Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München
aus dem
Institut für Schlaganfall- und Demenzforschung
des Klinikums der Universität München

From Phenotype to Function via Mass Spec-Based Proteomics: An LC-MS/MS DNA-Protein Pulldown Approach Applied to Functional Stroke Genetics

submitted by
Manuel Lehm
30th April 2019

Supervisor
Prof. Dr. med. Martin Dichgans
Institut für Schlaganfall- und Demenzforschung (ISD)
Klinikum der Universität München

First Reviewer:	Prof. Dr. med. Martin Dichgans
Second Reviewer:	Prof. Dr. rer. nat. Matthias Mann
Third Reviewer:	Prof. Dr. med. Tim Magnus
Date of Submission:	30th April 2019
Date of Defense:	4th November 2019

To son and wife.

ABBREVIATIONS

CAD	Coronary Artery Disease
CCS	Causative Classification of Stroke
DHS	DNase I hypersensitive site
ECG	Electrocardiogram
ENCODE	Encyclopedia of DNA Elements
ESUS	Embolic Stroke of Undetermined Source
GWAS	Genome-Wide Association Study
HDAC	Histone Deacetylase
HPLC	High Performance Liquid Chromatography
IP	Immunoprecipitation
IP-MS	Immunoprecipitation Followed by Mass Spectrometric Analysis
IS	Ischemic Stroke
LAS	Large-Artery Atherosclerotic Stroke
LC-MS/MS	Liquid Chromatography Coupled to Tandem Mass Spectrometry
MI	Myocardial Infarction
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MT	Mechanical Thrombectomy
NINDS	National Institute of Neurological Disorders and Stroke
PWAS	Proteome-Wide Analysis of Disease-Associated SNPs
rt-PA	Recombinant Tissue-Type Plasminogen Activator
SiGN	Stroke Genetics Network (part of the NINDS)
SILAC	Stable Isotope Labeling by Amino Acids in Cell Culture
SNP	Single Nucleotide Polymorphism
TOAST	Trial of Org 10172 in Acute Stroke Treatment

TABLE OF CONTENTS

1 SUMMARY

Stroke is one of the leading causes of morbidity, disability and mortality in humans worldwide. Accurate classification of etiologic stroke subtypes is of utmost importance for adequate primary and secondary prevention as well as for sufficiently powered genetical studies. In contrast, clinical phenotypization of ischemic stroke is often limited due to evidence of concurrent etiologies. For large-artery atherosclerotic stroke, genome-wide association studies have recently identified *HDAC9* as a strong risk locus. However, as in the case of the *HDAC9* locus, most common variants associated with disease risk locate to non-coding regions of the genome, making it notoriously hard to determine the causative molecular mechanisms. Our objective was to (1) test the reliability of established ischemic stroke classification systems in the setting of a research consortium, (2) identify novel risk loci for ischemic stroke and its distinct subtypes, and (3) elucidate the molecular mechanism mediating ischemic stroke risk at the *HDAC9* locus.

First, applying the ischemic stroke classification systems CCS and TOAST to a research cohort of more than 16.000 patients from the US and Europe, we found higher interrater reliability for CCS (κ = 0.72) than with the traditional TOAST classification system. Second, classifying ischemic stroke based on CCS and TOAST, we performed a two-stage genome-wide association study with more than 37.000 patients and almost 400.000 controls. We identified a novel locus near *TSPAN2* significantly associating with large-artery atherosclerotic stroke and replicated the 4 previously identified loci *PITX2*, *ZFHX3*, *ALDH2*, and *HDAC9* with subtype-specific associations. Third, utilizing a modern DNA-protein pulldown approach for high resolution mass spectrometry-based proteomics, we identified preferential binding of an E2F3-TFDP1-Rb1 complex to the common allele of rs2107595. Additional functional follow-up studies imply allele-specific regulation of HDAC9 expression via E2F3 and Rb1 as the molecular mechanism mediating disease risk at the *HDAC9* locus.

Collectively, we demonstrate improvements in ischemic stroke phenotypization, identification of a new risk locus for ischemic stroke and application of mass spectrometry-based proteomics to functional genetics to uncover the molecular mechanism mediating stroke risk. This approach is virtually applicable to any complex disease and may aid in the functional follow-up of non-coding variants.

2 INTRODUCTION

2.1 Fundamentals of Stroke

Sudden onset neurological deficits such as facial palsy, arm weakness and/or impaired speech are typical symptoms in patients suffering a stroke.[1] While the causative disruption of cerebral blood flow may be due to intracranial hemorrhage or blockage of cerebral blood vessels, the vast majority of strokes (approximately 72 %)[2] are based on the latter with ischemia subsequently leading to infarction of brain parenchyma.

Epidemiology and Risk Factors of Ischemic Stroke

Ischemic stroke (IS) is one of the most frequent causes of morbidity and mortality worldwide, in developed countries also being the most frequent cause of disability and care dependency in the elderly.[2,3] In 2016, the global lifetime risk of IS was approximately 18 %,[2] with two thirds of IS usually occurring above the age of 65.

Besides age several other risk factors for IS are well established, such as smoking, atrial fibrillation, hypertension, hypercholesterolemia, obesity and diabetes.[4] Most of these risk factors are shared with other cardiovascular diseases like coronary artery disease (CAD)[5] or myocardial infarction (MI),[6] and are accountable for about 60 % of IS lifetime risk. The remaining IS risk is largely attributed to specific genetic heritability[7] (see section 2.2).

Pathophysiology and Treatment of Ischemic Stroke

In case of IS cerebral blood flow is impaired via narrowing or complete occlusion of cerebral arteries. The brain parenchyma itself can be described as a “metabolic powerhouse”, as evidenced by its high extraction fraction of blood oxygen and glucose.[8] In case of ischemia failure of Na^+/K^+-ATPase results in rapid breakdown of neuronal membrane potential and loss of function of neurotransmitter reuptake may additionally cause glutamatergic excitotoxicity.[9,10] This low tolerance to ischemia results in high vulnerability of brain parenchyma, with neuronal damage becoming irreversible within minutes after onset of severe ischemia.[11]

Therefore, treatment of acute IS aims to restore cerebral blood flow as quickly as possible in order to minimize the extent of infarction of brain parenchyma. Systemic thrombolysis via intravenous administration of recombinant tissue-type plasminogen activator (rt-PA)[12] within

4.5 hours after symptom onset increases the chance of a good neurological outcome with an odds ratio of 1.34.[13] However, in cases of large vessel occlusions with high thrombus burden, systemic thrombolysis via rt-PA on its own only has a minor chance of vessel recanalization.[14] For large vessel occlusions of the anterior circulation, endovascular therapy with mechanical thrombectomy (MT) is the therapy of choice,[15] within 6 hours after symptom onset MT dramatically improves chances of good neurological outcome with an odds ratio of 2.49,[16-23] and MT was shown to be effective up to 24 hours after symptom onset in select cases.[24,25]
Secondary prevention of IS includes potential lifestyle changes like physical exercise and quitting smoking as well as medication like statins, antiplatelet agents or anticoagulation.[4] The choice of medication for secondary prevention is highly dependent on the patient's IS etiology.

Etiology and Classification of Ischemic Stroke

Both for secondary prevention as well as for stroke research it is important to determine an individual patient's IS etiology, for which several classification systems have been developed. The most widespread classification system of IS in clinical routine is the Trial of Org 10172 in Acute Stroke Treatment (TOAST),[26] recognizing the following 5 distinct stroke etiologies: (1) large-artery atherosclerosis, (2) cardioembolism, (3) small-vessel occlusion, (4) other determined etiologies such as arterial dissection, and (5) undetermined.
However, by applying the TOAST criteria about 30 % of stroke etiologies still remain undetermined (category 5).[27] One focus of the present work was to improve the phenotypization of IS by assessing the discriminatory power of a new IS classification system, the Causative Classification of Stroke (CCS)[28,29] in comparison to the established TOAST system (see section 3). Since IS per se is a complex phenotype with multiple possible etiologies, improving its causative phenotypization also has direct implications for stroke research such as genetics.

2.2 Genetics of Ischemic Stroke

Genetic Heritability of Ischemic Stroke

Apart from the aforementioned risk factors, about 40 % of IS lifetime risk is accounted to genetical risk factors, as evidenced by the fact that prior IS among first-degree relatives constitutes an increase in IS risk of up to 30 % when compared to the general population.[30] Generally, heritability is usually conferred via distinct genetical mechanisms:[31] (1) non-

synonymous single gene mutations leading to mendelian diseases, (2) rare variants with moderate to high effect size, and/or (3) common variants of so-called single nucleotide polymorphisms (SNPs). The most common monogenic stroke syndrome due to single gene mutations is cerebral autosomal dominant arteriopathy with subcortical infarcts and leukencephalopathy (CADASIL).[32] However, the vast majority of IS incidents in the general population are caused by the combination of environmental factors, lifestyle and common variants, i.e. SNPs, with a low to moderate effect size.[7] Method of choice for dissecting the genetical component of such complex diseases are genome-wide association studies (GWAS),[33] where typically several thousands of SNPs across the whole genome are tested for their association with a certain phenotype,[34] e.g. IS. With DNA sample numbers in local stroke cohorts usually limited to the range of several thousands, identification of risk loci with genome-wide significance does require a very clean phenotype.[35,36] Hence, another focus of the present work was the identification of novel risk loci for IS via performing a GWAS based on the CCS classification[28,29] (see section 4).

Risk Loci for Ischemic Stroke

The first risk locus for IS to be discovered via a GWAS was 4q25, with two SNPs close to *PITX2* significantly associating with IS in general and most strongly with cardioembolic stroke.[37] Meanwhile, as a result of two large scale GWAS meta-analyses of several international stroke cohorts (combined to MEGASTROKE and International Stroke Genetics Consortium) both with European and non-European ancestry in more than 70.000 stroke patients and 800.000 controls,[38,39] a total of 35 risk loci for stroke were identified. Out of these 35 risk loci, 20 reached genome-wide significance for IS. These IS risk loci predominantly showed stroke subtype specificity, with all known risk loci only reaching genome-wide significance for one specific IS subtype.[38,40] For e.g. large-artery atherosclerotic stroke (LAS), the following risk loci reached genome-wide significance: *TSPAN2*, *TM4SF4-TM4SF1*, *EDNRA*, *LINC01492*, *MMP12* and *HDAC9.* Since LAS and other cardiovascular traits such as CAD and MI share a lot of genetic heritability through their underlying pathophysiology of atherosclerosis,[5,6] LAS risk loci currently are of particular biomedical interest.

HDAC9 and Functional Genetics

The first risk locus to reach genome-wide significance for LAS was *HDAC9*, as identified in three independent European samples via the framework of the International Stroke Genetics Consortium and the Wellcome Trust Case Control Consortium 2.[41] Besides LAS this risk locus is also strongly associated with CAD[5,6] as well as peripheral artery disease and increased intima-media-thickness.[42,43] With a p-value of 3.65×10^{-15} and an odds ratio of 1.21 the initially published lead SNP rs2107595 reached the highest level of confidence for the *HDAC9* locus in the MEGASTROKE meta-analysis as well.[38,41] The genomic location of rs2107595 is just 7.5 kb 3' of the *HDAC9* gene.

HDAC9 is a member of the histone deacetylases family and critically involved in gene regulation and transcriptional activity, both via deacetylation of histones as well as direct interaction with transcription factors.[44-46] HDAC9 was found to be overexpressed in human atherosclerotic vessel lesions,[42] whereas HDAC9 deficiency led to significant reduction of atherosclerotic lesion load in a mouse model of atherosclerosis.[47,48] Furthermore, HDAC9 was implicated in inflammatory processes via T-cell homeostasis.[49-51] Taken together, HDAC9 seems a plausible candidate for mediating the risk effect of the *HDAC9* locus with respect to LAS and CAD.

rs2107595, like most common variants in the human genome,[52,53] is located in an intergenic region and therefore might most likely impact on HDAC9 activity levels via gene regulatory mechanisms, i.e. via disrupting *cis*-regulatory elements such as promoter, enhancer or suppressor sequences,[54] resulting in the allele-specific binding of transcription factors. Indeed, rs2107595 was found to match both with a DNase I hypersensitive site (DHS) as well as enhancer histone marks H3K4me1 and H3K27ac (ENCODE[55,56]), indicating colocalization with a *cis*-regulatory element.[57-61]

Ultimately, a major focus of the present work was to apply a proteomics approach (see section 2.3) to the lead SNP rs2107595 in order to detect allele-specific binding of transcription factors and thus enabling identification of the causative molecular mechanism by which the *HDAC9* locus mediates LAS risk (see section 5).

2.3 Mass Spectrometry-Based Proteomics

Principles of Mass Spectrometry-Based Proteomics

Proteomics, the large-scale identification and quantification of proteins from a complex sample, started out via 2D separation of proteins via SDS-PAGE and subsequent analysis of individually excised gel bands.[62] Modern proteomics, however, is predominantly based on recently developed high-resolution quantitative mass spectrometry,[63,64] allowing for precise identification and quantification of thousands of proteins as well as its interactions or posttranslational modifications.[65-67]

The proteomics method of choice for high coverage of complex biological samples is the "bottom-up" approach:[68] during sample preparation the biochemical isolation of the protein fraction of choice is performed and may or may not include further steps of fractionation and/or affinity purification.[69,70] Next, the entire proteomic sample is digested into peptides with the addition of proteases such as trypsin and lysyl endopeptidase.[71] After subsequent purification steps the sample is then subjected to high performance liquid chromatography (HPLC), where the highly complex peptide mixture is separated and coupled to a mass spectrometer (LC-MS) via electrospray ionization of the eluting peptides.[72,73] For tandem mass spectrometry (LC-MS/MS), the ionized peptides' mass is analyzed, followed by fragmentation and subsequent analysis of the fragment ions.[74-76] From this data the peptide sequences as well as protein identities and abundances can be inferred via searching against a human proteome sequence database.[77,78]

Interaction Proteomics for Functional Genetics

In the context of functional genetics mass spectrometry-based proteomics techniques are increasingly suitable for the identification of protein-protein and protein-DNA interactions, it is now possible to accurately identify and quantify e.g. proteins interacting with specific chromatin marks or the allele-specific binding of transcription factors.[79-82] Following immunoprecipitation (IP) of a certain protein or DNA bait of interest, the resultant proteomic sample is digested into peptides, purified and subjected to LC-MS/MS.[83]

In order to perform accurate quantifications of interacting proteins two main strategies are available: metabolic labeling e.g. via "stable isotope labeling by amino acids in cell culture"[84,85] or label-free methods such as MaxLFQ,[86] involving stringent bioinformatic normalization

strategies for deduction of protein intensities.[82] While both strategies have their advantages, one benefit of metabolic labeling is the robust exclusion of confounding background binders or contaminants.[87]

In case of metabolic labeling via SILAC, the cell culture-based input material is generated in two states, either with heavy-labeled or light-labeled amino acids.[84] These differentially labeled input materials are then subjected to two sets of IP each, once with the candidate bait being assigned the heavy-labeled input, once with the control bait being assigned the heavy-labeled input. Due to label-switching between the two sets of IP, a typical SILAC experiment will generate two different heavy-over-light ratios for identified proteins, with specific interactors showing inverse ratios between both sets of IPs.[88,89]

Proteome-Wide Analysis of Disease-Associated SNPs (PWAS)

Interaction proteomics can be used as an unbiased DNA-centric method for detecting DNA-protein interactions.[90] If applied to functional genetics, concatemerized synthetic DNA oligos containing the risk allele or its corresponding wildtype allele plus some of the SNPs flanking genomic sequence can be used as candidate and control bait, respectively. Using nuclear extracts which have been metabolically labeled via SILAC as proteomic input material, this DNA-protein pulldown followed by quantitative LC-MS/MS enables robust identification of differentially binding transcription factors in an allele-specific manner.[91]

This method was called proteome-wide analysis of disease-associated SNPs (PWAS), its first application in the context of clinical genetics was the identification of allele-specific binding factors for SNPs at the *IL2RA* locus, a risk locus associated with type 1 diabetes.[91] In the present study, this approach was specifically applied to rs2107595 for identification of the molecular mechanism conferring LAS risk at the *HDAC9* locus (see section 5).

3 MANUSCRIPT 1

“Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network.”

Ay, H., (…), **Lehm, M.**, (…), Meschia, J.

This manuscript has been peer-reviewed and published under the following reference:

Ay et al. (2014). Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network. *Stroke.* 2014;45:3589-3596; originally published online November 6, 2014; doi: https://doi.org/10.1161/STROKEAHA.114.007362.

Author contribution:

Recruitment of IS patients, classification of IS patients according to their probable etiology.

Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network

Hakan Ay, MD; Ethem Murat Arsava, MD; Gunnar Andsberg, MD; Thomas Benner, PhD; Robert D. Brown Jr, MD; Sherita N. Chapman, MD; John W. Cole, MD, MS; Hossein Delavaran, MD; Martin Dichgans, MD; Gunnar Engström, MD; Eva Giralt-Steinhauer, MD; Raji P. Grewal, MD; Katrina Gwinn, MD; Christina Jern, MD; Jordi Jimenez-Conde, MD; Katarina Jood, MD; Michael Katsnelson, MD; Brett Kissela, MD; Steven J. Kittner, MD; Dawn O. Kleindorfer, MD; Daniel L. Labovitz, MD; Silvia Lanfranconi, MD; Jin-Moo Lee, MD; Manuel Lehm, BSc; Robin Lemmens, MD; Chris Levi, MD; Linxin Li, PhD; Arne Lindgren, MD; Hugh S. Markus, DM; Patrick F. McArdle, PhD; Olle Melander, MD; Bo Norrving, MD; Leema Reddy Peddareddygari, MD; Annie Pedersén, MD; Joanna Pera, MD; Kristiina Rannikmäe, MD; Kathryn M. Rexrode, MD; David Rhodes, MPH; Stephen S. Rich, PhD; Jaume Roquer, MD, PhD; Jonathan Rosand, MD, MSc; Peter M. Rothwell, MD; Tatjana Rundek, MD, PhD; Ralph L. Sacco, MD, MS; Reinhold Schmidt, MD; Markus Schürks, MD; Stephan Seiler, MD; Pankaj Sharma, MD; Agnieszka Slowik, MD; Cathie Sudlow, MD; Vincent Thijs, MD; Rebecca Woodfield, MD; Bradford B. Worrall, MD, MSc*; James F. Meschia, MD*

Background and Purpose—NINDS (National Institute of Neurological Disorders and Stroke)-SiGN (Stroke Genetics Network) is an international consortium of ischemic stroke studies that aims to generate high-quality phenotype data to identify the genetic basis of pathogenic stroke subtypes. This analysis characterizes the etiopathogenetic basis of ischemic stroke and reliability of stroke classification in the consortium.

Received September 5, 2014; final revision received October 2, 2014; accepted October 3, 2014.

From the Department of Radiology, AA Martinos Center for Biomedical Imaging (H.A., E.M.A., T.B.), Stroke Service, Department of Neurology (H.A., J.R.), and Center for Human Genetic Research (J.R.), Massachusetts General Hospital, Harvard Medical School, Boston; Department of Neurology and Rehabilitation Medicine, Skåne University Hospital, Lund, Sweden (G.A., H.D., A.L., O.M., B.N.); Department of Neurology, Mayo Clinic Rochester, MN (R.D.B.); Department of Neurology (S.N.C., B.B.W.), Center for Public Health Genomics (S.S.R.), and Department of Public Health Sciences (B.B.W.), University of Virginia, Charlottesville; Department of Neurology (J.W.C., S.J.K.) and Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine (P.F.M.), University of Maryland School of Medicine, Baltimore; Institute for Stroke and Dementia Research (ISD), Klinikum der Universität München, Ludwig-Maximilians-University, München, Germany (M.D., M.L.); Department of Clinical Science, Lund University, Malmö, Sweden (G.E., B.N.); Department of Neurology, Neurovascular Research Group, IMIM-Hospital del Mar, Universitat Autonoma de Barcelona/DCEXS-Universitat Pompeu Fabra, Spain (E.G.-S., J.J.-C., J.R.); Neuroscience Institute, Saint Francis Medical Center, Trenton, NJ (R.P.G., L.R.P.); National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD (K.G.); Institute of Biomedicine (C.J., A.P.) and Institute of Neuroscience and Physiology (K.J.), Sahlgrenska Academy at University of Gothenburg, Sweden; Department of Neurology, University of Miami Miller School of Medicine, FL (M.K., T.R., R.L.S.); Department of Neurology, University of Cincinnati College of Medicine, OH (B.K., D.O.K.); Department of Neurology, Stern Stroke Center, Albert Einstein College of Medicine, Bronx, NY (D.L.L.); Department of Neuroscience and Sensory Organs, Policlinico Hospital Foundation IRCCS Cà Granda, Milan, Italy (S.L.); Department of Neurology, Washington University, St. Louis, MO (J.-M.L.); Department of Neurology, University Hospitals Leuven, Belgium (R.L., V.T.); Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), KU Leuven-University of Leuven, Belgium (R.L., V.T.); VIB, Vesalius Research Center, Laboratory of Neurobiology, Leuven, Belgium (R.L., V.T.); Department of Neurology, John Hunter Hospital, The University of Newcastle, New South Wales, Australia (C.L.); Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Oxford University, United Kingdom (L.L., P.M.R.); Department of Clinical Neurosciences, University of Cambridge, United Kingdom (H.S.M.); Department of Neurology, Jagiellonian University, Medical College, Krakow, Poland (J.P., A.S.); Centre for Clinical Brain Sciences, University of Edinburgh, United Kingdom (K.R., C.S., R.W.); Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA (K.M.R.); School of Public Health, University of Alabama, Birmingham (D.R.); Department of Neurology, Medical University Graz, Austria (R.S., S.S.); Department of Neurology, University Hospital Essen, Essen, Germany (M.S.); Cerebrovascular Research Unit, Department of Medicine, Imperial College London, United Kingdom (P.S.); and Department of Neurology, Mayo Clinic Jacksonville, FL (J.F.M.).

Guest Editor for this article was Anthony J. Furlan, MD.

*Drs Worrall and Meschia are joint co-senior authors and co-chairs of the SiGN Phenotype Committee.

The online-only Data Supplement is available with this article at http://stroke.ahajournals.org/lookup/suppl/doi:10.1161/STROKEAHA.114.007362/-/DC1.

Correspondence to Hakan Ay, MD, AA Martinos Center for Biomedical Imaging and Stroke Service, Departments of Neurology and Radiology, Massachusetts General Hospital, Harvard Medical School, 149 13th St, Room 2301, Charlestown, MA 02129. E-mail hay@mgh.harvard.edu

Stroke **is available at http://stroke.ahajournals.org** **DOI: 10.1161/STROKEAHA.114.007362**

Methods—Fifty-two trained and certified adjudicators determined both phenotypic (abnormal test findings categorized in major pathogenic groups without weighting toward the most likely cause) and causative ischemic stroke subtypes in 16 954 subjects with imaging-confirmed ischemic stroke from 12 US studies and 11 studies from 8 European countries using the web-based Causative Classification of Stroke System. Classification reliability was assessed with blinded readjudication of 1509 randomly selected cases.

Results—The distribution of pathogenic categories varied by study, age, sex, and race ($P<0.001$ for each). Overall, only 40% to 54% of cases with a given major ischemic stroke pathogenesis (phenotypic subtype) were classified into the same final causative category with high confidence. There was good agreement for both causative (κ 0.72; 95% confidence interval, 0.69–0.75) and phenotypic classifications (κ 0.73; 95% confidence interval, 0.70–0.75).

Conclusions—This study demonstrates that pathogenic subtypes can be determined with good reliability in studies that include investigators with different expertise and background, institutions with different stroke evaluation protocols and geographic location, and patient populations with different epidemiological characteristics. The discordance between phenotypic and causative stroke subtypes highlights the fact that the presence of an abnormality in a patient with stroke does not necessarily mean that it is the cause of stroke. **(*Stroke*. 2014;45:3589-3596.)**

Key Words: classification ■ pathogenesis ■ phenotype

Successful identification of genes that modify ischemic stroke risk relies on accurate delineation of pathogenic stroke phenotypes.[1–5] Determination of pathogenic stroke subtypes requires integration of several clinical, diagnostic, and imaging features and, therefore, is inherently subject to variability. Reproducible data on frequency of pathogenic stroke subtypes based on large multicenter data sets using well-defined and evidence-based criteria do not exist. Published studies on pathogenic stroke subtypes are largely constrained by poor to moderate reliability of the classification system,[6–8] suboptimal or uncertain diagnostic workup,[9,10] small sample size,[7,8,11] single center design,[8] and use of stringent selection criteria.[10,11]

This analysis sought to better understand the pathogenic basis of ischemic stroke. We prospectively identified pathogenic stroke subtypes using the rule- and evidence-based Causative Classification of Stroke (CCS) system within the context of the NINDS (National Institute of Neurological Disorders and Stroke)-Stroke Genetics Network (SiGN).[12–15] CCS automatically provides both phenotypic and causative stroke subtypes in each case. The former is a summary of positive test findings, whereas the latter requires integration of clinical, laboratory, and imaging stroke features and diagnostic test results to identify a single most likely causative subtype for each case. Hence, they provide different information. Here, we report distribution characteristics of various CCS-defined ischemic stroke subtypes and inter-rater reliability of pathogenic subtype assignments in the SiGN data set.

Methods

Contributing Studies and Patient Population

SiGN is a large international consortium of ischemic stroke studies aiming to generate high-quality phenotype data to assist in the identification of the genetic basis of ischemic stroke subtypes. This analysis included ischemic stroke cases from the initial 12 US and 11 European ischemic stroke studies in SiGN from 9 countries. Imaging confirmation of the absence of hemorrhagic stroke was required in each subject. Details about the individual contributing studies have been described previously in a separate publication.[15] Seventeen studies recruited consenting cases without using any selection criteria. In contrast, 6 studies were conducted in selected populations based on age, sex, and family history.[15] Recruitment to contributing studies occurred during a 23-year period between 1989 and 2012.

Stroke Subtyping

Pathogenic stroke classification in SiGN started in July 2010. The current study included 16 954 cases for whom pathogenic subtype information was available in the SiGN database as of March 2014. SiGN used the web-based CCS system for stroke subtyping (available at https://ccs.mgh.harvard.edu).[13] The details of CCS were published elsewhere.[13] For the purpose of SiGN, we customized CCS by generating a confidential, password-protected data collection platform. We also made a modification in the online CCS form by separating the single data entry field for small artery occlusion (SAO) in the original CCS into 2 separate data entry fields: one to indicate whether there is a typical lacunar infarct on neuroimaging and the second one to rule out whether there is an accompanying parent artery disease at the origin of the penetrating artery supplying the site of the lacunar infarct. Thus, it became possible to collect phenotypic data on lacunar infarcts for which vascular imaging for parent artery disease was not available. No modification was made in the decision-making code of the CCS; both customized and original CCS algorithms provided the same subtype for each given test condition.

We determined phenotypic subtypes in each subject.[13,14] Phenotypic subtypes referred to abnormal test findings categorized in major pathogenic groups without weighting toward the most likely cause in the presence of multiple causes.[14] There were 4 main phenotypic categories: large artery atherosclerosis (LAA), cardiac embolism (CE), lacunar infarction, and other uncommon causes. There were 4 possible states for LAA and CE (major, minor, absent, and incomplete evaluation), 3 for lacunar infarction (major, absent, and incomplete evaluation), and 2 for other uncommon causes (major and absent), giving rise to a total of 96 phenotypic categories. We collapsed these 96 categories into the following 7 subtypes: LAA-major, CE-major, lacunar infarction-major, other-major, no major pathogenesis, multiple competing major pathogeneses, and incomplete investigation. We further collapsed the last 3 categories into undetermined category and generated a 5-subtype phenotypic categorization.

We also recorded causative subtypes in each case. In contrast to phenotypic subtypes, causative subtyping requires integration of multiple aspects of ischemic stroke evaluation in a probabilistic and objective manner.[12,13] The causative subtype differs from the phenotypic subtype in certain occasions. For instance, in a patient with internal carotid artery stenosis, ipsilateral internal borderzone infarcts, and atrial fibrillation, the causative subtype is LAA, whereas the phenotypic subtype is multiple competing pathogeneses because of coexistence of LAA and CE. Major causative categories included LAA, CE, SAO, other uncommon causes, and undetermined causes. The undetermined group was further divided into 4 subcategories as

cryptogenic embolism, cryptogenic-other, incomplete evaluation, and multiple competing causes (unclassified). We grouped cardiac pathologies with uncertain risk of stroke (minor sources) into the undetermined cryptogenic-other category. This allowed us to generate a more refined cardioembolic category (CE-major). Each causative category in CCS (except for undetermined category) was subdivided based on the weight of available data as evident, probable, or possible to identify the level of confidence in assigning an pathogenesis.[12] Overall, CCS generated 17 causative subtypes.

CCS did not require a minimum level of investigations. In cases with missing tests, the system still assigned a subtype based on results of available tests but with a lower level of confidence. For instance, in a patient with typical lacunar infarct in the internal capsule and missing intracranial vascular imaging to rule out a parent artery disease, the level of confidence in attributing lacunar infarct to SAO was reduced from evident to possible. A subtype (both causative and phenotypic) was considered to be incomplete evaluation only when brain imaging, vascular imaging, or cardiac evaluation was not performed in the absence of an identified pathogenesis.

Data Adjudication and Quality Control

A total of 52 adjudicators (13 stroke neurologists, 17 stroke fellows, 13 neurology residents, and 9 non-neurologists) performed stroke subtyping. A centralized Phenotype Committee of 4 expert stroke neurologists met weekly to monitor data quality and site performance. The same committee blindly readjudicated a randomly selected 10% of cases recruited from the US studies for quality control. Similarly, 10% of cases from European studies were readjudicated by blinded European investigators (n=20). Each adjudicator and readjudicator had to complete an interactive online training module. The Phenotype Committee members provided training to adjudicators/readjudicators on data entry, data submission, and archiving at scheduled study meetings and via webinars. Every investigator was required to pass an online certification examination available at the CCS website.

Data Source

Study-specific case report forms and unabstracted medical records served as data source for subtyping. Readjudicators used the same data source available to adjudicators to determine the CCS subtypes. Data sources varied in length and detail among the study sites. Subtype assignments were done based on data available at the time of discharge in the majority. Prolonged ambulatory cardiac monitoring was obtained after discharge in 14% of the subjects. In such cases, we used postdischarge cardiac monitoring findings for stroke subtyping. All data entered into CCS and the system output were saved in a confidential SiGN database. In addition to subtype-related data, each study site provided baseline variables such as age, sex, race, and vascular risk factors, using a structured data collection form.

Statistics

Our primary objective was to determine the distribution of CCS subtypes within the SiGN cohort. We also determined pathogenic subtype distribution in a subset with complete diagnostic investigation. We defined complete investigation as the presence of brain imaging, intracranial and extracranial vascular imaging, and cardiac evaluation with echocardiography if ECG and clinical assessment did not reveal a source. We assessed the heterogeneity among centers in utilization of diagnostic tests using the χ^2 test. We used χ^2 test and Student t test to evaluate differences between cohorts with and without complete investigation for categorical and continuous variables, respectively. We assessed the correlation between causative and phenotypic subtypes by calculating how often CCS classified a given major abnormal evaluation finding (phenotypic subtype) as the causative stroke mechanism (causative subtype). We performed multinomial logistic regression to evaluate associations between causative CCS subtypes and age, sex, and race. In regression models, undetermined category served as the reference. We assessed the concordance between paired ratings by adjudicators and readjudicators by calculating crude agreement rates and unweighted κ values for both 5-subtype causative and phenotypic classification.[16] We expressed associations as odds ratios and 95% confidence intervals (CIs). We considered P values <0.05 as statically significant.

Results

Study Cohort

Table 1 presents characteristics of the study population. Complete diagnostic investigation was available in 46%. Cases with complete investigation were similar to those with incomplete investigation except that they were slightly younger and more likely to be men (P<0.001; Table 1). The proportion of cases with complete investigation varied across the 23 studies (P<0.001; Table 2). US studies had higher complete investigation rate as compared with European centers (53% versus 40%; P<0.001).

Stroke Subtypes

Figure 1 shows the distribution of phenotypic and causative subtypes. Compared with the overall population, subtype distribution differed in the cohorts with complete investigation (Figure 1; P<0.001) and after exclusion of the 6 studies that used selection rules (Figure I in the online-only Data Supplement; P<0.001).

Vascular investigations revealed an atherosclerotic lesion causing ≥50% stenosis (LAA-major phenotype) in 3392 of the 16954 cases (20%); among these, 2093 (62%) had extracranial stenosis, 962 (28%) had intracranial stenosis, and 337 (10%) had both extra- and intracranial stenoses. LAA-major was an isolated finding in 2536 (75%); in the remaining 856 (25%), there was another major pathogenesis such as a major cardioembolic source. Diagnostic tests for other pathogeneses were missing in 972 (29%). Overall, 1719 (51%) cases with a major LAA had either a missing test or another competing pathogenesis. The final causative subtype was LAA-evident in only 1815 (54%) cases (Figure 2A). The remaining individuals were either classified into the category of LAA but with a

Table 1. Patient Characteristics

	Overall Study Population (n=16954)	Complete Investigation (n=7748)	Incomplete Investigation (n=9206)
Age (mean±SD), y	67.1 (14.9)	64.7 (15.7)	69.1 (13.9)
Female (%)	48.8	44.5	52.3
Race (%)			
Black	9.7	11.0	8.7
White	79.3	77.5	80.7
Other	11.0	11.5	10.6
Hypertension (%)	67.8	66.0	69.3
Diabetes mellitus (%)	25.0	25.7	24.4
Atrial fibrillation (%)	21.4	23.2	19.9
Coronary artery disease (%)	22.9	21.3	24.3
Current smoking (%)	24.1	24.4	23.8

Complete investigation is defined as the presence of brain imaging, cardiac evaluation with electrocardiography, echocardiography if other investigations did not reveal a source, and intracranial and extracranial vascular evaluation.

Table 2. Complete Investigation Rates Across the Contributing Studies

Study	No. of Cases	Complete Cardiac Investigation, %	Complete Vascular Investigation, %	Complete Cardiac and Vascular Investigation, %
1	684	40.9	20.2	4.8
2	578	89.6	75.3	70.4
3	840	71.8	39.4	30.1
4	1072	79.5	98.5	78.7
5	331	98.8	97.9	96.7
6	876	94.5	79.3	75.9
7	598	58.2	35.8	21.9
8	675	80.0	79.4	64.3
9	1088	64.0	97.6	61.7
10	626	45.7	13.9	5.9
11	642	81.2	50.9	43.3
12	891	87.8	68.2	62.3
13	470	57.7	20.9	13.0
14	555	73.9	42.9	34.8
15	643	95.2	32.7	30.6
16	407	51.4	40.8	29.0
17	1085	78.6	47.3	37.4
18	686	83.4	92.7	80.2
19	554	58.3	31.0	19.0
20	685	74.9	92.3	69.5
21	524	93.9	85.7	80.5
22	957	55.3	31.2	16.7
23	1487	85.3	33.8	29.1
		$P<0.001$	$P<0.001$	$P<0.001$
Europe	9360	68.8	54.7	39.7
USA	7594	81.6	60.6	53.1
		$P<0.001$	$P<0.001$	$P<0.001$

The heterogeneity among studies was assessed by χ^2 test.

lower level of confidence (either probable or possible) or into another category.

There was a major cardiac source in 4496 of the 16 954 (27%) cases. Atrial fibrillation accounted for the largest proportion of the major cardiac source of embolism (3735; 83%). There was another competing major vascular or systemic abnormality in 816 (18%) cases. Diagnostic investigations were incomplete in 2229 (50%) cases. The final causative subtype was CE-evident in 2011 (45%) cases with a major cardiac source of embolism (Figure 2B).

There were 2458 (15%) cases with a typical lacunar infarct on neuroimaging. Among those with lacunar infarction, intracranial vascular imaging was available in 1567 (64%). An abnormality in the parent artery at the origin of the penetrating artery supplying the territory of the lacunar infarct was reported in 300 (19%). Cardiac investigations were performed in 1617 (66%) and revealed a major cardiac source in 208 (13%). Overall, vascular and cardiac evaluations revealed another major pathogenesis in 492 (20%). Investigations were incomplete in 1344 (55%). The final causative subtype was SAO-evident in 992 (40%) cases with a typical lacunar infarct on neuroimaging (Figure 2C).

Diagnostic investigations revealed a major uncommon pathogenesis in 1109 (7%) cases. The most frequent uncommon pathogenesis was acute arterial dissection (397 cases, 36%). Overall, incomplete evaluation and completing major pathogenesis rate was 53% in this category. The final causative subtype was evident other uncommon causes in 526 (47%) patients with a major uncommon pathogenesis (Figure 2D).

The largest pathogenic category was undetermined pathogenesis (7272 cases; 43%). This category included 3947 (55%) with incomplete evaluation, 1333 (18%) with minor cardiac emboli sources, 655 (9%) with multiple competing pathogeneses, 1118 (15%) with cryptogenic-other, and 219 (3%) with cryptogenic embolism. Of note, there were a total of 3257 (19%) cases in the entire study cohort with multiple competing pathogeneses (major or minor). Nevertheless, the final causative subtype was unclassified in only 655 (4%), suggesting that the CCS algorithm was able to identify a probable pathogenesis in the vast majority of patients with overlapping pathogeneses.

There was a significant relationship between stroke subtypes and age, sex, and race ($P<0.001$ for each; Figure II in the online-only Data Supplement). The association was most obvious for age. Subjects ≥50 years were ≈4× more likely to have cardioembolic stroke and 6× less likely to have stroke because of other uncommon causes as compared with those <50 years. In a further analysis where age was classified by decades, we found a continuous increase in LAA and SAO with increasing age with peak values in the age ≈50 to 70 years (Figure 3). There was no such age peak in major CE; instead, the probability of major CE continuously increased by increasing age. In contrast, there was a steady decrease in cryptogenic and other uncommon strokes by increasing age.

Reliability

There were 1509 paired ratings by 52 adjudicators and 24 readjudicators. The crude agreement for 5-subtype causative system was 80% (Table I in the online-only Data Supplement). The corresponding κ value was 0.72 (95% CI, 0.69–0.75). The crude agreement rate for 5-subtype phenotypic system was 81% with a corresponding κ value of 0.73 (95% CI, 0.70–0.75; Table II in the online-only Data Supplement). Crude agreements rates for causative system varied between 65% and 99% across the study sites except for 1 site where the agreement rate was 40% (Figure III in the online-only Data Supplement). After excluding that one outlier, the κ value increased to 0.75 (95% CI, 0.72–0.77) for causative and 0.75 (95% CI, 0.72–0.78) for phenotypic classifications.

Discussion

This is a large study of systematic ischemic stroke subtyping using an evidence- and rule-based system. Because of its size, patterns of subtype distribution across age groups are more readily discernible. It is also the largest study of the inter-rater reliability of ischemic stroke subtyping published thus far, based on 1509 paired ratings by a total of 76 trained and certified adjudicators and readjudicators. There was simultaneous

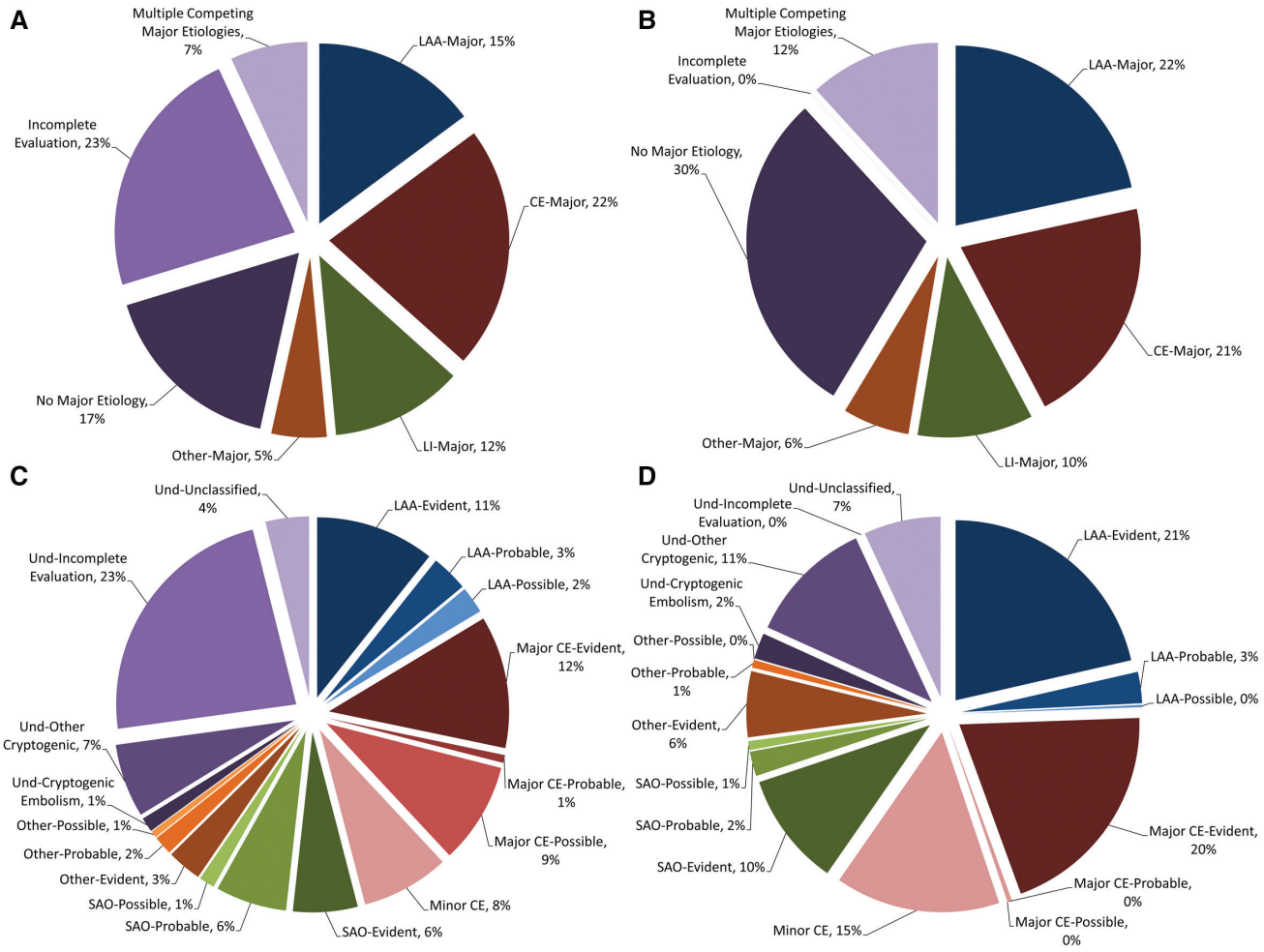


Figure 1. Distribution of phenotypic and causative stroke subtypes. **A**, Phenotypic subtypes in the entire population (n=16954); (**B**) phenotypic subtypes in the subset with complete vascular and cardiac investigation (n=7748); (**C**) causative subtypes in the entire population; and (**D**) causative subtypes in the subset with complete vascular and cardiac investigation. Please note that the term incomplete evaluation in **A** and **C** designates a pathogenic subgroup under undetermined (Und) category that is considered when diagnostic investigations are not performed in the absence of an identified pathogenesis. According to this definition, a case with atrial fibrillation in history is not classified as incomplete evaluation when vascular and cardiac investigations are not done. The term complete investigation in **B** and **D**, however, is solely based on availability of diagnostic tests indicating that brain imaging, vascular imaging, and cardiac evaluation are available. CE indicates cardiac embolism; LAA, large artery atherosclerosis; LI, lacunar infarction; and SAO, small artery occlusion.

assessment of phenotypic and causative subtypes allowing examination of subtype distribution and reliability separately for these 2 types of classification. Our finding of discordance between the causative and phenotypic classifications is expected and reflects the fact that the presence of a phenotypic characteristic in a given patient, such as a vascular or cardiac abnormality, does not necessarily mean that it is the cause of stroke in that patient.

The extent of diagnostic evaluation was heterogeneous for a variety of reasons. Some studies used single site recruitment where strokes were evaluated at tertiary medical centers by vascular neurologists with a highly consistent diagnostic approach, whereas other studies were regional or national in scope with strokes evaluated primarily at community hospitals by physicians with diverse backgrounds with a less consistent diagnostic approach.[15] This variation in extent of diagnostic evaluation motivated us to provide data separately for the subset with complete vascular and cardiac investigations. The results in this subset are the highest quality data available in the literature on the distribution of stroke subtypes. Of note, the subset with complete investigations resembled the overall study population with respect to the majority of baseline characteristics, suggesting no substantial selection bias.

In the present study, inter-rater reliability was slightly lower (κ=0.72) than previously reported for CCS ($\kappa \geq 0.80$).[12–14] Prior studies had smaller number of raters (n=2–20) and smaller number of cases (n=50). As the number of cases and the number of raters increase, the variance introduced to stroke classification also increases and hence the reliability decreases. In contrast to prior studies that used abstracted case summaries, reliability analysis in this study was based on reviews of unabstracted case report forms and patient charts. Differences in raters' ability to pinpoint the medical record data that is critical for subtyping, ambiguities in the source data (for instance, inconsistencies in interpretation of test finding between physician notes), variance in raters' interpretation of the diagnostic data, and lack of data or incomplete data

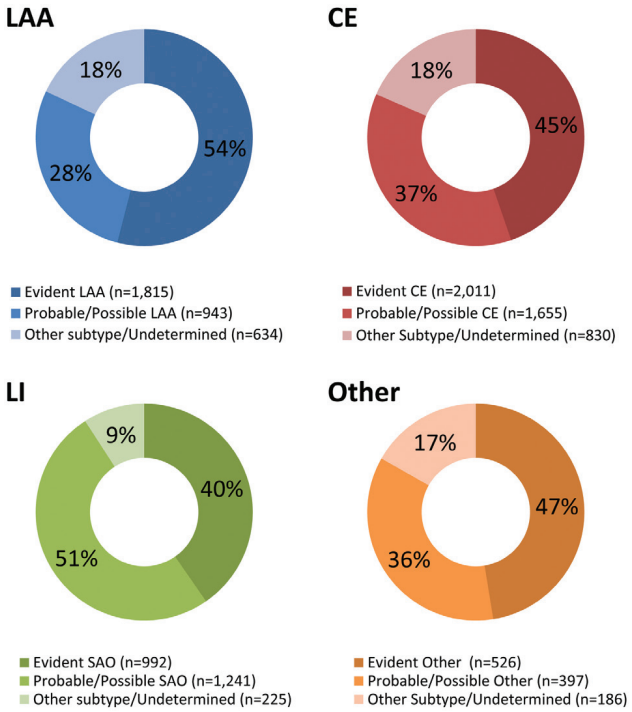


Figure 2. Correlation between causative and phenotypic subtypes. Segments in each circle indicate proportion of causative subtypes in each major phenotypic category (circles). CE indicates cardiac embolism; LAA, large artery atherosclerosis; LI, lacunar infarction; and SAO, small artery occlusion.

(for instance, unavailability of radiographic images for visual assessment) likely contributed to lower agreement. Despite all these factors, κ for CCS still exceeds reported κs for conventional classification systems. The Women's Health Study is the largest reliability study of Trial of ORG 10172 in Acute Stroke Treatment (TOAST) including 133 cases and 2 raters and reporting a κ of 0.49.[9] The Siblings with Ischemic Stroke Study (SWISS) assessed the reliability of TOAST using the largest number of raters (6 raters and 30 cases) and reporting a κ of 0.54.[11] The κ for conventional classification would be expected to be much lower when tested in the same test conditions with the present study.

Several limitations merit further discussion. We did not include a specific minimum standard for quality of source data used for phenotyping. The 23 studies included in this analysis represent a broad range of methodologies (hospital-based case–control, pedigree-based, observational cohorts, and population-based studies) and using a broad range of criteria for inclusion ranging from no restriction to targeted recruitment by age, sex, family history, etc. Source documents varied from secondary notes of test results to computerized data repositories where access to source data such as radiographic images was possible. This diversity strengthens the confidence in our findings by capturing the vagaries that may occur in the real world as opposed to the rigors of a structured clinical research setting. Moreover, CCS provides refined subtypes by integrating quality and completeness of source data into level of confidence for each subtype, minimizing the impact of diversity in source data on validity of classifications. Insufficient representation of certain racial and ethnic groups (for instance, Asian population) in SiGN may have caused underestimation of certain mechanisms such as intracranial atherosclerosis. Finally, because majority of studies were hospital-based, the study population was vulnerable to survival, severity, or consent bias in addition to the impact of specific inclusion and exclusion criteria.

A major strength of this study was systematic adjudication of stroke subtypes using a rule- and evidence-based system. CCS offers several advantages such as good to excellent reliability and web-based interface.[13] In addition, CCS retains and standardizes individual data points such as atrial fibrillation or arterial dissection that underlie subtype classification. Furthermore, its ability to provide both phenotypic and causative subtypes would allow one to separately explore the genetic basis of the presence of a potential pathogenesis (phenotypic subtype) and the presence of a causative pathogenesis (causative subtype). A gene for LAA could be different from a gene that makes an atherosclerotic plaque rupture and cause

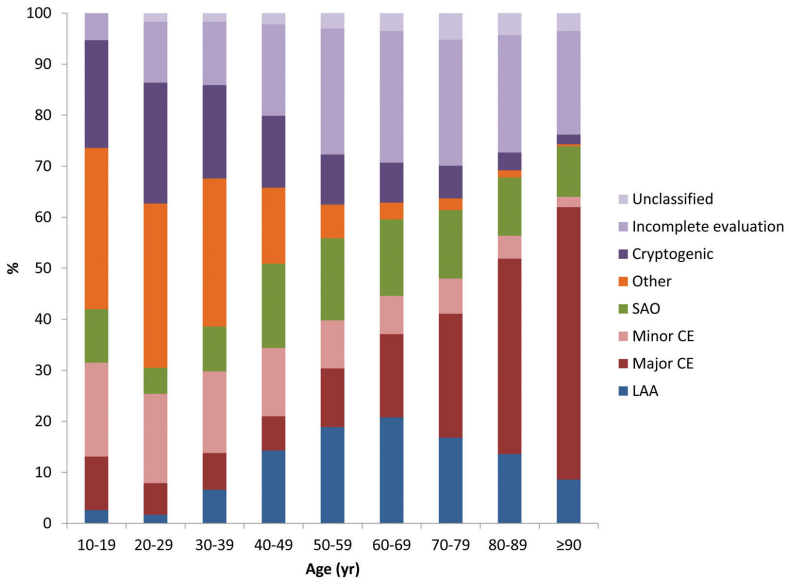


Figure 3. The relationship between age and causative stroke subtypes. CE indicates cardiac embolism; LAA, large artery atherosclerosis; and SAO, small artery occlusion.

stroke. The ability to study such differential genetic associations would facilitate our understanding of the pathophysiological basis of ischemic stroke.

Sources of Funding

The Stroke Genetics Network (SiGN) study was funded by a cooperative agreement grant from the National Institute of Neurological Disorders and Stroke (NINDS) U01 NS069208. The Base de Datos de Ictus del Hospital del Mar (BASICMAR) Genetic Study was supported by the Ministerio de Sanidad y Consumo de España, Instituto de Salud Carlos III (ISC III) with the grants: Registro BASICMAR Funding for Research in Health (PI051737); GWA Study of Leukoaraiosis (GWALA) project from Fondos de Investigación Sanitaria ISC III (PI10/02064) and (PI12/01238); and Fondos European Regional Development Funding (FEDER/EDRF) Red de Investigación Cardiovascular (RD12/0042/0020). Additional support was provided by the Fundació la Marató TV3 with the grant GOD's project. Genestroke Consortium (76/C/2011) Recercaixa'13 (JJ086116). Assistance with data cleaning was provided by the Research in Cardiovascular and Inflammatory Diseases Program of Institute Hospital del Mar of Medical Investigations, Hospital del Mar, and the Barcelona Biomedical Research Park. The Bio-Repository of DNA in Stroke (BRAINS) was supported by the British Council (UKIERI), Henry Smith Charity, and the UK Stroke Research Network. Dr Sharma was supported by a Department of Health (United Kingdom) Senior Fellowship. Center for Inherited Disease Research (CIDR): genotyping services were provided by the Johns Hopkins University CIDR, which is fully funded through a federal contract from the National Institutes of Health (NIH) to the Johns Hopkins University (contract No. HHSN268200782096C). The Edinburgh Stroke Study was supported by the Wellcome Trust (clinician scientist award to Dr Sudlow) and the Binks Trust. Sample processing occurred in the Genetics Core Laboratory of the Wellcome Trust Clinical Research Facility, Western General Hospital, Edinburgh. Much of the neuroimaging occurred in the Scottish Funding Council Brain Imaging Research Centre (www.sbirc.ed.ac.uk), Division of Clinical Neurosciences, University of Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the Scottish Imaging Network–A Platform for Scientific Excellence collaboration (www.sinapse.ac.uk), funded by the Scottish Funding Council and the Chief Scientist Office. Genotyping was performed at the Wellcome Trust Sanger Institute in the United Kingdom and funded by the Wellcome Trust as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA). The Massachusetts General Hospital Stroke Genetics Group was supported by the NIH Genes Affecting Stroke Risks and Outcomes Study grant K23 NS042720, the American Heart Association/Bugher Foundation Centers for Stroke Prevention Research 0775010N, and NINDS K23NS042695, R01NS059727, the Deane Institute for Integrative Research in Atrial Fibrillation and Stroke, and by the Keane Stroke Genetics Fund. Genotyping services were provided by the Broad Institute Center for Genotyping and Analysis, supported by grant U54 RR020278 from the National Center for Research Resources. The Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS) was supported by the NIH (NS 030678). The Genetics of Early Onset Stroke (GEOS) Study was supported by the NIH Genes, Environment, and Health Initiative (GEI) grant U01 HG004436, as part of the Gene Environment Association Studies (GENEVA) consortium under GEI, with additional support provided by the Mid-Atlantic Nutrition and Obesity Research Center (P30 DK072488) and the Office of Research and Development, Medical Research Service, and the Baltimore Geriatrics Research, Education, and Clinical Center of the Department of Veterans Affairs. Genotyping services were provided by the Johns Hopkins University CIDR, which is fully funded through a federal contract from the NIH to the Johns Hopkins University (contract No. HHSN268200782096C). Assistance with data cleaning was provided by the GENEVA Coordinating Center (U01 HG 004446; PI Bruce S Weir). Study recruitment and assembly of data sets were supported by a Cooperative Agreement with the Division of Adult and Community Health, Centers for Disease Control and by grants from the NINDS and the NIH Office of Research on Women's Health (R01 NS45012, U01 NS069208-01). GRAZ: The Austrian Stroke Prevention Study was supported by the Austrian Science Fund (FWF) grant Nos. P20545-P05 and P13180 and I904-B13 (Era-Net). The Medical University of Graz supports the databases of the Graz Stroke Study and the Austrian Stroke Prevention Study. The Ischemic Stroke Genetics Study (ISGS) was supported by the NINDS (R01 NS42733; PI Dr Meschia). The Sibling with Ischemic Stroke Study (SWISS) was supported by the NINDS (R01 NS39987; PI Dr Meschia). Both SWISS and ISGS received additional support, in part, from the Intramural Research Program of the National Institute on Aging (Z01 AG000954-06; PI Andrew Singleton). SWISS and ISGS used samples and clinical data from the NIH-NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds), human subject protocol Nos. 2003-081 and 2004-147. SWISS and ISGS used stroke-free participants from the Baltimore Longitudinal Study of Aging (BLSA) as controls with the permission of Dr Luigi Ferrucci. The inclusion of BLSA samples was supported, in part, by the Intramural Research Program of the National Institute on Aging (Z01 AG000015-50), human subject protocol No. 2003-078. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the NIH (http://biowulf.nih.gov). Phenotypic data and genetic specimens collection were funded by the grant from the Polish Ministry of Science and Higher Education for Leading National Research Centers (KNOW) and by the grant from the Medical College, Jagiellonian University in Krakow, Poland: K/ZDS/002848. The Leuven Stroke genetics study was supported by personal research funds from the Department of Neurology of the University Hospitals Leuven. Dr Thijs is supported by a Fundamental Clinical Research grant from FWO Flanders (Nos. 1.8.009.08.N.00 and 1800913N). Dr Lemmens is a Senior Clinical Investigator of FWO Flanders (FWO 1841913N) and is supported through Fonds Annie Planckaert-Dewaele. The Lund Stroke Register was supported by the Swedish Research Council (K2010-61X-20378-04-3), The Swedish Heart-Lung Foundation, Region Skåne, the Freemasons Lodge of Instruction EOS in Lund, King Gustaf V's and Queen Victoria's Foundation, Lund University, and the Swedish Stroke Association. Biobank services were provided by Region Skåne Competence Centre (RSKC Malmö), Skåne University Hospital, Malmö, Sweden, and Biobank, Labmedicin Skåne, University and Regional Laboratories Region Skåne, Sweden. The Middlesex County Ischemic Stroke Study was supported by intramural funding from the New Jersey Neuroscience Institute/JFK Medical Center, Edison, NJ, and The Neurogenetics Foundation, Cranbury, NJ. The Northern Manhattan Study was supported by grants from the NINDS (R37 NS029993, R01 NS27517). The Cerebrovascular Biorepository at University of Miami/Jackson Memorial Hospital (The Miami Stroke Registry, Institutional Review Board No. 20070386) was supported by the Department of Neurology at University of Miami Miller School of Medicine and Evelyn McKnight Brain Institute. Biorepository and DNA extraction services were provided by the Hussmann Institute for Human Genomics at the Miller School of Medicine. The MUNICH study was supported by the Vascular Dementia Research Foundation and the Jackstaedt Stiftung. The Nurses' Health Study work on stroke is supported by grants from the NIH, including HL088521 and HL34594 from the National Heart, Lung, and Blood Institute, as well as grants from the National Cancer Institute funding the questionnaire follow-up and blood collection: CA87969 and CA49449. The Oxford Vascular Study was supported by the Stroke Association, Medical Research Council, Wellcome Trust, Dunhill Medical Trust, NIH Research (NIHR), and NIHR Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford. Dr Rothwell is in receipt of Senior Investigator Awards from the Wellcome Trust and the NIHR. The Reasons for Geographic and Racial Differences in Stroke Study (REGARDS) was supported by a cooperative agreement U01 NS041588 from the NINDS, NIH, and Department of Health and Human Service. A full

list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org. The Sahlgrenska Academy Study of Ischemic Stroke was supported by the Swedish Research Council (K2011-65X-14605-09-6), the Swedish Heart and Lung Foundation (20100256), the Swedish state/Sahlgrenska University Hospital (ALFGBG-148861), the Swedish Stroke Association, the Swedish Society of Medicine, and the Rune and Ulla Amlöv Foundation. SPS3: The Secondary Prevention of Small Subcortical Strokes trial was funded by the US National Institute of Health and Neurological Disorders and Stroke grant No. U01NS38529-04A1 (principal investigator, Oscar R. Benavente; coprincipal investigator, Robert G. Hart). The SPS3 Genetic Substudy (SPS3-GENES) was funded by R01 NS073346 (coprincipal investigators, Julie A. Johnson, Oscar R. Benavente, and Alan R. Shuldiner). ST. GEORGE'S: The principal funding for this study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z and WT084724MA). Collection of some of the St George's stroke cohort was supported by project grant support from the Stroke Association. The Women's Health Initiatives program was funded by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118 to 32119, 32122, 42107-26, 42129-32, and 44221. The Hormones and Biomarkers Predicting Stroke was supported by a grant from the National Institutes of Neurological Disorders and Stroke (R01NS042618). Washington University St. Louis Stroke Study (WUSTL): The collection, extraction of DNA from blood, and storage of specimens were supported by 2 NINDS NIH grants (P50 NS055977 and R01 NS8541901). Basic demographic and clinical characterization of stroke phenotype was prospectively collected in the Cognitive Rehabilitation and Recovery Group registry. The Recovery Genomics after Ischemic Stroke study was supported by a grant from the Barnes-Jewish Hospital Foundation.

Disclosures

Drs Brown, Kittner, Markus, Rexrode, Sacco, and Meschia received research grant from National Institutes of Health (NIH). Dr Engström has an employment position in Astra Zeneca R&D. Dr Rosand received research grant from NIH and has a consultant or advisory relationship with Boehringer Ingelheim. Dr Worrall received research grant from NIH and has an associate editor affiliation with American Academy of Neurology. The other authors report no conflicts.

References

1. Gschwendtner A, Bevan S, Cole JW, Plourde A, Matarin M, Ross-Adams H, et al; International Stroke Genetics Consortium. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. *Ann Neurol*. 2009;65:531–539.
2. International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez C, Bevan S, Gschwendtner A, Spencer CC, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet*. 2012;44:328–333.
3. Holliday EG, Maguire JM, Evans TJ, Koblar SA, Jannes J, Sturm JW, et al; Australian Stroke Genetics Collaborative; International Stroke Genetics Consortium; Wellcome Trust Case Control Consortium 2. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet*. 2012;44:1147–1151.
4. Lubitz SA, Yi BA, Ellinor PT. Genetics of atrial fibrillation. *Heart Fail Clin*. 2010;6:239–247.
5. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*. 2007;448:353–357.
6. Goldstein LB, Jones MR, Matchar DB, Edwards LJ, Hoff J, Chilukuri V, et al. Improving the reliability of stroke subgroup classification using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria. *Stroke*. 2001;32:1091–1098.
7. Gordon DL, Bendixen BH, Adams HP Jr, Clarke W, Kappelle LJ, Woolson RF. Interphysician agreement in the diagnosis of subtypes of acute ischemic stroke: implications for clinical trials. The TOAST Investigators. *Neurology*. 1993;43:1021–1027.
8. Lindley RI, Warlow CP, Wardlaw JM, Dennis MS, Slattery J, Sandercock PA. Interobserver reliability of a clinical classification of acute cerebral infarction. *Stroke*. 1993;24:1801–1804.
9. Atiya M, Kurth T, Berger K, Buring JE, Kase CS; Women's Health Study. Interobserver agreement in the classification of stroke in the Women's Health Study. *Stroke*. 2003;34:565–567.
10. Selvarajah JR, Glaves M, Wainwright J, Jha A, Vail A, Tyrrell PJ. Classification of minor stroke: intra- and inter-observer reliability. *Cerebrovasc Dis*. 2009;27:209–214.
11. Meschia JF, Barrett KM, Chukwudelunzu F, Brown WM, Case LD, Kissela BM, et al; Siblings with Ischemic Stroke Study (SWISS) Investigators. Interobserver agreement in the trial of org 10172 in acute stroke treatment classification of stroke based on retrospective medical record review. *J Stroke Cerebrovasc Dis*. 2006;15:266–272.
12. Ay H, Furie KL, Singhal A, Smith WS, Sorensen AG, Koroshetz WJ. An evidence-based causative classification system for acute ischemic stroke. *Ann Neurol*. 2005;58:688–697.
13. Ay H, Benner T, Arsava EM, Furie KL, Singhal AB, Jensen MB, et al. A computerized algorithm for etiologic classification of ischemic stroke: the Causative Classification of Stroke System. *Stroke*. 2007;38:2979–2984.
14. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al; International Stroke Genetics Consortium. The Causative Classification of Stroke system: an international reliability and optimization study. *Neurology*. 2010;75:1277–1284.
15. Meschia JF, Arnett DK, Ay H, Brown RD Jr, Benavente OR, Cole JW, et al; NINDS SiGN Study. Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke*. 2013;44:2694–2702.
16. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.

Stroke
Journal of the American Heart Association

Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network

Hakan Ay, Ethem Murat Arsava, Gunnar Andsberg, Thomas Benner, Robert D. Brown Jr, Sherita N. Chapman, John W. Cole, Hossein Delavaran, Martin Dichgans, Gunnar Engström, Eva Giralt-Steinhauer, Raji P. Grewal, Katrina Gwinn, Christina Jern, Jordi Jimenez-Conde, Katarina Jood, Michael Katsnelson, Brett Kissela, Steven J. Kittner, Dawn O. Kleindorfer, Daniel L. Labovitz, Silvia Lanfranconi, Jin-Moo Lee, Manuel Lehm, Robin Lemmens, Chris Levi, Linxin Li, Arne Lindgren, Hugh S. Markus, Patrick F. McArdle, Olle Melander, Bo Norrving, Leema Reddy Peddareddygari, Annie Pedersén, Joanna Pera, Kristiina Rannikmäe, Kathryn M. Rexrode, David Rhodes, Stephen S. Rich, Jaume Roquer, Jonathan Rosand, Peter M. Rothwell, Tatjana Rundek, Ralph L. Sacco, Reinhold Schmidt, Markus Schürks, Stephan Seiler, Pankaj Sharma, Agnieszka Slowik, Cathie Sudlow, Vincent Thijs, Rebecca Woodfield, Bradford B. Worrall and James F. Meschia

Stroke. 2014;45:3589-3596; originally published online November 6, 2014;
doi: 10.1161/STROKEAHA.114.007362
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

Correction

The version of the article, "Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network" by Ay et al that published online ahead-of-print on November 6, 2014, and appears in the December issue (*Stroke*. 2014;45:3589–3596) contained incomplete author affiliations. The following affiliations have been added for authors Robin Lemmens and Vincent Thijs:

KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), B-3000 Leuven, Belgium (R.L., V.T.).

VIB, Vesalius Research Center, Laboratory of Neurobiology, B-3000 Leuven, Belgium (R.L., V.T.).

This correction has been made to the online version of the article, which is available at http://stroke.ahajournals.org/content/45/12/3589.

(*Stroke*. 2015;46:e17. DOI: 10.1161/STR.0000000000000057.)

SUPPLEMENTAL MATERIAL

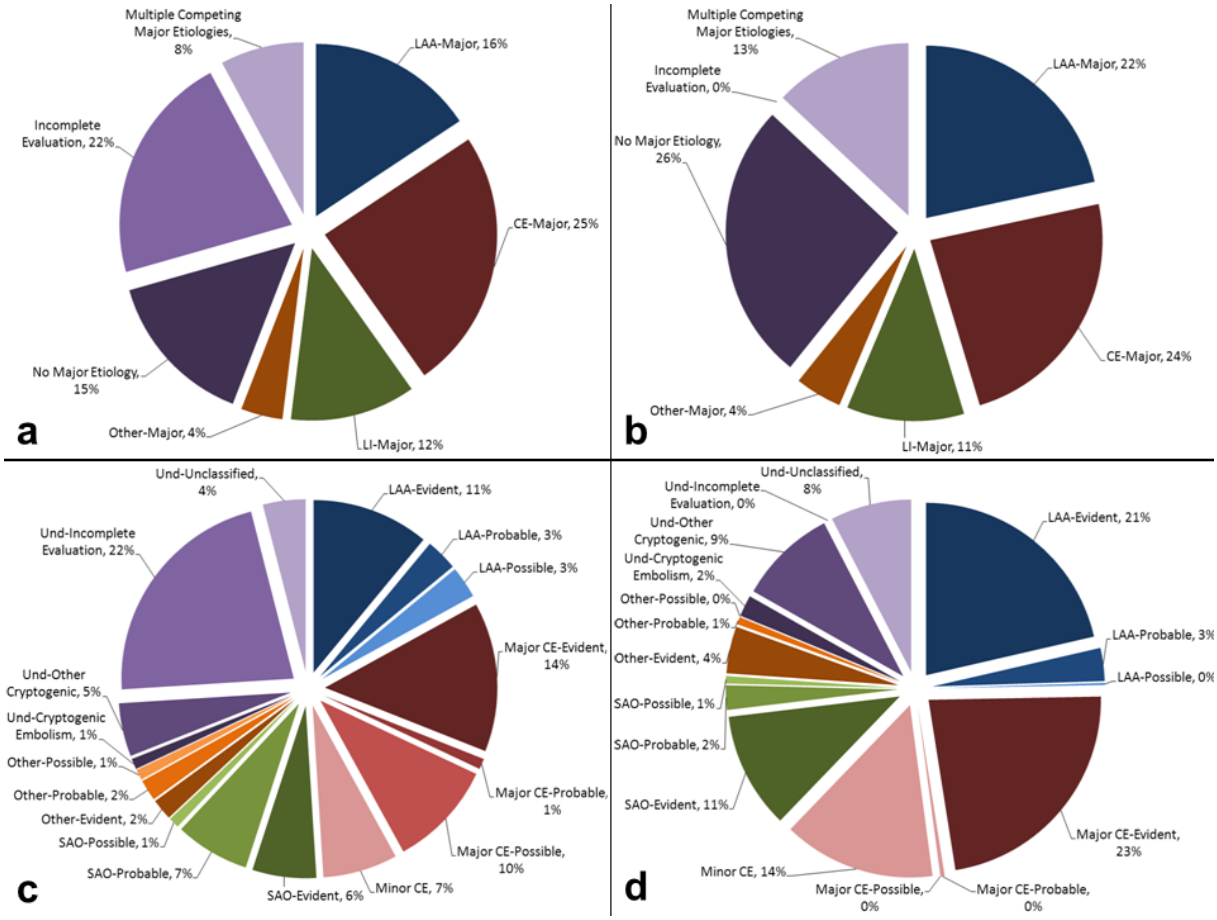
Supplemental Table I: Causative subtypes by adjudicators and readjudicators. The numbers indicate number of stroke cases evaluated.

		Readjudicator				
		LAA	CE	SAO	Other	Undetermined
Adjudicator	LAA	**186**	9	3	1	39
	CE	4	**296**	7	2	31
	SAO	6	8	**125**	3	62
	Other	1	0	2	**56**	10
	Undetermined	23	36	40	13	**546**

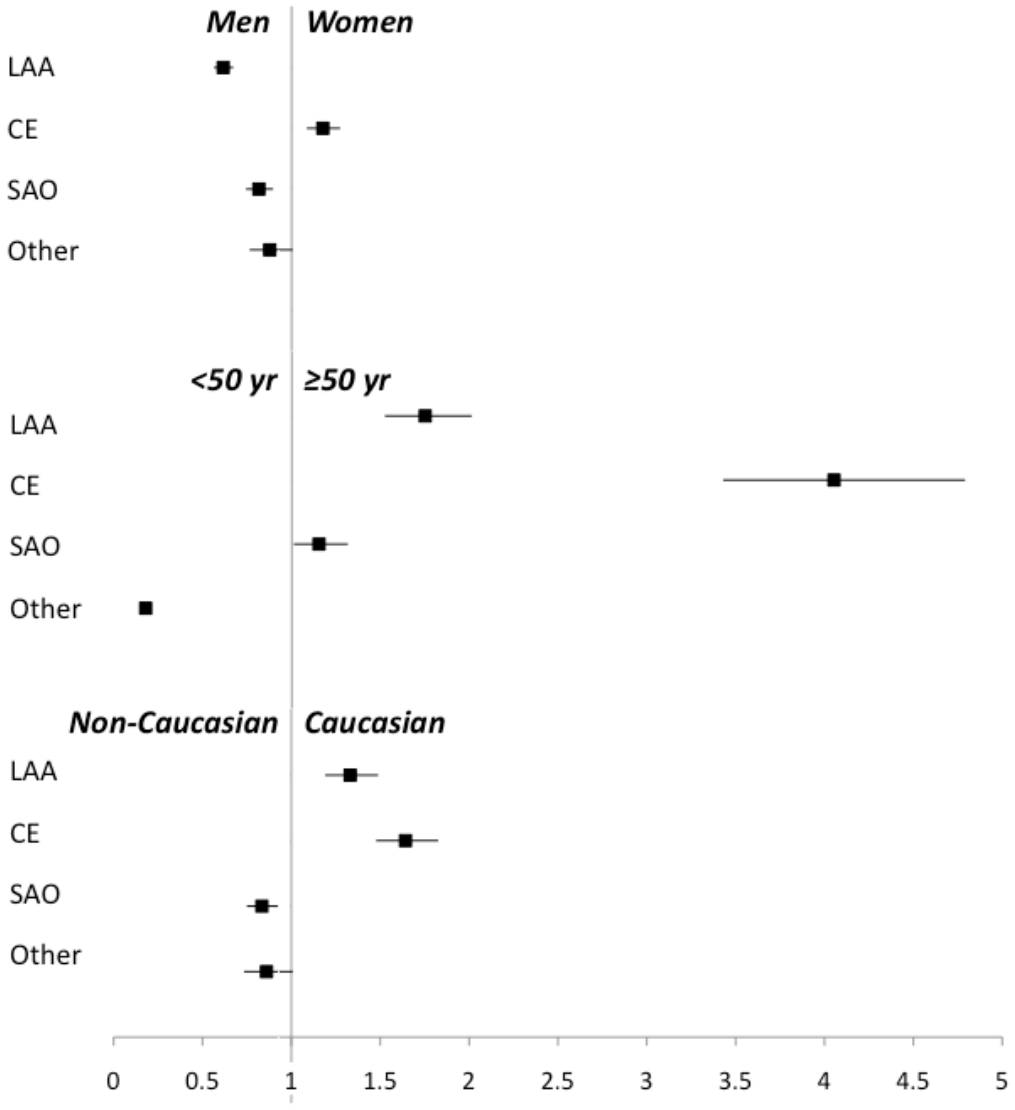
Supplemental Table II: Phenotypic subtypes by adjudicators and readjudicators. The numbers indicate number of stroke cases evaluated.

		Readjudicator				
		LAA-major	CE-major	LI-major	Other-major	Undetermined
Adjudicator	LAA-major	**177**	4	1	1	36
	CE-major	4	**307**	4	1	26
	LI-major	5	3	**108**	3	56
	Other-major	1	0	2	**51**	12
	Undetermined	28	45	40	15	**579**

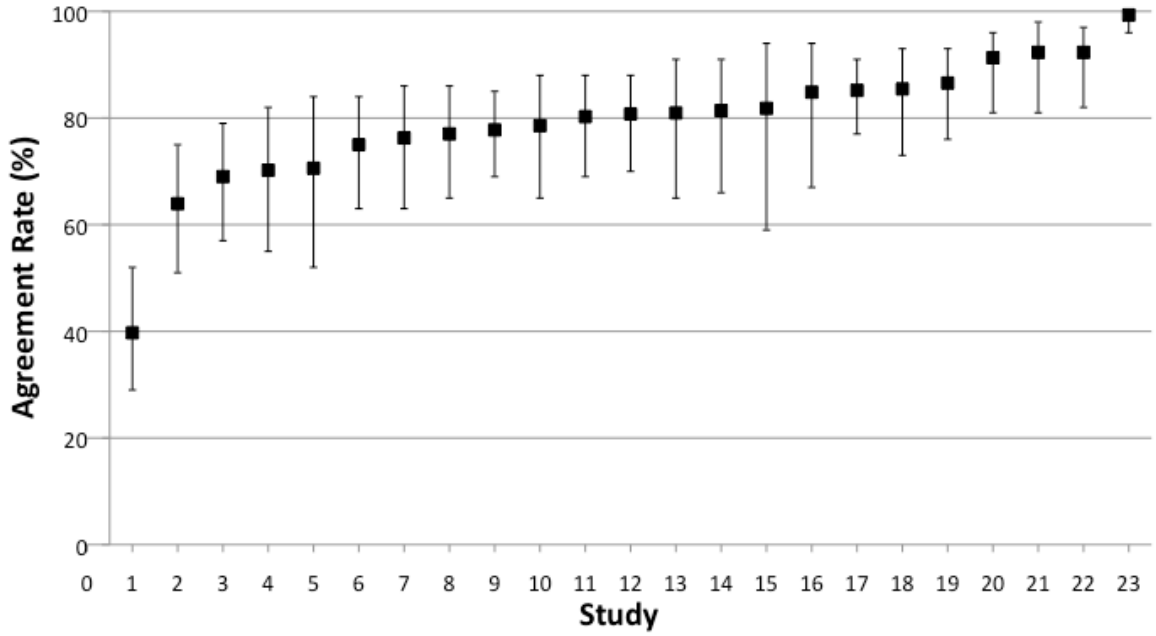
Supplemental figure I:



Supplemental Figure II:



Supplemental Figure III:



Supplemental Figure Legends:

Supplemental figure I: Distribution of causative and phenotypic stroke subtypes in studies with unselected populations: 1(a), phenotypic subtypes in the entire population; 1(b), phenotypic subtypes in the subset with complete vascular and cardiac investigation; 1(c), causative subtypes in the entire population; 1(d), causative subtypes in the subset with complete vascular and cardiac investigation. Und: undetermined

Supplemental figure II: Association between causative stroke subtypes and patient characteristics. Multinomial logistic regression was used to calculate odds ratios and 95% CI with the "Undetermined" group as the reference category.

Supplemental figure III: Crude agreement rates for causative classification between adjudicators and readjudications across the contributing studies.

4 MANUSCRIPT 2

“Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study.”

Pulit, S., (…), **Lehm, M.**, (…), Worrall, B.

This manuscript has been peer-reviewed and published under the following reference:
Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *The Lancet Neurology*, Volume 15, Issue 2, 174 - 184; doi: http://dx.doi.org/10.1016/S1474-4422(15)00338-5.

Author contribution:
Recruitment of IS patients, collection of DNA samples, classification of IS patients according to their probable etiology.

Supplemental material:
The full supplemental material can be accessed via:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4912948/bin/NIHMS750481-supplement-1.pdf

Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study

*NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium (ISGC)**

Summary

Background The discovery of disease-associated loci through genome-wide association studies (GWAS) is the leading genetic approach to the identification of novel biological pathways underlying diseases in humans. Until recently, GWAS in ischaemic stroke have been limited by small sample sizes and have yielded few loci associated with ischaemic stroke. We did a large-scale GWAS to identify additional susceptibility genes for stroke and its subtypes.

Methods To identify genetic loci associated with ischaemic stroke, we did a two-stage GWAS. In the first stage, we included 16 851 cases with state-of-the-art phenotyping data and 32 473 stroke-free controls. Cases were aged 16 to 104 years, recruited between 1989 and 2012, and subtypes of ischaemic stroke were recorded by centrally trained and certified investigators who used the web-based protocol, Causative Classification of Stroke (CCS). We constructed case-control strata by identifying samples that were genotyped on nearly identical arrays and were of similar genetic ancestral background. We cleaned and imputed data by use of dense imputation reference panels generated from whole-genome sequence data. We did genome-wide testing to identify stroke-associated loci within each stratum for each available phenotype, and we combined summary-level results using inverse variance-weighted fixed-effects meta-analysis. In the second stage, we did in-silico lookups of 1372 single nucleotide polymorphisms identified from the first stage GWAS in 20 941 cases and 364 736 unique stroke-free controls. The ischaemic stroke subtypes of these cases had previously been established with the Trial of Org 10 172 in Acute Stroke Treatment (TOAST) classification system, in accordance with local standards. Results from the two stages were then jointly analysed in a final meta-analysis.

Findings We identified a novel locus (G allele at rs12122341) at 1p13.2 near *TSPAN2* that was associated with large artery atherosclerosis-related stroke (first stage odds ratio [OR] 1·21, 95% CI 1·13–1·30, $p=4·50\times10^{-8}$; joint OR 1·19, 1·12–1·26, $p=1·30\times10^{-9}$). Our results also supported robust associations with ischaemic stroke for four other loci that have been reported in previous studies, including *PITX2* (first stage OR 1·39, 1·29–1·49, $p=3·26\times10^{-19}$; joint OR 1·37, 1·30–1·45, $p=2·79\times10^{-32}$) and *ZFHX3* (first stage OR 1·19, 1·11–1·27, $p=2·93\times10^{-7}$; joint OR 1·17, 1·11–1·23, $p=2·29\times10^{-10}$) for cardioembolic stroke, and *HDAC9* (first stage OR 1·29, 1·18–1·42, $p=3·50\times10^{-8}$; joint OR 1·24, 1·15–1·33, $p=4·52\times10^{-9}$) for large artery atherosclerosis stroke. The 12q24 locus near *ALDH2*, which has previously been associated with all ischaemic stroke but not with any specific subtype, exceeded genome-wide significance in the meta-analysis of small artery stroke (first stage OR 1·20, 1·12–1·28, $p=6·82\times10^{-8}$; joint OR 1·17, 1·11–1·23, $p=2·92\times10^{-9}$). Other loci associated with stroke in previous studies, including *NINJ2*, were not confirmed.

Interpretation Our results suggest that all ischaemic stroke-related loci previously implicated by GWAS are subtype specific. We identified a novel gene associated with large artery atherosclerosis stroke susceptibility. Follow-up studies will be necessary to establish whether the locus near *TSPAN2* can be a target for a novel therapeutic approach to stroke prevention. In view of the subtype-specificity of the associations detected, the rich phenotyping data available in the Stroke Genetics Network (SiGN) are likely to be crucial for further genetic discoveries related to ischaemic stroke.

Funding US National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Lancet Neurol 2015

Published Online December 18, 2015 http://dx.doi.org/10.1016/S1474-4422(15)00338-5

See Online/Comment http://dx.doi.org/10.1016/S1474-4422(15)00400-7

*Members listed at end of the paper

Correspondence to: Dr Jonathan Rosand, Department of Neurology and Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA
jrosand@partners.org

Introduction

Worldwide, stroke is the second leading cause of death[1] and a major contributor to dementia and age-related cognitive decline. About 15 million people have a stroke each year.[1] Most survivors are left with a permanent disability, which makes stroke the world's leading cause of adult incapacity.[2] Strokes result from the sudden occlusion or rupture of a blood vessel supplying the brain, and so are categorised accordingly as ischaemic (vessel occlusion) or haemorrhagic (vessel rupture) on the basis of neuroimaging results. Up to 85% of all strokes are ischaemic.

Although hypertension, atrial fibrillation, diabetes mellitus, and cigarette smoking are known risk factors for stroke,[3] a substantial proportion of the risk remains unexplained and might be attributable to inherited genetic variation. Discovery of genetic variants that predispose to stroke is a crucial first step toward the development of improved diagnostic tests for stroke and novel therapies that might reduce the disease burden. Genome-wide association studies (GWAS) have thus far identified only a few confirmed loci,[4–7] which together account for a small proportion of the heritable risk.[8]

Research in context

Evidence before this study

We searched PubMed with the search terms "stroke" and "genome wide association study" for reports published before Oct 19, 2015. We only included peer-reviewed reports in English. Compared with the rapid pace of genetic discovery for other common disorders, only four loci (PITX2, HDAC9, ZFHX3, and 12q24.2) have been convincingly implicated by genome-wide association studies (GWAS) in ischaemic stroke. GWAS of stroke have been limited by small sample sizes and concerns about phenotypic heterogeneity.

Added value of this study

To our knowledge, the National Institute of Neurological Disorders and Stroke (NINDS)-Stroke Genetics Network (SiGN) project is the largest and most comprehensive study of ischaemic stroke so far. Discovery analyses were done in 16 851 cases and 32 473 controls and findings were followed up in an additional 20 941 cases and 364 736 controls. Furthermore, the project implemented the Causative Classification of Stroke (CCS) system to subtype cases and generated a rich phenotypic database. Trial of Org 10 172 in Acute Stroke Treatment (TOAST)-based subtypes were also available, allowing for the first ever analysis of the genetic overlap between TOAST and CCS subtypes.

Implications of all the available evidence

Our data show that increasing sample size and applying a standardised subtyping method can reveal additional information about the underlying genetic architecture of stroke. Because we had access to phenotype information generated by two different subtyping methods, we also showed that there is moderate to strong genetic correlation between the CCS and TOAST subtyping methods, suggesting that future studies might benefit from liberal inclusion of cases, regardless of subtyping approach. Also, our results show that all discovered loci, including the 12q24.12 locus, which was previously implicated in all ischaemic stroke, are specific to a single subtype, suggesting that these subtypes will have at least partly distinct genetic signatures. Because of the subtype-specificity of genetic associations in stroke, substantially larger samples of stroke subtypes will probably be needed to expand the number of identified stroke loci to that of other common diseases.

Ischaemic stroke occurs when the blood flow to a region of the brain is interrupted because of blockage of a blood vessel. Because vessel occlusion can occur through different mechanisms, ischaemic stroke can be subtyped on the basis of the presumed mechanism: large artery atherosclerosis, cardioembolism, or small artery occlusion. With one exception, all associations for ischaemic stroke detected in GWAS have been subtype-specific, suggesting the need for studies that are powered to detect subtype-specific associations. The National Institute of Neurological Disorders and Stroke (NINDS) Stroke Genetics Network (NINDS-SiGN)[9] is the largest and most comprehensive GWAS of stroke and its subtypes to date. We sought to detect new associations of polymorphisms with risk of ischaemic stroke and its subtypes and to provide evidence for previously reported associations.

See Online for appendix

Methods

Study design

We did a two-stage joint association analysis of ischaemic stroke and its subtypes. The first stage consisted of a GWAS, and the second stage was an in-silico association analysis of the top single nucleotide polymorphisms (SNPs) identified in the first stage in a set of independent samples of cases and controls. We then analysed both stages together to identify loci that exceeded the threshold for genome-wide significance (1×10^{-8}). Compared with separate discovery and replication analyses, this two-stage approach has been shown to improve the power for discovery without altering the type I error.[10]

Study sample

For the first stage, we assessed 31 existing collections that included cases of ischaemic stroke with either available genotypic data or DNA for genotyping, neuroimaging confirmation of stroke, and adequate clinical data to enable phenotypic classification. The cases of ischaemic stroke in the second stage met similar requirements, except that we used pre-existing Trial of Org 10 172 in Acute Stroke Treatment (TOAST)[11] subtyping data for the phenotypic classification. The appendix contains details about each collection, including their study design.

For each collection, approval for inclusion in the SiGN analysis complied with local ethical standards and with local institutional review board and ethics committee oversight. All people included as cases and controls provided written informed consent for genetic studies either directly or by a legally authorised representative.

Classification of stroke subtype

In the NINDS-SiGN,[9] we used two subtyping systems: the Causative Classification of Stroke (CCS) system, which is a standardised web-based subtype classification system,[12] and the more widely used TOAST subtype classification system.[11,13] Both of these systems are based on a similar conceptual framework but are operationalised differently. The TOAST subtyping system is based on the application of written rules requiring clinician judgment; patients with conflicting potential causes are placed into an undetermined category. The CCS subtyping system uses two web-based algorithms that classify patients with conflicting potential causes. Causative (CCSc) categorisation uses historical examination and test data

from each ischaemic stroke subject to assign the most probable cause in the presence of competing aetiologies, while phenotypic (CCSp) categorisation uses abnormal test findings to assign each case into one or more major groups without using rules to determine the most likely aetiology. In addition to the generation of both CCSc and CCSp subtype categories, the advantages of the CCS system are improved inter-observer and intra-observer reliability[12,14,15] and the ability to capture and store individual data elements from the clinical evaluation of the subject.

In the first stage of our study, each site assigned stroke subtypes with the CCS system (appendix). All of the CCS data were collected, subjected to quality control, and analysed centrally. Most sites had previously generated TOAST subtype classifications. In the second stage, we identified additional sites that had GWAS data for subtyped stroke cases. Because we included all available CCS-classified cases in the first stage, we used the corresponding subtype categories from TOAST in the second stage.

For both CCS and TOAST, each case was assigned to one of five ischaemic stroke subtypes: cardioembolic, large artery atherosclerosis, small artery occlusion, undetermined, and other. Although the classification of other was available, we did not analyse it because of low sample counts and insufficient power. In CCS, the classification undetermined was used to refer to cryptogenic cases in which no cause was identified after adequate assessment, whereas in TOAST, undetermined cases were those with incomplete assessment, more than one possible cause, and cryptogenic.

Quality control

Full details of the genotyping and quality control processes are provided in the appendix (p 4). Briefly, newly genotyped cases and about 1150 controls were genotyped on the Illumina 5M array (Illumina, San Diego, CA, USA) so we could include them in the analyses for the first stage. All other cases had been previously genotyped on various Illumina platforms (appendix). We selected publicly available external controls to match cases on the basis of ancestral background and genotyping array.

The cases and controls that were newly genotyped formed separate analysis groups (Krakow, Poland, and Leuven, Belgium; table 1). The remaining cases and controls were matched based on cohort, geographic region of the sample collection site, and genotyping platform (table 1). We assigned matched cases and controls into ancestry-specific analysis strata in two steps (appendix). We projected samples onto HapMap 3[16] data using principal component analysis to establish a group of European ancestry samples. Then, we implemented a hyper-ellipsoid clustering technique based on principal components within self-reported groups of non-Hispanic black and Asian participants. We used the hyper-ellipsoid

	Location of sample collection	Genotyping platform	Ancestry groups	Cases	Controls
First stage					
Case-control group 1					
BRAINS	UK	650Q	European	267	..
MGH-GASROS	USA	610	European	111	..
ISGS	USA	610	European	351	..
SWISS	USA	610	European	25	..
HABC	USA	1M	European	..	1586
Case-control group 2					
EDIN	UK	660	European	566	..
MUNICH	UK	660	European	1131	..
OXVASC	UK	660	European	457	..
STGEORGE	UK	660	European	418	..
KORA	Germany	550	European	..	804
WTCCC	UK	660	European	..	5150
Case-control group 3					
GEOS	USA	1M	African, European	843	880
Case-control group 4					
BRAINS	UK	5M	European, Hispanic	110	..
MGH-GASROS	USA	5M	African, European, Hispanic	456	..
GCNKSS	USA	5M	African, European, Hispanic	482	..
ISGS	USA	5M	African, European, Hispanic	178	..
MCISS	USA	5M	African, European, Hispanic	619	..
MIAMISR	USA	5M	African, European, Hispanic	294	..
NHS	USA	5M	European, Hispanic	314	..
NOMAS	USA	5M	African, European, Hispanic	358	..
REGARDS	USA	5M	African, European, Hispanic	304	..
SPS3	The Americas, Spain	5M	African, European, Hispanic	949	..
SWISS	USA	5M	African, European, Hispanic	181	..
WHI	USA	5M	African, European, Hispanic	454	..
WUSTL	USA	5M	African, European, Hispanic	449	..
HRS	USA	2·5M	African, European, Hispanic	..	11174
OAI	USA	2·5M	African, European	..	3882
HCHS/SOL	USA	2·5M	Hispanic	..	1214
Case-control group 5					
Krakow	Poland	5M	European, Hispanic	880	717
Case-control group 6					
Leuven	Belgium	5M	European, Hispanic	460	453
Case-control group 7					
BASICMAR	Spain	5M	European, Hispanic	890	..
ADHD	Spain	1M	European	..	411
INMA	Spain	1M	European	..	807
Case-control group 8					
GRAZ	Austria	610	European	..	815
GRAZ	Austria	5M	European	609	..
Case-control group 9					
SAHLSIS	Sweden	5M	European, Hispanic	783	..
LUND	Sweden	5M	European, Hispanic	613	..
MDC*	Sweden	610	European, Hispanic	211	1362
Case-control group 10					
ASGC	Australia	610	European	1109	1200
Case-control group 11					
VISP	USA, Canada, UK	1M	African, European	1979	..

(Table 1 continues on next page)

	Location of sample collection	Genotyping platform	Ancestry groups	Cases	Controls
(Continued from previous page)					
Melanoma Study	USA	1M	European	..	1047
HANDLs	USA	1M	African	..	971
Total	..	..	..	16851	32473
Second stage					
ARIC	USA	Affy 6.0	African	263	2466
CADISP†	Multi-cohort	Illumina 610	European	555	9259
CHARGE†	Multi-cohort	Multi-chip	European	3100	75530
CHS	USA	Illumina Omni 1M	African	110	623
deCODE	Iceland	Multi-chip	European	5291	228512
Glasgow	UK	ImmunoChip	European	599	1775
HVH	USA	Illumina 370CNV	European	577	1330
INTERSTROKE†	Multi-cohort	Cardio-metabochip	African, East Asian, European, Hispanic	1771	2103
LUND	Sweden	635	European	546	528
MDC	Sweden	5M	European	1304	3504
METASTROKE†	Multi-cohort	Multi-chip	European	1729	7925
RACE	Pakistan	660	South Asian	2385	5193
SAHLSIS	Sweden	750	European	299	596
SIFAP	Germany	2·5M	European	981	1825
SIGNET-REGARDS	USA	Affy 6.0	African	258	2094
SWISS/ISGS	USA	Illumina 610 or 660	African	173	389
UTRECHT	The Netherlands	ImmunoChip	European	556	1145
VHIR-FMT-BARCELONA	Spain	HumanCore and ExomeChip	European	545	320
WGHS‡	USA	Human Hap300 and custom array	European	440	22725
Total	..	..	..	21482	367842
Joint					
Total	..	..	..	38333	400315

Case cohorts in the first stage were matched to external controls based on genotyping array, cohort, ancestry, and location of sample collection. Case-control groups were constructed for the first stage analyses from contributing cohorts, which were mainly case-only or control-only cohorts. Hispanic samples were an exception and are not shown as a separate group here, because the small number of samples required that we pool all available Hispanic samples into a single analysis stratum. The second stage consisted of in-silico SNP lookups of summary-level results in previously analysed case-control sets. Totals represent the number of unique samples, accounting for partial sample overlap between two sites (CHARGE and WGHS). NINDS-SiGN=National Institute of Neurological Disorders and Stroke Stroke Genetics Network. BRAINS=Biorepository of DNA in Stroke. MGH-GASROS=Massachusetts General Hospital—Genes Affecting Stroke Risk and Outcome Study. ISGS=Ischemic Stroke Genetics Study. SWISS=Siblings with Ischemic Stroke Study. HABC=Health ABC. EDIN=Edinburgh Stroke Stoke. OXVASC=Oxford Vascular Study. STGEORGE=St George's Hospital. KORA=MONICA/KORA Ausburg Study. WTCCC=Wellcome Trust Case Control Consortium. GEOS=Genetics of Early Onset Stroke. GCNKSS=Greater Cincinnati/Northern Kentucky Stroke Study. MCISS=Middlesex County Ischemic Stroke Study. MIAMISR=Miami Stroke Registry and Biorepository. NHS=Nurses' Health Study. NOMAS=Northern Manhattan Study. REGARDS=Reasons for Geographic and Racial Differences in Stroke. SPS3=Secondary Prevention of Small Subcortical Strokes. WHI=Women's Health Initiative. WUSTL=Washington University St Louis. HRS=Health and Retirement Study. OAI=Osteoarthritis Initiative. HCHS/SOL=The Hispanic Community Health Study/Study of Latinos. LEUVEN=Leuven Stroke Genetics Study. BASICMAR=Base de Datos de Ictus del Hospital del Mar. ADHD=Attention-deficit Hyperactivity Disorder. INMA=Infancia y medio ambiente. SAHLSIS=Sahlgrenska Academy Study of Ischemic Stroke. LUND=Lund Stroke Registry. MDC=Malmo Diet and Cancer Study. ASGC=Australian Stroke Genetics Collaborative. VISP=Vitamin Intervention for Stroke Prevention. HANDLs=Health/Aging in Neighborhoods of Diversity across the Lifespan Study. ARIC=Atherosclerosis Risk in Communities Study. CADISP=Cervical Artery Dissection and Ischemic Stroke Patients. CHARGE=Cohorts for Aging and Research in Genetic Epidemiology. CHS=Cardiovascular Health Study. HVH=Heart and Vascular Health Study. GLASGOW=Glasgow ImmunoChip Study. RACE=Risk Assessment of Cardiovascular Events. SIFAP=Stroke in Young Fabry Patients. SIGNET=The Sea Island Genetics Network. UTRECHT=Utrecht ImmunoChip Study/PROMISe Study. WGHS=Women's Genome Health Study. *Only TOAST subtypes available for the first stage. †Contributing cohorts are described in the appendix. ‡Not included in the ischaemic stroke and cerebroembolism analyses because of overlap with CHARGE.

Table 1: **Case and control cohorts in NINDS-SiGN**

analysis to establish a group of non-Hispanic black (African ancestry) participants and a group of participants of Asian ancestry. Samples that were not grouped as European, African, or Asian ancestry formed the Hispanic stratum. We excluded samples of Asian ancestry from further analysis because of the small number. After establishing the ancestry-based composite groups, we did principal component analysis again to confirm the ancestral homogeneity within each case-control stratum. Case-control strata then underwent extensive quality control (appendix). Finally, each stratum was prephased[17] and imputed. We imputed samples of European ancestry using a merged reference panel that included the 1000 Genomes Project Phase I[18] and the Genome of the Netherlands.[19] We imputed samples in the African and Hispanic groups using the 1000 Genomes Project Phase I reference panel only. We added summary-level imputed data from an additional cohort (Vitamin Intervention for Stroke Prevention) to the first stage meta-analysis.

First stage genome-wide association analysis

After quality control and imputation, 16851 cases and 32473 controls from 15 ancestry-specific groups were available for genome-wide testing (table 1, appendix). Within each stratum, we analysed all ischaemic stroke phenotypes and the four main subtypes (cardioembolism, large artery atherosclerosis, small artery occlusion, and undetermined) as established with CCSc, CCSp, and TOAST, which were available for 12612 (74·8%) cases. All GWAS were adjusted for sex and the top ten principal components; genome-wide testing was not corrected for age, because age information was missing for most of the controls.

After the GWAS, we removed SNPs with frequency of less than 1% because they showed excessive genomic inflation. We checked the frequencies of imputed SNPs for consistency with the continental populations represented in the 1000 Genomes Project Phase I, and we removed SNPs with a difference in frequency of more than 30%. After quality control, 9·3 million to 15·4 million SNPs were available in the study strata for the meta-analysis. We did inverse variance-weighted fixed-effects meta-analysis across the 15 ancestry-specific strata using MANTEL[20] in each of the 15 traits. The genomic inflation factor λ of the 15 meta-analyses for each trait ranged from 0·936 to 1·005 (appendix pp 5–8).

Second stage analysis

In the second stage, we performed in-silico lookups of association results in 18 independent studies that contained 20941 TOAST-subtyped cases and 364736 controls, using the nominally significant SNPs identified in the first stage (table 1 and appendix p 51). The SNPs selected for the second stage for each subtype were aggregated such that, for example, SNPs with $p<1\times10^{-6}$ from the three cardioembolism GWAS (CCSc,

CCSp, and TOAST) were all selected for lookup in the independent TOAST cardioembolism cases and matched controls. This process was repeated for the other subtypes.

Joint analysis

We did a meta-analysis of the results from the in-silico lookups from the second stage and the results from the first stage. We set the threshold for genome-wide significance in the joint analysis at $p<1\times10^{-8}$, after correction for testing of the five phenotypes (all stroke, cardioembolic, large artery atherosclerosis, small artery occlusion, and undetermined). λ in the ischaemic stroke joint analysis was 1·005 and ranged from 0·936 to 0·998 in the subtype analyses (appendix pp 9–12).

Role of the funding source

The funder participated in the design of the study. The study investigators were solely responsible for the data collection, analysis, and interpretation. An employee of NINDS (KG) was a member of the writing committee The analysis team had full access to all data included in the study. The steering committee had final responsibility for the decision to submit the report for publication.

Results

After data quality control (appendix p 4 and pp 114–26), we included 16 851 stroke cases and 32 473 controls in the first stage of our analyses. The first stage GWAS revealed 1372 SNPs in 268 loci associated with ischaemic stroke or a specific subtype in any of the CCS or TOAST traits at $p<1\times10^{-6}$. We included an additional independent set of 20 941 cases and 364 736 controls in the second stage, which enabled the joint analysis of 37 893 cases and 397 209 controls across five primary independent traits (ischaemic stroke and the four subtypes).

Genome-wide *Z* scores (SNP β values divided by their respective SE) from the CCSc, CCSp, and TOAST GWAS were checked for correlation (Pearson's *r*) between each possible pair of traits. The analysis revealed moderate to strong genetic correlation (figure 1) between the standardised SNP effects in CCSc, CCSp, and TOAST, despite the modest phenotypic correlation noted previously.[21] The moderate to strong genetic correlation between CCS and TOAST within subtype-specific clusters suggested that TOAST subtyping was appropriate for inclusion in the second stage of the analysis. Phenotypic correlations were also strong within subtype-specific clusters (figure 1).

In the joint analysis of CCS (first stage) and TOAST (second stage) results, SNPs in two novel loci exceeded genome-wide significance. Four common SNPs in linkage disequilibrium ($r^2>0\cdot57$ in the 1000 Genomes Project samples of European ancestry) near the *TSPAN2* locus on chromosome 1 were associated at genome-wide significance with large artery atherosclerosis. The lead

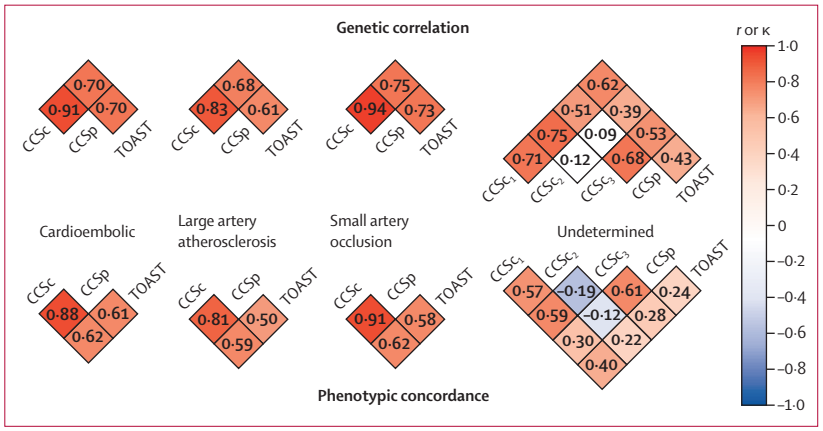


Figure 1: **Genetic and phenotypic correlation between subtyping methods in the first stage analysis**
All cases with an available CCS subtype were included in the first stage analyses. Genome-wide *Z* scores from the CCSc, CCSp, and TOAST GWAS were checked for correlation between each possible pair of traits. Pearson's *r* correlation coefficient (mathematically equivalent in this scenario to the Lin's concordance correlation coefficient) within each square shows genetic correlation. Cohen's κ within each square shows phenotypic agreement. $CCSc_1$ includes all undetermined strokes; $CCSc_2$ includes all incomplete and unclassified strokes; and $CCSc_3$ includes all cryptogenic and cardioembolic minor strokes. The $CCSc_2$ and $CCSc_3$ classifications are mutually exclusive. CCS=Causative Classification of Stroke. CCSc=CCS causative. CCSp=CCS phenotypic. TOAST=Trial of Org 10172 in Acute Stroke Treatment classification system. GWAS=genome-wide association study.

SNP in the associated locus was rs12122341 (odds ratio [OR] for the G allele 1·19, 95% CI 1·12–1·26, $p=1\cdot3\times10^{-9}$; figure 2, table 2).

A second locus emerged as having a genome-wide significant association with ischaemic stroke, but only in samples of African ancestry. In view of the small sample size in which it was identified, the association must be interpreted with caution. rs74475935 in *ABCC1* on chromosome 16 was associated with the undetermined phenotype (table 2, appendix p 14), driven by a variant with rare frequency (minor allele frequency [MAF] about 0·01%) in European-ancestry samples and low frequency (MAF about 1·5%) in African-ancestry samples.

We also identified associations for the previously reported loci *PITX2*[4] and *ZFHX3*[5] for cardioembolic stroke, and *HDAC9*[6] for large artery atherosclerotic stroke, all of which exceeded genome-wide significance in our samples (table 2). The 12q24.12 locus near *ALDH2*, previously reported to be associated with all ischaemic stroke, but not with any specific subtype,[7] exceeded genome-wide significance in the joint analysis of all ischaemic stroke (OR for the T allele 1·07, 95% CI 1·5–1·09, $p=4\cdot20\times10^{-9}$). However, the association was even stronger for small artery occlusion in the joint analysis of CCSp in the first stage and TOAST in the second stage (OR 1·17, 95% CI 1·11–1·23, $p=2\cdot92\times10^{-9}$); the association was not genome-wide significant in the joint analysis of CCSc (first stage) and TOAST (second stage; OR 1·16, 95% CI 1·10–1·22, $p=2\cdot77\times10^{-8}$). Evidence of associations with other subtypes was reduced in our study (OR<1·1 and $p>4\times10^{-3}$ for cardioembolism, large artery atherosclerosis, and undetermined in the combined CCSp and TOAST analysis; appendix p 15). Systematic testing that accounted for shared controls

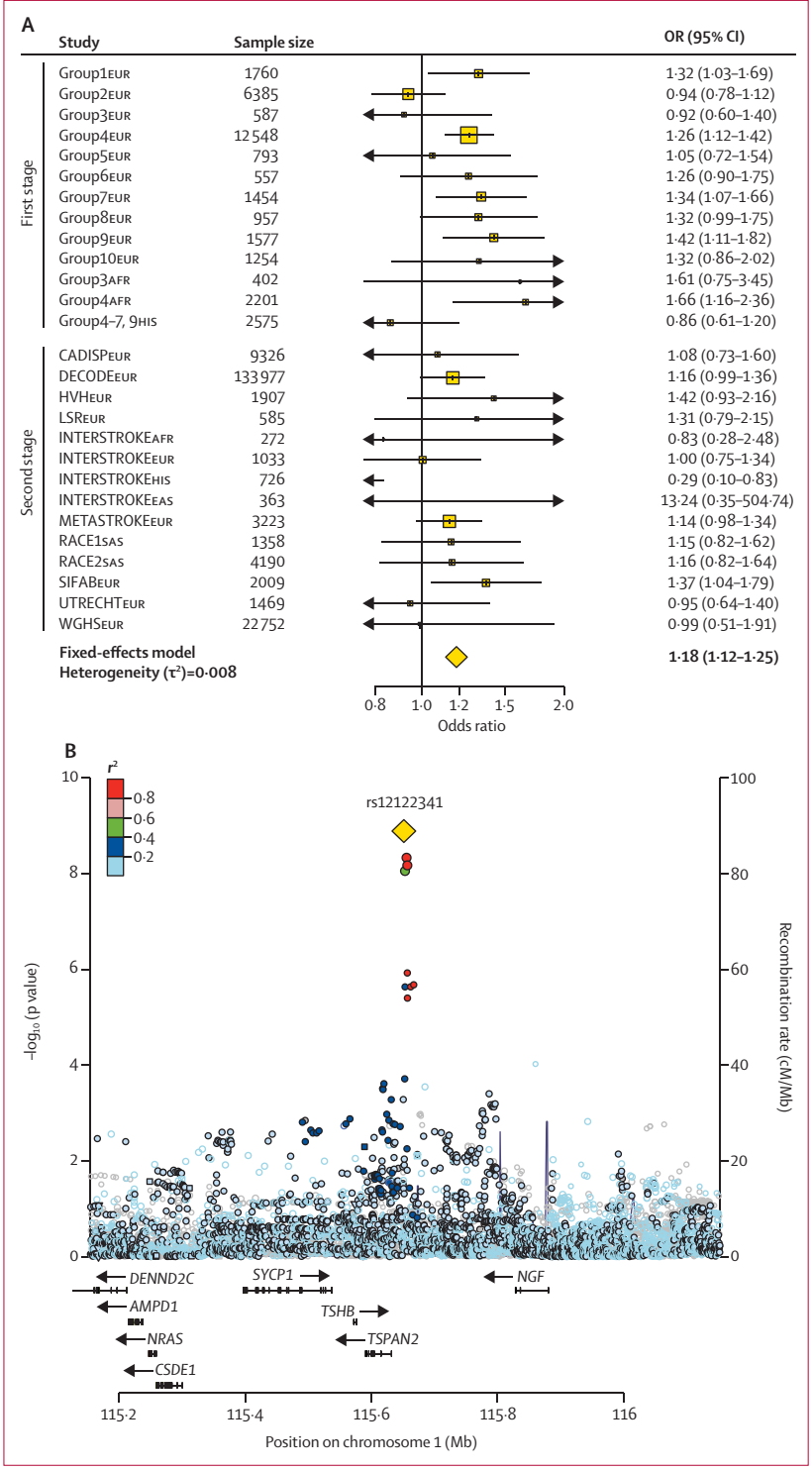


***Figure 2*: Forest plot (A) and regional association plot (B) of rs12122341**
Plot of effect size of the association of rs12122341 with large artery atherosclerosis-related stroke across the case–control groups included in the fist and second stage analyses (A). Association of rs12122341 and other SNPs in the region with large artery atherosclerosis-related stroke (B). Point shading shows linkage disequilibrium (r^2) to rs12122341 as calculated in 1000 Genomes Project Phase I European samples. Purple lines show recombination rate. EUR=European ancestry. AFR=African ancestry. HIS=Hispanic samples. EAS=east Asian ancestry. SAS=south Asian ancestry.

(appendix p 15) showed a significant difference in the magnitude of ORs between small artery occlusion and the combined non-small artery occlusion subtypes (p=0·048, appendix p 15), suggesting that the effect of 12q24.12 might be specific for small artery occlusion.

By contrast, we did not find any evidence for the previously reported association between ischaemic stroke and *NINJ2* (rs34166160, OR for the A allele 1·20, 95% CI 0·96–1·48, p=0·106; table 2), even though our sample size had 100% power to detect an association (p<0·05) at this locus. In the full first stage analysis, evidence for association was weak for both the 6p21[22] and *CDKN2B-AS1*[23] loci in large artery atherosclerosis, and for the *ABO*[24] locus in all ischaemic stroke, large artery atherosclerosis, and cardioembolism (table 2). When we restricted our analysis to only the samples not used for the initial discovery (appendix p 52), *CDKN2B-AS1* was associated with large artery atherosclerosis (OR for the G allele 1·09, 95% CI 1·02–1·17, p=0·009) and *ABO* was associated with all ischaemic stroke (OR for the C allele 1·07, 95% CI 1·03–1·10, $p=2·5\times10^{-4}$), large artery atherosclerosis (OR 1·15, 95% CI 1·07–1·24, $p=2·5\times10^{-4}$), and cardioembolism (OR 1·09, 95% CI 1·02–1·16, p=0·007). For 6p21, however, we detected no evidence for any association with large artery atherosclerosis (OR for the T allele 1·04, 95% CI 0·96–1·12, p=0·304).

Discussion

Our results show a novel association between a genetic locus and large artery atherosclerosis. The lead SNP, rs12122341, is located in an intergenic region 23·6 kb upstream of *TSPAN2*, the gene encoding tetraspanin-2 (figure 2) This SNP is in linkage disequilibrium with intronic and untranslated region variants in *TSPAN2* (r^2>0·3 in 1000 Genomes Project samples of European ancestry), but is located in a DNA sequence immediately adjacent to *TSPAN2* that can be bound by several transcription factor proteins, including CTCF. This sequence is a promotor and enhancer site that is marked by histone modification and DNase hypersensitivity according to experimental data from ENDCODE and ROADMAP Epigenomics (appendix p 16),[25,26] suggesting a potential role for rs12122341 in gene regulation. An intergenic SNP near rs12122341 has been reported to be associated with migraine,[27] but the two SNPs are not in linkage disequilibrium (r^2=0·03 in 1000 Genomes Project samples of European ancestry).

TSPAN2, the gene closest to rs12122341, is a member of the transmembrane 4 (tetraspanin) superfamily. This family of proteins can mediate signal transduction to regulate cell development, activation, growth, and motility. *TSPAN2* knock-out mice have increased neuroinflammation, shown by activation of microglia and astrocytes with no effect on myelination and axon integrity.[28] Notably, *TSPAN2* is highly expressed in artery tissue and whole blood cells (appendix p 16), which accords with the association we detected between

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
Novel loci																		
Large artery atherosclerosis																		
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	CCSc	2454	1·20 (1·12–1·29)	$3·38×10^{-7}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	CCSc	4703	1·18 (1·12–1·25)	$8·32×10^{-9}$
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	CCSp	2715	1·21 (1·13–1·30)	$4·50×10^{-8}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	CCSp	4964	1·19 (1·12–1·26)	$1·30×10^{-9}$
rs12122341	1	G	25·7	8·8	19·5	*TSPAN2*	TOAST	2346	1·15 (1·07–1·24)	$1·61×10^{-4}$	TOAST	2249	1·15 (1·04–1·26)	$5·25×10^{-3}$	TOAST	4595	1·15 (1·08–1·22)	$2·70×10^{-6}$
Undetermined																		
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	CCSc	2392*	5·17 (2·99–8·92)	$3·69×10^{-9}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	CCSc	5861	4·63 (2·77–7·72)	$4·70×10^{-11}$
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	CCSp	1062*	8·68 (4·55–16·58)	$5·94×10^{-11}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	CCSp	4531	6·89 (3·80–12·47)	$1·85×10^{-10}$
rs74475935	16	G	0·2	1·8	0·6	*ABCC1*	TOAST	3593	2·18 (1·16–4·10)	$1·58×10^{-2}$	TOAST	3469	1·87 (0·55–6·41)	$3·16×10^{-1}$	TOAST	7062	2·11 (1·20–3·70)	$9·22×10^{-3}$
Previously identified loci, first stage $p<1×10^{-6}$																		
All ischaemic stroke																		
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	..	16 851	1·10 (1·06–1·13)	$3·07×10^{-8}$	..	21042	1·05 (1·01–1·08)	$6·55×10^{-3}$		37893	1·07 (1·5–1·09)	$4·20×10^{-9}$
rs2634074	4	T	2·1	4·8	4·1	*PITX2*	..	16 851	1·09 (1·06–1·13)	$2·56×10^{-7}$	..	21042	1·10 (1·07–1·14)	$2·00×10^{-8}$		37893	1·10 (1·07–1·12)	$2·68×10^{-14}$
rs2107595	7	A	15·7	2·2	2·2	*HDAC9*	..	16 851	1·10 (1·06–1·14)	$7·74×10^{-7}$	..	21042	1·07 (1·03–1·11)	$1·70×10^{-4}$		37893	1·09 (1·06–1·12)	$8·60×10^{-10}$
Cardioembolism																		
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	CCSc	3071	1·39 (1·28–1·50)	$1·24×10^{-16}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	CCSc	7062	1·37 (1·30–1·45)	$1·04×10^{-29}$
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	CCSp	3695	1·39 (1·29–1·49)	$3·26×10^{-19}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	CCSp	7686	1·37 (1·30–1·45)	$2·79×10^{-32}$
rs2200733	4	T	12·0	2·2	2·6	*PITX2*	TOAST	3427	1·37 (1·27–1·48)	$1·02×10^{-16}$	TOAST	3991	1·36 (1·26–1·46)	$1·21×10^{-16}$	TOAST	7418	1·36 (1·29–1·44)	$8·05×10^{-30}$
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	CCSc	3071	1·17 (1·09–1·26)	$1·12×10^{-5}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	CCSc	7062	1·17 (1·10–1·22)	$7·28×10^{-9}$

(Table 2 continues on next page)

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
(Continued from previous page)																		
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	CCSp	3695	1·19 (1·11–1·27)	$2·93×10^{-7}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	CCSp	7686	1·17 (1·11–1·23)	$2·29×10^{-10}$
rs7193343	16	T	17·4	2·4	18·9	*ZFHX3*	TOAST	3427	1·17 (1·09–1·25)	$1·45×10^{-5}$	TOAST	3991	1·15 (1·07–1·23)	$7·93×10^{-5}$	TOAST	7418	1·16 (1·10–1·22)	$8·88×10^{-9}$
Large artery atherosclerosis																		
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	CCSc	2454	1·30 (1·18–1·42)	$8·46×10^{-8}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	CCSc	4703	1·23 (1·15–1·33)	$1·10×10^{-8}$
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	CCSp	2715	1·29 (1·18–1·42)	$3·50×10^{-8}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	CCSp	4964	1·24 (1·15–1·33)	$4·52×10^{-9}$
rs11984041	7	T	9·3	2·2	6·7	*HDAC9*	TOAST	2346	1·30 (1·17–1·43)	$3·62×10^{-7}$	TOAST	2249	1·15 (1·03–1·29)	$1·16×10^{-2}$	TOAST	4595	1·23 (1·14–1·33)	$4·48×10^{-8}$
Small artery occlusion																		
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	CCSc	2736	1·19 (1·11–1·27)	$9·10×10^{-7}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	CCSc	5162	1·16 (1·10–1·22)	$2·77×10^{-8}$
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	CCSp	2734	1·20 (1·12–1·28)	$6·82×10^{-8}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	CCSp	5160	1·17 (1·11–1·23)	$2·92×10^{-9}$
rs10744777	12	T	66·7	4·5	5·2	*ALDH2*	TOAST	3147	1·13 (1·06–1·21)	$1·05×10^{-4}$	TOAST	2426	1·12 (1·03–1·21)	$4·66×10^{-3}$	TOAST	5573	1·13 (1·07–1·18)	$1·62×10^{-6}$
Previously identified loci, first stage $p>1×10^{-6}$																		
All ischaemic stroke																		
rs34166160	12	A	0·9	0·0	0·3	*NINJ2*	..	16 851	1·20 (0·96–1·48)	$1·06×10^{-1}$	..	..	..	..	..	..	..	..
rs11833579	12	G	75·8	79·4	68·0	*NINJ2*	..	16 851	1·02 (0·95–1·01)	$2·15×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	..	16 851	1·07 (1·04–1·10)	$2·03×10^{-5}$	..	..	..	..	..	..	..	..

(Table 2 continues on next page)

	Chromosome	Risk allele	Risk allele frequency (%)			Nearest gene	First stage				Second stage				Joint analysis			
			European	African	The Americas		Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value	Subtyping system	Cases	OR (95% CI)	p value
(Continued from previous page)																		
Cardioembolism																		
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSc	3071	1·04 (0·98–1·10)	$1·88×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSp	3695	1·04 (0·98–1·10)	$1·62×10^{-1}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	TOAST	3427	1·08 (1·02–1·15)	$5·66×10^{-3}$	..	..	..	..	..	..	..	..
Large artery atherosclerosis																		
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSc	2454	1·09 (1·02–1·17)	$6·93×10^{-3}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	CCSp	2715	1·11 (1·04–1·18)	$1·29×10^{-3}$	..	..	..	..	..	..	..	..
rs505922	9	C	35·1	32·6	23·5	*ABO*	TOAST	2346	1·14 (1·06–1·21)	$2·15×10^{-4}$	..	..	..	..	..	..	..	..
rs556621	6	T	29·1	8·1	40·7	6p21	CCSc	2454	1·04 (0·97–1·11)	$3·18×10^{-1}$	..	..	..	..	..	..	..	..
rs556621	6	T	29·1	8·1	40·7	6p21	CCSp	2715	1·02 (0·95–1·19)	$6·36×10^{-1}$	..							
rs556621	6	T	29·1	8·1	40·7	6p21	TOAST	2346	1·11 (1·04–1·19)	$2·55×10^{-3}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	CCSc	2454	1·12 (1·05–1·19)	$4·34×10^{-4}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	CCSp	2715	1·11 (1·05–1·19)	$7·93×10^{-4}$	..	..	..	..	..	..	..	..
rs2383207	9	G	49·9	4·5	41·3	*CDKN2B-AS1*	TOAST	2346	1·09 (1·02–1·17)	$8·13×10^{-3}$	..	..	..	..	..	..	..	..

For subtype-specific loci, ORs and their corresponding p values are reported for the CCSc, CCSp, and TOAST subtypes. Risk allele frequency was calculated with 1000 Genomes (Phase I) European-ancestry samples, African-ancestry samples, and samples from the Americas. Association results were looked up in TOAST-subtyped cases and their matched controls and meta-analysed with the first stage results from CCSc, CCSp, and TOAST cases. CCS=Causative Classification of Stroke. TOAST=Trial of Org 10172 in Acute Stroke Treatment classification system. CCSc=CCS causative. CCSp=CCS phenotypic. OR=odds ratio. *Results from the CCS cryptogenic phenotype.

Table 2: **Novel and previously identified loci implicated in ischaemic stroke and its subtypes through genome-wide testing**

TSPAN2 with large artery atherosclerosis stroke. Whether the association of rs12122341 is caused by the locus' regulation of *TSPAN2* or other nearby genes will need further functional assessment.

The additional locus that we identified as being associated with undetermined stroke (rs74475935) is in a gene-rich region with linkage-disequilibrium-paired SNPs (r^2>0·1 in 1000 Genomes Project samples of African ancestry) of up to 4 Mb. Because of the small sample size for rs74475935 (610 cases) and the shortage of samples from people with African ancestry, studies with large samples from people of African descent will be necessary to fully assess the robustness of this signal.

So far, only four loci—*PITX2*,[4] *ZFHX3*,[5] *HDAC9*,[6] and 12q24.12[7]—have been repeatedly identified in GWAS of ischaemic stroke, all of which are subtype specific except for 12q24.12. Although the 12q locus association was originally identified for all ischaemic stroke, our analysis suggests that it is probably specific to small artery occlusion. These findings suggest that ischaemic stroke subtypes have distinct genetic signatures. Our analysis of genetic correlation across the traits also showed that the subtypes share subtle genetic associations (appendix p 17 and p 53). This finding is supported by the results of another study, which identified genetic overlap between the large artery atherosclerosis and small artery occlusion subtypes.[29] Future efforts will help to clarify both the shared and unique genetic architectures within and between subtypes.

Until now, GWAS of ischaemic stroke subtypes have used far smaller sample sizes than studies of other complex traits. The SiGN study, the largest GWAS of ischaemic stroke so far, was well powered (75·1%) to detect common SNP subtype-specific associations of larger effect (MAF 25% and OR 1·2 in 3000 cases and 30 000 controls) but was substantially less powered to identify lower frequency or lower effect SNPs (13·8% power for MAF 10% and OR 1·2; 1·1% power for MAF 25% and OR 1·1). Because of the almost linear relation that exists between sample size and discovered loci,[30] and because large-scale GWAS in other complex traits have yielded hundreds of SNP-disease associations,[31–33] studying ischaemic stroke subtypes in larger samples will probably yield additional associated common variants. Furthermore, the implementation of whole genome sequencing studies of stroke will begin to test whether rare alleles in the population account for a substantial proportion of disease heritability.

The SiGN study has several other limitations. First, sample inclusion was heavily biased towards individuals of European descent; inclusion of non-European populations will improve power for locus discovery[34] and will be especially informative for future fine-mapping efforts.[35] Second, the inclusion of TOAST-based classification for samples in the second stage probably added phenotypic heterogeneity (figure 1, appendix p 53), which potentially reduced power.[21] Third, many of the participating studies within SiGN (especially the publicly available controls) had little or no stroke-specific risk factor data available. Such data are key to disentangling potential gene–environment interactions. Future genetic studies of stroke will continue to face challenges related to the disease phenotype, including high prevalence of the disease (lifetime risk about 20%), its late onset (mainly in individuals >65 years), the contribution of other cardiovascular diseases and environment as causative factors, and difficulties of subtyping (in SiGN 12·6–22·3% of all cases analysed were ultimately classified as undetermined by CCS or TOAST).

Our use of CCS enabled identification of candidate SNPs that were not significant for the second stage follow-up in TOAST, including those SNPs at the *TSPAN2* locus. This refinement might represent a reduction in phenotypic heterogeneity that CCS introduces through its capture of clinical stroke features, completeness of diagnostic investigations, and, where possible, classification of cases with different potential causes into the most probable causes. The association signal of the *TSPAN2* locus identified with CCS was, however, improved by the inclusion of TOAST-classified samples, suggesting that making use of the genetic correlation underlying the subtyping methods and allowing for broader inclusion of cases, regardless of subtyping system, can lead to the discovery of more susceptibility loci. Further studies will help to establish whether the rich repository of individual-level data created through the use of the CCS will help to uncover novel phenotypes and thus reveal biological mechanisms and broaden the understanding of the genetic architecture in patients with stroke.

Contributors

JR, BDM, HA, PIWdB, SJK, AL, JFM, SLP, CLMS, VT, DW, and BBW contributed to data collection and provided critical review of the manuscript. JR, BDM, HA, PIWdB, KG, SJK, AL, SLP, CLMS, VT, DW, and BBW made critical decisions regarding study design and conduct. JR, BDM, HA, PIWdB, KG, SJK, AL, SLP, CLMS, VT, DW, and BBW participated in literature search and writing of the paper. BDM, PIWdB, and SLP did the statistical analysis and data interpretation.

NINDS Stroke Genetics Network (SiGN) and International Stroke Genetics Consortium Writing committee Jonathan Rosand (chair), Braxton D Mitchell (co-chair), Hakan Ay, Paul I W de Bakker, Katrina Gwinn, Steven J Kittner, Arne Lindgren, James F Meschia, Sara L Pulit, Cathie L M Sudlow, Vincent Thijs, Daniel Woo, Bradford B Worrall.

Steering committee Donna K Arnett, Oscar Benavente, John W Cole, Martin Dichgans, Raji P Grewal, Christina Jern, Jordi Jiménez Conde, Julie A Johnson, Steven J Kittner, Jin-Moo Lee, Christopher Levi, Arne Lindgren, Hugh S Markus, Olle Melander, James F Meschia, Kathryn Rexrode, Jonathan Rosand, Peter M Rothwell, Tatjana Rundek, Ralph L Sacco, Reinhold Schmidt, Pankaj Sharma, Agnieszka Slowik, Cathie L M Sudlow, Vincent Thijs, Sylvia Wassertheil-Smoller, Daniel Woo, Bradford B Worrall.

Declaration of interests

KG is an employee of NINDS. The other members of the writing committee declare no competing interests.

Acknowledgments

The SiGN study was funded by a cooperative agreement grant from the US National Institute of Neurological Disorders and Stroke, National Institutes of Health (U01 NS069208). The information about funding for each collection is reported in the appendix (pp 159–70).

References

1 World Health Organization. The top 10 causes of death. May 2014. http://www.who.int/mediacentre/factsheets/fs310/en/ (accessed April 23, 2015).

2 Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation* 2014; **131:** e29–322.

3 Meschia JF, Bushnell C, Boden-Albala B, et al. Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2014; **45:** 3754–832.

4 Gretarsdottir S, Thorleifsson G, Manolescu A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* 2008; **64:** 402–09.

5 Gudbjartsson DF, Holm H, Gretarsdottir S, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet* 2009; **41:** 876–78.

6 Bellenguez C, Bevan S, Gschwendtner A, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* 2012; **44:** 328–33.

7 Kilarski LL, Achterberg S, Devan WJ, et al. Meta-analysis in more than 17,900 cases of ischemic stroke reveals a novel association at 12q24.12. *Neurology* 2014; **83:** 678–85.

8 Bevan S, Traylor M, Adib-Samii P, et al. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke* 2012; **43:** 3161–67.

9 Meschia JF, Arnett DK, Ay H, et al. Stroke Genetics Network (SiGN) study: design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke* 2013; **44:** 2694–702.

10 Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38:** 209–13.

11 Adams H, Bendixen B, Kappelle L, 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:** 35–41.

12 Ay H, Benner T, Arsava EM, et al. A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke* 2007; **38:** 2979–84.

13 Kolominsky-Rabas PL, Weber M, Gefeller O, Neundoerfer B, Heuschmann PU. Epidemiology of ischemic stroke subtypes according to TOAST criteria: incidence, recurrence, and long-term survival in ischemic stroke subtypes: a population-based study. *Stroke* 2001; **32:** 2735–40.

14 Arsava EM, Ballabio E, Benner T, et al. The causative classification of stroke system: an international reliability and optimization study. *Neurology* 2010; **75:** 1277–84.

15 Ay H, Arsava EM, Andsberg G, et al. Pathogenic ischemic stroke phenotypes in the NINDS-Stroke Genetics Network. *Stroke* 2014; **45:** 3589–96.

16 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38:** 904–09.

17 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012; **44:** 955–59.

18 Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491:** 56–65.

19 Francioli LC, Menelaou A, Pulit SL, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; **46:** 818–25.

20 de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17:** R122–28.

21 McArdle PF, Kittner SJ, Ay H, et al. Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. *Neurology* 2014; **83:** 1653–60.

22 Holliday EG, Maguire JM, Evans T-J, et al. Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet* 2012; **44:** 1147–51.

23 Smith JG, Melander O, Lövkvist H, et al. Common genetic variants on chromosome 9p21 confers risk of ischemic stroke a large-scale genetic association study. *Circ Cardiovasc Genet* 2009; **2:** 159–64.

24 Williams FMK, Carter AM, Hysi PG, et al. Ischemic stroke is associated with the ABO locus: the EuroCLOT study. *Ann Neurol* 2013; **73:** 16–31.

25 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518:** 317–30.

26 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012; **40:** D930–34.

27 Esserlind AL, Christensen AF, Le H, et al. Replication and meta-analysis of common variants identifies a genome-wide significant locus in migraine. *Eur J Neurol* 2013; **20:** 765–72.

28 de Monasterio-Schrader P, Patzig J, Möbius W, et al. Uncoupling of neuroinflammation from axonal degeneration in mice lacking the myelin protein tetraspanin-2. *Glia* 2013; **61:** 1832–47.

29 Holliday EG, Traylor M, Malik R, et al. Genetic overlap between diagnostic subtypes of ischemic stroke. *Stroke* 2015; **46:** 615–19.

30 Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet* 2012; **90:** 7–24.

31 Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511:** 421–27.

32 Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46:** 1173–86.

33 Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013; **45:** 1274–83.

34 Pulit SL, Voight BF, de Bakker PIW. Multiethnic genetic association studies improve power for locus discovery. *PLoS One* 2010; **5:** e12600.

35 Zaitlen N, Paşaniuc B, Gur T, Ziv E, Halperin E. Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet* 2010; **86:** 23–33.

5 MANUSCRIPT 3

"Functional characterization of an atherosclerosis associated noncoding variant at the *HDAC9* locus."

Prestel, M., Prell-Schicker, C., Webb, T., Malik, R., Lindner, B., Ziesch, N., Rex-Haffner, M., Röh, S., Viturawong, T., **Lehm, M.**, Mokry, M., den Ruijter, H., Haitjema, S., Asare, Y., Söllner, F., Najafabadi, M., Civelek, M., Samani, N., Mann, M., Haffner, C., Dichgans, M.

This manuscript is being prepared for submission at a peer-reviewed journal.

Author contribution:

MP, CPS, TW, RM, ML, CH and MD designed, performed and analyzed the experiments. MP, CPS, RM, ML, CH and MD interpreted the results. MP, CPS, RM, ML and MD wrote and edited the manuscript.

Functional characterization of an atherosclerosis associated noncoding variant at the *HDAC9* locus

Short title: Functional noncoding variant at HDAC9

Authors:
Matthias Prestel, PhD[1*], Caroline Prell-Schicker, PhD[1*], Tom Webb, PhD[2], Rainer Malik, PhD[1], Barbara Lindner[1], Natalie Ziesch[1], Monika Rex-Haffner, B.Sc.[3], Simone Röh, Dipl.[3], Thanatip Viturawong, PhD[4], Manuel Lehm, MD[1,4,5], Michal Mokry, MD, PhD[6], Hester den Ruijter, PhD[7], Saskia Haitjema, MD[7], Yaw Asare, PhD[1], Flavia Söllner, M.A., M.Sc.[1,8], Maryam Ghaderi Najafabadi, M.Sc.[2], Rédouane Aherrahrou, PhD[10], Mete Civelek, PhD[10], Nilesh J. Samani, MD[2], Matthias Mann, PhD[4], Christof Haffner, PhD[1], Martin Dichgans, MD[1,9]

Affiliation:
[1]Institute for Stroke and Dementia Research, Klinikum der Universität München, 81377 Munich, Germany
[2]Department of Cardiovascular Sciences, University of Leicester and National Institute for Health Research Leicester Biomedical Research Centre, Leicester, United Kingdom, Glenfield Hospital, Leicester, LE3 9QP, UK
[3]Department of Translational Research in Psychiatry, Max-Planck-Institute for Psychiatry, 80804 Munich, Germany
[4]Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, 82152 Martinsried, Germany
[5]Abteilung für Diagnostische und Interventionelle Neuroradiologie, Klinikum rechts der Isar, 81675 Munich, Germany
[6]Department of Pediatrics, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands
[7]Laboratory of Experimental Cardiology, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands
[8]Department of Physiological Chemistry, Biomedical Center Munich, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany
[9]Munich Cluster for Systems Neurology (SyNergy), 81377 Munich, Germany
[10] Center for Public Health Genomics, Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22904-4741

* These authors contributed equally to this article.

Address of correspondence:
Prof. Dr. med. Martin Dichgans
Institute for Stroke and Dementia Research
Feodor-Lynen-Straße 17
81377 Munich
Germany
martin.dichgans@med.uni-muenchen.de
phone: +49-89-4400-46019
fax: +49-89-4400-46010

Total word count: 7992(8000)

Subjects codes:
Gene Expression and Regulation, Ischemic Stroke, Atherosclerosis, Mechanisms

ABSTRACT

Rationale: Genome-wide association studies (GWAS) have identified the histone deacetylase 9 (*HDAC9*) gene region as a major risk locus for atherosclerotic stroke and coronary artery disease in humans. Gene expression studies and data from mouse models of atherosclerosis suggest a role of altered HDAC9 expression levels as the underlying disease mechanism. rs2107595, the lead single nucleotide polymorphism (SNP) in recent GWAS for stroke and coronary artery disease resides in noncoding DNA and colocalizes with histone modification marks suggestive of enhancer elements.
Objective: To determine the mechanisms by which genetic variation at rs2107595 regulates HDAC9 expression and thus vascular risk.
Methods and Results: Targeted resequencing of the HDAC9 locus in patients with atherosclerotic stroke and controls supported candidacy of rs2107595 as the causative SNP at this locus. A search for nuclear binding partners by proteome-wide analysis revealed preferential binding of the E2F3/TFDP1/Rb1 complex to the rs2107595 common allele, consistent with the disruption of an E2F3 consensus site by the risk allele. Gain- and loss-of-function studies showed a regulatory effect of E2F and Rb proteins on HDAC9 gene expression. Compared to the common allele the rs2107595 risk allele exhibited higher transcriptional capacity in luciferase assays in Jurkat cells and THP-1 macrophages and was associated with higher HDAC9 mRNA levels in primary macrophages and genome-edited Jurkat cells. Circularized chromosome conformation capture revealed a genomic interaction of the rs2107595 region with the *HDAC9* promoter, which was stronger for the common allele as was the *in vivo* interaction with E2F3 and Rb1 determined by chromatin immunoprecipitation. Gain-of-function experiments in isogenic Jurkat cells demonstrated a key role of E2F3 in mediating rs2107595-dependent transcriptional regulation of HDAC9.
Conclusions: Collectively, our findings imply allele-specific transcriptional regulation of HDAC9 via E2F3 and Rb1 as a major mechanism mediating vascular risk at rs2107595.

KEY WORDS

Gene regulation, large artery stroke, HDAC9

INTRODUCTION

Stroke is the leading cause of permanent disability and the second most common cause of death worldwide.[1, 2] Genome-wide association studies (GWAS) have mapped more than 35 genomic loci for stroke most residing in noncoding DNA.[3, 4] However, at many loci the causal variant, gene, and mechanism remain undetermined[5, 6] thus impeding the identification of novel pathways and possible targets for intervention. The histone deacetylase 9 (*HDAC9*) gene region on chromosome 7p21.1 represents the strongest risk locus for atherosclerotic stroke (large artery stroke)[3, 7] and has further been established as a major risk locus for myocardial infarction, coronary artery disease,[8] and peripheral artery disease,[9] thus implying a broader involvement in atherosclerosis and a major impact on human health.[10]

rs2107595, the lead single nucleotide polymorphism (SNP) in recent GWAS for stroke[3, 11, 12] and coronary artery disease[8] resides in noncoding DNA 3' to the *HDAC9* gene. rs2107595 colocalizes with DNase I hypersensitive sites (DHS) and histone modification marks H3K27ac and H3K4me1 (ENCODE,[13] genome build hg19) indicating a possible involvement in gene regulatory mechanisms.[14-16]

We and others recently provided evidence for a central role of HDAC9 expression levels in atherogenesis and stroke: first, *Hdac9* deficiency attenuates atherogenesis in mouse models of atherosclerosis.[17, 18] Second, HDAC9 expression levels were found to be elevated in human atherosclerotic plaques.[19] Third, gene expression studies in peripheral blood mononuclear cells (PBMCs) revealed an association between the rs2107595 risk allele and elevated levels of HDAC9 mRNA expression with a gene dosage effect.[18] The same variant further associates with both carotid intima media thickness and the presence of atherosclerotic plaques in the common carotid artery.[19, 20] Collectively, these findings point to the possibility that the rs2107595 region mediates disease risk by influencing HDAC9 expression levels.

In the current study, we aimed to elucidate the molecular mechanisms linking genetic variation in the rs2107595 region to HDAC9 expression. For this we employed targeted resequencing of the HDAC9 locus, proteome-wide search for allele-specific nuclear binding partners, chromatin immunoprecipitation (ChIP), genome-editing, reporter assays, circularized chromosome conformation capture (4C), and gain- and loss-of-function experiments in cultured human cell lines and primary vascular and immune cells. We provide evidence for a regulatory effect of rs2107595 on HDAC9 expression involving a direct physical interaction between the rs2107595 region and the *HDAC9* promoter. We further demonstrate a role of the E2F3 and Rb1 proteins in mediating allele-specific effects of rs2107595 on HDAC9 transcription.

METHODS

Targeted Resequencing

192 patients with large artery stroke and 192 age- and gender-matched controls were chosen for targeted resequencing of the *HDAC9* gene locus. Barcoded libraries were generated from fragmented genomic DNA (~200 bp, Covaris S2 sonifier) using a fragment library preparation kit (Applied Biosystems). Libraries were amplified on a Gene Amp PCR System 9700 (Applied Biosystems), subjected to a quality control on a 2100 Bioanalyzer (Agilent Technologies) and hybridized twice to a pool of biotinylated capture probes from a TargetSeq™ Custom Enrichment Kit (Applied Biosystems) designed to cover the region 18,123,000-19,188,000 at chromosome 7 (GRCh37/hg19). Bound DNA fragments were recovered by streptavidin Dynabeads (Thermo Fisher Scientific), quantified by qPCR using a TaqMan Quantification kit (Applied Biosystems), subjected to an emulsion PCR on a SOLiD EZ Bead System (E120 scale), enriched to a concentration of approximately 1.5 million beads and sequenced on a 5500 SOLiD xl system (Applied Biosystems).

Coverage of the sequence was over 85% in 352 individuals (176 cases, 176 controls), 32 individuals were removed from the analysis due to low coverage (Supplemental Figure 3). Mean alignment rate of reads was generally >90% and approximately one third of reads were ultimately used after removal of duplicates. After alignment and post-processing, the GATK software suite was used for quality control, re-alignment and re-calibration. SNPs were subsequently called using GATK and SAMtools. Data from rs11984041 and rs2107595 were compared to prior genotyping efforts (microarray/TaqMan SNP genotyping) demonstrating an overlap of 99.8%.

From the 9,427 single nucleotide polymorphisms (minor allele count >= 1) and 1,040 insertions/deletions (InDels), 931 SNPs and 108 InDels were filtered out due to quality control issues (e.g. Hardy-Weinberg equilibrium). Of the remaining 8,496 SNPs, 2,939 variants had previously not been reported in dbSNP (Version dbSNP142). Of those, 2,008 had a minor allele count of one in the whole dataset. SNPs were analyzed using logistic regression followed by Bonferroni correction. Haplotype blocks were reconstructed using Haploviewer and association analyses performed using logistic regression. As variant-collapsing methods we used SKAT and SKAT-O and calculated values for multiple p-value thresholds: all SNPs, common SNPs (10%<MAF<50%), low frequency SNPs (1%<MAF<10%) and rare SNPs (0.1%<MAF<1%).

Proteome-Wide Analysis of SNPs (PWAS)

PWAS was conducted as previously described[21] with minor modifications as described in the supplemental material and methods.

Chromatin immunoprecipitation (ChIP)

$1x10^6$ HeLa cells were crosslinked with 1% formaldehyde at room temperature, lysed (50 mmol/L Tris-HCl pH 8, 10 mmol/L EDTA, 1% SDS, complete EDTA-free protease inhibitor, Roche) and sheared by sonication on a Covaris S220 (Covaris). The supernatant was diluted 1:10 with ChIP-RIPA Buffer (10 mmol/L Tris-HCl pH 7.5, 1 mmol/L EDTA, 0.5 mmol/L EGTA, 1% Triton X-100, 0.1% SDS, 0.1% Na-deoxycholate, 140 mmol/L NaCl). Immunoprecipitation was performed using an E2F3 antibody (Santa Cruz, C-18 sc-878) or IgG antibody (abcam, rabbit IgG, ab37415). qPCR data are normalized to the IgG control.

Genome-edited Jurkat cells were grown with or without 1 mmol/L HU (200,000 cells/ml) and harvested after 24 h. Formaldyde fixation and nuclei preparation was conducted according to TruChIP™ protocol (Covaris). Nuclei were washed in MNase digestion buffer (1% Triton X-100, 0.1 % Na-deoxycholate, 0.1 % SDS, 140 mmol/L NaCl, 10 mmol/L Tris-HCl pH 8.0, 1 mmol/L EDTA). Chromatin was digested using Micrococcal Nuclease (MNase digestion buffer supplemented with 2 mmol/L $CaCl_2$).[22] ChIP was performed as described above using E2F3 and Rb1 antibodies (Santa Cruz, C-18 sc-878, sc-50 X).

Cell culture and Transfection

HeLa cells were maintained in DMEM-GlutaMAX™-I, and THP-1 and Jurkat (clone E6-1) cells were cultured in RPMI 1640 medium, both supplemented with 10% fetal bovine serum, 100 U/ml penicillin/100 µg/ml streptomycin (reagents from Gibco, Life Technologies). Human aortic smooth muscle cells (HAoSMC) and human aortic endothelial cells (HAoEC) obtained from PromoCell were cultured according to manufacturer's instructions. HeLa cells were transfected using Lipofectamine 2000 (Invitrogen). HAoSMCs and HAoECs, THP-1 and Jurkat cells were transfected by appropriate Amaxa® Cell Line Nucleofector® Kits. Small interfering RNAs (siRNAs) were obtained from Dharmacon Thermo Fisher Scientific (Supplemental Table 1). THP-1 cells were seeded in RPMI1640 supplemented with 100ng/ml Phorbol 12-myristate 13-acetate (PMA) immediately after transfection to induce THP-1 macrophage (MΦ) differentiation.

RNA isolation and cDNA synthesis

Total RNA was isolated using either Qiazol or the RNeasy Mini kit (Qiagen) according to the manufacturer`s instructions including the RNase-free DNase Set (Qiagen). An equal amount of RNA was used for each Oligo dT(15) or random primed cDNA synthesis (Omniscript RT kit Qiagen).

Protein isolation and Immunoblotting

Cells were washed with PBS and lysed with RIPA buffer (50 mmol/L Tris-HCl pH 7.5, 150 mmol/L NaCl, 1% NP-40, 0.5% desoxycholate, 0.1% SDS) containing complete Protease Inhibitor (Roche) for 30 min on ice. Protein concentrations were determined by using BCA Protein assay kit (Pierce, Thermo Fisher Scientific). Primary antibodies: rabbit anti-E2F3 1:1000 (Santa Cruz), rabbit anti-E2F4 1:1000 (Santa Cruz), rabbit anti-Rb1 1:1000 or 1:4000 (Santa Cruz), rabbit anti-Rbl1 1:1000 (Santa Cruz), rabbit anti-Rbl2 1:1000 (Santa Cruz), rabbit anti-actin 1:1000 (Sigma Aldrich) and mouse anti-HA (clone 16B12) 1:1000 (Covance). Secondary antibodies: goat anti-mouse 1:10000 (Dako) and goat anti-rabbit 1:10000 (Dako). For Western Blot analysis of the PWAS samples 16 mmol/L biotin was eluted in 40µl PBB (150 mmol/L NaCl, 50 mmol/L Tris/HCl pH 8.0, 10 mmol/L MgCl2, 0.5% NP40, Complete Protease Inhibitor-EDTA, Roche) after the DNA pull-down and separated SDS-PAGE.

Gene expression analysis

Gene expression analysis was performed using quantitative PCR applying SYBR Green or TaqMan technology. Gene specific primers are listed in supplemental table 2. TaqMan probes were obtained from Applied Biosystems: HDAC9 (Hs00206843 m1), Twist1 (Hs00361186_m1), CD68 (Hs02836816_g1). For normalization RPLP0 (4326314E) and HPRT (4326321E) probes were used.

Cell cycle synchronization

HeLa cells or genome-edited Jurkat cells were seeded in T80 cell culture flasks with a density of 35,000 or 200,000 cells per ml, respectively. After 24 h hydroxyurea (HU) was added in a final concentration of 5 mmol/L or 1 mmol/L, respectively. 24 h after synchronization represents the baseline time point and media were changed allowing progression of the cell cycle and subsequently harvested any other hour.

Human primary Aortic Smooth Muscle cells (HAoSMCs) and human blood-derived MΦ cultures

Experiments in primary human cells were approved by the local institutional review board (project #17-693). Primary human blood-derived MΦ were obtained from healthy volunteers. PBMCs purified over a 15 ml Ficoll Paque Premium cushion at room temperature by centrifugation (400xg 30 min, acceleration 3, deceleration 0). PBMCs were washed (PBS, 2 mmol/L EDTA) at 400xg 10min and

resuspended in 1 ml Cyro-SFM (Promo Cell). Cells were frozen at -80°C overnight (Mr. Frosty, ThermoFisher) and subsequently stored in liquid nitrogen.

To isolate monocytes from PBMCs and differentiate them into MΦ, cryopreserved aliquots of each genotype (rs2107595, GG genotype: n=9; AA: n=6) age- and gender-matched were plated in two wells (6 well plate) in monocyte attachment medium (Promo Cell). After 1 h, cells medium was replaced with M1-Macrophage Generation Medium DFX (Promo Cell). After 8 days in culture, cells were incubated in medium with reduced supplement (1%) either with or without 50 ng/ml human TNFα and 10 ng/ml IFNγ and incubated for 24 h.

Primary HAoSMC were obtained from Dr. Civelek (University of Virginia) (rs2107595 GG: n=9; AA: n=6) and cultured in Smooth Muscle Basal Medium (SMBM, Lonza) according to manufacturer's instructions on 0.1% Gelatine coated flasks. For stimulation, 100,000 cells were seeded and cultured overnight in 6 well plates and treated with 20 ng/ml human TNFα in temperature and pH equilibrated FBS-free SMBM for 0, 4 h and 8 h. Cells were lysed with Qiazol in the plate for RNA isolation.

Dual luciferase reporter assay

For the dual luciferase reporter assay we used the pGL3 vector (Promega) carrying the minimal murine HSP68 promoter. 41-bp oligonucleotides of the genomic SNP sequences containing either the common or risk allele (Supplemental Table 2) were generated by annealing ssDNA oligo nucleotides (flanked by single-stranded overhangs for KpnI and SacI restriction sites) and subsequently cloned into the pGL3-mHSP68 plasmid using KpnI and SacI restriction sites. For normalization the phRL-TK Renilla vector (Promega) was used. The luciferase constructs were transiently transfected and measured using a Glomax-Multi Detection System (Promega) after 24 h.

Generation of genome-edited Jurkat cell lines

Edited cell lines were generated as previously described.[23] In brief, targeting vectors were designed and produced by Horizon Discovery (Cambridge, UK) and rAAV produced by co-transfection of HEK293T cells with targeting and helper vectors. Viruses were purified using an AAV purification kit (Virapur, San Diego, USA). Jurkat cells heterozygous for rs2107595 were infected with rAAV carrying either the common or the risk allele of rs2107595. For selection a loxP-flanked neomycin resistance cassette was included in the vector and genome-edited single cell clones were identified by genotyping. The neomycin cassette was removed using Cre recombinase and its absence verified in single cell clones by PCR. Successful editing was subsequently confirmed by sequencing.

Circular Chromosome Conformation Capture

4C-chromatin was prepared as described previously.[24] Further details are explained in supplemental material and methods.

Cell proliferation assay

To determine cell proliferation pulse-chase experiments were performed in unsynchronized genome-edited Jurkat cells using Click-iT™ Plus EdU (5-ethynyl-2'-deoxyuridine) technology (# C10634, Thermo Fisher). Cells were seeded at a density of 200,000 cells per ml and grown in the presence of 10 µmol/L EdU for 4 h. Afterward, cells were washed 3 times with PBS, resuspended in RPMI 1640 and analysed by flow cytometry (BD FACSVerse™) at 0, 24 h, 48 h and 72 h.

Statistical analysis

The Shapiro-Wilks Test was utilized to determine the distribution of data sets. Normally distributed data were statistically analysed with the parametric T-Test, else a Wilcoxon Rank-Sum Test or Wilcoxon Signed-Rank Test were applied. Data are represented as mean values and standard error of the mean unless specified otherwise. Significance is depicted as follows; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. HDAC9 regional plots (**Figure 1A**) were constructed using locuszoom. The upper panel uses data from the large artery stroke analysis of the MEGASTROKE collaboration.[3]

RESULTS

Targeted resequencing of the HDAC9 region supports candidacy of rs2107595 as the causal variant for large artery stroke

rs2107595 gave the strongest signal in previous GWAS for atherosclerotic phenotypes,[3, 11, 12, 19, 25, 26] (**Figure 1A, upper panel**) and had a >95% posterior probability of being the only causal SNP at this locus in the most recent stroke GWAS.[3] To further examine the candidacy of rs2107595 as the causal variant at this locus while also capturing rare variants, low-frequency variants, and haplotypes, we performed targeted resequencing of the *HDAC9* gene region including the nearby *TWIST1* and *FERD3L* genes in 176 patients with large artery stroke and 176 stroke-free controls (**Figure 1A**, **middle panel**; **Figure S1**). Genotypes for rs2107595 showed 99.8% agreement with previously obtained microarray and TaqMan SNP genotyping data demonstrating the reliability of our sequencing approach. Overall, we identified 9,428 variants (8,496 SNPs, 932 insertions/deletions) and 169 haplotype blocks but no rare or low-frequency variants in the rs2107595 haplotype block. Following correction for multiple testing, none of the variants or haplotypes significantly associated with large artery stroke thus arguing against variants with large effect sizes in this region. Next, we used variant-collapsing methods (SKAT and SKAT-O) to analyse the 2.5-kb sequence block around rs2107595, which is conserved in mammals, the intergenic region between *HDAC9* and *TWIST1*, and the *HDAC9*, *TWIST1*, and *FERD3L* genes (**Figure 1A**, **lower panel**). SKAT-O analyses revealed a significant association (p=0.017) for the conserved sequence block encompassing rs2107595, while all other equally-sized sequence blocks showed higher p-values. Of note, all proxy SNPs (r^2 with rs2107595 >0.8) localize outside the conserved sequence block. Collectively, these findings support rs2107595 as the causative variant at this locus. Hence, we focused on this variant in our functional analyses.

The rs2107595 risk variant interferes with E2F3 binding

The rs2107595 region shows enrichment for marks of regulatory chromatin (DHS, H3K27ac, H3K4me1, H3K9me3) in various cell types listed in HaploReg,[27] Roadmap Epigenomics,[28] and ENCODE[13] (**Supplemental Table 3 and 4**, **Figure S2**) suggesting a potential involvement of rs2107595 in transcriptional regulation. To identify transcription factors with allele-specific binding at rs2107595 and hence a possible role in transcriptional regulation, we performed proteome-wide analysis of SNPs (PWAS). This approach is based on the interaction of synthetic oligonucleotides with metabolically labelled nuclear factors (Stable Isotopic Labelling with Amino acids in Cell culture, SILAC) that are subsequently identified by mass spectrometry.[21] 41-bp-SNP-centered oligonucleotides differing only at rs2107595 (**Supplemental Table S2**) were incubated either with light or heavy isotope labelled nuclear factors from HeLa cells. A comparison of the heavy/light ratios of all identified binding proteins revealed six factors surpassing the predefined false discovery rate of 0.01: NFATC2, a member of the nuclear factor of activated T-cells,[29] L3MBTL3, a putative polycomb group protein functioning as transcriptional regulator in large protein complexes,[30] SAMD1, a protein with a potential role in immobilizing low density lipoprotein (LDL) in the arterial wall,[31] and all constituents of the E2F3/TFDP1/Rb1 complex (**Figure 1B**).

E2F3 and TFDP1 represent transcription factors of the E2F and DP1 families known to complex with Rb proteins.[32] The observed enrichment of E2F3 at the common allele is supported by the prediction of an E2F3 consensus site[33] within the common allele sequence which is disrupted by the risk allele (**Figure 1C**). To validate allele-specific binding of E2F3 we further incubated biotinylated synthetic oligonucleotides with nuclear extracts from HeLa cells and purified the assembled allele-specific nucleoprotein complexes by DNA pull-down. Subsequent immunoblotting revealed enriched binding of E2F3 to the common allele (**Figure 1D**). Finally, we performed ChIP experiments in HeLa cells, which are homozygous for the rs2107595 common allele and thus suited to explore E2F3 binding *in vivo*. ChIP revealed a significant occupancy of E2F3 at rs2107595 (**Figure 1E**). Given these results and the known role of E2F and Rb proteins in transcriptional regulation[34, 35] we considered these proteins to be strong candidates for regulating HDAC9 expression.

E2F3 and Rb1 regulate HDAC9 expression

To determine the effect of E2F and Rb proteins on endogenous HDAC9 expression we next conducted gain- and loss-of-function experiments in HeLa cells. Overexpression of E2F3a resulted in a 6-fold increase in HDAC9 mRNA levels compared to empty vector control. In contrast, overexpression of Rb1 led to a reduction in HDAC9 expression which however did not reach significance (**Figure 2A and S3A, B**). siRNA-mediated knockdown of E2F3, E2F4, or both resulted in a significant decrease of HDAC9 mRNA compared to non-targeting control (**Figure 2B and S3C, D**). In contrast, knockdown of Rb proteins caused a significant increase in HDAC9 expression for Rb1, Rbl1 and the triple knockdown (**Figure 2C and S3E, F**).

E2F and Rb act as transcriptional regulators of cell cycle genes. At the G_1/S boundary repressive Rb proteins become phosphorylated by cyclin-dependent kinases and dissociate from E2F proteins, which then activate the expression of target genes.[34-36] Hence, we analysed cell cycle-dependent variations in HDAC9 expression. Synchronization of HeLa cells by hydroxyurea- (HU)-induced cell cycle arrest at the G_1/S boundary led to a significant increase in HDAC9 mRNA expression compared to untreated cells (**Figure 2D and 2E**). Following release of the cell cycle arrest HDAC9 mRNA expression further increased during progression through S phase and declined upon reaching G2, thus paralleling the activity of E2F proteins across the cell cycle.[37, 38]

The rs2107595 risk variant is associated with elevated HDAC9 transcription

To examine the association between rs2107595 and HDAC9 gene expression in cells relevant to atherosclerosis we first examined primary human MΦ and human aortic smooth muscle cells (HAoSMCs) with defined carrier status at rs2107595. Proinflammatory MΦ were isolated from PBMCs obtained from healthy donors (GG genotype: n=7; GA: n=7; AA: n=5, matched for age and gender) and differentiated *in vitro* (**Figure S4A**). Upon stimulation with TNFα and IFNγ, MΦ homozygous for the risk allele showed significantly elevated HDAC9 expression levels compared to MΦ homozygous for the common allele (**Figure 3A**). Gene expression analysis in cultured HAoSMC (GG genotype: n=9; AA: n=6) revealed no allele-specific differences in HDAC9 expression before and after 4 or 8 h of TNFα stimulation (**Figure 3B**). Also, there was no allele-specific effect on TWIST1 expression in HAoSMCs and MΦ (**Figure S4B and** results not shown).

To examine the effects of rs2107595 on transcriptional regulation, we further performed luciferase reporter assays in T-lymphoid Jurkat cells, THP-1 monocytes and PMA-induced THP-1 MΦ, HAoEC, and HAoSMC. 41-bp-SNP centered fragments containing either the rs2107595 common or risk variant were cloned into a firefly luciferase reporter vector[21] (**Figure 3C**) and tested for a *cis*-regulatory function by measuring luciferase activity after transient transfection. Transcriptional activity was significantly higher for the risk allele compared to the common allele both in Jurkat cells and PMA-induced THP-1 MΦ[39] (**Figure 3C**) whereas we found no allele-specific differences in HAoEC, HAoSMCs, and THP-1 monocytes, (**Figure S5A-C**). SNPs in high (rs57301765, r^2=0.99) and low (rs10255384, r^2=0.47) linkage disequilibrium (LD) with rs2107595 showed no consistent results in Jurkat cells and PMA-induced THP1 MΦ (**Figure S5D-G**).

Next, we specifically genome-edited rs2107595 in Jurkat cells using recombinant adeno-associated virus[23] (rAAV) resulting in isogenic cells differing solely at rs2107595. Jurkat cells were chosen because of (1) their immunological origin, (2) the presence of open chromatin marks both in the rs2107595 region (**Figure S2**) and *HDAC9* promoter, and (3) their diploidy and heterozygosity for rs2107595[13, 28, 40] allowing a one-step editing procedure in either direction. Successful editing was confirmed by direct sequencing (**Figure 3D**). Cells homozygous for the risk allele exhibited 2-fold higher mRNA levels of HDAC9 compared to cells carrying the common allele (**Figure 3E**). Heterozygous cells displayed intermediate mRNA levels compatible with a gene dosage effect. TWIST1 and FERDL3 expression levels were below detection limit in these cells (data not shown). Collectively, these results show that rs2107595 regulates HDAC9 transcription in an allele-specific manner. We further examined allele-specific effects of rs2107595 on HDAC9 transcription across the cell cycle. Following synchronization at the G_1/S-boundary, HDAC9 levels were significantly elevated in risk allele cells compared to common allele cells (time point zero, **Figure 3F**) in accordance with the results obtained in unsynchronized cells (**Figure 3E**). This difference was sustained for 6 h following

release of the HU block. Because of the allele-specific effects on cell cycle associated HDAC9 expression we analysed the effect of rs2107595 on cell proliferation in genome-edited Jurkat cells. Pulse-chase labeling with the thymidine analogue EdU and detection by flow cytometry revealed no allele-specific differences for rs2107595 (**Figure S6 A and B**).

E2F3 mediates allele-specific effects of rs2107595 on HDAC9 transcription

Given the observed effect of rs2107595 on HDAC9 transcription we next tested for physical interactions of the rs2107595 region with the HDAC9 promoter by circularized chromosome conformation capture (4C) in isogenic Jurkat cells. Based on Jurkat cell-specific open chromatin structure (DHS) and promoter information (H3K4me3)[13] we selected the promoter viewpoint at nt ~18,330,000. Mapping the 4C-seq signals to the HDAC9 gene region revealed a significant interaction between rs2107595 and the promoter region in common allele ("GG" in **Figure 4A**) but not in risk allele cells ("AA") indicating allele-specific differences in chromatin organisation. Analyses for an alternative HDAC9 promoter lacking detectable chromatin marks in Jurkat cells showed lower significance for allele-specific interactions at both viewpoints (**Figure S7**). These results provide further mechanistic evidence for a role of the rs2107595 region in regulating HDAC9 transcription.

To determine whether the *in vivo* binding of E2F3 and Rb1 at rs2107595 observed in HeLa cells occurs in a truly allele-specific manner we next performed ChIP experiments in the genome-edited isogenic Jurkat cells. Since E2F3 and Rb1 control cell cycle progression at the G1/S boundary,[38] we arrested these cells with HU. Upon synchronization, we found a significantly enriched occupancy of E2F3 and Rb1 proteins at the common allele compared to the risk allele (**Figure 4B and 4C**), which was not present in unsynchronized cells (**Figure S8 A and B**) possibly reflecting cell cycle-dependent binding of E2F3 and Rb1 to the common allele.

Finally, to examine whether the allele-specific effects on HDAC9 transcription at rs2107595 are mediated by allele-specific binding of E2F3 and Rb1, we tested the influence of exogenous E2F3a and Rb1 expression in isogenic Jurkat cells. Compared to empty vector control, overexpression of E2F3a but not Rb1 resulted in a significant increase of the ratio between HDAC9 expression in cells homozygous for the common allele vs cells homozygous for the risk allele (**Figure 4D and S8C**). Collectively, these results suggest allele-specific interactions between rs2107595 and the HDAC9 promotor and show a mediating effect of E2F3 on HDAC9 expression via rs2107595 (see proposed model in **Figure 4E**).

DISCUSSION

We present a mechanism by which a noncoding variant at the large artery stroke and coronary artery disease risk locus on 7p21.1 regulates HDAC9 transcription. We show that rs2107595, the likely causal variant at this locus, has allele-specific transcriptional capacity and that the risk allele associates with elevated HDAC9 expression levels in cell types relevant to atherosclerosis. We further identify a physical interaction of the rs2107595 region with the HDAC9 promoter, demonstrate preferential binding of the E2F3/ TFDP1/Rb1 cell-cycle complex to the common allele, and show that E2F3 mediates HDAC9 transcription in an allele-specific manner. Together, our data imply transcriptional regulation of HDAC9 via E2F3/Rb1 complexes as a major mechanism linking genetic variation at rs2107595 with disease risk.

A transcriptional effect of rs2107595 on HDAC9 expression is demonstrated by our data in genome-edited T-lymphoid Jurkat cells and in primary proinflammatory MΦ, and is further substantiated by the 4C results, which showed a physical interaction between the rs2107595 region and the HDAC9 promoter. The directionality of the transcriptional effect was consistent with the results from luciferase assays for rs2107595 in Jurkat cells and THP-1 MΦ. It was further consistent with the effects on HDAC9 transcription reported previously for PBMCs[18] in that the risk allele was associated with higher HDAC9 expression levels. Of note, however, our earlier observations in PBMCs did not allow attributing allele-specific effects to a specific genetic variant. As such, the current findings represent a major advance. While we cannot exclude that other variants in the rs2107595 region also contribute to transcriptional regulation of HDAC9, several observations point to rs2107595 as the causal variant mediating vascular risk: rs2107595 was the lead SNP in GWAS for stroke[3, 11, 12] and coronary artery disease.[8] It was the only variant contained in the 95% credible SNP set in the most recent stroke GWAS,[3] and here using targeted resequencing we found no variants with large effect sizes in the HDAC9 region with SKAT-O analyses further favoring rs2107595 as the causal variant.

The observed allele-specific interaction between rs2107595 and E2F3/Rb1 complexes is supported by four independent lines of evidence: an unsupervised approach using proteome-wide analysis of allele-specific binding partners, DNA pull-down experiments in combination with immunoblotting, ChIP, and bioinformatics data showing a consensus-binding site for E2F3 at the common allele. Again, the directionality was consistent across all approaches in that the risk allele disrupted binding to E2F3. Importantly, allele-specific interaction was also demonstrated *in vivo* using ChIP. Our proteome-wide experiment identified differential interactors aside from E2F3 and Rb proteins and we cannot exclude a role of these factors in mediating allele-specific effects.[30, 41] However, the binding of three proteins belonging to the same complex (E2F3, TFDP1, and Rb1) together with our functional results strongly support a major role of these factors in mediating the effects of rs2107595 on HDAC9 expression.

An involvement of E2F and Rb proteins in regulating HDAC9 transcription is evidenced by our gain- and loss-of-function experiments in HeLa cells and by the mediating effect of E2F on allele specific HDAC9 expression in isogenic Jurkat cells. In accord with this, we found the expression of HDAC9 in HeLa and genome-edited Jurkat cells to be cell cycle-dependent in a manner paralleling E2F3 activity. Despite these observations and the proposed role of HDAC9 in cell proliferation and cancer,[42-45] we found no allele-specific effect on cell proliferation in isogenic Jurkat cells. However, this might relate to Jurkat cells lacking functional p53,[46] which is transcriptionally regulated by HDAC9.[45]

Our findings provide some indication that the effects of rs2107595 on HDAC9 expression might be cell-type dependent. While the rs2107595 risk allele was associated with higher HDAC9 expression levels in proinflammatory human MΦ and genome-edited T-lymphoid Jurkat cells we found no indication for an allele-specific effect in cultured HAoSMC. Similarly, luciferase assays showed a higher transcriptional activity with the risk allele in Jurkat cells and proinflammatory THP-1 MΦ but not in undifferentiated THP-1 monocytes, HAoSMCs and HAoECs. However, these cell lines also vary in terms of their source and cell senescence. For instance, primary MΦ were isolated from healthy young adults, whereas HAoSMCs were isolated from heart transplant donors with propagation for multiple passages. Future studies using genome-editing in inducible pluripotent stem cells (IPSCs) with

differentiation into different cell lineages might allow better delineating the biological effects of rs2107595 in specific cell-types relevant to human atherosclerosis.[47]

Our observations are reminiscent of a previous study that found a single risk variant associated with LDL levels and myocardial infarction to create a CCAAT enhancer binding protein transcription factor binding site and alter the expression of SORT1, a transporter protein involved in LDL secretion.[48, 49] Similarly, the lead SNP at the coronary artery disease locus on 4q32.1 has been shown to influence GUCY1A3 expression levels by allele-specific binding to the transcription factor ZEB1.[50] To our knowledge, the current study is the first to provide a gene regulatory mechanism for a common variant associated with risk of atherosclerotic stroke.[5]

In conclusion, our current findings imply transcriptional regulation of HDAC9 via E2F3 and Rb1 as a major mechanism mediating disease risk at rs2107595. HDAC9 has emerged as a potential target for drug development. For one, there is evidence from different mouse models of atherosclerosis that lowering HDAC9 expression may attenuate atherogenesis.[17, 18] Second, rs2107595 has been associated with early stages of atherogenesis,[19, 20] which makes HDAC9 an attractive target for early intervention. Third, recent drug discovery programs have resulted in the development of selective class IIa HDAC inhibitors with reasonable specificity and inhibitory activity against HDAC9.[51-53] Interest in HDAC9 further emerges from the observation that the HDAC9 locus is implicated in three major manifestations of atherosclerosis: stroke, coronary artery disease, and peripheral artery disease. More work is needed to better understand the mechanisms linking genetic variation in the rs2107595 region to atherosclerosis.

ACKNOWLEDGMENTS

We thank Joseph R. Nevins and Alexander Brehm for providing reagents, Arthur Liesz und Stefan Roth for technical support and advice, Horizon Discovery, Cambridge, UK for support in generating genome-edited cell lines, Noortje A. M. van den Dungen for technical assistance and the Utrecht Sequencing Facility for performing sequencing of the 4C libraries. Tom Webb and Nilesh J. Samani are supported by the British Heart Foundation.

SOURCES OF FUNDING

This work was supported by grants from the Deutsche Forschungsgemeinschaft (CRC 1123 [B3] and Munich Cluster for Systems Neurology [SyNergy]), the German Federal Ministry of Education and Research (BMBF, e:Med programme e:AtheroSysMed), the FP7/2007-2103 European Union project CVgenes@target (grant agreement No Health-F2-2013-601456), the Leducq Foundation CADgenomics programme, European Union Horizon2020 projects SVDs@target (grant agreement No 66688) and CoSTREAM (grant agreement No 667375), the Vascular Dementia Research Foundation and the Friedrich Baur Stiftung. Nilesh J. Samani is a NIHR senior investigator.

DISCLOSURES

None.

REFERENCES

1. Collaborators GBDLRoS, Feigin VL, Nguyen G, et al. Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016. *N Engl J Med*. 2018;379:2429-2437. doi: 10.1056/NEJMoa1804492.
2. Collaborators GBDCoD. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2017;390:1151-1210. doi: 10.1016/S0140-6736(17)32152-9.
3. Malik R, Chauhan G, Traylor M, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet*. 2018;50:524-537. doi: 10.1038/s41588-018-0058-3.
4. Malik R, Rannikmae K, Traylor M, Georgakis MK, Sargurupremraj M, Markus HS, Hopewell JC, Debette S, Sudlow CLM, Dichgans M, consortium M, the International Stroke Genetics C. Genome-wide meta-analysis identifies 3 novel loci associated with stroke. *Ann Neurol*. 2018;84:934-939. doi: 10.1002/ana.25369.
5. Dichgans M, Pulit SL, Rosand J. Stroke genetics: discovery, biology, and clinical applications. *Lancet Neurol*. 2019;in press. doi.
6. Nurnberg ST, Zhang H, Hand NJ, Bauer RC, Saleheen D, Reilly MP, Rader DJ. From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease. *Circ Res*. 2016;118:586-606. doi: 10.1161/CIRCRESAHA.115.306464.
7. International Stroke Genetics C, Wellcome Trust Case Control C, Bellenguez C, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet*. 2012;44:328-333. doi: 10.1038/ng.1081.
8. Consortium CAD, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45:25-33. doi: 10.1038/ng.2480.
9. Matsukura M, Ozaki K, Takahashi A, et al. Genome-Wide Association Study of Peripheral Arterial Disease in a Japanese Population. *PLoS One*. 2015;10:e0139262. doi: 10.1371/journal.pone.0139262.
10. Dichgans M, Malik R, Konig IR, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*. 2014;45:24-36. doi: 10.1161/STROKEAHA.113.002707.
11. Traylor M, Farrall M, Holliday EG, et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2012;11:951-962. doi: 10.1016/S1474-4422(12)70234-X.
12. Malik R, Traylor M, Pulit SL, et al. Low-frequency and common genetic variation in ischemic stroke: The METASTROKE collaboration. *Neurology*. 2016;86:1217-1226. doi: 10.1212/WNL.0000000000002528.
13. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2018;46:D794-D801. doi: 10.1093/nar/gkx1081.
14. Marsman J, Horsfield JA. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. *Biochim Biophys Acta*. 2012;1819:1217-1227. doi: 10.1016/j.bbagrm.2012.10.008.
15. Meng H, Bartholomew B. Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. *J Biol Chem*. 2018;293:13786-13794. doi: 10.1074/jbc.R117.813485.
16. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*. 2014;15:272-286. doi: 10.1038/nrg3682.
17. Cao Q, Rong S, Repa JJ, St Clair R, Parks JS, Mishra N. Histone deacetylase 9 represses cholesterol efflux and alternatively activated macrophages in atherosclerosis development. *Arterioscler Thromb Vasc Biol*. 2014;34:1871-1879. doi: 10.1161/ATVBAHA.114.303393.
18. Azghandi S, Prell C, van der Laan SW, Schneider M, Malik R, Berer K, Gerdes N, Pasterkamp G, Weber C, Haffner C, Dichgans M. Deficiency of the stroke relevant HDAC9 gene attenuates atherosclerosis in accord with allele-specific effects at 7p21.1. *Stroke*. 2015;46:197-202. doi: 10.1161/STROKEAHA.114.007213.
19. Markus HS, Makela KM, Bevan S, Raitoharju E, Oksala N, Bis JC, O'Donnell C, Hainsworth A, Lehtimaki T. Evidence HDAC9 genetic variant associated with ischemic stroke increases risk via promoting carotid atherosclerosis. *Stroke*. 2013;44:1220-1225. doi: 10.1161/STROKEAHA.111.000217.

20. Franceschini N, Giambartolomei C, de Vries PS, et al. GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat Commun*. 2018;9:5141. doi: 10.1038/s41467-018-07340-5.
21. Butter F, Davison L, Viturawong T, Scheibe M, Vermeulen M, Todd JA, Mann M. Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet*. 2012;8:e1002982. doi: 10.1371/journal.pgen.1002982.
22. Cappabianca L, Thomassin H, Pictet R, Grange T. Genomic footprinting using nucleases. *Methods Mol Biol*. 1999;119:427-442. doi: 10.1385/1-59259-681-9:427.
23. Jones PD, Kaiser MA, Ghaderi Najafabadi M, McVey DG, Beveridge AJ, Schofield CL, Samani NJ, Webb TR. The Coronary Artery Disease-associated Coding Variant in Zinc Finger C3HC-type Containing 1 (ZC3HC1) Affects Cell Cycle Regulation. *J Biol Chem*. 2016;291:16318-16327. doi: 10.1074/jbc.M116.734020.
24. van de Werken HJ, de Vree PJ, Splinter E, Holwerda SJ, Klous P, de Wit E, de Laat W. 4C technology: protocols and data analysis. *Methods Enzymol*. 2012;513:89-112. doi: 10.1016/B978-0-12-391938-0.00004-5.
25. Shroff N, Ander BP, Zhan X, Stamova B, Liu D, Hull H, Hamade FR, Dykstra-Aiello C, Ng K, Sharp FR, Jickling GC. HDAC9 Polymorphism Alters Blood Gene Expression in Patients with Large Vessel Atherosclerotic Stroke. *Transl Stroke Res*. 2018. doi: 10.1007/s12975-018-0619-x.
26. Wang XB, Han YD, Sabina S, Cui NH, Zhang S, Liu ZJ, Li C, Zheng F. HDAC9 Variant Rs2107595 Modifies Susceptibility to Coronary Artery Disease and the Severity of Coronary Atherosclerosis in a Chinese Han Population. *PLoS One*. 2016;11:e0160449. doi: 10.1371/journal.pone.0160449.
27. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res*. 2016;44:D877-881. doi: 10.1093/nar/gkv1340.
28. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317-330. doi: 10.1038/nature14248.
29. Modiano JF, Johnson LD, Bellgrau D. Negative regulators in homeostasis of naive peripheral T cells. *Immunol Res*. 2008;41:137-153. doi: 10.1007/s12026-008-8017-1.
30. Meier K, Brehm A. Chromatin regulation: how complex does it get? *Epigenetics*. 2014;9:1485-1495. doi: 10.4161/15592294.2014.971580.
31. Lees AM, Deconinck AE, Campbell BD, Lees RS. Atherin: a newly identified, lesion-specific, LDL-binding protein in human atherosclerosis. *Atherosclerosis*. 2005;182:219-230. doi: 10.1016/j.atherosclerosis.2005.01.041.
32. Girling R, Partridge JF, Bandara LR, Burden N, Totty NF, Hsuan JJ, La Thangue NB. A new component of the transcription factor DRTF1/E2F. *Nature*. 1993;362:83-87. doi: 10.1038/362083a0.
33. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, Kolpakov FA, Makeev VJ. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46:D252-D259. doi: 10.1093/nar/gkx1106.
34. Korenjak M, Brehm A. E2F-Rb complexes regulating transcription of genes important for differentiation and development. *Curr Opin Genet Dev*. 2005;15:520-527. doi: 10.1016/j.gde.2005.07.001.
35. Blais A, Dynlacht BD. E2F-associated chromatin modifiers and cell cycle control. *Curr Opin Cell Biol*. 2007;19:658-662. doi: 10.1016/j.ceb.2007.10.003.
36. Sachdeva UM, O'Brien JM. Understanding pRb: toward the necessary development of targeted treatments for retinoblastoma. *J Clin Invest*. 2012;122:425-434. doi: 10.1172/JCI57114.
37. Leone G, Nuckolls F, Ishida S, Adams M, Sears R, Jakoi L, Miron A, Nevins JR. Identification of a novel E2F3 product suggests a mechanism for determining specificity of repression by Rb proteins. *Mol Cell Biol*. 2000;20:3626-3632. doi.
38. Stevaux O, Dyson NJ. A revised picture of the E2F transcriptional network and RB function. *Curr Opin Cell Biol*. 2002;14:684-691. doi.
39. Lund ME, To J, O'Brien BA, Donnelly S. The choice of phorbol 12-myristate 13-acetate differentiation protocol influences the response of THP-1 macrophages to a pro-inflammatory stimulus. *J Immunol Methods*. 2016;430:64-70. doi: 10.1016/j.jim.2016.01.012.
40. Schneider U, Schwenk HU, Bornkamm G. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int J Cancer*. 1977;19:621-626. doi.

41. Baksh S, Widlund HR, Frazer-Abel AA, Du J, Fosmire S, Fisher DE, DeCaprio JA, Modiano JF, Burakoff SJ. NFATc2-mediated repression of cyclin-dependent kinase 4 expression. *Mol Cell*. 2002;10:1071-1081. doi.
42. Yang R, Wu Y, Wang M, Sun Z, Zou J, Zhang Y, Cui H. HDAC9 promotes glioblastoma growth via TAZ-mediated EGFR pathway activation. *Oncotarget*. 2015;6:7644-7656. doi: 10.18632/oncotarget.3223.
43. Yuan Z, Peng L, Radhakrishnan R, Seto E. Histone deacetylase 9 (HDAC9) regulates the functions of the ATDC (TRIM29) protein. *J Biol Chem*. 2010;285:39329-39338. doi: 10.1074/jbc.M110.179333.
44. Zhang Y, Wu D, Xia F, Xian H, Zhu X, Cui H, Huang Z. Downregulation of HDAC9 inhibits cell proliferation and tumor formation by inducing cell cycle arrest in retinoblastoma. *Biochem Biophys Res Commun*. 2016;473:600-606. doi: 10.1016/j.bbrc.2016.03.129.
45. Zhao YX, Wang YS, Cai QQ, Wang JQ, Yao WT. Up-regulation of HDAC9 promotes cell proliferation through suppressing p53 transcription in osteosarcoma. *Int J Clin Exp Med*. 2015;8:11818-11823. doi.
46. Gioia L, Siddique A, Head SR, Salomon DR, Su AI. A genome-wide survey of mutations in the Jurkat cell line. *BMC Genomics*. 2018;19:334. doi: 10.1186/s12864-018-4718-6.
47. Gupta RM, Hadaya J, Trehan A, et al. A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell*. 2017;170:522-533 e515. doi: 10.1016/j.cell.2017.06.049.
48. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466:714-719. doi: 10.1038/nature09266.
49. Wang X, Raghavan A, Peters DT, Pashos EE, Rader DJ, Musunuru K. Interrogation of the Atherosclerosis-Associated SORT1 (Sortilin 1) Locus With Primary Human Hepatocytes, Induced Pluripotent Stem Cell-Hepatocytes, and Locus-Humanized Mice. *Arterioscler Thromb Vasc Biol*. 2018;38:76-82. doi: 10.1161/ATVBAHA.117.310103.
50. Kessler T, Wobst J, Wolf B, et al. Functional Characterization of the GUCY1A3 Coronary Artery Disease Risk Locus. *Circulation*. 2017;136:476-489. doi: 10.1161/CIRCULATIONAHA.116.024152.
51. Lobera M, Madauss KP, Pohlhaus DT, et al. Selective class IIa histone deacetylase inhibition via a nonchelating zinc-binding group. *Nat Chem Biol*. 2013;9:319-325. doi: 10.1038/nchembio.1223.
52. Choi SY, Kee HJ, Jin L, Ryu Y, Sun S, Kim GR, Jeong MH. Inhibition of class IIa histone deacetylase activity by gallic acid, sulforaphane, TMP269, and panobinostat. *Biomed Pharmacother*. 2018;101:145-154. doi: 10.1016/j.biopha.2018.02.071.
53. Guerriero JL, Sotayo A, Ponichtera HE, Castrillon JA, Pourzia AL, Schad S, Johnson SF, Carrasco RD, Lazo S, Bronson RT, Davis SP, Lobera M, Nolan MA, Letai A. Class IIa HDAC inhibition reduces breast tumours and metastases through anti-tumour macrophages. *Nature*. 2017;543:428-432. doi: 10.1038/nature21409.

Figure Legends

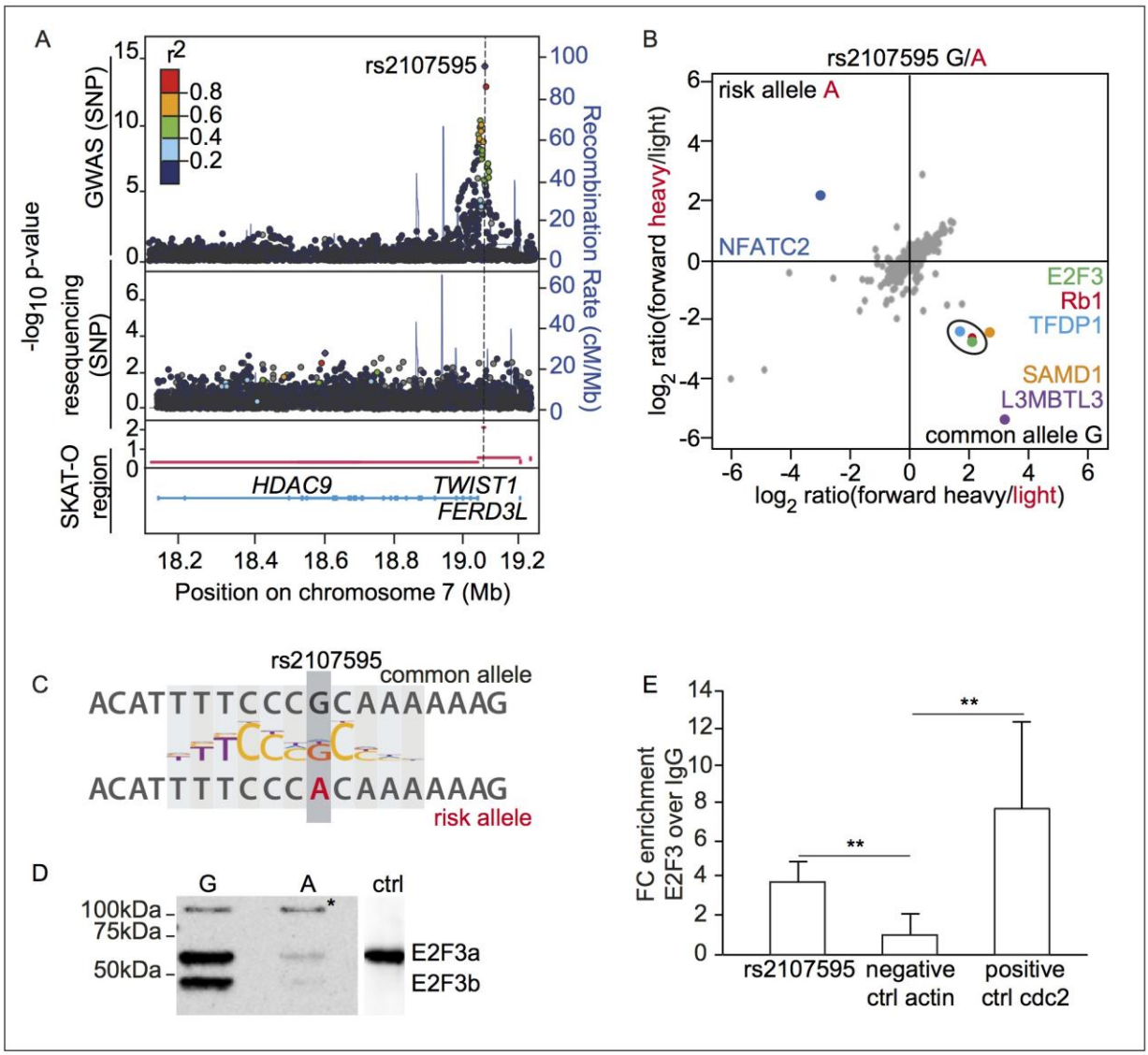
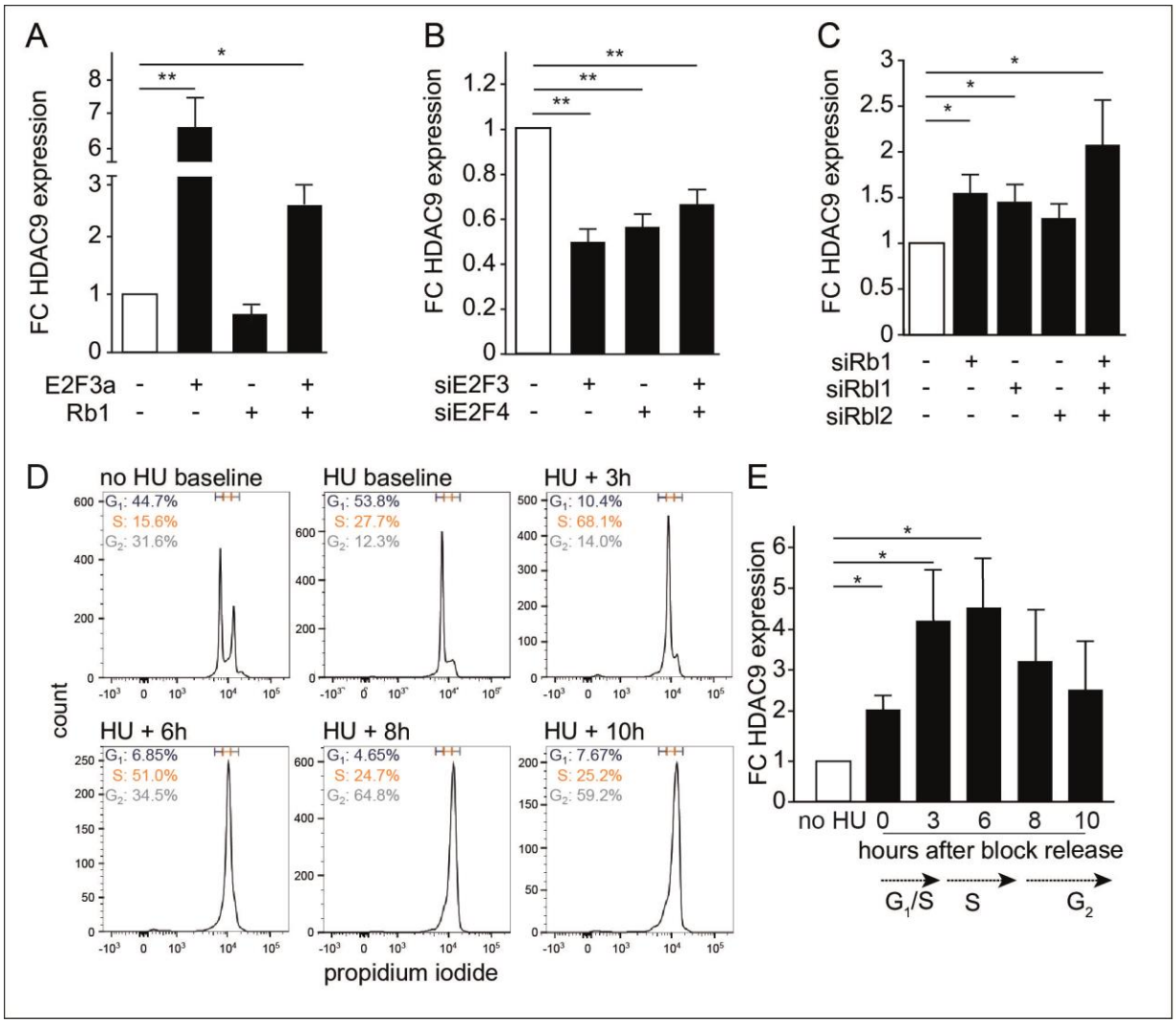


Figure 1: The rs2107595 risk variant interferes with E2F3 binding. (**A**) Top: regional association plot of the HDAC9 gene region (18123000-19188000, GRCh37/hg19) showing association signals around rs2107595 for large artery stroke in the MEGASTROKE dataset.[3] rs2107595 is located in noncoding DNA 3' to HDAC9 and 5' to TWIST1 and FERD3L. Middle: association plot of the same region showing variants identified by targeted resequencing of 176 cases of large artery stroke and 176 stroke-free controls. Bottom: -$\log_{10}$ p-values for the conserved sequence element around rs2107595, the intergenic region between HDAC9 and TWIST1, and the HDAC9, TWIST1, and FERD3L genes, calculated by variant-collapsing methods (SKAT and SKAT-O). The conserved 2.5-kb sequence block around rs2107595 (position marked by the dashed line) significantly associated with large artery stroke (p=0.017). (**B**) Identification of allele-specific binding partners of rs2107595 using PWAS. E2F3, Rb1, TFDP1, SAMD1 and L3MBTL3 preferentially interacted with the common allele (G) whereas NFATC2 preferentially bound to the risk allele (A). (**C**) Position Weight Matrix[33] for the consensus site of the human E2F3 protein aligned to the genomic sequence surrounding rs2107595. (**D**) Preferential binding of both E2F3a and E2F3b to the rs2107595 common allele (G) as revealed by immunoblotting. Overexpressed E2F3a (right panel) was used as a positive control. The asterisk marks an unspecific band that served as a loading control. (**E**) ChIP experiments showing *in vivo* binding of E2F3 to the

675 rs2107595 region in HeLa cells (E2F3 FC enrichment over IgG). The CDC2 promoter served as a
676 positive control for E2F3 binding while β-actin served as a negative control. n=7-8, mean±SD.
677 Wilcoxon Signed-Rank Test.
678
679



680
681
682
683 **Figure 2: E2F3 and Rb1 regulate HDAC9 expression**. **(A-C)** Fold changes (FC) in HDAC9 mRNA
684 expression assessed by quantitative real-time PCR analysis in HeLa cells after (**A**) overexpression of
685 E2F3a and Rb1, (**B**) siRNA mediated knockdown of E2F3 and E2F4 and (**C**) siRNA mediated
686 knockdown of Rb1, Rbl1 and Rbl2. n=7. (**D**) Cell cycle analysis by flow cytometry and propidium
687 iodide staining in HeLa cells following cell cycle arrest at the G1/S boundary by hydroxurea (HU).
688 Panel headlines indicate treatment modality and time points after release of the cell cycle arrest. (**E**)
689 Quantitative real-time PCR analysis of HDAC9 at different stages of the cell cycle shows an increase
690 of HDAC9 expression at the G_1/S boundary and during S phase. n=6-7. FC: fold change. mean±SEM.
691 Wilcoxon Signed-Rank Test.
692
693

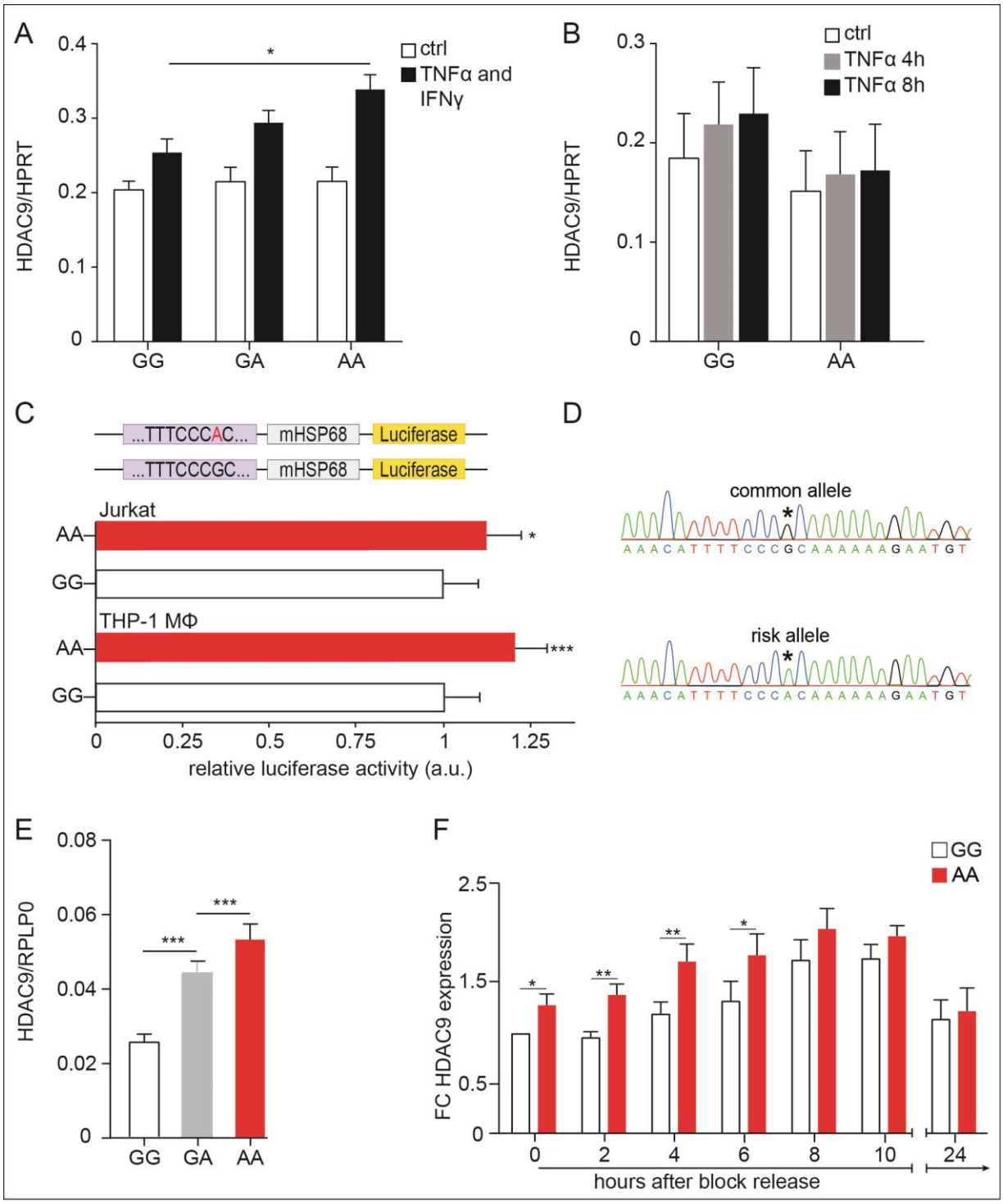


Figure 3: The rs2107595 risk variant is associated with elevated HDAC9 transcription. (**A**) Human blood-derived monocytes were isolated and differentiated *in vitro* to proinflammatory MΦ. Upon TNFα and IFNγ stimulation MΦ homozygous for the risk allele displayed significantly higher HDAC9 expression levels compared to common allele carriers. No allele-specific effects were seen in unstimulated MΦ. (GG: n=5; GA: n=5; AA: n=7). (**B**) Post mortem-derived HAoSMC were cultured *in vitro* and harvested for transcriptional analysis. No significant expression differences were measured in unstimulated or TNFα stimulated HAoSMCs (4h or 8h). (GG: n=9; AA: n=6). (**C**) Luciferase reporter assays using constructs containing a 41bp genomic region carrying the common (G) or risk (A) allele of

rs2107595. The risk allele showed a significant increase in luciferase activity compared to the common allele in T-lymphoid Jurkat and PMA-induced THP-1 MΦ. Increased HDAC9 mRNA expression in heterozygous and homozygous genome-edited Jurkat cells carrying the rs2107595 risk allele (A) compared to cells homozygous for the common allele (G). n=16 or 24, mean±SD. (**D**) Sanger sequencing of genome-edited Jurkat cells containing either the (*) common allele (G) or risk allele (A). (**E**) Increased HDAC9 mRNA expression in heterozygous and homozygous genome-edited Jurkat cells carrying the rs2107595 risk allele (A) compared to cells homozygous for the common allele (G). n=16 or 24, mean±SD. T-test. (**F**) Comparative expression analysis during cell cycle progression in isogenic Jurkat cells carrying either the common (G) or risk allele (A). HU arrested cells were relieved and harvested every 2 h until 10 h and after 24 h. HDAC9 expression levels increased during the first 8 h after HU removal. Risk allele carrying cells showed a significantly increased expression of HDAC9 until 6 h. mean±SD. T-test.

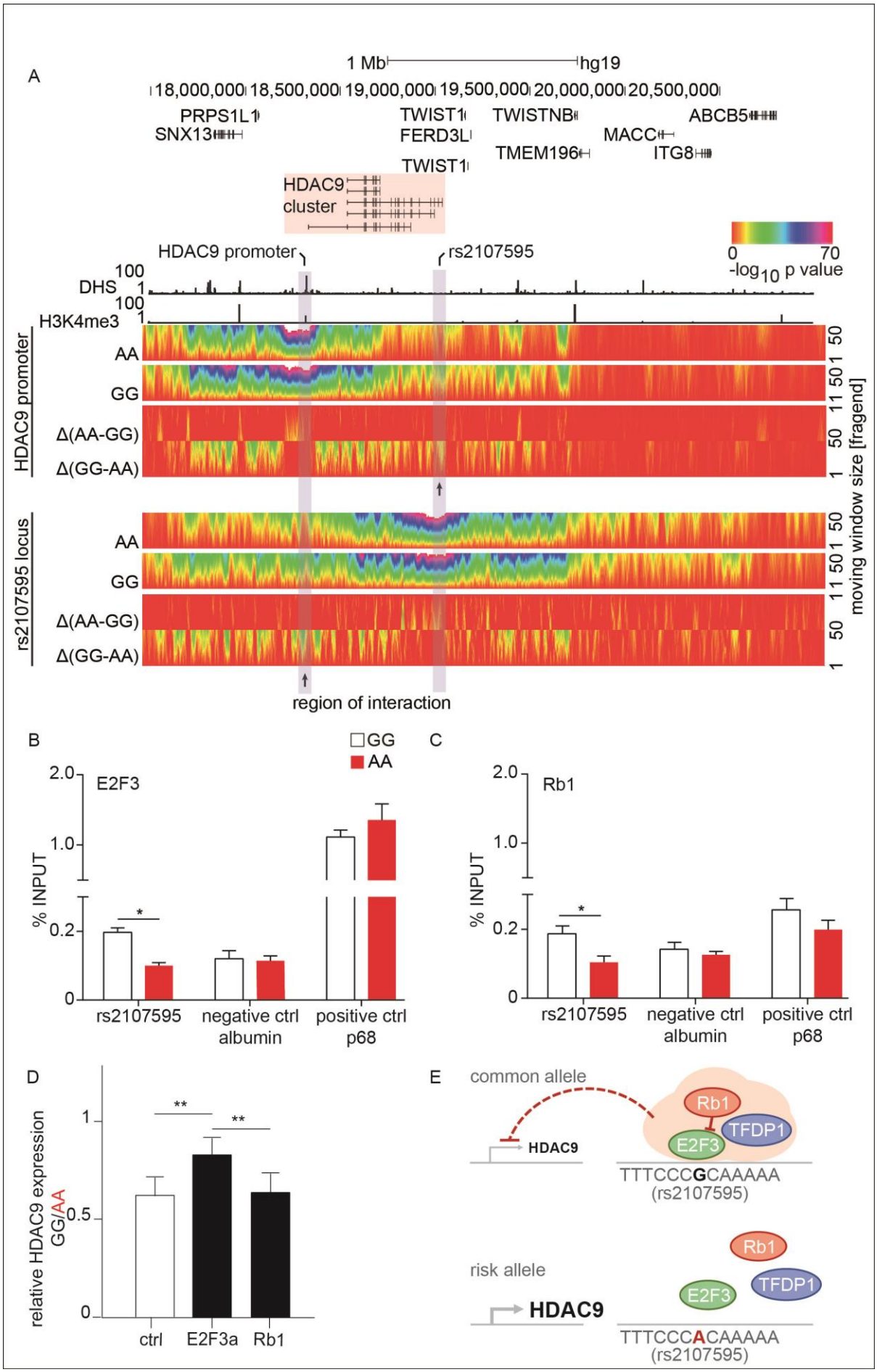

A
1 Mb
hg19
18,000,000
18,500,000
19,000,000
19,500,000
20,000,000
20,500,000
PRPS1L1
SNX13
TWIST1
FERD3L
TWIST1
TWISTNB
TMEM196
MACC
ITG8
ABCB5
HDAC9
cluster
HDAC9 promoter
rs2107595
0
70
-log10 p value
DHS
H3K4me3
HDAC9 promoter
AA
GG
Δ(AA-GG)
Δ(GG-AA)
rs2107595 locus
moving window size [fragend]
region of interaction
B
GG
AA
E2F3
% INPUT
rs2107595
negative ctrl albumin
positive ctrl p68
C
Rb1
D
relative HDAC9 expression GG/AA
ctrl
E2F3a
Rb1
E
common allele
HDAC9
Rb1
TFDP1
E2F3
TTTCCCGCAAAAA
(rs2107595)
risk allele
TTTCCCACAAAAA

Figure 4: E2F3 mediates allele-specific effects of rs2107595 on HDAC9 transcription. (**A**) Domain plot of the 4C-seq results obtained in isogenic Jurkat cells homozygous for the common (G) or risk allele (A). Shown are the significance levels of the 4C-seq signal coverage with viewpoints in the HDAC9 promoter (top) and rs2107595 region (bottom). For both viewpoints results for individual alleles are depicted in the upper panels with difference plots depicted below. Region of interactions (arrows) are defined by an enrichment of covered fragends within a running window of *1 to 50* fragends. Grey boxes represent the location of the 4C viewpoints. DHS and H3K4me3 histone marks are displayed at the top. (**B and C**) Comparative ChIP experiments in isogenic Jurkat cells homozygous for the common (G) or risk allele (A). G1/S boundary arrested cells showed an enrichment for E2F3 (**B**) and Rb1 (**C**) in common allele cells but not in risk allele cells at rs2107595. (n=6, mean±SEM, Wilcoxon Rank-Sum Test). (**D**) Influence of exogenous E2F3 and Rb1 expression in isogenic Jurkat cells. Compared to empty vector control (ctrl), overexpression of E2F3a but not Rb1 resulted in a significant increase of the ratio between HDAC9 expression in cells homozygous for the rs2107595 common allele (A) vs cells homozygous for the risk allele (G). (n=8-10, mean±SD, T-test). (**E**) Proposed model for the regulatory effect of rs2107595 on HDAC9 expression by allele-specific binding of the E2F3/Rb1/TFDP1 complex. In the presence of the common allele (G) the E2F3/Rb1/TFDP1 complex is recruited to the rs2107595 region and mediates a repressive effect on HDAC9 transcription. The risk allele (A) disrupts binding of the E2F3/Rb1/TFDP1 complex thus resulting in elevated HDAC9 expression.

Figure S1: Coverage Plot of the targeted resequencing experiment of the *HDAC9* region.
10 fold coverage calculated as N × L / G (N: number of reads,L: average read length, G: locus size), was achieved on average in this experiment. Individual samples with a percent coverage of < 80% were discarded from analysis. BC represents the barcode number.

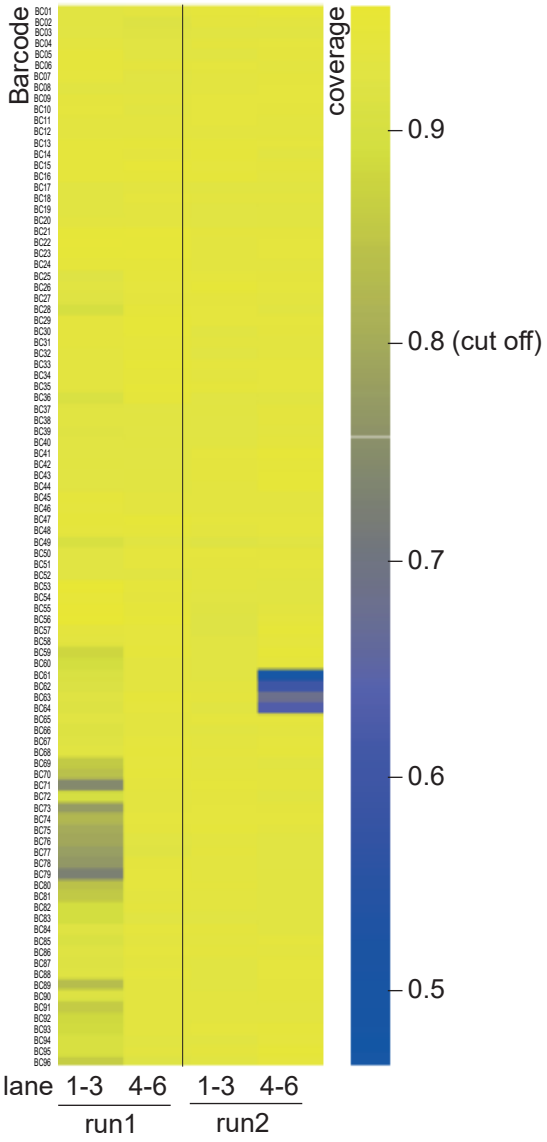


Figure S2: Chromatin landscape of the intergenic rs2107595 SNP region. Depicted are DHS sites, and the histone marks H3K4me1, H3K27ac and H3K9me3 surrounding the rs2107595 region in HeLa, HUVEC, $CD14^+$ monocytes, $CD3^+$ and Jurkat cells as well as aorta, if available (ENCODE, Roadmap Epigenomics).

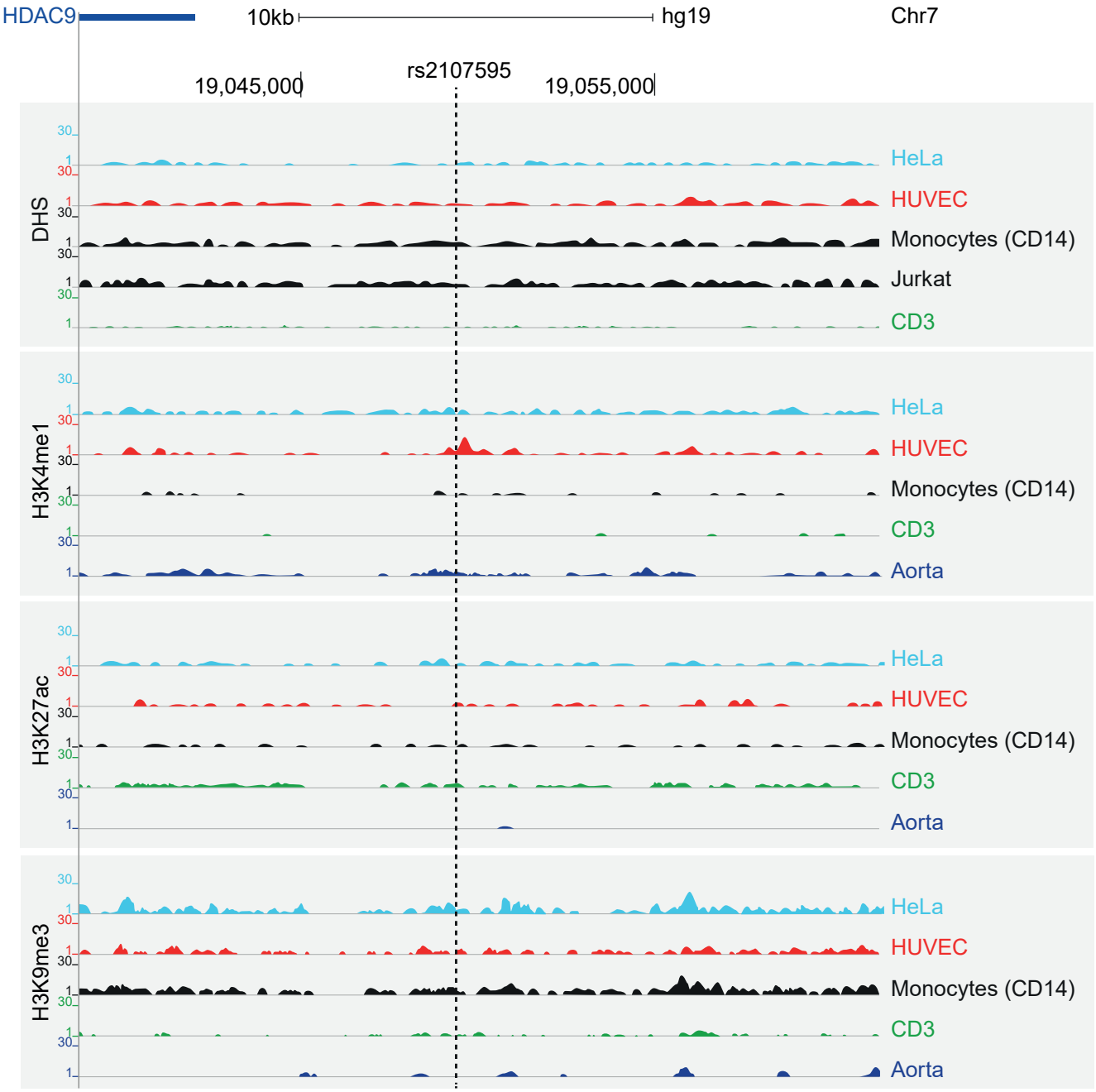


Figure S3: Control experiments for the gain- and loss-of-function approaches in HeLa cells. Overexpression of E2F3a and Rb1 was confirmed with quantitative real-time PCR analysis (**A**), (n=8), and immunoblotting (**B**). Expression levels of E2F3 and E2F4 after siRNA mediated knockdowns were analyzed on mRNA level (**C**), n=7-8, and protein level (**D**). Using siRNAs against Rb1, Rbl1 and Rbl2 expression levels were quantified on mRNA level (**E**), n=7, and controlled on protein level (**F**).

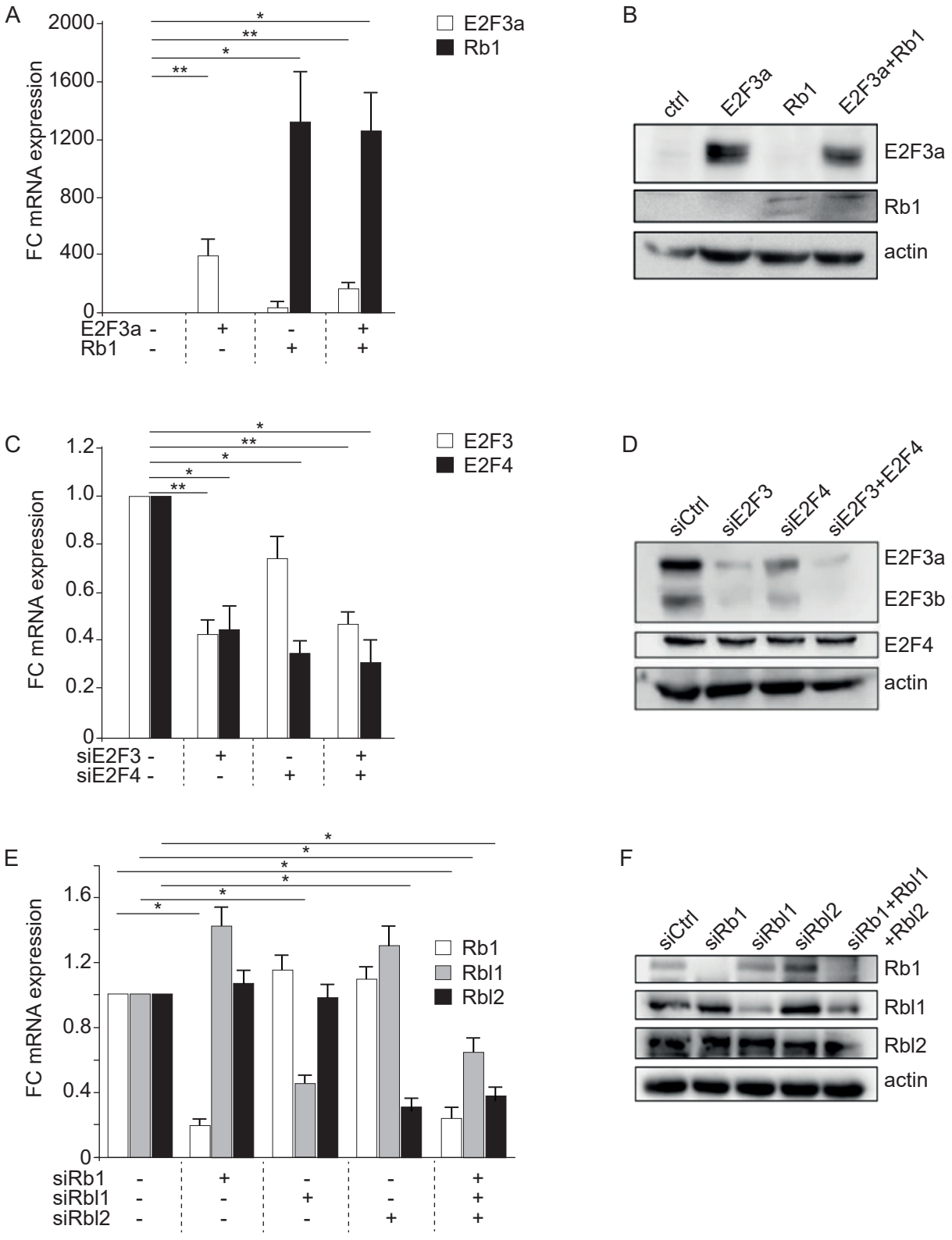


Figure S4: Gene expression analysis in primary human Mϕ and HAoSMCs. (**A**) Verification of monocyte to Mϕ differentiation by measuring gene expression of the Mϕ marker CD68. After 9 days in culture unstimulated and TNFα and IFNγ-stimulated Mϕ showed a 3-5 fold increase of CD68 expression compared to monocytes cultured for one day or primary PBMCs. (**B**) Shown are results for TWIST expression in HAoSMCs.There was no allele-specific effects on TWIST1 expression in HAoSMCs were observed. TWIST1 expression in human Mϕ was under the detection limit (results not shown).

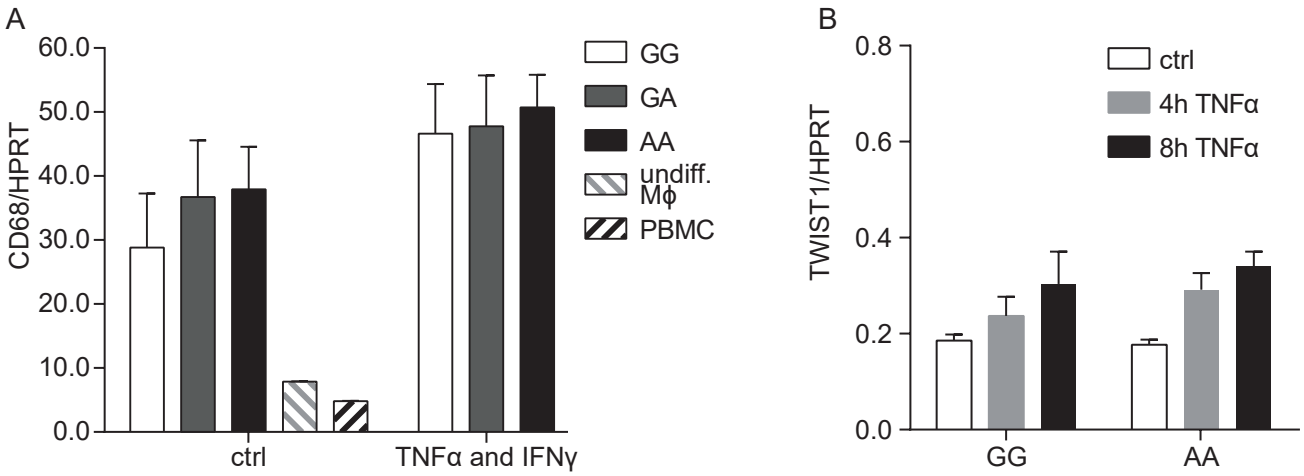


Figure S5: Luciferase assays in THP-1 cells, HAoECs and HAoSMCs. Shown are results for rs2107595, rs57301765 (r2=0.99) and rs10255384 (r2=0.47). (**A-C**) results for rs2107595 in THP-1 monocytes (**A**), HAoEC (**B**) and HAoSMC (**C**). (**D-E**) results for rs57301765 in Jurkat (**D**) and THP-1 MΦ (**E**). (**F-G**) results for rs10255384 in Jurkat (**F**) and THP-1 MΦ (**G**).

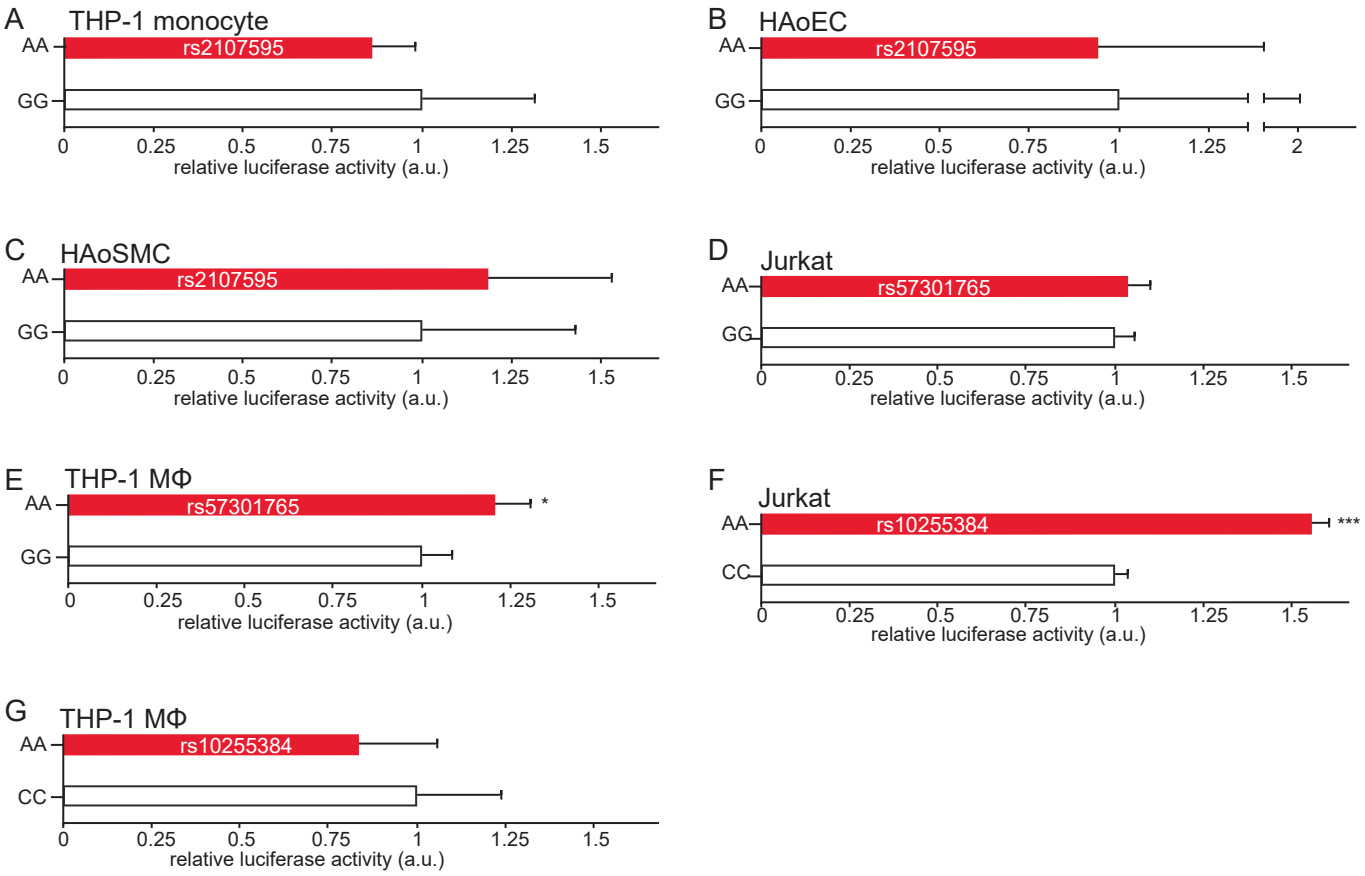


Figure S6: rs2107595 and cell proliferation. (**A**) Shown are representative figures from FACS analyses following EdU pulse-chase labelling in isogenic Jurkat cells at variable time points after pulse labelling. (**B**) Quantification showing the relative number of EdU positive cells in cells homozygous for the common (GG) and risk allele (AA).

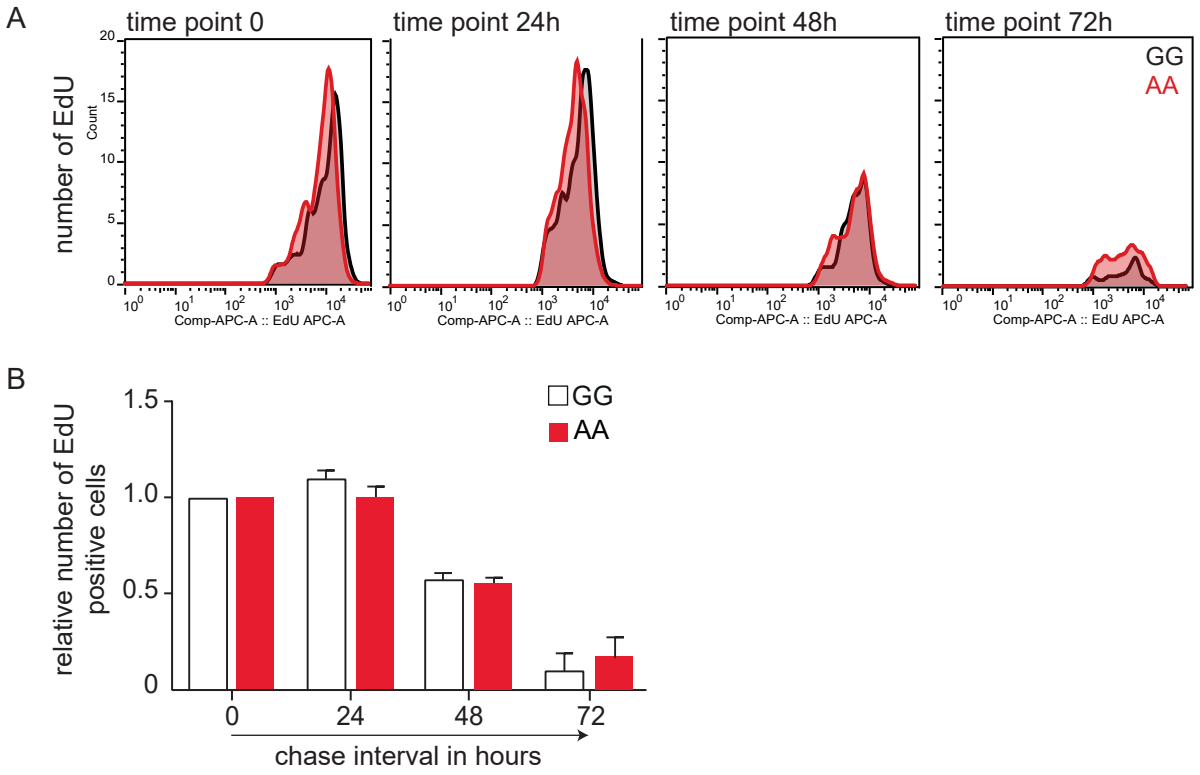


Figure S7: 4C analysis at the alternative HDAC9 promoter. Significance level of the 4C-seq signal from the second HDAC9 promoter - lacking detectable chromatin marks - and rs2107595 in Jurkat cells carrying the common (G) or risk allele (A). Shaded boxes beneath arrows represent the regions of interaction at the 4C viewpoints.

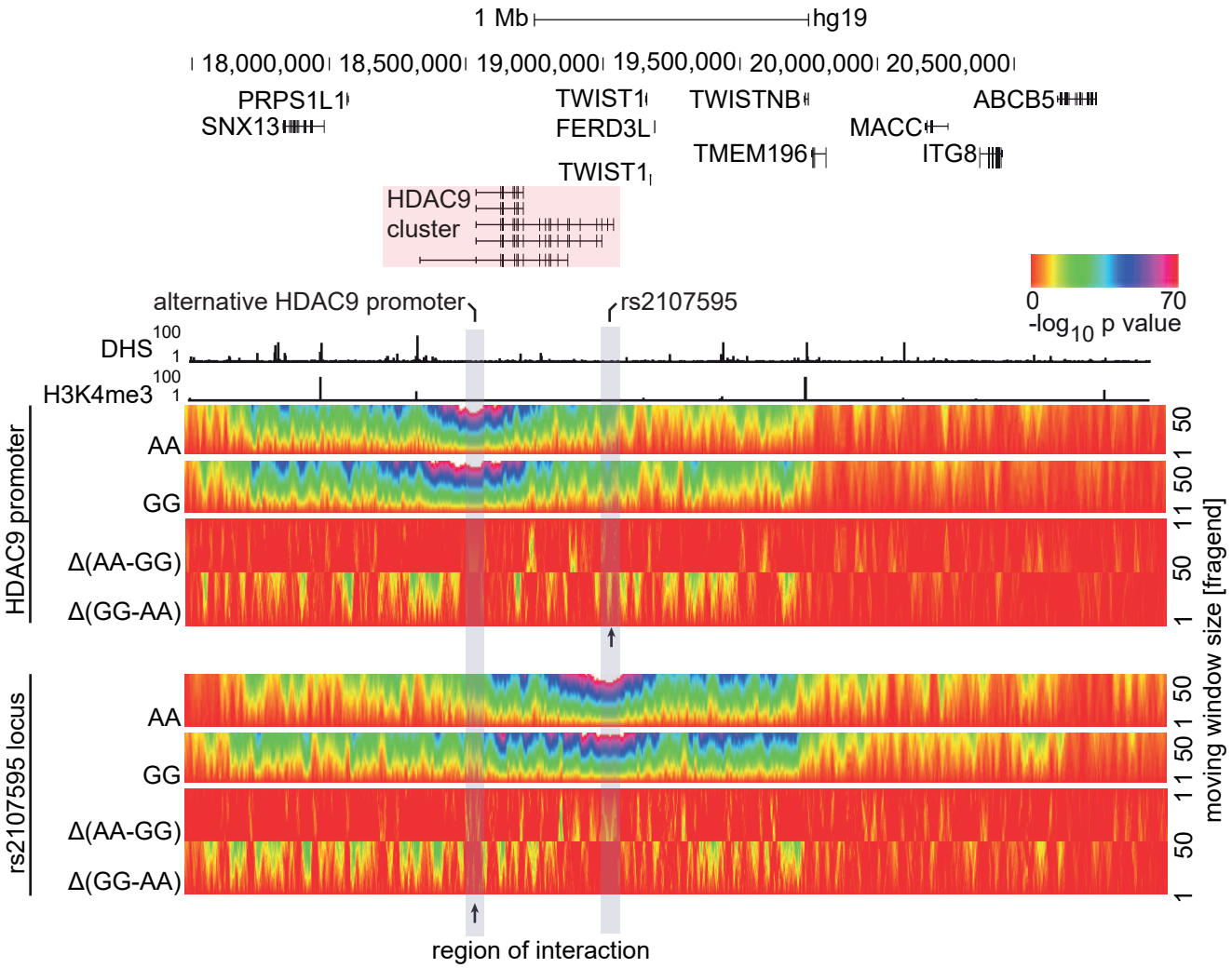
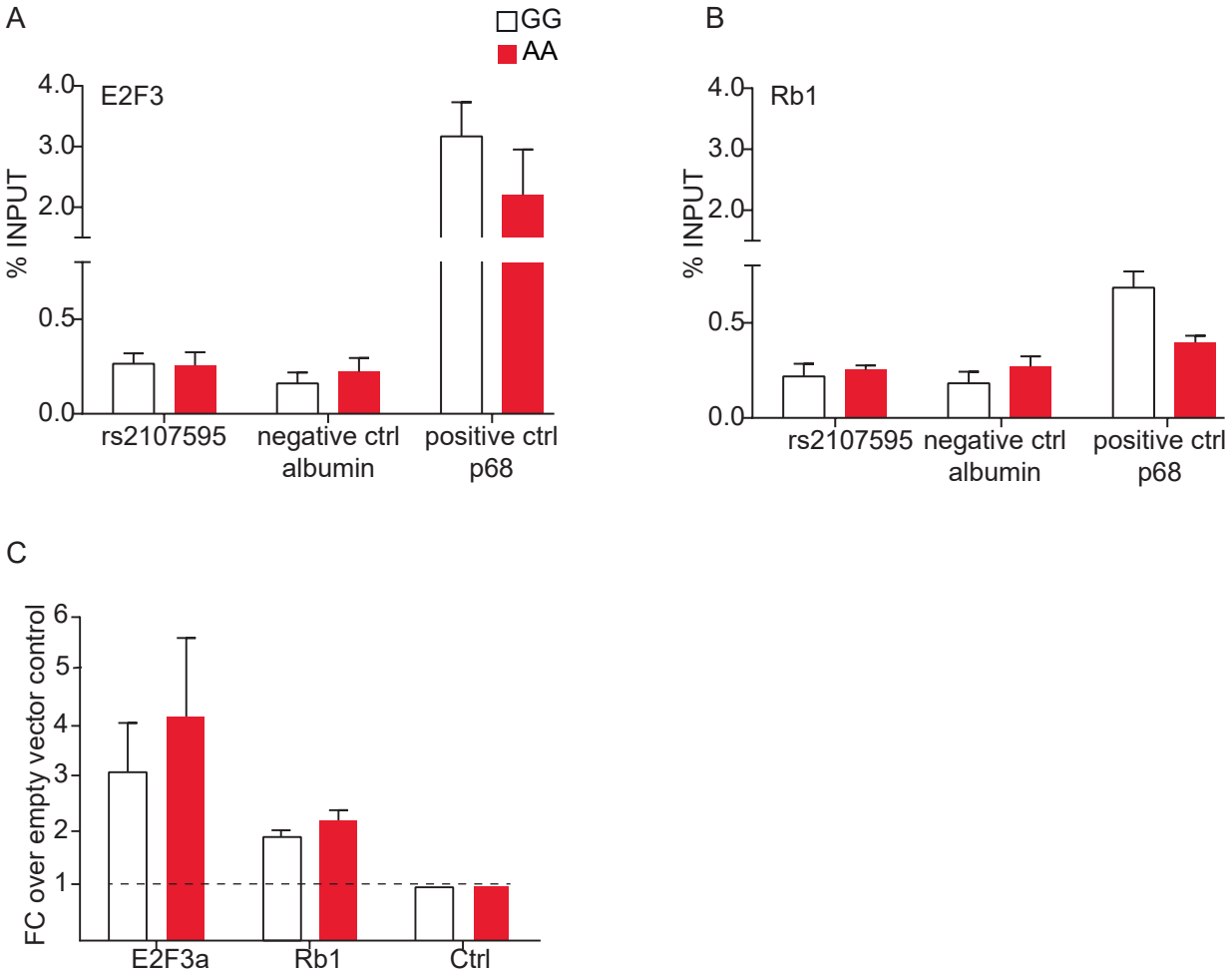


Figure S8: Role of E2F3 in mediating HDAC9 expression. Comparative ChIP experiments in isogenic Jurkat cells carrying either the common (G) or risk allele (A). Unsynchronized cells display background levels of E2F3 (**A**) and RB1 (**B**) occupancy at both alleles. n=4, mean+/-SEM. Albumin and p68 served as a negative and positive control, respectively. (**C**) Control experiment showing the transfection efficiency of E2F3a and Rb1 in isogenic Jurkat cells. Empty vector control was set to 1 (dashed line). n=4, mean +/-SEM.



Supplemental Material and Methods

Proteome-Wide Analysis of Disease-Associated SNPs (PWAS)

PWAS was conducted as previously described[1] with the following minor modifications:

11.25 nmol each of the corresponding pair of bait oligonucleotides (Supplemental Table 2) were annealed and phosphorylated using polynucleotide kinase (Fermentas). Concatemers were generated by ligating the bait oligonucleotides using T4 ligase (20 U, Fermentas) overnight at room temperature. 0.5 nmol of pre-annealed desthiobiotinylated adapter oligonucleotides were ligated to the concatemer ends and subsequently purified using a G50 column (GE Healthcare). SNP pull-down and LC-MS/MS samples processing was performed as described[1] and subsequently analyzed by nanoflow liquid chromatography on an EASY-nLC system from Proxeon Biosystems into a Q Exactive or Q Exactive HF (Thermo Fisher Scientific). Peptides were separated via HPLC on a C18-reversed phase 200 mm column packed with Reprosil (Dr. Maisch).

Raw files were processed with MaxQuant[2] (version 1.5.0.0) and searched against the human UniProt database. Search results were processed with MaxQuant filtered with a false discovery rate of 0.01. Data handling and outlier definition were performed using the Perseus software package (version 1.4.2.35).[3] Heavy over light ratios of identified proteins were filtered for contaminants, logarithmized, and the results of both forward and reverse experiments were plotted against each other. For outlier definition the heavy over light ratios were z-scored and a 1% cutoff was applied. Proteins were considered statistically significant outliers if this cutoff level was met in both forward and reverse experiments.

For E2F3 consensus site prediction the Match™ tool (Gene-Regulation)[4] was used.

Circular Chromosome Conformation Capture

4C-chromatin was prepared as described previously.[5] In brief, 10 million cells were used for crosslinking in 2% formaldehyde (Sigma), lysed in 50mmol/L TRIS pH 7.5, 150mmol/L NaCl, 5mmol/L EDTA, 0.5% NP-40, 1% Triton X-100. Isolated chromatin was subjected to a digestion with with DpnII (NEB, #R0543L), a ligation by T4 DNA ligase (Roche), a second digestion by CviQI (NEB, #R069S), a second ligation and a purification. Digestion and ligation efficiencies were checked on agarose gels. For 4C-sequencing library preparation PCR of 1.6 µg of 4C template per reaction was performed by multiplexing 4 to 10 primer pairs (Supplemental Table 2) in the initial PCR reaction and subsequent pooling according to primer efficiency. After an initial PCR reaction of 6 cycles (reaction volume = 200 µL) individual samples were divided among PCR reactions containing single primer pairs for another 26 cycles (reaction volume = 25 µL). PCR products (20 ng, 100 µL volume) derived from the same Jurkat cell genotype were pooled in equimolar amounts and a final 6-cycle PCR reaction was performed with primers containing sequencing adapter and barcode sequences. After size selection by agarose gel electrophoresis fragments <700 bp were sequenced using the NextSeq500 platform (Illumina), producing single end reads of 75 bp.

The raw sequencing reads were de-multiplexed based on viewpoint specific primer sequences. Reads were trimmed to 16 bases and mapped to an in silico generated library of fragends (fragment ends) neighbouring all DpnII sites in human genome (NCBI37/hg19), using custom Perl scripts. No mismatches were allowed during the mapping and the reads mapping to only one fragend were used for further analysis. For visualization of the 4C-signal the number of covered fragends mapping to chromosome 7 was calculated within a running window of k fragends (k was set increasingly from 1 to 50). The number of covered fragends in each running window was compared to the number of covered fragends expected by random distribution. When the number of the covered fragends was higher compared to the number expected from the random distribution the significance level was calculated using the binominal cumulative distribution function; R *pbinom*.

1. Butter F, Davison L, Viturawong T, et al. Proteome-wide analysis of disease-associated snps that show allele-specific transcription factor binding. *PLoS Genet*. 2012;8:e1002982. doi: 10.1371/journal.pgen.1002982.
2. Cox J, Mann M. Maxquant enables high peptide identification rates, individualized p.P.B.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26:1367-1372. doi: 10.1038/nbt.1511.

3. Tyanova S, Temu T, Sinitcyn P, et al. The perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*. 2016;13:731-740. doi: 10.1038/nmeth.3901.
4. Matys V, Fricke E, Geffers R, et al. Transfac: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31:374-378. doi.
5. van de Werken HJ, de Vree PJ, Splinter E, et al. 4c technology: Protocols and data analysis. *Methods Enzymol*. 2012;513:89-112. doi: 10.1016/B978-0-12-391938-0.00004-5.

6 DISCUSSION

In the present work disease phenotypization, genetics and proteomics are integrated into a comprehensive workflow, depicting the scientific discovery from phenotype to function.

First, we applied the novel web-based classification system CCS[28,29] to more than 16.000 IS patients from the US and Europe, considering clinical, laboratory, imaging as well as additional technical data for the delineation of phenotypic and causative stroke subtypes.[92] Interrater reliability was higher (κ = 0.72) than in comparison with the traditional TOAST classification system.[93] Importantly, via systematic probabilistic scoring of individual phenotypic findings, CCS also assigned the most probable cause of IS with good interrater reliability (κ = 0.75).
Second, we performed a two-stage GWAS with IS phenotypization based on CCS and TOAST, testing the associations of SNPs with a total of more than 37.000 IS patients and almost 400.000 control subjects.[94] A novel locus near *TSPAN2* significantly associated with LAS, 4 previously identified loci (*PITX2*, *ZFHX3*, *HDAC9* and *ALDH2*) were replicated and associated in a stroke subtype-specific manner as well.
Third, we applied a modern DNA-protein pulldown approach coupled to LC-MS/MS to the lead SNP rs2107595 at the LAS risk locus *HDAC9*, identifying preferential binding of an E2F3-TFDP1-Rb1 complex to the common allele of rs2107595 (see section 5, Prestel et al.). Additional gain- and loss-of-function-studies, reporter assays and chromosome conformation capture provide evidence for an interaction of rs2107595 with the *HDAC9* promoter and allele-specific regulation of HDAC9 expression via E2F3 and Rb1.

Taken together, this workflow represents a compelling application of modern mass spectrometry-based proteomics to stroke genetics and showcases its benefit for the functional follow-up of GWAS hits.

6.1 Ischemic Stroke, a Complex Phenotype of Complex Diseases

Besides being a complex disease caused by the interplay of environmental factors, lifestyle and a multitude of common variants, IS also is a complex phenotype with several distinct pathomechanisms potentially leading to the same endpoint, i.e. ischemic stroke. This interplay of complex disease and complex phenotype is critical in the understanding of IS, both for clinical routine as well as for stroke genetics.

In clinical routine, IS patients are etiologically classified according to the widespread TOAST system,[26] with the distinction between LAS and cardioembolism being one of the most important clinical differentiations to guide the choice of secondary prevention via antiplatelet agents or anticoagulants, respectively.[95] When applying the TOAST classification system to IS patients in a hospital setting, about 30-50 % of IS cases are rated "undetermined",[96,97] raising questions about the right secondary prophylaxis for affected IS patients. The CCS system on the other hand proves to be a relatively easy-to-use web-based classification algorithm for IS,[29,98] which can be helpful in the assignment of the causative phenotype via its probabilistic stratification of individual IS risk factors.[99]

However, even the application of CCS rates about 25 % of IS cases as "undetermined".[100] Two major obstacles for IS phenotypization are limited diagnostic sensitivity of today's routinely used tests as well as a high percentage of incomplete diagnostic workup in the setting of acute stroke care.[92,101] For instance, the likelihood of detecting intermittent atrial fibrillation with one 72 hour ECG is just about 20 %,[102,103] generating the necessity for more stringent diagnostic work-ups. With atrial fibrillation most likely being underdiagnosed in IS patients, the association of risk loci of cardioembolism with IS in general could be partially explained.[37]

Due to this diagnostic uncertainty a clinical concept is currently being evaluated for secondary prevention of IS, "embolic stroke of undetermined source" (ESUS).[104] Defined as a non-lacunar cerebral infarction without evidence for proximal arterial stenosis nor atrial fibrillation,[105] ESUS is likely to be a subset of the undetermined IS cases. For ESUS, the benefit of new oral anticoagulants vs. antiplatelet agents for prevention of recurrent IS is currently tested in randomized multicentre trials. While RE-SPECT ESUS[106] and NAVIGATE ESUS[107] did not find any benefit for dabigatran or rivaroxaban, respectively, vs. acetylsalicylic acid, the trials ATTICUS and ARCADIA comparing apixaban vs. acetylsalicylic acid are still ongoing.

Additionally, technological advancements that might improve the accuracy of causative IS classification are arriving in clinical stroke care as well.[108] Recently the use of event recorders, small implantable ECG devices, proved to be superior in detection of intermittent atrial fibrillation than conventional follow-up after stroke.[109] Such developments are of particular interest, since missing out on intermittent atrial fibrillation will unquestionably result both in inadequate medication for secondary prevention as well as in inferiorly curated IS phenotypization and consequently in less-powered GWAS.

As for stroke genetics, GWAS remain the state-of-the-art technology for elucidating genetic heritability of complex diseases. As evidenced in the present work for IS, precise phenotyping is crucial, particularly for IS with its heterogenous etiologies. Here, the application of the refined CCS classification system to a two-stage GWAS lead to the identification of *TSPAN2*, a previously unknown subtype-specific risk locus for LAS.[94] Since both classification systems, TOAST and CCS, showed moderate correlation in terms of their respective causative stroke classification,[100] it is not surprising that TOAST and CCS also showed moderate to strong genetic correlation.[94] Hence, for future GWAS in the setting of IS, the use of a well-curated phenotypization per se might be more important for identification of novel risk loci than the choice between TOAST and CCS itself.

Moreover, the study of larger IS sample sizes in a subtype-specific manner will possibly generate the biggest improvement in identification of additional risk loci.[110,111] This is evidenced by the success of two recent multicohort GWAS meta-analyses,[38,39] bringing the total number of significant risk loci for stroke to 35. IS subtype-specificity of the identified risk loci was demonstrated by the fact that all loci reaching genome-wide significance for one IS subtype did not reach genome-wide significance for an alternate IS subtype.[38] Additional risk loci might reach genome-wide significance for a specific IS subtype once subtype information becomes available for the UK biobank cohort.[39] Interestingly however, two loci showed evidence of a shared genetic influence both on LAS and cardioembolism (*SH2B3* and *ABO*), hinting at some degree of shared pathophysiology between these two etiologies. Furthermore, there was strong evidence for shared genetic variation with related vascular phenotypes such as bloodpressure, CAD and MI.[5,38]

As with all conventional GWAS, the aforementioned studies robustly uncover common variants with almost linear relation to their sample sizes.[110,111] However, rare variants with moderate to high effect sizes and minor allele frequencies of < 1 % are typically not represented.[112,113] Future developments in stroke genetics will therefore most likely incorporate new technologies such as whole-exome sequencing or whole-genome sequencing in order to uncover additional layers of IS heritability.[114,115]

6.2 Modern Proteomics and its Application to Stroke Genetics

Modern proteomics have tremendous applications for both basic biology and functional genomics. Today, high resolution quantitative MS-based proteomics is the *de facto* standard for the unbiased study of global protein dynamics and protein-protein interactions from a complex sample.[116,117] Due to technological and methodological advancements in all aspects of the proteomics workflow, MS-based proteomics is increasingly capable to quantify low abundance proteins, DNA-protein interactions and even specific histone marks from complex samples.[118,119]

In clinical genetics, one major obstacle is the identification of causal variants and uncovering the molecular mechanisms mediating disease risk, because like in the case of IS the vast majority of disease-associated SNPs locate to non-coding regions of the genome.[38,39,52,53] Due to the large-scale multicentre effort Encyclopedia of DNA elements (ENCODE[55,56]), systematically employing functional genomics experiments such as ChIP-seq and DNase-seq, the concentration of disease-associated SNPs in regulatory DNA as marked by DHSs was shown.[53] Hence, disease-associated SNPs from non-coding regions might mediate their risk effects via gain- or loss-of-function mutations in regulatory DNA elements.[54,120,121] This was e.g. demonstrated for the *SORT1* locus associated with cholesterol levels and MI, where the common variant rs12740374 resides in a binding site for the transcription factor C/EBP, resulting in allele-specific hepatic expression of SORT1.[120]

In order to screen for allele-specific binding of transcription factors, we employed interaction proteomics via PWAS as an unbiased DNA-centric approach.[91] As demonstrated for the vascular risk locus *HDAC9*, this DNA-protein pulldown coupled to LC-MS/MS readily identified allele-specific binding of the E2F3/Rb1-complex to rs2107595 and thus facilitated the

discovery of the probable molecular mechanism mediating the risk effect towards CAD, MI, and among others, LAS (see section 5). In comparison to functional genomics experiments such as protein-centric ChIP-seq,[122,123] the DNA-centric PWAS approach does not require *a priori* knowledge of the transcription factors involved. This potentially allows the identification of allele-specific binding of transcription factors for which (1) no functional genomics data via ChIP-seq are available, (2) no consensus binding site is available for *in silico* predictions[124], or (3) non-DNA binding cofactors of transcription factor complexes exist.[125,126] In addition, PWAS is virtually applicable to any phenotype of interest with synthesis of custom DNA oligos being easily accessible.

However, one limitation of PWAS is the proteomic input material for the IP-MS experiment: depending on the cell line used for nuclear extract preparation,[91] for some highly differentially expressed transcription factors allele-specific binding might be missed due to extremely low abundance, potentially resulting in false-negative PWAS studies. One approach to overcome this limitation might be the application of label-free quantification[86] in combination with PWAS, as this would allow the use of biologically relevant primary cells or tissues as proteomic input material.[127]

Another important limitation of PWAS relates to its *in vitro* assay with synthetic DNA oligos, which do not necessarily resemble the precise *in vivo* chromatin context of the analyzed SNP. Since DNA-protein interactions *in vivo* are not only dependent on the isolated DNA sequence, but on histones, histone modifications and the 3D chromatin structure as well,[128] this might be another source of false-negative or false-positive PWAS results. Here, additional studies like locus-specific ChIP-MS[129] in combination with genome editing[130] could have added value in corroborating *in vitro* PWAS hits, albeit locus-specific ChIP-MS currently still being severely limited by its sensitivity and not being feasible for a multi-variant screening approach.[131]

In due consideration of its limitations, PWAS does provide a scalable DNA-centric assay that can be used to screen multiple SNPs in a reasonable amount of time.[87,90] This is especially important when considering the number of SNPs that have to be screened per disease-associated risk locus. For significantly associated risk loci, fine mapping e.g. via resequencing or 1000G imputation generates insight into its haplotype structure, usually revealing multiple SNPs that could have causative effects.[132,133] While prioritization of SNPs for follow-up via

epigenetic features seems promising,[53,134] at least for some loci a summation effect of multiple SNPs cooperatively modulating gene expression seems entirely possible.[91,110] Once a prioritization and selection of target SNPs for follow-up has been performed, PWAS may then be applied to gain additional biological insight into target SNP functionality, as demonstrated with the analysis of rs2107595 at the *HDAC9* locus (see section 5). With current MS technology and methodology, the throughput for PWAS analysis is about 5-6 SNPs per day and mass spectrometer. Due to modern advances in mass spectrometer technology[135] and methodological improvements,[136,137] this throughput will further increase, possibly allowing genome-wide analyses of disease-associated SNPs in the near future.

6.3 From Bed to Bench and Back

This work integrates all aspects from phenotypization to genetics and proteomics into a comprehensive workflow, exemplifying how to get from phenotype to function in the case of IS, famously corresponding to the phrase “from bed to bench”. However, to suffice for the modern label of “translational medicine”,[138] one question remains: how to get back to bed?

With our improving knowledge of human genetic variation and its relevance for disease risk, personalised medicine has become increasingly important.[139] So far, however, attempts in personalised risk stratification or causative classification based on the knowledge of individual’s common variants have been unrewarding in the setting of IS,[140] most likely due to the SNPs’ relatively small effect sizes. Today, one prominent example of stroke genetics and its successful application in clinical routine is Fabry’s disease.[141] Identification of such mendelian disorders is especially important among young patients with stroke of undetermined cause,[142] as recurrence of IS is typically high and enzyme replacement therapy might improve longterm outcome of these patients.[143]

Another relevant example of stroke genetics is its application in pharmacogenetic decision making. A variety of SNPs have been shown to critically affect metabolization of drugs, thus possibly decreasing efficacy or resulting in an increase of drug-related adverse events, including warfarin,[144] clopidogrel[145] and dabigatran.[146] For warfarin, genotype-dependent dosage even showed a reduction in drug-related adverse events and was implemented into FDA guidelines.[147,148]

For common genetic variants, the focus of the present work, identification of risk loci and their molecular mechanisms will ideally result in drug target identification and drug development for primary, secondary or tertiary prevention of human diseases such as IS. In the case of *HDAC9*, risk locus fine mapping taken together with functional data from ENCODE strongly hinted towards a causative role of rs2107595.[40,41,56] Via LC-MS/MS and extensive follow-up experiments, we generated conclusive evidence for an interaction of rs2107595 with the *HDAC9* promoter and allele-specific regulation of HDAC9 expression via E2F3 and Rb1 (see section 5; Prestel et al.). These results are in line with previous studies which found HDAC9 to be overexpressed in human atherosclerotic plaques,[42] while HDAC9 deficiency in mice led to a reduction of atherosclerotic lesions.[47] Besides LAS HDAC9 is also strongly implicated in CAD and MI,[5,6,38] rendering HDAC9 a promising drug target for a variety of vascular phenotypes. Indeed, data suggests that valproate, typically used as an antiepileptic drug, functions as a non-specific HDAC inhibitor[149] and results in reduction of atherosclerotic lesions in apoE-deficient mice.[150] Also, exposure to valproate was linked to reduced stroke recurrence rate.[151] Despite that, little is known about human HDAC9 physiology, which will become increasingly important with the advent of more specific and more potent HDAC9 inhibitors. Besides its involvement in the pathogenesis of atherosclerosis,[42,47] HDAC9 has been implicated in schizophrenia[152] as well as increased Treg activity in the setting of autoimmune disorders.[49,50] Once a specific and potent HDAC9 inhibitor is available, further animal studies and ultimately clinical studies will be required in order to determine efficacy and safety for primary or secondary prevention of stroke.

On a more general scheme, research on one risk locus including functional follow-up studies clearly is a multi-year effort, as evidenced by the present work on IS. Ultimately, stroke genetics and its subsequent translation into clinical neurology can only be as effective as the whole chain of translational research. As such, further improvements in data acquisition during clinical routine, stringent probabilistic phenotypization of stroke cases, large multi-cohort collaborative GWAS consortia, next-generation sequencing technologies and modern functional follow-up of associated risk loci will unquestionably result in a deeper understanding of stroke heritability and additional gene discovery for medical treatment.

7 REFERENCES

1. Sacco, R. L. *et al.* An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **44,** 2064–2089 (2013).
2. GBD 2016 Lifetime Risk of Stroke Collaborators *et al.* Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016. *N. Engl. J. Med.* **379,** 2429–2437 (2018).
3. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390,** 1151–1210 (2017).
4. Meschia, J. F. *et al.* Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **45,** 3754–3832 (2014).
5. Dichgans, M. *et al.* Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* **45,** 24–36 (2014).
6. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* **45,** 25–33 (2013).
7. Malik, R. & Dichgans, M. Challenges and opportunities in stroke genetics. *Cardiovasc. Res.* **114,** 1226–1240 (2018).
8. Magistretti, P. J. & Allaman, I. A cellular perspective on brain energy metabolism and functional imaging. *Neuron* **86,** 883–901 (2015).
9. Chamorro, Á., Dirnagl, U., Urra, X. & Planas, A. M. Neuroprotection in acute stroke: targeting excitotoxicity, oxidative and nitrosative stress, and inflammation. *Lancet Neurol* **15,** 869–881 (2016).
10. Khoshnam, S. E., Winlow, W., Farzaneh, M., Farbood, Y. & Moghaddam, H. F. Pathogenic mechanisms following ischemic stroke. *Neurol. Sci.* **38,** 1167–1186 (2017).
11. Lo, E. H., Dalkara, T. & Moskowitz, M. A. Mechanisms, challenges and opportunities in stroke. *Nat. Rev. Neurosci.* **4,** 399–415 (2003).
12. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *N. Engl. J. Med.* **333,** 1581–1587 (1995).
13. Hacke, W. *et al.* Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N. Engl. J. Med.* **359,** 1317–1329 (2008).
14. Riedel, C. H. *et al.* The importance of size: successful recanalization by intravenous thrombolysis in acute anterior stroke depends on thrombus length. *Stroke* **42,** 1775–1777 (2011).
15. Powers, W. J. *et al.* 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke* **49,** e46–e110 (2018).

16. Jovin, T. G. *et al.* Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N. Engl. J. Med.* **372,** 2296–2306 (2015).

17. Berkhemer, O. A. *et al.* A randomized trial of intraarterial treatment for acute ischemic stroke. *N. Engl. J. Med.* **372,** 11–20 (2015).

18. Saver, J. L. *et al.* Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N. Engl. J. Med.* **372,** 2285–2295 (2015).

19. Goyal, M. *et al.* Randomized assessment of rapid endovascular treatment of ischemic stroke. *N. Engl. J. Med.* **372,** 1019–1030 (2015).

20. Campbell, B. C. V. *et al.* Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N. Engl. J. Med.* **372,** 1009–1018 (2015).

21. Mocco, J. *et al.* Aspiration Thrombectomy After Intravenous Alteplase Versus Intravenous Alteplase Alone. *Stroke* **47,** 2331–2338 (2016).

22. Bracard, S. *et al.* Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol* **15,** 1138–1147 (2016).

23. Goyal, M. *et al.* Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet* **387,** 1723–1731 (2016).

24. Nogueira, R. G. *et al.* Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct. *N. Engl. J. Med.* **378,** 11–21 (2018).

25. Albers, G. W. *et al.* Thrombectomy for Stroke at 6 to 16 Hours with Selection by Perfusion Imaging. *N. Engl. J. Med.* **378,** 708–718 (2018).

26. Adams, H. P. Jr.*, 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* **24**, 35-41 (1993).

27. Ay, H. Advances in the diagnosis of etiologic subtypes of ischemic stroke. *Curr Neurol Neurosci Rep* **10,** 14–20 (2010).

28. Ay, H. *et al.* An evidence-based causative classification system for acute ischemic stroke. *Ann. Neurol.* **58,** 688–697 (2005).

29. Ay, H. *et al.* A computerized algorithm for etiologic classification of ischemic stroke: the Causative Classification of Stroke System. *Stroke* **38,** 2979–2984 (2007).

30. Falcone, G. J., Malik, R., Dichgans, M. & Rosand, J. Current concepts and clinical applications of stroke genetics. *Lancet Neurol* **13,** 405–418 (2014).

31. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

32. Di Donato, I. *et al.* Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL) as a model of small vessel disease: update on clinical, diagnostic, and management aspects. *BMC Med* **15,** 41 (2017).

33. Dehghan, A. Genome-Wide Association Studies. *Methods Mol. Biol.* **1793,** 37–49 (2018).

34. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43,** 519–525 (2011).

35. Dichgans, M. & Markus, H. S. Genetic association studies in stroke: methodological issues and proposed standard criteria. *Stroke* **36,** 2027–2031 (2005).

36. Bevan, S. *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke* **43,** 3161–3167 (2012).

37. Gretarsdottir, S. *et al.* Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann. Neurol.* **64,** 402–409 (2008).

38. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* **50,** 524–537 (2018).

39. Malik, R. *et al.* Genome-wide meta-analysis identifies 3 novel loci associated with stroke. *Ann. Neurol.* **84,** 934–939 (2018).

40. Traylor, M. *et al.* Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol* **11,** 951–962 (2012).

41. International Stroke Genetics Consortium (ISGC) *et al.* Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* **44,** 328–333 (2012).

42. Markus, H. S. *et al.* Evidence HDAC9 genetic variant associated with ischemic stroke increases risk via promoting carotid atherosclerosis. *Stroke* **44,** 1220–1225 (2013).

43. Matsukura, M. *et al.* Genome-Wide Association Study of Peripheral Arterial Disease in a Japanese Population. *PLoS ONE* **10,** e0139262 (2015).

44. Falkenberg, K. J. & Johnstone, R. W. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat Rev Drug Discov* **13,** 673–691 (2014).

45. Parra, M. Class IIa HDACs - new insights into their functions in physiology and pathology. *FEBS J.* **282,** 1736–1744 (2015).

46. Verdin, E., Dequiedt, F. & Kasler, H. G. Class II histone deacetylases: versatile regulators. *Trends Genet.* **19,** 286–293 (2003).

47. Azghandi, S. *et al.* Deficiency of the stroke relevant HDAC9 gene attenuates atherosclerosis in accord with allele-specific effects at 7p21.1. *Stroke* **46,** 197–202 (2015).

48. Cao, Q. *et al.* Histone deacetylase 9 represses cholesterol efflux and alternatively activated macrophages in atherosclerosis development. *Arterioscler. Thromb. Vasc. Biol.* **34,** 1871–1879 (2014).

49. Tao, R. *et al.* Deacetylase inhibition promotes the generation and function of regulatory T cells. *Nat. Med.* **13,** 1299–1307 (2007).

50. de Zoeten, E. F., Wang, L., Sai, H., Dillmann, W. H. & Hancock, W. W. Inhibition of HDAC9 increases T regulatory cell function and prevents colitis in mice. *Gastroenterology* **138,** 583–594 (2010).

51. Li, X. *et al.* Methyltransferase Dnmt3a upregulates HDAC9 to deacetylate the kinase TBK1 for activation of antiviral innate immunity. *Nat. Immunol.* **17,** 806–815 (2016).

52. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).

53. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

54. Spielmann, M. & Klopocki, E. CNVs of noncoding cis-regulatory elements in human disease. *Curr. Opin. Genet. Dev.* **23,** 249–256 (2013).

55. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).

56. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

57. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

58. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).

59. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 21931–21936 (2010).

60. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).

61. Henikoff, S. & Greally, J. M. Epigenetics, cellular memory and gene regulation. *Curr. Biol.* **26,** R644–8 (2016).

62. Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N.Y.)* **14,** 61–65 (1996).

63. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem (Palo Alto Calif)* **8,** 61–80 (2015).

64. Nolting, D., Malek, R. & Makarov, A. Ion traps in modern mass spectrometry. *Mass Spectrom Rev* **38,** 150–168 (2019).

65. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).

66. Mann, M. Origins of mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **17,** 678–678 (2016).

67. Ebhardt, H. A., Root, A., Sander, C. & Aebersold, R. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* **15,** 3193–3208 (2015).

68. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif)* **9,** 449–472 (2016).

69. Thakur, S. S. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics* **10,** M110.003699 (2011).

70. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793 (2011).

71. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690 (2001).

72. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).

73. Whitehouse, C. M., Dreyer, R. N., Yamashita, M. & Fenn, J. B. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679 (1985).

74. Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80,** 273–299 (2011).

75. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162 (2000).

76. Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **17,** 1–35 (1998).

77. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

78. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–740 (2016).

79. Wierer, M. & Mann, M. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Hum. Mol. Genet.* **25,** R106–R114 (2016).

80. Vermeulen, M. *et al.* Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142,** 967–980 (2010).

81. Eberl, H. C., Mann, M. & Vermeulen, M. Quantitative proteomics for epigenetics. *Chembiochem* **12,** 224–234 (2011).

82. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M. & Mann, M. A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics. *Mol. Cell* **49,** 368–378 (2013).

83. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12,** 1576–1590 (2012).

84. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1,** 376–386 (2002).

85. Ong, S.-E. & Mann, M. Stable isotope labeling by amino acids in cell culture for quantitative proteomics. *Methods Mol. Biol.* **359,** 37–52 (2007).

86. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13,** 2513–2526 (2014).

87. Viturawong, T., Meissner, F., Butter, F. & Mann, M. A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. *Cell Rep* **5,** 531–545 (2013).

88. Scheibe, M. *et al.* Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. *Genome Res.* **23,** 2149–2157 (2013).

89. Tyanova, S., Mann, M. & Cox, J. MaxQuant for in-depth analysis of large SILAC datasets. *Methods Mol. Biol.* **1188,** 351–364 (2014).

90. Mittler, G., Butter, F. & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res.* **19,** 284–293 (2009).

91. Butter, F. *et al.* Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet.* **8,** e1002982 (2012).

92. Ay, H. *et al.* Pathogenic ischemic stroke phenotypes in the NINDS-stroke genetics network. *Stroke* **45,** 3589–3596 (2014).

93. Atiya, M. *et al.* Interobserver agreement in the classification of stroke in the Women's Health Study. *Stroke* **34,** 565–567 (2003).

94. NINDS Stroke Genetics Network (SiGN)International Stroke Genetics Consortium (ISGC). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *Lancet Neurol* **15,** 174–184 (2016).

95. Hankey, G. J. Secondary stroke prevention. *Lancet Neurol* **13,** 178–194 (2014).

96. Grau, A. J. *et al.* Risk factors, outcome, and treatment in subtypes of ischemic stroke: the German stroke data bank. *Stroke* **32,** 2559–2566 (2001).

97. Arsava, E. M. *et al.* Assessment of the Predictive Validity of Etiologic Stroke Classification. *JAMA Neurol* **74,** 419–426 (2017).

98. Arsava, E. M. *et al.* The Causative Classification of Stroke system: an international reliability and optimization study. *Neurology* **75,** 1277–1284 (2010).

99. Ay, H. *et al.* Pathogenic ischemic stroke phenotypes in the NINDS-stroke genetics network. *Stroke* **45,** 3589–3596 (2014).

100. McArdle, P. F. *et al.* Agreement between TOAST and CCS ischemic stroke classification: the NINDS SiGN study. *Neurology* **83,** 1653–1660 (2014).

101. Diener, H.-C., Bernstein, R. & Hart, R. Secondary Stroke Prevention in Cryptogenic Stroke and Embolic Stroke of Undetermined Source (ESUS). *Curr Neurol Neurosci Rep* **17,** 64 (2017).

102. Sanna, T., Ziegler, P. D. & Crea, F. Detection and management of atrial fibrillation after cryptogenic stroke or embolic stroke of undetermined source. *Clin Cardiol* **41,** 426–432 (2018).

103. Haeusler, K. G., Tütüncü, S. & Schnabel, R. B. Detection of Atrial Fibrillation in Cryptogenic Stroke. *Curr Neurol Neurosci Rep* **18,** 66 (2018).

104. Hart, R. G., Catanese, L., Perera, K. S., Ntaios, G. & Connolly, S. J. Embolic Stroke of Undetermined Source: A Systematic Review and Clinical Update. *Stroke* **48,** 867–872 (2017).

105. Hart, R. G. *et al.* Embolic strokes of undetermined source: the case for a new clinical construct. *Lancet Neurol* **13,** 429–438 (2014).

106. Paciaroni, M. & Kamel, H. Do the Results of RE-SPECT ESUS Call for a Revision of the Embolic Stroke of Undetermined Source Definition? *Stroke* **50,** 1032–1033 (2019).

107. Hart, R. G. *et al.* Rivaroxaban for Stroke Prevention after Embolic Stroke of Undetermined Source. *N. Engl. J. Med.* **378,** 2191–2201 (2018).

108. Saver, J. L. CLINICAL PRACTICE. Cryptogenic Stroke. *N. Engl. J. Med.* **374,** 2065–2074 (2016).

109. Sanna, T. *et al.* Cryptogenic stroke and underlying atrial fibrillation. *N. Engl. J. Med.* **370,** 2478–2486 (2014).

110. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).

111. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46,** 1173–1186 (2014).

112. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).

113. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45,** 197–201 (2013).

114. Shendure, J. Next-generation human genetics. *Genome Biol.* **12,** 408 (2011).

115. Malik, R. *et al.* Low-frequency and common genetic variation in ischemic stroke: The METASTROKE collaboration. *Neurology* **86,** 1217–1226 (2016).

116. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48 (2013).

117. Wierer, M. *et al.* Compartment-resolved Proteomic Analysis of Mouse Aorta during Atherosclerotic Plaque Formation Reveals Osteoclast-specific Protein Expression. *Mol Cell Proteomics* **17,** 321–334 (2018).

118. Huang, H., Lin, S., Garcia, B. A. & Zhao, Y. Quantitative proteomic analysis of histone modifications. *Chem. Rev.* **115,** 2376–2418 (2015).

119. Torrente, M. P. *et al.* Proteomic interrogation of human chromatin. *PLoS ONE* **6,** e24747 (2011).

120. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2010).

121. Kessler, T. *et al.* Functional Characterization of the GUCY1A3 Coronary Artery Disease Risk Locus. *Circulation* **136,** 476–489 (2017).

122. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–680 (2009).

123. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **13,** 840–852 (2012).

124. Farrel, A. & Guo, J.-T. An efficient algorithm for improving structure-based prediction of transcription factor binding sites. *BMC Bioinformatics* **18,** 342 (2017).

125. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147,** 1270–1282 (2011).

126. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulyk, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7,** 555–555 (2011).

127. Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J. Proteome Res.* **14,** 1315–1329 (2015).

128. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).

129. Byrum, S. D., Raman, A., Taverna, S. D. & Tackett, A. J. ChAP-MS: a method for identification of proteins and histone posttranslational modifications at a single genomic locus. *Cell Rep* **2,** 198–205 (2012).

130. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32,** 347–355 (2014).

131. Fujita, T. & Fujii, H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem. Biophys. Res. Commun.* **439,** 132–136 (2013).

132. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24,** R111–9 (2015).

133. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

134. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

135. Scheltema, R. A. *et al.* The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol Cell Proteomics* **13,** 3698–3708 (2014).

136. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15,** 440–448 (2018).

137. Hosp, F. *et al.* A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. *Mol Cell Proteomics* **14,** 2030–2041 (2015).

138. Woolf, S. H. The meaning of translational research and why it matters. *JAMA* **299,** 211–213 (2008).

139. Lu, Y.-F., Goldstein, D. B., Angrist, M. & Cavalleri, G. Personalized medicine and human genetic diversity. *Cold Spring Harb Perspect Med* **4,** a008581 (2014).

140. Markus, H. S. Stroke genetics: prospects for personalized medicine. *BMC Med* **10,** 113 (2012).

141. Schiffmann, R. Fabry disease. *Handb Clin Neurol* **132,** 231–248 (2015).

142. Munot, P., Crow, Y. J. & Ganesan, V. Paediatric stroke: genetic insights into disease mechanisms and treatment targets. *Lancet Neurol* **10,** 264–274 (2011).

143. Wanner, C. *et al.* European expert consensus statement on therapeutic goals in Fabry disease. *Mol. Genet. Metab.* **124,** 189–203 (2018).

144. Belley-Cote, E. P. *et al.* Genotype-guided versus standard vitamin K antagonist dosing algorithms in patients initiating anticoagulation. A systematic review and meta-analysis. *Thromb. Haemost.* **114,** 768–777 (2015).

145. Anderson, C. D., Biffi, A., Greenberg, S. M. & Rosand, J. Personalized approaches to clopidogrel therapy: are we there yet? *Stroke* **41,** 2997–3002 (2010).

146. Pare, G. *et al.* Genetic determinants of dabigatran plasma levels and their relation to bleeding. *Circulation* **127,** 1404–1412 (2013).

147. Epstein, R. S. *et al.* Warfarin genotyping reduces hospitalization rates results from the MM-WES (Medco-Mayo Warfarin Effectiveness study). *J. Am. Coll. Cardiol.* **55,** 2804–2812 (2010).

148. Johnson, J. A. *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clinical pharmacology and therapeutics* **90,** 625–629 (2011).

149. Lv, L. *et al.* Valproic acid improves outcome after rodent spinal cord injury: potential roles of histone deacetylase inhibition. *Brain Res.* **1396,** 60–68 (2011).

150. Bowes, A. J., Khan, M. I., Shi, Y., Robertson, L. & Werstuck, G. H. Valproate attenuates accelerated atherosclerosis in hyperglycemic apoE-deficient mice: evidence in support of a role for endoplasmic reticulum stress and glycogen synthase kinase-3 in lesion development and hepatic steatosis. *Am. J. Pathol.* **174,** 330–342 (2009).

151. Brookes, R. L. *et al.* Sodium Valproate, a Histone Deacetylase Inhibitor, Is Associated With Reduced Stroke Risk After Previous Ischemic Stroke or Transient Ischemic Attack. *Stroke* **49,** 54–61 (2018).

152. Lang, B. *et al.* HDAC9 is implicated in schizophrenia and expressed specifically in post-mitotic neurons but not in adult neural stem cells. *Am J Stem Cells* **1,** 31–41 (2012).

ACKNOWLEDGEMENTS

At the beginning and in the end, plus several times in between, there was a thesis advisory committee: I would like to thank for the stimulating supervision, guidance and inspiration by Prof. Martin Dichgans, Prof. Matthias Mann and Prof. Christian Haass on this fascinating and interdisciplinary topic.

Projects were carried out in truly stellar working environments, both in the Dichgans group at the Institute for Stroke and Dementia Research, and in the Mann department at the Max Planck Institute of Biochemistry. Matthias Prestel and Michael Wierer provided expert knowledge on molecular biology, gene regulation and mass spectrometry; Rainer Malik and Jürgen Cox delivered magical guidance for the dark art of statistics in genetics and proteomics.

There have been many more people involved in living through the daily ups and downs of lab work and LC-MS technicalities. In particular, I am very thankful to Matthias M. himself both for extraordinary scientific and social meetings, to Alison D. and Theresa S. for cordial administration, to Korbi M., Igor P., and Gabi S. for technical support, as well as to Tar V., Dirk W., Chris E., and Fabian H. for sharing our renowned office.

Special thanks go to Dr. G., Dr. W., Dr. D., and Dr. B. for countless spontaneous coffee sessions and Balint groups, which definitely helped with refocusing medium- and long-term plans.

Very special thanks are due to all members of the GSN as well. Without the likes of Benedikt Grothe, Lena Bittl, Maj-Catherine Botheroyd-Hobohm, Stefanie Bosse, Birgit Reinbold and Nadine Hamze, this would not have been possible.

And last but not least, to what matters most, my family. Once again I want to thank my parents for all their unconditional support, decade after decade. Likewise, I am very glad to pass on unconditional support to Oskar and Moni, and to see you grow and/or grow old.

CURRICULUM VITAE

LIST OF PUBLICATIONS

2019

Mönch, S., Boeckh-Behrens, T., Kreiser, K., Blüm, P., Hedderich, D., Maegerlein, C., Berndt, M., **Lehm, M.**, Wunderlich, S., Zimmer, C., Friedrich, B. Thrombocytopenia and declines in platelet counts: predictors of mortality and outcome after mechanical thrombectomy. *J Neurol*, (2019) Mar 27.

Hedderich, D., Reess, T., Thaler, M., Berndt, M., Mönch, S., **Lehm, M.**, Andrisan, T., Maegerlein, C., Meyer, B., Ryang, Y., Zimmer, C., Wostrack, M., Friedrich, B. Hippocampus subfield volumetry after microsurgical or endovascular treatment of intracranial aneurysms - an explorative study. *Eur Radiol Exp*, (2019) Mar 21;3(1):13.

Maegerlein, C., Fischer, J., Mönch, S., Berndt, M., Wunderlich, S., Seifert, CL., **Lehm, M.**, Boeckh-Behrens, T., Zimmer, C., Friedrich, B. Automated Calculation of the Alberta Stroke Program Early CT Score: Feasibility and Reliability. *Radiology*, (2019) Apr;291(1):141-148.

Boeckh-Behrens, T., Pree, D., Lummel, N., Friedrich, B., Maegerlein, C., Kreiser, K., Kirschke, J., Berndt, M., **Lehm, M.**, Wunderlich, S., Mosimann, P., Fischer, U., Zimmer, C., Kaesmacher, J. Vertebral Artery Patency and Thrombectomy in Basilar Artery Occlusions: Is There a Need for Contralateral Flow Arrest? *Stroke*, (2019) Feb;50(2):389-395.

2018

Maegerlein, C., Berndt, M., Mönch, S., Kreiser, K., Boeckh-Behrens, T., **Lehm, M.**, Wunderlich, S., Zimmer, C., Friedrich, B. Further Development of Combined Techniques Using Stent Retrievers, Aspiration Catheters and BGC: The PROTECT^{PLUS} Technique. *Clin Neuroradiol*, (2018) Nov 9.

Berndt, M., Kaesmacher, J., Friedrich, B., Maegerlein, C., Mönch, S., Hedderich, D., **Lehm, M.**, Zimmer, C., Straeter, A., Poppert, H., Wunderlich, S., Schirmer, L., Oberdieck, P., Boeckh-Behrens, T. Thrombus permeability in admission CT imaging as a marker for stroke etiology. *Stroke*, (2018) Nov;49(11):2674-2682.

Mönch, S., **Lehm, M.**, Maegerlein, C., Hedderich, D., Berndt, M., Boeckh-Behrens, T., Wunderlich, S., Kreiser, K., Zimmer, C., Friedrich, B. Worse endovascular mechanical recanalization results for patients with in-hospital onset acute ischemic stroke. *J Neurol*, (2018) Nov;265(11):2525-2530.

Friedrich, B., Kempf, F., Boeckh-Behrens, T., Fischer, J., **Lehm, M.**, Bernd, M., Wunderlich, S., Mönch, S., Zimmer, C., Maegerlein, C. Presence of the Posterior Communicating Artery Contributes to the Clinical Outcome After Endovascular Treatment of Patients with MCA Occlusions. *Cardiovasc Intervent Radiol*, (2018) Dec;41(12):1917-1924.

Friedrich, B., Maegerlein, C., Lobsien, D., Mönch, S., Berndt, M., Hedderich, D., Wunderlich, S., Michalski, D.**, Lehm, M.**, Boeckh-Behrens, T., Zimmer, C., Kreiser, K. Endovascular Stroke Treatment on Single-Plane vs. Bi-Plane Angiography Suites: Technical Considerations and Evaluation of Treatment Success. *Clin Neuroradiol*, (2018) Jan 2.

Maegerlein, C., Mönch, S., Boeckh-Behrens, T., **Lehm, M.**, Hedderich, D., Berndt, M., Wunderlich, S., Zimmer, C., Kaesmacher, J., Friedrich, B. PROTECT: PRoximal balloon Occlusion TogEther with direCt Thrombus aspiration during stent retriever thrombectomy - evaluation of a double embolic protection approach in endovascular stroke treatment. *J Neurointerv Surg*, (2018) Aug;10(8):751-755.

2017

Kaesmacher, J., Huber, T., **Lehm, M.**, Zimmer, C., Bernkopf, K., Wunderlich, S., Boeckh-Behrens, T., Manning, NW., Kleine, JF. Isolated Striatocapsular Infarcts after Endovascular Treatment of Acute Proximal Middle Cerebral Artery Occlusions: Prevalence, Enabling Factors, and Clinical Outcome. *Front Neurol*, (2017) Jun 19;8:272.

2016

Pulit, S., (…), **Lehm, M.**, (…), Worrall, B. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *Lancet Neurol*, (2016) Volume 15, Issue 2, 174-184.

2014

Ay, H., (…), **Lehm, M.**, (…), Meschia, J. Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network. *Stroke*, *(*2014) 45:3589-3596.

AFFIDAVIT

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „**From Phenotype to Function via MassSpec-Based Proteomics: An LC-MS/MS DNA-Protein Pulldown Approach Applied to Functional Stroke Genetics“** selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation “**From Phenotype to Function via MassSpec-Based Proteomics: An LC-MS/MS DNA-Protein Pulldown Approach Applied to Functional Stroke Genetics”** is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, 30.04.2019

_____

Manuel Lehm

DECLARATION OF AUTHOR CONTRIBUTIONS

Manuscript 1

Pathogenic Ischemic Stroke Phenotypes in the NINDS-Stroke Genetics Network. Ay, H., (…), **Lehm, M.**, (…), Meschia, J.

ML collected and analyzed the phenotypic data for the Munich cohort (524 cases) and revised the manuscript.

Manuscript 2

Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Pulit, S., (…), **Lehm, M.**, (…), Worrall, B.

ML conducted DNA sample isolation, collected and analyzed the phenotypic data for the Munich cohort (1131 cases) and revised the manuscript.

Manuscript 3

Functional characterization of an atherosclerosis associated noncoding variant at the *HDAC9* locus. Prestel, M., Prell-Schicker, C., Webb, T., Malik, R., Lindner, B., Ziesch, N., Rex-Haffner, M., Röh, S., Viturawong, T., **Lehm, M.**, Mokry, M., den Ruijter, H., Haitjema, S., Asare, Y., Söllner, F., Najafabadi, M., Civelek, M., Samani, N., Mann, M., Haffner, C., Dichgans, M.

ML designed, performed and analyzed the PWAS experiments, prepared corresponding figure 1B, wrote the PWAS supplemental material and methods and revised the entire manuscript.

Thesis

ML conceptualized, wrote and revised the entire thesis.

I hereby confirm the author contributions.

München, 30.04.2019

_____	_____
Manuel Lehm	Prof. Dr. Martin Dichgans