association was mostly driven by males SVO, the top variant in the SVO-only males' analysis: rs59970332-T: beta (se) = 0.135 (0.025), p-value = 5.19×10^{-8}. These two replicated in MEGASTROKE-SVO analysis were not present in the GIGASTROKE-SVO analyses. The other ten loci had a p-value > 0.05 for the index variant and/or of variants in high LD in 1000G European data (r^2>0.8) in MEGASTROKE and GIGASTROKE analyses.

TABLE 8. Replicated loci in MEGASTROKE-SVO data.

Analysis	SNP	A1	A2	AF1	B	SE	P	Z_m	P_m
SVO only females	**rs146966463**	C	T	0.015	0.09	0.02	6.60×10^{-09}	2.7	**0.007**
SVO	rs141286387	C	T	0.012	0.07	0.02	1.32×10^{-06}	2.34	**0.019**
SVO	rs112512503	C	T	0.012	0.07	0.01	7.05×10^{-07}	2.35	**0.019**
SVO	rs113371540	G	C	0.012	0.08	0.02	2.42×10^{-07}	2.36	**0.019**
SVO	rs111342433	T	C	0.012	0.07	0.02	7.25×10^{-07}	2.27	**0.023**
SVO	rs187154912	A	G	0.012	0.07	0.02	7.25×10^{-07}	2.27	**0.023**
SVO	rs113414410	T	C	0.012	0.07	0.02	6.58×10^{-07}	2.06	**0.04**
SVO	rs138814010	T	C	0.012	0.07	0.01	2.76×10^{-07}		
SVO	rs147550316	C	G	0.013	0.07	0.01	6.57×10^{-07}		
SVO	rs113435657	G	T	0.012	0.08	0.01	2.96×10^{-08}		
SVO	**rs59970332**	T	C	0.013	0.08	0.01	1.99×10^{-08}		

Bold SNPs are those lead variants for loci 8p11.22 and 5p15.2. AF1: Allele frequency of allele 1; A1: Allele 1; A2: Allele 2; B: Beta effect of A1; SE: Standard Error; SNP: Single Nucleotide Polymorphism; P: p-value; P_m: p-value for SVO analysis in MEGASTROKE cohort; SVO: Small-Vessel Occlusion; Z_m: Z effect for SVO analysis in MEGASTROKE cohort.

5.1.2.5 Proteomic and pathway analysis

A total of 98 different proteins were found to be regulated (p-value < 0.05) in plasma by rs59970332-T (Table 9) and 127 by rs146966463-C (Table 10).

TABLE 9. Top ten proteins associated with rs59970332-T.

Protein	B	SE	P
Immediate early response 3-interacting protein 1	-0.453	0.130	4.79×10^{-04}
Protein DEPP	-0.432	0.130	8.71×10^{-04}
HEPACAM family member 2	-0.382	0.130	3.24×10^{-03}
Protein phosphatase 1 regulatory subunit 3B	-0.381	0.130	3.31×10^{-03}
Delta-like protein 1	-0.368	0.130	4.57×10^{-03}
Retinol dehydrogenase 16	-0.368	0.130	4.57×10^{-03}
T-cell surface protein tactile	-0.368	0.130	4.57×10^{-03}
Probable inactive ribonuclease-like protein 13	-0.356	0.130	6.17×10^{-03}
40S ribosomal protein SA	-0.346	0.130	7.59×10^{-03}
Peroxisomal membrane protein PEX14	0.345	0.130	7.76×10^{-03}

Beta value is indicated for allele T, risk allele of SVO analysis. B: beta value; SE: Standard Error; P: P-value.

TABLE 10. Top ten proteins associated with rs146966463-C.

Protein	B	SE	p
Transmembrane protease serine 11D	0.311	0.098	1.48×10^{-03}
Macrophage mannose receptor 1	-0.308	0.098	1.58×10^{-03}
Interleukin-17 receptor A	-0.297	0.098	2.34×10^{-03}
Engulfment and cell motility protein 1	0.292	0.098	2.75×10^{-03}
Tumor necrosis factor receptor superfamily member 1B	-0.293	0.098	2.75×10^{-03}
Zinc finger protein 175	0.290	0.098	2.95×10^{-03}
Cytokine receptor-like factor 2	-0.283	0.098	3.72×10^{-03}
Sialic acid-binding Ig-like lectin 12	0.281	0.098	4.07×10^{-03}
Calcium-binding protein 8	0.272	0.098	5.37×10^{-03}
Kremen protein 1	0.270	0.098	5.75×10^{-03}

Beta value is indicated for allele C, risk allele of SVO only females' analysis. B: beta value; SE: Standard Error; P: P-value.

In ORA analyses no biological process was found significantly associated after FDR correction. Some nominally associated biological processes modulated by rs59970332-T involve response to nutrients and cytokine production involved in inflammatory response (Figure 16). In the same way, some nominally associated biological processes modulated by rs146966463-C correspond to blood vessel remodeling and regulation of cell aging among others (Figure 17).

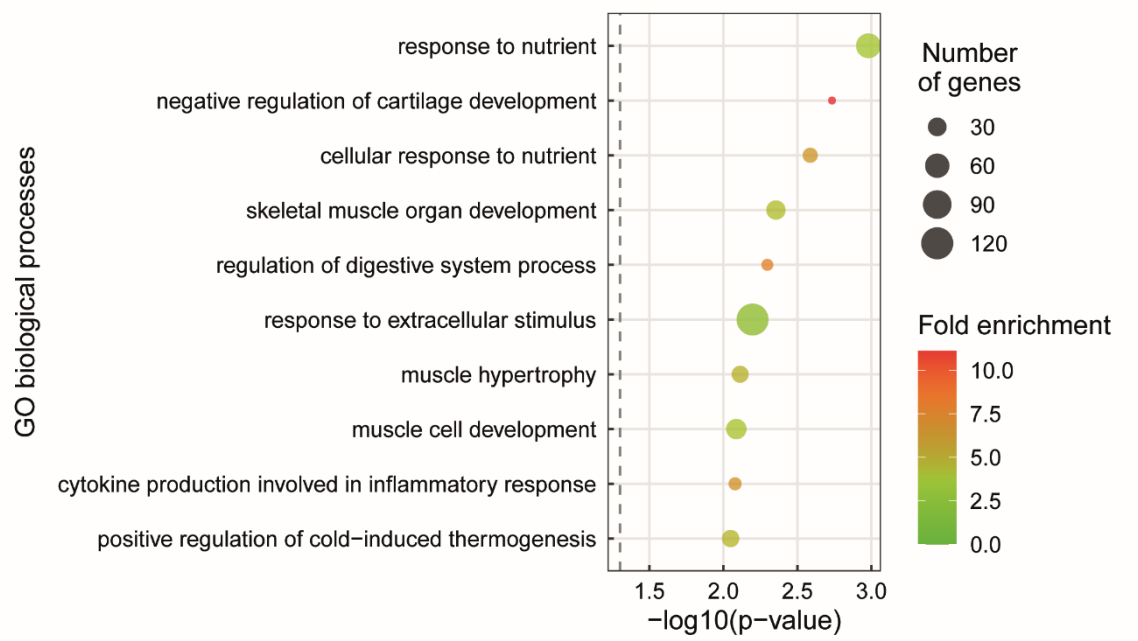


FIGURE 16. Top ten biological processes predicted to be altered by proteins modulated by rs59970332-T in plasma. GO: Gene Ontology.

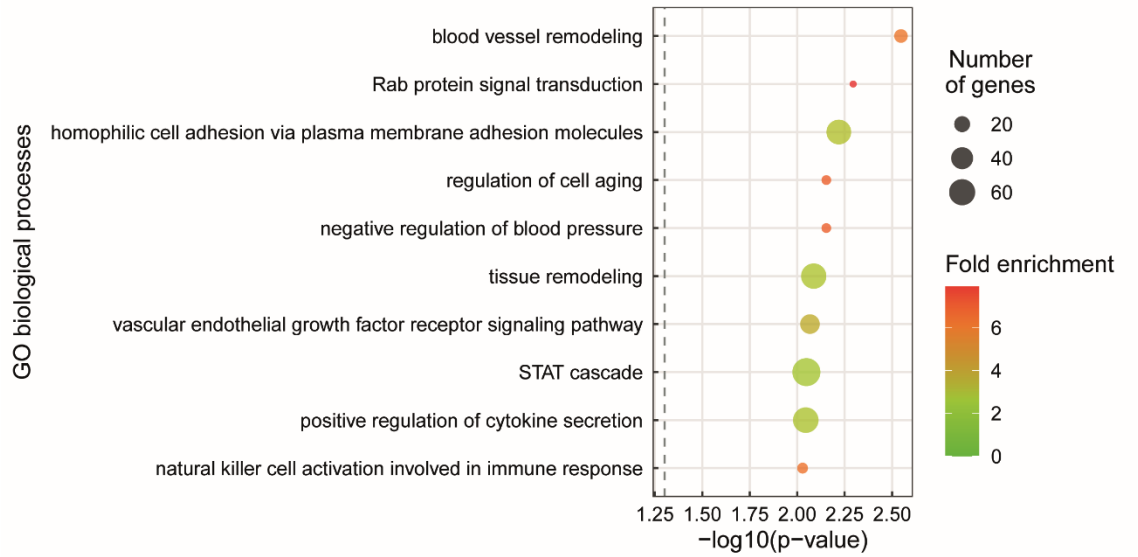


FIGURE 17. Top ten biological processes predicted to be altered by proteins modulated by rs146966463-C in plasma. GO: Gene Ontology.

Regarding the GSEA results, for proteins modulated by rs59970332-T no biological processes was found significant after FDR correction. Some nominally significant processes correspond to amyloid-beta clearance and microtubule polymerization among others (Figure 18). Concerning the GSEA for proteins modulated by rs146966463-C, only production of interleukin 5 (IL-5) biological process was found to be significantly decreased after FDR correction (p-value <

2x×10^{-6}; q-value = 0.01). Other nominally associated processes include the amyloid-beta process and macrophage activation (Figure 19).

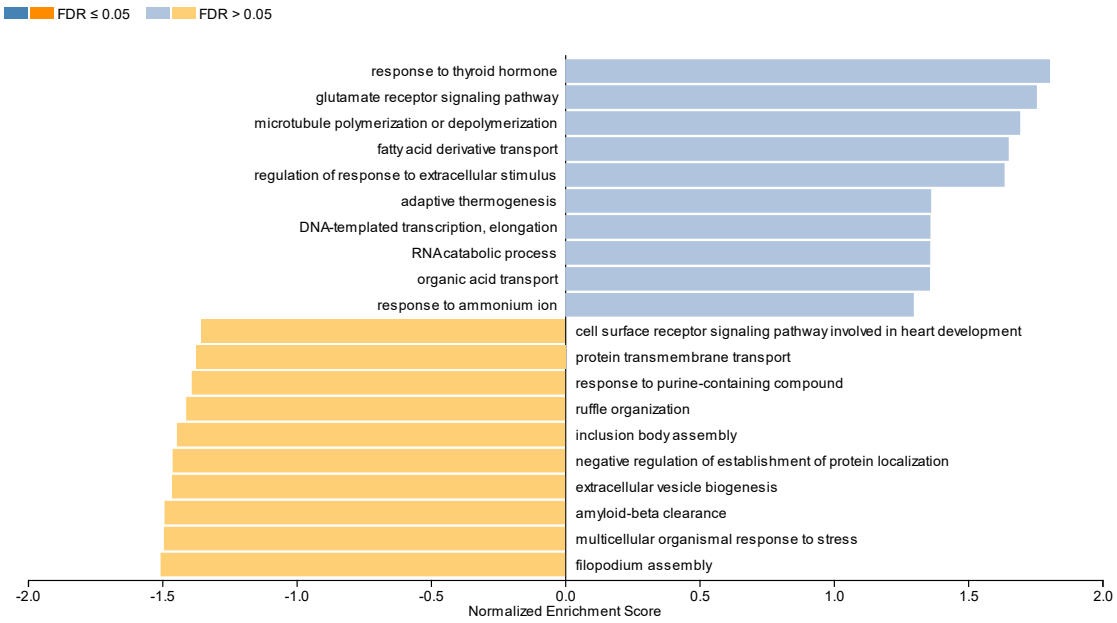


FIGURE 18. GSEA results for proteins associated with rs59970332-T, associated with small vessel occlusion risk. FDR: False Discovery Rate.

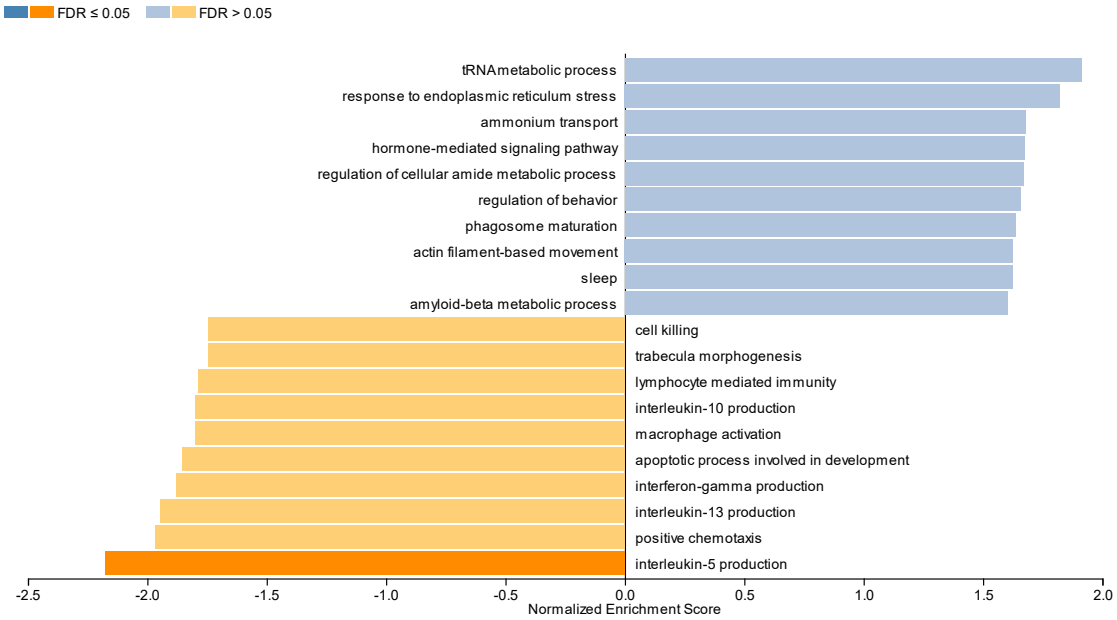


FIGURE 19. GSEA results for proteins associated with rs146966463-C, associated with small vessel occlusion risk. FDR: False Discovery Rate.

5.1.2.6 Replication of previously known loci

A total 87 independent significant loci from GIGASTROKE were evaluated for replication in the GENERACION cohort (Appendix Table A1). Eighty-one loci were present in our cohort due to imputation quality measures. Using similar directionality of effect, we were able to replicate 53 to 73 of these loci, depending on replication in any stroke subtype. Checking for both directionality of effect and nominal significance at p-value < 0.05, we observed association of 14 loci (*THADA*, *F11*, *ANKRD33*, *MESDC1*, *FGG*, *ABO*, *LIPA*, *PRDM16*, *PROCR*, *LDLR*, *OVOL1*, *WNT2B*, *ZFHX3*, *NOA1*) with at least 1 of the stroke subtypes. None of them were significant after accounting for multiple comparison adjustment (p-value < 0.05/81).

5.1.3 Appendix

TABLE A1. Significant loci from GIGASTROKE study evaluated in the GENERACION cohort.

Locus	Phenotype	SNP	EA	OA	OR	P	OR IS	P IS	OR LAA	P LAA	OR CE	P CE	OR SVO	P SVO	OR UE	P UE
PRDM16	SVO, AS, IS	rs2455132	C	T	1.10	1.2×10^{-08}	1.01	0.20	1.01	0.05	1.01	0.13	1.00	0.26	1.01	0.10
PRRX1	CE	rs680084	G	A	1.10	8×10^{-11}	1.00	0.62	1.00	0.65	1.01	0.13	1.00	0.63	1.00	0.95
LAMC1	AS	rs2877984	G	A	1.03	1×10^{-08}	1.00	0.77	1.00	0.73	1.00	0.87	1.00	0.64	1.00	0.94
THADA	AS	rs6722806	A	T	1.04	2×10^{-08}	1.01	0.02	1.01	0.02	1.01	0.00	1.01	0.04	1.00	0.57
FIGN	AS	rs11691032	C	G	1.03	2.5×10^{-08}	1.00	0.93	1.00	0.94	1.00	0.95	1.00	0.85	1.00	0.34
NBEAL1	IS, AS	rs2351524	C	T	1.07	3.8×10^{-10}	1.00	0.55	1.00	0.83	1.00	0.80	1.00	0.74	0.99	0.40
3p12	IS	rs73852583	A	G	1.44	3.4×10^{-08}										
PLSCR5	AS	rs4681330	C	T	1.03	2.8×10^{-09}	0.99	0.07	1.00	0.57	0.99	0.00	1.00	0.22	1.00	0.38
FGF5	IS, AS	rs16998073	T	A	1.04	1.1×10^{-11}	1.00	0.46	1.00	0.52	1.00	0.62	1.00	0.18	0.99	0.08
F11	CE	rs4444878	A	C	1.10	1.3×10^{-10}	1.01	0.01	1.01	0.12	1.01	0.08	1.00	0.31	1.01	0.01
TAP1	IS	rs36229526	T	G	1.08	4.8×10^{-08}	0.99	0.15	0.98	0.03	1.00	0.60	1.01	0.31	0.99	0.46
CENPQ	IS, AS	rs2501966	A	G	1.04	6.2×10^{-11}	1.00	0.43	1.00	0.42	1.00	0.91	1.00	0.30	1.00	0.84
LPA	LAA	rs56393506	T	C	1.16	2.1×10^{-08}	1.00	0.47	1.00	0.37	1.01	0.19	1.00	0.80	1.00	0.58
EVX1	AS	rs7800053	G	A	1.05	9.6×10^{-09}	1.00	0.73	0.99	0.46	0.99	0.18	1.00	0.48	0.99	0.32
COBL	IS, AS	rs113916171	A	G	1.87	3.5×10^{-08}	1.01	0.22	1.00	0.79	1.01	0.37	1.01	0.17	1.01	0.27
PIK3CG	IS, AS	rs12539561	C	T	1.04	4.9×10^{-08}	1.01	0.39	1.00	0.55	1.00	0.89	1.00	0.81	1.01	0.32
THAP5	Inc_IS	rs114798023	G	A	1.29	3.2×10^{-08}	0.99	0.65	0.99	0.44	1.00	0.93	1.01	0.51	0.98	0.32
DEFB1	CE	rs2738158	G	A	1.14	6.6×10^{-09}										
EBF2	AS	rs4298492	A	G	1.04	1.7×10^{-08}	1.00	0.41	0.99	0.07	1.00	0.48	1.00	0.56	1.00	0.42
BNC2	IS	rs1487504	A	G	1.06	4.7×10^{-08}	0.99	0.43	1.01	0.48	1.00	0.85	1.00	0.43	1.00	0.85
PTCH1	SVO	rs2405068	A	G	1.54	4.6×10^{-08}	1.00	0.44	1.00	0.37	1.01	0.05	1.00	0.40	1.00	0.97
GRK5	IS, AS	rs10886430	G	A	1.08	2×10^{-11}	1.01	0.21	1.00	0.59	1.01	0.39	1.00	0.97	1.01	0.11
HTRA1	SVO, AS, IS	rs60401382	C	T	1.08	3.7×10^{-08}	0.99	0.17	1.00	0.40	0.99	0.09	1.00	0.50	1.00	0.57
LSP1	IS, AS	rs1973765	C	T	1.04	2×10^{-08}	1.00	0.81	1.01	0.10	1.00	0.69	1.00	0.20	1.00	0.42
SWAP70	IS	rs415895	G	C	1.04	7×10^{-09}	0.99	0.06	1.00	0.62	0.99	0.01	1.00	0.84	0.99	0.29

Locus	Phenotype	SNP	EA	OA	OR	P	OR IS	P IS	OR LAA	P LAA	OR CE	P CE	OR SVO	P SVO	OR UE	P UE
ANKRD33	AS	rs7973143	T	A	1.03	5.8×10^{-09}	1.01	0.02	1.00	0.74	1.01	0.02	1.00	0.19	1.00	0.35
HOXC4	IS, AS	rs12426667	A	C	1.04	1.9×10^{-09}	1.00	0.53	1.00	0.35	1.00	0.34	1.00	0.33	1.00	0.95
ATP2B1	AS	rs12579302	A	G	1.04	1.3×10^{-09}	1.01	0.12	1.01	0.08	1.00	0.47	1.00	0.38	1.01	0.20
PTPN11	IS, AS	rs7974266	T	C	1.04	1.4×10^{-09}	1.00	0.59	1.00	0.61	1.00	0.80	1.00	0.72	1.01	0.26
MESDC1	AS	rs2663905	G	A	1.03	2.3×10^{-08}	1.02	0.00	1.01	0.02	1.01	0.01	1.01	0.20	1.01	0.03
LINC00924	IS, AS	rs2397816	A	G	1.04	4.5×10^{-10}	1.00	0.88	1.00	0.68	1.00	0.52	1.00	0.92	1.00	0.62
RPRML	AS	rs2316757	A	G	1.03	1.5×10^{-08}	1.01	0.21	1.00	0.75	1.01	0.09	1.00	0.22	1.00	0.54
MTMR4	IS, AS	rs2108911	C	T	1.05	4×10^{-09}	0.99	0.26	0.99	0.19	1.00	0.36	1.00	0.85	1.00	0.55
PROCR	IS, AS	rs11907011	C	T	1.06	9.9×10^{-10}	1.01	0.08	1.01	0.05	1.01	0.25	1.00	0.81	1.00	0.65
CASZ1	IS, AS	rs880315	C	T	1.05	1.9×10^{-13}	1.00	0.31	1.00	0.71	1.00	0.56	1.00	0.32	1.00	0.30
WNT2B	IS, AS	rs3790607	C	A	1.05	8.7×10^{-11}	1.02	0.10	1.01	0.24	1.02	0.05	1.00	0.72	1.02	0.05
PMF1	SVO, AS, IS	rs2251636	G	C	1.09	2.9×10^{-10}	1.00	0.89	1.00	0.67	1.00	0.41	1.00	0.71	1.00	0.86
KCNK3	IS, AS	rs11694327	C	T	1.04	3.5×10^{-08}	1.00	0.65	1.00	0.68	1.00	0.80	1.00	0.55	1.01	0.31
PITX2	CE, AS, IS	rs6847935	T	A	1.32	9.4×10^{-59}	1.00	0.37	0.99	0.12	1.01	0.10	1.00	0.34	1.00	0.77
FGG	CE, AS, IS	rs6536024	C	T	1.09	3.2×10^{-08}	1.01	0.03	1.01	0.08	1.01	0.02	1.00	0.76	1.00	0.64
LOC100505841	IS, AS	rs17148926	A	C	1.07	2.8×10^{-12}	1.00	0.50	1.00	0.41	1.00	0.65	1.00	0.94	1.01	0.14
FOXF2	IS, AS	rs79318212	G	A	1.09	3.6×10^{-14}	1.00	0.72	1.00	0.87	0.99	0.25	1.00	0.43	1.01	0.48
SLC22A7	AS	rs1574430	A	C	1.03	8.6×10^{-10}	1.00	0.87	1.00	0.60	1.00	0.76	1.00	0.73	1.00	0.61
HDAC9	LAA, AS, IS	rs2107595	A	G	1.16	5.3×10^{-13}	1.00	0.64	1.01	0.35	0.99	0.14	1.00	0.94	0.99	0.18
CDK6	IS	rs42035	A	G	1.06	3.3×10^{-12}	1.00	0.43	0.99	0.04	1.00	0.36	1.00	0.65	1.00	0.74
CDKN2B-AS1	IS	rs7859362	C	T	1.04	3.6×10^{-11}										
ABO	IS	rs649129	T	C	1.06	7.5×10^{-10}	1.01	0.03	1.00	0.38	1.00	0.46	1.00	0.78	1.01	0.02
SH3PXD2A	IS, AS, CE	rs55983834	C	T	1.05	1.6×10^{-14}	1.01	0.16	1.00	0.72	1.01	0.12	1.00	0.89	1.00	0.50
MMP12	LAA, AS, IS	rs72985562	G	T	1.23	4.4×10^{-08}	1.01	0.44	1.00	0.58	0.99	0.21	1.00	0.63	1.01	0.52
PDE3A	IS, AS	rs7304841	A	C	1.05	4.6×10^{-14}	1.00	0.79	1.01	0.23	1.00	0.68	1.00	0.24	1.00	0.87
SH2B3	IS, AS	rs10774625	A	G	1.07	2.6×10^{-20}	1.01	0.14	1.00	0.27	1.00	0.54	1.00	0.50	1.01	0.18

Locus	Phenotype	SNP	EA	OA	OR	P	OR IS	P IS	OR LAA	P LAA	OR CE	P CE	OR SVO	P SVO	OR UE	P UE
12q24	IS, AS	rs35429	A	G	1.04	1.9×10^{-08}	1.00	0.38	1.01	0.06	1.00	0.33	1.00	0.22	1.00	0.49
LRCH1	IS	rs842365	A	G	1.05	3.6×10^{-08}	1.00	0.83	1.00	0.71	1.00	0.68	1.00	0.20	1.00	0.99
FURIN	IS, AS	rs1573644	C	T	1.04	3.6×10^{-09}	1.00	0.47	1.01	0.07	1.00	0.34	1.00	0.53	1.00	0.48
ZFHX3	CE, AS, IS	rs2359171	A	T	1.17	7.5×10^{-18}	1.01	0.06	1.01	0.22	1.02	0.02	1.00	0.84	1.00	0.45
ZCCHC14	SVO, AS, IS	rs12445022	A	G	1.11	1.3×10^{-09}	1.01	0.06	1.00	0.48	1.00	0.39	1.00	0.61	1.01	0.11
SLC44A2	IS, AS	rs28860769	G	A	1.05	1.6×10^{-10}	1.00	0.69	1.00	0.33	1.00	0.79	1.00	0.73	1.00	0.62
LDLR	IS	rs8106503	T	C	1.07	7.9×10^{-09}	1.01	0.13	1.01	0.05	1.00	0.60	1.01	0.12	1.00	0.61
NOS3	IS, AS	rs1549758	T	C	1.04	2×10^{-08}	1.00	0.31	0.99	0.18	1.00	0.47	1.00	0.16	0.99	0.22
COL4A2	SVO, AS, IS	rs9515201	A	C	1.11	2.1×10^{-08}	0.99	0.23	1.00	0.38	0.99	0.15	1.00	0.45	1.00	0.34
TSPAN19	AS	rs78015967	C	T	1.01	1.9×10^{-17}	1.02	0.15	1.02	0.13	1.02	0.11	1.01	0.47	1.01	0.31
DAZL	AS	rs55683442	G	A	1.01	5.3×10^{-10}	1.00	0.87	1.00	0.79	1.00	0.96	1.01	0.15	1.00	0.88
SHOC1	AS	rs7021485	G	C	1.03	4.8×10^{-08}	1.01	0.15	1.00	0.96	1.01	0.07	1.00	0.92	1.00	0.93
INPP5B	IS, AS	rs28673728	G	A	1.03	9.2×10^{-09}	0.99	0.25	1.00	0.83	0.99	0.11	1.00	0.29	1.00	0.57
USP34	IS, AS	rs72811469	C	A	1.04	4.1×10^{-08}	0.99	0.36	1.00	0.85	0.99	0.34	1.00	0.99	0.99	0.17
NOA1	IS, AS	rs7687767	G	A	1.03	4.8×10^{-09}	1.01	0.18	1.00	0.69	1.01	0.07	1.01	0.03	1.00	0.82
PCDH18	IS, AS	rs13148045	C	T	1.04	5.3×10^{-10}	1.00	0.84	1.01	0.34	1.00	0.70	1.00	0.71	1.00	0.82
FBN2	IS, AS	rs55670004	C	T	1.03	3.5×10^{-09}	1.00	0.62	1.00	0.75	1.00	0.94	1.00	0.98	1.00	0.82
ARHGAP26	IS, AS	rs3776307	G	A	1.03	7.3×10^{-09}	1.00	0.60	1.00	0.63	1.00	0.68	1.00	0.39	1.00	0.88
TMEM41B	IS, AS	rs11042273	G	T	1.05	2×10^{-09}	1.00	0.99	1.00	0.75	1.00	0.69	1.00	0.37	1.00	0.71
OVOL1	IS, AS	rs557675	T	G	1.03	1.7×10^{-08}	1.01	0.08	1.01	0.02	1.00	0.40	1.00	0.15	1.01	0.00
HHIPL1	IS, AS	rs8014986	G	A	1.03	1×10^{-08}	1.01	0.16	1.00	0.41	1.01	0.19	1.00	0.41	1.01	0.17
CDH13	IS, AS	rs7500448	A	G	1.03	1.4×10^{-09}	1.00	0.49	1.00	0.70	1.01	0.12	1.00	0.92	1.00	0.89
DRC3	IS, AS	rs4471742	T	C	1.03	2.4×10^{-09}	1.00	0.93	1.00	0.88	1.00	0.35	1.00	0.56	1.00	0.79
DYM	IS, AS	rs833509	C	T	1.03	2.1×10^{-11}	1.00	0.43	1.00	0.89	1.00	0.32	1.01	0.07	1.00	0.42
PTK2	AS	rs10111852	C	T	1.02	4.8×10^{-09}										
RMC1	AS	rs1788820	A	G	1.02	3.1×10^{-08}	1.00	0.65	1.00	0.82	1.00	0.72	1.00	0.25	1.00	0.48

Locus	Phenotype	SNP	EA	OA	OR	P	OR IS	P IS	OR LAA	P LAA	OR CE	P CE	OR SVO	P SVO	OR UE	P UE
RBM20	CE	rs10749053	T	C	1.08	5.8×10^{-09}										
NCOR2	CE	rs11057583	A	G	1.09	1.3×10^{-08}	1.00	0.66	0.99	0.40	1.00	0.63	1.00	0.39	1.00	0.95
POLR2A	CE	rs11078685	C	T	1.05	2.3×10^{-08}	1.01	0.09	1.01	0.19	1.00	0.55	1.00	0.46	1.01	0.26
IL6R	LAA, AS	rs11265613	T	C	1.05	1.4×10^{-08}	1.00	0.38	1.00	0.93	1.01	0.20	1.00	0.21	1.00	0.71
FN1	LAA	rs17517928	T	C	1.11	8.7×10^{-10}	0.99	0.02	0.99	0.07	0.99	0.04	1.00	0.72	0.99	0.11
FGD5	LAA	rs748431	G	T	1.05	2.6×10^{-09}	1.00	0.45	1.00	0.89	1.00	0.42	1.00	0.24	0.99	0.19
EDNRA	LAA	rs1878406	T	C	1.08	1.3×10^{-09}	0.99	0.44	1.00	0.58	1.00	0.46	0.99	0.05	0.99	0.21
JCAD	LAA	rs2487928	A	G	1.06	7.2×10^{-12}	1.00	0.32	1.00	0.72	1.00	0.66	1.00	0.71	1.00	0.50
LIPA	LAA	rs1412444	T	C	1.07	4.4×10^{-15}	1.01	0.04	1.00	0.36	1.01	0.03	1.00	0.19	1.01	0.02
FLT1	LAA	rs1924981	T	C	1.05	1.9×10^{-08}										

Bold OR is indicating a consistent direction of the effect between the two studies for at least one phenotype, and bold p-value is indicating a nominal p-value < 0.05 for those with consistent direction of the effects. OR: Odds Ratio; EA: Effect Allele; OA: Other Allele; P: P-value.

5.2. Multitrait analysis of GWAS of cardioembolic stroke

5.2.1 Introduction

About 25% of ischemic strokes are of undetermined etiology (166): patients with multiple stroke etiologies, incomplete diagnostic work-up or embolic stroke of undetermined source (ESUS). Up to 17% of all ischemic strokes are ESUS, with a stroke recurrence rate of 4-5% despite antiplatelet therapy (167).

ESUS encompasses different entities. Atrial cardiopathy, occult atrial fibrillation (AF) and left ventricular disease might benefit from anticoagulation, but atherosclerotic plaques might benefit from low-dose anticoagulation with antiplatelets in ESUS patients (167). The subgroup of patients >75 years in RE-SPECT ESUS (Dabigatran Etexilate for Secondary Stroke Prevention in Patients With Embolic Stroke of Undetermined Source) had a significant benefit of lower-dose dabigatran over aspirin, suggesting occult AF as a triggering cause (168). Different studies indicate that the prevalence of occult AF among ESUS patients is of 11-30%(169,170).

A tool capable of better stratifying patients is needed to offer them appropriate treatment regarding its potential stroke cause to decrease its recurrence.

On the other hand, not all patients with AF will develop a stroke, and decision of anticoagulation for stroke prevention in AF patients is carried out based on a clinical scale: CHA2DS2-VASc. The rates of stroke vary considerably in patients with CHA2DS2-VASc 1-2 (171), hence the need for a more accurate scale in these cases.

Cardioembolic strokes (CE) are mostly caused by an onset/already known AF. Understanding CE genetic architecture will provide tools to select ESUS or AF patients who would benefit from anticoagulant therapy and develop specific and more effective therapies with fewer side effects.

Therefore, we aimed to discover novel loci associated with CE by performing a Multitrait Analysis of Genome Wide Association Study (MTAG) of CE-AF and create a polygenetic risk score (PRS) that allowed a more efficient stratification of stroke patient risk of having a CE.

5.2.2 Results

5.2.2.1 MTAG analysis of CE

After QCs (Figure 20), there were 6,808,676 common qualified SNPs from the AF-2018 and MEGASTROKE-CE cohorts. MTAG software revealed mean χ^2 for AF-2018 and MEGASTROKE-CE in 1.39 and 1.12 respectively. The estimated equivalent GWAS sample size of the MTAG analysis for CE was of 861,823 individuals. A Manhattan plot of the MTAG-CE analysis is shown in Figure 21; little evidence of genomic inflation was observed with a lambda of 1.02.

The MTAG-CE results revealed a total of 44 associated loci (p-value $<5\times10^{-8}$). 40 significant loci associated with CE were novel and four were previously found known associations (Table 11). All loci significantly associated in MEGASTROKE-CE (*ABO*, *NKX2-5*, *PITX2* and *ZFHX3*) were genome-wide associated in this MTAG-CE, except for the locus belonging to RGS7 gene (top SNP rs146390073 MTAG p-value=0.001, AF-2018 p-value=0.98).

Other loci found significant in previously GWAS of CE different from MEGASTROKE, were: *PHF20*, *GNAO1* and 5q22.3 region. For two of them, the association was more significant in our analysis. For the 5q22.3 region, top SNP rs2169955 MTAG-CE p-value=4.76×10^{-7} vs MEGASTROKE-CE p-value=6.13×10^{-3}, and for mapped gene PHF20, top SNP rs11697087 MTAG-CE p-value=6.62×10^{-5} vs MEGASTROKE-CE p-value=6.05×10^{-4}. *GNAO1* was not evaluated in our study due to absence of the index SNP in AF-2018.

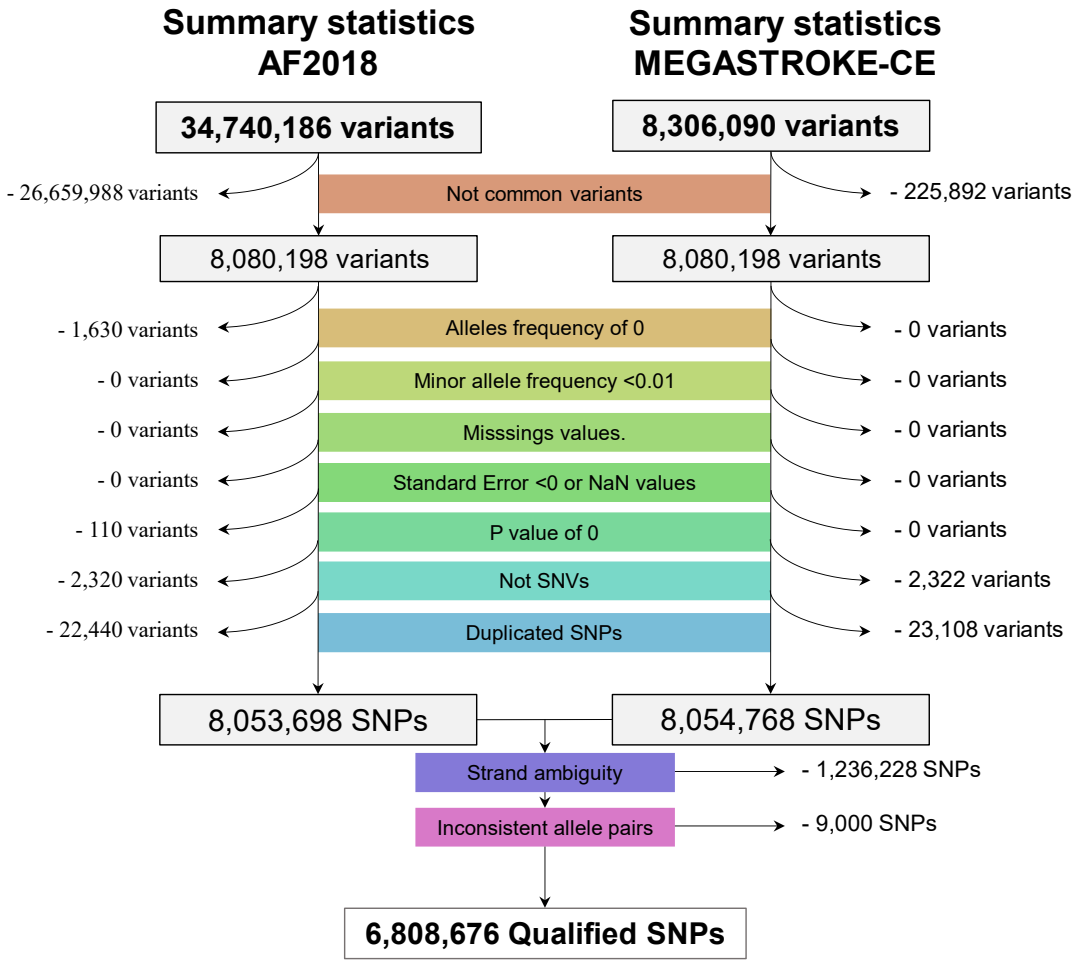


FIGURE 20. Workflow of the SNPs for the AF-2018 and MEGASTROKE-CE datasets. NaN: Not a number. SNP: Single Nucleotide Polymorphism.

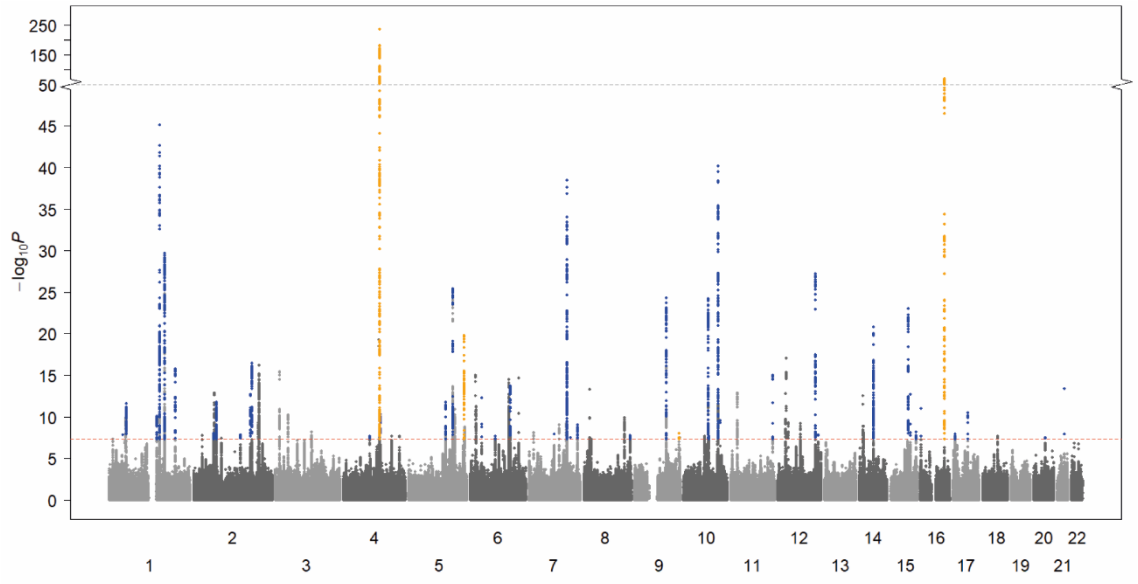


FIGURE 21. **Manhattan plot of MTAG-CE.** The x axis represents chromosome location, and the y axis represents the minus logarithm on base 10 of p-value.

TABLE 11. MTAG-CE results of the independent and significant loci

				MEGASTROKE CE		AF-2018		MTAG-CE		MTAG-AF		
SNP	**Locus**	**Gene**	**Novelty**	**Z**	**P**	**Z**	**P**	**Z**	**P**	**Z**	**P**	**PPA**
rs17042098-A	4q25	*PITX2*	Known	12.72	3.72×10^{-37}	38.01	8.97×10^{-318}	32.86	7.73×10^{-237}	37.93	0	1
rs2106261-T	16q22.3	*ZFHX3*	Known	6.75	1.63×10^{-11}	20.23	4.97×10^{-91}	17.48	2.02×10^{-68}	20.19	1.20×10^{-90}	1
rs11264280-T	1q21.3	*ADAM15*	Novel	2.49	1.28×10^{-02}	18.97	3.07×10^{-79}	14.22	7.17×10^{-46}	18.44	5.88×10^{-76}	0.65
rs11598047-A	10q24.33	*NEURL1*	Novel	-3.27	1.06×10^{-03}	-17.08	8.95×10^{-66}	-13.38	7.50×10^{-41}	-16.73	7.42×10^{-63}	0.93
rs3807989-A	7q31.2	*CAV1*	Novel	-4.5	6.75×10^{-06}	-15.63	1.24×10^{-54}	-13.1	3.43×10^{-39}	-15.5	3.30×10^{-54}	1
rs680084-A	1q24.2	*GORAB*	Novel	-4.1	4.02×10^{-05}	-13.54	3.31×10^{-42}	-11.46	2.10×10^{-30}	-13.46	2.84×10^{-41}	0.99
rs883079-T	12q24.21	*TBX5*	Novel	3.55	3.95×10^{-04}	13.26	2.84×10^{-40}	10.96	5.97×10^{-28}	13.12	2.60×10^{-39}	0.97
rs17171711-T	5q31.2	*FAM13B*	Novel	3.79	1.56×10^{-04}	12.48	1.95×10^{-35}	10.57	4.04×10^{-26}	12.41	2.35×10^{-35}	0.99
rs4385527-A	9q22.32	*AOPEP*	Novel	3.92	8.56×10^{-05}	12.03	6.16×10^{-33}	10.34	4.69×10^{-25}	11.99	3.89×10^{-33}	0.95
rs78249997-T	10q22.2	*MYOZ1*	Novel	-3.82	1.37×10^{-04}	-12.08	8.75×10^{-34}	-10.32	5.80×10^{-25}	-12.03	2.45×10^{-33}	0.98
rs7172038-T	15q24.1	*NEO1*	Novel	-2.73	6.43×10^{-03}	-12.58	4.78×10^{-36}	-10.04	1.02×10^{-23}	-12.37	3.76×10^{-35}	0.74
rs2738413-A	14q23.2	*ESR2*	Novel	2.97	3.03×10^{-03}	11.61	2.55×10^{-31}	9.52	1.67×10^{-21}	11.47	1.80×10^{-30}	0.85
rs6891790-T	5q35.1	*NKX2-5*	Known	-4.97	6.67×10^{-07}	-9.59	4.53×10^{-22}	-9.29	1.57×10^{-20}	-9.8	1.16×10^{-22}	1
rs2857265-A	2q31.2	*FKBP7*	Novel	2.76	5.85×10^{-03}	10.16	4.58×10^{-24}	8.42	3.65×10^{-17}	10.06	8.29×10^{-24}	0.7
rs10753933-T	1q32.1	*PPFIA4*	Novel	3.72	2.09×10^{-04}	9.09	9.84×10^{-20}	8.24	1.70×10^{-16}	9.16	5.30×10^{-20}	0.99
rs74399915-T	11q24.3	*C11orf45*	Novel	3.38	7.20×10^{-04}	9.05	1.23×10^{-19}	8.03	1.01×10^{-15}	9.08	1.11×10^{-19}	0.92
rs13191450-A	6q22.31	*HSF2*	Novel	2.93	3.51×10^{-03}	8.88	9.97×10^{-19}	7.65	1.95×10^{-14}	8.86	7.96×10^{-19}	0.81
rs2834618-T	21q22.12	*RUNX1*	Novel	3.31	9.48×10^{-04}	8.43	3.41×10^{-17}	7.56	3.96×10^{-14}	8.47	2.37×10^{-17}	0.96
rs56181519-T	2q31.1	*WIPF1*	Novel	-2.73	6.31×10^{-03}	-8.6	6.46×10^{-18}	-7.35	1.98×10^{-13}	-8.56	1.11×10^{-17}	0.82
rs12908004-A	15q25.1	*ARNT2*	Novel	-3.28	1.03×10^{-03}	-8.13	4.12×10^{-16}	-7.35	2.03×10^{-13}	-8.19	2.65×10^{-16}	0.96
rs3176326-A	6p21.2	*CDKN1A*	Novel	-3.98	6.66×10^{-05}	-7.36	1.42×10^{-13}	-7.23	5.01×10^{-13}	-7.54	4.59×10^{-14}	1
rs6747542-T	2p13.3	*GMCL1*	Novel	2.61	8.86×10^{-03}	8.27	1.10×10^{-16}	7.06	1.63×10^{-12}	8.23	1.82×10^{-16}	0.75
rs337705-T	5q22.3	*KCNN2*	Novel	-2.56	1.04×10^{-02}	-8.29	1.63×10^{-16}	-7.05	1.77×10^{-12}	-8.25	1.57×10^{-16}	0.72
rs41292535-A	1p32.2	*EPS15*	Novel	3.81	1.39×10^{-04}	7.16	7.42×10^{-13}	6.99	2.73×10^{-12}	7.33	2.34×10^{-13}	0.96

rs140185678-A	16p13.3	*RPL3L*	Novel	2.92	3.54×10^{-03}	7.61	2.43×10^{-14}	6.79	1.13×10^{-11}	7.64	2.13×10^{-14}	0.89
rs76774446-A	17q21.32	*GOSR2*	Novel	4.38	1.20×10^{-05}	6.15	8.72×10^{-10}	6.63	3.32×10^{-11}	6.43	1.25×10^{-10}	0.99
rs55754224-T	4q26	*CAMK2D*	Novel	2.8	5.05×10^{-03}	7.39	2.15×10^{-13}	6.57	4.92×10^{-11}	7.41	1.22×10^{-13}	0.8
rs79187193-A	1q21.2	*GJA5*	Novel	-2.37	1.78×10^{-02}	-7.59	3.15×10^{-14}	-6.47	9.79×10^{-11}	-7.56	4.08×10^{-14}	0.71
rs12260801-T	10q25.2	*PDCD4*	Novel	4.56	5.22×10^{-06}	5.53	2.94×10^{-08}	6.31	2.79×10^{-10}	5.86	4.62×10^{-09}	1
rs2269001-A	7q36.1	*KCNH2*	Novel	-2.89	3.94×10^{-03}	-6.63	4.01×10^{-11}	-6.11	1.00×10^{-09}	-6.7	2.06×10^{-11}	0.73
rs6598541-A	15q26.3	*IGF1R*	Novel	2.69	7.17×10^{-03}	6.35	2.22×10^{-10}	5.81	6.21×10^{-09}	6.41	1.48×10^{-10}	0.72
rs11125871-T	2p15	*C2orf74*	Novel	-3.24	1.17×10^{-03}	-5.79	6.42×10^{-09}	-5.75	9.19×10^{-09}	-5.95	2.71×10^{-09}	0.87
rs635634-T	9q34.2	*ABO*	Known	5.1	3.31×10^{-07}	4.2	2.74×10^{-05}	5.72	1.04×10^{-08}	4.66	3.13×10^{-06}	1
rs10272350-A	7q11.23	*TMEM60*	Novel	4.53	6.13×10^{-06}	4.62	3.41×10^{-06}	5.68	1.32×10^{-08}	4.99	5.93×10^{-07}	0.99
rs116600817-A	17p13.1	*TNFSF12*	Novel	2.88	3.92×10^{-03}	6	2.45×10^{-09}	5.68	1.33×10^{-08}	6.1	1.07×10^{-09}	0.85
rs13010313-T	2q22.3	*ZEB2*	Novel	3.52	4.21×10^{-04}	5.45	5.72×10^{-08}	5.67	1.41×10^{-08}	5.65	1.57×10^{-08}	0.9
rs2885697-T	1p34.2	*SCMH1*	Novel	-2.54	1.09×10^{-02}	-6.27	2.88×10^{-10}	-5.67	1.41×10^{-08}	-6.32	2.70×10^{-10}	0.72
rs11057583-A	12q24.31	*NCOR2*	Novel	3.82	1.35×10^{-04}	5.19	2.10×10^{-07}	5.67	1.46×10^{-08}	5.45	5.13×10^{-08}	0.74
rs11782313-T	8q24.3	*PTK2*	Novel	-3.05	2.34×10^{-03}	-5.81	5.71×10^{-09}	-5.64	1.65×10^{-08}	-5.94	2.94×10^{-09}	0.78
rs12211255-A	6q14.1	*FILIP1*	Novel	3.42	6.24×10^{-04}	5.44	5.06×10^{-08}	5.61	2.06×10^{-08}	5.63	1.78×10^{-08}	0.96
rs1898096-A	10q22.3	*LRMDA*	Novel	-2.64	8.30×10^{-03}	-6.08	1.39×10^{-09}	-5.6	2.13×10^{-08}	-6.15	7.96×10^{-10}	1
rs11099098-T	4q21.21	*FGF5*	Novel	2.77	5.67×10^{-03}	5.94	2.96×10^{-09}	5.58	2.38×10^{-08}	6.03	1.62×10^{-09}	0.78
rs55985730-T	7q32.1	*CALU*	Novel	-2.84	4.47×10^{-03}	-5.82	5.24×10^{-09}	-5.54	3.06×10^{-08}	-5.92	3.19×10^{-09}	0.87
rs3746471-A	20q11.23	*KIAA1755*	Novel	3.56	3.58×10^{-04}	5.18	2.30×10^{-07}	5.51	3.50×10^{-08}	5.4	6.59×10^{-08}	0.96

SNP: Index Single Nucleotide Polymorphism and effect allele, Z: Z-score; P: P-value. Loci highlighted in bold are the ones not previously associated with AF. PPA: posterior probability of model 3 of GWAS-PW.

5.2.2.2 New candidate loci associated with CE

After gene prioritization, 40 genes were selected from the 40 novel loci (Appendix Table 2). Novel loci showed a high degree of functionality of the SNPs as missense variants, eQTL, pQTLs, and Hi-C physical interaction (Appendix Table 2).

Replication analysis was performed in an independent cohort of IS patients and controls, GENERACION cohort, n = 9,105. Information of recruiting from each participating hospital, array and projects information as well as clinical description can be assessed in subsection 5.1.2.1. Evaluation of the index SNPs and SNPs in high LD belonging to genome-wide significant loci from the MTAG-CE revealed that 11 loci were replicated, as SNPs had a p-value <0.05 in MEGASTROKE-CE, a PPA-3 >0.6 and a p-value <0.05 in the replication cohort (GENERACION). Among these 11 different loci, *PITX2*, *ZFHX3*, *NKX2-5* were already known. Eight loci were novel associations whose prioritized genes were: *CAV1*, *IGF1R*, *KIAA1755*, *NEURL1*, *GORAB*, *ESR2*, *ZEB2* and *WIPF1* (Appendix Table 3).

Interestingly, we found loci not previously described in AF or CE with four prioritized genes: *TMEM60*, *KIAA1755*, *NCOR2*, and *FILIP1*. Functional annotation of the index SNPs revealed rs3746471 as a missense variant of the *KIAA1755* gene coding for R1045W, and it was predicted to be deleterious with a SIFT score of 0.007.

5.2.2.3 New candidate locus associated with AF

FILIP1 locus reached genome-wide significance in the MTAG-AF, a p-value <0.05 in AF-2018 and a PPA-3 >0.6. The FILIP1 index variant was additionally evaluated in the GWAS of AF in the independent cohort (GENERACION), revealing a suggestive p-value with consistent direction of the effect of this novel association with AF (rs12211255-A, beta(se)=0.013(0.007), p-value=0.09).

The study of the 111 AF-2018 significant loci using GWAS-pairwise strategy suggested 51 loci that have an exclusive association with AF risk and a lack of association with CE (Appendix Table 4).

5.2.2.4 Biological processes of loci associated with CE and AF and biological processes of loci associated exclusively to AF

GO of biological processes from the Genome-Wide loci of the MTAG-CE analysis revealed 98 enriched gene sets (Appendix Figure 1); the top biological processes

were: cardiac conduction, cardiac muscle cell contraction and cardiac muscle contraction.

A biological processes analysis of the genes associated exclusively with AF revealed 41 biological processes exclusive to AF risk and mainly associated with cardiac development processes (Appendix Figure 2).

5.2.2.5 Polygenic Risk Score

The training set was composed of 1,212 CE patients and 4,501 controls and the test set of 303 CE patients and 1,125 controls from GENERACION. No significant differences in clinical variables were found between the training and the test sets (Appendix Figure 2).

For model-1, the PRS with the highest r^2 in the training set (r^2 = 0.018) was obtained with a SNP p-value threshold of 5×10-8, comprising a total of 93 SNPs (Figure 22). Age, sex, and hypertension were the only variables for which information was available for >90% of the patients, and therefore the only ones considered in the multivariable model as mentioned in the "Methods" section. The three variables were significantly associated and therefore included in model-2. For model-3 we added the PRS to the model-2, remaining all variables significant (Appendix Table 5), including the PRS with a Z value of 4.33 and a p-value of 1.28×10^{-5}.

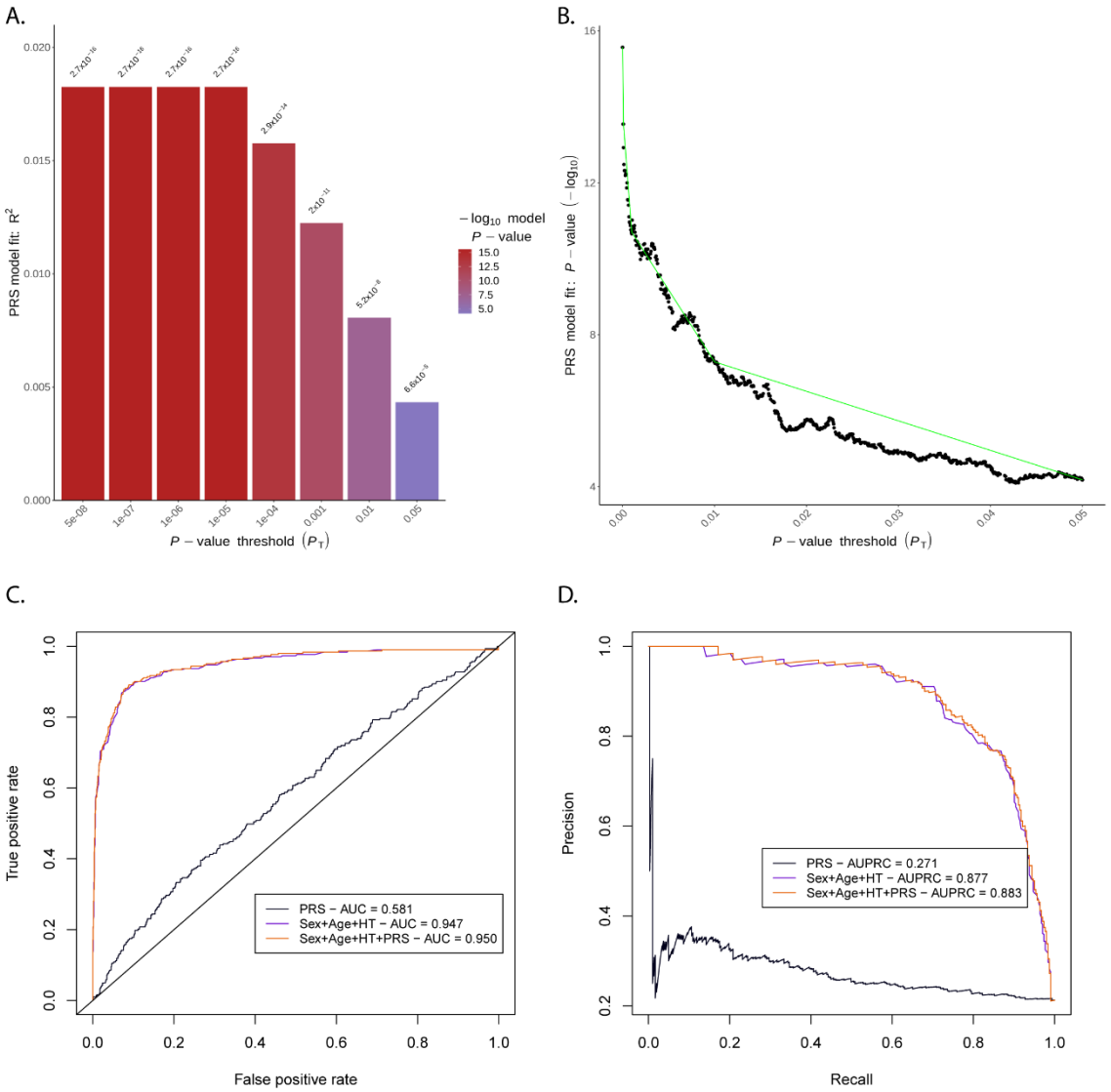


FIGURE 22. Polygenic risk score (PRS) performance. Panel A is a bar plot of the r^2 for the PRS models of eight different thresholds in the training set. Panel B represents the p-value variation along the full range of thresholds evaluated in the training set. Panel C shows ROC curves and panel D Precision-Recall curves for the PRS performance in the independent test set. AUC: area under the ROC curve; AUPRC: area under the precision recall curve; HT: hypertension.

AUC in the test set for the different models were: 0.581 in model-1, 0.947 in model-2 and 0.950 in model-3. AUPRC were: 0.271 in model-1, 0.877 in model-2 and 0.883 in model-3 (Figure 22). Comparing AUC, there was significantly better discrimination in the model-3 than model-2 (Z-score = -2.50, p-value = 0.01). AUC and AUPRC for each individual predictor can be found in Appendix Figure 3.

Additionally, the NRI categorical and quantitative and IDI showed a significant reclassification when quartiles of score risk was analyzed (Table 12).

TABLE 12. Reclassification table comparing CE models with and without PRS adding.

	Risk Category Age+Sex+HT+PRS model				
	Q1	Q2	Q3	Q4	
Risk Category Age+Sex+HT model	No.	No.	No.	No.	% Reclassified
Non cases					
Q1	4062	23	0	0	1
Q2	25	174	16	0	19
Q3	0	23	120	4	18
Q4	0	0	3	50	0
Cases					
Q1	149	7	0	0	4
Q2	7	124	10	0	12
Q3	0	12	171	34	21
Q4	0	0	18	677	3

NRI(Categorical) [95% CI]: 0.0134 [-0.0024 - 0.0291]; p-value: 0.09688

NRI(Continuos) [95% CI]: 0.1416 [0.0782 - 0.2049]; p-value: 1×10^{-5}

IDI [95% CI]: 0.0029 [9×10^{-04} - 0.0049]; p-value: 0.00431

5.2.3 Appendix

5.2.3.1 Appendix Tables

APPENDIX TABLE A2. Variant-to-Gene prioritization results of the 40 novel loci in the MTAG-CE analysis.

Index variant	**Gene**	**Overall V2G**	**Distance (Canonical TSS)**	**pQTL**	**eQTL**	**Enhancer-TSS corr** (172)	**PCHi-C** (173)	**DHS-promoter corr** (174)	**PCHi-C** (175)	**VEP (Ensembl)**
rs2885697	*SCMH1*	0.233	163547		0.9					intron_variant
rs41292535	*EPS15*	0.250	46299		0.9					intron_variant
rs79187193	*GJA5*	0.199	10358		0.7					
rs11264280	*ADAM15*	0.208	160090		0.9					
rs680084	*GORAB*	0.158	127295		0.6					
rs10753933	*PPFIA4*	0.266	30595		1					intron_variant
rs11125871	*C2orf74*	0.233	97923		1					
rs6747542	*GMCL1*	0.250	50040		0.9					3_prime_UTR_variant
rs13010313	*ZEB2*	0.008	442900						0	
rs56181519	*WIPF1*	0.216	8070		0.8					
rs2857265	*FKBP7*	0.283	55496	0.8	0.4		0.1			
rs11099098	*FGF5*	0.183	17881		0.6					
rs55754224	*CAMK2D*	0.083	254369		0.1					intron_variant
rs337705	*KCNN2*	0.050	345439							intron_variant
rs17171711	*FAM13B*	0.191	22855		0.5					intron_variant
rs3176326	*CDKN1A*	0.258	3052		0.9					intron_variant
rs12211255	*FILIP1*	0.233	15186		0.7		0.1			intron_variant
rs13191450	*HSF2*	0.108	328561		0.5					
rs10272350	*TMEM60*	0.316	63638		1		0.9		0	
rs3807989	*CAV1*	0.241	21193		0.8					intron_variant
rs55985730	*CALU*	0.191	37698		0.7					
rs2269001	*KCNH2*	0.266	23593		0.7			0.5	0	intron_variant
rs11782313	*PTK2*	0.283	48132		0.9		0.4		0	intron_variant

rs4385527	*AOPEP*	0.200	159606		0.7					intron_variant
rs78249997	*MYOZ1*	0.233	21082		0.9					
rs1898096	*LRMDA*	0.025								intron_variant
rs11598047	*NEURL1*	0.175	89210		0.5					intron_variant
rs12260801	*PDCD4*	0.150	57151				0.9		0	
rs74399915	*C11orf45*	0.150	6929		0.4					
rs883079	*TBX5*	0.100	53007							3_prime_UTR_variant
rs11057583	*NCOR2*	0.133	241806		0.4					intron_variant
rs2738413	*ESR2*	0.233	124870		0.9					intron_variant
rs7172038	*NEO1*	0.166	323204		0.8					
rs12908004	*ARNT2*	0.249	19767		1				0	
rs6598541	*IGF1R*	0.108	79367			0.2			0	intron_variant
rs140185678	*RPL3L*	0.249	4591							missense_variant
rs116600817	*TNFSF12*	0.424	6420	0.9	1					intron_variant
rs76774446	*GOSR2*	0.166	45927		0.4					intron_variant
rs3746471	*KIAA1755*	0.241	47260							missense_variant
rs2834618	*RUNX1*	0.066					0.3	0.5		

APPENDIX TABLE A3. Results of CE analysis on the GENERACION cohort.

									MTAG CE				CE analysis GENERACION			
Index Variant	**Gene**	**SNP**	**r^2**	**D'**	**CHR**	**BP**	**EA**	**OA**	**B**	**SE**	**Z**	**P**	**B**	**SE**	**Z**	**P**
rs680084	*GORAB*	rs503706	0.9	0.99	1	170635084	T	C	0.02	0.00	11.27	1.79×10^{-29}	0.01	0.00	2.06	0.040
rs13010313	*ZEB2*	rs11679718	0.93	0.97	2	145681935	A	G	0.01	0.00	5.61	1.99×10^{-08}	0.02	0.01	2.14	0.032
rs56181519	*WIPF1*	rs56181519	1	1	2	175555714	T	C	-0.01	0.00	-7.35	1.98×10^{-13}	-0.01	0.01	-2.16	0.031
rs17042098	*PITX2*	rs17042098	1	1	4	111664158	A	G	0.08	0.00	32.86	7.73×10^{-237}	0.01	0.01	1.96	0.049
rs6891790	*NKX2-5*	rs2277923	0.88	0.97	5	172662024	T	C	0.02	0.00	9.10	9.03×10^{-20}	0.01	0.01	2.03	0.043
rs3807989	*CAV1*	rs3807989	1	1	7	116186241	A	G	-0.02	0.00	-13.10	3.43×10^{-39}	-0.02	0.00	-3.19	0.001
rs11598047	*NEURL1*	rs11598047	1	1	10	105342672	A	G	-0.03	0.00	-13.38	7.50×10^{-41}	-0.02	0.01	-3.20	0.001
rs2738413	*ESR2*	rs2738413	1	1	14	64679960	A	G	0.01	0.00	9.52	1.67×10^{-21}	0.01	0.00	2.55	0.011
rs6598541	*IGF1R*	rs12898337	0.87	0.94	15	99294355	T	C	0.01	0.00	5.47	4.53×10^{-08}	0.01	0.00	2.05	0.040
rs2106261	*ZFHX3*	rs2106261	1	1	16	73051620	T	C	0.04	0.00	17.48	2.02×10^{-68}	0.01	0.01	2.15	0.032
rs3746471	*KIAA1755*	rs3746471	1	1	20	36841914	A	G	0.01	0.00	5.51	3.50×10^{-08}	0.01	0.00	2.24	0.025

EA: effect allele, B: beta; CHR: chromosome, BP: base pairs; OA: Other allele; P: p-value; SE: Standard Error; TSS: Transcription Start Site; Z: z-score. The genomic location is in Hg19.

APPENDIX TABLE A4. Results of MTAG for previous and novel loci associated with AF that are significantly associated exclusively to atrial fibrillation.

			MEGASTROKE CE		AF2018				
Index Variant	**Locus**	**Prioritzyed genes**	**Z**	**P**	**Z**	**P**	**PPA-1**	**PPA-2**	**PPA-3**
rs7529220-T	1p36.12	*HSPG2*	-0.43	6.68×10^{-01}	-6.34	1.98×10^{-10}	0.00	0.84	0.15
rs1545300-T	1p13.2	*KCND3*	1.07	2.85×10^{-01}	-7.64	1.48×10^{-14}	0.00	0.78	0.20
rs4073778-A	1p13.1	*CASQ2*	0.56	5.75×10^{-01}	7.25	4.96×10^{-13}	0.00	0.88	0.12
rs4951258-A	1q32.1	*NUCKS1,SLC41A1*	-0.83	4.05×10^{-01}	5.61	2.10×10^{-08}	0.00	0.84	0.13
rs6546620-T	2p23.3	*KIF3C*	-1.62	1.07×10^{-01}	-7.00	3.19×10^{-12}	0.00	0.71	0.29
rs2723064-T	2p14	*CEP68*	1.60	1.11×10^{-01}	9.56	1.43×10^{-21}	0.00	0.72	0.28
rs72926475-A	2p11.2	*REEP1*	-1.71	8.83×10^{-02}	-6.70	2.37×10^{-11}	0.00	0.68	0.32
rs113949548-T	2q14.3	*GYPC*	0.76	4.45×10^{-01}	6.44	1.09×10^{-10}	0.00	0.80	0.19
rs6738011-T	2q34	*ERBB4*	1.53	1.26×10^{-01}	6.33	3.05×10^{-10}	0.00	0.70	0.30
rs73041705-T	3p24.2	*THRB*	0.34	7.34×10^{-01}	6.07	1.55×10^{-09}	0.00	0.88	0.11
rs10428132-T	3p22.2	*SCN10A,SCN5A*	-0.60	5.52×10^{-01}	-9.21	2.72×10^{-20}	0.00	0.89	0.11
rs34080181-A	3p14.1	*LRIG1,SLC25A26*	-1.81	7.07×10^{-02}	-6.46	1.28×10^{-10}	0.00	0.66	0.34
rs17005647-T	3p14.1	*FRMD4B*	-0.89	3.73×10^{-01}	5.99	2.70×10^{-09}	0.00	0.86	0.14
rs6771054-T	3p11.1	*EPHA3*	0.88	3.81×10^{-01}	6.72	2.42×10^{-11}	0.00	0.81	0.19
rs10804493-A	3q13.2	*PHLDB2,PLCXD2*	0.72	4.71×10^{-01}	7.97	1.63×10^{-15}	0.00	0.87	0.12
rs1278493-A	3q22.3	*PPP2R3A*	-0.23	8.21×10^{-01}	-5.72	8.77×10^{-09}	0.00	0.87	0.11
rs75880040-T	3q26.33	*GNB4*	-1.07	2.88×10^{-01}	-5.84	5.21×10^{-09}	0.00	0.84	0.16
rs60902112-T	3q29	*XXYLT1*	1.07	2.84×10^{-01}	5.63	1.72×10^{-08}	0.00	0.80	0.17
rs3960788-T	4q24	*SLC9B1*	-2.41	1.58×10^{-02}	-5.45	5.98×10^{-08}	0.00	0.66	0.33
rs6839459-A	4q31.23	*ARHGAP10*	-1.25	2.13×10^{-01}	-6.72	1.64×10^{-11}	0.00	0.67	0.32
rs74500426-T	4q34.1	*HAND2,HAND2-AS1*	0.33	7.40×10^{-01}	-7.25	4.29×10^{-13}	0.00	0.83	0.17
rs6596717-A	5q21.3	*LOC102467213*	-1.15	2.53×10^{-01}	-5.94	3.00×10^{-09}	0.00	0.82	0.18
rs2012809-A	5q23.3	*SLC27A6*	-1.03	3.04×10^{-01}	-6.19	4.92×10^{-10}	0.00	0.74	0.25
rs6580277-A	5q31.1	*NR3C1*	-1.24	2.15×10^{-01}	-8.48	1.64×10^{-17}	0.00	0.83	0.16

			MEGASTROKE CE		AF2018				
Index Variant	**Locus**	**Priorityzed genes**	**Z**	**P**	**Z**	**P**	**PPA-1**	**PPA-2**	**PPA-3**
rs73366713-A	6p22.3	*ATXN1*	-1.70	8.80×10^{-02}	-10.45	1.53×10^{-25}	0.00	0.64	0.36
rs34969716-A	6p22.3	*KDM1B,DEK*	1.44	1.50×10^{-01}	9.00	1.60×10^{-19}	0.00	0.73	0.27
rs13210074-A	6q14.3	*CGA*	-1.48	1.38×10^{-01}	-6.41	1.48×10^{-10}	0.00	0.79	0.20
rs12154315-T	7p21.2	*DGKB*	1.86	6.28×10^{-02}	7.00	3.38×10^{-12}	0.00	0.66	0.34
rs6948592-T	7p15.1	*CREB5*	1.06	2.86×10^{-01}	6.09	9.99×10^{-10}	0.00	0.86	0.13
rs35005436-T	7q11.23	*GTF2I,LOC101926943,GTF2IRD2*	0.02	9.84×10^{-01}	-6.31	3.34×10^{-10}	0.00	0.85	0.14
rs35620480-A	8p23.1	*GATA4*	-0.48	6.29×10^{-01}	-5.87	5.15×10^{-09}	0.00	0.79	0.20
rs62521286-A	8q24.13	*FBXO32*	-0.71	4.75×10^{-01}	-8.90	4.50×10^{-19}	0.00	0.81	0.19
rs10822156-T	10q21.3	*REEP3,NRBF2*	-1.44	1.50×10^{-01}	-6.94	4.21×10^{-12}	0.00	0.79	0.21
rs2156664-T	11q24.1	*SORL1*	0.15	8.82×10^{-01}	5.86	4.93×10^{-09}	0.00	0.88	0.12
rs17380837-T	12p12.1	*SSPN*	0.75	4.55×10^{-01}	-6.96	4.80×10^{-12}	0.00	0.87	0.13
rs2860482-A	12q13.3	*NACA*	0.33	7.45×10^{-01}	7.11	1.21×10^{-12}	0.00	0.88	0.11
rs71454237-A	12q15	*LRRC10*	-0.50	6.14×10^{-01}	-7.38	1.78×10^{-13}	0.00	0.82	0.17
rs12426679-T	12q21.2	*PHLDA1*	-0.71	4.77×10^{-01}	-5.84	4.95×10^{-09}	0.00	0.85	0.13
rs6560886-T	12q24.33	*FBRSL1*	-1.31	1.90×10^{-01}	-5.67	1.49×10^{-08}	0.00	0.65	0.33
rs35569628-T	13q34	*CUL4A*	1.40	1.62×10^{-01}	5.65	1.38×10^{-08}	0.00	0.74	0.26
rs28631169-T	14q11.2	*MYH6,MYH7*	1.38	1.69×10^{-01}	6.21	5.35×10^{-10}	0.00	0.85	0.15
rs73241997-T	14q13.1	*CFL2*	1.29	1.96×10^{-01}	7.88	2.94×10^{-15}	0.00	0.76	0.24
rs74884082-T	14q24.2	*DPF3*	-0.22	8.25×10^{-01}	-6.32	3.48×10^{-10}	0.00	0.86	0.14
rs10873298-T	14q24.3	*IRF2BPL*	-0.15	8.85×10^{-01}	-5.81	7.07×10^{-09}	0.00	0.88	0.11
rs7225165-A	17p13.3	*YWHAE,CRK,MYO1C*	-0.53	5.98×10^{-01}	-5.90	3.20×10^{-09}	0.00	0.76	0.23
rs55941572-T	17p12	*MYOCD*	1.18	2.39×10^{-01}	6.60	4.45×10^{-11}	0.00	0.76	0.24
rs7359623-T	17q12	*ZPBP2,GSDMB,ORMDL3*	-0.31	7.59×10^{-01}	-6.64	4.39×10^{-11}	0.00	0.89	0.10
rs7224711-T	17q25.3	*CYTH1,USP36*	-1.40	1.60×10^{-01}	-5.53	3.72×10^{-08}	0.00	0.79	0.19
rs8088085-A	18q21.2	*MEX3C*	0.05	9.65×10^{-01}	5.45	4.79×10^{-08}	0.00	0.80	0.17
rs464901-T	22q11.21	*TUBA8*	-1.06	2.89×10^{-01}	7.06	1.53×10^{-12}	0.00	0.83	0.16

			MEGASTROKE CE		AF2018				
Index Variant	**Locus**	**Priorityzed genes**	**Z**	**P**	**Z**	**P**	**PPA-1**	**PPA-2**	**PPA-3**
rs133902-T	22q12.1	*MYO18B*	1.91	5.65×10^{-02}	6.16	9.14×10^{-10}	0.00	0.63	0.36

SE: Standard Error, B: beta, Z: z-score; P: p-value.

APPENDIX TABLE A5. Clinical characteristics of the training and test set used for the polygenic risk score analysis.

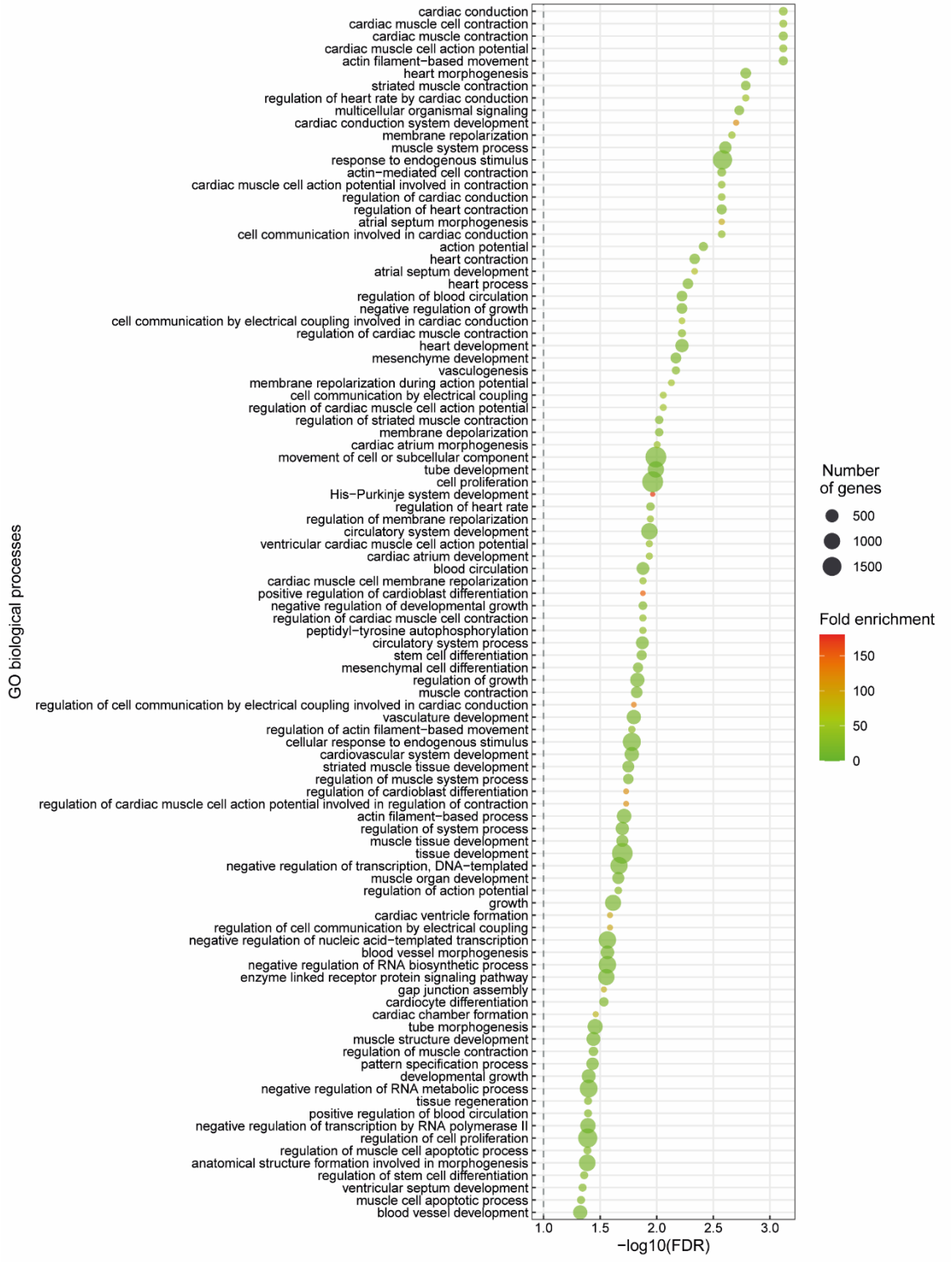
	Training set			Test set			
Variable	**Total = 5713**	**CE = 1212**	**controls = 4501**	**Total = 1428**	**CE = 303**	**controls = 1125**	**P**
Sex (% f)	3124 (54.7)	643 (53.1)	2481 (55.1)	780 (54.6)	154 (50.8)	626 (55.6)	0.99
DM (%)	407 (22.8)	315 (26.1)	92 (16.0)	88 (21.3)	70 (23.3)	18 (15.9)	0.53
HTN (%)	1892 (33.1)	875 (72.4)	1017 (22.6)	459 (32.1)	214 (70.6)	245 (21.8)	0.49
DL (%)	490 (42.5)	278 (39.0)	212 (48.0)	121 (46.4)	74 (41.6)	47 (56.6)	0.28
AF (%)	789 (14.2)	770 (64.1)	19 (0.4)	188 (13.4)	181 (59.9)	7 (0.6)	0.49
Age (Y. IQR)	55 (48-65)	78 (70-83)	52 (47-59)	55 (48-64)	78 (69-83)	52 (46-59)	0.44

APPENDIX TABLE A6. Model-3 for the prediction of cardioembolic stroke.

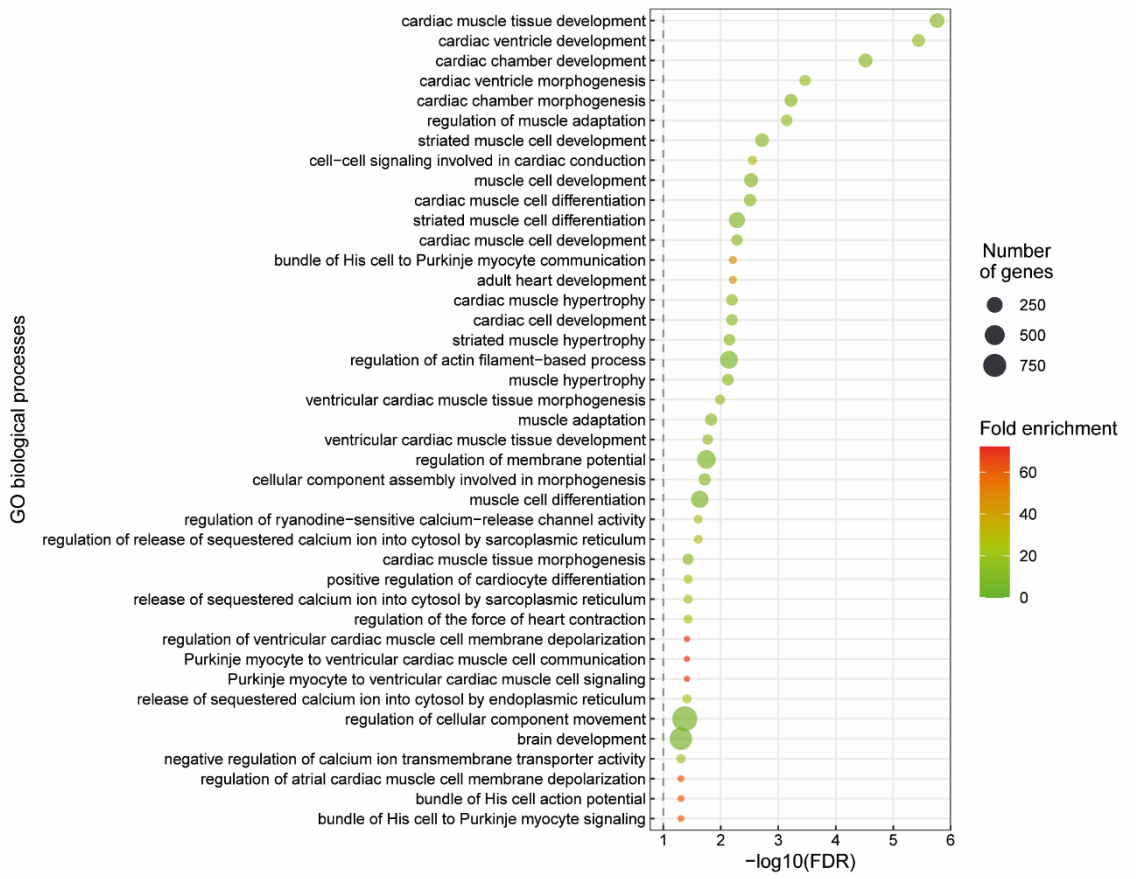
	β coefficients	**SE**	**z value**	**p-value**
Intercept	-13.93	0.42	-33.52	$<2\times10^{-16}$
PRS	348.82	80.52	4.33	1.48×10^{-05}
Age	0.19	0.01	32.51	$<2\times10^{-16}$
Sex	0.5	0.1	4.79	1.68×10^{-06}
Hypertension	0.66	0.11	6.17	6.98×10^{-10}

SE: standard error, PRS: polygenic risk score

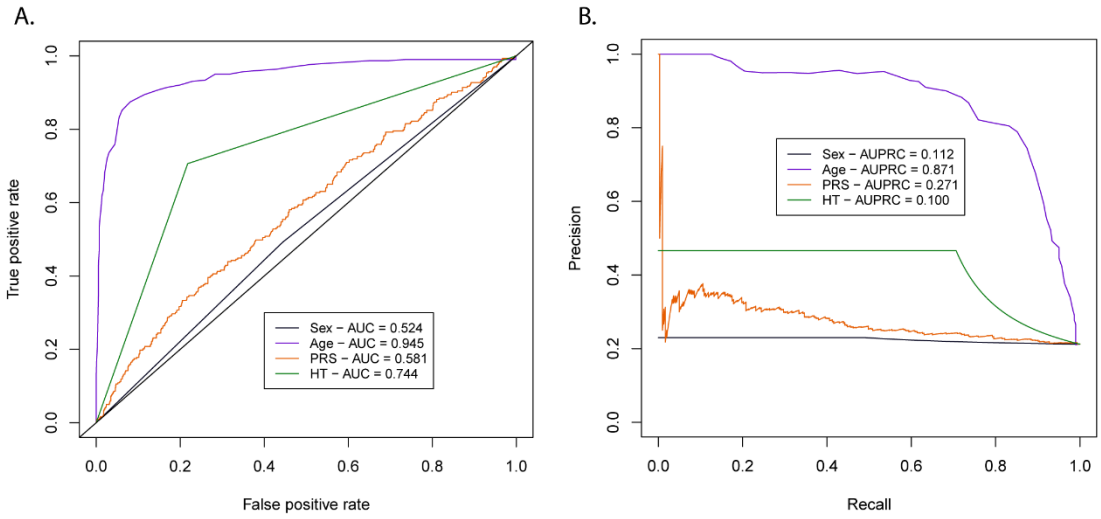
5.2.3.2 Appendix Figures



APPENDIX FIGURE A1. GO Biological processes enriched in CE prioritized gene set.



APPENDIX FIGURE A2. GO Biological processes enriched exclusively in analysis of AF associated genes independently of CE risk.



APPENDIX FIGURE A3. Polygenic risk score (PRS) performance for the individual predictors. Panel A shows ROC curves and panel B Precision-Recall curves for the PRS performance in the independent test set. AUC: area under the ROC curve; AUPRC: area under the precision recall curve; HT: hypertension.

5.3. Metalloproteinases levels and ischemic stroke

5.3.1 Introduction

Matrix metalloproteinases (MMPs) are a diverse group of endopeptidases. They are known to mediate degradation or remodeling of the extracellular matrix and may be responsible for physiological processes such as wound healing and angiogenesis(175). However, they are also responsible for pathophysiological processes like fibrotic disease and atherosclerosis(176). Observational studies have found a correlation between MMP levels and risk of atheromatous plaque instability(13) and ischemic stroke (IS)(177). Several studies have also suggested that MMPs may play a key role in the outcome of IS(178). Despite that, the causality of MMP levels has only recently been proved regarding MMP-12, lower plasma levels of which are associated with a risk of large-artery atherosclerosis stroke (LAA)(49). This has prompted us to study other members of the MMP family in the context of IS.

Mendelian randomization (MR) is a technique that leverages genetic variants associated with heritable risk factors and diseases. This tool is particularly important in the study of complex diseases that may involve multiple factors and, therefore, a strategy to find the drivers behind the disease is needed.

We conducted a two-sample MR study to test the hypothesis that serum/plasma levels of MMPs are causally associated with risk and long-term outcome of IS.

5.3.2 Results

5.3.2.1 Data sources

After searching the literature, we found a total of six studies; only three of them performed GWAS on MMP levels and were selected, evaluating MMP-1(152), MMP-8(179) and MMP-12(155) levels. The SNPs used are specified in appendix tables A7-A9.

5.3.2.2 Primary MR analysis

For MMP-1, a significant association with LAA was observed: odds ratio (OR) 0.95 (95% CI 0.92-0.98; q-value=0.01). MMP-8 showed a significant association with SVO, OR 1.24 (95% CI 1.06-1.45; q-value=0.03). Regarding MMP-12, a significant association with IS and LAA was observed: OR IS: 0.90 (95% CI 0.86-0.94; q-value=7.43×10^{-5}) and OR LAA: 0.71 (95% CI 0.65–0.77; q-value=5.11×10^{-14}). No significant associations with stroke outcome were observed (Table 13).

TABLE 13. Inverse variance weighted Mendelian randomization results.

Exposure	Outcome	nSNPs	Beta	SE	OR (CI 95%)	p-value	q-value
MMP-1	IS	50	-0.01	0.01	0.99 (0.97-1.00)	3.47×10^{-02}	7.79×10^{-02}
MMP-1	LAA	50	-0.05	0.02	0.95 (0.92-0.98)	2.09×10^{-03}	1.04×10^{-02}
MMP-1	CE	50	0.00	0.01	1.00 (0.98-1.02)	9.72×10^{-01}	9.72×10^{-01}
MMP-1	SVO	50	-0.03	0.01	0.97 (0.94-1.00)	3.82×10^{-02}	7.79×10^{-02}
MMP-1	mRS	46	-0.04	0.02	0.96 (0.93-1.00)	2.86×10^{-02}	7.79×10^{-02}
MMP-8	IS	6	0.02	0.03	1.02 (0.96-1.07)	5.98×10^{-01}	6.41×10^{-01}
MMP-8	LAA	6	-0.06	0.07	0.94 (0.82-1.08)	3.90×10^{-01}	4.88×10^{-01}
MMP-8	CE	6	-0.08	0.11	0.93 (0.75-1.14)	4.74×10^{-01}	5.47×10^{-01}
MMP-8	SVO	6	0.21	0.08	1.24 (1.06-1.45)	7.69×10^{-03}	2.88×10^{-02}
MMP-8	mRS	6	-0.08	0.08	0.92 (0.79-1.07)	2.82×10^{-01}	4.23×10^{-01}
MMP-12	IS	7	-0.11	0.02	0.90 (0.86-0.94)	4.96×10^{-05}	7.43×10^{-05}
MMP-12	LAA	6	-0. 35	0.04	0.71 (0.65-0.77)	3.41×10^{-15}	5.11×10^{-14}
MMP-12	CE	8	-0.04	0.04	0.96 (0.89-1.03)	2.72×10^{-01}	4.23×10^{-01}
MMP-12	SVO	8	-0.07	0.04	0.93 (0.86-1.00)	4.15×10^{-02}	7.79×10^{-02}
MMP-12	mRS	7	0.04	0.04	1.04 (0.96-1.12)	3.27×10^{-01}	4.46×10^{-01}

CI: Confidence Interval, IS: Ischemic Stroke, LAA: Large-Artery Atherosclerosis, CE: Cardioembolism, OR: Odds Ratio, SVO: Small-vessel Occlusion, and mRS: modified Rankin Scale.

5.3.2.3 Sensitivity tests

The complementary analysis proved consistency between the different tests (Figure 1) (Appendix table A10). Pleiotropy and heterogeneity were found to be significant for MMP-12 (Appendix table A11). The MR-PRESSO outlier test revealed two SNP outliers for LAA analysis and one SNP outlier for IS analysis (Appendix table A12-A13). No significant pleiotropy or heterogeneity were detected after removal of the outlier (Appendix table A11). The leave-one-out analysis showed that causality was not driven by a SNP (Appendix Figures A4-A7). The results for MMP-1 and MMP-8 were not biased by heterogeneity, pleiotropy or by a SNP (Appendix table A11, Appendix Figures A4-A7).

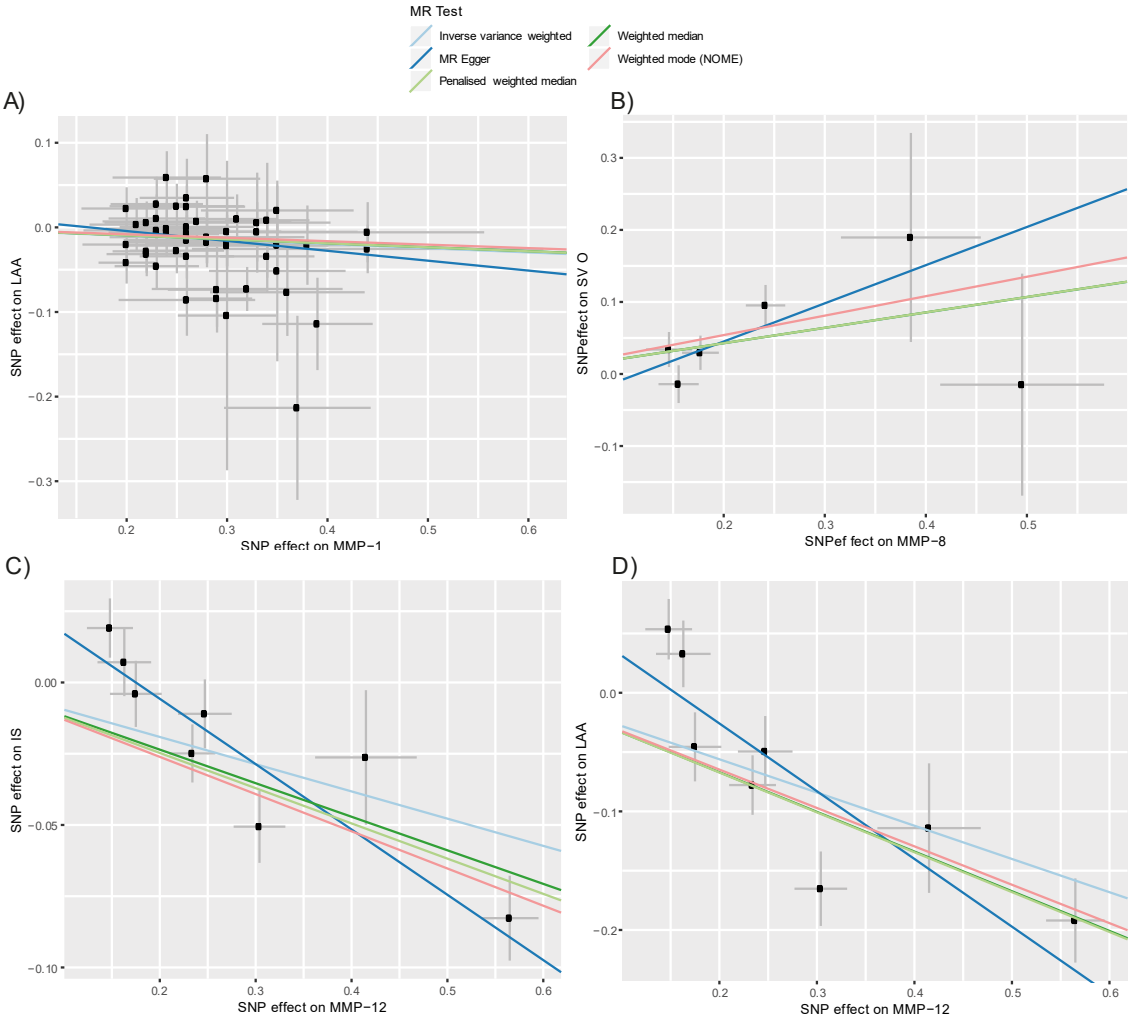


FIGURE 23. Scatterplots of significant Mendelian randomization results. A) Analysis of MMP-1 and Large-Artery Atherosclerosis (LAA), B) Analysis of MMP-8 and Small-Vessel Occlusion, C) Analysis of MMP-12 and Ischemic Stroke (IS), D) Analysis of MMP-12 and Large-Artery Atherosclerosis (LAA).

5.3.3 Appendix

5.3.3.1 Appendix Tables

Appendix tables: A7-A13:

APPENDIX TABLE A7. SNPs used in the analysis of MMP-1 serum levels as exposure.

SNP	EA	OA	AF	B	SE	P value	Analysis
rs495366	A	G	0.36	-0.44	0.18	5.73×10^{-34}	Risk and mRS
rs1942518	C	A	0.37	0.35	0.13	4.03×10^{-21}	Risk and mRS
rs7115014	A	G	0.35	0.32	0.10	3.99×10^{-19}	Risk and mRS
rs11226373	G	A	0.15	0.44	0.12	1.38×10^{-18}	Risk and mRS
rs685395	T	C	0.26	-0.33	0.10	3.22×10^{-17}	Risk and mRS
rs1944432	T	C	0.42	0.30	0.10	4.87×10^{-17}	Risk and mRS
rs10895597	A	G	0.40	-0.28	0.10	8.73×10^{-16}	Risk and mRS
rs666825	C	T	0.42	-0.28	0.08	2.40×10^{-15}	Risk and mRS
rs11602707	T	C	0.20	0.35	0.08	6.52×10^{-15}	Risk and mRS
rs1939052	C	T	0.40	0.27	0.09	1.14×10^{-14}	Risk and mRS
rs3819089	T	C	0.26	-0.30	0.08	1.07×10^{-13}	Risk and mRS
rs7939072	T	C	0.18	-0.33	0.07	6.31×10^{-13}	Risk and mRS
rs2466912	A	G	0.48	-0.26	0.07	1.06×10^{-12}	Risk and mRS
rs1940054	C	A	0.38	0.26	0.06	1.24×10^{-12}	Risk and mRS
rs613804	A	G	0.12	-0.39	0.06	1.91×10^{-12}	Risk and mRS
rs17710616	T	C	0.13	-0.37	0.07	5.29×10^{-12}	Risk and mRS
rs11225649	A	C	0.21	0.30	0.05	5.30×10^{-12}	Risk and mRS
rs2515081	T	A	0.14	-0.36	0.08	8.48×10^{-12}	Risk
rs7125424	T	G	0.13	-0.35	0.07	1.31×10^{-11}	Risk and mRS
rs7930146	C	T	0.22	-0.29	0.04	2.39×10^{-11}	Risk and mRS
rs480950	A	C	0.44	0.24	0.06	2.69×10^{-11}	Risk and mRS
rs12805072	T	C	0.27	0.26	0.07	2.94×10^{-11}	Risk and mRS
rs1939008	A	G	0.29	0.25	0.07	3.04×10^{-11}	Risk and mRS
rs1301783	G	A	0.50	0.22	0.03	6.58×10^{-11}	Risk and mRS
rs10791643	A	G	0.22	0.28	0.06	9.32×10^{-11}	Risk and mRS
rs523332	G	T	0.36	-0.23	0.07	1.25×10^{-10}	Risk and mRS
rs168636	G	A	0.24	0.26	0.05	2.99×10^{-10}	Risk and mRS
rs523519	A	C	0.39	0.23	0.06	3.93×10^{-10}	Risk and mRS
rs10502070	A	G	0.28	0.24	0.06	4.00×10^{-10}	Risk and mRS
rs11226865	A	G	0.16	0.31	0.04	4.64×10^{-10}	Risk and mRS
rs1531751	C	G	0.25	-0.24	0.05	6.09×10^{-10}	Risk
rs7946913	A	G	0.49	-0.22	0.04	9.66×10^{-10}	Risk and mRS
rs3181174	A	T	0.16	-0.29	0.05	1.38×10^{-09}	Risk
rs11220658	A	G	0.24	0.25	0.05	2.37×10^{-09}	Risk and mRS
rs3019723	C	T	0.20	-0.26	0.06	3.04×10^{-09}	Risk and mRS
rs4320978	G	A	0.22	0.26	0.04	3.27×10^{-09}	Risk and mRS
rs11224733	A	T	0.17	-0.28	0.05	3.50×10^{-09}	Risk
rs1815913	C	T	0.47	0.20	0.04	6.14×10^{-09}	Risk and mRS
rs502318	C	T	0.31	-0.23	0.05	6.66×10^{-09}	Risk and mRS
rs17106456	C	A	0.14	-0.30	0.05	7.99×10^{-09}	Risk and mRS

SNP	EA	OA	AF	B	SE	P value	Analysis
rs668285	T	C	0.24	0.23	0.05	8.76×10^{-09}	Risk and mRS
rs480846	A	G	0.09	0.34	0.05	1.19×10^{-08}	Risk and mRS
rs10502062	G	T	0.44	0.20	0.05	1.83×10^{-08}	Risk and mRS
rs4342998	A	C	0.38	0.22	0.04	1.89×10^{-08}	Risk and mRS
rs11217456	T	C	0.26	0.23	0.04	2.80×10^{-08}	Risk and mRS
rs17094347	T	C	0.19	-0.26	0.05	3.28×10^{-08}	Risk and mRS
rs11226791	C	T	0.08	0.34	0.04	3.31×10^{-08}	Risk and mRS
rs1052313	A	G	0.40	-0.20	0.03	3.35×10^{-08}	Risk and mRS
rs10890667	C	T	0.08	0.38	0.04	3.78×10^{-08}	Risk and mRS
rs7946787	C	A	0.36	-0.21	0.05	4.19×10^{-08}	Risk and mRS

AF: Allele Frequency of the EA; EA: Effect Allele; OA: Other Allele; SE: Standard Error; SNP: Single Nucleotide Polymorphism.

APPENDIX TABLE A8. SNPs used in the analysis of MMP-8 as exposure.

SNP	EA	OA	AF	B	SE	P value	Analysis
rs800292	A	G	0.30	-0.24	0.02	2.42×10^{-35}	Risk and Mrs
rs1409153	T	C	0.55	-0.18	0.02	1.81×10^{-22}	Risk and mRS
rs1560833	A	G	0.28	-0.16	0.02	5.31×10^{-15}	Risk and mRS
rs10922198	C	G	0.56	0.15	0.02	1.61×10^{-10}	Risk
rs193201657	T	C	0.98	0.50	0.08	1.02×10^{-09}	Risk and mRS
rs2184850	T	C	0.55	-0.12	0.02	2.14×10^{-08}	mRS
rs148136314	T	C	0.03	-0.39	0.07	2.57×10^{-08}	Risk and mRS

AF: Allele Frequency of the EA; EA: Effect Allele; OA: Other Allele; SE: Standard Error; SNP: Single Nucleotide Polymorphism.

APPENDIX TABLE A9. SNPs used in the analysis of MMP-12 as exposure.

SNP	EA	OA	AF	B	SE	P value	Analysis
rs499459	A	G	0.18	-0.57	0.03	8.26×10^{-76}	Risk and Mrs
rs1892971	A	G	0.26	0.30	0.03	7.97×10^{-29}	Risk and mRS
rs671188	C	T	0.44	-0.23	0.02	1.02×10^{-21}	Risk and mRS
rs2186789	G	T	0.24	-0.25	0.03	4.98×10^{-18}	Risk and mRS
rs613804	C	T	0.06	0.42	0.05	3.80×10^{-15}	Risk and mRS
rs1942524	C	T	0.27	0.18	0.03	1.38×10^{-10}	Risk and mRS
rs484915	A	T	0.41	0.15	0.02	1.65×10^{-09}	Risk
rs650108	A	G	0.25	0.16	0.03	5.39×10^{-09}	Risk and mRS

AF: Allele Frequency of the EA; EA: Effect Allele; OA: Other Allele; SE: Standard Error; SNP: Single Nucleotide Polymorphism.

APPENDIX TABLE A10. MR results of all analyses and methods.

Exposure	Outcome	Method	no SNP	B	SE	OR	p-value
MMP-1	IS	IVW	50	-0.01	0.01	0.99 (0.97-1.00)	3.47×10^{-02}
MMP-1	IS	MR Egger	50	-0.06	0.03	0.94 (0.88-1.00)	5.09×10^{-02}
MMP-1	IS	WM	50	-0.01	0.01	0.99 (0.97-1.00)	1.36×10^{-01}
MMP-1	IS	PWM	50	-0.01	0.01	0.99 (0.97-1.01)	1.58×10^{-01}
MMP-1	IS	WMode	50	-0.02	0.02	0.98 (0.95-1.02)	3.21×10^{-01}
MMP-1	LAA	IVW	50	-0.05	0.02	0.95 (0.92-0.98)	2.09×10^{-03}
MMP-1	LAA	MR Egger	50	-0.12	0.08	0.89 (0.76-1.04)	1.45×10^{-01}
MMP-1	LAA	WM	50	-0.05	0.02	0.95 (0.91-1.00)	3.73×10^{-02}
MMP-1	LAA	PWM	50	-0.05	0.02	0.95 (0.91-1.00)	4.42×10^{-02}
MMP-1	LAA	WMode	50	-0.04	0.04	0.96 (0.89-1.04)	3.01×10^{-01}
MMP-1	CE	IVW	50	0.00	0.01	1.00 (0.98-1.02)	9.72×10^{-01}
MMP-1	CE	MR Egger	50	0.04	0.06	1.04 (0.92-1.17)	5.35×10^{-01}
MMP-1	CE	WM	50	0.01	0.02	1.01 (0.97-1.04)	6.24×10^{-01}
MMP-1	CE	PWM	50	0.01	0.02	1.01 (0.97-1.05)	6.15×10^{-01}
MMP-1	CE	WMode	50	0.03	0.03	1.03 (0.96-1.10)	4.40×10^{-01}
MMP-1	SVO	IVW	50	-0.03	0.01	0.97 (0.94-1.00)	3.82×10^{-02}
MMP-1	SVO	MR Egger	50	-0.15	0.07	0.86 (0.75-0.99)	4.75×10^{-02}
MMP-1	SVO	WM	50	-0.02	0.02	0.98 (0.93-1.02)	2.55×10^{-01}
MMP-1	SVO	PWM	50	-0.02	0.02	0.98 (0.94-1.02)	3.12×10^{-01}
MMP-1	SVO	WMode	50	-0.03	0.04	0.97 (0.90-1.05)	4.97×10^{-01}
MMP-1	mRS	IVW	46	-0.04	0.02	0.96 (0.93-1.00)	2.86×10^{-02}
MMP-1	mRS	MR Egger	46	0.07	0.08	1.08 (0.92-1.26)	3.56×10^{-01}
MMP-1	mRS	WM	46	-0.04	0.02	0.96 (0.92-1.01)	9.97×10^{-02}
MMP-1	mRS	PWM	46	-0.04	0.02	0.96 (0.92-1.01)	9.19×10^{-02}
MMP-1	mRS	WMode	46	-0.02	0.05	0.98 (0.89-1.07)	6.12×10^{-01}
MMP-8	IS	IVW	6	0.02	0.03	1.02 (0.96-1.07)	5.98×10^{-01}
MMP-8	IS	MR Egger	6	0.04	0.11	1.04 (0.84-1.29)	7.54×10^{-01}
MMP-8	IS	WM	6	0.03	0.03	1.03 (0.96-1.10)	3.77×10^{-01}
MMP-8	IS	PWM	6	0.03	0.03	1.03 (0.96-1.10)	3.80×10^{-01}
MMP-8	IS	WMode	6	0.03	0.04	1.03 (0.95-1.11)	5.58×10^{-01}
MMP-8	LAA	IVW	6	-0.06	0.07	0.94 (0.82-1.08)	3.90×10^{-01}
MMP-8	LAA	MR Egger	6	0.00	0.28	1.00 (0.58-1.74)	9.99×10^{-01}
MMP-8	LAA	WM	6	-0.08	0.09	0.93 (0.78-1.10)	3.70×10^{-01}
MMP-8	LAA	PWM	6	-0.08	0.09	0.93 (0.78-1.09)	3.62×10^{-01}
MMP-8	LAA	WMode	6	-0.09	0.11	0.91 (0.74-1.13)	4.35×10^{-01}
MMP-8	CE	IVW	6	-0.08	0.11	0.93 (0.75-1.14)	4.74×10^{-01}
MMP-8	CE	MR Egger	6	-0.71	0.28	0.49 (0.29-0.84)	6.16×10^{-02}
MMP-8	CE	WM	6	-0.04	0.07	0.96 (0.83-1.10)	5.32×10^{-01}
MMP-8	CE	PWM	6	-0.04	0.07	0.96 (0.83-1.11)	5.84×10^{-01}
MMP-8	CE	WMode	6	-0.03	0.08	0.97 (0.83-1.12)	6.81×10^{-01}
MMP-8	SVO	IVW	6	0.21	0.08	1.24 (1.06-1.45)	7.69×10^{-03}
MMP-8	SVO	MR Egger	6	0.53	0.31	1.70 (0.92-3.15)	1.68×10^{-01}
MMP-8	SVO	WM	6	0.21	0.09	1.24 (1.03-1.49)	2.26×10^{-02}
MMP-8	SVO	PWM	6	0.21	0.09	1.24 (1.04-1.48)	1.91×10^{-02}

Exposure	Outcome	Method	no SNP	B	SE	OR	p-value
MMP-8	SVO	WMode	6	0.27	0.11	1.31 (1.06-1.62)	5.59×10^{-02}
MMP-8	mRS	IVW	6	-0.08	0.08	0.92 (0.79-1.07)	2.82×10^{-01}
MMP-8	mRS	MR Egger	6	-0.15	0.25	0.86 (0.53-1.41)	5.91×10^{-01}
MMP-8	mRS	WM	6	-0.04	0.09	0.96 (0.80-1.15)	6.39×10^{-01}
MMP-8	mRS	PWM	6	-0.04	0.09	0.96 (0.80-1.15)	6.49×10^{-01}
MMP-8	mRS	WMode	6	-0.05	0.10	0.95 (0.78-1.16)	6.55×10^{-01}
MMP-12	IS	IVW	7	-0.11	0.02	0.90 (0.86-0.94)	4.96×10^{-06}
MMP-12	IS	MR Egger	7	-0.21	0.04	0.81 (0.75-0.89)	4.80×10^{-03}
MMP-12	IS	WM	7	-0.12	0.02	0.89 (0.85-0.92)	2.65×10^{-08}
MMP-12	IS	PWM	7	-0.12	0.02	0.88 (0.85-0.92)	2.81×10^{-08}
MMP-12	IS	WMode	7	-0.14	0.02	0.87 (0.83-0.91)	1.20×10^{-03}
MMP-12	LAA	IVW	6	-0.34	0.04	0.71 (0.65-0.77)	3.41×10^{-15}
MMP-12	LAA	MR Egger	6	-0.38	0.13	0.68 (0.54-0.88)	3.92×10^{-02}
MMP-12	LAA	WM	6	-0.34	0.05	0.71 (0.64-0.79)	5.52×10^{-10}
MMP-12	LAA	PWM	6	-0.34	0.05	0.71 (0.64-0.79)	3.30×10^{-10}
MMP-12	LAA	WMode	6	-0.33	0.06	0.72 (0.64-0.81)	2.83×10^{-03}
MMP-12	CE	IVW	8	-0.04	0.04	0.96 (0.89-1.03)	2.72×10^{-01}
MMP-12	CE	MR Egger	8	-0.22	0.07	0.81 (0.7-0.93)	2.43×10^{-02}
MMP-12	CE	WM	8	-0.04	0.04	0.96 (0.89-1.04)	3.14×10^{-01}
MMP-12	CE	PWM	8	-0.05	0.04	0.95 (0.88-1.04)	2.57×10^{-01}
MMP-12	CE	WMode	8	-0.02	0.05	0.98 (0.89-1.08)	7.09×10^{-01}
MMP-12	SVO	IVW	8	-0.07	0.04	0.93 (0.86-1.00)	4.15×10^{-02}
MMP-12	SVO	MR Egger	8	-0.15	0.09	0.86 (0.73-1.02)	1.38×10^{-01}
MMP-12	SVO	WM	8	-0.08	0.05	0.92 (0.84-1.01)	8.83×10^{-02}
MMP-12	SVO	PWM	8	-0.08	0.05	0.92 (0.84-1.01)	7.85×10^{-02}
MMP-12	SVO	WMode	8	-0.08	0.06	0.92 (0.82-1.03)	2.09×10^{-01}
MMP-12	mRS	IVW	7	0.04	0.04	1.04 (0.96-1.12)	3.27×10^{-01}
MMP-12	mRS	MR Egger	7	0.04	0.10	1.04 (0.86-1.25)	7.28×10^{-01}
MMP-12	mRS	WM	7	0.03	0.05	1.03 (0.94-1.13)	5.39×10^{-01}
MMP-12	mRS	PWM	7	0.03	0.05	1.03 (0.94-1.13)	5.30×10^{-01}
MMP-12	mRS	WMode	7	0.03	0.05	1.03 (0.93-1.14)	6.03×10^{-01}

LAA: Large-Artery Atherosclerosis, SVO: Small-Vessel Occlusion, CE: Cardioembolism. IS: Ischemic Stroke, IVW: Inverse Variance Weighted, WM: Weighted Median, PWM: Penalized Weighted Median, WMode: Weighted Mode, OR: Odds ratio, no SNP: Number of SNPs.

APPENDIX TABLE A11. Pleiotropy and heterogeneity tests of all Mendelian randomization analyses.

		Pleiotropy			Heterogeneity	
Exposure	**Outcome**	**Intercept**	**SE**	**P-value**	**Q**	**P-value**
MMP-1	IS	0.014	0.01	0.12	51.61	3.72×10^{-01}
MMP-1	LAA	0.019	0.02	0.39	48.02	5.13×10^{-01}
MMP-1	CE	-0.011	0.02	0.52	45.95	5.98×10^{-01}
MMP-1	SVO	0.033	0.02	0.11	45.04	6.34×10^{-01}
MMP-1	mRS	-0.031	0.02	0.16	32.83	9.11×10^{-01}
MMP-8	IS	-0.004	0.02	0.85	3.02	6.96×10^{-01}
MMP-8	LAA	-0.012	0.05	0.83	1.96	8.54×10^{-01}
MMP-8	CE	0.126	0.05	0.07	17.53	3.59×10^{-03}
MMP-8	SVO	-0.061	0.06	0.36	6.91	2.27×10^{-01}
MMP-8	mRS	0.012	0.05	0.81	2.53	7.72×10^{-01}
MMP-12*	IS	0.04	0.01	0.01	13.11	1.63×10^{-03}
MMP-12	IS	0.031	0.01	0.05	12.33	5.50×10^{-02}
MMP-12*	LAA	0.088	0.04	0.08	29.67	1.04×10^{-04}
MMP-12	LAA	0.012	0.04	0.78	5.68	3.38×10^{-01}
MMP-12	CE	0.051	0.02	0.04	11.54	1.17×10^{-01}
MMP-12	SVO	0.022	0.02	0.39	5.2	6.36×10^{-01}
MMP-12	mRS	0.001	0.03	0.97	3.47	7.47×10^{-01}

IS: Ischemic Stroke, LAA: Large-Artery Atherosclerosis, CE: Cardioembolism, SVO: Small Vessel Occlusion, mRS: modified Rankin Scale. *Prior removal of significant outliers detected with MR-PRESSO outlier test.

APPENDIX TABLE A12. MR-PRESSO Outlier test for MMP-12 and IS.

SNP	**RSSobs**	**P-value**
rs1892971	6.39×10^{-04}	0.60
rs1942524	1.83×10^{-04}	1.00
rs2186789	1.99×10^{-04}	1.00
rs484915	1.23×10^{-03}	**0.01**
rs499459	2.08×10^{-03}	0.25
rs613804	2.11×10^{-04}	1.00
rs650108	5.68×10^{-04}	0.49
rs671188	8.61×10^{-06}	1.00

RSSobs: Observed Residual Sum of Squares, SNP: Single Nucleotide Polymorphism.

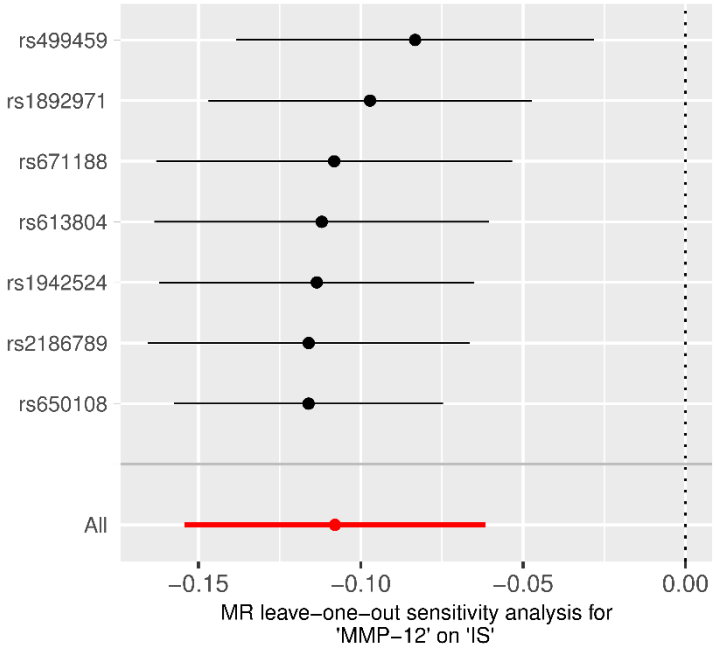
APPENDIX TABLE A13. MR-PRESSO Outlier test for MMP-12 and LAA.

SNP	RSSobs	P-value
rs1892971	8.68×10^{-03}	0.06
rs1942524	1.32×10^{-05}	1.00
rs2186789	4.83×10^{-04}	1.00
rs484915	1.00×10^{-02}	**0.01**
rs499459	2.97×10^{-03}	1.00
rs613804	5.79×10^{-06}	1.00
rs650108	6.83×10^{-03}	**0.04**
rs671188	1.98×10^{-04}	1.00

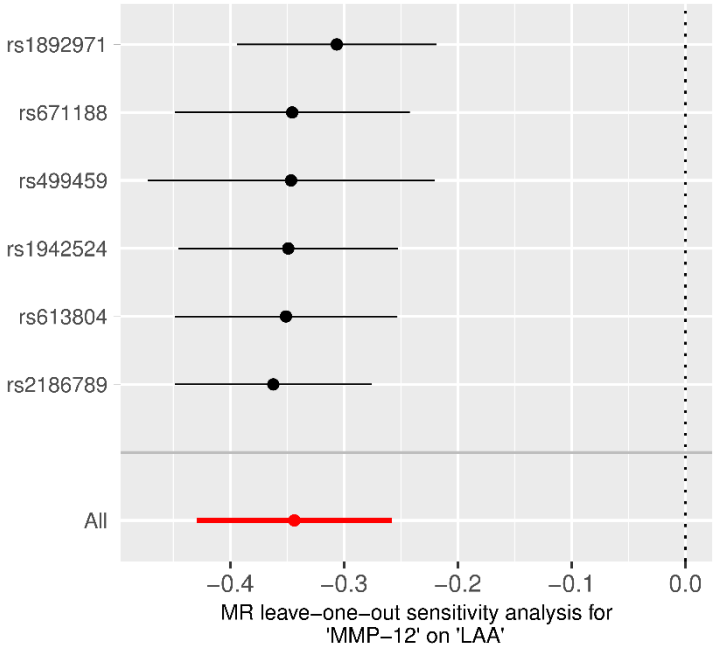
RSSobs: Observed Residual Sum of Squares, SNP: Single Nucleotide Polymorphism.

5.3.3.2 Appendix Figures

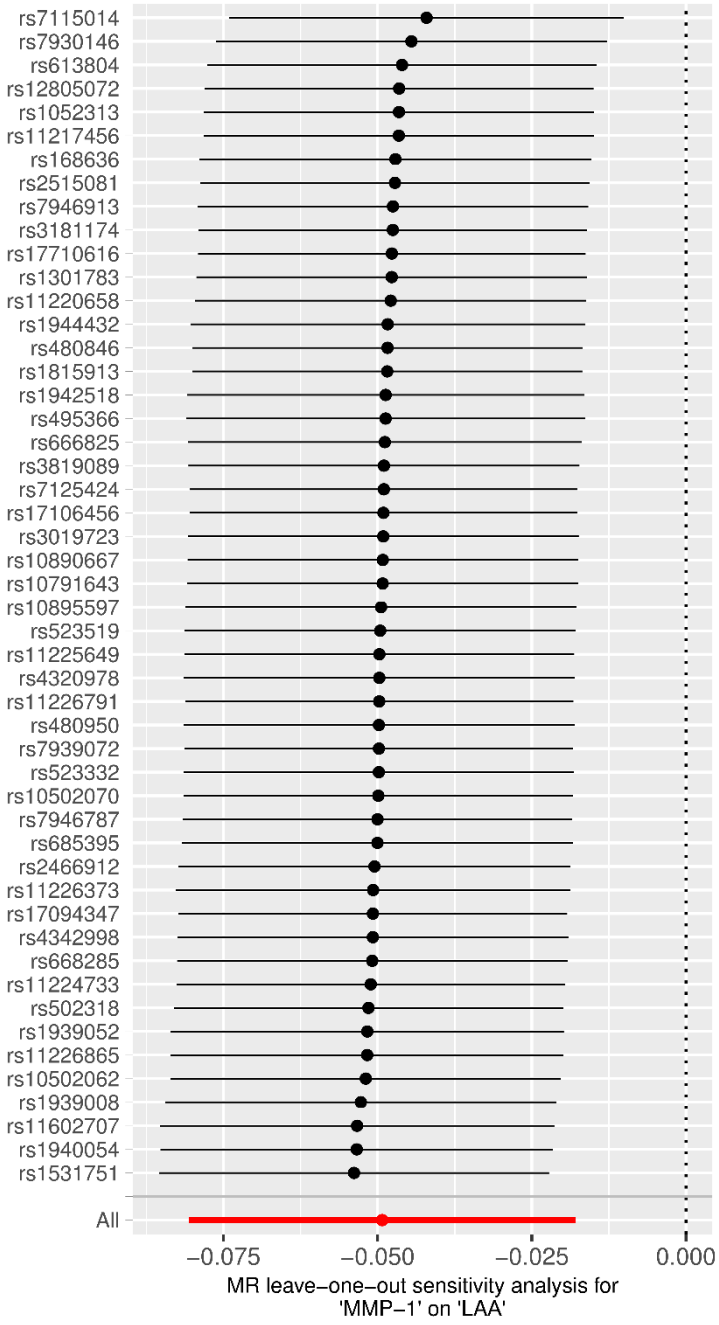
Appendix Figures: A3-A6:



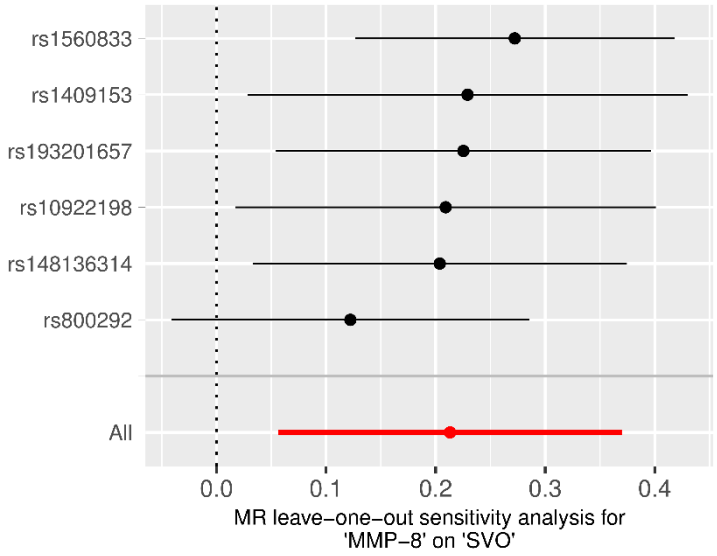
APPENDIX FIGURE A4. Leave-one-out analysis for MMP-12 on all ischemic strokes (IS).



APPENDIX FIGURE A5. Leave-one-out analysis for MMP-12 on Large-Artery Atherosclerosis (LAA) stroke.



APPENDIX FIGURE A6. Leave-one-out analysis for MMP-1 on Large-Artery Atherosclerosis (LAA) stroke.



APPENDIX FIGURE A7. Leave-one-out analysis for MMP-8 on Small-Vessel Occlusion (SVO) stroke.

6. Discussion

6. Discussion

6.1. GENERACION study

Using a GWAS approach we have performed the first analyses on ischemic stroke risk in the Spanish population, as well as the first analyses using sex stratification on stroke risk and subtypes. Following this approach, we gathered a total sample size of 9,111 individuals. We have been able to detect different significant loci and replicate two of them associated with SVO in an international cohort.

We have found a locus in chromosome 5, 5p15.2, lead variant rs59970332-T, significantly associated in the SVO both sexes' analysis. The nearest gene is *CTNND2*, it encodes for delta catenin 2. Replication was assessed for this locus in MEGASTROKE-SVO data (68). Variants within this gene have been found significantly associated with cognitive performance(180), which is a trait with a significant relation with small vessel disease(181). Interestingly, CTNND2 has also been also suggested as a potential plasma biomarker to discern between ischemic stroke patients and stroke mimics (182).

Following a stratified by sex approach, we could find locus 8p11.22, lead variant rs146966463-C, associated in the SVO only-females' analysis. The nearest gene is *TACC1*, it encodes for transforming Acidic Coiled-Coil-Containing Protein 1. Even though *TACC1* is the nearest gene, *FGFR1* gene was prioritized according to V2G. *FGFR1* encodes for fibroblast growth factor receptor 1 protein. Interestingly, different variants in *TACC1* and *FGFR1* genes have been described associated with cortical thickness evaluated with vertex-wise method (183). This measurement of the grey matter of the human cortex have been previously linked to white matter hyperintensities in the context of cerebral small-vessel disease, which play a crucial role in SVO (184).

In addition, we also wanted to explore which proteins are modified by the lead variants found in our study in relation to lacunar stroke. Unfortunately, proteins CTNND2, TACC1 and FGFR1, were not evaluated in the INTERVAL project, so we could not prove a direct effect as cis-pQTL of the associated loci. In contrary, we identified numerous proteins trans modulated, and we have been able to perform different pathway analysis. Derived from our results, rs59970332-T locus

could be potentially regulating cytokine production in blood, cytokines are well known as key part of pathophysiological processes such as inflammation that are gaining interest in the context or cSVD, because of promotion of brain-blood barrier disruption and endothelial disfunction, key processes in cSVD (185). In addition, another remarkable altered biological process is amyloid-beta clearance. Amyloid-beta accumulation has been described associated with enlarged perivascular spaces(186); an MRI marker described associated with SVO(187).

Regarding the biological processes modulated by rs146966463-C, locus associated with SVO in only females' analysis, interestingly our results suggest a modulation of response to endoplasmic reticulum stress. Process that has been associated in a CADASIL mice model(188). CADASIL is a monogenic cerebral small-vessel disease, which patients are prone to suffer from SVO at younger ages. Additionally, we found a decrease production of the anti-inflammatory IL-5 biological process significantly associated after FDR correction, as well as a reduction of other interleukins; IL-10 and IL-13; suggestively associated. Some immunological mechanisms have been described to differ between females and males hypothesized as modulation of different hormones' levels (189). This could be an explanation of the difference between genetic risk factor rs146966463-C in males and females SVO patients.

Regarding the replication of previously known genomic risk factors associated with ischemic stroke risk, we could replicate from 53 to 73 of the 81 evaluated loci, depending on replication in any stroke subtype. There is a caveat in this approximation that is that we have collaborated within the GIGASTROKE group with a substantial part of the GENERACION cohort. Despite of that, it is still informative to get a better understand about how many of the known loci are associated in each specific collaborative project and to understand if there are some loci that are not associated in the Spanish population.

Our study has limitations, the main one is the sample size. We are underpowered to detect small size effects. Even though, we were capable to find some genome-wide significant loci and replicate two of them in an international cohort. Another limitation is the lack of availability of a genetically equivalent cohort to the

GENERACION to replicate the significant loci. No other Spanish cohort have been described in terms of stroke genomics, so we cannot confirm if some of the genomic risk factors found in our study are specific of Spanish individuals.

To sum up, we conducted the first study on ischemic stroke genomics in the Spanish population, leading to the discovery of different genomic risk factors associated with risk. This is the first GWAS of ischemic stroke risk and subtypes stratified by sex. Replication was assessed for two loci associated with SVO, one of them specific of females' SVO. Further studies are guaranteed to fully understand the complexity of ischemic stroke genomics in Spain as well as the role of loci 5p15.2 and 8p11.22 in small-vessel occlusion risk and in sex differences.

6.2. Multitrait analysis of CE and AF

Using a MTAG with the two biggest cohorts of CE (68) and AF (66) to date, we found 44 genome-wide significant loci associated with CE. The prioritized genes of this loci were involved in biological processes such as cardiac conduction and contraction. Nevertheless, the 51 loci associated exclusively to AF (not associated to CE as showed the GWAS-pairwise) were mainly associated with cardiac development processes. This highlights the possible role in the risk of stroke due to AF of genes related to cardiac conduction and contraction instead of cardiac development process, and thereby would help to develop more specific prevention drugs.

Eleven loci significantly associated with CE were replicated in the independent GENERACION cohort. Their prioritized genes are the followings: *PITX2*, *ZFHX3*, *NKX2-5*, *CAV1*, *IGF1R*, *KIAA1755*, *NEURL1*, *GORAB*, *ESR2*, *ZEB2* and *WIPF1*. Of the genes associated with these loci, *PITX2*, *ZFHX3* and *NKX2-5* were already known associations with CE and AF. Eight were new CE associations: seven of them previously associated with AF: *CAV1*, *ESR2*, *GORAB*, *IGF1R*, *NEURL1*, *WIPF1* and *ZEB2*; and *KIAA1755*, a completely new association to CE, not being previously associated to AF.

One could think that by increasing the statistical power to find CE-associated SNPs through enrichment of AF patients, part of the associations is due to actually being associated only with AF. For this reason, we ensure that SNPs belonged to genomic regions associated with AF and CE through GWAS-pairwise (PPA-3 >0.6). Therefore, these 11 SNPs could be markers of stroke risk among patients with ESUS or among AF patients, as they are SNPs located in genomic regions that are not exclusively associated with either CE or AF, but with both.

Of the new loci associations with CE, we could highlight some genes. *CAV1* encodes caveolin-1, the principal structural component of caveolae organelles in smooth muscle cells and endothelial cells (190). Caveolin-1confers an anti-AF effect by mediating atrial structural remodeling through its antifibrotic action (191). Also, it plays a key role in how gas6 exerts its prothrombotic role in the vasculature (192). Genetic disruption of caveolin-1 in mice induces a severe biventricular hypertrophy with systolic and diastolic heart failure (193). That

supports the relevance that caveolin-1 might have in other causes of CE as symptomatic congestive heart failure with reduced ejection fraction (194), or its importance in ESUS as marker of an occult AF or left ventricular disfunction, which could benefit from anticoagulant treatment.

ESR2 encodes for the estrogen receptor beta, one of the receptors that mediates the biological effects of estrogens, which increase the levels of procoagulant factors VII, IX, X, XII and XIII, and reduce concentrations of the anticoagulant factor's protein S and antithrombin (195). Therefore, it might be a stroke risk marker.

IGF1R encodes the insulin-like growth factor (IGF) 1 receptor, that is the main receptor mediating IGF signaling in the heart (196). Inhibition of the IGF receptor decreases proliferation of cardiomyocytes in murine embryonic stem cells (197). ZEB2 encodes the zinc finger E-box-binding homeobox 2 protein that regulates cardiac fibroblast activation. An aberrant activation could lead to structural changes prone to develop AF.

KIAA1755 has not previously been found associated with AF. The index variant of this locus, rs3746471-A, encodes for R1045W amino acid change, predicted to be deleterious according to SIFT. rs3746471-A has been previously described associated with heart rate (198–200) and PR interval (201), and remarkably suggestively associated with stroke infarct volume (p-value=6.80×10^{-7}) (202,203). *KIAA1755* is predicted to encode an uncharacterized protein and is only characterized at the transcriptional level. The transcript is highly expressed in the brain and nerves and is also expressed in the heart.

We also found a novel locus suggestive to be associated with AF: 6q14.1, being *FILIP1* the prioritized gene linked with the leading SNP of the locus. This gene encodes a filamin A binding protein and has been identified as a regulator of myogenesis differentiation in human cells and in an in vivo mouse model (204). In the replication stage, this SNP was found suggestive (p-value=0.09), highly probable due to the small sample size in comparison to MTAG analysis.

The PRS generated with the SNPs from MTAG-CE here created was associated with CE independently of age, sex and hypertension, being simpler than other PRS that needs a major number of SNPs for association (205). We found that

the addition of our PRS to a model with age, sex and hypertension significantly improves the discriminatory power to detect CE.

Interestingly, the quantitative NRI was estimated in 14.16%, which is the proportion of cases correctly assigned to a higher probability of CE, among controls correctly assigned to a lower probability by an updated model adding our PRS compared with the initial model without it.

As limitations, the difference in the sample size between the two original studies could lead significant results to SNPs that are truly null for one trait but not for another, biasing effect-size estimates for the first trait and increasing the false discovery rate (and inflated type I error rate) (206). Nevertheless, MTAG estimation of χ^2 revealed a scenario expected to be strong against false positives, as tested in the original publication (206) and little evidence of genomic inflation was observed. Besides, we used GWAS-pairwise to ensure that the novel loci were not associated with only one of the traits, but with both at the same time, having a PPA-3 >0.6. But even more important, as usually in this kind of studies, we validated the significant loci found in this MTAG-CE and MTAG-AF in a GWAS of an independent European cohort. The small size of this last cohort underpower the ability to find significant results. However, we were able to replicate 11 leading SNPs from the total number of significant loci in the MTAG-CE and suggest one new potential locus in the MTAG-AF.

Another limitation is that we have only found loci associated with CE risk due to AF. Therefore, further multitrait analysis should be performed with different traits to uncover the different high-risk sources of CE. Nevertheless, our aim was to better characterize patients with CE due to AF as it is the most frequent cause of this type of stroke, for subsequently being able to find tools to detect those patients with a higher risk of developing a stroke due to an occult AF among ESUS for guiding future clinical trials with anticoagulant therapy.

In conclusion, we found and replicate eleven loci associated with CE, being eight of them new associations. We showed that their leading SNPs are in genomic regions related with both, CE and AF, suggesting that them, together with the creation of a PRS that improves the predictive models of CE, might allow to better

stratify the risk of stroke and its possible etiology for guide future clinical trials of anticoagulant therapy in AF or ESUS patients for a personalized medicine.

6.3. Mendelian Randomization of MMPs

The goal of this study was to clarify the causality of MMP serum levels and the risk of ischemic stroke, its subtypes and functional recovery after stroke evaluated as mRS at three months. Following an MR approach, we explored causality of MMP-1, MMP-8 and MMP-12, leading to significant and robust results in the sensitivity analysis for MMP-1 serum levels and the risk of LAA, MMP-8 serum levels and risk of SVO and MMP-12 serum levels and risk of LAA and IS.

Lower serum levels of MMP-1 were found to be a risk factor for LAA. Beneficial higher levels of MMP-1 in the biology of IS may be due to its role in extracellular matrix remodeling and repair by degradation of components such as collagen I, II and III(152).

With regard to MMP-8, a causal relationship of higher serum levels as a risk factor for SVO was detected. MMP-8 cleaves collagen I three times more effectively than MMP-1 or MMP-13, which makes this protein a potential disruptor of extracellular matrix compounds(207). Dysregulation of the matrisome by mutation in collagenase genes has been found to be the cause of small vessel diseases(208) such as SVO.

Regarding MMP-12, we found a causal relationship between lower plasma levels of MMP-12 and the risk of IS and LAA, which has been observed previously(49).

In our study, no significant association was observed between serum levels of MMP-1, MMP-8 and MMP-12 and mRS. These results suggest that these metalloproteinases do not play a crucial role in the process of stroke recovery, although other members of the MMP family may be involved. However, we must consider that the GODS study has a limited number of samples, and they are mainly of Spanish ancestry.

A strength of the present study is the consistency of the results across different MR tests and sensitivity analyses. Furthermore, our findings are unlikely to be confounded by population stratification because the analyses included individuals of European ancestry only.

In conclusion, using a Mendelian randomization approach our study suggests causality between lower serum levels of MMP-12 and the risk of ischemic stroke, lower serum levels of MMP-1 and MMP-12 and the risk of large-artery stroke and

higher serum levels of MMP-8 and the risk of lacunar stroke. Therefore, we believe our results suggest the potential use of these as biomarkers and therapeutic targets for stroke risk. However, further research is needed, in addition to an analysis of the 26 human MMPs, to understand the full biology of of these proteins in IS pathology.

7. Conclusions

7. Conclusions

Conclusions derived from the results obtained in the present doctoral thesis dissertation are the following:

1. Two genomic loci have been described associated with small-vessel occlusion in the GENERACION cohort. Locus 5p15.2 found associated in both sexes' analysis and locus 8p11.22 associated exclusively with females' SVO patients. These loci are suggested to be modulating plasma proteins involved in different pathological processes.
2. In the multitratit analysis of CE and AF, eleven loci were found replicated its association with CE, being eight of them new associations. We showed that their leading SNPs are in genomic regions jointly related with CE and AF.
3. The generated PRS has showed association to CE in GENERACION cohort as well as independence of age, sex and hypertension. Suggesting that PRS could allow to better stratify the risk of stroke patients.
4. Using a Mendelian Randomization approach results suggests causality between lower serum levels of MMP-12 and the risk of ischemic stroke, lower serum levels of MMP-1 and MMP-12 and the risk of large-artery atherosclerosis stroke and higher serum levels of MMP-8 and the risk of lacunar stroke.

8. Future perspectives

8. Future perspectives

Although, this thesis has answered several questions in relation to the genetic of ischemic stroke risk, it has undoubtedly left more questions unanswered that I hope the scientific community can continue studying in the future. From the results obtained in this thesis, the following lines of research can be derived.

Regarding the GENERACION project there is ample room for improvement. It will of great interest to generate an equivalent Spanish cohort to replicate the results obtained, as well as the increasing of the sample size which is already planned in Dr. Israel Fernandez's group. Additionally, different PRSs and could be generated with ischemic stroke genetics and related phenotypes to explore potential associtions in this unique cohort. In relation to the results of SVO analyses, functional studies are also needed to perform a precise prioritization of the causal genes involved in SVO risk, functional studies would involve RT-PCR to estimate if the variants are altering transcription of the near genes, as well as ELISAS to determine if it is a protein level differences the key in SVO risk.

In regards with the results obtained using multitrait analysis of GWAS to study the relation of cardioembolic stroke and atrial fibrillation. As mentioned before, cardioembolic stroke is not a homogenous group rather than several alterations can lead to a cardioembolic stroke. To fully uncover the genetic of CE in a multitrait approach different multitrait analyses should be performed, candidate traits considered risk sources of cardioembolic stroke would involve: mechanical prosthetic valve, dilated myocardiopathy, infective endocarditis, atrial myxoma and mitral rheumatic stenosis (209). In multitrait results, a novel missense variant in KIAA1755 associated with CE was obtained and replicated in the GENERACION cohort. Further functional analysis should be performed to evaluate if R1045W amino acid change in KIAA1755 has a deleterious effect in the activity of the protein. Additionally, in relation with the PRS results for CE other strategies could be evaluated, such as the development of PRS using machine learning algorithms for variant selection. At last, an evaluation of the CE PRS should be performed in ESUS patients with long term cardiac monitoring as it is planned in the financed project GENEtic and Clinical variables in the HUMT-AF Score to predict AF conversion risk in ESUS patients (GENECS-ESUS Study)

in this data it could be tested the hypothesis that PRS of CE and AF is useful to detect those CE due to AF patients in the ESUS category.

Finally in relation to the MMPs results, the results of the three MMPs studied showing a causal link to ischemic stroke risk, encourage the study of the rest MMPs that have been described, as well as their inhibitors. It also encourages the study of Mendelian Randomization using covariates to see if there are some MMPs that are acting independently or there is a balance of all of them. Additionally, a study that plan the follow up of individuals with altered levels of MMP-1, MMP-12 and MMP-8 would be of interest to quantify the exact risk of dysregulation of these MMPs and the risk of disease.

9. References

9. References

1. Overview of Stroke - Neurologic Disorders - MSD Manual Professional Edition [Internet]. [cited 2022 Aug 18]. Available from: https://www.msdmanuals.com/professional/neurologic-disorders/stroke/overview-of-stroke

2. Purroy F, Montalà N. Epidemiology of stroke in the last decade: a systematic review. Rev Neurol. 2021 Nov 1;73(9):321–36.

3. EL ATLAS DEL ICTUS ESPAÑA 2019.

4. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020;141(9):E139–596.

5. HP A, BH B, LJ K, J B, BB L, DL G, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. 1993;24(1):35–41.

6. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al. The Causative Classification of Stroke system: An international reliability and optimization study. Neurology. 2010 Oct 10;75(14):1277.

7. Hart RG, Diener HC, Coutts SB, Easton JD, Granger CB, O'Donnell MJ, et al. Embolic strokes of undetermined source: the case for a new clinical construct. Lancet Neurol [Internet]. 2014 [cited 2022 Oct 12];13(4):429–38. Available from: https://pubmed.ncbi.nlm.nih.gov/24646875/

8. Boehme AK, Esenwa C, Elkind MSV. Stroke Risk Factors, Genetics, and Prevention. Circ Res. 2017 Feb 2;120(3):472.

9. Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, et al. Executive summary: heart disease and stroke statistics--2012 update: a report from the American Heart Association. Circulation. 2012 Jan 3;125(1):188–97.

10. Gallego-Fabrega C, Muiño E, Cullell N, Cárcel-Márquez J, Lazcano U, Soriano-Tárraga C, et al. Biological Age Acceleration Is Lower in Women With Ischemic Stroke Compared to Men. Stroke [Internet]. 2022 Jul 1 [cited 2022 Oct 26];53(7):2320–30. Available from: https://pubmed.ncbi.nlm.nih.gov/35209739/

11. Kapral MK, Fang J, Hill MD, Silver F, Richards J, Jaigobin C, et al. Sex differences in stroke care and outcomes: results from the Registry of the Canadian Stroke Network. Stroke. 2005 Apr;36(4):809–14.

12. O'Donnell MJ, Denis X, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22

countries (the INTERSTROKE study): a case-control study. Lancet. 2010;376(9735):112–23.

13. Chen L, Yang Q, Ding RUI, Liu DAN, Chen Z. Carotid thickness and atherosclerotic plaque stability, serum inflammation, serum mmp-2 and mmp-9 were associated with acute cerebral infarction. Exp Ther Med. 2018;16(6):5253–7.

14. Ma F, Rodriguez S, Buxo X, Morancho A, Riba-Llena I, Carrera A, et al. Plasma Matrix Metalloproteinases in Patients With Stroke During Intensive Rehabilitation Therapy. Arch Phys Med Rehabil. 2016 Nov 1;97(11):1832–40.

15. Meschia JF, Bushnell C, Boden-Albala B, Braun LT, Bravata DM, Chaturvedi S, et al. Guidelines for the primary prevention of stroke: A statement for healthcare professionals from the American heart association/American stroke association. Stroke [Internet]. 2014 Dec 11 [cited 2022 Oct 27];45(12):3754–832. Available from: https://professional.heart.org/en/science-news/guidelines-for-the-primary-prevention-of-stroke

16. Hankey GJ. Stroke. Lancet. 2017 Feb 11;389(10069):641–54.

17. Katsanos AH, Psychogios K, Turc G, Sacco S, de Sousa DA, de Marchis GM, et al. Off-Label Use of Tenecteplase for the Treatment of Acute Ischemic Stroke: A Systematic Review and Meta-analysis. JAMA Netw Open [Internet]. 2022 Mar 1 [cited 2022 Oct 9];5(3):e224506–e224506. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2790581

18. Toul M, Mican J, Slonkova V, Nikitin D, Marek M, Bednar D, et al. Hidden Potential of Highly Efficient and Widely Accessible Thrombolytic Staphylokinase. Stroke. 2022 Aug 30;101161STROKEAHA122040219.

19. Chen HS, Cui Y, Li XQ, Wang XH, Ma YT, Zhao Y, et al. Effect of Remote Ischemic Conditioning vs Usual Care on Neurologic Function in Patients With Acute Moderate Ischemic Stroke: The RICAMIS Randomized Clinical Trial. JAMA. 2022 Aug 16;328(7):627–36.

20. Kapil N, Datta YH, Alakbarova N, Bershad E, Selim M, Liebeskind DS, et al. Antiplatelet and Anticoagulant Therapies for Prevention of Ischemic Stroke. Clin Appl Thromb Hemost. 2017 May 1;23(4):301–18.

21. Kwah LK, Diong J. National Institutes of Health Stroke Scale (NIHSS). J Physiother. 2014 Mar 1;60(1):61.

22. Heitsch L, Ibanez L, Carrera C, Binkley MM, Strbian D, Tatlisumak T, et al. Early Neurological Change After Ischemic Stroke Is Associated With 90-Day Outcome. Stroke [Internet]. 2021 [cited 2021 Nov 22];52(1):132–41. Available from: https://pubmed.ncbi.nlm.nih.gov/33317415/

23. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. Stroke. 2007 Mar;38(3):1091–6.

24. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.

25. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001 409:6822. 2001 Feb 15;409(6822):860–921.

26. Del Giacco L, Cattaneo C. Introduction to genomics. Methods Mol Biol. 2012;823:79–88.

27. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. Journal of Human Genetics 2010 55:7 [Internet]. 2010 May 20 [cited 2022 Sep 1];55(7):403–15. Available from: https://www.nature.com/articles/jhg201055

28. Ewens WJ. Genetic Variation. Brenner's Encyclopedia of Genetics: Second Edition. 2013 Jan 1;290–1.

29. What is genetic variation | Human genetic variation [Internet]. [cited 2022 Sep 1]. Available from: https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/

30. Day INM. dbSNP in the detail and copy number complexities. Hum Mutat [Internet]. 2010 Jan 1 [cited 2022 Oct 9];31(1):2–4. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.21149

31. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Vol. 526, Nature. Nature Publishing Group; 2015. p. 68–74.

32. Sukhumsirichart W. Polymorphisms. Genetic Diversity and Disease Susceptibility. 2018 Oct 17;

33. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017 Oct 11;550(7675):204–13.

34. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, et al. Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. Am J Hum Genet. 2018 Nov 1;103(5):654–65.

35. Currin KW, Erdos MR, Narisu N, Rai V, Vadlamudi S, Perrin HJ, et al. Genetic effects on liver chromatin accessibility identify disease regulatory variants. The American Journal of Human Genetics. 2021 Jul 1;108(7):1169–89.

36. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature [Internet]. 2018 Jun 6 [cited 2022 Sep 1];558(7708):73. Available from: /pmc/articles/PMC6697541/

37. X M, S C, R D, Z Z, X Z, S G, et al. ipaQTL-atlas: an atlas of intronic polyadenylation quantitative trait loci across human tissues. Nucleic Acids Res. 2022;1(1256879):13–4.

38. Del Rosario RCH, Poschmann J, Rouam SL, Png E, Khor CC, Hibberd ML, et al. Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. Nat Methods. 2015 Apr 29;12(5):458–64.

39. Ma L, Jia P, Zhao Z. Splicing QTL of human adipose-related traits. Scientific Reports 2017 8:1. 2018 Jan 10;8(1):1–5.

40. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics 2019 20:8. 2019 May 8;20(8):467–84.

41. Illumina Microarray Technology [Internet]. [cited 2022 Sep 6]. Available from: https://www.illumina.com/science/technology/microarray.html

42. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct 1;48(10):1279–83.

43. McMahon A, Lewis E, Buniello A, Cerezo M, Hall P, Sollis E, et al. Sequencing-based genome-wide association studies reporting standards. Cell Genomics. 2021 Oct 13;1(1):100005.

44. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016 Mar 23;17(1):53.

45. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016 May 2;534(7606):200–5.

46. Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Feb 25;47(3):291–5.

47. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. Am J Hum Genet. 2017 Dec 7;101(6):939–64.

48. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019 Jul 1;8(7).

49. Chong M, Sjaarda J, Pigeyre M, Mohammadi-Shemirani P, Lali R, Shoamanesh A, et al. Novel Drug Targets for Ischemic Stroke Identified

Through Mendelian Randomization Analysis of the Blood Proteome. Circulation. 2019 Sep 3;140(10):819–30.

50. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res [Internet]. 2019 Jan 8 [cited 2022 Oct 27];47(D1):D1005–12. Available from: https://pubmed.ncbi.nlm.nih.gov/30445434/

51. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. bioRxiv [Internet]. 2020 Aug 10 [cited 2022 Oct 27];2020.08.10.244293. Available from: https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1

52. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018 Feb 1;50(2):229–37.

53. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020 May 18;12(1).

54. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols 2020 15:9. 2020 Jul 24;15(9):2759–72.

55. Steinthorsdottir V, McGinnis R, Williams NO, Stefansdottir L, Thorleifsson G, Shooter S, et al. Genetic predisposition to hypertension is associated with preeclampsia in European and Central Asian women. Nat Commun. 2020 Dec 1;11(1).

56. Carrera C, Cullell N, Torres-Águila N, Muiño E, Bustamante A, Dávalos A, et al. Validation of a clinical-genetics score to predict hemorrhagic transformations after rtPA. Neurology. 2019;93(9):e851–63.

57. Davies NM, Holmes M V., Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362.

58. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nature Genetics 2015 47:8. 2015 Jun 29;47(8):856–60.

59. Konuma T, Ogawa K, Okada Y. Integration of genetically regulated gene expression and pharmacological library provides therapeutic drug candidates. Hum Mol Genet. 2021 Apr 26;30(3–4):294–304.

60. Sakaue S, Okada Y. GREP: genome for REPositioning drugs. Bioinformatics. 2019 Oct 1;35(19):3821–3.

61. Bevan S, Traylor M, Adib-Samii P, Malik R, Paul NLM, Jackson C, et al. Genetic heritability of ischemic stroke and the contribution of previously

reported candidate gene and genomewide associations. Stroke. 2012 Dec;43(12):3161–7.

62. Matarín M, Brown WM, Scholz S, Simón-Sánchez J, Fung HC, Hernandez D, et al. A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Lancet Neurology. 2007;6(5):414–20.

63. Malik R, Traylor M, Pulit SL, Bevan S, Hopewell JC, Holliday EG, et al. Low-frequency and common genetic variation in ischemic stroke: The METASTROKE collaboration. Neurology. 2016 Mar 29;86(13):1217–26.

64. Williams FMK, Carter AM, Hysi PG, Surdulescu G, Hodgkiss D, Soranzo N, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. Ann Neurol. 2013;73(1):16–31.

65. Das S, Natarajan R. HDAC9: An inflammatory link in atherosclerosis. Circ Res. 2020 Aug 28;127:824–6.

66. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet [Internet]. 2018 Sep 1 [cited 2020 Mar 9];50(9):1234–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30061737

67. (ISGC) NSGN (SiGN) and ISGC, Pulit SL, McArdle PF, Wong Q, Malik R, Gwinn K, et al. The NINDS Stroke Genetics Network: a genome-wide association study of ischemic stroke and its subtypes. Lancet Neurol. 2016 Feb 1;15(2):174.

68. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018 Apr 1;50(4):524–37.

69. Malik R, Rannikmäe K, Traylor M, Georgakis MK, Sargurupremraj M, Markus HS, et al. Genome-wide meta-analysis identifies 3 novel loci associated with stroke. Ann Neurol [Internet]. 2018 Dec 1 [cited 2022 Oct 10];84(6):934–9. Available from: https://pubmed.ncbi.nlm.nih.gov/30383316/

70. Jeanne M, Gould DB. Genotype-phenotype correlations in pathology caused by collagen type IV alpha 1 and 2 mutations. Matrix Biol [Internet]. 2017 Jan 1 [cited 2022 Oct 10];57–58:29–44. Available from: https://pubmed.ncbi.nlm.nih.gov/27794444/

71. Rozen EJ, Roewenstrunk J, Barallobre MJ, di Vona C, Jung C, Figueiredo AF, et al. DYRK1A Kinase Positively Regulates Angiogenic Responses in Endothelial Cells. Cell Rep [Internet]. 2018 May 8 [cited 2022 Oct 10];23(6):1867–78. Available from: https://pubmed.ncbi.nlm.nih.gov/29742440/

72. Farah C, Michel LYM, Balligand JL. Nitric oxide signalling in cardiovascular health and disease. Nat Rev Cardiol [Internet]. 2018 May 1 [cited 2022 Oct 10];15(5):292–316. Available from: https://pubmed.ncbi.nlm.nih.gov/29388567/

73. Zhou W, Kanai M, Wu KHH, Rasheed H, Tsuo K, Hirbo JB, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genomics [Internet]. 2022 Oct 12 [cited 2022 Nov 4];2(10):100192. Available from: http://www.cell.com/article/S2666979X22001410/fulltext

74. Zhou W, Kanai M, Wu KHH, Humaira R, Tsuo K, Hirbo JB, et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases. medRxiv [Internet]. 2021 Nov 21 [cited 2022 Oct 10];2021.11.19.21266436. Available from: https://www.medrxiv.org/content/10.1101/2021.11.19.21266436v1

75. Mishra A, Malik R, Hachiya T, Jürgenson T, Namba S, Posner DC, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. Nature 2022 [Internet]. 2022 Sep 30 [cited 2022 Oct 10];1–15. Available from: https://www.nature.com/articles/s41586-022-05165-3

76. Lyden P, Pryor KE, Coffey CS, Cudkowicz M, Conwit R, Jadhav A, et al. Final Results of the RHAPSODY Trial: A Multi-Center, Phase 2 Trial Using a Continual Reassessment Method to Determine the Safety and Tolerability of 3K3A-APC, A Recombinant Variant of Human Activated Protein C, in Combination with Tissue Plasminogen Activator, Mechanical Thrombectomy or both in Moderate to Severe Acute Ischemic Stroke. Ann Neurol [Internet]. 2019 Jan 1 [cited 2022 Oct 10];85(1):125–36. Available from: https://pubmed.ncbi.nlm.nih.gov/30450637/

77. Mola-Caminal M, Carrera C, Soriano-Tárraga C, Giralt-Steinhauer E, Díaz-Navarro RM, Tur S, et al. *PATJ* Low Frequency Variants Are Associated With Worse Ischemic Stroke Functional Outcome. Circ Res. 2019;124(1):114–20.

78. Ibanez L, Heitsch L, Carrera C, Farias FHG, del Aguila JL, Dhar R, et al. Multi-ancestry GWAS reveals excitotoxicity associated with outcome after ischaemic stroke. Brain [Internet]. 2022 Jul 1 [cited 2022 Sep 12];145(7):2394–406. Available from: https://pubmed.ncbi.nlm.nih.gov/35213696/

79. Carrera C, Cárcel-Márquez J, Cullell N, Torres-Águila N, Muiño E, Castillo J, et al. Single nucleotide variations in ZBTB46 are associated with post-thrombolytic parenchymal haematoma. Brain. 2021 Aug 1;144(8):2416.

80. Traylor M, Persyn E, Tomppo L, Klasson S, Abedi V, Bakker MK, et al. Genetic basis of lacunar stroke: a pooled analysis of individual patient data and genome-wide association studies. Lancet Neurol. 2021 May 1;20(5):351–61.

81. Chung J, Marini S, Pera J, Norrving B, Jimenez-Conde J, Roquer J, et al. Genome-wide association study of cerebral small vessel disease reveals established and novel loci. Brain. 2019 Oct 1;142(10):3176–89.

82. Wassertheil-Smoller S, Qi Q, Dave T, Mitchell BD, Jackson RD, Liu S, et al. Polygenic risk for depression increases risk of ischemic stroke: From the stroke genetics network study. Stroke [Internet]. 2018 [cited 2022 Sep 13];49(3):543–8. Available from: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.117.018857

83. Harris SE, Malik R, Marioni R, Campbell A, Seshadri S, Worrall BB, et al. Polygenic risk of ischemic stroke is associated with cognitive ability. Neurology [Internet]. 2016 Feb 16 [cited 2022 Sep 13];86(7):611–8. Available from: https://pubmed.ncbi.nlm.nih.gov/26695942/

84. Hou L, Xu M, Yu Y, Sun X, Liu X, Liu L, et al. Exploring the causal pathway from ischemic stroke to atrial fibrillation: a network Mendelian randomization study. Mol Med. 2020 Jan 15;26(1).

85. Park S, Lee S, Kim Y, Cho S, Kim K, Kim YC, et al. Causal effects of atrial fibrillation on brain white and gray matter volume: a Mendelian randomization study. BMC Med. 2021 Dec 1;19(1).

86. Larsson SC, Mason AM, Bäck M, Klarin D, Damrauer SM, Michaëlsson K, et al. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. Eur Heart J. 2020 Sep 14;41(35):3304–10.

87. Larsson SC, Scott RA, Traylor M, Langenberg CC, Hindy G, Melander O, et al. Type 2 diabetes, glucose, insulin, BMI, and ischemic stroke subtypes: Mendelian randomization study. Neurology. 2017 Aug 1;89(5):454–60.

88. Gan W, Bragg F, Walters RG, Millwood IY, Lin K, Chen Y, et al. Genetic Predisposition to Type 2 Diabetes and Risk of Subclinical Atherosclerosis and Cardiovascular Diseases Among 160,000 Chinese Adults. Diabetes. 2019 Nov 1;68(11):2155–64.

89. Georgakis MK, Harshfield EL, Malik R, Franceschini N, Langenberg C, Wareham NJ, et al. Diabetes Mellitus, Glycemic Traits, and Cerebrovascular Disease: A Mendelian Randomization Study. Neurology. 2021 Mar 30;96(13):e1732–42.

90. Liu J, Rutten-Jacobs L, Liu M, Markus HS, Traylor M. Causal Impact of Type 2 Diabetes Mellitus on Cerebral Small Vessel Disease: A Mendelian Randomization Analysis. Stroke. 2018;49(6):1325–31.

91. Fatumo S, Karhunen V, Chikowore T, Sounkou T, Udosen B, Ezenwa C, et al. Metabolic Traits and Stroke Risk in Individuals of African Ancestry: Mendelian Randomization Analysis. Stroke. 2021;52(8):2680–4.

92. Dale CE, Fatemifar G, Palmer TM, White J, Prieto-Merino D, Zabaneh D, et al. Causal Associations of Adiposity and Body Fat Distribution With

Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. Circulation. 2017 Jun 13;135(24):2373–88.

93. Marini S, Merino J, Montgomery BE, Malik R, Sudlow CL, Dichgans M, et al. Mendelian Randomization Study of Obesity and Cerebrovascular Disease. Ann Neurol. 2020 Apr 1;87(4):516–24.

94. Qian Y, Ye D, Wu DJH, Feng C, Zeng Z, Ye L, et al. Role of cigarette smoking in the development of ischemic stroke and its subtypes: a Mendelian randomization study. Clin Epidemiol. 2019;11:725–31.

95. Harshfield EL, Georgakis MK, Malik R, Dichgans M, Markus HS. Modifiable Lifestyle Factors and Risk of Stroke: A Mendelian Randomization Analysis. Stroke. 2021;(March):931–6.

96. Zhuang Z, Gao M, Yang R, Li N, Liu Z, Cao W, et al. Association of physical activity, sedentary behaviours and sleep duration with cardiovascular diseases and lipid profiles: A Mendelian randomization analysis. Lipids Health Dis. 2020 May 8;19(1).

97. Bahls M, Leitzmann MF, Karch A, Teumer A, Dörr M, Felix SB, et al. Physical activity, sedentary behavior and risk of coronary artery disease, myocardial infarction and ischemic stroke: a two-sample Mendelian randomization study. Clin Res Cardiol. 2021 Oct 1;110(10):1564–73.

98. Lu H, Wu PF, Li RZ, Zhang W, Huang GX. Sleep Duration and Stroke: A Mendelian Randomization Study. Front Neurol. 2020 Oct 7;11:976.

99. Titova OE, Michaëlsson K, Larsson SC. Sleep Duration and Stroke: Prospective Cohort Study and Mendelian Randomization Analysis. Stroke. 2020;3279–85.

100. Cai H, Liang J, Liu Z, Fang L, Zheng J, Xu J, et al. Causal effects of sleep traits on ischemic stroke and its subtypes: A mendelian randomization study. Nat Sci Sleep. 2020 Oct 21;12:783–90.

101. Larsson SC, Burgess S, Mason AM, Michaëlsson K. Alcohol Consumption and Cardiovascular Disease: A Mendelian Randomization Study. Circ Genom Precis Med. 2020;13(3).

102. Lankester J, Zanetti D, Ingelsson E, Assimes TL. Alcohol use and cardiometabolic risk in the UK Biobank: A Mendelian randomization study. PLoS One. 2021 Aug 1;16(8).

103. Qian Y, Ye D, Huang H, Wu DJH, Zhuang Y, Jiang X, et al. Coffee Consumption and Risk of Stroke: A Mendelian Randomization Study. Ann Neurol. 2020 Apr 1;87(4):525–32.

104. Zhang Z, Wang M, Yuan S, Cai H, Zhu SG, Liu X. Genetically Predicted Coffee Consumption and Risk of Alzheimer's Disease and Stroke. J Alzheimers Dis. 2021;83(4):1815–23.

105. Li J, Guasch-Ferré M, Chung W, Ruiz-Canela M, Toledo E, Corella D, et al. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk. Eur Heart J. 2020 Jul 21;41(28):2645–56.

106. Xiuyun W, Qian W, Minjun X, Weidong L, Lizhen L. Education and stroke: evidence from epidemiology and Mendelian randomization study. Sci Rep. 2020;10(1):1–11.

107. Carter AR, Gill D, Davies NM, Taylor AE, Tillmann T, Vaucher J, et al. Understanding the consequences of education inequality on cardiovascular disease: Mendelian randomisation study. The BMJ. 2019;365:1–12.

108. Gill D, Efstathiadou A, Cawood K, Tzoulaki I, Dehghan A. Education protects against coronary heart disease and stroke independently of cognitive function: Evidence from Mendelian randomization. Int J Epidemiol. 2019;48(5):1468–77.

109. Wang M, Bai Y, Wang Z, Zhang Z, Liu D, Lian X. Higher tea consumption is associated with decreased risk of small vessel stroke. Clin Nutr. 2021 Mar 1;40(3):1430–5.

110. Lin J, Wang Y, Wang Y, Pan Y. Inflammatory biomarkers and risk of ischemic stroke and subtypes: A 2-sample Mendelian randomization study. Neurol Res. 2020 Feb 1;42(2):118–25.

111. Georgakis MK, Malik R, Gill D, Franceschini N, Sudlow CLM, Dichgans M. Interleukin-6 Signaling Effects on Ischemic Stroke and Other Cardiovascular Outcomes: A Mendelian Randomization Study. Circ Genom Precis Med. 2020;13(3).

112. Song L, Sun J, Söderholm M, Melander O, Orho-Melander M, Nilsson J, et al. Association of TIM-1 (T-Cell Immunoglobulin and Mucin Domain 1) With Incidence of Stroke. Arterioscler Thromb Vasc Biol. 2020;40(7):1777–86.

113. Yuan S, Lin A, He Q qiang, Burgess S, Larsson SC. Circulating interleukins in relation to coronary artery disease, atrial fibrillation and ischemic stroke and its subtypes: A two-sample Mendelian randomization study. Int J Cardiol. 2020 Aug 15;313:99–104.

114. Georgakis MK, Gill D, Rannikmäe K, Traylor M, Anderson CD, Lee JM, et al. Genetically Determined Levels of Circulating Cytokines and Risk of Stroke. Circulation. 2019 Jan 8;139(2):256–68.

115. Huang T, Afzal S, Yu C, Guo Y, Bian Z, Yang L, et al. Vitamin D and cause-specific vascular disease and mortality: a Mendelian randomisation study involving 99,012 Chinese and 106,911 European adults. BMC Med. 2019 Aug 30;17(1).

116. Larsson SC, Traylor M, Markus HS. Circulating Vitamin K_1 Levels in Relation to Ischemic Stroke and Its Subtypes: A Mendelian Randomization Study. Nutrients. 2018 Nov 1;10(11).

117. Yuan S, Zheng JS, Mason AM, Burgess S, Larsson SC. Genetically predicted circulating vitamin C in relation to cardiovascular disease. Eur J Prev Cardiol. 2021 May 31;

118. Zhu J, Ling Y, Tse LA, Kinra S, Li Y. Circulating vitamin C and the risk of cardiovascular diseases: A Mendelian randomization study. Nutr Metab Cardiovasc Dis. 2021 Jul 22;31(8):2398–406.

119. Chen L, Sun X, Wang Z, Lu Y, Chen M, He Y, et al. The impact of plasma vitamin C levels on the risk of cardiovascular diseases and Alzheimer's disease: A Mendelian randomization study. Clin Nutr. 2021 Oct 1;40(10):5327–34.

120. Chan YH, Schooling CM, Zhao J, Au Yeung SL, Hai JJ, Thomas GN, et al. Mendelian Randomization Focused Analysis of Vitamin D on the Secondary Prevention of Ischemic Stroke. Stroke. 2021 Dec;52(12):3926–37.

121. Yuan T, Si S, Li Y, Li W, Chen X, Liu C, et al. Roles for circulating polyunsaturated fatty acids in ischemic stroke and modifiable factors: a Mendelian randomization study. Nutr J. 2020 Jul 11;19(1).

122. Harshfield EL, Sims MC, Traylor M, Ouwehand WH, Markus HS. The role of haematological traits in risk of ischaemic stroke and its subtypes. Brain. 2020 Jan 1;143(1):210–21.

123. Gill D, Georgakis MK, Laffan M, Sabater-Lleal M, Malik R, Tzoulaki I, et al. Genetically Determined FXI (Factor XI) Levels and Risk of Stroke. Stroke. 2018;49(11):2761–3.

124. Gill D, Monori G, Tzoulaki I, Dehghan A. Iron Status and Risk of Stroke. Stroke. 2018;49(12):2815–21.

125. Heitsch L, Ibanez L, Carrera C, Pera J, Jimenez-Conde J, Slowik A, et al. Meta-analysis of Transethnic Association (MANTRA) Reveals Loci Associated With Neurological Instability After Acute Ischemic Stroke. In: International Stroke Conference. 2017.

126. Mola-Caminal M, Carrera C, Soriano-Tárraga C, Giralt-Steinhauer E, Díaz-Navarro RM, Tur S, et al. PATJ Low Frequency Variants Are Associated with Worse Ischemic Stroke Functional Outcome: A Genome-Wide Meta-Analysis. Circ Res. 2019 Jan 4;124(1):114–20.

127. Domingues-Montanari S, Fernández-Cadenas I, Del Río-Espinola A, Mendioroz M, Fernandez-Morales J, Corbeto N, et al. KCNK17 genetic variants in ischemic stroke. Atherosclerosis. 2010 Jan;208(1):203–9.

128. Obón-Santacana M, Vilardell M, Carreras A, Duran X, Velasco J, Galván-Femenía I, et al. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open. 2018 Mar;8(3):e018324.

129. Domingues-Montanari S, Fernández-Cadenas I, Del Río-Espinola A, Mendioroz M, Fernandez-Morales J, Corbeto N, et al. KCNK17 genetic variants in ischemic stroke. Atherosclerosis. 2010 Jan;208(1):203–9.

130. Fernández-Cadenas I, Mendióroz M, Giralt D, Nafria C, Garcia E, Carrera C, et al. GRECOS Project (Genotyping Recurrence Risk of Stroke). Stroke. 2017 May;48(5):1147–53.

131. Riba I, Jarca CI, Mundet X, Tovar JL, Orfila F, Nafría C, et al. Cognitive assessment protocol design in the ISSYS (Investigating Silent Strokes in hYpertensives: A magnetic resonance imaging Study). J Neurol Sci. 2012 Nov;322(1–2):79–81.

132. Heitsch L, Ibanez L, Carrera C, Pera J, Jimenez-Conde J, Slowik A, et al. Meta-analysis of Transethnic Association (MANTRA) Reveals Loci Associated With Neurological Instability After Acute Ischemic Stroke. In: International Stroke Conference. 2017.

133. Fernández-Cadenas I, Mendióroz M, Giralt D, Nafria C, Garcia E, Carrera C, et al. GRECOS Project (Genotyping Recurrence Risk of Stroke). Stroke. 2017 May;48(5):1147–53.

134. Riba I, Jarca CI, Mundet X, Tovar JL, Orfila F, Nafría C, et al. Cognitive assessment protocol design in the ISSYS (Investigating Silent Strokes in hYpertensives: A magnetic resonance imaging Study). J Neurol Sci. 2012 Nov;322(1–2):79–81.

135. Obón-Santacana M, Vilardell M, Carreras A, Duran X, Velasco J, Galván-Femenía I, et al. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open. 2018 Mar;8(3):e018324.

136. Larrue V, Von Kummer R, Del Zoppo G, Bluhmki E. Hemorrhagic transformation in acute ischemic stroke: Potential contributing factors in the European Cooperative Acute Stroke Study. Stroke. 1997;28(5):957–60.

137. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.

138. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019 Dec 1;51(12):1749–55.

139. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol [Internet]. 2015 Apr 1 [cited 2022 Oct 20];11(4). Available from: https://pubmed.ncbi.nlm.nih.gov/25885710/

140. Zhang B, Kirov S, Snoddy J. WebGestalt: An integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res [Internet]. 2005 Jul [cited 2020 Nov 20];33(SUPPL. 2). Available from: https://pubmed.ncbi.nlm.nih.gov/15980575/

141. Kumar A, Chauhan G, Sharma S, Dabla S, Sylaja PN, Chaudhary N, et al. Association of SUMOylation Pathway Genes With Stroke in a Genome-Wide Association Study in India. Neurology [Internet]. 2021 Jul 27 [cited 2022 Oct 13];97(4):e345–56. Available from: https://n.neurology.org/content/97/4/e345

142. S Krokstad, A Langhammer, K Hveem, T L Holmen, K Midthjell, T R Stene, G Bratberg, J Heggland JH. Cohort Profile: The HUNT Study [Internet]. International Journal of Epidemiology. 2013 [cited 2020 Jul 10]. p. 968–77. Available from: https://academic.oup.com/ije/article/42/4/968/655743

143. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: An electronic health record-linked biobank for precision medicine research. Genetics in Medicine. 2016 Sep 1;18(9):906–13.

144. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015 Mar 1;12(3).

145. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018 Feb 1;50(2):229–37.

146. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016 Jul 1;48(7):709–17.

147. Williams FMK, Carter AM, Hysi PG, Surdulescu G, Hodgkiss D, Soranzo N, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. Ann Neurol [Internet]. 2013 [cited 2021 Dec 13];73(1):16–31. Available from: https://pubmed.ncbi.nlm.nih.gov/23381943/

148. von Berg J, van der Laan SW, McArdle PF, Malik R, Kittner SJ, Mitchell BD, et al. Alternate approach to stroke phenotyping identifies a genetic risk locus for small vessel stroke. European Journal of Human Genetics. 2020;28(7):963–72.

149. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nature Genetics 2021 53:11. 2021 Oct 28;53(11):1527–33.

150. Bennett DA. How can I deal with missing data in my study? Aust N Z J Public Health. 2001;25(5):464–9.

151. Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, et al. The genetic response to short-term interventions affecting cardiovascular function: Rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J. 2008 May;155(5):823–8.

152. Cheng YC, Kao WHL, Mitchell BD, O'Connell JR, Shen H, McArdle PF, et al. Genome-wide association Scan Identifies variants near matrix metalloproteinase (MMP) genes on chromosome 11q21-22 strongly associated with serum MMP-1 levels. Circ Cardiovasc Genet. 2009;2(4):329–37.

153. Vaara S, Nieminen MS, Lokki ML, Perola M, Pussinen PJ, Allonen J, et al. Cohort profile: The corogene study. Int J Epidemiol. 2012;41(5):1265–71.

154. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, et al. Forty-year trends in cardiovascular risk factors in Finland. Eur J Public Health. 2015;25(3):539–46.

155. Mahdessian H, Perisic Matic L, Lengquist M, Gertow K, Sennblad B, Baldassarre D, et al. Integrative studies implicate matrix metalloproteinase-12 as a culprit gene for large-artery atherosclerotic stroke. J Intern Med. 2017 Nov 1;282(5):429–44.

156. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One [Internet]. 2014 Apr 22 [cited 2021 Mar 26];9(4). Available from: https://pubmed.ncbi.nlm.nih.gov/24755770/

157. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol. 2017;46(6):1985–98.

158. Bowden J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. Int J Epidemiol. 2017;46(6):2097–9.

159. Bowden J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. Int J Epidemiol. 2017;46(6):2097–9.

160. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018 May 1;50(5):693–8.

161. Rasooly D, Patel CJ. Conducting a Reproducible Mendelian Randomization Analysis Using the R Analytic Statistical Environment. Curr Protoc Hum Genet. 2019;(1):1–13.

162. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing Mendelian randomization investigations. Wellcome Open Res. 2020;4.

163. Salminen A, Vlachopoulou E, Havulinna AS, Tervahartiala T, Sattler W, Lokki ML, et al. Genetic Variants Contributing to Circulating Matrix Metalloproteinase 8 Levels and Their Association with Cardiovascular Diseases: A Genome-Wide Analysis. Circ Cardiovasc Genet [Internet]. 2017 Dec 1 [cited 2021 Mar 26];10(6). Available from: https://pubmed.ncbi.nlm.nih.gov/29212897/

164. Holliday EG, Maguire JM, Evans TJ, Koblar SA, Jannes J, Sturm JW, et al. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. Nat Genet [Internet]. 2012 Oct [cited 2022 Oct 16];44(10):1147–51. Available from: https://pubmed.ncbi.nlm.nih.gov/22941190/

165. Jiménez-Sánchez L, Hamilton OKL, Clancy U, Backhouse E v., Stewart CR, Stringer MS, et al. Sex Differences in Cerebral Small Vessel Disease: A Systematic Review and Meta-Analysis. Front Neurol [Internet]. 2021 Oct 28 [cited 2022 Oct 19];12:756887. Available from: /pmc/articles/PMC8581736/

166. Hart RG, Diener HC, Coutts SB, Easton JD, Granger CB, O'Donnell MJ, et al. Embolic strokes of undetermined source: The case for a new clinical construct. Lancet Neurol. 2014;13(4):429–38.

167. Ntaios G. Embolic Stroke of Undetermined Source: JACC Review Topic of the Week. J Am Coll Cardiol. 2020;75(3):333–40.

168. Diener HC, Sacco RL, Easton JD, Granger CB, Bernstein RA, Uchiyama S, et al. Dabigatran for Prevention of Stroke after Embolic Stroke of Undetermined Source. New England Journal of Medicine. 2019;380(20):1906–17.

169. Ntaios G. Embolic Stroke of Undetermined Source: JACC Review Topic of the Week. J Am Coll Cardiol. 2020;75(3):333–40.

170. Flint AC, Banki NM, Ren X, Rao VA, Go AS. Detection of paroxysmal atrial fibrillation by 30-day event monitoring in cryptogenic ischemic stroke: The stroke and monitoring for PAF in real time (SMART) registry. Stroke. 2012;43(10):2788–90.

171. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J. 2016;37(38):2893–962.

172. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol [Internet]. 2015 Jan 5 [cited 2022 Oct 25];16(1):1–14. Available from:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0560-6

173. Choy MK, Javierre BM, Williams SG, Baross SL, Liu Y, Wingett SW, et al. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. Nat Commun. 2018 Dec 1;9(1):1–10.

174. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012 489:7414 [Internet]. 2012 Sep 5 [cited 2022 Oct 25];489(7414):75–82. Available from: https://www.nature.com/articles/nature11232

175. Rempe RG, Hartz AMS, Bauer B. Matrix metalloproteinases in the brain and blood-brain barrier: Versatile breakers and makers. Journal of Cerebral Blood Flow and Metabolism. 2016;36(9):1481–507.

176. Rempe RG, Hartz AMS, Bauer B. Matrix metalloproteinases in the brain and blood-brain barrier: Versatile breakers and makers. Journal of Cerebral Blood Flow and Metabolism. 2016;36(9):1481–507.

177. Chen L, Yang Q, Ding RUI, Liu DAN, Chen Z. Carotid thickness and atherosclerotic plaque stability, serum inflammation, serum mmp-2 and mmp-9 were associated with acute cerebral infarction. Exp Ther Med. 2018;16(6):5253–7.

178. Ma F, Rodriguez S, Buxo X, Morancho A, Riba-Llena I, Carrera A, et al. Plasma Matrix Metalloproteinases in Patients With Stroke During Intensive Rehabilitation Therapy. Arch Phys Med Rehabil. 2016 Nov 1;97(11):1832–40.

179. Salminen A, Vlachopoulou E, Havulinna AS, Tervahartiala T, Sattler W, Lokki ML, et al. Genetic Variants Contributing to Circulating Matrix Metalloproteinase 8 Levels and Their Association with Cardiovascular Diseases: A Genome-Wide Analysis. Circ Cardiovasc Genet. 2017;10(6).

180. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet [Internet]. 2018 Jul 23 [cited 2022 Oct 18];50(8):1112–21. Available from: https://europepmc.org/articles/PMC6393768

181. Zhou LW, Panenka WJ, Al-Momen G, Gicas KM, Thornton AE, Jones AA, et al. Cerebral small vessel disease, risk factors, and cognition in tenants of precarious housing. Stroke [Internet]. 2020 [cited 2022 Oct 18];3271–8. Available from: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.120.030446

182. Simats A, Ramiro L, García-Berrocoso T, Briansó F, Gonzalo R, Martín L, et al. A Mouse Brain-based Multi-omics Integrative Approach Reveals Potential Blood Biomarkers for Ischemic Stroke. Mol Cell Proteomics

[Internet]. 2020 Dec 1 [cited 2022 Oct 18];19(12):1921–35. Available from: https://pubmed.ncbi.nlm.nih.gov/32868372/

183. van der Meer D, Kaufmann T, Shadrin AA, Makowski C, Frei O, Roelfs D, et al. The genetic architecture of human cortical folding. Sci Adv [Internet]. 2021 Dec 1 [cited 2022 Oct 18];7(51). Available from: https://pubmed.ncbi.nlm.nih.gov/34910505/

184. Lambert C, Sam Narean J, Benjamin P, Zeestraten E, Barrick TR, Markus HS. Characterising the grey matter correlates of leukoaraiosis in cerebral small vessel disease. Neuroimage Clin. 2015 Jan 1;9:194–205.

185. Gao Y, Li D, Lin J, Thomas AM, Miao J, Chen D, et al. Cerebral small vessel disease: Pathological mechanisms and potential therapeutic targets. Front Aging Neurosci [Internet]. 2022 Aug 12 [cited 2022 Oct 19];14. Available from: https://pubmed.ncbi.nlm.nih.gov/36034144/

186. Perosa V, Oltmer J, Munting LP, Freeze WM, Auger CA, Scherlek AA, et al. Perivascular space dilation is associated with vascular amyloid-β accumulation in the overlying cortex. Acta Neuropathol [Internet]. 2022 Mar 1 [cited 2022 Oct 19];143(3):331–48. Available from: https://pubmed.ncbi.nlm.nih.gov/34928427/

187. Potter GM, Doubal FN, Jackson CA, Chappell FM, Sudlow CL, Dennis MS, et al. Enlarged perivascular spaces and cerebral small vessel disease. Int J Stroke [Internet]. 2015 Apr 1 [cited 2022 Oct 19];10(3):376–81. Available from: https://pubmed.ncbi.nlm.nih.gov/23692610/

188. Neves KB, Harvey AP, Moreton F, Montezano AC, Rios FJ, Lopes RA, et al. ER stress and Rho kinase activation underlie the vasculopathy of CADASIL. JCI Insight [Internet]. 2019 Dec 5 [cited 2022 Oct 19];4(23). Available from: https://pubmed.ncbi.nlm.nih.gov/31647781/

189. Klein SL, Flanagan KL. Sex differences in immune responses. Nature Reviews Immunology 2016 16:10 [Internet]. 2016 Aug 22 [cited 2022 Oct 19];16(10):626–38. Available from: https://www.nature.com/articles/nri.2016.90

190. Murata T, Lin MI, Huang Y, Yu J, Bauer PM, Giordano FJ, et al. Reexpression of caveolin-1 in endothelium rescues the vascular, cardiac, and pulmonary defects in global caveolin-1 knockout mice. Journal of Experimental Medicine. 2007;204(10):2373–82.

191. Yi SL, Liu XJ, Zhong JQ, Zhang Y. Role of caveolin-1 in atrial fibrillation as an anti-fibrotic signaling molecule in human atrial fibroblasts. PLoS One. 2014;9(1).

192. Laurance S, Aghourian MN, Jiva Lila Z, Lemarié CA, Blostein MD. Gas6-induced tissue factor expression in endothelial cells is mediated through caveolin-1-enriched microdomains. Journal of Thrombosis and Haemostasis. 2014 Mar;12(3):395–408.

193. Wunderlich C, Schober K, Lange SA, Drab M, Braun-Dullaeus RC, Kasper M, et al. Disruption of caveolin-1 leads to enhanced nitrosative stress and severe systolic and diastolic heart failure. Biochem Biophys Res Commun. 2006;340(2):702–8.

194. Ay H, Benner T, Arsava EM, Furie KL, Singhal AB, Matt B, et al. A Computerized Algorithm for Etiologic Classification of. 2007;

195. Aléssio AM, Höehr NF, Siqueira LH, Ozelo MC, de Pádua Mansur A, Annichino-Bizzacchi JM. Association between estrogen receptor alpha and beta gene polymorphisms and deep vein thrombosis. Thromb Res. 2007;120(5):639–45.

196. Díaz Del Moral S, Benaouicha M, Muñoz-Chápuli R, Carmona R. The Insulin-like Growth Factor Signalling Pathway in Cardiac Development and Regeneration. Int J Mol Sci. 2021;23(1).

197. Díaz Del Moral S, Benaouicha M, Muñoz-Chápuli R, Carmona R. The Insulin-like Growth Factor Signalling Pathway in Cardiac Development and Regeneration. Int J Mol Sci. 2021;23(1).

198. Nolte IM, Munoz ML, Tragante V, Amare AT, Jansen R, Vaez A, et al. Genetic loci associated with heart rate variability and their effects on cardiac disease risk. Nat Commun. 2017 Jun;8.

199. van den Berg ME, Warren HR, Cabrera CP, Verweij N, Mifsud B, Haessler J, et al. Discovery of novel heart rate-associated loci using the Exome Chip. Hum Mol Genet. 2017;26(12):2346–63.

200. Common Metabolic Diseases Knowledge Portal. rs3746471. 2021 Jun 17. https://hugeamp.org/variant.html?variant=rs3746471.

201. Common Metabolic Diseases Knowledge Portal. rs3746471. 2021 Jun 17. https://hugeamp.org/variant.html?variant=rs3746471.

202. Common Metabolic Diseases Knowledge Portal. KIAA1755. 2021 Jun 17. https://hugeamp.org/gene.html?gene=KIAA1755.

203. Pirruccello JP, Bick A, Wang M, Chaffin M, Friedman S, Yao J, et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat Commun. 2020 Dec;11(1).

204. Militello G, Hosen MR, Ponomareva Y, Gellert P, Weirick T, John D, et al. A novel long non-coding RNA Myolinc regulates myogenesis through TDP-43 and Filip1. J Mol Cell Biol. 2018;10(2):102–17.

205. Pulit SL, Weng LC, McArdle PF, Trinquart L, Choi SH, Mitchell BD, et al. Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. Neurol Genet. 2018;4(6):1–8.

206. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018 Feb;50(2):229–37.

207. Joutel A, Haddad I, Ratelade J, Nelson MT. Perturbations of the cerebrovascular matrisome: A convergent mechanism in small vessel disease of the brain? Journal of Cerebral Blood Flow and Metabolism. 2016;36(1):143–57.

208. Joutel A, Haddad I, Ratelade J, Nelson MT. Perturbations of the cerebrovascular matrisome: A convergent mechanism in small vessel disease of the brain? Journal of Cerebral Blood Flow and Metabolism. 2016;36(1):143–57.

209. Arboix A, Alioc J. Cardioembolic Stroke: Clinical Features, Specific Cardiac Disorders and Prognosis. Curr Cardiol Rev [Internet]. 2010 Jul 2 [cited 2022 Oct 24];6(3):150. Available from: /pmc/articles/PMC2994107/

10. Annexes

10.1. Funding

This thesis dissertation was possible thanks to an AGAUR grant co-financed by Font Social Europeu (FSE). Grant numbers: 2019FI_B 00853, 2020FI_B1 00157 and 2021FI_B2 00216.

During my PhD project I was awarded with an EMBO Scientific Exchange Grant for a research stay at Università degli Studi di Milano Informatics Department, at Anacleto Lab leaded by Dr Giorgio Valentini.

GENERACION project was possible thanks to the funding of Carlos III Institute: PI15/01978, PI17/02089, PI18-01338, and RETICS RICORS RD21/0006; Fundació Docència i Recerca FMT grant for the Epigenesis project, Marató TV3 support of the Epigenesis study, Eranet-Neuron funding of the Ibiostroke project, AC19/00106, Boehringer Ingelheim funding of the SEDMAN Study.

10.2. Original published “A Polygenic Risk Score Based on a Cardioembolic Stroke Multitrait Analysis Improves a Clinical Prediction Model for This Stroke Subtype”

Publications details:

Journal: FRONTIERS IN CARDIOVASCULAR MEDICINE

Volume: 9

Article Number: 940696

DOI: 10.3389/fcvm.2022.940696

Published: JUL 8 2022

Indexed: 2022-08-11

Document Type: Article

Metrics:

IF-2021 = 5.848

Journal Citation Category: CARDIAC & CARDIOVASCULAR SYSTEMS - 43/143 (Q2)

frontiers | Frontiers in Cardiovascular Medicine

BRIEF RESEARCH REPORT
published: 08 July 2022
doi: 10.3389/fcvm.2022.940696

OPEN ACCESS

Edited by:
Zhihua Wang,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

Reviewed by:
Xiao Chang,
Children's Hospital of Philadelphia, United States
Qingqing Yan,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Georgios Tsivgoulis,
National and Kapodistrian University of Athens, Greece

***Correspondence:**
Israel Férnandez-Cadenas
israelcadenas@yahoo.es

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cardiovascular Genetics and Systems Medicine, a section of the journal Frontiers in Cardiovascular Medicine

Received: *10 May 2022*
Accepted: *06 June 2022*
Published: *08 July 2022*

Citation:
Cárcel-Márquez J, Muiño E, Gallego-Fabrega C, Cullell N, Lledós M, Llucià-Carol L, Sobrino T, Campos F, Castillo J, Freijo M, Arenillas JF, Obach V, Álvarez-Sabín J, Molina CA, Ribó M, Jiménez-Conde J, Roquer J, Muñoz-Narbona L, Lopez-Cancio E, Millán M, Diaz-Navarro R, Vives-Bauza C, Serrano-Heras G, Segura T, Ibañez L, Heitsch L, Delgado P, Dhar R, Krupinski J, Delgado-Mederos R, Prats-Sánchez L, Camps-Renom P, Blay N, Sumoy L, de Cid R, Montaner J, Cruchaga C, Lee J-M, Martí-Fàbregas J and Fernandez-Cadenas I (2022) A Polygenic Risk Score Based on a Cardioembolic Stroke Multitrait Analysis Improves a Clinical Prediction Model for This Stroke Subtype. Front. Cardiovasc. Med. 9:940696. doi: 10.3389/fcvm.2022.940696

A Polygenic Risk Score Based on a Cardioembolic Stroke Multitrait Analysis Improves a Clinical Prediction Model for This Stroke Subtype

***Jara Cárcel-Márquez*[1,2†], *Elena Muiño*[1†], *Cristina Gallego-Fabrega*[1,3], *Natalia Cullell*[1,4], *Miquel Lledós*[1], *Laia Llucià-Carol*[1,5], *Tomás Sobrino*[6], *Francisco Campos*[6], *José Castillo*[6], *Marimar Freijo*[7], *Juan Francisco Arenillas*[8], *Victor Obach*[9], *José Álvarez-Sabín*[10], *Carlos A. Molina*[10], *Marc Ribó*[10], *Jordi Jiménez-Conde*[11], *Jaume Roquer*[11], *Lucia Muñoz-Narbona*[12], *Elena Lopez-Cancio*[13], *Mònica Millán*[12], *Rosa Diaz-Navarro*[14], *Cristòfol Vives-Bauza*[14], *Gemma Serrano-Heras*[15], *Tomás Segura*[15], *Laura Ibañez*[16], *Laura Heitsch*[17,18], *Pilar Delgado*[19], *Rajat Dhar*[18], *Jerzy Krupinski*[20], *Raquel Delgado-Mederos*[3], *Luis Prats-Sánchez*[3], *Pol Camps-Renom*[3], *Natalia Blay*[21], *Lauro Sumoy*[22], *Rafael de Cid*[21], *Joan Montaner*[23], *Carlos Cruchaga*[16,24], *Jin-Moo Lee*[18], *Joan Martí-Fàbregas*[3] *and Israel Férnandez-Cadenas*[1*]
on behalf of GeneStroke Consortium and International Stroke Genetics Consortium

[1] *Stroke Pharmacogenomics and Genetics Group, Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain,* [2] *Departament de Medicina, Universitat Autònoma de Barcelona, Barcelona, Spain,* [3] *Stroke Unit, Department of Neurology, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain,* [4] *Stroke Pharmacogenomics and Genetics Laboratory, Fundación Docència i Recerca Mútua Terrassa, Hospital Mútua Terrassa, Terrassa, Spain,* [5] *Department de Genética i de Microbiologia, Universitat Autónoma de Barcelona, Barcelona, Spain,* [6] *Clinical Neurosciences Research Laboratory, Hospital Clínico Universitario de Santiago de Compostela, Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain,* [7] *Department of Neurology, Biocruces-Bizkaia Health Research Institute, Bilbao, Spain,* [8] *Stroke Unit, Department of Neurology, University Hospital of Valladolid, Valladolid, Spain,* [9] *Department of Neurology, Hospital Clínic de Barcelona, IDIBAPS, Barcelona, Spain,* [10] *Stroke Unit, Department of Neurology, Hospital Universitari Vall d'Hebron, Barcelona, Spain,* [11] *Department of Neurology, IMIM-Hospital del Mar; Neurovascular Research Group, IMIM (Institut Hospital del Mar d'Investigacions Mèdiques), Universitat Autònoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Barcelona, Spain,* [12] *Department of Neurosciences, Hospital Germans Trias i Pujol, Universitat Autònoma de Barcelona, Barcelona, Spain,* [13] *Department of Neurology, University Hospital Central de Asturias (HUCA). Oviedo, Spain,* [14] *Department of Neurology, Son Espases University Hospital, Illes Balears Health Research Institute (IdISBa), Palma, Spain,* [15] *Department of Neurology, University Hospital of Albacete, Albacete, Spain,* [16] *Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO, United States,* [17] *Department of Emergency Medicine, Washington University School of Medicine, Saint Louis, MO, United States,* [18] *Department of Neurology, Washington University School of Medicine, Saint Louis, MO, United States,* [19] *Neurovascular Research Laboratory, Vall d'Hebron Institute of Research, Universitat Autònoma de Barcelona, Barcelona, Spain,* [20] *Neurology Service, Hospital Universitari Mútua Terrassa, Terrassa, Spain,* [21] *GenomesForLife-GCAT Lab, Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain,* [22] *High Content Genomics and Bioinformatics Unit, Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain,* [23] *Institute de Biomedicine of Seville, IBiS/Hospital Universitario Virgen del Rocío/CSIC/University of Seville and Department of Neurology, Hospital Universitario Virgen Macarena, Seville, Spain,* [24] *Neurogenomics and Informatics Center at Washington University in St. Louis, Saint Louis, MO, United States*

Background: Occult atrial fibrillation (AF) is one of the major causes of embolic stroke of undetermined source (ESUS). Knowing the underlying etiology of an ESUS will reduce stroke recurrence and/or unnecessary use of anticoagulants. Understanding cardioembolic strokes (CES), whose main cause is AF, will provide tools to select patients who would benefit from anticoagulants among those with ESUS or AF. We aimed to discover novel loci associated with CES and create a polygenetic risk score (PRS) for a more efficient CES risk stratification.

Methods: Multitrait analysis of GWAS (MTAG) was performed with MEGASTROKE-CES cohort (n = 362,661) and AF cohort (n = 1,030,836). We considered significant variants and replicated those variants with MTAG p-value $< 5 \times 10^{-8}$ influencing both traits (GWAS-pairwise) with a p-value < 0.05 in the original GWAS and in an independent cohort (n = 9,105). The PRS was created with PRSice-2 and evaluated in the independent cohort.

Results: We found and replicated eleven loci associated with CES. Eight were novel loci. Seven of them had been previously associated with AF, namely, *CAV1*, *ESR2*, *GORAB*, *IGF1R*, *NEURL1*, *WIPF1,* and *ZEB2*. *KIAA1755* locus had never been associated with CES/AF, leading its index variant to a missense change (R1045W). The PRS generated has been significantly associated with CES improving discrimination and patient reclassification of a model with age, sex, and hypertension.

Conclusion: The loci found significantly associated with CES in the MTAG, together with the creation of a PRS that improves the predictive clinical models of CES, might help guide future clinical trials of anticoagulant therapy in patients with ESUS or AF.

Keywords: polygenic risk score, GWAS, multi-trait analysis, stroke, ESUs

INTRODUCTION

About 25% of ischemic strokes are of undetermined etiology (1): patients with multiple stroke etiologies, incomplete diagnostic work-up, or embolic stroke of undetermined source (ESUS). Up to 17% of all ischemic strokes are ESUS, with a stroke recurrence rate of 4–5% despite antiplatelet therapy (2).

The ESUS encompasses different entities. Atrial cardiopathy, occult atrial fibrillation (AF), and left ventricular disease might benefit from anticoagulation, but atherosclerotic plaques might benefit from low-dose anticoagulation with antiplatelets in ESUS patients (2). The subgroup of patients >75 years in RE-SPECT ESUS (Dabigatran Etexilate for Secondary Stroke Prevention in Patients With Embolic Stroke of Undetermined Source) had a significant benefit of lower-dose dabigatran over aspirin, suggesting occult AF as a triggering cause (3). Different studies indicate that the prevalence of occult AF among ESUS patients is 11–30% (2, 4).

A tool capable of better stratifying patients is needed to offer them appropriate treatment regarding its potential stroke cause to decrease its recurrence.

On the contrary, not all patients with AF will develop a stroke, and the decision of anticoagulation for stroke prevention in AF patients is carried out based on a clinical scale: CHA_2DS_2-VASc. The rates of stroke vary considerably in patients with CHA_2DS_2-VASc 1–2 (5) and hence the need for a more accurate scale in these cases.

Cardioembolic strokes (CES) are mostly caused by an onset/already known AF. Understanding CES genetic architecture will provide tools to select ESUS or AF patients who would benefit from anticoagulants and develop specific and more effective therapies with fewer side effects.

Therefore, we aimed to discover novel loci associated with CES by performing a Multitrait Analysis of Genome Wide Association Study (MTAG) of CES-AF and create a polygenetic risk score (PRS) that allowed a more efficient stratification of stroke patient risk of having a CES.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Cohorts' Description

The summary statistics for CES were obtained from the MEGASTROKE analysis (MEGASTROKE-CES) through the Cerebrovascular Disease Knowledge Portal (http://cerebrovascularportal.org). This cohort was composed of 7,193 CES patients and 355,468 controls of European ancestry. The summary statistics for AF were obtained from the Atrial Fibrillation 2018 (AF-2018) analysis through the GWAS catalog portal (https://www.ebi.ac.uk/gwas/). The AF-2018 cohort was composed of 60,620 AF cases and 970,216 controls. The characteristics of the individuals in both studies are listed in the **Supplementary Material** and their respective publications (6, 7).

Single-Nucleotide Variation Quality Controls

A series of standard quality controls (QC) was applied to select the single-nucleotide variants (SNVs) for the analysis. Variant exclusion criteria include the following (1): Not common to the summary statistics of the traits (2), Minor allele frequency lower or equal to 0.01 (3), Missing values (4), Negative standard error or not a number value (5), p-value of 0, 6 Not SNVs (7), Duplicated SNVs (8), Strand ambiguity, and (8) Inconsistent allele pairs. Locus 15q21.3, which prioritized genes *GCOM1* and *MYZAP* from AF-2018, was not evaluated due to the absence

of the significant SNVs of AF-2018 in the MEGASTROKE-CES analysis.

Multitrait Analysis of GWAS

We applied MTAG (8) of MEGASTROKE-CES and AF-2018 summary statistics. We considered loci to be significantly associated with the trait of interest when the p-value was $< 5 \times 10^{-8}$ in the MTAG result and the p-value was < 0.05 in the original GWAS. We considered replicating the SNVs with a p-value < 0.05 in the GWAS of our independent cohort.

To avoid an increase in the type I error rate due to the presence of SNVs that are not associated with CES but with AF or vice versa, we used GWAS-pairwise (9). This is a Bayesian pleiotropy association test to identify genetic variants that influence pairs of traits (9). We used it to ensure that the leading SNV of a significant locus belongs to a genomic region influenced by both traits evaluated (9), since SNPs that are not really associated with one trait, but are associated with the other one, could bias effect-size estimates for the first trait and increase false-positive rate (8). The posterior probability for model-3 (PPA-3) >0.6 suggests that a specific genomic region is associated with both traits. A PPA-1 >0.6 will suggest that the genomic region is associated only with CES, and a PPA-2 >0.6 is associated only with AF. Genomic inflation was estimated as lambda.

Identification of Independent and Novel Loci Associated With CES

Independent loci were defined as those >1 megabase (Mb) apart in the physical distance among SNVs with a genome-wide significance threshold of p-value $< 5 \times 10^{-86}$. Loci were defined as novel when SNVs had an $r^2 < 0.1$ compared with the index SNVs of the loci, namely, *PITX2* [7], *ZFHX3* [7], *NKX2-5* [7], *RGS7* [7], *ABO* [7 10], *PHF20* [11], *GNAO1* [11], and 5q22.3 [11], that were GWAS significant in previous studies.

Replication Stage in an Independent European Cohort

We performed GWAS in an independent cohort of 9,105 individuals [GENERACION cohort: 3,479 ischemic stroke (IS) patients and 5,625 controls]. IS patients over 18 years were recruited *via* hospital-based studies, between 2003 and 2020 in Spain, if they had a measurable neurologic deficit on the NIHSS within 6 h of the last known asymptomatic status and had been diagnosed with stroke by an experienced neurologist and confirmed by neuroimaging (10, 11). Controls were subjects over 18 years recruited in Spain, without a history of IS, who declared they were free of neurovascular diseases before enrollment. An Institutional Review Board or Ethics Committee approved the study at each participating site. All patients or their relatives provided written informed consent. Further description of the cohorts is present in **Supplementary Material**, as well as the array information, the contribution of hospitals, and the clinical description (**Supplementary Tables 1–3**).

Quality Control and Imputation

The DNA samples were genotyped on commercial arrays from Illumina® (San Diego, CA) and Axiom™ Spain Biobank array (**Supplementary Table 2**). Standard QCs were performed using the PLINK v1.9 and KING v2.1.3 software. Imputation was performed in the Michigan Imputation Server Pipeline (12) using Minimac4 and HRC r1.1 2016 panel. Further descriptions of QCs and imputation are present in the **Supplementary Material**.

GWAS Analysis

We performed two different GWAS in the same cohort (for the two different traits here studied), with an additive genetic model using fastGWA from GCTA (13). We studied the association with CES (CES = 1,515; controls = 5,626) and AF (AF patients = 1,110; controls = 7,791). Age, sex, and the first 10 principal components were used as covariates.

The results of these two GWAS were used to evaluate replicability. We studied those index variants from significant loci with a p-value $< 5 \times 10^{-8}$ in the MTAG, a p-value < 0.05 in the original GWAS used for performing the MTAG, and PPA-3 > 0.6 that suggests that the genomic region is associated with CES and AF. We considered the replicated SNVs with a p-value < 0.05 and a consistent direction of the effect on this analysis.

Functional Annotation and Gene Prioritization

Gene prioritization was performed for the novel loci using Variant-to-Gene tool from Open targets Genetics Version 7 (14). This tool integrates biological evidence of four main data types, namely, (1) molecular phenotype quantitative trait loci experiments (QTLs), (2) chromatin interaction experiments, e.g., Promoter Capture Hi-C (PCHi-C), 3) *in silico* functional predictions, e.g., Variant Effect Predictor (VEP) from Ensembl and (4) distance between the variant and each gene's canonical transcription start site (TSS). Additionally, we used the HaploReg database to determine the functional annotation of the most strongly associated SNVs per locus. For the missense SNVs, we determined the likelihood that amino acid substitution has a deleterious effect on protein function using SIFT.

Gene Set Analysis

We conducted a WebGestalt Overrepresentation Analysis of the selected prioritized genes associated with MTAG-CES. Gene Ontology (GO) of biological processes was performed, as well as a Benjamini Hochberg correction of the association p-value. We defined a biological process with a p-value < 0.05 as statistically significant.

Polygenic Risk Score Development

A PRS was conducted through the PRSice-2 software version 2.3.3 (15), where the estimation is based on the risk alleles of having a CES and their effect size extracted from the regions with PPA-3 > 0.6 of the MTAG summary statistics created in this study.

GENERACION cohort was randomly split into training and test sets in 80:20 proportion. Best score threshold selection was performed based on the major variance explained by the score (PRS r^2) in the training set. The evaluation of this score was performed in the independent test set.

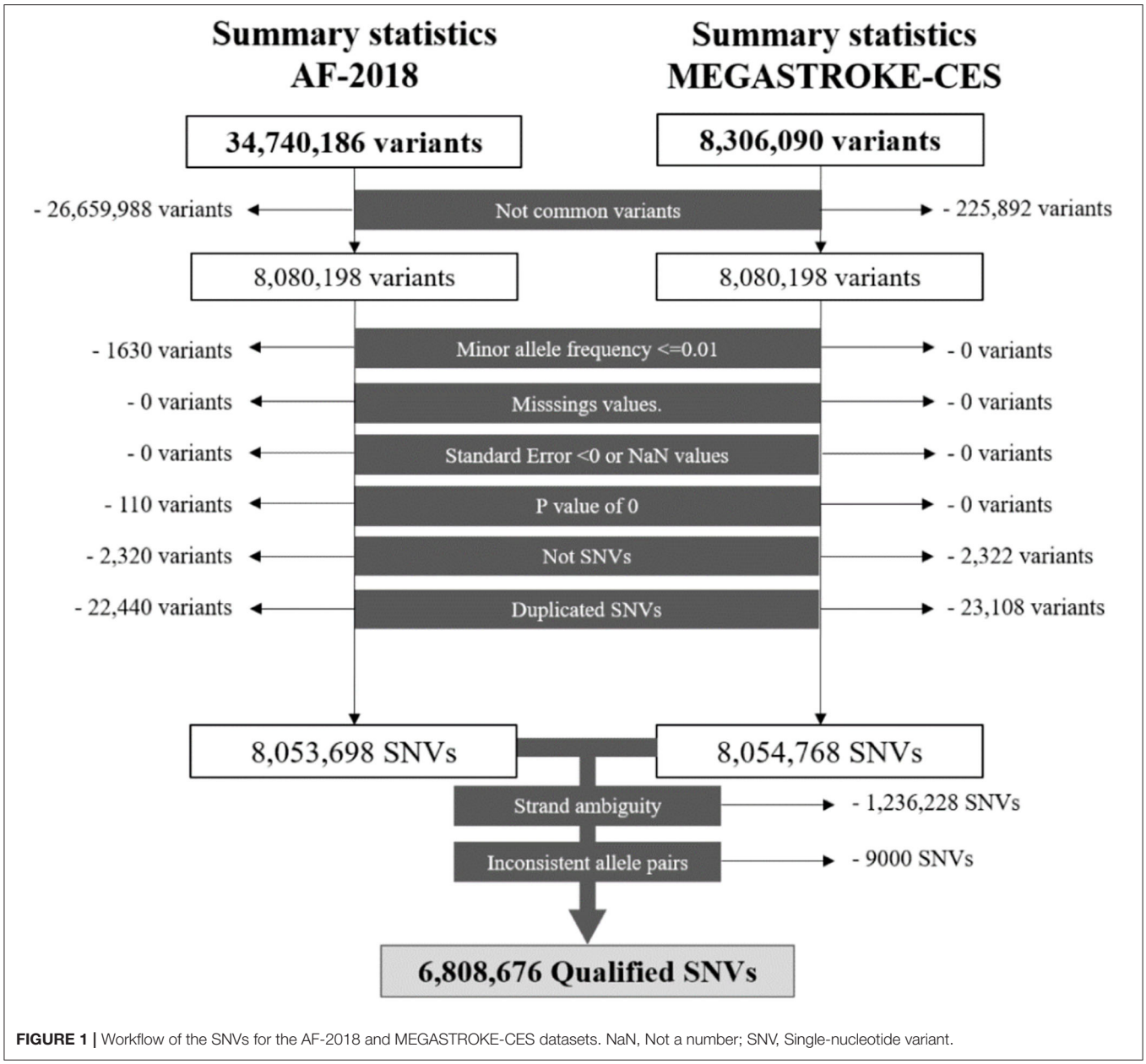


FIGURE 1 | Workflow of the SNVs for the AF-2018 and MEGASTROKE-CES datasets. NaN, Not a number; SNV, Single-nucleotide variant.

We used R version 4.1.3 and Bioconductor packages to evaluate the clinical relevance of this PRS. We calculated three models, namely, model-1 including only the PRS; model-2 including statistically significant clinical variables with <10% missing values, since a high rate of missing values might bias the results of subsequent statistical analyses (16); and model-3 adding the PRS to the model-2. Model discrimination was assessed with the area under the ROC curve (AUC) and the area under the precision recall curve (AUPRC). We used DeLong's test for two correlated ROC curves to find out whether there are significant differences between the discrimination of the models. The net reclassification index (NRI) and integrated discrimination index (IDI) were performed to evaluate model-2 and model-3. Additionally, we estimated the AUC and AUPRC for each individual predictor.

RESULTS

MTAG Analysis of CES

After QCs (**Figure 1**), there were 6,808,676 common qualified SNVs from the AF-2018 and MEGASTROKE-CES cohorts. The MTAG software revealed mean χ^2 for AF-2018 and MEGASTROKE-CES in 1.39 and 1.12, respectively. The estimated equivalent GWAS sample size of the MTAG analysis for CES was 861,823 individuals. A Manhattan plot of the MTAG-CES analysis is shown in **Figure 2**; less evidence of genomic inflation was observed with a lambda of 1.02.

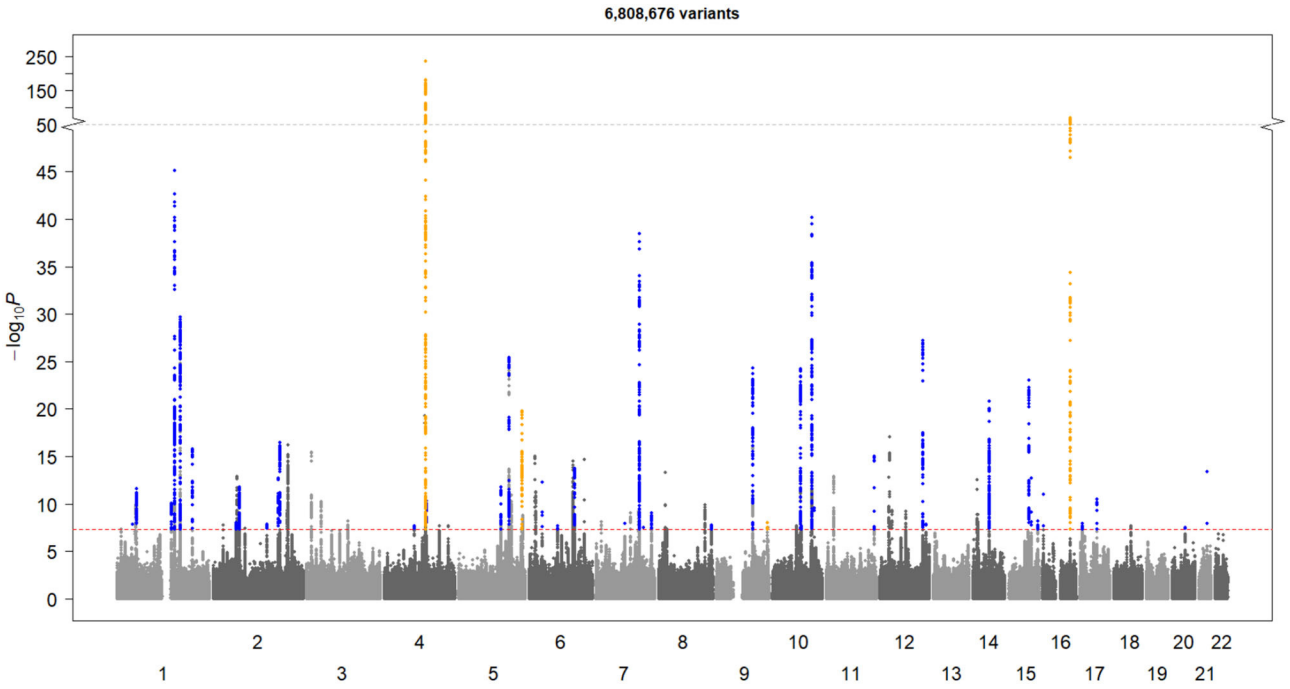


FIGURE 2 | Manhattan plot of MTAG-CES. The X-axis represents chromosome location, and the Y-axis represents the minus logarithm on base 10 of *p*-value. The red line represents the GWAS-significance threshold. The novel loci are shown in blue and the established loci are shown in yellow.

The MTAG-CES results revealed a total of 44 associated loci (p-value $< 5 \times 10^{-8}$); 40 significant loci associated with CES were novel, and four were previously found known associations (**Table 1**). All loci significantly associated with MEGASTROKE-CES (*ABO*, *NKX2-5*, *PITX2,* and *ZFHX3*) were genome-wide associated in this MTAG-CES, except for the locus belonging to *RGS7* gene (top SNV rs146390073 MTAG p-value = 0.001, AF-2018 p-value = 0.98).

Other loci found significant in previous GWAS of CES different from MEGASTROKE were: *PHF20*, *GNAO1,* and 5q22.3 region. For two of them, the association was more significant in our analysis. For the 5q22.3 region, top SNV rs2169955 MTAG-CES p-value $= 4.76 \times 10^{-7}$ vs. MEGASTROKE-CES p-value $= 6.13 \times 10^{-3}$, and for mapped gene *PHF20,* top SNV rs11697087 MTAG-CES p-value $= 6.62 \times 10^{-5}$ vs. MEGASTROKE-CES p-value $= 6.05 \times 10^{-4}$. *GNAO1* was not evaluated in our study due to the absence of the index SNV in AF-2018.

New Candidate Loci Associated With CES

After gene prioritization, 44 genes were selected from the 44 loci (**Supplementary Table 4**). Novel loci showed a high degree of functionality of the SNVs as missense variants, eQTL, pQTLs, and HiC physical interaction (**Supplementary Table 4**).

Replication analysis was performed in a new cohort of IS patients and controls (GENERACION cohort, $n = 9{,}105$). Evaluation of the index SNVs and SNVs in high LD belonging to genome-wide significant loci from the MTAG-CES revealed that 11 loci were replicated, as SNVs had a p-value < 0.05 in MEGASTROKE-CES, a PPA-3 > 0.6, and a p-value < 0.05 in the replication cohort (GENERACION). Among these 11 different loci, *PITX2*, *ZFHX3*, and *NKX2-5* were already known. Eight loci were novel associations whose prioritized genes were *CAV1*, *IGF1R*, *KIAA1755*, *NEURL1*, *GORAB*, *ESR2*, *ZEB2,* and *WIPF1* (**Supplementary Table 5**).

Interestingly, we found loci not previously described in AF or CES with four prioritized genes, namely, *TMEM60*, *KIAA1755*, *NCOR2,* and *FILIP1*. Functional annotation of the index SNVs revealed rs3746471 as a missense variant of the *KIAA1755* gene coding for R1045W, and it was predicted to be deleterious with a SIFT score of 0.007.

New Candidate Locus Associated With AF

FILIP1 locus reached genome-wide significance in the MTAG-AF, a p-value < 0.05 in AF-2018 and a PPA-3 > 0.6. The FILIP1 index variant was additionally evaluated in the GWAS of AF in the independent cohort (GENERACION), revealing a suggestive p-value with a consistent direction of the effect of this novel association with AF rs12211255-A, beta(se) = 0.013(0.007), p-value = 0.09.

The study of the 111 AF-2018 significant loci (**Supplementary Table 6**) using GWAS-pairwise strategy suggested 51 loci that have an exclusive association with AF risk and a lack of association with CES.

TABLE 1 | MTAG-CES results of the independent and significant loci.

				MEGASTROKE CES		AF-2018		MTAG-CES		MTAG-AF		
SNV	**Locus**	**Gene**	**Novelty**	**Z**	**P**	**Z**	**P**	**Z**	**P**	**Z**	**P**	**PPA**
rs17042098-A	4q25	*PITX2*	(Malik et al.) [7]	12.72	3.72×10^{-37}	38.01	8.97×10^{-318}	32.86	7.73×10^{-237}	37.93	0	1.00
rs2106261-T	16q22.3	*ZFHX3*	(Malik et al.) [7]	6.75	1.63×10^{-11}	20.23	4.97×10^{-91}	17.48	2.02×10^{-68}	20.19	1.20×10^{-90}	1.00
rs11264280-T	1q21.3	*ADAM15*	Novel	2.49	1.28×10^{-02}	18.97	3.07×10^{-79}	14.22	7.17×10^{-46}	18.44	5.88×10^{-76}	0.65
rs11598047-A	10q24.33	*NEURL1*	Novel	−3.27	1.06×10^{-03}	−17.08	8.95×10^{-66}	−13.38	7.50×10^{-41}	−16.73	7.42×10^{-63}	0.93
rs3807989-A	7q31.2	*CAV1*	Novel	−4.50	6.75×10^{-06}	−15.63	1.24×10^{-54}	−13.10	3.43×10^{-39}	−15.50	3.30×10^{-54}	1.00
rs680084-A	1q24.2	*GORAB*	Novel	−4.10	4.02×10^{-05}	−13.54	3.31×10^{-42}	−11.46	2.10×10^{-30}	−13.46	2.84×10^{-41}	0.99
rs883079-T	12q24.21	*TBX5*	Novel	3.55	3.95×10^{-04}	13.26	2.84×10^{-40}	10.96	5.97×10^{-28}	13.12	2.60×10^{-39}	0.97
rs17171711-T	5q31.2	*FAM13B*	Novel	3.79	1.56×10^{-04}	12.48	1.95×10^{-35}	10.57	4.04×10^{-26}	12.41	2.35×10^{-35}	0.99
rs4385527-A	9q22.32	*AOPEP*	Novel	3.92	8.56×10^{-05}	12.03	6.16×10^{-33}	10.34	4.69×10^{-25}	11.99	3.89×10^{-33}	0.95
rs78249997-T	10q22.2	*MYOZ1*	Novel	−3.82	1.37×10^{-04}	−12.08	8.75×10^{-34}	−10.32	5.80×10^{-25}	−12.03	2.45×10^{-33}	0.98
rs7172038-T	15q24.1	*NEO1*	Novel	−2.73	6.43×10^{-03}	−12.58	4.78×10^{-36}	−10.04	1.02×10^{-23}	−12.37	3.76×10^{-35}	0.74
rs2738413-A	14q23.2	*ESR2*	Novel	2.97	3.03×10^{-03}	11.61	2.55×10^{-31}	9.52	1.67×10^{-21}	11.47	1.80×10^{-30}	0.85
rs6891790-T	5q35.1	*NKX2-5*	(Malik et al.) [7]	−4.97	6.67×10^{-07}	−9.59	4.53×10^{-22}	−9.29	1.57×10^{-20}	−9.80	1.16×10^{-22}	1.00
rs2857265-A	2q31.2	*FKBP7*	Novel	2.76	5.85×10^{-03}	10.16	4.58×10^{-24}	8.42	3.65×10^{-17}	10.06	8.29×10^{-24}	0.70
rs10753933-T	1q32.1	*PPFIA4*	Novel	3.72	2.09×10^{-04}	9.09	9.84×10^{-20}	8.24	1.70×10^{-16}	9.16	5.30×10^{-20}	0.99
rs74399915-T	11q24.3	*C11orf45*	Novel	3.38	7.20×10^{-04}	9.05	1.23×10^{-19}	8.03	1.01×10^{-15}	9.08	1.11×10^{-19}	0.92
rs13191450-A	6q22.31	*HSF2*	Novel	2.93	3.51×10^{-03}	8.88	9.97×10^{-19}	7.65	1.95×10^{-14}	8.86	7.96×10^{-19}	0.81
rs2834618-T	21q22.12	*RUNX1*	Novel	3.31	9.48×10^{-04}	8.43	3.41×10^{-17}	7.56	3.96×10^{-14}	8.47	2.37×10^{-17}	0.96
rs56181519-T	2q31.1	*WIPF1*	Novel	−2.73	6.31×10^{-03}	−8.60	6.46×10^{-18}	−7.35	1.98×10^{-13}	−8.56	1.11×10^{-17}	0.82
rs12908004-A	15q25.1	*ARNT2*	Novel	−3.28	1.03×10^{-03}	−8.13	4.12×10^{-16}	−7.35	2.03×10^{-13}	−8.19	2.65×10^{-16}	0.96

(Continued)

TABLE 1 | Continued

SNV	Locus	Gene	Novelty	MEGASTROKE CES		AF-2018		MTAG-CES		MTAG-AF		PPA
				Z	P	Z	P	Z	P	Z	P	
rs3176326-A	6p21.2	*CDKN1A*	Novel	−3.98	6.66×10^{-05}	−7.36	1.42×10^{-13}	−7.23	5.01×10^{-13}	−7.54	4.59×10^{-14}	1.00
rs6747542-T	2p13.3	*GMCL1*	Novel	2.61	8.86×10^{-03}	8.27	1.10×10^{-16}	7.06	1.63×10^{-12}	8.23	1.82×10^{-16}	0.75
rs337705-T	5q22.3	*KCNN2*	Novel	−2.56	1.04×10^{-02}	−8.29	1.63×10^{-16}	−7.05	1.77×10^{-12}	−8.25	1.57×10^{-16}	0.72
rs41292535-A	1p32.2	*EPS15*	Novel	3.81	1.39×10^{-04}	7.16	7.42×10^{-13}	6.99	2.73×10^{-12}	7.33	2.34×10^{-13}	0.96
rs140185678-A	16p13.3	*RPL3L*	Novel	2.92	3.54×10^{-03}	7.61	2.43×10^{-14}	6.79	1.13×10^{-11}	7.64	2.13×10^{-14}	0.89
rs76774446-A	17q21.32	*GOSR2*	Novel	4.38	1.20×10^{-05}	6.15	8.72×10^{-10}	6.63	3.32×10^{-11}	6.43	1.25×10^{-10}	0.99
rs55754224-T	4q26	*CAMK2D*	Novel	2.80	5.05×10^{-03}	7.39	2.15×10^{-13}	6.57	4.92×10^{-11}	7.41	1.22×10^{-13}	0.80
rs79187193-A	1q21.2	*GJA5*	Novel	−2.37	1.78×10^{-02}	−7.59	3.15×10^{-14}	−6.47	9.79×10^{-11}	−7.56	4.08×10^{-14}	0.71
rs12260801-T	10q25.2	*PDCD4*	Novel	4.56	5.22×10^{-06}	5.53	2.94×10^{-08}	6.31	2.79×10^{-10}	5.86	4.62×10^{-09}	1.00
rs2269001-A	7q36.1	*KCNH2*	Novel	−2.89	3.94×10^{-03}	−6.63	4.01×10^{-11}	−6.11	1.00×10^{-09}	−6.70	2.06×10^{-11}	0.73
rs6598541-A	15q26.3	*IGF1R*	Novel	2.69	7.17×10^{-03}	6.35	2.22×10^{-10}	5.81	6.21×10^{-09}	6.41	1.48×10^{-10}	0.72
rs11125871-T	2p15	*C2orf74*	Novel	−3.24	1.17×10^{-03}	−5.79	6.42×10^{-09}	−5.75	9.19×10^{-09}	−5.95	2.71×10^{-09}	0.87
rs635634-T	9q34.2	*ABO*	(Williams et al.) [10]; Malik et al. [7]	5.10	3.31×10^{-07}	4.20	2.74×10^{-05}	5.72	1.04×10^{-08}	4.66	3.13×10^{-06}	1.00
rs10272350-A	7q11.23	*TMEM60*	Novel	4.53	6.13×10^{-06}	4.62	3.41×10^{-06}	5.68	1.32×10^{-08}	4.99	5.93×10^{-07}	0.99
rs116600817-A	17p13.1	*TNFSF12*	Novel	2.88	3.92×10^{-03}	6.00	2.45×10^{-09}	5.68	1.33×10^{-08}	6.10	1.07×10^{-09}	0.85
rs13010313-T	2q22.3	*ZEB2*	Novel	3.52	4.21×10^{-04}	5.45	5.72×10^{-08}	5.67	1.41×10^{-08}	5.65	1.57×10^{-08}	0.90
rs2885697-T	1p34.2	*SCMH1*	Novel	−2.54	1.09×10^{-02}	−6.27	2.88×10^{-10}	−5.67	1.41×10^{-08}	−6.32	2.70×10^{-10}	0.72
rs11057583-A	12q24.31	*NCOR2*	Novel	3.82	1.35×10^{-04}	5.19	2.10×10^{-07}	5.67	1.46×10^{-08}	5.45	5.13×10^{-08}	0.74

(Continued)

TABLE 1 | Continued

SNV	Locus	Gene	Novelty	MEGASTROKE CES Z	MEGASTROKE CES P	AF-2018 Z	AF-2018 P	MTAG-CES Z	MTAG-CES P	MTAG-AF Z	MTAG-AF P	PPA
rs11782313-T	8q24.3	*PTK2*	Novel	−3.05	2.34×10^{-03}	−5.81	5.71×10^{-09}	−5.64	1.65×10^{-08}	−5.94	2.94×10^{-09}	0.78
rs12211255-A	6q14.1	*FILIP1*	Novel	3.42	6.24×10^{-04}	5.44	5.06×10^{-08}	5.61	2.06×10^{-08}	5.63	1.78×10^{-08}	0.96
rs1898096-A	10q22.3	*LRMDA*	Novel	−2.64	8.30×10^{-03}	−6.08	1.39×10^{-09}	−5.60	2.13×10^{-08}	−6.15	7.96×10^{-10}	1.00
rs11099098-T	4q21.21	*FGF5*	Novel	2.77	5.67×10^{-03}	5.94	2.96×10^{-09}	5.58	2.38×10^{-08}	6.03	1.62×10^{-09}	0.78
rs55985730-T	7q32.1	*CALU*	Novel	−2.84	4.47×10^{-03}	−5.82	5.24×10^{-09}	−5.54	3.06×10^{-08}	−5.92	3.19×10^{-09}	0.87
rs3746471-A	20q11.23	*KIAA1755*	Novel	3.56	3.58×10^{-04}	5.18	2.30×10^{-07}	5.51	3.50×10^{-08}	5.40	6.59×10^{-08}	0.96

SNV, Index Single Nucleotide Variant and assessed allele; Z, Z-score; P, P-value. Loci highlighted in bold are the ones not previously associated with AF. PPA, posterior probability of model 3 of GWAS-PW.

Biological Processes of Loci Associated With CES and AF and Biological Processes of Loci Associated Exclusively With AF

The GO of biological processes from the Genome-Wide loci of the MTAG-CES analysis revealed 98 enriched gene sets (**Supplementary Figure 1**, **Supplementary Table 7**); the top biological processes were cardiac conduction, cardiac muscle cell contraction, and cardiac muscle contraction.

A biological process analysis of the genes associated exclusively with AF (**Supplementary Figure 2**, **Supplementary Table 8**) revealed 41 biological processes exclusive to AF risk and mainly associated with cardiac development processes (**Supplementary Figure 1**, **Supplementary Table 9**).

Polygenic Risk Score

The training set was composed of 1,212 CES patients and 4,501 controls and the test set of 303 CES patients and 1,125 controls from GENERACION. No significant differences in clinical variables were found between the training and the test sets (**Supplementary Table 10**).

For model-1, the PRS with the highest r^2 in the training set ($r^2 = 0.018$) was obtained with an SNV p-value threshold of 5×10^{-8}, comprising a total of 93 SNVs (**Figure 3**). Age, sex, and hypertension were the only variables for which information was available for >90% of the patients, and therefore, the only ones considered in the multivariable model as mentioned in the Methods" section. The three variables were significantly associated and therefore included in model-2. For model-3, we added the PRS to model-2, and all remaining variables were significant (**Supplementary Table 11**), including the PRS with a Z-value of 4.33 and a p-value of 1.28×10^{-5}.

The AUC in the test set for the different models was 0.581 in model-1, 0.947 in model-2, and 0.950 in model-3. AUPRC was 0.271 in model-1, 0.877 in model-2, and 0.883 in model-3 (**Figure 3**). Comparing AUC, there was significantly better discrimination in model-3 than model-2 (Z-score $= -2.50$, p-value $= 0.01$). AUC and AUPRC for each individual predictor can be found in **Supplementary Figure 3**.

Additionally, the NRI categorical and quantitative and IDI showed a significant reclassification when quartiles of score risk were analyzed (**Table 2**).

DISCUSSION

Using an MTAG with the two biggest cohorts of CES (7) and AF (6) to date, we found 44 genome-wide significant loci associated with CES. The prioritized genes of this loci were involved in biological processes such as cardiac conduction and contraction. Nevertheless, the 51 loci associated exclusively with AF (not associated with CES as shown in the GWAS-pairwise) were mainly associated with cardiac development processes. This highlights the possible role in the risk of stroke due to AF of genes related to cardiac conduction and contraction instead of the cardiac development process and thereby would help to develop more specific prevention drugs.

TABLE 2 | Reclassification table comparing CES models with and without PRS addition.

	Risk Category Age+Sex+HT+PRS model				
Risk Category Age+Sex+HT model	**No.**	**No.**	**No.**	**No.**	**% Reclassified**
Non cases					
Q1	4,062	23	0	0	1
Q2	25	174	16	0	19
Q3	0	23	120	4	18
Q4	0	0	3	50	0
Cases					
Q1	149	7	0	0	4
Q2	7	124	10	0	12
Q3	0	12	171	34	21
Q4	0	0	18	677	3
NRI (Categorical) [95% CI]: 0.0134 [−0.0024–0.0291]; *p*-value: 0.09688					
NRI (Continuos) [95% CI]: 0.1416 [0.0782–0.2049]; *p*-value: 1e-5					
IDI [95% CI]: 0.0029 [9e-04−0.0049]; *p*-value: 0.00431					

Q, quartile; NRI, Net Reclassification Index; IDI, integrated discrimination improvement.

Eleven loci significantly associated with CES were replicated in the independent cohort. Their prioritized genes are listed as follows: *PITX2*, *ZFHX3*, *NKX2-5*, *CAV1*, *IGF1R*, *KIAA1755*, *NEURL1*, *GORAB*, *ESR2*, *ZEB2,* and *WIPF1*. Of the genes associated with these loci, *PITX2*, *ZFHX3,* and *NKX2-5* were already known to be associated with CES and AF. Eight were new CES associations; seven of them were previously associated with AF, namely, *CAV1*, *ESR2*, *GORAB*, *IGF1R*, *NEURL1*, *WIPF1,* and *ZEB2*; and KIAA1755 was a completely new association with CES, not being previously associated with AF.

One could think that by increasing the statistical power to find CES-associated SNVs through enrichment of AF patients, part of the associations is due to actually being associated only with AF. For this reason, we ensure that SNVs belonged to genomic regions associated with AF and CES through GWAS-pairwise (PPA-3 $>$ 0.6). Therefore, these 11 SNVs could be markers of stroke risk among patients with ESUS or among AF patients, as they are SNVs located in genomic regions that are not exclusively associated with either CES or FA, but with both.

Of the new loci associations with CES, we could highlight some genes. *CAV1* encodes caveolin-1, the principal structural component of caveolae organelles in smooth muscle cells and endothelial cells (17). Caveolin-1 confers an anti-AF effect by mediating atrial structural remodeling through its antifibrotic action (18). Also, it plays a key role in how gas6 exerts its prothrombotic role in the vasculature (19). Genetic disruption of caveolin-1 in mice induces a severe biventricular hypertrophy with systolic and diastolic heart failure (20). That supports the relevance that caveolin-1 might have in other causes of CES as symptomatic congestive heart failure with reduced ejection fraction (21), or its importance in ESUS as a marker of an occult FA or left ventricular disfunction, which could benefit from anticoagulant treatment.

ESR2 encodes for the estrogen receptor beta, one of the receptors that mediates the biological effects of estrogens, which increase the levels of procoagulant factors VII, IX, X, XII, and XIII and reduce the concentrations of the anticoagulant factors protein S and antithrombin (22). Therefore, it might be a stroke risk marker.

IGF1R encodes the insulin-like growth factor (IGF) 1 receptor, that is, the main receptor mediating IGF signaling in the heart (23). Inhibition of the IGF receptor decreases the proliferation of cardiomyocytes in murine embryonic stem cells (23). *ZEB2* encodes the zinc finger E-box-binding homeobox 2 protein that regulates cardiac fibroblast activation. An aberrant activation could lead to structural changes prone to develop AF.

KIAA1755 has not previously been found associated with AF. The index variant of this locus, rs3746471-A, encodes for R1045W amino acid change, predicted to be deleterious according to SIFT. rs3746471-A has been previously described as associated with heart rate (24–26) and PR interval (26) and is remarkably suggestively associated with stroke infarct volume (p-value $= 6.80 \times 10^{-7}$) (27, 28). *KIAA1755* is predicted to encode an uncharacterized protein and is only characterized at the transcriptional level. The transcript is highly expressed in the brain and nerves and is also expressed in the heart.

We also found a novel locus suggestive to be associated with AF: 6q14.1, being *FILIP1* the prioritized gene linked with the leading SNV of the locus. This gene encodes a filamin A binding protein and has been identified as a regulator of myogenesis differentiation in human cells and in an *in vivo* mouse model (29). In the replication stage, this SNV was found suggestive (p-value $= 0.09$), highly probable due to the small sample size in comparison with MTAG analysis.

The PRS generated with the SNVs from MTAG-CES was associated with CES independently of age, sex, and hypertension, being simpler than other PRS that needs a major number of SNVs for association (30). We found that the addition of our PRS to a model with age, sex, and hypertension significantly improves the discriminatory power to detect CES.

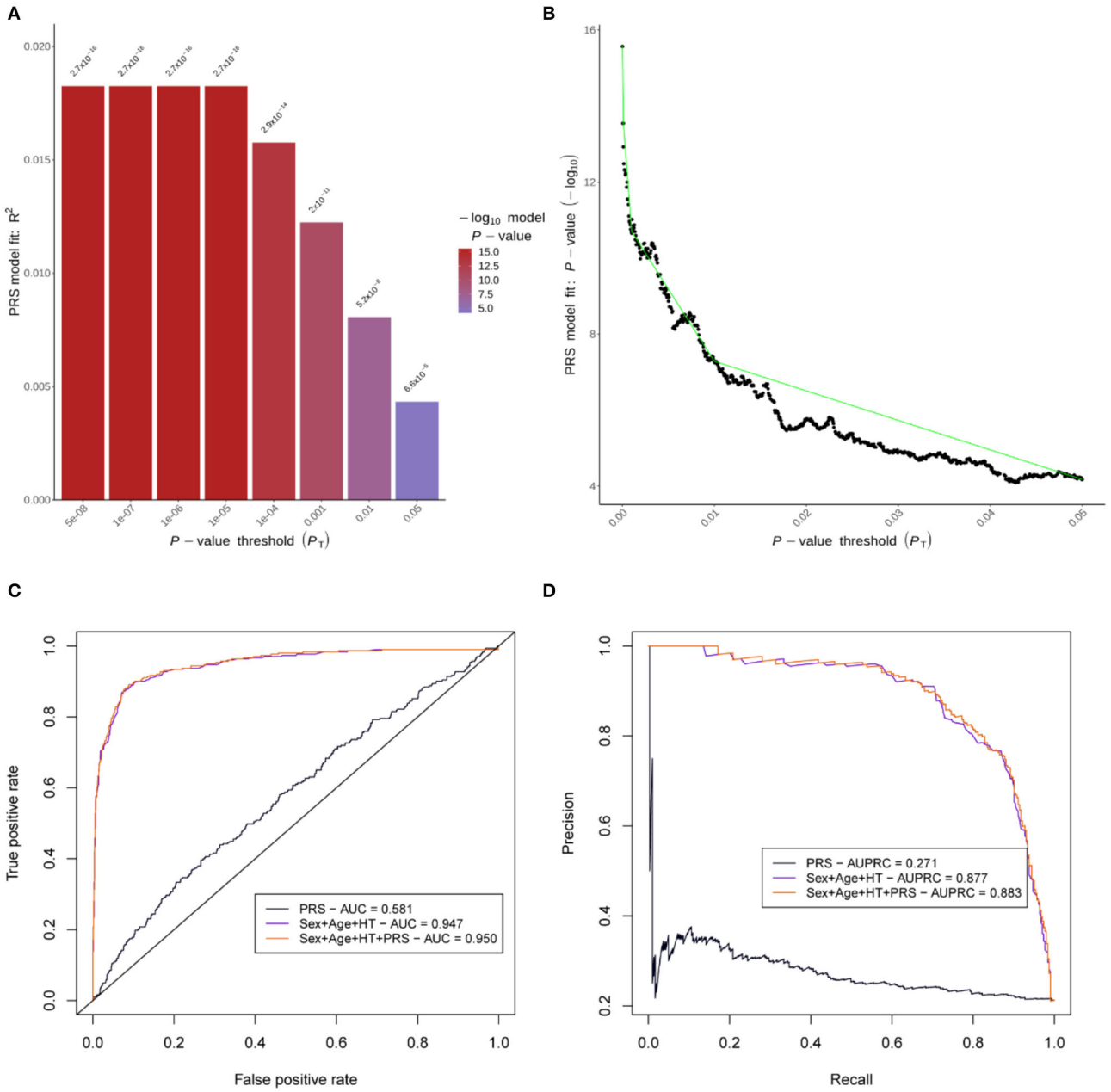


FIGURE 3 | Polygenic risk score (PRS) performance. **(A)** is a bar plot of the r^2 for the PRS models of eight different thresholds in the training set. **(B)** represents the *p*-value variation along the full range of thresholds evaluated in the training set. **(C)** shows ROC curves, and **(D)** shows precision-recall curves for the PRS performance in the independent test set. AUC, area under the ROC curve; AUPRC, area under the precision-recall curve; HT, hypertension.

Interestingly, the quantitative NRI was estimated in 14.16%, which is the proportion of cases correctly assigned to a higher probability of CES, among controls correctly assigned to a lower probability by an updated model adding our PRS compared with the initial model without it.

As limitations, the difference in the sample size between the two original studies could lead to significant results for SNVs that are truly null for one trait but not for another, biasing effect-size estimates for the first trait and increasing the false discovery rate (and inflated type I error rate) (8). Nevertheless, MTAG estimation of χ^2 revealed a scenario expected to be strong against false positives, as tested in the original publication (8), and less evidence on genomic inflation was observed. Besides, we used GWAS-pairwise to ensure that the novel loci were not associated with only one of the traits, but with both at the same time, having a PPA-3 > 0.6. But even more important, as usually

in this kind of studies, we validated the significant loci found in this MTAG-CES and MTAG-AF in a GWAS of an independent European cohort. The small size of this last cohort underpowers the ability to find significant results. However, we were able to replicate 11 leading SNVs from the total number of significant loci in the MTAG-CES and suggest one new potential locus in the MTAG-AF.

Another limitation is that we have only found loci associated with CES risk due to AF. Therefore, further multitrait analysis should be performed with different traits to uncover the different high-risk sources of CES. Nevertheless, our aim was to better characterize patients with CES due to AF as it is the most frequent cause of this type of stroke, for subsequently being able to find tools to detect those patients with a higher risk of developing a stroke due to an occult AF among ESUS for guiding future clinical trials with anticoagulant therapy.

In conclusion, we found and replicate 11 loci associated with CES, with eight of them having new associations. We showed that their leading SNVs are in genomic regions related to both, CES and AF, suggesting that they, together with the creation of a PRS that improves the predictive models of CES, might allow to better stratify the risk of stroke and its possible etiology to guide future clinical trials of anticoagulant therapy in AF or ESUS patients for a personalized medicine.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comité ético de la Fundación Docència I Recerca Mútua Terrassa. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Conception and design of the work: JC-M, EM, and IF-C. Data acquisition: TS, FC, JC, JA, VO, LL-C, MF, JÁ-S, CM, MR, JJ-C, JR, LM-N, EL-C, RD-N, CV-B, GS-H, TS, LI, LH, PD, JK, RD, RD-M, LP-S, PC-R, NB, LS, RdC, JM, CC, J-ML, JM-F, and IF-C. Formal analysis and methodology: JC-M, EM, CG-F, NC, ML, and IF-C. Interpretation and supervision and writing—original draft: JC-M, EM, CG-F, NC, ML, JM, CC, J-ML, JM-F, and IF-C. All authors have been involved in drafting the article or revising it critically for intellectual content, writing—review and editing, and approved the submitted version.

FUNDING

J. Cárcel-Márquez has received funding through an AGAUR Contract (Agència de Gestió d'Ajuts Universitaris i de Recerca; FI_DGR 2019, grant number 2020FI_B1 00157) co-financed with Fons Social Europeu (FSE) (https://agaur.gencat.cat). From Instituto de Salud Carlos III: E. Muiño is funded by a Río Hortega Contract (CM18/00198), M. Lledós is funded by a PFIS Contract (Contratos Predoctorales de Formación en Investigación en Salud, FI19/00309), C. Gallego-Fabrega is supported by a Sara Borrell Contract (CD20/00043) and Fondo Europeo de Desarrollo Regional (ISCIII-FEDER), T. Sobrino (CPII17/00027), and F. Campos (CPII19/00020) are recipients of research contracts from the Miguel Servet Program (https://www.isciii.es). This study has been funded by Instituto de Salud Carlos III PI15/01978, PI17/02089, PI18-01338, and RICORS-ICTUS RD21/0006/0006 (Instituto de Salud Carlos III), by Marató TV3 support of the Epigenesis study (https://www.ccma.cat/tv3/marato/), by the Fundació Docència i Recerca FMT grant for the Epigenesis project (https://www.mutuaterrassa.com), by Eranet-Neuron of the Ibiostroke project (AC19/00106) (https://www.neuron-eranet.eu), by Boehringer Ingelheim of the SEDMAN Study (https://www.boehringer-ingelheim.it), and GCAT Cession Research Project PI-2018-01 (http://www.gcatbiobank.org). GCAT was funded by Acción de Dinamización del ISCIII-MINECO and the Ministry of Health of the Generalitat of Catalunya (ADE 10/00026); and have additional suport by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (2017-SGR 529).

ACKNOWLEDGMENTS

The Genotype-Tissue Expression (GTEx) Project was funded by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 03/30/20. This study uses data generated by the GCAT, Genomes for Life. Cohort study of the Genomes of Catalonia, IGTP. A full list of the investigators who contributed to the generation of the data is available from http://www.genomesforlife.com/. IGTP is part of the CERCA Program/Generalitat de Catalunya. This study was carried out using anonymized data provided by the Catalan Agency for Quality and Health Assessment, within the framework of the PADRIS Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.940696/full#supplementary-material

REFERENCES

1. Hart RG, Diener HC, Coutts SB, Easton JD, Granger CB, O'Donnell MJ, et al. Embolic strokes of undetermined source: the case for a new clinical construct. *Lancet Neurol.* (2014) 13:429–38. doi: 10.1016/S1474-4422(13)70310-7
2. Ntaios G. Embolic stroke of undetermined source: JACC review topic of the week. *J Am Coll Cardiol.* (2020) 75:333–40. doi: 10.1016/j.jacc.2019.11.024
3. Diener H-C, Sacco RL, Easton JD, Granger CB, Bernstein RA, Uchiyama S, et al. Dabigatran for prevention of stroke after embolic stroke of undetermined source. *N Engl J Med.* (2019) 380:1906–17. doi: 10.1056/NEJMoa1813959
4. Flint AC, Banki NM, Ren X, Rao VA, Go AS. Detection of paroxysmal atrial fibrillation by 30-day event monitoring in cryptogenic ischemic stroke: the stroke and monitoring for PAF in real time (SMART) registry. *Stroke.* (2012) 43:2788–90. doi: 10.1161/STROKEAHA.112.665844
5. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J.* (2016) 37:2893–962. doi: 10.1093/eurheartj/ehw210
6. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet.* (2018) 50:1234–9. doi: 10.1038/s41588-018-0171-3
7. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet.* (2018) 50:524–37. doi: 10.1038/s41588–018–0058–3
8. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet.* (2018) 50:229–37. doi: 10.1038/s41588-017-0009-4
9. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet.* (2016) 48:709–17. doi: 10.1038/ng.3570
10. Williams FMK, Carter AM, Hysi PG, Surdulescu G, Hodgkiss D, Soranzo N, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. *Ann Neurol.* (2013) 73:16–31. doi: 10.1002/ana.23838
11. von Berg J, van der Laan SW, McArdle PF, Malik R, Kittner SJ, Mitchell BD, et al. Alternate approach to stroke phenotyping identifies a genetic risk locus for small vessel stroke. *Eur J Hum Genet.* (2020) 28:963–72. doi: 10.1038/s41431-020-0580-5
12. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet.* (2016) 48:1284–7. doi: 10.1038/ng.3656
13. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet.* (2019) 51:1749–55. doi: 10.1038/s41588-019-0530-8
14. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet.* (2021) 53:1527–33. doi: 10.1038/s41588-021-00945-5
15. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience.* (2019) 8:1–6. doi: 10.1093/gigascience/giz082
16. Bennett DA. How can I deal with missing data in my study? *Aust NZJ Public Health.* (2001) 25:464–9. doi: 10.1111/j.1467-842X.2001.tb00294.x
17. Murata T, Lin MI, Huang Y, Yu J, Bauer PM, Giordano FJ, et al. Reexpression of caveolin-1 in endothelium rescues the vascular, cardiac, and pulmonary defects in global caveolin-1 knockout mice. *J Exp Med.* (2007) 204:2373–82. doi: 10.1084/jem.20062340
18. Yi SL, Liu XJ, Zhong JQ, Zhang Y. Role of caveolin-1 in atrial fibrillation as an anti-fibrotic signaling molecule in human atrial fibroblasts. *PLoS ONE.* (2014)9:e85144. doi: 10.1371/journal.pone.0085144
19. Laurance S, Aghourian MN, Jiva Lila Z, Lemarié CA, Blostein MD. Gas6-induced tissue factor expression in endothelial cells is mediated through caveolin-1-enriched microdomains. *J Thromb Haemost.* (2014) 12:395–408. doi: 10.1111/jth.12481
20. Wunderlich C, Schober K, Lange SA, Drab M, Braun-Dullaeus RC, Kasper M, et al. Disruption of caveolin-1 leads to enhanced nitrosative stress and severe systolic and diastolic heart failure. *Biochem Biophys Res Commun.* (2006) 340:702–8. doi: 10.1016/j.bbrc.2005.12.058
21. Ay H, Benner T, Arsava EM, Furie KL, Singhal AB, Matt B. et al. A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke.* (2007) 38:2979–84. doi: 10.1161/STROKEAHA.107.490896
22. Aléssio AM, Höehr NF, Siqueira LH, Ozelo MC, de Pádua Mansur A, Annichino-Bizzacchi JM. Association between estrogen receptor alpha and beta gene polymorphisms and deep vein thrombosis. *Thromb Res.* (2007) 120:639–45. doi: 10.1016/j.thromres.2006.10.019
23. Díaz Del Moral S, Benaouicha M, Muñoz-Chápuli R, Carmona R. The insulin-like growth factor signalling pathway in cardiac development and regeneration. *Int J Mol Sci.* (2021) 23:234. doi: 10.3390/ijms23010234
24. Nolte IM, Munoz ML, Tragante V, Amare AT, Jansen R, Vaez A, et al. Genetic loci associated with heart rate variability and their effects on cardiac disease risk. *Nat Commun.* (2017) 8:15805. doi: 10.1038/ncomms15805
25. van den Berg ME, Warren HR, Cabrera CP, Verweij N, Mifsud B, Haessler J. et al. Discovery of novel heart rate-associated loci using the Exome Chip. *Hum Mol Genet.* (2017) 26:2346–63. doi: 10.1093/hmg/ddx113
26. *Common Metabolic Diseases Knowledge Portal*. 2021 Jun 17. Available online at: https://hugeamp.org/variant.html?variant=rs3746471 (accessed June 17, 2021)
27. *Common Metabolic Diseases Knowledge Portal*. Available online at: https://hugeamp.org/gene.html?gene=KIAA1755 (accessed June 17, 2021).
28. Pirruccello JP, Bick A, Wang M, Chaffin M, Friedman S, Yao J, et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat Commun.* (2020) 11:2254. doi: 10.1038/s41467-020-15823-7
29. Militello G, Hosen MR, Ponomareva Y, Gellert P, Weirick T, John D, et al. A novel long non-coding RNA Myolinc regulates myogenesis through TDP-43 and Filip1. *J Mol Cell Biol.* (2018) 10:102–17. doi: 10.1093/jmcb/mjy025
30. Pulit SL, Weng LC, McArdle PF, Trinquart L, Choi SH, Mitchell BD, et al. Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. *Neurol Genet.* (2018) 4:1–8. doi: 10.1212/NXG.0000000000000293

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

1 Supplementary Data

1.1 Detailed methods

1.1.1 Cohorts used in Multitrait analysis of GWAS (MTAG)

MEGASTROKE-CES

For the European ancestry analysis of MEGASTROKE consortium 16 different cohorts were analyzed, comprising up to 34,217 cases of ischemic stroke and 405,111 healthy controls. Stroke was defined according to the World Health Organization (WHO) as rapidly developing signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin. Strokes were defined as ischemic stroke (IS) or intracerebral hemorrhage (ICH) based on clinical and imaging criteria. IS was further subdivided into the following categories mostly using the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria): i) large vessel ischemic stroke; ii) cardioembolic ischemic stroke; iii) small vessel ischemic stroke. Specifically, for European analysis of CES they analyzed 7,193 cases of CES and 355,468 healthy controls. Further details of the cohorts in the original publication (Malik et al., 2018).

AF-2018 study

A total of 60,620 cases of AF and 970,216 controls were analyzed, these patients were recruited as part of main cohorts:

HUNT.—The Nord-Trøndelag Health Study (HUNT) is a population-based health survey conducted in the county of Nord-Trøndelag, Norway from 1984 to 2009 (S Krokstad, A Langhammer, K Hveem, T L Holmen, K Midthjell, T R Stene, G Bratberg, J Heggland, 2013). They used a combination of hospital, out-patient, and emergency room discharge diagnoses (ICD-9 and ICD-10) to identify 6,493 atrial fibrillation cases and 63,142 atrial fibrillation-free controls with genotype data. Participation in the HUNT Study is based on informed consent, and the study was approved by the Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway.

deCODE.—The Icelandic atrial fibrillation population consisted of all patients diagnosed with atrial fibrillation (ICD-10 code I48 and ICD-9 code 427.3) at Landspitali, The National University Hospital, in Reykjavik, and Akureyri Hospital (the two largest hospitals in Iceland) from 1987 to 2015. All atrial fibrillation cases, a total of 13,471, were included. Controls were 358,161 Icelanders recruited through different genetic research projects at deCODE genetics, excluding those in the atrial fibrillation cohort. The study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee of Iceland (no. VSNb2015030021).

MGI.—MGI is a hospital-based cohort collected at Michigan Medicine, USA. Atrial fibrillation cases (n = 1,226) were defined as patients with ICD-9 billing code 427.31, and controls were individuals without atrial fibrillation, atrial flutter, or related phenotypes (ICD-9 426–427.99). MGI was reviewed and approved by the Institutional Review Board of the University of Michigan Medical School.

DiscovEHR.—The DiscovEHR collaboration cohort is a hospital-based cohort including 58,124 genotyped individuals of European ancestry from the ongoing MyCode Community Health Initiative of the Geisinger Health System, USA(Carey et al., 2016). Atrial fibrillation cases (n = 6,679) were defined as DiscovEHR participants with at least one electronic health record problem list entry or at least two diagnosis code entries for two separate clinical encounters on separate calendar days for ICD-10 I48: atrial fibrillation and flutter. Corresponding controls (n = 41,803) were defined as

individuals with no electronic health record diagnosis code entries (problem list or encounter codes) for ICD-10 I48. The Study was approved by the Geisinger Institutional Review Board.

UK Biobank.—The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom(Sudlow et al., 2015). Cases of atrial fibrillation were selected using ICD-9 and ICD-10 codes for atrial fibrillation or atrial flutter (ICD-9 427.3 and ICD-10 I48). Controls were participants without any ICD-9 or ICD-10 codes specific for atrial fibrillation, atrial flutter, other cardiac arrhythmias, or conduction disorders.

AFGen Consortium.—Published atrial fibrillation association summary statistics from 31 cohorts representing 17,931 atrial fibrillation cases and 115,142 controls were obtained from the authors.

1.1.2 Single Nucleotide Variants (SNVs) quality control.

A series of standard quality controls were applied to select the single nucleotide variants (SNVs) for the MTAG analysis. Variant exclusion criteria: 1) Not common to the summary statistics of the traits, 2) Minor allele frequency lower or equal to 0.01, 3) Missing values, 4) Negative standard error or not a number value. 5) P-value of 0, 6) Not SNVs, 7) Duplicated SNVs, 8) Strand ambiguity, and 8) Inconsistent allele pairs. After QCs, a total of 6,808,676 SNVs were selected (eFigure 1). Locus 15q21.3 prioritized genes *GCOM1* and *MYZAP* for AF-2018 was not evaluated due to absent of the significant SNVs of AF-2018 in the MEGASTROKE-CES analysis.

1.1.3. Replication GWAS in GENERACION cohort

1.1.3.1 Study population

We performed analysis in GENERACION study cohort (Spain). 9,105 individuals (3,479 IS and 5,626 controls). IS patients were recruited via hospital-based studies between 2003 and 2020. The participants were part of the Genetics of Early Neurological Instability After Ischemic Stroke (GENISIS), Genetic contribution to Functional Outcome and Disability after Stroke (GODS), the Genetic Study in Ischemic Stroke Patients treated with tPA (GenoTPA), the CONtrol ICtus (CONIC), and SEDMAN studies. Controls were subjects without a history of ischemic stroke, aged over 18 years, who declared they were free of neurovascular diseases before recruitment. The control cohort was collected in blood donation at primary care centers in Barcelona and in hospitals throughout Spain as a part of the GCAT, CONtrol ICtus (CONIC), Investigating Silent Stroke in hYpertensives: A magnetic resonance imaging Study (ISSYS) and the Genotyping Recurrence Risk of Stroke (GRECOS) projects. Array information, contribution of hospitals and clinical description of the cohort are present in Online Tables I-III.

The ischemic stroke patients were recruited if they had a measurable neurologic deficit on the NIHSS within 6 hours of the last known asymptomatic status, had been diagnosed with stroke by an experienced neurologist, which had been confirmed by neuroimaging and were over 18 years of age.

These patients were recruited as part of the GENISIS(Heitsch et al., 2017), GODS(Mola-Caminal et al., 2019a), and CONIC(Domingues-Montanari et al., 2010) projects.

Controls were subjects without a history of ischemic stroke, aged over 18 years, who declared they were free of neurovascular diseases before recruitment. The control cohort was collected in blood donation and primary care centers in Barcelona and in hospitals throughout Spain as a part of the GCAT(Obón-Santacana et al., 2018) (Galván-Femenia, I, 2018), CONIC(Domingues-Montanari et al., 2010), GRECOS(Fernández-Cadenas et al., 2017), and ISSYS(Riba et al., 2012) projects.

Description of the cohorts included:

GENISIS cohort.—Genetics of Early Neurological Instability after Ischemic Stroke (GENISIS)(Heitsch et al., 2017) is an international study currently recruiting patients from four different locations: United States, Finland, Poland, and Spain. The inclusion criteria for the GENISIS study are IS patients (age ≥ 18 years) Collected from 2003 to 2016 with a measurable neurologic deficit on the NIHSS within 6 hours of last known normal. Patients who received endovascular thrombectomy, or for whom consent and/or a blood sample could not be obtained were excluded. For our study we only include Spanish patients. Genotyping was performed with Human Core Exome chip (Illumina®).

GODS cohort.—Genetic contribution to functional Outcome and Disability after Stroke (GODS)(Mola-Caminal et al., 2019b) project is a study that aim find genetic factors associated to stroke outcome. All participants met the following criteria: (1) European descent, aged >18 years, diagnosis of IS in the anterior vascular territory; (2) assessed by a neurologist during the acute phase of stroke; (3) initial stroke severity >4, according to the National Institutes of Health Stroke Scale (NIHSS); (4) information on post-stroke functional status at 3 months (or alternatively between 3-6 months); (5) evidence of acute IS in a neuroimaging study; (6) lack of concomitant pathology. Individuals with stroke recurrence during the follow-up period were excluded, as well as, posterior vascular territory and lacunar strokes. Samples were genotyped at the Genetic and Molecular Epidemiology Laboratory of McMaster University (David Braley Research Institute) in Ontario, Canada, with Human Core Exome chip (Illumina®).

CONIC cohort.—CONtrol ICtus (CONIC) study(Domingues-Montanari et al., 2010) is a national study that recruited controls and IS cases participants in Vall d'Hebron Hospital between 2007 and 2008. All controls were older than 65 years of age and declared free of dementia, neurovascular and/or cardiovascular disease, as evaluated by self-description during a direct interview before recruitment. Subjects with a history of first and/or second-degree neurovascular disorder were also excluded from the study.

The IS cases were admitted to the emergency department of a university hospital who had a documented middle cerebral artery (MCA) occlusion on transcranial Doppler ultrasonography (TCD) and received tPA in a standard 0.9-mg/kg dose (10% bolus, 90% continuous infusion during 1 hour) within 3 hours of symptom onset following National Institute of Neurological Disorders and Stroke (NINDS) recommendations. Cases and controls were genotyped with Human Core Exome chip (Illumina®).

GRECOS cohort.—Genotyping RECurrence Risk Of Stroke (GRECOS)(Fernández-Cadenas et al., 2017) project is a national study that aim find genetic factors associated with the recurrence after stroke. Control participants were selected from relatives of patients (wife or husband, without any consanguinity among cases and controls) and healthy volunteers visiting the same hospital for routine testing. They were >65 years of age and classified as free of neurovascular and cardiovascular history and familial history of stroke by direct interview before recruitment. All samples were genotyped with Human Core Exome chip (Illumina®).

ISSYS cohort.—Investigating Silent Stroke in hYpertensives: A magnetic resonance imaging Study (ISSYS)(Riba et al., 2012) is an observational prospective study in hypertensive participants to determine the prevalence of silent or magnetic resonance imaging (MRI)–defined brain infarcts and cognitive impairment. This cohort comprises 1000 non-demented individuals, aged 50 to 70 years old, and diagnosed of essential hypertension at least one year before inclusion in the ISSYS study. Those individuals were genotyped with Human Core Exome chip (Illumina®).

GCAT cohort.—GCAT health databank is a collection of health data and samples from participants of the "GCAT/Genomes for Life. Cohort Study of the Genomes of Catalonia Study"(Obón-Santacana et al., 2018). The aim of the GCAT project was study the genetic and environmental factors that lead to the appearance of chronic diseases in the general population. The study is conducted in several waves of data gathering, namely GCAT1, the baseline Survey from 2014-2017 and GCAT2, the GCAT follow-up in the second year. Data collection is done with web-based self-questionnaires, direct interviews, clinical data, and analyses of DNA blood derived samples. Genome-wide genotypes have been generated using Illumina Infinium SNV-bead array technology using the Multi-Ethnic Global (MEGAEX, V.2) consortium array. We used only GCAT1 genotyped patients and we exclude individuals with heart infarct or heart diseases or with non-Caucasian ancestry.

genotPA—Consecutive Caucasian patients with acute ischemic stroke who were admitted to the emergency room and received recombinant tissue-type plasminogen activator (r-tPA) within 4.5 hours of symptom onset were recruited. Patients were enrrolled at Spanish hospitals (Vall d'Hebron University Hospital, Hospital Clinic, Hospital Universitari de Girona Doctor Josep Trueta, Hospital de la Santa Creu i Sant Pau, Hospital Universitari Germans Trias I Pujol, Hospital Universitari del Mar, Hospital de Basurto) between 2002 to 2012. The study protocol was approved by the Ethics Committee of each center, all patients or relatives signed the informed consent.

Patients were identified by medical evaluation at emergency room arrival; stroke diagnosis was performed by trained neurologists and confirmed by neuroimaging. There were no exclusion criteria regarding age, sex or ethnicity. Follow-up CT scan at 24 hours after onset of symptoms or if neurological worsening occurred were performed and was classified according to European Cooperative Acute Stroke Study (ECASS) (Larrue et al., 1997).

SEDMAN cohort. - patients ≥ 18 years old, treated with acenocoumarol or dabigatran for stroke or systemic embolism prevention following the local recommendations. All patients had a stroke or transient ischemic attack (TIA) during the previous 14 days before the initiation of anticoagulation treatment and had a diagnostic of non-valvular atrial fibrillation. Only patients with mild to moderate stroke (less than 2/3 of the vascular territory) with initial Alberta Stroke Program Early CT Score (ASPECTS) in the first CT/MRI > 6 and National Institute of Health Stroke Scale (NIHSS) < 25 were included. All patients had a general condition which allowed the 12 months' follow-up. Only patients with stablished stroke were used in this analysis.

Quality control and imputation

DNA samples were genotyped on commercial arrays from Illumina (San Diego, CA) (Online Table II). Quality controls were performed using PLINK v1.9 and KING v2.1.3 software. For all datasets, samples were excluded if there was a mismatch between the genetic and reported sex, genotype call rate lower than 95%, excess or loss of heterozygosity, non-European detected as outliers of 1000 Genomes Project Phase 3 dataset (1000G), duplicate samples or relatedness at a PI-HAT>0.20. SNPs were excluded if call rate lower than 95%, located in non-autosomes, non-biallelic, strand ambiguous, monomorphic or were deviated from Hardy-Weinberg equilibrium (p-value $<10^{-6}$ in controls, $<10^{-10}$ in IS).

Imputation was performed in the Michigan Imputation Server Pipeline(Das et al., 2016) using Minimac4. HRC r1.1 2016 (GRCh37/hg19) was the reference panel used, with European population, and for phasing Eagle v2.4 was used. After imputation, we removed SNV with imputation score <0.6 and MAF <1%. SNVs that were not present in at least 95% of the individuals were removed.

Samples

DNA for all subjects were obtained from whole blood samples. A total of 10,066 samples were genotyped in the study. After QCs 9,105 samples fulfilled the QC criteria and were not missing for the phenotypes and covariates analyzed: CES, AF, sex, and age.

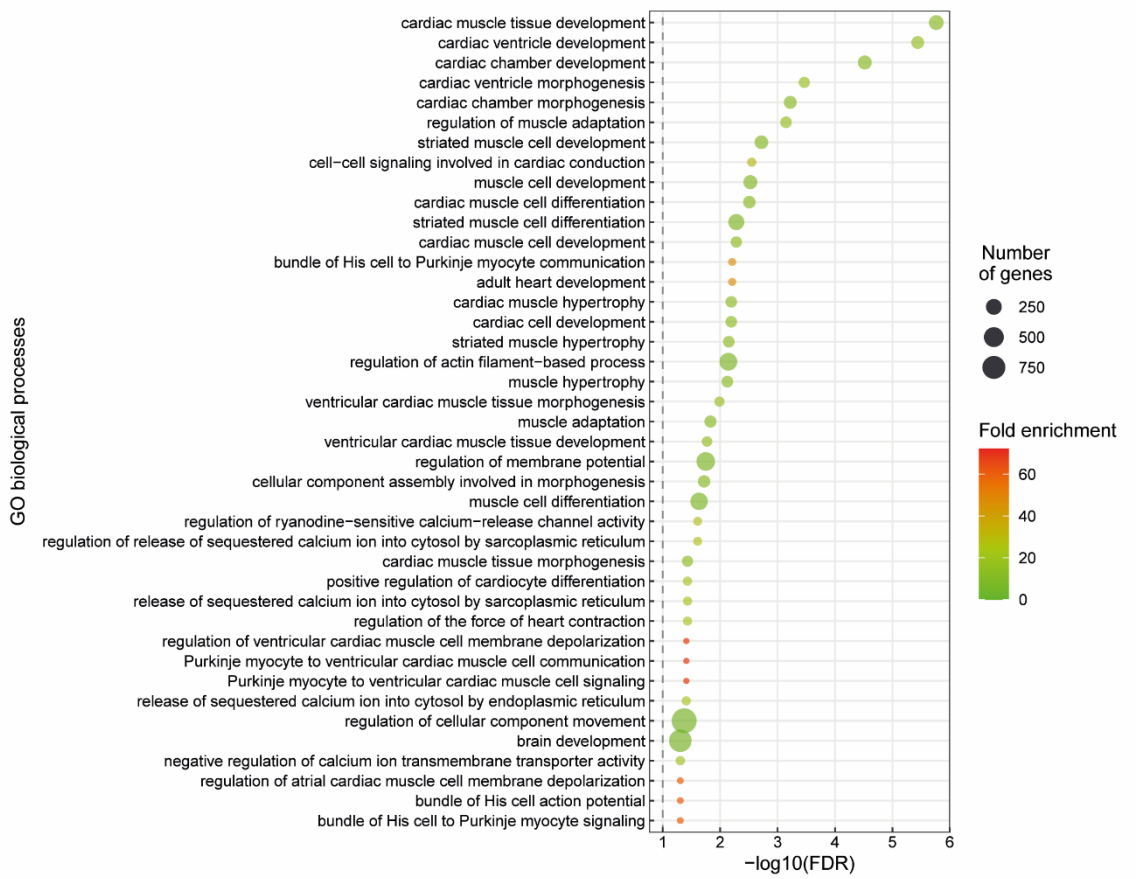
Replication-stage in an independent European cohort

Post-analysis quality controls

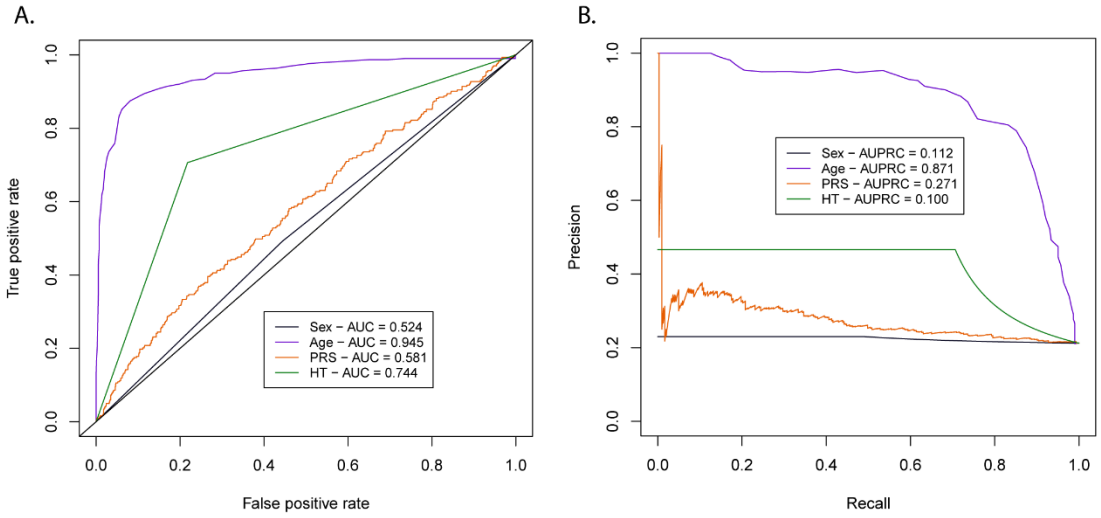
After analysis using fastGWA we removed variants with minor allele frequency < 1%, minor allele count in cases or controls < 6 and variants that deviated from Hardy Weinberg equilibrium p-value; $< 1x10^{-6}$. Additionally, genomic inflation was estimated as lambda.

2 Supplementary Figures and Tables

Supplementary Figure 1. GO Biological processes enriched in CES prioritized gene set.



Supplementary Figure 2. GO Biological processes enriched exclusively in analysis of AF associated genes independently of CES risk.



Supplementary Figure 3. Polygenic risk score (PRS) performance for the individual predictors. Panel A shows ROC curves and panel B Precision-Recall curves for the PRS performance in the independent test set. AUC: area under the ROC curve; AUPRC: area under the precision recall curve; HT: hypertension.

2.1 Supplementary Tables

Supplementary Tables are annexed in an online data Excel file.

Supplementary Table 1. Detailed number of individuals included form each participant Hospital on the independent cohort.

Supplementary Table 2. Detailed number of participants (IS and controls) included from each project. IS: ischemic strokes patients.

Supplementary Table 3. Clinical findings and univariate analysis of the additional cohort.

Supplementary Table 4. Variant-to-Gene prioritization of the 40 novel loci in the MTAG-CES analysis.

Supplementary Table 5. Results of CES analysis on the European independent cohort. AA: Assessed allele, OA: Other allele, CHR: chromosome, BP: base pairs, SE: Standard Error, B: beta, Z: z-score; P: p-value. Genomic location is in Hg19

Supplementary Table 6. Results of MTAG for previous and novel loci associated with AF. SE: Standard Error, B: beta, Z: z-score; P: p-value.

Supplementary Table 7. Gene Ontology of biological systems results for the gene sets of prioritized genes associated with a risk of cardioembolic stroke.

Supplementary Table 8. Gene Ontology of biological systems results for the gene sets of prioritized genes associated with atrial fibrillation risk exclusive.

Supplementary Table 9. Gene Ontology of biological systems results for the gene sets of prioritized genes associated with atrial fibrillation risk that do not overlap with those associated with CES.

Supplementary Table 10. Clinical characteristics of the training and test set used for the polygenic risk score analysis.

3. Supplementary References

Carey, D. J., Fetterolf, S. N., Davis, F. D., Faucett, W. A., Kirchner, H. L., Mirshahi, U., et al. (2016). The Geisinger MyCode community health initiative: An electronic health record-linked biobank for precision medicine research. *Genet. Med.* 18, 906–913. doi:10.1038/gim.2015.187.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi:10.1038/ng.3656.

Domingues-Montanari, S., Fernández-Cadenas, I., Del Río-Espinola, A., Mendioroz, M., Fernandez-Morales, J., Corbeto, N., et al. (2010). KCNK17 genetic variants in ischemic stroke. *Atherosclerosis* 208, 203–9. doi:10.1016/j.atherosclerosis.2009.07.023.

Fernández-Cadenas, I., Mendióroz, M., Giralt, D., Nafria, C., Garcia, E., Carrera, C., et al. (2017). GRECOS Project (Genotyping Recurrence Risk of Stroke). *Stroke* 48, 1147–1153. doi:10.1161/STROKEAHA.116.014322.

Heitsch, L., Ibanez, L., Carrera, C., Pera, J., Jimenez-Conde, J., Slowik, A., et al. (2017). Meta-analysis of Transethnic Association (MANTRA) Reveals Loci Associated With Neurological Instability After Acute Ischemic Stroke. in *International Stroke Conference*.

Larrue, V., Von Kummer, R., Del Zoppo, G., and Bluhmki, E. (1997). Hemorrhagic transformation in acute ischemic stroke: Potential contributing factors in the European Cooperative Acute Stroke Study. *Stroke* 28, 957–960. doi:10.1161/01.STR.28.5.957.

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* 50, 524–537. doi:10.1038/s41588-018-0058-3.

Mola-Caminal, M., Carrera, C., Soriano-Tárraga, C., Giralt-Steinhauer, E., Díaz-Navarro, R. M., Tur, S., et al. (2019a). PATJ Low Frequency Variants Are Associated with Worse Ischemic Stroke Functional Outcome: A Genome-Wide Meta-Analysis. *Circ. Res.* 124, 114–120. doi:10.1161/CIRCRESAHA.118.313533.

Mola-Caminal, M., Carrera, C., Soriano-Tárraga, C., Giralt-Steinhauer, E., Díaz-Navarro, R. M., Tur, S., et al. (2019b). *PATJ* Low Frequency Variants Are Associated With Worse Ischemic Stroke Functional Outcome. *Circ. Res.* 124, 114–120. doi:10.1161/CIRCRESAHA.118.313533.

Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., et al. (2018). GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* 8, e018324. doi:10.1136/bmjopen-2017-018324.

Riba, I., Jarca, C. I., Mundet, X., Tovar, J. L., Orfila, F., Nafría, C., et al. (2012). Cognitive assessment protocol design in the ISSYS (Investigating Silent Strokes in hYpertensives: A magnetic resonance imaging Study). *J. Neurol. Sci.* 322, 79–81. doi:10.1016/j.jns.2012.06.015.

S Krokstad, A Langhammer, K Hveem, T L Holmen, K Midthjell, T R Stene, G Bratberg, J Heggland, J. H. (2013). Cohort Profile: The HUNT Study. *Int. J. Epidemiol.*, 968–77. Available at: https://academic.oup.com/ije/article/42/4/968/655743 [Accessed July 10, 2020].

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* 12. doi:10.1371/journal.pmed.1001779.

Supplementary tables can be assessed from the following link:

https://www.frontiersin.org/articles/10.3389/fcvm.2022.940696/full

10.3. Original published “Causal Effect of MMP-1 (Matrix Metalloproteinase-1), MMP-8, and MMP-12 Levels on Ischemic Stroke: A Mendelian Randomization Study”

Publications details:

Journal: STROKE

Volume: 52

Issue: 7

Page: E316-E320

DOI: 10.1161/STROKEAHA.120.033041

Published: JUL 2021

Indexed: 2021-07-21

Document Type: Article

Metrics:

IF-2021 = 10.17

Journal Citation Category: CLINICAL NEUROLOGY - 14/212 (Q1)

Journal Citation Category: PERIPHERAL VASCULAR DISEASE - 7/67 (Q1)

Stroke

BRIEF REPORT

Causal Effect of MMP-1 (Matrix Metalloproteinase-1), MMP-8, and MMP-12 Levels on Ischemic Stroke

A Mendelian Randomization Study

Jara Cárcel-Márquez, MSc; Natalia Cullell, MSc; Elena Muiño, PhD; Cristina Gallego-Fabrega, PhD; Miquel Lledós, MSc; Laura Ibañez, PhD; Jerzy Krupinski, PhD; Joan Montaner, PhD; Carlos Cruchaga, PhD; Jin-Moo Lee, PhD; Dipender Gill, BMBCh, PhD; Guillaume Paré, MD, MSc; Marina Mola-Caminal, PhD; Jaume Roquer, PhD; Jordi Jimenez-Conde, PhD; Joan Martí-Fàbregas, PhD; Israel Fernandez-Cadenas, PhD

BACKGROUND AND PURPOSE: MMP (matrix metalloproteinase) levels have been widely associated with ischemic stroke risk and poststroke outcome. However, their role as a risk factor or as a subeffect because of ischemia is uncertain.

METHODS: We performed a literature search of genome-wide studies that evaluate serum/plasma levels of MMPs. We used a 2-sample Mendelian randomization approach to evaluate the causality of MMP levels on ischemic stroke risk or poststroke outcome, using 2 cohorts: MEGASTROKE (n=440 328) and GODs (n=1791).

RESULTS: Genome-wide association studies of MMP-1, MMP-8, and MMP-12 plasma/serum levels were evaluated. A significant association, which was also robust in the sensitivity analysis, was found with all ischemic strokes: MMP-12 (odds ratio=0.90 [95% CI, 0.86–0.94]; q value=7.43×10^{-5}), and with subtypes of stroke, large-artery atherosclerosis: MMP-1 (odds ratio=0.95 [95% CI, 0.92–0.98]; q value=0.01) and MMP-12 (odds ratio=0.71 [95% CI, 0.65–0.77]; q value=5.11×10^{-14}); small-vessel occlusion: MMP-8 (odds ratio=1.24 [95% CI, 1.06–1.45]; q value=0.03). No associations were found in relation to stroke outcome.

CONCLUSIONS: Our study suggests a causal link between lower serum levels of MMP-12 and the risk of ischemic stroke, lower serum levels of MMP-1 and MMP-12 and the risk of large-artery stroke and higher serum levels of MMP-8 and the risk of lacunar stroke.

GRAPHIC ABSTRACT: An online graphic abstract is available for this article.

Key Words: atherosclerosis ■ ischemic stroke ■ metalloproteinase ■ modified Rankin Scale ■ risk factor

MMPs (matrix metalloproteinases) are a diverse group of endopeptidases. They are known to mediate degradation or remodeling of the extracellular matrix and may be responsible for physiological processes such as wound healing and angiogenesis.[1] However, they are also responsible for pathophysiological processes like fibrotic disease and atherosclerosis.[1] Observational studies have found a correlation between MMP levels and risk of atheromatous plaque instability[2] and ischemic stroke (IS).[2] Several studies have also suggested that MMPs may play a key role in the outcome of IS.[3] Despite that, the causality of MMP levels has only recently been

Correspondence to: Israel Fernandez-Cadenas, PhD, Sant Pau Hospital Research Institute, Sant Antoni Maria Claret, 167, 08025, Barcelona, Spain. Email israelcadenas@yahoo.es

This manuscript was sent to Jean-Claude Baron, Guest Editor, for review by expert referees, editorial decision, and final disposition.

The Data Supplement is available with this article at https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.120.033041.

For Sources of Funding and Disclosures, see page XXX.

Stroke is available at www.ahajournals.org/journal/str

BRIEF REPORT

Nonstandard Abbreviations and Acronyms

IS	ischemic stroke
LAA	large-artery atherosclerosis
MMP	matrix metalloproteinase
MR	Mendelian randomization
OR	odds ratio
SVO	small vessel occlusion

proved regarding MMP-12, lower plasma levels of which are associated with a risk of large-artery atherosclerosis stroke (LAA).[4] This has prompted us to study other members of the MMP family in the context of IS.

Mendelian randomization (MR) is a technique that leverages genetic variants associated with heritable risk factors and diseases. This tool is particularly important in the study of complex diseases that may involve multiple factors and, therefore, a strategy to find the drivers behind the disease is needed.

We conducted a 2-sample MR study to test the hypothesis that serum/plasma levels of MMPs are causally associated with risk and long-term outcome of IS.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

SNV Selection and Data Sources

A literature search was conducted in PubMed in May 2020. The keywords used were: "matrix metalloproteinase", "serum" or "plasma", "levels" or "concentration", and "GWAS". Genome-wide significant ($P<5\times10^{-8}$) and independent ($r^2<0.2$) single nucleotide variants (SNVs) were used as instruments for MR analysis. To evaluate a causal effect on risk of stroke, we used summary-level data from the European analysis of MEGASTROKE[5] (IS subjects=34 217, controls=406 111) for IS and its subtypes: LAA (n=4373), cardioembolism (n=7193), and small-vessel occlusion (SVO) (n=5386). To evaluate poststroke functional outcome with the modified Rankin Scale at 3 months, we extracted summary-level data from the GODs (Genetic contribution to functional Outcome and Disability after Stroke) project in which modified Rankin Scale was analyzed as a continuous variable (n=1791).[6]

For each cohort, all aspects of the studies were approved by the local institutional review board and ethics committee. All the participants included, or their approved representative, provided written informed consent for participation.

The full methodology can be found in the Data Supplement, including a description of the cohorts evaluated in the study.

Statistical Analysis

The main MR method was inverse-variance weighted.[7] We applied the Benjamini-Hochberg procedure to control the false positive rate. Horizontal pleiotropy was assessed using Egger regression,[8] and heterogeneity was analyzed with Cochran Q statistic. Complementary MR approaches were applied: MR-Egger, weighted median, penalized weighted median, and weighted mode approaches.[8] Finally, when significant MR results were found, we used the MR-PRESSO (Mendelian Randomization Pleiotropy Residual Sum and Outlier) outlier test[9] and the leave-one-out analysis to explore the presence of outliers that could bias the results and performed the analysis extracting them.

Statistical analysis was conducted in R using TwoSampleMR, version 0.4.22.[10] The methodology followed in this study was in accordance with the latest guidelines for Mendelian randomization studies.[11] A checklist of questions to consider for MR studies can be found in the Data Supplement.

Table. Inverse-Variance Weighted Mendelian Randomization Results

Exposure	Outcome	Number of SNVs	Beta	SE	OR (CI 95%)	*P* value	*Q* value
MMP-1	IS	50	−0.01	0.01	0.99 (0.97–1.00)	3.47×10^{-02}	7.79×10^{-02}
MMP-1	LAA	50	−0.05	0.02	0.95 (0.92–0.98)	2.09×10^{-03}	1.04×10^{-02}
MMP-1	CE	50	0.00	0.01	1.00 (0.98–1.02)	9.72×10^{-01}	9.72×10^{-01}
MMP-1	SVO	50	−0.03	0.01	0.97 (0.94–1.00)	3.82×10^{-02}	7.79×10^{-02}
MMP-1	mRS	46	−0.04	0.02	0.96 (0.93–1.00)	2.86×10^{-02}	7.79×10^{-02}
MMP-8	IS	6	0.02	0.03	1.02 (0.96–1.07)	5.98×10^{-01}	6.41×10^{-01}
MMP-8	LAA	6	−0.06	0.07	0.94 (0.82–1.08)	3.90×10^{-01}	4.88×10^{-01}
MMP-8	CE	6	−0.08	0.11	0.93 (0.75–1.14)	4.74×10^{-01}	5.47×10^{-01}
MMP-8	SVO	6	0.21	0.08	1.24 (1.06–1.45)	7.69×10^{-03}	2.88×10^{-02}
MMP-8	mRS	6	−0.08	0.08	0.92 (0.79–1.07)	2.82×10^{-01}	4.23×10^{-01}
MMP-12	IS	7	−0.11	0.02	0.90 (0.86–0.94)	4.96×10^{-05}	7.43×10^{-05}
MMP-12	LAA	6	−0.35	0.04	0.71 (0.65–0.77)	3.41×10^{-15}	5.11×10^{-14}
MMP-12	CE	8	−0.04	0.04	0.96 (0.89–1.03)	2.72×10^{-01}	4.23×10^{-01}
MMP-12	SVO	8	−0.07	0.04	0.93 (0.86–1.00)	4.15×10^{-02}	7.79×10^{-02}
MMP-12	mRS	7	0.04	0.04	1.04 (0.96–1.12)	3.27×10^{-01}	4.46×10^{-01}

CE indicates cardioembolism; IS, ischemic stroke; LAA, large-artery atherosclerosis; MMP, matrix metalloproteinase; mRS, modified Rankin Scale; OR, odds ratio; SVO, small vessel occlusion; and SNV, single nucleotide variant.

RESULTS

Data Sources

After searching the literature, we found a total of 6 studies; only 3 of them performed Genome-wide association study on MMP levels and were selected, evaluating MMP-1,[12] MMP-8,[13] and MMP-12[14] levels. The SNVs used are specified in Tables I through III in the Data Supplement.

Primary MR Analysis

For MMP-1, a significant association with LAA was observed: odds ratio (OR), 0.95 ([95% CI, 0.92–0.98]; q value=0.01). MMP-8 showed a significant association with SVO, OR, 1.24 ([95% CI, 1.06–1.45]; q value=0.03). Regarding MMP-12, a significant association with IS and LAA was observed: OR IS: 0.90 ([95% CI, 0.86–0.94]; q value=7.43×10^{-5}) and OR LAA: 0.71 ([95% CI, 0.65–0.77]; q value=5.11×10^{-14}). No significant associations with stroke outcome were observed (Table).

Sensitivity Tests

The complementary analysis proved consistency between the different tests (Figure; Table IV in the Data

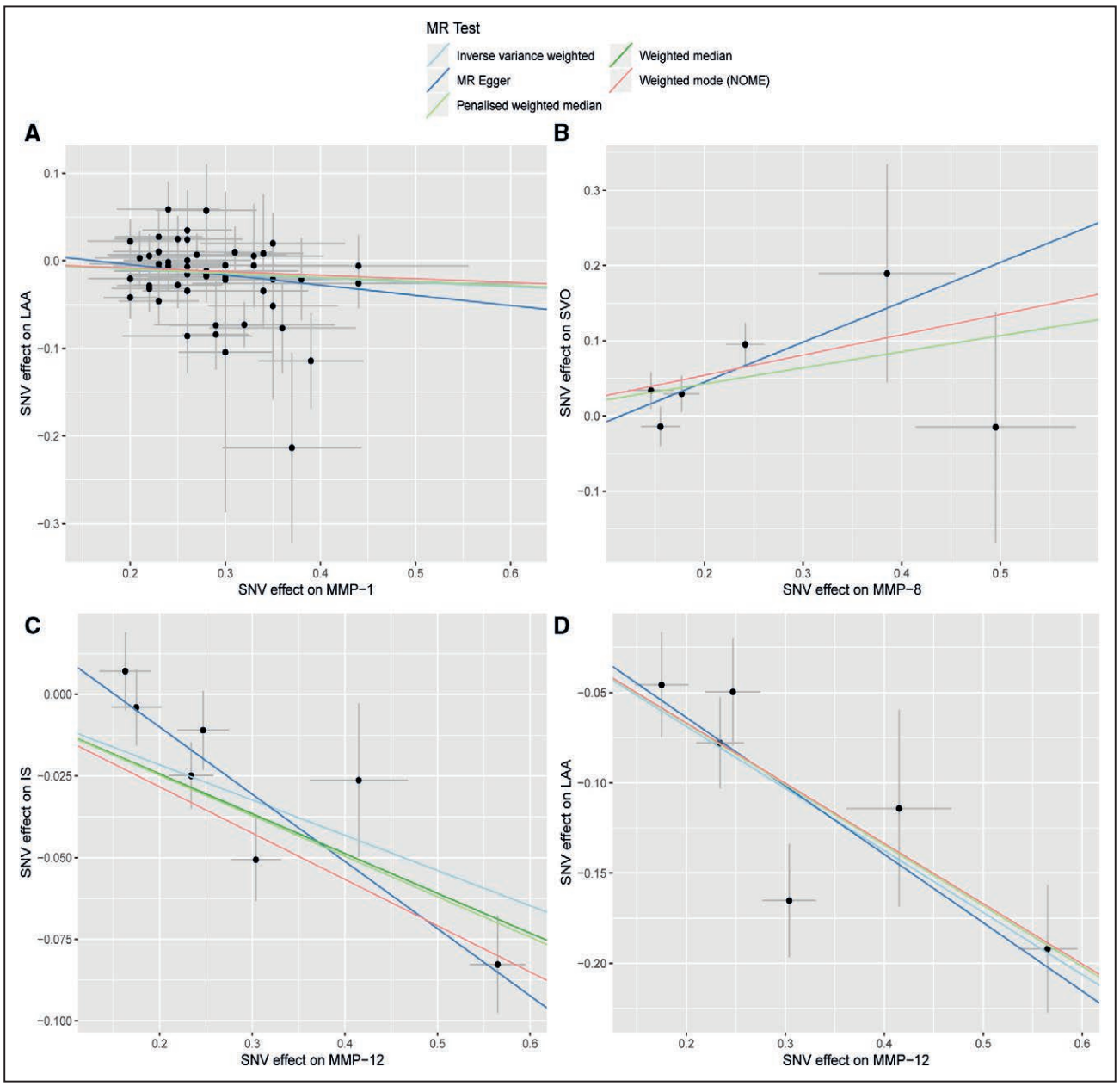


Figure. Scatterplots of significant Mendelian randomization results.
A, Analysis of MMP-1 (matrix metalloproteinase-1) and large-artery atherosclerosis (LAA); (**B**) analysis of MMP-8 and small vessel occlusion; (**C**) analysis of MMP-12 and ischemic stroke (IS); (**D**) analysis of MMP-12 and LAA. SNV indicates single nucleotide variant.

Supplement). Pleiotropy and heterogeneity were found to be significant for MMP-12 (Table V in the Data Supplement). The MR-PRESSO outlier test revealed 2 single nucleotide variant outliers for LAA analysis and one single nucleotide variant outlier for IS analysis (Tables VI and VII in the Data Supplement). No significant pleiotropy or heterogeneity were detected after removal of the outlier (Table V in the Data Supplement). The leave-one-out analysis showed that causality was not driven by a single nucleotide variant (Figures I and II in the Data Supplement). The results for MMP-1 and MMP-8 were not biased by heterogeneity, pleiotropy or by a Single nucleotide variant (Table V and Figures III and IV in the Data Supplement).

DISCUSSION

The goal of this study was to clarify the causality of MMP serum levels and the risk of IS, its subtypes, and functional recovery after stroke evaluated as modified Rankin Scale at 3 months. Following an MR approach, we explored causality of MMP-1, MMP-8, and MMP-12, leading to significant and robust results in the sensitivity analysis for MMP-1 serum levels and the risk of LAA, MMP-8 serum levels and risk of SVO and MMP-12 serum levels and risk of LAA and IS.

Lower serum levels of MMP-1 were found to be a risk factor for LAA. Beneficial higher levels of MMP-1 in the biology of IS may be because of its role in extracellular matrix remodeling and repair by degradation of components such as collagen I, II, and III.[12]

With regard to MMP-8, a causal relationship of higher serum levels as a risk factor for SVO was detected. MMP-8 cleaves collagen I 3 times more effectively than MMP-1 or MMP-13, which makes this protein a potential disruptor of extracellular matrix compounds.[15] Dysregulation of the matrisome by mutation in collagenase genes has been found to be the cause of small vessel diseases[15] such as SVO.

Regarding MMP-12, we found a causal relationship between lower plasma levels of MMP-12 and the risk of IS and LAA, which has been observed previously.[4]

In our study, no significant association was observed between serum levels of MMP-1, MMP-8, and MMP-12 and mRS. These results suggest that these metalloproteinases do not play a crucial role in the process of stroke recovery, although other members of the MMP family may be involved. However, we must consider that the GODs study has a limited number of samples and they are mainly of Spanish ancestry.

A strength of the present study is the consistency of the results across different MR tests and sensitivity analyses. Furthermore, our findings are unlikely to be confounded by population stratification because the analyses included individuals of European ancestry only.

In conclusion, using a Mendelian randomization approach, our study suggests causality between lower serum levels of MMP-12 and the risk of IS, lower serum levels of MMP-1 and MMP-12 and the risk of large-artery stroke and higher serum levels of MMP-8 and the risk of lacunar stroke. Therefore, we think our results suggest the potential use of these as biomarkers and therapeutic targets for stroke risk. However, further research is needed, in addition to an analysis of the 26 human MMPs, to understand the full biology of these proteins in IS pathology.

ARTICLE INFORMATION

Received June 18, 2020; final revision received January 24, 2021; accepted March 22, 2021.

Affiliations

Stroke Pharmacogenomics and Genetics Laboratory, Sant Pau Research Institute, Barcelona, Spain (J.C.-M., N.C., E.M., C.G.-F., M.L., I.F.-C.). Department of Medicine, Universitat Autònoma de Barcelona, Spain (J.C.-M.). Stroke Pharmacogenomics and Genetics Laboratory, Fundación Docència I Recerca Mútua Terrassa, Hospital Mútua de Terrassa, Spain (N.C., C.G.-F., J.K., I.F.-C.). Department of Psychiatry (L.I., C.C.) and Department of Neurology (J.-M.L.), Washington University School of Medicine, Saint Louis, MO. Institute de Biomedicine of Seville, IBiS/Hospital Universitario Virgen del Rocío/CSIC/University of Seville, Department of Neurology, Hospital Universitario Virgen Macarena, Spain (J.M.). Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, United Kingdom (D.G.). McMaster University, Hamilton, Ontario, Canada (G.P.). Department of Neurology, IMIM-Hospital del Mar, Neurovascular Research Group, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, Spain (M.M.-C., J.R., J.J.-C.). Stroke Unit, Department of Neurology, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain (J.M.-F.).

Acknowledgments

We thank the MEGASTROKE and International Stroke Genetics consortia for providing summary statistics data for ischemic stroke.

Sources of Funding

This research was funded by agència de gestió d'ajuts universitaris i de recerca (AGAUR; grant number 2019_FIB00853) co-financed with Fons Social Europeu (FSE), by Fondo Europeo de Desarrollo Regional (FEDER) and the Carlos III Health Institute (grant numbers PI15/01978, PI17/02089, PI18/01338, CM18/00198, RD16/0019/0011, FI19/00309), by Marató TV3 through the Epigenesis study and the GODs study, by Fundació Docència i Recerca Mutua Terrassa grant for the Epigenesis project, and INVICTUS PLUS groups: RD0016/0019/0011, RD0016/0019/0010 and RD0016/0019/0002. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgments.html.

Disclosures

Dr Gill is employed part-time by Novo Nordisk. The other authors report no conflicts.

Supplemental Materials

Online Methods
Online Tables I–VII
Online Figures I–IV
Online Note
References16–20

REFERENCES

1. Rempe RG, Hartz AMS, Bauer B. Matrix metalloproteinases in the brain and blood-brain barrier: versatile breakers and makers. *J Cereb Blood Flow Metab*. 2016;36:1481–1507. doi: 10.1177/0271678X16655551
2. Chen L, Yang Q, Ding R, Liu D, Chen Z. Carotid thickness and atherosclerotic plaque stability, serum inflammation, serum MMP-2 and MMP-9 were

associated with acute cerebral infarction. *Exp Ther Med*. 2018;16:5253–5257. doi: 10.3892/etm.2018.6868
3. Ma F, Rodriguez S, Buxo X, Morancho A, Riba-Llena I, Carrera A, Bustamante A, Giralt D, Montaner J, Martinez C, et al. Plasma matrix metalloproteinases in patients with stroke during intensive rehabilitation therapy. *Arch Phys Med Rehabil*. 2016;97:1832–1840. doi: 10.1016/j.apmr.2016.06.007
4. Chong M, Sjaarda J, Pigeyre M, Mohammadi-Shemirani P, Lali R, Shoamanesh A, Gerstein HC, Paré G. Novel drug targets for ischemic stroke identified through Mendelian randomization analysis of the blood proteome. *Circulation*. 2019;140:819–830. doi: 10.1161/CIRCULATIONAHA.119.040180
5. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van der Laan SW, Gretarsdottir S, et al; AFGen Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium; International Genomics of Blood Pressure (iGEN-BP) Consortium; INVENT Consortium; STARNET; BioBank Japan Cooperative Hospital Group; COMPASS Consortium; EPIC-CVD Consortium; EPIC-InterAct Consortium; International Stroke Genetics Consortium (ISGC); METASTROKE Consortium; Neurology Working Group of the CHARGE Consortium; NINDS Stroke Genetics Network (SiGN); UK Young Lacunar DNA Study; MEGASTROKE Consortium. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet*. 2018;50:524–537. doi: 10.1038/s41588-018-0058-3
6. Mola-Caminal M, Carrera C, Soriano-Tárraga C, Giralt-Steinhauer E, Díaz-Navarro RM, Tur S, Jiménez C, Medina-Dols A, Cullell N, Torres-Aguila NP, et al. PATJ low frequency variants are associated with worse ischemic stroke functional outcome: a genome-wide meta-analysis. *Circ Res*. 2019;124:114–120.
7. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. 2017;46:1985–1998. doi: 10.1093/ije/dyx102
8. Bowden J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *Int J Epidemiol*. 2017;46:2097–2099. doi: 10.1093/ije/dyx192
9. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. 2018;50:693–698. doi: 10.1038/s41588-018-0099-7
10. Rasooly D, Patel CJ. Conducting a reproducible Mendelian randomization analysis using the R analytic statistical environment. *Curr Protoc Hum Genet*. 2019;101:e82. doi: 10.1002/cphg.82
11. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, Hartwig FP, Holmes MV, Minelli C, Relton CL, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res*. 2020;4:186.
12. Cheng YC, Kao WH, Mitchell BD, O'Connell JR, Shen H, McArdle PF, Gibson Q, Ryan KA, Shuldiner AR, Pollin TI. Genome-wide association scan identifies variants near Matrix Metalloproteinase (MMP) genes on chromosome 11q21-22 strongly associated with serum MMP-1 levels. *Circ Cardiovasc Genet*. 2009;2:329–337. doi: 10.1161/CIRCGENETICS.108.834986
13. Salminen A, Vlachopoulou E, Havulinna AS, Tervahartiala T, Sattler W, Lokki ML, Nieminen MS, Perola M, Salomaa V, Sinisalo J, et al. Genetic variants contributing to circulating matrix metalloproteinase 8 levels and their association with cardiovascular diseases: a genome-wide analysis. *Circ Cardiovasc Genet*. 2017;10:e001731.
14. Mahdessian H, Perisic Matic L, Lengquist M, Gertow K, Sennblad B, Baldassarre D, Veglia F, Humphries SE, Rauramaa R, de Faire U, et al; IMPROVE Study Group. Integrative studies implicate matrix metalloproteinase-12 as a culprit gene for large-artery atherosclerotic stroke. *J Intern Med*. 2017;282:429–444. doi: 10.1111/joim.12655
15. Joutel A, Haddad I, Ratelade J, Nelson MT. Perturbations of the cerebrovascular matrisome: a convergent mechanism in small vessel disease of the brain? *J Cereb Blood Flow Metab*. 2016;36:143–157.
16. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
17. Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, Jaquish C, Douglas JA, Roy-Gagnon MH, Sack P, et al. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. *Am Heart J*. 2008;155:823–828. doi: 10.1016/j.ahj.2008.01.019
18. Vaara S, Nieminen MS, Lokki ML, Perola M, Pussinen PJ, Allonen J, Parkkonen O, Sinisalo J. Cohort profile: the Corogene study. *Int J Epidemiol*. 2012;41:1265–1271. doi: 10.1093/ije/dyr090
19. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, Männistö S, Salomaa V, Sundvall J, Puska P. Forty-year trends in cardiovascular risk factors in Finland. *Eur J Public Health*. 2015;25:539–546. doi: 10.1093/eurpub/cku174
20. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, Eriksson A, Rennel Dickens E, Ohlsson S, Edfeldt G, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One*. 2014;9:e95192. doi: 10.1371/journal.pone.0095192

SUPPLEMENTAL MATERIAL

1. Supplementary Methods

Selection of instrumental variables

Single Nucleotide Variants (SNVs) were selected from the data on MMP-1, MMP-8 and MMP-12 serum/plasma levels in GWAS analyses[12–14]. Significantly associated (p-value $< 5x10^{-8}$) SNVs for each metalloproteinase: 156 associated with serum levels of MMP-1, 499 associated with serum levels of MMP-8 and 12 associated with plasma levels of MMP-12. We checked for the presence of the associated variants in the MEGASTROKE[5] and GODs[6] summary statistics, then performed a clumping of the variants with a threshold of $r^2 < 0.2$ using the 1000 G European reference panel[16]. This led to a different number of SNVs; for risk analysis using MEGASTROKE we selected a total of 50 independent variants for MMP-1, 6 independent variants for MMP-8 and 8 independent variants for MMP12, and for post-stroke functional outcome we selected 46 independent variants for MMP-1, 6 independent variants for MMP-8 and 7 independent variants for MMP-12 (Online Tables I-III). These variants were used to perform the different Mendelian randomization analyses for the 5 outcomes used, namely ischemic stroke and ischemic stroke subtypes: large-artery atherosclerosis, cardioembolism, small-vessel occlusion and post-stroke outcome evaluated by modified Rankin scale three months after stroke.

Checklist of questions to consider for Mendelian randomization studies [11]

1. What is the primary hypothesis of interest?

Serum/plasma levels of MMPs are causally associated with ischemic stroke risk and long-term outcome.

2. Data sources

We performed two-sample MR analyses. Summary-level GWAS data from European ancestry studies evaluating MMP-1[12] serum levels, MMP-8[13] serum levels, MMP-12[14] plasma levels, modified Rankin scale 3 months post-stroke[6] and ischemic stroke risk and subtypes[5] were used. No sample overlap is expected after evaluation of the cohorts descripted in each study.

3. Selection of genetic variants

Genetic variants were selected reaching a p-value $< 5 \cdot 10^{-8}$ at a clump of $r^2 < 0.2$ from summary-level GWAS data. When pleiotropy was detected MR-PRESSO outliers test was performed in order to detect and remove the pleiotropic variants.

4. Variant harmonization

The correct orientation of the variants' alleles was performed for all studies.

5. Primary analysis

The main MR method was inverse-variance weighted. We applied the Benjamini-Hochberg procedure to control the false positive rate.

6. and 7. Supplementary and sensitivity analyses

Horizontal pleiotropy was assessed using Egger regression[8] and heterogeneity was analyzed with Cochran's Q statistic. Complementary MR approaches were applied: MR-Egger, weighted median, penalized weighted median and weighted mode approaches[8]. Finally, when significant MR results were found, we used the MR-PRESSO outlier test[9] and the leave-one-out analysis to explore the presence of outliers that could bias the results and performed the analysis extracting them.

8. Data presentation

The results of IVW MR primary analysis are presented in table 1. For the significant and robust results, we performed scatterplots showed in figure 1. Supplementary and sensitivity analyses are presented in supplemental material in tables but leave-one-out test that are showed in forest plots.

9. Interpretation

A significant causal association was considered to exist when the adjusted p-value (q-value) was lower than 0.05, complementary methods showed same direction of the association, no pleiotropy or heterogeneity was detected and no single SNV was driving the causal association.

2. Online Tables

SNV	EA	OA	Freq EA	Beta	SE	P value	Analysis
rs495366	A	G	0.36	-0.44	0.18	5.73E-34	Risk and mRS
rs1942518	C	A	0.37	0.35	0.13	4.03E-21	Risk and mRS
rs7115014	A	G	0.35	0.32	0.10	3.99E-19	Risk and mRS
rs11226373	G	A	0.15	0.44	0.12	1.38E-18	Risk and mRS
rs685395	T	C	0.26	-0.33	0.10	3.22E-17	Risk and mRS
rs1944432	T	C	0.42	0.30	0.10	4.87E-17	Risk and mRS
rs10895597	A	G	0.40	-0.28	0.10	8.73E-16	Risk and mRS
rs666825	C	T	0.42	-0.28	0.08	2.40E-15	Risk and mRS
rs11602707	T	C	0.20	0.35	0.08	6.52E-15	Risk and mRS
rs1939052	C	T	0.40	0.27	0.09	1.14E-14	Risk and mRS
rs3819089	T	C	0.26	-0.30	0.08	1.07E-13	Risk and mRS
rs7939072	T	C	0.18	-0.33	0.07	6.31E-13	Risk and mRS
rs2466912	A	G	0.48	-0.26	0.07	1.06E-12	Risk and mRS
rs1940054	C	A	0.38	0.26	0.06	1.24E-12	Risk and mRS
rs613804	A	G	0.12	-0.39	0.06	1.91E-12	Risk and mRS
rs17710616	T	C	0.13	-0.37	0.07	5.29E-12	Risk and mRS
rs11225649	A	C	0.21	0.30	0.05	5.30E-12	Risk and mRS
rs2515081	T	A	0.14	-0.36	0.08	8.48E-12	Risk
rs7125424	T	G	0.13	-0.35	0.07	1.31E-11	Risk and mRS
rs7930146	C	T	0.22	-0.29	0.04	2.39E-11	Risk and mRS
rs480950	A	C	0.44	0.24	0.06	2.69E-11	Risk and mRS
rs12805072	T	C	0.27	0.26	0.07	2.94E-11	Risk and mRS
rs1939008	A	G	0.29	0.25	0.07	3.04E-11	Risk and mRS
rs1301783	G	A	0.50	0.22	0.03	6.58E-11	Risk and mRS
rs10791643	A	G	0.22	0.28	0.06	9.32E-11	Risk and mRS
rs523332	G	T	0.36	-0.23	0.07	1.25E-10	Risk and mRS
rs168636	G	A	0.24	0.26	0.05	2.99E-10	Risk and mRS
rs523519	A	C	0.39	0.23	0.06	3.93E-10	Risk and mRS
rs10502070	A	G	0.28	0.24	0.06	4.00E-10	Risk and mRS
rs11226865	A	G	0.16	0.31	0.04	4.64E-10	Risk and mRS
rs1531751	C	G	0.25	-0.24	0.05	6.09E-10	Risk
rs7946913	A	G	0.49	-0.22	0.04	9.66E-10	Risk and mRS
rs3181174	A	T	0.16	-0.29	0.05	1.38E-09	Risk
rs11220658	A	G	0.24	0.25	0.05	2.37E-09	Risk and mRS
rs3019723	C	T	0.20	-0.26	0.06	3.04E-09	Risk and mRS
rs4320978	G	A	0.22	0.26	0.04	3.27E-09	Risk and mRS
rs11224733	A	T	0.17	-0.28	0.05	3.50E-09	Risk
rs1815913	C	T	0.47	0.20	0.04	6.14E-09	Risk and mRS
rs502318	C	T	0.31	-0.23	0.05	6.66E-09	Risk and mRS
rs17106456	C	A	0.14	-0.30	0.05	7.99E-09	Risk and mRS
rs668285	T	C	0.24	0.23	0.05	8.76E-09	Risk and mRS
rs480846	A	G	0.09	0.34	0.05	1.19E-08	Risk and mRS
rs10502062	G	T	0.44	0.20	0.05	1.83E-08	Risk and mRS
rs4342998	A	C	0.38	0.22	0.04	1.89E-08	Risk and mRS

rs11217456	T	C	0.26	0.23	0.04	2.80E-08	Risk and mRS
rs17094347	T	C	0.19	-0.26	0.05	3.28E-08	Risk and mRS
rs11226791	C	T	0.08	0.34	0.04	3.31E-08	Risk and mRS
rs1052313	A	G	0.40	-0.20	0.03	3.35E-08	Risk and mRS
rs10890667	C	T	0.08	0.38	0.04	3.78E-08	Risk and mRS
rs7946787	C	A	0.36	-0.21	0.05	4.19E-08	Risk and mRS

Online Table I. SNVs used in the analysis of MMP-1 serum levels as exposure. SNV: Single Nucleotide Variant. EA: Effect Allele. OA: Other Allele. Freq: frequency. SE: Standard Error.

SNV	EA	OA	Freq EA	Beta	SE	P value	Analysis
rs800292	A	G	0.30	-0.24	0.02	2.42E-35	Risk and mRS
rs1409153	T	C	0.55	-0.18	0.02	1.81E-22	Risk and mRS
rs1560833	A	G	0.28	-0.16	0.02	5.31E-15	Risk and mRS
rs10922198	C	G	0.56	0.15	0.02	1.61E-10	Risk
rs193201657	T	C	0.98	0.50	0.08	1.02E-09	Risk and mRS
rs2184850	T	C	0.55	-0.12	0.02	2.14E-08	mRS
rs148136314	T	C	0.03	-0.39	0.07	2.57E-08	Risk and mRS

Online Table II. SNVs used in the analysis of MMP-8 as exposure. SNV: Single Nucleotide Variant. EA: Effect Allele. OA: Other Allele. Freq: frequency. SE: Standard Error.

SNV	EA	OA	Freq EA	Beta	SE	P value	Analysis
rs499459	A	G	0.18	-0.57	0.03	8.26E-76	Risk and mRS
rs1892971	A	G	0.26	0.30	0.03	7.97E-29	Risk and mRS
rs671188	C	T	0.44	-0.23	0.02	1.02E-21	Risk and mRS
rs2186789	G	T	0.24	-0.25	0.03	4.98E-18	Risk and mRS
rs613804	C	T	0.06	0.42	0.05	3.80E-15	Risk and mRS
rs1942524	C	T	0.27	0.18	0.03	1.38E-10	Risk and mRS
rs484915	A	T	0.41	0.15	0.02	1.65E-09	Risk
rs650108	A	G	0.25	0.16	0.03	5.39E-09	Risk and mRS

Online Table III. SNVs used in the analysis of MMP-12 as exposure. SNV: Single Nucleotide Variant. EA: Effect Allele. OA: Other Allele. Freq: frequency. SE: Standard Error.

Exposure	Outcome	Method	no SNV	Beta	Standard Error	OR	p-value
MMP-1	IS	IVW	50	-0.01	0.01	0.99 (0.97-1.00)	3.47E-02
MMP-1	IS	MR Egger	50	-0.06	0.03	0.94 (0.88-1.00)	5.09E-02
MMP-1	IS	WM	50	-0.01	0.01	0.99 (0.97-1.00)	1.36E-01
MMP-1	IS	PWM	50	-0.01	0.01	0.99 (0.97-1.01)	1.58E-01
MMP-1	IS	WMode	50	-0.02	0.02	0.98 (0.95-1.02)	3.21E-01
MMP-1	LAA	IVW	50	-0.05	0.02	0.95 (0.92-0.98)	2.09E-03
MMP-1	LAA	MR Egger	50	-0.12	0.08	0.89 (0.76-1.04)	1.45E-01
MMP-1	LAA	WM	50	-0.05	0.02	0.95 (0.91-1.00)	3.73E-02
MMP-1	LAA	PWM	50	-0.05	0.02	0.95 (0.91-1.00)	4.42E-02
MMP-1	LAA	WMode	50	-0.04	0.04	0.96 (0.89-1.04)	3.01E-01
MMP-1	CE	IVW	50	0.00	0.01	1.00 (0.98-1.02)	9.72E-01
MMP-1	CE	MR Egger	50	0.04	0.06	1.04 (0.92-1.17)	5.35E-01
MMP-1	CE	WM	50	0.01	0.02	1.01 (0.97-1.04)	6.24E-01
MMP-1	CE	PWM	50	0.01	0.02	1.01 (0.97-1.05)	6.15E-01
MMP-1	CE	WMode	50	0.03	0.03	1.03 (0.96-1.10)	4.40E-01
MMP-1	SVO	IVW	50	-0.03	0.01	0.97 (0.94-1.00)	3.82E-02
MMP-1	SVO	MR Egger	50	-0.15	0.07	0.86 (0.75-0.99)	4.75E-02
MMP-1	SVO	WM	50	-0.02	0.02	0.98 (0.93-1.02)	2.55E-01
MMP-1	SVO	PWM	50	-0.02	0.02	0.98 (0.94-1.02)	3.12E-01
MMP-1	SVO	WMode	50	-0.03	0.04	0.97 (0.90-1.05)	4.97E-01
MMP-1	mRS	IVW	46	-0.04	0.02	0.96 (0.93-1.00)	2.86E-02
MMP-1	mRS	MR Egger	46	0.07	0.08	1.08 (0.92-1.26)	3.56E-01
MMP-1	mRS	WM	46	-0.04	0.02	0.96 (0.92-1.01)	9.97E-02
MMP-1	mRS	PWM	46	-0.04	0.02	0.96 (0.92-1.01)	9.19E-02
MMP-1	mRS	WMode	46	-0.02	0.05	0.98 (0.89-1.07)	6.12E-01
MMP-8	IS	IVW	6	0.02	0.03	1.02 (0.96-1.07)	5.98E-01
MMP-8	IS	MR Egger	6	0.04	0.11	1.04 (0.84-1.29)	7.54E-01
MMP-8	IS	WM	6	0.03	0.03	1.03 (0.96-1.10)	3.77E-01
MMP-8	IS	PWM	6	0.03	0.03	1.03 (0.96-1.10)	3.80E-01
MMP-8	IS	WMode	6	0.03	0.04	1.03 (0.95-1.11)	5.58E-01
MMP-8	LAA	IVW	6	-0.06	0.07	0.94 (0.82-1.08)	3.90E-01
MMP-8	LAA	MR Egger	6	0.00	0.28	1.00 (0.58-1.74)	9.99E-01
MMP-8	LAA	WM	6	-0.08	0.09	0.93 (0.78-1.10)	3.70E-01
MMP-8	LAA	PWM	6	-0.08	0.09	0.93 (0.78-1.09)	3.62E-01
MMP-8	LAA	WMode	6	-0.09	0.11	0.91 (0.74-1.13)	4.35E-01
MMP-8	CE	IVW	6	-0.08	0.11	0.93 (0.75-1.14)	4.74E-01
MMP-8	CE	MR Egger	6	-0.71	0.28	0.49 (0.29-0.84)	6.16E-02
MMP-8	CE	WM	6	-0.04	0.07	0.96 (0.83-1.10)	5.32E-01
MMP-8	CE	PWM	6	-0.04	0.07	0.96 (0.83-1.11)	5.84E-01
MMP-8	CE	WMode	6	-0.03	0.08	0.97 (0.83-1.12)	6.81E-01
MMP-8	SVO	IVW	6	0.21	0.08	1.24 (1.06-1.45)	7.69E-03
MMP-8	SVO	MR Egger	6	0.53	0.31	1.70 (0.92-3.15)	1.68E-01
MMP-8	SVO	WM	6	0.21	0.09	1.24 (1.03-1.49)	2.26E-02
MMP-8	SVO	PWM	6	0.21	0.09	1.24 (1.04-1.48)	1.91E-02

MMP-8	SVO	WMode	6	0.27	0.11	1.31 (1.06-1.62)	5.59E-02
MMP-8	mRS	IVW	6	-0.08	0.08	0.92 (0.79-1.07)	2.82E-01
MMP-8	mRS	MR Egger	6	-0.15	0.25	0.86 (0.53-1.41)	5.91E-01
MMP-8	mRS	WM	6	-0.04	0.09	0.96 (0.80-1.15)	6.39E-01
MMP-8	mRS	PWM	6	-0.04	0.09	0.96 (0.80-1.15)	6.49E-01
MMP-8	mRS	WMode	6	-0.05	0.10	0.95 (0.78-1.16)	6.55E-01
MMP-12	IS	IVW	7	-0.11	0.02	0.90 (0.86-0.94)	4.96E-06
MMP-12	IS	MR Egger	7	-0.21	0.04	0.81 (0.75-0.89)	4.80E-03
MMP-12	IS	WM	7	-0.12	0.02	0.89 (0.85-0.92)	2.65E-08
MMP-12	IS	PWM	7	-0.12	0.02	0.88 (0.85-0.92)	2.81E-08
MMP-12	IS	WMode	7	-0.14	0.02	0.87 (0.83-0.91)	1.20E-03
MMP-12	LAA	IVW	6	-0.34	0.04	0.71 (0.65-0.77)	3.41E-15
MMP-12	LAA	MR Egger	6	-0.38	0.13	0.68 (0.54-0.88)	3.92E-02
MMP-12	LAA	WM	6	-0.34	0.05	0.71 (0.64-0.79)	5.52E-10
MMP-12	LAA	PWM	6	-0.34	0.05	0.71 (0.64-0.79)	3.30E-10
MMP-12	LAA	WMode	6	-0.33	0.06	0.72 (0.64-0.81)	2.83E-03
MMP-12	CE	IVW	8	-0.04	0.04	0.96 (0.89-1.03)	2.72E-01
MMP-12	CE	MR Egger	8	-0.22	0.07	0.81 (0.7-0.93)	2.43E-02
MMP-12	CE	WM	8	-0.04	0.04	0.96 (0.89-1.04)	3.14E-01
MMP-12	CE	PWM	8	-0.05	0.04	0.95 (0.88-1.04)	2.57E-01
MMP-12	CE	WMode	8	-0.02	0.05	0.98 (0.89-1.08)	7.09E-01
MMP-12	SVO	IVW	8	-0.07	0.04	0.93 (0.86-1.00)	4.15E-02
MMP-12	SVO	MR Egger	8	-0.15	0.09	0.86 (0.73-1.02)	1.38E-01
MMP-12	SVO	WM	8	-0.08	0.05	0.92 (0.84-1.01)	8.83E-02
MMP-12	SVO	PWM	8	-0.08	0.05	0.92 (0.84-1.01)	7.85E-02
MMP-12	SVO	WMode	8	-0.08	0.06	0.92 (0.82-1.03)	2.09E-01
MMP-12	mRS	IVW	7	0.04	0.04	1.04 (0.96-1.12)	3.27E-01
MMP-12	mRS	MR Egger	7	0.04	0.10	1.04 (0.86-1.25)	7.28E-01
MMP-12	mRS	WM	7	0.03	0.05	1.03 (0.94-1.13)	5.39E-01
MMP-12	mRS	PWM	7	0.03	0.05	1.03 (0.94-1.13)	5.30E-01
MMP-12	mRS	WMode	7	0.03	0.05	1.03 (0.93-1.14)	6.03E-01

Online Table IV. MR results of all analyses and methods. LAA: Large-Artery Atherosclerosis, SVO: Small-Vessel Occlusion, CE: Cardioembolism. IS: Ischemic Stroke, IVW: Inverse-Variance Weighted, WM: Weighted Median, PWM: Penalized Weighted Median, WMode: Weighted Mode, OR: Odds ratio, no SNV: Number of SNVs.

		Pleiotropy			Heterogeneity	
Exposure	**Outcome**	**Intercept**	**Standard Error**	**P-value**	**Q**	**P-value**
MMP-1	IS	0.014	0.01	0.12	51.61	3.72E-01
MMP-1	LAA	0.019	0.02	0.39	48.02	5.13E-01
MMP-1	CE	-0.011	0.02	0.52	45.95	5.98E-01
MMP-1	SVO	0.033	0.02	0.11	45.04	6.34E-01
MMP-1	mRS	-0.031	0.02	0.16	32.83	9.11E-01
MMP-8	IS	-0.004	0.02	0.85	3.02	6.96E-01
MMP-8	LAA	-0.012	0.05	0.83	1.96	8.54E-01

MMP-8	CE	0.126	0.05	0.07	17.53	3.59E-03
MMP-8	SVO	-0.061	0.06	0.36	6.91	2.27E-01
MMP-8	mRS	0.012	0.05	0.81	2.53	7.72E-01
MMP-12*	IS	0.040	0.01	0.01	13.11	1.63E-03
MMP-12	IS	0.031	0.01	0.05	12.33	5.50E-02
MMP-12*	LAA	0.088	0.04	0.08	29.67	1.04E-04
MMP-12	LAA	0.012	0.04	0.78	5.68	3.38E-01
MMP-12	CE	0.051	0.02	0.04	11.54	1.17E-01
MMP-12	SVO	0.022	0.02	0.39	5.20	6.36E-01
MMP-12	mRS	0.001	0.03	0.97	3.47	7.47E-01

Online Table V. Pleiotropy and heterogeneity tests of all Mendelian randomization analyses. IS: Ischemic Stroke, LAA: Large-Artery Atherosclerosis, CE: Cardioembolism, SVO: Small Vessel Occlusion, mRS: modified Rankin Scale. *Prior removal of significant outliers detected with MR-PRESSO outlier test.

SNV	RSSobs	P-value
rs1892971	6.39E-04	0.60
rs1942524	1.83E-04	1.00
rs2186789	1.99E-04	1.00
rs484915	1.23E-03	**0.01**
rs499459	2.08E-03	0.25
rs613804	2.11E-04	1.00
rs650108	5.68E-04	0.49
rs671188	8.61E-06	1.00

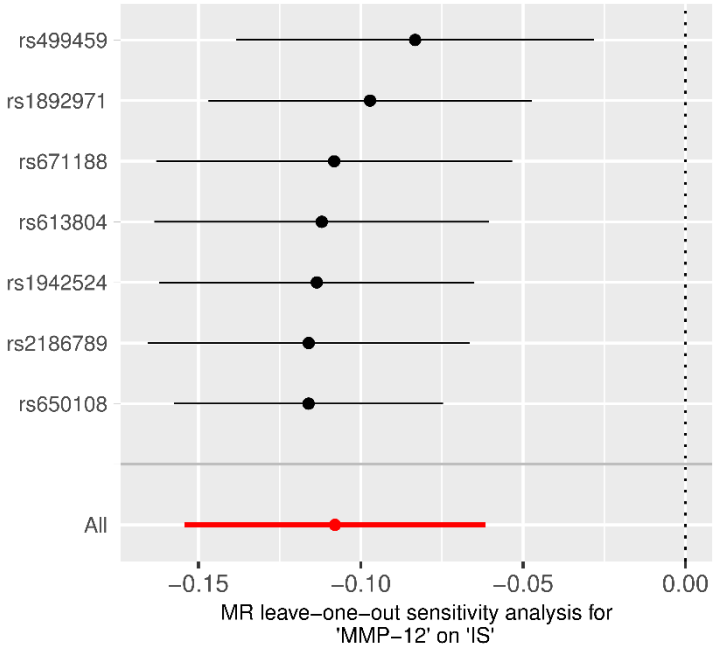
Online Table VI. MR-PRESSO Outlier test for MMP-12 and IS. RSSobs: Observed Residual Sum of Squares, SNV: Single Nucleotide Variant.

SNV	RSSobs	P-value
rs1892971	8.68E-03	0.06
rs1942524	1.32E-05	1.00
rs2186789	4.83E-04	1.00
rs484915	1.00E-02	**0.01**
rs499459	2.97E-03	1.00
rs613804	5.79E-06	1.00
rs650108	6.83E-03	**0.04**
rs671188	1.98E-04	1.00

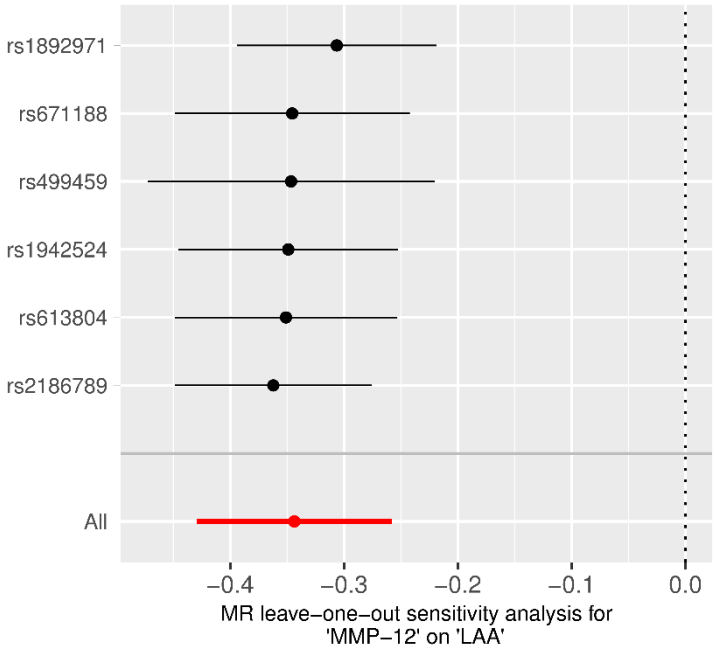
Online Table VII. MR-PRESSO Outlier test for MMP-12 and LAA. RSSobs: Observed Residual Sum of Squares, SNV: Single Nucleotide Variant.

3. Online Figures

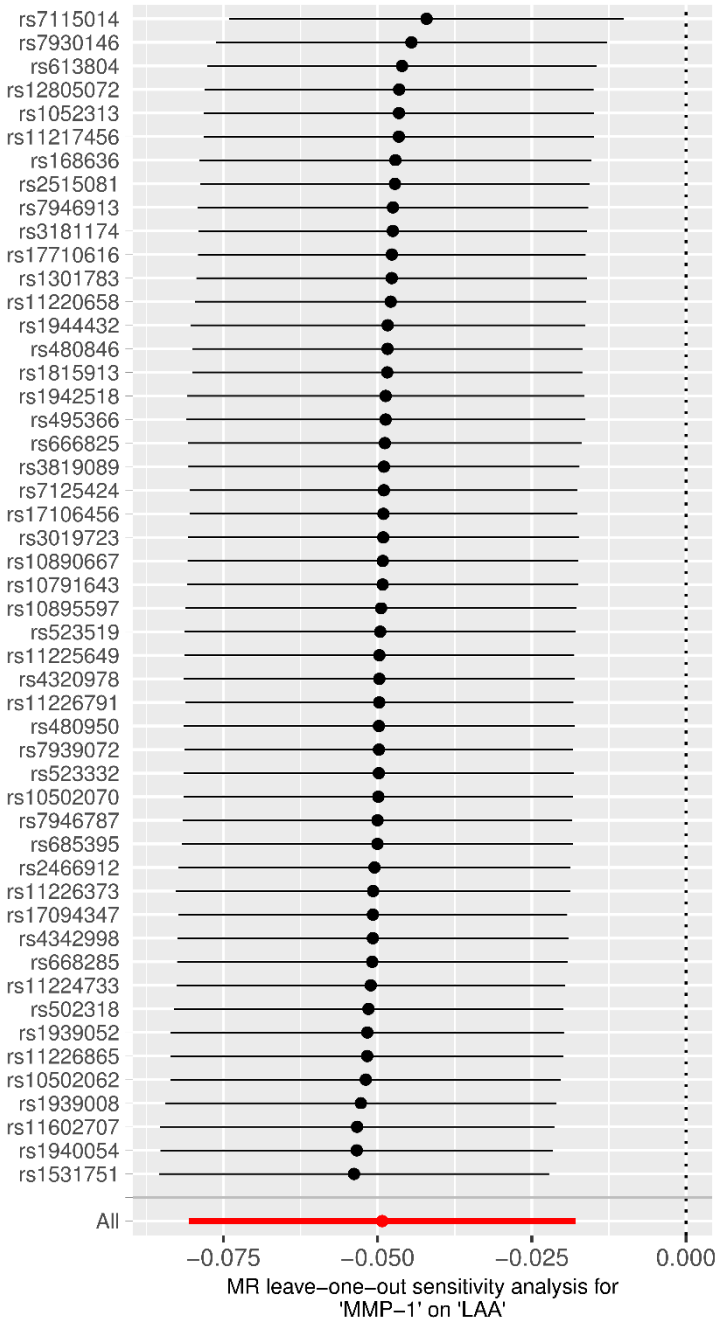
Online Figure I. Leave-one-out analysis for MMP-12 on all ischemic strokes (IS).



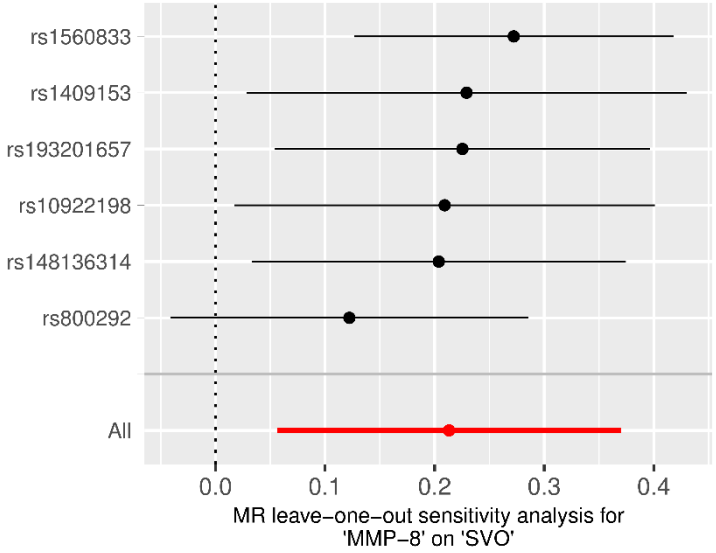
Online Figure II. Leave-one-out analysis for MMP-12 on Large-Artery Atherosclerosis (LAA) stroke.



Online Figure III. Leave-one-out analysis for MMP-1 on Large-Artery Atherosclerosis (LAA) stroke.



Online Figure IV. Leave-one-out analysis for MMP-8 on Small-Vessel Occlusion (SVO) stroke.



4. Supplementary Note

4.1 Description of study populations

For each cohort, all aspects of the studies were approved by the local institutional review board and ethics committee. All the participants included, or their approved representative, provided written informed consent for participation.

MEGASTROKE consortium GWAS

For the European ancestry analysis of the MEGASTROKE consortium, 16 different cohorts were analyzed, comprising 34,217 cases of ischemic stroke and 406,111 healthy controls. Stroke was defined according to the World Health Organization (WHO) definition, i.e. rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer, or leading to death, with no apparent cause other than of vascular origin. Strokes were defined as ischemic stroke (IS) or intracerebral hemorrhage (ICH) based on clinical and imaging criteria. IS was further subdivided into the following categories, mostly using the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria: i) large vessel ischemic stroke; ii) cardioembolic ischemic stroke; and iii) small vessel ischemic stroke.

GODs GWAS

European ancestry patients with a diagnosis of IS according to World Health Organization criteria were selected from the Spanish Stroke Genetics Consortium (GeneStroke) and the International Stroke Genetics Consortium (ISGC).

Selection criteria

All participants met the following criteria: (1) European descent, aged >18 years, diagnosis of IS in the anterior vascular territory; (2) assessed by a neurologist during the acute phase of stroke; (3) initial stroke severity >4, according to the National Institutes of Health Stroke Scale (NIHSS); (4) information on post-stroke functional status at 3 months (or alternatively between 3 and 6 months); (5) evidence of acute IS in a neuroimaging study; (6) lack of concomitant pathology. Individuals with stroke recurrence during the follow-up period were excluded. Posterior vascular territory and lacunar strokes were also excluded, as these locations show a poor correlation between infarct size and clinical symptoms, and thus, functional outcome. In these cases, the recovery process could be masked by the random location effect and lead to imprecision in measuring the genetic contribution to the degree of recovery.

Stringent analysis was the analysis used in this study. They only included cohorts that fulfilled the above criteria and had complete information for the following variables: (1) 3-month post-stroke functional status; (2) NIHSS score at hospital discharge; and (3) previous functional independence (mRS<3) before the stroke.

MMP-1 levels GWAS

For the MMP-1 levels GWAS, they used the Heredity and Phenotype Intervention (HAPI) Heart Study that was initiated in 2002. Participants in this study comprised adults from the Old Order Amish community of Lancaster County, PA, USA, who were recruited over a three-year period. Study participants were aged 20 years and older and relatively healthy based on a variety of exclusion criteria, including severe hypertension (blood pressure > 180/105 mm Hg), among others. The study aims, recruitment procedures and ascertainment criteria have been described previously[17]. The study included 868 participants, 792 of whom had available DNA and serum MMP-1 measurements. Fourteen additional individuals were excluded from the final analysis due to genotyping issues, leaving 778 subjects[12].

MMP-8 levels GWAS

For this study, the Corogene and FINRISK 1997 cohorts were used:

The Corogene study included 5295 Finnish patients who were assigned for coronary angiography in the Helsinki University Central Hospital, Finland, between 2006 and 2008[18]. Blood samples were drawn from the arterial line during the coronary angiography into serum and citrate plasma vacuum tubes. The samples were handled according to the laboratory standards of Helsinki University Central Hospital (accredited laboratory).

FINRISK is a Finnish national population-based study[19] that has been conducted at 5-year intervals since 1972. In 1992, 1997, 2002 and 2007, the surveys were performed in 5 geographical areas of Finland, and they included a clinical examination, questionnaire and laboratory analyses. Blood samples were collected during the health examination. Serum samples were taken without coagulation activators. Cases with prevalent coronary heart disease (CHD) or CVD at baseline were identified using (1) the questionnaire as a doctor-diagnosed disease, (2) the disease-associated drug reimbursement records from the Social Insurance Institution of Finland, including purchased medications and entitled reimbursements, and (3) the National Hospital Discharge register for hospitalizations. The study includes up to 20 years of follow-up data. The incident CHD events (acute myocardial infarction [AMI], bypass surgery and angioplasty), CVD events (CHD events and stroke) and all-cause deaths during follow-up were identified using (1) the drug reimbursement records from the Social Insurance Institution of Finland, (2) the National Hospital Discharge register for hospitalizations, and (3) the National Causes-of-Death Register.

In total, genotype data and serum concentrations of MMP-8 were available for 2203 subjects in the Corogene cohort and 3846 subjects in the FINRISK 1997 cohort.

MMP-12 levels GWAS

For this study, they used the IMPROVE cohort, which consists of a multicenter, longitudinal, observational study that recruited 3711 individuals, aged 54-79 years, with at least three cardiovascular risk factors but who were asymptomatic for cardiovascular disease. Subjects were considered to be exposed to a vascular risk factor when one of the following criteria was met: male sex or at least 5 years after menopause for women; hypercholesterolemia (mean calculated LDL-C blood levels > 160 mg/dL or treatment with lipid-lowering drugs); hypertriglyceridemia (triglycerides levels > 200 mg/dL after diet or treatment with triglyceride-lowering drugs); hypoalphalipoproteinemia (HDL-C < 40 mg/dL); hypertension (diastolic blood pressure > 90 mmHg and/or systolic blood pressure >140 mmHg or treatment with anti-hypertensive drugs); diabetes or impaired fasting glucose (blood glucose level > 110 mg/dL or treatment with insulin or oral hypoglycemic drugs); smoking habit (at least 10 cigarettes/day for at least thirty months); family history of cardiovascular disease. Exclusion criteria were: age under 54 or over 79 years; abnormal anatomical configuration of neck and muscles, marked tortuosity and/or depth of the carotid vessels and/or uncommon location of arterial branches; personal history of myocardial infarction (MI), angina pectoris, stroke, transient ischemic attack, aortic aneurysm, intermittent claudication, surgical revascularization in carotid, coronary or peripheral arterial territories, congestive heart failure (III-IV NYHA Class); and history of serious medical conditions that might limit longevity (e.g. cancer)[14].

4.2 Evaluation of metalloproteinase levels

The protocols of the three MMPs studied were:

Serum levels of MMP-1, measured in the fasting state, were determined by an enzyme-linked immunosorbent assay (ELISA) (R&D Systems, Minneapolis, MN, USA) in the University of Maryland Cytokine Core Laboratory in Baltimore, Maryland, USA. MMP-1 levels were measured in duplicate, and the mean values of duplicates were used for data analysis. Because study participants were recruited over a three-year period, MMP-1 levels were measured in five different batches throughout the study period, and the batch effect was included in the regression model using dummy variables. The detection range of MMP-1 values was between 0.16 and 10 ng/ml, and the intra-assay coefficient of variation was 7.5%. MMP-1 values that were above (n = 33) and below (n = 2) the detection range were assigned the maximum and minimum values of the detection range, respectively.

MMP-8 was measured from serum samples of the Corogene and FINRISK 1997 subjects with a time-resolved immunofluorometric assay (Medix Biochemica, Kauniainen, Finland) according to the manufacturer's instructions. The MMP-8 concentrations were log-transformed before analysis because they were not normally distributed.

Plasma concentrations of MMP12 were measured at baseline using the Olink ProSeek CVD I array (Olink Proteomics, Uppsala, Sweden), according to the standard protocol[20]. The interplate coefficient of variation for the MMP12 assay was 14%, as estimated using a pooled plasma control across all plates. The measurements were carried out in 3394 IMPROVE participants in whom genotype information post-quality control (QC) was also present. Plasma MMP12 activity was measured in duplicates in 20 plasma samples from the IMPROVE cohort. Ten samples were randomly chosen from the 75th to 95th percentile of plasma MMP12 concentration and ten from the 5th to 25th percentile. Activity was measured with the Fluorimetric Sensolyte 520 MMP12 Assay Kit (Anaspec), which specifcally detects elastin degradation products (desmosine).